


12-2018

A SECONDARY DATA ANALYSIS FOR THE REFINEMENT OF EMPIRICAL MODELLING FOR THE ESTIMATION OF ENVIRONMENTAL HEALTH & SAFETY PROGRAM RESOURCING FOR COLLEGES AND UNIVERSITIES

SETH MICHAEL PARKER
UTHealth School of Public Health

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen

 Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

PARKER, SETH MICHAEL, "A SECONDARY DATA ANALYSIS FOR THE REFINEMENT OF EMPIRICAL MODELLING FOR THE ESTIMATION OF ENVIRONMENTAL HEALTH & SAFETY PROGRAM RESOURCING FOR COLLEGES AND UNIVERSITIES" (2018). *UT School of Public Health Dissertations (Open Access)*. 236.

https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/236

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digcommons@library.tmc.edu.

A SECONDARY DATA ANALYSIS FOR THE REFINEMENT OF EMPIRICAL
MODELLING FOR THE ESTIMATION OF ENVIRONMENTAL
HEALTH & SAFETY PROGRAM RESOURCING
FOR COLLEGES AND UNIVERSITIES

by

SETH MICHAEL PARKER, BS MS

APPROVED:

KRISTINA D. MENA, MSPH PhD

ROBERT J. EMERY, MS MPH DrPH

PATRICK M. TARWATER, MS PhD

DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

© Copyright
by
Seth Michael Parker, BS, MS, DrPH
2018

DEDICATION

This dissertation is dedicated to my lovely wife, Tiffany Michelle Parker, and to my son, Cooper Ellis Parker.

A SECONDARY DATA ANALYSIS FOR THE REFINEMENT OF EMPIRICAL
MODELLING FOR THE ESTIMATION OF ENVIRONMENTAL
HEALTH & SAFETY PROGRAM RESOURCING
FOR COLLEGES AND UNIVERSITIES

by

SETH MICHAEL PARKER
BS, The University of Texas at Austin, 2004
MS, The University of Texas at San Antonio, 2009

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PUBLIC HEALTH

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
December, 2018

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to my dissertation supervisor, Professor Robert Emery, for his guidance and support throughout my research at The University of Texas Health Science Center at Houston School of Public Health. This dissertation would not be possible without his help.

I would like to thank my advisor, Dr. Kristina Mena, for her encouragement to strive for this degree.

I wish to thank Dr. Patrick Tarwater for agreeing to be a member of the dissertation committee and for his expert statistical advice.

I would also like to thank Dr. Bruce Brown for his permission to access the data assembled for his research and the Campus Safety, Health, and Environmental Management Association (CSHEMA) for their enthusiastic support for this effort.

A SECONDARY DATA ANALYSIS FOR THE REFINEMENT OF EMPIRICAL
MODELLING FOR THE ESTIMATION OF ENVIRONMENTAL
HEALTH & SAFETY PROGRAM RESOURCING
FOR COLLEGES AND UNIVERSITIES

Seth Michael Parker, BS, MS, DrPH
The University of Texas
School of Public Health, 2018

Dissertation Chair: Robert J. Emery, MS, MPH, DrPH

The use of potentially hazardous physical, chemical, biological, and radiological agents is inherent to the teaching, research, and services missions of any university. To manage the risks associated with these agents, it is common for universities to host environmental health & safety (EH&S) programs to protect the safety of the institution's students, faculty, staff, and visitors. However, since EH&S programs in universities are primarily focused on prevention, it is difficult to estimate the appropriate "industry average" in terms of budget and staffing resources a particular university needs for such programs.

Historically, the Campus Safety, Health, and Environmental Management Association (CSHEMA) has collected data on a multitude of statistical measures for benchmarking purposes. CSHEMA currently collects data on a subset of likely predictors using the "vital statistics" survey. Cross-validation and information criteria were used to objectively identify which of the collected statistical measures are critical to the prediction of industry average EH&S program resourcing. The purpose of this project is to pinpoint the

variables that explain the majority of variance in the model, thereby minimizing unnecessary resource allocation dedicated to the collection of irrelevant data and illuminating predictors critical to CSHEMA's "vital statistics" survey.

A total of 109 members of the CSHEMA organization participated in this research project. The dependent variables were: (1) environmental health and safety expenditures and (2) environmental health and safety full-time employees; the independent variables were: (1) total institutional net assignable square footage, (2) total institutional expenditures, (3) research net assignable square footage, (4) institutional research expenditures, (5) total number of enrolled students, and (6) total institutional full-time employees.

Based on cross-validation and information criteria followed by robust regression methods: M-estimation, LTS-estimation, S-estimation, and MM-estimation, the findings of the present study indicate that institutional research expenditures, institutional research net assignable square footage, and institutional full-time employees are the optimal set of potential predictors for EH&S expenditures. The optimum predictors for EH&S full-time employees are total institutional net assignable square footage, institutional research expenditures, and total institutional full-time employees. The results indicate that these independent predictor variables should be considered critical for CSHEMA's future vital statistics survey.

TABLE OF CONTENTS

List of Tables	i
List of Figures	iii
List of Appendices	v
Background.....	1
Introduction.....	2
The College and University Population at Risk and Health and Safety Outcomes	4
Historical and Current Approaches to Estimating Safety Program Resourcing	8
Public Health Significance	10
Research Objectives	11
Statistical Methodologies	12
Information Criteria for Model Selection	14
Cross-Validation.....	16
Outliers, Influence, and Leverage	17
Efficiency and Breakdown Point.....	18
Equivariance.....	21
Robust Regression Estimators	22
M-Estimators.....	22
Andrew’s Sine	24
Huber’s Method.....	24
Tukey’s Bisquare.....	24
Algorithm for M-Estimation	25
Least Trimmed Squares (LTS) Estimators	26
Algorithm for LTS-Estimation.....	27
S-Estimators.....	27
Algorithm for S-Estimation	28
MM-Estimators	29
Algorithm for MM-Estimation.....	29
Study Design.....	31
Sample Size Calculation and/or Study Power	32
Data Set.....	34
Dependent Variables.....	35
Independent Variables	35
Data Analysis	38

Human Subjects, Animal Subjects, or Safety Considerations	40
Results	41
Discussion.....	71
Model Validation.....	81
Limitations and Suggestions	84
Conclusion	86
Appendices.....	95

LIST OF TABLES

Table 1. Descriptive Statistics for the Campus Safety, Health, and Environmental Management Association (CSHEMA) Data.....	45
Table 2. Descriptive Statistics for Natural Logarithmic and Square Root Transformed Campus Safety, Health, and Environmental Management Association (CSHEMA) Data.	46
Table 3. Multiple Regression Models Selected with Environmental Health and Safety Expenditures as the Dependent Variable.....	50
Table 4. Natural Logarithmically Transformed Parameter Estimates and Associated Variance Inflation Factor (VIF) with p-values for Model A.	51
Table 5. Natural Logarithmically Transformed OLS and Robust Parameter Estimates for Model A.	54
Table 6. Multiple Regression Models Selected with Environmental Health and Safety Full-time Employees as the Dependent Variable.	62
Table 7. Natural Logarithmically Transformed Parameter Estimates and Associated Variance Inflation Factor (VIF) with p-values for Model D.	63
Table 8. Natural Logarithmically Transformed OLS and Robust Parameter Estimates for Model D.	65
Table 9. Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 56,029ft ²	74
Table 10. Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 511,000ft ² ..	75
Table 11. Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 26,324,500ft ²	76
Table 12. Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 56,029ft ²	77
Table 13. Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 511,000ft ²	78

Table 14. Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 26,324,500ft ²	79
Table 15. Information Criteria for Competing Models.	80
Table 16. Reported and Modeled Values for Environmental Health and Safety Full-time Employees.....	82
Table 17. Reported and Modeled Values for Environmental Health and Safety Expenditures.	83

LIST OF FIGURES

Figure 1. Power vs. Sample Size for Multiple Regression Analysis.....	33
Figure 2. Histogram Overlay for Institutional Research Expenditures for Years 2011, 2013 and 2015.....	42
Figure 3. Histogram Overlay for Institutional Research Net Assignable Square Footage for Years 2011, 2013 and 2015.	43
Figure 4. Histogram Overlay for Institutional Full-time Employees for Years 2011, 2013 and 2015.	44
Figure 5. Fit Criteria using 5-Fold Cross-Validation for the Natural Logarithm of Environmental Health and Safety Expenditures as the Dependent Variable.	48
Figure 6. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Full-time Employees.....	55
Figure 7. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Research Net Assignable Square Footage.....	56
Figure 8. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Research Expenditures.....	57
Figure 9. 3D Surface Plot for Institutional Research Net Assignable Square Footage and Institutional Research Expenditures as the Independent Variables and Environmental Health and Safety Expenditures as the Dependent Variable.	58
Figure 10. 3D Surface Plot for Institutional Research Net Assignable Square Footage and Institutional Full-time Employees as the Independent Variables and Environmental Health and Safety Expenditures as the Dependent Variable.	59
Figure 11. Fit Criteria using 5-fold Cross-Validation for the Square Root of Environmental Health and Safety Full-time Employees as the Dependent Variable.	61
Figure 12. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Total Institutional Net Assignable Square Footage.	66
Figure 13. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Institutional Full-time Employees.	67

Figure 14. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Institutional Research Expenditures.68

Figure 15. 3D Surface Plot for Total Institutional Net Assignable Square Footage and Institutional Research Expenditures as the Independent Variables and Environmental Health and Safety Full-time Employees as the Dependent Variable.69

Figure 16. 3D Surface Plot for Total Institutional Net Assignable Square Footage and Institutional Full-time Employees as the Independent Variables and Environmental Health and Safety Full-time Employees as the Dependent Variable.70

LIST OF APPENDICES

Appendix A: Campus Safety, Health, and Environmental Management Association (CSHEMA) Approval Letter.....	95
Figure A1. CSHEMA Approval Letter.....	95
Appendix B: Histograms and Boxplots for Transformed Dependent and Independent Variables.	96
Figure B1. Histogram and Boxplot for the Natural Logarithm of Environmental Health and Safety Expenditures.....	96
Figure B2. Histogram and Boxplot for the Square Root of Environmental Health and Safety Full-time Employees.....	97
Figure B3. Histogram and Boxplot for the Natural Logarithm of Research Net Assignable Square Footage.	98
Figure B4. Histogram and Boxplot for the Natural Logarithm of Institutional Full-time Employees.	99
Figure B5. Histogram and Boxplot for the Natural Logarithm of Institutional Research Expenditures.	100
Figure B6. Histogram and Boxplot for the Natural Logarithm of Total Institutional Net Assignable Square Footage.	101
Figure B7. Histogram and Boxplot for the Natural Logarithm of Total Institutional Expenditures.....	102
Figure B8. Histogram and Boxplot for the Natural Logarithm of Total Number of Enrolled Students.	103
Appendix C: Box-Cox Analysis for Dependent Variables.....	104
Figure C1. Box-Cox Analysis for Environmental Health and Safety Expenditures.	104
Figure C2. Box-Cox Analysis for Environmental Health and Safety Full-time Employees.	105
Appendix D: Fit diagnostics for the Natural Logarithm of Environmental Health and Safety Expenditures.....	106
Figure D1. Fit Diagnostics for Model A.....	106
Figure D2. Q-Q Plot of Residuals with Shapiro-Wilk’s test for Model A.	107
Figure D3. Residual by Predicted for Model A.	108
Figure D4. Residual by Regressor for Model A.....	109
Appendix E: Outlier Analysis for Model A.....	110
Figure E1. Cook’s D for Model A.....	110
Figure E2. Difference in Fits (DFFITS) Influence Diagnostics for Model A.....	111
Figure E3. Difference in Betas (DFBETAS) Influence Diagnostics for Model A.....	112
Figure E4. Outlier and Leverage Diagnostics using OLS for Model A.	113
Appendix F: Robust Regression Estimation for Model A.....	114

Figure F1. M-estimation RDPLOT for Model A.	114
Figure F2. M-estimation DDPlot for Model A.	115
Figure F3. M-estimation Histogram of Standardized Robust Residuals for Model A.	116
Figure F4. M-estimation Q-Q Plot for Standardized Robust Residuals for Model A.	117
Figure F5. LTS-estimation RDPlot for Model A.	118
Figure F6. LTS-estimation DDPlot for Model A.	119
Figure F7. LTS-estimation Histogram of Standardized Robust Residuals for Model A.	120
Figure F8. LTS-estimation Q-Q Plot for Standardized Robust Residuals for Model A.	121
Figure F9. S-estimation RDPlot for Model A.	122
Figure F10. S-estimation DDPlot for Model A.	123
Figure F11. S-estimation Histogram of Standardized Robust Residuals for Model A.	124
Figure F12. S-estimation Q-Q Plot for Standardized Robust Residuals for Model A.	125
Figure F13. MM-estimation RDPlot for Model A.	126
Figure F14. MM-estimation DDPlot for Model A.	127
Figure F15. MM-estimation Histogram of Standardized Robust Residuals for Model A.	128
Figure F16. MM-estimation Q-Q Plot for Standardized Robust Residuals for Model A.	129
Appendix G: Fit Diagnostics for the Square Root of Environmental Health and Safety Full-time Employees.	130
Figure G1. Fit Diagnostics for Model D.	130
Figure G2. Q-Q Plot of Residuals with Shapiro-Wilk's Test for Model D.	131
Figure G3. Residuals by Predicted for Model D.	132
Figure G4. Residuals by Regressor for Model D.	133
Appendix H: Outlier Analysis for Model D.	134
Figure H1. Cook's D for Model D.	134
Figure H2. Difference in Fits (DFFITs) Influence Diagnostics for Model D.	135
Figure H3. Difference in Betas (DFBETAS) Influence Diagnostics for Model D.	136
Figure H4. Outlier and Leverage Diagnostics using OLS for Model D.	137
Appendix I: Robust Regression Estimation for Model D.	138
Figure I1. M-estimation RDPlot for Model D.	138
Figure I2. M-estimation DDPlot for Model D.	139
Figure I3. M-estimation Histogram of Standardized Robust Residuals for Model D.	140

Figure I4. M-estimation Q-Q Plot for Standardized Robust Residuals for Model D.....	141
Figure I5. LTS-estimation RDPLOT Model D.	142
Figure I6. LTS-estimation DDPLOT for Model D.	143
Figure I7. LTS-estimation Histogram of Standardized Robust Residuals for Model D.....	144
Figure I8. LTS-estimation Q-Q Plot for Standardized Robust Residuals for Model D.....	145
Figure I9. S-estimation RDPLOT for Model D.	146
Figure I10. S-estimation DDPLOT for Model D.	147
Figure I11. S-estimation Histogram of Standardized Robust Residuals Model D.....	148
Figure I12. S-estimation Q-Q plot for Standardized Robust Residuals for Model D.....	149
Figure I13. MM-estimation RDPLOT for Model D.....	150
Figure I14. MM-estimation DDPLOT for Model D.	151
Figure I15. MM-estimation Histogram of Standardized Robust Residuals for Model D.....	152
Figure I16. MM-estimation Q-Q Plot for Standardized Robust Residuals for Model D.....	153

BACKGROUND

The CSHEMA has historically collected membership benchmark data describing various aspects of EH&S program operations. Examples include number of research labs, amount of hazardous waste generated, number of persons trained, number of spills, and number of accidents. Over time, it became salient to assess which of these operational parameters were predictive of EH&S program staffing and resourcing. Based on a series of unpublished works (which will be addressed later in this introduction), CSHEMA identified six key variables deemed the vital statistics. These metrics include: (1) total institutional net assignable square footage, (2) total institutional expenditures, (3) research net assignable square footage, (4) institutional research expenditures, (5) total number of enrolled students, and (6) total institutional full-time employees. The present study is focused on objective analysis of the vital statistics that explain the most variance in the models when regressed on an EH&S program's expenditures and full-time employees.

The use of non-traditional statistical techniques that conform to a better selection of CSHEMA vital statistics variables was accomplished through the use of cross-validation and information criteria, performed simultaneously during the model selection process to find the most parsimonious models. Robust regression methods—M-estimation, LTS-estimation, S-estimation and MM-estimation—were used in the presence of outliers and influential observations to optimize the parameter estimates.

INTRODUCTION

Over time, colleges and universities have transformed from institutions consisting of traditional classroom spaces, for the provision of academic instruction, to enormous and complex research enterprises (Bush, 1945). This transformation has caused an increase in the incidence and prevalence of employment-related accidents. Due to the lack of jurisdiction of city health departments on university campuses, institutions of higher education must take responsibility for their own EH&S. Therefore, in order to manage the risks associated with the increase in hazardous conditions, academic institutions have created environmental health and safety programs.

University EH&S programs face a multitude of potential occupational health risks, as well as the possibility of simultaneous exposure to several hazardous agents (Emery, 1998). Academic EH&S programs typically comprise safety departments charged with biological, chemical, radiation, environmental, fire prevention, and occupational health. The complexity of environmental health and safety programs tends to vary with the size of the institution, population density, and types of academic departments present (Deroos, 1977).

In 1945, the Campus Safety Association (CSA) was founded in order to provide guidance and benchmarking efforts to address the complexity of safety challenges facing various health and safety programs across different universities. In fact, the founders of CSA noted that the university health and safety setting is incredibly unique in that it faces challenges similar to those of industry, business, research, and municipal spheres as well as those specific to an educational setting. In 1995, the CSA's name was changed to CSHEMA:

the preeminent organization for providing guidance to the environmental health and safety community in higher education.

Presently, CSHEMA is responsible for assisting the safety culture of academic institutions through instructional training, surveys, and symposia. This programming allows CSHEMA members to discuss relevant topics related to the multitude of issues facing the university health and safety community. The biannual benchmarking survey is used to collect data on general safety topics, fire safety, biosafety, radiation safety, waste disposal, regulatory compliance, and other areas related to best practices; the results of the survey allow member institutions to compare these critical metrics and develop effective benchmarking. The campus climate survey allows for member institutions to assess the integration of a safety climate over time. The safety advancement program addresses health and safety performance by assessing safety management system topics including general health and safety requirements, legal considerations, and environmental health and safety monitoring.

THE COLLEGE AND UNIVERSITY POPULATION AT RISK AND HEALTH AND SAFETY OUTCOMES

Developing surveillance methods for university laboratory settings presents a challenge for EH&S programs, as most incidents go unreported and central databases require improvement in terms of tracking the incidence and prevalence of laboratory associated hazards. The Bureau of Labor Statistics (BLS) has developed a surveillance program for academic institutions, termed the Injuries, Illnesses, and Fatalities (IIF) program. The IIF program is responsible for reporting the incidence rates and numbers of fatal and nonfatal occupational injuries and illnesses for colleges, universities, and professional schools. The IIF accomplishes this through the Survey of Occupational Injuries and Illnesses (SOII) and the Census of Fatal Occupational Injuries (CFOI). The SOII reported an incidence rate of 1.8 per 100 full-time employees with 22,000 total cases for the year 2016 (BLS, 2017a). The CFOI estimated that fatal occupational injuries for colleges, universities, and professional schools totaled 30 for the year 2016 (BLS, 2017b).

Laboratory work can be a daunting setting, in which routine pipetting, centrifuging, and autoclaving is coupled with the pressure to meet proposal and other administrative deadlines; as such, it is easy to overlook laboratory safety. However, failing to deal proactively with laboratory personnel health and safety issues can lead to dire consequences. In order to reduce morbidity and mortality associated with hazards in the university laboratory setting, individuals are required to understand Occupational Safety and Health Administrative (OSHA) guidelines and effectively increase active awareness of the potential hazards associated with institutions of higher education among laboratory personnel, faculty,

and staff. While personnel complete basic laboratory training before being allowed to enter most academic labs, students do not always receive sufficient training in the hazards and can therefore begin to take on a complacent attitude. Complacency, accompanied by insufficient laboratory hazard training, can lead to incidents of morbidity and mortality in the university laboratory setting.

Laboratory associated hazards typically fall under the categories of biological, radiological, chemical, and physical. Biological hazards are especially salient concerns, as a plethora of pathogenic microorganisms are currently being studied in university research laboratories. Harding and Byers (2006) classified these microorganisms as viral, parasitic, bacterial, and fungal in nature. Cumulative data from multiple cross-sectional studies by Sulkin and Pike, which spanned 1950 to 1978, reveals a total of 4,079 documented cases of Laboratory Associated Infections (LAI), of which 168 resulted in mortality. Interestingly, more than two-thirds of these lethal and nonlethal infections were associated with viruses and bacteria, namely *Coxiella burnetii*, *Brucella*, hepatitis B virus, *Salmonella typhi*, *Francisella tularensis*, and *Mycobacterium tuberculosis*.

These silent killers can go undetected, resulting in morbidity or mortality after days or even weeks have passed; therefore, root cause analysis can be a daunting task. Often, failure of proper containment while sampling biological specimens creates an increased risk for LAIs among safety professionals. In fact, findings from a cross-sectional study among biological safety personnel conducted by Patlovich et al. (2015) suggest that the majority of biological sampling occurs at the institutional level. Patlovich et al. (2015) also proposed a field research toolkit and advanced training for biosafety personnel.

Radiological hazards also pose potential hazards to university health and safety professionals. Harmful levels of radiation exposure can have devastating consequences in the form of deoxyribonucleic acid breakage, chromosomal aberrations, and genetic mutations. As such, containment of radiation is critical in any university laboratory setting. The potential risks and side effects of radiation exposure vary depending on the amount of radiation absorbed. The ionizing radiation found in the university setting typically comes in the form of isotopes used in diagnostic settings: for example, the use of cyclotrons and x-rays to diagnose certain ailments. Geiger counters and film badges are typically used to monitor dangerous levels of radiation in the university laboratory setting. Based on the linear no threshold hypothesis, there is no safe dose of radiation; therefore, when dealing with radiation, it is useful to remember the acronym ALARA: “As Low As Reasonably Achievable” (Kathren, 2002). Amis et al. (2007) have provided recommendations concerning factors that contribute to unsafe and unnecessary radiation exposure, and suggested that the majority of hazardous radiation exposure incidents are due to overuse and misuse of radiation technology. The specific recommendations were designed to educate stakeholders in employing radiation safety principles, appropriately utilizing imaging to minimize any associated radiation risks, and standardizing radiation dose data to be archived during imaging for defining good practice.

Chemical hazards in the university setting can be extremely dangerous if not dealt with according to OSHA safety guidelines. Hazardous chemicals are used in a wide variety of operations and activities in university laboratory settings. Therefore, laboratory personnel must pay close attention to proper usage, handling and storage of these materials. These

activities are essential to mitigating chemically induced hazards in the work environment. Hazardous waste is a significant chemical hazard in the university setting; the volume of hazardous chemicals used in universities makes effective chemical waste disposal a daunting task. The Environmental Protection Agency (EPA) has specifically identified approximately 500 chemicals as hazardous waste. The EPA classifies these chemicals into lists: F-list, K-list, P-list, and U-list (EPA, 2018). These chemicals generally fall under the category of being either corrosive, toxic, reactive, or ignitable. In order to combat and prevent chemical exposure to hazardous materials, laboratory personnel must attend mandatory meetings covering OSHA standards 1910.1450 and 29 CFR 1910.1200.

Issues pertaining to physical safety are also a concern in the higher education environment. These include, but are not limited to, elevated work surfaces, compressed gases, electrical equipment, seismic considerations, and thermal hazards. Compressed gases are at risk of leaks and ruptures, which can lead to asphyxiation or turn a gas cylinder into a projectile. Electrical hazards from improper wiring and faulty or malfunctioning Ground-Fault Circuit Interrupters (GFCIs) also present a risk to physical safety. Improper loading of electrical outlets can lead to fire hazards, placing an entire building at risk. Minimizing risk in these areas is accomplished via the development of fire safety programs which train students and staff on proper application and physical safety.

HISTORICAL AND CURRENT APPROACHES TO ESTIMATING SAFETY PROGRAM RESOURCING

University environmental health and safety departments struggle with optimizing performance while maintaining proper resource allocation and a sufficient number of staff to effectively prevent workplace associated hazards. Maintaining adequate resource allocation and staffing levels while minimizing downtime is a difficult task, which most environmental professionals find challenging. Health and safety professionals are charged with the task of weighing the budget and optimizing environmental health and safety workload throughout the university setting, with the ultimate goal of achieving the optimum amount of staff to minimize downtime and still maintain a sound EH&S program.

Historically, mathematical modeling of university health and safety programs has been concerned with factors that are theoretical and qualitative in nature. Although the heuristic approach to resource allocation and staffing is less time consuming, quantitative assessments provide for a pragmatic and more statistically sound approach. Ultimately, quantitative methods for benchmarking resource allocation and staffing in university health and safety programs were lacking.

Fine et al. (1982) pioneered mathematical modeling for health and safety programs, using a model aimed at developing an environmental health and safety staffing formula. However, after a thorough review by the American Society of Safety Engineers, Council on Practices and Standards (ASSE, 2010), the mathematical formula presented by Fine et al. (1982) was found to lack specificity for the academic health and safety setting due to its generalizability to health and safety programs in multiple areas of industry. The

mathematical model developed by Fine et al. (1982) was based primarily on weighting certain variables, and was not substantiated by any quantitative reasoning; instead, it was primarily grounded in theoretical assumptions and heuristic in nature. Ultimately, the Fine et al. (1982) mathematical model was simply based on qualitative assessments from previous on-the-job experience. Such heuristic approaches are ineffective in the proper allocation of university health and safety program resources. Therefore, health and safety programs should emphasize quantitative data-based assessments rather than qualitative and heuristic assessments in addressing resource allocation.

Significantly, Brown (2014) developed a quantitative approach for assessing resource allocation for university health and safety programs. Brown's (2014) initial model included total institutional net assignable square footage and total institutional expenditures as the primary predictors for both environmental health and safety expenditures and environmental health and safety staffing. Brown (2014) also included subsets during analysis, including whether institutions were members of the Association of Academic Health Centers (AAHC) and/or members of the Carnegie classified institutions. Brown et al. (2014) discovered that members of AAHC institutions required more resources than non-AAHC member institutions, and suggested that this increase in resources among AAHC member institutions may be caused by increased proportion of institutional research square footage to total institutional net assignable square footage. Similarly, J. Wang (personal communication [Practicum project], (2002)) developed a quantitative approach for EH&S staffing, which included total institutional net assignable square footage, institutional research net assignable

square footage, presence of medical or veterinary school, and whether there was a biosafety level-3 facility as part of the institution.

PUBLIC HEALTH SIGNIFICANCE

Reducing the incidence and prevalence of workplace related hazards is the primary focus of any university environmental health and safety program. An informed decision-making process to facilitate this reduction can only be accomplished through proper allocation of resources for university environmental health and safety programs. The purpose of this dissertation was to assist health and safety management stakeholders in making informed decisions on the allocation of resources by using a quantitative data set—the CSHEMA “vital statistics” survey—to create a valid statistical model. The present study included the use of applied statistical reasoning methods, cross-validation, and information criteria with robust regression to build upon the existing body of knowledge and develop a more efficacious statistical model for the appropriation of health and safety resources. The results of this dissertation hold great potential for enabling university leaders and other key stakeholders in predicting “industry average” safety program resourcing.

RESEARCH OBJECTIVES

Identification of environmental health and safety program resources, to optimize field performance, is critical to management as well as stakeholders. The models produced in this analysis will create an industry average for environmental health and safety personnel to benchmark performance.

- 1) Create an optimal regression equation simultaneously using information criteria and cross-validation techniques.
- 2) Field test the regression equation using external model validation.
- 3) Investigate if the previous analysis of this data could have utilized other important factors to supplement the estimated value and strength of the factors found utilizing different robust and resistant regression methodologies.

STATISTICAL METHODOLOGIES

The method of least squares was first introduced by Gauss (1795) for the study of geodesy and astronomy; however, the application of least squares was first published by Legendre (1805), which set off the “priority dispute over the discovery of the method of least squares” (Stigler, 1981). Today, the method of least squares is the workhorse in the world of regression analysis, which is by far the most widely used statistical technique (Takeaki & Horoshi, 2004). It is important to note that, if the error terms are i.i.d; $\epsilon_i \sim N(0, \sigma^2)$, the least squares estimator is the maximum likelihood estimator for β .

$$\prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e\left(-\frac{\epsilon_i^2}{2\sigma^2}\right) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e\left(-\frac{(y_i - x_i^T \beta)^2}{2\sigma^2}\right)$$

This is equivalent to maximizing the logarithm.

$$\sum_{i=1}^n \left(-\frac{1}{2} \ln(2\pi\sigma^2) - \frac{\epsilon_i^2}{2\sigma^2} \right)$$

Since σ is constant, this simply corresponds to minimizing:

$$\sum_{i=1}^n \epsilon_i^2$$

and, therefore, the least squares estimate, $\hat{\beta}$, is that which results in the minimal sum of squares of the residuals, $\sum_{i=1}^n r_i^2$.

The general linear model is defined by:

$$Y_{nx1} = X_{n \times p} \beta_{n \times p} + \epsilon_{nx1}$$

where: Y is the response vector of observations on the dependent variable; β is the parameter vector; X is a $n \times p$ design matrix of observations on p variables for n units; and ϵ is a vector of identical and independent errors i.i.d. $N(0, \sigma_i^2)$.

