Thesis submitted in fulfillment of the requirements for the degree

Dr. rer. pol.

to the topic

THE USE OF DATA-DRIVEN TRANSFORMATIONS AND THEIR APPLICABILITY IN SMALL AREA ESTIMATION

in the

Chair of Statistics and Econometrics

School of Business and Economics

Freie Universität Berlin

submitted by

Natalia Rojas-Perilla

Bogotá-Colombia

Berlin, 2018

Natalia Rojas-Perilla, *The Use of Data-driven Transformations and Their Applicability in Small Area Estimation*©,
May, 2018

Supervisors:
Prof. Dr. Timo Schmid (Freie Universität Berlin)
Prof. Dr. Nikos Tzavidis (University of Southampton)

Location:
Berlin

Dissertation day:
June 4th, 2018

Acknowledgements

First and foremost, I would like to extend my heartfelt and deepest gratitude to my supervisor, Prof. Dr. Timo Schmid (Freie Universität Berlin, Germany). His invaluable guidance and profound understanding were key tools for the success of this project. Thanks for showing me an exemplary motivation, giving me lasting patience in tough times and constructive advice on my research and career.

I am also very thankful to Prof. Dr. Nikos Tzavidis (Southampton University, England) for his insightful and fruitful discussions, which greatly improved the realization of this thesis. My special thanks also go to the great joint work of Prof. Dr. Li-Chun Zhang (Southampton University, England), Dr. Angela Luna Hernandez (Southampton University, England), and Dr. Matthias Templ (Zurich University of Applied Sciences, Switzerland).

I gratefully acknowledge the support by the Foundation of German Economy (SDW -Stiftung der Deutschen Wirtschaft) and the UK multi-institutional grant funded by the Economic and Social Research Council (ESRC) and the National Centre for Research Methods (NCRM). I am also very thankful for the transnational visiting grants offered by the Inclusive Growth Research Infrastructure Diffusion (InGRID) program and the German Academic Exchange Service (DAAD - Deutscher Akademischer Austauschdienst). Special thanks go the National Council for the Evaluation for Social Polic (CONEVAL - Consejo Nacional de Evaluación de la Política de Desarrollo Social) for providing the data used in this thesis.

My sincere thanks also goes to Prof. Dr. Ulrich Rendtel (Freie Universität Berlin, Germany) and my colleagues at the Chair of Statistics and the Statistical Consulting Unit *fu:stat*, for providing me with an enjoyable personal and working environment. I am especially grateful to my friend and co-author Ann-Kristin Kreutzmann, who has unconditionally supported me in all dimensions of life in this challenging time with her unique and unselfish manner. My thanks also go to Sören Pannier for his constant input and invaluable and always creative ideas. My thanks also go to Lily Medina and Piedad Castro for our great research collaboration and their constructive comments. My sincere thanks also go to Paul Walter for his precious review and friendly feedback in this project. I am also very thankful to Brian Hose for the careful reading of this thesis.

I want to express my deep gratitude to my family in Colombia for their hearty, permanent, and unconditional support. They have continued to encourage me to pursue my aspirations in life by the life motto “yes you can.”

Last but not least, I thank Santiago Rodríguez for joining me up to here in this adventure, and for his support with this always present, colorful energy.

Publication List

The publications listed below are the result of the research carried out in this thesis titled, "The Use of Data-driven Transformations and Their Applicability in Small Area Estimation."

1. Rojas-Perilla, N., Kreutzmann, A.-K., and Medina, L., (2018). **A Guideline of Transformations in Linear and Linear Mixed Regression Models.** Working paper, to be submitted. The work is an extension of Medina (2017).

2. Medina, L., Rojas-Perilla, N., Kreutzmann, A.-K., and Castro, P., (2017). **The R Package trafo for Transforming Linear Regression Models.** Working paper, to be submitted.

3. Tzavidis, N., Zhang, L.-C., Luna Hernandez, A., Schmid, T., and Rojas-Perilla, N., (2018). **From Start to Finish: A Framework for the Production of Small Area Official Statistics.** *Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4), pp. 927-979.* Accepted and published.

4. Rojas-Perilla, N., Pannier, S., Schmid, T., and Tzavidis, N., (2018). **Data-Driven Transformations in Small Area Estimation.** *Journal of the Royal Statistical Society: Series A (Statistics in Society).* Under revision.

5. Kreutzmann, A.-K., Pannier, S., Rojas-Perilla, N., Schmid, T., Templ, M., and Tzavidis, N., (2018). **The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators.** *Journal of Statistical Software*. Accepted. Preliminary work was done in Kreutzmann (2016).

6. Rojas-Perilla, N., (2018). **Should we Transform Count Data Sets? Generalized Linear Models vs. Count Data Transformations.** Working paper, to be submitted.

Contents

Introduction **7**

I Modeling Guidelines for Practitioners on Transformations **9**

1 A Guideline of Transformations in Linear and Linear Mixed Regression Models **10**
1.1 Introduction 10
1.2 Transformations Step Framework 12
1.2.1 Choose the model and be aware of the corresponding assumptions . . . 12
1.2.2 Choose a suitable transformation that addresses assumption violations . 13
1.3 Further Issues 39
1.4 Conclusions and Future Research Directions 44

2 The R Package trafo for Transforming Linear Regression Models **46**
2.1 Introduction 46
2.2 Transformations 47
2.3 Study Case 50
2.3.1 Finding a suitable transformation 50
2.3.2 Comparing the untransformed model with a transformed model 52
2.3.3 Compare two transformed models 55
2.4 Customized Transformation 56
2.5 Conclusions and Future Research Directions 57

Appendices **58**
.1 Likelihood Derivation of the Transformations 59
.1.1 Log (shift) transformation 59
.1.2 Glog transformation 59
.1.3 Neglog transformation 61
.1.4 Reciprocal transformation 61
.1.5 Box-Cox (shift) transformation 62
.1.6 Log-shift opt transformation 63
.1.7 Bickel-Docksum transformation 64
.1.8 Yeo-Johnson transformation 65
.1.9 Square root-shift opt transformation 67

.1.10 Manly transformation 68
.1.11 Modulus transformation 69
.1.12 Dual power transformation 70
.1.13 Gpower transformation 72

II Transformations in the Context of Small Area Estimation 74

3 From Start to Finish: A Framework for the Production of Small Area Official Statistics 75
3.1 Introduction 75
3.2 Specification 78
3.2.1 Specify user needs: Targets of estimation and target geography 78
3.2.2 Data availability and geographical coverage 79
3.2.3 Illustration using the ENIGH data 80
3.3 Analysis/Adaptation 81
3.3.1 Initial triplet of estimates 81
3.3.2 Use of models for small area estimation 82
3.3.3 Model building, residual diagnostics and transformations in practice . . 84
3.4 Evaluation 92
3.4.1 Uncertainty assessment 93
3.4.2 Method evaluation 95
3.4.3 Illustrating aspects of SAE evaluation using the ENIGH data 97
3.5 An Update on SAE Software 101
3.6 Conclusions and Future Research Directions 103

4 Data-driven Transformations in Small Area Estimation 105
4.1 Introduction 105
4.2 The Empirical Best Prediction (EBP) Method 107
4.3 The Guerrero Case Study: Data Source and Initial Analysis 108
4.4 Use of Transformations 110
4.4.1 EBP under transformations 111
4.4.2 Likelihood-based approach for estimating λ 111
4.4.3 Alternative approaches for estimating λ 113
4.5 MSE Estimation Under Transformations 114
4.6 The Guerrero Case Study: Application of Data-driven Transformations 116
4.6.1 Model checking and residual diagnostics 116
4.6.2 Deprivation and inequality indicators for municipalities in Guerrero . . 118
4.7 Model-Based Simulation Study 120
4.7.1 Behavior of the data-driven transformation parameters 121
4.7.2 Performance of the EBP under data-driven transformations 122
4.7.3 Impact of alternative estimation methods for λ 123
4.8 Conclusions and Future Research Directions 125

Appendices **127**
.1 Derivation of Scaled Transformations . . . 128
.1.1 Log-shift transformation . . . 128
.1.2 Box-Cox transformation . . . 129
.1.3 Dual power transformation . . . 129

5 The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators **131**
5.1 Introduction . . . 131
5.2 Statistical Methodology . . . 133
5.2.1 Direct estimation . . . 134
5.2.2 Model-based estimation . . . 136
5.3 Data Sets . . . 138
5.4 Basic Design and Core Functionality . . . 140
5.4.1 Estimation of domain indicators . . . 141
5.4.2 Summary statistics and model diagnostics . . . 144
5.4.3 Selection and comparison of indicators . . . 148
5.4.4 Mapping of the estimates . . . 149
5.4.5 Exporting the results . . . 151
5.5 Additional Features . . . 152
5.5.1 Incorporating an external indicator . . . 152
5.5.2 Parallelization . . . 153
5.6 Conclusion and Future Developments . . . 155

Appendices **157**
.1 Semi-parametric Wild Bootstrap . . . 158
.2 Reproducibility . . . 158

III Discussion on the Applicability of Transformations **161**

6 Should we Transform Count Data Sets? Generalized Linear Models vs. Count Data Transformations **162**
6.1 Introduction . . . 162
6.2 Count Data Regression Models . . . 163
6.3 Count Data Transformations . . . 165
6.4 Methodological Differences . . . 167
6.5 Simulation Study for the Mean . . . 169
6.6 Conclusions and Further Research Directions . . . 172

Bibliography **174**

Summary **200**
Abstracts in English . . . 201
Kurzfassungen in deutscher Sprache . . . 203

Introduction

One of the goals of data analysts is to establish relationships between variables using regression models. Standard statistical techniques for linear and linear mixed regression models are commonly associated with interpretation, estimation, and inference. These techniques rely on basic assumptions underlying the working model, listed below:

1. Normality: Transforming data to create symmetry in order to correctly use interpretation and inferential techniques

2. Homoscedasticity: Creating equality of spread as a means to gain efficiency in estimation processes and to properly use inference processes

3. Linearity: Linearizing relationships in an effort to avoid misleading conclusions for estimation and inference techniques.

Different options are available to the data analyst when the model assumptions are not met in practice. Researchers could formulate the regression model under alternative and more flexible parametric assumptions. They could also use a regression model that minimizes the use of parametric assumptions or under robust estimation. Another option would be to parsimoniously redesign the model by finding an appropriate transformation such that the model assumptions hold. A standard practice in applied work is to transform the target variable by computing its logarithm. However, this type of transformation does not adjust to the underlying data. Therefore, some research effort has been shifted towards alternative data-driven transformations, such as the Box-Cox, which includes a transformation parameter that adjusts to the data.

The literature of transformations in theoretical statistics and practical case studies in different research fields is rich and most relevant results were published during the early 1980s. More sophisticated and complex techniques and tools are available nowadays to the applied statistician as alternatives to using transformations. However, simplification is still a gold nugget in statistical practice, which is often the case when applying suitable transformations within the working model. In general, researchers have been using data transformations as a go-to tool to assist scientific work under the classical and linear mixed regression models instead of developing new theories, applying complex methods or extending software functions. However, transformations are often automatically and routinely applied without considering different aspects on their utility.

In Part I of this work, some modeling guidelines for practitioners in transformations are each presented. An extensive guideline and an overview of different transformations and esti-

mation methods of transformation parameters in the context of linear and linear mixed regression models are presented in Chapter 1. Furthermore, in order to provide an extensive collection of transformations usable in linear regression models and a wide range of estimation methods for the transformation parameter, the package **trafo** is presented in Chapter 2. This package complements and enlarges the methods that exist in R so far, and offers a simple, user-friendly framework for selecting a suitable transformation depending on the research purpose.

In the literature, little attention has been paid to the study of techniques of the linear mixed regression model when working with transformations. This becomes a challenge for users of small area estimation (SAE) methods, since most commonly used SAE methods are based on the linear mixed regression model which often relies on Gaussian assumptions. In particular, the empirical best predictor is widely used in practice to produce reliable estimates of general indicators for areas with small sample sizes. The issue of data transformations is addressed in the current SAE literature in a fairly ad-hoc manner. Contrary to standard practice in applied work, recent empirical work indicates that using transformations in SAE is not as simple as transforming the target variable by computing its logarithm.

In Part II of the present work, transformations in the context of SAE are applied and further developed. Chapter 3 proposes a protocol for the production of small area official statistics that is based on three stages, namely (i) Specification, (ii) Analysis/Adaptation and (iii) Evaluation. In this chapter, the use of some adaptations of the working model by using transformations is showed as a part of the (ii) stage. In Chapter 4 we extended the use of data-driven transformations under linear mixed model-based SAE methods; In particular, the estimation method of the transformation parameter under maximum likelihood theory. First, we analyze how the performance of SAE methods are affected by departures from normality and how such transformations can assist with improving the validity of the model assumptions and the precision of small area prediction. In particular, attention has been paid to the estimation of poverty and inequality indicators, due to its important socio-economical relevance and political impact. Second, we adapt the mean squared error estimator to account for the additional uncertainty due to the estimation of transformation parameters. Finally, as in Chapter 3, the methods are illustrated by using real survey and census data from Mexico. In order to improve some features of existing software packages suitable for the estimation of indicators for small areas, the package **emdi** is developed in Chapter 5. This package offers a methodological and computational framework for the estimation of regionally disaggregated indicators using SAE methods as well as providing tools for assessing, processing, and presenting the results.

Finally, in Part III, a discussion of the applicability of transformations is made in the context of generalized linear models (GLMs). In Chapter 6, a comparison is made in terms of precision measurements between using count data transformations within the classical regression model and applying GLMs, in particular for the Poisson case. Therefore, some methodological differences are presented and a simulation study is carried out. The learning from this analysis focuses on the relevance of knowing the research purpose and the data scenario in order to choose which methodology should be preferable for any given situation.

Part I

Modeling Guidelines for Practitioners on Transformations

Chapter 1

A Guideline of Transformations in Linear and Linear Mixed Regression Models

1.1 Introduction

The linear regression model is perhaps the simplest and most common model used in statistical analysis. The linear mixed regression model is similarly useful for cluster or longitudinal data types. The estimation and inference methods employed with these kinds of models typically rely on a set of assumptions; some of them inherent to the functional form of the model (e.g., linearity), and others related to the nature of the error terms, the response variable, and the covariates (e.g., homoscedasticity). However, empirical data does not always satisfy these assumptions and, therefore, one must decide how to carry on with the analysis. According to Sakia (1992), there are many available options for such cases, which may be summarized as: (i) ignore the violation(s) and proceed; (ii) use a method that allows for the corresponding violation(s); (iii) redesign the model e.g., by properly transforming the data, and (iv) use a distribution-free method. Instead of developing new theories, applying complex methods or extending software functions, using transformations (option (iii)) is a parsimonious way to deal with model assumption violations under both linear and linear mixed regression models. The set of model assumptions that are commonly satisfied by properly transforming the data are normality, homoscedasticity, and linearity. Furthermore, using transformations allows practitioners to apply the most powerful methods available for parametric statistics and to make analysis simpler than otherwise possible. For instance, transformations can allow us to easily get rid of high order terms and work only with first-order linear relationships, which is often preferred in several branches of knowledge (Draper and Hunter, 1969). But how and where are transformations usually used in practice?

The use of transformations has received much attention in the last century in both theoretical knowledge and practical applications (e.g., Edgeworth (1900); Bartlett (1947); Box and Cox (1964)), and is still of great concern in many investigations (e.g., Gurka et al. (2006); Lakhana (2014)). In the literature of transformations, we find linear, monotonic, accelerating,

and decelerating, power and two-bend transformations, among others. The most discussed type of transformations is the power family, which includes as a particular case both the Box-Cox transformation and the logarithmic function. General overviews about applying transformations under the linear regression model are published by Kruskal (1968); Hoyle (1973); Tukey (1977); Sakia (1992) and Fink (2009). Zarembka (1974b) provides an overview of variable transformations in econometrics. He paid special attention to the problem of heteroscedasticity and illustrated the transformations theory employing elasticity and demand studies. Volatility studies, functional form of demand equations, and economic depreciations have been analyzed mainly using the logarithmic transformation, and also the Box-Cox method (Gemmill et al., 1980; Hulten and Wykoff, 1981; Boylan et al., 1982; Goncalves and Meddahi, 2011). Hossain (2011) gives an analytical review in economic sciences about the importance of the Box-Cox transformation regarding estimation, model selection, and testing. In education, social, biological, and ecological studies, the logarithm is certainly the most relevant transformation and the Box-Cox is also becoming a standard method for variable transformations in these fields (Buchinsky, 1995). In the medical sciences, special attention is paid to dealing with non-normal data (Bland and Altman, 1996). Snedecor and Cochran (1989); Sokal (1995); Keene (1995); Zar (1999) and Armitage et al. (2008) give an introductory literature for medical researches about using transformations, focusing on the logarithmic, Box-Cox, square root, and arcsine transformations. Since biological and medical studies often use longitudinal data, linear mixed regression models for repeated measures analysis are commonly applied (Miller, 2010). In order to deal with model assumption violations under these models, the logarithmic and Box-Cox transformations are preferred (Gurka et al., 2006; Maruo et al., 2017). Furthermore, renowned applications of the Box-Cox transformation in this context are described in Solomon (1985); Sakia (1988); Gurka et al. (2006); Piepho and McCulloch (2004) and Lo and Andrews (2015).

As we can see, the literature of transformations in theoretical statistics and practical case studies is very rich. However, some important considerations for using them in linear and linear mixed regression models are still broadly discussed: for example, at which stage of the analysis a transformation should be applied, which transformation is suitable for a specific problem and how the results should be interpreted. Practitioners often automatically and routinely apply transformations without considering the above mentioned questions. For this purpose, the present work proposes a framework that seeks to help the researcher to decide if and how a transformation should be applied in practice. It combines a set of pertinent steps, tables, and flowcharts that guide the practitioner through the analysis of transformations in a friendly and practical manner. This guideline is structured as follows:

- Defining relevant assumptions depending on the research goals
- Choosing a suitable transformation and estimation method according to model assumption violations
- Providing a proper inference analysis and interpreting model results more carefully

Furthermore, the paper points out briefly a selection of special issues that need to be considered when using transformations. To the best of our knowledge, none of the existing reviews for transformations provides such a comprehensive overview of transformations in the context

of linear and linear mixed regression models, as well as developing a practical guideline for researchers.

The remainder of this paper is structured as follows. Section 1.2 guides the reader through the steps of the framework. Each transformation and estimation method is introduced to its corresponding model assumption. Section 1.3 discusses further issues that can arise in modelling and how these interact with the transformations. We conclude the paper in Section 1.4.

1.2 Transformations step framework

"Although we often hear that data speak for themselves,
their voices can be soft and sly."

—Frederick Mosteller

1.2.1 Choose the model and be aware of the corresponding assumptions

Linear regression models are one of the most widely used statistical methods in most branches of knowledge, in particular, the social and natural sciences. It can be expressed in a general form:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + e_i, \quad e_i \overset{iid}{\sim} N(0, \sigma_e^2), \tag{1.1}$$

where y_i is the target variable defined for the ith individual, with $i = 1, \ldots, n$; $\boldsymbol{x}_i^\top$ is a vector containing deterministic auxiliary information with dimension $1 \times (p+1)$ and $\boldsymbol{X}$ would be the corresponding $n \times (p+1)$ matrix where p is equal to the number of predictors; $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients defined as $\boldsymbol{\beta}^\top = (\beta_0, \ldots, \beta_p)$ and e_i is the unit-level error term.

In social, behavioral, educational, and medical sciences, data is commonly hierarchically collected, for instance, as a clustered or longitudinal design (Raudenbush and Bryk, 2002). To appropriately take this type of data structure into account, the so-called linear mixed regression models are typically used. These models, handled as a special extension of the linear regression model, contain additional random-effects depending on the case study and can be written as follows:

$$\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\beta} + \boldsymbol{Z}_j \boldsymbol{u}_j + \boldsymbol{e}_j, \tag{1.2}$$

where $\boldsymbol{y}_j$ is a $n_j \times 1$ vector of the dependent variable, n_j is the sample size in each cluster j with $j = 1, ..., m$ cluster, $\boldsymbol{X}_j$ is a $n_j \times (p+1)$ matrix, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients, $\boldsymbol{Z}_j$ is the $n_j \times (q+1)$ matrix with $(q+1)$ random effects, $\boldsymbol{u}_j$ is a $(q+1) \times 1$ vector of random effects and $\boldsymbol{e}_j$ is the vector of residuals of size $n_j \times 1$. The distribution of the random effects is given by:

$$\boldsymbol{u}_j \sim N(\boldsymbol{0}, \boldsymbol{G}), \qquad \text{where} \qquad \boldsymbol{G} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \ldots & \sigma_{0q} \\ \sigma_{10} & \sigma_1^2 & \ldots & \sigma_{1q} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{q0} & \sigma_{q1} & \ldots & \sigma_q^2 \end{bmatrix},$$

and the residuals are distributed with $\boldsymbol{e}_j \sim N(\boldsymbol{0}, \boldsymbol{R})$ with $\boldsymbol{R} = \boldsymbol{I}_{n_j}\sigma_e^2$ where $\boldsymbol{I}_{n_j}$ is the $n_j \times n_j$ identity matrix and σ_e^2 is the residual variance. The random effects $\boldsymbol{v}_j$ and the residuals $\boldsymbol{e}_j$ are assumed to be independent.

Typically the set of assumptions upon which these models rely can be summarized as:

(i) The error terms are normally distributed.

(ii) The error terms have (conditional) homoscedastic variances.

(iii) The response and explanatory variables have a linear and an additive relationship.

(iv) The error terms are (conditionally) independent.

(v) The error terms have (conditional) mean equal to zero.

Two potential problems that will also be taken into account are multicollinearity and outliers. However, these are not listed as assumptions for these regression models, since they are not seen as theoretical constraints (Barry, 1993). As we shall discuss in more detail below, if any of these assumptions is violated, estimations, predictions, and scientific insights produced by the linear and linear mixed regression models may be inefficient or, in some cases, severely biased and misleading (Nau, 2017). This work mainly focuses on the relevance of assumptions (i) - (iii). For readers interested in the assumptions (iv) and (v), discussions, diagnostics and potential solutions are presented in econometric books such as Johnston and DiNardo (1972) and Spanos (1986).

1.2.2 Choose a suitable transformation that addresses assumption violations

The usage of data transformations is directed towards a twofold aim: to create a useful metric or to improve model regression assumptions. For the first aim, linear transformations help in the following ways: information can be easier to understand (e.g. percentage); standardization can be applied in order to change the scale (e.g. covariances into correlation); and a shift can be added to the set of points to make variables positive. Furthermore, these can be useful when transforming qualitative ordinal data into a more convenient and continuous scale, for which normal scores are recommended (for further details see Hoyle (1973) and Fink (2009)). However, such linear transformations do no attempt to correct violations of the regression model assumptions presented in Section 1.2.1. A linear transformation will change only the intercept of the regression equation. For instance, using this type of transformation does not help to linearize non-linear relations (Brown, 2015). In this work, we focus on transformations that attempt to correct violations of the assumptions of the linear and linear mixed regression model. These non-linear transformations are monotonic and shrink or stretch a topological space in an inhomogeneous way. That is, the order of the points lying on this space remain unchanged, but the relative distance between them will be altered (Cohen et al., 2014). For defining such a transformation, the following notation is used consistently through the present work. We denote y as the response variable with expected value denoted by $E(y) = \mu_y$ and variance by $V(y) = \sigma_y^2$. For a single untransformed observation we use y_i, y_{ij} where an additional symbol * denotes that the observation is transformed. The untransformed vector

of the response variable is defined as $\boldsymbol{y}$. Furthermore, $\boldsymbol{y}^*(\Theta)$ represents the vector of the transformed observations of the response variable and Θ represents the set of parameters upon which the transformation depends. The transformation parameter is generally denoted by λ, but it depends on the functional form of the transformation. Some transformations also include additional parameters. The relationship between original and transformed data is denoted by $T(y) = y^*$.

For this section, the following structure is used: we describe the model assumption and its relevance, we introduce assessment tools to check its fulfilment, we mention alternative methods to transformations, and we discuss the range of possibilities using transformations with corresponding estimation methods.

1.2.2.1 Transformations to achieve normality

Why is the normality assumption important?

The fulfillment of the normality assumption is usually twofold: it builds confidence intervals and for computing statistical tests and appropriately uses the percentage points of customary tables of χ^2, t, F distributions. When this assumption is not fulfilled, practical problems can arise; as for estimation, the ordinary least square method does not provide best estimators in terms of efficiency, in case the true distribution of the error term is skewed or has heavy tails. When the interest lies in inference hypothesis testing, such as a t-test for significance of the coefficients, the results of this test seem to be fairly robust for large enough samples. However, its power may be somewhat affected when, for instance, the true distribution has heavy tails, as σ_e^2 is very sensitive to values at the tails of the distribution (Wilcox, 2005). The most common departures from normality are skewed, heavy-tailed, and light-tailed distributions. Additionally, human errors can contribute to the presence of non-random aspects which lessen the strength of the assumption that the error term is normally distributed (Zeckhauser and Thompson, 1970). Some papers related to the consequences when Gaussian assumptions are not satisfied are published by Fisher (1922b); Pearson (1931); Bartlett (1935); Hey (1938); Finney (1941), and Cochran (1947).

How can we check the normality assumption?

Due to the importance of the normality assumption, many methods have been developed to check its validity: visual methods such as the normal probability plot of the residuals (Chambers et al., 1983), histograms, and probability plots. The normal probability plot, also known as normal scores plot, quantile-quantile (Q-Q) plot, quantile comparison plot or rankit plot can be useful for comparing two probability distributions in terms of the location, scale, and skewness parameters (Weisberg, 1980; Bock, 1985; Fox, 1997; Hutcheson and Sofroniou, 1999; Johnson, 2009). The histogram is a standard visualization of the empirical distribution form. The probability plot, also known as probability-probability (P-P) plot or percent-percent plot, is suitable for analyzing the skewness of a distribution, by plotting two cumulative distribution functions. Numerical analysis of the distribution moments, such as skewness and kurtosis, is a common rule-of-thumb for checking the normality assumption. The skewness and kurtosis for a nor-

mal distribution are equal to zero and three, respectively. Therefore, a comparison with this distribution is often made in practice. Additionally, normality tests such as the Kolmogorov-Smirnov test (Smirnov, 1948), Anderson-Darling test (Anderson and Darling, 1954) and the Shapiro-Wilk test (Shapiro and Wilk, 1965) are also widely used.

What are the alternative methods to overcome non-normality?

If any of the aforementioned techniques suggests that the data is not normally distributed, we could move to non-normal methods or redesign the model. In this case, there are some typically recommended solutions. The first method and perhaps the most common one is to allow a more flexible model where the conditions imposed over the error term and independent variables can be relaxed. This method is known as the generalized linear or generalized linear mixed model (Nelder and Wedderburn, 1972). A second solution is to work with more robust tests such as the Kruskal-Wallis (Kruskal and Wallis, 1952) or the Levene's test (Levene et al., 1960). Robust and more efficient estimators have been studied when the error term is not normally distributed (see, for instance, Huber (1964)). Among these approaches, we find non-parametric maximum likelihood theory (Aitkin, 1999; Agresti et al., 2004; Litière et al., 2008), more flexible parametric distributions (Peng Zhang and Greene, 2008), marginalized mixed effects models (Heagerty and Zeger, 2000), and h-likelihood approaches that can be adapted to fit different distributions (Lee et al., 2004). Also possible are methods based on mixtures of normal distributions (Lesaffre and Molenberghs, 1991) and "smooth" non-parametric fits (Zhang and Davidian, 2001).

How can transformations help to improve normality?

The use of transformations is considered as a parsimonious alternative to complex methodologies when dealing with the departure from normality, a feature seldom observed in raw data. A significant part of the effort put into transformations has been focused on achieving approximate normally distributed errors. To ensure normality, it is common to use a proper one-to-one transformation on the target variable (Thoni, 1969; Hoyle, 1973). A standard practice in applied work is transforming the target variable by computing its logarithm. That means using a transformation of the form $\log(y)$. Due to its effectiveness in turning highly right-skewed or log-normal distributions into more symmetrical ones, it is commonly used in practice for this purpose. Furthermore, the logarithmic transformation is used in parallel for achieving normality, homoscedasticity, and linearity (Bartlett and Kendall, 1946; Bartlett, 1947; Anscombe, 1948; Kleczkowski, 1949; Moore, 1958). However, the ease of its use and its popularity often induce an imprudent application (Changyong et al., 2014). One drawback of the logarithmic transformation is the lack of ability to deal with negative values. Thus, some adjustments based on the logarithm have been proposed. A simple shifted version includes a fixed term s such that $y + s > 0$. The logarithmic transformation is often recommended when dealing with substantially positive skewness. For a left-skewed distribution, the log neg transformation is suggested. It includes a fixed parameter p for which every observation of the target variable is subtracted so that the smallest score is 1 (Tabachnick and Fidell, 2007). Furthermore, the generalized logarithm, also known as the glog transformation allows for negative values, but it is recom-

mended for low values rather than high ones (Durbin et al., 2002; Huber et al., 2003). Even though it is suitable for correcting non-normality, it is more widely used as a variance stabilizing transformation. Another transformation used particularly for dealing with non-negative variables such as the non-central chi-square is suggested by Moschopoulos (1983). He bases his work on the theory developed by Jensen and Solomon (1972), including the moments of the distribution as transformation parameters.

Square roots and inverse transformations are commonly used for dealing with right-skewed distributions (Bartlett, 1937). The square root is also used for dealing with data having zero inflated problems or containing extremely small values. The cube-root transformation, also known as the Wilson-Hilferty (Wilson and Hilferty, 1931) transformation, is particularly suitable for symmetrizing gamma-distributed data forms. The exponential, square, and cube root transformations are commonly used for negative skewed data. A quasi generalization of this problem is made in practice in the transformation exponent: right-skewed distributions tend to be more symmetrical by applying a transformation with an exponent smaller than one, and left-skewed distributions, with an exponent greater than one (Hoaglin et al., 2000). When comparing the square-root transformation with the logarithm, Garson (2012) states that the latter is more useful in case symmetry in the central distribution is needed. Meanwhile the square root is suggested in case symmetry in the tails is more important. Finally, in the case of negative skewness, the reciprocal transformations may be useful as an appropriate variance stabilizing transformation (Hoyle, 1973) for certain distributions.

The transformations mentioned so far have in common that they do not adjust to the underlying data. To find a data-driven transformation, an adjustment is done by including a data-driven transformation parameter, denoted by λ. This parameter should be estimated and this estimate changes according to the data, the assumption violations or to a specific researcher criteria. For instance, an advanced log-shift opt transformation used in practice (e.g. Feng et al. (2016)) includes an optimal transformation parameter as follows $y^*(\lambda) = \log(y + \lambda)$. Tukey (1957) proposed a family of power transformations based on monotonic functions. The general form of this family is defined as: y^λ if $\lambda \neq 0$ and $\log(y)$ if $\lambda = 0$. The power transformations are also commonly denoted as single- or one-bend transformations (Box and Cox, 1964; Montgomery, 2008; Fink, 2009; Cohen et al., 2014). To avoid the discontinuity at $\lambda = 0$, Box and Cox (1964) modified this family. The straightforward manner in which the interpretation of this parameter is made makes the Box-Cox method one of the most widely used transformations. For instance, when $\lambda = -1$, it means the reciprocal transformation is needed, $\lambda = 0$ means the logarithmic transformation is recommended, $\lambda = 1/2$ implies the use of the square root and $\lambda = 1$ suggests that no transformation is necessary. The Box-Cox transformation is the simplest single-bend transformation (Fink, 2009) and is more appropriate when dealing with skewed distributions than symmetric but non-normal distributions. It has been extensively implemented in different branches of knowledge. For detailed information about renowned applications, see Draper and Cox (1969), Mills (1978), Poirier (1978), Machado and Mata (2000), Chen (2002), Chen and Deo (2004) and Yang and Tsui (2004).

Since the Box-Cox transformation is not defined for negative values, the data must be shifted to the positive side by incorporating a shift parameter. This method is known as the

shifted power transformation. It overcomes the difficulties encountered in the Box-Cox transformation due to the restriction $y > 0$. This is done by incorporating a constant, denoted by s, for accommodating negative values of the target variable. The parameter s is chosen such that $y + s > 0$. Moore (1957) studies the benefits of adding this shift parameter in the power family of transformations. However, Hill (1963); Atkinson (1987) and Yeo and Johnson (2000) state that shifting the data is not always an optimal way to deal with negative values. Different modifications have been proposed in the literature to address this issue. The first proposal to avoid this difficulty was made by Manly (1976), who proposes the Manly transformation, an exponential power transformation family. This transformation family is considered to nearly normalize unimodal skewed distributions, but it is not suitable for bimodal or U-shape distributions. In case the data also presents a symmetric but non-normal error distribution, the modulus power transformation proposed by John and Draper (1980) should be used. It can manage negative values and is claimed to be effective for somewhat symmetrical or bimodal distributions. In the same way, the neglog transformation, proposed by Whittaker et al. (2005) is developed especially to deal with negative values. In order to avoid the non-negativity restriction of the Box-Cox transformation, Bickel and Doksum (1981) introduced the Bickel-Docksum power transformation which is defined on the whole real line. This transformation is especially useful for handling kurtosis rather than skewness, in particular for leptokurtic and platykurtic distributions. However, as Yeo and Johnson (2000) point out, one should avoid the use of this transformation when dealing with skewed data that takes negative and positive values. As another alternative to the Box-Cox transformation, Kelmansky et al. (2013); Kelmansky and Ricci (2017) recently proposed an extension of the glog transformation, also known as gpower transformation. It allows for negative values, heavier tails and peaked sample modes (Tsai et al., 2017). The work of MacKinnon and Magee (1990) proposes a scale-invariant family of transformations, which deals with variables with zero or negative values.

Zwet (1964) emphasizes that for reaching near symmetry when the response variable has positive and negative values, the transformation should be concave. One could say that a transformation has the quality of reducing left-skewness if such a transformation is non-decreasing convex or upward bending, and a transformation is needed to symmetrize right-skewness if such a transformation is non-decreasing concave or downward bending. Under this motto, different transformations have been proposed for kurtosis adjustments in order to deal with non-normality. This is also achieved by the convex-to-concave Yeo-Johnson transformation (Yeo and Johnson, 2000) for different ranges of λ. The transformation is convex in y for $\lambda > 1$, and concave for $\lambda < 1$. Nevertheless, this transformation is not suitable when data has a platykurtic, leptokurtic or bimodal form. Analogously, the power transformations family is convex in case $\lambda > 1$ and concave when $\lambda < 1$. Following Tsai et al. (2017), transformations that are suitable for data with a peaked mode are the signed power (Bickel and Doksum, 1981), the modulus (John and Draper, 1980), the sinh-arcsinh (Jones and Pewsey, 2009), the gpower (Kelmansky et al., 2013) and the hyperbolic sine (Burbidge et al., 1988). The signed transformation is convex-concave as the outcome variable changes the sign, which is an effect that is difficult to predict. Therefore, it is recommended to use it for a kind of symmetric distribution in order to deal with the kurtosis, rather than skewness (Zwet, 1964; Oja, 1981).

Another difficulty of the Box-Cox transformation is the truncation on the transformation parameter determined by λ. If λ is positive, y^* has an upper-bound at $\frac{-1}{\lambda}$ and if λ negative y^* has also a lower-bound at $\frac{-1}{\lambda}$. Unless $\lambda = 0$ this transformation has a compatibleness problem with the exact normality distribution. In order to deal with this problem, Yang (2006) recently proposed the dual power transformation. It is defined only for strictly positive values. In the case that the outcome variable is bounded above as well as below, the previous transformations are not suitable. Therefore, the appropriate transformation based on an interval $[0, b]$ is the folded-power transformation (Mosteller and Tukey, 1977; Atkinson, 1982). However, if the outcome scores are close to 0 or b the behavior would be like the Box-Cox transformation (Cook and Weisberg, 1982). The shifted version of the dual transformation can also be applied in practice (see, e.g., Rojas-Perilla et al. (2017)).

Besides the power transformations presented above, the multi-parameter transformation families have been suggested in order to estimate different transformation parameters, accounting for scale, location, and shape (skewness and tailweight). For this purpose, Johnson (1949) proposes three normalizing transformations, which include shape, scale, and location parameters, where a system of curves represents the empirical distributions (Edgeworth, 1900). Furthermore, for continuous empirical forms, this method has the particular advantage that many distributions can be fitted into the system, which delivers a high flexibility that can be advantageous for dealing with complicated data sets (George, 2007). As a special case of the Johnson transformation, the one-parametric inverse hyperbolic sine is suitable for dealing with negative and positive values (Burbidge et al., 1988). This transformation contains the Pearson system of frequency curves (Pearson, 1894). These curves properly represent data which exhibit departures from normality or with considerable skewness, that means non-normal forms. In contrast, the sinh-arcsinh transformations are applied for heavy-tailed and light-tailed distributions (Jones and Pewsey, 2009).

As mentioned before, in many branches of knowledge, cross-sectional data are widely used. However, little attention has been paid to the study of techniques in the literature of linear mixed regression models, which assess or improve the validity of the multiple distributional assumptions by departures from normality of the error terms expressed in Equation 1.2. In order to improve the assumptions of the model by parametrically transforming the outcome variable in linear mixed regression models, single-bend transformations, such as the logarithmic and square root transformations, have been applied in particular case studies (McCulloch and Neuhaus, 2001; Piepho and McCulloch, 2004; West et al., 2007; Lo and Andrews, 2015). Solomon (1985); Sakia (1988) and Lipsitz et al. (2000) have furthermore studied the application of the Box-Cox transformation to cover all linear mixed regression models and some longitudinal datasets, while the work of Gurka et al. (2006) formally extended the use of the Box-Cox method for these kinds of models. Finally, as Box and Cox (1964) state, several transformations are suitable to improve not only one model assumption, but many. This is also expressed in Table 1.1, which contains transformations that help to achieve normality. Additionally, it is indicated which further assumption can be often improved by these transformations. We exhaustively examine the literature on transformations and present it in Table 1.1 and subsequent tables as a condensed version of the research work.

Table 1.1: Transformations for achieving normality

Transformation	Source	Formula	Support	N	H	L
Log	Tukey (1977)	$\log(y)$	$y > 0$	✗	✗	✗
Log (shift)	Box and Cox (1964)	$\log(y + s)$	$y \in \mathbb{R}$	✗	✗	✗
Log neg	Tabachnick and Fidell (2007)	$\log(p - y)$	$y \in \mathbb{R}$	✗	✗	✗
Glog	Durbin et al. (2002)	$\log(y + \sqrt{y^2 + 1})$	$y \in \mathbb{R}$	✗	✗	✗
Moschopoulos	Moschopoulos (1983)	$\left(\frac{y+a}{\mu}\right)^b$	$y > 0$	✗		
Square Root	Bartlett (1937)	$\sqrt{y}$	$y > 0$	✗	✗	
Square root neg	Tabachnick and Fidell (2007)	$\sqrt{p - y}$	$y \in \mathbb{R}$	✗	✗	
Wilson-Hilferty	Wilson and Hilferty (1931)	$y^{1/3}$	$y \in \mathbb{R}$	✗	✗	
Reciprocal	Tukey (1977)	$\frac{1}{y}$	$y \neq 0$	✗	✗	
Log-shift opt	Feng et al. (2016)	$\log(y + \lambda)$	$y \in \mathbb{R}$	✗	✗	✗
Folded	Mosteller and Tukey (1977)	$y^\lambda - (1 - y)^\lambda$ if $\lambda \neq 0$.	$y > 0$	✗	✗	
Box-Cox	Box and Cox (1964)	$\begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗	✗	✗
Box-Cox (shift)	Box and Cox (1964)	$\begin{cases} \frac{(y+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + s) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	✗
Manly	Manly (1976)	$\begin{cases} \frac{e^{\lambda y} - 1}{\lambda} & \text{if } \lambda \neq 0; \\ y & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Modulus	John and Draper (1980)	$\begin{cases} \text{Sign}(y) \frac{(\lvert y \rvert + 1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \text{Sign}(y) \log(\lvert y \rvert + 1) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Neglog	Whittaker et al. (2005)	$\text{Sign}(y) \log(\lvert y \rvert + 1)$	$y \in \mathbb{R}$	✗	✗	
Bickel-Docksum	Bickel and Doksum (1981)	$\frac{\lvert y \rvert^{\lambda \text{Sign}(y)} - 1}{\lambda}$ for $\lambda > 0$	$y \in \mathbb{R}$	✗	✗	
Gpower	Kelmansky et al. (2013)	$\begin{cases} \frac{(y + \sqrt{y^2+1})^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + \sqrt{y^2 + 1}) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Mackinnon-Magee	MacKinnon and Magee (1990)	$\frac{h(\lambda y)}{\lambda}$	$y \in \mathbb{R}$	✗		✗
Yeo-Johnson	Yeo and Johnson (2000)	$\begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0; \\ \log(y + 1) & \text{if } \lambda = 0, y \geq 0; \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y < 0; \\ -\log(1 - y) & \text{if } \lambda = 2, y < 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Dual	Yang (2006)	$\begin{cases} \frac{(y^\lambda - y^{-\lambda})}{2\lambda} & \text{if } \lambda > 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗		
Tukey	Tukey (1957)	$\begin{cases} y^\lambda & \text{if } \lambda \neq 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗	✗	
Johnson	Johnson (1949)	$\kappa + \nu h\left(\frac{y - \xi}{\eta}\right)$	$y \in \mathbb{R}$	✗	✗	
Sinh-arcsinh	Burbidge et al. (1988)	$\sinh[\theta \sinh^{-1}(y - \gamma_1)]$	$y \in \mathbb{R}$	✗		

Note: Normality, homoscedasticity, and linearity are denoted as N,H,L, respectively. Additional to the notation that is used throughout the paper, for some transformations further parameters need to be defined. The parameters s and p are fixed parameter and chosen such that the smallest score is equal to 1. In the Moschopoulos transformation μ is the first moment of the distribution, and a and b are determined from the first three moments of the distribution. The known fixed values that work for this transformation are $b = 1/3$ (Wilson and Hilferty, 1931) and $b = 1/2$ (Fisher, 1922a). In the transformation by MacKinnon and Magee (1990), $h(\cdot)$ is a monotonically increasing function that satisfies the following properties: $h(0) = 0$, $h`(0) = 1$ and $h``(0) \neq 0$. One common function is defined as $h(\cdot) = sinh^{-1}(y)$. According to Johnson (1949), η and ν are the scale parameters and κ and ξ the location parameters. $h(\cdot)$ is a monotonic function of y. In the sin-arcsinh transformation, $\gamma_1 \in \mathbb{R}$ represents the skewness parameter and $\theta > 0$ controls the tail weight.

How can we estimate the transformation parameter to normality?

In addition to the selection of a suitable transformation, different methodologies for the estimation parameter have been introduced. The estimation method partly depends on which model assumption we want to enforce. Please notice that some of the estimation methods are, so far, only developed for the Box-Cox transformation. In general, the approaches for estimating the optimal transformation parameter to normality are classified in maximum likelihood-based approaches (A), analytical considerations (B), robust adaptations (C), and Bayesian approaches (D). The methods are described below and the mathematical formulation is presented in detail for these ones, which are more commonly applied.

A: Maximum likelihood-based approaches

A.1: Maximum likelihood (ML) approach

The ML-based method is also known as the profile log-likelihood approach. It is the most commonly cited approach under the linear regression model and is described in detail in Box and Cox (1964). It has been studied by Draper and Cox (1969); Andrews (1971); Atkinson (1973); Carroll (1980) and Bickel and Doksum (1981). The goal is to find the transformation parameter λ for which the expected value $E[\boldsymbol{y}^*(\lambda)]$ is equal to $\boldsymbol{X\beta}$ meeting the model assumptions listed in the previous chapter. If the normality assumption $y_i^*(\lambda) \sim N(\boldsymbol{x}_i\boldsymbol{\beta}, \sigma_e^2)$ is fulfilled, the probability density function for $y_i^*(\lambda)$ is written as

$$f(y_i^*(\lambda)) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left\{ - \frac{(y_i^*(\lambda) - \boldsymbol{x}_i\boldsymbol{\beta})^2}{2\sigma_e^2} \right\}. \tag{1.3}$$

The probability density function for the untransformed observations, and thus the likelihood for the whole (transformed) model in relation to those observations, is computed as the likelihood of Equation 1.3 multiplied by the Jacobian of the transformation, explicitly:

$$L(\boldsymbol{y}, \lambda \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}} \exp\left\{ - \frac{(\boldsymbol{y}^*(\lambda) - \boldsymbol{X\beta})^\top (\boldsymbol{y}^*(\lambda) - \boldsymbol{X\beta})}{2\sigma_e^2} \right\} J(\lambda, \boldsymbol{y}),$$

where

$$J(\lambda, \boldsymbol{y}) = \prod_{i=1}^{n} \left| \frac{\partial y_i^*(\lambda)}{\partial y_i} \right|$$

is the Jacobian of the transformation from y to $y^*(\lambda)$. and $\boldsymbol{\theta}$ are the unknown parameters $\boldsymbol{\beta}$ and σ_e^2. This property comes from the transformation theorem defined as:

Theorem 1 (Transformation theorem)**.** *Let y be a continuous random variable with density function $f(y)$, taking values in $\mathbb{R}^n$. Let $T(y) = y^*$ a continuous transformation $T(y) : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, for which the inverse $T^{-1}(y^*)$ is also continuous. Suppose that the inverse of the transformation is differentiable for all values of $\mathbb{R}^n$ and the Jacobian is not equal to zero. Then $f_{T(y)}(y)$, the density function of the transformed target variable, is given by:*

$$f_{T(y)}(y) = f\left[T^{-1}(y^*)\right] |J(y)|$$

The maximum likelihood estimates are found in two stages. First, for fixed λ, the estimates for $\boldsymbol{\beta}$ and σ_e^2 are computed. When the Jacobian does not depend on $\boldsymbol{\beta}$ or σ_e^2 this is the likelihood for a least-square problem with response $y^*(\lambda)$. Hence

$$\hat{\boldsymbol{\beta}}(\lambda) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}^*(\lambda),$$

and

$$\hat{\sigma}_e^2(\lambda) = \frac{\boldsymbol{y}^*(\lambda)^\top \mathbf{A} \boldsymbol{y}^*(\lambda)}{n} = \frac{S(\lambda)}{n},$$

where $\mathbf{A} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and $S(\lambda)$ is the residual sum square in the transformed model. Holding λ as fixed and substituting $\hat{\boldsymbol{\beta}}(\lambda)$ and $\hat{\sigma}_e^2(\lambda)$ into the logarithm, we obtain, apart from a constant,

$$l_{max}(\lambda) = -\frac{n}{2} \log \hat{\sigma}_e^2(\lambda) + \log J(\lambda, \boldsymbol{y}). \tag{1.4}$$

The λ that maximizes the profile log-likelihood in Equation 1.4 will be selected. For the underlying optimization process by using the ML estimation method, the Newton-Raphson iterative procedure and its modifications are commonly used (Nelder and Mead, 1965; Lagarias et al., 1998).

A.2: Restricted maximum likelihood estimation method (REML)

As mentioned before, little attention has been paid in the literature to the study of data-driven transformations for linear mixed regression models: in particular, the improvement of the validity of model assumptions by departures from normality of both sources of randomness and the transformation parameter estimation methods are still under research. The work of Gurka et al. (2006) extends the use of the Box-Cox transformation under maximum likelihood theory for the estimation of the transformation parameter to the linear mixed regression models theory.

For the estimation of λ under the linear mixed regression model presented in Equation 1.2 and described in Gurka et al. (2006), we assume that the vectors $\boldsymbol{y}_i^*$ are independent and normal distributed for some unknown λ as follows:

$$\boldsymbol{y}_j^*(\lambda) \sim N(\boldsymbol{\mu}_j, \mathbf{V}_j) \quad \text{for} \quad j = 1, \ldots, m,$$

with

$$\boldsymbol{\mu}_j = \boldsymbol{X}_j \boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_j = \boldsymbol{Z}_j \boldsymbol{G} \boldsymbol{Z}_j^\top + \boldsymbol{R}.$$

where $\mathbf{1}_{N_i}$, is a column vector of ones of size N_i and $\boldsymbol{I}_{N_i}$ is the $N_i \times N_i$ identity matrix.

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the transformation from y to $y_j^*(\lambda)$, defined as

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{j=1}^{m} \prod_{i=1}^{n_j} \left| \frac{dy_{ij}^*(\lambda)}{dy_{ij}} \right| \\ &= \prod_{j=1}^{m} \prod_{i=1}^{n_j} y_{ij}^{\lambda - 1}. \end{aligned}$$

The log-likelihood function in relation to the original observations is obtained by multiplying the normal density by $J(\lambda, \boldsymbol{y})$ as:

$$\begin{aligned} l_{\text{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{m}\log|\boldsymbol{V}_j| \\ &- \frac{1}{2}\sum_{j=1}^{m}[\boldsymbol{y}_j^*(\lambda) - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}]^\top \boldsymbol{V}_j^{-1}[\boldsymbol{y}_j^*(\lambda) - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}] + \log J(\lambda, \boldsymbol{y}). \end{aligned}$$

The maximization process of $l_{\text{ML}}(\boldsymbol{\theta})$ leads to ML estimators of the unknown parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma_e^2, \boldsymbol{G})$. However, the REML theory is recommended when more accurate estimators of the variance components are needed (Verbeke and Molenberghs, 2000). This function is calculated by maximizing the ML of a set of error contrasts stemming from the fixed effects design matrix (Gurka et al., 2006). As a result, the REML function, in which the maximum possible number of linearly independent contrasts is $n - p$ (Harville, 1974), does not depend on $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{j=1}^{m}\boldsymbol{X}_j^\top\boldsymbol{X}_j\right| - \frac{1}{2}\sum_{j=1}^{m}\log|\boldsymbol{V}_j| \\ &- \frac{1}{2}\log\left|\sum_{j=1}^{m}\boldsymbol{X}_j^\top\boldsymbol{V}_j^{-1}\boldsymbol{X}_j\right| - \frac{1}{2}\sum_{j=1}^{m}[\boldsymbol{y}_j^*(\lambda) - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}]^\top\boldsymbol{V}_j^{-1}[\boldsymbol{y}_j^*(\lambda) - \boldsymbol{X}_j\hat{\boldsymbol{\beta}}] \\ &+ n(\lambda - 1)\log(\bar{y}), \end{aligned}$$

in which $\bar{y}$, is the geometric mean, defined as

$$\bar{y} = \left(\prod_{j=1}^{m}\prod_{i=1}^{n_j} y_{ij}\right)^{1/n}.$$

Bickel and Doksum (1981) studied the estimation properties of the parameters while using the Box-Cox transformation, whereby the inference about $\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2$ is conditioned on $\lambda = \hat{\lambda}$. They conclude that the asymptotic marginal unconditional variance of $\hat{\boldsymbol{\beta}}$ can be inflated for a fixed λ. The standard solution to this problem is to include the geometric mean of the response variable in the denominator of the Box-Cox transformation $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$, which converts it in a scaled transformation $Z(\lambda)$, whereby the unit is preserved and the interpretation is simplified, due to the fact that the units do not change as λ changes and the conditional variance of $\boldsymbol{\beta}$ is reduced. The Jacobian of this transformation is equal to one and the ML theory can be used for the linear mixed regression model. It is defined as follows:

$$Z(\lambda) = \begin{cases} \frac{y_{ij}^\lambda - 1}{\bar{y}^{\lambda-1}\lambda} & \text{if } \lambda \neq 0; \\ \bar{y}\log(y_{ij}) & \text{if } \lambda = 0, \end{cases}$$

for $y_{ij} > 0$. Gurka et al. (2006) recommend this scaled transformation in order to take advantage of procedures for estimating λ already computationally implemented.

B: Analytical considerations

Other analytical considerations have also been proposed in the literature as alternatives to ML-based methods. It consists of the use of distances or divergence measures, fit tests, and distribution moments (Hernandez and Johnson, 1980; Yeo and Johnson, 2000; Vélez et al., 2015). These approaches have also been studied in the context of linear mixed regression models (see e.g., Rojas-Perilla et al. (2017)). For some multiparameter transformations, such as the Johnson's System (Johnson, 1949), the method of moments of percentile points is proposed (George, 2007; Forbes et al., 2011). It is based on a simple selection rule introduced by Slifker and Shapiro (1980). Therefore, Chung et al. (2007) recommend the method of percentiles over the profile log-likelihood due to its simplicity. The hyperbolic power transformation is another example of a multi-parameter transformation. For this, the matching quantile approach (Tsai et al., 2017) is used for estimating the transformation parameters. Finally, in case the outcome variable is truncated, Poirier (1978) introduced a methodology as alternative of the ML method.

B.1: Estimators based on goodness of fit tests

In simple words, a goodness of fit test compares the empirical distribution, g, of a random sample against a theoretical distribution, f. Typically, a null hypothesis, H_0, is tested that assumes that f and g are statistically equal. If the hypothesis is rejected, we say that there is ground for believing that the sample is not f distributed. If we fail to reject H_0, the hypothesis that the sample is f distributed cannot be discarded. In the frame of the present work, f is the density function of the normal distribution. The goodness of fit tests can be employed to estimate the transformation paramater. The main idea is to maximize the statistic of such tests. Rahman (1999) employs the Shapiro-Wilk test. Rahman and Pearson (2008) make use of the Anderson-Darling test. Both focus on the Box-Cox transformation and use the Newton-Raphson algorithm to estimate the transformation parameter. However, these methods can also be applied to all one-parameter transformations mentioned in the present work. Yang and Abeysinghe (2003) make use of two score tests to determine transformation parameter for the Box-Cox transformation. Applications for multiple parameters transformations such as the Johnson transformation need to be further studied. Asar et al. (2017) extend the work of Rahman (1999) and Rahman and Pearson (2008) by utilizing seven goodness of fit tests, proposing a new algorithm. For a more detailed description about their method see Asar et al. (2017). Ruppert and Aldershof (1989) introduce an estimator for λ, σ_e^2 and β based on a test which depends on the correlation of the fitted values with the squared residuals. Other versions of this type of estimator are based on the Levene's test and Anscombe test. Finally, the work of Vélez et al. (2015) makes a selection of different normality type tests, classified in regression/correlation-,empirical distribution function-,and measure of moments-based tests. They develop a grid-search method for choosing the transformation where the combined p-value is the highest.

B.2: Estimators based on distribution moments: skewness and kurtosis

Skewness and kurtosis are major characteristics of the shape of distributions (Rosenthal, 2011). The former is a measure of the degree to which a distribution departs from symmetry; if it is negative, the left tail is long and the right short and thick. Positive values of skewness mean the contrary: a large right tail and a stubby left tail. The normal distribution has a skewness equal to zero. For a random variable z with mean μ_z and variance σ_z^2 the skewness is defined as

$$\gamma_1(z|\mu_z, \sigma_z^2) = E\left[\left(\frac{z-\mu_z}{\sigma_z}\right)^3\right].$$

Kurtosis is a measure of the degree of "tailedness" or "peakedness" concerning the normal distribution. A leptokurtic distribution has high kurtosis, which means that the probability of falling in the center is greater compared to that of the normal distribution. In contrast, a platykurtic distribution has more area, and therefore, more probability in the tails. The kurtosis for a standard normal distribution is equal to 3. Typically, the interest lies in the excess of kurtosis, which for the random variable z is defined as follows:

$$\gamma_2(z|\mu_z, \sigma_z^2) = \left[\frac{\left(E[z-\mu_z]\right)^4}{\left(E[(z-\mu_z)^2]\right)^2}\right] - 3.$$

Even tough the skewness is considered more important than the kurtosis when dealing with model assumption violations (Royston et al., 2011), the optimization of the last is also relevant. The parameter of the transformation is then chosen so that the value of skewness or kurtosis for the error term e_i is as close as possible to that of the normal distribution (Carroll and Ruppert, 1987).

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\text{argmin}}|\gamma_1(e_i)|,$$

and

$$\hat{\lambda}_{\text{kurt}} = \underset{\lambda}{\text{argmin}}|\gamma_2(e_i)|.$$

where $\gamma_1(e_i)$ is the skewness and $\gamma_2(e_i)$ denotes the kurtosis of the unit-level error terms.

The parameter could be also selected with the help of a statistical test that accounts for kurtosis or skewness (see, for instance, Gaudard and Karson (2007)). In the context of linear mixed regression models, an additional problem arises as there are two independent error terms to be considered. Therefore, a pooled skewness approach is suggested by Rojas-Perilla et al. (2017), if skewness minimization is chosen as the target criteria. This ensures that the larger the error term variance is, the more importance its skewness in the optimization has.

B.3: Estimators based on divergence or distance optimization

Only considering skewness may ignore many other properties of the distribution. Hence, a measure describing the distance between two distribution functions as a total might be prefer-

able. A few of these alternatives are based on the minimization of the Kullback-Leibler (KL) divergence, based on Kullback (1997) and described in Yeo and Johnson (2000) and Hernandez and Johnson (1980), and on measures of symmetry as the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) distances (Carroll, 1980; Bickel and Doksum, 1981; Carroll, 1982a; Taylor, 1985). For this method the real distribution of the data needs to be known. The exact formulations of the target measures are given as follows:

$$\hat{\lambda}_{\text{KL}} = \underset{\lambda}{\text{argmin}} \int_{-\infty}^{+\infty} f(y^*(\lambda)) \log \left[\frac{f(y^*(\lambda))}{\phi_{\mu,\sigma^2}} \right],$$

with f the probability density function of the transformed target variable $y^*(\lambda)$. ϕ_{μ,σ^2} denotes the probability density function of a normal distribution with mean μ and variance σ^2.

$$\hat{\lambda}_{\text{KS}} = \underset{\lambda}{\text{argmin}} \ \sup \ |(F(e^{std}) - \Phi|,$$
$$\hat{\lambda}_{\text{CvM}} = \underset{\lambda}{\text{argmin}} \int_0^1 \left[F(e^{std}) - \Phi^2 \right].$$

$F(\cdot)$ denotes the empirical cumulative distribution function (ecdf) estimated on the normalized residuals e^{std} and Φ is the distribution function of a standard normal distribution.

C: Robust adaptations

Draper and Cox (1969) stated that the ML method is robust to non-normal error terms as long as they are reasonably symmetric. It depends on parametric distributional assumptions and it is not robust to outliers. Therefore, different robust adaptations are proposed in the literature. Hinkley (1975, 1977); Hinkley and Runger (1984) and Taylor (1985) introduce and discuss a non-parametric and symmetry-based adaptation method of the ML procedure. This quick-choice method uses a symmetric distribution of the error terms about zero rather than the normal, and is based on an asymmetry measure based on order statistics (Taylor, 1985). It is also known as the Hinkley's quick method or quantile-based method because it studies how the quantiles of the distribution are symmetrically placed about the median. While this approach is not sensitive to outliers and robust in case the interquartile range is used, it is an inefficient method. Another similar quantile-based method for assessing the need of transforming data is suggested in Velilla (1993). Leinhardt and Wasserman (1979) and Emerson and Stoto (1982) propose the symmetrization of the quartiles around the median. However, Cameron (1984) pointed out that the method of Emerson and Stoto (1982) is not suitable for highly skewed data.

In order to access the accuracy of the ML estimator, Carroll (1980) and Bickel and Doksum (1981) propose another robust modification, also studied in Carroll (1982a) and Hinkley and Runger (1984). It generates a famous controversy in the study of transformations (see Doksum, 1984; Rubin, 1984; Johnson, 1984) and Carroll and Ruppert (1984)). They propose a robustification against heavy-tailed distributions in case the normality assumption is not present in the data and the Box-Cox transformation is required. This method is based on the robust estimator defined by Huber (1981), but it is not consistent in terms of mean squared error. Please note

that these robust adaptations are made for handling outliers only in the outcome variable and not in the explanatory ones.

In order to find a consistent and efficient non-parametric method, Han (1987) suggests an estimator based on the Kendall's rank correlation (Kendall, 1938). With the aim of covering some heavy tailed distributions, Carroll and Ruppert (1985, 1987, 1988) proposed a robust bounded influence method based on Kruskal and Wallis (1952) to a moderate number of outlying points in the data. Foster et al. (2001) introduce a consistent semi-parametric estimation method without assuming parametric assumptions on the error distribution. In general, these robust adaptations are not suitable for heavy contamination and heteroscedasticity. Therefore, Marazzi and Yohai (2006) derive a consistent estimation method based on the minimization of a robust measure of residual autocorrelation with respect to a robust fit of the transformed outcome variable. This approach is robust to outliers, even if normality and homoscedasticity are not present in the data (Marazzi and Yohai, 2004).

