## Abstract Title Page

# A Bayesian Nonparametric Causal Model for Regression Discontinuity Designs[1]

**George Karabatsos**

*University of Illinois-Chicago*
*Department of Educational Psychology*
*e-mail*: gkarabatsos1@gmail.com

**and**

**Stephen G. Walker**

*The University of Texas at Austin*
*Division of Statistics and Scientific Computation*

September 30, 2013

First preference: Section on Research Methods.
Second preference: Section on Understanding the Effects of Education Policies.

[1] This research is supported by a National Science Foundation Grant SES-1156372, from the Program in Methodology, Measurement, and Statistics.

## 1 Background, Purpose and Novelty of Study

The regression discontinuity (RD) design (Thistlewaite & Campbell, 1960; Cook, 2008) provides a framework to identify and estimate causal effects from a non-randomized design. Each subject of a RD design is assigned to the treatment (versus assignment to a non-treatment) whenever her/his observed value of an the assignment variable equals or exceeds a cutoff value. The RD design provides a "locally-randomized experiment" under remarkably mild conditions, so that the causal effect of treatment outcomes versus non-treatment outcomes can be identified and estimated at the cutoff (Lee, 2008). Such effect estimates are similar to those of a randomized study (Goldberger, 2008/1972). As a result, since 1997, at least 74 RD-based empirical studies have emerged in the fields of education, political science, psychology, economics, statistics, criminology, and health science (see van der Klaauw, 2008; Lee & Lemieux, 2010; Bloom, 2012; Wong et al. 2013; Li et al., 2013). Polynomial and local linear models are standard for RD designs (Bloom, 2012; Imbens & Lemieux, 2008). However, these models can produce biased causal effect estimates, due to the presence of outliers of treatment outcomes; and/or due to incorrect choices of the bandwidth parameter for the local linear model. Currently, the correct choice of bandwidth has only been justified by large-sample theory (Imbens & Kalyanaraman, 2012), and the local linear model for quantile regression (Frandsen et al., 2012) suffers from the "quantile crossing" problem.

We introduce a novel formulation of our Bayesian nonparametric regression model (BLIND, 2012), which provides causal inference for RD designs. It is an infinite-mixture model, that allows the entire probability density of the outcome variable to change flexibly as a function of the assignment variable. Moreover, our Bayesian model can provide inferences of causal effects, in terms of how the treatment variable impacts the mean, variance, a quantile, distribution function, probability density, hazard function, and/or any other chosen functional of the outcome variable. Moreover, the accurate causal effect estimation relies on a predictively-accurate model for the data. The Bayesian nonparametric regression model attained best overall predictive performance, over many real data sets, compared to many other regression models (BLIND, 2012). Finally, we will illustrate our Bayesian model through the causal analysis of two real educational data sets.

## 2 Identifying Causal Effects in a RD Research Design

In a RD design, let $R_i$ be a continuous-valued assignment variable (Berk & Rauma, 1983) having a known cutoff $r_0$, for each subject $i$. Then in such a design, the treatment assignment mechanism is defined by $A_{r_0}^{(R_i)} = \mathbf{1}(R_i \geq r_0)$, with a realization denoted by $a_{r_0}^{(r_i)} = \mathbf{1}(r_i \geq r_0)$, where $\mathbf{1}(\cdot)$ is the indicator function. In the *sharp RD design* (Thistlewaite & Campbell, 1960), the treatment receipt probability function is defined by $\Pr(T = 1 | R = r) = \mathbf{1}(r \geq r_0)$, and thus it has a discontinuous jump of 1 at $r_0$. In a *fuzzy RD design* (Trochim, 1984), the probability function $\Pr(T = 1 | R = r)$ has a discontinuous jump that is smaller than 1, at $r_0$. This smaller jump is a result of imperfect treatment compliance, which can occur in settings where the assignment variable $R$ measures the eligibility to receive a treatment, and some ineligible subjects (with $R_i < r_0$) decided to receive treatment (i.e., $T_i = 1$), and some eligible subjects (with $R_i \geq r_0$) decided receive the non-treatment (i.e., $T_i = 0$).

