


1-1-2014

Robust Regression Methods For Massively Decayed Intelligence Data

Akiva Joachim Lorenz
Wayne State University,

Follow this and additional works at: http://digitalcommons.wayne.edu/oa_dissertations

 Part of the [Criminology and Criminal Justice Commons](#), [Political Science Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Lorenz, Akiva Joachim, "Robust Regression Methods For Massively Decayed Intelligence Data" (2014). *Wayne State University Dissertations*. Paper 900.

This Open Access Dissertation is brought to you for free and open access by DigitalCommons@WayneState. It has been accepted for inclusion in Wayne State University Dissertations by an authorized administrator of DigitalCommons@WayneState.

**ROBUST REGRESSION METHODS FOR
MASSIVELY DECAYED INTELLIGENCE DATA**

by

AKIVA JOACHIM LORENZ

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2014

MAJOR: EVALUATION AND RESEARCH

Approved by:

Advisor

Date

© COPYRIGHT BY
AKIVA JOACHIM LORENZ
2014
All Rights Reserved

DEDICATION

I would like to dedicate this dissertation to my family who supported and believed in me throughout the years. My late father, Wunibald Lorenz, served as a shining example to reach for my dreams. My mother, Christina Lorenz, taught me the importance of pursuing an education. Finally, a special thanks to my magnificent wife, Sarah Lorenz, for her inspiration, encouragement, and support during these trying times.

I would also like to acknowledge the late Dr. Gail Fahoome, my initial major professor. The proposal concept was developed prior to her demise. I am very grateful for Prof. Barry Markman, who was helpful in taking over the dissertation process. I would like to thank the other members of my committee, Drs. Monte Piliawsky, Irwin Jopps, Julie Smith, and Boris Shulkin, who had an important role in the dissertation process. I also benefited from conversations and advice from Dr. Robyn Mace in developing the concept. Finally, I am grateful to Prof. Shlomo Sawilowsky who had an important role in my educational and professional development.

TABLE OF CONTENTS

Dedication	II
List of Tables	IV
Chapter 1 “Introduction”	1
Security	2
Counter-Terrorism	3
Areas of Activity	4
Information and Intelligence	5
Quality of Data and its Analysis	5
Linear Regression	7
Purpose of the study	9
Research Hypothesis	9
Importance of the Study	9
Operation Definitions.....	10
Limitations.....	11
Chapter 2 “Literature Review”	13
Terrorist Motivation.....	13
Group Motivation	13
Individual Motivation	15
Terrorism Capabilities	17
Terrorists’ Modi Operandi	18
Terrorism: A Changing World	21
Suicide/Homicide Terrorism	22
Force Multipliers.....	23
Organizational Structures	28
Intelligence Cycle	30
Data Quality and Data Analysis in Intelligence	31
Data Decay	33
Missing Values – MAR, MCAR, MNAR	33

Missing Values - Censored Data.....	34
Simple Linear Regression	35
Resistant regression via Maximum Likelihood Methods.....	35
Least-Trimmed Squares Regression.....	36
Monte Carlo	36
Chapter 3 “Methodology”	38
Design.....	38
Sampling Plan.....	38
Type I Error Model Definitions.....	38
Type II Error (Comparative Power) Model Definitions	40
Analysis	40
Standard error of beta and the lqs() method	41
Tabulation of Results	41
Chapter 4 “Results”	42
Original Data Results.....	44
Type I Errors	47
Model 1 (Type I Right Censuring) Results.....	47
Model 2 (Middle Censuring) Results.....	49
Model 3 (Systematic Censuring) Results.....	50
Model 4 (Middle, Type I Right, and Systematic Censuring) Results	52
Type II Errors	54
Model 5 (Correlated) Results.....	54
Model 6 (4th Generation Correlation) Results	61
Chapter 5 “Conclusion”	68
Discussion.....	72
Type I Errors.....	72
Type II Errors (and Comparative Power).....	73
Conclusion	74
Implications for Further Research	75
References	77
Abstract.....	89
Autobiographical Statement.....	91

LIST OF TABLES

Table 1. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions	42
Table 2. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions.....	44
Table 3. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions.....	45
Table 4. Original Data; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions.....	45
Table 5. Original Data; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions.....	45
Table 6. Original Data; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions	45
Table 7. Original Data; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions.....	46
Table 8. Model 1; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions	47
Table 9. Model 1; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions	47
Table 10. Model 1; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions	48
Table 11. Model 1; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions	48
Table 12. Model 1; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions	48
Table 13. Model 2; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions	49
Table 14. Model 2; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions	49
Table 15. Model 2; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions	49
Table 16. Model 2; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions	50
Table 17. Model 2; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions	50
Table 18. Model 3; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions	50
Table 19. Model 3; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions	51
Table 20. Model 3; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions	51
Table 21. Model 3; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions	51
Table 22. Model 3; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions	52
Table 23. Model 4; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions	52
Table 24. Model 4; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions	52
Table 25. Model 4; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions	53
Table 26. Model 4; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions	53
Table 27. Model 4; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions	53
Table 28. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.2$; 100,000 repetitions	54
Table 29. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.2$; 100,000 repetitions	55
Table 30. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.2$; 100,000 repetitions	55

Table 31. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.2$; 100,000 repetitions.....	55
Table 32. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.2$; 100,000 repetitions.....	55
Table 33. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.4$; 100,000 repetitions.....	56
Table 34. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.4$; 100,000 repetitions.....	56
Table 35. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.4$; 100,000 repetitions.....	56
Table 36. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.4$; 100,000 repetitions.....	56
Table 37. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.4$; 100,000 repetitions.....	57
Table 38. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.6$; 100,000 repetitions.....	57
Table 39. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.6$; 100,000 repetitions.....	57
Table 40. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.6$; 100,000 repetitions.....	57
Table 41. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.6$; 100,000 repetitions.....	58
Table 42. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.6$; 100,000 repetitions.....	58
Table 43. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.8$; 100,000 repetitions.....	58
Table 44. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.8$; 100,000 repetitions.....	58
Table 45. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.8$; 100,000 repetitions.....	59
Table 46. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.8$; 100,000 repetitions.....	59
Table 47. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.8$; 100,000 repetitions.....	59
Table 48. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 1$; 100,000 repetitions.....	59
Table 49. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 1$; 100,000 repetitions.....	60
Table 50. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 1$; 100,000 repetitions.....	60
Table 51. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 1$; 100,000 repetitions.....	60
Table 52. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 1$; 100,000 repetitions.....	60
Table 53. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.2$; 100,000 repetitions.....	61
Table 54. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.2$; 100,000 repetitions.....	61
Table 55. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.2$; 100,000 repetitions.....	62
Table 56. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.2$; 100,000 repetitions.....	62
Table 57. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.2$; 100,000 repetitions.....	62
Table 58. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.4$; 100,000 repetitions.....	62
Table 59. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.4$; 100,000 repetitions.....	62
Table 60. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.4$; 100,000 repetitions.....	62
Table 61. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.4$; 100,000 repetitions.....	63
Table 62. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.4$; 100,000 repetitions.....	63

Table 63. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.6$; 100,000 repetitions	63
Table 64. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.6$; 100,000 repetitions	63
Table 65. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.6$; 100,000 repetitions	64
Table 66. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.6$; 100,000 repetitions	64
Table 67. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.6$; 100,000 repetitions	64
Table 68. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.8$; 100,000 repetitions	64
Table 69. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.8$; 100,000 repetitions	65
Table 70. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.8$; 100,000 repetitions	65
Table 71. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.8$; 100,000 repetitions	65
Table 72. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.8$; 100,000 repetitions	65
Table 73. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 1$; 100,000 repetitions	66
Table 74. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 1$; 100,000 repetitions	66
Table 75. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 1$; 100,000 repetitions	66
Table 76. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 1$; 100,000 repetitions	66
Table 77. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 1$; 100,000 repetitions	67

Chapter 1

Introduction

Decay or degradation of materials has been primarily discussed in the fields of engineering, chemistry, and biology. Topics include the failure points of materials used in construction, the capabilities of batteries to hold a charge, and the reduced efficiency of medications. Another example is digital imagery due to a loss of information points, because the decay of pixels in a digital graphic .jpg file leads to a blurred image (citation).

Data decay in the fields of education, social and behavioral science, and political sciences, however, remain uncharted areas. For example, there are no current models of data degradation as it relates to criminal justice or its subfields of counter-terrorism and intelligence. Consumers of counter-terrorism and intelligence information, such as the Department of Homeland Security in the United States, would benefit from an exploratory analysis of the impact of data decay when dealing with predictive analyses (i.e., regression methods).

Homeland security became a buzzword in the late twenty-first century. Fuelled by the technological advances and the vulnerabilities of an interconnected and interdependent global infrastructure, terrorism has become the “preferred tactic for ideological extremists” (Joint Publication 3-26, 2009). Terrorism has changed dramatically over the past decades in terms of perpetrator, ideology, tactics, and scope of operation, culminating in the death of thousands.

Combating the threat of terrorism, workers at security agencies rely on intelligence that consists of the collection of information, its analysis, and action on the gained knowledge. Information is collected from various open and covert sources and is stored by their representative agencies after it has been scanned for important or predictive markers.

Unfortunately, sometimes the stored information is faulty, incomplete, or is information that - at one point in time - was correct but subsequently decayed. Hence, quantitative research methodologies, regardless of their sophistication, may yield less than optimal outcomes. For example, intelligence failures, such as the inability to

predict or prevent the attacks of September 11, 2001 on the twin towers of the World Trade Center and the Pentagon in the United States exposed the limitation of compartmental intelligence warehousing. It was asserted that the limitations were fueled by the rivalry between the different services, and restrictive policies and laws (9/11 Commission, 2004). However, no information service, within its own jurisdiction, held sufficient (or sufficiently correct) information (9/11 Commission, 2004).

The creation of joint agency fusion centers and the Joint Terrorism Task Forces (JTTF) aimed to facilitate the sharing of intelligence, the collaboration between agencies and the pooling of different databases. This is referred to as terrorism informatics (Chen, et al., 2008). Informatics has also been developed in other arenas, including bioinformatics, health informatics, and human services informatics. Ultimately, inter-agency barricades and public policies should no longer prevent the information flow necessary to predict or prevent major terrorist acts. Nevertheless, exploratory intelligence analysis based on statistical analysis of raw data is still in its infancy and no published sources exist on the effectiveness of statistical models when confronted with massively decayed data.

Moreover, what constitutes intelligence analysis versus its predecessors as constituent parts is subject to debate, as Angrell (2002) stated that "information processing, information screening, and informatics are sometimes described as intelligence analysis, which they are not" (p. 6). Angrell (2002) also noted that veteran analysts work hard "to see, with an intuitive ability, potential intelligence values" (p. 5).

Security

The Department of Homeland Security (DHS) was created by an act of the United States Congress in November, 2002, in part, to ameliorate the flow in inter-agency information. Due to the urgency brought about due to global threats, it became operational only four months later. Its mission is to integrate twenty-two federal departments and agencies into a single, comprehensive department to secure the United States from threats.

Similarly, other departments and agencies exist that are charged with protection from threats. For example, the Federal Bureau of Investigation (FBI) and the United States Marshal Service (USMS) within the Department of Justice are concerned with domestic threats, and the Central Intelligence Agency (CIA) is concerned with foreign threats.

The security infrastructure in the United States is mirrored by similar structures around the world. For example, in Germany the overarching service for foreign threats is the Bundesnachrichtendienst (BND) which is under the auspices of the Chancellor's Office, and Bundesamt für Verfassungsschutz (BfV) domestic intelligence agency which is under the Federal Ministry of the Interior. In Israel, the Sherut haBitachon haKlali (Shin Bet) is charged with internal threats and the Mossad handles foreign intelligence services. In the United Kingdom, Military Intelligence (MI5) and the Secret Intelligence Service (MI6), both under the Joint Intelligence Committee (JIC), deal with internal and foreign security, respectively. Each country's internal security infrastructure also provides for state militias, state police forces, municipal sheriff and police forces, firefighters, park rangers, and emergency medical personnel. Their collective purposes are to provide front line or ancillary assistance in response to domestic security concerns.

Counter-Terrorism

Counter-terrorism has been defined as the "actions to inhibit terrorism attacks or curtail their consequences" (Enders & Sandler, 2012, p. 103). In this capacity, security agencies are confronted with a myriad of obstacles including the absence of a single definition of terrorism and constant change of threat scenarios. However, the definition of a terrorist act has not been universally accepted.

This confusion of definitional distinctions has not been without a debilitating impact. For example, in the United States, despite a pledge to work closely together, the different political, law enforcement, and defense agencies remain unable to agree on a single definition of terrorism. Hence, administrators of the agencies have developed their own definitions, interpretations, and methods of approaching terrorism,

which has been influenced by specific priorities and outlooks of the agency involved (Hoffman, 1999). This inconstancy limits the cooperation between the different local, state, and federal agencies.

On a global stage, the lack of a single definition of terrorism is further exacerbated with respect to the cooperation among the security agencies between different countries. Surveying scholars around the world, Schmid and Jongman (1988) found 109 definitions. More recently, Weinberg's et al. (2004) literature review found 73 definitions of terrorism.

Moreover, even defining terrorism among terrorist groups is a difficult task. In order to win and maintain a support base, terrorist groups have tried to rid themselves of any negative connotation with terrorism and hide behind semantic camouflage, describing themselves as freedom fighters or urban guerrillas. These semantic debates have been especially noticeable in the news media. In an effort to appear neutral, reporters in the news media often have used terms such as terrorist and freedom fighter interchangeably to describe the perpetrators. Due to this ambiguousness, Hoffman (1999) noted that there is no widely accepted definition of terrorism (p. 37).

Areas of Activity

Over the past century, security professionals witnessed a change in the ideology and tactics of terrorists and terrorist organizations ranging from the early Russian anarchists in the 1890s to the current religious terrorism around the globe. Rapoport (2004) indicated the existence of four defined terrorist ideologies (referred to as wave theory). Hoffman (1999) noted similar changes in terrorism ideology and tactics, defining them as: 1) ethno-nationalist/separatist, 2) international, 3) religious, and 4) state-sponsored terrorism.

Terrorist attacks can occur in four areas: 1) Land, 2) Air, 3) Maritime, and 4) Cyberspace. Terrorist strategies and tactics are governed by their respective ideologies and by the availability and vulnerability of desired targets. As rational decision makers, terrorists weigh the benefits of an action (probability of success and gained publicity)

while trying to minimize potential risks. Thus, with the hardening of high value targets such as government installations, terrorists move to more vulnerable targets of choice.