C.1: A robustified maximum likelihood estimator

Carroll (1980) develop a more robust version of the profile log-likelihood estimator motivated by a dilemma. On the one hand, as shown by Andrews (1971), the normal maximum likelihood is usually not robust to deviations from normality or outliers. Andrews (1971) proposes a more robust method to overcome the sensitivity to outliers of the likelihood methodology based on the F-test of significance. On the other hand, Atkinson (1973) shows in a Monte Carlo experiment that the original likelihood test proposed in Box and Cox (1964) is more powerful than the significance method introduced by Andrews (1971). Atkinson (1973) suggests a modified version of the ML approach that does not account for robustness. This leads to the situation where a powerful method delivers no robust results, while a more robust method seems not to be so powerful. Based on the Huber's method (Huber, 1992) and the profile log-likelihood methodology presented earlier, Carroll (1980) propose an estimator which considers not only the normal distribution but also distributions with "normal-centre" and "exponential-tails". The method is powerful for these types of distributions, but also relatively robust to Andrew's method of significance. The likelihood function for such distributions is given by

$$L(\lambda, \boldsymbol{\beta}, \sigma_e^2) = \frac{1}{(2\pi\sigma_e^2)^{\frac{n}{2}}} \sum_{i=1}^{n} \exp\left\{ -\rho\left(\frac{y_i^*(\lambda) - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{2\sigma_e^2}\right) + (1-\lambda)\log y_i \right\}, \tag{1.5}$$

where for some k and variable z

$$\rho(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq k; \\ k(|z| - \frac{k}{2}) & \text{if } |z| > k = 0. \end{cases}$$

Typical values of k are 1.5 or 2 (Carroll, 1980). Note that if $k = \infty$, Equation 1.5 is the normal likelihood for the Box-Cox transformation. λ, σ_e^2 and $\boldsymbol{\beta}$ are found in several stages. For further description of this algorithm please refer to Carroll (1980).

D: Bayesian approaches

As mentioned before, the ML estimator is not consistent in case non-normal errors are present and it is a non-robust methodology in the presence of outliers. Therefore, some research effort has been shifted towards alternative Bayesian estimation methods of the transformation parameter. The paper of Box and Cox (1964) propose a Bayesian estimation method for the transformation parameter, which uses a non-informative prior distribution of λ but is outcome-dependent. Pericchi (1981) introduced a solution for choosing a non-outcome-dependent a priori distribution, with a posterior log likelihood distribution similar in concept to the profile log likelihood-based on ML theory. Additionally, Sweeting (1984) suggested the use of a non-outcome-dependent family of non-informative priors distributions, which is closer in concept to that proposed by Box and Cox (1964).

1.2.2.2 Transformations to achieve homoscedasticity

Why is the homoscedasticity assumption important?

In linear regression analysis, the level of variance is assumed to be constant across the range of explanatory variables. This so-called homoscedasticity of the error term in the linear model can be formally written as $V(e_i|x_{i1}, ..., x_{ip}) = \sigma_e^2$. It means that the conditional variance of e_i, given the set of values $x_{i1}, ..., x_{ip}$, is not dependent on the xs (Wilcox, 2005). On the other hand, when the contrary occurs, we talk about heteroscedastic error terms, which can be expressed as $V(e_i|x_{i1}, ..., x_{ip}) = \sigma_{e_i}^2$. What happens when the assumption of homoscedasticity is violated? As stated in many econometrics textbooks, the ordinary least squares (OLS) estimators for the βs remain unbiased and consistent but are no longer efficient or best linear unbiased estimator (BLUE) (Williams et al., 2013). This means, the OLS estimator does not provide the smallest variance or the smallest standard error estimations (see Wooldridge (2000)). Therefore, if the interest lies only in the estimation of the βs, OLS can be used. However, if the focus is on inference, then t-tests, F-tests, and confidence intervals are no longer valid since there is a higher probability of y lying outside the confidence interval, for example, for large values of x. Heteroscedasticity can arise from different sources: first, as a result of a measurement error, for instance coming from the fact that some respondents give more precise answers (Berry, 1993); second, from misspecifications of the model, e.g., when an important variable is omitted and thus, the error term exhibits idiosyncratic variation (Wooldridge, 2000); third, when the population should be clustered and thus variance changes across subpopulations (Natrella, 2013); and fourth, if there are outliers which means one or a few observations severely affect the non-robust variance estimator and induce (apparent) heteroscedasticity (Carroll, 1980). Some papers related to the consequences when homoscedasticity assumptions are not satisfied are in Cochran (1947) and Eisenhart (1947).

How can we check the homoscedasticity assumption fulfillment?

To graphically explore the homoscedastic assumption, let us suppose that we want to regress y against a vector containing one single explanatory variable, x. If the error term is homoscedastic, we would expect the set of points $[x, y]$ to spread along the regression line on the scatter-

plot exhibiting the same level of variation. A visual inspection of heteroscedasticity is made by plotting the residuals against the fitted values and the residuals versus a predictor which is possibly generating the violation of this assumption. There is also a huge range of tests for assessing homoscedasticity in the literature (Kirk, 1968). For detecting any linear form of heteroscedasticity, the Glesjer, Breusch-Pagan, Goldfeld-Quandt and Cook-Weisberg tests are commonly used. Additionaly, the White's general test is useful when non-linear forms of heteroscedasticity need to be proved. Other suitable tests are the Hartley's Fmax (Hartley, 1950) and Cochran's C (Cochran, 1941), but they are sensitive to Gaussian assumptions, and the Bartlett's test (Bartlett, 1937) and Levene's test (Levene et al., 1960), among others. Additionally, the Ramsey Regression Equation Specification Error Test (RESET) test (Ramsey, 1969) can be used for the misspecification of the model.

What are the alternative methods to overcome heteroscedasticity?

If we have tested the correctness of the assumption and found statistical support to believe that the error term is heterosedastic, a pre-analysis should be carried out before jumping into methods to correct for heteroscedasticity. First, model misspecification should be left to field experts for methodological issues. This is because heteroscedasticity arising from model misspecification is not genuine heteroscedasticity, but model misspecification since, by re-specifying the model, one could get rid of it. The need of clustering should be examined as well. It is also recommended to remove or replace outliers, or just apply an outlier treatment and then test for heteroscedasticity to verify if the homoscedasticity assumption is being violated by the influence of one or a few observations.

Alternatively, if the error term exhibits heteroscedasticity, a more robust and efficient estimator can be achieved via modified OLS residuals or generalized least squares (see Wooldridge (2000)). It includes the use of feasible generalized least squares and weighted least squares regression by minimizing a weighted sum of squared residuals (Berry, 1993). The downside of the latter is that the form of the weights is often unknown. Secondly, techniques for estimating robust standard errors can be applied. They are known as Eicker-, Huber-, White-, Eicker-Huber-White-, heteroscedasticity-consistent-, Huber-White-standard errors or sandwich estimators (Eicker, 1967; Huber, 1967; White, 1980). Thirdly and most widely used, is the application of generalized linear regression models (Nelder and Wedderburn, 1972). These models take specific heteroscedasticity forms into account and contain different data structures; for instance, logistic regression for dichotomous (binary) variables or Poisson regression for count data. Additionally, Bayesian linear regression approaches can also account for the lack of homoscedasticity.

How can transformations help to improve homoscedasticity?

According to Johnson (1949) and based on Bartlett (1937) and Bartlett (1947), transformations might provide a fair correction for heteroscedasticity. When a functional dependence of the variance of the outcome variable on the mean is present in the data, we may gain the advantages of using variance-stabilizing transformations. This dependence mostly implies an underlying distributional process and determines the form of the suitable transformation. Table

1.2 shows different relations between these moments, the corresponding suitable transformations, some examples of appropriate distributions and the range of the outcome variable that the transformation supports. According to Ruppert (2001), populations which have larger means also exhibit the property of larger variances. If we denote the mean of the conditional distribution of the outcome variable given a vector of explanatory variables by $\mathrm{E}(y|x) = \mu_y(x)$, then it is possible that the conditional variance $\mathrm{Var}(y|x)$ is a function of $\mu_y(x)$. This relation is denoted by $\mathrm{Var}(y|x) = R[\mu_y(x)]$ for some function $R(\cdot)$. Without loss of generality and following Ruppert (2001), if we use a transformation $T(y)$, this relation holds the delta-method linear approximation and is denoted as follows (Bartlett, 1947):

$$Var\Big[T(y)|x\Big] \approx \Big\{T^{'}\big[\mu_y(x)\big]\Big\}^2 R\Big[\mu_y(x)\Big].$$

The transformation $T(y)$ will correct the variance assumptions, if $\big[T^{'}(y)\big]^2 R(y)$ is constant. For instance and following Ruppert (2001), if $R(y) \propto y^{\alpha}$, then $T(y) \propto y^{1-\frac{\alpha}{2}}$ would be a variance-stabilizing transformation, with $\alpha \neq 2$. The transformations that are used most for the issue of achieving homoscedasticity are square roots, logarithms, reciprocals, and trigonometrical transformations (Cohen et al., 2014). Some of these are also known as double bend transformations, because the data sets for which these are used, are bound at both top and bottom, such as data sets from the binomial distribution.

Bartlett (1937) proposes the use of the square root transformation to stabilize variances that are exactly proportional to the mean, which is the case for gamma and exponential distributed data, as for such a distribution in which the variance is exactly equal to the mean, which is the case of the Poisson distribution. In this case, $\alpha = 1$, that means, $g(y) \propto y^1$, then $T(y) \propto y^{1-\frac{1}{2}}$ is the root square, which is the variance-stabilizing transformation for Poisson data sets. In a later work, Bartlett (1947) and Anscombe (1948) suggest the use of $\sqrt{y + c_1}$ type transformations, where c_1 is a fixed constant. In case of a large sample size this transformation with a constant c_1 is more useful to achieve a constant variance. They propose handling heteroscedasticity by using $c_1 = 1/2$ or $c_1 = 3/8$, when y takes only small values or when zeros are common in the data, respectively. Freeman and Tukey (1950) proposes a more sophisticated twofold transformation, which is called the Freeman-Tukey deviate or the chordal transformation and is denoted by $\sqrt{y} + \sqrt{y+1}$. This transformation is particularly suitable in case y is very small or equal to 0. Similarly, the inverse transformation is recommended for stabilizing the variance for observations that are mostly close to zero. It stabilizes the variance when $n > 3$ (Mosteller and Bush, 1954; Mosteller and Youtz, 2006).

The negative binomial distribution is appropriate to represent for Poisson distributed data under overdispersion, that means, the variance greater than the mean. For this kind of data sets some transformations based on the logarithm and hyperbolic trigonometric functions are proposed (Bartlett, 1947; Chatterjee and Hadi, 2015). For instance, some modifications of the inverse hyperbolic sine function, such as $\sinh^{-1}\sqrt{\left(\frac{y+c_2}{k+c_3}\right)}$, are suitable for the negative binomial data. While Anscombe (1948) suggests values of $c_2 = 3/8$ and $c_3 = -3/4$, Beall (1942) proposes using $c_2 = 0$ and $c_3 = 0$. Especially recommended for small values is the adjustment $\frac{1}{\lambda}\sinh^{-1}(\lambda\sqrt{y + 1/2})$ (Chatterjee and Hadi, 2015).

In order to stabilize the variance of binomial distributed data, different trigonometric transformations are suggested. For instance, the inverse sine root square transformation of the form $\sin^{-1}\sqrt{y}$, also called the angular transformation (Fisher, 1922b; Bartlett, 1937), is analogous to the root square transformations for binary data. Thus, this variance-stabilizing transformation is widely used in practice. Some modifications based in this transformation have been proposed according to specific values of the parameters of the distribution based on the data set are proposed in Curtiss (1943) and Anscombe (1948). For instance, $\sin^{-1}\sqrt{\frac{y+c_4}{n+c_5}}$ is then suitable for data from the binomial distribution. Researchers commonly use $c_4 = c_5 = 0$. However Bartlett (1937) suggests $c_4 = 1/2$ and $c_5 = 0$ and Anscombe (1948) improves this transformation further by setting $c_4 = 3/8$ and $c_5 = 3/4$. Unlike similar transformations, the arcsine is defined for y between 0 and 1. However, research done by Wilson et al. (2013) and Warton and Hui (2011) have warned about employing this transformation. According to Warton and Hui (2011) one of the downsides of this transformation is that if the relation between the untransformed y and the independent variables $x_{ip}, ..., x_{ip}$ is e.g., always increasing, the same relation is not held after transformation due to the periodicity of $\arcsin$. Sofisticated twofold transformations are also suggested by Laubscher (1961) and Freeman and Tukey (1950). To correct for heteroscedasticity of variables contained to a bounded interval, such as proportions and percentages, two-bend transformations families can be appropriate. For instance, the most common transformations are the logit, probit, Guerrero-Johnson, Aranda-Oraz, beta, angular and arsine transformations. For detailed information about transformations for these kinds of data sets please refer to Kruskal (1968); Atkinson (1987) and Piepho and McCulloch (2004). This topic falls out of the scope of the present work.

The ordinary power transformations family, in which different powers of the target variable are applied, are defined according to the functional dependence of the variance on the mean. If the variance increases proportional to the mean on a square root scale, the stabilization is made on a logarithmic scale (Bartlett, 1947). This is the case of the log-normal distribution. Fisher and Yates (1949) proposed some modifications of the logarithmic transformation in case the values are less than 10 and for larger values. For distributions with constant coefficient of variation, such the exponential or gamma with constant shape parameter distribution, the logarithm is also recommended (Ruppert, 2001). This transformation is generally suggested when the range of the outcome variable is very broad but not negative (Fink, 2009). For data from other distributions as the Gamma or Weibull distribution a variance stabilizing transformation is recently proposed by Lakhana (2014). When data is very bunched to the minimum and maximum of the distribution the transformation presented by Fink (2009) can be used for stretching the data. For selecting the parameters λ and k of this transformation we refer to Fink (2009); Erickson and Nosanchuk (1977) and McNeil (1977). Additionally, if the data presents heteroscedasticity problems and the distributional form is not clear or there are other violations of assumptions, some of the already mentioned transformations in the beginning of Section 1.2.2 also help to correct heteroscedasticity, since stabilizing variance and normalizing errors often goes together (Johnson, 1949). These transformations include in particular the logarithm, gpower, Box-Cox, Johnson, Manly, and Yeo-Johnson transformations. That means, the researcher should empirically find the most appropriate transformation that stabilizes the

variance of the data regardless the mean value (Montgomery, 2008). Finally, transforming both sides helps for both, stabilizing the variance and create more symmetric distributions (see Section 1.3 for the both sides methodology).

Table 1.2: Transformations for achieving homoscedasticity

Dependence	Source	Formula	Example	Support
$\sigma_y^2 \propto \mu_y$	Bartlett (1937)	$\sqrt{y}$	Poisson(λ)	$y \geq 0$
$\sigma_y^2 \propto \mu_y$	Bartlett (1947)	$\sqrt{y + c_1}$	Poisson(λ)	$y \geq -c$
$\sigma_y^2 \propto \mu_y$	Freeman and Tukey (1950)	$\sqrt{y} + \sqrt{y+1}$	Poisson(λ)	$y \geq -1$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\log(y)$	lognormal(μ, σ^2)	$y > 0$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\log_{10}(y)$	lognormal(μ, σ^2)	$y > 0$
$\sigma_y^2 \propto \mu_y^2$	Fisher and Yates (1949)	$\frac{1}{3}\sqrt{y}$	lognormal(μ, σ^2)	$y \geq 0$
$\sigma_y^2 \propto 2\mu_y$	Freeman and Tukey (1950)	$\sqrt{2y}$	$\chi^2(k)$	$y \geq 0$
$\sigma_y^2 \propto 2\mu_y$	Wilson and Hilferty (1931)	$y^{1/3}$	$\chi^2(k)$	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y + \lambda^2\mu_y^2$	Bartlett (1947)	$\lambda \log(y)$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y > 0$
$\sigma_y^2 \propto \mu_y + \lambda^2\mu_y^2$	Bartlett (1947)	$\log(y)$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y > 0$
$\sigma_y^2 \propto \mu_y + \lambda^2\mu_y^2$	Bartlett (1947)	$\lambda^{-1}\sinh^{-1}(\lambda\sqrt{y})$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y \geq 0$
$\sigma_y^2 \propto \mu_y + \lambda^2\mu_y^2$	Bartlett (1947)	$\lambda^{-1}\sinh^{-1}\left(\lambda\sqrt{y+\frac{1}{2}}\right)$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y \geq -\frac{1}{2}$
$\sigma_y^2 \propto \mu_y + \lambda^2\mu_y^2$	Beall (1942)	$\sinh^{-1}\sqrt{\frac{y+c_2}{r+c_3}}$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y \geq 0$
$\sigma_y^2 \propto \mu_y$	Ruppert (2001)	$\log(y)$	BN$\left(r, p, \lambda = \frac{1}{\sqrt{r}}\right)$	$y > 0$
$\sigma_y^2 \propto \mu_y$	Lakhana (2014)	$\begin{cases} \frac{(\sqrt{y+1})^\lambda}{\lambda} & \text{if } \lambda \neq 0; \\ \log(\sqrt{y+1}) & \text{if } \lambda = 0. \end{cases}$	$\Gamma(\alpha, \beta)$, Weibull(l, k)	$y \geq -1$
$\sigma_y^2 \propto \mu_y$	Wilson and Hilferty (1931)	$y^{1/3}$	$\Gamma(\alpha, \beta)$	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y$	Curtiss (1943)	$\sqrt{y + c_1}$	$\Gamma(\alpha, \beta)$	$y \geq$ -c
$\sigma_y^2 \propto \mu_y$	Ruppert (2001)	$\log(y)$	exp(λ)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Bartlett (1937)	$\sin^{-1}\sqrt{y}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Bartlett (1937)	$\sin^{-1}\sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Anscombe (1948)	$\sqrt{n+c_6}\sin^{-1}\sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Laubscher (1961)	$\sqrt{n}\sin^{-1}\sqrt{\frac{y}{n}} + \sqrt{n+1}\sin^{-1}\sqrt{\frac{y+\frac{3}{4}}{n+\frac{3}{2}}}$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Freeman and Tukey (1950)	$\sqrt{n+\frac{1}{2}}\left(\sin^{-1}\sqrt{\frac{y}{n+1}} + \sin^{-1}\sqrt{\frac{y+1}{n+1}}\right)$	Bin(n, p)	$y \geq 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Fisher (1922b)	$\sin^{-1} y$	Bin(n, p)	$0 \leq y \leq 1$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Fisher (1922b)	$\sin^{-1}\sqrt{\frac{y+c_4}{n+c_5}}$	Bin(n, p)	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Curtiss (1943)	$\sqrt{n}\sin^{-1}\sqrt{y + \frac{c_7}{n}}$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Curtiss (1943)	$\sqrt{n}\log(y)$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y(1-\mu_y)$	Curtiss (1943)	$\frac{1}{2}\sqrt{n}\log\left(\frac{y}{1-y}\right)$	Bin(n, p)	$y > 0$
$\sigma_y^2 \propto \mu_y^3$	Draper and John (1981)	$\frac{1}{\sqrt{y}}$	-	$y > 0$
$\sigma_y^2 \propto \frac{1}{\mu}$	Draper and John (1981)	y^2	-	$y \in \mathbb{R}$
$\sigma_y^2 \propto \mu^2$	Draper and John (1981)	$\log(y)$	-	$y > 0$
$\sigma_y^2 \propto \mu_y$	Draper and John (1981)	$\sqrt{y}$	-	$y \geq 0$
$\sigma_y^2 \propto \mu_y^4$	Draper and John (1981)	$\frac{1}{y}$	-	$y \neq 0$

Note: e*0.1cm Note: Please note that due to lack of different parameter names and conventional definitions of the distributions in column Example the parameter names can conflict with the notation in the rest of the paper. The parameter $c_1 = 1$ is widely used in practice. However, Bartlett (1937) and Anscombe (1948) recommend $c_1 = \frac{1}{2}$ and $c_1 = \frac{1}{3}$, respectively. Beall (1942) suggests $c_2 = c_3 = 0$ and Anscombe (1948) $c_2 = \frac{3}{8}$, $c_3 = \frac{-3}{4}$. This author recommends $c_4 = \frac{3}{8}$ and $c_5 = \frac{3}{4}$, meanwhile Bartlett (1937) $c_4 = \frac{1}{2}$ and $c_5 = 0$. However, $c_4 = c_5 = 0$ are often used in practice. Anscombe (1948) suggests $c_6 = \frac{1}{2}$. Curtiss (1943) suggests c_7 equal to 0 or $\frac{1}{2}$, depending on the values of p.

How can we estimate the transformation parameters to homoscedasticity?

In general, the approaches for estimating the optimal transformation parameter to homoscedasticity are ML-based or analytical considerations. Therefore, in case a transformation for simultaneous correcting non-normality and heteroschedasticity is selected, then the ML-based approaches presented already for normality can be used (see Section 1.2.2). However, Zarembka (1974a) pointed out that this method is not robust in the presence of heteroscedastic error terms. Therefore, Blaylock and Smallwood (1985) propose an alternative adaptation, the robustified maximum likelihood estimator. Since it is especially useful, we explain it in detail below. Hinkley (1985) suggests the use of an analytical likelihood-based method for analyzing local deviations. This procedure for estimating the transformation parameter considers both the homoscedasticity model violation of residuals and the lack of additivity. Ruppert and Aldershof (1989) propose a method which attempts to deal with non-normality and heteroscedasticity. It is based on the minimization of the correlation between the fitted values and the squared residuals.

A.3: Robustified maximum likelihood estimator

Blaylock and Smallwood (1985) propose a more robust version of the profile log-likelihood, which allows for unequal variances across observations, but considers all elements off the diagonal of variance-covariance matrix Σ as zero. The functional form of σ^2_{ei} is chosen as:

$$\sigma^2_{ei} = \exp\{\delta w_i^*(\lambda_w)\},$$

where w is an instrumental variable upon which the error term depends, λ_w is the transformation parameter for w, and δ allows for different forms of heteroscedaticity. Indeed, when $\delta = 0$, the homoscedastic form is obtained. This means that the homoscedasticity case is nested in the form of the variance σ^2_{ei}. Thus, the likelihood ratio test can be employed to compare the model with and without heteroscedasticity. The estimates for Σ, β, λ, and δ are obtained by employing the profile log-likelihood function. In the first stage, Σ and β are estimated using a nonlinear optimizing algorithm; afterwards, values for λ and δ are selected so that the profile log-likelihood is maximized.

1.2.2.3 Transformations to achieve linearity and additivity

Why is the linearity assumption important?

As it is implied in its name, the linear regression is an approach to model linear relationships. The linear regression model is linear in two senses: first, the model is linear in the variables because each response y is expressed as a weighted sum of the independent variables where the parameters are the weights (Dougherty, 2011); second, the model is also linear in the parameters where, this time, the independent variables are the weights. If non-linearity is present and we decide to follow through with the use of linear techniques as in OLS, the consequences would be misrepresenting the actual relationship. Therefore, when non-linearity occurs, it is very likely that estimation and inference techniques based on the linearity of the model yield

misleading conclusions. In addition to linearity, it is important that the additivity assumption is met. This assumption ensures that the independent variables multiplied by their regressors can be added together to provide an estimate (Berry, 1993). However, given the complexity of many empirical relationships, it is sometimes expected that the effect of an independent variable x_1 on y may be influenced by a third variable, x_2. This interaction not only violates the implicit assumption of additivity, but it also becomes a practical problem since it leads to multicollinearity (Friedrich, 1982). Moreover, when a non-additive relationship takes place, and it is not detected or is ignored, the linear regression yields unreliable results since the relationship that is being represented fails to account for the interaction between the independent variables. Again, as in the presence of non-linearity, estimation and inference techniques based on the linear regression model provide non-accurate results (Williams et al., 2013).

How can we check the linearity assumption fulfillment?

A useful visual method to examine non-linearity is using scatterplots between the outcome variable and the explanatory variables, which is called added variable plot, also known as partial-regression- or adjusted plot (Atkinson, 1982). Additionally, a scatter plot of the standardized residuals and the standardized predicted values of y is also useful. If the relationship appears to take a line-like form, we do not need to occupy ourselves with correcting for non-linearity. Additionally, the RESET test, a general test for functional form misspecification proposed by Ramsey (1969, 1974) can be used as an indicator of lack of linearity.

A technique to detect non-additivity effects is the Tukey's test (Tukey, 1949; Moore and Tukey, 1954). As an alternative to Tukey's test, Barry (1993) introduces a Bayesian test to check the validity of this assumption.

What are the alternative methods to overcome non-linearity?

If the assessment tools provide evidence for non-linearity and/or non-additivity, a model restructuring is a possible solution. For instance, if the relation between the dependent and independent variables seems to be curvilinear, a curve component could be added and tested on significance (Osborne and Waters, 2012). For receiving additivity, Friedrich (1982) favors the use of multiplicative models over dropping interactive variables to use linear regression techniques. If non-linearity or non-additivity is still present, ridge regression, also known as linear regularization, is particularly useful. Other alternative methods are Tikhonov regularization, Tikhonov-Miller method, Phillips-Twomey method, constrained linear inversion method or weight decay (Hoerl and Kennard, 1970), lasso regression (Tibshirani, 1996) and Bayesian linear regression.

How can transformations help to improve linearity?

In general, transformations to linearize data can be divided into two classes: in one class, the expected response is related to the independent variables by a known non-linear function; in the other, the relationship between the expected response and the explanatory variables is not exactly known (Cook and Weisberg, 1982). For the first class, transformations can be easily

selected. Wood and Gorman (1971) show plots for a comprehensive number of non-linear functions that can be transformed into linear ones. In the second class fall transformations such as the Box-Cox transformation, which have the potential to correct non-normality, heteroscedasticity, and non-linearity, so that, after the data is transformed, normal theory methods and linear regression techniques can be employed. An approach for selecting a suitable power transformation is given by Mosteller and Tukey (1977), who introduce a trial-and-error heuristic to linearize data based on the ladder of powers, called the "ladder of transformations", as shown in Table 1.4 and Table 1.5. This is also known as the "bulging rule" and it determines the value of the power employed for both the outcome variable and the explanatory variables within the model. Please follow Brown (2015) for more details about the bulging rule. Power transformations are useful if the relationship between x and y is a simple monotone. Table 1.3 summarizes transformations for regression forms that can be linearized since the relation is known for the simple linear regression model, and is based on Weisberg (1980); Fink (2009); Johnson (2009) and Chatterjee and Hadi (2015). The generalization of this table for the multiple regression form can be find in Fink (2009). Additionally, Box and Tidwell (1962) propose an iterative methodology known as the Box-Tidwell transformation to linearize the relationship between the dependent variable and the explanatory variables. It is basically based on individually finding the optimal power transformation to transform the set of explanatory variables. A power transformation test can help to determine which variable should be transformed or not (Brown, 2015). Finally, both sides methodology is also suitable for dealing with non-linearity problems in the regression model (see Section 1.3). Nevertheless, one should be careful when transforming both sides to induce linearization, since it may produce heteroscedasticity of the error term (Carroll and Ruppert, 1988). A tentative transformation to linearize multiplicative models is the logarithmic transformation. For non-additivity, Tukey (1949) recommends the use of the t-score of added non-linear terms as the transformation criteria. Without loss of generality, the transformations that are suitable for correcting non-additivity have a restricted form and the works of Elston (1961) and Anscombe and Tukey (1963) concentrate on the selection of the power. Rocke (1993) suggests the use of the t-score as a criteria to linearize proportional data.

Table 1.3: Transformations to achieve linearity when the relation is known for the simple linear regression model

Reference	Regression form	Transformation	Linear model
Weisberg (1980)	$y = \beta_0 x_1^{\beta}$	$y^* = \log y, x^* = \log x$	$y^* = \log \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \beta_0 e^{\beta_1 x}$	$y^* = \log y$	$y^* = \log \beta_0 + \beta_1 x$
Weisberg (1980)	$y = \beta_0 + \beta_1 \log x$	$x^* = \log x$	$y^* = \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \frac{x}{\beta_0 x - \beta_1}$	$y^* = \frac{1}{y}, x^* = \frac{1}{x} x$	$y^* = \beta_0 - \beta_1 x^*$
Chatterjee and Hadi (2015)	$y = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$	$y^* = \log\left(\frac{y}{1-y}\right)$	$y^* = \beta_0 + \beta_1 x$
Fink (2009)	$y = \beta_0 + \beta_1\left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	$y^* = \beta_0 + \beta_1 x^*$
Weisberg (1980)	$y = \frac{1}{\beta_0+\beta_1 x}$	$y^* = \frac{1}{y}$	$y^* = \beta_0 + \beta_1 x$
Johnson (2009)	$y = \beta_0 + \beta_1\sqrt{x}$	$x^* = \sqrt{x}$	$y^* = \beta_0 + \beta_1 x^*$

How can we estimate the transformation to linearity?

For the transformations that fall in the second class, the ML-based methods and analytical considerations that we already introduced are equally applicable for achieving linearity. A special approach to find the correct power when the regression form is known is given by Mosteller and Tukey (1977), who introduce a trial-and-error heuristic to linearize data based on the ladder of powers shown in Table 1.4 and Table 1.5. Tukey (1949) introduces the minimization of the F-value for the degree of freedom for non-additivity as an estimation method of a transformation.

The Tukey and Mosteller estimation algorithm

As mentioned before, Mosteller and Tukey (1977) propose a graphical bulging rule for selecting a power transformation, which is based on power of ladders. This seeks to guide practitioners to simply select a linearizing relationship transformation for any random variable z. The ladders are tabulated as follows:

Table 1.4: The ladder of powers

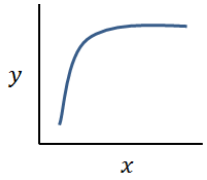
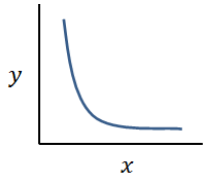
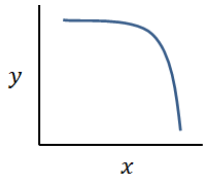
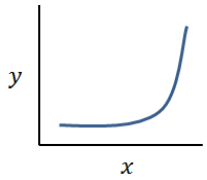
λ_i	-2	-1	-0.5	0	0.5	1	2
z	$\frac{1}{z^2}$	$\frac{1}{z}$	$\frac{1}{\sqrt{z}}$	$\log z$	$\sqrt{z}$	z	z^2

They can be generalized and formally expressed as:

$$y_i^*(\lambda) = \begin{cases} y_i^{\lambda_1} = \beta_0 + \beta_1 x_i^{\lambda_2} & \text{if } \lambda_1, \lambda_2 \neq 0; \\ \log yi = \beta_0 + \beta_1 x_i^{\lambda_2} & \text{if } \lambda_1 = 0, \lambda_2 \neq 0; \\ y_i^{\lambda_1} = \beta_0 + \beta_1 \log x_i & \text{if } \lambda_1 \neq 0, \lambda_2 = 0; \\ \log yi = \beta_0 + \beta_1 \log x_i & \text{if } \lambda_1, \lambda_2 = 0. \end{cases}$$

The parameters λ_1 and λ_2 are chosen according to Table 1.4 and Table 1.5. Examining a scatterplot of y against x leads us to select a power transformation based on the pattern of the curvature. We have two options for transforming: transform y by moving up/down the ladder or up/down the ladder for x depending on the pattern. That means in case the pattern is hollow upward, one should go down the ladder; and if hollow downward go up the ladder.

Table 1.5: The ladder of transformations

Pattern	Transformation	Parameter
	$y^* = y^{\lambda_1>1}, x^* = x^{\lambda_2<1}$	λ_1 up and/or λ_2 down
	$y^* = y^{\lambda_1<1}, x^* = x^{\lambda_2<1}$	λ_1 down and/or λ_2 down
	$y^* = y^{\lambda_1>1}, x^* = x^{\lambda_2>1}$	λ_1 up and/or λ_2 up
	$y^* = y^{\lambda_1<1}, x^* = x^{\lambda_2>1}$	λ_1 down and/or λ_2 up

Mosteller and Tukey (1977) present a simple numerical algorithm, which is explained as follows:

1. Plot x against y.
2. Based on Table 1.4, choose λ_1 and λ_2 according to the shape exhibited by the points on the scatter plot of x against y.
3. Transform y by y^{λ_1} and x by x^{λ_2}.
4. Plot the transformed predictor against the transformed response variable.
5. If the relationship appears to be linear: stop.
6. Otherwise, choose new values for λ_1 and/or λ_2 by going up or down the power ladder based on Table 1.4.

1.2.2.4 Parameter inference and interpretation

Does a transformation influence the inference on the model parameters?

The inference analysis is a controversial question that arises when a transformation, and especially a transformation with a transformation parameter, is used under the linear and linear mixed regression model. One question is whether we should treat the transformation parameters as fixed in case we are making inferences on the model parameters. If the transformation does not contain a data-driven transformation parameter common model inference can be conducted. In contrast, when using data-driven transformations, one point of discussion concerns if the transformation parameter can be treated as known or not. Ruppert (2001) and Box and Cox (1982) stated that the regression parameter estimates strongly depend on the chosen transformation parameter λ. Box and Cox (1964) further pointed out that after selecting a value for λ via e.g., ML-based methods, this should be treated as known and inference can be carried out as usual. However, Bickel and Doksum (1981) made a remark on this by studying the joint distribution of $\hat{\lambda}$ and $\hat{\beta}$. They found that when the real value of λ is unknown, the estimates for the variance of the $\hat{\beta}$s are inflated and highly dependent on the $\hat{\lambda}$ estimate. Box and Cox (1982) replied by saying that this was not only obvious, but also irrelevant, since "the gross correlation effects would be avoided if, following [their] paper, the investigation had been conducted in terms of [the normalized transformation]". Note that the normalized transformation is equivalent to the scaled transformation presented in Section 1.2.2. Furthermore, Hinkley and Runger (1984) carried out a sensitivity analysis where they found that the estimates of contrast and scale parameters are quite stable on the scale of the normalized transformation, whereas the estimates of location parameters, such as the mean, are more dependent on the value of $\hat{\lambda}$.

Research on the accuracy of the estimation and inference on the random effects after applying a transformation under a linear mixed regression model is still necessary. Under this scenario, the works of Verbeke and Lesaffre (1996) and Gurka et al. (2006) discussed, in a simulated scenario, the effects of a transformation on the inference process. Gurka et al. (2006) suggests including a correction factor from the Jacobian of the Box-Cox transformation in the estimated coefficients.

How is the inference process on the transformation parameters?

Inference about the transformation parameters is also a fundamental step in the transformation selection process. For testing the hypothesis $H_0 : \lambda = \lambda_0$, we could use the standard likelihood-based methods for getting a likelihood ratio test. The test statistic would be $W = 2[L_{\max}(\hat{\lambda}) - L_{\max}(\lambda)]$, which is chi-squared asymptotically distributed. Box and Cox (1964) extend this theory and propose two approaches to make inferences about the parameters after applying a transformation. In the first approach, large sample maximum likelihood theory is applied, which delivers point estimates of the parameters and provides an approximate test and confidence intervals based on the chi-squared distribution. In the second approach, Bayesian theory is applied. For that, the prior distributions for β and σ^2 are assumed to be uniform, obtaining a posterior distribution for λ. For more details about the Bayesian method please see Box and Cox (1964) and Jeffreys (1998).

Following Box and Cox (1964), an approximate $100(1-\alpha)$ per cent confidence interval is

$$L_{\max}(\hat{\lambda}) - L_{\max}(\lambda) < \frac{1}{2}\chi^2_{\nu}(\alpha),$$

$$L_{\max}(\lambda) = -\frac{1}{2}n\log\left[\hat{\sigma}^2(\lambda)\right] + \log\left[J(\lambda, y)\right],$$

where ν is the number of independent components in λ, α denotes the significance level, and $\hat{\sigma}^2(\lambda)$ represents the residual sum of squares in the transformed outcome variable.

In the same way, in order to test $H_0 : \lambda = \lambda_0$, Andrews (1971) proposes a test which ignores the Jacobian of the applied transformation. However, Atkinson (1973) re-introduces the Jacobian, developing a score-type statistic, which is not a maximum likelihood-based method. This is also known as the Atkinson's score statistic and was further standardized by Lawrance (1987a,b), the result of which is called Lawrance's statistic. Some robust versions of these tests are proposed by Carroll (1980) and Wang (1987).

Last but not least, some studies regarding the consistency and efficiency properties, as well as the the asymptotic variances of the estimated λ in the Box-Cox transformation, have been published. See Bickel and Doksum (1981); Carroll and Ruppert (1981); Carroll (1982a); Doksum and Wong (1983) and Hinkley and Runger (1984) for detailed information.

How are the model results interpreted when a transformation is applied?

One of the biggest challenges that researchers face when working with transformations is the interpretation of the results. It implies choosing the scale in which we need to present the results, depending on the research question. O'Hara and Kotze (2010) summarized this issue by pointing out that transformations comes at some cost to the trade-off between accuracy and interpretability. When working with the logarithmic transformation, an approximation helps to obtain a meaningful interpretation of the coefficients as percentages. However, this is a feature rarely observed when working with other non-linear transformations, such as the Box-Cox transformation family. In the words of Box and Cox (1964), transformation parameters that are obtained by maximum likelihood-based methods, which are widely used in practice for finding a suitable transformation, are "useful as a guide" but "not to be followed blindly". The selection of transformation parameters could be made based only on the information provided by the data. However, if a particular value for λ in the Box-Cox transformation is more convenient regarding interpretability, the selection of the parameter could be adjusted. For instance, if the output of an estimation suggests that λ should be equal to 0.25 one could work instead with $\lambda = 0$ i.e., the logarithmic transformation, which has an easier interpretation, especially when this choice is common in the specific research field.

Does the back-transforming process lead to bias in the predictions?

Researchers interested in predictions face another challenge which is to deal with the back-transforming bias when applying non-linear transformations. In case, a back-transformation is used for getting the values in their original measurement scale, an artificial bias comes from

this re-transforming process. Without loss of generality:

$$T[E(y|x)] \neq E[T(y)|x]$$

for all non-linear transformations, $T(\cdot)$, applied on the target variable. Although this effect is not always severe (Sakia, 1992), ignoring the magnitude of the generated bias may lead to misleading conclusions. Therefore, several methods and empirical work for removing the back-transforming bias after applying a power transformation, in particular, the logarithmic and Box-Cox transformations have been proposed in the literature for the linear and the linear mixed regression model (Neyman and Scott, 1960; Hoyle, 1973; Lee, 1982; Sakia, 1988; Rothery, 1988; Sakia, 1990, 1992; Newman, 1993; Gurka et al., 2006; da Costa and Crepaldi, 2014).

1.3 Further issues regarding to variable transformations

Additionally to the model assumptions that we discussed in the previous section, special features in the data can interact with the transformations or have effects on the usage of transformations. Thus, this section discusses issues such as model selection, the presence of outliers, incomplete responses, multimodal data, zero inflated data, and the range of the variable when using transformations. Note that these issues are a selection of the most common possible interactions. Furthermore, this section explains how to decide which variables in the model should be transformed.

How is the model selection process under transformations?

The strategy for selecting the working model under different transformation is still under discussion. Sakia (1992) states "The selection of a transformation may be properly viewed as model selection". However, comparing regression models for variable selection under different scale levels has some difficulties. The model selection criterion should be invariant to a change of scale in the target variable, which is not the case for the Akaike information criterion (AIC) or the Bayesian information criterion (BIC), two commonly used information criteria for the linear and linear mixed regression models (Burnham and Anderson, 2004; Müller et al., 2013). Therefore, the coefficients of determination and their extensions to the linear mixed regression models are a first approximation for comparing the models in terms of general fitting, since they are scale invariant. Additionally, the working model always depends on which procedure is done first, variable or transformation selection. Some procedures that have been implemented for the linear regression model include the combination of these two procedures in one (Laud and Ibrahim, 1995; Hoeting and Ibrahim, 1998; Hoeting et al., 2002).

How should transformations be used in the presence of outliers?

Without loss of generality an outlier is defined as an atypical observation among a data set, which can be representative or non-representative (Chambers, 1986). A discussion of the definition of outlying observations for the linear mixed regression models can be find in Bell and

Huang (2006) and Warnholz (2016b). Outliers are not themselves a violation of model assumptions. However, their presence could induce skewed distributions and heteroscedasticity that lead to problems already examined in Section 1.2. Simply excluding an outlier is not always the right answer since they may contain valuable information about the distribution of our data (Belsley et al., 2005). If the presence of an outlier has a disproportionate influence on the estimated model, an analysis with and without such observation is usually recommended.

Carroll and Ruppert (1985) state that proper transformation decision and identification of outlying observations are interconnected. Thus, Figure 1.1 summarizes the stages in the analysis where the detection and the handling of outliers interconnects with the usage of a transformation. For the detection of outliers, scatterplots between the outcome variable and the explanatory variables or a box plot of the outcome variable can be sufficient. More methodologies can be found for instance in Cook and Weisberg (1982) and Barnett and Lewis (1984). The most popular measures of influence are the Cook's distance (Cook, 1977), the Welsch and Kuh measure (Belsley et al., 2005) and the Hadi's Influence Measure (Hadi, 1992). In the case that the use of a transformation seems suitable after checking model assumptions and outliers are detected, a sensitivity analysis is suggested. This includes finding out if the outlying case in the original scale is also an outlying observation in the transformed scaled. Furthermore, it is important to have an idea about how these observations can influence the need or utility of a transformation. For instance, if the outliers cause heteroscedasticity, the deletion of the outlier could make the usage of the transformation unnecessary. Some diagnostics for studying the contribution of single observations on the need of transformations are presented in Cheng (2005) and Atkinson and Riani (2012). A sensitivity analysis under a Box-Cox power transformation model has been discussed by Bickel and Doksum (1981); Box and Cox (1982); Hinkley and Runger (1984); Atkinson (1986) and Duan (1993). Atkinson (1986) proposes a sensitivity analysis by eliminating outlying observations after applying a Box-Cox transformation. Atkinson (1982) studied the reduction of influential cases and outliers after applying transformations in some examples. However, Cook and Wang (1983) proposed a method to detect influential observations under the Box-Cox transformation that is superior to the method of Atkinson (1982) (Cook and Wang, 1983; Sakia, 1992). Tsai and Wu (1990) and Kim et al. (1996) studied the influence on the Jacobian of the transformation when single observations are deleted. If the outliers influence the need of the transformation, different methods are suitable to treat the outliers (Hawkins, 1980; Cook and Prescott, 1981; Cook and Weisberg, 1982; Cook and Wang, 1983; Barnett and Lewis, 1984; Hawkins et al., 1984). In a model context and for the estimation process, different procedures have been proposed: model reformulations, downweighting outlying observations (Rousseeuw and Leroy, 2005), use of the winsorization method (Yale and Forsythe, 1976) and use extreme-value distributions (Withers and Nadarajah, 2007). Furthermore, there has been considerable growing interest in using robust techniques in recent years for incorporating this effect into the model structure and fitting or bounding outliers and influential observations (Huber, 1964; Krasker and Welsch, 1982; Hampel et al., 1986; Rousseeuw and Van Zomeren, 1990). For instance, the M-estimation (Huber, 1964) and the least trimmed squares (Anscombe and Guttman, 1960) are examples of robust models that can be used when outliers are present in the data. As alternatives, it is common in practice to

use Bayesian methods (Gelman et al., 2014) and quantile regression (Koenker, 2005).

In the other case, transformations can be useful in the presence of outliers since all information can be kept in the data set and, at the same time, skewness and error variance can be reduced (Osborne and Overbay, 2004). Furthermore, the rank transformation (Conover and Iman, 1981) replaces the data for their corresponding ranks, and it can be seen as an outlying observations handling. When a transformation is used without a previous outliers treatment, it is recommended to use robust methods for the estimation of the transformation parameters because the maximum likelihood theory is sensitive to outliers. In particular, Carroll (1980) Carroll (1982b) , Carroll and Ruppert (1985), Bickel and Doksum (1981) and most recently Marazzi and Yohai (2004) propose different robust methods for the Box-Cox transformation and Burbidge et al. (1988) for the inverse hyperbolic sine transformation parameters. These approaches are concerned with a modified likelihood function (see e.g. Krasker and Welsch (1982)). Gottardo and Raftery (2009) developed a Bayesian estimation method for the Box-Cox transformation that accounts for outlying values. Pericchi (1981) and Sweeting (1984) study different choices of prior distributions for the Box-Cox linear model. For the same model, Shin (2008) develops a semi-parametric estimation method. Note that all mentioned methods only handle outliers in the outcome variable.

How do incomplete responses affect the usage of transformations?

The problem of missing data becomes a fundamental part of almost every research setting.

Rubin (1976) introduced a classification system of missing data which describes the probability of missing values in relation to the data. The missing data mechanisms are missing completely at random (MCAR), missing at random (MAR) and missing not at random data (MNAR). The effects of missing responses in the data set on the usage of transformations have not yet been extensively studied. However, if it is reasonable that the missing data mechanism is MAR, the missing values can be ignored and maximum likelihood theory can be used in combination with a transformation (Rubin, 1976; Lipsitz et al., 2000).

How can the transformations be used when the data is multimodal?

The transformations presented in this paper most likely do not ensure the correction of assumptions when data is multimodal. For instance, a specific variable (e.g. gender or income) can generate different groups in the data with distinct distributions (Bradley, 1977). Therefore, before an appropriate transformation is selected, this effect should be removed or corrected by including a factor as an explanatory variable in the regression model. After this, the residuals should be unimodal. This conventional technique in general modelling theory is also a type of transformation (Fink, 2009).

How does the range of the variable limit the choice of transformations?

One of the most important features that we have to know when choosing a transformation is the range of the variable. Most of the transformations are not mathematically defined for zero or negative values. In order to deal with this problem, three general solutions regarding the use

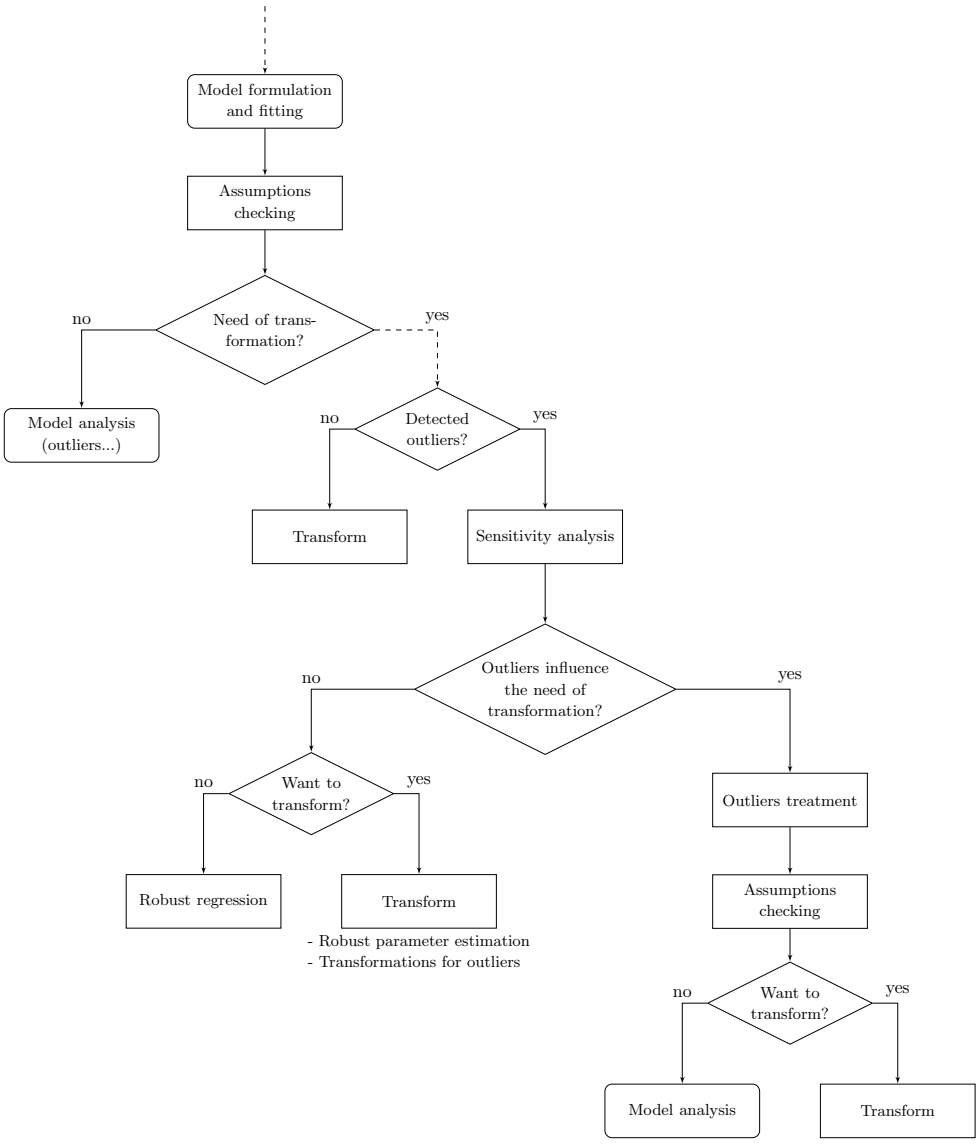


Figure 1.1: A guide how to handle the interactions between transformations and outliers

of transformations have been published on this topic. Firstly, the researcher can shift the data with a fixed constant (usually equal to one) or a fixed parameter that makes the data positive. However, using an arbitrary parameter for making the data positive affects the analysis results (Fletcher et al., 2005). Osborne (2002) suggests that adding a constant to the outcome variable only changes the mean and not the other moments of the distribution, and he recommends its use. Atkinson (1987) dedicates a whole chapter to discussing the implications of using this family of transformations with a shifted parameter on model fitting, in particular in the constant parameter and the estimation transformation parameter. Additionally, Hill (1963) and Yeo and Johnson (2000) suggest that asymptotic results of maximum likelihood theory may not hold including a shift parameter. Therefore, the second solution is to use a transformation that includes in its functional form the possibility of using negative and non-negative responses. Finally, Burbidge and Robb (1985) propose to shrink any zero values toward forward zero, while holding the rest constant and applying the maximum likelihood theory.

How are the effects of many zeros in the variable on the transformation?

Data containing a substantial proportion of zeros is commonly known as a zero inflation problem or as an excess zeros problem. If this phenomenon is not correctly handled, the relation between the conditional variance and the dependent variable is not equal, but greater. This problem is called overdispersion and this can lead to an underestimation of the standard errors. Furthermore, when the zero inflated problem is present, transformations may not be applicable to achieve linearity. Another typical situation occurs when changing negative values in the outcome variable to numbers close to zero. Magee (1988) studied the effects of this change in the outcome variable for the Box-Cox transformation. In this case, the Jacobian of the transformation (see estimation methods in Section 1.2) usually tends to plus or minus infinity (MacKinnon and Magee, 1990) and the transformation parameter tends to be also zero. Furthermore, if a Box-Cox transformation is applied under this condition, the transformed variable will be bounded from below, which is not optimal, if the aim is to deal with non Gaussian assumptions.

How can we decide which variables should be transformed?

Mosteller and Tukey (1977) propose a ladder of transformations to guide the selection of a transformation that helps to fulfill the linearity assumption (see Section 1.2). If it becomes evident that a serious problem in the residuals is present, a transformation in the dependent variable is suggested. Otherwise, if the residuals are well behaved, transforming the outcome can artificially lead to a violation of assumptions, especially to heteroscedasticity. In this case, one or more of the explanatory variables should be transformed (Cohen et al., 2014; Brown, 2015). Box and Tidwell (1962) suggest the use of a power transformation in the explanatory variables in order to linearize the relationship with the outcome variable. The method is known as the Box-Tidwell transformation and seeks to find the optimal transformation parameter under a Box-Cox transformation for each variable that can be transformed (e.g not for dummy variables). The estimation process is based on maximum likelihood theory and is iterated until convergence. Furthermore, it does not affect the variance stabilization and the Gaussian assumptions of the error term distribution (Box and Cox, 1964).

It is also possible to transform both sides of the regression model. This can be useful when there is a fair certainty that the regression model already describes well the studied interaction, but the assumptions over the error terms are not yet met. In this case, a transformation family T can be applied on both sides of the equation, which leads to the transform both sides (TBS) model:

$$T(y_i) = T\big[f(x_i, \beta), \lambda\big] + e_i,$$

and for $f(\cdot)$, the error terms are usually assumed to be additive. The transformation function may take different functional forms. It can simply be the logarithmic transformation, but can also be a more elaborate family of power transformations, such as the Box-Cox. For instance, when the logarithmic transformation is applied on both sides, the level-level regression specification is known as log-log transformation. If done properly, transforming both sides makes the estimation of β more efficient (Carroll and Ruppert, 1988). If the transformation relies on a transformation parameter, adjustments for the estimation of this parameter are suggested. Regarding the Box-Cox transformation, Carroll and Ruppert (1988) propose writing the maximum likelihood function in terms of β, σ_e^2 and λ, and then maximizing it by employing an optimization technique such as a the Newton algorithm. As they also acknowledge, it is not always possible to carry out this procedure as it can become computationally expensive. Carroll and Ruppert (1988) suggest two alternatives. One of them is known as the profile likelihood, which is based on the same theory proposed in Box and Cox (1964). The second method is the use of the pseudo-regression model. In terms of parsimony, Carroll and Ruppert (1988) favor the use of the pseudo-regression model method over the profile likelihood. However, the pseudo-regression model method can have irremediable convergence problems, and when that happens the profile likelihood method is more reliable. Further estimation methods for the TBS method are also studied by Ruppert and Aldershof (1989), Kettl (1991), Nychka and Ruppert (1995) and Wang and Ruppert (1995). In order to calculate standard errors, Carroll and Ruppert (1988) classify six techniques according to the estimation method employed for σ_e^2, λ and β and which model is fitted to the data.

1.4 Conclusions and Future Research Directions

As this review of transformations shows, the application of transformations is a helpful tool for achieving model assumptions for the linear and linear mixed regression models. In this work, special attention has been paid to the wide range of transformations useful for achieving model assumptions and estimation methods that can be used for the estimation of transformations parameters. We explored the implications of these assumptions, their importance, and the consequences of their violation in terms of estimation and inference. Moreover, an attempt was made to present possible solutions to correct in the case that any of these assumptions is violated. By doing so we showed that transformations can work as a solution for some of these violations; particularly, for non-normality, heteroscedasticity, and non-linearity. In order to combat the misuse of transformations, this work also provides a guide for the correct and thoughtful application.

Because an increasing number of researchers are using the linear and linear mixed regression models, more theory of transformations for these models should be developed in future. For instance, one drawback of transformations is still the interpretation of model results. Interpreting estimations in the transformed scale is not always desired, and most researchers prefer to take decisions on the original scale. Manning (1998) summarized this issue by pointing out that "First Bank will not cash a check for log dollars". Therefore, further research is needed to investigate the bias of back-transforming into the original scale an the interpretation of model results under transformations. Nonetheless, these limitations should be seen as future opportunities. Finally, more effort should be put into the comparison of different estimations under diverse data circumstances.

Chapter 2

The R Package trafo for Transforming Linear Regression Models

2.1 Introduction

To study the relation between two or more variables, the linear regression model is one of the most employed statistical methods. For an appropriate usage of this model, a set of assumptions needs to be fulfilled. These assumptions are, among others, related to the functional form and to the error terms, such as linearity and homoscedasticity. However, in practical applications, these assumptions are not always satisfied. This leads to the question of how the practitioner can move on with the analysis in such case. One way to proceed is to conduct the analysis ignoring the model assumption violations which is, of course, not recommended as it would likely yield misleading results. Another solution is to use more complex methods such as generalized linear regression models or non-parametric methods, as they might fit the data and problem better. A third method which also constitutes the focus of the present paper is the application of suitable transformations. Transformations have the potential to correct certain violations and by doing so, enable to continue the analysis with the known (linear) regression model. Due to its convenience, transformations such as the logarithm or the Box-Cox are commonly applied in many branches of sciences; for example in economics (Hossain, 2011) and neuroscience (Morozova et al., 2016). In order to simplify the choice and the usage of transformations in the linear regression model, the R (R Core Team, 2017) package **trafo** (Medina et al., 2017) is developed. The present work is inspired by the framework proposed in Rojas-Perilla et al. (2017) and extends other existing R packages that provide transformations.

Many packages that contain transformations do not focus especially on the usage of transformations (Venables and Ripley, 2002; Fox and Weisberg, 2011; Molina and Marhuenda, 2015; Ribeiro Jr. and Diggle, 2016; Fife, 2017). They often only include popular transformations like the logarithmic or the Box-Cox transformation family. The package **car** (Fox and Weisberg, 2011) expands the selection of transformations. It includes the Box-Cox, the basic power, and the Yeo-Johnson transformation families, and uses the maximum likelihood approach for the estimation of the transformation parameter. An exponential transformation proposed by Manly (1976) is provided in the package **caret** (Kuhn, 2008) and the multiple parameter Johnson transformation in the packages **Johnson** (Fernandez, 2014) and **jtrans** (Wang,

2015). While package **MASS** (Venables and Ripley, 2002) and package **car** (Fox and Weisberg, 2011) only provide the maximum likelihood approach for the estimation of the transformation parameter for the Box-Cox family, the estimation can be conducted by a wide range of methods in the **AID** package (Dag et al., 2017). Most of the provided methods are based on goodness of fit tests like the Shapiro-Wilk or the Anderson-Darling test. However, the **AID** package only contains the Box-Cox transformation.

It is noticeable that none of the above-mentioned packages helps the user in the process of deciding which transformation is actually suitable according to his needs. Furthermore, most packages do not provide tools to see at the first sight if the transformation improves the untransformed model. Therefore, package **trafo** combines and extends the features provided by the packages mentioned above. Additionally to transformations that are already provided by existing packages, the **trafo** package includes, among others, the Bickel-Doksum (Bickel and Doksum, 1981), Modulus (John and Draper, 1980), the neglog (Whittaker et al., 2005) and glog (Durbin et al., 2002) transformations that are modifications of the Box-Cox and the logarithmic transformation, respectively, in order to deal with negative values in the response variable. Furthermore, the selection of estimation methods for the transformation parameter is enlarged by methods based on moments and divergence measures. The main benefits of the package **trafo** can be summarized as follows:

- An initial check can be conducted that helps to decide if and which transformation is useful for the researchers needs.
- The untransformed model and a model with a transformed dependent variable as well as two transformed models can be run simultaneously, and thus the models can be easily compared with regard to the model assumptions.
- Extensive diagnostics are provided in order to check if the transformation helps to fulfill the model assumptions normality, homoscedasticity, and linearity.

The remainder of this paper is structured as follows. In Section 2.2, the transformations included in the package are presented. Section 2.3 demonstrates in form of a case study the functionality of the package. Section 2.4 summarizes the user-defined function feature of the package. In Section 2.5, some concluding remarks and potential extensions of the package are discussed. Finally, Appendix .1 presents the mathematical derivations underlying the package.

2.2 Transformations and estimation methods

The equation describing and summarizing the relationship between a continuous outcome variable Y and different covariates X (either discrete or continuous) is defined by $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + e_i$, with $i = 1, \ldots, n$. This is also known as the linear regression model and is composed by a deterministic and a random component, which rely on different assumptions. Among others, these assumptions can be summarized as follows:

- Normality (N): The conditional distribution of Y given X follows a normal distribution.

- Homoscedasticity (H): The conditional variance of Y given X is constant.
- Linearity (L): The conditional expectation of the outcome variable Y given the continuous covariates X is a linear function in X.

As already mentioned, different approaches have been proposed for achieving these model assumptions. Some of them include using alternative estimation methods of the regression terms or applying more complex regression models. In this paper, we focus on defining a parsimonious re-specification for the model, such as the usage of non-linear transformations of the outcome variable. The transformations implemented in the package **trafo** basically help to achieve normality. However, most of them simultaneously correct other assumptions (see also Table 2.1 and Table 2.2).
The transformations can be classified into transformations without a transformation parameter and data-driven transformations with a transformation parameter that needs to be estimated. The first set of transformations presented in Table 2.1 comprises, among others, the logarithmic transformation and some variations, which is considered due to its popularity and straightforward application. The reciprocal transformation is one of the well-known ladder of powers, which is a family of power transformations (Tukey, 1977; Emerson and Stoto, 1983). The

Table 2.1: Transformations without transformation parameter

Transformation	Source	Formula	Support	N	H	L
Log (shift)	Box and Cox (1964)	$\log(y+s)$	$y \in \mathbb{R}$	✗	✗	✗
Glog	Durbin et al. (2002)	$\log(y+\sqrt{y^2+1})$	$y \in \mathbb{R}$	✗	✗	✗
Neglog	Whittaker et al. (2005)	$\text{Sign}(y)\log(\lvert y\rvert+1)$	$y \in \mathbb{R}$	✗	✗	
Reciprocal	Tukey (1977)	$\frac{1}{y}$	$y \neq 0$	✗	✗	

data-driven transformations presented in Table 2.2 are dominated by the Box-Cox transformation and its modifications or alternatives, e.g. the modulus or Bickel-Doksum transformation. However, more flexible versions of the logarithmic transformation, as the log-shift opt or the Manly transformation which is an exponential transformation, are also included in the package **trafo**.