For each subject of a given RD study, indexed by $i = 1, \ldots, n$, let $T_i(A_{r_0}^{(R_i)}) = 1$ indicate receipt of the treatment, and let $T_i(A_{r_0}^{(R_i)}) = 0$ indicates receipt of the non-treatment, when assigned treatment $A_{r_0}^{(R_i)} \in \{t = 0, 1\}$. Also, denote $Y_i(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_n)}, \boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_n)}))$ as the $2^{2n}$ potential outcomes to treatments that defined a common time point, for all $\boldsymbol{R}_n = (R_1, \ldots, R_n)^\top$, $\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_n)} = (A_{r_0}^{(R_1)}, \ldots, A_{r_0}^{(R_n)})^\top$, and all $\boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_n)}) = (T_1(A_{r_0}^{(R_1)}), \ldots, T_n(A_{r_0}^{(R_n)}))^\top$ (e.g., Angrist, et al., 1996). Now, suppose that data of a RD design satisfies the following five assumptions: *RD:* the existence of limits $\lim_{r\uparrow r_0} \mathrm{E}(T|r) \neq \lim_{r\downarrow r_0} \mathrm{E}(T|r)$ (Hahn, et al. 2001); *Local SUTVA (LS):* $Y_i(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})}, \boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})})) = Y_i(A_{r_0}^{(R_i)}, T_i(A_{r_0}^{(R_i)}))$ for all $n_0$ subjects with $r_i$ near $r_0$ (Cattaneo et al. 2013); *Local Exclusion Restriction (LER):* $Y_i(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})}, \boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})})) = Y_i(\boldsymbol{A}_{r_0}^{(\underline{\boldsymbol{R}}_{n_0})}, \boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\underline{\boldsymbol{R}}_{n_0})}))$ for all $(\boldsymbol{R}_{n_0}, \underline{\boldsymbol{R}}_{n_0})$ and for all $n_0$ subjects with $r_i$ near $r_0$ (e.g., Angrist, et al., 1996); *Local Monotonicity (LM)*: $T_i(A_{r_0}^{(r_0+\epsilon)}) \geq T_i(A_{r_0}^{(r_0-\epsilon)})$ for some $\epsilon > 0$ and for every subject $i$ with $r_i \in (r_0 - \epsilon, r_0 + \epsilon)$ (Hahn, et al. 2001); *Local Randomization (LR)*: Each subject, indexed by $w$, has "imprecise control" over $R$, i.e., $F_R(r|w) = \Pr(R \leq r|w)$ is continuous in $r$ at $r_0$, with $0 < F_R(r_0|w) < 1$ (Lee, 2008). Then, for the subgroup of subjects with assignment variables $r_i$ near $r_0$, a complier is a subject with $(T_i(1), T_i(0)) = (0, 1)$; the $2^{2n_0}$ potential outcomes $Y_i(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})}, \boldsymbol{T}(\boldsymbol{A}_{r_0}^{(\boldsymbol{R}_{n_0})}))$ reduce to two potential outcomes $Y_i(t), t = 0, 1$; and then $Y(1) - Y(0)$ is a causal effect of $T$ on $Y$; and for any functional $h\{\cdot\}$ of $Y$, an estimator of the causal effect of $T$ on $h\{Y\}$, at the cutoff $r_0$, is given by:

$$\tau = \mathrm{E}[h\{Y_i(1)\} - h\{Y_i(0)\}|r_0 \text{ and } i \text{ is a complier}] = \frac{\lim_{r\downarrow r_0} \mathrm{E}[h\{Y\}|r] - \lim_{r\uparrow r_0} \mathrm{E}[h\{Y\}|r]}{\Pr[i \text{ is a complier}|r_0]} \tag{1}$$