Information and Intelligence

Intelligence consists of the collection of information, its analysis, and action on the gained knowledge (Gilboa, 2012). Information is collected from various overt and covert sources. Classified information is what a government deems sensitive and vital for its operation. Gaining access to these protected information requires the use of covert interception by either an agent (HUMINT, or human intelligence) or through technological means (e.g., SIGINT, or signal intelligence).

Open Source intelligence (OSINT) and information are publicly available and can include media sources such as newspapers, television, and user generated content (e.g. social networking and sharing sites). Another important source for OSINT are databases from governmental, business, academic, and non-profit organizations (e. g., Census, Equifax).

Automatic data interception and processing rates have quadrupled over the past decade. It is estimated to occur at a rate of over 20 terabytes per minute (Bamford, 2012). Fed by geostationary satellites and domestic and international listening posts, the National Security Authority (NSA) alone requires five substations and more than a million square feet of digital storage. The new datacenter in Bluffdale, Utah will serve as the center of the NSA's cloud-based management strategy. It is estimated to cost US\$ 2 billion (Bamford, 2012). This increasing availability of information changed the intelligence analysis from "a process of stitching together parse data to derive conclusions to a process of extracting conclusions from aggregation and distillation of massive data and data reflections" (Farber, et al., n.d.).

Quality of Data and its Analysis

The concern with corruption in data is "one of the oldest and most fruitful lines of statistical investigation," (Fisher, 1925, p. i). Grace and Sawilowsky (2009) noted that the principle of Garbage In – Garbage Out (GIGO), which emerged from the early days

of programming electronic computers, is a major threat to meaningful data analysis. Corruption emerges from a variety of sources and for a variety of reasons. For example, a datum may be transformed via a keystroke error from a meaningful value into an outlier (i. e., beyond expected minimum or maximum values) or an inlier (i. e., hidden by being placed toward the median), it may become auto-correlated, or it may split into mixed-distributions.

Although there were sporadic attempts to handle corrupt data in the early 20th century, the most comprehensive practical treatment of the statistical analysis of corrupt data was known as the "1972 Princeton Study" (Andrews, et al., 1972). It was a collection of Monte Carlo studies on a variety of statistical procedures to determine (1) the impact with regard to Type I error and other statistical properties due to corrupt data, and (2) the initiation of the search for methods that were robust to corrupt data. With regard to linear regression, for example, their modest conclusion was "next to nothing is known about how to robustize regression procedures with respect to errors in the C_{ij} (Huber, 1972, p. 1062)." Robustize, or to make robust, referred to the ability of the regression method to preserve the false positive error rate to the threshold set by nominal alpha when the data are sampled from a source that does not meet the distribution (e.g., normality) or other underlying requirements (e.g., homoscedasticity) of the statistic.

However, the "Princeton Study" served as a Sputnik moment, and propelled many workers toward the development of solutions to this problem. Statistical methods were developed that could be shown to be robust, at least according to some local definition, to the impact of corrupt data. For example, with regard to linear regression, within a decade, Brown (1982) reviewed the flurry of studies, numbering over a dozen, conducted to determine methods of making regression robust to outliers. Brown offered the BML (or β maximum likelihood) method, which was specific to the presence of a form of an outlier called "one-wild" (p. 74). More modern approaches are the least-trimmed squares and resistant regression methods (Verzani, 2004, p. 100). (These forms of corrupt data are further discussed in Chapter 2.)

Linear Regression

An entry level objective method that is useful in explaining variance, or in predicting future status, value, or location is simple linear regression. This is a least squares method of the form

$$Y' = a + \beta X_1, \quad (1)$$

where Y' is the predicted outcome or dependent variable, a is the Y-intercept, β is a standardized weight (although in the case of simple linear regression such as (1), the unstandardized weight b is equivalent), and X is the independent variable. A test of β_1 is essentially a test of the veracity of X_1 . Thus, testing the null hypothesis $H_0: \beta = 0$ against the alternative hypothesis $H_a: \beta \neq 0$ is a test to determine if Y can be regressed by X in order to either explain or predict variance. The $n-2$ df t-test is applied to β , and if the null is rejected, then X is considered useful as an explanatory or a predictive independent variable for Y . (Note that the independent samples t test on X and Y can be accomplished by dummy coding group membership and then conducting Eq. (1).) After testing for the significance of β , the next step in linear regression is to evaluate the explained R^2 .

It is a straightforward matter to extend simple linear regression to multiple linear regression through the introduction of a second, third, or more independent variables (see, e. g., Hair, et al., 2006, Chapter 4). This introduces a variety of complexities, such as the order of entry (i. e., if the independent variables are hierarchical), method of entry (e. g., stepwise if no a priori hypothesis exists, forward entry, backwards removal), the degree that various independent variables are correlated with the dependent variable and with each other, homogeneity of regression slopes, independence of error terms, and residual analyses.

All of these complexities become exacerbated due to the presence of decayed data. Therefore, a necessary first step prior to evaluating the robustness (e. g., with respect to Type I error for departures from population normality) of a regression technique is to first ascertain whether the method is successful in simple linear regression when the underlying distribution is normally distributed prior to decay.

In terms of outliers, in simple linear regression there are “outliers in the univariate sense – a data point that doesn’t fit the pattern set by the bulk of the data” (Verzani, 2004, p. 98). There is also a second type, which are “outliers in the regression model” that are “data points that are far from the trend or pattern of the data” (Verzani, 2004, p. 98).

The presence of a one-wild outlier is an example of an outlier in the univariate sense. Outliers in the regression can be modeled by more sophisticated types of decay. Although there is not a uniform definition of data decay, this form of corruption can be expressed as a censoring or missing values, as well as any other conceivable way to weaken or stress the initial variable with respect to Type I error.

In terms of robustness with respect to Type II error, a straightforward method of decaying data can be invoked when creating correlated data with methods that do not preserve distributional properties. For example, the algorithm

$$y = rx + z\sqrt{1-r^2} \quad (2),$$

given in Sawilowsky and Fahoome (2003, p. 295) for producing correlated data “does not provide for controlling skew, γ_1 and kurtosis, γ_2 ” (p. 300). Note that (2) pertains to data samples from the Gaussian distribution. In this sense, those descriptive statistics (skew and kurtosis) in the original data set have decayed during the production of correlated data. Hence, (2) is a desirable method for correlating data if the objective is to model decay, because it will produce the desired degree of correlation while arbitrarily modifying the initial values of skew, kurtosis, and higher moments.

Unfortunately, a review of the development of robust regression methods by Brown (1982) showed that none of the techniques were evaluated for an onslaught of severe corruption and decay. For example, the simultaneous introduction of censoring, missing values, and poorly correlated data were not the subject of any of the investigations mentioned by Brown (1982).

New robust regression techniques have been developed that are far more sophisticated than adaptations to regression models based on minor perturbations;

these are called least-trimmed and resistant regression. However, their robustness properties in the presence of massively decayed data are not yet known.

Purpose of the study

Given the potential state of massive and varied data decay models in security studies, and the historical fact that robust regression techniques were developed under simple or singular forms of data corruption or decay, the purpose of this study is to compare least-trimmed and resistant regression in the simple linear regression model. This will provide useful information in determining which - if either - method is useful in predicting future status, value, or location of assets in security studies. If one or both are successful, recommendations can then be made to extend the technique(s) to more complicated general linear models, beginning with multiple linear regression, to informatics in other arenas (e.g., education, social and behavioral sciences).

Research Hypothesis

This study is designed to investigate the use of least-trimmed ($lqs()$) and resistant ($rlm()$) simple linear regression in the $N-2$ df t -test test of β as methods to preserve the Type I error rate and Type II error rate for data decay models that potentially may appear in small samples terrorism informatics and security data. Although an exhaustive comparative power analysis would be premature, the introduction of various levels of correlated data will simulate the impact of simple effects, and will provide a glimpse of the competitiveness of $lqs()$ and $rlm()$ with the ordinary least squares ($lm()$) regressions.

Importance of the Study

Simple linear regression, and its extension to multiple linear regression, is the initial choice when the data are known not to be curvilinear for modeling. However, there are classical data distribution requirements that are rarely met that may adversely impact the general linear model. The backdrop of this study is on modeling in the field

of homeland security intelligence and informatics, but the principles apply equally to education, social and behavioral sciences, and related disciplines.

Currently, the state of the art in regression modeling is limited to data that are noncompliant in the sense that they contain mild perturbations from normality. In real intelligence situations, as in related disciplines, however, the data rarely arrive in such a pristine condition. Therefore, the importance of this study is to determine if modern robust and resistant regression methods are robust to realistic data decay in the simple linear layout. If so, the next step would be to investigate more complex regression models.

Operation Definitions

Least-trimmed mean. Least trimmed means regression is a technique that follows the least-squares regression method, except the sum of squared residuals are replaced with the “sum of the q smallest squared residuals, where q is roughly $n/2$ ” (Verzani, 2004, p. 100), which essentially is an M (maximum likelihood) estimator. It is invoked in R via the `lqs()` function located in the MASS package.

Maximum likelihood regression. This is a form of resistant regression, invoked via the `rlms()` function located in the MASS package, is equivalent to the `lqs()` function, except that the method can be changed from M (maximum likelihood) to other probability models.

Power. Power is the ability to reject a false null hypothesis. Although this study does not present a systematic comparative power analysis, the comparison of Type II error rates will give an indication of typical power comparisons.

Terrorism informatics. Terrorism informatics is defined as “the application of advanced methodologies and information fusion and analysis techniques to acquire, integrate, process, analyze, and manage the diversity of terrorism-related information for national/international and homeland security-related applications” (Chen, 2000, p. xv).

Type I error. A Type I error is defined as the false positive rate (Sawilowsky & Fahoome, 2003, p. 157.) It refers to rejecting the null hypothesis when in fact it is true.

Type II error. A Type II error is the failure to reject a false null hypothesis. A parametric statistic is considered robust with respect to Type II error when the rejection rate under assumption violations produces approximately the same rejection rate in the absence of those violations.

Limitations

A limitation of this study pertains to the models of decayed data. In terms of Type I errors it will be limited to various models of censoring and missing data in terms of Type II errors (and comparative power) it will be limited to treatments model as correlated data, and the multi-generational correlation (i. e, invoking of Eq. (2) four times to produce Y values) using Eq. (1) (meaning the data will originate from a Gaussian distribution prior to being decayed).

The `lqs` and `rlm` routines available in R's MASS library produce the y intercept, beta, and other summary statistics. However, neither produces the p value associated with beta. The reason is because the t test (or Z test for large samples) on beta is defined as beta divided by the standard error of beta, which is then associated with the $df = N - 2$ for the t distribution (which asymptotically converges with the Z distribution). It is generally not optimal to use the normal theory formula for the standard error (i.e., the standard deviation divided by the sample size) because it is not robust to non-normally distributed data (including decayed data). There are potential alternatives, such as the winsorized sample standard deviation, or a jackknife or bootstrap approximation (see, e.g., Sawilowsky & Fahoome, 2003, p. 22, 376 - 382). However, there are many limitations to those alternatives. Although Wilcox (1996), for example, provided such alternatives in computing the standard error for other hypothesis tests (e.g., the sample median), he first presented a test using the robust estimator combined with the normal curve theory standard error (see, e.g., p. 120). Hence, in this dissertation, the p value associated with beta will be determined with the normal curve theory standard error, despite the fact robust methods (i.e., `lqs` and `rlm`) will be used to determine the value of beta. When the statistical literature settles on an

optimal robust standard error, this study should be replicated using it to determine the p value associated with beta.

Chapter 2

Literature Review

Whether terrorist organizations conduct attacks depends on two factors: motivation and capabilities. If both indexes are high (i.e., above a certain threshold), it is likely that a terrorist organization will conduct an attack. Attacking when only one of these two factors is present poses a security services risk, that is, a boomerang effect in which an insufficient attack on a group's capabilities results in an escalation of terror activities as the motivation level rises, and vice versa. Thus, a successful counter-terrorism policy will concurrently address both the motivation and capabilities of a terrorist organization (Ganor, 2005).

Terrorist Motivation

Group Motivation

From a historical perspective, the prominence of the long term (strategic) motivation of terrorist groups has changed over time. Rapoport (2004) described this condition as waves and coined the term "wave theory", in which terrorist groups' ideologies and tactics evolve based on the given socio-economic environment. Although Rapoport (2004) indicated the existence of four distinct waves (anarchist, anti-colonial, international, religious), other researchers have developed their own terrorist typographies based on actor, purpose, motivational, or geographic factors. Complicated by the lack of a universal definition and the multitude of variables found in terrorism, Ganor (2011) noted that "very few typologies actually meet this goal and succeed in forging a connection between a certain category and terrorists' behavior" (p. 270). Although Thornton (1964) and Shultz (1978) distinguished between two groups employing terrorism, those in power (i.e., enforcement terror) and those aspiring power (i.e., agitational terror), most terrorist typologies primarily deal with sub-state actors only. Examples of terrorist typologies include:

- Crenshaw (1981): revolutionist, nationalists, separatists, reformists, anarchists, and reactionaries;
- Gurr (1989): vigilante, insurgent, single-issue, separatist, and revolutionary;

- Hoffman (1999): ethno-nationalist/separatist, international, religious, and state-sponsored;
- Laqueur (1999): far rightist, religious, state, exotic, and criminal;
- Wilkinson (2001): nationalism, separatism, racism, vigilantism, ultra-left ideology, religious fundamentalism, millennialism, and single-issue;
- Barkan and Snowden (2001): vigilante, insurgent, transnational, and state;
- Cronin (2003): leftist, rightist, ethno-nationalist/separatist, and scared;
- Rapoport (2004): anarchist, anti-colonial, international, and religious;
- Vasilenko (2004): political, separatist, nationalist, religious, and criminal;
- Post (2008): social-revolutionary, right-wing, nationalist-separatist, religious-extremist, and single-issue;
- Martin (2009): state, dissident, religious, criminal, and international; and
- Ganor (2011): revolutionary, national liberation, social, separatist, radical ideological, and religious.

State or state-sponsored terrorism, also known as enforcement terrorism, is the use of force or the threat of violence by a state or state sponsored organization to pursue specific policy objectives. State terrorism describes the direct involvement of a state's agency in the terrorism activities against internal and external opposition. State-sponsored terrorism describes indirect involvement by either providing save heavens, financial, and operational resources, or using third party terrorist groups to conduct attacks (and take responsibility for them). State-sponsored activities provide the advantage that - not only can the state actor now bring pressure against its opponent without being directly implicit in the act, but terrorist groups also have fewer constraints because they do not have to rely on local support of the population (Shultz, 1978; Ganor, 1997; Hoffman, 1999).