Therefore, the least squares estimator is calculated as:

$$\hat{\beta}_{OLS} = \mathop{\text{arg min}}_{\beta} \sum_{i=1}^n (Y_i - X_i^T \beta)^2$$

since the sum of squared errors (SSE) is equal to the inner product of the residuals vector with itself, $\sum_{i=1}^n r_i^2 = r^T r$:

$$\begin{aligned} \hat{\beta}_{OLS} &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= (Y^T - X^T \hat{\beta}^T) (Y - X\hat{\beta}) \\ &= Y^T Y - Y^T X \hat{\beta} - X^T \hat{\beta}^T Y - X^T \hat{\beta}^T X \hat{\beta} \end{aligned}$$

In order to find the smallest possible values of β , it is necessary to differentiate $r^T r$ with respect to β and set the derivative equal to zero: $\frac{\partial r^T r}{\partial \beta} = \mathbf{0}$. Since $X^T \hat{\beta}^T Y = (X\hat{\beta})^T Y = Y^T X \hat{\beta}$.

$$= Y^T Y - 2X^T Y \hat{\beta} - X^T \hat{\beta}^T X \hat{\beta}$$

$$\frac{\partial r^T r}{\partial \beta} = 2X^T X \hat{\beta} - 2X^T Y = \mathbf{0}$$

If the matrix $X^T X$ is invertible, the normal equation for β is:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T Y$$

where the least squares equation satisfies the following classical assumptions. The sum of the residuals is equal to zero, $\sum r_i = \mathbf{0}$. The sample mean of the residuals must also be

equal to zero: $\bar{\mathbf{r}} = \frac{\sum r_i}{n} = \mathbf{0}$. The predicted values of y are uncorrelated with the residuals:

$$\hat{\mathbf{y}}^T \mathbf{r} = (\mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{r} = \mathbf{b}^T \mathbf{X}^T \mathbf{r} = \mathbf{0}.$$

Information Criteria for Model Selection

The focus of multiple regression analysis is establishing a regression equation that can be applied to external data to predict an unknown criterion variable or dependent variable with known data on predictor variables. After obtaining the parameter estimates, the sample squared coefficient of multiple correlation is often used in determining the proportion of variation in the criterion variable, which is explained by the predictors in the model. The coefficient of multiple correlation or R_j^2 is defined as:

$$R_j^2 = 1 - \left(\frac{SSE}{SST} \right)$$

where sum of squares errors (SSE) and total sum of squares (SST). The adjusted squared coefficient of multiple correlation:

$$R_{adj}^2 = 1 - \frac{n-1}{n-p-1} (1 - R_j^2)$$

R_{adj}^2 is typically used as criterion for model selection because it adjusts for the number of parameters in the model. The model subsets which produce the highest value of R_{adj}^2 are considered as those which produce the best fit. The Mallows' C_p statistic is also used among the criterion for model selection (Mallows, 1973):

$$C_p = \frac{SSE}{MSEP} + 2(p+1) - n$$

where MSEP is the mean squared prediction error, $\hat{d}^2 = \sum(Y - \hat{Y})/n$ (Brown, 1975). The best model is that in which the number of predictors in the model are close to the C_p statistic. The predicted error sum of squares (PRESS) statistic, which is an alternative of MSEP, is also commonly used in best subset selection (Allen, 1971, 1974).

$$PRESS = \sum (Y - \hat{Y})^2$$

where the smaller values for the PRESS statistic are favored, indicating a better subset of predictors. Akaike's Information Criterion (AIC) is also of importance in best subset selection (Akaike, 1973).

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2(p + 1)$$

The model with the lowest AIC value is considered the best subset model. The Schwarz Bayesian Criteria, SBC, is similar to AIC. They both penalize the number of parameters in the model; however, SBC imposes a greater penalty on additional parameters in the model by using the multiplier of $\ln \cdot n$ for the number of parameters in the model, instead of the constant 2 (Schwarz, 1978).

$$SBC = -2L_p + p \ln \cdot n$$

Where n is the sample size, L_p is the maximum log-likelihood of the model and p is the number of parameters in the model. Another model selection criterion derived from AIC is Sawa's Bayesian Information Criteria (BIC), which is a function of the number of observations, number of predictor variables, the sum of squared errors (SSE), and the error variance for fitting the full model σ^2 (Sawa, 1978).

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{2(p+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2}$$

As shown, the complexity of the penalty term for the number of parameters in the model for BIC is greater than AIC and is a function of n , SSE, and σ^2 , in addition to p .

Cross-Validation

After determining which parameters to include in the model. Model selection techniques generally conclude with cross-validating the regression model to adequately determine the predictive accuracy of the model. Snee (1977) argued that these two processes are not independent of one another and should be considered simultaneously to enhance the model selection process.

Cross-validation originated in the context of multiple regression analysis. The use of cross-validation is important in terms of selecting a reliable model that will fit the data available in other samples. In other words, cross-validation is useful for determining how the model will generalize to unseen data sets not used in determining the initial parameters. In K-fold cross-validation, the sample is partitioned into K sub-samples of equal size. Where one of the K sub-samples is retained as the model validation set, the remaining K-1 sub-samples are used as the model training set. The cross-validation procedure is then repeated K times, where each of the K folds in the process is used as the validation set.

Outliers, Influence, and Leverage

Outliers and influence diagnostics are an inherent component of model building; therefore, dealing with outliers and influential observations requires careful consideration. With the advent of advanced computational resources, the method of least squares seems rather attractive to data analysts; however, this method is unlikely to result in perfect linear relationships that satisfy classical assumptions. The most important limitation of using least squares regression is its sensitivity to outliers. This problem arises from the fact that the term is squared, $\sum_{i=1}^n r_i^2$, which exaggerates the magnitude of the difference. The squaring places more emphasis on discordant or rogue observations. There are numerous definitions of outliers in the literature. Johnson et al. (1992) defined an outlier as “an observation in a dataset which appears to be inconsistent with the remainder of that set of data.” In the presence of outliers and influential points, the least squares estimates can be biased and unreliable, producing erroneous results.

Several types of outliers potentially plague least squares estimates. Outliers in the covariate space are referred to as leverage points. Rousseeuw and Leroy (1987) have defined leverage as an observation $(\mathbf{x}_k, \mathbf{y}_k)$ where \mathbf{x}_k lies far away from the bulk of the observed data \mathbf{x}_i . It is important to understand that \mathbf{y}_k does not change, meaning that a leverage point does not have to be an outlier because we are just dealing with x space. If an observation resides close to the regression line, it is considered a “good” leverage point; conversely, if an observation falls far from the regression line, it is considered a “bad” leverage point. The influence of a leverage point on regression coefficients depends on how far away from the

regression line they reside. In order to deal with influential and outlying observations, more robust techniques are required.

Effectively dealing with outliers is an important component of the model building process; as an alternative to least squares regression, robust regression operates with less restriction on the classical OLS assumptions. Robustness is defined by Huber (1996) as, “insensitivity to small deviations from the assumptions made”. Regression using robust estimators is capable of providing better regression coefficients when corruption or discrepant data are present. The goal of robust regression is to limit the influence of outliers while providing stout results in their presence.

Efficiency and Breakdown Point

The efficiency and breakdown point are two of the most popular criteria to measure the robustness of a statistical procedure. The breakdown point (BDP) is defined as the smallest fraction of contamination in a dataset that can cause the estimator to produce arbitrary results. When an estimator “breaks down,” it fails to adequately represent the general trend in the dataset. Hodges (1967) introduced the idea of BDP restricted to one dimensional estimation of location. Further work by Hampel (1968, 1971) extended the breakdown point to include estimation for location functionals. Finite sample breakdown points (FSBDP) were later introduced by Donoho (1982) and Donoho and Huber (1983). Given a sample of n data points,

$$\mathbf{z} = \{(x_{11}, \dots, x_{1p}, y_1), \dots, (x_{n1}, \dots, x_{np}, y_n)\}$$

where T is a regression estimator. Applying T to the sample z yields a vector of regression coefficients:

$$T(Z) = \hat{\theta}.$$

It is also necessary to consider all possible corrupted samples z' where m of the original dataset is replaced with arbitrary values. Then, the maximum bias that could be obtained by these substitutions is:

$$\mathbf{bias}(m; T, Z) = \sup_{z'} ||T(Z') - T(Z)||$$

where the supremum is over all possible Z' . If the $\mathbf{bias}(m; T, Z)$ is infinite, then the corrupted sample space by m outliers can have a substantial impact on the regression estimator T ; in other words, the estimator “breaks down.” Thus, the finite sample BDP for the regression estimator T at the sample Z is defined as:

$$\epsilon_n^*(T, Z) = \min\left\{\frac{m}{n}; \mathbf{bias}(m; T, Z) \text{ is infinite}\right\}.$$

Therefore, the finite sample breakdown point of the regression estimator T , $\epsilon_n^*(T, Z)$, is the smallest fraction or percentage of corrupted sample space that can cause the estimator T to take on values arbitrarily far from $T(Z)$.

The ordinary least squares estimator is extremely sensitive to outliers, with a breakdown point of:

$$\epsilon_n^*(T, Z) = \frac{1}{n}$$

and, as the sample size increases, $1/n$ tends to zero, so the OLS has a BDP of 0%. This means that just one outlier is enough for the least squares estimator to reach its BDP.

The highest possible value of the BDP for an estimator is 50%; in this case, half of the observations in the dataset are contaminated. BDPs higher than 50% are unreasonable, because the estimates depend on less than half the data and, therefore, it would be difficult to distinguish between the uncontaminated and contaminated sample spaces. Therefore, a reasonable BDP for an estimator should be chosen. Hampel (1986) argued that, in any given data set, 10% of the observations deviate from the general trend in the data. Thus, a reasonable estimator should have a BDP of at least 10%.

If making inferences concerning a larger population, it is desirable to have the Best Linear Unbiased Estimator (BLUE) possible, which is incredibly efficient (Anderson, 2008). The relative efficiency of an estimator is defined as the ratio of its minimum variance to its actual variance. An estimator is only deemed “efficient” when the ratio is equal to one. This means that the sampling variance and standard errors are small. Estimators that reach acceptable levels of efficiency with large sample sizes are deemed asymptotically efficient. Anderson (2008) argued that it is critical for an estimator to have high efficiency if a researcher is seeking inferences about a larger population. Given T_1 and T_2 as estimators for the population parameter θ , where T_1 has maximum efficiency and T_2 is less efficient. The relative efficiency of T_2 is calculated by the ratio of its mean squared error to the mean squared error of T_1 (Anderson, 2008). The relative efficiency between two estimators is expressed as:

$$\text{Efficiency}(T_1, T_2) = \frac{E[(T_1 - \theta)^2]}{E[(T_2 - \theta)^2]}$$

OLS is used as a benchmark, to which other robust estimators are compared, when the assumptions of linearity, constant error variance, and uncorrelated errors are met.

Equivariance

Rousseeuw and Leroy (1987) addressed three critical equivariance properties concerning robust regression estimators. Listed in order of importance, these are: regression, scale, and affine equivariance (Rousseeuw & Leroy, 1987).

Regression equivariance. An estimator T is regression equivariant if

$$T(\{(x_i^T, y_i + x_i^T v); i = 1, \dots, n\}) = T(\{(x_i^T, y_i); i = 1, \dots, n\}) + v$$

meaning that any additional linear dependence is reflected in the regression vector accordingly.

Scale equivariance. Scale equivariance for a robust estimator means that the fit is independent of the measurement unit for the response variable, y . An estimator is scale equivariant if it satisfies

$$T(\{(x_i^T, cy_i); i = 1, \dots, n\}) = cT(\{(x_i^T, y_i); i = 1, \dots, n\})$$

If an estimator does not have the property of equivariance of scale, it must be standardized.

We will show later that M-estimators are not scale equivariant and the studentization of residuals has to be performed by an error scale estimate, $\hat{\sigma}$.

Affine equivariance. An estimator has the property of affine equivariance

$$T(\{(x_i^T A, cy_i); i = 1, \dots, n\}) = A^{-1}T(\{(x_i^T, y_i); i = 1, \dots, n\})$$

Since $\hat{y}_i = (\mathbf{x}_i^T \mathbf{T}) = (\mathbf{x}_i^T \mathbf{A})(\mathbf{A}^{-1} \mathbf{T})$, affine equivariance allows for the use of another coordinate system for the independent variables, without altering the estimated \hat{y}_i .

ROBUST REGRESSION ESTIMATORS

M-ESTIMATORS

Maximum likelihood type estimators, or M-estimators, made an important addition to robust statistics. M-estimators were introduced by Huber (1964) as a measure of estimating the location of a distribution and later generalized to regression analysis (Huber, 1973). The motivation behind M-estimation is to minimize the residual function ρ , where ρ is known as the “objective function.” The objective function should be continuous, symmetric, $\rho(\mathbf{r}) = \rho(-\mathbf{r})$, and have a unique minimum at zero (Rousseeuw & Leroy, 1987).

$$\hat{\beta}_m = \mathop{\text{arg min}}_{\beta} \sum_{i=1}^n \rho(Y_i - \mathbf{X}_i^T \beta)$$

M-estimators are regression equivariant but not scale equivariant; thus, the studentization of residuals must be performed by an error scale estimate of $\hat{\sigma}$:

$$\hat{\beta}_m = \mathop{\text{arg min}}_{\beta} \sum_{i=1}^n \frac{\rho(y_i - \mathbf{X}_i^T \beta)}{\hat{\sigma}}$$

which is minimized by taking the first partial derivative of ρ with respect to β and setting it to zero. Where the derivative of ρ , denoted as $\psi = \rho'$, gives the influence function, ψ .

$$\sum_{i=1}^n \mathbf{X}_i \psi \left(\frac{y_i - \mathbf{X}_i^T \beta}{\hat{\sigma}} \right) = \sum_{i=1}^n \psi \left(\frac{r_i}{\hat{\sigma}} \right) \mathbf{X}_i$$

The coefficients can now be calculated from a system of $K + 1$ estimating equations, where ψ is replaced by weights; as the size of the residuals increases, the size of the weights decrease. The system of estimating equations can now be written as:

$$\hat{\beta}_m = \arg \min_{\beta} \sum_{i=1}^n w(X_i) \frac{\rho(Y_i - X_i^T \beta)}{\hat{\sigma}}$$

and in matrix notation:

$$XW_i X \beta = XW_i Y$$

$$\hat{\beta} = (XW_i X)^{-1} XW_i Y$$

where W_i is defined as the weight matrix.

In SAS® 9.4, the default value for the estimate of scale, $\hat{\sigma}$, is the median absolute deviation (MAD), which is defined as:

$$MAD = \mathit{median} |r_i| = \mathit{median} |y_i - x_i^T \hat{\beta}|$$

which can be selected by the SCALE=MED option. The MAD is based on the median and is highly resistant to outliers, with a BDP equal to 50% (Anderson, 2008).

M-estimation uses a weight function, w , in order to bound the influence of outlying X_i 's. M-estimators are extremely vulnerable to leverage points and have a BDP of 0%; M-estimators are popular when leverage points are not an issue.

Some of the most commonly used objective, score, and weight functions for M-estimation are listed below. Where c is the tuning constant.

Andrew's Sine:
$$\rho(r) = \begin{cases} c[1 - \cos(r/c)], & \text{if } |r| < \pi c \\ 2c, & \text{if } |r| \geq \pi c \end{cases}$$

$$\psi(r) = \begin{cases} \sin(r/c), & \text{if } |r| < \pi c \\ 0, & \text{if } |r| \geq \pi c \end{cases}$$

$$w(r) = \begin{cases} \frac{\sin(r/c)}{r/c}, & \text{if } |r| < \pi c \\ 0, & \text{if } |r| \geq \pi c \end{cases}$$

Huber's Method:
$$\rho(r) = \begin{cases} r^2, & \text{if } |r| < c \\ |2r|c - c^2, & \text{if } |r| \geq c \end{cases}$$

$$\psi(r) = \begin{cases} r, & \text{if } |r| < c \\ c|\text{sign}(r)|, & \text{if } |r| \geq c \end{cases}$$

$$w(r) = \begin{cases} 1, & \text{if } |r| < c \\ c/|r|, & \text{if } |r| \geq c \end{cases}$$

Tukey's Bisquare:
$$\rho(r) = \begin{cases} \frac{c^2}{3} \left\{ 1 - \left[1 - \left(\frac{r}{c} \right)^2 \right]^3 \right\}, & \text{if } |r| < c \\ 2c, & \text{if } |r| \geq c \end{cases}$$

$$\psi(r) = \begin{cases} z \left[1 - \left(\frac{r}{c} \right)^2 \right]^2, & \text{if } |r| < c \\ 0, & \text{if } |r| \geq c \end{cases}$$

$$w(r) = \begin{cases} \left[1 - \left(\frac{r}{c} \right)^2 \right]^2, & \text{if } |r| < c \\ 0, & \text{if } |r| \geq c \end{cases}$$

Algorithm for M-Estimation

M-estimation requires that the residuals and scale be calculated simultaneously; therefore, an iterative procedure must be used that is capable of converging on an estimate for both (Anderson, 2008).

Step 1) A least squares regression is run, calculating the initial regression coefficient estimates $\hat{\beta}^{(0)}$.

Step 2) The residuals are calculated from the least squares regression in step 1, and the initial weight estimates are calculated from the residuals.

Step 3) A weight function is then applied to the initial least squares residuals to create preliminary weights, $w(r_i^{(0)})$.

Step 4) The first iteration uses weighted least squares (WLS) to minimize $\sum w_i^{(1)} r_i^2$ to obtain $\hat{\beta}^{(1)}$. In matrix notation:

$$\hat{\beta}^{(1)} = (XW_iX)^{-1}XW_iY.$$

Step 5) The process continues by using the residuals from the initial least squares regression to calculate new weights, $w_i^{(2)}$.

Step 6) The new weights $w_i^{(2)}$ are used in the next iteration, $I = 2$, of the least squares regression to estimate $\hat{\beta}^{(2)}$.

Steps 4-6 are repeated until the estimate of $\hat{\beta}$ stabilizes.

The iteration procedure can also be modified for different convergence criteria other than the change in the coefficient estimates. The relative change in the scaled residuals and

the relative change in weights can also be used; however, in SAS[®], the default for the iteration process continues until $\widehat{\boldsymbol{\beta}}^{(l)} - \widehat{\boldsymbol{\beta}}^{(l-1)} \cong \mathbf{10}^{-8}$.

LEAST TRIMMED SQUARES (LTS) ESTIMATORS

Least trimmed squares (LTS) estimators, introduced by Rousseeuw (1984), are high breakdown estimators with a breakdown point of up to 50%, meaning that, if the volume of outliers in the data set is less than 50%, the LTS estimator will not break down. The method of LTS is defined as:

$$\widehat{\boldsymbol{\theta}}_{LTS} = \mathbf{arg\,min}_{\boldsymbol{\theta}} Q_{LTS}(\boldsymbol{\theta})$$

with

$$Q_{LTS}(\boldsymbol{\theta}) = \sum_{i=0}^h r_{(i)}^2$$

where $r_{(1)}^2 \leq r_{(2)}^2 \leq \dots \leq r_{(n)}^2$ are the ordered squared residuals $r_{(i)}^2 = (\mathbf{y}_i - \mathbf{x}_i' \boldsymbol{\beta}^T)^2$ and $i = 1, \dots, n$. The subset of h observations, used in the least trimmed squares (LTS) regression, is defined within the range $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, where n is the total number of observations and p is the number of independent variables. The subset h of observations is then used to find the least squares fit that possesses the smallest sum of squared residuals.

SAS[®] uses the default value of $\frac{3n+p+1}{4}$ for h ; by using the H=option in the MODEL statement, it is possible to choose any number for h within the range

$$\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}.$$

Algorithm for LTS-Estimation

The ROBUSTREG procedure uses the FAST-LTS algorithm (Rousseeuw & Van Driessen, 2000). The LTS algorithm was adapted from SAS/STAT[®] 9.2 User's Guide.

Step 1) Choose a value for h , within $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, using the H = option in the Model statement.

Step 2) Determine the number of regressors in the model, p . If $p = 1$, then the exact algorithm suggested (Rousseeuw & Leroy, 1987). If $p \geq 2$, then use the algorithm in step 3.

Step 3) Draw a random p subset and compute the regression coefficients.

Step 4) Compute the absolute residuals from all the observations in the data set.

Step 5) Select the first h points with the smallest absolute residuals.

Step 6) Carry out the Concentration steps or C-steps from the selected h subset. Redraw p subsets and repeat $nrep$ times to find the solutions with the lowest sums of h squared residuals.

S-ESTIMATORS

Another high breakdown estimator was first introduced by Rousseeuw and Yohai (1984), aptly named s-estimators based on estimates of scale. S-estimators have the equivariant properties of regression, scale, and affine. S-estimators minimize the dispersion of the residuals.

$$\hat{\beta} = \underset{\beta}{\mathit{arg\ min}}(r_1(\beta), \dots, r_n(\beta))$$

where the final scale estimates

$$\hat{\sigma} = s(r_1(\hat{\beta}), \dots, r_n(\hat{\beta}))$$

The dispersion, s , of the residuals, according to Rousseeuw and Leroy (1987), is defined as:

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{y_i - x_i^T \theta}{s}\right) = \beta$$

where β being a constant which is often set equal to $E_{\Phi}[\rho(r)]$, and Φ is denoted as the standard normal distribution. The objective function, ρ , should satisfy the following requirements:

- 1) ρ is symmetric, $\rho(r) = \rho(-r)$ and continuously differentiable and the objective function equals zero when its argument is zero, $\rho(0) = 0$; and
- 2) there exists $c > 0$ such that ρ is strictly increasing on $[0, c]$ and constant on $[c, \infty)$.

Algorithm for S-Estimation

The S-estimate, implemented by Marazzi (1993), uses the ROBUSTREG procedure in SAS[®].

Step 1) The regression coefficients are calculated by the random q -subset, where $q \geq p$.

Step 2) Compute the residuals: $r_i = y_i - \sum_{j=1}^p x_{ij}\beta_j$ for $i = 1, \dots, n$. If iteration = 1, set

$s^* = 2\text{median}\{|r_i|, i = 1, \dots, n\}$; if $s^* = 0$, set $s^* = \min\{|r_i|, i = 1, \dots, n\}$; while

$\sum_{i=1}^n \rho(r_i/s^*) > (n-p)\beta$, set $s^* = 1.5s^*$; go to step 3. If iteration > 1 and

$\sum_{i=1}^n \rho(r_i/s^*) \leq (n-p)\beta$, go to step 3; otherwise go to step 5.

Step 3) Solve for s using iteratively reweighted least squares.

$$\frac{1}{n-p} \sum_{i=1}^n \rho(r_i/s) = \beta$$

Step 4) If iteration > 1 and $s > s^*$, go to step 5. Otherwise, set $s = s^*$ and $\theta^* = \theta$. If $s^* < \text{TOLS}$, where TOLS is the tolerance for the S estimate with a default = 0.001, return s^* and θ^* ; otherwise, go to step 5.

Step 5) If iteration < NREP, where NREP is the number of repeats, set iteration = iteration + 1 and return to step 1; otherwise, return s^* and θ^* .

MM-ESTIMATORS

First proposed by Yohai (1987), MM-estimators simultaneously possess high breakdown and high asymptotic relative efficiency. MM-estimation is a further development of M-estimation. MM-estimation is accomplished by finding the initial regression parameter estimates using S-estimation followed by M-estimation. These properties allow for MM-estimation to obtain a higher statistical efficiency than S-estimation.

Algorithm for MM-Estimation

Step 1) The initial estimates of the coefficients $\hat{\beta}^{(1)}$ and the residual values are calculated from a highly resistant robust regression method. S-estimation with Huber or Tukey's bisquare weights are typically chosen for this step.

Step 2) The residuals from step 1 are used to compute an M-estimation of the scale of the residuals, $\hat{\sigma}_e$.

Step 3) The initial estimates of the residuals from step 1 and of the residual scale $\hat{\sigma}_r$ are used in the first iteration of the weighted least squares to determine the M-estimates of the regression coefficients, where w_i are usually Huber or Tukey's bisquare weights.

$$\sum_{i=1}^n w_i \left(\frac{r_i^{(1)}}{\hat{\sigma}_r} \right) x_i = \mathbf{0}$$

Step 4) New weights are calculated, $w_i^{(2)}$, with the residuals from the initial weighted least squares (step 3).

Step 5) The scale of the residuals from steps 2-4 are kept constant and are continually reiterated until convergence.

STUDY DESIGN

The descriptive nature of this study is in survey format collected via a survey of the CSHEMA members, combined with data from publicly available sources during 2011, 2013 and 2015. Therefore, a cross-sectional study design for this dissertation was deemed appropriate. The convenience sample covers a large geographic region and will provide trends among U.S. colleges and universities pertaining to the outcomes of interest:

- 1) Environmental Health and Safety Expenditures
- 2) Environmental Health and Safety Full-time Employees

SAMPLE SIZE CALCULATION AND/OR STUDY POWER

In multiple regression analysis a sufficient number of observations, n , is necessary. Literature cites different formulas for recommended minimum sample sizes. The confidence interval and prediction interval are dependent upon the number of observations thus a solution for the number of observations exists. Consideration to the number of independent variables, k , in the model must also be given. Sample size is determined by the desired significance level, power and variance in the population (Hicks and Turner, 1999). Stevens (1995) recommends that the formula $n \geq 15k$ be used in calculating the appropriate sample size for multiple regression analysis. In this analysis the number of participating colleges and universities is $n = 109$. The total number of predictors $k = 6$. Based on the formula provided by Stevens (1995) a sufficient number of observations are present in the data set used in this analysis. Figure 1 shows the type III F test for the multiple regression used during the model building process. The power used for the three-parameter model is 0.90 which is above 0.80 indicating that there is sufficient power and that the sample size is adequate for the analysis.

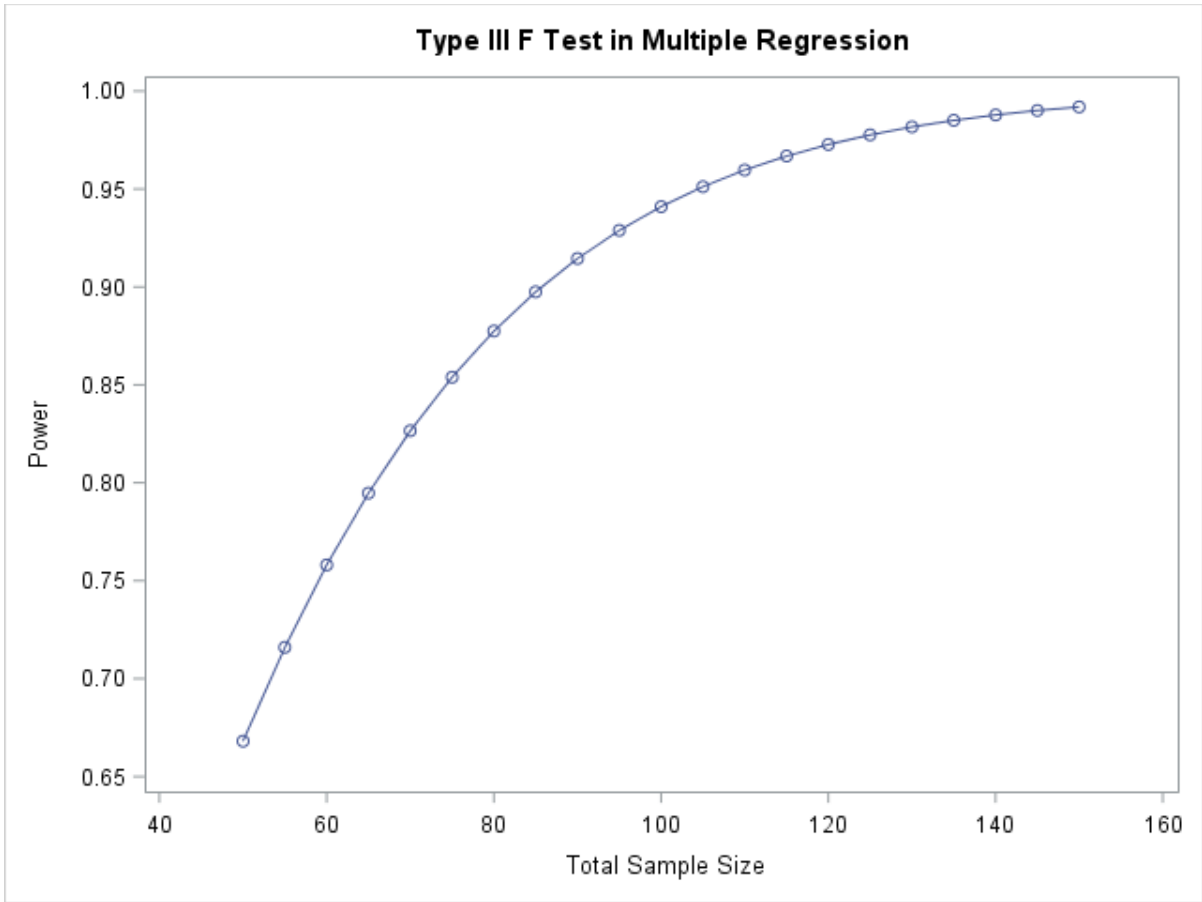


Figure 1. Power vs. Sample Size for Multiple Regression Analysis.

DATA SET

The data set used in this project was extracted from the Campus Safety, Health, and Environmental Management Association (CSHEMA) survey taken during 2013. The data set consisted of survey data from publicly available databases collected during 2011. Additional data was collected for the years 2013 and 2015 for the independent variables: institutional research expenditures, total institutional full-time employees and institutional research net assignable square footage. The CSHEMA data set was complemented with information obtained from the following publicly available databases:

1. National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS) Survey
2. National Science Foundation (NSF) Survey of Science and Engineering Research Facilities
3. National Science Foundation Higher Education Research and Development (HERD) Survey

The detailed questionnaire was produced with the primary purpose of identifying resource drivers for environmental health and safety programs at 109 U.S. colleges and universities. Influential variables were also analyzed to determine their significance in determining institutional environmental health and safety staffing and expenditures. This data set holds the potential to produce an industry average to be used in environmental health and safety staffing and expenditures.