Both tables provide information about the range of the dependent variable that is supported by the transformation. Some transformations are only suitable for positive values of y. This is generally true for the logarithmic and Box-Cox transformations. However, in case that the dependent variable contains negative values, the values are shifted by a deterministic shift s such that $y + s > 0$ by default in package **trafo**. Furthermore, the tables emphasize which assumptions the transformation helps to achieve. These are general suggestions and the actual success always also depends on the data. For specific properties of each transformation we refer to the original references.
Since the transformations in Table 2.2 contain transformation parameters that need to be estimated, package **trafo** contains different methodologies for this estimation. The benefit of each estimation method depends on the research analysis and the underlying data. They can be summarized as follows:

Table 2.2: Data-driven transformations

Transformation	Source	Formula	Support	N	H	L
Box-Cox (shift)	Box and Cox (1964)	$\begin{cases} \frac{(y+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y+s) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	✗
Log-shift opt	Feng et al. (2016)	$\log(y + \lambda)$	$y \in \mathbb{R}$	✗	✗	✗
Bickel-Docksum	Bickel and Doksum (1981)	$\frac{\lvert y\rvert^\lambda \text{Sign}(y) - 1}{\lambda}$ for $\lambda > 0$	$y \in \mathbb{R}$	✗	✗	
Yeo-Johnson	Yeo and Johnson (2000)	$\begin{cases} \frac{(y+1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, y \geq 0; \\ \log(y+1) & \text{if } \lambda = 0, y \geq 0; \\ \frac{(1-y)^{2-\lambda} - 1}{\lambda - 2} & \text{if } \lambda \neq 2, y < 0; \\ -\log(1-y) & \text{if } \lambda = 2, y < 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Square Root (shift)	Emerson and Stoto (1983)	$\sqrt{y+s}$	$y \in \mathbb{R}$	✗	✗	
Square Root (shift)	as Rojas-Perilla et al. (2017)	$\sqrt{y+\lambda}$	$y \in \mathbb{R}$	✗	✗	
Manly	Manly (1976)	$\begin{cases} \frac{e^{\lambda y} - 1}{\lambda} & \text{if } \lambda \neq 0; \\ y & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗	✗	
Modulus	John and Draper (1980)	$\begin{cases} \text{Sign}(y) \frac{(\lvert y\rvert + 1)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \text{Sign}(y) \log(\lvert y\rvert + 1) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		
Dual	Yang (2006)	$\begin{cases} \frac{(y^\lambda - y^{-\lambda})}{2\lambda} & \text{if } \lambda > 0; \\ \log(y) & \text{if } \lambda = 0. \end{cases}$	$y > 0$	✗		
Gpower	Kelmansky et al. (2013)	$\begin{cases} \frac{(y+\sqrt{y^2+1})^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y + \sqrt{y^2+1}) & \text{if } \lambda = 0. \end{cases}$	$y \in \mathbb{R}$	✗		

- Maximum likelihood theory
- Distribution moments optimization: Skewness or kurtosis
- Divergence minimization: Following Kolmogorov-Smirnov (KS), Cramér-von-Mises (KM) or Kullback-Leibler (KL) measurements

The maximum likelihood estimation method finds the set of values for the transformation parameter that maximizes the likelihood function of the dataset under the selected transformation. This is a standard approach that is also implemented in several of the mentioned R packages (Venables and Ripley, 2002; Fox and Weisberg, 2011). However, since the maximum likelihood estimation is rather sensitive to outliers, the skewness or kurtosis optimization might be preferable for the estimation of the transformation parameter in the presence of such outliers. These methods are especially favorable when it is important in the analysis to meet these moments. For instance, skewness minimization should be used when it is important to get a symmetric distribution. Additionally, if the focus lies on comparing the whole distribution of the transformed data with a normal distribution, and not only some moments, different divergence measures as the KS, KM or KL can be used. For all estimation methods a lambda range on which the functions are evaluated needs to be proposed. Therefore, default values are set for the predefined transformations.

Since the user can only decide if the transformation is helpful by checking the above mentioned assumptions, the package **trafo** contains a wide range of diagnostic checks. A smaller selection is used in the fast check that helps to decide if a transformation might be useful. Table 3 summarizes the implemented diagnostic checks that are simultaneously returned for the untransformed and a transformed model or two differently transformed models and indicates

Table 2.3: Diagnostic checks provided in the package **trafo**

Assumption	Diagnostic check	Fast check
Normality	Skewness and kurtosis	X
	Shapiro-Wilk test	X
	Quantile-quantile plot	
	Histograms	
Homoscedasticity	Breusch-Pagan test	X
	Residuals vs. fitted plot	
	Scale-location	
Linearity	Scatter plots between y and x	X
	Observed vs. fitted plot	

which diagnostics are conducted in the fast check. Additionally, plots are provided that help to detect outliers such as the Cook's distance plot and influential observations by the residuals vs leverage plot.
Another feature of the package **trafo** is the possibility of defining a customized transformation. Thus, a user can also use the infrastructure of the package for a transformation that suits the individuals needs better than the predefined transformations. However, in this version of the package **trafo** the user needs to define the transformation and the standardized transformation in order to use this feature.

2.3 Case Study

In order to show the functionality of the package **trafo**, we present in form of a case study the steps a user faces when checking the assumptions of the linear model. For this illustration, we use the data set called `University` from the R package **Ecdat** (Croissant, 2016). This data set contains variables about the equipment and costs of university teaching and research and can be obtained as follows:

```
R> library(Ecdat)
R> data(University)
```

A practical question for the head of a university could be how study fees (`stfees`) raise the universities net assets (`nassets`). Both variables are metric. Thus, a linear regression could help to explain the relation between these two variables. A linear regression model can be conducted in R using the `lm` function.

```
R> linMod <- lm(nassets ~ stfees, data = University)
```

The features in the package **trafo** that help to find a suitable transfomation for this model and to compare different models are summarized in Table 2.4 and illustrated in the next subsections.

2.3.1 Finding a suitable transformation

It is well known that the reliability of the linear regression model depends on assumptions. Amongst others, normality, homoscedasticity, and linearity are assumed. In this section, we

Table 2.4: Core functions of package **trafo**

Function	Description
`assumptions()`	Enables a fast check which transformation is suitable.
`trafo_lm()`	Compares the untransformed model with a transformed model.
`trafo_compare()`	Compares two differently transformed models.
`diagnostics()`	Returns information about the transformation and different diagnostics checks in form of tests.
`plot()`	Returns graphical diagnostics checks.

focus on presenting how the user can decide and assess, if and which, transformations help to fulfill these model assumptions. Thus, a first fast check of these model assumptions can be used in the package **trafo** in order to find out if the untransformed model meets these assumptions or if using a transformation seems suitable. The fast check can be conducted by the function `assumptions`. This function returns the skewness, the kurtosis and the Shapiro-Wilk test for normality, the Breusch-Pagan test for homoscedasticity and scatter plots between the dependent and the explanatory variables for checking the linear relation. All possible arguments of the function `assumptions` are summarized in Table 2.5. In the following, we only show the returned normality and homoscedasticity tests. The results are ordered by the highest p value of the Shapiro-Wilk and Breusch-Pagan test.

```
R> assumptions(linMod)

The default lambdarange for the log shift opt transformation
is calculated dependent on the data range. The lower value is
set to -2035.751 and the upper value to 404527.249

The default lambdarange for the square root shift
transformation is calculated dependent on the data range. The
lower value is set to -2035.751 and the upper value to
404527.249

Test normality assumption
              Skewness Kurtosis Shapiro_W Shapiro_p
logshiftopt    -0.4201   4.0576    0.9741    0.2132
boxcox         -0.4892   4.2171    0.9621    0.0527
bickeldoksum   -0.4892   4.2171    0.9621    0.0527
gpower         -0.4892   4.2171    0.9621    0.0527
modulus        -0.4892   4.2171    0.9621    0.0527
yeojohnson     -0.4892   4.2171    0.9621    0.0527
dual           -0.4837   4.2180    0.9619    0.0519
sqrtshift       0.6454   5.2752    0.9504    0.0139
log            -1.1653   5.1156    0.9140    0.0004
neglog         -1.1651   5.1150    0.9140    0.0004
```

```
glog           -1.1653   5.1156    0.9140    0.0004
untransformed   2.4503  12.7087    0.7922    0.0000
reciprocal     -3.7260  19.0487    0.5676    0.0000

Test homoscedasticity assumption
              BreuschPagan_V BreuschPagan_p
modulus               0.1035         0.7477
yeojohnson            0.1035         0.7477
boxcox                0.1035         0.7476
bickeldoksum          0.1036         0.7476
gpower                0.1035         0.7476
dual                  0.1128         0.7369
logshiftopt           0.1154         0.7341
neglog                0.7155         0.3976
log                   0.7158         0.3975
glog                  0.7158         0.3975
reciprocal            1.6109         0.2044
sqrtshift             5.4624         0.0194
untransformed         9.8244         0.0017
```

Following the Shapiro-Wilk test, the best transformation to fulfill the normality assumption is the log-shift opt transformation followed by the Box-Cox, Bickel-Doksum, gpower, modulus and Yeo-Johnson transformation. For improving the homoscedasticity assumption, all transformations help except the square root (shift) transformation. As mentioned before, default values for the lambda range for all transformations are predefined and these are used in this fast check. Since the default values for the log-shift opt and square root (shift) transformation depend on the range of the response variable, the chosen range is reported in the return. The Manly transformation is not in the list since the default lambda range for the estimation of the transformation parameter is not suitable for this data set. For such a case, the user can change the lambda range for the transformations manually. Similarly, the user can change the estimation methods for the transformation parameter. For instance, if symmetry is of special interest for the user the skewness minimization might be a better choice than the default maximum likelihood method. In this study case all assumptions are equally important. Thus, we choose the Box-Cox transformation for the further illustrations even though some other transformations would be suitable as well.

2.3.2 Comparing the untransformed model with a transformed model

For a more detailed comparison of the transformed model with the untransformed model, a function called `trafo_lm` (for the arguments see Table 2.6) can be used as follows:

```
R> linMod_trafo <- trafo_lm(linMod)
```

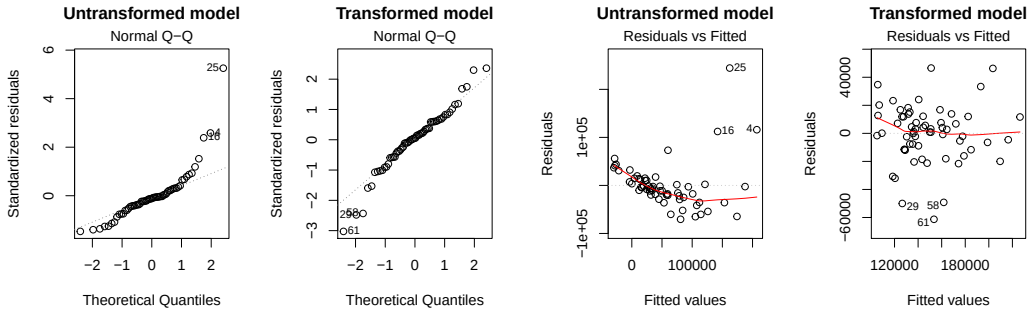
Table 2.5: Arguments of function `assumptions`

Argument	Description	Default
`object`	Object of class `lm`.	
`method`	Estimation method for the transformation parameter.	Maximum likelihood
`std`	Normal or scaled transformation.	Normal
`...`	Addtional arguments can be added, especially for changing the lambda range for the estimation of the parameter, e.g. `manly_lr = c(0.000005,0.00005)`	Default values of lambda range of each transformation

The Box-Cox transformation is the default option such that only the `lm` object needs to be given to the function. The object `linMod_trafo` is of class `trafo_lm` and the user can conduct the methods `print`, `summary` and `plot` in the same way as for an object of class `lm`. The difference is that the new methods simultaneously return the results for both models, the untransformed model and the transformed model. Furthermore, a method called `diagnostics` helps to compare results of normality and homoscedasticity tests. In the following, we will show the return of the `diagnostics` method and some selected plots in order to check the normality, homoscedasticity and the linearity assumption of the linear model.

```
R> diagnostics(linMod_trafo)

Diagnostics: Untransformed vs transformed model

Transformation:  boxcox
Estimation method:  ml
Optimal Parameter:  0.1894257

Residual diagnostics:

Normality:
Pearson residuals:
                     Skewness  Kurtosis Shapiro_W    Shapiro_p
Untransformed model 2.4503325 12.708681 0.7921672 6.024297e-08
Transformed model  -0.4892222  4.217105 0.9620688 5.267566e-02

Heteroscedasticity:
                   BreuschPagan_V BreuschPagan_p
Untransformed model      9.8243555     0.00172216
Transformed model        0.1035373     0.74762531
```

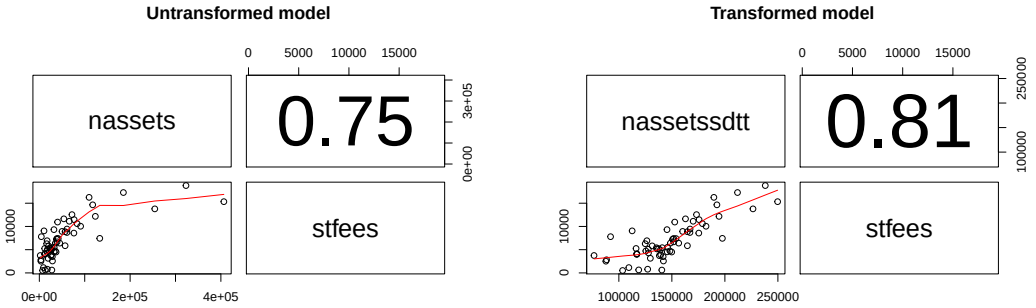
The first part of the return shows information from the applied transformation. As chosen, the Box-Cox transformation is used with the optimal transformation parameter around 0.19 which is estimated using the maximum likelihood approach that is also set as default. The optimal transformation parameter differs from 0, which would be equal to the logarithmic transformation, and 1, which means that no transformation is optimal. The Shapiro-Wilk



(a) Normal Q-Q plots of the error terms.

(b) Residuals versus fitted.

Figure 2.1: Selection of diagnostic plots obtained by using `plot(linMod_trafo)`. (a) shows Normal Q-Q plots error terms of the untransformed and the transformed model. (b) shows the residuals against the fitted values of the untransformed and the transformed model.



(a)

(b)

Figure 2.2: Selection of obtained diagnostic plots by using `plot(linMod_trafo)`. (a) shows the scatter plot of the untransformed net assets and the study fees (b) shows scatter plot of the transformed net assets and the study fees. The numbers specify the correlation coefficient between the dependent and independent variable.

test rejects normality of the residuals of the untransformed model but it does not reject normality for the residuals of the transformed model on a 5% level of significance. Furthermore, the skewness shows that the residuals in the transformed model are more symmetric and the kurtosis is closer to 3, the value of the kurtosis of the normal distribution. The results of the Breusch-Pagan test clearly show that homoscedasticity is rejected in the untransformed model but not in the transformed model. These two findings can be supported by diagnostic plots shown in Figure 2.1.

```
R> plot(linMod_trafo)
```

In order to evaluate the linearity assumption, scatter plots of the dependent variable against the explanatory variable can help. Figure 2.2 shows that the assumption of linearity is violated in the untransformed model. In contrast, the relation between the transformed net assets and the study fees seems to be linear. As demonstrated above, the user can receive diagnostics for an untransformed and a transformed model with only a little more effort in comparison to fitting the standard linear regression model without transformation. While we only show the example

Table 2.6: Arguments of function `trafo_lm`

Argument	Description	Default
`object`	Object of class `lm`.	
`trafo`	Selected transformation.	Box-Cox
`lambda`	Estimation or a self-selected numeric value.	Estimation
`method`	Estimation method for the transformation parameter.	Maximum likelihood
`lambdarange`	Determines lambdarange for the estimation of the transformation parameter.	Default lambdarange for each transformation.
`std`	Normal or scaled transformation.	Normal
`custom_trafo`	Add customized transformation.	None

with the default transformation, the user can also easily change the transformation and the estimation method. For instance, the user could choose the log-shift opt transformation with the skewness minimization as estimation method.

```
R> linMod_trafo2 <- trafo_lm(object = linMod, trafo =
+    "logshiftopt", method = "skew")
```

2.3.3 Compare two transformed models

The user can also compare different transformations with regard to meet the model assumptions. In many present-day applications, the logarithm is often used without longer considerations about its usefulness. In order to compare the logarithm, e.g., with the selected Box-Cox transformation, the user needs to specify two objects of class `trafo` as follows:

```
R> boxcox_uni <- boxcox(linMod)
R> log_uni <- logtrafo(linMod)
```

The utility of `trafo` objects is twofold. First, the user can use the functions for each transformation in order to simply receive the transformed vector. The `print` method gives first information about the vector and the method `as.data.frame` returns the whole data frame with the transformed variable in the last column. The variable is named as the dependent variable with an added t.

```
R> head(as.data.frame(boxcox_uni))

    nassets stfees nassetst
1   3669.71   2821 19.71248
2  12156.00   4037 26.07723
3 185203.00  17296 47.24867
4 323100.00  18800 53.08840
5  32154.00   9314 32.42140
6  41669.00   7388 34.31882
```

Second, the objects can be used to compare linear models with differently transformed dependent variable using function `trafo_compare`. The arguments of this functions are shown in Table 2.7. The user creates an object of class `trafo_compare` by:

Table 2.7: Arguments of function `trafo_compare`

Argument	Description	Default
`object`	Object of class `lm`.	
`trafos`	List of objects of class `trafo`.	
`std`	Normal or scaled transformation.	Normal

```
R> linMod_comp <- trafo_compare(object = linMod,
+   trafos = list(boxcox_uni, log_uni))
```

For this object, the user can use the same methods as for an object of class `trafo_lm`. In this work, we only want to show the return of method `diagnostics`.

```
R> diagnostics(linMod_comp)

Diagnostics of two transformed models

Transformations:  Box-Cox and Log
Estimation methods:  ml  and no estimation
Optimal Parameters:  0.1894257  and no parameter

Residual diagnostics:

Normality:
Pearson residuals:
          Skewness Kurtosis Shapiro_W    Shapiro_p
Box-Cox -0.4892222 4.217105 0.9620688 0.0526756632
Log     -1.1653028 5.115615 0.9140135 0.0003534879

Heteroscedasticity:
        BreuschPagan_V BreuschPagan_p
Box-Cox      0.1035373      0.7476253
Log          0.7158162      0.3975197
```

The first part of the return points out that the Box-Cox transformation is a data-driven transformation with a transformation parameter while the logarithmic transformation does not adapt to the data. Furthermore, we can see that normality is rejected for the model with a logarithmic transformed dependent variable, while it is not rejected when the Box-Cox transformation is used. The violation of the homoscedasticity assumption can be fixed by both transformations.

2.4 Customized transformation

An additional user-friendly feature in the package **trafo** is the possibility of using the framework also for self-defined transformations. In the following we show this option for the glog transformation.

In a first step, the transformation and the standardized or scaled transformation need to be defined. The mathematical expression of these two functions is presented in the Appendix .1.2.

```
R> glog_trafo <- function(y) {
+   yt <-  log(y + sqrt(y^2 + 1))
+   return(y = yt)}

R> glog_std <- function(y) {
+   zt <-  log(y + sqrt(y^2 + 1)) *
+   sqrt(geometric.mean(1 + y^2))
+   return(zt = zt)}
```

Second, the user inserts the two functions as a list argument to the `trafo_lm` function. Furthermore, the user needs to specify for the `trafo` argument if the transformation is without a parameter (`"custom_wo"`) or with one parameter (`"custom_one"`). The glog transformation does not rely on a transformation parameter.

```
R> linMod_custom <- trafo_lm(linMod, trafo = "custom_wo",
+   custom_trafo = list(glog_trafo = glog_trafo,
+   glog_std = glog_std))
```

One limitation of this feature is the necessity to insert both the transformation and the scaled transformation since the latter is often not known. Furthermore, the framework is only suitable for transformations without and with only one transformation parameter.

2.5 Conclusions and Future Research Directions

Even though the development in computing enables the use of complex methods nowadays, transformations are still a parsimonious way to meet model assumptions in a linear regression model. In the Section 2.3, we demonstrated how the package **trafo** helps the user to decide easily if and which transformation is suitable to fulfill the model assumptions normality, homoscedasticity and linearity. To the best of our knowledge **trafo** is the only R package that supports this decision process. Furthermore, the package **trafo** provides an extensive collection of transformations usable in linear regression models and a wide range of estimation methods for the transformation parameter. In future versions, we plan to enlarge this collection constantly, also for other types of data, e.g, count data. Additionally, more of the methods that are available for the class `lm` could be developed for objects of class `trafo_lm`. We would also like to expand the infrastructure for linear mixed regression models.

Appendices

.1 Likelihood Derivation of the Transformations

.1.1 Log (shift) transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^*. In order to obtain z_i^*, the log (shift) transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{LS}}$. Therefore, the Jacobian, the scaled, and the inverse of the log (shift) transformation are given bellow.

The log (shift) transformation presented in Table 2.1 is defined as:

$$y_i^* = \log(y_i + s).$$

In case, the shifted and fixed parameter s would not be necessary, the standard logarithm function (logarithmic transformation with $s = 0$) is applied.

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{LS} = \left[\prod_{i=1}^{n} y_i + s\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^{n} \frac{1}{y_i + s} \\ &= \bar{y}_{LS}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \log(y_i + s)\bar{y}_{LS}.$$

The inverse function of the log (shift) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \log(y_i + s) \\ x_i &= \log(y_i + s) \\ y_i &= e^{x_i} - s \\ \Rightarrow f^{-1}(y_i) &= e^{y_i} - s. \end{aligned}$$

.1.2 Glog transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^*. In order to obtain z_i^*, the glog transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{GL}}$. Therefore, the Jacobian, the scaled, and the inverse of the glog transformation are given bellow.

The glog transformation presented in Table 2.1 is defined as:

$$y_i^* = \log\left(y_i + \sqrt{y_i^2 + 1}\right) \text{ if } \lambda = 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{GL} = \left[\prod_{i=1}^{n} 1 + y_i^2\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as this defined for the inverse hyperbolic sine (arsinh) function:

$$\begin{aligned}
J(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\
&= \prod_{i=1}^{n} \frac{1}{y_i + \sqrt{y_i^2 + 1}} \left(1 + \frac{2y_i}{2\sqrt{y_i^2 + 1}}\right) \\
&= \prod_{i=1}^{n} \frac{1}{y_i + \sqrt{y_i^2 + 1}} \left(\frac{y_i + \sqrt{y_i^2 + 1}}{\sqrt{y_i^2 + 1}}\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{y_i^2 + 1}} \\
&= \bar{y}_{GL}^{\frac{-n}{2}}.
\end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \log\left(y_i + \sqrt{y_i^2 + 1}\right) \bar{y}_{GL}^{\frac{1}{2}}.$$

The inverse function of the glog transformation is denoted as:

$$\begin{aligned}
f(y_i) &= \log\left(y_i + \sqrt{y_i^2 + 1}\right) \\
x_i &= \log\left(y_i + \sqrt{y_i^2 + 1}\right) \\
e^{x_i} - y_i &= \sqrt{y_i^2 + 1} \\
(e^{x_i} - y_i)^2 &= y_i^2 + 1 \\
e^{x_i^2} - 2e^{x_i} y_i &= 1 \\
y_i &= -\frac{(1 - e^{x_i^2})}{2e^{x_i}} \\
\Rightarrow f^{-1}(y_i) &= -\frac{(1 - e^{y_i^2})}{2e^{y_i}}.
\end{aligned}$$

.1.3 Neglog transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^*. In order to obtain z_i^*, the scaled neglog transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{NL}}$. Therefore, the Jacobian, the scaled, and the inverse of the neglog transformation are given bellow.

The neglog transformation presented in Table 2.1 is defined as:

$$y_i^* = \text{sign}(y_i) \log\left(|y_i| + 1\right).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{NL} = \left[\prod_{i=1}^{n}(|y_i| + 1)\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^{n} \text{sign}(y_i) \frac{1}{|y_i| + 1} \\ &= \text{sign}\left(\prod_{i=1}^{n} y_i\right)\left(\prod_{i=1}^{n} |y_i| + 1\right)^{-1} \\ &= \text{sign}\left(\prod_{i=1}^{n} y_i\right)\bar{y}_{NL}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = \text{sign}(y_i) \log\left(|y_i| + 1\right) \text{sign}\left(\prod_{i=1}^{n} y_i\right)\bar{y}_{NL}.$$

The inverse function of the neglog transformation is denoted as:

$$\begin{aligned} f(y_i) &= \text{sign}(y_i) \log\left(|y_i| + 1\right) \\ x_i &= \text{sign}(y_i) \log\left(|y_i| + 1\right) \\ |y_i| &= e^{\text{sign}(x_i)x_i} - 1 \\ \Rightarrow f^{-1}(y_i) &= \pm\left[e^{\text{sign}(y_i)y_i} - 1\right]. \end{aligned}$$

.1.4 Reciprocal transformation

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^*. In order to obtain z_i^*, the reciprocal transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{R}}$. Therefore, the Jacobian, the scaled, and the inverse of the reciprocal transformation are given bellow.

The reciprocal transformation presented in Table 2.1 is defined as:

$$y_i^* = \frac{1}{y_i}.$$

The definition of the geometric mean is:

$$\bar{y}_R = \left[\prod_{i=1}^{n} y_i\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^{n} -\frac{1}{y_i^2} \\ &= -\bar{y}_R^{-2n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\frac{1}{y_i}\bar{y}_R^2.$$

The inverse function of the reciprocal transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{1}{y_i} \\ x_i &= \frac{1}{y_i} \\ y_i &= \frac{1}{x_i} \\ \Rightarrow f^{-1}(y_i) &= \frac{1}{y_i}. \end{aligned}$$

.1.5 Box-Cox (shift) transformation

$$y_i^*(\lambda) = \begin{cases} \frac{(y_i+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \log(y_i + s) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Box-Cox (shift) transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Box-Cox (shift)(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{BC}}$. Therefore, the Jacobian, the scaled, and the inverse of the Box-Cox (shift)(A) transformation are given bellow.

The Box-Cox (shift)(A) transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \frac{(y_i + s)^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0.$$

In case, the shifted and fixed parameter s is not necessary for making the dataset positive, the standard Box-Cox transformation (with $s = 0$) is applied.

The definition of the geometric mean is:

$$\bar{y}_{BC} = \left[\prod_{i=1}^{n} y_i + s\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\lambda(y_i + s)^{\lambda-1}}{\lambda} \\ &= \prod_{i=1}^{n} (y_i + s)^{\lambda-1} \\ &= \bar{y}_{BC}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{(y_i + s)^\lambda - 1}{\lambda} \frac{1}{\bar{y}_{BC}^{\lambda-1}}.$$

The inverse function of the Box-Cox (shift)(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ x_i &= \frac{(y_i + s)^\lambda - 1}{\lambda} \\ y_i &= (\lambda x_i + 1)^{\frac{1}{\lambda}} - s \\ \Rightarrow f^{-1}(y_i) &= (\lambda y_i + 1)^{\frac{1}{\lambda}} - s. \end{aligned}$$

Box-Cox (shift) transformation case (B)

This case is exactly equal to the log (shift) case.

.1.6 Log-shift opt transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$ to $\boldsymbol{y}_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the log-shift opt transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{LSO}}$. Therefore, the Jacobian, the scaled, and the inverse of the log-shift opt transformation are given bellow.

The log-shift opt transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \log(y_i + \lambda).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{LSO} = \left[\prod_{i=1}^{n} y_i + \lambda\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{1}{y_i + \lambda} \\ &= \bar{y}_{LSO}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \log(y_i + \lambda)\bar{y}_{LSO}.$$

The inverse function of the log-shift opt transformation is denoted as:

$$\begin{aligned} f(y_i) &= \log(y_i + \lambda) \\ x_i &= \log(y_i + \lambda) \\ y_i &= e^{x_i} - \lambda \\ \Rightarrow f^{-1}(y_i) &= e^{y_i} - \lambda. \end{aligned}$$

.1.7 Bickel-Docksum transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Bickel-Docksum transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{BD}}$. Therefore, the Jacobian, the scaled, and the inverse of the Bickel-Docksum transformation are given bellow.

The Bickel-Docksum transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \frac{|y_i|^{\lambda}\text{sign}(y_i) - 1}{\lambda} \text{ if } \lambda > 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{BD} = \left[\prod_{i=1}^{n} |y_i|\right]^{\frac{1}{n}}.$$

Therefore, the expression of the jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\text{sign}(y_i)\lambda|y_i|^{\lambda-1}}{\lambda} \\ &= \text{sign}\Big(\prod_{i=1}^{n} y_i\Big)\Big(\prod_{i=1}^{n}|y_i|\Big)^{\lambda-1} \\ &= \text{sign}\Big(\prod_{i=1}^{n} y_i\Big)\bar{y}_{BD}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{|y_i|^{\lambda}\text{sign}(y_i) - 1}{\lambda} \frac{1}{\text{sign}\Big(\prod_{i=1}^{n} y_i\Big)\bar{y}_{BD}^{(\lambda-1)}}.$$

The inverse function of the Bickel-Docksum transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{|y_i|^{\lambda}\text{sign}(y_i) - 1}{\lambda} \\ x_i &= \frac{|y_i|^{\lambda}\text{sign}(y_i) - 1}{\lambda} \\ |y_i| &= \big[\text{sign}(x_i)(x_i\lambda + 1)\big]^{\frac{1}{\lambda}} \\ \Rightarrow f^{-1}(y_i) &= \pm\big[\text{sign}(y_i)(y_i\lambda + 1)\big]^{\frac{1}{\lambda}}. \end{aligned}$$

.1.8 Yeo-Johnson transformation

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_i+1)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0, y_i \geq 0 \quad (A); \\ \log(y_i + 1) & \text{if } \lambda = 0, y_i \geq 0 \quad (B); \\ -\frac{(1-y_i)^{2-\lambda}-1}{2-\lambda} & \text{if } \lambda \neq 2, y_i < 0 \quad (C); \\ -\log(1 - y_i) & \text{if } \lambda = 0, y_i < 0 \quad (D). \end{cases}$$

Yeo-Johnson transformation case (A)

This case is exactly equal to the Box-Cox (shift) case (A), with $s = 1$.

Yeo-Johnson transformation case (B)

This case is exactly equal to the log (shift) case, with $s = 1$.

Yeo-Johnson transformation case (C)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the Yeo-Johnson(C) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modifi-

cation of the definition of the geometric mean, denoted by $\bar{y}_{\text{YC}}$. Therefore, the Jacobian, the scaled, and the inverse of the Yeo-Johnson(C) transformation are given bellow.

The Yeo-Johnson(C) transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = -\frac{(1-y_i)^{2-\lambda}-1}{2-\lambda} \text{ if } \lambda \neq 2 \text{ and } y_i < 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{YC} = \left[\prod_{i=1}^{n} 1 - y_i\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{(2-\lambda)(1-y_i)^{1-\lambda}}{2-\lambda} \\ &= \prod_{i=1}^{n} (1-y_i)^{1-\lambda} \\ &= \bar{y}_{YC}^{n(1-\lambda)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = -\frac{(1-y_{ij})^{2-\lambda}-1}{2-\lambda}\bar{y}_{YC}^{n(1-\lambda)}.$$

The inverse function of the Yeo-Johnson(C) transformation is denoted as:

$$\begin{aligned} f(y_i) &= -\frac{(1-y_i)^{2-\lambda}-1}{2-\lambda} \\ x_i &= -\frac{(1-y_i)^{2-\lambda}-1}{2-\lambda} \\ -x_i(2-\lambda) &= (1-y_i)^{2-\lambda} - 1 \\ y_i &= 1 - \big[-x_i(2-\lambda)+1\big]^{\frac{1}{2-\lambda}} \\ \Rightarrow f^{-1}(y_i) &= 1 - \big[-y_i(2-\lambda)+1\big]^{\frac{1}{2-\lambda}}. \end{aligned}$$

Yeo-Johnson transformation case (D)

Let $J(y)$ denote the Jacobian of a transformation from y_i to y_i^*. In order to obtain z_i^*, the Yeo-Johnson(D) transformation, given by $\frac{y_i^*}{J(y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{YD}}$. Therefore, the Jacobian, the scaled, and the inverse of the Yeo-Johnson(D) transformation are given bellow.

The Yeo-Johnson(D) transformation presented in Table 2.2 is defined as:

$$y_i^* = -\log(1-y_i).$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{YD} = \left[\prod_{i=1}^{n} 1 - y_i\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^{n} \frac{1}{1 - y_i} \\ &= \bar{y}_{YD}^{-n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\log(1 - y_i)\bar{y}_{YD}.$$

The inverse function of the Yeo-Johnson(D) transformation is denoted as:

$$\begin{aligned} f(y_i) &= -\log(1 - y_i) \\ x_i &= -\log(1 - y_i) \\ y_i &= -e^{-x_i} + 1 \\ \Rightarrow f^{-1}(y_i) &= -e^{-y_i} + 1. \end{aligned}$$

.1.9 Square root-shift opt transformation

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain z_i^*, the square root-shift opt transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{SR}}$. Therefore, the Jacobian, the scaled, and the inverse of the square root-shift opt transformation are given bellow.

The square root-shift opt transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \sqrt{y_i + \lambda}.$$

The definition of the geometric mean is:

$$\bar{y}_{SR} = \left[\prod_{i=1}^{n} y_i + \lambda\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian is defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*}{dy} \\ &= \prod_{i=1}^{n} -\frac{1}{2\sqrt{y_i + \lambda}} \\ &= \frac{1}{2}\bar{y}_{SR}^{\frac{-n}{2}}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^* = -\frac{1}{y_i}\bar{y}_{SR}^2.$$

The inverse function of the square root-shift opt transformation is denoted as:

$$\begin{aligned} f(y_i) &= \sqrt{y_i + \lambda} \\ x_i &= \sqrt{y_i + \lambda} \\ y_i &= x_i^2 - \lambda \\ \Rightarrow f^{-1}(y_i) &= y_i^2 - \lambda. \end{aligned}$$

.1.10 Manly transformation

$$y_i^*(\lambda) = \begin{cases} \frac{e^{\lambda y_i} - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ y_i & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Manly transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled Manly(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{M}}$. Therefore, the Jacobian, the scaled, and the inverse of the Manly(A) transformation are given bellow.

The Manly(A) transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \frac{e^{\lambda y_i} - 1}{\lambda} \text{ if } \lambda \neq 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\begin{aligned} \bar{y}_M &= \left[\prod_{i=1}^{n} e^{y_i}\right]^{\frac{1}{n}} \\ &= \left[e^{\sum_{i=1}^{n} y_i}\right]^{\frac{1}{n}} \\ &= e^{\bar{y}}. \end{aligned}$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\lambda e^{\lambda y_i}}{\lambda} \\ &= \left(\prod_{i=1}^{n} e^{y_i} \right)^{\lambda} \\ &= \bar{y}_M^{\lambda n} \\ &= e^{\lambda n \bar{y}}. \end{aligned}$$

The scaled transformation is given by:

$$\begin{aligned} z_i^*(\lambda) &= \frac{e^{\lambda y_i} - 1}{\lambda} \frac{1}{\bar{y}_M^{\lambda}} \\ &= \frac{e^{\lambda y_i} - 1}{\lambda} \frac{1}{e^{\lambda \bar{y}}}. \end{aligned}$$

The inverse function of the Manly(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{e^{\lambda y_i} - 1}{\lambda} \\ x_i &= \frac{e^{\lambda y_i} - 1}{\lambda} \\ \lambda x_i + 1 &= e^{\lambda y_i} \\ y_i &= \frac{\log(\lambda x_i + 1)}{\lambda} \\ \Rightarrow f^{-1}(y_i) &= \frac{\log(\lambda y_i + 1)}{\lambda}. \end{aligned}$$

Manly transformation case (B)

The dataset remains exactly equal.

.1.11 Modulus transformation

$$y_i^*(\lambda) = \begin{cases} \text{sign}(y_i) \frac{(|y_i|+1)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \text{sign}(y_i) \log(|y_i| + 1) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Modulus transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled modulos(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{MA}}$. Therefore, the Jacobian, the scaled, and the inverse of the modulus(A) transformation are given bellow.

The modulus(A) transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \text{sign}(y_i)\frac{(|y_i|+1)^\lambda - 1}{\lambda} \text{ if } \lambda \neq 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{MA} = \left[\prod_{i=1}^{n} |y_i| + 1\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\text{sign}(y_i)\lambda(|y_i|+1)^{\lambda-1}}{\lambda} \\ &= \text{sign}\left(\prod_{i=1}^{n} y_i\right)\left(\prod_{i=1}^{n} |y_i| + 1\right)^{\lambda-1} \\ &= \text{sign}\left(\prod_{i=1}^{n} y_i\right)\bar{y}_{MA}^{n(\lambda-1)}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \text{sign}(y_i)\frac{(|y_i|+1)^\lambda - 1}{\lambda}\frac{1}{\text{sign}\left(\prod_{i=1}^{n} y_i\right)\bar{y}_{MA}^{(\lambda-1)}}.$$

The inverse function of the modulus(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \text{sign}(y_i)\frac{(|y_i|+1)^\lambda - 1}{\lambda} \\ x_i &= \text{sign}(y_i)\frac{(|y_i|+1)^\lambda - 1}{\lambda} \\ |y_i| &= \left[\text{sign}(x_i)\lambda + 1\right]^{\frac{1}{\lambda}} - 1 \\ \Rightarrow f^{-1}(y_i) &= \pm\left[\left(\text{sign}(y_i)\lambda + 1\right)^{\frac{1}{\lambda}} - 1\right]. \end{aligned}$$

Modulus transformation case (B)

This case is exactly equal to the neglog transformation case.

.1.12 Dual power transformation

$$y_i^*(\lambda) = \begin{cases} \frac{y_i^\lambda - y_i^{-\lambda}}{2\lambda} & \text{if } \lambda > 0 \quad (A); \\ \log(y_i) & \text{if } \lambda = 0 \quad (B). \end{cases}$$

Dual power transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the scaled dual power(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{DA}}$. Therefore, the Jacobian, the scaled, and the inverse of the dual power(A) transformation are given bellow. The dual power(A) transformation presented in Table 2.2 is defined as:

$$y_i^*(\lambda) = \frac{y_i^{\lambda} - y_i^{-\lambda}}{2\lambda} \text{ if } \lambda > 0.$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{DA} = \left[\prod_{i=1}^{n}\left(y_i^{\lambda-1} + y_i^{-\lambda-1}\right)\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\lambda y_i^{\lambda-1} + \lambda y_i^{-\lambda-1}}{2\lambda} \\ &= \frac{1}{2}\bar{y}_{DA}^{n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{y_i^{\lambda} - y_i^{-\lambda}}{2\lambda}\frac{2}{\bar{y}_{DA}}.$$

The inverse function of the dual power(A) transformation is found by solving the quadratic by completing the square as:

$$
\begin{aligned}
f(y_i) &= \frac{y_i^{\lambda} - y_i^{-\lambda}}{2\lambda} \\
x_i &= \frac{y_i^{\lambda} - y_i^{-\lambda}}{2\lambda} \\
2\lambda x_i &= y_i^{\lambda} - y_i^{-\lambda} \\
2\lambda x_i &= y_i^{\lambda} - \frac{1}{y_i^{\lambda}} \\
2\lambda x_i &= \frac{y_i^{2\lambda} - 1}{y_i^{\lambda}} \\
2\lambda x_i y_i^{\lambda} &= y_i^{2\lambda} - 1 \\
1 + \lambda^2 x_i^2 &= y_i^{2\lambda} - 2\lambda x_i y_i^{\lambda} + \lambda^2 x_i^2 \\
1 + \lambda^2 x_i^2 &= (y_i^{\lambda} - \lambda x_i)^2 \\
\sqrt{1 + \lambda^2 x_i^2} + \lambda x_i &= y_i^{\lambda} \\
y_i &= \left[\sqrt{1 + \lambda^2 x_i^2} + \lambda x_i\right]^{\frac{1}{\lambda}} \\
\Rightarrow f^{-1}(y_i) &= \left[\sqrt{1 + \lambda^2 y_i^2} + \lambda y_i\right]^{\frac{1}{\lambda}}.
\end{aligned}
$$

Dual power transformation case (B)

This case is exactly equal to the Box-Cox (shift) transformation, case (B).

.1.13 Gpower transformation

$$
y_i^*(\lambda) = \begin{cases} \dfrac{\left(y_i + \sqrt{y_i^2+1}\right)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \quad (A); \\ \log\left(y_i + \sqrt{y_i^2+1}\right) & \text{if } \lambda = 0 \quad (B). \end{cases}
$$

Gpower transformation case (A)

Let $J(\lambda, y)$ denote the Jacobian of a transformation from y_i to $y_i^*(\lambda)$. In order to obtain $z_i^*(\lambda)$, the gpower(A) transformation, given by $\frac{y_i^*(\lambda)}{J(\lambda,y)^{1/n}}$, and for simplicity, we use a modification of the definition of the geometric mean, denoted by $\bar{y}_{\text{GA}}$. Therefore, the Jacobian, the scaled, and the inverse of the gpower(A) transformation are given bellow.

The gpower(A) transformation presented in Table 2.2 is defined as:

$$
y_i^*(\lambda) = \frac{\left[y_i + \sqrt{y_i^2+1}\right]^{\lambda} - 1}{\lambda} \text{ if } \lambda \neq 0.
$$

The modification of the definition of the geometric mean for this transformation is:

$$\bar{y}_{GA} = \left[\prod_{i=1}^{n}\left(y_i + \sqrt{y_i^2+1}\right)^{\lambda-1}\left(1 + \frac{y_i}{\sqrt{y_i^2+1}}\right)\right]^{\frac{1}{n}}.$$

Therefore, the expression of the Jacobian comes to:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{n} \frac{dy_i^*(\lambda)}{dy} \\ &= \prod_{i=1}^{n} \frac{\lambda\left(y_i + \sqrt{y_i^2+1}\right)^{\lambda-1}\left(1 + \frac{2y_i}{2\sqrt{y_i^2+1}}\right)}{\lambda} \\ &= \bar{y}_{GA}^{n}. \end{aligned}$$

The scaled transformation is given by:

$$z_i^*(\lambda) = \frac{\left[y_i + \sqrt{y_i^2+1}\right]^{\lambda} - 1}{\lambda}\frac{1}{\bar{y}_{GA}}.$$

The inverse function of the gpower(A) transformation is denoted as:

$$\begin{aligned} f(y_i) &= \frac{\left[y_i + \sqrt{y_i^2+1}\right]^{\lambda} - 1}{\lambda} \\ x_i &= \frac{\left[y_i + \sqrt{y_i^2+1}\right]^{\lambda} - 1}{\lambda} \\ \lambda x_i + 1 &= \left[y_i + \sqrt{y_i^2+1}\right]^{\lambda} \\ (\lambda x_i + 1)^{\frac{1}{\lambda}} &= y_i + \sqrt{y_i^2+1} \\ \left[(\lambda x_i + 1)^{\frac{1}{\lambda}} - y_i\right]^2 &= \left[\sqrt{y_i^2+1}\right]^2 \\ (\lambda x_i + 1)^{\frac{2}{\lambda}} - 2y_i(\lambda x_i + 1)^{\frac{1}{\lambda}} + y_i^2 &= y_i^2 + 1 \\ -y_i(\lambda x_i + 1)^{\frac{1}{\lambda}} &= \frac{1 - (\lambda x_i + 1)^{\frac{2}{\lambda}}}{2} \\ y_i &= -\left[\frac{1 - (\lambda x_i + 1)^{\frac{2}{\lambda}}}{2(\lambda x_i + 1)^{\frac{1}{\lambda}}}\right] \\ \Rightarrow f^{-1}(y_i) &= -\left[\frac{1 - (\lambda y_i + 1)^{\frac{2}{\lambda}}}{2(\lambda y_i + 1)^{\frac{1}{\lambda}}}\right]. \end{aligned}$$

Gpower transformation case (B)

This case is exactly equal to the glog transformation case.

Part II

Transformations in the Context of Small Area Estimation

Chapter 3

From Start to Finish: A Framework for the Production of Small Area Official Statistics

3.1 Introduction

Small area (or domain) estimation has been and still is a very fertile area of theoretical and applied research in official statistics. Although the term domain is more general as it may include non-geographic dimensions, the term small area estimation (SAE) is the established one. We shall follow the custom in this paper and use the terms area and domain interchangeably. In the last decades an increasing number of national statistical institutes (NSIs) and other organisations across the world have recognised the potential of producing small area (SA) statistics and their use for informing policy decisions. Some SA estimates have gained accreditation as national official statistics. Two examples in the UK are the annual set of unemployment estimates for unitary authorities and local authority districts (UALADs) by gender and age groups, and the estimates of average income for electoral wards. Other organisations and research groups have promoted the use of SAE techniques via the development of new methodologies and computational tools available for public use. An excellent example is the work by the World Bank (WB) and the use of its software PovMap (The World Bank, 2013). In collaboration with country teams, the WB has used SAE techniques for producing poverty maps in more than twenty developing countries. This is perhaps the most widespread application of SAE to date. Case studies can be found in The World Bank (2007).

Over time users' needs have surpassed the limits of what can be achieved with traditional SAE methods. Nowadays in addition to simple linear statistics such as averages and proportions, users request the estimation of more complex indicators, for example measures of deprivation and inequality. Meeting the increasing complexity of users' needs requires specialised methodology and software beyond conventional survey operations within NSIs. This has created opportunities for closer collaboration between researchers and NSIs and for transferring research into practice. Given the fast development of SAE methods and software researchers (or analysts) and users of small area statistics can benefit from having practical guidelines for

the SAE process. This can help to improve the understanding of what is achievable and to ensure that the methods adopted or developed are appropriate for the actual users' needs. In this paper we propose a framework based on three broadly defined stages, namely (i) specification, (ii) analysis/adaptation and (iii) evaluation, which are summarised in Figure 3.1. A description of user needs, the available data and existing SAE methods are the most important inputs to the first, specification, stage. With the help of the analyst, the user defines a set of possible target geographies and indicators and identifies potential existing small area methods that are applicable given the available data. These are the necessary inputs for the second stage.

The second stage, analysis and adaptation, is where the estimators are developed. In our view it is helpful if this process is governed by the principle of parsimony. That is, one should be looking to use the simplest possible method that achieves acceptable precision. Parsimony may be defined in terms of a hierarchy of estimation methods in increasing order of complexity. It is always possible to start by producing initial estimates that are easy to compute as part of the usual survey process within an NSI without involving explicit modelling or additional data sources. This can include direct, synthetic and composite estimators (see Section 3.3.1). Typically, these estimators can be improved by the use of standard unit/area level models (see Section 3.3.2). Clearly this is a more complex step as it involves model building and diagnostics. Finally, elaborations of the model may include use of transformations, correlated random effects over time and space, non-normal random effects and robust estimators, semi or non-parametric model specifications. The principle of parsimony dictates that such endeavour should only be introduced to overcome specific shortcomings which have been identified in the more basic methods, and the potential improvement must be weighed against the extra complexity and possible drawbacks. While such a definition of parsimony is not exact, we believe it provides a useful framework for guiding the process of producing small area estimates.

The aim of the third stage, evaluation, is to evaluate the multiple sets of estimates produced at the previous stage. This involves both uncertainty assessment and method evaluation (see Sections 3.4.1 and 3.4.2). Hopefully, the SAE process is finalised provided that at least one set of estimates is considered of acceptable precision. It is common practice for NSIs to have guidelines about precision thresholds for publishing estimates. Such thresholds can be used to define the basis of what is acceptable. However, what constitutes acceptable precision should also be defined relatively by comparing a range of methods in terms of precision gains, sensitivity to underlying model assumptions, additional investment in resources for implementing the methods and subsequent operational costs and risks. If after following these steps no set of acceptable small area estimates is found, the process may need to return to the specification stage for defining alternative geographies, target indicators and/or data sources.

To keep a practical focus it is important to illustrate the application of the proposed framework using real data. The data we use in this paper come from Mexico. While being one of the largest economies in Latin America, according to the World Bank Mexico is also among the most unequal countries in the world. Developing policies against deprivation therefore requires a detailed description of the spatial distribution of income deprivation and inequality. The National Council for the Evaluation of Social Development Policy (CONEVAL *Consejo Nacional de Evaluación de la Política de Desarrollo Social*) is responsible for estimating measures of

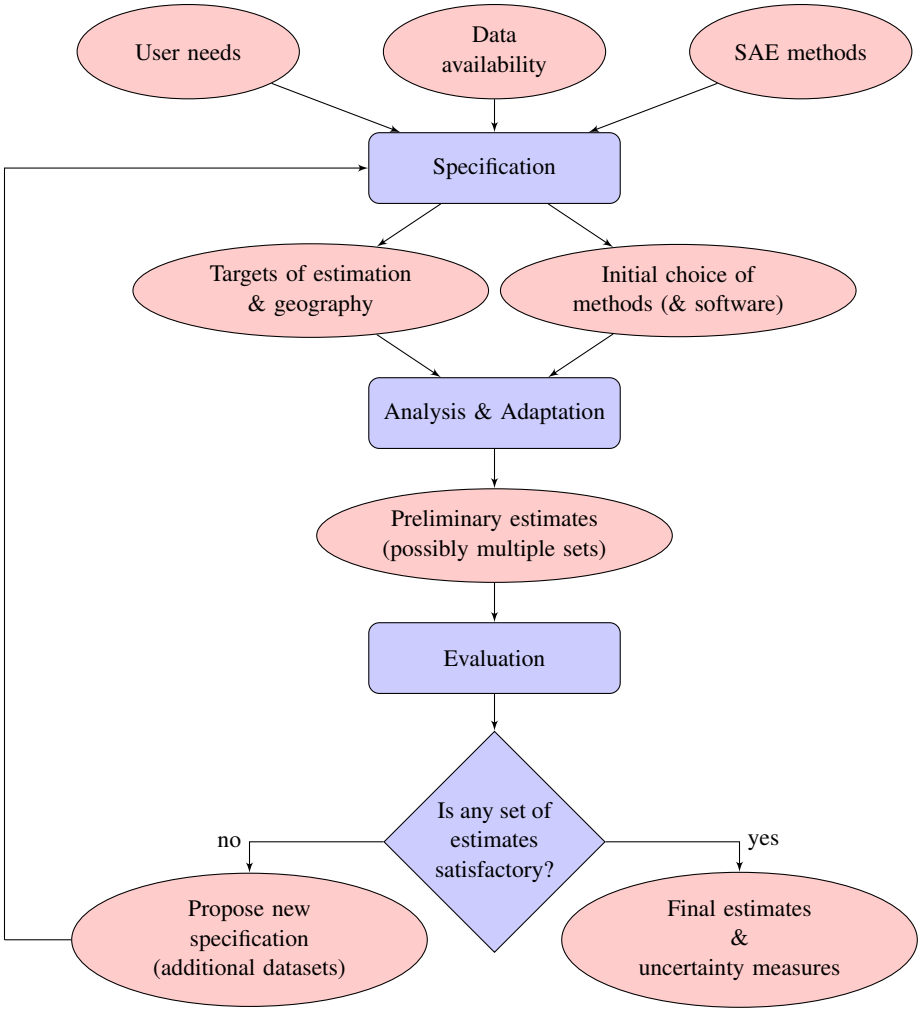


Figure 3.1: Framework for the production of SA statistics: Stages of the project are represented by blocks. Inputs and outputs of each stage are represented by ellipses. Decisions to be made are represented by diamonds. Arrows indicate the direction of the relationship. Text in parenthesis indicates optional items

poverty, social deprivation and inequality in Mexico. Furthermore, the general social development law (LGDS *Ley General de Desarrollo Social*) requires measures at the national and state levels to be obtained every two years and measures at the municipal level every five years. For the purposes of empirical analysis in this paper we use a sample from the household income and expenditure survey, ENIGH (*Encuesta Nacional de Ingreso y Gasto de los Hogares*) and a large sample of census micro-data. Both datasets are produced by the National Institute of Statistics and Geography (INEGI *Instituto Nacional de Estadística y Geografía*) and were provided to the authors by CONEVAL. In the present paper we shall illustrate the SAE process for estimating linear and non-linear indicators based on continuous outcomes, recognising that in practice discrete and categorical variables may also be of interest.

The paper is structured as follows. Sections 3.2 - 3.4 describe the three stages of the SAE process, one for each stage. Section 3.5 provides a review of open source software for SAE. In Section 3.6 we conclude the paper with some final remarks and comments on open areas for research.

3.2 Specification

In this section we describe the elements of the first stage in our framework. This includes specifying the user needs, the targets of estimation, the target geography and reviewing the data sources available and their geographical coverage.

3.2.1 Specify user needs: Targets of estimation and target geography

Sample surveys are designed to provide estimates with acceptable precision at national and specific sub-national levels but usually have insufficient sizes to allow for precise estimation at lower levels of aggregation. An important task at this stage is the specification of the target level of geography and the targets of estimation, which will impact upon all the subsequent SAE process. It is very tempting for the user to target a geography that is unrealistically low. As we will see later, doing so will affect the methods and the assumptions required for computing the estimates and evaluating their precision. It is also becoming increasingly common that the user is interested in more than simple linear indicators such as averages and proportions, and aims for more complex, non-linear indicators, for example estimating the percentiles of the income distribution locally. As will be explained in the next section, increasing the complexity of the targets of estimation increases the granularity of the data one needs to have access to. Hence, the recommended approach is to start from a relatively high level of geographical aggregation, at which direct estimation with acceptable precision is supported by the survey data, and move on to more disaggregated levels of geography after assessing the feasibility of producing small area estimates at each level in turn. It would be ideal if a level can be chosen which both serves the user needs and is well supported by the data available. Sometimes, however, the user may have a non-negotiable target level of geography - as is the case in Mexico - dictated by specific policy needs or predetermined by law. Even in this case, it is still the responsibility of the statistician to explain to the user the consequences of the different choices and the extent to which the results will depend on finding a good enough predictive model for the level of interest.

Besides the target level of geography and the targets of estimation, the most important properties of the estimation method also need to be clarified. For instance, whether the user is more interested in cross-sectional estimates or estimates of change over time will affect both the data required and the models used. For purposes such as fund allocation, policy evaluation and monitoring, it may be important to pay attention to the various ensemble characteristics of the estimates such as the range, the rank and order statistics. The standard approach to deriving model-based small area estimates is to minimise the squared prediction error for each given area subject to unbiased prediction. This is intuitive for area-specific cross-sectional estimation but is generally not optimal if there are other properties that are more important to the actual use of the small area estimates. A clear understanding of the most desirable properties of the estimates is therefore necessary in order to ensure that the user needs are served in the best possible way.

3.2.2 Data availability and geographical coverage

Identifying what data are needed affects not only the estimation results but also the workload of staff at NSIs and similar organisations. Small area estimation is a prediction problem and typically relies on the use of survey data and data from the census or administrative/register data sources. The census data contain auxiliary information that is potentially correlated with the target variable and can be used to improve the estimation. Access to census and administrative data sources is usually challenging due to confidentiality constraints. Commonly, access to census aggregate (area/domain) level data is possible but access to census micro-data may not be possible. The question is how the type of census data available affects small area estimation. If the user is interested in estimating linear statistics, for example small area averages, access to area level census or administrative data will be sufficient for small area estimation. To illustrate this, suppose we have data on an outcome variable y_{ik} and a set of covariates $\boldsymbol{x}_{ik}$ for individuals i in domains k. The target of estimation is the domain average and for now let us assume that estimation is assisted by a regression model with model parameters $\boldsymbol{\beta}$. An estimator of the small area average is defined as follows,

$$\hat{\bar{\theta}}_k = N_k^{-1}\left[\sum_{i=1}^{n_k} y_{ik} + \sum_{i=n_k+1}^{N_k} \boldsymbol{x}_{ik}^T\hat{\boldsymbol{\beta}}\right], \tag{3.1}$$

where n_k (N_k) denotes the sample (population) size in domain k and $\boldsymbol{x}_{ik}^T$ is the transpose of the vector $\boldsymbol{x}_{ik}$. The first summation in (3.1) is computed by using the survey data in domain k, assuming that sample data are available in the domain. The second summation in (3.1) represents the out-of-sample model predictions. It is easy to see that in order to compute (3.1), there is no need to have access to covariate micro-data. Instead, access to domain-level totals $\sum_{i=1}^{N_k} \boldsymbol{x}_{ik}$ will be sufficient. If the interest is however in estimating non-linear indicators, then access to census or administrative micro-data is needed. Access to such data is very challenging and has implications for staff resources, in for example ensuring appropriate use of the data and respecting confidentiality constraints. Hence, the complexity of the targets of estimation determines the data requirements for small area estimation. Although the illustration of methods in this paper assumes the availability of census/administrative micro-data for covariates, it is important to discuss briefly what options are available when such data are not available. One possibility is to assume a model for the observed covariates and impute the missing values from that model (e.g. Sverchkov and Pfeffermann, 2004). With many covariates this might be too cumbersome and Pfeffermann and Sikov (2011) develop a simple non-parametric alternative that is shown to work well. An alternative approach would be to use area level models. Fabrizi and Trivisano (2016) consider hierarchical Bayes approaches to fitting area level models for estimating non-linear indicators. Schmid et al. (2017) present a first attempt to use sources of big data, in particular mobile data, as covariate information in area level models. We believe that researchers should invest more effort on developing methodologies and software that can be used when population micro-data for the covariates are not available or are available only for a sample from the target population.

It is also necessary to examine the data coverage at the specified level of geography. The

analyst should explore whether sample observations are available for every small area and also check the distribution of the sample size across areas. For example, if many of the target areas have no sample data (out-of-sample areas), the user must realise that small area estimation will heavily rely on model assumptions. Even when data are available for every domain one may still decide to use models in an attempt to improve the precision of direct estimation. Deciding whether to use models and which model to use is a complex process which is governed by a trade-off between improved efficiency and dependence on model assumptions. Our recommendation is for users to be open to alternative methodologies and for researchers to place emphasis on diagnostic analysis for evaluating small area estimates. The process of model building will be illustrated later in the paper.

3.2.3 Illustration using the ENIGH data

In this case the targets of estimation and the required geography are specified by the LGDS (see Section 3.1). The Mexican government is interested in estimates of proportions and totals of social and economic deprivation, as well as more complex, non-linear, indicators such as estimates of the Gini coefficient (Gini, 1912; Ceriani and Verme, 2012) and income ratio. Methodologists in CONEVAL have access to micro-data from the most recent census and survey data from the ENIGH. Hence, the estimation of the target indicators specified by the LGDS is feasible at least in principle.

Let us now look in more detail at the data available and their geographic coverage. Mexico is divided into 32 federal entities (states). The State of Mexico (EDOMEX *Estado de México*) has the highest population density, and is also regarded by the United Nations Development Programme (UNDP) as being one of the states that most contribute to inequality in Mexico. EDOMEX is made up of 125 municipalities, which by their geographical and demographic characteristics are further grouped into 16 districts. The pilot data we have available were provided by CONEVAL and come from the 2010 ENIGH survey and the 2010 census in EDOMEX. The ENIGH survey data comprise 2748 households in 58 out of 125 municipalities. The census micro-data covers all EDOMEX municipalities. The survey and census data sources include a large number of socio-demographic variables, many of which are common and are measured in similar ways in both datasets. Total equivalised household income is an example of a variable that is available in the ENIGH survey but not in the census.

For the ENIGH survey more than 50% of municipalities are out-of-sample, making direct estimation for these municipalities impossible. For in-sample municipalities, the median sample size is 21 households and the mean is 47.4 households. The case here illustrates the situation where the user has a non-negotiable target geography predetermined by legal requirements, which clearly poses challenges for estimation. On the one hand, the use of SAE methods can be justified if (a) they can produce municipal estimates that are more efficient than direct estimates and (b) they can produce acceptable estimates for non-sampled municipalities. On the other hand, it is important that the analyst carefully communicates the potential impact of model assumptions and appropriately evaluates the methods and the estimates.

3.3 Analysis/Adaptation

The second stage in small area estimation involves the analysis of the data and the adaptation of the models. As explained earlier, in our view the process should be governed by the principle of parsimony. Section 3.3.1 presents a triplet of small area estimates described in the Eurostat document ESSnet SAE (2012). As we shall explain, these estimators can always be obtained as by-products of the original sample survey estimation set-up without any additional modelling effort. Ideally this triplet of estimates should be provided by the user to the analyst as an input to the analysis and adaption stage but this is hardly ever the case. The analyst will most likely need to extend the triplet of estimates, by developing suitable models for small area estimation, both to improve the method of estimation and to be able to handle more complicated target parameters. Sections 3.3.2 and 3.3.3 use the ENIGH data to describe and illustrate the core activities of analysis and adaption including the relevant issues of how to use a model for prediction, model building, model testing, diagnostic analysis and finally adaptations of the model that are informed by the diagnostic analysis.

3.3.1 Initial triplet of estimates

The initial triplet of estimates for the small area parameter θ_k are the direct, synthetic and composite estimates. The direct estimator, denoted by $\hat{\theta}_k^{Direct}$, uses only the data from area k, so it is available only for an in-sample area. For areas with small sample sizes we expect that the direct estimator will have low precision. The synthetic estimator, denoted by $\hat{\theta}_k^{Synthetic}$, uses the data from a broader area that includes area k and so it can be derived for any out-of-sample area as well. Use of a synthetic estimator reduces uncertainty but at the cost of possibly introducing bias. Let us make things more specific and distinguish between two situations of standard design-based sample survey estimation. The first is when no auxiliary data are available and the estimation is based on the design weights directly. For example, let $\bar{\theta}_k$ be the area population mean. The Hajek-Brewer Ratio estimator is defined by

$$\hat{\bar{\theta}}_k^{Direct} = \Big(\sum_{i=1}^{n_k} y_{ik}/\pi_{ik}\Big)/\Big(\sum_{i=1}^{n_k} 1/\pi_{ik}\Big), \tag{3.2}$$

where π_{ik} is the corresponding sample inclusion probability (Hájek, 1958; Brewer, 1963). A synthetic estimator of the mean $\hat{\bar{\theta}}_k^{Synthetic}$ is given similarly, based on the sub-sample from a broad area including area k, denoted by $\hat{\theta}_k^{Synthetic} = \hat{\theta}$, where $\hat{\theta}$ is a broad area estimate. The second situation is when auxiliary data are available, in which case the estimation is based on model-assisted weights (Särndal et al., 1992), denoted by w_{ik}, for unit i in area k. In this case the direct estimator of the area population mean is given by

$$\hat{\bar{\theta}}_{k,GREG}^{Direct} = \frac{1}{N_k}\sum_{i=1}^{n_k} w_{ik} y_{ik},$$

where $w_{ik} = g_{ik}/\pi_{ik}$, and $g_{ik} = 1 + (X - \sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik}/\pi_{ik})^T (\sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^T/\pi_{ik})^{-1}\boldsymbol{x}_{ik}$, and X is the population total of $\boldsymbol{x}_{ik}$. A synthetic estimator $\hat{\theta}_k^{Synthetic} = \bar{\boldsymbol{x}}_k^T\hat{\boldsymbol{\beta}}$ is obtained by the linear model $E(y_{ik}|\boldsymbol{x}_{ik}) = \boldsymbol{x}_{ik}^T\boldsymbol{\beta}$, with $\hat{\boldsymbol{\beta}} = (\sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik}\boldsymbol{x}_{ik}^T/\pi_{ik})^{-1}(\sum_k \sum_{i=1}^{n_k} \boldsymbol{x}_{ik}y_{ik}/\pi_{ik})$

and $\bar{\boldsymbol{x}}_k = N_k^{-1} \sum_{i=1}^{N_k} \boldsymbol{x}_{ik}$. One approach to reconciling the possibly large bias of a synthetic estimator and the possibly large variance of a direct estimator is to define a composite estimator, which is a linear combination of the two. This defines the last estimator in the triplet of initial estimators:

$$\hat{\theta}_k^{Composite} = \alpha_k \hat{\theta}_k^{Direct} + (1 - \alpha_k)\hat{\theta}_k^{Synthetic}, \tag{3.3}$$

for some chosen coefficient $\alpha_k \in [0, 1]$, where by definition $\alpha_k = 0$ for any out-of-sample area.