Criminal terrorism, defined as either the use of terror to eliminate rivals in a profit-oriented environment (Vasilenko, 2004) or the use of criminal activity to sustain a terrorist organization (Martin, 2009), do not fit the *casus belli* (i.e., definition) of a political goal. Therefore, discussing the different ideologies, Crenshaw's (1981) and

Ganor's (2011) frameworks provided useful starting points as their extensive and broad categories allow for the absorption of most of the motivations listed above.

1. *Revolutionary Organizations* use terrorism to change the fundamental socio-political foundation of a state and its government. Often influenced by radical ideologies, these organizations act to change a nation's regime or government, for example, Peru's Shining Path.
2. *Social Organizations* use terrorism to change in the socio-economic structure of a nation, such as El Salvador's Farabundo Martí National Liberation Front (Ganor, 2011). However, socio-economic ideas are often based in radical ideologies which require a change of government. Therefore, social organizations are often viewed as sub-group of revolutionary organizations
3. *National Liberation Organizations* use terrorism to achieve national independence from a colonial or occupying force, such as the National Liberation Front in Algeria and the National Organization of Cypriot Fighters (EOKA).
4. *Separatist Organizations* use terrorism to achieve independence (i.e., secession) for an ethnic or religious minority from a state. Examples include the attacks of the National Liberation Front of Corsica against the French and those of the Euskadi Ta Askatasuna (ETA) against the government of Spain.
5. *Radical Ideological Organizations* use terrorism to spread their extremist ideologies such as anarchism, communism, and fascism. Groups belonging to this category include the Red Army Faction in Germany and the Italian Red Brigade.
6. *Reformist and Single Issue* groups or individuals use terrorism to pressure an organization or government to change a specific policy (but not to overthrow the government). Bombings such as by the "right to life" (Bowie, 2005) or against nuclear constructions sites (Crenshaw, 1981) are examples.
7. *Reactionary Organizations* use terrorism to prevent any change to the current political, territorial or socio-economic structure, such as Northern Ireland's Ulster Defense Association.
8. *Religious Organizations* use terrorism to fulfill a divine duty (e.g., the will of G-d) in order to dissemination of their religion, advance religious interests including the creation of an area/state governed by religious beliefs, or to defend their religion from perceived hostile sources. Although religious terrorist groups can be found in many religions (e.g. Muslim, Christian) and cults (e.g., Aum Supreme), it is primarily associated with Islamic groups such as Al Qaeda, Hezbollah, and the Palestinian Islamic Jihad.

Individual Motivation

The pathways for a person to join a terrorist organization can be quite diverse due to the different types of catalysts, including: perceived injustice or humiliation,

need for identity, and a need for belonging (Borum, 2011; McCormick, 2003). Therefore, no one-size-fits-all approach exists in understanding the different psychological factors that influence violent behaviors (Rogers, 2011). Among the different schools of thought are: the psychoanalytical approach argues that terrorism is an abnormal activity caused by a psychological disorder such as Absolutist/Apocalyptic, Narcissism, and Paranoia. The cognitive approach views the use of terrorism as a rational or logical choice of an individual based on their prior behaviors that can be understood through Rational Choice Theory or Humiliation-Revenge Theory. The social approach suggests that extremist behaviors are primarily based on group memberships and group identity (Rogers, 2011).

During the radicalization and indoctrination process a cognitive restructuring occurs through de-individuation (i.e., diffusion of responsibility). By dehumanizing the target (i.e., cultural devaluation), the recruit builds his/her moral justification for violent acts. McCormick (2003) noted that the radicalization process can be viewed as the "result of a dialectical process that gradually pushes an individual toward a commitment to violence over time" (p. 492). Thus, the identity of the terrorist group becomes the same as the person's own, thereby fulfilling his/her need for belonging and becoming homogeneous with the group. The person leaves any doubts behind and uses feelings of alienation and grievances to morph into a committed group member by accepting the justifications of the terrorist group. The possible radicalization process (see Figure 1) was visualized by Lorenz (2011).

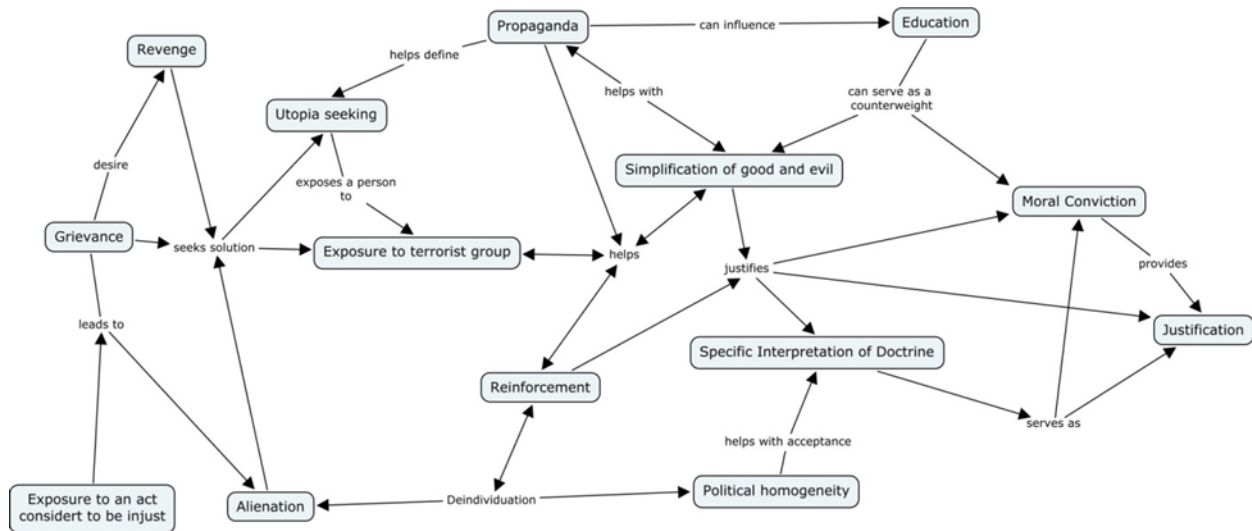


Figure 1: Radicalization Pathways (Lorenz, 2011).

Religious extremist groups that have risen to the level of terrorism are one of the main benefactors of crisis, such as civil wars or oppression by a state, as people are attracted to their simplified definitions of good and evil, moral conviction, and religious rituals. However, terrorists will require a constant reminder (i.e., reinforcement) to stay on course (Borum, 2011; Bartol & Bartol, 2009; McCormick, 2003). Terrorist groups have found that an early indoctrination process will yield stronger loyalty and commitment to the terrorist organization. Therefore, some organizations such as Hezbollah or Liberation Tigers of Tamil Eelam (LTTE) have established educational programs that range from baby brigades to college groups (Love, 2010; Ramasubramanian, 2004).

Terrorism Capabilities

Although long-term goals of terrorist groups or individual terrorists can be quite diverse, all use terrorism to gain attention and/or recognition for their cause (Thornton 1964; Crenshaw 1981; Hoffman, 1998). Attacks can also occur for the following reasons: remind the world of their existence and cause, disorientate the population by interrupting their day to day life, eliminate internal or external opposition, provoke reprisal by the state, or to build the moral of their own constituency (Thornton 1964).

The timing and style of attack - such as demonstrative, destructive, and suicide terrorism (Pape, 2003) - depend on the terrorists' motivational index (e.g., cause and desired outcome) and their capabilities.

Terrorists' Modi Operandi

Acts of terrorism are perpetrated in four realms: land, air, maritime, and cyberspace. These acts may be contained entirely within a designation, or may cross boundaries from one *modus operandi* (*M. O.*) to another.

Land

Terrorist tactics on land include assassinations of high value political and business executives. Assassinations occur through the use of explosives, small arms, and poison. Targeted assassinations were the primary tactics of the Nechaev anarchists (Rapoport, 2004). Hostage taking and kidnapping are two means by which a terrorist (or terrorist group) takes one or several individuals to be used as pawns in political negotiations or to raise funds. The difference between a hostage situation and a kidnapping is that, in the latter, the location of the hostage is unknown. Because of their publicity, hostage and kidnapping scenarios were among the preferred means of ethno-nationalist/separatist, and international terrorist organizations such as the Palestine Liberation Organization (PLO) and the Red Army faction (RAF).

Bombings can be used in two tactical situations. Although both seek to gain publicity for the cause of the perpetrator, in the first case the bombings of governmental or iconic buildings by ethno-nationalist/separatist terrorist organizations often only cause property, idealistic and religious fanatics often use bombings to cause nondiscriminatory mass damage and loss of life among the population. The rise of religious terrorism expanded the target definition and calls for the use of unlimited and unconstrained violence. The use of unconventional weapons such as chemical or biological agents (e.g., Tokyo attacks by Aum Supreme), and the use of improvised explosive devices together with suicide terrorism tactics have resulted in higher numbers of casualties (Hoffman, 2006).

Air

Terrorists have long realized the vulnerabilities of the global mass transportation system. Aircraft have several vulnerabilities. Commercial aircraft carry a large number of passengers in a confined space and rely on airports for fuel and to load and unload passengers and cargo. Airports employ thousands to facilitate the operation of commercial and private airplanes, the shipment of cargo, and the handling of passengers. Security agencies are challenged daily to strike a balance between the requirements of keeping an airport running efficiently and providing security for travelers. Moreover, no one universal airport security mechanism exists, potentially compromising aircrafts and its passengers.

Four types of terrorism attacks have occurred via air: Hijacking of aircraft by international terrorist organizations such as the 1976 hijacking of Air France Flight 139 to Entebbe by terrorists of the Popular Front for the Liberation of Palestine (PFLP-EO) and the German Revolutionäre Zellen; this was intended to gain the attention of the international media as well as socio-economic concessions. Vulnerabilities in airports security enabled Libyan agents to smuggle explosives onto Pan Am flight 103 (1988) in an act of state-sponsored terrorism that cost the lives of 270 people. Aircraft are especially vulnerable to Rocket Propelled Grenades (RPGs) and surface to air missiles (SAMs) during takeoff and landing (e.g., Arkia flight in 2002) due to the low speed and flight level of the aircraft. A hijacked aircraft can be turned into a human controlled missile and potentially attack targets of high value political, economic and defense such as the attacks in the United States on the World Trade Center and the Pentagon on 9/11.

Maritime

Since the successful attack by the Provincial Irish Republican Army on the yacht of Lord Mountbatten in 1979, maritime terrorism attacks have accounted for only 2% of all terrorist attacks worldwide (Lorenz, 2007). Notable successes were perpetrated by Al Qaeda on the USS Cole in 2000 and MV Limburg in 2002, by Abu Sayyaf on Super Ferry

14 in 2004, and by Hezbollah on the INS Hanit in 2006. This demonstrated that neither military nor commercial vessels are insulated from a well-designed maritime attack.

Lorenz (2007) noted that both maritime vessels and maritime installations (e.g., ports) are vulnerable from several types of attacks. Water-Borne Improvised Explosive Device (WBIEDs) are small craft that are loaded with explosives and ramped into the target by remote control or suicide bomber. Freighters with explosive laden containers could be sunk in port or in a maritime channel (e.g., Straits of Malacca), costing the port to close or requiring lengthy detours. Mines lain in busy maritime channels. Underwater demolition teams such as divers using Swimmer Delivery Vehicles (SDVs). Aircraft used as human missiles. Rocket attacks from sea to shore or reverse. Terrorists using the maritime transportation means to board vessels to take hostages or attack targets on shore.

Cyberspace

The internet is the new terrorist frontier. It has enabled instant, global access and exchange of information for governments, business, and private individuals. Although this digital frontier has enabled increased productivity and new academic and commercial heights, it has also provided terrorist organizations the ideal breeding and operational grounds.

Terrorist organizations use the internet for multiple purposes. The internet can serve as a mean to disseminate propaganda, radicalize and indoctrinate followers, or shape the media coverage about the organization or its actions. The increasing availability and (potential) anonymity of the internet means terrorists are not required to learn their craft in person from a teacher, but can gain the required knowledge via guidebooks and videos hosted on what is referred to as the Dark-Web - hidden or non-indexed websites (Chen et al., 2004). The worldwide financial secondary service providers (e. g., Western Union, PayPal) and virtual money markets (e.g., Bit Coin) are not required to follow federal banking and security standards to prevent money laundering. This often guarantees anonymity, requiring only a screen-name that enables terrorists to collect and move funds worldwide. In addition, the internet has

brought about a decrease in vulnerability within the command and file structure of the terrorist organization itself. This has come about due to decentralization, and the ability to communicate rapidly and accurately.

Given the reliance on the interconnectivity of the internet for business (from the supply chain command to the stock market) and governmental needs, it is understandable that cyberspace is not only a means to gather intelligence on possible targets but also an inviting target in itself for terrorist organizations. Indeed, successes of non-political hackers and "hacktivism," defined as the use of hacking by political activists (Weimann, 2005) have shown that these systems are vulnerable to attacks. Moreover, given the technical knowledge and human capital requirements, cyber terrorism is considerably less expensive than regular forms of terrorism.

Terrorism: A Changing World

Terrorist groups based on ethnic grounds such as nationalist and separatist terrorist organizations gain their strength from their local ethnic communities and those abroad (i.e., diaspora) by forging a singular identity based on history, national myths, hero worship, and the use of language. This "built-in audience among their own communal groups" (Byman, 1998, p. 151) poses both a strength as well as weakness, as they cannot afford to alienate their support base. This often constricts their operational and tactical options to achieve their clearly stated political and social objectives (i.e., independence). Tactics include the attack of symbolic national targets such as government, political, and economic buildings and personnel, while regularly releasing warnings of imminent attacks. Most attacks have occurred within national borders and were intended to demoralize the local government and to win foreign support to pressure on the local government to give into their demands. Winning the hearts and minds is therefore of the utmost importance. Thus, using high-cost strategies such as suicide terrorism were only used (e.g., LTTE) when high interests were at stake that outweigh the potential alienation of their supporters and displayed as the last mean of the weak (Byman, 1998; Hoffmann, 1998; Pape, 2003).