Dependent Variables

Environmental Health and Safety Expenditures – EH&S expenditures is defined as the total fiscal year funds for operations, maintenance and salary for each university. The EH&S expenditures data were obtained from the CSHEMA benchmarking survey.

Environmental Health and Safety Full-time Employee – EH&S full-time employee is defined as the number of EH&S employees on full-time schedules. Part-time employees' schedules are converted to full-time basis if staff devotes half-time or greater (≥ 0.5 FTE) to the institution's environmental health and safety department. The number of EH&S full-time equivalents for each institution was obtained from the CSHEMA benchmarking survey.

Independent Variables

Total Institutional Net Assignable Square Feet – Reported as “Net Usable Area” by the National Center for Education Statistics and defined as “The sum of all areas on all floors of a building either assigned to, or available for assignment to, an occupant or specific use, or necessary for the general operation of a building” (U.S. Department of Education, National Center for Education Statistics, Postsecondary Education Facilities Inventory and Classification Manual, 2006). Total institutional net assignable square feet were obtained from the CSHEMA benchmarking survey.

Research Net Assignable Square Feet – The academic research area reported by the National Science Foundation is contained in the following academic departments, “agricultural sciences and natural resources sciences, biological and biomedical sciences, computer and information sciences, engineering, health and clinical sciences, mathematics and statistics, physical sciences, psychology, social sciences, and other science and engineering fields” (National Science Foundation, National Institutes of Health Survey of Science and Engineering Research Facilities, 2009). Research net assignable square feet was obtained from the National Science Foundation Survey of Science and Engineering Research Facilities WebCASPAR database.

Research Expenditures – Research expenditures data was obtained from the National Science Foundation, Higher Education Research and Development Survey. The survey is taken annually, which includes universities which expend at least \$150,000 in research and development. The survey collects information on R&D expenditures by field of research and source of funds. (National Science Foundation, Higher Education Research and Development Survey, 2010).

Total Institutional Expenditures – Total institutional expenditures are the financial statements for an institution’s fiscal year. This data is recorded by the National Center for Education Statistics Integrated Postsecondary Education Data System (U.S. Department of Education, Integrated Postsecondary Education Data System, Finance for Public Institutions 2011).

Total Number of Enrolled Students – Total enrolled students are recorded by the National Center for Education Statistics Integrated Postsecondary Education Data System Survey. The total number of enrolled students is defined as the sum of an unduplicated count of students enrolled for credit during a 12-month period regardless of when the student enrolled (U.S. Department of Education, Integrated Postsecondary Education Data System Survey, 12-month enrollment 2011).

Total Full-time Employees – The total full-time employee per institution was obtained from the U.S. Department of Education, Integrated Postsecondary Education Data System Survey. Total full-time employees are defined by the National Center for Education Statistics as “The full-time-equivalent of staff is calculated by summing the total number of full-time staff from the Employees by Assigned Position component and adding one-third of the total number of part-time staff.” (U.S. Department of Education, Integrated Postsecondary Education Data System, IPEDS glossary, 2012).

DATA ANALYSIS

Data analysis was performed using SAS[®] version 9.4. A general observation of the data was implemented, detailing any types of numerical instability that could interrupt the analysis. Descriptive statistics including calculated mean, minimum, maximum, median, standard deviation and interquartile ranges were included in the analysis. This was accomplished using the MEANS and UNIVARIATE procedures in SAS[®].

Analysis of Variance (ANOVA) was performed to examine the distribution of the year-to-year differences among the variables: institutional research net assignable square footage, total institutional full-time employees; institutional research expenditures. Data for the independent variable total institutional net assignable square footage was not available.

Box plots and histograms were modified for each of the independent and dependent variables using the TEMPLATE procedure in SAS[®]. The box plots assessed the center and spread of the distributions giving insight into the issues of non-normal or unusual distributions. The box plots also revealed the skewness of the individual response and predictor variables in the analysis. Histograms of each of the independent and dependent variables were created to check for normality in the data by visually inspecting if the data is similar in appearance to that of the Gaussian distribution (Stigler, 1981). The independent variables were subsequently transformed using the natural logarithm and the Box-Cox family of power transformations were performed on both of the dependent variables.

The GLMSELECT procedure, with the cross-validation (CV) option in the model statement, was used for model selection. The recommended number of k-folds was

determined to be 5 (Hastie, Tibshirani and Friedman, 2001). The optimum values of Akaike's Information Criteria (AIC), corrected Akaike's Information Criteria (AICC), Sawa Bayesian Information Criteria (BIC), Schwarz Bayesian Information Criteria (SBC), R_{adj}^2 , predicted residual sum of squares (cvPRESS), as well as Mallows (C)_p were used as fit criteria for the model selection.

Quantile-quantile Q-Q plots were also used in the univariate analysis to check for normal distributions, outliers as well as skewness in the data. The statistical techniques used to check for normality in this project included the Shapiro-Wilk test statistic (Shapiro, 1965). The Variance Inflation Factor (VIF) was used to quantify multicollinearity and "the rule of 10" was determined to be the defining factor in detecting severe multicollinearity (Hair et. al., 1995). Visual representation of the dependent variables and the selected independent variables were performed using bivariate histograms, and 3D surface plots.

Outliers and influential observations were examined graphically by looking at residual plots. Residuals vs fitted values and plots of residuals vs predictors were used to assess the presence of outliers. The outliers were examined statistically using Cook's Distance, DFFITS and DFBETAS. The ROBUSTREG procedure in SAS[®] was used to perform M, LTS, S and MM robust regression techniques. The RDPLOT and DDLOT were used to check for outliers and influential observations using the robust estimation methods. Histograms and Q-Q plots of the residuals for the robust regression estimations are also displayed.

HUMAN SUBJECTS, ANIMAL SUBJECTS, OR SAFETY CONSIDERATIONS

This dissertation has been reviewed and approved by The University of Texas School of Public Health at Houston Office of Academic Affairs and Student Services. This work was determined to be exempt by The University of Texas Health Science Center at Houston (UTHealth) Committee for the Protection of Human Subjects as study # HSC-SPH-17-0836.

RESULTS

A total of 109 members of the Campus Safety, Health, and Environmental Management Association participated in the submission of data for this research project in 2013. The data set consisted of survey data from publicly available databases collected during 2011 from: National Center for Education Statistics (NCES), Integrated Postsecondary Education Data System (IPEDS) survey, National Science Foundation (NSF) Survey of Science and Engineering Research Facilities, and National Science Foundation Higher Education Research and Development (HERD) survey. Additional data was collected for the years 2013 and 2015 for the independent variables: institutional research expenditures, total institutional full-time employees, and institutional research net assignable square footage.

The dependent and independent variables that compose this data set are all continuous variables. The sole purpose of this research project was to examine the relationship between environmental health and safety expenditures and full-time employees with the institutional predictors: total institutional net assignable square footage, total institutional expenditures, research net assignable square footage, institutional research expenditures, total number of enrolled students and total institutional full-time employees. The full definition for each of the dependent and independent variables is located in the data set section of this research project.

Examining the year-to-year distributions of the independent variables—institutional research net assignable square footage, institutional full-time employees, and institutional

research expenditures—using one-way analysis of variance (ANOVA) produced p values > 0.05. This indicates that the means are not significantly different from year-to-year. The histogram overlays for these variables are located in Figures 2-4.

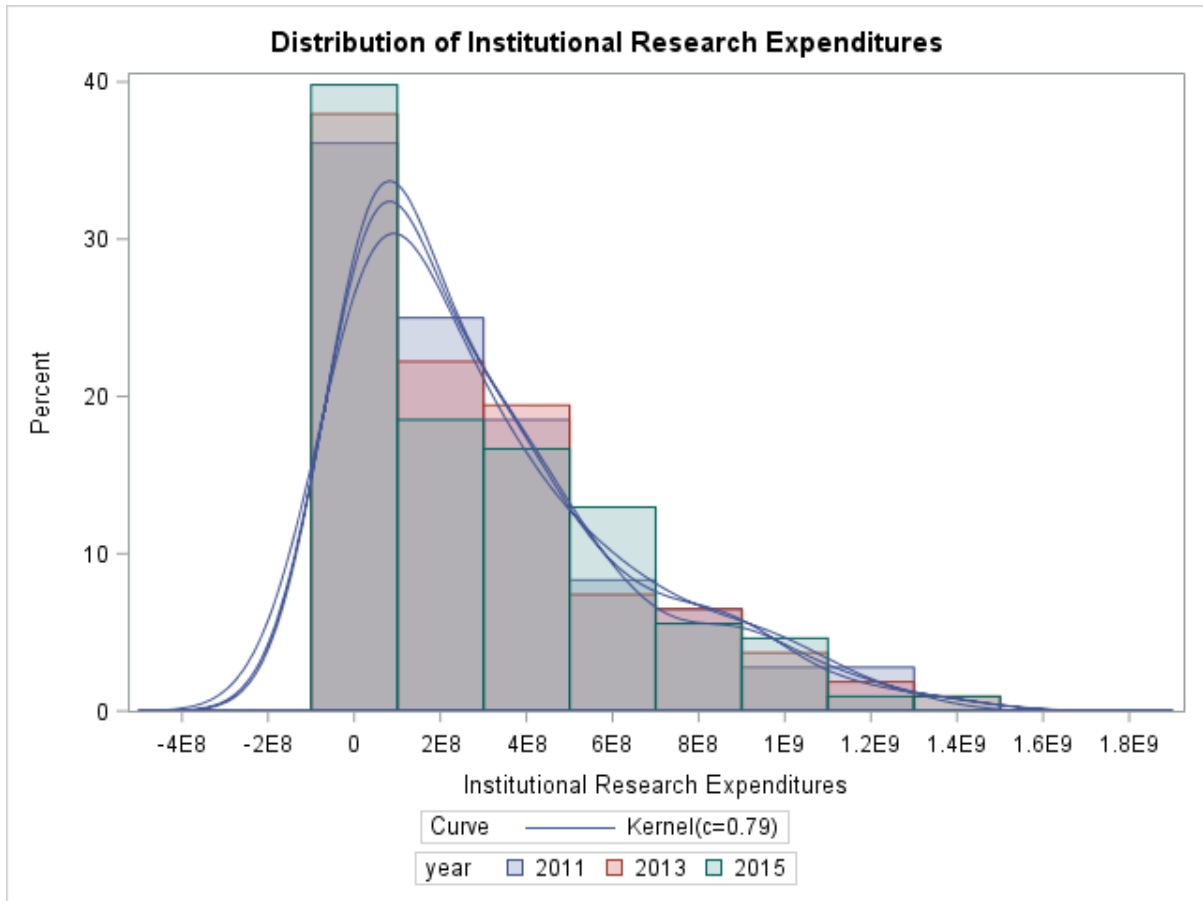


Figure 2. Histogram Overlay for Institutional Research Expenditures for Years 2011, 2013 and 2015.

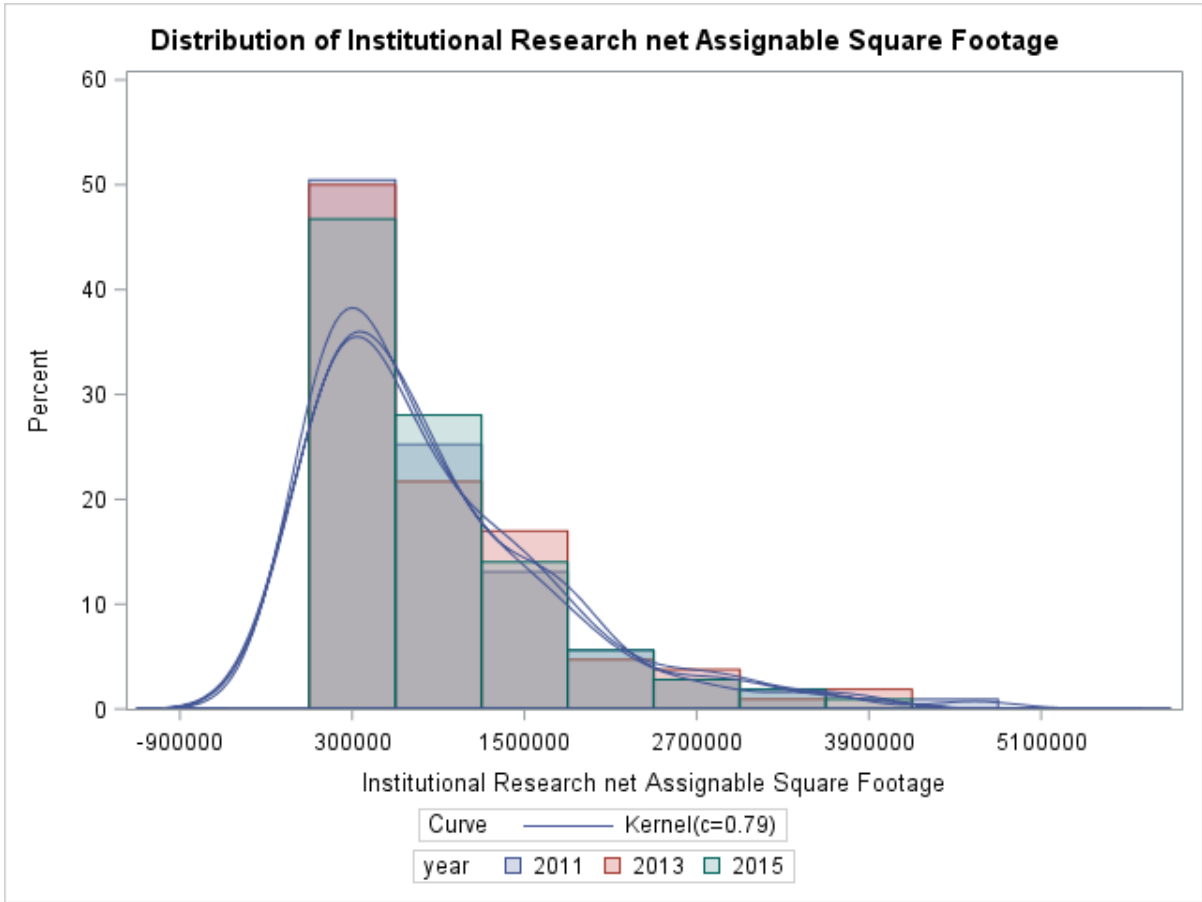


Figure 3. Histogram Overlay for Institutional Research Net Assignable Square Footage for Years 2011, 2013 and 2015.

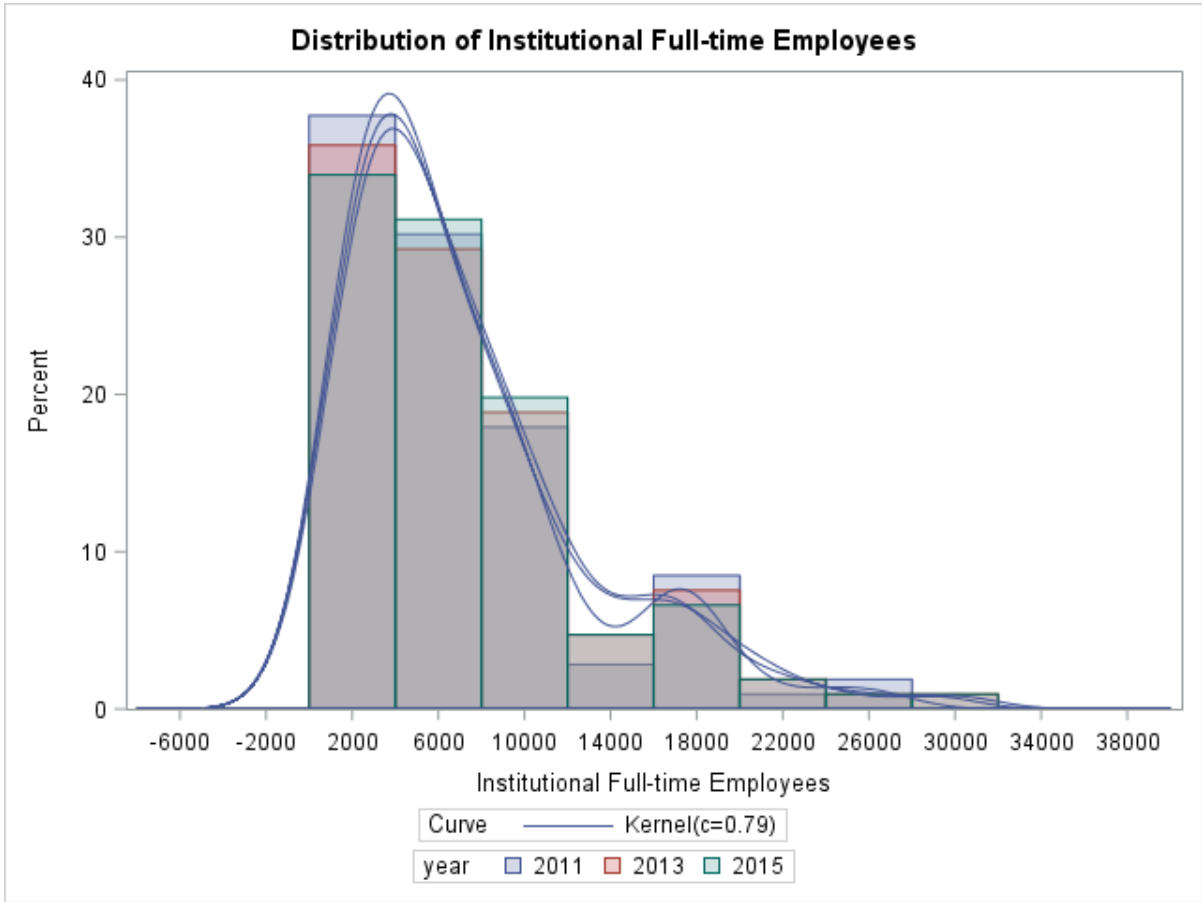


Figure 4. Histogram Overlay for Institutional Full-time Employees for Years 2011, 2013 and 2015.

The descriptive statistics for each of the dependent and independent variables include the mean, standard deviation, median, minimum, and maximum values. The descriptive analytics for each of the untransformed and transformed variables are located in Tables 1 and 2.

Table 1.

Descriptive Statistics for the Campus Safety, Health, and Environmental Management Association (CSHEMA) Data.

	Mean	Std Dev	Median	Min	Max
Dependent Variables					
Environmental Health and Safety Expenditures	\$2,252,608	\$2,091,973	\$1,650,000	\$48,500	\$10,102,788
Environmental Health and Safety Full-time Employees	23.89	20.30	20.00	1.00	100.00
Independent Variables					
Total Institutional Net Assignable Square Footage	6,695,880.63ft ²	4,912,633.07ft ²	5,298,610.00ft ²	562,104.00ft ²	25,320,731.00ft ²
Research Net Assignable Square Footage	815,820.31ft ²	854,278.22ft ²	511,000.00ft ²	5,400.00ft ²	4,631,400.00ft ²
Total Institutional Expenditures	\$981,720,608	\$944,428,769	\$667,265,752	\$75,341,410	\$5,249,817,000
Institutional Research Expenditures	\$289,473,431	\$304,233,668	\$198,655,000	\$1,232,000	\$1,279,123,000
Institutional Full-time Employees	6,976.72	5,383.14	5,281.00	376.00	26,489.00
Total Number of Enrolled Students	25,799.35	15,431.44	23,854.00	193.00	79,274.00

Table 2.

Descriptive Statistics for Natural Logarithmic and Square Root Transformed Campus Safety, Health, and Environmental Management Association (CSHEMA) Data.

	Mean	Std Dev	Median	Min	Max
Dependent Variables					
Environmental Health and Safety Expenditures	14.155	1.080	14.316	10.789	16.128
Environmental Health and Safety Full-time Employees	4.426	2.083	4.472	1.000	10.000
Independent Variables					
Total Institutional Net Assignable Square Footage	15.435	0.809	15.483	13.239	17.047
Research Net Assignable Square Footage	12.933	1.394	13.144	8.594	15.348
Total Institutional Expenditures	20.358	0.835	20.319	18.137	22.381
Institutional Research Expenditures	18.594	1.695	19.107	14.024	20.969
Institutional Full-time Employees	8.557	0.810	8.572	5.929	10.184
Total Number of Enrolled Students	9.871	0.953	10.080	5.263	11.281

In order to find the optimal fitting models, the dependent variables were transformed using different families of power transformations. Ultimately, the Box-Cox family of power transformations was used to facilitate selecting the optimum lambda values for the dependent variables in the regression models. The natural logarithmic transformation for the dependent variable, environmental health and safety expenditures, and the square root transformation for the dependent variable, environmental health and safety full-time employees, proved to be successful in normalizing the residuals in each of the models. The $F = t^2$ plots for the dependent variable, environmental health and safety expenditures, shows that the value of F is at its maximum in the vicinity of the optimal Box-Cox transformation. The lambda value selected for the dependent variable, environmental health and safety full-time employees, was 0.23; however, a square root transformation was used in this analysis. The plots for each of the dependent variables with the Box-Cox selected lambda values are located in Appendix C. The natural logarithmic transformation of the independent variables transformed the unimodal skewed distributions into normally distributed data. Univariate analysis after transformation of each variable, with histogram and box-plot, are located in Appendix B.

Cross-validation was used simultaneously with information criteria to find the “best” fitting model. K-fold cross-validation for environmental health and safety expenditures as the dependent variable was performed where the number of folds (K) was equal to 5. The fit criteria for environmental health and safety expenditures as the dependent variable are shown in Figure 5.

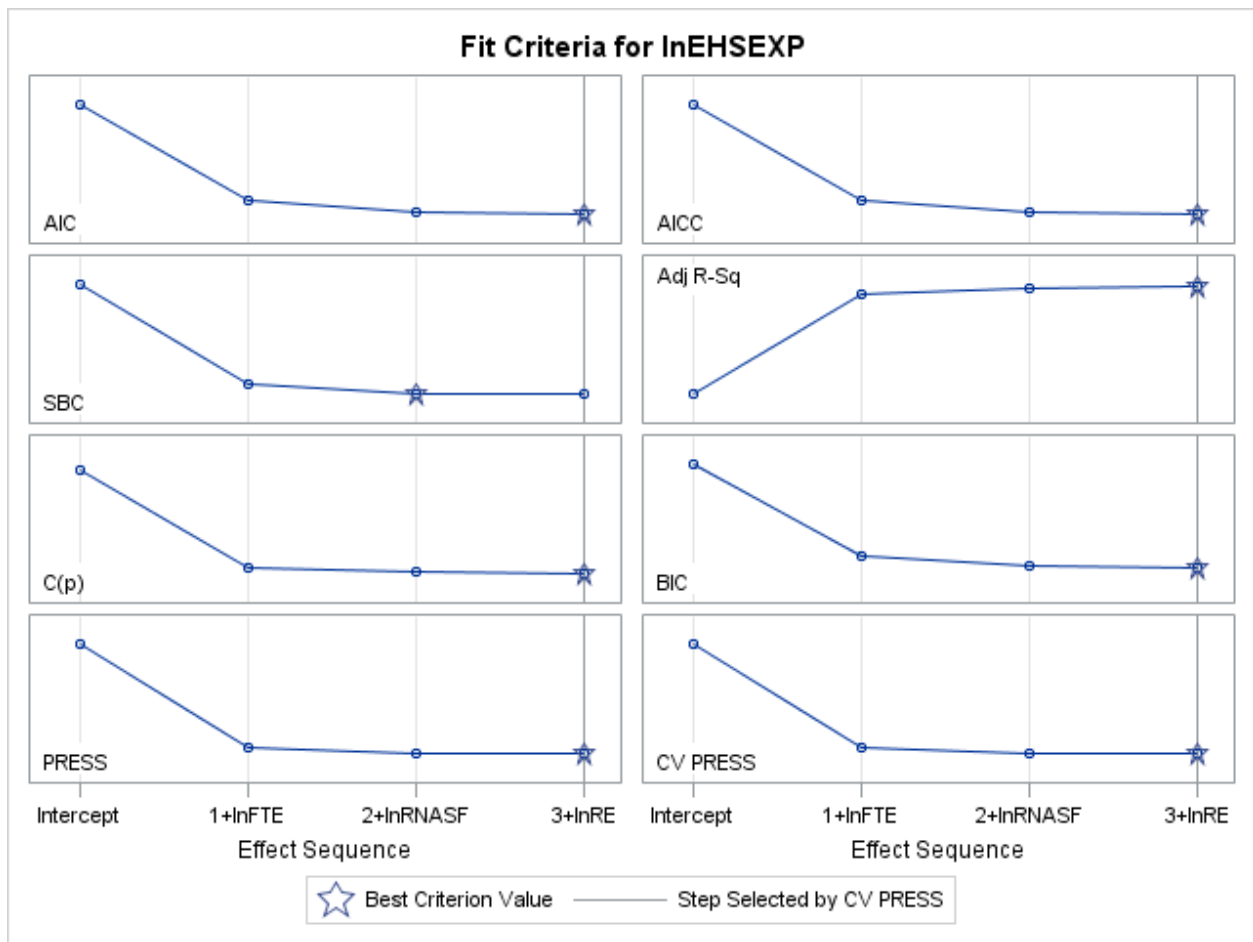


Figure 5. Fit Criteria using 5-Fold Cross-Validation for the Natural Logarithm of Environmental Health and Safety Expenditures as the Dependent Variable.

The selection of the appropriate regression models in this research project is based upon parsimony, which is defined by Montgomery (2001) as: “the simplest possible model that is consistent with the data and knowledge of the problem environment.” Model selection was facilitated by the use of information criteria. The information criteria used in model selection were AIC, AICC, SBC, BIC, PRESS, R_{adj}^2 , and Mallow’s (C)p. The model

selection using the maximum of R_{adj}^2 values coincided with the minimum value of the PRESS statistic. The prediction residual error sum of squares (PRESS) statistic is a measure of the predictive capability of the regression model. Smaller values of the PRESS statistic were chosen in examining the right model. Comparing the adjusted coefficient of determination R_{adj}^2 with the coefficient of determination R^2 is a good indication of whether the regression model contains too many predictor variables, hinting at possibly overfitting the model. The Schwarz Bayesian Information Criteria (SBC) selected the simpler model with institutional research net assignable square footage and institutional full-time employees as the independent variables. However, the SBC criteria chose the simpler two independent variable model; all of the other information criteria suggested the three independent variable model. Therefore, the model based on the SBC criteria has an increased capacity to generalize to the rest of the population, whereas the SBC model is possibly underfitted. Ultimately, the three-parameter model was chosen, and the comparison criteria were chosen as the R_{adj}^2 values. The models selected for the dependent variable, environmental health and safety expenditures, are listed in Table 3. After the regression analysis, a significant model emerged ($F = 126.89$, $p < 0.0001$, $R_{adj}^2 = 0.7776$); an alpha level of 0.05 was used throughout all statistical tests. The results from the regression analysis indicate that the research net assignable square footage, institutional full-time employees, and institutional research expenditures account for 77.76% of the variance in the model when regressed on environmental health and safety expenditures. Conversely, total institutional expenditures,

total number of enrolled students, and total institutional net assignable square footage were not significant in the model, and hence were excluded from the final model.

The parameter estimates with variance inflation factors (VIF) for Model A are located in Table 4. Values of VIF that exceed 10 tend to indicate that multicollinearity is present in the model and instability among the beta coefficients is present. The analysis revealed that none of the predictor variables had a VIF value greater than 10, suggesting that multicollinearity is not an issue among the independent variables. Therefore, instability among the beta coefficients is not present in the model.

Table 3.

Multiple Regression Models Selected with Environmental Health and Safety Expenditures as the Dependent Variable.

Model A:	$EHSEXP = e^{(\beta_0 + \beta_1 \ln RNASF + \beta_2 \ln FTE + \beta_3 \ln RE)}$
Model B:	$EHSEXP = e^{(\beta_0 + \beta_1 \ln RNASF + \beta_2 \ln RE)}$
Model C:	$EHSEXP = e^{(\beta_0 + \beta_1 \ln RNASF + \beta_2 \ln FTE)}$

Table 4.

Natural Logarithmically Transformed Parameter Estimates and Associated Variance Inflation Factor (VIF) with p-values for Model A.

Parameter	Parameter Estimate	Standard Error	t Value	Pr > t	VIF
Intercept	3.78236	0.55553	6.81	<.0001	0
Institutional Research Net Assignable Square Footage	0.17980	0.08147	2.21	0.0295	5.36367
Institutional Research Expenditures	0.16249	0.07695	2.11	0.0371	7.08199
Institutional Full-time Employees	0.58742	0.12597	4.66	<.0001	4.33693

Residual analysis for model A does not display any curves or cyclical patterns in the plots of the residuals over each of the independent variables in the model; this suggests that all the structure in the data is essentially captured by the multiple regression model. The quantile-quantile plot, Q-Q plot, shows light tails with a description of the point pattern, showing all but a few points fall on the line; this suggests the presence of outliers in Model A. The Shapiro-Wilk statistic for normality has a p-value of 0.2650; since the p-value is > 0.05, it is reasonable to assume that the error terms follow a normal distribution. The residual analysis graphs and Shapiro-Wilk criterion are located in Appendix D.

Model A diagnostic statistics—DFFITS, DFBETAS, and Cook’s Distance—were used to evaluate the effect on the parameter estimates of the regression for deleting a single observation. The Cook’s distance plot showed the presence of multiple outliers. Multiple observations exceed the cutoff, $4/n$, 0.036. The results of the DFBETAS statistics for Model A are graphically presented in Appendix E. Several of the observations exceed the cutoff, $2/\sqrt{n}$, 0.191565. The DFFITS also showed multiple observations exceeding the cutoff value of, $2\sqrt{p/n}$, 0.331801. The DFFITS, DFBETAS, and Cook’s Distance results indicate that these observations are influential in estimating the given parameters and should be further examined.