There are several choices of α_k for the composite estimator (3.3), including the James-Stein estimator that uses a common α in all areas, and the area-specific minimizer of the mean squared error (MSE). The latter is not very practical and Rao and Molina (2015) discuss different approaches for selecting α_k. One alternative approach is to define α_k as a function of the domain sample size such that for domains with larger sample size a higher weight is given to the direct estimator. It is worth noting that the composite estimator appears more intuitive for target parameters that are linear statistics of the $\{y_{ik}\}$, like domain averages. However, estimators of more complex statistics for example percentiles of the domain-specific distribution function and non-linear indicators have recently attracted some interest in the small area literature (Tzavidis et al., 2010; Alfons and Templ, 2013). Regardless of how the initial triplet of estimates is produced, it provides useful input to the analysis and adaptation stages and possibly to the specification stage too.

The initial triplet estimates would certainly be more useful if some appropriate measure of the associated uncertainty can be produced in addition. However, it can be challenging to obtain a stable estimate of the potential bias of the synthetic and composite estimator, as we shall discuss in Section 3.4. At the very minimum, the direct estimates need to be analysed and their uncertainty quantified as this will offer an indication of the improvement required for producing small area estimates. It is common that the analyst will subsequently consider the use of more complex model-dependent SAE methods. In this case juxtaposing the direct, synthetic and composite estimates provides a tangible appreciation of the between-area variation of the target parameter, i.e. the heterogeneity across the areas, as well as possibly the predictive power of the auxiliary variables already in use.

3.3.2 Use of models for small area estimation

Small area estimation is one of the areas in survey sampling where the use of models is widely accepted as necessary. Model-based methods assume a model for the population and sample data and construct optimal predictors of the target parameters under the model. The term predictor instead of estimator is conventionally used as, under the model, the target parameters are assumed to be random. Here we describe how to use a model to estimate both linear and non-linear small area parameters of interest. In Section 3.3.3 we describe model building, diagnostic analysis and model adaptations in more detail.

Users of small area statistics in Mexico are interested in the estimation of key income-related indicators such as the Head Count Ratio (HCR) and the Gini coefficient. To this set we add average income, which is also of interest for NSIs. The most widely used approaches

for estimating non-linear indicators require the use of unit-level survey data for the outcome variable and the covariates, and unit-level census micro-data for the covariates. Area-level models for non-linear indicators have been proposed in the literature (Fabrizi and Trivisano, 2016) but these models lie outside the scope of the present paper.

Two predominant approaches for estimating non-linear indicators are the World Bank method (Elbers et al., 2003) and the Empirical Best Predictor (EBP) method (Molina and Rao, 2010). To start with both methods make use of a unit-level nested error regression model (Battese et al., 1988). The response variable is a welfare variable that is only available in the survey, e.g. income or consumption. The explanatory variables, used for modelling the welfare variable, are available both in the survey and in the census datasets. After the model is fitted using the survey data, the estimated model parameters are combined with census micro-data to form unit-level synthetic census predictions of the welfare variable. The synthetic values of the welfare variable along with a defined poverty line are then used for estimating non-linear indicators, for example the HCR or the Gini coefficient. Linear statistics such as average income can also be estimated by using the same synthetically generated values.

Let us first describe the EBP approach, before we provide a brief discussion of the similarities and differences from the World Bank method. Under the EBP approach census predictions of the welfare outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The starting point is the following unit-level nested error regression model,

$$y_{ik} = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + u_k + \epsilon_{ik}, u_k \sim N(0, \sigma_u^2); \epsilon_{ik} \sim N(0, \sigma_\epsilon^2), \tag{3.4}$$

where u_k denotes the domain random effect. A random effect is necessary when the covariates we include in the model do not fully explain the between-domain variability. Assuming normality for the unit-level error and the domain random effects, the conditional distribution of the out-of-sample data given the sample data is also normal. The synthetic values of the welfare variable for the entire area population (of size N_k) are then generated from the following model,

$$y_{ik}^* = \boldsymbol{x}_{ik}^T\boldsymbol{\beta} + \tilde{u}_k + u_k^* + \epsilon_{ik}^*, u_k^* \sim N(0, \sigma_u^2 \times (1-\gamma_k)); \epsilon_{ik}^* \sim N(0, \sigma_\epsilon^2); \gamma_k = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{n_k}}, \tag{3.5}$$

where $\tilde{u}_k = E(u_k|y_s)$ is the conditional expectation of u_k given the sample data y_s. In (3.5), $\boldsymbol{x}_{ik}^T\boldsymbol{\beta} + \tilde{u}_k$ is the conditional mean of y_{ik} in the population given the sample data, whereas $u_k^* + \epsilon_{ik}^*$ are simulated from the conditional normal distribution of y_{ik} for the units outside the sample. Implementation of (3.5) requires replacing the unknown quantities $\boldsymbol{\beta}, \sigma_u, \sigma_\epsilon$, with estimates and simulating L synthetic populations of the welfare outcome, $\boldsymbol{y}^*$. Linear and non-linear indicators are computed in each domain k for each replication and the estimates are averaged over L. A moderate number of Monte-Carlo simulations, $L = 50$ or $L = 100$, is used in practice. MSE estimation for model-based small area estimation will be discussed in Section 3.4.1. For now we notice that evaluation of the uncertainty both for in-sample and out-of-sample domains is usually performed using parametric bootstrap under (3.4) and (3.5). Alternatively, protection against model misspecification can be offered by wild bootstrap. In

this case bootstrap for the unit level error term uses the empirical distribution of scaled residuals instead of a normal distribution.

We now briefly compare the World Bank and EBP methods. Although both methods use a nested error regression model, one key difference in practice is that in the World Bank method it is common to specify the random effect at a much finer geography (cluster) level (indexed by l) whereas in the EBP method the random effect is specified at the domain level. A second key difference is that the EBP method simulates population realisations of the outcome from the estimated conditional distribution (3.5) whereas the World Bank method simulates from the marginal distribution,

$$y_{il}^* = \boldsymbol{x}_{il}^T\boldsymbol{\beta} + u_l^* + \epsilon_{il}^*, u_l^* \sim N(0, \sigma_u^2); \epsilon_{il}^* \sim N(0, \sigma_\epsilon^2), \tag{3.6}$$

with all parameters replaced by their estimates. We now distinguish two cases. When clusters coincide with the target domains, Molina and Rao (2010) demonstrate the superior performance of the EBP method for in-sample domains. For out-of-sample domains the predicted random effect u_k and the shrinkage factor γ_k in (3.5) are both zero by default so that (3.5) reduces to (3.6) and the two methods yield the same estimates. Next, consider the more common case where clusters and target domains do not coincide. Since in most applications the between-domain variation tends to be small compared to the between-household variation, the conditional distribution (3.5) may not differ much from the unconditional distribution, as long as the variance of $\tilde{u}_k$ is small compared to the total variance of $y_{ik} - \boldsymbol{x}_{ik}^T\boldsymbol{\beta}$. Meanwhile, since the World Bank method is applied at the cluster level, it is possible to capture much of the variability beyond the between-household variability at the cluster level, provided relevant cluster level covariates are included in the fixed part of the model (3.4). Moreover, the use of the conditional distribution (3.5) may be impossible in most of the clusters due to the absence of sample units. The World Bank method is then well suited in practice, despite the use of the marginal distribution (3.6). Having said this, Marhuenda et al. (2017) recently proposed EBP methodology that allows for a two-fold nested error regression model that can accommodate both cluster and domain random effects.

3.3.3 Model building, residual diagnostics and transformations in practice

Before considering model-based estimation, an assessment of initial estimates produced with the ENIGH data is necessary for motivating the use of more complex methods. The data provider did not supply the initial triplet of estimates described in Section 3.3.1. Producing appropriate sets of initial estimates and their corresponding coefficients of variation (CV) would require access to data about the sampling design beyond our reach. The analysis below, obtained using the function direct of the sae package in R (Molina and Marhuenda, 2015), attempts to replicate such initial estimates in a way that can inform the subsequent stages of the process. Figure 3.2 (left) presents point estimates of average equivalised household income at the municipality level calculated from the ENIGH survey data using the final weights supplied. Figure 3.2 (right) shows estimated CVs, obtained under the assumption of a single-stage Poisson sampling of households in each municipality, with first order inclusion probabilities

given by the inverse of the final weights. The assumption of single stage Poisson sampling is made for convenience. We expect the CVs estimated under this assumption to be overly optimistic considering that the actual sampling design of the ENIGH includes stratification and two stages of selection, and has a design effect around 3.3 for the income variable (ENIGH, 2010). However, even under this optimistic scenario it can be seen that, with the exception of few municipalities, the CVs are clearly above usual publication thresholds of $20\% - 25\%$. Notice also that direct estimates cannot be produced for the out-of-sample municipalities (white coloured areas). Hence, in order to satisfy the current user needs we should explore the use of model-based methods.

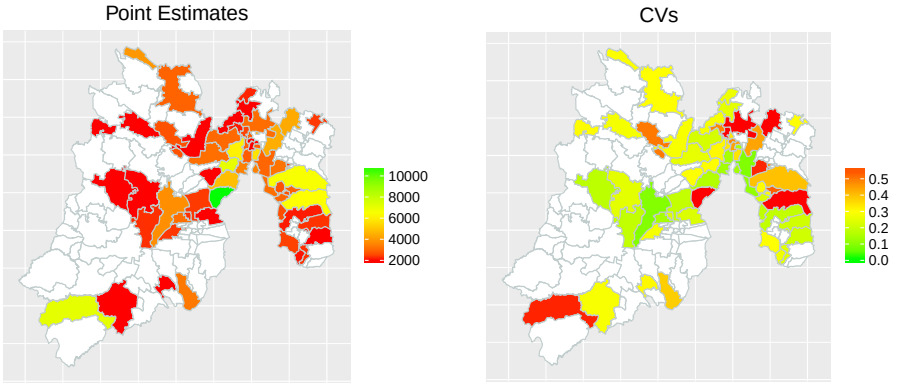


Figure 3.2: Direct estimates of average household equivalised income and CVs in EDOMEX municipalities

The use of models aims to improve the precision of small area estimates by making optimal use of the data available. Hence, model building, model diagnostics, sensitivity analysis and validation take central stage in model-based small area estimation. There is no single approach to model building. Here we describe some best practice guidelines one could follow, and illustrate these guidelines for estimating income related indicators with the ENIGH data.

Model-based estimation requires the use of a model that usually includes area random effects. However, before discussing the use of random effects, the most important step in building the model remains the specification of the fixed effects part. Ideally, one should aim to explain as much between-domain variation as possible by using the available covariates so that random effects can potentially be avoided in the spirit of parsimony. A reasonable starting point for building the model is therefore to use a standard regression model with uncorrelated errors. Alternatively, if one suspects that despite the inclusion of covariates there is unexplained between-domain variability which can affect inference for the regression parameters, the analyst can consider a regression model with correlated errors for example, an exchangeable correlation structure in the simplest case. In order to decide whether to include a covariate in the fixed part of the model one can use simple t-statistics -computed using the correct variance under the model- or information criteria, for example the Akaike or the Bayesian Information Criteria (AIC, BIC) computed under the standard linear model with uncorrelated errors. In the case of the ENIGH data and following the recommendation by the data provider (CONEVAL), y is defined to be the total household per capita income (*ictpc*) measured in Mexican pesos, which is the current monetary and non-monetary income of households adjusted by equivalent

scales and economies of scales. Using the AIC and a standard linear regression model the following covariates that are available both in the survey and census data have been identified as good predictors of *ictpc*:

1. Percentage of employees older than 14 years in the household;
2. Highest degree of education completed by the head of household;
3. Social class of the household;
4. Percentage of income earners and employees in the household;
5. Total number of communication assets in the household;
6. Total number of goods in the household.

To investigate whether the use of a mixed effects model is necessary, we estimated a linear model with an exchangeable correlation structure using generalized least squares (GLS) (Pinheiro and Bates, 2000). The model is estimated in R with function gls within the nlme package (Pinheiro et al., 2017). The class of GLS models contains the standard linear model that assumes independence as a special case. Therefore, given the fixed effects, the standard linear model is nested within the model with exchangeable correlation structure and a likelihood-ratio test or other information criteria can be used to decide whether the latter fits the data better. First, we compared the GLS with an exchangeable correlation structure against a standard linear model where both models included only an intercept term. This allows us to quantify how much of the between-municipality variability is explained by the model covariates. We conclude that the model with exchangeable correlation structure fits the data better than the standard linear model (AIC for GLS with an exchangeable correlation structure: 54239 vs. AIC for the standard linear model: 54275). One could also use a likelihood-ratio test for comparing the two models which produces a p-value for testing. The value of this test statistic is 37.52, with a p-value $4.521 \cdot 10^{-10}$, which provides evidence of significant unobserved heterogeneity between municipalities. Care must be taken with using a likelihood-ratio test when a parameter like a random effects variance is on the boundary of the parameter space (see e.g. Snijders and Bosker, 2012). In the second step the GLS and standard linear regression models with the set of six covariates identified above were compared against each other. The AIC and the likelihood-ratio test (p-value: 0.029) suggest that the model with the exchangeable correlation structure fits the data marginally better (AIC for GLS with an exchangeable correlation structure: 53077 vs. AIC for the standard linear model: 53079). The difference between these AIC values is very small, indicating that the covariates we included in the model explain a substantial part of the between municipalities variability. In particular, the intra cluster correlation (ICC) for the empty GLS model is 0.054 and for the GLS model that includes the six significant predictors it reduces to 0.015. In light of the marginally better fit of the GLS model, the benefits of a random effects model are likely to be small. We discuss this in Section 3.4 where we compare indirect and regression synthetic estimates. Although not used in the case study, model selection and testing procedures under the random effects model have been proposed in the literature. Here we refer to the use of a conditional AIC criterion (Vaida and Blanchard, 2005) that accounts for the prediction of random effects in selecting covariates to be included in the model. We further refer to a test for the inclusion of random effects proposed by Datta et al. (2011). The

authors show that if random effects are not needed and are removed from the model, the precision of point and interval estimators is improved. Additional testing procedures are proposed by El-Horbaty (2015) and reviewed by Pfeffermann (2013).

After the best possible set of covariates has been identified, the inclusion or not of random effects has been decided and the model has been fitted, the next step in model selection uses residual diagnostics and assessment of the predictive power of the model. Despite the inclusion of a number of significant covariates, the model may have low predictive power. The user must remember that SAE is concerned with prediction and not with discovering associations and causal mechanisms between the explanatory variables and the outcome. Hence, assessing the overall predictive power of the model is important. One can use simple measures such as the coefficient of determination (R^2) of the model without random effects. Alternative, computer intensive methods such as cross-validation can be used. Cross-validation is mentioned by Pfeffermann (2013) and consists of leaving some areas out of the model fitting process and comparing model-based predictors for these areas with corresponding design-based estimates. For example, one may use as a validation benchmark design-based estimates for larger areas which can be trusted. For residual diagnostics we propose the use of graphical diagnostics such as normal Q-Q plots of the residuals (unit-level and domain-level) for checking the model assumptions and plots of standardised residuals against fitted values for testing the assumptions of constant variance. If residual diagnostics indicate that the model assumptions hold, the analyst can proceed to the production of point and MSE estimates. However, in most applications some adaptations of the model will be needed.

To illustrate the use of diagnostic analysis and model adaptation let us focus on the EBP method we described in Section 3.3.2 which relies on the normality of the residual terms. Figure 3.3 shows normal Q-Q plots of household-level and municipal-level residuals (random effects) obtain by fitting model (3.4) to income, using the six covariates we identified above and including municipality-specific random effects. There are notable departures from normality. This can be seen both from the shape of the normal Q-Q plots and from Table 3.1 where the skewness and kurtosis of the two sets of residuals are clearly different from that expected for normal data.

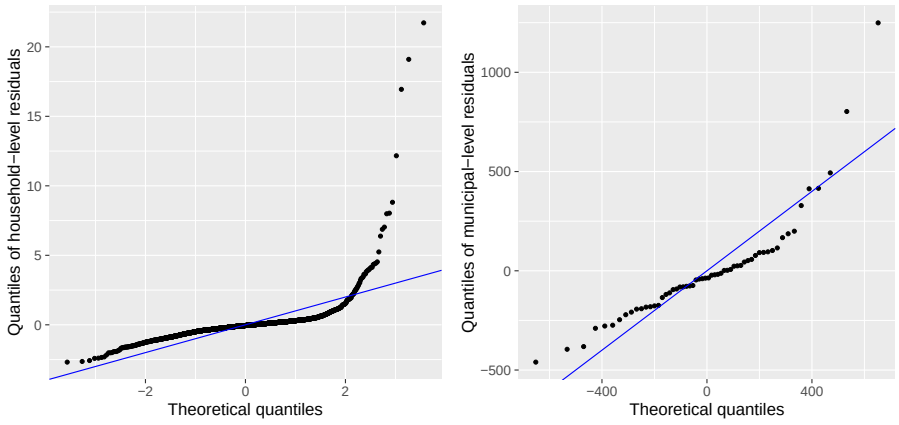


Figure 3.3: Normal Q-Q plots for household-level residuals (left) and municipal-level residuals (right) obtained from the model that uses raw income as the response variable

When residual diagnostics indicate that there are departures from normality, the analyst has several options. The first option is to use alternative parametric specifications that are more realistic. In the case of income data two possible distributions are the Pareto distribution or the Generalised Beta distribution of the second kind. The complication with using alternative distributions is that the analyst may need to develop new estimation and inference theory for each new application. Alternative semi-parametric approaches to model-based small area estimation have also been proposed (Weidenhammer et al., 2014). Use of semi-parametric methods also requires new theory and additional training for the users. There is also a large body of literature on extensions of the nested error regression model to better handle real data challenges. Examples include outlier robust estimation (Datta and Lahiri, 1995; Ghosh et al., 2008; Sinha and Rao, 2009; Chambers et al., 2014; Fabrizi et al., 2014), models with non-parametric instead of linear signal specification (Opsomer et al., 2008; Ugarte et al., 2009) and models that extend the covariance structure of the model by allowing for spatially correlated domain random effects (Pratesi and Salvati, 2009; Schmid et al., 2016) or for complex variance structures (Jiang and Nguyen, 2012). An option- when diagnostic analysis shows departures from the model assumptions- and one that is based on the principle of parsimony is to find a transformation of the data such that the normality assumptions of the EBP are met. Doing so means that the analyst can keep using standard estimation tools and software for small area estimation. The challenge in this case is in finding the most appropriate transformation. This adds another layer of complexity to the model building process. We now discuss the use of transformations in some detail as an example of adapting the model. This is something we encourage prospective users to explore before deciding to use more complex models.

The papers by Elbers et al. (2003) and Molina and Rao (2010) considered the use of a logarithmic or a logarithmic-shift transformation, which are popular for income data. A better approach is to use data-driven transformations with optimally chosen parameters. Data-driven transformations may offer better predictive power and hence small area estimates with improved precision. For an illustration using the ENIGH data we consider the log-shift -with an optimally chosen shift- and on the Box-Cox transformation (Box and Cox, 1964; Gurka et al., 2006). One key difference between the logarithmic and these additional transformations is that in the latter case the choice of transformation is adaptive i.e. driven by the data. This is achieved by a transformation parameter, denoted by λ, which must be estimated. The logarithmic transformation is then a special case of this family of transformations when $\lambda = 0$. Denoting by $\mathrm{T}_\lambda(y_{ik})$ the transformed outcome, the log-shift transformation is defined by

$$\mathrm{T}_\lambda(y_{ik}) = log(y_{ik} + \lambda). \tag{3.7}$$

The Box-Cox transformation is defined by

$$\mathrm{T}_\lambda(y_{ik}) = \begin{cases} \frac{(y_{ik}+c)^\lambda - 1}{\kappa^{\lambda-1}\lambda}, & \lambda \neq 0 \\ \kappa \log(y_{ik}+c), & \lambda = 0 \end{cases}, \tag{3.8}$$

for $y_{ik} > -c$, where c is a fixed parameter, which makes the data positive to enable the use of the Box-Cox transformation and κ is the geometric mean of y_{ik} (Box and Cox, 1964; Gurka

et al., 2006). This is an example of a scaled transformation. Conditional on κ, the Jacobian of the transformation is 1. Using the scaling by the geometric mean allows for the use of the likelihood function under the nested error regression model and as a result standard software for fitting this model with the transformed data can be used. This is consistent with the principle of parsimony. Different approaches have been proposed in the literature for estimating the optimal transformation parameter in linear models. These methods are mainly based on maximum-likelihood theory. However, little attention has been paid to the use of these techniques with linear mixed models. Gurka et al. (2006) used Box-Cox transformations based on restricted maximum likelihood theory for the estimation of the power transformation parameter in linear mixed models. In addition, the minimization of a measure of the asymmetry such as the skewness of the residuals for the log-shift transformation has been discussed by Feng et al. (2016). An empirical approach for choosing λ in (3.7) is to define a grid of values for λ, fit the nested error regression model by using each of the transformed outcomes $\mathrm{T}_\lambda(y_{ik})$ and select the transformation that makes distribution of the residuals as close as possible to normal. Note, however, that here we deal with two sets of residuals and to our knowledge there is no formal approach to defining the distance from normality. Recent work by Rojas-Perilla et al. (2017) studies the use of different scaled transformations and estimation methods for λ in small area estimation. A general algorithm for implementing the EBP method with power transformations is as follows:

1. Define a parameter interval for λ;
2. Set λ to a value inside the interval;
3. Maximize the restricted log-likelihood function with respect to the vector of model parameters conditional on the fixed value of λ;
4. Repeat 3 and 4 until the value of λ that maximises the likelihood is found;
5. Apply the EBP method with the chosen value of λ.

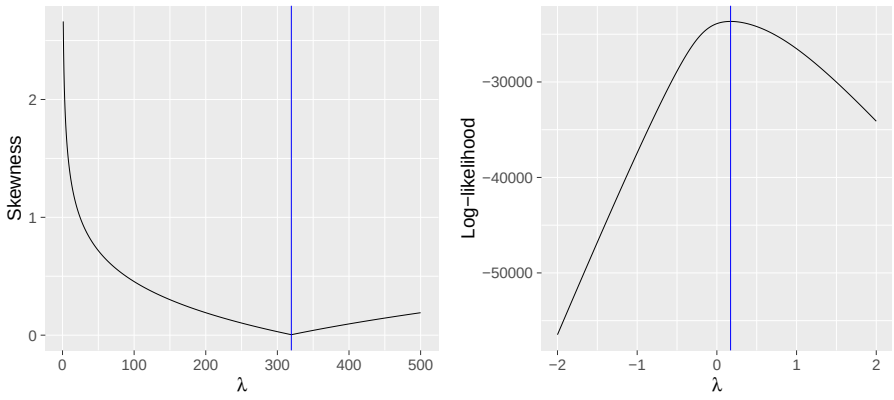


Figure 3.4: Shift parameter for the log-shift transformation (left) and optimal λ for the Box-Cox transformation (right)

Using the ENIGH data we apply the EBP method with three transformations for the outcome, namely log, log-shift and scaled Box-Cox. Figure 4.2 on the right shows the graphical representation of the maximization of the restricted maximal log-likelihood on a grid $\lambda \in [-2; 2]$

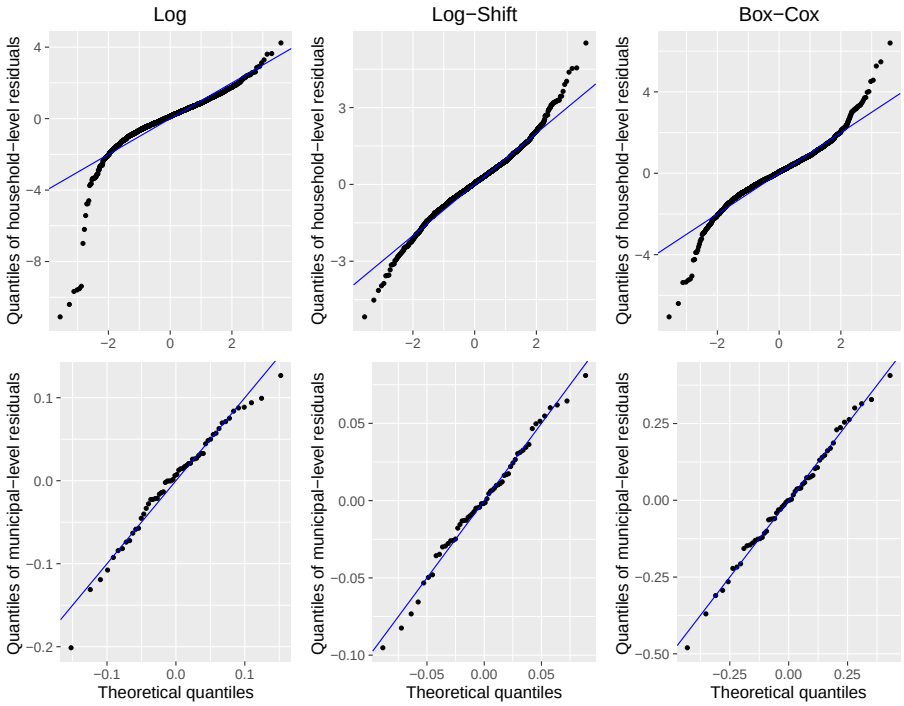


Figure 3.5: Normal Q-Q plots for household-level residuals and municipal-level residuals under three transformations for income

in the case of the Box-Cox transformation. In this case the optimal λ is approximately equal to 0.17. A similar graph on the left shows the shift parameter that minimises the skewness of the household-level error term. The resulting parameter is equal to 319.52. The question is whether the use of the transformations identified above improve the diagnostic analysis and the predictive power of the model. We start with comments on the normal Q-Q plots (Figure 3.5) and the distribution of the residuals in Table 3.1. For municipality random effects, all three transformations offer a good approximation to normality (see also Table 3.1). The picture is different for household-level. In particular, the household-level residuals under the log model show severe departures from normality. The situation is clearly improved when using the log-shift and power transformations (see also Table 3.1) with the log-shift transformation leading to less extreme and more symmetrical tails than the other transformations.

In order to assess the assumption of homoscedasticity, we produce plots of the fitted values (x-axis) against the standardised residuals (y-axis) obtained by fitting model (3.4) using the raw income data (left) and the Box-Cox power transformation (right) in Figure 3.6. It can be observed that using transformations helps to stabilise the variance of the residuals. The corresponding plots for the log and the log-shift transformations are similar.

The proportion of variability explained under each model is quantified by the coefficients of determination R^2 summarised in Table 3.1. Note that as R^2 is computed based on the transformed outcomes, the R^2 values are not directly comparable. As pointed out before, using the raw values of income in the EBP nested error regression model produces clearly unsatisfactory normal Q-Q plots and a R^2 equal to 31%. The use of transformations improves the predictive power of the model for the transformed variables.

Based on the results from the diagnostic analysis we conclude that two transformations,

Table 3.1: Coefficients of determination, skewness and kurtosis for household-level residuals and municipal-level residuals of the working models for EBP with and without transformations

	Household-level residuals		Municipal-level residuals		
Transformation	Skewness	Kurtosis	Skewness	Kurtosis	R^2
Without	10.10	177.00	2.09	9.87	0.31
Log	-2.71	26.50	-0.60	3.52	0.43
Log-shift	0.00	4.91	-0.24	3.03	0.51
Box-Cox	-0.24	7.95	-0.12	3.00	0.49

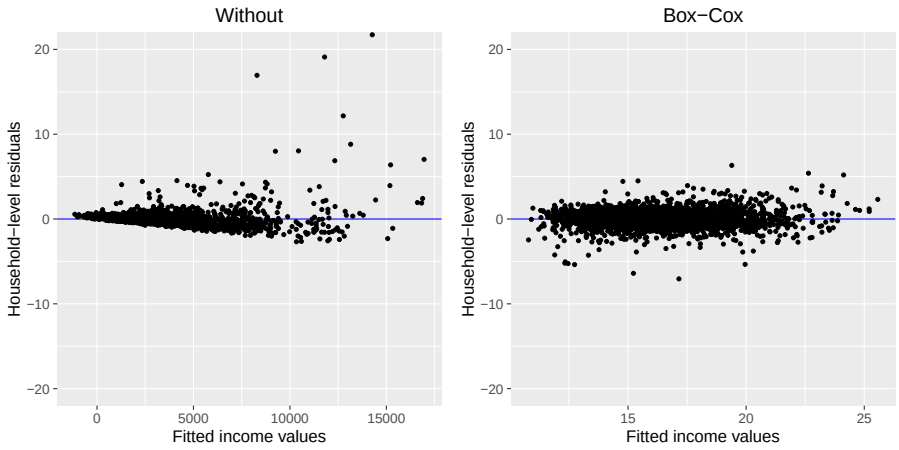


Figure 3.6: Standardized household-level residuals against fitted values without (left) and with Box-Cox transformation (right) for income

namely log-shift with shift parameter $\lambda = 319.52$ and Box-Cox with $\lambda = 0.17$ provide a better approximation to normality than the logarithmic transformation or the no transformation cases, albeit not perfect. In particular, the symmetry of the distribution of the residuals is improved but the tails of this distribution remain heavier than those of the standard normal one. The following questions are raised at this stage. How important is the choice of transformation in small area estimation? Does the improvement in the predictive power of the model with transformation and less severe departures from the model assumptions translate to more precise small area estimates on the original scale? Is the choice of transformation equally important for parameters associated with the centre of the distribution and parameters associated with tails of the distribution? We attempt to address these questions in Section 3.4 that presents an evaluation framework for SAE. For now, we comment on Figure 3.7 that show maps of point estimates of average income, Gini coefficients and HCR for municipalities in EDOMEX produced by the EBP approach using different transformations.

The maps for average income, Gini coefficient and HCR clearly indicate regional differences. As mentioned before, EDOMEX has 125 municipalities which by their geographic and demographic characteristics are grouped into 16 districts. The maps of the estimated income-based indicators for all transformations suggest intra-regional differences of poverty and inequality within and between the districts. Estimates of average income and HCR show that some of the wealthiest districts are concentrated in the central-east and northern zones of EDOMEX. The most unequal municipalities are located in the central and south-west parts

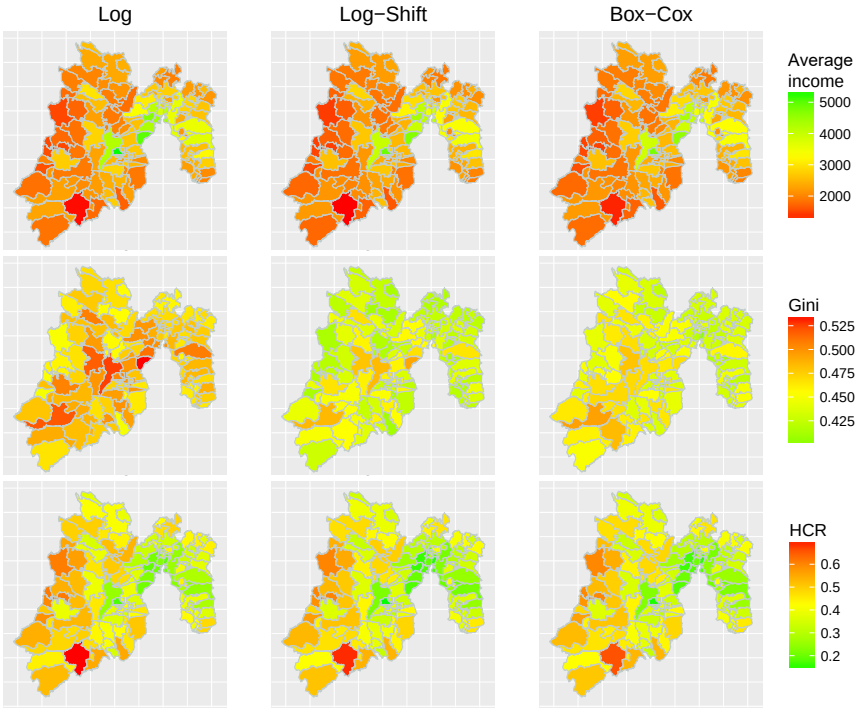


Figure 3.7: Map of municipal estimates of average income, Gini coefficients and HCR in EDOMEX using the EBP method under the log, log-shift and Box-Cox transformations

of EDOMEX. There are, however, some differences in the maps of point estimates produced with different transformations. Estimates of average income appear not to be affected by the choice of transformation. The same holds true to a large extent for estimates of HCR. On the other hand, estimates of the Gini coefficient appear to be more sensitive to the choice of transformation. These results suggest that the user should be very careful with the choice of transformation as this can have an impact on point estimation especially when interest is in non-linear indicators that depend on the entire distribution. We will return to this discussion at the end of Section 3.4.

3.4 Evaluation

The small area estimates are a set of numbers of identical definition and simultaneous interest. Evaluating the small area estimates is a relevant question for which there are hardly any definitive answers. For example, whether to measure the uncertainty using a design or a model-based MSE causes lively debates among researchers and practitioners. Comparing sets of optimal small area estimates produced under alternative models and deciding whether one set is better than another can be also a challenging task. Assessing ensemble properties of small area estimates such as the range or ranks of the estimates is relevant topic which has been largely overlooked. A detailed discussion on evaluation is beyond the scope of this paper. Our approach below is to describe some aspects of evaluation, which we believe should be taken into consideration in any application. In particular, we highlight the distinction between uncertainty assessment and method evaluation, which in our experience is a matter that is often

either misunderstood or overlooked. The purposes of each and the most common uses in SAE are described in Section 3.4.1 and 3.4.2, respectively. Some illustrations with the ENIGH data are given in Section 3.4.3.

3.4.1 Uncertainty assessment

Let θ_k be the target parameter of area k, for $k = 1, ..., m$. Let $\boldsymbol{\theta} = \{\theta_1, ..., \theta_m\}$ be the collection of them. Let $\hat{\theta}_k$ be the estimator of θ_k and $\hat{\boldsymbol{\theta}}$ the collection of them. We assume that one is equally interested in all elements of $\boldsymbol{\theta}$ and cannot fix only on one particular θ_k, or a few of them, and disregard how estimators perform in the rest of the areas.

The first question for uncertainty assessment is, "what is the target of estimation?", which refers back to the specification of the problem. Generally speaking, in small area estimation one may distinguish between the area-specific and ensemble targets of $\boldsymbol{\theta}$. An ensemble characteristic of $\boldsymbol{\theta}$ is defined by using all θ_k's. For example, let $\bar{\theta}_w = \sum_{k=1}^{m} N_k\theta_k/N$ be the population mean, where N_k is the population size in area k and $N = \sum_{k=1}^{m} N_k$, or let $G = \sum_{k=1}^{m}(\theta_k - \bar{\theta})^2/(m-1)$ be the dispersion (i.e. population variance) of $\boldsymbol{\theta}$, where $\bar{\theta} = \sum_{k=1}^{m} \theta_k/m$. Other examples include the range, the order statistics and the ranks of $\boldsymbol{\theta}$. Although the various ensemble target parameters may be very important for purposes such as benchmarking, subgroup analysis, fund allocation, evaluation and monitoring (see e.g. Ghosh, 1992; Shen and Louis, 1998), area-specific prediction seems to have been the focus in the majority of applications. The most common uncertainty measure for area-specific prediction is MSE. Below we explain the three types of MSE in use after which interval estimation will be briefly described.

Let y_k denote generically all the observed data in area k, for $k = 1, ..., m$. Let $\boldsymbol{y} = \{y_1, ..., y_m\}$ be the collection of them. Given a population model for $\boldsymbol{\theta}$, the (unconditional) MSE is given by $E[(\hat{\theta}_k - \theta_k)^2]$, where the expectation is over both $\boldsymbol{\theta}$ and $\boldsymbol{y}$. Prasad and Rao (1990) develop a second-order accurate analytic MSE estimator under the linear mixed model, which corrects the bias of the direct plug-in MSE estimator. Jackknife methods have been developed for the same purpose under a wider range of models (Jiang et al., 2002). Bootstrap (most commonly parametric) is more generally applicable, especially if either the target parameter or the performance measure is non-differentiable (Hall and Maiti, 2006; Pfeffermann and Correa, 2012), such as when the target parameter is a population quantile.

Using bootstrap is particularly relevant for uncertainty estimation of indicators such as the Gini coefficient and the HCR. For example, for the EBP method described in Section 3.3.2, simple unconditional MSE estimation uses the following parametric bootstrap, where the unknown model parameters are replaced by their estimates and treated as fixed. Generate B bootstrap populations using the fitted marginal model (3.6). Compute the population value of the target parameter from each bootstrap population, denoted by θ_k^*. From each bootstrap population select a bootstrap sample and compute bootstrap estimates of the target parameter, $\hat{\theta}_k^*$, by using the same method as used with the original sample. Finally, compute the average of the B squared bootstrap errors – defined as the difference between $\hat{\theta}_k^*$ and θ_k^* – as an estimate of the unconditional MSE. Notice that the procedure here is not second-order accurate, unlike the more sophisticated, but more computer intensive, bootstrap methods cited above. In case

of using a transformation, the bootstrap populations are generated using the model fitted to the transformed data but MSE estimates are computed at the end by back-transforming to the original scale. Estimation of the transformation parameter λ should be implemented for each bootstrap sample, hence capturing the variability due to its estimation.

According to Booth and Hobert (1998), the conditional MSE of prediction (CMSEP) is given by $E[(\hat{\theta}_k - \theta_k)^2|y_k]$, where the corresponding within-area y_k is held fixed, and the pairs (u_j, y_j) are independent across the areas, for $j = 1, ..., m$. They argue particularly for its use under the generalised linear mixed models, and elaborate their approach in terms of the linear predictor. When the model parameters are known, denoted by $\boldsymbol{\psi}$, the best predictor is $\tilde{\theta}_k = E(\theta_k|y_k; \boldsymbol{\psi})$, and the only natural measure of its uncertainty is the CMSEP that reduces to the variance $V(\theta_k|y_k; \boldsymbol{\psi})$. When the model parameters are estimated, denoted by $\hat{\boldsymbol{\psi}}$, the CMSEP is decomposed into two terms $V(\theta_k|y_k; \boldsymbol{\psi})$ and $E[(\hat{\theta}_k - \tilde{\theta}_k)^2|y_k; \boldsymbol{\psi}]$, where $\hat{\theta}_k = E(\theta_k|y_k; \hat{\boldsymbol{\psi}})$. The first term is evaluated with respect to u_k given y_k, and the second one with respect to $\hat{\boldsymbol{\psi}}$ that varies only with the rest y_j's, for $j \neq k$, given y_k, where u_k and $\hat{\boldsymbol{\psi}}$ are conditionally independent (Booth and Hobert, 1998). Lohr and Rao (2009) propose a second-order accurate jackknife estimator of the conditional MSE. For a practical example, Zhang (2009) applies the CMSEP to estimates of small area compositions subjected to informative missing data.

The third type of MSE we describe is given by $E[(\hat{\theta}_k - \theta_k)^2|\boldsymbol{\theta}]$, where only the observed data $\boldsymbol{y}$ are allowed to vary but the values of $\boldsymbol{\theta}$ are treated as fixed. The key difference from the two types of MSE above is that the set of small area parameters $\boldsymbol{\theta}$ are now held fixed, and for this reason one may refer to this MSE as the finite-population (FP) MSE. There are several variations of the FP-MSE in practice, where $\boldsymbol{\theta}$ may either be the actual population values or the theoretical values under a model, and the MSE may be evaluated with respect to the sampling design or a model for $\boldsymbol{y}|\boldsymbol{\theta}$. The FP-MSE becomes the well-known design-based MSE, when $\boldsymbol{\theta}$ are population quantities such as the area means and $\boldsymbol{y}$ vary according to the sampling design (e.g. Rivest and Belmonte, 2000). Often, however, simplifying assumptions are adopted, e.g. by assuming area-stratified simple random sampling with the observed area sample sizes treated as fixed, because one may not have access to the details required to implement the sampling design. Chambers et al. (2011) calculate the FP-MSE under the model for $\boldsymbol{y}|\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ are the theoretical area means rather than the population area means. Notice that these authors use the term "conditional" MSE, where it is the θ_k's that are treated as fixed not y_k as under the CMSEP. Finally, because the FP-MSE is a small area parameter itself, unbiased estimation is unstable whether it is with respect to the sampling design or model. Hence, one needs to treat the estimation of FP-MSE as a small area estimation problem in its own right.

Deciding which MSE to use is important. Tukey's remark on this matter is that one should "focus on the questions, not models" (Discussion of Nelder, 1977). There are times when the target parameter θ_k is of a theoretical nature. It is then quite appropriate to consider the u_k's as random variables, and to use the unconditional MSE or the CMSEP as the uncertainty measure. For instance, in life expectancy calculation one would first smooth the actual known death rates, which could only make sense if one considers the actual population death rate as an estimate of some unknown hypothetical parameter called mortality rate. But there are also many other situations, such as when θ_k is the area unemployment rate, where it is clearly defined as a

descriptive statistic of the given population. One can still treat u_k as a random effect in order to achieve a sensible bias-variance trade-off, e.g. using model (4) to motivate a choice of α_k in the composite estimator (3.3). Without introducing the random effects model, one would have to resort to other means for deriving α_k. However, we believe that while it is inferentially consistent to report the model-based MSE here, which treats $\boldsymbol{\theta}$ as random, one is entitled to question its relevance when θ_k is a descriptive statistic and the assumption $E(u_k) = 0$ may be doubtful for a given k. In such a case, the FP-MSE is attractive for many survey practitioners. However, as explained above, the estimation of the FP-MSE needs to be treated as a small area problem in its own right.

Finally, interval estimation may be considered in addition to MSE estimation. Let $C_k = (\hat{\theta}_{kL}, \hat{\theta}_{kU})$ be an interval estimator of θ_k, where $\hat{\theta}_{kL} < \hat{\theta}_{kU}$. The simplest procedure is to set the bounds such as $\hat{\theta}_k \pm 1.96 \cdot \hat{\text{MSE}}(\hat{\theta}_k)^{1/2}$, aimed at the 95% nominal confidence level. See Pfeffermann (2013, Section 6.2) for a review of interval estimation methods. Let $\delta_k = 1$ if $\theta_k \in C_k$ and 0 otherwise. Analogously to the unconditional MSE, the unconditional coverage of C_k is given by $\varsigma_k = E(\delta_k) = P(\theta_k \in C_k)$, where both $\boldsymbol{\theta}$ and $\boldsymbol{y}$ are allowed to vary. Similarly, one can speak about conditional coverage of C_k given by $E(\delta_k|y_k)$, and FP-coverage given by $E(\delta_k|\boldsymbol{\theta})$. Notice that any model-based C_k that treats θ_k as random can have rather erratic area-specific FP-coverage compared to the nominal level of confidence. Zhang (2007) defines $\varsigma = \sum_{k=1}^{m} E(\delta_k|\boldsymbol{\theta})/m$ to be the FP simultaneous coverage of all C_k, each aimed at the same nominal confidence level. For the population from which the sample is selected, this gives the proportion of area parameters that are expected to be covered by their interval estimates without specifying which areas these are. It is shown that, as m increases, ς converges to the nominal level, provided the underlying population model of $\boldsymbol{\theta}$ is correct.

3.4.2 Method evaluation

In the previous section we described different uncertainty measures. In addition to measuring the uncertainty associated with $\hat{\boldsymbol{\theta}}$ under the assumed model, an analyst may be interested in method evaluation. This might include comparing different point estimators, assessing how a MSE estimator performs in reality when approximations are used in its derivation, or assessing how a small area estimator behaves under departures from the underlying model assumptions. Method evaluation is generally a different matter from uncertainty assessment.

As we describe below, broadly speaking method evaluation can be design-based or model-based. It is also possible to combine both sources of uncertainty, where the distribution of $\boldsymbol{\theta}$ follows from a population model and the distribution of $\boldsymbol{y}$ from the sampling design. The evaluation can be performed analytically provided the required closed-form expressions can be derived. More often, both design-based and model-based simulation studies are used for method evaluation.

Conducting a design-based simulation study is very common in practice. Indeed, it is hard to imagine that an NSI will produce any small area statistics on a regular basis without validating the design-based performance of the adopted method under realistic conditions. Typically, a census or similar population dataset is fixed as the population from which samples are repeatedly taken. When such population data are unavailable, there are various proposals in

the literature on how one can generate a pseudo-population for the in-sample areas from the sample data at hand (e.g. Sverchkov and Pfeffermann, 2004). However, a model will be necessary in order to generate a pseudo-population for the out-of-sample areas. For each simulated sample, a given estimation method is applied to obtain a replicate set of small area estimates. Within a design-based simulation study different estimation methods or models can be directly compared to each other in terms of their design-based performances. We consider this to be a suitable approach for method evaluation, which establishes how a method is expected to perform over repeated sampling from a finite population, regardless of whether the underlying model is correct or not. Using the ENIGH data in Section 3.4.3 we provide a detailed description of how one can design and implement a design-based simulation that mimics the design and characteristics of the survey data.

Unlike in a design-based simulation study, where the different estimation methods are subjected to the same sampling variation and the population may be based on real data, model-based method evaluation generally requires the use of a model for generating the population. This is common when researchers develop new methods and they are interested in evaluating the properties of estimators. The design of model-based studies requires careful thinking about the choice of the evaluation model used for generating the population. A general question is whether it is meaningful to compare directly the MSE of an estimator $\hat{\theta}_{kA}$ of θ_k derived under model M_A to that of another estimator $\hat{\theta}_{kB}$ of θ_k under model M_B, which may involve different random effects or correlation structure. Notice that it is always possible to evaluate the MSE of $\hat{\theta}_{kA}$ under model M_B even though the estimator is motivated and computed under model M_A and vice versa. Since the MSE of $\hat{\theta}_{kA}$ will differ according to whether the evaluation model is M_A or M_B, there is a need to level the ground in order to avoid misleading comparisons. One may, for example, carry out simulation of both $\hat{\theta}_{kA}$ and $\hat{\theta}_{kB}$ under the model M_B if M_A is nested in M_B. When M_A and M_B are not nested in each other but are from the same class of models, one may use for the evaluation a model M_C which encompasses both. But it may not be obvious how to find an encompassing model when M_A and M_B belong to different classes of models.

It should be mentioned that, in addition to the methods described above, there are several informal evaluation approaches that are of relevance to practitioners, such as compatibility with external data, evaluation by subject-matter experts, bias and goodness of fit diagnostics, as described in Brown et al. (2001). Finally, a set of small area estimates is expected to be numerically consistent and more efficient than unbiased direct estimates. One can compare the aggregated area estimates to the corresponding direct estimates for the same purpose. If aggregated model-based (indirect) estimates do not agree with the corresponding direct estimates, an analyst can use benchmarking techniques to achieve consistency. Benchmarked small area estimates offer an attractive property for NSIs (see Ghosh and Steorts, 2013; Pfeffermann, 2013; Pfeffermann et al., 2014, for a discussion on benchmarking methods). A more challenging issue is benchmarking of aggregated ensemble properties, such as the population quantiles, which can be derived from the collection of within-area quantiles.

3.4.3 Illustrating aspects of SAE evaluation using the ENIGH data

In this section we illustrate some of the aspects of SAE evaluation we discussed in Sections 3.4.1 and 3.4.2. In particular, using the results of model selection and diagnostics we described in Section 3.3.3, we present results for the estimation of average household equivalised income, HCR and Gini coefficients for municipalities with the original sample in EDOMEX. We then show how the analyst can prepare a design-based simulation study that can be used for method evaluation. We discuss how the design-based simulation results can guide the production of the final set of SAE estimates.

Analysis with the original sample

Table 3.2 presents summaries over municipalities of point, root MSE (RMSE) and CV estimates computed using the original data supplied to us by CONEVAL and estimated MSEs under the assumed model. To start with, direct estimation is not considered because survey data cover only part of the target geography and - as we discussed in Section 3.3.3 - direct estimates have higher than acceptable estimated CVs. Results are presented separately for in-sample and out-of-sample areas. For in-sample areas we produce estimates using four versions of the EBP method i.e. with untransformed income and three transformations (Log, Log-shift and Box-Cox). For out-of-sample areas we use the four above-mentioned versions of the EBP, which in this case corresponds to synthetic estimation. MSE estimates are obtained by using the parametric bootstrap under the unit-level mixed models (see Section 3.4.1) and different transformations. The synthetic estimates are produced under the marginal model (3.6).

The results in Table 3.2 show that the EBP Log-shift and EBP Box-Cox produce small area estimates that are clearly more efficient than the corresponding estimates produced with the untransformed income model and more efficient than the log-income model. Hence, using the methods suggested by model building and diagnostic analysis results in estimates with better efficiency. It is also clear that failing to use transformations, when needed, has an impact on point estimation. The impact of transformations on point estimation is less pronounced for indicators that relate to the centre of the income distribution (average income) than for non-linear indicators such as the HCR and the Gini coefficient. However, even for average income, failing to transform has a substantial effect on the efficiency of the estimates. These results illustrate the importance of model diagnostics in SAE. A final comment about these results relates to MSE estimation. MSE estimates are produced by computing the parametric bootstrap estimator with the original sample. Parametric bootstrap relies on the belief that the model assumptions (after transformation) are met. In reality there are always departures from the model assumptions, the risk of which is uncontrollable for the out-of-sample areas in particular. One question is whether departures can have an impact on MSE estimation. Another question is whether the impact of model misspecification on MSE estimation is different for linear and non-linear indicators. The question becomes relevant when looking at the RMSE estimates for the Gini coefficient which are quite small. Evaluating MSE estimation subject to model misspecification is not easy. Using evaluation methods such as design or model-based simulations is essential. However, this can be very computer intensive because it requires bootstrap techniques to be embedded within a Monte-Carlo simulation framework. We discuss

this issue again in the next section.

Table 3.2: One sample analysis of income data. Median of point estimates, estimated RMSEs and CVs over municipalities in EDOMEX

	Municipalities	58 In-sample			67 Out-of-sample		
	Indicator	Mean	HCR	Gini	Mean	HCR	Gini
Point Estimates	EBP	2730	0.380	0.949	2042	0.436	1.261
	EBP Log	2699	0.363	0.477	2244	0.439	0.474
	EBP Log-shift	2600	0.329	0.433	2151	0.409	0.432
	EBP Box-Cox	2617	0.336	0.435	2171	0.409	0.440
RMSE	EBP	449.2	0.040	0.177	523.4	0.048	0.400
	EBP Log	249.7	0.039	0.011	256.1	0.050	0.013
	EBP Log-shift	202.3	0.036	0.010	209.3	0.048	0.011
	EBP Box-Cox	185.2	0.034	0.010	188.4	0.043	0.011
CV	EBP	0.163	0.104	0.187	0.251	0.114	0.313
	EBP Log	0.095	0.108	0.024	0.111	0.119	0.027
	EBP Log-shift	0.080	0.112	0.022	0.095	0.122	0.025
	EBP Box-Cox	0.071	0.103	0.022	0.085	0.110	0.025

Method evaluation using design-based simulation

In Section 3.4.3 above the MSE was calculated under the model estimated based on the ENIGH survey data. Naturally the user might be interested in knowing how the estimates will be affected if the model assumptions do not hold. Using design-based method evaluation that does not depend on the model assumptions can help with investigating this. We now illustrate an approach for setting up a design-based simulation that involves repeated sampling from a fixed population.

In a design-based simulation the first and possibly the most important step is deciding how to generate the fixed population from which we draw repeated samples. Sverchkov and Pfeffermann (2004) suggest generating a pseudo-population by using the sample data. In some cases a variable that is highly correlated with the target variable is available in the census. This is the case with the census data from Mexico for which we identified variable *inglabpc* - earned per capita income from work as being highly correlated with the variable of interest *ictpc*, which is only available in the survey data. Variable *inglabpc* does not have the desired income definition and this is why SAE using *ictpc* is needed. However, for the purposes of method evaluation we are interested in using a variable that has similar distributional characteristics as the target variable and *inglabpc* can play this role. A first reason as to why we decided not to include *inglabpc* as a covariate in our small area model is because we wanted to use this variable for evaluation purposes. A second reason is that we wanted to illustrate method evaluation in a situation where the covariates explain a moderate part of the variance. Table 3.3 presents summary statistics for $inglabpc$ (used in the design-based simulation) and $ictpc$ (used in the one sample analysis). The distribution of both variables is similar and the total per-capita income $ictpc$ is generally higher compared to per-capita income from work $inglabpc$. In fact, if anything, the census variable $inglabpc$ is even more skewed than the survey variable

$ictpc$, which seems reassuring with respect to the robustness of the evaluation using the census variable. Our design-based simulation will be based on repeated sampling from the Mexican census micro-data and modelling of proxy household income *inglabpc*.

Table 3.3: Summary statistics over municipalities

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
$inglabpc$ (census)	0	1000	1700	2717	3000	100000
$ictpc$ (survey)	0	1310	2142	3243	3518	98070
Population size	394	2759	6852	24820	16440	349100
Sample size	3	17	21	47.4	42	527

From the fixed population we independently drew $T = 500$ samples. The samples are selected by using a single-stage stratified random sampling with strata defined by the 58 in-sample municipalities in the ENIGH survey. The number of households in each in-sample municipality is the same as the number of households in the ENIGH survey. This leads to a sample size of 2748 households with 58 in-sample municipalities and 67 out-of-sample municipalities as is the case with the ENIGH survey. Summary statistics of the sample and population sizes -over municipalities- are provided in Table 3.3.

Using each sample selected from the fixed population we compute estimates of average equivalised household income from work, HCR and Gini coefficient. For in-sample areas we calculate the direct estimator (3.2), the EBP based on different transformations and the World Bank estimator (Section 3.3.2), which is denoted by WB in Table 3.4. As we mentioned in Section 3.3.2, for out-of-sample areas and when domains coincide with clusters, the EBP and the World Bank method coincide. All the models use the same six covariates identified in Section 3.3.3. The R^2 from linear regression models under different transformations (log, log-shift and Box-Cox) is around $40 - 50\%$ over the 500 samples, which is consistent with the results we obtained with the original sample.

The performance of these estimators is evaluated by computing the relative bias (RB) and root mean squared error (RMSE) given by

$$\text{Relative Bias}(\hat{\theta}_k) = \frac{1}{T}\sum_{t=1}^{T}\frac{\hat{\theta}_{tk} - \theta_k}{\theta_k}; \qquad \text{RMSE}(\hat{\theta}_k) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\theta}_{tk} - \theta_k\right)^2},$$

where $\hat{\theta}_k$ is generic notation to denote an estimator of the target parameter in municipality k, θ_k denotes the true population parameter in municipality k and t is an index for repeated sampling with $T = 500$ in this case. We further report CV as an additional performance indicator.

Table 3.4 reports the results split by the 58 in-sample and the 67 out-of-sample municipalities. The table presents median values of RMSE, relative bias and CV over municipalities. In line with the model diagnostics and the one sample analysis, the performance of the EBP estimates without transformation is inferior to the EBP estimates with transformations (log-shift and Box-Cox) for all indicators. The design-based simulation results confirm that transformations are necessary for improved small area estimation. As expected, the direct estimator is less efficient than model-based estimators, which justifies the use of indirect methods in this case.

Table 3.4: Performance of predictors over municipalities in design-based simulations

	Municipalities	58 In-sample			67 Out-of-sample		
	Indicator	Mean	HCR	Gini	Mean	HCR	Gini
RMSE	EBP	180.2	0.095	0.497	210.6	0.073	0.846
	EBP Log	187.5	0.049	0.026	216.3	0.061	0.032
	EBP Log-shift	156.6	0.038	0.022	200.7	0.062	0.031
	EBP Box-Cox	171.7	0.045	0.025	212.6	0.060	0.032
	WB	188.2	0.093	0.486	—	—	—
	WB Log	160.7	0.054	0.026	—	—	—
	WB Log-shift	159.4	0.041	0.022	—	—	—
	WB Box-Cox	168.5	0.051	0.025	—	—	—
	Direct	543.6	0.097	0.083	—	—	—
RB [%]	EBP	2.39	34.77	109.6	11.28	-0.69	152.6
	EBP Log	2.96	12.54	3.89	12.43	-5.27	2.25
	EBP Log-shift	0.93	6.49	0.08	11.19	-9.86	-0.21
	EBP Box-Cox	1.98	11.18	2.32	11.91	-6.60	1.09
	WB	2.79	34.45	110.1	—	—	—
	WB Log	1.84	16.65	3.89	—	—	—
	WB Log-shift	0.80	9.59	0.10	—	—	—
	WB Box-Cox	1.41	14.67	2.35	—	—	—
	Direct	-0.13	-0.35	-7.92	—	—	—
CV	EBP	0.082	0.262	0.534	0.109	0.179	0.693
	EBP Log	0.078	0.145	0.058	0.112	0.146	0.071
	EBP Log-shift	0.073	0.123	0.048	0.107	0.166	0.068
	EBP Box-Cox	0.076	0.137	0.056	0.110	0.154	0.071
	WB	0.088	0.260	0.530	—	—	—
	WB Log	0.072	0.174	0.058	—	—	—
	WB Log-shift	0.078	0.144	0.049	—	—	—
	WB Box-Cox	0.074	0.161	0.055	—	—	—
	Direct	0.239	0.291	0.203	—	—	—

A closer look at the EBP-based results with transformations shows that the EBP Log-shift and the EBP Box-Cox perform somewhat better compared to the EBP Log in terms of bias and efficiency for all indicators. This indicates that the log-shift and the Box-Cox transformations adapt better to the shape of the underlying distribution, which appears to be consistent with the results we obtained from diagnostic analysis (Section 3.3.3). Comparing the EBP Box-Cox and the EBP log-shift in detail we note that in general neither transformation has superior performance over the other. Additional (model-based) simulation studies are necessary for comparing the performance of the Box-Cox transformation and the log-shift transformation. However, this is beyond the scope of the present paper but we refer to some research in this direction by Rojas-Perilla et al. (2017). For in-sample areas we note that the WB estimates are somewhat less efficient than the EBP estimates. On the one hand, despite the relatively small between-area variability, including random effects is recommended for the in-sample municipalities. This can be seen from the increased biases of synthetic estimation for the out-of-sample areas. On the other hand, the relatively small difference between the WB and EBP

estimates highlights the importance of building a model that has a good fixed effects predictor. Doing so is of course also critical for the out-of-sample areas.

It is important to evaluate the performance of MSE estimators. Formal evaluation requires using parametric bootstrap with each of the 500 samples, which is very computer intensive and beyond the scope of the present paper. Nevertheless, practitioners must be particularly careful when using parametric MSE estimation methods and, in our view, they should always employ design-based method evaluation.

Finally we would like to give an illustration of informal evaluation. Comparing model-based estimates with corresponding design-based estimates for aggregated geographical levels can provide an indication about the quality of model-based estimates. As the Gini coefficient cannot be split into a weighted sum of sub-area Gini coefficients, we focus on average income. The State of Mexico consists of 125 municipalities and 16 districts. The maximum sample size in a district is 749 households, the minimum is 18 households, the mean is 172 households and the median is 150 households per district. As the sample size is still quite small for some districts, we compare model-based estimates with design-based estimates only for 13 districts for which design-based estimates have a CV below 30%. Figure 3.8 shows point estimates for district-level average household equivalised income using the direct estimator (black line) and the EBP estimators with log (blue line), log-shift (orange line) and Box-Cox (red line) transformations. The direct estimates are produced by using the district-specific samples. In contrast, the district-specific model-based estimates are aggregated from the corresponding municipality level estimates. For the aggregation we used weights defined by N_i/N, where N_i denotes the municipality population size. On the x-axis, districts are ordered by the CVs of the direct estimates (descending order from left to right). We observe that for districts where the direct estimates are more unreliable (left part of the plot), the model-based estimates are further from the direct estimates whereas for districts where the design-based estimates are more reliable (right part of the plot), the EBP Box-Cox and EBP Log-shift tend to be closer to the direct estimates. The correlation between the direct and the EBP Box-Cox and EBP Log-shift estimates is also slightly higher than the correlation between the direct and the EBP Log estimates. We should emphasise that this is an informal approach to evaluating the quality of model-based estimates and there is no rule of thumb as to what is an acceptable level of correlation between model and design-based estimates. An alternative is to average the direct estimates and the corresponding model-based estimates over the smallest 8 districts, and the largest 8 districts, and compare the numbers, as an indication of the potential bias. The use of cross-validation, where some areas are left out of fitting the model and model-based estimates for these areas are compared with design-based estimates, offers a more structured approach to evaluation.