However, not all terrorists experience the same constraints. Religious terrorists, also known as the fourth wave (Rapoport, 2004), derive their legitimacy from divine commandments that are sanctioned by the (fundamentalist) clergy. Although religious terrorist groups exist within many religions (e.g., Christian, Jewish, Muslim) and cults (e.g., Aum Supreme), in recent history, it is primarily associated with Islamic groups. From their point of view, Islam is threatened by perceived foreign (e.g., neo-colonialism, secularism, and modernism) and internal (e.g., moderate Muslims) influences that justify a reactive and defensive jihad. Guided by divine decree, Islamic terrorist groups interpret their fight as an all-out-war without constraints based on principles such as muqawamah (i.e., active resistance) and istishhad, which serves as religious justification for self-sacrificing actions taken by a shaheed (i.e. martyr) on the battlefield. Based on this world view (e.g., morals, legitimation), Islamic terrorist often target large groups indiscriminately in order to cause mass casualties (Martin, 2006; Hoffmann, 1998). According to White (2003), because religious terrorists answer to a divine power, they “are not constrained by social norms” (p. 17).

Suicide/Homicide Terrorism

A hallmark of high casualty attacks are multiple, coordinated suicide bombings in which a terrorist either uses explosives strapped to the body as a delivering vehicle or is actively driving a vehicle improvised explosive device (e.g., car, truck, boat, airplane) into a target before igniting the charge. By definition, a suicide attack requires the death of the suicide terrorist. The fact that the terrorist (willingly) accepts their death, is a so-called thinking bomb can adapt to situational changes, and does not require an escape plan, makes the attack more likely to succeed and therefore has clear advantages over other forms of terrorism (Pape. 2003; Ganor, 2000). Shay (2004) added that upon recruitment, a potential shaheed (i.e., suicide terrorist) goes through several conditioning stages, including: physical and emotional training, operational preparations, and a farewell ceremony. After the attack, the terrorist organization will use the media to disseminate propaganda videos and messages of the shaheed in order to increase the psychological effects of the attack.

Although suicide terrorism is primarily discussed within the realm of religious terrorism, Pape (2006; 2010) argued that it does not depend on religion (noting the numerous suicide attacks by secular groups such as the LTTE in Sri Lanka or Christian groups in Lebanon) but is primarily a strategic response to a foreign occupation. In other words, suicide terrorism occurs when a foreign military force of a different religious denomination occupies places of high importance to the local population (2010, p. 85). Pape (2006) based these conclusions on the analysis of a suicide terrorism database he established. However, as Modghadam (2006) noted, Pape (2006) made critical design errors and data omissions that shifted the data to support the stated hypotheses. Pape (2010) maintained the argument that occupation causes suicide terrorism based on a high correlation between the two, disregarding that some of the groups that supported the argument (e.g., LTTE) ceased to exist, the absence of any data analyzing sectarian violence within Muslim countries, the fact that statistically a strong correlation does not indicate causality and the presence of confounding variables, such as religion.

When the focus is on the victims, the phrase suicide terrorist is reconfigured to homicide terrorist. It has been defined as “the deliberate death of others, the death of the perpetrator being incidental to the act” (Khan, Goldney, & Hassan, 2010, p. 481). Currently, the former term is preferred by most government and media public relations outlets.

Force Multipliers

As the primary motivation of terrorist groups has shifted from secular (e.g., anarchism, anti-colonialism, and separatism) to religious, so have their organizational foundations, strategies and tactics become more efficient and destructive. This change was primarily facilitated by four force multipliers: technology, transnational support, media, and religion (White, 2003). Force multipliers are defined as factors that allow governments, terrorist groups and individuals (e.g., lone wolves) to dramatically increase their combat potential without simultaneously increasing their force strength, (DTIC, n.d.; White, 2003).

Technology

Technological developments in civilian and military industries over the past century have changed the way people, organizations, businesses, and governments understand their surroundings and conduct their daily business. Technological advances, partly influenced by the cold war (arms) race, have not only brought the development and proliferation of new weapon systems, including weapons of mass destruction such as chemical, biological, radiological and nuclear (CBRN) agents, but also new means of transportation, communications, navigation, and the internet that allowed for the creation of modern, interconnected and interdependent global infrastructures and communications networks.

Technology allows for the worldwide exchange of people, goods, and information, enabling social exchanges and opening new business opportunities. Globalization, together with a post-World War II mindset, led to the creation of political and economic unions (e.g., European Union) and trade areas that reduced border checks and allowed the free flow of goods and people.

Current technology includes products on the market today that can be used for both civilian and military purposes. These types of dual-use technologies include radio controllers, satellite phones, global positioning systems (GPS), satellite imagery, encryption software, etc. Although states (with some exceptions) have used technology (including weapons) to preserve this way of life, terrorists have also embraced and adapted technology to improve their operational and tactical operations. For example, garage door openers or cell phones can be used to trigger improvised implosive devices (IEDs), while the internet can - among other things - be used for recruitment, training (by imitation), fundraising and money laundering, and to collect intelligence (Bockstette, 2008; Weimann, 2004). In many cases, vulnerabilities of potential targets can be identified by examining posted schedules, building and security plans, and recognizing attack and escape routes by examining satellite and aerial photographs (e.g., Google & Bing maps) available on the internet.

Terrorists also found that social sites, such as Facebook, provide excellent remote reconnaissance opportunities and so-called honey-traps. As a result, the Israel Security Service noted that “terror organizations are using these sites to tempt Israelis to meet up in person in order to either abduct them, kill them or recruit them as spies” (Deitch, n.d.; IPT, 2010).

Although cybercrime and “hacktivism” (Weimann, 2005) have shown the potential devastation cyber terrorism could cause, Stohl (2007) noted that no cyber-attack so far has matched the description of cyber-terrorism. Post, et al. (2000) noted that cyber terrorism is not a new concept. It was envisioned by the Italian Red Brigades in their 1978 Strategic Directions Resolution. Further, Post, et al. (2000) found that the 1998 Milworm attack on the Indian Bhabha Atomic Research Centre (BARC) was an act of cyber terrorism because it was politically motivated, targeted the digital information systems, and intended to influence and coerce an audience to change their policy.

Although these technological advances allow individuals and societies to be more productive, it also made them vulnerable to attack, as witnessed by the use of chemical agents in the Tokyo subway attack (1995), attacks against the transportation infrastructure in the Philippines (2004), Madrid (2004), and London (2005), and the proliferation of technological advanced weapon systems from state sponsors to terrorist organization, as witnessed in the attack on the INS Hanit (2006) by Hezbollah (Hoffman, 1998; Lorenz, 2007). Technology has also played a vital role in the next two force multipliers: transnational support and media.

Transnational Support and Operations

Transnational support or transnational operations is defined by White (2012) as “the ability of terrorist groups to move and hide across nations” (p. 136) and to “strike transnational economic targets” (p. 149). In an effort to increase operational security, terrorist groups have found that the establishment of a foreign support base (with or without the support of a state sponsor) provided them to the opportunity to organize, train, and plan attacks in relative safety. Terrorists profited from the lack of an universal

definition of terrorism, and the lack of global cooperation between law enforcement and intelligence services (Hofman, 1998).

Support bases were established in areas where the terrorist group had a constituency (e. g., Tamil Diaspora) and could raise funds through social pressure or through criminal activities (Jayasekara, 2007). Transnational support bases also allow terrorists to connect with other likeminded groups and exchange knowledge and support, thus increasing their reach. Attacks on transnational economic targets such as the tourism industry (Bali, 2002), oil industry (Iraq, 2006), and the transportation industry (MV Limburg, 2002) displayed the vulnerabilities of the global economic system and proved to be a means by which terrorists could influence policy making (Lorenz, 2007; White, 2012).

Media

The traditional public media (print, television, radio) has changed modern life due to the technological revolution (Biagi, 2011). In a race for audience numbers, media outlets publicize every aspect of life, including terrorism threats or attacks. Modern digital media (e.g., websites) have exacerbated this race. Publishing (uncensored) terrorist propaganda materials, such as video statements of terrorist leaders that feature subtitles in the language of the targeted audience, in prime-time news casts, provide terrorist groups the opportunity to not only to shape news coverage but also to explain and justify their violence.

The inexpensive development and maintenance of websites, together with the ability to publish multimedia content (i.e., video and audio recordings), in addition to the written content, provides terrorists an additional means to directly reach their intended audiences (Weimann, 2004). Therefore, the media plays an essential role in the battle over the hearts and minds of people (i.e., propaganda) and in gaining soft (i.e., tolerance) and hard support (i.e., practical assistance) for the terrorist groups.

The media has essentially introduced another avenue for terrorists to influence a wider audience beyond their direct victims of an attack though the means of fear - psychological warfare. Terrorist groups learned that some tactics generated more media

coverage than others and thus they were more likely to imitate those in the future (i.e., contagion effect). Moreover, increased media coverage can be associated with a subsequent increase in attacks; and vice versa, more attacks lead to more media coverage (Hofmann, 1998; Ganor, 2005; White, 2003; Rohner & Frey, 2007).

Terrorist groups such as Hezbollah designed operations not to achieve traditional military goals, but to have the maximum psychological impact. Messages addressed to the enemy included speeches from terrorist leaders, videos of successful attacks and beheadings and emphasized their determination in this long (and divine) struggle while demonizing the enemy (Schleifer, 2006; 2009). This continues exposure through internet and media outlets, resulting in additional psychological trauma, including nightmares, anxiety, depression, and post-traumatic stress disorder (PTSD), for victims of terrorist attacks and regular viewers (Bockstette, 2008; Ganor, 2005; Rohner & Frey, 2007; Silke, 2011).

There are two classifications of fear within the realm of terrorism, rational and irrational (Ganor 2005). An individual's legitimate reaction to the chance of being harmed, as calculated by the scope of the threat and the probability of its success, is considered to be a rational fear. However, what is considered to be irrational fear (or anxiety) is the event of a terrorist (or perceived terrorist) attack in which a person fears for his/her own welfare (and/or that of his/her family) and therefore changes his/her belief system (i.e., importance of national objectives). This irrational anxiety paralyzes the individual and he/she cannot further contribute to society, which is the goal of the terrorist organization. Therefore, the media plays an important role in the decision making process of terrorist organizations and is considered one of four force multipliers.

The media is not only used by terrorists to address their enemy, but also their home (i.e., constituencies) and neutral (3rd party) audiences. In any conflict, the party that is able to maintain its constituency's morale the longest has a higher chance of success. Therefore, connecting with and reminding their audience of the cause and justification for the violence and struggle is an important task for any terrorist organization. Messages addressed to the neutral audiences often highlight the "unjust" suffering or reasons for the attacks, hoping these 3rd parties will influence their enemy

to give in, or give concessions (Schleifer, 2006; 2009). The importance of modern international media networks and the internet to terrorist organizations becomes apparent considering that prior to their appearance local and regional news stations played only a subordinate role for terrorists. This is because the government was able to control and censor their appearances (Bockstette, 2008).

Organizational structures

Terrorist groups in the 20th and the beginning of the 21st century featured a hierarchical organization structure in which a charismatic leader maintained tight lines of command and control. Due to its centralized operations (i.e., direct involvement of the leadership in day to day decision making process), this organizational structure was able to undertake long-term operations and conduct negotiations. However, hierarchical (pyramid like) organizations also suffered from the constant need to communicate commands between the different hierarchies (e.g., from top to bottom), large resources, and a secure base to maintain its operation (Zelinsky, et al., 2006). Although hierarchical organizations such as the former Irish Republican Army (IRA), the Liberation Tigers of Tamil Eelam (LTTE), or Hezbollah tried to reduce their vulnerability by restricting information and operational access of the lower cadres, the centralized command and control structure were their Achilles' heels, making them susceptible to information interception which could lead to the targeted killing of their leadership and the confiscation (i.e., loss) of centralized funding.

Although embracing the global transportation and communication networks to move beyond the confinements of national borders to establish operational and support networks abroad and to connect with like-minded people (e.g., based on issues, ethnicity, or religious grounds), terrorist organizations found that security forces were able to use technology formerly used to intercept communications during the cold-war (e.g., ECHELON) or the Carnivore internet wiretapping program (post 1997) to intercept and track their operations. Together with the displayed vulnerabilities and associated costs of maintaining a hierarchical command and control structure, some terrorist organizations (including Al Qaeda) morphed to decentralize their operations and

resources. Similar to a brand (i.e., based on a common ideology), this organizational structure features self-financed, loosely knit, and independently operating cells. This organizational structure, together with ad-hoc cooperation and limited communication between groups makes it difficult for security services to penetrate and disrupt these kinds of decentralized terrorist networks (Zelinsky, et al., 2006; Zanini, et al., 2001).

The ability of terrorist organizations or networks to publish propaganda and training materials via the internet provides the opportunity for self-radicalization and training through imitation, and introduces a new facet to the counter-terrorism world, which is homegrown terrorism. Although some try to connect with their terrorist organization of choice and travel to current conflict areas, some homegrown terrorists act as lone wolves and are therefore often difficult to identify and track because of their pursuit of social isolation. Examples include: Younes Tsouli, also known as Irhabi 007 (i.e., Terrorist 007), a Moroccan-born resident of the United Kingdom who used technical expertise to post propaganda materials, secure online communications, and connected people with terrorist organizations, and Michael Adebolajo and Michael Adebowale, two converts to Islam who in 2013 attacked and murdered Lee Rigby in Woolwich, UK. These and other examples demonstrate the diversity of threats (Kohlmann, 2008; SITE, 2013).

Despite the latest commotion caused by the revelation of the PRISM program by the NSA (i.e., the collection of user information from different websites and online services) and British Tempora program (interception of all communications going through British sea cables by the GCHQ) through Edward Snowden, terrorists have long known about the vulnerabilities of using electronic communications and the internet. Katz and Raisman (2013) noted that in 2006 the Technical Mujahid Magazin released recommendations on how to stay safe while using modern technology. Subjects included the use of proxies, email encryption and data security. Similarly, the Global Islamic Media Front (GIMF) released their first encryption program in 2007, the latest version "Asrar al-Mujahideen 2" boasts modern anti-symmetric RSA 2048-bit encryption revealing western programs (without the potential of a hidden backdoor). Other groups,

such as the Technical Research and Studies Center, released guides on how to encrypt cell phones as early as 2009 (Katz & Raisman, 2013).

Intelligence Cycle

The United States Intelligence system consists of 16 different organizations, each having their unique priorities and outlooks on the challenges at hand. Although efforts were made to foster cooperation between foreign and domestic intelligence (Intelligence Reform and Terrorism Prevention Act, 2004) that introduced the post of director of national intelligence (DNI) and strove to blend law enforcement and intelligence information in the form of fusion centers (DHS, 2008) and Joint Terrorism Task Forces (JTTF), pooling different databases are still a work in progress due to the lack of definition, rivalries between the different services, and restrictive U.S. laws. However, with the estimated collection of over one billion pieces of raw data per day, collaboration is essential in order to translate, analyze, and act on the information collected (Hoffman, 1998; McConnell, 2007; Monahan, 2010).