The outlier and leverage diagnostics plot for Model A displays observations (16, 28, 51, 53, 73, and 108) as outliers and observations (6, 9, 11, 19, 35, 45, 75, 100, and 102) as leverage points. Due to the presence of outliers and leverage points in the regression model, a more robust model was chosen based on robust regression methods.

The robust regression parameter estimates for LTS-estimation, maximum likelihood or M-estimation; S-estimation and MM-estimation were obtained using a residual cutoff value of 3 and a leverage cutoff value of 3.058. The R_{adj}^2 values are highest for LTS-estimation; however, the standardized robust residuals did not follow a normal distribution; outliers (28, 51, and 16) were present in the model and observation 19 proved to be an outlier with leverage as well, because 25% of the data is lost using this robust regression technique, LTS-estimation was not selected to optimize the parameter estimates. M-estimation was not selected as the optimum robust regression estimation method due to the fact that an outlier

was still present in the model: observation 73. MM-estimation did eliminate all of the outliers in the model and the standardized robust residuals appear to follow a normal distribution. However, employing this method produced a low R_{adj}^2 value: 0.6015. S-estimation was successful in eliminating all of the outliers and influential observations from the model. The standardized robust residuals follow a normal distribution and the R_{adj}^2 value: 0.8022. Ultimately, S-estimation was chosen as the optimum robust regression estimation method. The plots used for detecting outliers and leverage points for the robust regression were the robust residuals against the robust distances (RD PLOT), robust distances against the classical Mahalanobis distances (DD PLOT) i.e. (RD plot, DD plot; QQ plot), using the different estimation methods is located in Appendix F. The robust regression parameter estimates are located in Table 5.

Table 5.

Natural Logarithmically Transformed OLS and Robust Parameter Estimates for Model A.

Parameter	OLS	LTS	S	M	MM
Intercept	3.78236	5.0544	4.6236	4.1117	4.3851
Institutional Research Net Assignable Square Footage	0.17980	0.0350	0.1583	0.1825	0.1788
Institutional Research Expenditures	0.16249	0.2626	0.1805	0.1600	0.1636
Institutional Full-time Employees	0.58742	0.4473	0.4891	0.5524	0.5203
R_{adj}^2	0.7838	0.8420	0.8022	0.6659	0.6015

Visual assessment of the model was conducted by using bivariate histograms comparing environmental health and safety expenditures with each of the independent variables: institutional full-time employees, institutional research net assignable square footage, and institutional research expenditures; these are located in Figures 6-8. The bivariate histograms show that the data is primarily centered in the middle portion of the plot with a peak point near the center; this means that the data is linearly related, and a linear model should be used in the regression analysis. The 3D surface plots are located in Figures

9 and 10. The surface plots are close to a flat plane, but do show disturbances in the plane; however, these surface plots show that a linear model is best in the regression analysis.

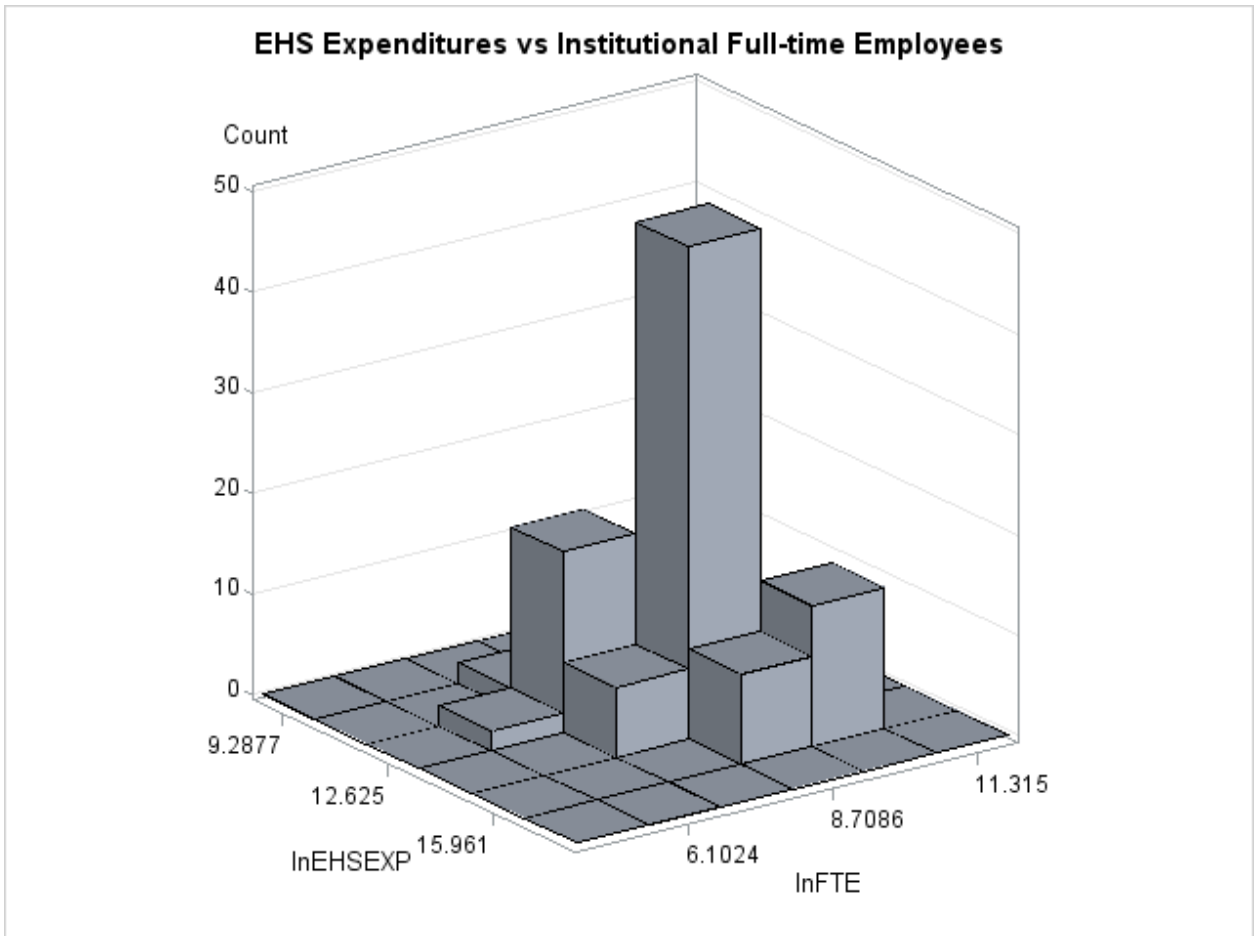


Figure 6. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Full-time Employees.

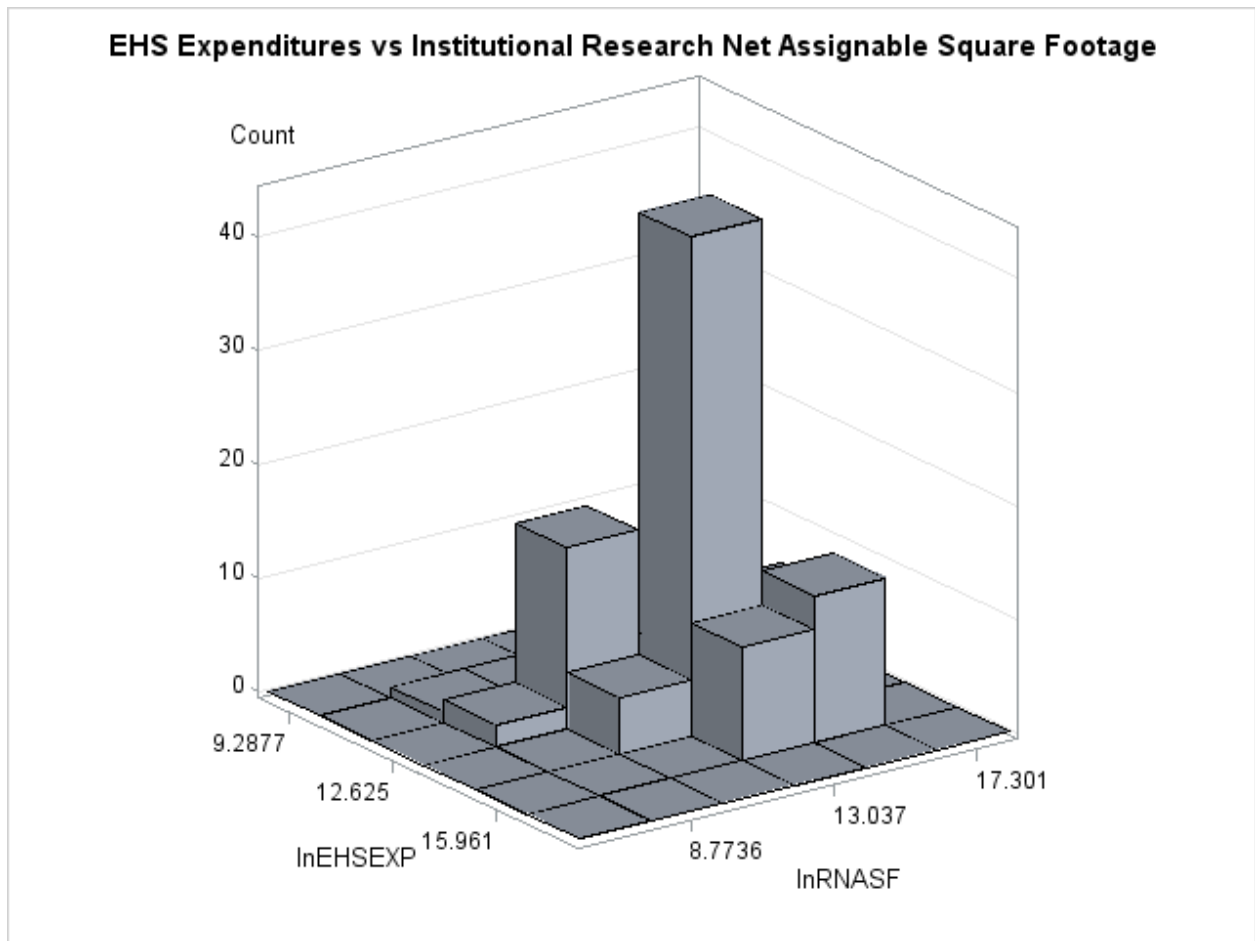


Figure 7. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Research Net Assignable Square Footage.

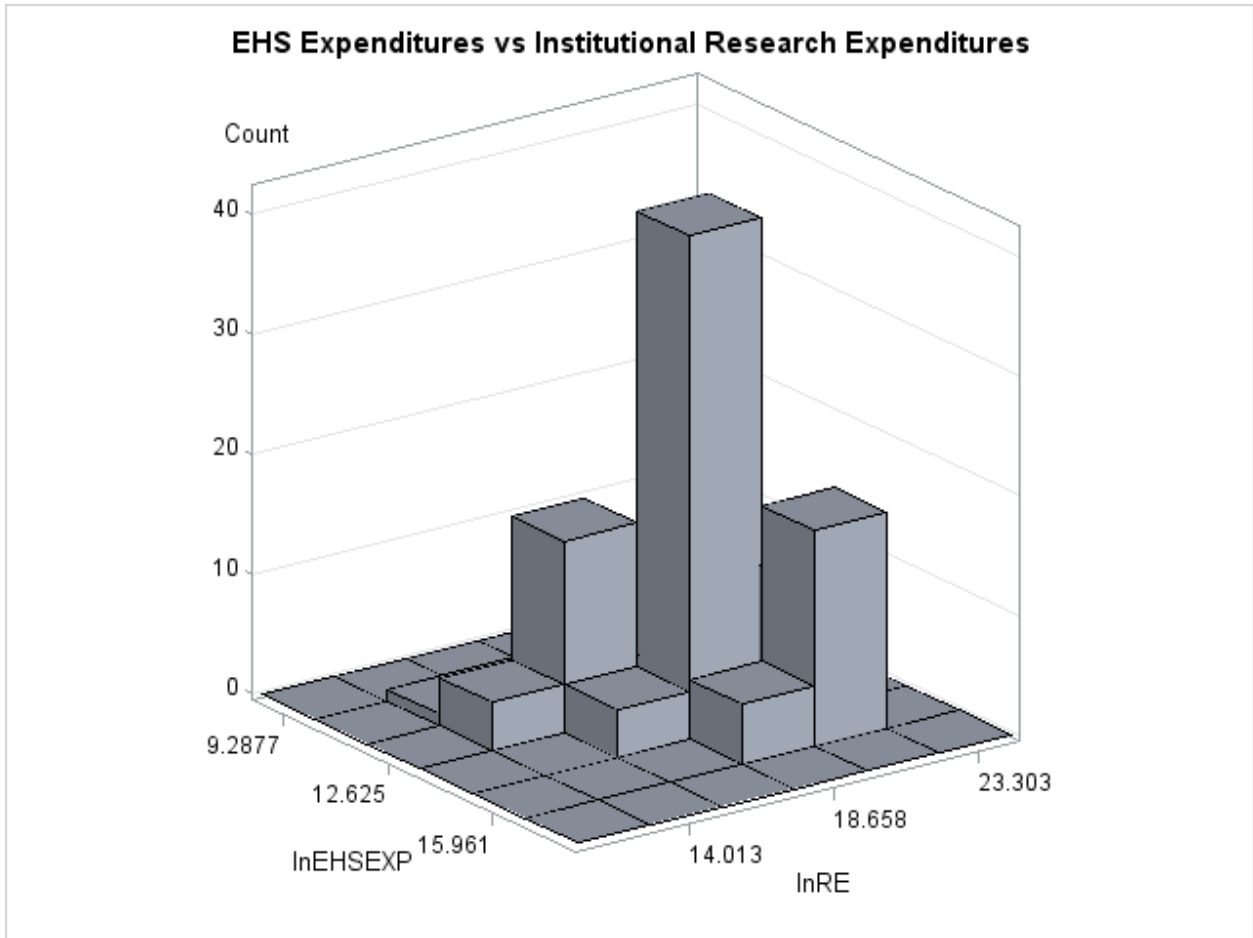


Figure 8. Bivariate Histogram for the Natural Logarithm of Environmental Health and Safety Expenditures and the Natural Logarithm of Institutional Research Expenditures.

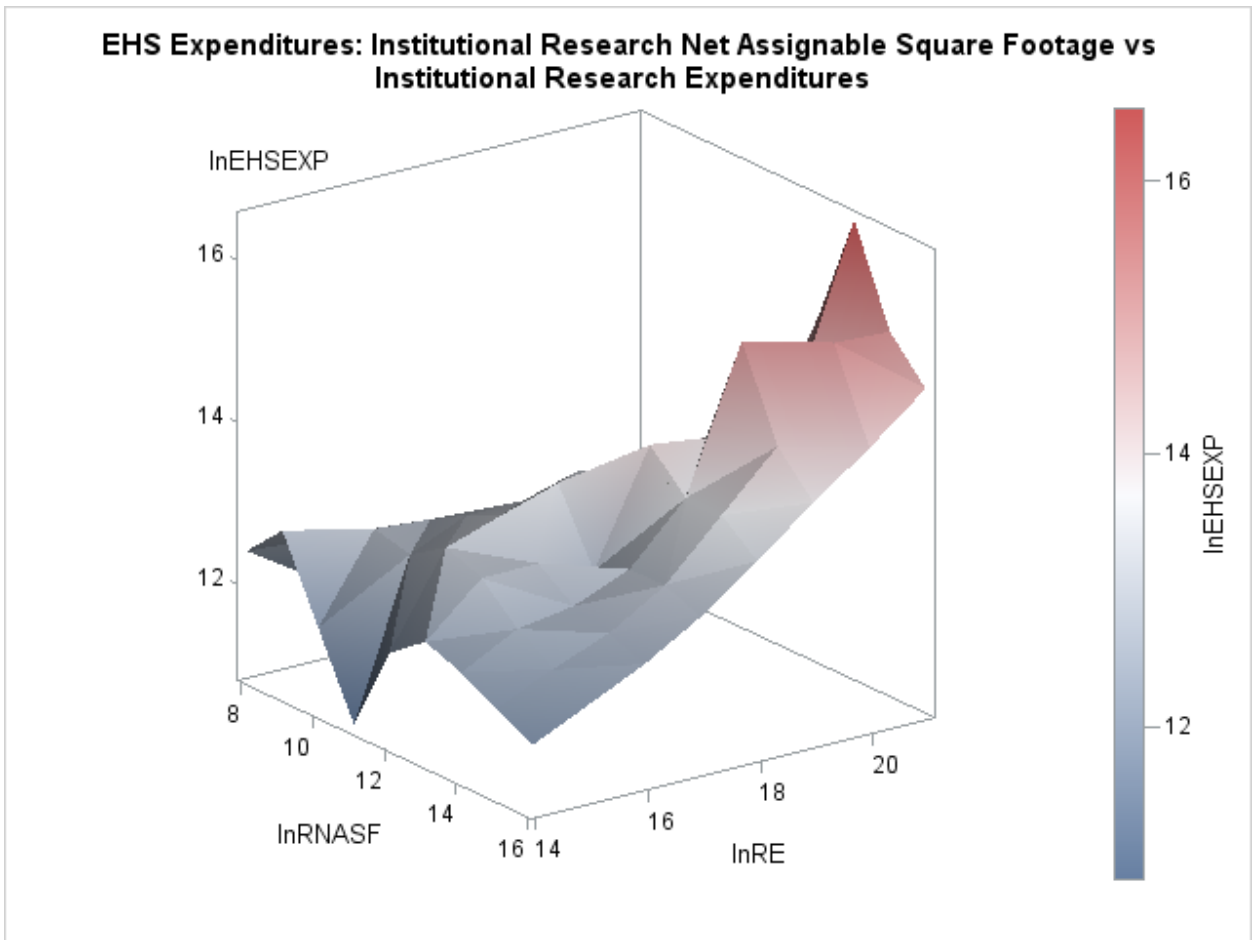


Figure 9. 3D Surface Plot for Institutional Research Net Assignable Square Footage and Institutional Research Expenditures as the Independent Variables and Environmental Health and Safety Expenditures as the Dependent Variable.

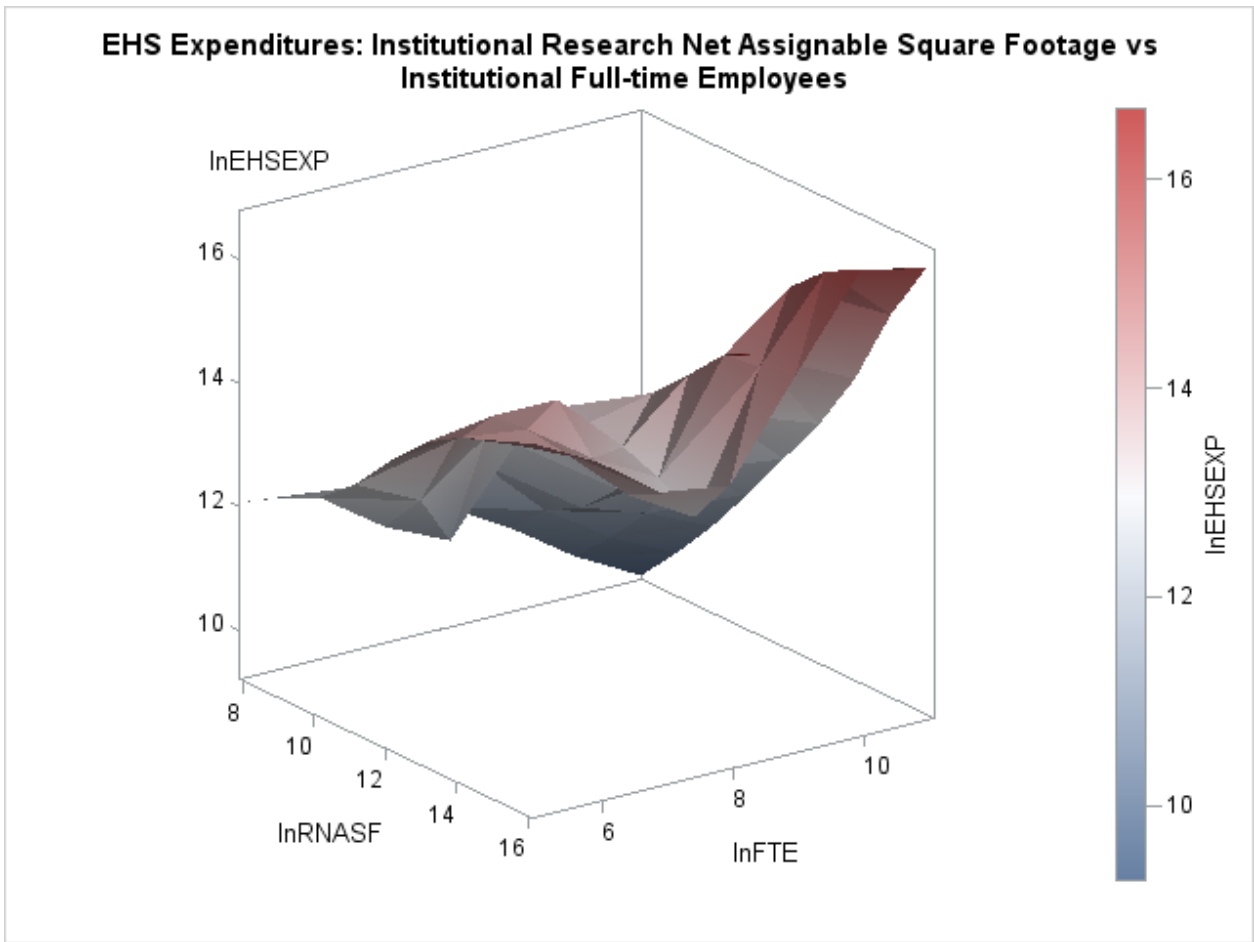


Figure 10. 3D Surface Plot for Institutional Research Net Assignable Square Footage and Institutional Full-time Employees as the Independent Variables and Environmental Health and Safety Expenditures as the Dependent Variable.

Cross-validation was simultaneously used, along with information criteria, to find the “best” fitting model for environmental health and safety full-time employees. Cross-validation for environmental health and safety full-time employees as the dependent variable was performed based on K-fold cross-validation where the number of folds (K) was equal to 5. The fit criteria for environmental health and safety full-time employees as the dependent variable are shown in Figure 11.

Similarly, a regression model was selected for the outcome variable environmental health and safety full-time employees using total institutional net assignable square footage, total institutional expenditures, research net assignable square footage, research expenditures, total number of enrolled students and total institutional full-time employees. Parsimony, along with information criteria, was used to produce the final models. The models selected for the dependent variable environmental health and safety full-time employees are listed in Table 6.

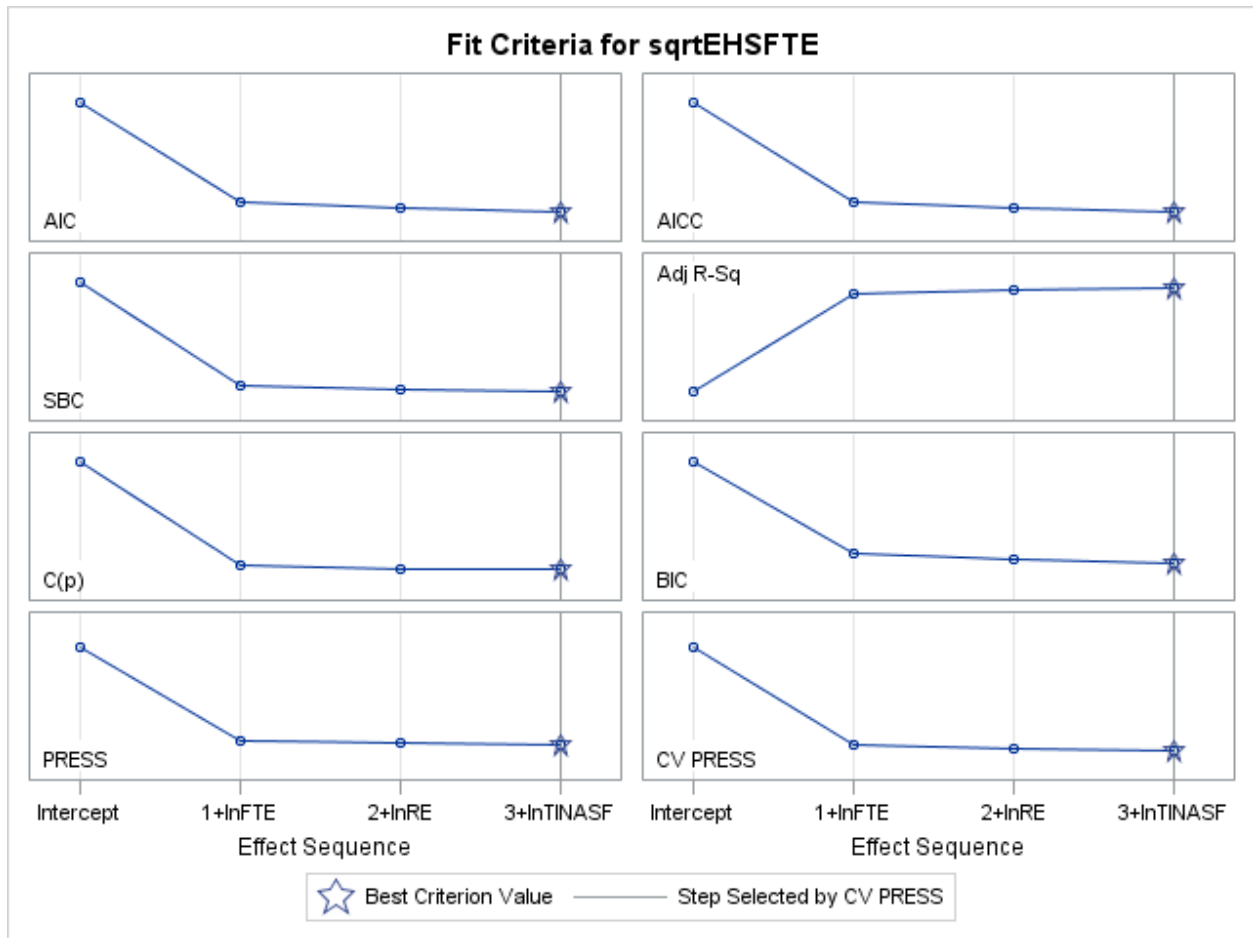


Figure 11. Fit Criteria using 5-fold Cross-Validation for the Square Root of Environmental Health and Safety Full-time Employees as the Dependent Variable.

Table 6.

Multiple Regression Models Selected with Environmental Health and Safety Full-time Employees as the Dependent Variable.

Model D:	$EHSFTE = (\beta_0 + \beta_1 \ln TINASF + \beta_2 \ln FTE + \beta_3 \ln RE)^2$
----------	--

Model E:	$EHSFTE = (\beta_0 + \beta_1 \ln TINASF + \beta_2 \ln RE)^2$
----------	--

Model F:	$EHSFTE = (\beta_0 + \beta_1 \ln TINASF + \beta_2 \ln FTE)^2$
----------	---

Once the regression algorithm concluded, a significant model emerged ($F = 112.10$; $p < 0.0001$; $R_{adj}^2 = 0.7553$). Indicating that the total institutional net assignable square footage, institutional full-time employees and institutional research expenditures accounts for 75.53% of the variance for the dependent variable environmental health and safety full-time employees. Conversely, total institutional expenditures, total number of enrolled students and total institutional research net assignable square footage were not significant in the model and hence excluded from the final model. The parameter estimates with variance inflation factors (VIF) for Model D are located in Table 7. The analysis revealed that none of the predictor variables had a VIF value greater than 10; this suggests that multicollinearity is not an issue among the independent variables.

Table 7.

Natural Logarithmically Transformed Parameter Estimates and Associated Variance Inflation Factor (VIF) with p-values for Model D.

Parameter	Parameter Estimates	Standard error	t Value	Pr > t	VIF
Intercept	20.36161	2.14381	9.50	<.0001	0
Total Institutional Net Assignable Square Footage	0.57887	0.21631	2.68	0.0086	3.11128
Institutional Research Expenditures	0.42650	0.11937	3.57	0.0005	4.16569
Institutional Full-time Employees	0.92579	0.31181	2.97	0.0037	6.49526

Residual analysis for model D does not display any curves or cyclical patterns in the plots of the residuals over each of the independent variables in the model; this suggests that all the structure in the data is essentially captured by the multiple regression model. The quantile-quantile plot, Q-Q plot, shows light tails with a description of the point pattern. All but a few points fall on the line, suggesting the presence of outliers in Model D. The Shapiro-Wilk statistic for normality has a p-value of 0.7043; since the p-value is > 0.05, it is reasonable to assume that the error terms follow a normal distribution. The residual analysis graphs and Shapiro-Wilk criterion are located in Appendix G.

Model D diagnostic statistics—DFFITs, DFBETAS, and Cook’s Distance—were used to evaluate the effect on the parameter estimates of the regression for deleting a single observation. The Cook’s distance plot also showed some outliers present. The results of the DFBETAS statistics for Model D are graphically presented in Appendix H. Several of the observations exceed the cutoff, $2/\sqrt{n}$, 0.191565. The DFFITS also showed multiple observations exceeding the cutoff value of $2\sqrt{p/n}$ 0.331801, which indicates that these observations are influential in estimating the given parameters and should be further examined. The outlier and leverage diagnostic plots for Model D display observations (33, 37, 46, 52, 53 and 96) as outliers and observations (15, 19, 24, 28, 35, 45 and 102) as leverage points. Observation 100 had both outlier and leverage on the multiple regression. Due to the presence of outliers and leverage points in the regression model, a more robust model (S-estimation) was chosen.