3.5 An Update on SAE Software

In this section we provide a update on the availability of SAE software. Although from an applied point of view many NSIs have a preference for software such as SAS, most of the recent developments in SAE are implemented in the open-source software R (R Core Team,

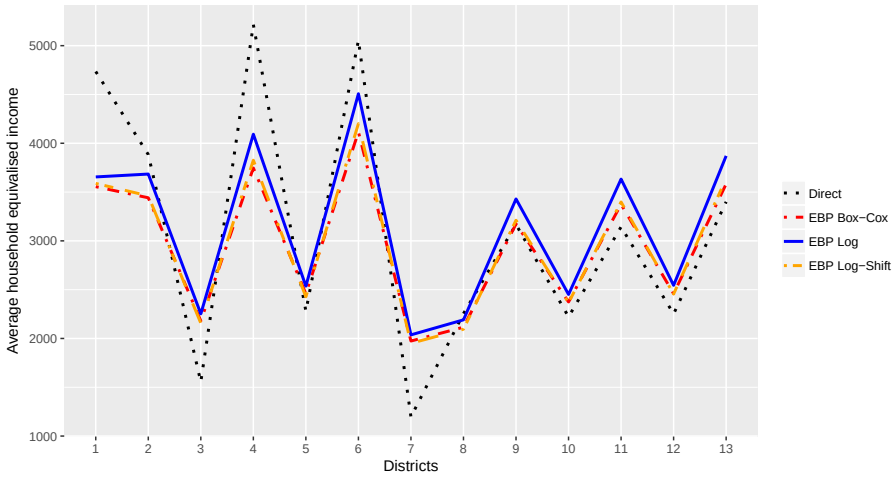


Figure 3.8: Estimates for average household equivalised income at district level.

2017) via R packages.

A comprehensive review of relevant software is included in the CRAN task view on *Official Statistics and Survey Methodology* (Templ, 2015) with specific categories on *Complex Survey Designs*, *Small Area Estimation* and *Microsimulations*. In particular, the section on *Complex Survey Designs* includes packages, like survey (Lumley, 2012) and sampling (Tillé and Matei, 2012) that can be used for point and variance estimation of direct estimators of means, totals, ratios, and quantiles under complex survey designs. Package laeken by Alfons and Templ (2013) provides functions for the estimation of different poverty and inequality indicators such as the at-risk of poverty-rate, Gini coefficient and quintile share ratio and the corresponding estimates of the variance. The sae package by Molina and Marhuenda (2015) can be used for computing synthetic and composite estimators and for implementing SAE with unit-level and area (Fay-Herriot) models that allow for complex correlations structures. A code in R for computing EBP estimates we discussed in Section 3.3.2 that includes an option for using the transformations discussed in the present paper, visualization and export of the results to Excel is proposed in the package emdi by Kreutzmann et al. (2018). Collections of R functions for implementing a wide range of SAE methods are available in the documentations of National and European funded research projects. Here we refer to the BIAS project (BIAS, 2005) which includes code for the unit-level EBLUP and spatial EBLUP with correlated random effects (Pratesi and Salvati, 2009). The SAMPLE project (SAMPLE, 2007) also provides a very wide range of code for implementing parametric, semi-parametric and outlier-robust small area estimation and allows for models with spatial and temporal correlations. We refer to Molina et al. (2010) for additional details. Small area estimation from a Bayesian perspective is provided in the packages hbsae (Boonstra, 2012) and BayesSAE (Shi and with contributions from Peng Zhang, 2013). It is also important to mention two packages namely, simPop (Meindl et al., 2016) and saeSim (Warnholz and Schmid, 2016) that support the prospective user in the setup of design- or model-based simulations that enable method evaluation at the evaluation stage.

In addition to software written in R, alternative SAE software is also available. The World

Bank provides open-source software for poverty estimation called PovMap (The World Bank, 2013). PovMap implements the small area estimation procedure developed in Elbers et al. (2003) and is stand-alone software solution. The European funded project EURAREA (2001) delivered SAS codes for the computation of direct and indirect small area methods. For additional procedures in SAS we refer to Mukhopadhyay and McDowell (2011). Finally, all methods discussed in the paper are implemented by computationally efficient algorithms using R. The codes are available from the authors upon request.

3.6 Conclusions and Future Research Directions

In this paper we propose a general framework for the production of SA statistics and illustrate the SAE process in practice. As part of this framework we have touched upon three interrelated topics, namely specification of the problem, analysis of the data/ adaptation of the model, and method evaluation. While much can be said for each of these three areas, it is the interplay between them that provides the key to the successful application of SAE methods. There are no clear-cut ways of trading between them in a formal manner and mastering a balance between these three stages is in many ways the wisdom of applied statistics, which holds true also for SAE. We have illustrated some practical ways of keeping this balance. It is shown that specifying a sensible geography and defining targets of estimation that are supported by the data available are the first important steps for successful SAE. Careful model building using the principle of parsimony, model diagnostics and model adaptations are crucial steps for improving estimation without the need for additional data sources. Finally, obtaining uncertainty measures of good quality and designing method evaluation studies are of paramount importance for reassuring the users especially if interest is in using the estimates for official purposes, for example in the design of policy interventions. SAE is of course a large research area and hence it is not possible to capture all of its aspects in a single paper. Production of SA statistics with discrete outcomes and use of area level models are not covered although the proposed framework can be applied in most cases.

Nevertheless, there are questions that remain unresolved and which we would like to raise at this stage. Within the context of sample surveys there exists currently an apparent contrast between the prevalent preference for design-based approaches to statistics at the higher levels of aggregation and model-based approaches at the lower levels. This seems to imply that at some intermediate level of aggregation the choice between the two approaches may be somewhat blurred. Where are these intermediate levels of aggregation? Is it possible to develop a coherent framework for the different levels in the aggregation hierarchy? Should benchmarking towards aggregate-level estimates of acceptable quality actively drive the development of SAE methods or should benchmarking, as often it is, remain a side issue that one only pays attention to at the last stage of estimation?

Both area-specific and ensemble properties of a set of small area estimates are undoubtedly of interest. This is a distinctive feature of SA statistics in comparison to the national estimate that is a single number. Small area estimation is a simultaneous rather than a point estimation problem. Multi-purpose (multiple-goal) SAE aims to provide a compromise in a theoretical

manner. However, the usefulness of such an approach can only be explored together with users if the solution is to have an impact in practice. Can users ever be ready or willing to accept multiple sets of estimates, each optimal for a particular purpose? How can one avoid or limit the misuses of a particular set of estimates in practice? For now we leave these questions open, hoping that they will inform future discussions.

Acknowledgements

First of all, the authors are indebted to the Editor, Associate Editor and referees for comments that significantly improved the paper. Tzavidis, Zhang, Luna, Schmid and Rojas-Perilla gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council. The authors are grateful to CONEVAL for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods.

Chapter 4

Data-driven Transformations in Small Area Estimation

4.1 Introduction

Model-based methods for small area estimation (SAE) are now widely used in practice for producing reliable estimates of linear and non-linear indicators for areas/domains with small sample sizes. Examples of indicators that are estimated by using model-based methods include poverty (income deprivation) and inequality measures such as the head count ratio, the poverty gap and the income quintile share ratio. Two popular small area methods in this case are the empirical best predictor (EBP), proposed by Molina and Rao (2010) and the World Bank method, proposed by Elbers et al. (2003). Both approaches are based on the use of unit-level linear mixed regression models. Although estimation of complex indicators can be also implemented with area-level linear mixed regression models (Fabrizi and Trivisano, 2016; Schmid et al., 2017), in this paper we focus on unit-level linear mixed regression models. In the original paper, Molina and Rao (2010) assumed that the error terms of the linear mixed regression model follow a Gaussian distribution. In case the model error terms significantly deviate from normality, the EBP estimator can be biased. What are the options available to the data analyst when the normality assumptions are not met? One option is to formulate the EBP under alternative and more flexible parametric assumptions. Graf et al. (2014) study an EBP method under the generalized beta distribution of the second kind (GB2), whereas Diallo and Rao (2014) propose the use of skewed-normal distributions in applications with income data. One complication with using the EBP under alternative parametric distributions is that new tools for estimation must be developed and training for the data analyst is needed. In addition, misspecification of the model assumptions is still possible. Another option when the Gaussian assumptions are not satisfied is to use a methodology that minimizes the use of parametric assumptions. For instance, Elbers and van der Weide (2014) proposed an EBP method based on normal mixture models. With this method the distribution of the error terms is described by normal mixtures. Weidenhammer et al. (2014) recently proposed a method that aims at estimating the quantiles of the empirical distribution function of the data. The estimation of the quantiles is facilitated by a nested error regression model using the asymmetric Laplace distribution for the unit-level error terms as a working assumption. The estimation of the random effects can be made com-

pletely non-parametric by using a discrete mixture proposed by Marino et al. (2016). Another option, and the one we study in this paper, is to find an appropriate transformation such that the model assumptions (in this paper the Gaussian assumptions of the EBP method) hold. The aim is to find transformations that (a) are data-driven and optimal according to some criterion and (b) can be implemented by using standard software. To the best of our knowledge, the use and choice of transformations in SAE has not been extensively studied or it has been studied in fairly ad-hoc manner. Elbers et al. (2003) and Molina and Rao (2010) suggested the use of logarithmic-type transformations for income data. However, are such transformations the most appropriate choice? Can alternative transformations offer improved estimation? In order to answer these research questions, the paper investigates data-driven transformations for small area estimation.

The choice of transformations when modelling income-type outcomes - as is the case with poverty mapping applications - presents different challenges. Transformations should be suitable for dealing with unimodal, leptokurtic and positively skewed data that may include zero and negative values. Besides the logarithmic transformation and its modifications (e.g. the log-shift transformation) a popular family of data-driven transformations that includes the logarithmic one as a special case is the Box-Cox family (Box and Cox, 1964). Since the Box-Cox transformation is not defined for negative values, when negative values are present, the data must be shifted to the positive range. Another difficulty with the use of the Box-Cox transformation is the truncation on the transformation parameter described later in Section 4.4. A solution to this problem can be offered by using of the dual power transformation. Although extensive literature on the use of transformations exists, see for example, John and Draper (1980), Bickel and Doksum (1981) and Yeo and Johnson (2000) among others. In this paper we focus on three types of transformations, namely log-shift, Box-Cox and dual power transformations.

In addition to selecting the type of transformation, estimating the transformation parameter adds another layer of complexity. To the best of our knowledge the use of transformations in recent applications of SAE has employed visual residual diagnostics for finding a suitable transformation parameter. In this paper we propose a structured, data-driven approach for estimating the transformation parameter. In particular, we introduce maximum likelihood and residual maximum likelihood methods for estimating the transformation parameter under the linear mixed regression model following Gurka et al. (2006). Alternative estimation approaches based on the minimization of distances (Cramér, 1928; Chakravarti and Laha, 1967) and on the minimization of the skewness (Carroll and Ruppert, 1987) are also discussed.

We study how the performance of the EBP method is affected by departures from normality and how data-driven transformations can assist with improving the validity of the model assumptions and estimation. Emphasis is given to the estimation of poverty and inequality indicators due to their important socio-economic relevance and policy impact. We further study whether the impact of departures from Gaussian assumptions is different depending on the target of estimation. For instance, departures from normality may have lesser impact on estimates of median income compared to estimates indicators that are more sensitive in the data distribution. The estimation for the latter indicator heavily depends on the entire distribution of the data. A parametric bootstrap for mean squared error (MSE) estimation under transformation is

studied and a wild-type bootstrap that may offer protection in the presence of departures from the Gaussian assumptions after transformations is also proposed.

The rest of the paper is structured as follows. The EBP approach is introduced in Section 4.2. Section 4.3 presents the survey data we use in this paper and makes the case, via the use of residual diagnostics, for using transformations. In Section 4.4 selected transformations are introduced and extended for their use with model-based SAE methods under the linear mixed regression model. This section includes the theoretical details about the choice of an appropriate scale and estimation of the transformation parameter. MSE estimation is discussed in Section 4.5. In Section 4.6 the proposed methods are applied to data from Guerrero in Mexico for estimating a range of deprivation and inequality indicators and corresponding estimates of uncertainty. In Section 4.7 the proposed methods are further evaluated by realistic - for income data - model-based simulations. Section 4.8 summarizes the main findings and outlines further research.

4.2 The Empirical Best Prediction (EBP) Method

Let U denote a finite population of size N partitioned into D areas or domains (representing the small areas) $U_1, U_2, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, where $i = 1, \ldots, D$ refers to the ith area. Let y_{ij} be the target variable defined for the jth individual belonging to the ith area, with $j = 1, \ldots, N_i$. Denote by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^T$ the design matrix containing p explanatory variables and define by s as the set of sample units, with s_i the in-sample units in area i and by r be the set of non-sampled units, with r_i the out-of-sample units in area i. Let n_i denote the sample size in area i with $n = \sum_{i=1}^{D} n_i$. Hence, we define by $\mathbf{y}_i$ a vector with population elements of the target outcome for area i partitioned as $\mathbf{y}_i^T = \left(\mathbf{y}_{is}^T, \mathbf{y}_{ir}^T\right)$, where $\mathbf{y}_{is}$ and $\mathbf{y}_{ir}$ denote the sample elements s and the out-of-sample elements r in area i respectively. Let us now describe in more detail the EBP approach by Molina and Rao (2010), which is the methodology we focus on in this paper. Under this approach census predictions of the target outcome are generated by using the conditional predictive distribution of the out-of-sample data given the sample data. The point of departure is the standard parametric unit-level linear mixed regression model, which is also known as the unit-level nested error regression model. This is defined by Battese et al. (1988) as:

$$y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2) \quad \text{and} \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \tag{4.1}$$

where u_i, the area-specific random effects, and e_{ij}, the unit-level error, are assumed to be independent. Assuming normality for the unit-level error and the area-specific random effects, the conditional distribution of the out-of-sample data given the sample data are also normal. A Monte Carlo approach is used to obtain a numerically efficient approximation to the expected value of this conditional distribution as follows:

1. Use the sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and the weighting factors $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

2. For $l = 1, \ldots, L$:

(a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:

$$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)},$$

where the predicted random effect $\hat{u}_i$ is defined as $\hat{u}_i = E(u_i|\mathbf{y}_{is})$.

(b) Calculate the indicator of interest $I_i^{(l)}$ in each area.

3. Finally, take the mean over the L Monte Carlo runs in each area to obtain a point estimate of the indicator of interest:

$$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}.$$

As is common in real applications, some areas are out-of-sample. For those areas, we cannot estimate an area-specific random effect, and hence the corresponding area-specific random effect is set equal to zero. Synthetic values of the outcome for the out-of-sample areas are then generated under the linear mixed regression model as follows:

$$y_{ij}{}^{(l)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_i^{(l)} + e_{ij}^{(l)},$$

with $u_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$. Finally, a parametric bootstrap - under the assumed model - is used for the MSE estimation. This is discussed in some detail in Section 4.5. Assuming normality for the error terms is a convenient assumption as allows the conditional distribution of $\mathbf{y}_r|\mathbf{y}_s$ to be derived. However, in applications that involve modelling an income-type outcome, as is the case in this paper, assuming normality is unrealistic. If our primary target is to develop a methodology that can easily be used in practice, finding appropriate data transformations is important.

4.3 The Guerrero Case Study: Data Source and Initial Analysis

In this section, we describe the data sources used in the application and provide a motivation for the use of transformations. The case study was carried out in the open-source software R (R Core Team, 2017).

The data used in this paper come from Mexico, which has one of the largest economies in Latin America and is still among the most unequal countries in the world (The World Bank, 2017). For tailored policies against deprivation it is necessary to have a detailed description of the spatial distribution of inequality and income deprivation. According to the general social development law in Mexico, the National Institute of Statistics and Geography (INEGI) has to provide measures at the national, state and municipal-levels. For carrying out the analysis in this paper, the statistical and geographical information was provided by INEGI through the Household Income and Expenditure Survey (ENIGH) 2010 and the National Population and Housing Census of 2010. Looking in more detail at the data available and their geographic coverage, Mexico is divided into 32 federal entities (states). The state Guerrero has been considered by the World Bank to be one of the entities that mostly contributes to inequality

in Mexico, presenting a high inequality in human development (Bedoya et al., 2013). Additionally, according to the United Nations Development Programme (UNDP), this region has one of the highest rates of poverty and lack of infrastructural development (Tortajada, 2006). Guerrero comprises 81 administrative divisions, known as municipalities. From the 81 municipalities 40 municipalities with 1611 households are in-sample (in the sample of the ENIGH survey), leaving the remaining 41 municipalities out-of-sample. For the in-sample municipalities the maximum sample size in a municipality is 511, the minimum is 9 and the median is 24 households. Note that more than 30% of the sample is from a single municipality, the capital (Chilpancingo de los Bravo).

The survey and census data include a large number of socio-demographic variables, which are common and are measured similarly in both data sources. The total household per capita income (*hciw*, measured in pesos) is a variable recorded for households and is available in the survey but not in the census. We used this variable as a proxy that best approximates the living standard in Guerrero and as the outcome variable in our models. Socio-economic variables available for the households both in the survey and census data are used as explanatory variables. The underlying linear mixed regression model (4.1) of the EBP has two levels, households and municipalities. The variables available in the survey and census data, which are identified by using Bayesian information criterion (BIC) as good predictors of *hciw*, are described in Table 4.1. From now on, the working model is assumed to be known and fixed.

Table 4.1: Description of the explanatory variables used in the working model

Determinant	Variable
Occupation	1) Indicator if the head of household and the spouse are employed
	2) Type of household occupation
	3) Total number of employees older than 14 years in a household
	4) Percentage of employees older than 14 years in a household
Sources of income	5) Indicator of a household receiving remittances
Socioeconomic level	6) Availability of assets in the household
	7) Total number of goods in the household
Education	8) Average standardized years of schooling (by age and sex) within the household relative to the population

The next step after the identification of a possible set of covariates is assessing the predictive power of the model. Nakagawa and Schielzeth (2013) propose the use of two coefficients of determination suitable for linear mixed regression models: (a) the marginal R^2_m, which is a measure for the variance explained by fixed effects and (b) the conditional R^2_c, which measures the variance explained by both, the fixed and random effects. Without using any transformation, these measures are both around 34% and the corresponding intraclass correlation (ICC) under the model is 0.02.

In order to explore the validity of the Gaussian assumptions underlying the linear mixed regression model, it is common practice to perform normality tests and some residual diagnostics. The p-values of the Shapiro-Wilk (S-W) test statistic are equal to $2.2 \cdot 10^{-16}$ for the household-level and 0.002 for the municipal-level. These results indicate that the null hypothesis of normality for both terms is rejected. As normality tests like Shapiro-Wilk have some

problems we also present some visual approaches in addition. Figure 4.1 presents the Normal probability quantile-quantile (Q-Q) plots for household-level and municipal-level residuals. As expected, in the case of using the non-transformed *hciw* variable, the shape of the Q-Q plots is clearly different from what would be expected under normality. In addition, the analysis of skewness and kurtosis for both error terms is also informative. The skewness and kurtosis for a Normal distribution are equal to zero and three, respectively. The skewness and kurtosis of the household-level are equal to 7.980 and 110.700, and for the municipal-level equal to 1.298 and 5.596. These results indicate severe departures from the Gaussian assumptions when modelling the non-transformed income.

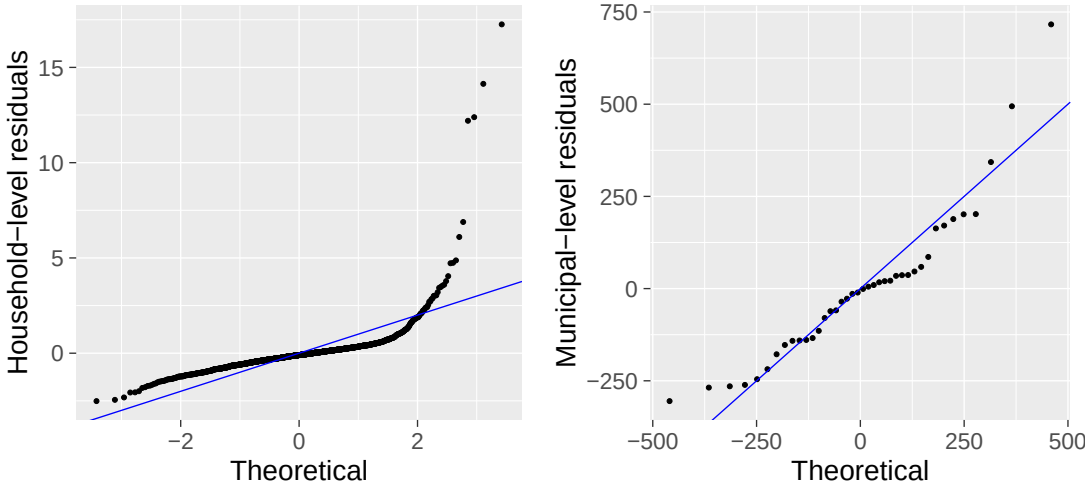


Figure 4.1: Q-Q plots of the household- and municipal-level error terms

4.4 Use of Transformations

In order to get closer to normality, it is common to use a one-to-one transformation $T(y_{ij}) = y^*_{ij}$ of the target variable. The application of the natural logarithmic transformation, which is a popular choice for income data, leads in many cases from right-skewed to more symmetric distributions. This is the most frequently used transformation in different research fields for dealing with non-normality due to its simplicity. However, can an alternative transformation with data-driven parameter λ, $T_\lambda(y_{ij}) = y^*_{ij}(\lambda)$, possibly offer small area estimates with improved precision?

The structure of the section is as follows. In Section 4.4.1 we introduce the EBP approach with data-driven transformations. In Section 4.4.2 we propose likelihood-based approaches for estimating the transformation parameter, λ, in general and discuss three particular subcases - the log-shift, Box-Cox and dual power transformations - in detail. Finally, in Section 4.4.3 we discuss alternative to likelihood-based approaches for estimating the transformation parameter.

4.4.1 EBP under transformations

In order to apply the EBP method by using transformations, the linear mixed regression model is re-defined as follows:

$$y_{ij}^*(\lambda) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}, \quad u_i \overset{iid}{\sim} N\big(0, \sigma_u^2\big) \text{ and } e_{ij} \overset{iid}{\sim} N\big(0, \sigma_e^2\big). \tag{4.2}$$

The EBP approach under transformations can be re-written as follows:

1. Select a transformation and obtain $T_\lambda(y_{ij}) = y_{ij}^*(\lambda)$.
2. Use the transformed sample data to obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2$ and calculate the weighting factors, $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.
3. For $l = 1, \ldots, L$:
   (a) Generate $v_i^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i)\big)$ and $e_{ij}^{(l)} \overset{iid}{\sim} N\big(0, \hat{\sigma}_e^2\big)$ and obtain a pseudo-population of the target variable by:
   $$y_{ij}{}^{*(l)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^{(l)} + e_{ij}^{(l)}.$$
   (b) Back-transform $y_{ij}^{*(l)}$ to the original scale $y_{ij}^{(l)} = T_\lambda^{-1}\left(y_{ij}^{*(l)}\right)$.
   (c) Calculate the indicator of interest $I_i^{(l)}$ in each area.
4. Finally, take the mean over the L Monte Carlo generations in each area to obtain an estimate of the indicator of interest:
   $$\hat{I}_i^{EBP} = \frac{1}{L}\sum_{l=1}^{L} I_i^{(l)}.$$

4.4.2 Likelihood-based approach for estimating λ

For estimating the transformation parameter λ, the linear mixed regression model defined in (4.2) is used. Assume that the transformed vectors $\boldsymbol{y}_i^*$ are independent and normally distributed for some unknown λ,

$$\boldsymbol{y}_i^*(\lambda) \sim N(\boldsymbol{\mu}_i, \mathbf{V}_i) \quad \text{for} \quad i = 1, \ldots, D,$$

where

$$\boldsymbol{\mu}_i = \mathbf{X}_i\boldsymbol{\beta} \quad \text{and} \quad \mathbf{V}_i = \sigma_u^2\mathbf{1}_{N_i}\mathbf{1}_{N_i}^{'} + \sigma_e^2\mathbf{I}_{N_i},$$

with $\mathbf{1}_{N_i}$ a column vector of ones of size N_i and $\boldsymbol{I}_{N_i}$ the $N_i \times N_i$ identity matrix, the vector of unknown model parameters is $\boldsymbol{\theta}^T = (\boldsymbol{\beta}, \sigma_u^2, \sigma_e^2, \lambda)$. The log-likelihood function under the

model is defined as follows:

$$\begin{aligned} l_{\text{ML}}(\boldsymbol{y}^*, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]. \end{aligned}$$

The log-likelihood function in relation to the original observations is obtained by multiplying the normal density by the log of the Jacobian of the transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$. The Jacobian $J(\lambda, \boldsymbol{y})$ is defined as $\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right|$ and is incorporated as follows:

$$\begin{aligned} l_{\text{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] + \log J(\lambda, \boldsymbol{y}). \end{aligned}$$

The maximization of $l_{\text{ML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta})$ produces maximum likelihood (ML) estimates of the unknown parameters $\boldsymbol{\theta}$. However, in the theory of linear mixed regression models, when interest focuses on accurate estimators of the variance components, restricted maximum likelihood (REML) theory is recommended (Verbeke and Molenberghs, 2000). The REML is defined as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] + \log J(\lambda, \boldsymbol{y}). \end{aligned} \quad (4.3)$$

To take advantage of existing algorithms for fitting mixed linear regression models, we use a scaled transformation defined by $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{\frac{1}{n}}} = z_{ij}^*(\lambda)$. The Jacobian of the scaled transformation is equal to 1 and hence standard software for mixed models can be used for maximizing $l_{\text{REML}}(\boldsymbol{z}^*, \lambda|\boldsymbol{\theta})$. The use of scaled transformations aids the implementation of the methods in practice. However, appropriate scaling factors must be developed depending on the type of transformation used.

Although the theory is applicable to data-driven transformations in general, we focus on three types of transformations, namely log-shift, Box-Cox and dual power transformations as particular subcases. The log-shift transformation (Yang, 1995) extends the logarithmic transformation by including the transformation parameter λ as follows:

$$y_{ij}^*(\lambda) = \log(y_{ij} + \lambda).$$

When $\lambda = 0$, a logarithmic transformation is obtained. The Box-Cox transformation (Box and Cox, 1964) is defined as follows:

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_{ij}+s)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s denotes a fixed parameter such that $y_{ij} + s > 0$. If $\lambda = 0$, the logarithmic transformation is then a special case and if $\lambda = 1$, the data are only shifted. One difficulty with the Box-Cox type transformations is the long-standing truncation, i.e. $y_{ij}^*(\lambda)$ is bounded, from below by $\frac{1}{\lambda}$ if $\lambda > 0$ and from above by $\frac{-1}{\lambda}$ if $\lambda < 0$. This is the key motivation for the third type of transformation. The dual power transformation, introduced by Yang (2006), is defined as follows:

$$y_{ij}^*(\lambda) = \begin{cases} \frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0, \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

where s is defined as in the case of the Box-Cox transformation.

The corresponding Jacobian used in (4.3) and scaled versions of the log-shift, Box-Cox and dual power transformations are presented in Table 4.2. For more details we refer to the developments in Appendix .1.

Table 4.2: Jacobian and scaled data-driven transformations for log-shift, Box-Cox and dual

Transformation	Jacobian J	Scaled transformation $z_{ij}^*(\lambda)$
Log-Shift	$\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+\lambda)^{-1}$	$J^{\frac{-1}{n}}\log(y_{ij}+\lambda)$
Box-Cox	$\prod_{i=1}^{D}\prod_{j=1}^{n_i} y_{ij}^{\lambda-1}$	$J^{\frac{-1}{n}}\frac{(y_{ij}+s)^\lambda-1}{\lambda}$, if $\lambda \neq 0$ $J^{\frac{-1}{n}}\log(y_{ij}+s)$ if $\lambda = 0$
Dual	$\frac{\prod_{i=1}^{D}\prod_{j=1}^{n_i}\left((y_{ij}+s)^{\lambda-1}+(y_{ij}+s)^{-\lambda-1}\right)}{2}$	$J^{\frac{-1}{n}}\frac{(y_{ij}+s)^\lambda-(y_{ij}+s)^{-\lambda}}{2\lambda}$, if $\lambda \neq 0$ $J^{\frac{-1}{n}}\log(y_{ij}+s)$ if $\lambda = 0$

4.4.3 Alternative approaches for estimating λ

The ML and REML approaches introduced in 4.4.2 rely on parametric assumptions that may be influenced by outliers in the data. The kurtosis and skewness are crucial features for defining the shape of a normal distribution and a proximity measure can be minimized in order to find a transformation parameter under which the empirical distribution of residuals has skewness and kurtosis as close as possible to zero and three respectively. In general, skewness is considered more important than kurtosis, therefore, minimizing the skewness is an approach already considered in the literature (Royston et al., 2011) for linear models as follows:

$$\hat{\lambda}_{\text{skew}} = \underset{\lambda}{\text{argmin}}\, |S_{e_\lambda}|,$$

where S_{e_λ} is the skewness and $\sigma^2_{e_\lambda}$ denotes the variance of the unit-level error terms. Note that the index λ is used to emphasize that the skewness and the variance parameters depend on the transformation parameter. In the context of linear mixed regression models, an additional problem arises as there are two independent error terms to be considered. We propose a pooled skewness approach that uses a weight w to ensure that the larger the error term variance $\sigma^2_{e_\lambda}$ is, the more weight its skewness will have in the minimization. Let S_{u_λ} be the skewness and $\sigma^2_{u_\lambda}$ be the variance of the area-specific random effects u_i of the linear mixed regression model. The estimation criteria in the pooled skewness approach is defined as follows:

$$\hat{\lambda}_{\text{poolskew}} = \underset{\lambda}{\text{argmin}} \Big(w|S_{e_\lambda}| + (1-w)|S_{u_\lambda}| \Big),$$
$$\text{where } w = \frac{\hat{\sigma}^2_{e_\lambda}}{\hat{\sigma}^2_{u_\lambda} + \hat{\sigma}^2_{e_\lambda}}.$$

Considering only the skewness may ignore other properties of the distribution. Hence, a measure describing the distance between two distribution functions is another alternative. Two distance measures, the Kolmogorov-Smirnov (KS) and the Cramér-von Mises (CvM) are used,

$$\hat{\lambda}_{\text{KS}} = \underset{\lambda}{\text{argmin}} \ \sup \ |F_n(\cdot) - \Phi(\cdot)|,$$
$$\hat{\lambda}_{\text{CvM}} = \underset{\lambda}{\text{argmin}} \int_{-\infty}^{\infty} [F_n(\cdot) - \Phi(\cdot)]^2 \, \phi(\cdot),$$

where $F_n(\cdot)$ is the empirical cumulative distribution function estimated by using the normalized residuals, $\Phi(\cdot)$ is the distribution function of a standard normal distribution and $\phi(\cdot)$ its density. The impact of using alternative approaches for estimating λ is studied in a model-based simulation study in Section 4.7.3.

4.5 MSE Estimation Under Transformations

Estimating the MSE of small area estimates is a challenging problem. In the case of the EBP Molina and Rao (2010) propose a parametric bootstrap procedure following González-Manteiga et al. (2008). In this section we propose two bootstrap schemes for estimating the MSE under transformations. These bootstrap MSE estimators are extended to capture the additional uncertainty due to the estimation of the transformation parameter λ. The difference between the two bootstrap schemes is the mechanism used for generating the bootstrap population. In particular, the first bootstrap generates bootstrap realisations of the random effects and unit-level error terms parametrically. In contrast, the second one is a semi-parametric wild bootstrap which aims to protect against departures from the assumptions of the model in particular, those of the unit-level error term.
The steps of the proposed parametric bootstrap are as follows:

1. For $b = 1, ..., B$

   (a) Using the sample estimates, $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2_u, \hat{\sigma}^2_e, \hat{\lambda}$, generate $u_i^{(b)} \overset{iid}{\sim} N\big(0, \hat{\sigma}^2_u\big)$ and $e_{ij}^{(b)} \overset{iid}{\sim}$

$N\left(0, \hat{\sigma}_e^2\right)$ and simulate a bootstrap super-population $y_{ij}^{*(b)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_i^{(b)} + e_{ij}^{(b)}$.

(b) Back-transform $y_{ij}^{*(b)}$ to the original scale $y_{ij}^{(b)} = T_\lambda^{-1}\left(y_{ij}^{*(b)}\right)$ and compute the population value of the indicator of interest $I_{i,b}$.

(c) Extract the bootstrap sample in ${y_{ij}}^{(b)}$ and perform the EBP method, as described in Section 4.4.1. Note, as the back-transformed sample data are used, the transformation parameter λ is re-estimated in each bootstrap replication b.

(d) Obtain $\hat{I}_{i,b}^{EBP}$.

2. $\widehat{MSE}\left(\hat{I}_i^{EBP}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

As mentioned before, the proposed parametric bootstrap allows for the additional uncertainty due to the estimation of the transformation parameter. Although the use of an optimal transformation may reduce the deviation from normality, there may still be departures from normality especially in the tails of the distribution of the unit-level error term. To overcome this problem, we propose a semi-parametric bootstrap that relies on the normality of the random effects but generates the unit-level error terms by using the empirical distribution of suitably scaled unit-level residuals. The proposed wild bootstrap scheme is described below:

1. Fit the model 4.1 using an appropriate transformation $T(y_{ij}) = y_{ij}^*$ and obtain $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$.

2. Calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} - \hat{u}_i$.

3. Scale and center the residuals using $\hat{\sigma}_e$. The scaled and centered residuals are denoted by $\hat{\epsilon}_{ij}$.

4. For $b = 1, ..., B$

(a) Generate $u_i^{(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$.

(b) Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_i^{(b)}$.

(c) Match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | \eta \in n\}$ from the sample by using

$$\min_{k \in n} \left|\eta_{ij}^{(b)} - \hat{\eta}_k\right|$$

and define $\tilde{k}$ as the corresponding index.

(d) Generate weights w from a distribution satisfying the conditions in Feng et al. (2011) where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$, respectively.

(e) Calculate the bootstrap population as $y_{ij}^{*(b)} = \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}} + u_i^{(b)} + w_k|\hat{\epsilon}_{\tilde{k}}^{(b)}|$.

(f) Back-transform $T({y_{ij}}^{*(b)})$ to the original scale and compute the population value $I_{i,b}$.

(g) Extract the bootstrap sample in ${y_{ij}}^{(b)}$ and use the EBP method, as described in Section 4.4.

(h) Obtain $\hat{I}_{i,b}^{EBP}$.

5. $\widehat{MSE}_{wild}\left(\hat{I}_i^{EBP}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

The performance of both MSE estimators is compared in a model-based simulation study in Section 4.7.

4.6 The Guerrero Case Study: Application of Data-driven Transformations

The benefits of using the proposed EBP approach with data-driven transformations for estimating deprivation and inequality indicators are illustrated in an application using the data from the ENIGH survey 2010 and the National Population and Housing Census 2010 we introduced in Section 4.3. The aim is to estimate the head count ratio (HCR) and the poverty gap (PGAP) as well as the income quintile share ratio (QSR) for the 81 municipalities in Guerrero.

The indicators HCR and PGAP are special cases of the Foster-Greer-Thorbecke (FGT) indicators (Foster et al., 1984) and they depend on a poverty line t which is equal to 0.6 times the median of the target variable. The FGT index of type α for an area i is defined by

$$F_i(\alpha, t) = \frac{1}{N_i}\sum_{j=1}^{N_i}\left(\frac{t - y_{ij}}{t}\right)^{\alpha}\mathbb{I}(y_{ij} \leq t), \quad \text{for } \alpha = 0, 1, 2,$$

where $\mathbb{I}(\cdot)$ denotes an indicator function which returns 1 if $(\cdot)$ holds and 0 otherwise. When $\alpha = 0$, $F_i(\alpha, t)$ is the HCR and represents the proportion of the population whose income is below the poverty line t. Taking $\alpha = 1$, $F_i(\alpha, t)$ defines the PGAP which is a measure of poverty intensity and quantifies the degree, to which the average income of people living under the poverty line differs from the poverty line. Next to the two deprivation indicators we investigate inequality by a modified QSR -suitable for developing countries with high unemployment rates- defined by

$$\text{QSR}_i = \frac{\sum_{j=1}^{N_i}\mathbb{I}\left(y_{ij} \geq \mathbf{y}_{0.6}\right)y_{ij}}{\sum_{j=1}^{N_i}\mathbb{I}\left(y_{ij} \leq \mathbf{y}_{0.4}\right)y_{ij}},$$

where $\mathbf{y}_{0.6}$ and $\mathbf{y}_{0.4}$, denote the 60% and 40% quantiles of the target variable respectively. The QSR is a widely used inequality indicator due to its simplicity and straightforward interpretation (Eurostat, 2004).

Before focusing on the state of Guerrero, we briefly illustrate the need for data-driven transformations in different states in Mexico. Figure 4.2 represents the estimated data-driven Box-Cox transformation parameters $\hat{\lambda}$ (by REML) for each state in Mexico. These estimates vary between 0.13 and 0.37, showing the adaptive feature of data-driven transformations for each state in Mexico. Furthermore, we observe that a fixed logarithmic transformation is not suitable for any of the states.

4.6.1 Model checking and residual diagnostics

In Section 4.3 we show that the model assumptions of the linear mixed regression model are not met. We now discuss the use of the proposed data-driven transformations for adapting

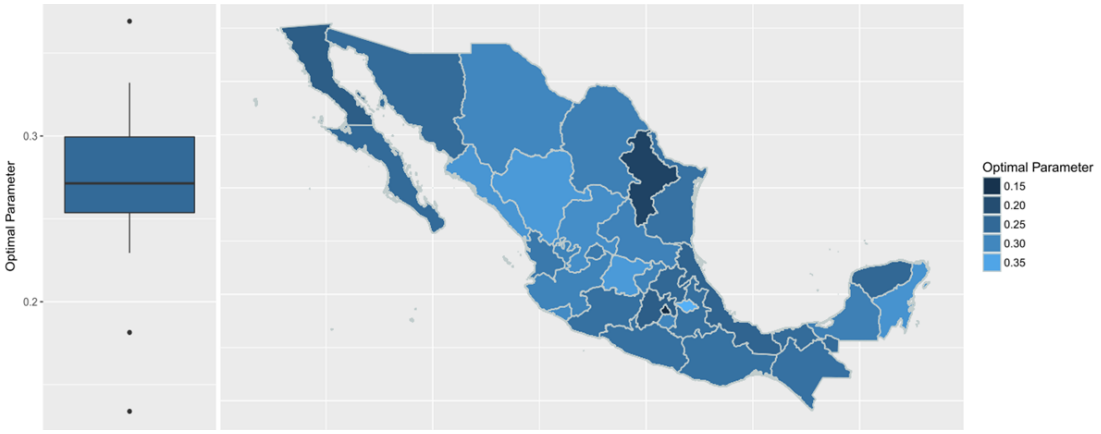


Figure 4.2: Estimated transformation parameters of the Box-Cox transformation in the different states of Mexico

the working model. In particular, we focus on the three data-driven transformations presented in Section 4.4.2, denoted by *Log-Shift*, *Box-Cox* and *Dual* power transformations and their comparison to (a) a model that use a logarithmic transformation (*Log*) and (b) a model that uses the untransformed income variable (*No*).

To start with, Figure 4.3 provides a graphical representation of the REML maximization for the transformation parameter λ for log-shift, Box-Cox and dual power transformations. In this case the optimal λs are approximately equal to 68.16, 0.26 and 0.30, respectively (cf. Table 4.3).

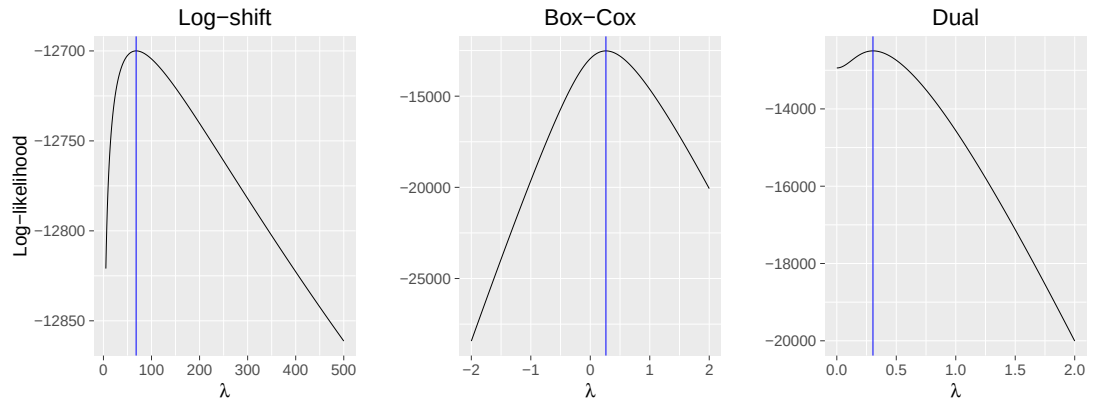


Figure 4.3: Optimal transformation parameter λs for the log-shift, Box-Cox and dual power transformations

In order to analyze whether the use of transformations improves the predictive power of the model, Table 4.3 reports the percentage of variability explained for each model and its corresponding ICC. As the ICC is larger than 0 in all cases, there appears to be unexplained between area variability and hence the use of the mixed model may be appropriate. Using the untransformed $hciw$ outcome leads to a marginal (R^2_m) and conditional (R^2_c) coefficients of determination of 0.33 and 0.35, respectively. The use of a logarithmic transformation improves the predictive power of the model in terms of the conditional R^2_c but it loses in terms of marginal R^2_m. However, it can clearly be noted that the use of data-driven transformations increases the

predictive power of the model.

Table 4.3: R^2_m, R^2_c, λ_s, and ICC for the working model under the different transformations

	R^2_m	R^2_c	λ	ICC
No	0.331	0.346	-	0.023
Log	0.263	0.416	-	0.207
Log-Shift	0.419	0.517	68.159	0.169
Box-Cox	0.439	0.517	0.263	0.140
Dual	0.443	0.517	0.304	0.132

A detailed analysis of the Gaussian assumptions of the working models corresponding to each transformation is now carried out. The results summarizing the skewness, kurtosis and S-W normality tests are presented in Table 4.4 and the Q-Q plots are presented in Figure 4.4. It should be noted, that at municipal-level, all three data-driven transformations perform similarly and yield good approximations to the normal distribution. In contrast, the household-level residuals show clear departures from normality, especially under the model with a fixed logarithmic transformation and without a transformation. The picture considerably improves for the data-driven transformations. The log-shift, Box-Cox and dual power transformations lead to very similar results in terms of skewness and kurtosis. We note that the log-shift transformation performs slightly better in terms of kurtosis, but not in terms of skewness compared to the Box-Cox and dual power transformation. These findings are supported by the Q-Q plots displayed in Figure 4.4. The data-driven transformations lead to similar Q-Q plots with more symmetrical and less extreme tails compared to the fixed log transformation. Overall, it appears that the proposed data-driven transformations improve the predictive power of the model and clearly give better approximations to the underlying model assumptions of the linear mixed regression model compared to the use of a fixed logarithmic transformation.

Table 4.4: Skewness, kurtosis and values of the S-W p-values for the municipal- and household-level error terms of the working models for EBP under the different transformations

	Household-level residuals			Municipal-level residuals		
Transformation	Skewness	Kurtosis	p-value	Skewness	Kurtosis	p-value
No	7.981	110.697	0.000	1.298	5.596	0.002
Log	-1.480	6.653	0.000	-0.576	2.336	0.025
Log-Shift	-0.346	3.895	0.000	-0.057	1.969	0.226
Box-Cox	-0.118	5.311	0.000	-0.023	2.181	0.484
Dual	-0.024	5.809	0.000	-0.005	2.242	0.627

4.6.2 Deprivation and inequality indicators for municipalities in Guerrero

Based on the analysis in Section 4.6.1, estimates for the deprivation and inequality indicators presented in Section 4.2 are calculated by using the EBP method under the three data-driven transformations and the fixed logarithmic transformation. MSE estimation is implemented with the wild bootstrap we introduced in Section 4.5 with $B = 500$ bootstrap replications.

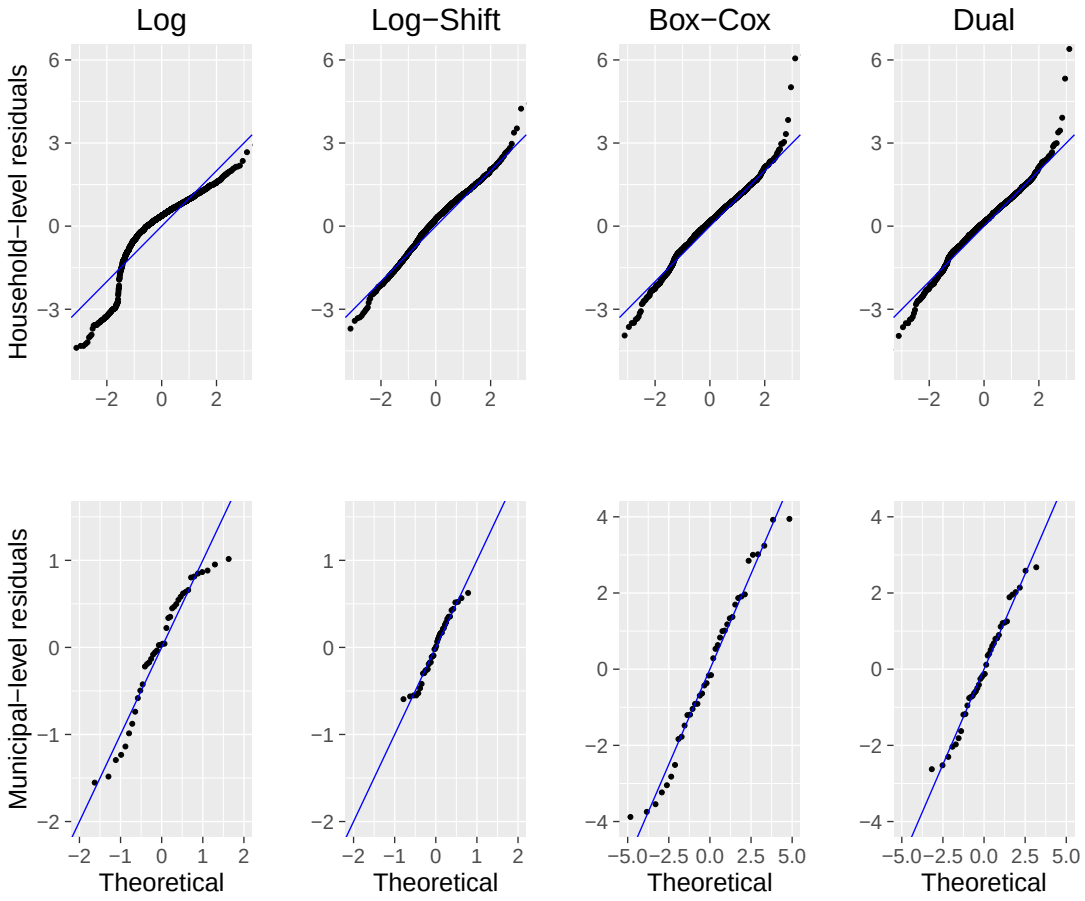


Figure 4.4: Q-Q-plots for the Pearson household-level (upper panels) and municipal-level (lower panels) residuals of the working model for EBP under the different transformations

Table 4.5 shows summaries over municipalities of point estimates and root MSEs (RMSEs) under the different transformations. We observe that the estimates based on the EBP with data-driven transformations are more efficient (in terms of RMSE) than the corresponding estimates based on a fixed logarithmic transformation. The effect is especially pronounced for indicators that rely on the tail of the distribution like the QSR. Furthermore, the use of data-driven transformations also has an effect on the point estimates of the indicators. For the HCR and PGAP, the three data-driven transformations result in very similar estimates, that are different to the EBP estimates under the model that uses the logarithmic transformation.

Having assessed the estimates from a statistical perspective, we investigate the results in the context of the spatial distribution of poverty and inequality in the state of Guerrero. Figure 4.5 presents the point estimates of HCR, PGAP and QSR at municipal-level. As the point estimates based on the three data-driven transformations are almost identical, we only show the results for the EBP with the log-shift transformation. We observe clear regional differences between the municipalities. Having a closer look to the coastal area in the south-west of Guerrero, where the largest city Acapulco is located, we observe lower levels of poverty (HCR and PGAP) and inequality (QSR) compared to other parts of the state. The coastline to the Pacific Ocean is wealthier due to several tourist destinations like Acapulco, Ixtapa and Zihuatanejo. In contrast, there is also a clear deprivation hotspot in the eastern part of the state Guerrero (e.g.

Table 4.5: Summaries of point estimates and corresponding RMSEs over municipalities in Guerrero

Point Estimation	HCR		PGAP		QSR	
Transformation	Mean	Median	Mean	Median	Mean	Median
Log	0.64	0.66	0.46	0.47	56.03	54.64
Log-Shift	0.56	0.59	0.35	0.36	18.06	15.83
Box-Cox	0.55	0.57	0.37	0.38	23.53	22.71
Dual	0.54	0.57	0.37	0.38	27.79	25.11
RMSE	HCR		PGAP		QSR	
Transformation	Mean	Median	Mean	Median	Mean	Median
Log	0.12	0.12	0.11	0.13	90.96	86.23
Log-Shift	0.10	0.11	0.09	0.09	8.73	5.92
Box-Cox	0.10	0.10	0.09	0.09	7.03	6.11
Dual	0.09	0.10	0.09	0.09	7.71	6.55

municipalities: Metlatnoc, Malinaltepec and Atlixtac) with high poverty and inequality rates. These municipalities are home to indigenous populations living in isolated mountain areas.

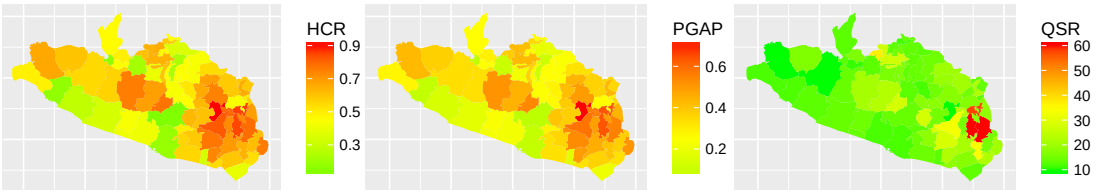


Figure 4.5: Maps of the HCR, PGAP and QSR in Guerrero for the EBP method under the log-shift transformation at municipal-level

4.7 Model-Based Simulation Study

In this section, we present results from a model-based simulation study that aims to evaluate the performance of the proposed methods. in Section 4.7.1 we analyze the behaviour of the data-driven transformation parameter under four scenarios for the distributions of the area and unit-level error terms. In Section 4.7.2 we investigate the ability of the proposed methods to provide more precise small area estimates than the EBP with a fixed logarithmic transformation or without a transformation and assess the performance of the proposed MSE estimators. Finally, in Section 4.7.3 we evaluate the methods for estimating the transformation parameter.

We generate finite populations U of size $N = 10000$, partitioned into $D = 50$ areas $U_1, U_2, \ldots, U_D$ of sizes $N_i = 200$. The samples are selected by a stratified random sampling with strata defined by the 50 small areas. This leads to a sample size of $n = \sum_{i=1}^{D} n_i = 921$ whereby the area-specific sample sizes n_i vary between 8 and 29. We chose the sample sizes mainly because of two reasons. First, we want to assess the data-driven transformations under extreme but realistic cases. Second, the sample sizes are similar in the case study.

Four scenarios, denoted by *Normal*, *Log-scale*, *Pareto* and *GB2*, are considered. Details about the data generating mechanisms of the different scenarios are provided in Table 4.6. Under scenario *Normal*, data are generated by using Normal distributions for the random effects and unit-level errors. Under the second scenario random effects and unit-level errors are generated under a log-normal distribution such that a fixed logarithmic transformation is suitable. Scenarios *Pareto* and *GB2* are settings that attempt to replicate realistic situations for income data. In particular, random effects are generated by using a Normal distribution and unit-level error terms are generated under a *Pareto* and *GB2* scenario respectively. Each setting was repeated independently $M = 500$ times. We focus on the three data-driven transformations, namely log-shift, Box-Cox and dual power transformations, and compare these to the case of a fixed logarithmic transformation and the case of using untransformed data.

Table 4.6: Model-based simulation settings for the analysis of the MSE

Scenario	Model	x_{ij}	z_{ij}	μ_i	u_i	e_{ij}
Normal	$4500 - 400x_{ij} + u_i + e_{ij}$	$N(\mu_i, 3)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$N(0, 1000^2)$
Log-scale	$\exp(10 - x_{ij} - 0.5z_{ij} + u_i + e_{ij})$	$N(\mu_i, 2)$	$N(0, 1)$	$U[2, 3]$	$N\left(0, 0.4^2\right)$	$N(0, 0.8^2)$
Pareto	$12000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 7.5)$	-	$U[-3, 3]$	$N\left(0, 500^2\right)$	$\sqrt{2}\text{Pareto}(3, 2000^2)$
GB2	$8000 - 400x_{ij} + u_i + e_{ij} - \bar{e}$	$N(\mu_i, 5)$	-	$U[-1, 1]$	$N\left(0, 500^2\right)$	$\text{GB2}(2.5, 1700, 18, 1.46)$

4.7.1 Behavior of the data-driven transformation parameters

Figure 4.6 shows box plots of the estimated transformation parameters λ for the log-shift, Box-Cox and dual power transformations (over $M = 500$ replications) under the four simulation settings. The data-driven transformation parameters are estimated by REML. Under the *Normal* setting the parameters of the Box-Cox and dual power transformations are close to one indicating that no transformation is needed. In the *Log-scale* scenario, the data was generated in such a way that normality may be achieved by applying the logarithmic transformation. In this case the log-shift transformation parameter is close to zero and the same holds for the parameters of Box-Cox and dual power transformations. For the other two scenarios (*Pareto* and *GB2*), the data-driven parameters are between 0.25 and 0.5, so neither using a logarithmic transformation nor ignoring the need for a transformation is appropriate. Overall, the results indicate that the data-driven transformations behave as expected in the four scenarios and adapt to the shapes of the data distributions.

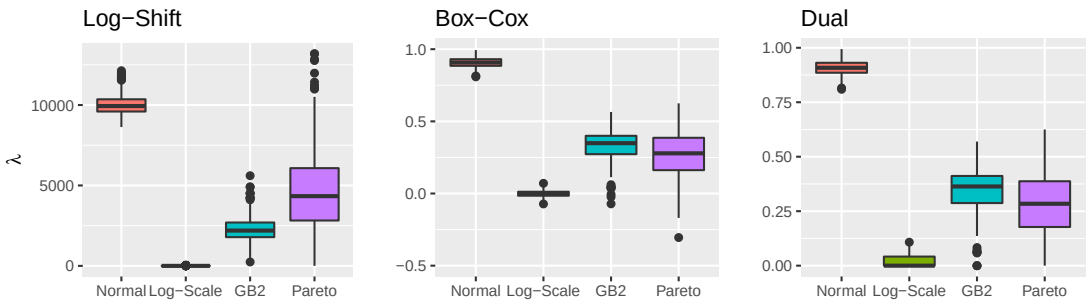


Figure 4.6: Estimated transformation parameters for the log-shift, Box-Cox and dual power transformations under the different settings.

4.7.2 Performance of the EBP under data-driven transformations

In this section we compare the performance of the proposed methods to the case of (a) fixed logarithmic transformation and (b) no transformation. We then assess the performance of the MSE estimators. Five estimators of small area deprivation and inequality indicators (HCR, PGAP and QSR) are evaluated. The EBP and the corresponding MSE estimators are implemented using $L = 100$ and $B = 500$. The following quality measures averaged over Monte-Carlo replications M are used to assess the performance of a small area estimator in area i:

$$\text{RMSE}\left(\hat{I}_i^{\text{method}}\right) = \left[\frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right)^2\right]^{1/2},$$

$$\text{Bias}\left(\hat{I}_i^{\text{method}}\right) = \frac{1}{M}\sum_{m=1}^{M}\left(\hat{I}_i^{\text{method}(m)} - I_i^{(m)}\right),$$

where $\hat{I}_i^{\text{method}}$ denotes the estimated indicator in area i based on any of the five methods under consideration and I_i denotes the corresponding true value in area i.

Table 4.7 presents the results split by the four scenarios. It shows median and mean values of RMSE and bias averaged over small areas. Under the *Normal* scenario the EBP without transformation is the gold, but the EBP with data-driven transformations (log-shift, Box-Cox and dual power) perform similarly in terms of RMSE and bias. The same picture emerges in the *Log-scale* scenario where the EBP with a logarithmic transformation is the gold standard, but again the EBP with data-driven transformations perform well both in terms of RMSE and bias. These results confirm our expectations that the EBP with data-driven transformations adapt to the shape of the data distribution. Under the *GB2* and *Pareto* scenarios we notice that the EBP with a fixed transformation or without transformation is inferior to the EBP with data-driven transformations both in terms of RMSE and Bias. The differences are especially pronounced for QSR which is very sensitive to the tails of the distribution. Furthermore, the estimates based on data-driven transformations are almost unbiased or have a small bias. A closer look at the data-driven transformations indicates that EBP with a log-shift transformation performs better than the EBP with Box-Cox and dual power transformations in these particular settings. Overall, it appears that the proposed EBP method with data-driven transformations adapts to the underlying distribution of the data, and hence improves the precision of small area estimates.

We now turn our attention to the performance of the MSE estimators. We denote by *parametric* and *wild* the proposed parametric bootstrap and proposed semi-parametric wild bootstrap respectively. The aim of this part is twofold. Firstly, we assess the performance of the two proposed MSE estimators we introduced in Section 4.5. Secondly, we investigate the ability of the wild bootstrap to protect against departures from the assumptions of the unit-level error term. Starting with the first aim, Table 4.8 reports the results for the two MSE estimators and presents the mean and median values of relative RMSE and relative bias -over Monte-Carlo replications and areas- of the EBP with Box-Cox transformation. For calculating the RMSE and relative bias we treat the empirical MSE (over Monte-Carlo replications) as the true MSE. The results for the EBP with a log-shift transformation and dual power transformation are available on request from the authors. We note that, on average, the proposed *parametric* and *wild*

bootstrap approaches for the EBP with a Box-Cox transformation have small positive relative bias (HCR and PGAP indicators) in the *Normal* and *Log-scale* settings. However, the *parametric* bootstrap shows some underestimation in the case of QSR. In this latter case *wild* bootstrap appears to be associated with smaller relative bias. For the *Normal* and *Log-scale* scenarios parametric bootstrap also appears to be more stable. Nevertheless, *wild* bootstrap improves MSE estimation as -in most cases- it has smaller relative bias and relative RMSE. These results indicate that departures from the model assumptions -even after using data transformations- affects MSE estimation with parametric methods. The problem is more pronounced when estimating parameters that depend on the tails of the distribution as is the case with the QSR. In those cases, the use of semi-parametric MSE estimation methods offers some protection against misspecification.

4.7.3 Impact of alternative estimation methods for λ

In this last section we explore the use of non-parametric alternatives to the REML approach for estimating data-driven transformation parameters (see Section 4.4.3). Here, we study five estimation methods. These are the REML approach, the minimization of the skewness (*Skew*) and the pooled skewness (*poolSkew*), and the distance-based criteria Kolomogorov-Smirnov (*KS*) and Cramér-von Mises (*CvM*) we introduced in Section 4.4.3.

The five methods estimate transformation parameters close to the theoretically correct ones, in the scenarios those are known. For instance, in the Log-scale scenario, the estimated transformation parameters under the different estimation methods are shown in Figure 4.7 and Table 4.9. We observe that although the five methods provide similar estimates of λ, the REML method has smaller variability. In our model-based simulations we further studied the impact of the estimation method of the transformation parameter on point and MSE estimation and we conclude that this only marginally influences the quality of small area estimates. These results are available from the authors upon request.

Overall, these results suggest that for the scenarios we considered in this paper the method used to estimate the transformation parameter does not have a noticeable impact on small area estimation and REML appears to be the most stable method.