(Imbens & Lemieux, 2008). Depending on the choice of $h\{\cdot\}$, the causal effect estimator (1) describes how much the treatment $T$ impacts either the mean, variance, distribution function, a quantile, probability density, or any other feature of the outcome $h\{Y\}$. The denominator of (1) is identical to $\lim_{r\downarrow r_0} \mathrm{E}[T|r] - \lim_{r\uparrow r_0} \mathrm{E}[T|r]$; and a sharp RD design has $\Pr[i \text{ is a complier}|r_0] = 1$, and trivially satisfies assumptions RD, LER, and LM. Also, the two data sets that will analyzed in the the present study, satisfies assumption LR, because arguably for each data set, each subject has imprecise control over the assignment variable.

## 3 A Statistical (Causal) Model for RD Designs

For the sharp RD design, our Bayesian nonparametric model is given by:

$$\begin{aligned}
f(y_i|r_i, a_{r_0}^{(r_i)}) &= \sum_{j=-\infty}^{\infty} \mathrm{n}(y_i|\mu_j, \sigma_j^2)\omega_j(\eta_\omega(r_i), \sigma_\omega(r_i)),\ i = 1, \ldots, n, && \text{(2a)}\\
\omega_j(\eta_\omega(r), \sigma_\omega(r)) &= \Phi(\{j - \eta_\omega(r)\}/\sigma_\omega(r)) - \Phi(\{j - 1 - \eta_\omega(r)\}/\sigma_\omega(r)) && \text{(2b)}\\
\eta_\omega(r) &= \beta_{0\omega} + \beta_{\omega 1} r + \beta_{\omega 2} a_{r_0}^{(r)} && \text{(2c)}\\
\sigma_\omega(r) &= \exp(\lambda_{\omega 0} + \lambda_{\omega 1} r + \lambda_{\omega 2} a_{r_0}^{(r)})^{1/2} && \text{(2d)}\\
(\mu_j, \sigma_j^2) &\sim \mathrm{normal}(\mu_j|\mu_\mu, \sigma_\mu^2)\mathrm{inverseGamma}(\sigma_j^2|1, b_\sigma) && \text{(2e)}\\
(\mu_\mu, \sigma_\mu^2) &\sim \mathrm{normal}(\mu_\mu|\mu_0, \sigma_0^2)\mathrm{uniform}(\sigma_\mu|0, b_{\sigma\mu}) && \text{(2f)}\\
(b_\sigma, \boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega) &\sim \mathrm{gamma}(b_\sigma|a_0, b_0)\mathrm{normal}(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega|\mathbf{0}, v\mathbf{I}_{p+1}) && \text{(2g)}
\end{aligned}$$

where $\Phi(\cdot)$ is the Normal$(\cdot|0,1)$ c.d.f.; and the mixture weights $\omega_j(\eta_\omega(r), \sigma_\omega(r))$ sum to 1 at every value of $r$. Our model (2) has infinite-dimensional parameter $\Gamma = ((\mu, \sigma_j^2)_{j=-\infty}^{\infty}, \mu_\mu, \sigma_\mu^2, b_\sigma, \boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega)$, and the degree of multimodality of $f(y|r, a_{r0}^{(r)})$ depends on the size of $\sigma_\omega(r)$ (BLIND, 2012). Obviously, there is a discontinuity in the regression at $r_0$, when either of the parameters $(\beta_{\omega 2}, \lambda_{\omega 2})$ is non-zero. Moreover, when prior information is limited, we may specify vague prior hyper-parameters $\mu_0 = 0, \sigma_0^2 \to \infty$, $a_0 \to 0$, $b_0 \to 0$, and $v = 10^5$, along with a choice of prior parameter $b_{\sigma\mu}$ that reflects prior knowledge about the range of $Y$.