Based on robust regression methods—namely least trimmed squares estimation, maximum likelihood or M-estimation, S-estimation and MM-estimation—S-estimation proved to be the ideal method for the parameter estimates. The output for the robust regression (i.e., RD plot, DD plot; QQ plot) using the different estimation methods is located in Appendix I. The robust regression parameter estimates are located in Table 8.

Table 8.

Natural Logarithmically Transformed OLS and Robust Parameter Estimates for Model D.

Parameter	OLS	LTS	S	M	MM
Intercept	20.36161	22.4097	20.3919	20.1610	20.1756
Total Institutional Net Assignable Square Footage	0.57887	0.7558	0.6437	0.5970	0.6178
Institutional Full-time Employees	0.92579	1.1026	1.0039	0.9576	0.9831
Institutional Research Expenditures	0.42650	0.3086	0.3368	0.3848	0.3555
R^2_{adj}	0.7621	0.8336	0.7857	0.6818	0.6229

Visual assessment of the model using bivariate histograms with environmental health and safety full-time employees compared with each of the independent variables— institutional full-time employees, total institutional net assignable square footage, and institutional research expenditures—are located in Figures 12-14. The bivariate histograms show that the data is primarily centered in the middle portion of the plot, with a peak point near the center; this means that the data is linearly related, and a linear model should be used in the regression analysis. The 3D surface plots are located in Figures 15 and 16. While the surface plots are close to a flat plane, they do show disturbances in the plane; however, these surface plots show that a linear model is best in the regression analysis.

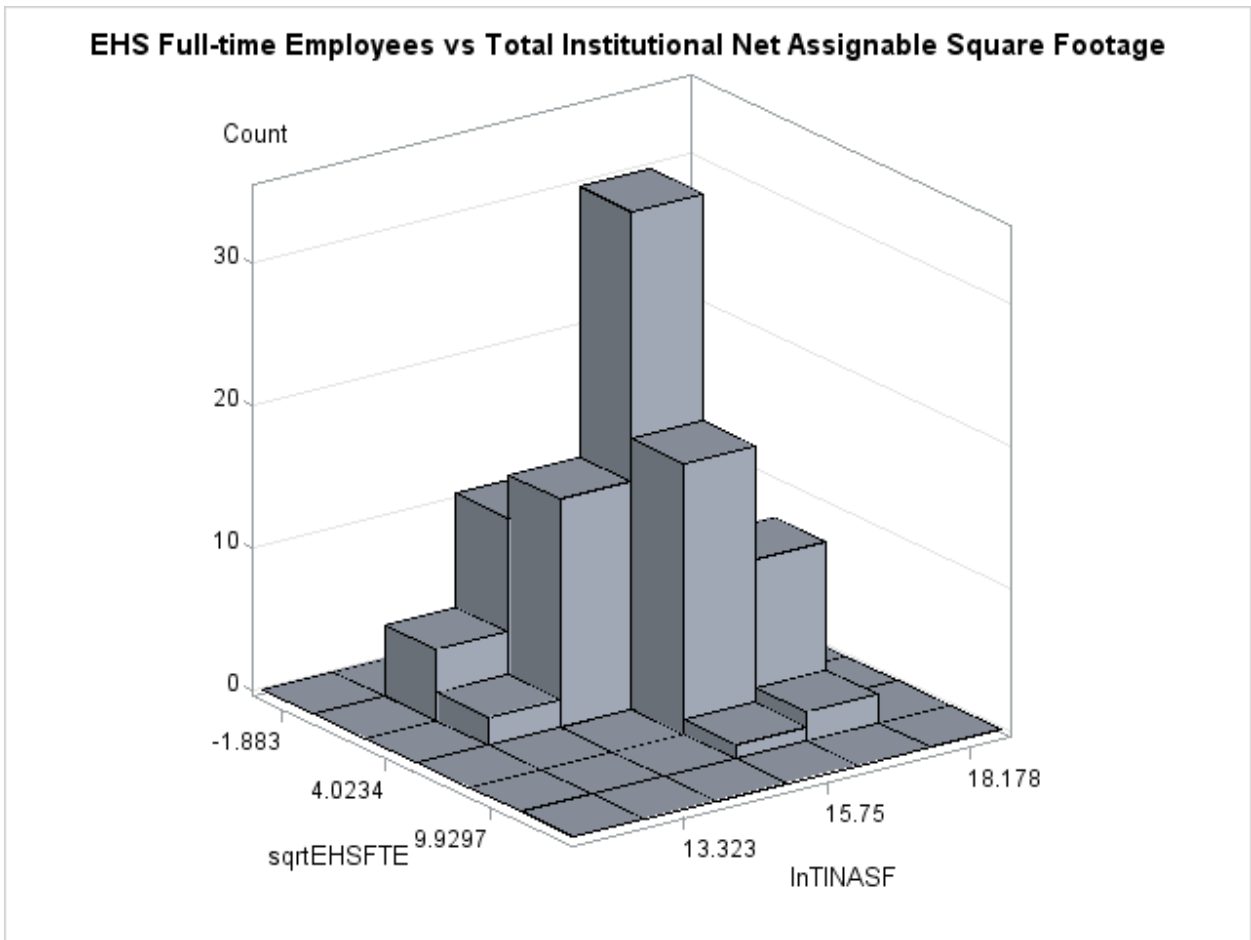


Figure 12. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Total Institutional Net Assignable Square Footage.

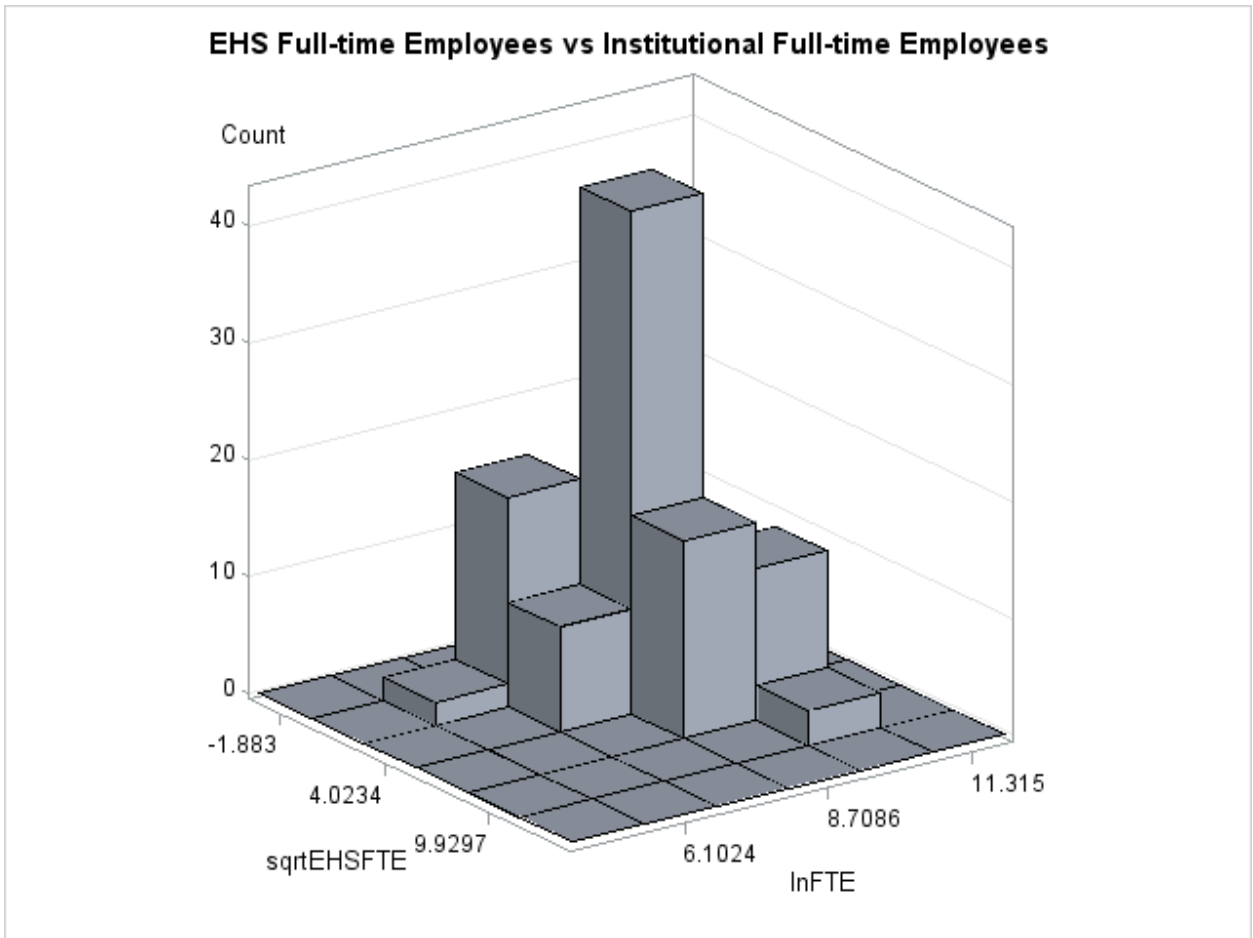


Figure 13. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Institutional Full-time Employees.

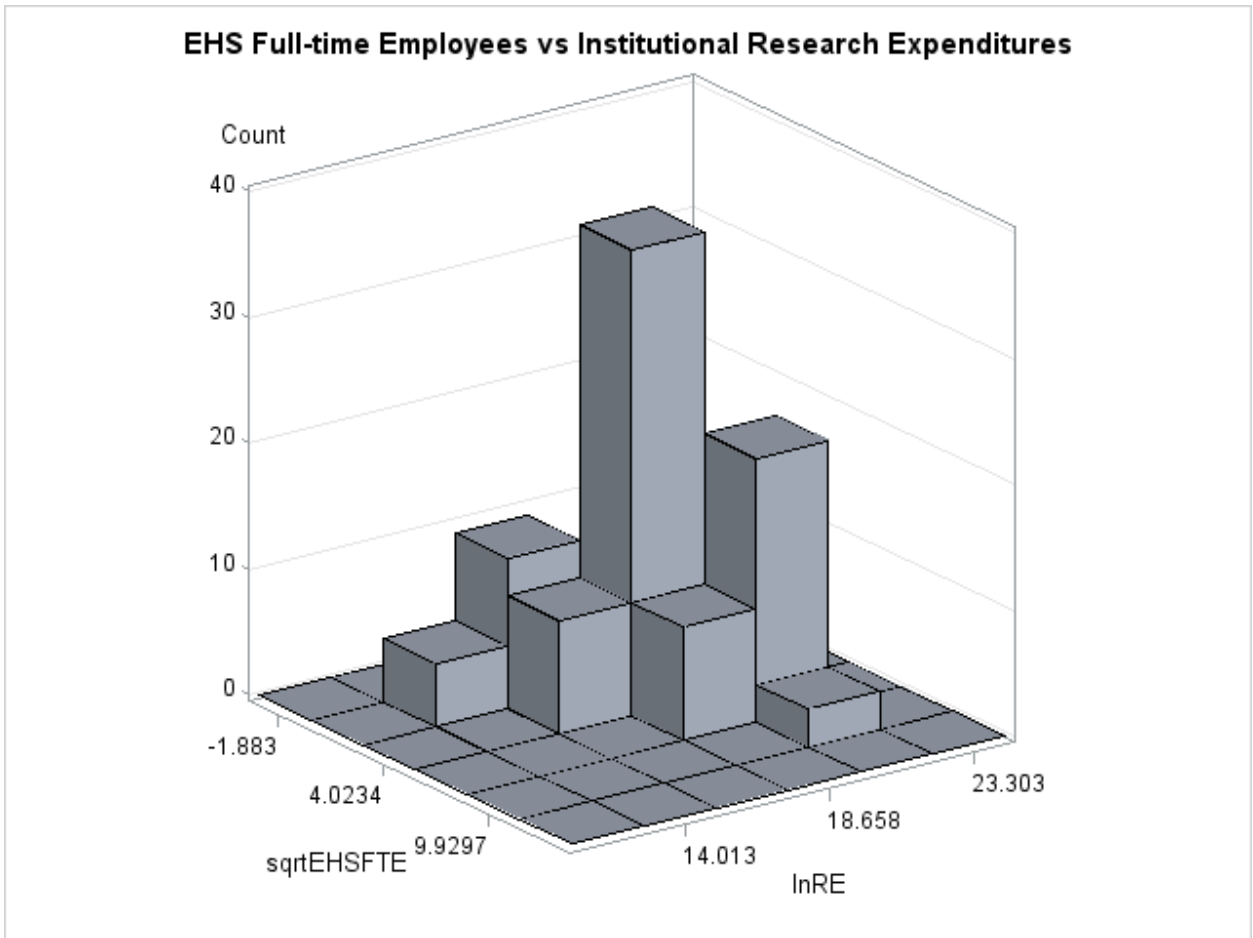


Figure 14. Bivariate Histogram for the Square Root of Environmental Health and Safety Full-time Employees and the Natural Logarithm of Institutional Research Expenditures.

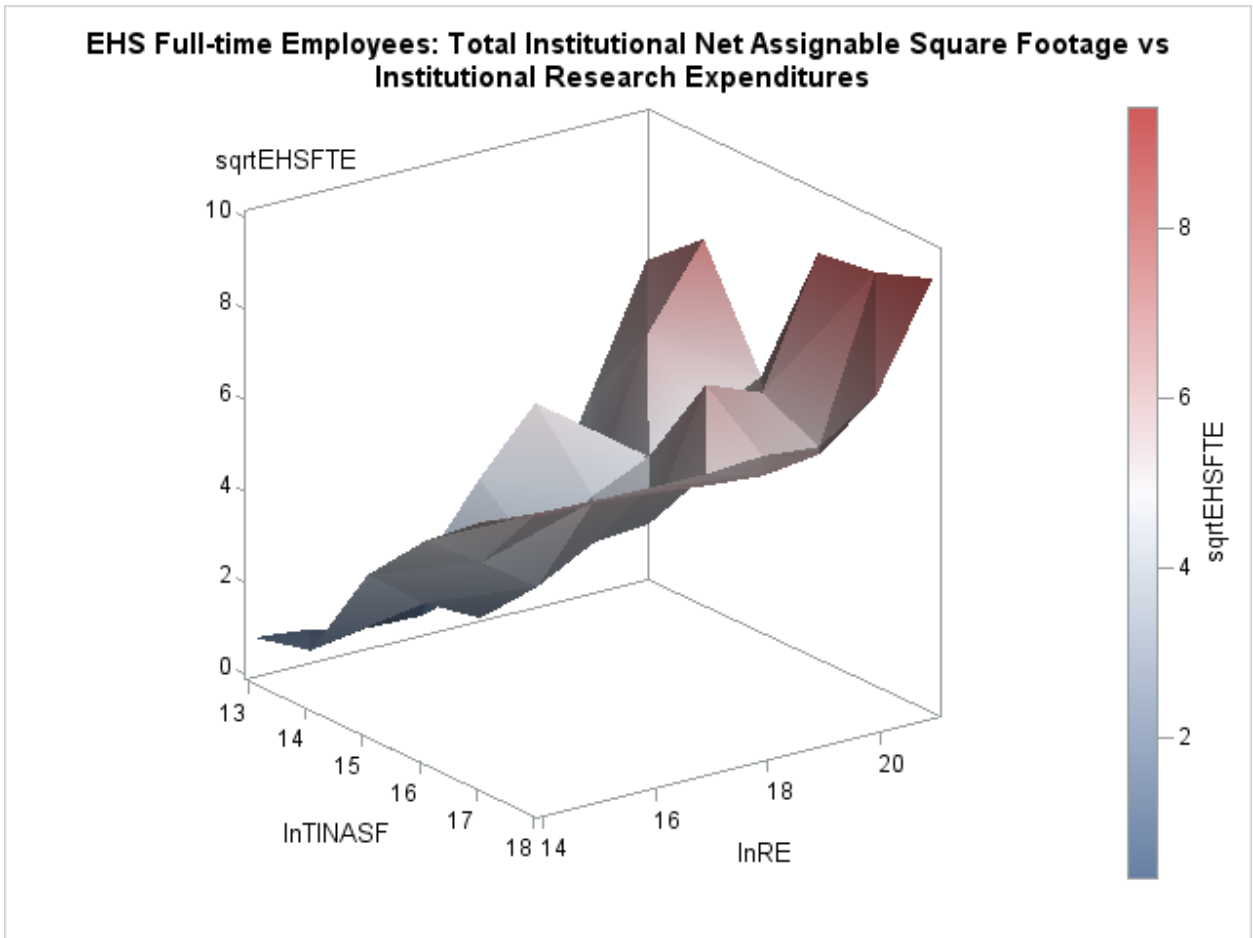


Figure 15. 3D Surface Plot for Total Institutional Net Assignable Square Footage and Institutional Research Expenditures as the Independent Variables and Environmental Health and Safety Full-time Employees as the Dependent Variable.

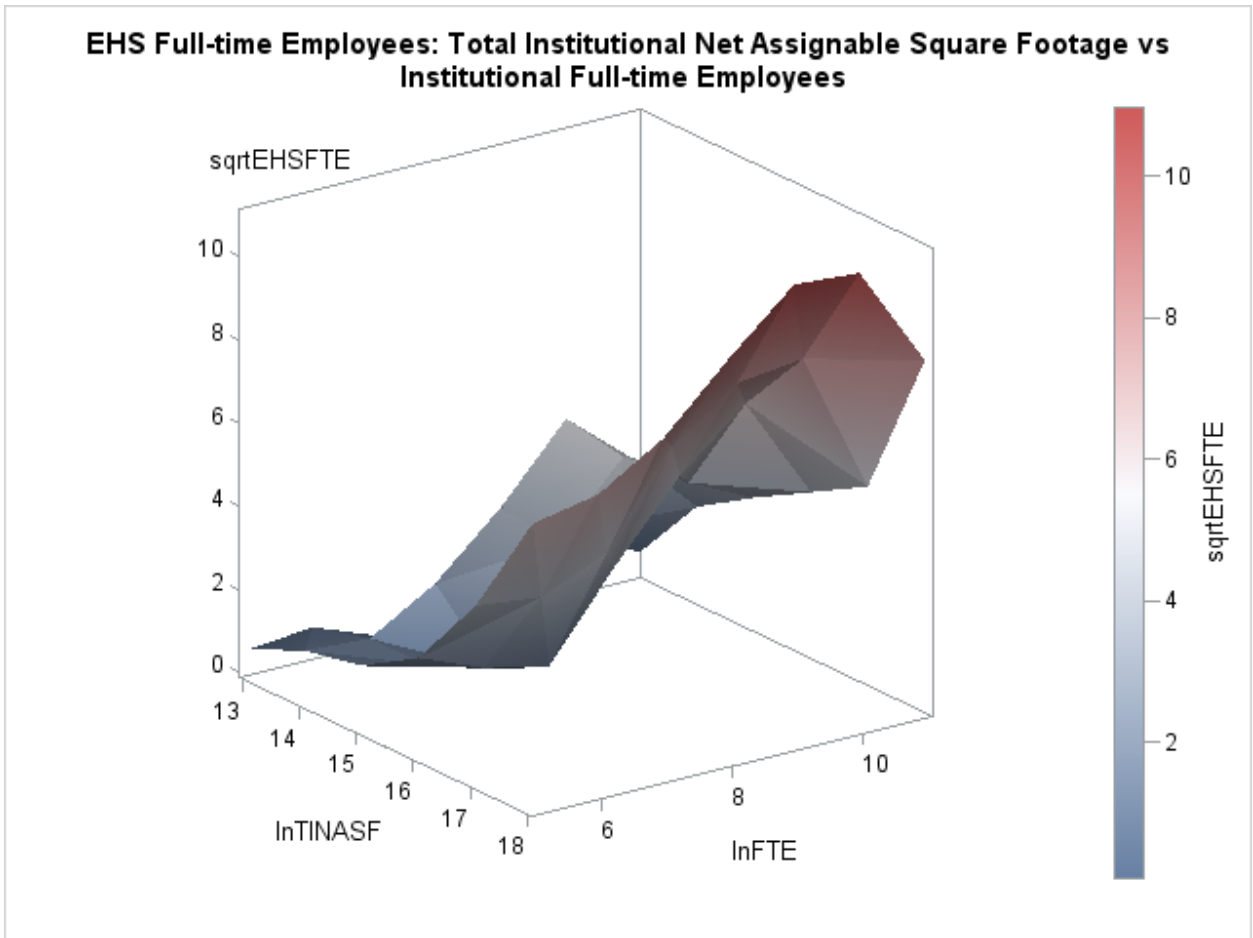


Figure 16. 3D Surface Plot for Total Institutional Net Assignable Square Footage and Institutional Full-time Employees as the Independent Variables and Environmental Health and Safety Full-time Employees as the Dependent Variable.

DISCUSSION

The purpose of this research project was to identify the factors that affect university environmental health and safety expenditures as well as environmental health and safety full-time employees. The research objectives were accomplished by simultaneously using cross-validation and information criteria to select the “best” fitting models based upon the independent variables: total institutional net assignable square footage, institutional research net assignable square footage, institutional research expenditures, total institutional expenditures, institutional full-time employees and total institutional number of enrolled students. The presence of outliers and influential observations in the selected models was then followed by using more robust and resistant regression methods to optimize parameter estimates. The robust regression estimators used in finding the optimum parameter estimates include maximum likelihood or M-estimation, LTS-estimation, S-estimation, and MM-estimation.

The residual analysis for the dependent variable, environmental health and safety expenditures, shows that a linear model is appropriate for the data; the Shapiro-Wilk test had a p-value of 0.2650, which was not significant, indicating that the residuals are normally distributed. Expressing environmental health and safety expenditures (Y_1) as a function of the natural logarithm of institutional research net assignable square footage ($\ln\text{RNASF}$), natural logarithm of institutional research expenditures ($\ln\text{RE}$), and natural logarithm of institutional full-time employees ($\ln\text{FTE}$), results in the following regression model:

$$\hat{y}_1 = \text{EHSEXP} = e^{(\beta_0 + \beta_1 \ln\text{RNASF} + \beta_2 \ln\text{FTE} + \beta_3 \ln\text{RE})}$$

Similarly, the model with environmental health and safety full-time employees had a non-significant p-value, 0.7043, for the Shapiro-Wilk test, displaying normality among the residuals. Expressing environmental health and safety full-time employees (Y_2) as a function of the natural logarithm of total institutional net assignable square footage ($\ln\text{TINASF}$), natural logarithm of institutional full-time employees ($\ln\text{FTE}$), and natural logarithm of institutional research expenditures ($\ln\text{RE}$) results in the following regression equation:

$$\hat{y}_2 = \text{EHSFTE} = (\beta_0 + \beta_1 \ln\text{TINASF} + \beta_2 \ln\text{FTE} + \beta_3 \ln\text{RE})^2.$$

However, the assumptions of normality were met with the multiple regression models; numerous outliers and influential observations were present in the CSHEMA vital statistics data set. Robust regression methods were employed to facilitate the optimum coefficients for the model. Among the robust regression procedures, M-Estimation, LTS-estimation, S-estimation, and MM-estimation (for the model with environmental health and safety expenditures as the dependent variable using LTS-estimators) produced the highest R_{adj}^2 value: 0.8420. However, LTS-estimation was not selected as the best robust estimator, due to the fact, that several outliers and influential observations remained present; further, 25% of the data was trimmed from the model. Therefore, the models based on LTS-estimation would not possess enough sample data to have adequate power for a three variable regression model. S-estimation was eventually settled on as the optimum robust regression method to obtain the parameter estimates. The outliers and influential observations were eliminated after applying S-estimators. The S-estimation robust regression method produced an R_{adj}^2 value: 0.8022. The revised parameter estimates using S-estimators resulted in the following regression equation.

Resourcing Model

$$\text{EHS expenditures} = e^{[4.6236 + 0.1583(\ln \text{ research net assignable sf}) + 0.4891(\ln \text{ institutional full-time employees}) + 0.1805(\ln \text{ institutional research expenditures})]}$$

Among the robust regression estimators, with environmental health and safety full-time employees as the dependent variable, LTS-estimators produced the highest R_{adj}^2 value: 0.8336. Again, LTS-estimation was not chosen due to the fact that 25% of the data was eliminated from the model. S-estimators were chosen to find the optimum parameter estimates. The S-estimation robust regression method produced an R_{adj}^2 value: 0.7857. Revised parameter estimates using S-estimators for the model with environmental health and safety full-time employees.

Staffing Model

$$\begin{aligned} \text{EHS FTE staff} = & [-20.3919 + (0.6437 \times (\ln \text{ total institutional assignable sf})) \\ & + (1.0039 \times (\ln \text{ institutional full - time employees})) \\ & + (0.3368 \times (\ln \text{ institutional research expenditures}))]^2 \end{aligned}$$

The summary of reported and modeled values for environmental health and safety expenditures with values for institutional research net assignable square footage, representing small, medium, and large academic institutions, are listed in Tables 9-11. Similarly, the summary of reported and modeled values for environmental health and safety full-time employees is located in Tables 12-14. A comparison between the information criteria used in model selection between the current and previous models is located in Table 15.

Table 9.

Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 56,029ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
1,152,180ft ²	56,029ft ²	\$246,443,516	\$1,159,100	1,689
Reported EHS Expenditures	\$322,000			
Modeled EHS Expenditures				
Model A				
OLS	LTS	S	M	MM
\$346,937.52	\$457,119.40	\$410,569.29	\$367,691.66	\$387,666.20
Model B				
OLS	LTS	S	M	MM
\$371,870.14	\$391,288.30	\$427,154.05	\$396,527.23	\$408,072.40
Model C				
OLS	LTS	S	M	MM
\$347,111.03	\$426,428.50	\$395,616.26	\$366,626.90	\$383,425.30

Table 10.

Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 511,000ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
9,211,656ft ²	511,000ft ²	\$956,296,016	\$230,411,000	6,547
Reported EHS Expenditures	\$1,855,870			
Modeled EHS Expenditures				
Model A				
OLS	LTS	S	M	MM
\$1,859,838.10	\$1,985,140.00	\$1,938,645.32	\$1,877,027.56	\$1,899,688.00
Model B				
OLS	LTS	S	M	MM
\$1,870,282.41	\$2,156,506.00	\$1,972,672.93	\$1,916,861.94	\$1,936,320.00
Model C				
OLS	LTS	S	M	MM
\$1,765,068.40	\$1,823,193.00	\$1,812,649.24	\$1,779,779.43	\$1,796,409.00

Table 11.

Summary of Reported and Modeled Values for Environmental Health and Safety Expenditures with an Institutional Research Net Assignable Square Footage of 26,324,500ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
15,053,079ft ²	26,324,500ft ²	\$956,296,016	\$982,357,000	18,057
Reported EHS Expenditures	\$8,000,000			
Modeled EHS Expenditures				
Model A				
OLS	LTS	S	M	MM
\$8,677,985.72	\$5,250,322.00	\$7,721,351.72	\$8,512,106.48	\$8,261,362.00
Model B				
OLS	LTS	S	M	MM
\$8,559,052.05	\$6,844,801.00	\$8,795,053.33	\$8,902,119.92	\$8,951,216.00
Model C				
OLS	LTS	S	M	MM
\$11,587,065.60	\$10,042,132.00	\$10,984,546.70	\$11,201,967.60	\$11,023,060.00

Table 12.

Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 56,029ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
1,152,180ft ²	56,029ft ²	\$246,443,516	\$11,591,000	1,689
Reported EHS Full-time Employees	4.0			
EHS Full-time Employees				
Model D				
OLS	LTS	S	M	MM
2.357	1.831	2.345	2.394	2.359
Model E				
OLS	LTS	S	M	MM
1.99	1.857	1.971	2.035	2.02
Model F				
OLS	LTS	S	M	MM
2.79	1.647	2.509	2.772	2.676

Table 13.

Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 511,000ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
9,211,656(ft ²)	511,000(ft ²)	\$956,296,016	\$230,411,000	6,547
Reported EHS Full-time Employees	35.0			
EHS Full-time Employees				
Model D				
OLS	LTS	S	M	MM
27.7	28.5	27.4	27.4	27.2
Model E				
OLS	LTS	S	M	MM
29.7	28.2	29.8	29.5	29.6
Model F				
OLS	LTS	S	M	MM
26.5	27.3	26.3	26.4	26.3

Table 14.

Summary of Reported and Modeled Values for Environmental Health and Safety Full-time Employees with an Institutional Research Net Assignable Square Footage of 26,324,500ft².

Total Institutional Net Assignable Square Feet(ft ²)	Research Net Assignable Square Feet(ft ²)	Total Institutional Expenditures	Research Expenditures	Full-time Employees
15,053,079ft ²	26,324,500ft ²	\$956,296,016	\$982,357,000	18,057
Reported EHS Full-time Employees	63			
EHS Full-time Employees				
Model D				
OLS	LTS	S	M	MM
50.5	52.9	49.8	49.8	49.4
Model E				
OLS	LTS	S	M	MM
47.8	43.8	46.9	47.0	46.8
Model F				
OLS	LTS	S	M	MM
49.4	52.4	50.3	49.2	49.5

Table 15.

Information Criteria for Competing Models.