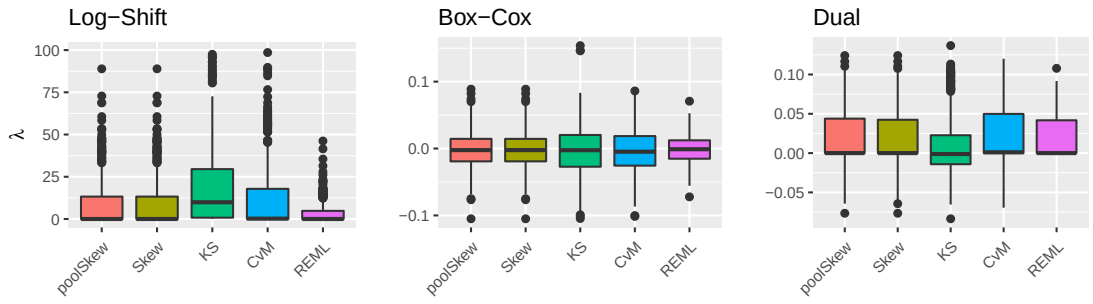


Figure 4.7: Box-plots of estimated transformation parameters for the log-scale scenario using different estimation methods

Table 4.7: Summaries of estimated RMSEs and Bias over the model-based settings

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
RMSE	No	0.0338	0.0357	0.0136	0.0154	0.3259	1.2765
	Log-Shift	0.0344	0.0363	0.0155	0.0175	0.3898	0.6710
	Box-Cox	0.0343	0.0358	0.0134	0.0156	0.3348	1.1178
	Dual	0.0343	0.0358	0.0134	0.0156	0.3346	0.5797
BIAS	No	0.0000	0.0007	0.0002	0.0009	0.0049	0.0899
	Log-Shift	0.0029	0.0039	-0.0067	-0.0076	-0.1000	-0.2190
	Box-Cox	0.0016	0.0027	-0.0021	-0.0025	-0.0396	-0.0807
	Dual	0.0016	0.0027	-0.0021	-0.0024	-0.0458	-0.1193
Log-Scale							
RMSE	Log	0.0583	0.0605	0.0358	0.0367	4.9100	4.8969
	Log-Shift	0.0583	0.0605	0.0358	0.0367	4.9024	4.8985
	Box-Cox	0.0581	0.0604	0.0358	0.0367	4.9731	4.9717
	Dual	0.0584	0.0605	0.0359	0.0367	4.9025	4.9093
BIAS	Log	-0.0011	-0.0009	-0.0007	-0.0003	0.0394	0.1143
	Log-Shift	-0.0020	-0.0017	-0.0011	-0.0007	-0.0873	-0.0072
	Box-Cox	-0.0009	-0.0006	-0.0008	-0.0004	0.1499	0.2106
	Dual	-0.0024	-0.0021	-0.0009	-0.0005	-0.1610	-0.0992
GB2							
RMSE	No	0.0650	0.0656	0.0552	0.0552	17.7364	32.0686
	Log	0.0912	0.0908	0.0272	0.0270	1.8979	1.9002
	Log-Shift	0.0418	0.0415	0.0127	0.0132	0.4286	0.4411
	Box-Cox	0.0471	0.0469	0.0136	0.0139	0.4708	0.4753
	Dual	0.0472	0.0470	0.0137	0.0140	0.4715	0.4760
BIAS	No	0.0471	0.0477	0.0481	0.0479	1.8355	2.0825
	Log	0.0746	0.0747	0.0169	0.0169	1.4718	1.4692
	Log-Shift	0.0176	0.0179	-0.0008	-0.0013	0.0546	0.0523
	Box-Cox	0.0274	0.0274	0.0035	0.0031	0.1780	0.1721
	Dual	0.0275	0.0274	0.0037	0.0034	0.1800	0.1747
Pareto							
RMSE	No	0.0448	0.0444	0.0622	0.0613	1.6814	3.6057
	Log	0.0304	0.0306	0.0082	0.0084	0.3887	0.3994
	Log-Shift	0.0185	0.0196	0.0060	0.0063	0.1661	0.1779
	Box-Cox	0.0192	0.0202	0.0059	0.0062	0.1786	0.1901
	Dual	0.0192	0.0203	0.0059	0.0062	0.1782	0.1902
BIAS	No	0.0277	0.0287	0.0166	0.0160	0.3173	0.3132
	Log	0.0086	0.0081	-0.0030	-0.0037	0.2068	0.2034
	Log-Shift	0.0003	-0.0001	-0.0034	-0.0041	0.0305	0.0300
	Box-Cox	0.0030	0.0026	-0.0031	-0.0037	0.0525	0.0530
	Dual	0.0030	0.0027	-0.0031	-0.0037	0.0522	0.0530

Table 4.8: Performance of MSE estimators in model-based simulations: EBP with Box-Cox transformation

	Indicator	HCR		PGAP		QSR	
	Estimator	Median	Mean	Median	Mean	Median	Mean
Normal							
rel. RMSE[%]	Parametric	8.30	9.22	9.15	9.47	15.25	21.23
	Wild	14.57	14.77	14.21	14.61	17.46	20.93
rel. Bias[%]	Parametric	6.64	7.27	-1.17	-0.12	-7.72	-12.61
	Wild	8.05	8.04	2.17	3.23	-1.01	-1.46
Log-Scale							
rel. RMSE[%]	Parametric	11.14	12.00	19.19	19.57	19.10	19.75
	Wild	16.82	17.00	22.70	22.95	25.34	25.62
rel. Bias[%]	Parametric	6.10	6.29	5.70	6.36	7.91	7.92
	Wild	7.69	7.82	7.34	7.39	6.58	6.78
GB2							
rel. RMSE[%]	Parametric	21.71	21.86	20.89	20.57	43.75	43.58
	Wild	19.01	19.39	14.76	15.12	26.21	27.23
rel. Bias[%]	Parametric	-20.04	-19.74	-16.88	-15.92	-42.90	-42.74
	Wild	-14.59	-14.64	-5.45	-5.75	-21.72	-22.53
Pareto							
rel. RMSE[%]	Parametric	11.31	12.60	35.60	34.78	50.04	51.63
	Wild	26.18	28.44	23.58	26.04	28.60	33.40
rel. Bias[%]	Parametric	2.43	3.38	-33.82	-31.16	-49.51	-51.06
	Wild	19.21	21.37	-8.28	-3.28	-23.02	-26.79

4.8 Conclusions and Future Research Directions

In this paper we investigate data-driven transformations for small area estimation. In particular, we propose an EBP approach with data-driven transformations estimated with likelihood-based methods. The use of scaled transformations (conditional on the Jacobian) allows for the use of standard software for fitting the mixed linear regression model. Three types of transformations are discussed log-shift, Box-Cox and dual power transformations. We further explore the use of parametric and semi-parametric wild bootstrap for MSE estimation that also captures the uncertainty from estimating the data driven transformation parameter. Semi-parametric bootstrap is used for protecting against departures from the model assumptions. Model-based simulations demonstrate the ability of the proposed EBP method to adapt to the shape of the data distribution and hence provide more efficient estimates than a fixed logarithmic transformation or the case where no transformation is used. Although the paper focuses on the EBP the proposed methods are applicable to other small area estimators for example, the ELL approach (Elbers et al., 2003).

Table 4.9: Mean and median of estimated transformation parameters under the log-scale scenario using different estimation methods

	Log-Shift		Box-Cox		Dual	
	Mean	Median	Mean	Median	Mean	Median
poolSkew	9.381	0.000	-0.002	-0.002	0.016	0.000
Skew	9.381	0.000	-0.002	-0.002	0.015	0.000
KS	23.906	10.816	-0.003	-0.003	0.009	-0.001
CvM	11.954	0.211	-0.004	-0.005	0.025	0.001
REML	3.349	0.000	-0.002	-0.001	0.021	0.000

Acknowledgements

Rojas-Perilla, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council. The authors are grateful to CONEVAL for providing the data used in empirical work. The views set out in this paper are those of the authors and do not reflect the official opinion of CONEVAL. The numerical results are not official estimates and are only produced for illustrating the methods.

Appendices

.1 Derivation of Scaled Transformations

In this appendix we derive the Jacobian and the corresponding scaling factors presented in Table 4.2 for the log-shift, Box-Cox, and dual power transformations.

.1.1 Log-shift transformation

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the log-shift transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} \left| \frac{dy_{ij}^*(\lambda)}{dy_{ij}} \right| \\ &= \prod_{i=1}^{D} \prod_{j=1}^{n_i} (y_{ij} + \lambda)^{-1}. \end{aligned}$$

The log-likelihood function in (4.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda | \boldsymbol{\theta}) &= -\frac{n-p}{2} \log(2\pi) + \frac{1}{2} \log \left| \sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{X}_i \right| - \frac{1}{2} \sum_{i=1}^{D} \log |\boldsymbol{V}_i| \\ &- \frac{1}{2} \log \left| \sum_{i=1}^{D} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right| \\ &- \frac{1}{2} \sum_{i=1}^{D} [\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i} \hat{\boldsymbol{\beta}}]^T \boldsymbol{V}_i^{-1} [\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i} \hat{\boldsymbol{\beta}}] - n \log \underbrace{\left(\prod_{i=1}^{D} \prod_{j=1}^{n_i} (y_{ij} + \lambda) \right)^{\frac{1}{n}}}_{=\bar{y}_\lambda}. \end{aligned}$$

In order to obtain the scaled log-shift transformation, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda, \boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda, \boldsymbol{y})^{\frac{1}{n}} = J(\lambda, \boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D} \prod_{j=1}^{n_i} (y_{ij} + \lambda)^{-1} \right]^{-\frac{1}{n}} \\ &= \bar{y}_\lambda. \end{aligned}$$

Therefore, the scaled log-shift transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda, \boldsymbol{y})^{1/n}} = \bar{y}_\lambda \log(y_{ij} + \lambda)$$

for $y_{ij} > -\lambda$.

.1.2 Box-Cox transformation

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the Box-Cox transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}. \end{aligned}$$

The log-likelihood function in (4.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] + n(\lambda-1)\log\underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)\right)^{\frac{1}{n}}}_{=\bar{y}}. \end{aligned}$$

In order to obtain the scaled transformation of the Box-Cox family, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda, \boldsymbol{y})^{\frac{1}{n}} = J(\lambda, \boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}(y_{ij}+s)^{\lambda-1}\right]^{-\frac{1}{n}} \\ &= \bar{y}^{-(\lambda-1)}. \end{aligned}$$

Therefore, the scaled Box-Cox transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda, \boldsymbol{y})^{1/n}} = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\bar{y}^{\lambda-1}\lambda}, & \lambda \neq 0, \\ \bar{y}\log(y_{ij}+s), & \lambda = 0, \end{cases}$$

for $y_{ij} > -s$.

.1.3 Dual power transformation

Let $J(\lambda, \boldsymbol{y})$ be the Jacobian of the dual power transformation from $\boldsymbol{y}_i$ to $\boldsymbol{y}_i^*(\lambda)$, defined as:

$$\begin{aligned} J(\lambda, \boldsymbol{y}) &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\left|\frac{dy_{ij}^*(\lambda)}{dy_{ij}}\right| \\ &= \prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}. \end{aligned}$$

The log-likelihood function in (4.3) can be rewritten as follows:

$$\begin{aligned} l_{\text{REML}}(\boldsymbol{y}, \lambda|\boldsymbol{\theta}) &= -\frac{n-p}{2}\log(2\pi) + \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{X}_i\right| - \frac{1}{2}\sum_{i=1}^{D}\log|\boldsymbol{V}_i| \\ &- \frac{1}{2}\log\left|\sum_{i=1}^{D}\boldsymbol{X}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{X}_i\right| \\ &- \frac{1}{2}\sum_{i=1}^{D}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}]^T\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i^*(\lambda) - \boldsymbol{X_i}\hat{\boldsymbol{\beta}}] \\ &+ n\log\underbrace{\left(\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}\right)^{\frac{1}{n}}}_{=\bar{y}_\lambda}. \end{aligned}$$

In order to obtain the scaled dual transformation, $z_{ij}^*(\lambda)$, the denominator of the term $\frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}}$ is given by:

$$\begin{aligned} 1/J(\lambda,\boldsymbol{y})^{1/n} = J(\lambda,\boldsymbol{y})^{-\frac{1}{n}} &= \left[\prod_{i=1}^{D}\prod_{j=1}^{n_i}\frac{(y_{ij}+s)^{\lambda-1} + (y_{ij}+s)^{-\lambda-1}}{2}\right]^{-\frac{1}{n}} \\ &= \bar{y}_\lambda^{-1}. \end{aligned}$$

Therefore, the scaled dual transformation is defined as follows:

$$z_{ij}^*(\lambda) = \frac{y_{ij}^*(\lambda)}{J(\lambda,\boldsymbol{y})^{1/n}} = \begin{cases} \bar{y}_\lambda^{-1}\frac{(y_{ij}+s)^\lambda - (y_{ij}+s)^{-\lambda}}{2\lambda} & \text{if } \lambda > 0; \\ \bar{y}_\lambda^{-1}\log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases}$$

for $y_{ij} > -s$.

Chapter 5

The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators

5.1 Introduction

In recent years an increased number of policy decisions has been based on statistical information derived from indicators estimated at disaggregated geographical levels using small area estimation methods. Clearly, the more detailed the information provided by official statistics estimates, the better the basis for targeted policies and evaluating intervention programs. The United Nations suggest further disaggregation of statistical indicators for monitoring the Sustainable Development Goals (SDGs). National Statistical Institutes (NSIs) and other organizations across the world have also recognized the potential of producing small area statistics and their use for informing policy decisions. Examples of NSIs with well-developed programs in the production of small area statistics include the US Bureau of Census, the UK Office for National Statistics (ONS) and the Statistical Office of Italy (ISTAT). Although the term domain is more general as it may include non-geographic dimensions, the term small area estimation (SAE) is the established one. We shall follow the custom in this paper and use the terms area/geography and domain/aggregation interchangeably.

Without loss of generality in this paper we will assume that the primary data sources used to estimate statistical indicators are national socio-economic household sample surveys. Sample surveys are designed to provide estimates with acceptable precision at national and possibly sub-national levels but usually have insufficient sizes to allow for precise estimation at lower geographical levels. Therefore, direct estimation that relies only on the use of survey data can be unreliable or even not possible for domains that are not represented in the sample. In the absence of financial resources for boosting the sample size of surveys, using model-based methodologies can help to obtain reliable estimates for the target domains.

Model-based SAE methods (Pfeffermann, 2013; Rao and Molina, 2015) work by using statistical models to link survey data, that are only available for a part of the target population, with administrative or census data that are available for the entire population. Despite the wide

range of SAE methods that have been proposed by academic researchers, these are so far applied only by a fairly small number of NSIs or other practitioners. This gap between theoretical advances and applications may have several reasons one of which is the lack of suitable, user friendly statistical software. More precisely, software needs not only to be available but it also needs to simplify the application of the methods for the user. The R (R Core Team, 2017) package **emdi** (Kreutzmann et al., 2018) aims to improve the user experience by simplifying the estimation of small area indicators and corresponding precision estimates. Furthermore, the user benefits from support that extends beyond estimation in particular, evaluating, processing, and presenting the results.

Traditionally model-based SAE methods have been employed for estimating simple, linear indicators for example, means and totals using for example, mixed (random) effects models and empirical best linear unbiased predictors (EBLUPs). Several software packages exist. In R, the package **JoSAE** (Breidenbach, 2015) includes functions for EBLUPs using unit-level models. Functions in the package **hbsae** (Boonstra, 2012) enable the use of unit- and area-level models and can be estimated either by using restricted maximum likelihood (REML) or hierarchical Bayes methods. The package **BayesSAE** (Shi and with contributions from Peng Zhang, 2013) also allows for Bayesian methods. The **rsae** package by Schoch (2012) and package **saeRobust** by Warnholz (2016a) provide functions for outlier robust small area estimation using unit- or area-level models. Gaussian area-level multinomial mixed-effects models for SAE can be done with the **mme** package (Lopez-Vizcaino et al., 2014). In addition, resources in R are available for Bayesian SAE from the BIAS (Bayesian methods for combining multiple Individual and Aggregate data Sources) project (Gómez-Rubio et al., 2010) and from the package **SAE2** (Gómez-Rubio et al., 2008) that provides likelihood-based methods. In Stata, functions `xtmixed` and `gllamm` support the estimation of linear mixed models, which is a popular basis for model-based SAE. EBLUPs can be derived using these functions (West et al., 2007). Similarly, `PROC MIXED` and `PROC IML` can be used for fitting unit- and area-level models in SAS as shown in Mukhopadhyay and McDowell (2011). Furthermore, several SAS macros for SAE are provided by the EURAREA (Enhancing Small Area Estimation Techniques to meet European Needs) project (EURAREA Consortium, 2004).

More recently widespread application of SAE methods involves the estimation of poverty and inequality indicators and distribution functions (The World Bank, 2007). In this case the use of methodologies for estimating means and totals is no longer appropriate since such indicators are complex, non-linear functions of the data. As an example, we refer to the Foster-Greer-Thorbecke indicators (Foster et al., 1984), the Gini coefficient (Gini, 1912) and the quantiles of the income distribution. Popular SAE approaches for estimating complex indicators include the Empirical Best Predictor (EBP) (Molina and Rao, 2010), the World Bank method (Elbers et al., 2003) and the M-Quantile method (Chambers and Chandra, 2006; Tzavidis et al., 2010). Although in this paper we focus exclusively on software for implementing the EBP method (Molina and Rao, 2010), a future version of the package will include the M-Quantile and World Bank methods. The World Bank provides a free software for using the World Bank method called PovMap (The World Bank, 2013). However, this focuses exclusively on poverty mapping. Creating a more general open-source software can help to accelerate the uptake of

modern model-based methods. Currently, the best known package that also includes the EBP method is the R package **sae** (Molina and Rao, 2010). Although the **sae** package implements a range of small area methods, it lacks the necessary functionality for supporting the user beyond estimation for example, for performing model diagnostic analyses, visualising, and exporting the results for further processing. In contrast, **emdi** supports the user by providing more options and greater flexibility. In particular, package **emdi** offers the following attractive features that distinguish it from the **sae** package and other R packages for SAE:

- The estimation functions return by default estimates for a set of predefined indicators, including the mean, the quantiles of the distribution of the response variable and poverty and inequality indicators. Additionally, self-defined indicators or indicators available from other packages can be included.
- The user can select the type of data transformation to be used in **emdi**. Data-driven transformation parameters are estimated automatically.
- In contrast to other packages that include only a parametric bootstrap for mean squared error (MSE) estimation, package **emdi** includes two bootstrap methods, a parametric bootstrap and a semi-parametric wild bootstrap (see Appendix .1) for MSE estimation. Both incorporate the uncertainty due to the estimation of the transformation parameter. The use of wild bootstrap (Thai et al., 2013; Flachaire, 2005) protects the user against departures from the distributional assumptions of the nested error linear regression model. This offers additional protection against possible misspecification of the model assumptions.
- Customized parallel computing is offered for reducing the computational time associated with the use of bootstrap.
- Package **emdi** provides predefined functions for diagnostic analyses of the model assumptions. A mapping tool for plotting the estimated indicators enables the creation of high quality visualization. The output summarizing the most relevant results can be exported to Excel™ and to OpenDocument Spreadsheets for presentation and reporting purposes.

The remainder of this paper is structured as follows. Section 5.2 gives information about the estimation methods that are included in the package. In Section 5.3 we present the data sets that we used for illustrating the use of the **emdi** package. Section 5.4 describes the core functionality of the package. Examples demonstrate the use of the methods for computing, assessing and presenting the estimates. Section 5.5 shows how users can extend the set of indicators to be estimated by including customized options and describes the parallelization features of the package. Finally, Section 5.6 discusses future potential extensions.

5.2 Statistical Methodology

In order to obtain regionally disaggregated indicators, package **emdi** includes direct estimation and currently model-based estimation using the EBP approach by Molina and Rao (2010).

Measurement	Indicator I_i	Expression	Range
Location	Mean_i	$\frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$	$\mathbb{R}$
	$\text{Q}_{i,q}$	$F_i^{-1}(q) = \inf\{y_i \in \mathbb{R} : F_i(y_i) \geq q\}$	$\mathbb{R}$
Poverty	HCR_i	$\frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq z)$	$[0, 1]$
	PG_i	$\frac{1}{N_i} \sum_{j=1}^{N_i} \left(\frac{z-y_{ij}}{z}\right) \mathbf{I}(y_{ij} \leq z)$	$[0, 1]$
Inequality	Gini_i	$\frac{2\sum_{j=1}^{N_i} j y_{ij}}{N_i \sum_{j=1}^{N_i} y_{ij}} - \frac{(N_i+1)}{N_i}$	$[0, 1]$
	QSR_i	$\frac{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} > Q_{i,0.8}) y_{ij}}{\sum_{j=1}^{N_i} \mathbf{I}(y_{ij} \leq Q_{i,0.2}) y_{ij}}$	$\mathbb{R}$

Table 5.1: List of predefined population indicators in **emdi**. Note that $F_i(y_i)$ denotes the empirical distribution function of the population in domain i and quantiles are generally defined for $q \in (0, 1)$. The predefined quantiles in **emdi** are $q \in (0.1, 0.25, 0.5, 0.75, 0.9)$.

The predefined indicators returned by the estimation functions in **emdi** include the mean and quantiles Q_q (10%, 25%, 50%, 75%, 90%) of the target variable as well as non-linear indicators of the target variable. A widely used family of indicators measuring income deprivation and inequality is the Foster-Greer-Thorbecke (FGT) one (Foster et al., 1984). Package **emdi** includes the FGT measures of Head Count Ratio (HCR) and Poverty Gap (PG). In order to compute the HCR and PG indicators one must use a threshold z, also known as poverty line. This line is a minimum level of income deemed adequate for living in a particular country and can be defined in terms of absolute or relative poverty. For instance, the international absolute poverty line is currently set to \$1.90 per day by the World Bank (The World Bank, 2017). Relative poverty means a low income relative to others in a particular country - for instance, below a percentage of the median income in that country. Another family of indicators of interest is the so-called Laeken indicators, endorsed by the European Council in Laeken, Belgium (Council of the European Union, 2001). Two examples of Laeken indicators that are well-known for measuring inequality are the Gini coefficient (Gini, 1912) and the Income Quintile Share Ratio (QSR) (Eurostat, 2004). These two inequality indicators are also available in **emdi**. Therefore, in total **emdi** includes ten predefined indicators I_i - summarized in Table 5.1 - that are estimated at domain level i using a) direct estimation introduced in Section 5.2.1 and b) model-based estimation via the EBP method introduced in Section 5.2.2.

In the following sections the notation denotes by U a finite population of size N, partitioned into D domains $U_1, U_2, \ldots, U_D$ of sizes $N_1, \ldots, N_D$, where $i = 1, \ldots, D$ refers to an ith domain and $j = 1, \ldots, N_i$ to the jth household/individual. From this population a random sample of size n is drawn. This leads to $n_1, \ldots, n_D$ observations in each domain. If n_i is equal to 0 the domain is not in the sample. The target variable is denoted by $\mathbf{y}$.

5.2.1 Direct estimation

Direct estimation relies on the use of sample data only. The definition of direct (point and variance) estimators in **emdi** follows Alfons and Templ (2013). The mean and the quantiles help to describe the level and the distribution of a target variable. Especially for target variables

with a skewed distribution, quantiles can be more appropriate summary statistics than the mean, since these are robust to extreme values. Direct estimators of the mean and the quantiles are defined as follows,

$$\widehat{\text{Mean}}_i = \frac{\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{j=1}^{n_i} w_{ij}},$$

$$\widehat{\text{Q}}_{i,q} = \begin{cases} \frac{1}{2}\left(y_{ik} + y_{ik+1}\right) & \text{if } \sum_{j=1}^{k} w_{ij} = q \sum_{j=1}^{n_i} w_{ij}; \\ y_{ik+1} & \text{if } \sum_{j=1}^{k} w_{ij} \leq q \sum_{j=1}^{n_i} w_{ij} \leq \sum_{j=1}^{k+1} w_{ij}, \end{cases}$$

where w denotes the sample weights and $q \in (0, 1)$ defines the corresponding quantile.
FGT measures HCR and PG are estimated by package **emdi** as follows,

$$\widehat{\text{HCR}}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} \mathbf{I}(y_{ij} \leq z),$$

$$\widehat{\text{PG}}_i = \frac{1}{\sum_{j=1}^{n_i} w_{ij}} \sum_{j=1}^{n_i} w_{ij} \left(\frac{z - y_{ij}}{z}\right) \mathbf{I}(y_{ij} \leq z),$$

where the indicator function $\mathbf{I}(\cdot)$ equals 1 if the target variable y_{ij} is below the threshold z and 0 otherwise. As already mentioned, for the computation of the HCR and PG indicators one must use a threshold z, also known as the poverty line. Package **laeken** (Alfons and Templ, 2013) uses relative poverty lines defined as 60% of median equivalized disposable income, which corresponds to the EU definition for poverty lines and thus in this case the HCR is the At-risk-of-poverty rate. In contrast, package **emdi** allows both for absolute and relative poverty lines and the user is free to set the poverty line. Therefore, the threshold can be given as an argument in **emdi** or, alternatively, the user can define an arbitrary function depending on the target variable and sampling weights. As a default, a relative threshold defined as 60% of the median of the target variable is used. The HCR describes the proportion of the population below the poverty line and the PG further takes into account how far, on average, this proportion falls below the threshold. Both indicators are between 0 and 1.
The two inequality indicators Gini and QSR are estimated in **emdi** by

$$\widehat{\text{Gini}}_i = \left[\frac{2 \sum_{j=1}^{n_i} \left(w_{ij} y_{ij} \sum_{k=1}^{j} w_{ik} \right) - \sum_{j=1}^{n_i} w_{ij}^2 y_{ij}}{\sum_{j=1}^{n_i} w_{ij} \sum_{j=1}^{n_i} w_{ij} y_{ij}} - 1 \right],$$

$$\widehat{\text{QSR}}_i = \frac{\sum_{j=1}^{n_i} \mathbf{I}(y_{ij} > Q_{i,0.8}) w_{ij} y_{ij}}{\sum_{j=1}^{n_i} \mathbf{I}(y_{ij} \leq Q_{i,0.2}) w_{ij} y_{ij}},$$

where $\mathbf{I}(\cdot)$ is an indicator function that equals 1 if the target variable is above the weighted 80% quantile or below the 20% quantile and 0 otherwise. The Gini coefficient is between 0 and 1, and the higher the value, the higher the inequality is. The extreme values of 0 and 1 indicate perfect equality and inequality, respectively. QSR is typically used when the target variable is income and in this case it is defined as the ratio of total income of the 20% richest households to the 20% poorest households. The higher the value of QSR, the higher the inequality is.
While variance estimation in package **laeken** (Alfons and Templ, 2013) is only available for

the poverty and inequality indicators, package **emdi** also provides a non-parametric bootstrap method (Alfons and Templ, 2013) for estimating the variance of estimates of the mean and the quantiles. The variance is, on the one hand, an important measure for measuring the precision of estimates. On the other hand, it is also important to compute the coefficient of variation (CV) which is one measure for showing the extent of the variability of the estimate. The CV is used, for instance, by NSIs for quantifying the uncertainty associated with the estimates and is defined as follows,

$$CV = \frac{\sqrt{\widehat{MSE}(\hat{I}_i)}}{\hat{I}_i},$$

where $\hat{I}_i$ is an estimate of an indicator I_i for domain i and $\widehat{MSE}(\hat{I}_i)$ is the corresponding mean squared error.

5.2.2 Model-based estimation

The implementation of the EBP method in package **emdi** is based on the theory described in Molina and Rao (2010) and Rao and Molina (2015). The underlying model is a unit-level mixed model also known in SAE literature as the nested error linear regression model (Battese et al., 1988). In its current implementation the EBP method is based on a two-level nested error linear regression model that includes a random area/domain-specific effect and a unit (household or individual)-level error term.

In addition to the notation above, here we assume that $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)^\top$ is the design matrix, containing p explanatory variables. The nested error linear regression model is defined by

$$T(y_{ij}) = \mathbf{x}_{ij}^\top \boldsymbol{\beta} + u_i + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, D, \quad u_i \overset{iid}{\sim} N(0, \sigma_u^2), \quad e_{ij} \overset{iid}{\sim} N(0, \sigma_e^2), \tag{5.1}$$

where T denotes a transformation of the target variable $\mathbf{y}$, $\mathbf{x}_{ij}$ is a vector of unit-level auxiliary variables of dimension $(p+1) \times 1$, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of regression coefficients and u_i and e_{ij} denote the random area and unit-level error terms.

The EBP approach works by using at least two data sources, namely a sample data set used to fit the nested error linear regression model and a population (e.g., census or administrative) data set used for predicting - under the model - synthetic values of the outcome for the entire population. Both data sources must share identically defined covariates but the target variable is only available in the sample data set.

Use of data transformations

Under model (5.1) we assume that the model error terms follow a Gaussian distribution. However, in certain applications - as is the case when analyzing economic variables - this assumption may be unrealistic. Package **emdi** includes the option of using a one-to-one transformation $T(y_{ij})$ of the target variable $\mathbf{y}$ aiming to make the Gaussian assumptions more plausible. A logarithmic-type transformation is very often used in practice (Elbers et al., 2003; Molina and Rao, 2010). However, this is not necessarily the optimal transformation, for example, when the

model error terms do not follow exactly a log-normal distribution. In addition to a logarithmic transformation, package **emdi** allows the use of a data-driven Box-Cox transformation. The Box-Cox transformation is denoted by

$$T(y_{ij}) = \begin{cases} \frac{(y_{ij}+s)^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0; \\ \log(y_{ij}+s) & \text{if } \lambda = 0, \end{cases} \tag{5.2}$$

where λ is an unknown transformation parameter and s denotes the shift parameter, which is a constant and chosen automatically such that $y_{ij}+s > 0$. A general algorithm for estimating the transformation parameter λ is the residual maximum likelihood (REML), which is described in detail in Rojas-Perilla et al. (2017). One advantage of using the Box-Cox transformation is that it includes the logarithmic and no transformation as cases for specific values of λ.

Package **emdi** currently includes the following options: no transformation, logarithmic transformation and Box-Cox transformation.
The EBP method is implemented using the following algorithm:

1. For a given transformation obtain $T(y_{ij})$. If the user selects the Box-Cox transformation, the transformation parameter λ is automatically estimated by the **emdi** package.

2. Use the sample data to fit the nested error linear regression model and estimate $\boldsymbol{\theta}$ denoted by $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2)$. The variance components are estimated by REML using the function `lme` from the package **nlme** (Pinheiro et al., 2017). Also compute $\hat{\gamma}_i = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_e^2}{n_i}}$.

3. For $l = 1, \ldots, L$:

   (a) For in-sample domains (domains that are part of the sample data set), generate a synthetic population of the target variable by $T(y_{ij}^{*(l)}) = \mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}} + \hat{u}_i + v_i^* + e_{ij}^*$, with $v_i^* \overset{iid}{\sim} N(0, \hat{\sigma}_u^2(1-\hat{\gamma}_i))$, $e_{ij}^* \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$ and $\hat{u}_i$, the conditional expectation of u_i given y_i.
   For out-of-sample domains (domains with no data in the sample) the conditional expectation of u_i cannot be computed, hence for these domains generate a synthetic population by using $T(y_{ij}^{*(l)}) = \mathbf{x}_{ij}^{\top}\hat{\boldsymbol{\beta}} + v_i^* + e_{ij}^*$, with $v_i^* \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$, $e_{ij}^* \overset{iid}{\sim} N(0, \hat{\sigma}_e^2)$.
   For additional details we refer to Molina and Rao (2010).

   (b) Back-transform to the original scale $\mathbf{y}_i^{(l)} = T^{-1}(\mathbf{y}_i^{*(l)})$ and calculate the target indicator $I_i^{(l)}(\mathbf{y}_i^{(l)})$ in each domain. Note that $I_i^{(l)}$ is used here as a generic notation for any indicator of interest.

4. Compute the final estimates by taking the mean over the L Monte Carlo simulations in each domain, $\hat{I}_i^{EBP} = 1/L \sum_{l=1}^{L} I_i^{(l)}(\mathbf{y}_i^{(l)})$.

The **emdi** package fits the nested error linear regression model by using the **nlme** package and currently does not permit the use of an alternative package for example **lme4** (Bates et al., 2015). The reason for this choice is that in future developments of **emdi** we plan to allow for

more complex covariance structures for the unit-level error term and the random effect for example, allowing for spatially correlated errors (Pratesi and Salvati, 2009; Schmid et al., 2016). To the best of our knowledge, the **nlme** package offers sufficient flexibility for incorporating such models.

Measuring the uncertainty of the EBP estimates is done by using bootstrap methods. Here the uncertainty is quantified by the MSE. Package **emdi** includes two bootstrap schemes. One is parametric bootstrap under model (1) following Molina and Rao (2010), which additionally includes the uncertainty due to the estimation of the transformation parameter (Rojas-Perilla et al., 2017).

Using an appropriate transformation often mitigates the departures from normality. However, even after transformations, departures from normality may still exist in particular for the unit-level error term. For this reason, **emdi** also includes a variation of semi-parametric wild bootstrap (Flachaire, 2005; Thai et al., 2013; Rojas-Perilla et al., 2017) to protect against departures from the model assumptions. The semi-parametric wild bootstrap is presented in detail in Appendix .1. A simulation study comparing the performance of both MSE estimators is presented in Rojas-Perilla et al. (2017). Since the bootstrap schemes presented here are computationally intensive, **emdi** includes an option for parallelization that is described in detail in Section 5.5.2.

5.3 Data Sets

The main idea of SAE is to combine multiple data sources. Typically, one data set is obtained from a survey on unit-level and the other one from census or administrative/register data. The target variable is available in the survey but not in the census data. The administrative data contains explanatory variables that are potentially correlated with the target variable and hence they can be used to assist the estimation. Depending on the model type and the indicator of interest, census information is needed at the unit-level, i.e., information is available for every unit in each domain, or it is required at the area-level which means that aggregated data for each domain is given. If the user is interest in estimating non-linear functions of the target variable (like indicators discussed in Section 5.2), then access to unit-level census data is needed. As the EBP approach in package **emdi** is suitable for estimating non-linear indicators, one population data set (`eusilcA_pop`) and one survey data set (`eusilcA_smp`) are provided at the household level such that the method can be illustrated. The two data sets are based on the use of `eusilcP` from the package **simFrame** (Alfons et al., 2010). This data set is a simulated close-to-reality version of the European Union Statistics on Income and Living Conditions (EU-SILC) in Austria from 2006. Austria is a federal republic in Central Europe made up of nine states and 94 districts (79 districts headed by commissions and 15 statutory cities) with a total population of about 8.8 million in 2018. The original EU-SILC data is obtained from an annual household survey that is nowadays conducted in all EU member states and six other European countries and enables the analysis of income, socio-demographic factors and living conditions.

For practical reasons, we need to modify the `eusilcP` data set used in package **sim-**

Frame. Due to confidentiality constraints the lowest geographical level in this data set includes the nine states and identifiers for lower regional levels, like the 94 districts, are not included. However, in the context of SAE the interest is on lower geographical levels like districts or municipalities. Therefore, we assigned households to Austrian districts for illustrating the methodology better. The modified synthetic population is called `eusilcA_pop`. The assignment is based on two criteria available from external sources: a) the population sizes at state and district level and b) the income level in each district. From the last register-based census in 2011 the population sizes in each district and in each state are known and publicly available (Statistik Austria, 2013). We defined the district population sizes in relation to the state population sizes in the `eusilcA_pop` data set such that their population ratios mimic the *true* ratios in Austria. Furthermore, the Austrian Chamber of Commerce published a ranking of the districts within the states based on the net per capita income (Wirtschaftskammer Österreich, 2017). Based on this ranking we assigned households to districts such that the ordering of the districts within states is maintained. One drawback of the population data set is the small number of households in some districts. For instance, the number of households is only 5 in Rust (Stadt). This is, however, partly due to the fact that it is also in reality a really small district with only 1896 inhabitants (Statistik Austria, 2013). Although the `eusilcA_pop` data set in **emdi** mimics some real characteristics in Austria, it is a synthetic population data set for demonstrating the functionality of the package and conclusions about the levels of inequality and poverty in the Austrian districts observed from this data are not official estimates. The full documented code for the assignment of the households to the districts is available from the authors' `GitHub` folder (`https://github.com/SoerenPannier/districtAssignment.git`).

The target variable in the example is the equivalized household income (`eqIncome`), which is defined as the total household disposable income divided by the equivalized household size determined by the modified OECD scale (Hagenaars et al., 1994). Thus, the indicators in our illustration describe the distribution of income, poverty and inequality similarly to the analysis in Alfons and Templ (2013). The remaining variables in `eusilcA_pop` are variables that identify the regional levels (`state` and `district`) and auxiliary variables that can be used for modeling income. These explanatory variables are, among others, gender (`gender`), the equivalized household size (`eqsize`) as well as financial resources like the employees cash (`cash`) or unemployment benefits (`unempl_ben`). Table 5.2 gives an overview of possible model covariates.

The sample data set `eusilcA_smp` is a household sample from the `eusilcA_pop` population that includes 1945 observations. The sample is drawn by stratified random sampling where the districts define the strata. For the 75% largest districts (in terms of number of households) 10% of the households were selected and the maximum number of sampled households is equal to 200 in any given district. Consequently, the 25% smallest districts do not have any observation in the sample. Summaries of state and district-specific sample sizes are given below.

```
R> data("eusilcA_smp")
R> table(eusilcA_smp$state)

Burgenland   Carinthia   Lower Austria   Salzburg   Styria
        31         162             387        163      337
    Tyrol      Upper Austria      Vienna      Vorarlberg
      173                392         200             100

R> summary(as.numeric(table(eusilcA_smp$district)))

 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
14.00   17.00   22.50   27.79   29.00  200.00
```

District-specific sample sizes (in contrast to state-specific) are quite small with 25% of districts having no sample data at all. Hence, the use of SAE methods may be useful in this case. In Section 5.4 we discuss the estimation of regional indicators based on these data sets in detail.

In addition to SAE methods, package **emdi** provides a function called `map_plot` that produces maps of the estimated indicators. In order to demonstrate the use of the function `map_plot` package **emdi** contains a shape file for the 94 Austrian districts which is downloaded from the SynerGIS website (Bundesamt für Eich- und Vermessungswesen, 2017). This shape file is saved in `.rda` format and the object `shape_austria_dis` is a `SpatialPolygons DataFrame`. For more information about this class we refer to Bivand et al. (2013).

5.4 Basic Design and Core Functionality

Section 5.2 presented the statistical methodology that uses either direct estimation or the model-based EBP approach. In package **emdi** direct and model-based estimation are implemented with functions `direct` and `ebp`, respectively. A key benefit offered by **emdi** is the flexibility for producing, assessing, presenting and exploring the estimates. This is achieved by using the following commands:

1. Estimate domain indicators including MSE estimation: `direct` and `ebp`
2. Get summary statistics and model diagnostics: `summary` and `plot`
3. Extract and compare the indicators of interest: `estimators` and `compare`
4. Visualize the estimated indicators: `map_plot`
5. Export the results to Excel™: `write.excel`

The package **emdi** uses the S3 object system (Chambers and Hastie, 1992). All objects created in the package **emdi** by an estimation function (`direct` and `ebp`) share the class `emdi`. Objects of class `emdi` comprise ten components, which are presented in Table 5.3. Some of these components are specific only to one of the estimation methods, such that they

Variable	Meaning	Scale level
Target variable		
`eqIncome`	The equivalized household income.	Numeric
Domain identifiers		
`state`	Austrian states.	Factor
`district`	Austrian districts.	Factor
Explanatory variables		
`eqsize`	The equivalized household size according to the modified OECD scale.	Numeric
`gender`	The person's gender (levels: `female` and `male`).	Factor
`cash`	Employee cash or near cash income.	Numeric
`self_empl`	Cash benefits or losses from self-employment (net).	Numeric
`unempl_ben`	Unemployment benefits (net).	Numeric
`age_ben`	Old-age benefits (net).	Numeric
`surv_ben`	Survivor's benefits (net).	Numeric
`sick_ben`	Sickness benefits (net).	Numeric
`dis_ben`	Disability benefits (net).	Numeric
`rent`	Income from rental of a property or land (net).	Numeric
`fam_allow`	Family/children related allowances (net).	Numeric
`house_allow`	Housing allowances (net).	Numeric
`cap_inv`	Interest, dividends, profit from capital investments in unincorporated business (net).	Numeric
`tax_adj`	Repayments/receipts for tax adjustment (net).	Numeric
Design variable		
weight	Sampling weight.	Numeric

Table 5.2: Variables of the two data sets in package **emdi**. Note that the population data set does not contain a variable for the sampling weights.

are `NULL` for the other one. These components are indicated in the second column of Table 5.3. Depending on the estimation method, the `emdi` object is also of class `direct` or `model`.

Thus, the commands can be tailored to the estimation method, e.g., model diagnostics (provided by the command `plot`) are only suitable when a model-based approach is used. In what follows the estimation functions are presented and **emdi** functionalities are illustrated.

5.4.1 Estimation of domain indicators

As far as possible, the two estimation functions (`direct` and `ebp`) have the same structure and variable names, which helps to simplify their use. For function `direct`, the user has to specify three arguments (see Table 5.4), that include the target variable, the sample data set, and the variable name that defines the domain identifier in the sample data. For the remaining arguments suitable defaults are defined. The EBP approach is implemented in **emdi**, using function `ebp`. As shown in Table 5.5, the user has to specify five arguments that include the structure of the fixed effects of the nested error linear regression model, the two data sets (population and sample), and the variable names that define the domain identifiers in each data set. For the remaining arguments suitable defaults are defined. Following Molina and Rao (2010), the number of Monte Carlo iterations L and the number of bootstrap populations B are set to 50 by

Position	Name	Short description	Available for `direct`	Available for `model`
1	`ind`	Point estimates for indicators per domain	✓	✓
2	`MSE`	Variance/MSE estimates per domain	✓	✓
3	`transform_param`	Transformation and shift parameters		✓
4	`model`	Fitted linear mixed-effects model as `lme` object		✓
5	`framework`	List with 8 components describing the data	✓	✓
6	`transformation`	Type of transformation		✓
7	`method`	Estimation method for transformation parameter		✓
8	`fixed`	Formula of fixed effects used in the nested error linear regression model		✓
9	`call`	Image of the function call that produced the object	✓	✓
10	`successful_bootstraps`	A matrix with domains as rows, indicators as columns and the number of corresponding successful bootstraps	✓	

Table 5.3: Components of `emdi` objects. All explanations can be found in the documentation of the `emdi` object in the package.

default. In practice, we recommend using larger values for example, $L \geq 200$ and $B \geq 200$. The choice of a transformation is simplified since the user only has to choose the type of transformation. The shift parameter s and the optimal transformation parameter λ in the case of using the Box-Cox transformation are automatically estimated. This distinguishes **emdi** from package **sae** (Molina and Marhuenda, 2015) where the user has to select the transformation parameters manually. Since the Box-Cox transformation includes the no transformation and logarithmic transformation as special cases, this family of transformations is chosen as the default option.

Example using Austrian districts:

For illustrating the functions of package **emdi** we estimate indicators using the data sets described in Section 5.3. The target variable is the equivalized income (`eqIncome`) and the regional level of interest are Austrian districts included in variable `district`. For direct estimation of the indicators the user has to specify these two arguments and the sample data set called `eusilcA_smp`. In addition, several other arguments are defined as shown below. We account for the sampling design by including the sampling weights in the estimation. Furthermore, we set the threshold argument to 60% of the median of equivalized income that - in this example - equals 10885.33 and we are also interested in obtaining the variance estimates of the indicators.

```
R> emdi_direct <- direct(y = "eqIncome", smp_data =
```

Arguments	Short description	Default
y	Target variable	
smp_data	Survey data	
smp_domains	Domain identifier	
weights	Sampling weights	No weights
design	Variable indicating strata	No design
threshold	Threshold for poverty indicators	60% of the median of the target variable
var	Variance estimation	No variance estimation
boot_type	Type of bootstrap: naive or calibrate	Naive
B	Number of bootstrap populations	50
seed	Seed for random number generator	123
X_calib	Calibration variables	None
totals	Population totals	None
custom_indicator	Customized indicators	None
na.rm	Deletion of observations with missing values	No deletion

Table 5.4: Input arguments for function direct. All explanations can also be found in the documentation of the direct function in the package.

```
+    eusilcA_smp, smp_domains = "district", weights = "weight",
+    threshold = 10885.33, var = TRUE)
```

The R object emdi_direct is of classes emdi and direct.
An example of using the ebp method for computing point and MSE estimates for the predefined indicators and two custom indicators, namely the minimum and maximum equivalized income is provided below:

```
R> emdi_model <- ebp(fixed = eqIncome ~ gender + eqsize + cash
+    self_empl + unempl_ben + age_ben + surv_ben + sick_ben +
+    dis_ben + rent + fam_allow + house_allow + cap_inv +
+    tax_adj, pop_data = eusilcA_pop, pop_domains = "district",
+    smp_data = eusilcA_smp, smp_domains = "district",
+    threshold = 10885.33, MSE = TRUE, custom_indicator =
+    list(my_max = function(y, threshold){max(y)}, my_min =
+    function(y, threshold){min(y)}))
```

In contrast to the direct estimation, the user also has to choose the auxiliary variables to be included in the nested error linear regression model. The variables that are chosen to explain the equivalized income are demographics as gender and the equivalized household size but also financial benefits and allowances as for example cash income, unemployment benefits and capital investement. Furthermore, model-based estimation requires the use of both, population (eusilcA_pop) and sample (eusilcA_smp) data and the domain identifiers. For enabling the comparison between direct and model-based estimates of the indicators of interest we use the same threshold as in the direct estimation. MSE estimates are returned by setting the MSE argument to TRUE. The final R object emdi_model is of classes emdi and model. For this object we show in the following subsections the **emdi** functionalities.

Arguments	Short description	Default
`fixed`	Fixed effects formula of the nested error regression model	
`pop_data`	Census or administrative data	
`pop_domains`	Domain identifier for population data, `pop_data`	
`smp_data`	Survey data	
`smp_domains`	Domain identifier for sample data, `smp_data`	
`L`	Number of Monte Carlo iterations	50
`threshold`	Threshold for poverty indicators	60% of the median of the target variable
`transformation`	Type of transformation: no, log or Box-Cox	Box-Cox
`interval`	Interval for the estimation of the optimal transformation parameter	(-1,2)
`MSE`	Mean Squared Error (MSE) estimation	No MSE estimation
`B`	Number of bootstrap populations	50
`seed`	Seed for random number generator	123
`boot_type`	Type of bootstrap: parametric or wild	Parametric
`parallel_mode`	Mode of parallelization	Automatic
`cpus`	Number of kernels for parallelization	1
`custom_indicator`	Customized indicators	None
`na.rm`	Deletion of observations with missing values	No deletion

Table 5.5: Input arguments for function `ebp`. All explanations can also be found in the documentation of the `ebp` function in the package.

5.4.2 Summary statistics and model diagnostics

R-users typically use a `summary` method for summarizing the results. For `emdi` objects the summary outputs differ depending on the two classes. The summary for objects obtained by direct estimation gives information about the number of domains in the sample, the total and domain-specific sample sizes. The summary for model-based objects is more extensive. In addition to information about the sample sizes, information about the population size and the number of out-of-sample domains is provided. Since model-based SAE relies on prediction under the model, including model diagnostics in **emdi** is important for users. A first measure to consider when evaluating the working model is the well known R^2. Nakagawa and Schielzeth (2013) provide a generalization of this measure for linear mixed models. A marginal R^2 and a conditional (a measure that accounts for the random effect) R^2 are implemented via function `r.squaredGLMM` in package **MuMIn** (Barton, 2018). The `summary` method uses this function to calculate and present both measures. For the EBP and model-based SAE methods in general the validity of parametric assumptions is crucial. Therefore, **emdi** also outputs residual diagnostics. In particular, results include the skewness and kurtosis of both sets of residuals (random effects and unit-level) and the results from using the Shapiro-Wilk test for normality (test statistic and p-value). The intra-cluster correlation (ICC) coefficient is further used for assessing the remaining unobserved heterogeneity. Finally, the `summary` command gives information about the selected transformation. If the user opts for a Box-Cox transformation, the transformation parameter λ and the shift parameter s are reported.

In addition to the diagnostics provided by `summary`, **emdi** enables the use of graphical diagnostics (see Figure 5.1). The `plot` method outputs graphics of residual diagnostics.

The first set of plots (Figure 5.1a) are Normal Quantile-Quantile (Q-Q) plots of Pearson unit-level residuals and standardized random effects. Figure 5.1b and 5.1c are kernel density plots of the distribution of the two sets of residuals contrasted against a standard normal distribution. Outliers can have a significant impact on the model fit and hence on prediction. Hence, a Cook's distance plot is also available (Figure 5.1d), in which the three largest values of the standardized residuals are identified alongside the case identification number for further investigation. Finally, if a Box-Cox transformation is used, a plot of the profile log-likelihood that shows the value of the transformation parameter for which the likelihood is maximized is also produced (Figure 5.1e). The user can customize the format of all plots. Method `plot` accepts the parameter `label` with the predefined values `blank` (deletes all labels) and `no_title` (axis labels are given, but no plot titles). In addition, a user-defined list that contains specific labels for each plot list can be given. Another parameter available is `color` which accepts a vector with two color specifications. The first color defines the lines in Figure 5.1a, 5.1d and 5.1e and the second one specifies the color of the shapes in Figure 5.1b and 5.1c. For the likelihood plot the range in which the likelihood should be computed can be specified by using the parameter `range`. The appearance of the plots benefits from the use of the **ggplot2** package (Wickham, 2009). Hence, `plot` accepts a `gg_theme` argument that allows for all customization options of `theme` that is a tool for modifying non-data components of a plot.

Example using Austrian districts:
In order to check the diagnostics in our example we use the `summary` and the `plot` methods. The summary output of the object `emdi_model` is presented below.

```
R> summary(emdi_model)

Empirical Best Prediction

Call:
ebp(fixed = eqIncome ~ gender + eqsize + cash + self_empl +
unempl_ben + age_ben + surv_ben + sick_ben + dis_ben +
rent + fam_allow + house_allow + cap_inv + tax_adj,
pop_data = eusilcA_pop, pop_domains = "district",
smp_data = eusilcA_smp, smp_domains = "district",
threshold = 10885.33, MSE = TRUE, custom_indicator =
list(my_max = function(y, threshold) {
max(y)
}, my_min = function(y, threshold) {
min(y)
}))

Out-of-sample domains:  24
```

```
In-sample domains:  70

Sample sizes:
Units in sample:  1945
Units in population:  25000
                   Min. 1st Qu. Median      Mean 3rd Qu. Max.
Sample_domains      14    17.0   22.5  27.78571   29.00  200
Population_domains  5   126.5  181.5 265.95745  265.75 5857

Explanatory measures:
Marginal_R2 Conditional_R2
    0.6325942       0.709266

Residual diagnostics:
               Skewness Kurtosis Shapiro_W    Shapiro_p
Error         0.7523871 9.646993 0.9619824 3.492626e-22
Random_effect 0.4655324 2.837176 0.9760574 1.995328e-01

ICC:  0.2086841

Transformation:
Transformation Method Optimal_lambda Shift_parameter
       box.cox   reml      0.6046901               0
```

This output helps to justify the use of a model-based approach for SAE in this specific example. On the one hand, 24 out of 94 districts are out-of-sample such that direct estimates cannot be produced for these districts. Furthermore, the sample sizes in the districts are rather small with a median of 22.5 households and vary between a minimum of 14 households and a maximum of 200 households. The explanatory power of the selected covariates is high with the conditional R^2, the measure that jointly considers the fixed and the random effect, of around 71%. The ICC of 20.9% further justifies the inclusion of a random effect. The normality tests show that normality is rejected for the unit-level error term but not for the random effect. The use of transformations helps to reduce the skewness of the distribution of the error terms. The optimal transformation parameter is 0.6 indicating that neither using the untransformed income or the logarithmic transformation of income would be appropriate for this data set. The plots in Figure 5.1 used for residual analyses of the object `emdi_model` can be produced as follows,

```
R> plot(emdi_model, label = "no_title", color =
     + c("red3", "red4"))
```

The Q-Q plots and the densities of the two error terms confirm that normality seems to be reasonable for the random effect but not for the unit-level error term. Furthermore, the Cook's distance plot identifies possible outliers. The last plot shows the optimal transformation parameter, which is the maximum of the profile log-likelihood.

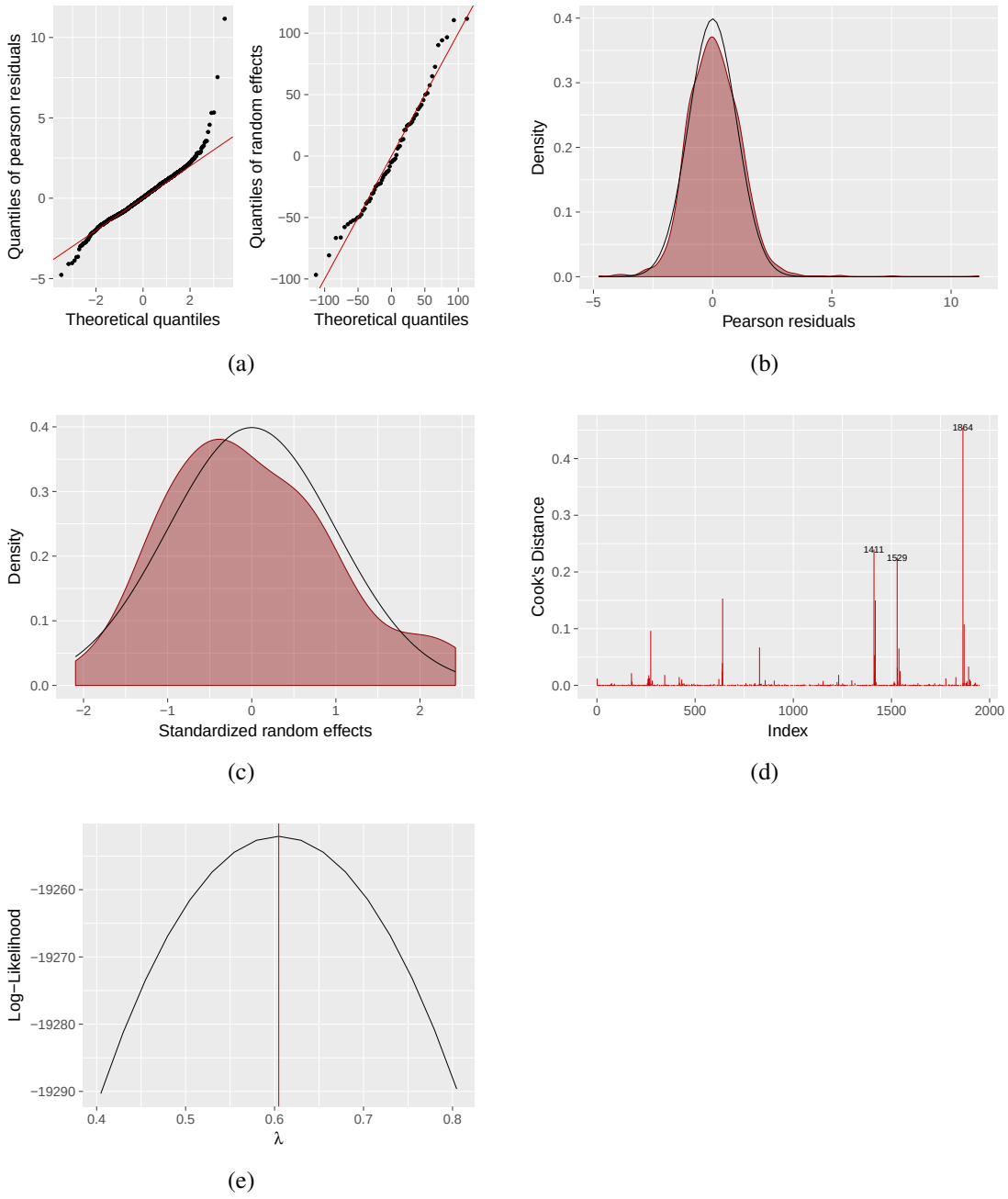


Figure 5.1: Graphics obtained by using `plot(emdi_model)`. (a) shows Normal Q-Q plots of the unit-level errors and the random effects. (b) and (c) show kernel density estimates of the distributions of standardized unit-level errors and standardized random effects compared to a standard normal distribution (black density). The Cook's distance plot is displayed in (d) whereby the index of outliers is labeled. The profile log-likelihood for the optimal parameter value of the Box-Cox transformation is shown in (e).

5.4.3 Selection and comparison of indicators

Package **emdi** returns a set of predefined and customized indicators. The ten predefined indicators are summarized in Table 5.1. However, the user may only be interested in some of these or only in individually defined (customized) indicators. A function called `estimators` helps the user to select the indicator or indicators of interest. This is done by using the `indicator` argument that takes a vector of indicator names as an argument, but in addition also accepts keywords defining predefined groups; for example, the keyword `custom` returns only user-defined indicators. In addition to variance and MSE estimates, NSIs often use the CV as an additional measure of the quality of the estimates. Estimated CVs as defined in Section 5.2 can be returned alongside MSE estimates.

It is often important to compare model-based and direct estimates. Direct estimates do not depend on the use of a model and hence the analyst should be interested in deriving model-based estimates that are close to direct estimates. Comparing model-based to direct estimates offers an internal validation procedure for checking whether the use of a model leads to unreasonable estimates. Package **emdi** provides a function called `compare_plot` that returns two plots, a scatter plot according to Brown et al. (2001) and a line plot. The scatter plot shows the direct and model-based point estimates, the fitted regression line, and the identity line. The closer the regression line is to the identity line, the closer the estimates are. The line plot is shown for domains ordered by the sample size. Thus, the user can see how the model-based estimates track the direct estimates across domains. In accordance with the function `estimators` the user can choose which indicators are compared by using the `indicator` argument. Similarly to the diagnostic plots, the user can modify the layout of the two plots. The label options are also `blank` (deletes all labels) and `no_title` (axis labels are given, but no plot titles). The color, the shape of the points and the type of the lines can be changed by using arguments `color`, `shape` and `line_type`, respectively.

Example using Austrian districts:

We illustrate how to estimate the median of equivalized income and the Gini coefficient and the corresponding CV estimates for the first 6 districts in Austria.

```
R> head(estimators(emdi_model, indicator = c("Gini", "Median"),
+    MSE = FALSE, CV = TRUE))
```

```
Domain      Gini    Gini_CV   Median  Median_CV
1 Eisenstadt-Umgebung 0.2214688 0.09790984 25414.07 0.10381883
2  Eisenstadt (Stadt) 0.2872751 0.06110093 49274.84 0.07673551
3             Güssing 0.1906263 0.13046770 16718.13 0.12732081
4         Jennersdorf 0.2098103 0.15371048 12869.55 0.17815504
5         Mattersburg 0.2091353 0.10851693 20102.09 0.12764578
6     Neusiedl am See 0.1865026 0.05934130 18386.83 0.06346778
```

For these districts, the Gini coefficient and the median income are highest in Eisenstadt (Stadt). The lowest Gini is in Neusiedl am See and the lowest median in Jennersdorf. Furthermore, it can be noted that none of the CVs is above 20%. This threshold is used by the ONS in

pop_data_id	shape_id
ID of domain 1 in the emdi obj	ID of domain 1 in the shape file
ID of domain 2 in the emdi obj	ID of domain 2 in the shape file
ID of domain 3 in the emdi obj	ID of domain 3 in the shape file
⋮	⋮

Table 5.6: Example of a mapping table for argument `map_tap` in function `map_plot` in **emdi**.

UK in order to decide if estimates can be reported.
The plots in Figure 5.2 are obtained by

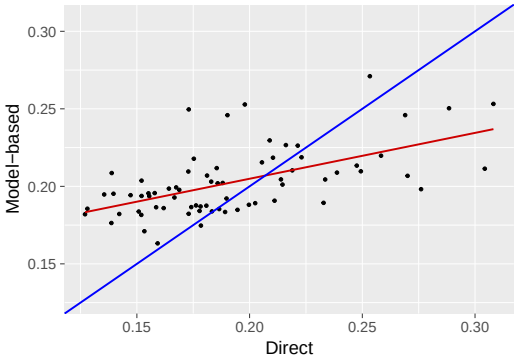
```
R> compare_plot(emdi_direct, emdi_model, indicator =
 +   c("Gini", "Median"), label = "no_title", color =
 +   c("red3", "blue"))
```

The scatter plots highlight that the disparity of the fitted regression line from the identity line is higher for the Gini coefficient than for the median. The model-based estimates do not track the direct estimates and show also a lower variability across the domains. In contrast, the direct and model-based estimates for the median are close to each other. Especially for large domains the difference is negligible.

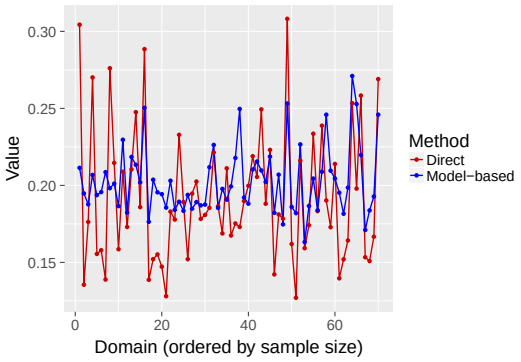
5.4.4 Mapping of the estimates

In SAE maps are a natural way to present the estimates as they help describing the spatial distribution of issues like poverty and inequality. Creating maps can be demanding or laborious in practice. Package **emdi** includes function `map_plot` that simplifies the creation of maps. Given a spatial polygon provided by a shape file and a corresponding `emdi` object `map_plot` produces maps of selected indicators and corresponding MSE and CV estimates. The parameters `MSE`, `CV` and `indicator` correspond to those in the `estimators` function. As Wickham (2009) points out the matching of domain identifiers in the statistical data to the corresponding identifiers in the spatial data (shape file) is challenging and general solutions are hard to obtain. The function `map_plot` in **emdi** allows for an argument `map_tab` when the identifiers do not match. The user must define a mapping table (cf. Table 5.6) for the argument `map_tab` in the form of a data frame that matches the domain variable in the population data set with the domain variable in the shape file. If the domain identifiers in both data sources match, this table is not required. The handling of the spatial shape files can be done using package **maptools** (Bivand and Lewin-Koh, 2017) in combination with package **rgeos** (Bivand and Rundel, 2017). Alternative approaches are provided by the packages **rgdal** (Bivand et al., 2018) and **sf** (Pebesma, 2018). For general information on how to work with spatial data and shape files we refer the reader to Bivand et al. (2013).