A set of data $\mathcal{D}_n = \{y_i, r_i, a_{r0}^{(r_i)}\}_{i=1}^n$ updates the prior density $\pi(\Gamma)$ to a posterior density, given by $\pi(\Gamma|\mathcal{D}_n) \propto \prod_{i=1}^n f(y_i|r_i, a_{r0}^{(r_i)}; \Gamma)\pi(\Gamma)$ up to a proportionality constant. Then $\mathrm{E}_n(y|r, a_{r0}^{(r)}) = \int \left\{ \int y f(y|r, a_{r0}^{(r)}; \Gamma)\mathrm{d}y \right\} \mathrm{d}\Pi(\Gamma|\mathcal{D}_n)$ gives the posterior predictive expectation of $Y$ conditionally on $(r, a_{r0}^{(r)})$. If all five assumptions hold for the sharp RD design, then a posterior estimate of the causal effect of $T$ on $Y$ is given by $\widehat{\tau}_h = \mathrm{E}_n(h\{y\}|r_0, 1) - \mathrm{E}_n(h\{y\}|r_0, 0)$, for any choice of functional $h\{\cdot\}$. Existing Markov Chain Monte Carlo (MCMC) methods can be used to estimate the posteriors $\pi(\Gamma|\mathcal{D}_n)$ and $\mathrm{E}_n(h\{y\}|r, a_{r0}^{(r)})$ (BLIND, 2012).

Our causal model (2) can be extended to a fuzzy RD design, where it is only known that Pr[$i$ is a complier$|r_0$] $< 1$. This extension involves the estimation of bounds of the causal effects $\tau_h$, over a plausible ranges of Pr[$i$ is a complier$|r_0$] and of LM and ER violation magnitudes (Angrist et al., 1996). It is prudent to estimate such bounds, because in the fuzzy RD design, the identifying LM and LER assumptions are empirically falsifiable and unverifiable, and because the estimation of Pr[$i$ is a complier$|r_0$] is also empirically unverifiable because the estimator (1) does not identify the compliers (e.g., Balke & Pearl, 1997).

## 4 Applicability of Model; Setting; Interventions; Subjects; Data Collection and Analysis; Results; Conclusions

Two data sets were collected under a partnership between four Chicago university schools of education, which implemented a new curriculum that aims to train and graduate teachers to improve Chicago public school education. Using Windows-based software developed by the first author, we analyzed each of the two data sets using the Bayesian model, specified by the vague priors mentioned earlier, along with prior specification $b_{\sigma\mu} = 5$. All posterior estimates, reported below, are based on 40K MCMC samples, which led to accurate posterior estimates according to standard convergence assessments (Geyer, 2011).

For the first data set, the aim is to estimate the effect of the new teacher education curriculum on math teaching ability, among undergraduate teacher education students attending one of four Chicago universities. This data set involves a sharp RD design, specifically an interrupted time-series design (Cook & Campbell, 1979, Ch. 5), with the assignment variable of time, ranging from fall semester 2007 through spring semester 2013. The new curriculum (treatment) was instituted in Fall 2010 (the cutoff, $r_0$), and the old teacher curriculum (non-treatment) was active before that time. The outcome variable is the number-correct score on the 25-item Learning Math for Teaching (LMT) test (University of Michigan). The LMT score was obtained from each student, who had just completed a course on teaching algebra. A total of $n = 347$ students completed the LMT test (89.9% female; 135 and 212 students under the old and new curriculum). Among these students, the Cronbach's alpha reliability of the LMT test score is .63, and the average LMT score was 12.9 (s.d.= 3.44).