Information criteria	Model A	Model X	Model D	Model Z
R ²	0.7838	0.6740	0.7621	0.6880
R _{adj} ²	0.7776	0.6679	0.7553	0.6821
AIC	-143.0695	-100.3081	10.4801	-98.6025
BIC	-140.7542	-98.1399	12.7679	-98.7975
SBC	-132.3041	-92.23404	21.24549	-90.52846
SSE	27.25915	41.10149	111.50853	41.74968
RMSE	0.50952	0.62270	1.03053	0.62759
Mallow's (C)p	3.8266	3.0000	4.1813	59.5939

Note. Model A: $\text{EHSEXP} = e^{(\beta_0 + \beta_1 \ln \text{RNASF} + \beta_2 \ln \text{FTE} + \beta_3 \ln \text{RE})}$; Model X: $\text{EHSEXP} = e^{(\beta_0 + \beta_1 \ln \text{TINASF} + \beta_2 \ln \text{TIE})}$; Model D: $\text{EHSFTE} = (\beta_0 + \beta_1 \ln \text{TINASF} + \beta_2 \ln \text{FTE} + \beta_3 \ln \text{RE})^2$; and Model Z: $\text{EHSFTE} = e^{(\beta_0 + \beta_1 \ln \text{TINASF} + \beta_2 \ln \text{TIE})}$.

The model selected for environmental health and safety expenditures as the dependent variable—based on fit criteria AIC, SBC, BIC, R_{adj}², PRESS, and Mallow's (C)p—showed significant improvement from the former model. The Mallow's (C)p values were small and close to the number of predictors plus the intercept in the model, 4, which is an indicator that the model has small variance and greater precision in estimating the beta coefficients, therefore minimizing the bias in the model. The R_{adj}² value from the previous model, 0.6679, increased to 0.7776 with the new model. The difference of 0.1097 indicates that the newer model explains 10% more of the variation in the dependent variable than the older model; this suggests that the two models are not statistically comparable. The AIC

values from the previous model, -100, decreased to -143 with the new model. Examining the models with the dependent variable environmental health and safety full-time employees, the Mallows's (C)_p value of 59.5939, which is substantially greater than 3, in model Z would suggest lack-of-fit and bias in estimating the regression coefficients. The Mallows's (C)_p value of 4.1813 in model D, given three independent variables including the constant, suggests that the lack-of-fit and bias have been reduced in the model, since the Mallows's (C)_p value is slightly greater than 4.

Model Validation

Empirical evaluation of the multivariable prediction models was based on performance evaluations with data that were not used in model development. Therefore, external model validation was used to assess the predictive ability of the multivariable models. Institutions used in the external model validation were de-identified. It should be made clear that the performance measures were based on the full regression models, rather than on the simplified versions. The models compared in the validation process include those created by Wang, Emery and Brown. The Wang models emphasize non-lab and lab net assignable square footage, presence of medical or veterinary school, and presence of biosafety level 3 facility. Brown's models are based on total institutional net assignable square footage, total institutional expenditures, and whether the institution is part of the Association for Academic Health Centers. The reported and modeled values comparing the various models developed by Emery and Brown for environmental health and safety full-time employees and expenditures are located in Tables 16 and 17, respectively.

Table 16.

Reported and Modeled Values for Environmental Health and Safety Full-time Employees.

De-identified Universities	Reported	Modeled Parker	Modeled Wang	Modeled Brown
University I	44	41.54	35	44*
University II	16	14	10.4	N/A
University III	25	40	42	35*
University IV	19	18.24	N/A	20.7
University V	1	2.57	2.4	3
University VI	N/A	20.114	12.5	19.7
University VII	49	35.15	39	43
University VIII	18	23.2	25.8	35.2

Note.

Wang model: Based on Staffing Model relying on non-lab and lab NASF, presence of Med or Vet School, and presence of BSL3 labs $EHS\ FTE = e^{[(0.516 \times Med/Vet\ School) + (0.357 \times (\ln\ Lab\ sq.\ ft)) + (0.398 \times (\ln\ Non\ lab\ sq.\ ft)) + (0.371 \times BSL) - 8.618]}$

Brown model: Based on most recent resourcing model:

$EHS\ expenditures = e^{(-7.8093 + (0.50506 \times \ln\ TNASF) + [0.69283 \times (\ln\ total\ institutional\ expenditures)])}$

Parker model: Based on S-estimation:

$EHS\ FTE\ Staff = (-20.3919 + 0.6437 \ln TNASF + 1.0039 \ln FTE + 0.3368 \ln RE)^2$

*Brown model: Based on new AAHC staffing model: $EHS\ FTE\ staff = e^{(-9.039 + 0.7899 \times \ln\ (total\ institutional\ assignable\ sf))}$.

Table 17.

Reported and Modeled Values for Environmental Health and Safety Expenditures.

De-identified Universities	Reported	Modeled Parker	Modeled Emery	Modeled Brown
University I	\$2,770,520	\$3,178,257	\$5,170,203	\$4,114,385****
University II	\$1,200,000	\$985,682	\$1,563,000**	N/A
University III	\$1,669,597	\$2,935,480	\$3,356,335**	\$3,210,976****
University IV	N/A	\$1,345,062	\$1,613,900**	\$1,837,653
University V	\$162,100	\$425,346	\$244,892***	\$276,509
University VI	N/A	\$1,406,789	\$1,918,131**	\$1,789,953
University VII	\$3,972,334	\$2,983,272	\$4,326,012*	\$3,913,724****
University VIII	\$1,806,268	\$1,764,538	\$2,708,892	\$3,210,976****

Note. Emery model: Based on TNASF model: TNASF x \$0.45/square feet (for higher density lab square footage peak on frequency histogram); Brown model: Based on new non-AAHC resourcing model: EHS expenditures = $e^{-7.8093 + 0.50506 * \ln(\text{total institutional assignable sf}) + 0.69283 * \ln(\text{total institutional expenditures})}$.

*Emery model: Based on TNASF model: TNASF x \$0.40/square feet (for higher than average density lab square footage peak on frequency histogram).

**Emery model: Based on TNASF model: TNASF x \$0.30/square feet (for average lab density square footage).

***Emery model: Based on TNASF model: TNASF x \$0.20/square feet (for lower density lab square footage peak on frequency histogram).

****Brown model: Based on new AAHC resourcing model: EHS expenditures = $e^{(1.587 + 0.8397 * \ln(\text{total institutional assignable sf}))}$

Parker model: Based on S-estimation: $\text{EHSEXP} = e^{(4.6236 + 0.1583 \ln \text{RNASF} + 0.4891 \ln \text{FTE} + 0.1805 \ln \text{RE})}$.

Limitations and Suggestions

The data set used in this analysis contains variables that did not exhibit significant p-values when one-way analysis of variance was performed. These independent variables include institutional research net assignable square footage, institutional full-time employees, and institutional research expenditures. The similarities in the distribution can also be visually assessed via the histogram overlays for the years 2011, 2013, and 2015. Although this research project provides insight into the environmental health and safety full-time employees and expenditures, these numbers are only industry average values and give sparse insight into the quality of the institutions' environmental health and safety programs. For example, the regression model might produce an operating budget of \$2,000,000; however, that same institution may maintain compliance with an operating budget of \$1,500,000. Institutional predictors related to operating efficiency should therefore be included in future regression modeling.

Ideally, future work should focus on institutional outcome measures related to injury and illness among the at-risk population (e.g., environmental health and safety full-time employees, medical residents, students, and institutional staff). Identifying the key predictors that provide insight into morbidity and mortality rates across different institutions would allow for a quantitative approach for identifying problematic areas. This systematic approach would give the institutions an opportunity to pinpoint the areas that need remediation and potentially function as a preventive technique for eliminating future incidents related to morbidity and mortality. This quantitative approach could be addressed by identifying the areas and locations in which each of the institutional full-time employees

primarily work. For example, the Integrated Postsecondary Education Data System Survey tends to focus primarily on outcome measures related to staff and student retention and graduation rates. This survey could also be used to address issues related to morbidity and mortality by recording information on injury and illness rates among students and staff. For example, the IPEDS data center should not only record data on institutional full-time staff but should also create records for time spent in each of the research laboratory areas.

Similarly, future regression models should include the component parts of the variable institutional research net assignable square footage and link them to morbidity and mortality rates across various institutions. The academic research area reported by the National Science Foundation is contained in the following academic departments:

“agricultural sciences and natural resources sciences, biological and biomedical sciences, computer and information sciences, engineering, health and clinical sciences, mathematics and statistics, physical sciences, psychology, social sciences, and other science and engineering fields.” Evidently, this would allow for future regression models to focus on areas which explain the majority of the variance in the models.

CONCLUSION

The final models were selected simultaneously using cross-validation with information criteria: AIC, SBC, BIC, PRESS, R^2 , R^2_{adj} , and Mallows' $(C)_p$. The model with research expenditures, research net assignable square footage, and institutional full-time employees regressed on environmental health and safety expenditures showed significant p-values. The presence of outliers and influential observations in a data set can wreak havoc on the least squares estimates when the error distribution is not normal. In the model with environmental health and safety expenditures as the dependent variable, the outliers and influential observations were corrected by the implementation of robust regression methods: M-estimation, S-estimation, LTS-estimation, and MM-estimation. Out of the robust regression methods, the high breakdown point robust estimates of scale, S-estimators, were chosen to optimize the parameter estimates.

Similarly, the model with institutional research expenditures, total institutional net assignable square footage, and institutional full-time employees regressed on environmental health and safety full-time employees showed significant p-values. The presence of outliers in the model with environmental health and safety full-time employees as the dependent variable was corrected by using robust S-estimators.

The results indicate that institutional research net assignable square footage possesses a more substantive predictive value for estimating environmental health and safety expenditures than total institutional net assignable square footage. Further, when estimating environmental health and safety full-time employees, the total institutional net assignable square footage should be used instead of institutional research net assignable square footage.

These findings suggest that members of CSHEMA should primarily focus on vital statistics pertaining to total institutional net assignable square footage, institutional research net assignable square footage, institutional research expenditures, and institutional full-time employees, as these variables explain most of the variance in the models. This will allow for the proper allocation of resources among environmental health and safety departments at member institutions.

References

- Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3), 469-475.
- Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1), 125-127.
- American Society of Safety Engineers, Council on Practices and Standards. (2010). *Technical report: Staffing issues and SH&E professionals*. Retrieved from http://www.asse.org/practicespecialties/bosc/bosc_article_StaffPaper.php (Accessed September 16, 2012).
- Amis, E. S., Jr, Butler P. F., Applegate, K. E., Birnbaum, S. B., Brateman, L. F., Hevezi J. M., Zeman, R. K. (2007). American College of Radiology white paper on radiation dose in medicine. *J Am Coll Radiol*, 4, 272–284. doi: 10.1016/j.jacr.2007.03.002.
- Andersen, R. (2008). *Modern methods for robust regression*. Thousand Oaks, CA: SAGE Publications.
- Brown, B. (2014). *An empirical model for estimating environmental health and safety program resourcing for colleges and universities* (Unpublished doctoral dissertation). The University of Texas Health Science Center School of Public Health, Houston.
- Bush, V. (1945). Science: The endless frontier. *Transactions of the Kansas Academy of Science* (1903), 231-264.

- Donoho, D. L. (1982). *Breakdown properties of multivariate location estimators*
(Unpublished Ph.D. qualifying paper). Department of Statistics, Harvard University,
Cambridge, MA.
- Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In P. J. Bickel, K.
Doksum, & J. L. Hodges, Jr. (Eds.), *A festschrift for Erich L. Lehmann* (pp. 157-184).
Belmont, CA: Wadsworth. MR689745
- Emery, R. J., Delclos, G., Cooper S. P., & Hardy, R. (1998). Evaluating the relative status of
health and safety programs for minority academic and research institutions. *American
Industrial Hygiene Association Journal*, 59, 882-888.
- Fine, W. T. (1982). Proper staffing of an occupational safety and health office. *Professional
Safety*, 20-24.
- Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate
data analysis* (3rd ed.). New York, NY: Macmillan.
- Hampel, F. R. (1968). *Contributions to the theory of robust estimation* (Unpublished doctoral
dissertation). University of California, Berkeley.
- Hampel, F. R. (1971). A general qualitative definition of robustness. *Ann. Math. Statist.*, 42
1887–1896. MR301858

- Harding, A.L., & Byers K. B. Epidemiology of Laboratory-Associated Infections, In D. O. Fleming & D. L. Hunt (Eds.), *Biological safety: Principles and practices* (4th ed., pp. 53-77). ASM Press.
- Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer-Verlag.
- Hicks, C. R., & Turner, K. V. (1999). *Fundamental concepts in the design of experiments* (5th ed.). New York, NY: Oxford University Press.
- Hodges, J. L., Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *In Proc. Fifth Berkeley Symp. Math. Statist. Probab., 1* (pp. 163-186). Berkeley, CA: University of California Press. MR214251.
- Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist., 35*, 73-101. MR161415.
- Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *The Annals of Statistics, 1*(5), 799-821.
- Huber, P. J. (1996). *Robust statistical procedures* (2nd ed.). Philadelphia, PA: SIAM.
- Johnson R. (1992). *Applied multivariate statistical analysis*. Prentice Hall.

Kathren, R. L. (2002). Historical development of the Linear Nonthreshold Dose-Response Model as applied to radiation. *University of New Hampshire Law Review*, 1.

Retrieved from http://scholars.unh.edu/unh_lr/vol1/iss1/5

Mallows, C. L. (1973). Some comments on Cp. *Technometrics*, 15(4), 661-675.

Marazzi. (1993). *Algorithms, routines and S functions for robust statistics*. Wadsworth and Brooks/Cole.

Montgomery, D. C., Peck, E. A., & Vining, C. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York, NY: John Wiley & Sons.

National Science Foundation / National Institutes of Health. (2009). *FY2009 Survey of Science and Engineering Research Facilities*. Retrieved from

http://www.nsf.gov/statistics/srvyfacilities/surveys/srvyfacilities_2009.pdf

National Science Foundation, Higher Education Research and Development. (2010). *Higher education research and development: Fiscal year 2010*. Retrieved from

<http://www.nsf.gov/statistics/nsf12330/pdf/nsf12330.pdf>.

Patlovich, S. T., et al. (2015). Assessing the biological safety profession's evaluation and control of risks associated with the field collection of potentially infectious specimens. *Appl Biosaf*, 20(1), 27-40.

- Pike, R. M., Sulkin, S. E., & Schulze, M. L. (1965). Continuing importance of laboratory-acquired infections. *American Journal of Public Health and the Nations Health*, 55(2), 190-199.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Hoboken, NJ: Wiley.
- Rousseeuw, P. J., & Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. In W. Gaul, O. Opitz, and M. Schader (Eds.), *Data analysis: Scientific modeling and practical application* (pp. 335-346). New York, NY: Springer-Verlag.
- SAS Institute Inc. 2009. *SAS/STAT® 9.2 User's Guide, Second Edition*. Cary, NC: SAS Institute Inc
- Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, 46, 1273-1282.
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461-464.
- Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4), 591-611.
- Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics*, 19(4), 415-428.

- Stevens, J. P. (1995). *Applied multivariate statistics for the social sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Stigler, S. M. (1981). Gauss and the invention of least squares. *Ann. Statist.*, 9(3), 465-474.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.
- Takeaki, K., & Hiroshi, K. (2004). *Generalized least squares*. Chichester, UK: Wiley.
- U.S. Bureau of Labor Statistics, U.S. Department of Labor. (2017). *2016 survey of occupational injuries & illnesses charts package*. Retrieved from <https://www.bls.gov/iif/osch0060.pdf>
- U.S. Bureau of Labor Statistics, U.S. Department of Labor. (2017). *Fatal occupational injuries to private sector wage and salary workers, government workers, and self-employed workers by industry, all United States, 2016*. Retrieved from <https://www.bls.gov/iif/oshcfoi1.html2016>
- U.S. Department of Education. *National Center for Education Statistics postsecondary education facilities inventory and classification manual (FICM)* (NCES 2006-160). Retrieved from <http://nces.ed.gov/pubs2006/2006160.pdf>.

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System. (2011). *IPEDS finance for public institutions using GASB reporting standards*. Retrieved from http://nces.ed.gov/ipeds/surveys/2010/pdf/f_1_2010.pdf.

U.S. Department of Education, National Center for Education Statistics, Integrated Postsecondary Education Data System. (2011). *IPEDS 12-month enrollment 2009-2010*. Retrieved from http://nces.ed.gov/ipeds/surveys/2010/pdf/e12_2010.pdf.

U.S. Department of Education, National Center for Education Statistics. (2012). *Integrated Postsecondary Education Data System. IPEDS glossary*. Retrieved from <http://nces.ed.gov/ipeds/glossary>.

U.S. Environmental Protection Agency. *Defining hazardous waste: Listed, characteristic and mixed radiological wastes*. Retrieved from <https://www.epa.gov/hw/defining-hazardous-waste-listed-characteristic-and-mixed-radiological-wastes#FandK>

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, 15, 642-656.

APPENDICES

APPENDIX A: CAMPUS SAFETY, HEALTH, AND ENVIRONMENTAL MANAGEMENT ASSOCIATION (CSHEMA) APPROVAL LETTER.

UNIVERSITY OF CALIFORNIA, IRVINE

BERKELEY • DAVIS • IRVINE • LOS ANGELES • RIVERSIDE • SAN DIEGO • SAN FRANCISCO



SANTA BARBARA • SANTA CRUZ

Facilities Management
Office of the Assistant Vice Chancellor

104A Interim Office Building
Irvine, CA 92697-5444
Phone: (949) 924-5444
Fax: (949) 824-4868

December 16, 2016

Robert Emery, DrPH
Vice President for Safety, Health, Environment & Risk Management
Professor of Occupational Health
The University of Texas Health Science Center at Houston
1851 Crosspoint Drive, OCB 1.330
Houston, TX 77054

Dear Dr. Emery:

On behalf of the Campus Safety, Health, and Environmental Management Association (CSHEMA) Research and Survey Community of Practice Committee, I would like to voice our enthusiastic support for Mr. Seth Parker's dissertation research proposal entitled "Secondary Data Analysis of an Empirical Model for Estimating Environmental Health and Safety Program Resourcing for Colleges and Universities". We are in full agreement that the development of a statistically-valid model for predicting industry average loss control resource outlays for colleges and universities would be very helpful to the profession.

In support of Mr. Parker's research we will be providing the 2012 CSHEMA Benchmarking dataset originally assembled as part of Dr. Bruce Brown's dissertation to be used in this secondary data analysis effort. The data set is de-identified, with the names of participating institutions removed.

We hope that the analysis serves to validate the previous pilot work that you and your students have undertaken at the University of Texas School of Public Health. The preliminary work in the identification of possible key predictors has been warmly received by our membership, and we would hope that if this effort is successful that Seth would consider submitting a proposal to present his findings at an upcoming CSHEMA national conference.

Again, we are very excited to be part of this project, and stand ready to provide any additional support that may be needed.

With best regards,

A handwritten signature in blue ink, appearing to read "Marc A. Gomez".

Marc A. Gomez, MPH, CIH, CSP, ARM
Chair, CSHEMA Benchmarking Committee
Assistant Vice Chancellor, EH&S and Facilities Management

Figure A1. CSHEMA Approval Letter.

APPENDIX B: HISTOGRAMS AND BOXPLOTS FOR TRANSFORMED DEPENDENT AND INDEPENDENT VARIABLES.

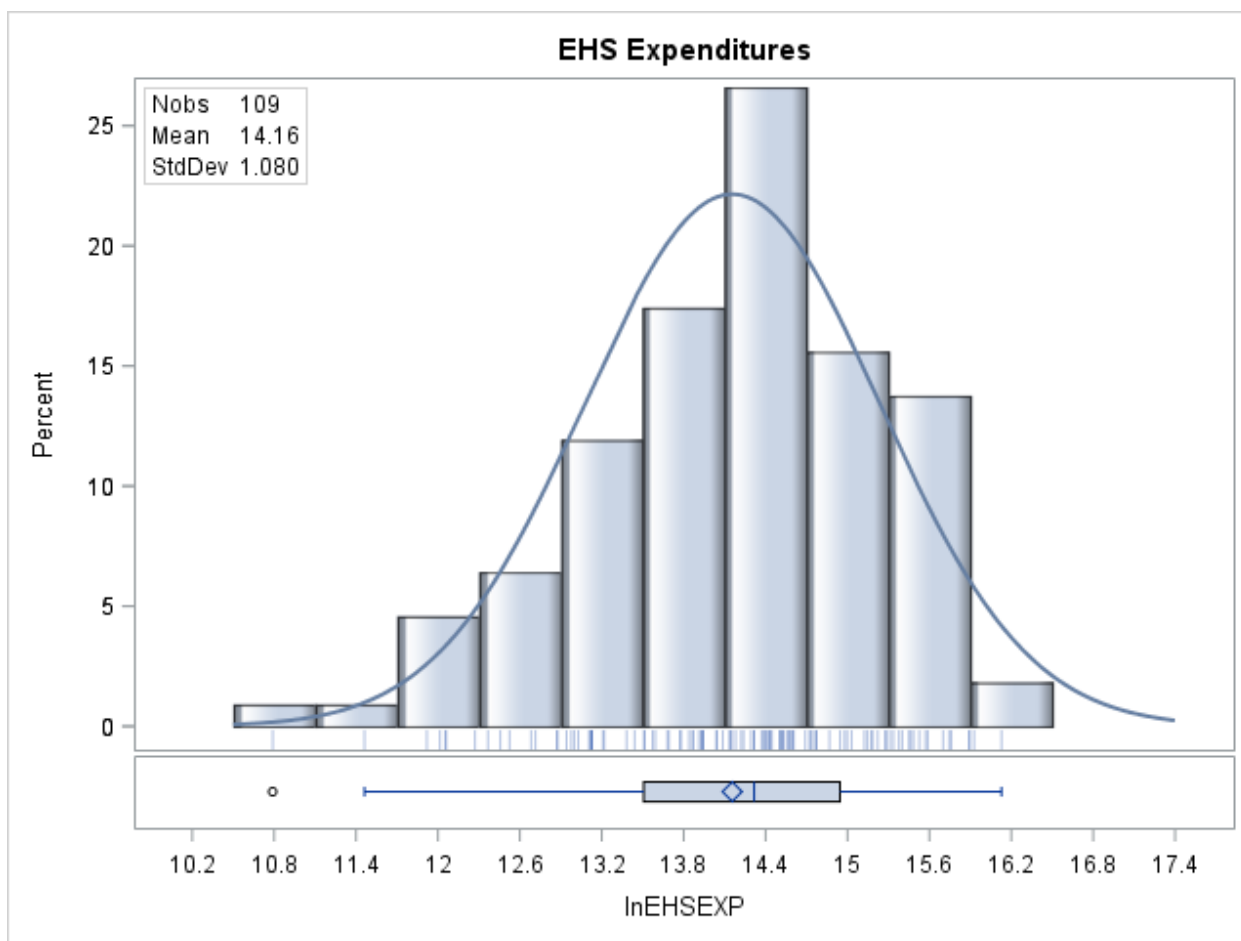


Figure B1. Histogram and Boxplot for the Natural Logarithm of Environmental Health and Safety Expenditures.

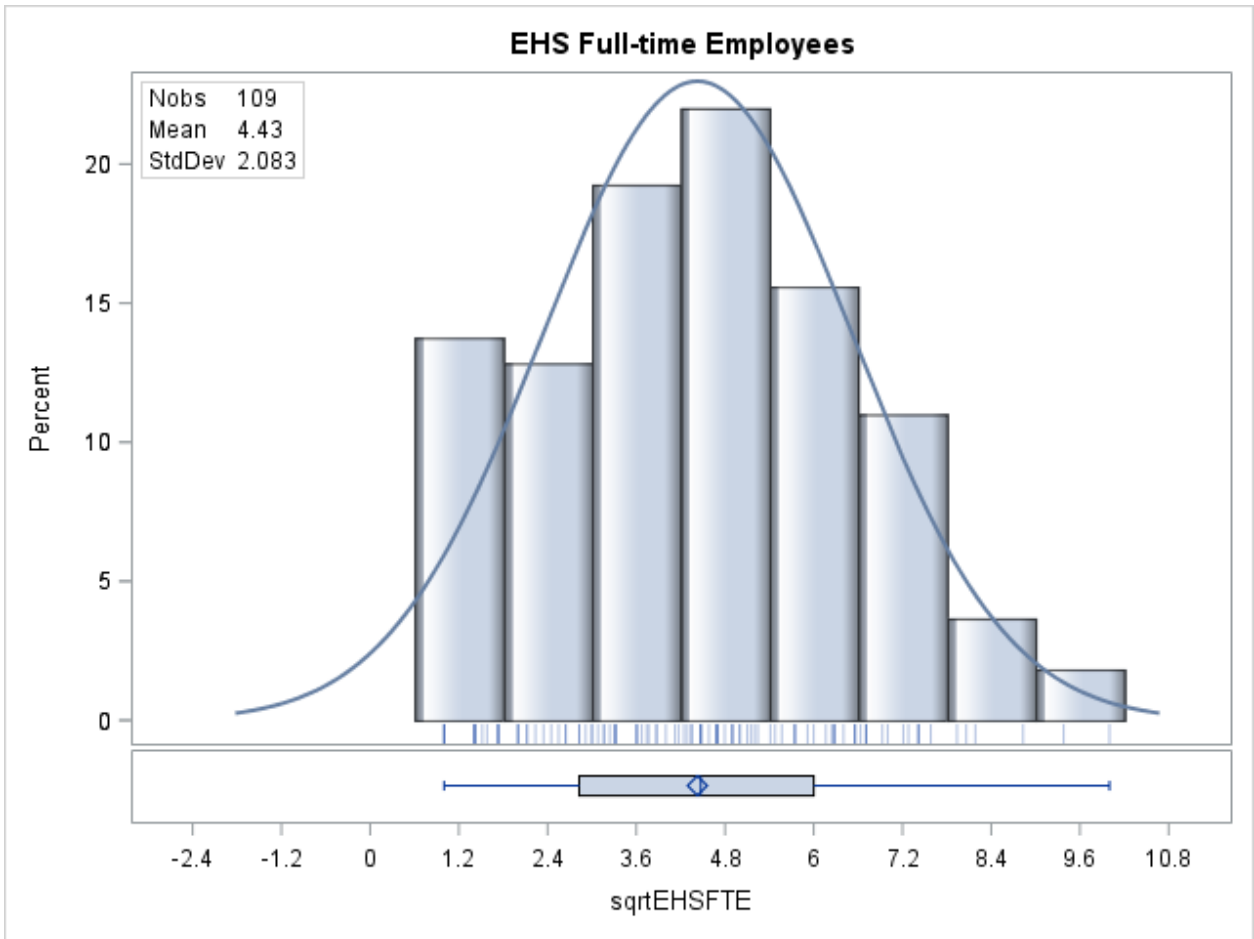


Figure B2. Histogram and Boxplot for the Square Root of Environmental Health and Safety Full-time Employees.

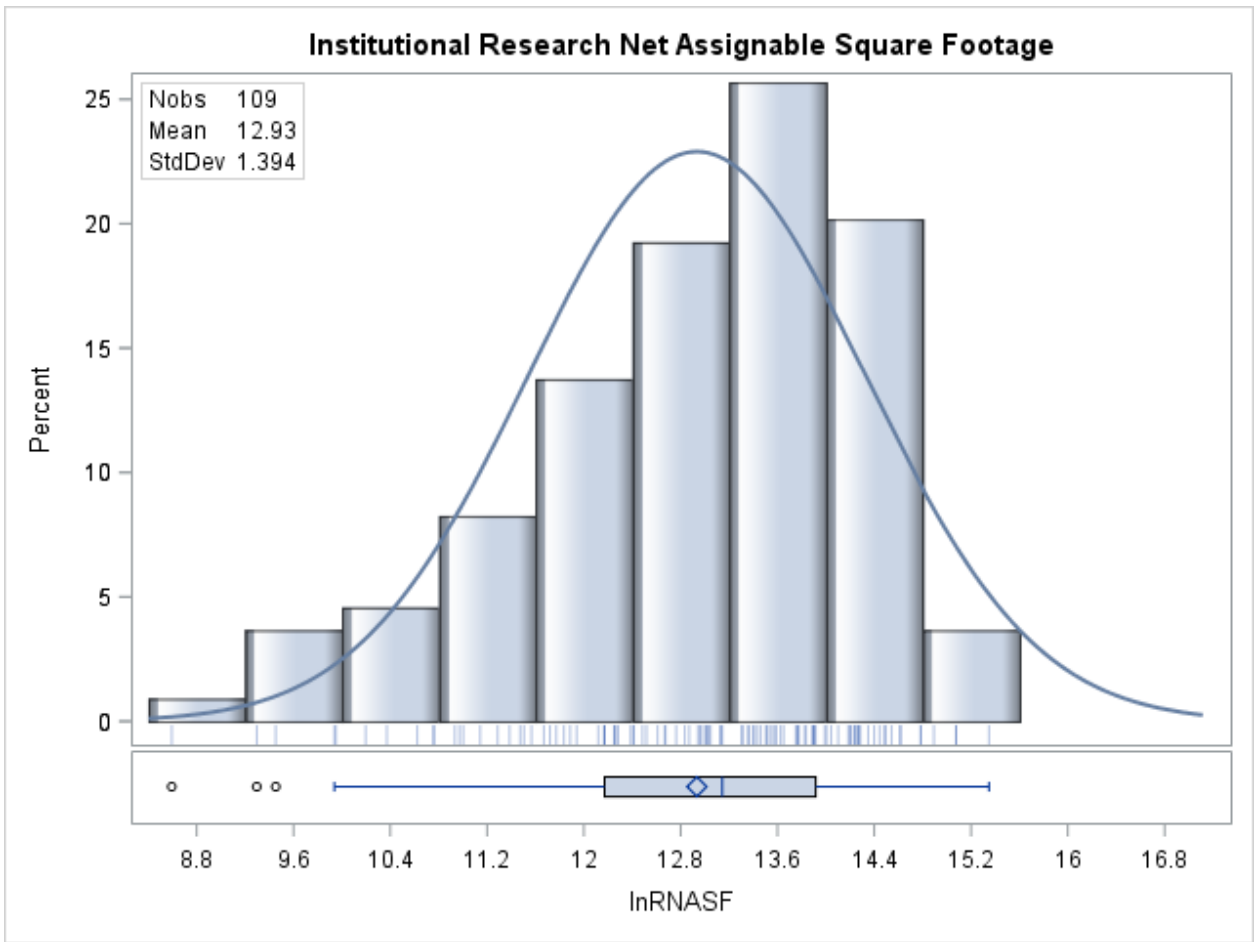


Figure B3. Histogram and Boxplot for the Natural Logarithm of Research Net Assignable Square Footage.