The Federal Bureau of Investigation (FBI) defined an intelligence cycle as the “process of developing unrefined data into polished intelligence for the use of policymakers” (FBI, n.d.). A process (see Figure 2) consist of five interdependent steps: (1) planning of an intelligence operation, (2) collection of information via overt and covert means, (3) processing of information including data entry, (4) analysis and production (e.g., transforming raw data to intelligence), and (5) dissemination of intelligence to the policymakers, based on requirements set forth by the director of national intelligence (FBI, n.d.; CIA, 2013).



Figure 2: Intelligence Cycle (Source: FBI)

Data Quality and Data Analysis in Intelligence

Data quality in part depends on the means by which it is being collected from human and material sources. Due to the nature of such information the quantity and quality of the raw intelligence differs widely.

Human Intelligence (HUMINT) describes the use of agents to: conduct surveillance missions, infiltrate a target organization, develop a human source (i.e., a mole) in the target organization, or persuade key personnel to defect. Information collected from such missions is often limited and fragmented due to the counter-intelligence efforts of the target organization. Knowledge of local languages, dialects and customs is essential to develop HUMINT assets. Thus, HUMINT is often time consuming and the quality of information must be constantly monitored and verified. Furthermore, data entry mistakes caused by human factors can corrupt the information, in addition, if information is available to parties not directly related to the intelligence operation it could endanger the source (CIA, n.d.; Gilboa, 2012; Richelson, 2012)

Signals Intelligence (SIGINT) consists of the interception of signals in form of communications intercepts between people (COMINT, or communication intelligence), or those of electronic signatures (ELINT, or electronic intelligence) which helps identify and differentiate between different maritime vessels or aircrafts among others. SIGINT installations, such as the U.S. ECHELON interception system or the Carnivore internet

wiretapping program, resulted in a massive amount of information that needs to be stored, decrypted, translated, and analyzed. Advances in cryptology and automated language translations constrain SIGINT. Should a target discover he/she is compromised, SIGINT can also be used to spread misinformation compromising the quality or usefulness of information (Gilboa, 2012; Richelson, 2012).

Visual Intelligence (VISINT), together with Geo Spatial Intelligence (GEOINT) and Imagery Intelligence (IMINT), provide analysis with large amounts of detailed, high resolution images of target areas. It can be used to identify (covert) installations, track the movement of equipment and personnel, and provide real-time updates on the battlefield. Storing and analyzing this information requires large amounts of computer resources. Despite new and improved technology (e. g., nanoscale) and analytical software that could limit distortions, interpretation is still primarily a human endeavor (Gilboa, 2012; Richelson, 2012).

Open Source Intelligence (OSINT) relies on public available information, and must overcome two obstacles: information collection and analysis. Although well-known OSINT sources such as governmental and business databases, newspapers, and television stations are readily available, the internet and its undiscovered, un-indexed and sometimes temporary Dark-Web offer a treasure trove of information (Chen, 2000). Finding these sources often requires the creation of custom search bots that scour the internet for hidden links. In its analysis it is important to be able to accurately identify the source (e. g., author) and their political, social and economic views that could influence how information is being presented. Moreover, each source needs to be monitored and counter-checked and given a quality score to accurately qualify the presented information (Gilboa, 2012; Richelson, 2012).

Quantitative analysis of intelligence data deals primarily with the analysis step of the intelligence cycle. Data mining is an essential tool in intelligence analysis to discover previously unknown patterns and relationships. Its usage can be impacted by mistakes made during the collection and processing steps (e.g., data quality), interoperability between different databases and analysis software, and privacy concerns (Seifert, 2007). Moreover, the time difference between analysis and a policy decision can lead to

a disconnect between policy and the situation on the ground. Therefore, when dealing with decayed data, analysts and policy makers should know whether and what impact data decay has had on the analysis/decision making process.

Data Decay

Data decay is defined in this study as a combination of missing values (especially censored data), corrupted data (including outliers), and faded (correlated less than 1.0) data. Although each component has been previously identified and defined, no comprehensive study exists that investigates the simultaneous impact of all three on data analysis techniques used to explain or predict behavior, such as the general linear model.

Missing Values – MAR, MCAR, MNAR

Missing data are often categorized as systematic, missing at random (MAR), or missing completely at random (MCAR). Heitjan and Basu (1996), based on the definitive work of Rubin (1976), defined MAR as the “probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the unobserved data” (p. 207). Ibrahim, et al. (2005) further described MCAR as “if the failure to observe a value does not depend on any data, either observed or missing” (page 333). Hair, et al. (2006) noted the distinction lies in the generalizability of the data to the population (p. 56). In the case of MAR, values are missing randomly within subgroups but are not representative of the population, while in the case of MCAR, missing values are “indistinguishable from cases with complete data” (p. 57), and therefore are considered completely at random.

Missing data fitting in neither missing at random or missing completely at random category are considered to be systematic or missing not at random (MNAR). Hence, missing values have distinct patterns that did not occur due to some random process. Therefore, these missing values are considered to be non-ignorable missing data; and can be found often in longitudinal studies with repeated measures (Hair, et al., 2006; Ibrahim, et al., 2005; Little & Rubin, 2002). Possible methods of dealing with

missing values were described by Hair et al. (2006), Heitjan & Basu (1996), Ibrahim et al. (2005), and Little & Rubin (2002). Techniques include complete cases (CC) analysis, maximum likelihood (ML), multiple imputations (MI), fully Bayesian (FB), and weighted estimating equations (WEE). According to all approaches, the listwise deletion of all subjects with missing values is inefficient (Ibrahim et al., 2005).

Missing Values - Censored Data

Censored data are observations that are known to exist but are out of reach. Therefore, censored data are another form of missing values (Hair, et al., 2006). As noted by Cook (2008), the data are missing due to some reason unrelated to the dependent variable. For example, consider a time series study on the impact of attention deficient hypertension disorder (ADHD) on functional life skill outcomes of students with disabilities. In most states, students with disabilities age out of the formal educational system at 26 (or some similar legislated age). As a result, ADHD scores are abruptly absent beginning with age 26, even though the (now) adults with disabilities obviously continue to exist. The key component is that the age (26) is not related (or it is random with respect) to the dependent variable of interest.

Research on the impact of censored data has especially been driven by survival in clinical trials, as noted by Buckley & James (1979), Miller & Halpern (1982), and Rabinowitz, et al. (1995). Three types of censoring exist: right censoring, left censoring, and interval censoring.

Right censoring describes a situation in which data are unavailable after some point in time. For instance, a subject is not observed anymore after a certain point in time without the event of interest having occurred. This could be due to the death (unrelated to the study) of the subject, the loss of the subject (e.g., moving, dropping out), or because the study ended before the event of interest happened. Two types of right censoring can occur: Type 1 describes a situation in which a study ends at a fixed time without the occurrence of an event; Type 2 describes a situation in which a study ends after a predetermined number of events occurred within the study group. In both cases, the remaining study participants are right censored. *Left censoring* describes a

situation in which data before a certain point are unavailable. In other words, an event of interest has already occurred before the study onset, but the exact moment (e.g., date and time) is unknown. *Interval censoring* describes a situation in which the exact time of an event is not known, but the interval in which the event happens is known (Cook, 2008).

Miller and Halpern (1982) discussed four statistical analyses of censored data in regression. They are the Cox estimator (1972), Miller estimator (1976), Buckley & James estimator (1979), and the Koul, Susarla & Van Ryzin estimator (1981). They concluded that "the Cox and Buckley & James estimators are the two most reliable regression estimators for use with censored data" (p. 527). However, the Cox estimator (1972) is not able to deal with data sets exhibiting both right and left censoring.

Simple Linear Regression

Regression is a valuable tool in predicting asset locations (Fiosina, 2012). Simple Linear Regression (Eq. 1) is used to explain or predict. As with most parametric methods, it has the following underlying assumptions: normality, homoscedasticity, and linearity. It is easily extended to multiple independent variables; when the number of dependent variables are increased to more than one the method is known as canonical correlations, which is the statistical engine of discriminate function analysis (the forerunner of logistic regression).

The violation of the assumptions may bring "undesirable repercussions" (Sawilowsky & Markman, 1990, page 425). Bradley (1978) noted that "any violation of a parametric test's assumptions alters the distribution of the test statistic and changes the probabilities of Type I and Type II errors" (p. 25). Hence, the presence of decayed data will try the robustness of this procedure.

Resistant regression via Maximum Likelihood Methods

According to Ripley (2004), resistant regression "is about non-disastrous behavior in the presence of incorrect data points." A natural, initial solution to the problem of noncompliant data in regression was developed in the early 1980s, which

was to simply replace the arithmetic mean in Eq. (1) with the median. However, as with all inferential techniques based on the median, this approach suffers from the sampling distribution of the median is intractable, and the fact that sample median is not the uniformly, best unbiased estimate of the population median (Shulkin & Sawilowsky, 2009). It is based on replacing the mean with M-estimators (maximum-likelihood), such as the Huber (1981) or Tukey (1960) estimator. (For other approaches, such as the Winsorized regression, see Wilcox, 1996, p. 324). Resistant regression can be conducted in R via the `rlm()` subroutine.

Least-trimmed squares regression

Rousseuw and Leroy (1987) suggested the least trimmed squares regression as an improvement over resistant regression. It is more resistant to outliers (Verzani, 2005, p. 100), because as opposed to accommodating outliers, it eliminates them. Ripley (2004) noted that least trimmed *squares* is based on minimizing “the sum of squares for the smallest q of the residuals,” where q takes on various values (e.g., $S+$ and R sets q to 90% as the default). The result is a regression model that maximizes accuracy to the $q\%$ of data. The quantile squared residual...[with] $\text{floor}((n+p+1)/2)$ ” (Ripley, n.d.), where n are data points and p are the regressors. `lqs()` is exact with one regressor. For further details, see Fox (2002).

However, least trimmed regression is ill equipped to recover in the null case (i.e., no outliers when ordinary regression should have been used), because once data are trimmed, they are removed from further calculations whether they should have been eliminated or not. Least trimmed squares can be conducted in R via the `lqs` subroutine in the MASS library.

Monte Carlo

Monte Carlo Simulation describes the “use of a computer program to simulate some aspect of reality, and making determinations of the nature of reality or change in reality through the repeated sampling” (Sawilowsky & Fahoome, 2003, p. 46), and was first used on the Manhattan Project during World War II to simulate nuclear fusion

(Spence, 1983). Although opinions differ to who invented of the computerized use of Monte Carlo simulation, Spence (1983) credited S. Ulam, while Sawilowsky & Fahoome (2003) credited Jerzy Newman. However, both agreed that Gosset (Student, 1908) used similar techniques. Sawilowsky & Fahoome (2003) defined Monte Carlo as:

Repeated sampling from a probability distribution to determine the long run average of some parameter or characteristic. Sampling is usually done with replacement, meaning that a subset of scores are obtained, they are analyzed, the results are recorded, and the scores are returned to the reservoir of data values. On the next iteration, the values just examined have the same probability of being selected as the values not yet examined (p. 46).

Chapter 3

Methodology

In order to explicate the impact of data decay on robust regression, a Monte Carlo simulation will be conducted via the open source R programming environment. The study will be conducted on a 2.2 Ghz AMD Athlon II P340 dual core computer. The number of iterations per experiment will be set to 10,000 due to the speed obstacle of R, which is an interpretive (as opposed to compiled) computing platform. This number of repetitions is sufficient to produce Type I error accuracy to four digits (Robey & Barcikowski, 1992).

Design

A simple linear regression layout will be used, defined as

$$Y' = a + \beta X_1, \quad (1)$$

where X_1 represents the initial location of an asset, and Y' represents the final location of an asset.

Sampling Plan

Data will be obtained from R's pseudo-random number command (i.e, rnorm), using the built-in Marsaglia-Multicarry or Mersenne-Twister algorithms. The seed subcommand will be left to the default to initialize the pseudo-random number generator. Sample sizes will be set to $n = 30, 90, 120, 240,$ and 480. The original data will then be subjected to the three regression methods, which will yield the Type I and II error rates.

Type I Error Model Definitions

Model 1. In this model, data will originate from a Gaussian distribution. They will then be subjected to Type I right censoring, which means that data are unavailable

after some point in time, either by predetermination or occurrence of some event. Right censoring means data points are known to exist beyond a specific point, but they are unavailable for some reason. This will be accomplished via truncation commands, set at 25% of censoring. (Due to symmetry, there is no need to also model left censoring in this study.)

Model 2. In this model, data will be obtained from the Gaussian distribution, and then subjected to a systematic amount of censoring from the middle of the data set. (In the statistical literature, only right and left censoring is defined.) This will be accomplished by removing the middle 25% of the data.

For example, consider the sample size of $n_1 = n_2 = 30$. First, pseudo-random numbers will be obtained via R's `rnorm` command, with thirty observations placed in each of two arrays (or vectors as they are referred to in R). Using the `matrix` command, the two arrays will be joined into a two-dimensional array (or two columns as they are referred to in R). Then, the first column will be sorted from low to high, keeping the original observation in the second column as its coordinate pair. Then, the final x and y arrays will be created by selecting the paired x and y scores numbered 1-11 and 20-30.

The result will retain $\frac{11+11}{30} = \frac{22}{30}$ or 73.33% of the original scores. This is as close to a 25% censoring that can be obtained with the given sample size. Also, note that if (a) the middle 25% of the original paired data were censored it would be tantamount to reducing the sample size, and (2) if both x and y were sorted and then the middle 25% of the original data were censored then r would be 1.0 instead of 0.

Model 3. In this model, systematically arbitrarily missing data will be obtained, which is distinguished from randomly missing data in that the latter may not represent any particular pattern of missingness, but nevertheless, occurs due to some random process. This will be accomplished by deleting every other value.

Model 4. This model will be a combination of the previous three models. The data will be subjected to center, right, and systematic censoring, in that order.

Type II Error (comparative power) Model Definitions

Model 5. In this model, correlated data will be created via:

$$y = rx + z\sqrt{1-r^2} \quad (2) \quad .$$

As a result, the descriptive statistics of correlated data will not be maintained between X_1 and Y values. (In order to preserve the mean, variance, skewness, kurtosis, and higher moments, methods such as the Fleishman power method, see e.g. Headrick & Sawilowsky, 1999, or the Gibbs sampler, see e.g. Casella & George, 1992, must be used.) Censuring and missing data will be created via R commands.