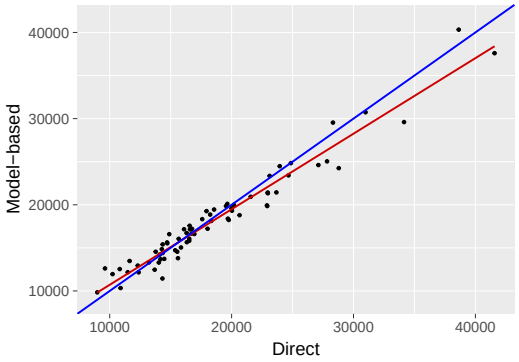
Example using Austrian districts:
The steps for obtaining a map of median income in Austrian districts and the corresponding CVs are outlined below. First, the shape file needs to be loaded.



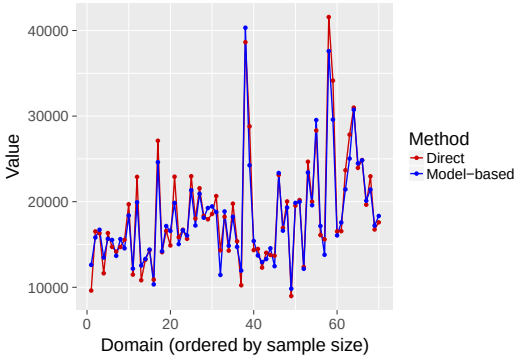
(a)



(b)

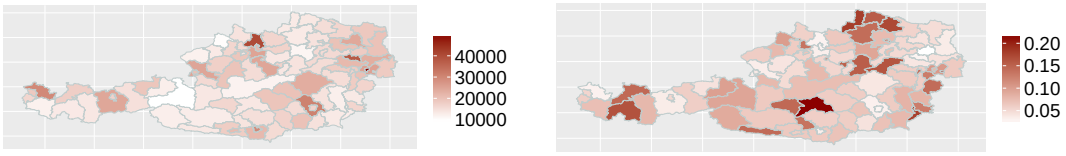


(c)



(d)

Figure 5.2: Graphics obtained by using `compare_plot(emdi_model)`. (a) and (c) show the scatter plots of the direct and model-based estimates for the Gini coefficient (top) and the median (bottom), respectively. (b) and (d) are line plots of the same estimates where the domains are ordered by increasing sample size.



(a) Median income. (b) CV of the median income estimates.

Figure 5.3: Maps of point estimates and CVs of the median income for 94 districts in Austria.

```
R> load_shapeaustria()
```

Then, two maps are created (cf. Figure 5.3).

```
R> map_plot(emdi_model, MSE = FALSE, CV = TRUE, map_obj =
+     shape_austria_dis, indicator = "Median",
+     map_dom_id = "PB")
```

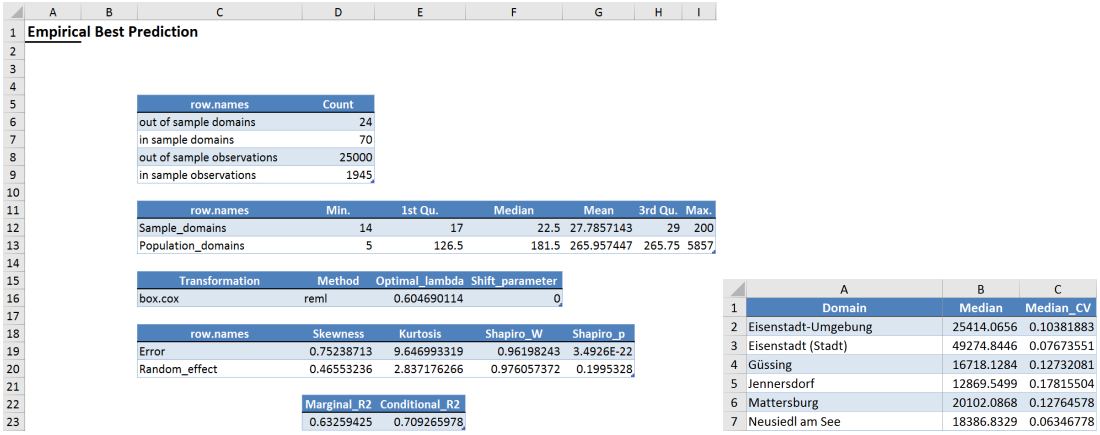
As the domain identifiers in the data set and shape file already match, the argument `map_tab` is not required. For an example where the argument `map_tab` needs to be specified, we refer the reader to `help(map_plot)`.

The map of the median equivalized income in Figure 5.3 indicates differences across Austrian districts. The richest district appears to be Eisenstadt (Stadt) followed by Urfahr-Umgebung. Furthermore, throughout the country some districts have a relatively low median income like Zell am See and Schärding. The map of the CVs shows that most districts have a CV below 20%. The highest CVs occur in the out-of-sample domains.

5.4.5 Exporting the results

Exporting the results from R to other widely used software such as Excel™ is important for users. For doing so a large set of well established tools already exists. Nevertheless, exporting all model information, including the information contained in the summary output is not straightforward. Function `write.excel` creates a new Excel™ file that contains the summary output in the first sheet and the results from the selected estimators in the following sheet. Again the parameters `MSE`, `CV` and `indicator` correspond to those in the `estimators` function. The link with the Excel™ file format is done by using the package **openxlsx** (Walker, 2017). This package does not require a Java™ installation, which offers an advantage over the use of the **xlsx** package (Dragulescu, 2014) because Java™ may be seen as a potential security threat. Nevertheless, package **openxlsx** (Walker, 2017) needs a zipping application available to R. Under Microsoft Windows™ this can be achieved by installing RTools while under macOS™ or Linux™ such an application is available by default. In addition to exporting the results to Excel™, **emdi** also provides an option to export output directly as OpenDocument Spreadsheets via the function `write.ods`.

Example using Austrian districts:



Empirical Best Prediction

row.names	Count
out of sample domains	24
in sample domains	70
out of sample observations	25000
in sample observations	1945

row.names	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Sample_domains	14	17	22.5	27.7857143	29	200
Population_domains	5	126.5	181.5	265.957447	265.75	5857

Transformation	Method	Optimal_lambda	Shift_parameter
box.cox	reml	0.604690114	0

row.names	Skewness	Kurtosis	Shapiro_W	Shapiro_p
Error	0.75238713	9.646993319	0.96198243	3.4926E-22
Random_effect	0.46553236	2.837176266	0.976057372	0.1995328

Marginal_R2	Conditional_R2
0.63259425	0.709265978

Domain	Median	Median_CV
Eisenstadt-Umgebung	25414.0656	0.10381883
Eisenstadt (Stadt)	49274.8446	0.07673551
Güssing	16718.1284	0.12732081
Jennersdorf	12869.5499	0.17815504
Mattersburg	20102.0868	0.12764578
Neusiedl am See	18386.8329	0.06346778

Figure 5.4: Export of the summary output and estimates to Excel™.

Excel™ outputs of model-based estimates for Austrian districts can be obtained by the following command.

```
R> write.excel(emdi_model, file = "excel_output.xlsx",
+     indicator = "Median", MSE = FALSE, CV = TRUE)
```

The output is presented in Figure 5.4 and shows that also the Excel™ user receives the same diagnostics from the summary and results for selected estimates. The summary output is described in detail in Section 5.4.2.

5.5 Additional Features

In addition to those features that are essential for estimating regional indicators, package **emdi** offers to incorporate external indicators and increases the computational efficiency of the MSE estimation by parallel computing. In this section we show how users can bring indicators from other R packages into **emdi** and how parallel computing can help with reducing the computational burden.

5.5.1 Incorporating an external indicator

A feature we should pay attention to is the ease by which indicators of other R packages can be brought into **emdi**. This is demonstrated by using the Theil index from the R package **ineq** (Zeileis, 2014). The Theil index describes economic inequality and thus can be also used in the application with the data of this paper. It belongs to a family of generalized entropy inequality measures and can be expressed by

$$\text{Theil}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \frac{y_{ij}}{\bar{y}} log\left(\frac{y_{ij}}{\bar{y}}\right),$$

where $\bar{y} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ (Cowell, 2011). The Theil index takes values from 0 to ∞ with 0 indicating equality and higher values increasing inequality (The World Bank, 2005). As the function `ineq` only requires a numeric vector of the target variable, it can be straightforwardly

wrapped into a form usable within the `direct` or `ebp` functions. Using the function `direct` the Theil index can be estimated as follows.

First, the package **ineq** needs to be installed and loaded.

```
R> install.packages("ineq")
R> library("ineq")
```

Subsequently, the function `ineq` with `type = "Theil"` can be given to the argument `custom_indicator`.

As the function `direct` needs the arguments `y`, `weights` and `threshold`, these arguments have to be also specified in the newly defined function.

```
R> my_theil <- function(y, weights, threshold) {
+    ineq(x = y, type = "Theil")
+    }
```

The argument `custom_indicator` needs to include a named list of self-defined indicators.

```
R> my_indicators <- list(theil = my_theil)
R> emdi_direct2 <- direct(y = "eqIncome", smp_data =
+    eusilcA_smp, smp_domains = "district", weights = "weight",
+    var = TRUE, custom_indicator = my_indicators)
```

As the Theil index is now part of the `emdi` object, all methods shown in Section 5.4 can be also used for this newly defined inequality indicator. For instance, by estimating a customized indicator via function `direct` a bootstrap variance estimator is used and the `subset` method can be applied in order to get results for certain districts.

```
R> select_theil <- estimators(emdi_direct2, indicator =
+    "theil", CV = TRUE)
R> subset(select_theil, Domain == "Wien")

   Domain     theil  theil_CV
67   Wien 0.1202542 0.1108617
```

5.5.2 Parallelization

Bootstrapping the MSE can be very costly in terms of computation time and the possibilities of speeding up are limited when staying within R. Nevertheless, as the bootstrap procedures described in Section 5.2.2 and Appendix .1 consist of B independent iterations, they are suitable for efficient parallel computing. In this particular case, parallelization may be described as follows:

1. The user predefines how many parallel processes (`cpus`) and bootstrap iterations (`B`) should be used in function `ebp`.

2. The bootstrap iterations are equally distributed on the parallel processes.

3. In each process the differences between EBP point estimates and the pseudo true values $\widehat{\Delta I}_{i,b} = \hat{I}_{i,b}^{EBP} - I_{i,b}$ (compare e.g., Appendix .1) are calculated. This is done on different central processing units (CPUs) at the same time (parallel computing).

4. The results $\widehat{\Delta I}_{i,b}$ from all processes are combined and the MSE is estimated by $\widehat{MSE}\left(\hat{I}_i^{EBP}\right) = B^{-1}\sum_{b=1}^{B}\left(\widehat{\Delta I}_{i,b}\right)^2$.

In R there are numerous ways and packages for implementing parallel computing. The most used package in this context is **parallel** (R Core Team, 2017), which mainly builds on the work of packages **snow** (Tierney et al., 2016) and **multicore** (Urbanek, 2014). These packages follow two different approaches for parallelization. Package **snow** launches a new version of R on each core. Those versions communicate with the master process through the so-called "socket". Therefore, we will proceed calling this way of parallelization the socket approach. The second approach is called "forking" and is the approach developed in the **multicore** package. Forking duplicates the entire current version of R and shifts it to a new core. Forking has one crucial advantage: all slave processes share the same memory with the master process for any object that is not modified. This feature makes it very fast. Its disadvantage is that it is not available on Microsoft Windows™ operating systems. The **parallel** package allows for both approaches but uses different functions. These functions are given an unified interface by the package **parallelMap** (Bischl and Lang, 2015). This interface for parallelization is used in **emdi**. In the `ebp` function the parallelization approach defaults to socket if a Microsoft Windows™ OS is detected and to forking otherwise. The parallelization is activated by setting the `cpus` argument to an integer value larger than 1. In the example below the computation time is measured when the number of CPUs is set equal to 1 and to 2, respectively:

```
R> system.time(emdi_model1 <- ebp(fixed = eqIncome ~ gender +
+    eqsize + cash + self_empl + unempl_ben + age_ben +
+    surv_ben + sick_ben + dis_ben + rent + fam_allow +
+    house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
+    pop_domains = "district", smp_data = eusilcA_smp,
+    smp_domains = "district", threshold = 10885.33, MSE =
+    TRUE, seed = 100, cpus = 1))

  user      system     elapsed
155.86        0.09      157.36

R> system.time(emdi_model2 <- ebp(fixed = eqIncome ~ gender +
+    eqsize + cash + self_empl + unempl_ben + age_ben +
+    surv_ben + sick_ben + dis_ben + rent + fam_allow +
+    house_allow + cap_inv + tax_adj, pop_data = eusilcA_pop,
+    pop_domains = "district", smp_data = eusilcA_smp,
+    smp_domains = "district", threshold = 10885.33, MSE =
+    TRUE, seed = 100, cpus = 2))

user      system     elapsed
3.62        0.45       89.45
```

The return value `elapsed` from function `system.time` informs the user about the real time that has passed from submitting the command until completion. Hence, the time comparison shows that two parallel processes reduce the time that is needed for the `ebp` function to run approximately by half. Please note that computation times are not replicable.

Despite the advantages in terms of computation time, parallelization comes with a major drawback. The reproducibility of results that depends on random number generations is non trivial. The usual `set.seed()` command that is used in R to ensure reproducibility is not sufficient due to the different R sessions used in parallel computing. In the socket approach, the function `clusterSetRNGStream()` from the **parallel** package is used to provide reproducible random number streams to each process that are far apart from each other. Therefore, all processes would produce different but reproducible random numbers. When using the forking approach, reproducibility can be more easily achieved by simply using a different random number generator. In the `ebp` function, `set.seed(seed, kind = "L'Ecuyer")` is used to set the random number generation to L'Ecuyer (L'Ecuyer et al., 2002) which is based on L'Ecuyer (1999). The multiple substreams of random numbers are created by the **rstream** package (Leydold, 2017) in both approaches. Please note that results obtained from parallel computation are only reproducible if the same number of processes and the same parallelization approach are used. The reproducibility is demonstrated below by reproducing the results with `cpus` equal to 2.

```
R> emdi_model22 <- ebp(fixed = eqIncome ~ gender + eqsize +
+    cash + self_empl + unempl_ben + age_ben + surv_ben +
+    sick_ben + dis_ben + rent + fam_allow + house_allow +
+    cap_inv + tax_adj, pop_data = eusilcA_pop, pop_domains =
+    "district", smp_data = eusilcA_smp, smp_domains =
+    "district", threshold = 10885.33, MSE = TRUE, seed =
+    100, cpus = 2)

R> all.equal(emdi_model2, emdi_model22)

[1] TRUE
```

5.6 Conclusion and Future Developments

In this paper we show how the **emdi** package can simplify the application of SAE methods. This package is, to the best of our knowledge, the first R SAE package that supports the user beyond estimation in the production of complex, non-linear indicators. Another important feature is that data-driven transformation parameters are estimated automatically. Estimating the uncertainty of small area estimates is achieved by using both parametric bootstrap and semi-parametric wild bootstrap. The additional uncertainty due to the estimation of the transformation parameter is also captured in MSE estimation. Customized parallel computing is included for reducing the computational time. The complexity in applying SAE methods is considerably reduced, useful diagnostic tools are incorporated and the user is also supported by the availability of tools for presenting, visualizing and further processing the results. For

instance, the model summary and results can be exported to Excel™ and to OpenDocument Spreadsheets. Since **emdi** makes the application of SAE methods in R almost as simple as fitting a linear or a generalized linear regression model, it also has the potential to close the gap between theoretical advances in SAE and their application by practitioners.

Additional features will be integrated in future versions of the package. Firstly, the implementation of alternative SAE methods will increase the usage of the package. For example, the World Bank (Elbers et al., 2003) and M-Quantile (Chambers and Chandra, 2006; Tzavidis et al., 2010) methods complement the EBP approach (Molina and Rao, 2010) for estimating disaggregated complex, non-linear indicators. Secondly, including additional evaluation and diagnostic tools for comparing direct and model-based estimates will assist the user with deciding which estimation method should be preferred. Thirdly, currently **emdi** includes only some possible types of transformations and one estimation method for the transformation parameter, namely REML. Future versions of the package will include a wider range of transformations (e.g., log shift and dual power transformations) and alternative estimation methods (minimization of the skewness or measures of symmetry) for the transformation parameter.

Acknowledgments

Rojas-Perilla, Schmid and Tzavidis gratefully acknowledge support by grant ES/N011619/1 - Innovations in Small Area Estimation Methodologies from the UK Economic and Social Research Council. The work of Kreutzmann and Schmid has been also supported by the German Research Foundation within the project QUESSAMI (SCHM 3113/2-1). The numerical results are not official estimates and are only produced for illustrating the methods.

Appendices

.1 Semi-parametric Wild Bootstrap

The semi-parametric wild bootstrap is implemented as follows,

1. Fit model 5.1 (using an appropriate transformation for $\mathbf{y}$) to obtain estimates $\hat{\boldsymbol{\beta}}, \hat{\sigma}_u^2, \hat{\sigma}_e^2, \hat{\lambda}$.
2. Calculate the sample residuals by $\hat{e}_{ij} = y_{ij} - \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} - \hat{u}_i$.
3. Scale and center these residuals using $\hat{\sigma}_e$. The scaled and centered residuals are denoted by $\hat{\epsilon}_{ij}$.
4. For $b = 1, ..., B$
   (a) Generate $u_i^{(b)} \overset{iid}{\sim} N(0, \hat{\sigma}_u^2)$.
   (b) Calculate the linear predictor $\eta_{ij}^{(b)}$ by $\eta_{ij}^{(b)} = \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} + u_i^{(b)}$.
   (c) Match $\eta_{ij}^{(b)}$ with the set of estimated linear predictors $\{\hat{\eta}_k | k \in n\}$ from the sample by using nn
   $$\min_{k \in n} \left| \eta_{ij}^{(b)} - \hat{\eta}_k \right|$$
   and define $\tilde{k}$ as the corresponding index.
   (d) Generate weights w from a distribution satisfying the conditions in Feng et al. (2011) where w is a simple two-point mass distribution with probabilities 0.5 at $w = 1$ and $w = -1$, respectively.
   (e) Calculate the bootstrap population as $T(y_{ij}^{(b)}) = \mathbf{x}_{ij}^\top \hat{\boldsymbol{\beta}} + u_i^{(b)} + w_{\tilde{k}} |\hat{\epsilon}_{\tilde{k}}^{(b)}|$.
   (f) Back-transform $T(y_{ij}^{(b)})$ to the original scale and compute the bootstrap population value $I_{i,b}$.
   (g) Select the bootstrap sample and use the EBP method as described above.
   (h) Obtain $\hat{I}_{i,b}^{EBP}$.
5. $\widehat{MSE}_{Wild}\left(\hat{I}_i^{EBP}\right) = B^{-1} \sum_{b=1}^{B} \left(\hat{I}_{i,b}^{EBP} - I_{i,b}\right)^2$.

A simulation study assessing the performance of the semi-parametric wild bootstrap is presented in Rojas-Perilla et al. (2017).

.2 Reproducibility

The results presented in this paper were obtained under R version 3.4.4 on a 64-bit platform under Microsoft Windows 7™. The installed packages are listed in Table 7. A snapshot of the corresponding repository was created with the package **packrat** (Ushey et al., 2018) and is available from the authors' `GitHub` folder (`https://github.com/SoerenPannier/emdi.git`). To make use of this repository Git must be installed. The authors recommend the following workflow:

- Use the new project functionality from RStudio.

- Choose checkout from version control and select Git.
- Enter the repository URL: `https://github.com/SoerenPannier/emdi.git`.
- Wait until **packrat** finishes the initialization process.
- Restart RStudio.
- Enter the R command `packrat::restore()`.
- After the package installation has finished all packages are installed as documented in Table 7.

Package	Version	Package	Version	Package	Version
assertthat	0.2.0	mgcv	1.8-23	stringi	1.1.7
backports	1.1.2	mime	0.5	stringr	1.3.0
BBmisc	1.11	minqa	1.2.4	testthat	2.0.0
BH	1.66.0-1	moments	0.14	tibble	1.4.2
boot	1.3-20	MuMIn	1.40.4	utf8	1.1.3
brew	1.0-6	munsell	0.4.3	viridisLite	0.3.0
cellranger	1.1.0	nlme	3.1-131.1	whisker	0.3-2
checkmate	1.8.5	nloptr	1.0.4	withr	2.1.2
cli	1.0.0	openssl	1.0.1	xml2	1.2.0
colorspace	1.3-2	openxlsx	4.0.17	base	3.4.4
commonmark	1.4	packrat	0.4.9-1	boot	1.3-20
crayon	1.3.4	parallelMap	1.3	class	7.3-14
curl	3.1	pillar	1.2.1	cluster	2.0.6
desc	1.1.1	pkgconfig	2.0.1	codetools	0.2-15
devtools	1.13.5	plyr	1.8.4	compiler	3.4.4
dichromat	2.0-0	praise	1.0.0	datasets	3.4.4
digest	0.6.15	R.cache	0.13.0	foreign	0.8-69
emdi	1.1.2	R.methodsS3	1.7.1	graphics	3.4.4
foreign	0.8-69	R.oo	1.21.0	grDevices	3.4.4
ggplot2	2.2.1	R.rsp	0.42.0	grid	3.4.4
git2r	0.21.0	R.utils	2.6.0	KernSmooth	2.23-15
glue	1.2.0	R6	2.2.2	lattice	0.20-35
gridExtra	2.3	RColorBrewer	1.1-2	MASS	7.3-49
gtable	0.2.0	Rcpp	0.12.16	Matrix	1.2-12
HLMdiag	0.3.1	RcppArmadillo	0.8.400.0.0	methods	3.4.4
hms	0.4.2	RcppEigen	0.3.3.4.0	mgcv	1.8-23
httr	1.3.1	readODS	1.6.4	nlme	3.1-131.1
ineq	0.2-13	readr	1.1.1	nnet	7.3-12
jsonlite	1.5	rematch	1.0.1	parallel	3.4.4
labeling	0.3	reshape2	1.4.3	rpart	4.1-13
laeken	0.4.6	rgeos	0.3-26	spatial	7.3-11
lattice	0.20-35	rlang	0.2.0	splines	3.4.4
lazyeval	0.2.1	RLRsim	3.1-3	stats	3.4.4
lme4	1.1-15	roxygen2	6.0.1	stats4	3.4.4
magrittr	1.5	rprojroot	1.3-2	survival	2.41-3
maptools	0.9-2	rstudioapi	0.7	tcltk	3.4.4
MASS	7.3-49	scales	0.5.0	tools	3.4.4
Matrix	1.2-12	simFrame	0.5.3	utils	3.4.4
memoise	1.1.0	sp	1.2-7		

Table 7: Packages installed while producing the results presented in this paper.

Part III

Discussion on the Applicability of Transformations

Chapter 6

Should we Transform Count Data Sets? Generalized Linear Models vs. Count Data Transformations

6.1 Introduction

We see and interpret the world as a set of discrete individual things that can be grouped: dogs, trees, countries and, thus, the act of counting is, usually, natural to all of us: two dogs, five trees, ten countries, among others. In statistics, these variables are known as counts and refer to enumerated events or observations often confined within a fixed time-interval or a defined area. Sometimes, one also may like to analyze variables that take only values within the interval $[0, 1]$, such as proportions or percentages: for instance, the proportion of animals habitating a specific area. Thus, if the aim is to model these non-continuous variables, linear regression may not be able to be directly used. In fact, it makes different key assumptions about the target variable, the explanatory variables, and their relationship. First, it is based on modeling the expected value of measurements from a continuous quantity (such as weights or income) as a linear function of quantitative and qualitative covariates. This is also called the linearity assumption. Second, the variability is attached by the normal distribution of the error regression terms (normality assumption), which are also assumed to be independent with constant variance (homoscedasticity assumption). If one aims to explain non-continuous variables using the classical linear regression model, a non-normal distributed error and heterogeneous variance structures arise and the above mentioned assumptions are not fulfilled. Typically, the conditional distribution of these data types can be skewed, their variances can be dependent on the mean, and they often contain many zero values (Blom, 1954). Even counts are easy to interpret: difficulties in the distribution of the observed variable can arise when the target variable is also bounded. Thus, directly using linear regression might yield inaccurate results and, moreover, might yield predictions for the target variable that lie outside the data range. Therefore, possible modifications in the response variable may be needed in order to apply the least squares estimation method and subsequent inference for the classical linear regression model. These modifications are known in the literature as transformations, and are broadly ap-

plied in this context in order to improve linearity, normality, and homosedasticity assumptions (Rocke (1993)). Proper transformations for non-continuous data often depend on the underlying assumed distribution of the target variable or on the variance structure inherent to the data. But even if no evidence of a model-specific process underlying the data is taken into account or can not be demonstrated, transformations can still be applied. The most prominent ones are the logarithmic function, the Box-Cox transformation, and different powers of roots, among others. However, is such a modification the only and most adequate device for modeling these variables?

A broad range of models suitable for the analysis of non-continuous data have emerged as an alternative approach. For instance, generalized linear models (GLMs) were proposed by Nelder and Wedderburn (1972) and extended by McCullagh and Nelder (1989). These models allow for directly modeling a target variable coming from the family of exponential distributions that includes in particular the Poisson, binomial, and negative binomial distributions. GLMs are broadly applied in a wide variety of disciplines, such as human biology, ecology, and social sciences. They are specified by a linear predictor; a link function, which describes how the mean of the target variable is related to the linear predictor; and a variance function, which describes the relationship between the variance and the mean. Furthermore, generalized linear mixed models (GLMMs) additionally account for dependency coming from repeated measurements made on the same statistical units. Therefore, the non-continuous variables mentioned above could be modeled by using GLMs and GLMMs. However, do these kinds of models remove the necessity of transforming non-continuous variables? In order to answer the research questions, the present paper compares these two approaches in terms of bias, root-mean-square error, and variance under count data sets, in particular the Poisson distribution. The performance of the generalized linear regression model and the classical linear regression model under different transformations, such as the Box-Cox and shifted square root, are studied in the present work.

The remainder of this paper is structured as follows. Detailed information about generalized linear regression models is given in Section 6.2. In Section 6.3, data transformations for count data sets are introduced. The most relevant comparison criteria are presented in Section 6.4. Section 6.5 presents a model-based simulation study under different scenarios. Finally, in Section 6.6, some concluding remarks and future research directions are presented.

6.2 Count Data Regression Models

Studying the relationship between explanatory variables and special response data types, such as counts, is a fundamental activity encountered in natural, social and medical sciences. For these data sets, a distribution from the exponential family of distributions is assumed for the response variable and is modeled by using GLMs. Additionally, defining the distribution of the response variable implicitly implies defining the relationship between the corresponding mean and variance. Therefore, GLMs are considered as an extension of the linear regression model for addressing the necessity of assuming a distributional form of the response variable, and in case specific variance structures are needed. Following Agresti (2015), GLMs are essentially

made up of the following components. Some of the most common practice and paper relevant GLMs are described in Table 6.1.

- *Random component (RC)*: Let y denote the target variable with n observations defined by $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$. The random component specifies the density function of y coming from the *exponential family of distributions*, which contains a set of probability distributions and takes the form:

$$f_y(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi) + c(y_i, \phi)}\right\},$$

  where θ is known as the *natural parameter* and ϕ the *dispersion parameter (DP)*. The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are assumed to be known. The most common distributions are the normal, binomial, and Poisson distributions.

- *Linear predictor (LP)*: Assume $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$, the values of the explanatory variables, with $\mathbf{X}$ the $n \times p$ design matrix. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, the vector of regression coefficients. The linear predictor reflects the linearity in the parameters, and is denoted by:

$$\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } i = 1, \ldots, n.$$

  In matrix form the linear predictor is expressed as $\mathbf{X}\boldsymbol{\beta}$.

- *Link function (LF)*: This is a monotonic and differentiable function, which is denoted by $g(\cdot)$ and describes how the linear predictor is related to the random component. In other words, how the mean $E(y_i) = \mu_i$ is related to the linear predictor as following:

$$g(\mu_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad \text{for } i = 1, \ldots, n.$$

  The most commonly used are the inverse, logarithmic, and logit link functions.

- *Variance function (VF)*: This function is given by $f(\cdot)$ and reflects how the variance and the mean are related:

$$\text{Var}(y_i) = \phi f(\mu_i).$$

Table 6.1: Generalized linear regression models common in practice

RC	Range	Mean	Variance	LF	VF	DP
Normal: $y_i \sim (\mu_i, \sigma^2)$	$(-\infty, \infty)$	μ_i	σ^2	Identity	1	σ^2
Poisson: $y_i \sim P(\lambda_i)$	$0, 1, \ldots$	λ_i	λ_i	$\log(\mu_i)$	μ_i	1
Negative Binomial: $y_i \sim NB(k_i, \pi_i)$	$0, 1, \ldots$	$\frac{k_i}{\pi_i}$	$\frac{k_i(1-\pi_i)}{\pi_i^2}$	$\log(\mu_i)$	$\frac{1-\mu_i}{\mu_i}$	$\frac{1}{k_i^2}$
Binomial: $y_i \sim \frac{B(k_i,\pi_i)}{k_i}$	$\frac{0,1,\ldots,k_i}{k_i}$	π_i	$\frac{\pi_i(1-\pi_i)}{k_i}$	$\text{logit}(\mu_i)$	$\mu_i(1-\mu_i)$	$\frac{1}{k_i}$

Poisson regression is the starting point of the analysis of count data. As McCullagh and Nelder (1989) states, "The Poisson distribution is the nominal distribution for counted data in much the same way that the normal distribution is the benchmark for continuous data".

The data coming from this discrete distribution generally describes the probability of a given number of events randomly occurring in time or space. Therefore, it takes on only integer values $0, 1, 2, \ldots$, without an upper limit. As a special case in GLMs, the Poisson regression model frequently uses a log-linear relationship between the linear predictor and the mean by including the logarithm as the link function, $g(\mu_i) = \log(\mu_i)$, and the dispersion parameter, $\phi = 1$. This model is known as the Poisson log-linear model. Typically, the iterative weighted least squares algorithm is used in order to estimate the vector of regression coefficients.

The Poisson distribution is entirely described by only one single parameter, namely the mean and denoted by λ, with $\lambda > 0$. This parameter is theoretically equal to the variance (Var$(y) = E(y)$). In practice, however, overdispersion and excess of zeros, are two frequent problems of empirical count data sets. The first issue occurs when, under a Poisson distribution, the Var$(y) > E(y)$. This may caused by heterogeneity between the units of observation or correlated responses. The Poisson distribution with underdispersion, i.e. Var$(y) < E(y)$, is less common in practice. There are three main ways of dealing with overdispersion. First, using a robust sandwich covariance matrix estimation method (Zeileis et al., 2008). Second, the dispersion parameter ϕ is not assumed to be equal to one and fixed, but rather, it is estimated from the data. It is also known as the quasi-Poisson model (Wedderburn, 1974). Finally, assuming a negative binomial distribution on the data set is the most common way in practice to accommodate overdispersion in count data regression modeling (McCullagh and Nelder, 1989). Meanwhile, hurdle count data regression models (Mullahy, 1986; Heilbron, 1994) deal with overdispersion and excess zero counts. Different approaches for directly modeling excess of zeros can be found in the literature: mixed Poisson distribution regression (Hinde and Demétrio, 1998), zero-inflated distribution models (Lambert, 1992; Greene, 1994), threshold models (Saei et al., 1996; Saei and McGilchrist, 1997), among others. Choosing one of these methodologies should be accomplished by graphical analysis and a sound scientific reasoning. For more detail about these and more research directions of modeling with count data sets, see for example Grogger and Carson (1991); Famoye (1993); Faddy (1997); Gurmu (1998) and Cameron and Trivedi (2013).

6.3 Count Data Transformations

Count data sets are characterized by having non-negative integers that can theoretically take values from zero to infinity, but may vary according to the nature of the regarded data. A count response exhibits inherent characteristics (non-negative and integer) that collide against essential aspects of the linear regression model, given that the latter models the target variable as a variable taking innumerable infinite values over the whole regression line. Researchers often ignore the discontinuity problem that arises with discrete data if the response variable takes many different distinct values and model it as continuous (Hilde, 2014). On the other hand, the truncation at zero rules out the simple and direct use of the normal distribution for computation of probabilities, in particular for data close to zero. As in standard statistical models, y is assumed to come from a certain probability distribution where each observation is independent from each other. The Poisson distribution, is commonplace, along with some of

their generalizations to represent this data type (Freeman and Tukey, 1949). For all of these distributions, the variance is a known function of the mean. This becomes a complication when applying classical linear techniques such as the analysis of variance. A further complication may arise since the relation between variability and mean level often suggests excessive skewness (Anscombe, 1948).

Transformations have received much attention as a method to stabilize variance and correct for normality for this type of data. Bartlett (1937) presents the square root transformation, of which the further developed transformations are based.

$$y_i^* = \sqrt{y_i}.$$

Due to the form of Poisson distributed data, this transformation is a natural suggestion for with a large mean (> 10). Bartlett (1937, 1947) suggests introducing a constant $c = \frac{1}{2}$ to the square root transformation: $\sqrt{y_i + \frac{1}{2}}$ when the target variable takes only small values and specially when zeros are common in the data set. Later, by allowing a more flexible form for the transformation by adding a constant parameter c, equal to

$$y_i^* = \sqrt{y_i + c}.$$

Anscombe (1948) finds that for large means the transformation in which $c = \frac{3}{8}$ produces the most nearly constant variance. This author suggests similar transformations related to the angular transformation for the negative binomial case, for both large and small values of the mean. Anscombe (1948) also demonstrates that a variance stabilizing transformation of the form $\sqrt{y + c}$, with $c \geq 0$ any fixed constant, and low sample mean values has the following feature

$$\lim_{\lambda \longrightarrow 0} \text{Var}(\sqrt{y_i + c}) = 0.$$

The work of Uddin et al. (2006) studies the relation between the mean and the parameter c using this family of transformations under different simulation settings. In this paper, the shifted root square transformation is applied. For this, an adaptive transformation parameter, denoted by s, is estimated according to the distribution features of the dataset with maximum likelihood theory, as in Rojas-Perilla et al. (2017). The adaptive root square transformation is defined as follows

$$y_i^*(s) = \sqrt{y_i + s}.$$

Freeman and Tukey (1950) carried out an empirical study on the use of transformations related to the arcsine and square root as a method to stabilize variance and normalize errors for Poisson distributed data. They suggest combining transformations, in particular using twofold transformations for both distributions and problems (see Table 6.2). Several of the distributions suitable for non-continuous data tend to normality, if n tends to infinity. This means, this will not always hold for small n. Curtiss (1943) studies the mathematical limitations of some of the above transformations under these situations. For these cases, Cornish and Fisher (1938) and Fisher and Cornish (1960) developed the *Cornish-Fisher expansions*, based on the quantiles from the empirical probability distribution. By using this approximation, Blom (1954) derives a

general transformation series form for Poisson and negative binomial distributed data so that the skewness correction is as small as possible. This solution is related to the beta transformation. They directly transform the mean of the distributions as shown in Table 6.2. Please notice that the selection of the parameters c, μ_0 are based on some cases, presented in detail in Blom (1954). However, the blom transformation series for count data sets are not commonly applied in practice.

Kendall et al. (1948) and Blom (1954) demonstrated that inducing homogeneity for some data sets also results in symmetry. However, it does not always mean perfect normality. The cube root transformation is used for count data in case symmetry is focus on the research. This means that the researcher should always have a criteria in the evaluation or requirements of model assumptions. Kendall et al. (1948) states that transformations which deal with heteroscedasticity problems in general also protect against non-symmetry. However, it always depends on the features of the data set. In order to improve the assumptions of the linear regression model, the logarithmic and Box-Cox transformations are commonly used for count data sets. The Box-Cox transformation includes the logarithm as a special case and adapts, like the shifted square root transformation, to the data set. Therefore, this paper focuses especially on these two data-driven transformations. The Box-Cox transformation is defined as

$$y_i^*(s) = \begin{cases} \frac{y_i^s - 1}{s} & \text{if } s \neq 0; \\ \log(y_i) & \text{if } s = 0. \end{cases}$$

Transformation	Functional form	Data range
Square root	$y_i^* = \sqrt{y_i}$	$y > 0$
	$y_i^* = \sqrt{y_i + 1}$	$y > -1$
Shifted square root	$y_i^*(s) = \sqrt{y_i + s}$	$y > -s$
Bartlett	$y_i^* = \sqrt{y_i + \frac{1}{2}}$	$y > -\frac{1}{2}$
Anscombe	$y_i^* = \sqrt{y_i + \frac{3}{8}}$	$y > -\frac{3}{8}$
Twofold	$y_i^* = \frac{1}{2}(\sqrt{y_i} + \sqrt{y_i + 1})$	$y > 0$
Logarithm	$y_i^* = \log(y_i)$	$y > 0$
Cube root	$y_i^* = y_i^{\frac{1}{3}}$	$y \in R$
Blom	$\mu^*(\lambda) = \begin{cases} \frac{1}{1-\lambda}\mu^{(1-\lambda)} + c & \text{if } \lambda \leq 1; \\ \log(\frac{\mu}{\mu_0}) & \text{if } \lambda = 1. \end{cases}$	$\mu > 0$
Box-Cox	$y_i^*(s) = \begin{cases} \frac{y_i^s - 1}{s} & \text{if } s \neq 0; \\ \log(y_i) & \text{if } s = 0. \end{cases}$	$y > 0$

Table 6.2: Transformations for count data sets

6.4 Methodological Differences

Choosing between the linear regression model or the generalized linear regression model for count data sets could be seen as a model selection problem. McCullagh and Nelder (1989) de-

scribe it also as a part of selecting the right scale for analysis, taking the research purpose into account. But should we use the original scale, or the transformation scale of the target variable? Jereys (1961) states, "It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude." In linear regression models it is crucial to analyze the fulfillment of model assumptions, such as normality, homoscedasticity, and linearity. In particular, combining Poisson-distributed data under this kind of models usually leads to different possible "good" scales. For instance, the square root transformation of the target variable often stabilizes the variance. Meanwhile, the cube root of the squared variable gives approximate symmetry or normality. In practice, finding a common scale may mean choosing a suitable transformation that simultaneously improves these model assumptions. This is a seldom feature obtained by applying a selected transformation for right skewed distributions with variance equal to the mean, as in the Poisson distribution case. The most suitable transformation to achieve homoscedasticity frequently differs from the best transformation to achieve symmetry or normality (Agresti, 2015). Furthermore, transformation comes at some cost to the trade-off between accuracy and interpretability (O'Hara and Kotze, 2010; Ives, 2015). The interpretability on the original scale of measurement is often preferable to the transformed scale. However, if the transformed scale is chosen, such as the logarithm scale, conclusions can be also presented in some cases on this scale and the subsequent inference of the linear regression models can be applied. However, applying a logarithmic transformation for count data often leads to results that are not defined on the original scale, particularly, if the data is highly skewed and contains many outliers. Additionally, in Poisson log-linear models, model parameters express the effects of the covariates on $\log[E(y|x)]$. In order to obtain the information on $E(y|x)$, these effects can be translated to an exponential model for the mean by using the inverse link function. In contrast, if a logarithmic transformation is applied to the linear regression model, the model parameters are defined only on $E[\log(y)|x]$, but not exactly on $E(y|x)$ (Agresti, 2015).

One of the biggest challenges that researchers face when working with the linear regression model under transformations is the bias problem. Often, after fitting such a model, it is common to want to return to an untransformed scale. The bias is produced in the inverse transformation process of a non-linear transformation. In general, a non-linear function has a non-linear inverse. In fact, $E[t(y)|x]$ is not equal to $t[E(y|x)]$, for most functions $t(\cdot)$ applied in the response variable. Expressing this issue for the linear regression model leads to:

$$\begin{aligned} t^{-1}\Big[E(\mathbf{x}\beta+\mathbf{e})\Big] &= t^{-1}\Big[E(\mathbf{x}\beta)+E(\mathbf{e})\Big] \\ &= t^{-1}\Big[E(\mathbf{x}\beta)\Big] \\ &= t^{-1}\Big[E(\mathbf{y}|\mathbf{x})\Big] \\ &\neq E\Big[t^{-1}(\mathbf{x}\beta+\mathbf{e})\Big] \end{aligned}$$

Therefore, it becomes a common problem in practice to determine the magnitude of the bias caused by applying a specific transformation. If no attention is paid to this problem, grossly misleading conclusions can be produced. On the contrary, GLMs directly model the conditional expectation of the target variable in the original scale. Therefore, using GLMs does

not produce this kind of bias. This is naturally one of the reasons why the researchers prefer fitting this model, instead of analyzing the possible bias problem inherent to the non-linear transformations by using the linear regression model.

GLMs are considered to be a unified theory of modeling some prominent continuous and non-continuous response variables, for which the random component is separately chosen from the choice of the link function. That means the probability distribution and the variance structure can be defined by the researcher in case they are known. However, Gelman and Hill (2006) and Ives (2015) point out that these distributional assumptions are not carefully analyzed in common practice. On the contrary, transformations may be useful when no evidence of the exact definition of the underlying probability distribution of counts is known, such as Poisson-like processes. Furthermore, in case the counts are large in the data set, the linear regression model can be useful as an alternative to GLMs (Warton et al., 2016). Additionally, GLMs and GLMMs have some mathematical limitations and computational complications in case other correlation structures or data-type of covariates are needed.

Finally, a challenge regarding the research purpose also arises when choosing between the linear regression model or the generalized linear regression model for count data. Is the focus paid on prediction or inference? If the research is only concerned with statements about the likely values of the target variable under a question of the form "What is the predicted value of the response under the selected model?", the prediction problem should be the key point. The analysis should be accompanied by some measures of precision, such as the root-mean-squared error and bias deviation, and some measures of goodness of fit.

6.5 Simulation Study for the Mean

In order to assess the performance of the linear regression model under specific transformations and the generalized linear regression model for count data, in terms of prediction, model-based evaluations are carried out in this paper. The simulation study was implemented in the open-source software R (R Core Team, 2017) by using different R packages. For instance, the glm() function from package stats and the function glm.nb() from the package MASS (Venables and Ripley, 2002). Transformations provided by the package trafo (Medina et al., 2017) are used.

For the simulation study: a fixed Poisson distributed population of size $N = 10000$ is used for generating $S = 200$ sub-samples for a fixed sample size n. This was repeated for $n = 50, 100, 150, \ldots, 1000$. For practical reasons, only the logarithmic function is used in this paper as the link function for Poisson data, as presented in Table 6.1. As mentioned before, the Poisson regression model under these specifications, is also known as the Poisson log-linear model. The population is generated as follows

$$\begin{aligned} \text{Population:}\quad \lambda_i &= \exp(2.5 + x_i), \\ x_i &\sim U(0, 1.2). \\ \Longrightarrow Y &\sim \text{Poisson}(\lambda), \text{ with } P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!}. \end{aligned}$$

Additionally, a weighted version of the least squares method is used for the maximum likelihood parameter estimates (Agresti, 2015). Additional simulation results for the binomial negative case are available from the author on request.

The Poisson log-linear regression model fitted in every sub-population is defined as follows

$$\text{Model 1 (}\textit{Poisson}\text{):} \qquad \log(\lambda_i) = \beta_0 + \beta_1 x_i.$$

In parallel and with the idea of finding the most suitable transformation, the shifted square root and the Box-Cox type transformations, described in Table 6.2, are applied for the linear regression model as

$$\text{Model 2 (}\textit{Box-Cox}\text{):} \qquad \text{Box-Cox}(y_i) = \beta_0 + \beta_1 x_i + e_i.$$

$$\text{Model 3 (}\textit{SQRT}\text{):} \qquad \text{Shifted square root}(y_i) = \beta_0 + \beta_1 x_i + e_i,$$

where e_i is the unit-level error term and is expected to be normal distributed under the applied transformation.

The results focus on the estimation of the average number of counts in the population. There are different quality measures for assessing the performance of mean estimators. This paper focuses on only two of them: the root-mean-squared error (RMSE) and the bias/relative bias (RB). Let $\hat{\lambda}$ be the estimated mean and λ the corresponding true value, known from the simulations. The RMSE is defined as:

$$\text{RMSE}\left(\hat{\lambda}\right) = \left[\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\lambda}_s - \lambda_s\right)^2\right]^{1/2}.$$

The RMSE measures the differences between the estimated mean and the respectively true value in each sub-population. It values are non negative and lie between 0 and ∞.

The RB provides a measure that indicates the relative differences of the estimated means and their respectively true values and lies between $-\infty$ and ∞.

$$\text{Bias}\left(\hat{\lambda}\right) = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\lambda}_s - \lambda_s\right),$$

$$\text{RB}\left(\hat{\lambda}\right) = \frac{1}{S}\sum_{s=1}^{S}\left(\frac{\hat{\lambda}_s - \lambda_s}{\lambda_s}\right) \times 100.$$

Figure 6.1 presents the RB, the RMSE, and the variances for the generated predictions, for which different sample sizes are used. As expected, the Poisson approach leads to more efficient results, in terms of RMSE, than the transformation-based approaches. For this, the RMSE always decreases as the sample size increases. In case of using the Box-Cox and SQRT transformations, the RMSE also decreases as the sample size increases, but it remains constant at one point. The Box-Cox transformation is well-known for leading both, stabilized variances

and more symmetric data. It is perhaps not surprising that the Box-Cox transformation leads to more accurate results than the SQRT approach. In fact, the SQRT could be seen as an special case of the Box-Cox transformation and therefore, the same picture is obtained for the negative binomial case and it is also expected for some other count distributions.

Let us now turn to the bias results. As we can see, the Poisson approach has negligible bias, whereas the transformed models lead to biased results. This is observed independent of the sample size. If we in parallel analyzed the variances, we notice that the bias problem describe in Section 6.4 is clearly present. The variances by using these three approaches are similar to each other. The differences in the RMSE are possibly due to the non-correction bias in this analysis. As noted in Section 6.2, after applying a specific transformation on the target variable, it is necessary to make an analysis of the bias, coming from the back-transforming process. However, how do we know for certain the magnitude of the bias under these transformations?

As mentioned in Section 6.4, bias incurred when back-transforming to the original scale is almost inherent to this process and we should be aware of this (Anscombe, 1948; Neyman and Scott, 1960). It always depends on which transformation we are using in practice. The statistical complications by back-transforming data to the original scales has been extensively discussed (Laurent, 1963; Patterson, 1966; Duan, 1983; Miller, 1984; Rothery, 1988; Smith, 1993; Rainey, 2017). For instance, the work of Rothery (1988) states that the magnitude of the bias mainly depends on the variance and less upon on the sample size. Different bias corrections for the regression parameters introduced by some specific transformations from Table 6.2 have been proposed in the literature. For instance, most relevant results for the logarithmic transformation were published until the early 1980s in different research fields. For example, one of the simplest analytical bias correction approach when using the logarithmic transformation is proposed by Sprugel (1983), in which the back-transformed estimates are multiplied by a constant equal to the half of the variance error. For further insights regarding this problem under the logarithmic transformation see Finney (1941); Meyer (1941); Neyman and Scott (1960); Goldberger (1968); Heien (1968); Aitchison and Brown (1969); Bradu and Mundlak (1970); Zellner (1971); Baskerville (1972); Beauchamp and Olson (1973); Sprugel (1983); Rukhin (1986) and Newman (1993). In case of using the Anscomble variance stabilizing transformation, the paper of Makitalo and Foi (2011) gives an approximation of the bias correction. For the Box-Cox transformation also different approaches have been proposed. For detailed information see Taylor (1986); Smallwood et al. (1986); Sakia (1988) and Sakia (1990).

The researchers can use some of these bias correction solutions according to the transformation type. However, these methods are mainly studied for the logarithmic and Box-Cox transformations under the linear regression model. This is still under research for the other transformations presented in Section 6.3. In particular, for data-driven transformations. If a bias correction is carried out for the analysis presented before, it is expected that the differences in the bias would not that much as noted in Figure 6.1.

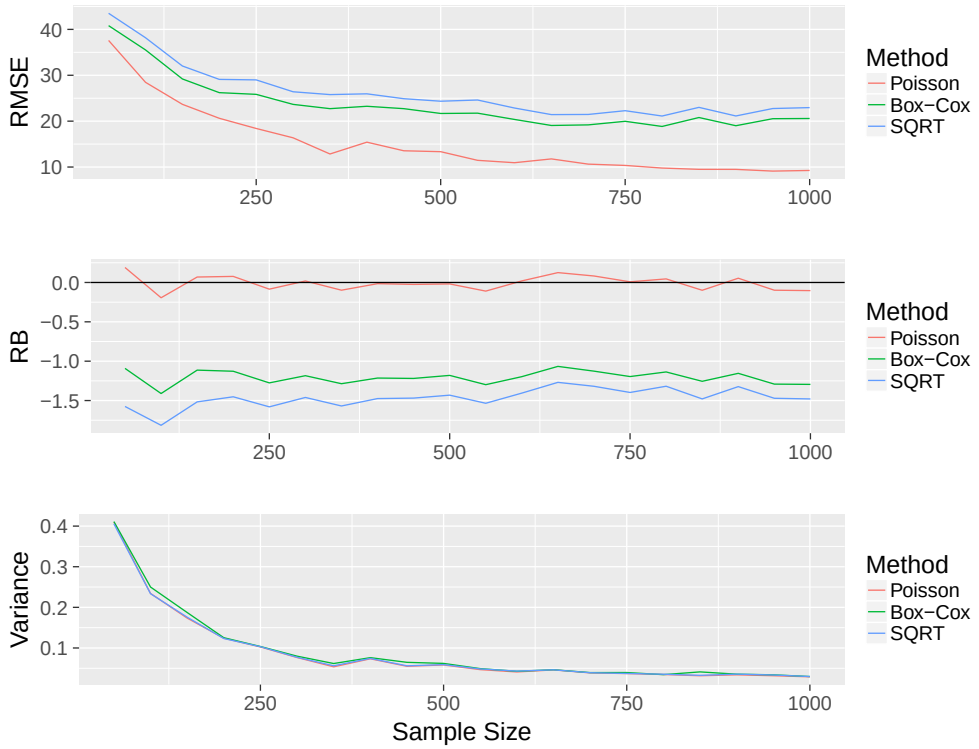


Figure 6.1: RMSE, relative bias, and conditional variances of the predictors for the estimation of the mean

6.6 Conclusions and Further Research Directions

The broad goal of this paper is to analyze the methodical differences between the linear regression model under transformations and the generalized linear regression model, in particular the Poisson log-linear regression model. Choosing which methodology should be preferable, always depends on the research question. As we already noted, using non-linear transformations for count data sets have different challenges for researches. First, the selection of a suitable transformation should be part of a previous careful analysis of the data to be studied. The distributional form of the underlying distributional process, the data range, and some features of distributional moments are some of the characteristics to be included in this previous analysis. For instance, in case the underling process of the data is not previously known, data transformations are able to adapt on different count data distributions. In the scenario studied in Section 6.5 the Poisson distribution was used for representing the underlying distributional process of the data. Thus, the exactly distribution of the target variable was applied in the context of GLMs. In such a scenario, the use of GLMs are usually recommended in practice. Second, selecting only one transformation that improves all distributional assumptions of the linear regression model is not always straightforward. Thus, it is not common to have in practice one transformation, which in parallel corrects the model assumptions in the same way. Therefore, the research should know in which scale is the analyses made or the criteria of selecting one suitable transformation. Third, if a selected transformation is applied on a target variable and the researcher needs to return to the original measurement scale, a bias correction analysis should be proposed.

According to the simulation settings, more realistic scenarios should be implemented. These include, count data sets presenting the overdispersion problem and/or an excess number of zeros. One research gap is analyzing the influence of different, particularly, of lower values of λ when one is working under the square root space. One would expect a more accurate stabilization of the variance. More research is needed for the comparison between the two approaches in terms of inference tools, assessment of model assumptions, and goodness of fit. For instance, the Pearson, Anscombe, and deviance residuals are widely used for assessing model fit under these models. For more information about the definition of residuals in non-continuous variables, see Anscombe (1953); Cox and Snell (1968) and Pierce and Schafer (1986). It would be interesting to incorporate a detailed analysis of the accurate estimation under other distributional processes. Finally, the analysis of bias correction approaches under different count data transformations should be addressed in further research.

Bibliography

Agresti, A. (2015). Foundations of linear and generalized linear models. John Wiley & Sons.

Agresti, A., B. Caffo, and P. Ohman-Strickland (2004). Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Computational Statistics & Data Analysis 47, 639–653.

Aitchison, J. and J. Brown (1969). The lognormal distribution: With special reference to its uses in economics. Cambridge University Press.

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. Biometrics 55, 117–128.

Alfons, A. and M. Templ (2013). Estimation of social exclusion indicators from complex surveys: The R package **laeken**. Journal of Statistical Software 54, 1–25.

Alfons, A., M. Templ, and P. Filzmoser (2010). Contamination models in the R package **simFrame** for statistical simulation. In S. Aivazian, P. Filzmoser, and Y. Kharin (Eds.), Computer Data Analysis and Modeling: Complex Stoachstic Data and Systems, Volume 2, pp. 178–181.

Anderson, T. W. and D. A. Darling (1954). A test of goodness of fit. Journal of the American Statistical Association 49, 765–769.

Andrews, D. F. (1971). A note on the selection of data transformations. Biometrika 58, 249–254.

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. Biometrika 35, 246–254.

Anscombe, F. J. (1953). Contribution to the discussion of H. Hotelling's paper. Journal of the Royal Statistical Society, Series B 15, 229–230.

Anscombe, F. J. and I. Guttman (1960). Rejection of outliers. Technometrics 2, 123–147.

Anscombe, F. J. and J. W. Tukey (1963). The examination and analysis of residuals. Technometrics 5, 141–160.

Armitage, P., G. Berry, and J. N. S. Matthews (2008). Statistical methods in medical research. John Wiley & Sons.

Asar, Ö., O. Ilk, and O. Dag (2017). Estimating Box-Cox power transformation parameter via goodness-of-fit tests. Communications in Statistics - Simulation and Computation 46, 91–105.

Atkinson, A. C. (1973). Testing transformations to normality. Journal of the Royal Statistical Society, Series B 35, 473–479.

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables. Journal of the Royal Statistical Society, Series B 44, 1–36.

Atkinson, A. C. (1986). Diagnostic tests for transformations. Technometrics 28, 29–37.

Atkinson, A. C. (1987). Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis. Clarendon Press.

Atkinson, A. C. and M. Riani (2012). Robust diagnostic regression analysis. Springer Science & Business Media.

Barnett, V. and T. Lewis (1984). Outliers in statistical data. John Wiley & Sons.

Barry, D. (1993). Testing for additivity of a regression function. The Annals of Statistics 21, 235–254.

Bartlett, M. S. (1935). The effect of non-normality on the t distribution. Mathematical Proceedings of the Cambridge Philosophical Society 31, 223–231.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. Proceedings of the Royal Society of London, Series A 160, 268–282.

Bartlett, M. S. (1947). The use of transformations. Biometrics 3, 39–52.

Bartlett, M. S. and D. Kendall (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. Supplement to the Journal of the Royal Statistical Society 8, 128–138.

Barton, K. (2018). **MuMIn**: Multi-Model Inference. R package version 1.40.4.

Baskerville, G. L. (1972). Use of logarithmic regression in the estimation of plant biomass. Canadian Journal of Forest Research 2, 49–53.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using **lme4**. Journal of Statistical Software 67, 1 – 48.

Battese, G. E., R. M. Harter, and W. A. Fuller (1988). An error component model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association 83, 28–36.

Beall, G. (1942). The transformation of data from entomological field experiments so that the analysis of variance becomes applicable. Biometrika 32, 243–262.

Beauchamp, J. J. and J. S. Olson (1973). Corrections for bias in regression estimates after logarithmic transformation. Ecology 54, 1403–1407.

Bedoya, H., S. Freije, L. Vila, G. Echeverria, D. Biller, G. M. Grandolini, R. Albisetti, E. Quintrell, and R. Vish (2013). Country partnership strategy for the United Mexican States (2014-2019).

Bell, W. R. and E. T. Huang (2006). Using the t-distribution to deal with outliers in small area estimation. In Proceedings of Statistics Canada Symposium 2006: Methodological Issues in Measuring Population Health.

Belsley, D. A., E. Kuh, and R. E. Welsch (2005). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley & Sons.

Berry, W. D. (1993). Understanding regression assumptions. SAGE Publications.

BIAS (2005). Bayesian methods for combining multiple individual and aggregate data sources in observational studies. `http://www.bias-project.org.uk/`. Accessed: 11.04.2016.

Bickel, P. J. and K. A. Doksum (1981). An analysis of transformations revisited. Journal of the American Statistical Association 76, 296 – 311.

Bischl, B. and M. Lang (2015). **parallelMap**: Unified interface to parallelization back-ends. R package version 1.3.

Bivand, R., T. Keitt, and B. Rowlingson (2018). **rgdal**: Bindings for the "Geospatial" data abstraction library. R package version 1.2-18.

Bivand, R. and N. Lewin-Koh (2017). **maptools**: Tools for reading and handling spatial objects. R package version 0.9-2.

Bivand, R., E. Pebesma, and V. Gómez-Rubio (2013). Applied spatial data analysis with R. Springer Science & Business Media.

Bivand, R. and C. Rundel (2017). **rgeos**: Interface to Geometry Engine - Open Source ("GEOS"). R package version 0.3-26.

Bland, J. M. and D. G. Altman (1996). Transformations, means, and confidence intervals. BMJ: British Medical Journal 312, 1078–1079.

Blaylock, J. R. and D. M. Smallwood (1985). Box-Cox transformations and a heteroscedastic error variance: Import demand equations revisited. International Statistical Review / Revue Internationale de Statistique 53, 91–97.

Blom, G. (1954). Transformations of the binomial, negative binomial, Poisson and χ^2 ditributions. Biometrika 41, 302–316.

Bock, R. (1985). Multivariate statistical methods in behavioral research. Scientific Software International.

Boonstra, H. (2012). **hbsae**: Hierarchical Bayesian small area estimation. R package version 1.0.

Booth, J. and J. Hobert (1998). Standard errors of prediction in generalized linear mixed models. Journal of the American Statistical Association 93, 262–272.

Box, G. E. and P. W. Tidwell (1962). Transformation of the independent variables. Technometrics 4, 531–550.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26, 211–252.

Box, G. E. P. and D. R. Cox (1982). An analysis of transformations revisited, rebutted. Journal of the American Statistical Association 77, 209–210.

Boylan, T. A., M. P. Cuddy, and I. G. O'Muircheartaigh (1982). Import demand equations: An application of a generalized Box-Cox methodology. International Statistical Review / Revue Internationale de Statistique 50, 103–112.

Bradley, J. V. (1977). A common situation conducive to bizarre distribution shapes. The American Statistician 31, 147–150.

Bradu, D. and Y. Mundlak (1970). Estimation in lognormal linear models. Journal of the American Statistical Association 65, 198–211.

Breidenbach, J. (2015). **JoSAE**: Functions for some unit-level Small Area Estimators and their variances. R package version 0.2.3.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. Australian Journal of Statistics 5, 93–105.

Brown, G., R. Chambers, P. Heady, and D. Heasman (2001). Evaluation of small area estimation methods - an application to unemployment estimates from the UK LFS. In Proceedings of Statistics Canada Symposium 2001: Achieving Data Quality in a Statistical Agency: A Methodological Perspective.

Brown, J. D. (2015). Linear models in matrix form: A hands-on approach for the behavioral sciences. Springer.

Buchinsky, M. (1995). Quantile regression, Box-Cox transformation model, and the US wage structure, 1963–1987. Journal of Econometrics 65, 109–154.

Bundesamt für Eich- und Vermessungswesen (2017). Verwaltungsgrenzen (VGD) - 1:250.000 Bezirksgrenzen, Daten vom 01.04.2017 von SynerGIS. [accessed: 07.02.2018].

Burbidge, J. B., L. Magee, and A. L. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. Journal of the American Statistical Association 83, 123–127.

Burbidge, J. B. and A. L. Robb (1985). Evidence on wealth-age profiles in Canadian cross-section data. Canadian Journal of Economics 18, 854–875.

Burnham, K. P. and D. R. Anderson (2004). Multimodel inference: Understanding AIC and BIC in model selection. Sociological Methods & Research 33, 261–304.

Cameron, A. C. and P. K. Trivedi (2013). Regression analysis of count data. Cambridge University Press.

Cameron, M. A. (1984). Choosing a symmetrizing power transformation. Journal of the American Statistical Association 79, 107–108.

Carroll, R. J. (1980). A robust method for testing transformations to achieve approximate normality. Journal of the Royal Statistical Society, Series B 42, 71–78.

Carroll, R. J. (1982a). Tests for regression parameters in power transformation models. Scandinavian Journal of Statistics 9, 217–222.

Carroll, R. J. (1982b). Two examples of transformations when there are possible outliers. Journal of the Royal Statistical Society, Series C 31, 149–152.

Carroll, R. J. and D. Ruppert (1981). On prediction and the power transformation family. Biometrika 68, 609–615.

Carroll, R. J. and D. Ruppert (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79, 312–313.

Carroll, R. J. and D. Ruppert (1985). Transformations in regression: A robust analysis. Technometrics 27, 1–12.