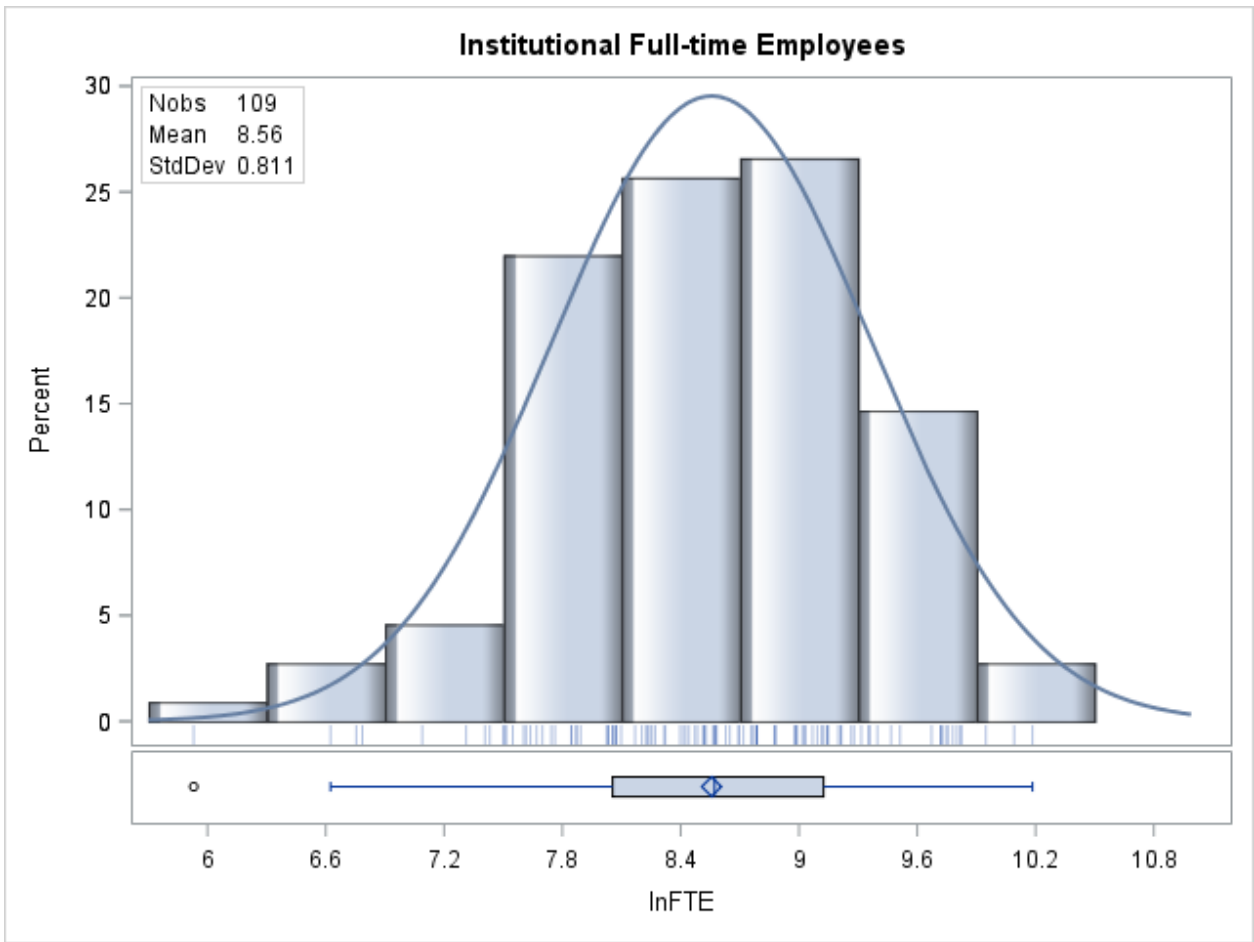


Figure B4. Histogram and Boxplot for the Natural Logarithm of Institutional Full-time Employees.

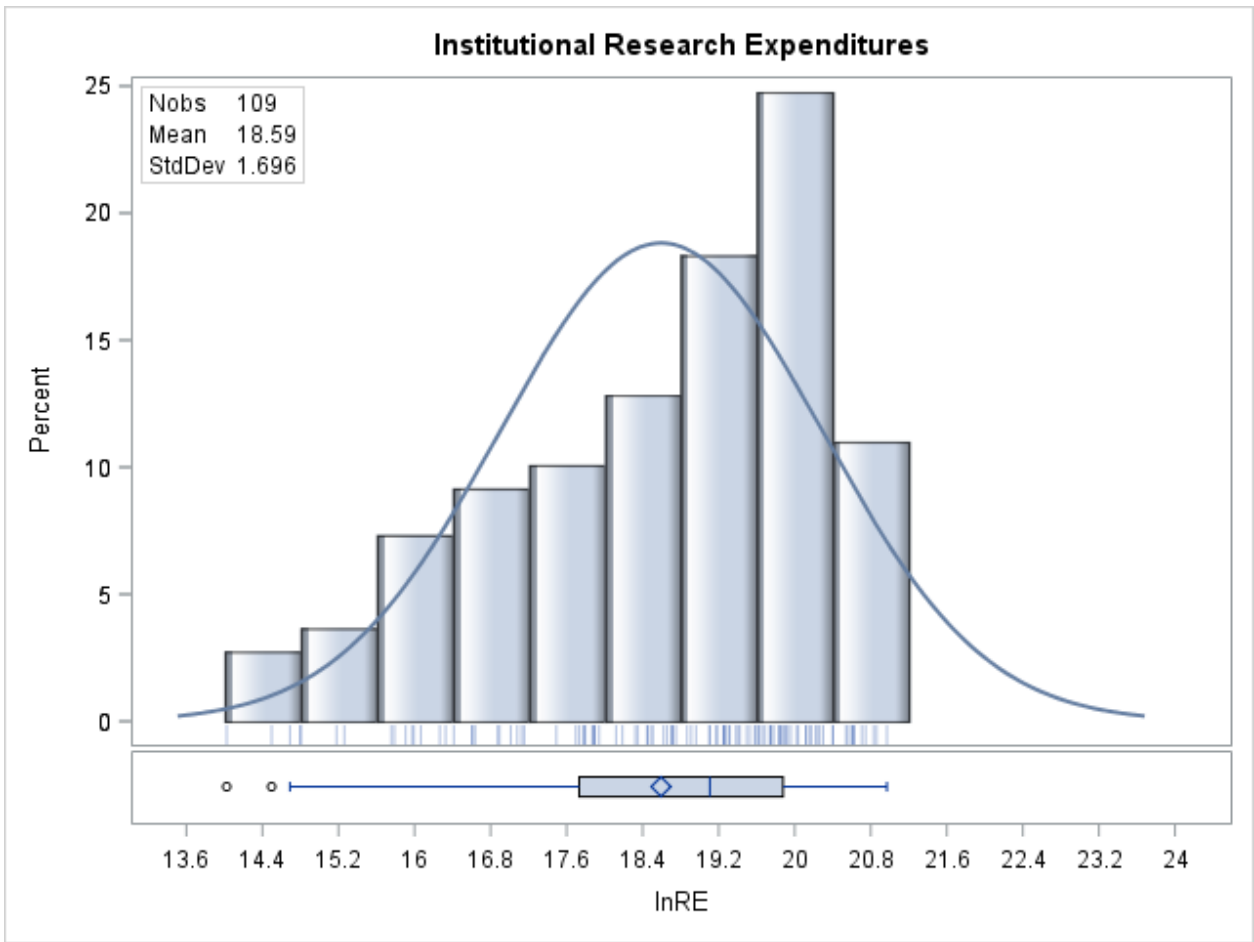


Figure B5. Histogram and Boxplot for the Natural Logarithm of Institutional Research Expenditures.

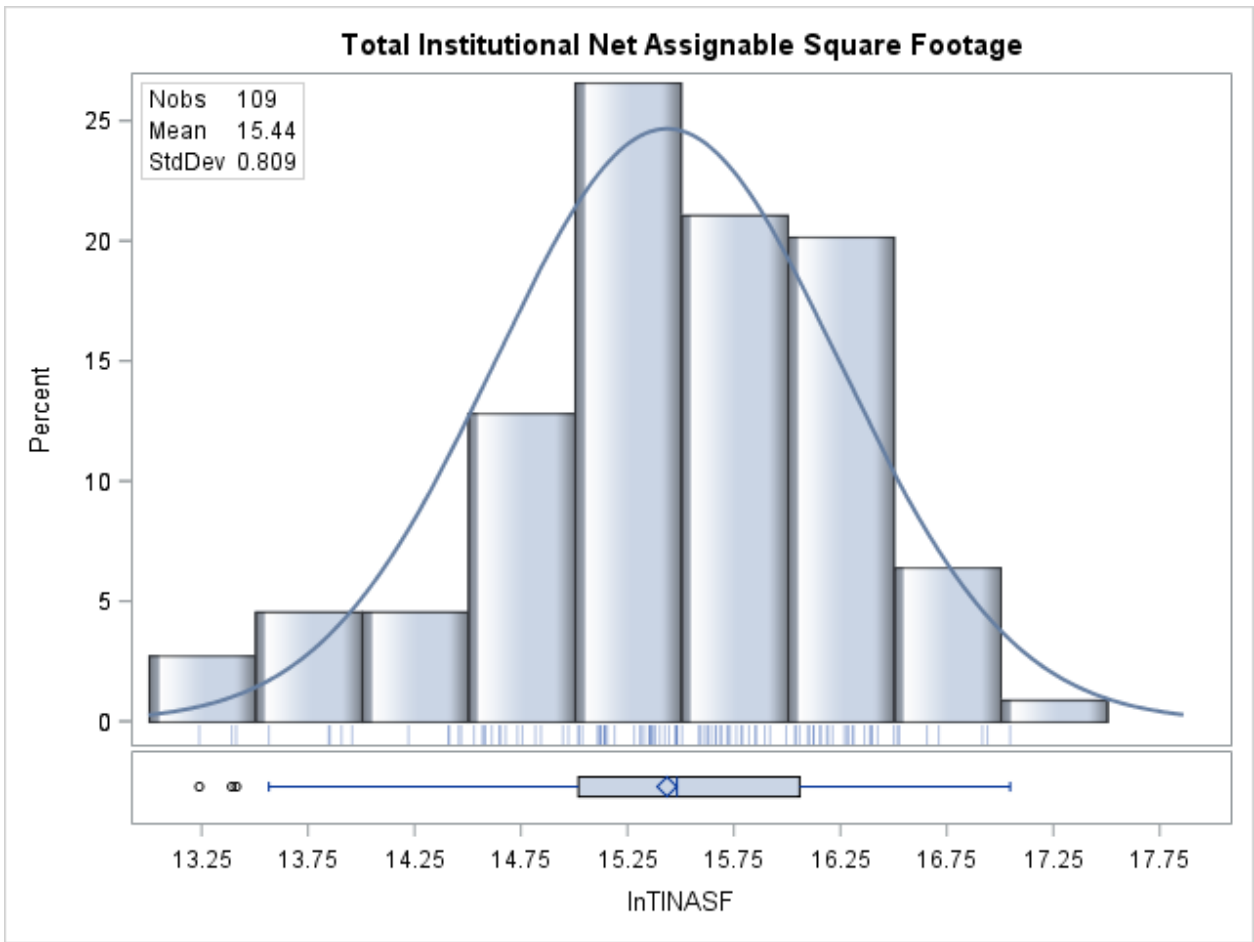


Figure B6. Histogram and Boxplot for the Natural Logarithm of Total Institutional Net Assignable Square Footage.

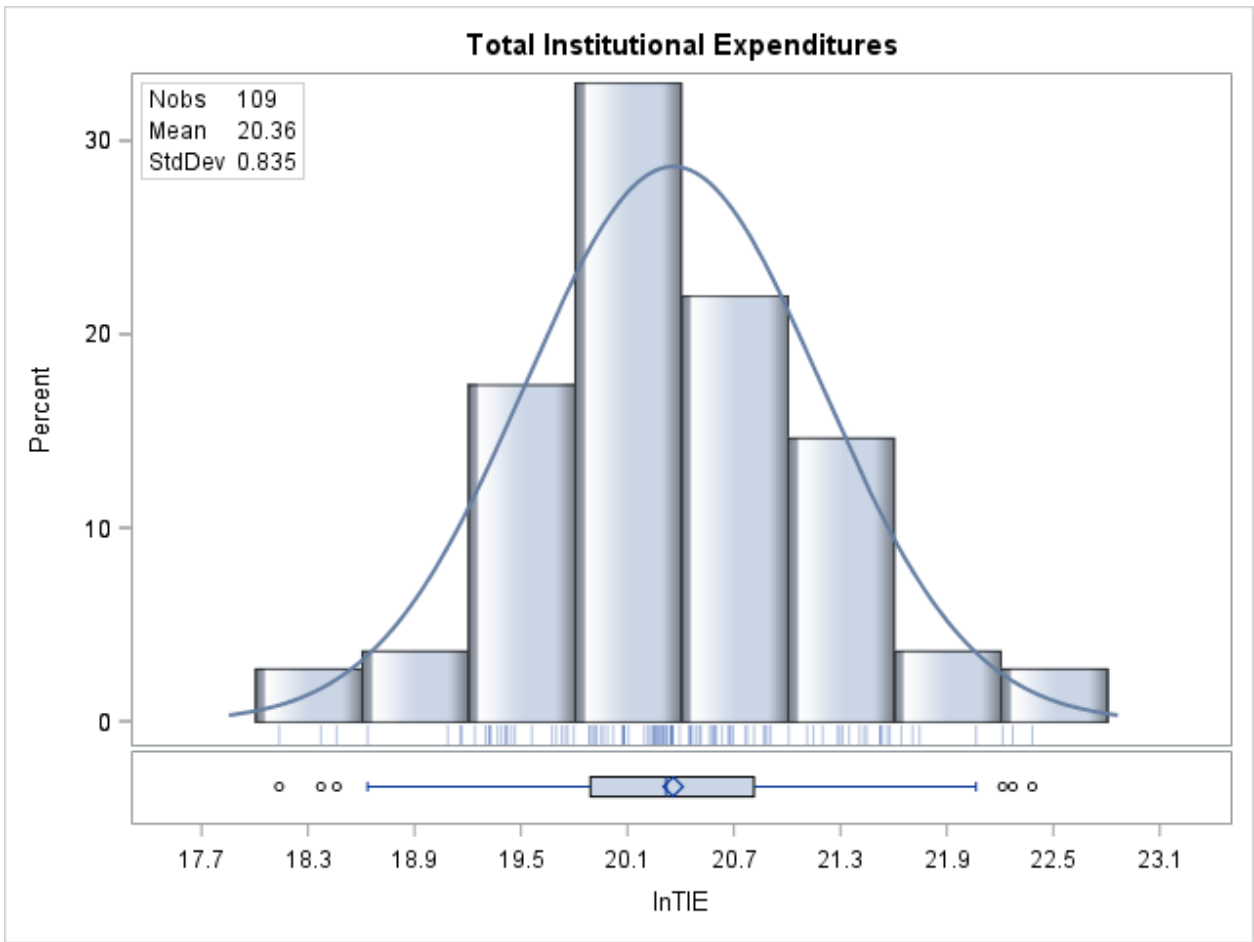


Figure B7. Histogram and Boxplot for the Natural Logarithm of Total Institutional Expenditures.

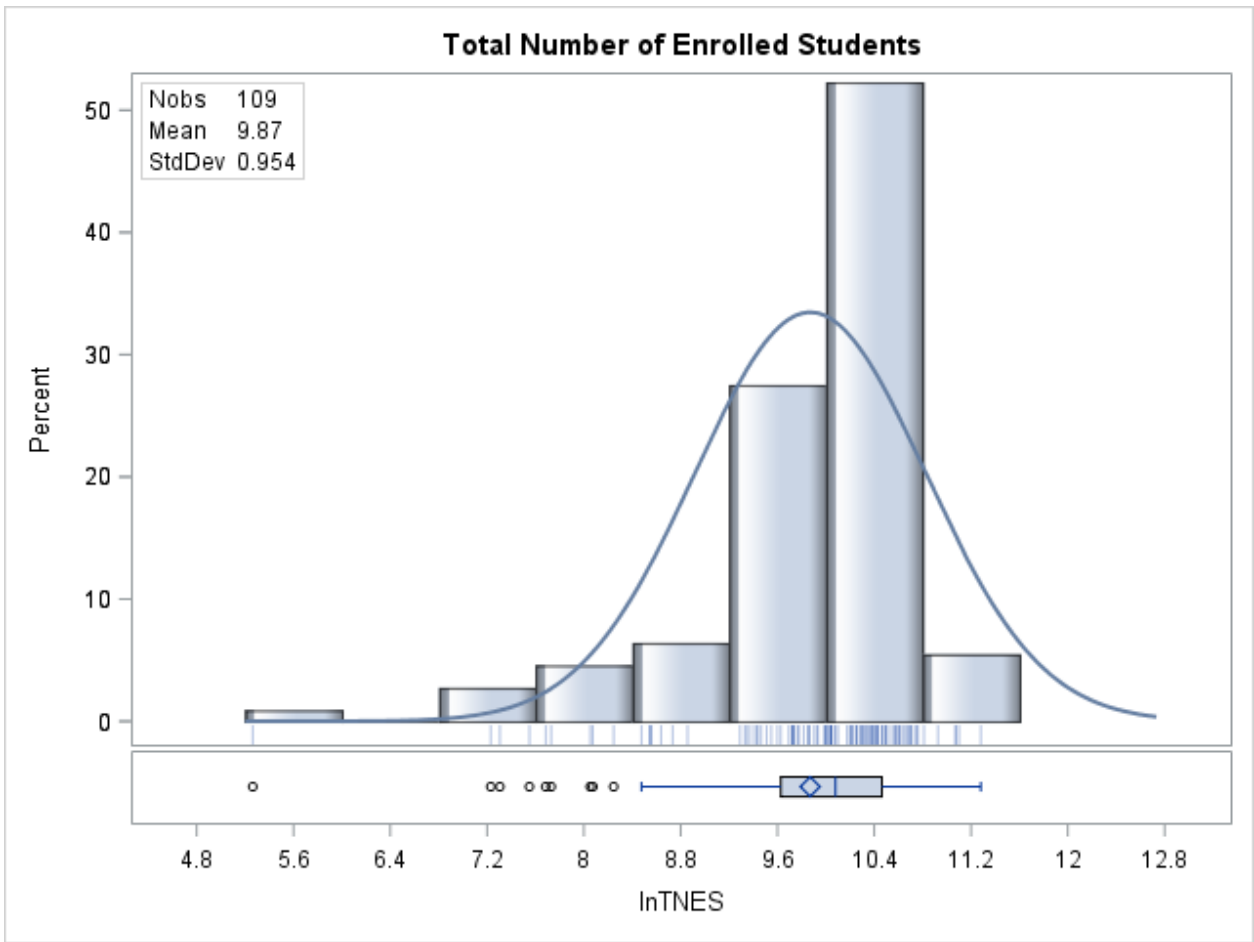


Figure B8. Histogram and Boxplot for the Natural Logarithm of Total Number of Enrolled Students.

APPENDIX C: BOX-COX ANALYSIS FOR DEPENDENT VARIABLES.

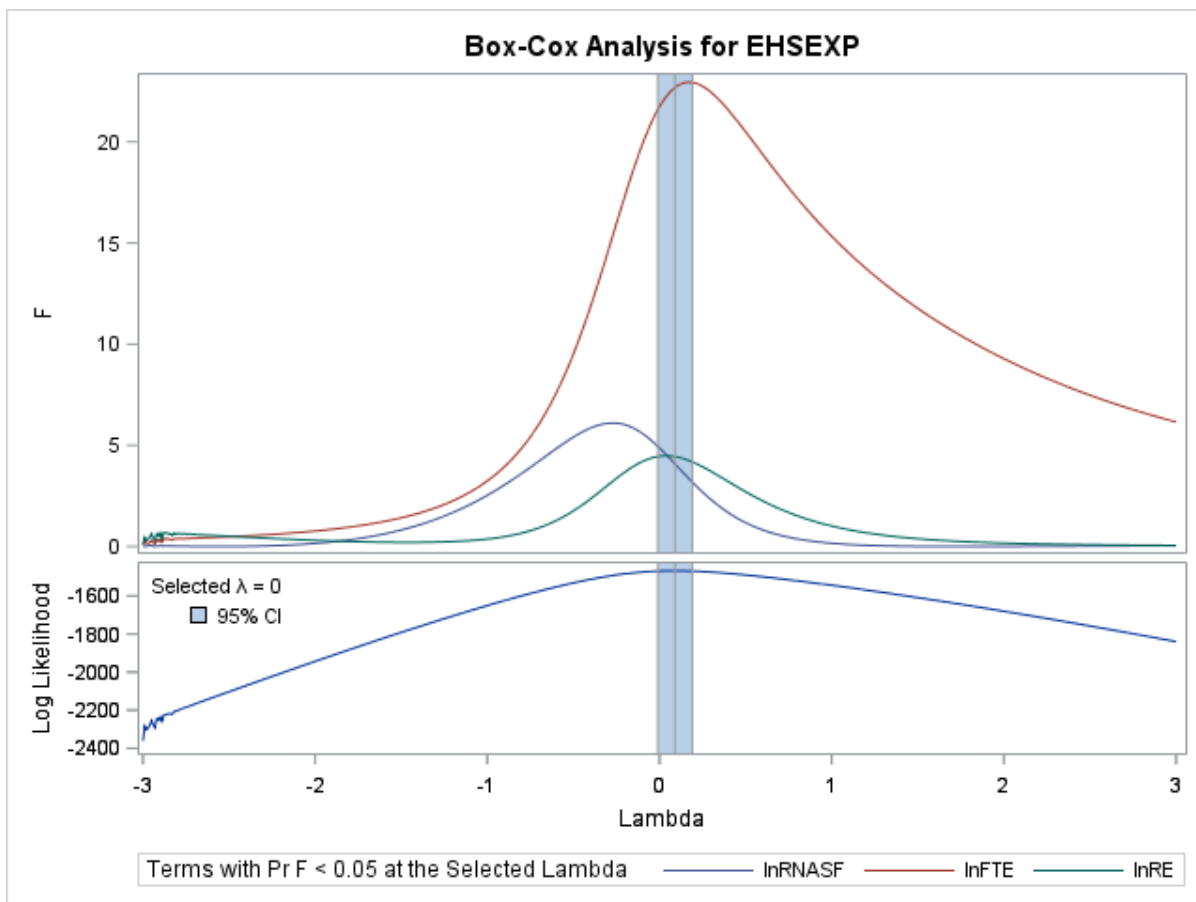


Figure C1. Box-Cox Analysis for Environmental Health and Safety Expenditures.

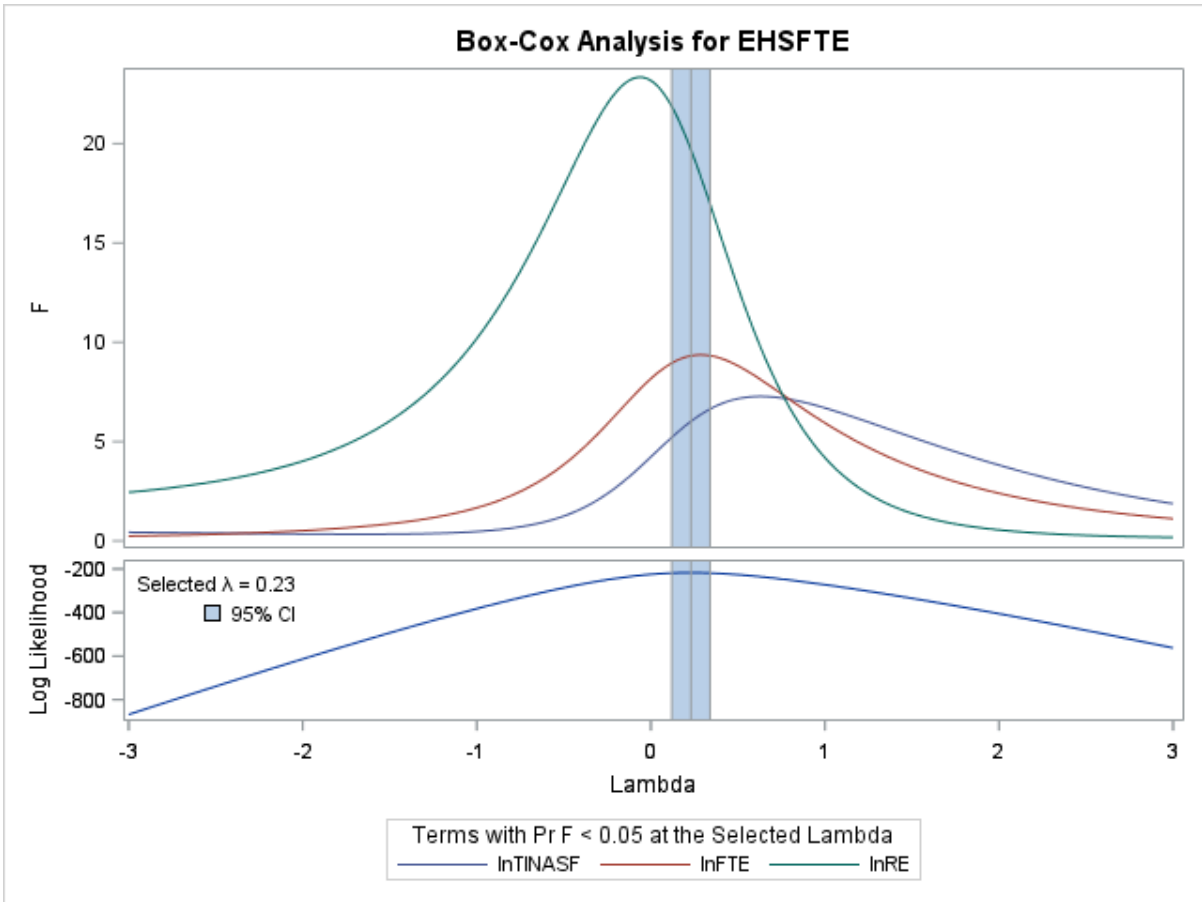


Figure C2. Box-Cox Analysis for Environmental Health and Safety Full-time Employees.

APPENDIX D: FIT DIAGNOSTICS FOR THE NATURAL LOGARITHM OF ENVIRONMENTAL HEALTH AND SAFETY EXPENDITURES.

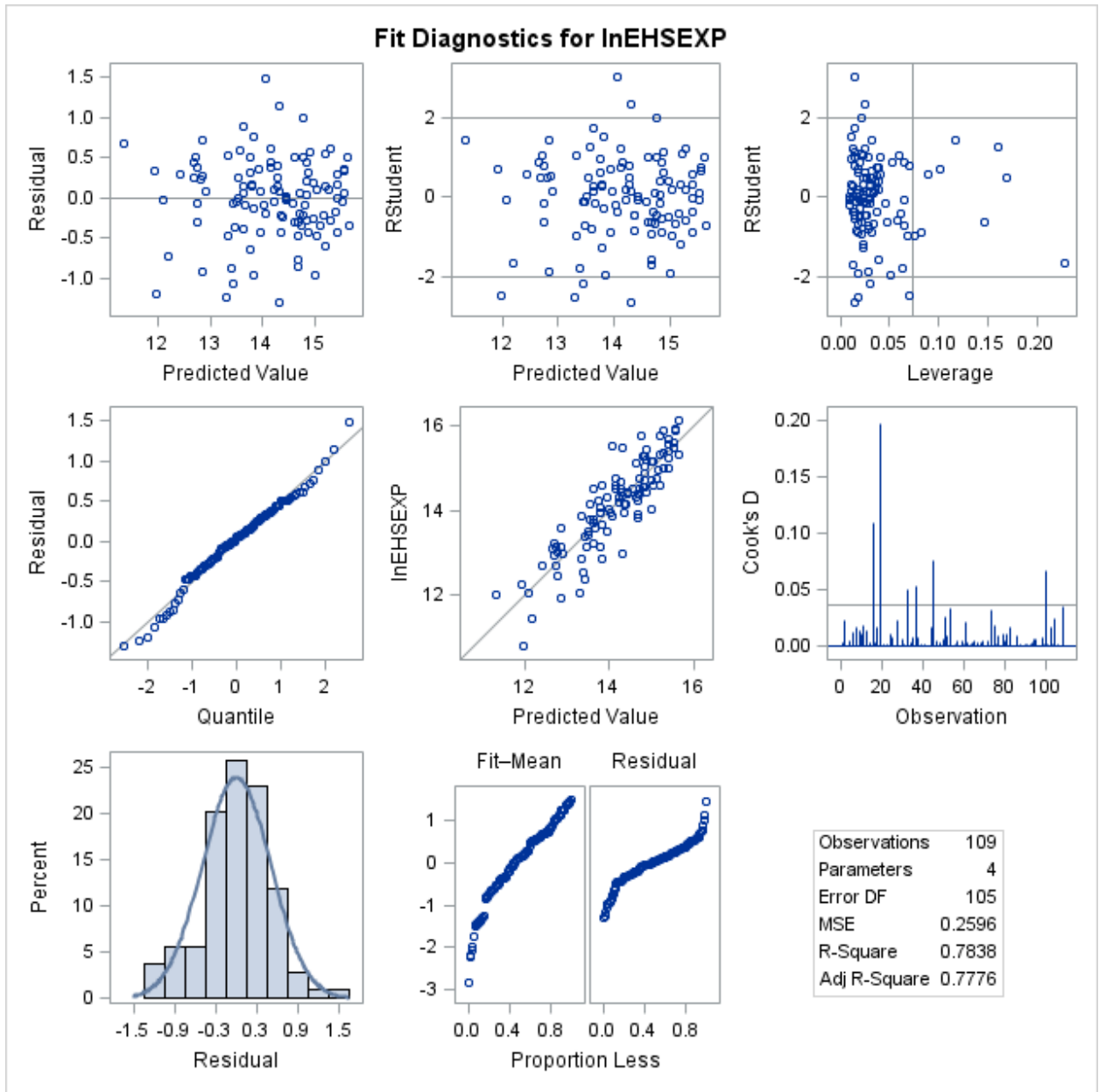


Figure D1. Fit Diagnostics for Model A.