Using our Bayesian model, we analyzed the data to estimate the effect of the new curriculum, versus the old curriculum, on student ability to teach math (LMT score), at the Fall semester 2010 cutoff. The model included the LMT test z-score as the outcome (dependent) variable, and included covariates of the assignment variable TimeF10 = (Year − 2010.9)/10 and of the treatment assignment variable CTPP $= A_{2010.9}^{(\text{TimeF10})} = \mathbf{1}(\text{TimeF10} \geq 0)$, with time 2010.9 referring to the Fall 2010 cutoff. Our model displayed good fit to these data. The standardized residuals ranged from −0.84 to 0.77 over the 347 observations, and R-squared was .92. Figure 1 presents the model's posterior predictive density estimate of the LMT outcome, for the new curriculum (treatment) and for the old curriculum (non-treatment), at Fall 2010. As shown, the new curriculum, compared to the old curriculum, increased the LMT scores, in terms of shifting the density of LMT scores to the right. This shift corresponds to an increase in the mean (from .17 to .20), the 10%ile (−1.43 to −1.35), the median (.07 to .15), and corresponds to a variance decrease (1.78 to 1.69).

The second data set, from another sharp RD design, involves $n = 205$ undergraduate teacher education students, each of whom enrolled into one of the four Chicago schools of education during either the year of 2010, 2011, or 2012 (90% female; mean age=22.5, s.d.=5.35, $n = 203$); 47%, 21%, 10%, and 22% attended the four universities; 49%, 41% and 10% enrolled in 2010, 2011, and 2012). It is of interest to investigate the causal effect of basic skills on teacher performance (e.g., Gitomer & Brown, 2011), because most U.S. schools of education based their undergraduate admissions decisions on the ability of individual applicants to pass basic skills tests. Here, the assignment variable is defined by the score on an Illinois test of reading basic skills, with minimum cutoff passing score of 240. The outcome variable is the total score on the 50-item Haberman (2008) Teacher Pre-screener assessment, which has a test-retest reliability of .93, and has a 95% accuracy rate in predicting which teachers will stay and succeed in the teaching profession (Haberman, 2008). A score in the 40-50 range indicates a very effective teacher, and many schools use the Haberman Pre-screener to assess applicants of teaching positions. Among all the 205 students of the RD design, the average reading basic skills score is 204.69 (s.d.=33.7); and the average Haberman Pre-screener score is 29.82 (s.d.=4.32).

Using the Bayesian model, we analyzed the data set to estimate the causal effect of passing the reading basic skills exam (treatment), versus not passing (non-treatment), on students' ability to teach in urban schools. The model included the Haberman z-score as the outcome (dependent) variable, and included covariates of the assignment variable Rd240 = (Read − 240)/10 and the reading (Read) score passing (assignment) indicator ReadPass $= A_{240}^{(\text{Read})} = \mathbf{1}(\text{Read} \geq 240)$. Our model fit the data well. The standardized residuals ranged from −1.7 to 1.22 over the 205 observations, and R-squared was .98. Figure 2 presents the model's posterior predictive density estimates, for treatment versus non-treatment. A detailed inspection revealed that passing the basic skills reading test causally increased the Haberman z-score, in terms of the mean (from .13 to .26), median (.05 to .28), 75%ile (.97 to 1.43), 90%ile (1.66 to 2.21), 95%ile (2.13 to 2.82), and variance (1.60 to 3.36); and causally decreased the 5%ile (−1.74 to −2.38) and 10%ile (−1.30 to −1.70). Also, the treatment density and the non-treatment density each has two modes (clusters) of students, with below-average and above-average Haberman z-scores, respectively.

## References

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444-455.

Balke, A., & Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, *92*, 1171-1176.

Berk, R., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, *78*, 21-27.

Bloom, H. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, *5*, 43-82.

Cattaneo, M., Frandsen, B., & Titiunik, R. (2013). *Randomization inference in the regression discontinuity design: An application to the study of party advantages in the U.S. senate* (Tech. Rep.). University of Michigan: Department of Statistics.

Cook, T. (2008). Waiting for life to arrive: A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*, 636-654.

Cook, T., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally.

Frandsen, B., Frölich, M., & Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, *168*, 382-395.