Model 6. In this model, the X_1 values will be obtained from a 4th generation correlation produced from Eq. (2). In other words, Y will first be used to produce X_a , X_a will then be used to produce X_b , X_b will then be used to produce X_c , and finally, X_c will be used to produce X_1 which will be used in the Monte Carlo simulation. By repeatedly invoking Eq. (2), the descriptive statistics will accrue additional degeneration with each iteration, while maintaining the post correlation to the set values of $r = 0.1$ (.2).

Analysis

In simple linear regression Eq. (2), β is a weighting coefficient. It is tested via a $N-2$ *df* t test. The significance of t will be evaluated at the nominal $\alpha = 0.05, 0.01$, and 0.001 levels, using the ordinary least squares regression technique in R called `lm()`. Then, the same data will be submitted to the R's `lqs()` and `rlm()` robust subroutines from the MASS library, in order to conduct the least trimmed-squares and the resistant regression.

The `lqs ()` and `rlm ()` routines produce the Y intercept, beta, and other summary statistics. However, neither produces the p value associated with beta, which are required to compare with the results from `lm()`. The `lqs()` provides beta, and its test of significance will be discussed below. The `rlm()` provides the t test on beta, which will be evaluated with the appropriate degrees of freedom to produce the associated p value.

Also, because the maximum number of iterations to resolution in `rlm()` is “maxit = 20”, it will be increased to “maxit = 1000” to help ensure the method resolves and to avoid warning messages.

Standard error of beta and the `lqs()` method.

The t test is defined as beta divided by the standard error of beta (Brase & Brase, 2013, p. 536; Mann, 1995, p. 667), which is then associated with the $df = N - 2$ for the t (or Z for large samples) distribution. It is generally not optimal to use the normal theory formula for the standard error (i.e., the standard deviation divided by the sample size) because it is not robust to non-normally distributed data (including decayed data). There are potential alternatives, such as the winsorized sample standard deviation, or a jackknife or bootstrap approximation (see, e.g., Sawilowsky & Fahoome, 2003, p. 22, 376 - 382). However, there are many limitations to those alternatives.

Wilcox (1996) provided alternatives in computing the standard error for other hypothesis tests (e.g., the sample median), but that was only after he presented a test using the robust estimator in the numerator combined with the normal curve theory standard error in the denominator (see, e.g., p. 120). The same approach will be used here, with the p value associated with beta obtained from `lqs()` determined via the normal curve theory standard error (i.e., which is produced by the `lm()` routine).

Tabulation of Results

A template for the tabulation of Model 1 results to be compiled and presented in Chapter 4 appears below:

Table X. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	lm()	Test	
		lqs()	rlm()
0.05			
0.01			
0.001			

Similar tables will be produced for the other combinations of the sampling plan.

Chapter 4

Results

Initially the research protocol called for 10,000 repetitions per experiment. However, that number was too small to produce accurate Type I error results. The most likely culprit is R's `rnorm` pseudo-random number generator algorithm. Similarly, Eq. 2 failed to produce sufficiently precise correlations for small sample sizes (e.g., $n_1 = n_2 = 30$). Therefore, the number of repetitions per experiment was increased to 100,000. In addition, the study was moved to an Intel Sandy Bridge i7-2600K 3.4GHz CPU-based computer, with ultra-high speed Corsair Vengeance Low Profile 4x4GB RAM, Crucial M4 256GB solid state hard drive, and the Windows 7 Ultimate 64 bit operating system. Nevertheless, confirming the well known lack of speed of the R platform, the results compiled in each table in this chapter took more than 45 minutes to complete.

Using the standard error under $\text{lm}(y \sim x)$ (i.e., beta associated with the ordinary least squares regression) as the denominator for the test of beta obtained from `lqs()` was unsatisfactory, with inflated Type I errors from between 7.3 and 104 times nominal alpha, as noted in Table 1 below:

Table 1. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	<code>lm()</code>	<code>lqs()</code>
0.05	0.04972	0.36455
0.01	0.01041	0.21966
0.001	0.00102	0.10248

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

The next attempt was to improve the standard error used in lqs() by replacing the original y values with the fitted values of y obtained from lqs(). In other words, the standard error of the estimate (SE_E , or residual standard deviation) was based on

$$SE_E = \sqrt{\sum_i^n \frac{(y - y')^2}{n - 2}}, \quad (3)$$

where y' was obtained as fitted values from lqs() instead of the fitted values from lm().

The standard error of beta (SE_b) is determined by

$$SE_b = \frac{SE_E}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (4)$$

Assembling the t test on beta as a ratio of beta divided by Eq. 4,

$$t = \frac{b}{SE_b}, \quad (5)$$

the obtained t is significant if

$$|t_{obt}| \geq t_{\frac{\alpha}{2}, n-2}.$$

Although, as noted in Table 2 there was improvement in the Type I error rates, the inflation from between 5.8 and 39.4 times nominal alpha is not acceptable. (Note the values for lm() differed slightly from those in Table 1 above due to the change in the seed number.)

Table 2. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	lm()	Test	lqs()
0.05	0.05029		0.29371
0.01	0.01061		0.14499
0.001	0.00109		0.04151

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Regarding the least median squares (lms) option (i.e., "method = lms" option in `lqs()`, which can be used to invoke a variety of robust methods), subsequent to a Monte Carlo simulation Paranagama (2010) concluded, "In practice, the use of LMS is limited by the absence of formulas for standard errors" (p. 35). This difficulty applies to the default method (least trimmed squares), and hence, `lqs()` must be abandoned if the purpose of conducting the linear model is to compute a t test on beta until an adequate standard error for the least squares regression algorithm can be found. Therefore, results in the balance of this chapter will be restricted to `lm()` and `rlm()`.

Original Data Results

The Type I error results are compiled in Table 3 below. Note that the `lm()` are slightly more accurate than the `rlm()`, but as expected both techniques produce the correct Type I error rate for data obtained from the Gaussian distribution. As the sample size increased from $n_1 = n_2 = 30$ to $n_1 = n_2 = 480$, both procedures' Type I errors converged with nominal alpha, as dictated by theory.

Table 3. Original Data; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04968	0.05377
0.01	0.00955	<i>0.01186</i>
0.001	0.00082	<i>0.00144</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 4. Original Data; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04988	0.05152
0.01	0.01021	<i>0.01095</i>
0.001	0.00103	<i>0.00120</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 5. Original Data; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05057	0.05119
0.01	0.01042	<i>0.01101</i>
0.001	0.00103	<i>0.00113</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 6. Original Data; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04979	0.05004
0.01	0.00977	0.00993
0.001	0.00098	<i>0.00115</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 7. Original Data; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05002	0.05116
0.01	0.01023	0.01051
0.001	0.00093	<i>0.00112</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Robustness with Respect to Type I and II Error

The balance of the tables in Chapter 4 pertains to the two error conditions, and is presented without further comment. Type I error rates in this section are based on decay models 1 through 4. Type II error rates (or comparative power), as represented by decay and correlated models 5 and section, are compiled in the following section.

Type I Errors

Model 1 (Type I Right Censuring) Results

Table 8. Model 1; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04979	<i>0.05501</i>
0.01	0.01002	<i>0.01305</i>
0.001	0.00000	<i>0.00167</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 9. Model 1; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04979	0.05162
0.01	0.00994	0.01092
0.001	0.00109	<i>0.00127</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 10. Model 1; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05008	0.05132
0.01	0.00986	0.01057
0.001	0.00098	<i>0.00115</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 11. Model 1; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05058	0.05104
0.01	0.00963	0.01012
0.001	0.00108	0.00104

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 12. Model 1; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05005	0.04992
0.01	0.00986	0.01012
0.001	0.00105	<i>0.00114</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Model 2 (Middle Censuring) Results

Table 13. Model 2; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05087	<i>0.05627</i>
0.01	0.01017	<i>0.01334</i>
0.001	0.00089	0.00189

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 14. Model 2; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05020	0.05168
0.01	0.00923	0.01043
0.001	0.00086	0.00100

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 15. Model 2; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05091	0.05164
0.01	0.00980	0.01045
0.001	0.00000	0.00096

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 16. Model 2; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05100	0.05048
0.01	0.01030	0.01046
0.001	0.00000	0.00098

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 17. Model 2; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04992	0.05007
0.01	0.01020	0.01034
0.001	0.00000	0.00104

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Model 3 (Systematic Censuring) Results

Table 18. Model 3; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05018	<i>0.05959</i>
0.01	0.01030	0.01520
0.001	0.00092	0.00239

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 19. Model 3; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04947	0.05246
0.01	0.00988	<i>0.01112</i>
0.001	0.00101	<i>0.00120</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness

Table 20. Model 3; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05032	0.05214
0.01	0.01026	<i>0.01121</i>
0.001	0.00103	<i>0.00123</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 21. Model 3; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05009	0.05154
0.01	0.00960	0.01008
0.001	0.00078	0.00101

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 22. Model 3; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04966	0.05049
0.01	0.01002	0.01021
0.001	0.00097	0.00098

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Model 4 (Middle, Type I Right, and Systematic Censuring) Results

Table 23. Model 4; Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04953	<i>0.07043</i>
0.01	0.01064	0.02168
0.001	0.01064	0.00465

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 24. Model 4; Type I error rates for $n_1 = n_2 = 90$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04935	0.05400
0.01	0.01020	<i>0.01208</i>
0.001	0.00086	0.00152

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 25. Model 4; Type I error rates for $n_1 = n_2 = 120$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05004	0.05356
0.01	0.01020	<i>0.01202</i>
0.001	0.00108	0.00153

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 26. Model 4; Type I error rates for $n_1 = n_2 = 240$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.04983	0.05128
0.01	0.01035	0.01075
0.001	0.00107	<i>0.00121</i>

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Table 27. Model 4; Type I error rates for $n_1 = n_2 = 480$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.05017	0.05081
0.01	0.01049	0.01038
0.001	0.00122	0.00123

Note: Values in italics exceed Bradley's (1978) conservative definition of robustness. Values in bold exceed Bradley's (1978) liberal definition of robustness.

Type II Errors

Model 5 (Correlated) Results

This series of tables pertains to the situation where the data are sampled from the Gaussian distribution, but the X and Y data are correlation from $.2 - 1$ (.2), meaning from .2 to 1.0 in increments of .2. This model represents the impact of an intervention or treatment, meaning the tabled values are power results. Because the referent distribution is Gaussian, the comparison of the values for `rlm()` with `lm()` are in indication of the former's robustness with respect to Type II errors, meaning because the tabled entries are very similar, the beta error properties of `rlm()` are nearly as good as the ordinary least squares regression.

Alternately, the two procedures are equally powerful under this study condition. Note that as the correlation increases, the statistical power increases. At a certain point, the combination of sample size and magnitude of correlation produces the maximum rejection rate of 1.0.

Table 28. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.2$; 100,000 repetitions

α	Test	
	<code>lm()</code>	<code>rlm()</code>
0.05	0.18334	0.18141
0.01	0.06103	0.06339
0.001	0.01087	0.01259

Table 29. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.47783	0.46028
0.01	0.24586	0.23212
0.001	0.07715	0.07378

Table 30. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.59674	0.57581
0.01	0.35234	0.33257
0.001	0.13455	0.12437

Table 31. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.87995	0.86293
0.01	0.70785	0.68081
0.001	0.43245	0.40105

Table 32. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.99323	0.99081
0.01	0.96784	0.95940
0.001	0.87289	0.84744

Table 33. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.4$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.60984	0.59052
0.01	0.35693	0.34225
0.001	0.12669	0.12552

Table 34. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.4$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.97864	0.97207
0.01	0.91943	0.90201
0.001	0.74592	0.71245

Table 35. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.4$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.99581	0.99456
0.01	0.97876	0.97223
0.001	0.90138	0.87971

Table 36. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.4$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	0.99998
0.01	0.99994	0.99991
0.001	0.99950	0.99912

Table 37. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.4$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 38. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.6$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.95553	0.94610
0.01	0.85578	0.83263
0.001	0.61467	0.58465

Table 39. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.6$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.99998	0.99999
0.01	0.99996	0.99994
0.001	0.99923	0.99858

Table 40. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.6$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	0.99999
0.001	0.99998	0.99997

Table 41. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 42. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 43. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 0.8$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.99992	0.99983
0.01	0.99921	0.99871
0.001	0.99222	0.98810

Table 44. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 0.8$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 45. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 46. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 47. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 48. Model 5; Rejection rates for $n_1 = n_2 = 30$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 49. Model 5; Rejection rates for $n_1 = n_2 = 90$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 50. Model 5; Rejection rates for $n_1 = n_2 = 120$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 51. Model 5; Rejection rates for $n_1 = n_2 = 240$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 52. Model 5; Rejection rates for $n_1 = n_2 = 480$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Model 6 (4th Generation Correlation) Results

Table 53. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	0.98124	0.97546
0.01	0.92540	0.90968
0.001	0.75067	0.71809

Table 54. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	0.99997	0.99992

Table 55. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 56. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.2$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 57. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.2$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 58. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.4$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	0.99999
0.01	0.99999	0.99999
0.001	0.99995	0.99991

Table 59. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.4$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 60. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.4$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 61. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.4$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 62. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.4$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 63. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 64. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 65. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 66. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 67. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.6$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 68. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 0.8$;
100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 69. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 70. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 71. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 72. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 0.8$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 73. Model 6; Rejection rates for $n_1 = n_2 = 30$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 74. Model 6; Rejection rates for $n_1 = n_2 = 90$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 75. Model 6; Rejection rates for $n_1 = n_2 = 120$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 76. Model 6; Rejection rates for $n_1 = n_2 = 240$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Table 77. Model 6; Rejection rates for $n_1 = n_2 = 480$; $r = 1$; 100,000 repetitions

α	Test	
	lm()	rlm()
0.05	1	1
0.01	1	1
0.001	1	1

Chapter 5

Conclusion

Terrorist attacks such as those on *USS Cole* (2000), *Twin-Towers* on 9/11 (2001), *MV Limburg* (2002), *Super Ferry 14* (2004), *Madrid train bombing* (2004), *London tube bombing* (2005), and the *Boston marathon bombing* (2013) with their devastating loss of life, together with the new forms of asymmetric warfare displayed by these groups, demonstrated the need for 1) better cooperation between security services, and 2) the need to change or adapt data collection sources, patterns, and analytic methods. Information is collected from various open and covert sources, analyzed for important or predictive markers, and then acted on by the appropriate security services.