Carroll, R. J. and D. Ruppert (1987). Diagnostics and robust estimation when transforming the regression model and the response. Technometrics 29, 287–299.

Carroll, R. J. and D. Ruppert (1988). Transformation and weighting in regression. CRC Press.

Ceriani, L. and P. Verme (2012). The origins of the gini index: Extracts from variabilità e mutabilità (1912) by corrado gini. The Journal of Economic Inequality 10, 421–443.

Chakravarti, I. M. and R. G. Laha (1967). Handbook of methods of applied statistics. John Wiley & Sons.

Chambers, J. M., W. S. Cleveland, B. Kleiner, P. A. Tukey, et al. (1983). Graphical methods for data analysis. Wadsworth Belmont, CA.

Chambers, J. M. and T. Hastie (1992). Statistical models in S. Wadsworth & Brooks California.

Chambers, R. and H. Chandra (2006). Improved direct estimators for small areas. Working paper.

Chambers, R., H. Chandra, N. Salvati, and N. Tzavidis (2014). Outlier robust small area estimation. Journal of the Royal Statistical Society, Series B 76, 47–69.

Chambers, R., H. Chandra, and N. Tzavidis (2011). On bias-robust mean squared error estimation for pseudo-linear small area estimators. Survey Methodology 37, 153–170.

Chambers, R. L. (1986). Outlier robust finite population estimation. Journal of the American Statistical Association 81, 1063–1069.

Chandra, H., U. Sud, and Y. Gharde (2014). Small area estimation using estimated population level auxiliary data. Communications in Statistics - Simulation and Computation.

Changyong, F., W. Hongyue, L. Naiji, C. Tian, H. Hua, L. Ying, et al. (2014). Log-transformation and its implications for data analysis. Shanghai Archives of Psychiatry 26, 105–109.

Chatterjee, S. and A. S. Hadi (2015). Regression analysis by example. John Wiley & Sons.

Chen, S. (2002). Rank estimation of transformation models. Econometrica 70, 1683–1697.

Chen, W. W. and R. S. Deo (2004). Power transformations to induce normality and their applications. Journal of the Royal Statistical Society, Series B 66, 117–130.

Cheng, T.-C. (2005). Robust regression diagnostics with data transformations. Computational Statistics & Data Analysis 49, 875–891.

Chung, S. H., W. L. Pearn, and Y. S. Yang (2007). A comparison of two methods for transforming non-normal manufacturing data. The International Journal of Advanced Manufacturing Technology 31, 957–968.

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. Annals of Human Genetics 11, 47–52.

Cochran, W. G. (1947). Some consequences when the assumptions for the analysis of variance are not satisfied. Biometrics 3, 22–38.

Cohen, J., P. Cohen, S. G. West, and L. S. Aiken (2014). Applied multiple regression/correlation analysis for the behavioral sciences. Psychology Press.

Conover, W. J. and R. L. Iman (1981). Rank transformations as a bridge between parametric and nonparametric statistics. The American Statistician 35, 124–129.

Cook, R. D. (1977). Detection of influential observation in linear regression. Technometrics 19, 15–18.

Cook, R. D. and P. Prescott (1981). On the accuracy of Bonferroni significance levels for detecting outliers in linear models. Technometrics 23, 59–63.

Cook, R. D. and P. Wang (1983). Transformations and influential cases in regression. Technometrics 25, 337–343.

Cook, R. D. and S. Weisberg (1982). Residuals and influence in regression. Chapman & Hall.

Cornish, E. A. and R. A. Fisher (1938). Moments and cumulants in the specification of distributions. Revue de l'Institut international de Statistique / Review of the International Statistical Institute 5, 307–320.

Council of the European Union (2001). Report on indicators in the field of poverty and social exclusions. Report, European Union.

Cowell, F. (2011). Measuring inequality. Oxford University Press.

Cox, D. R. and E. J. Snell (1968). A general definition of residuals. Journal of the Royal Statistical Society, Series B 30, 248–275.

Cramér, H. (1928). On the composition of elementary errors. Scandinavian Actuarial Journal 1928, 13–74.

Croissant, Y. (2016). **Ecdat**: Data sets for econometrics. R package version 0.3-1.

Curtiss, J. H. (1943). On transformations used in the analysis of variance. The Annals of Mathematical Statistics 14, 107–122.

da Costa, A. F. and A. F. Crepaldi (2014). The bias in reversing the Box-Cox transformation in time series forecasting: An empirical study based on neural networks. Neurocomputing 136, 281–288.

Dag, O., O. Asar, and O. Ilk (2017). **AID**: Box-Cox power transformation. R package version 1.1.0.

Datta, G. S., P. Hall, and A. Mandal (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. Journal of the American Statistical Association 106, 362–374.

Datta, G. S. and P. Lahiri (1995). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. Journal of Multivariate Analysis 54, 310–328.

Diallo, M. S. and J. N. K. Rao (2014). Small area estimation of complex parameters under unit-level models with skew-normal errors. JSM 2014, Survey Research Methods Section.

Doksum, K. A. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79, 316–319.

Doksum, K. A. and C.-W. Wong (1983). Statistical tests based on transformed data. Journal of the American Statistical Association 78, 411–417.

Dougherty, C. (2011). Introduction to econometrics. OUP Oxford.

Dragulescu, A. A. (2014). **xlsx**: Read, write, format Excel™2007 and Excel™97/2000/XP/2003 files. R package version 0.5.7.

Draper, N. R. and D. R. Cox (1969). On distributions and their transformation to normality. Journal of the Royal Statistical Society, Series B 31, 472–476.

Draper, N. R. and W. G. Hunter (1969). Transformations: Some examples revisited. Technometrics 11, 23–40.

Draper, N. R. and J. John (1981). Influential observations and outliers in regression. Technometrics 23, 21–26.

Duan, N. (1983). Smearing estimate: a nonparametric retransformation method. Journal of the American Statistical Association 78, 605–610.

Duan, N. (1993). Sensitivity analysis for Box-Cox power transformation model: Contrast parameters. Biometrika 80, 885–897.

Durbin, B. P., J. S. Hardin, D. M. Hawkins, and D. M. Rocke (2002). A variance-stabilizing transformation for gene-expression microarray data. Bioinformatics 18, 105–110.

Edgeworth, F. Y. (1900). On the representation of statistics by mathematical formulae. Journal of the Royal Statistical Society 63, 72–81.

Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1, 59–82.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. Biometrics 3, 1–21.

El-Horbaty, Y. E.-S. (2015). Model checking techniques for small area estimation. Ph. D. thesis, University of Southampton, School of Social Sciences.

Elbers, C., J. Lanjouw, and P. Lanjouw (2003). Micro-level estimation of poverty and inequality. Econometrica 71, 355–364.

Elbers, C. and R. van der Weide (2014). Estimation of normal mixtures in a nested error model with an application to small area estimation of poverty and inequality. Working paper.

Elston, R. C. (1961). On additivity in the analysis of variance. Biometrics 17, 209–219.

Emerson, J. D. and M. A. Stoto (1982). Exploratory methods for choosing power transformations. Journal of the American Statistical Association 77, 103–108.

Emerson, J. D. and M. A. Stoto (1983). Understanding robust and exploratory data analysis, Chapter Transforming data, pp. 97–128. John Wiley & Sons.

ENIGH (2010). Encuesta Nacional de Ingresos y Gastos de los Hogares 2010. ENIGH. diseño muestral. http://www.beta.inegi.org.mx/app/biblioteca/ficha.html?upc=702825002420. Accessed: 20.12.2017.

Erickson, B. H. and T. A. Nosanchuk (1977). Understanding data. McGraw-Hill Ryerson.

ESSnet SAE (2012). Small area estimation. http://ec.europa.eu/eurostat/cros/content/sae-finished_en. Accessed: 19.04.2016.

EURAREA (2001). Enhancing small area estimation techniques to meet european needs. http://www.ons.gov.uk/ons/guide-method/method-quality/general-methodology/spatial-analysis-and-modelling/eurarea/index.html. Accessed: 11.04.2016.

EURAREA Consortium (2004). Enhancing small area estimation techniques to meet European needs. Project Reference Volume, Deliverable 7.1.4.

Eurostat (2004). Common cross-sectional EU indicators based on EU-SILC; the gender pay gap. Unit D-2: Living conditions and social protection, Directorate D: Single Market, Employment and Social statistics, Eurostat, Luxembourg (EU-SILC 131-rev/04,).

Fabrizi, E., N. Salvati, M. Pratesi, and N. Tzavidis (2014). Outlier robust model-assisted small area estimation. Biometrical Journal 56, 157–175.

Fabrizi, E. and C. Trivisano (2016). Small area estimation of the Gini concentration coefficient. Computational Statistics & Data Analysis 99, 223–234.

Faddy, M. J. (1997). Extended Poisson process modelling and analysis of count data. Biometrical Journal 39, 431–440.

Famoye, F. (1993). Restricted generalized Poisson regression model. Communications in Statistics - Theory and Methods 22, 1335–1354.

Feng, Q., J. Hannig, and J. S. Marron (2016). A note on automatic data transformation.

Feng, X., X. He, and J. Hu (2011). Wild bootstrap for quantile regression. Biometrika 98, 995–999.

Fernandez, E. S. (2014). **Johnson**: Johnson transformation. R package version 1.1.0.

Fife, D. (2017). **fifer**: A biostatisticians toolbox for various activities, including plotting, data cleanup, and data analysis. R package version 1.1.0.

Fink, E. L. (2009). The FAQs on data transformation. Communication Monographs 76, 379–397.

Finney, D. (1941). On the distribution of a variate whose logarithm is normally distributed. Supplement to the Journal of the Royal Statistical Society 7, 155–161.

Fisher, R. and F. Yates (1949). Statistical tables for biological, agricultural and medical research. Hafner.

Fisher, R. A. (1922a). On the interpretation of χ^2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 85, 87–94.

Fisher, R. A. (1922b). On the mathematical foundations of theroetical statistics. Philosophical Transactions of the Royal Society of London, Series A 222, 309–368.

Fisher, R. A. and E. Cornish (1960). The percentile points of distributions having known cumulants. Technometrics 2, 209–225.

Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: Wild bootstrap vs. pairs bootstrap. Computational Statistics & Data Analysis 49, 361–376.

Fletcher, D., D. MacKenzie, and E. Villouta (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. Environmental and Ecological Statistics 12, 45–54.

Forbes, C., M. Evans, N. Hastings, and B. Peacock (2011). Statistical distributions. John Wiley & Sons.

Foster, A., L. Tian, and L. Wei (2001). Estimation for the Box-Cox transformation model without assuming parametric error distribution. Journal of the American Statistical Association 96, 1097–1101.

Foster, J., J. Greer, and E. Thorbecke (1984). A class of decomposable poverty measures. Econometrica 52, 761–766.

Fox, J. (1997). Applied regression analysis, linear models, and related methods. SAGE Publications.

Fox, J. and S. Weisberg (2011). An R Companion to applied regression (2 ed.). SA.

Freeman, M. and J. Tukey (1949). The uses and usefulness of binomial probability paper. American Statistical Association 4, 174–212.

Freeman, M. and J. Tukey (1950). Transformations related to the angular and the square root. The Annals of Mathematical Statistics 21, 607–611.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. American Journal of Political Science 26, 797–833.

Garson, G. D. (2012). Testing statistical assumptions. Statistical Associates Publishing.

Gaudard, M. and M. Karson (2007). On estimating the Box-Cox transformation to normality. Communications in Statistics - Simulation and Computation 29, 559–582.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2014). Bayesian data analysis. Taylor & Francis.

Gelman, A. and J. Hill (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Gemmill, G. et al. (1980). Using the Box-Cox form for estimating demand: A comment. The Review of Economics and Statistics 62, 147–48.

George, F. (2007). Johnson's system of distributions and microarray data analysis. University of South Florida.

Ghosh, M. (1992). Constrained Bayes estimation with applications. Journal of the American Statistical Association 87, 533–540.

Ghosh, M., T. Maiti, and A. Roy (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. Biometrika 95, 573–585.

Ghosh, M. and R. C. Steorts (2013). Two-stage benchmarking as applied to small area estimation. Methodology (stat.ME) 22, 670–687.

Gini, C. (1912). Variabilità e mutabilità : Contributo allo studio e delle distribuzioni e relazioni statistiche. Studi Economico-Giuridici della R, Universitá di Cagliari.

Goldberger, A. S. (1968). The interpretation and estimation of Cobb-Douglas functions. Econometrica 36, 464–472.

Gómez-Rubio, V., N. Best, S. Richardson, G. Li, and P. Clarke (2010). Bayesian statistics for small area estimation. Technical Report [accessed: 27.02.2018].

Gómez-Rubio, V., N. Salvati, et al. (2008). **SAE2**: Small Area Estimation with R. R package version 0.09.

Goncalves, S. and N. Meddahi (2011). Box-Cox transforms for realized volatility. Journal of Econometrics 160, 129–144.

González-Manteiga, W., M. Lombardía, I. Molina, D. Morales, and L. Santamaría (2008). Bootstrap mean squared error of a small-area eblup. Journal of Statistical Computation and Simulation 78, 443–462.

Gottardo, R. and A. Raftery (2009). Bayesian robust transformation and variable selection: a unified approach. The Canadian Journal of Statistics / La Revue Canadienne de Statistique 37, 361–380.

Graf, M., J. Marin, and I. Molina (2014). Estimation of poverty indicators in small areas under skewed distributions. Working paper.

Greene, W. H. (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working paper.

Grogger, J. T. and R. T. Carson (1991). Models for truncated counts. Journal of Applied Econometrics 6, 225–238.

Gurka, M., L. Edwards, K. Muller, and L. Kupper (2006). Extending the Box-Cox transformation to the linear mixed model. Journal of the Royal Statistical Society, Series A 169, 273–288.

Gurmu, S. (1998). Generalized hurdle count data regression models. Economics Letters 58, 263–268.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate data. Journal of the Royal Statistical Society, Series B 54, 761–771.

Hagenaars, A., K. de Vos, and M. Zaidi (1994). Poverty statistics in the late 1980s: Research based on mirco-data. Office for Official Publications of the European Communities.

Hájek, J. (1958). On the theory of ratio estimates. Aplikace Matematiky 3, 384–398.

Hall, P. and T. Maiti (2006). On parametric bootstrap methods for small area prediction. Journal of the Royal Statistical Society, Series B 68, 221–238.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel (1986). Robust statistics: The approach based on influence functions. John Wiley & Sons.

Han, A. K. (1987). A non-parametric analysis of transformations. Journal of Econometrics 35, 191–209.

Hartley, H. O. (1950). The use of range in analysis of variance. Biometrika 37, 271–280.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. Biometrika 61, 383–385.

Hawkins, D. (1980). Identification of ouliers. Chapman & Hall.

Hawkins, D. M., D. Bradu, and G. V. Kass (1984). Location of several outliers in multiple-regression data using elemental sets. Technometrics 26, 197–208.

Heagerty, P. and S. Zeger (2000). Marginalized multilevel models and likelihood inference. Statistical Science 15, 1–26.

Heien, D. M. (1968). A note on log-linear regression. Journal of the American Statistical Association 63, 1034–1038.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. Biometrical Journal 36, 531–547.

Hernandez, F. and R. A. Johnson (1980). The large-sample behavior of transformations to normality. Journal of the American Statistical Association 75, 855–861.

Hey, G. (1938). A new method of experimental sampling illustrated on certain non-normal populations. Biometrika 30, 68–80.

Hilde, J. M. (2014). Modeling count data, Chapter Varieties of count data, pp. 1–33. Cambridge University Press.

Hill, B. M. (1963). The three-parameter lognormal distribution and Bayesian analysis of a point-source epidemic. Journal of the American Statistical Association 58, 72–84.

Hinde, J. and C. G. Demétrio (1998). Overdispersion: Models and estimation. Computational Statistics & Data Analysis 27, 151–170.

Hinkley, D. (1975). On power transformations to symmetry. Biometrika 62, 101–111.

Hinkley, D. (1977). On quick choice of power transformation. Journal of the Royal Statistical Society, Series C 26, 67–69.

Hinkley, D. (1985). Transformation diagnostics for linear models. Biometrika 72, 487–496.

Hinkley, D. and G. Runger (1984). The analysis of transformed data. Journal of the American Statistical Association 79, 302–309.

Hoaglin, D., F. Mosteller, and W. Tukey (2000). Understanding robust and exploratory data analysis. John Wiley & Sons.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12, 55–67.

Hoeting, J. A. and J. G. Ibrahim (1998). Bayesian predictive simultaneous variable and transformation selection in the linear model. Computational Statistics & Data Analysis 28, 87–103.

Hoeting, J. A., A. E. Raftery, and D. Madigan (2002). Bayesian variable and transformation selection in linear regression. Journal of Computational and Graphical Statistics 11, 485–507.

Hossain, M. Z. (2011). The use of Box-Cox transformation technique in economic and statistical analyses. Journal of Emerging Trends in Economics and Management Sciences 2, 32–39.

Hoyle, M. H. (1973). Transformations: An introduction and a bibliography. International Statistical Review / Revue Internationale de Statistique 41, 203–223.

Huber, P. (1981). Robust statistics. John Wiley & Sons.

Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistical 35, 73–101.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1, 221–233.

Huber, P. J. (1992). Breakthroughs in statistics, Chapter Robust estimation of a location parameter, pp. 492–518. Springer.

Huber, W., A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron (2003). Parameter estimation for the calibration and variance stabilization of microarray data. Statistical Applications in Genetics and Molecular Biology 2, 1–24.

Hulten, C. R. and F. C. Wykoff (1981). The estimation of economic depreciation using vintage asset prices: An application of the Box-Cox power transformation. Journal of Econometrics 15, 367–396.

Hutcheson, G. and N. Sofroniou (1999). The multivariate social scientist: Introductory statistics using generalized linear models. SAGE Publications.

Ives, A. R. (2015). For testing the significance of regression coefficients, go ahead and log-transform count data. Methods in Ecology and Evolution 6, 828–835.

Jeffreys, H. (1998). The theory of probability. OUP Oxford.

Jensen, D. R. and H. Solomon (1972). A Gaussian approximation to the distribution of a definite quadratic form. Journal of the American Statistical Association 67, 898–902.

Jereys, H. (1961). Theory of probability. Clarendon Press.

Jiang, J., P. Lahiri, and S. Wan (2002). A unified jackknife theory for empirical best prediction with m-estimation. The Annals of Statistics 30, 1782–1810.

Jiang, J. and T. Nguyen (2012). Small area estimation via heteroscedastic nested-error regression. Canadian Journal of Statistics 40, 588–603.

John, J. A. and N. R. Draper (1980). An alternative family of transformations. Journal of the Royal Statistical Society, Series C 29, 190–197.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Johnson, R. A. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79, 314–315.

Johnson, R. A. (2009). Statistics - Principles and methods, 6th Edition. John Wiley & Sons.

Johnston, J. and J. DiNardo (1972). Econometric methods. McGraw-Hill Ryerson.

Jones, M. C. and A. Pewsey (2009). Sinh-arcsinh distributions. Biometrika 96, 761 – 780.

Keene, O. N. (1995). The log transformation is special. Statistics in Medicine 14, 811–819.

Kelmansky, D. M., E. J. Martínez, and V. Leiva (2013). A new variance stabilizing transformation for gene expression data analysis. Statistical Applications in Genetics and Molecular Biology 12, 653–666.

Kelmansky, D. M. and L. Ricci (2017). A new distribution family for microarray data. Microarrays 6, 1–5.

Kendall, M. G. (1938). A new measure of rank correlation. Biometrika 30, 81–93.

Kendall, M. G., A. Stuart, and J. K. Ord (1948). The advanced theory of statistics. Charles Griffin.

Kettl, S. (1991). Accounting for heteroscedasticity in the transform both sides regression model. Journal of the Royal Statistical Society, Series C 40, 261–268.

Kim, C., B. E. Storer, and M. Jeong (1996). Note on Box-Cox transformation diagnostics. Technometrics 38, 178–180.

Kirk, R. E. (1968). Experimental design: Procedures for the behavioral sciences. Brooks/Cole.

Kleczkowski, A. (1949). The transformation of local lesion counts for statistical analysis. Annals of Applied Biology 36, 139–152.

Koenker, R. (2005). Quantile regression. Cambridge University Press.

Krasker, W. S. and R. E. Welsch (1982). Efficient bounded-influence regression estimation. Journal of the American Statistical Association 77, 595–604.

Kreutzmann, A.-K. (2016). Poverty mapping using small area estimation: An application with R. Master's thesis, Freie Universität Berlin.

Kreutzmann, A.-K., S. Pannier, N. Rojas-Perilla, T. Schmid, M. Templ, and N. Tzavidis (2018). **emdi**: estimating and mapping disaggregated Indicators. R package version 1.1.2.

Kruskal, J. B. (1968). International encyclopaedia of the social sciences, Chapter Statistical analysis: Transformations of data, pp. 182–193. MacMillan.

Kruskal, W. H. and W. A. Wallis (1952). Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association 47, 583–621.

Kuhn, M. (2008). Building predictive models in R using the **caret** package. Journal of Statistical Software 28, 1–26.

Kullback, S. (1997). Information theory and statistics. Dover Publications.

Lagarias, J. C., J. A. Reeds, M. H. Wright, and P. E. Wright (1998). Convergence properties of the nelder-mead simplex method in low dimensions. SIAM Journal on optimization 9, 112–147.

Lakhana, W. (2014). A new family of transformations for lifetime data. Working Paper.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics 34, 1–14.

Laubscher, N. F. (1961). On stabilizing the binomial and negative binomial variances. Journal of the American Statistical Association 56, 143–150.

Laud, P. W. and J. G. Ibrahim (1995). Predictive model selection. Journal of the Royal Statistical Society, Series B 57, 247–262.

Laurent, A. G. (1963). The lognormal distribution and the translation method: description and estimation problems. Journal of the American Statistical Association 58, 231–235.

Lawrance, A. (1987a). A note on the variance of the Box-Cox regression transformation estimate. Journal of the Royal Statistical Society, Series C 36, 221–223.

Lawrance, A. (1987b). The score statistic for regression transformation. Biometrika 74, 275–379.

L'Ecuyer, P. (1999). Good parameters and implementations for combined multiple recursive random number generators. Operations Research 47, 159–164.

L'Ecuyer, P., R. Simard, E. J. Chen, and W. D. Kelton (2002). An object-oriented random-number package with many long streams and substreams. Operations Research 50, 1073–1075.

Lee, C. (1982). Comparison of two correction methods for the bias due to the logarithmic transformation in the estimation of biomass. Canadian Journal of Forest Research 12, 326–331.

Lee, Y., J. A. Nelder, et al. (2004). Conditional and marginal models: Another view. Statistical Science 19, 219–238.

Leinhardt, S. and S. S. Wasserman (1979). Exploratory data analysis: An introduction to selected methods. Sociological Methodology 10, 311–365.

Lesaffre, E. and G. Molenberghs (1991). Multivariate probit analysis: A neglected procedure in medical statistics. Statistics in Medicine 10, 1391–1403.

Levene, H. et al. (1960). Robust tests for equality of variances. Contributions to Probability and Statistics 1, 278–292.

Leydold, J. (2017). **rstream**: Streams of random numbers. R package version 1.3.5.

Lipsitz, S. R., J. Ibrahim, and G. Molenberghs (2000). Using a Box-Cox transformation in the analysis of longitudinal data with incomplete responses. Journal of the Royal Statistical Society, Series C 49, 287–296.

Litière, S., A. Alonso, and G. Molenberghs (2008). The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models. Statistics in Medicine 16, 3125–3144.

Lo, S. and S. Andrews (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. Frontiers in Psychology 6, 1171–1177.

Lohr, S. and J. Rao (2009). Jackknife estimation of mean squared error of small area predictors in nonlinear mixed models. Biometrika 96, 457–468.

Lopez-Vizcaino, E., M. Lombardia, and D. Morales (2014). **mme**: Multinomial mixed effects models. R package version 0.1-5.

Lumley, T. (2012). **survey**: Analysis of complex survey samples. R package version 3.28-2.

Machado, J. A. F. and J. Mata (2000). Box-Cox quantile regression and the distribution of firm sizes. Journal of Applied Econometrics 15, 253–274.

MacKinnon, J. G. and L. Magee (1990). Transforming the dependent variable in regression models. International Economic Review 31, 315–339.

Magee, L. (1988). The review of economics and statistics. The Review of Economics and Statistics 70, 362–366.

Makitalo, M. and A. Foi (2011). A closed-form approximation of the exact unbiased inverse of the anscombe variance-stabilizing transformation. IEEE Transactions on Image Processing 20, 2697–2698.

Manly, B. F. J. (1976). Exponential data transformations. Journal of the Royal Statistical Society, Series D 25, 37–42.

Manning, W. G. (1998). The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17, 283–295.

Marazzi, A. and V. J. Yohai (2004). Theory and applications of recent robust methods, Chapter Robust Box-Cox transformations for simple regression, pp. 173–182. Birkhäuser Basel.

Marazzi, A. and V. J. Yohai (2006). Robust Box-Cox transformations based on minimum residual autocorrelation. Computational Statistics & Data Analysis 50, 2752–2768.

Marhuenda, Y., I. Molina, D. Morales, and J. N. K. Rao (2017). Poverty mapping in small areas under a twofold nested error regression model. Journal of the Royal Statistical Society, Series A 180, 1111–1136.

Marino, M. F., N. Tzavidis, and M. Alfo (2016). Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. Statistical Methods in Medical Research, forthcoming.

Maruo, K., Y. Yamaguchi, H. Noma, and M. Gosho (2017). Interpretable inference on the mixed effect model with the Box-Cox transformation. Statistics in Medicine 36, 2420–2434.

McCullagh, P. and J. Nelder (1989). Generalized linear models (2 ed.). Taylor & Francis.

McCulloch, C. E. and J. M. Neuhaus (2001). Generalized linear mixed models. John Wiley & Sons.

McNeil, D. R. (1977). Interactive data analysis: A practical primer. John Wiley & Sons.

Medina, L. (2017). Transformations in the linear regression model: An overview. Master's thesis, Freie Universität Berlin.

Medina, L., N. Rojas-Perilla, A. Kreutzmann, and P. Castro (2017). The R package **trafo** for transforming linear regression models.

Meindl, B., M. Templ, A. Alfons, A. Kowarik, and with contributions from Mathieu Ribatet (2016). **simPop**: simulation of synthetic populations for survey data considering auxiliary information. R package version 0.3.0.

Meyer, H. (1941). A correction for a systematic error occurring in the application of the logarithmic volume equation. Pensylvania Forest School Research. 7, 905–912.

Miller, D. M. (1984). Reducing transformation bias in curve fitting. The American Statistician 38, 124–126.

Miller, J. P. (2010). Essential statistical methods for medical statistics. Elsevier.

Mills, T. C. (1978). The functional form of the U.K. demand for money. Journal of the Royal Statistical Society, Series C 27, 52–57.

Molina, I. and Y. Marhuenda (2015). **sae**: An R package for small area estimation. The R Journal 7, 81–98.

Molina, I., D. Morales, M. Pratesi, and N. Tzavidis (2010). Final small area estimation developments and simulations results. Research Project Report Deliverable D12 and D16, EU-FP7-SSH-2007-1 SAMPLE.

Molina, I. and J. N. K. Rao (2010). Small area estimation of poverty indicators. The Canadian Journal of Statistics 38, 369–385.

Montgomery, D. (2008). Design and analysis of experiments. John Wiley & Sons.

Moore, P. (1957). Transformations to normality using fractional powers of the variable. Journal of the American Statistical Association 52, 237–246.

Moore, P. (1958). Interval analysis and the logarithmic transformation. Journal of the Royal Statistical Society, Series B 20, 187–192.

Moore, P. G. and J. W. Tukey (1954). Answer to query 112. Biometrics 10, 562–568.

Morozova, M., K. Koschutnig, E. Klein, and G. Wood (2016). Monotonic non-linear transformations as a tool to investigate age-related effects on brain white matter integrity: A Box-Cox investigation. Neuroimage 125, 1119–1130.

Moschopoulos, P. G. (1983). On a new transformation to normality. Communications in Statistics - Theory and Methods 12, 1873–1878.

Mosteller, F. and R. R. Bush (1954). Handbook of Social Psychology, Chapter Selected quantitative techniques. Addison-Wesley.

Mosteller, F. and J. W. Tukey (1977). Data analysis and regression: A second course in statistics. Addison Wesley.

Mosteller, F. and C. Youtz (2006). Selected papers of Frederick Mosteller, Chapter Tables of the Freeman-Tukey transformations for the binomial and Poisson distributions, pp. 337–347. Springer.

Mukhopadhyay, P. K. and A. McDowell (2011). Small area estimation for survey data analysis using SAS software. Technical report, SAS Institute Inc.

Mullahy, J. (1986). Specification and testing of some modified count data models. Journal of Econometrics 33, 341–365.

Müller, S., J. L. Scealy, A. H. Welsh, et al. (2013). Model selection in linear mixed models. Statistical Science 28, 135–167.

Nakagawa, S. and H. Schielzeth (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. Methods in Ecology and Evolution 4, 133–142.

Natrella, M. G. (2013). Experimental statistics. Courier Corporation.

Nau, R. F. (2017). Regression diagnostics: Testing the assumptions of linear regression. Working paper.

Nelder, J. A. (1977). A reformulation of linear models. Journal of the Royal Statistical Society, Series A 140, 48–77.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. The Computer Journal 7, 308–313.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. Journal of the Royal Statistical Society, Series A 135, 370–384.

Newman, M. C. (1993). Regression analysis of log-transformed data: Statistical bias and its correction. Environmental Toxicology and Chemistry 12, 1129–1133.

Neyman, J. and E. L. Scott (1960). Correction for bias introduced by a transformation of variables. The Annals of Mathematical Statistics 31, 643–655.

Nychka, D. and D. Ruppert (1995). Nonparametric transformations for both sides of a regression model. Journal of the Royal Statistical Society, Series B 57, 519–532.

O'Hara, R. B. and D. J. Kotze (2010). Do not log-transform count data. Methods in Ecology and Evolution 1, 118–122.

Oja, H. (1981). On location, scale, skewness and kurtosis of univariate distributions. Scandinavian Journal of Statistics 8, 154–168.

Opsomer, J., G. Claeskens, M. Ranalli, G. Kauermann, and F. Breidt (2008). Nonparametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society, Series B 70, 265–283.

Osborne, J. W. (2002). The effects of minimum values on data transformations. Annual Meeting of the American Educational Research Association.

Osborne, J. W. and A. Overbay (2004). The power of outliers (and why researchers should always check for them). Practical Assessment, Research, & Evaluation 9, 1–8.

Osborne, J. W. and E. Waters (2012). Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research, & Evaluation 8, 1–5.

Patterson, R. (1966). Difficulties involved in the estimation of a population mean using transformed sample data. Technometrics 8, 535–537.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. Biometrika, 114–133.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London, Series A 185, 71–110.

Pebesma, E. (2018). **sf**: Simple features for R. R package version 0.6-0.

Peng Zhang, Peter X.-K. Song, A. Q. and T. Greene (2008). Efficient estimation for patient-specific rates of disease progression using nonnormal linear mixed models. Biometrics 64, 29–38.

Pericchi, L. R. (1981). A Bayesian approach to transformations to normality. Biometrika 68, 35–43.

Pfeffermann, D. (2013). New important developments in small area estimation. Statistical Science 28, 40–68.

Pfeffermann, D. and S. Correa (2012). Empirical bootstrap bias correction and estimation of prediction mean square error in small area estimation. Biometrika 99, 457–472.

Pfeffermann, D. and A. Sikov (2011). Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. Journal of Official Statistics 27, 181–209.

Pfeffermann, D., A. Sikov, and R. Tiller (2014). Single- and two-stage cross-sectional and time series benchmarking procedures for small area estimation. TEST 23, 631–666.

Piepho, H.-P. and C. E. McCulloch (2004). Transformations in mixed models: Application to risk analysis for a multienvironment trial. Journal of Agricultural, Biological, and Environmental Statistics 9, 123–137.

Pierce, D. A. and D. W. Schafer (1986). Residuals in generalized linear models. Journal of the American Statistical Association 81, 977–986.

Pinheiro, J. and D. Bates (2000). Mixed-effects models in S and S-Plus. Springer.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar, and R Core Team (2017). **nlme**: linear and nonlinear mixed effects models. R package version 3.1-131.1.

Poirier, D. J. (1978). The use of the Box-Cox transformation in limited dependent variable models. Journal of the American Statistical Association 73, 284–287.

Prasad, N. G. N. and J. N. K. Rao (1990). The estimation of the mean squared error of small area estimators. Journal of the American Statistical Association 85, 163–171.

Pratesi, M. and N. Salvati (2009). Small area estimation in the presence of correlated random area effects. Journal of Official Statistics 25, 37–53.

Rahman, M. (1999). Estimating the box-cox transformation via Shapiro-Wilk W statistic. Communications in Statistics-Simulation and Computation 28, 223–241.

Rahman, M. and L. Pearson (2008). Anderson-Darling statistic in estimating the Box-Cox. Journal of Applied Probability & Statistics 3, 45–57.

Rainey, C. (2017). Transformation-induced bias: Unbiased coefficients do not imply unbiased quantities of interest. Political Analysis 25, 402–409.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. Journal of the Royal Statistical Society, Series B 31, 350–371.

Ramsey, J. B. (1974). Classical model selection through specification error tests. Frontiers in Econometrics 1, 13–47.

Rao, J. N. K. and I. Molina (2015). Small area estimation (2 ed.). John Wiley & Sons.

Raudenbush, S. and A. Bryk (2002). Hierarchical linear models: Applications and data analysis methods. SAGE Publications.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Ribeiro Jr., P. and P. Diggle (2016). **geoR**: Analysis of geostatistical data. R News 1, 15–18.

Rivest, L.-P. and E. Belmonte (2000). A conditional mean squared error of small area estimators. Survey Methodology 26, 67–78.

Rocke, D. M. (1993). On the beta transformation family. Technometrics 35, 72–81.

Rojas-Perilla, N., A.-K. Kreutzmann, and L. Medina (2017). A guideline of transformations in linear and linear mixed regression models. Working Paper.

Rojas-Perilla, N., S. Pannier, T. Schmid, and N. Tzavidis (2017). Data-driven transformations in small area estimation. Working paper. Discussion Paper 30/2017, School of Business and Economics, Freie Universiät Berlin.

Rosenthal, J. A. (2011). Statistics and data interpretation for social work, Chapter Shape of distribution, pp. 51–61. Springer.

Rothery, P. (1988). A cautionary note on data transformation: Bias in back-transformed means. Bird Study 35, 219–221.

Rousseeuw, P. J. and A. M. Leroy (2005). Robust regression and outlier detection. John Wiley & Sons.

Rousseeuw, P. J. and B. C. Van Zomeren (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association 85, 633–639.

Royston, P., P. C. Lambert, et al. (2011). Flexible parametric survival analysis using Stata: Beyond the Cox model. Stata Press.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63, 581–592.

Rubin, D. B. (1984). The analysis of transformed data: Comment. Journal of the American Statistical Association 79, 309–312.

Rukhin, A. L. (1986). Improved estimation in lognormal models. Journal of the American Statistical Association 81, 1046–1049.

Ruppert, D. (2001). Transformations of data. The International Encyclopedia of the Social & Behavioral Sciences.

Ruppert, D. and B. Aldershof (1989). Transformations to symmetry and homoscedasticity. Journal of the American Statistical Association 84, 437–446.

Saei, A. and C. McGilchrist (1997). Random threshold models applied to inflated zero class data. Australian & New Zealand Journal of Statistics 39, 5–16.

Saei, A., J. Ward, and C. McGilchrist (1996). Threshold models in a methadone programme evaluation. Statistics in Medicine 15, 2253–2260.

Sakia, R. (1990). Retransformation bias: A look at the Box-Cox transformation to linear balanced mixed ANOVA models. Metrika 37, 345–351.

Sakia, R. M. (1988). Application of the Box-Cox transformation technique to linear balanced mixed analysis of variance models with a multi-error structure. Ph. D. thesis, University of Hohenheim, FRG.

Sakia, R. M. (1992). The Box-Cox transformation technique: A review. Journal of the Royal Statistical Society, Series D 41, 169–178.

SAMPLE (2007). Small area methods for poverty and living condition estimates. `http://www.sample-project.eu/`. Accessed: 11.04.2016.

Särndal, C.-E., B. Swensson, and J. Wretman (1992). Model assisted survey sampling. Springer.

Schmid, T., F. Bruckschen, N. Salvati, and T. Zbiranski (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: estimating literacy rates in Senegal. Journal of the Royal Statistical Society, Series A 180, 1163–1190.

Schmid, T., N. Tzavidis, R. Münnich, and R. Chambers (2016). Outlier robust small area estimation under spatial correlation. Scandinavian Journal of Statistics 43, 806–826.

Schoch, T. (2012). Robust unit-level small area estimation: A fast algorithm for large datasets. Austrian Journal of Statistics 41, 243–265.

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality. Biometrika 52, 591–611.

Shen, W. and T. Louis (1998). Triple-goal estimates in two-stage hierarchical models. Journal of the Royal Statistical Society, Series B 60, 455–471.

Shi, C. and with contributions from Peng Zhang (2013). **BayesSAE**: Bayesian analysis of Small Area Estimation. R package version 1.0-1.

Shin, Y. (2008). Semiparametric estimation of the Box-Cox transformation model. The Econometrics Journal 11, 517–537.

Sinha, S. K. and J. N. K. Rao (2009). Robust small area estimation. The Canadian Journal of Statistics 37, 381–399.

Slifker, J. F. and S. S. Shapiro (1980). The Johnson system: Selection and parameter estimation. Technometrics 22, 239–246.

Smallwood, D. M., J. R. Blaylock, et al. (1986). Forecasting performance of models using the Box-Cox transformation. Agricultural Economics Research 4, 14–24.

Smirnov, N. (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics 19, 279–281.

Smith, R. J. (1993). Logarithmic transformation bias in allometry. American Journal of Physical Anthropology 90, 215–228.

Snedecor, G. W. and W. G. Cochran (1989). Statistical methods. Iowa State Press.

Snijders, T. and R. Bosker (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling. SAGE Publications.

Sokal, R.R.; Rohlf, F. (1995). Biometry: The principles and practice of statistics in biological research. W.H. Freeman and Company.

Solomon, P. J. (1985). Transformations for components of variance and covariance. Biometrika 72, 233–239.

Spanos, A. (1986). Statistical foundations of econometric modelling. Cambridge University Press.

Sprugel, D. (1983). Correcting for bias in log-transformed allometric equations. Ecology 64, 209–210.

Statistik Austria (2013). Registerbasierte Statistiken Demographie (RS). Schnellbericht 10.7.

Sverchkov, M. and D. Pfeffermann (2004). Prediction of finite population total based on the sample distribution. Survey Methodology 30, 79–92.

Sweeting, T. J. (1984). On the choice of prior distribution for the Box-Cox transformed linear model. Biometrika 71, 127–134.

Tabachnick, B. G. and L. S. Fidell (2007). Using multivariate statistics. Pearson.

Taylor, J. M. (1986). The retransformed mean after a fitted power transformation. Journal of the American Statistical Association 81, 114–118.

Taylor, J. M. G. (1985). Power transformations to symmetry. Biometrika 72, 145–152.

Templ, M. (2015). Cran task view: Official statistics and survey methodology. https://cran.r-project.org/web/views/OfficialStatistics.html. Accessed: 11.04.2016.

Thai, H.-T., F. Mentré, N. H. Holford, C. Veyrat-Follet, and E. Comets (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. Pharmaceutical Statistics 12, 129–140.

The World Bank (2005). Introduction to Poverty Analysis.

The World Bank (2007). More than a pretty picture: Using poverty maps to design better policies and interventions. The international Bank for Reconstruction and Development - The World Bank.

The World Bank (2013). Software for poverty mapping. http://go.worldbank.org/QG9L6V7P20. Accessed: 11.04.2016.

The World Bank (2017). Measuring poverty. http://go.worldbank.org/QG9L6V7P20. [accessed: 27.04.2017].

Thoni, H. (1969). Transformation of variables used in the analysis of experimental and observational data: A review. Statistical Laboratory, Iowa State University.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58, 267–288.

Tierney, L., A. J. Rossini, N. Li, and H. Sevcikova (2016). **snow**: Simple network of workstations. R package version 0.4-2.

Tillé, Y. and A. Matei (2012). **sampling**: Survey sampling. R package version 2.5.

Tortajada, C. (2006). Who has access to water case study of Mexico City metropolitan area Human Development Report 2006.

Tsai, A. C., M. Liou, M. Simak, and P. E. Cheng (2017). On hyperbolic transformations to normality. Computational Statistics & Data Analysis 115, 250–266.

Tsai, C.-L. and X. Wu (1990). Diagnostics in transformation and weighted regression. Technometrics 32, 315–322.

Tukey, J. W. (1949). One degree of freedom for non-additivity. Biometrics 5, 232–242.

Tukey, J. W. (1957). On the comparative anatomy of transformations. The Annals of Mathematical Statistics 28, 602–632.

Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.

Tzavidis, N., S. Marchetti, and R. Chambers (2010). Robust estimation of small area means and quantiles. Australian and New Zealand Journal of Statistics 52, 167–186.

Uddin, M., M. Noor, A. Kabir, R. Ali, and M. Islam (2006). The transformations of random variables under symmetry. Journal of Applied Sciences 6, 1818–1821.

Ugarte, M., T. Goicoa, A. Militino, and M. Durban (2009). Spline smoothing in small area trend estimation and forecasting. Computational Statistics and Data Analysis 53, 3616–3629.

Urbanek, S. (2009 - 2014). **multicore**: Parallel processing of R code on machines with multiple cores or CPUs.

Ushey, K., J. McPherson, J. Cheng, A. Atkins, and J. Allaire (2018). **packrat**: A dependency management system for projects and their R package dependencies. R package version 0.4.9-1.

Vaida, F. and S. Blanchard (2005). Conditional Akaike information for mixed-effects models. Biometrika 92, 351–370.

Vélez, J. I., J. C. Correa, and F. Marmolejo-Ramos (2015). A new approach to the Box-Cox transformation. Frontiers in Applied Mathematics and Statistics 1, 1–12.

Velilla, S. (1993). Quantile-based estimation for the Box-Cox transformation in random samples. Statistics & probability Letters 16, 137–145.

Venables, W. N. and B. D. Ripley (2002). Modern applied statistics with S. Springer.

Verbeke, G. and E. Lesaffre (1996). A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association 91, 217–221.

Verbeke, G. and G. Molenberghs (2000). Linear mixed models for longitudinal data. Springer.

Walker, A. (2017). **openxlsx**: Read, write and edit XLSX files. R package version 4.0.17.

Wang, N. and D. Ruppert (1995). Nonparametric estimation of the transformation in the transform-both-sides regression model. Journal of the American Statistical Association 90, 522–534.

Wang, S. (1987). Improved approximation for transformation diagnostics. Communications in Statistics - Theory and Methods 16, 1797–1819.

Wang, Y. (2015). **jtrans**: Johnson transformation for normality. R package version 1.1.0.

Warnholz, S. (2016a). **saeRobust**: Robust small area estimation. R package version 0.1.0.

Warnholz, S. (2016b). Small area estimation using robust extensions to area level models. Ph. D. thesis, Freie Universität Berlin.

Warnholz, S. and T. Schmid (2016). Simulation tools for small area estimation: Introducing the R package saeSim. Austrian Journal of Statistics 45, 55–69.

Warton, D. and F. Hui (2011). The arcsine is asinine: The analysis of proportions in ecology. Ecology 92, 3–10.

Warton, D. I., M. Lyons, J. Stoklosa, and A. R. Ives (2016). Three points to consider when choosing a lm or glm test for count data. Methods in Ecology and Evolution 7, 882–890.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika 61, 439–447.

Weidenhammer, B., N. Tzavidis, T. Schmid, and N. Salvati (2014). Domain prediction for counts using microsimulation via quantiles. In Small Area Estimation 2014 Conference, Poznan, Poland.

Weisberg, S. (1980). Applied linear regression. John Wiley & Sons.

West, B. T., K. B. Welch, and A. T. Galecki (2007). Linear mixed models: A practical guide using statistical software. CRC Press.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48, 817–838.

Whittaker, J., C. Whitehead, and M. Somers (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. Journal of the Royal Statistical Society, Series C 54, 863–878.

Wickham, H. (2009). **ggplot2**: Elegant graphics for data analysis. Springer.

Wilcox, R. R. (2005). Introduction to robust estimation and hypothesis testing. Elsevier.

Williams, M., C. A. G. Grajales, and D. Kurkiewicz (2013). Assumptions of multiple regression: Correcting two misconceptions. Practical Assessment, Research & Evaluation 18, 1–14.

Wilson, E., M. Underwood, O. Puckrin, K. Letto, R. Doyle, H. Caravan, and S. C. . K. Bassett (2013). The arcsine transformation: Has the time come for retirement?

Wilson, E. B. and M. M. Hilferty (1931). The distribution of chi-square. Proceedings of the National Academy of Sciences 17, 684–688.

Wirtschaftskammer Österreich (2017). Wirtschaftsdaten auf bezirksebene. [accessed: 07.02.2018].

Withers, C. S. and S. Nadarajah (2007). Linear regression with extreme value residuals. Communications in Statistics - Simulation and Computation 37, 73–91.

Wood, F. S. and J. W. Gorman (1971). Fitting equations to data: Computer analysis of multifactor data for scientists and engineers. Wiley-Interscienc.

Wooldridge, J. (2000). Introductory econometrics: A modern approach. South-Western Cengage Learning.

Yale, C. and A. B. Forsythe (1976). Winsorized regression. Technometrics 18, 291–300.

Yang, L. (1995). Transformation-density estimation. Ph. D. thesis, University of North Carolina, Chapel Hill.

Yang, Z. (2006). A modified family of power transformations. Economics Letters 92, 14–19.

Yang, Z. and T. Abeysinghe (2003). A score test for Box-Cox functional form. Economics Letters 79, 107–115.

Yang, Z. and A. K. Tsui (2004). Analytically calibrated Box-Cox percentile limits for duration and event-time models. Insurance: Mathematics and Economics 35, 649–677.

Yeo, I.-K. and R. A. Johnson (2000). A new family of power transformations to improve normality or symmetry. Biometrika 87, 954–959.

Zar, J. (1999). Biostatistical analysis. Prentice Hall.

Zarembka, P. (1974a). Econometrics, Chapter Transformation of variables in econometrics. Springer.

Zarembka, P. (1974b). Frontiers in econometrics. Academic Press.

Zeckhauser, R. and M. Thompson (1970). Linear regression with non-normal error terms. The Review of Economics and Statistics 52, 280–286.

Zeileis, A. (2014). **ineq**: Measuring inequality, concentration, and poverty. R package version 0.2-13.

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in R. Journal of Statistical Software 27, 1–25.

Zellner, A. (1971). Bayesian and non-Bayesian analysis of the log-normal distribution and log-normal regression. Journal of the American Statistical Association 66, 327–330.

Zhang, D. and M. Davidian (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. Biometrics 57, 795–802.

Zhang, L.-C. (2007). Finite population small area interval estimation. Journal of Official Statistics 23, 223–237.

Zhang, L.-C. (2009). Estimates for small area compositions subjected to informative missing data. Survey Methodology 35, 191–201.

Zwet, W. R. (1964). Convex transformations of random variables. Mathematisch Centrum Amsterdam.

Summary

Summary in English

Abstract: A Guideline of Transformations in Linear and Linear Mixed Regression Models

Representing a relationship between a response variable and a set of covariates is an essential part of the statistical analysis. The linear regression model offers a parsimonious solution to this issue, and hence it is extensively used in nearly all science disciplines. In recent years the linear mixed regression model has become common place in the statistical analysis. Numerous assumptions are usually made whenever these models are employed in scientific research. If one or several of these assumptions are not met, the application of transformations can be useful. This work provides an extensive overview of different transformations and estimation methods of transformation parameters in the context of linear and linear mixed regression models. The main contribution is the development of a guideline that leads the practitioner working with data that does not meet model assumptions by using transformations.

Keywords: Transformations, model assumptions, linear regression models, linear mixed regression models, transformation parameters

Abstract: The R Package trafo for Transforming Linear Regression Models

The linear regression model has been widely used for descriptive, predictive, and inferential purposes. This model relies on highly restrictive set of assumptions, which are not always fulfilled when working with empirical data. In this case, one solution could be the use of more complex regression methods that do not strictly rely in the same assumptions. However, in order to improve the validity of model assumptions, transformations are a simpler approach and enable the user to keep using the well-known linear regression model. But how can a user find a suitable transformation? The R package **trafo** offers a simple user-friendly framework for selecting a suitable transformation depending on the user needs. The collection of selected transformations and estimation methods in the package **trafo** complement and enlarge the methods that are existing in R so far.

Keywords: Transformations, optimal parameter, power transformations, normality, linear regression model

Abstract: From start to finish: A framework for the production of small area official statistics

Small area estimation is a research area in official and survey statistics of great practical relevance for national statistical institutes and related organisations. Despite rapid developments in methodology and software, researchers and users would benefit from having practical guidelines for the process of small area estimation. In this paper we propose a general framework for the production of small area statistics that is governed by the principle of parsimony and is based on three broadly defined stages namely, specification, analysis/adaptation and evaluation. Emphasis is given to the interaction between a user of small area statistics and the statistician in specifying the target geography and parameters in light of the available data. Model-free and model-dependent methods are described with focus on model selection and testing, model diagnostics and adaptations such as use of data transformations. Uncertainty measures and the use of model and design-based simulations for method evaluation are also at the centre of the paper. We illustrate the application of the proposed framework using real data for the estimation of non-linear deprivation indicators. Linear statistics, for example averages, are included as special cases of the general framework.
Keywords: Census, design-based methods, diagnostics, inequality, model-based methods

Abstract: Data-driven Transformations in Small Area Estimation

Small area models typically depend on the validity of model assumptions. For example, a commonly used version of the Empirical Best Predictor relies on the Gaussian assumptions of the error terms of the linear mixed regression model, a feature rarely observed in applications with real data. The present paper proposes to tackle the potential lack of validity of the model assumptions by using data-driven scaled transformations as opposed to ad-hoc chosen transformations. Different types of transformations are explored, the estimation of the transformation parameters is studied in detail under the linear mixed regression model and transformations are used in small area prediction of linear and non-linear parameters. The use of scaled transformations is crucial as it allows for fitting the linear mixed regression model with standard software and hence it simplifies the work of the data analyst. Mean squared error estimation that accounts for the uncertainty due to the estimation of the transformation parameters is explored using parametric and semi-parametric (wild) bootstrap. The proposed methods are illustrated using real survey and census data for estimating income deprivation parameters for municipalities in the Mexican state of Guerrero. Simulation studies and the results from the application show that using carefully selected, data-driven transformations can improve small area estimation.
Keywords: Random effects, bootstrap, adaptive transformations, maximum likelihood estimation, poverty mapping

Abstract: The R Package emdi for Estimating and Mapping Regionally Disaggregated Indicators

The R package **emdi** enables the estimation of regionally disaggregated indicators using small area estimation methods and includes tools for processing, assessing, and presenting the results. The mean of the target variable, the quantiles of its distribution, the Head Count Ratio, the Poverty Gap, the Gini coefficient, the Quintile Share Ratio, and customized indicators are estimated using direct and model-based estimation with the Empirical Best Predictor (EBP) (Molina and Rao, 2010). The user is assisted by automatic estimation of data-driven transformation parameters. Parametric and semi-parametric, wild bootstrap for mean squared error estimation are implemented with the latter offering protection against possible misspecification of the error distribution. Tools for (a) customized parallel computing, (b) model diagnostic analyses, (c) creating high quality maps and (d) exporting the results to Excel™ and OpenDocument Spreadsheets are included. The functionality of the package is illustrated with example data sets for estimating the Gini coefficient and median income for districts in Austria.

Keywords: Official statistics, survey statistics, parallel computing, small area estimation, visualization

Abstract: Should we Transform Count Data Sets? Generalized Linear Models vs. Count Data Transformations

Count data sets are also typically analyzed by using classical linear regression models; sometimes, without a careful analysis of model assumptions inherent to these models, or sometimes, just by incorporating a transformation in the target variable to satisfy parametric assumptions. Generalized linear regression models are suitable for modeling non-continuous variables, such as the Poisson or binomial cases. They are an alternative approach for directly using count data as the target variable. Simulating data from different discrete distributions, these two approaches are compared: the generalized linear regression model and the classical linear regression model under suitable count data transformations. The analysis focuses on prediction, which is evaluated in terms of some uncertainty measures, such as the relative bias and root-mean-square error. A discussion of some relevant comparison criteria between these approaches is made in this paper on open areas for research.

Keywords: Transformations, generalized linear regression models, count data

Kurzfassungen in deutscher Sprache

Zusammenfassung: Ein Leitfaden für die Nutzung von Transformationen in linearen und linear gemischten Modellen

Ein großer Bestandteil statistischer Analysen besteht darin, den Zusammenhang zwischen einer abhängigen und mehreren erklärenden Variablen zu beschreiben. Da das lineare Regressionsmodell eine einfache Lösung für die Beschreibung dieses Zusammenhangs ist, wird es in vielen Wissenschaften angewandt. Seit einiger Zeit werden auch immer häufiger linear gemischte Regressionsmodell genutzt. Beide Modelltypen basieren auf einigen Annahmen, die

bei der Anwendung überprüft werden und erfüllt sein sollten. Wenn eine oder mehrere dieser Annahmen nicht erfüllt sind, können Transformationen helfen weiterhin die Modellklasse der linearen Modelle zu nutzen. Dafür bietet diese Arbeit einen umfassenden Überblick über verschiedene Transformationen und Schätzmethoden für die Schätzung eines optimalen Transformationsparameters basierend auf den zugrunde liegenden Daten im Kontext von linearen und linear gemischten Modellen. Der größte Beitrag der Arbeit liegt darin, dem Anwender Leitlinien an die Hand zu geben, wie man Transformationen nutzen kann, um die Modellannahmen des linearen Modells zu erfüllen, und was dabei beachtet werden muss.
Stichworte: Transformationen, Modellannahmen, lineare Modelle, linear gemischte Modelle

Zusammenfassung: Das R Paket trafo für die Transformation von linearen Modellen

Das lineare Regressionsmodell ist eine weit verbreitete statistische Methode, um Zusammenhänge zu beschreiben und Vorhersagen durchzuführen. Allerdings beruht das Modell auf einer Anzahl an Annahmen, die in der Anwendung nicht immer erfüllt sind. In diesen Fällen könnten zum einen komplexere Methoden genutzt werden, die nicht auf den gleichen Annahmen beruhen. Zum anderen können Transformationen helfen, um die Gültigkeit der Annahmen zu verbessern. Um eine passende Transformation zu finden, bietet das R Paket **trafo** einen anwenderfreundlichen Rahmen. Die Auswahl an Transformationen und Schätzmethoden für den Transformationsparameter in diesem Paket ergänzen die bisher angebotenen Methoden in R.
Stichworte: Transformationen, optimaler Transformationsparameter, Normalität, lineares Modell

Zusammenfassung: Vom Anfang bis zum Ende: eine Anleitung für die Produktion von amtlichen Statistiken für kleine Regionen

Small Area Estimation ist ein Forschungsgebiet im Bereich der Survey-Statistik mit großer praktischer Relevanz für statistische Ämter und ähnliche Institutionen. Aufgrund der schnellen Entwicklung von Methoden und Software können Forscher von einer praxisnahen Anleitung für den Umgang mit Small Area Estimation Methoden profitieren. Daher schlagen wir ein allgemeines, möglichst einfach gehaltenes Konzept für die Erstellung von Statistiken für kleine Regionen vor, das auf drei Schritten basiert: Spezifikation, Analyse und Anpassung, und Evaluierung. Insbesondere wird die Interaktion zwischen dem Anwender von Small Area Estimation Methoden und dem Statisiker bei der Festlegung der Zielregion und der Zielparameter unter Berücksichtigung von verfügbaren Daten beschrieben. Sowohl modellfreie, als auch modellabhängige Methoden werden erläutert, wobei der Fokus auf der Modellauswahl, dem Testen, der Diagnose und Anpassungen des Modells mit Hilfe von Datentransformationen liegt. Die Messung von Unsicherheit und die Evaluierung der Methode mittels modell- und Design-basierter Simulationen ist auch ein wichtiger Bestandteil der vorliegenden Arbeit. Das Konzept wird anhand der Schätzung von nicht-linearen Armutsindikatoren basierend auf realen Daten veranschaulicht und lineare Statistiken, wie der Mittelwert, werden als spezielle Fälle des allgemeinen Konzepts vorgestellt.

Stichworte: Zensus, Design-basierte Methoden, Diagnosemethoden, Ungleichheit, modellbasierte Methoden

Zusammenfassung: Datengetriebene Transformationen für Small Area Estimation

Methoden aus dem Bereich Small Area Estimation hängen im Allgemeinen von der Gültigkeit ihrer Modellannahmen ab. Die Standardversion des Empirical Best Predictor basiert auf der Normalverteilungsannahme der Fehlerterme des linear gemischten Modells. Diese Annahme wird in der Anwendung jedoch nur selten erfüllt. Die vorliegende Arbeit schlägt vor, das Problem der nicht erfüllten Normalität mit datengetriebenen Transformationen zu beheben. Dafür werden verschiedene Transformationen betrachtet, die Schätzung von Transformationsparametern im linear gemischten Modell detailliert untersucht und die Transformationen in der Schätzung/Prognose von linearen und nicht-linearen Indikatoren im Small Area Kontext angewandt. Die Nutzung von standardisierten Transformationen ist besonders sinnvoll, da dies ermöglicht das linear gemischte Modell weiterhin mit Standardsoftware zu schätzen. Somit wird die Arbeit des Datenanlysten vereinfacht. Des Weiteren werden zwei Methoden für die Schätzung von Unsicherheit eingeführt, ein parametrischer und ein semi-parametrischer (wild) Bootstrap, die die zusätzliche Unsicherheit durch die Schätzung des Transformationsparameters berücksichtigen. Die neuen Methoden werden anhand der Schätzung von einkommensbasierten Armutsindikatoren für die Gemeinden im mexikanischen Bundesstaat Guerrero basierend auf realen Survey und Zensus Daten veranschaulicht. Umfangreiche Simulationsstudien und die Ergebnisse der Anwendung zeigen, dass sinnvoll ausgewählte, datengetriebene Transformationen die Schätzung von Indikatoren in kleinen Regionen verbessern.
Stichworte: Small Area Estimation, linear gemischtes Regressionsmodell, Schätzung von Unsicherheit, Poverty Mapping, Maximum Likelihood Theorie

Zusammenfassung: Das R Paket emdi für die Schätzung und die Erstellung von Karten für regional disaggregierte Indikatoren

Das R Paket **emdi** ermöglicht die Schätzung von regional disaggregierten Indikatoren mittels Small Area Estimation Methoden und enthält Funktionen für die Erstellung, die Analyse und die Präsentation von Ergebnissen. Der Mittelwert, die Quantile der Verteilung, the Armutsquote, die Armutslücke, der Gini-Koeffizient und das Quintilsverhältnis, sowie individuell definierte Indikatoren können mit direkter Schätzung oder modellbasierten Verfahren, mit dem Empirical Best Predictor (Molina and Rao, 2010), geschätzt werden. Der Anwender wird dabei durch die automatische Schätzung von Transformations-parametern für datengetriebene Transformationen unterstützt. Ein parametrischer und ein semi-parametrischer wild Bootstrap für die Schätzung des mittleren quadratischen Fehlers sind implementiert, wobei der zweite zusätzlich gegen die mögliche Misspezifikation der Fehlerverteilung schützt. Das Paket ermöglicht (a) parallele Berechnungen, (b) die Analyse von Modellannahmen, (c) die Erstellung von Karten, (d) den Export von Ergebnissen zu Excel™ und zu OpenDocument Spreadsheets. Die Funktionalität des Pakets wird mit der Schätzung des Gini-Koeffizienten und des Medians für österreichische Bezirke basierend auf Beispieldatensätzen illustriert.

Stichworte: amtliche Statistik, Survey-Statistik, parallele Berechnungen, Small Area Estimation, Visualisierung

Zusammenfassung: Sollten Zähldaten transformiert werden? Generalisierte lineare Modelle vs. Transformationen für Zähldaten

Zähldaten werden auch typischerweise mit klassischen linearen Modellen analyisert und manchmal werden die Modellannahmen entweder nicht sorgfältig untersucht oder die abhängige Variable wird transformiert, um Verteilungsannahmen des Fehlerterms zu erfüllen. Generalisierte lineare Modelle sind hingegen die passenden Modelle zur Analyse von nicht stetigen Variablen, wie Poisson- oder binomialverteilte Variablen. Beide Methoden, die Transformation der abhängigen Variable und generalisierte lineare Modelle, sind Alternativen zur einfachen Lösung die Zählvariable als abhängige Variable im linearen Modell zu nutzen. Diese beiden Ansätze werden in der vorliegenden Arbeit in einer Simulationsstudie verglichen. Der Fokus liegt auf der Vohersage, sodass die Schätzmethoden anhand der relativen Verzerrung und der relativen mittleren quadratischen Abweichung evaluiert werden. Darüber hinaus werden unterschiedliche Vergleichskriterien und offene Forschungsfragen für die beiden Methoden diskutiert.

Stichworte: Transformationen, generalisierte lineare Regressionsmodelle, Zähldaten

Declaration of Authorship

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

Berlin, 14. November 2018

Natalia Rojas-Perilla
14. November 2018