Geyer, C. (2011). Introduction to MCMC. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov Chain Monte Carlo* (p. 3-48). Boca Raton, FL: CRC.

Gitomer, D., Brown, T., & Bonett, J. (2011). Useful signal or unnecessary obstacle? The role of basic skills tests in teacher preparation. *Journal of Teacher Education*, *62*, 431-445.

Goldberger, A. (2008/1972). Selection bias in evaluating treatment effects: Some formal illustrations. In D. Millimet, J. Smith, & E. Vytlacil (Eds.), *Modelling and evaluating treatment effects in economics* (p. 1-31). Amsterdam: JAI Press.

Haberman, M. (2008). *The Haberman Star Teacher Pre-Screener*. Houston: The Haberman Educational Foundation.

Hahn, J., Todd, P., & Klaauw, W. V. der. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*, 201-209.

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, *79*, 933-959.

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, *142*, 615-635.

Klaauw, W. V. der. (2008). Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, *22*, 219-245.

Lee, D. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, *142*, 675-697.

Lee, D., & Lemieux, T. (2010). Regression discontinuity designs in economics. *The Journal of Economic Literature*, *48*, 281-355.

Li, F., Mattei, A., & Mealli, F. (2013). *Bayesian inference for regression discontinuity designs with application to the evaluation of Italian university grants* (Tech. Rep.). Duke University: Department of Statistics.

Thistlewaite, D., & Campbell, D. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, *51*, 309-317.

Trochim, W. (1984). *Research design for program evaluation: The regression-discontinuity approach.* Newbury Park, CA: Sage.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics*, *38*, 107-141.

## Appendix B: Figures

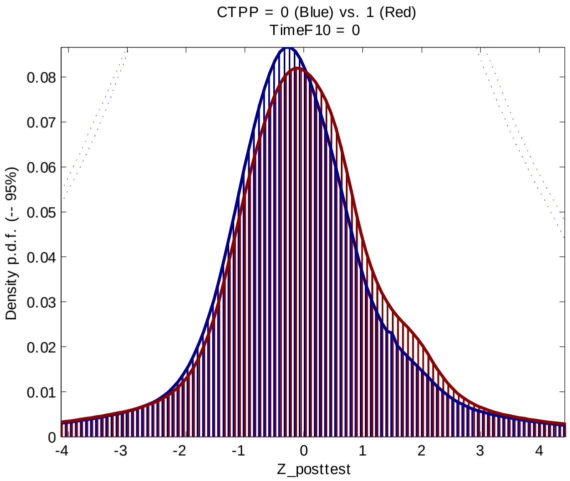

Figure 1: For the LMT z-score outcomes, the posterior predictive density estimates of $Y(1)$ (red) and of $Y(0)$ (blue).

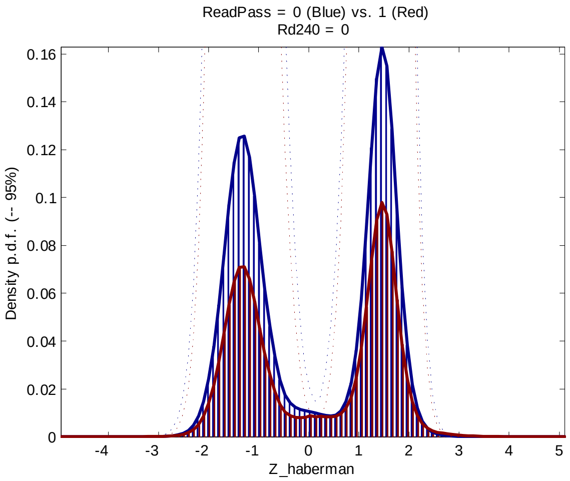

Figure 2: For the Haberman z-score outcomes, the posterior predictive density estimates of $Y(1)$ (red) and of $Y(0)$ (blue).