According to Bamford (2012), automatic data interception and processing rates have quadrupled to a rate of over 20 terabytes per minute, which required NSA to build a new digital storage and processing facility in Utah with more than a million square feet of digital storage and a cost of US \$2 Billion. It is not surprising therefore, that according to Monster.com the Homeland Security Industry is one of the fastest growing governmental sectors today, with an overall increase of 311% jobs since 2001.

The quality of this massive warehoused Homeland Security data is often subject to unspecified types of decay. As a result, it is not known how traditional statistical methods, such as ordinary least squares regression (conducted via R's `lm()` procedure), will fare. To begin the experimental process of assessing standard quantitative

methods, data sampled from a normal distribution should be subjected to various models of decay, and if and only if the normal theory statistic performs appropriately would it then be appropriate to consider decayed model originating from non-normally distributed data. (In other words, if [linear] regression cannot survive decayed data originating from normality, then there is no point if further complicating the study with nonnormality. For comparison purposes, a plethora of robust methods, such as least trimmed squares (via R's `lqs()` procedure) and maximum likelihood regression (via R's `rlm()` procedure), have been developed, which hold promise to provide correct statistical properties even when linear regression fails.

It is very important to be able to predict the location of assets, the movements of material, or even the likelihood of certain targets being compromised. Similar abilities pervade related disciplines, such as the prediction of man-made disasters. For example, the safety of highway-rail grade crossings has been the subject of study for many years. Prediction models abound in an attempt to reduce the likelihood of accidents between trains and highway vehicles (Oha, Washington, & Doohee, 2006; Schoppert & Hoyt, 1967).

In contradistinction to the relatively tame transportation data, the purpose of this dissertation was to begin the process of determining how ordinary least squares regression performs in the presence of massively decayed data presumed to applicable to Homeland Security, and to begin answering the question if `lqs()` and `rlm()` provide any advantages. This was accomplished by using R's `rnorm` (normal) pseudo-random number generator, after which data were subjected to various models of decay. Then,

the three regression methods `lm()`, `lqs()`, and `rlm()` were applied; sample sizes set to $n_1 = n_2 = 30, 90, 120, 240, \text{ and } 480$; and nominal α was set to 0.05, 0.01, and 0.001.

Four models of data decay were simulated in order to determine its impact on Type I errors:

- Model 1: 25% Type I Right Censuring
- Model 2: 25% Center Censuring
- Model 3: 25% Systematic Censuring
- Model 4: 25% each of Center, Type I Right, and Systematic Censuring

Two models of data decay were simulated in order to determine its impact on Type II errors (and comparative power):

- Model 5: First generation correlated data, with $r = .2 - 1 (.2)$
- Model 6: Fourth generation correlated data, with $r = .2 - 1 (.2)$.

Initially, the study protocol called for repeating each Monte Carlo experiment 10,000 times, conducting `lm()`, `lqs()`, and `rlm()`, and testing β via a $N-2$ *df* *t* test. Immediately, however, two issues arose. First, the number of repetitions, chosen in consideration of R's lack of speed, was insufficient to produce sufficiently precise Type I errors with non-decayed data obtained from R's `rnorm` procedure. Hence, the decision was made to increase the number of repetitions of each experiment to 100,000. As a result, the decision was made to migrate the study from a 2.2 Ghz AMD Athlon II P340 dual core machine to a 3.4 Ghz Intel Sandy Bridge i7-2600K CPU-based computer with ultra-high

speed RAM. Nevertheless, the approximate time necessary to produce each table in Chapter 4 was 45 minutes.

Second, it was immediately discovered that an appropriate standard error has not been derived for the `lqs()` method. Because the t test on β requires the standard error, various options were considered: (1) the p value associated with β obtained from `lqs()` was determined via the normal curve theory standard error via the `lm()` procedure, which failed because it produced Type I errors as large as 104 times nominal α , and (2) the standard error was obtained by replacing the original y values with the fitted values of y obtained from `lqs()`, which was an improvement, but also failed because it produced Type I errors as large as 39.4 times nominal α .

Because of this failure, even though the various resources cited in Chapter 2 use it to produce pretty regression equations, `lqs()` was omitted from further consideration in this study. Although the ability of this method to create a regression line that visually fits data with decay better than `lm()`, that feature is immaterial because the method is dangerous in not being even close to robust with respect to Type I errors. The lack of a robust test of beta in `lqs()` regression will become increasingly serious as applied researchers continue to be attracted to its highly publicized robustness regression lines and implement it into their applied work. For example, `lqs()` was used by Fan, Lub, Madnickc, and Cheungd (2001) in a study on data integration in information systems, Abo-Khalil and Abo-Zied (2012) in a study of sensorless control of wind turbines, and Gidnaa and Domínguez-Rodrigo (2013) in a study of human femoral length from fragmented specimens.

Discussion

Type I Errors

Normally Distributed Data

For the reasons outlined above, the discussion of `lqs()` will be omitted, because in the absence of an appropriate standard error, it produced wildly liberal Type I error rates. For the original data, regardless of the sample size, both `lm()` and `rlm()` produced correct Type I errors. R's `rlm()` produced trivially larger Type I error rates, but the results were well within Bradley's (1978) conservative definition (i.e., $\pm 1\alpha$) of robustness.

Decayed Data

There was no impact on `lm()` when the data were subjected to 25% Model 1 (right censoring) decay. However, for the smallest sample size ($n_1 = n_2 = 30$), `rlm()` produced Type I error rates that exceeded Bradley (1978) conservative definition of robustness. For example, with $\alpha = 0.05$, the Type I error rate rose to 0.055. With $\alpha = 0.001$, the Type I error rate rose to 0.0017, exceeding Bradley's (1978) liberal standard. However, with sample sizes of $n_1 = n_2 = 90$ or larger, `rlm()` produced correct Type I errors. Hence, `rlm()` should not be used with right censoring for small sample sizes.

The same pattern of Type I errors was repeated with Model 2 (center censoring) decay, except the inflations were slightly larger for `rlm()` (e.g., 0.0596 and .0024, respectively), and with Model 3 (systematic censoring) (e. g., 0.05959 and 0.00239, respectively). In the presence of massive decay as represented by Model 4 (center,

right, and systematic), $rlm()$ nearly exceeded Bradley's liberal definition (i.e., $\pm 5\alpha$) of robustness for the largest alpha levels and exceeded that standard for smaller alpha levels. For sample size of $n_1 = n_2 = 30$, with $\alpha = 0.05$, the Type I error rate rose to 0.0706, and with $\alpha = 0.001$ the Type I error rate rose to 0.00465. With sample size $n_1 = n_2 = 90$ and $\alpha = 0.05$, $rlm()$'s Type I error rate was barely inside Bradley's conservative definition (0.054), but with $\alpha = 0.01$, its Type I error rate excluded the liberal standard (0.00152). The Type I error rates displayed the same pattern for the remaining pattern for the remaining sample sizes. Hence, in the presence of massive decay as modeled by a combination of various types of censoring, $rlm()$ should not be used with extremely small nominal alpha levels if the intent to meet Bradley's (1978) conservative definition of robustness.

Type II Errors (and Comparative Power)

In order to examine Type II error properties of $lm()$ and $rlm()$, the data were subjected to two models of correlation: first and fourth generation of correlation, with the magnitudes of correlation spanning from $r = .2$ to 1 in increments of $.2$. As the correlation increases between X and Y in linear regression, the statistical significance of β increases. (Note that if Y represents group membership, such as belong to one terrorist group vs. another, is regressed on X , this is known as dummy coded regression and is equivalent to the ordinary independent samples t test.)

The first step in a comparative Type II error study is to determine if the competing statistics reject at the same rate for a given treatment alternative. The

second step is to determine the Type II error studies proceed to determine how the competing tests perform (i. e., if they obtain the same rejection rate if the data are sampled instead from a nonnormal distribution). Comparative power studies, a third step, are more comprehensive in modeling the treatment alternative throughout the power spectrum. However, the current study was restricted to first step, which was to sample data from a normal distribution and then apply the various models of decay to it. The treatment alternative was restricted to modeling various levels and types of correlated data.

Regardless of the model (first or fourth generation of correlated data), sample size, or magnitude of correlation, the Type II error results (i.e., rejection rates) were nearly equivalent for $lm()$ and $rlm()$. This is somewhat surprising for sample size of $n_1 = n_2 = 30$, because $rlm()$'s inflated Type I error should have given it a slight, albeit inappropriate, advantage. Overall, though, these two procedures performed nearly identically.

Conclusion

In conclusion, the least trimmed squares (R's $lqs()$ procedure) should be avoided, despite the pretty regression lines it produces, until such time that an appropriate standard error can be developed. In terms of the Type I error performance of ordinary least squares regression (via R's $lm()$ procedure) and maximum likelihood regression (R's $rlm()$ procedure), when data are massively decayed as modeled by various types of censoring, $rlm()$ should be avoided with sample sizes as small as $n =$

30 per group. In terms of Type II errors, however, the two procedures perform nearly identically. Interestingly, although it is known that the ordinary least squares ($lm()$) regression can be impacted by non-normality and other assumption violations, it is remarkable robust to normally distributed data that is subject to massive decay.

Implications for Further Research

R's $lqs()$ might become a suitable substitute for $lm()$ if further work on finding a better standard error is successful. As noted above, suggestions have been made to use a jackknife or bootstrap approach. However, those techniques are computationally intensive, add a layer of error because there are estimates, and would only be appropriate for the data at hand. Obviously, at such time that the statistical literature is settled on a robust standard error to use with $lqs()$ this study should be replicated using it to determine the p value associated with beta.)

It was concluded that the Type I error rate for $rlm()$ was unacceptable for a sample size of 30 per group. If Bradley's (1978) liberal definition of robustness is acceptable, $rlm()$ is useful when the sample size reaches 90 per group for the larger alpha level of 0.05. Based on the study parameters, however, the precise point after 30 and before 90 per group when $rlm()$ is acceptable is not known, which would require additional study.

Because $rlm()$ does not perform in an acceptable fashion when data are sampled from a normal distribution, it is pointless to continue promoting this method when data are obtained from nonnormal data. In terms of man-made disasters (i.e., highway-rail

accidents), Lord and Mannering (2010) considered sixteen non-linear prediction approaches: Poisson, Negative binomial/Poisson-gamma, Poisson-lognormal, Zero-inflated Poisson and negative binomial, Conway–Maxwell–Poisson, Gamma, Generalized estimating equation, Generalized additive, Random-effects, Negative multinomial, Random-parameters Bivariate/multivariate, Finite mixture/Markov switching, Duration, Hierarchical/multilevel, Neural network, and Bayesian neural network and support vector machine. Further study using those approaches may prove beneficial in the presence of the massively decayed data that is presumed to be present in Homeland Security data.

REFERENCES

Abo-Khalil, A. G., & Abo-Zied, H. (2012). Sensorless control for DFIG wind turbines based on support vector regression, *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, p. 3475 - 3480. 10.1109/IECON.2012.6389341

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rodgers, W., & Tukey, J. (1972). *Robust estimates of location: Survey and advances*. Princeton, NJ: Princeton University Press.

Barkan, S. E., & Snowden, L. (2001). *Collective violence*. Boston: Allyn & Bacon.

Bartol, C. & Bartol, A. (2011). *Criminal Behavior: A Psychological Approach*. NY: Prentice Hall.

Biagi, S. (2011). *Media Impact: An Introduction to Mass Media* (10th ed.). Belmont, USA: Wadsworth.

Bockstette, C. (2008). *Jihadist Terrorist Use of Strategic Communication Management Techniques*. Gorge C. Marshall European Center for Security Studies. Retrieved from http://www.marshallcenter.org/mcpublicweb/MCDocs/files/College/F_Publications/occPapers/occ-paper_20-en.pdf.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.

Borum, R. (2011). Understanding terrorist psychology. In Andrew Silke (Ed.), *The Psychology of Counter-Terrorism*. London, UK: Routledge.

Borum, R. (2003). Understanding the terrorist mind-set. *FBI Law Enforcement Bulletin*. 72 (July), 7-10.

Bowie, V. (2005). Organizational Violence: a trigger for reactive terrorism. In Bowie, Fischer, B., & Cooper, C. (Eds.). *Workplace Violence: Issues, trends, strategies*. Portland: William Publishing.

Brase, C. H., & Brase, C. P. (2013). *Understanding basic statistics*. 6th ed. Boston, MA: Brooks/Cole, Cengage Learning.

Brown, M. L. (1982). Robust line estimation with errors in both variables. *Journal of the American Statistical Association*, 77(377), 71-79.

Buckley, J. & James, I. (1979). Linear Regression with Censored Data. *Biometrika*, 66(3), 429-436.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, 46(3), p. 167-174.

Central Intelligence Agency (CIA). (2013). *The Intelligence Cycle*. Retrieved from <https://www.cia.gov/kids-page/6-12th-grade/who-we-are-what-we-do/the-intelligence-cycle.html>.

Central Intelligence Agency (CIA) (n.d.). *Intelligence: Human Intelligence*. Retrieved from <https://www.cia.gov/news-information/featured-story-archive/2010-featured-story-archive/intelligence-human-intelligence.html>.

Chen, H., Reid, E., Sinai, J., Sike, A., & Ganor, B. (2000). *Terrorism informatics: Knowledge management and data mining for homeland security*. NY, NY: Springer.

Cook, A. (2008). Censoring and Truncation, *Introduction to Survival Analysis* (ST3242). National University of Singapore.

Crenshaw, M. (1981). The causes of terrorism. *Comparative Politics*, 13(4), 379-399.

Cronin, A. K. (2003). Behind the curve: globalization and international terrorism. *International Security*, 27(3), 30-58.

Defense Technical Information Center - DTIC (n.d.). *Force Multiplier*. Retrieved from http://www.dtic.mil/doctrine/dod_dictionary/data/f/8037.html.

Deitch, I. (n.d.). Israeli intelligence issues Facebook warning. *Associated Press*. Retrieved on from <http://abcnews.go.com/Technology/story?id=7621098>.

Enders, W., & Sandler, T. (2012). *The political economy of terrorism* (2nd ed.). Cambridge: Cambridge University Press.

Fan, W, Lub, H., S. E. Madnickc, & Cheungd, D. (2001). Discovering and reconciling value conflicts for numerical data integration. *Information Systems*, 26, p. 635–656

Farber, M., Cameron, M, Ellis, C., and Sullivan, J. (n.d.). *Massive Data Analytics and the Cloud. A Revolution in Intelligence Analysis*. Booz, Allen, Hamilton. Retrieved from <http://www.boozallen.com/media/file/MassiveData.pdf>.

Federal Bureau of Investigation. (2006). (U//FOUO) The Radicalization Process: From Conversion to Jihad. *Intelligence Assessment*. Retrieved from <http://cryptome.sabotage.org/fbi-jihad.pdf>.

Federal Bureau of Investigation (n.d.). *Intelligence Cycle*. Retrieved from <http://www.fbi.gov/about-us/intelligence/intelligence-cycle>.

Fiosina, J. (2012). Decentralized Regression Model for Intelligent Forecasting in Multi-agent Traffic Networks. In Omatu, S. De Paz Santana, J, Molina, J.M., Bernardos, A.M., Corchado, J.M. (Eds.). *Distributed Computing and Artificial Intelligence. Advances in Intelligent and Soft Computing*. Springer Berlin Heidelberg, 151, 255-263.

Fisher, R. A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Fox, J. (2002). Robust Regression. *Appendix to An R and S-Plus Companion to Applied Regression*. R-Project.org. Retrieved from <http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-robust-regression.pdf>.

Ganor, B. (1997). *Countering State-Sponsored Terrorism*. Herzliya: International Policy Institute for Counter-Terrorism.

Ganor, B. (2005). *The Counter-Terrorism Puzzle*. New Brunswick: Transaction Publishers.

Ganor, B. (2008). Terrorist organization typologies and the probability of a boomerang effect. *Studies in Conflict & Terrorism*, 31(4), 269 – 283.

Gidnaa, A. O., & Domínguez-Rodrigo, M. (2013). A method for reconstructing human femoral length from fragmented shaft specimens. *HOMO - Journal of Comparative Human Biology*, 64(1), 29–41

Gilboa, A. (2012). *Israel's silent defender: an inside look at sixty years of Israeli intelligence*. Jerusalem: Gefen.

Grace, T. A., & Sawilowsky, S. S. (2009). Data error prevention and cleansing: A comprehensive guide for instructors of statistics and their students. *Model Assisted Statistics and Applications*, 4, 303-312.

Gurr, T. P. (1989): Political terrorism: historical antecedents and contemporary trends. In Guer, T.R. (ED.), *Violence in America*. Volume 2: Protest, Rebellion, Reform. Newbury Park, CA: Sage Publications.

Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. 6th ed. Upper Saddle River, NJ: Pearson/Prentice Hall.

Hand, D. (2000). *Data Mining: New Challenges for Statisticians*. Social Science Computer Review 18, 442

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method, *Psychometrika*, 64(2), p. 25-35.

Heitjan, D.F. & Basu, S. (1996). Distinguishing "Missing at Random" and "Missing Completely at Random." *The American Statistician*, 50(3), 207-213.

Hoffman, B. (1998). *Inside Terrorism*. NY: Columbia University Press.

Huber, P. J. (1981). *Robust statistics*. NY: Wiley.

Ibrahim, J.G., Chen, M., Lipsitz, S.R., & Herring, M. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469), 332-346.

Jackson, B. A. (2006). Groups, Networks, or Movements: A Command-and-Control-Driven Approach to Classifying Terrorist Organizations and Its Application to Al Qaeda. *Studies in Conflict & Terrorism*, 29(3), 241-262.

Jayasekara, S. (2007). LTTE fundraising & money transfer operations. *APG*. Retrieved from http://www.apgml.org/frameworks/docs/7/LTTE%20Fundraising%20%26%20Money%20Transfer_Oct07-Jayasekara.pdf.

Khan, M. M., Goldney, R., & Hassan, R. (2010). Homicide bombers: Life as a weapon. *Asian Journal of Social Science*, 38(3), 481-484.

Kohlmann, E. (2008). "Homegrown" terrorists: Theory and cases in the war on terror's newest front. *Annals of the American Academy of Political and Social Science*, 618, 95-109.

Laqueur, W. (1999). *The new terrorism, fanaticism and the arms of mass destruction*. NY: Oxford University Press.

Levin, J., & Fox, J. A. (2000). *Elementary statistics in social research*. 8th ed. Boston: Allyn & Bacon.

Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. NY: Wiley.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research, A* (44), 291-305.

Lorenz, A. J. (2007). *Al Qaeda's Maritime Threat*, Inter-American Committee against Terrorism (CICTE), http://www.cicte.oas.org/Database_/50637-Al%20Qaeda%27s%20MaritimeThreat.pdf.

Lorenz, A. J. (2011). *Radicalization Pathways*. The Maritime Terrorism Research Center (Archived Document #3), Maritimeterrorism.com.

Love, J. (2010). Hezbollah: Social Services as a Source of Power. *JSOU Report*, 10(5). Retrieved from <http://jsou.socom.mil>.

Magouirk, J., Atran, S., & Sageman, M. (2008). Connecting Terrorist Networks. *Studies in Conflict & Terrorism*, 31(1), 1–16.

Mann, P. S. (1995). *Introductory statistics*. 2nd ed. NY: Wiley.

Martin, G. (2009). *Understanding terrorism, challenges, perspectives, and issues*. (3rd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Mccauley, C. & Moskalenko, S. (2008). Mechanisms of political radicalization: pathways toward terrorism. *Terrorism and Political Violence*, 20(3), 415-433.

McConnell, M. (2007). Overhauling intelligence. *Foreign Affairs*, 86(4), 49-58.

McCormick, G. H. (2003). Terrorist decision making. *Annual Review of Political Science*, 6, 473-507.

Miller, R. & Halpern, J. (1982). Regression with censored data. *Biometrika*, 69(3), 521-531.

Moghadam, A. (2006). Suicide terrorism, occupation, and the globalization of martyrdom: A critique of dying to win. *Studies in Conflict & Terrorism*, 29(8), 707 – 729.

Monahan, T. (2010). The future of security? Surveillance operations at Homeland Security Fusion Centers. *Social Justice*, 37(2/3), 120-121.

Oha, J., Washington, S. P., & Doohee, N. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis and Prevention*, 38, 346-356.

Paranagapam, T. D. (2010). *A simulation study of the robustness of the least median of squares estimator of slope in a regression through the origin model* (Unpublished Master Thesis). Manhattan, KS: Kansas State University. Retrieved from <http://krex.k-state.edu/dspace/bitstream/handle/2097/7045/ThilankaParanagama2010.pdf>

Pape, R. A. (2003). The strategic logic of suicide terrorism. *American Political Science Review*, 97(3).

Pape, R. A. (2006). *Dying to win: the strategic logic of suicide terrorism*. NY: Random House Trade Paperbacks.

Pape, R. A., & Feldman, J. K. (2010). *Cutting the fuse the explosion of global suicide terrorism and how to stop it*. Chicago, IL: University of Chicago Press.

Post, J., Ruby, K., and E. Shaw. (2000). From Car bombs to logic bombs: the growing threat from information terrorism. *Terrorism and Political Violence*, 12(2), 97-122.

Post, J. M. (2008). *The mind of the terrorist: The psychology of terrorism from the IRA to Al-Qaeda*. NY: Palgrave Macmillan.

Rabinowitz, D., Tsiatis, A., & Aragon, J. (1995). Regression with interval-censored data. *Biometrika*, 82(3), 501-513.

Ramasubramanian, R. (2004). Suicide terrorism in Sri Lanka. *IPCS Research Papers, 5*.

Rapoport, D. C. (2004). The four waves of modern terrorism. In Audrey Kurth Cronin & James M. Ludes, (Eds.) *Attacking Terrorism : Elements of a Grand Strategy*. Washington, D.C.: Georgetown University Press.

Richelson, J. (2012). *The US intelligence community* (6th ed.). Boulder, Colo.: Westview Press.

Ripley, B. D. (n.d.). Resistant Regression. Retrieved from <http://astrostatistics.psu.edu/su07/R/html/MASS/html/lqs.html>

Ripley, B. D. (2004). *Robust statistics*. University of Oxford. Retrieved from <http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf>.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283–288.

Rohner, D. and Frey, B. (2007). Blood in INK! The common-interest-game between terrorists and the media. *Public Choice, 133*, 129-145.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. NY: Wiley.

Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.

Sawilowsky, S. S., & Markman, B. S. (1990). Rejoinder to Braver and Walton Braver. *Perceptual And Motor Skills, 71*, 424 – 426.

Schleifer, R. (2006): Psychological operations: A new variation on an age old art: Hezbollah versus Israel. *Studies in Conflict & Terrorism*, 29(1), 1-19.

Schleifer, R. (2009). Psyoping Hezbollah: The Israeli psychological warfare campaign during the 2006 Lebanon war. *Terrorism and Political Violence*, 21(2), 221-238.

Schoppert, D. W., & Hoyt, D. W., 1967; Factors influencing safety at highway-rail grade crossings, *NCHRP Report, Transportation Research Board*, 50, ISSN: 0077-5614.

Seifert, J. (2007). Data mining and homeland security: An overview. *CRS Report for Congress* (Item Code: RL31798).

Shulkin, B., & Sawilowsky, S. S. (2009). Estimating a population median with a small sample. *Model Assisted Statistics and Applications*, 4(2), 143-155.

Shultz, R. (1978). Conceptualizing political terrorism - a typology. *Journal of International Affairs*, 31(1), 5-17.

Silke, A. (2011). The psychology of counter-terrorism: critical issues and challenges. In Andrew Silke (Ed.), *The Psychology of Counter-Terrorism*, London, UK: Routledge.

SITE Intelligence Group (2013). *London Attack Echoes al-Qaeda Incitement to Target Soldiers in West*. Retrieved from <http://news.siteintelgroup.com/component/content/article/5-articles-a-analysis/3095-london-attack-echoes-al-qaeda-incitement-to-target-soldiers-in-west>.

Spence, I. (1983). Monte Carlo simulation studies. *Applied Psychological Measurement*, 7(4), 405-425.

Stohl, M. (2007). *Cyber Terrorism: A clear and present danger, the sum of all fears, breaking point or patriot games?* Center for Information Technology & Society, UC Santa Barbara. Retrieved from <http://blip.tv/cits/cyber-terrorism-a-clear-and-present-danger-the-sum-of-all-fears-breaking-point-or-patriot-games-cits-2007-212375>.

Student. (1908). The probable error of the mean. *Biometrika*, 6(1), 1-25.

The Investigative Project on Terrorism (IPT). (2010). *The Jihadist Social Network Underworld*. Retrieved from <http://www.investigativeproject.org/2398/the-jihadist-social-network-underworld>.

The 9/11 Commission report: final report of the National Commission on Terrorist Attacks upon the United States. (Official ed.). (2010). United States: Published and distributed by SOHO Books, as released by the U.S. Government.

Thornton, T.P. (1964). Terror as a weapon of political agitation. In H. Eckstein (Ed.) *Internal War: Problems and Approaches*, NY: Free Press, 71-99.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Maddow, and H. Mann (Eds.) *Contributions to probability and statistics*, p. 448-485. Stanford, CA: Stanford University Press.

Verzani, J. (2004). *Using R for introductory statistics*. Boca Raton: Chapman & Hall/ CRC.

Vasilenko, V.I. (2004). The concept and typology of terrorism. *Statutes and Decisions: The Laws of the USSR and its successor States*, 40(5), 46-56.

Weimann, G. (2004). How modern terrorism uses the internet. *United States Institute Report*, 116. Retrieved from <http://www.usip.org/files/resources/sr116.pdf>.

Weimann, G. (2005). Cyberterrorism: The sum of all fears? *Studies in Conflict & Terrorism*, 28, 129–149.

White, J.R. (2003). *Terrorism: An Introduction*. Belmont, CA: Wadsworth / Thompson Learning.

White, J. R. (2012). *Terrorism and homeland security* (7th ed.). Belmont, CA: Wadsworth Cengage Learning.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.

Wilkinson, P. (2001). Current and future trends in domestic and international terrorism: Implications for democratic government and the international community. *Strategic Review for Southern Africa*, 23(2).

Zanini, M. & Edwards, S (2001). Networking of terror in the information age. In J.Arquilla and D. Ronfeldt (Ed.s.), *Networks and netwars: the future of terror, crime, and militancy*, Santa Monica: RAND.

Zelinsky, A. & Shubik, M. (2006). *Terrorist Groups as Business Firms: A New Typological Framework*. Yale University. Retrieved at <http://ssrn.com/abstract=959258>.

ABSTRACT**ROBUST REGRESSION METHODS FOR MASSIVELY DECAYED INTELLIGENCE DATA**

by

AKIVA JOACHIM LORENZ**May 2014****Advisor:** Dr. Barry Markman**Major:** Evaluation and Research**Degree:** Doctor of Philosophy

Homeland Security, sponsored by governmental initiatives, has become a vibrant academic research field. However, most efforts were placed with the recognition of threats (e.g. theory) and response options. Less effort was placed in the analysis of the collected data through statistical modeling. In a field that collects more than 20 terabyte of information per minute through diverse overt and covert means and indexes it for future research, understanding how different statistical models behave when it comes to massively decayed data is of vital importance.

Using Monte Carlo methods, three regression techniques (ordinary least squares, least-trimmed, and maximum likelihood) were tested against different data decay models presumed to be found in homeland security research studies in order to test whether these techniques will preserve the Type I error rate in the t-test of standardized beta.

The results of these Monte Carlo simulations (sample size $n=30,90,120,240,480$ and 100,000 iterations) showed that the least trimmed squares method should be avoided under any circumstance due to the lack of a defined standard error, while the maximum likelihood technique should be avoided with smaller sample sizes due to the inflated Type I errors. Interestingly, although it is known that the ordinary least squares

regression can be impacted by non-normality and other assumption violations, it is remarkable robust to normally distributed data that is subject to massive decay.

Keywords: Homeland Security, Analysis, Data Decay, Monte Carlo, Regression

AUTOBIOGRAPHICAL STATEMENT

Akiva Joachim Lorenz received his Bachelor of Arts in Government (Cum Laude) in 2007 from the Interdisciplinary Center in Herzliya, Israel. Specializing in Security and Intelligence Studies, Akiva Lorenz utilized his academic knowledge working for several think tanks, including the Institute for Counter Terrorism (ICT). His research and policy papers were published by the United Nations, Intelligence Libraries, and are used by entities such as the Defense Academy of the UK in their Command and Staff Courses. During that time, data analysis has become an important factor in his life, resulting in the pursuit of his Doctor of Philosophy in Evaluation and Research (EER) at Wayne State University. During his doctoral studies, Akiva Lorenz received six academic awards for excellence including Wayne State's Graduate Research Fellowship. Akiva Lorenz has since applied his statistical and analytical skills also to the marketing and real estate fields as a licensed Realtor[®]. Currently, Akiva Lorenz is the President of Optimal Leads Inc., a consulting firm, and teaches as adjunct/part time faculty at several universities.