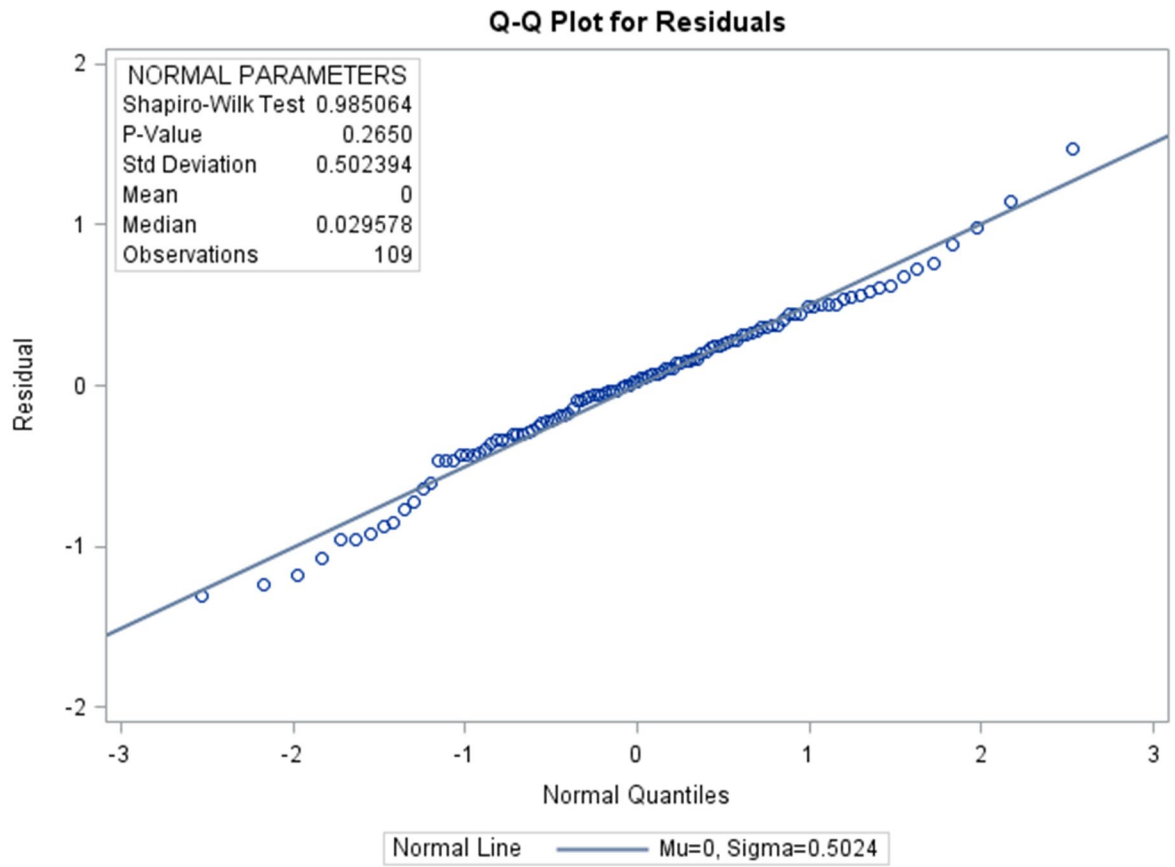


Figure D2. Q-Q Plot of Residuals with Shapiro-Wilk's test for Model A.

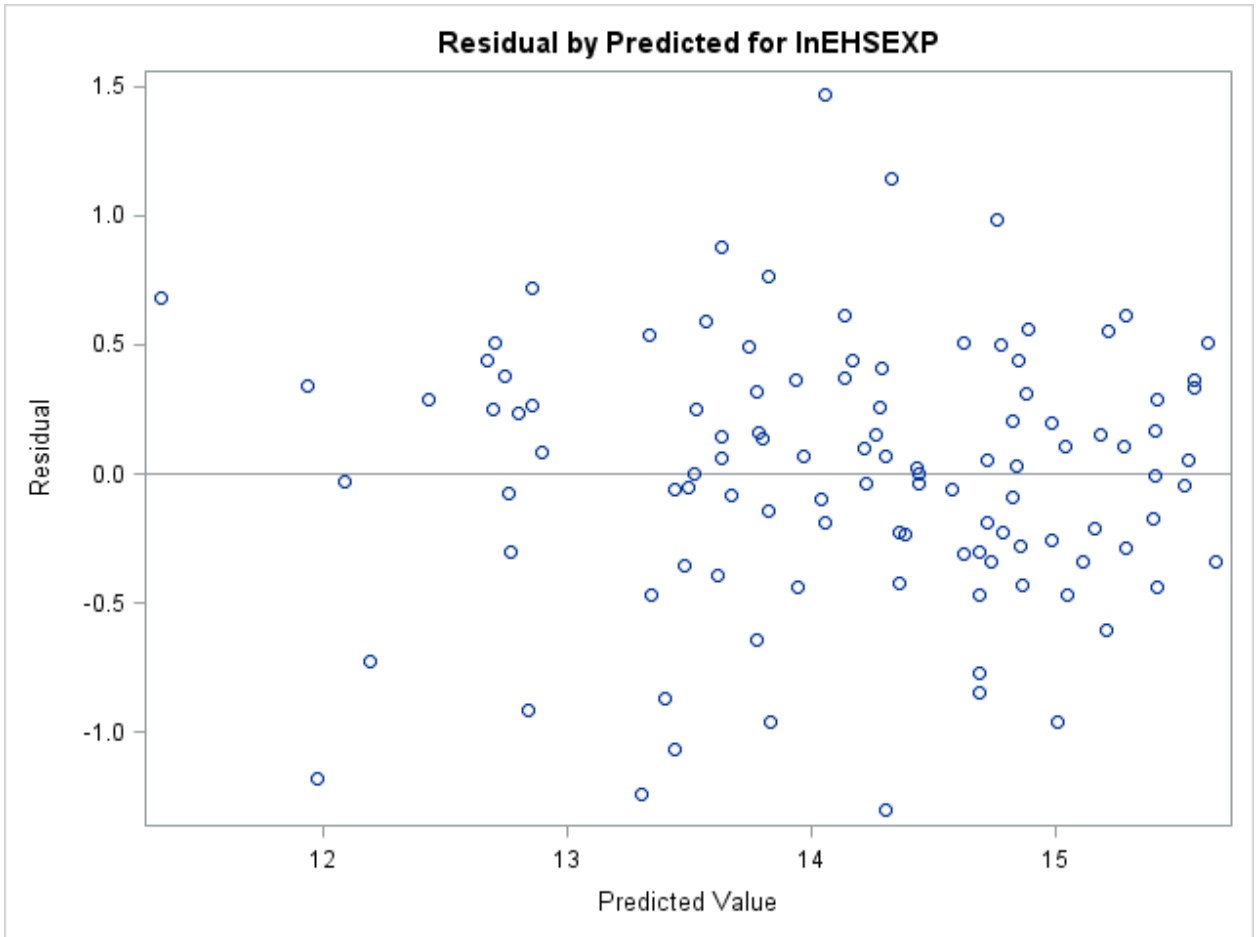


Figure D3. Residual by Predicted for Model A.

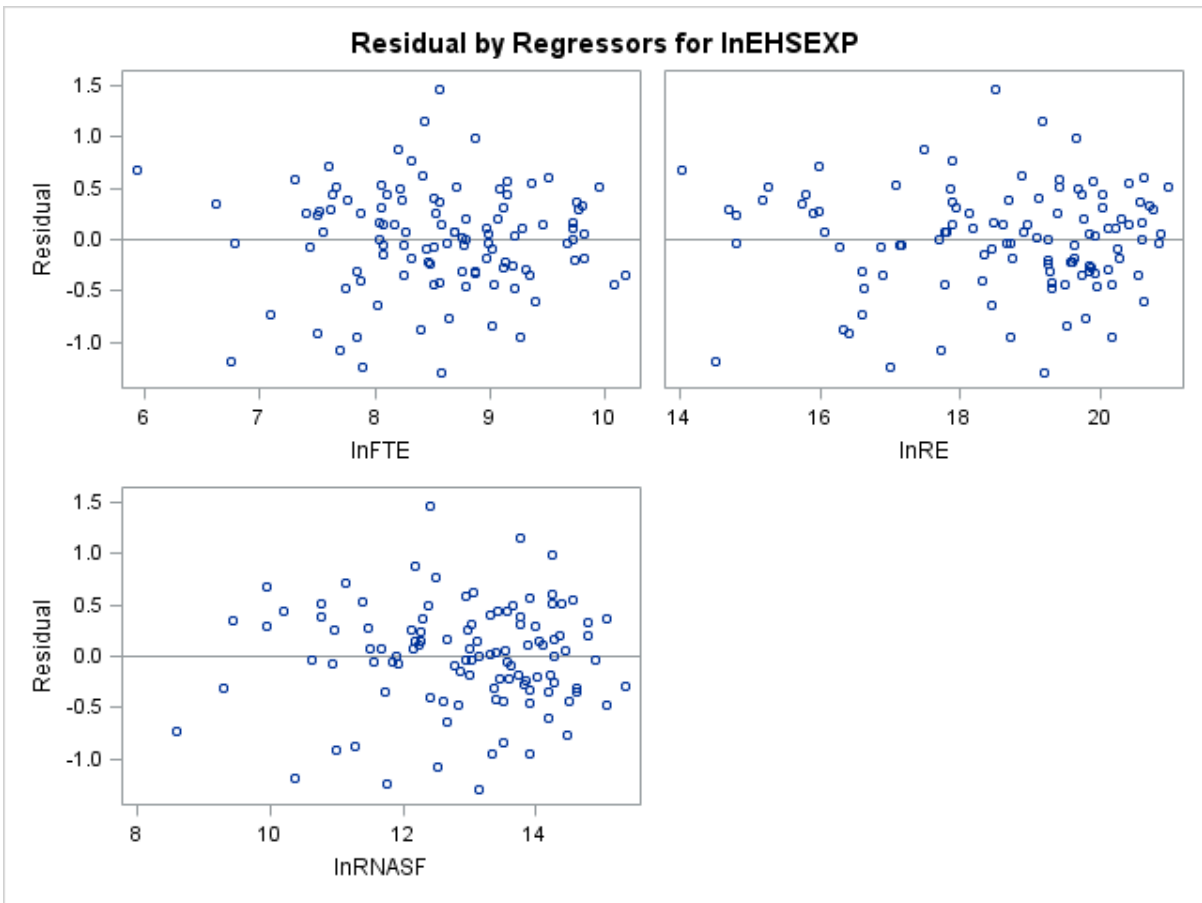


Figure D4. Residual by Regressor for Model A.

APPENDIX E: OUTLIER ANALYSIS FOR MODEL A.

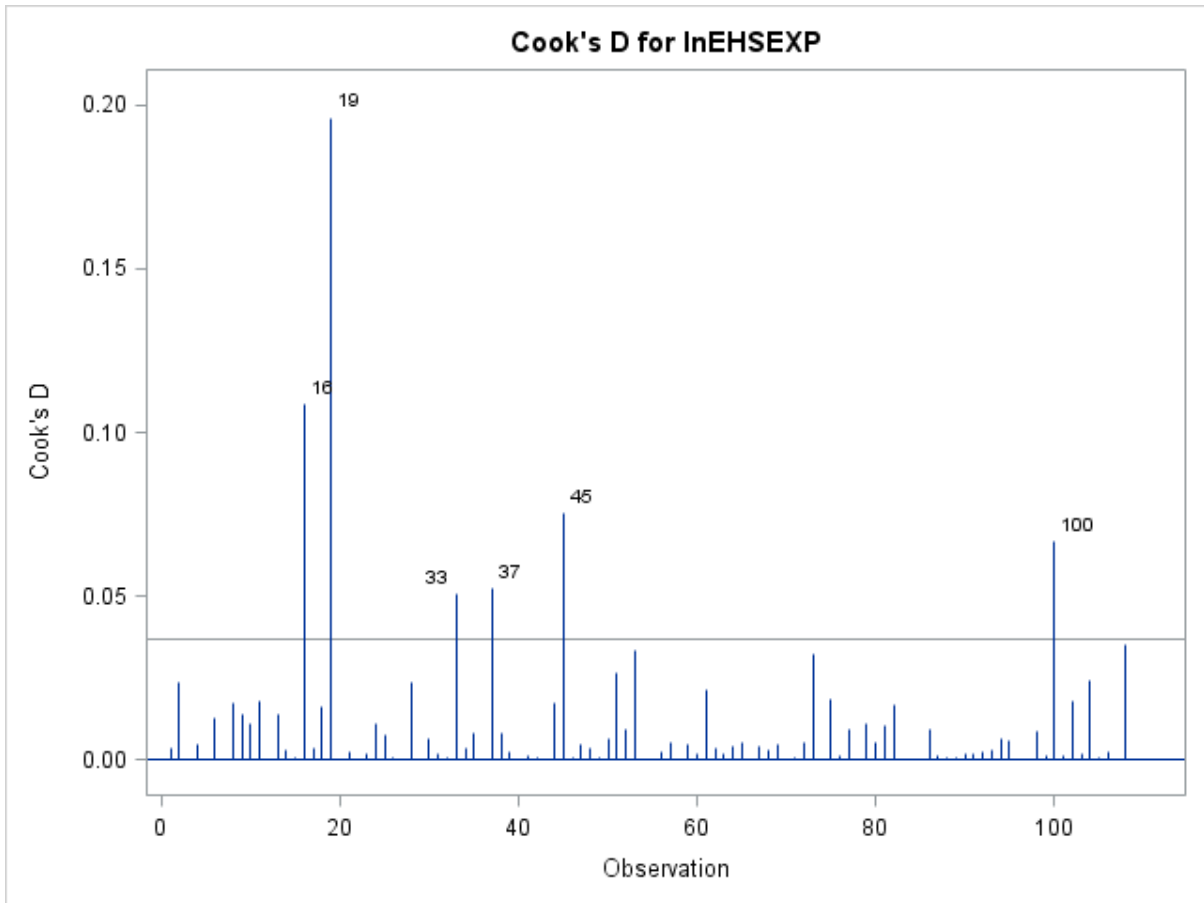


Figure E1. Cook's D for Model A.

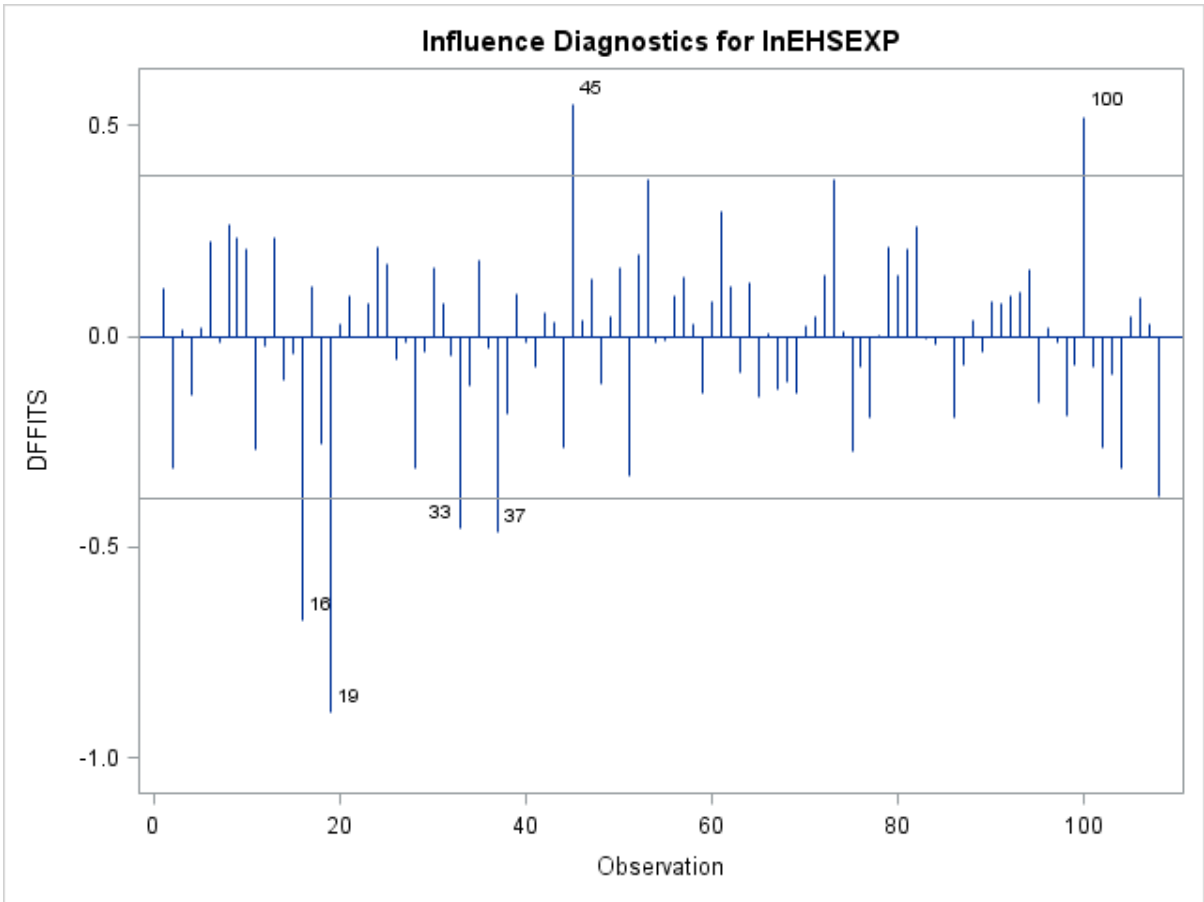


Figure E2. Difference in Fits (DFFITS) Influence Diagnostics for Model A.

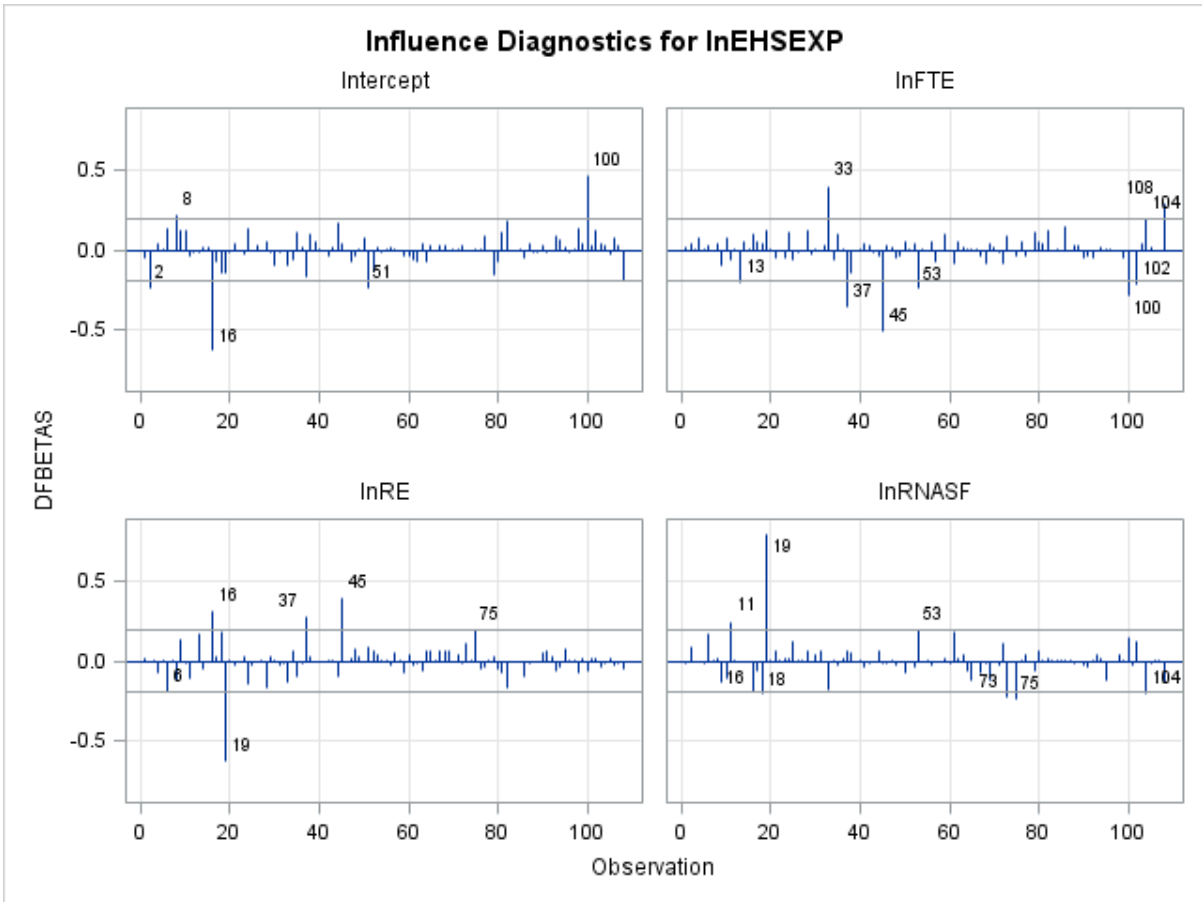


Figure E3. Difference in Betas (DFBETAS) Influence Diagnostics for Model A.

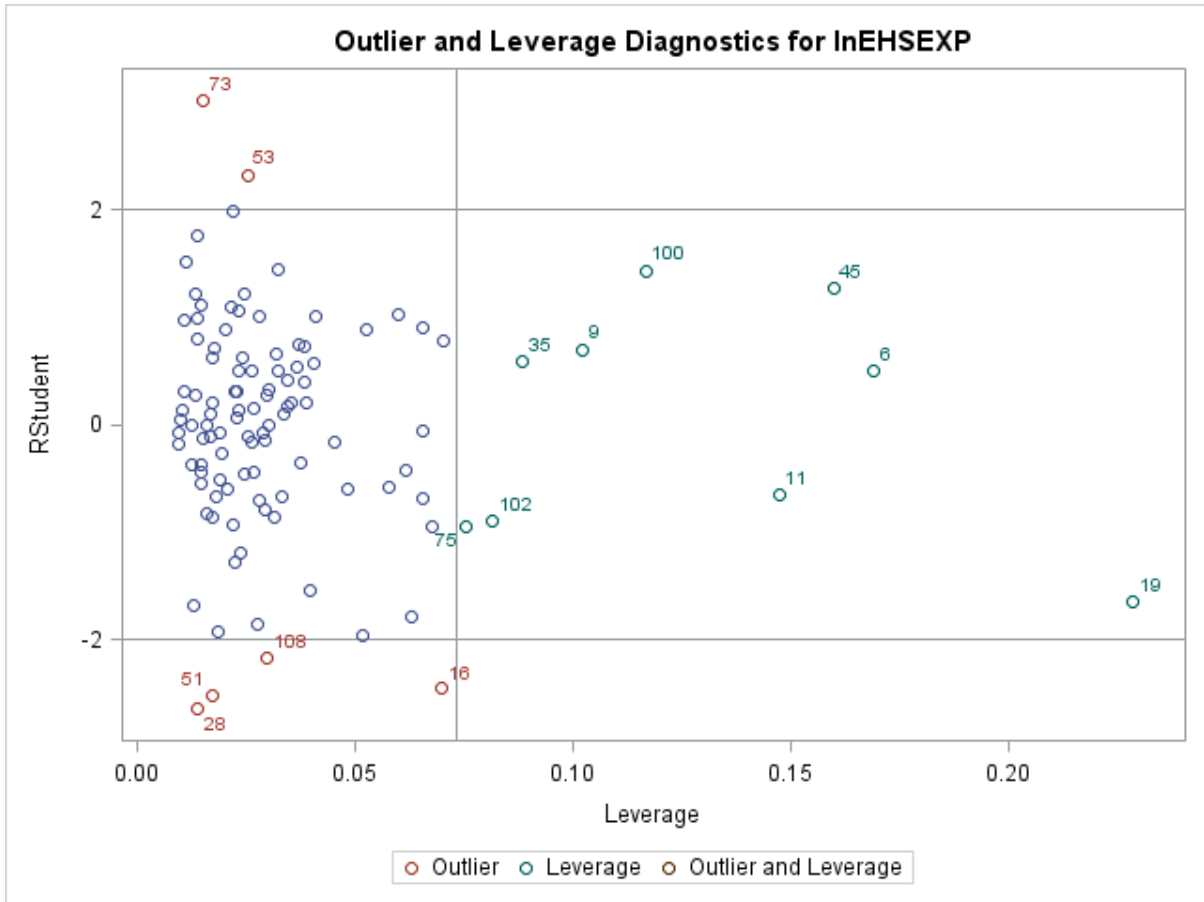


Figure E4. Outlier and Leverage Diagnostics using OLS for Model A.

APPENDIX F: ROBUST REGRESSION ESTIMATION FOR MODEL A.

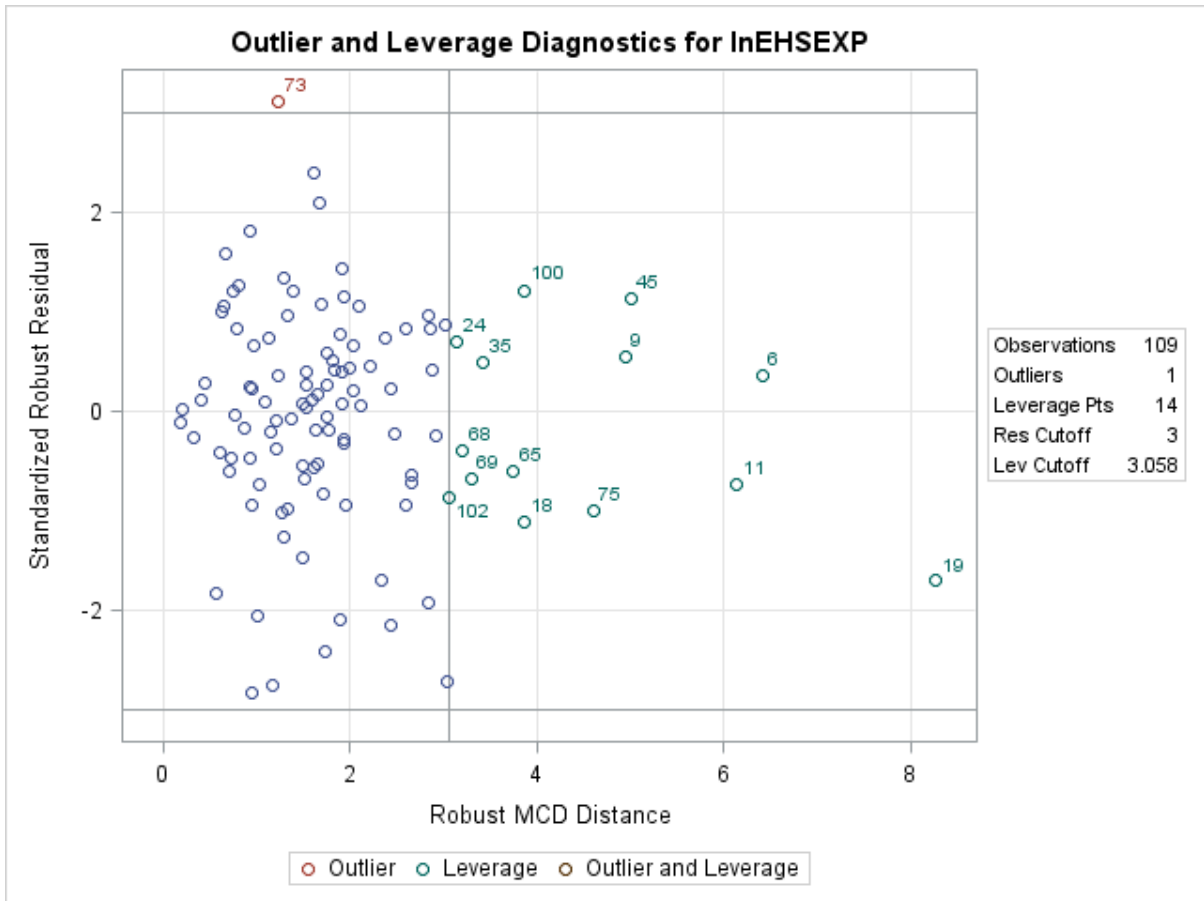


Figure F1. M-estimation RDPLLOT for Model A.

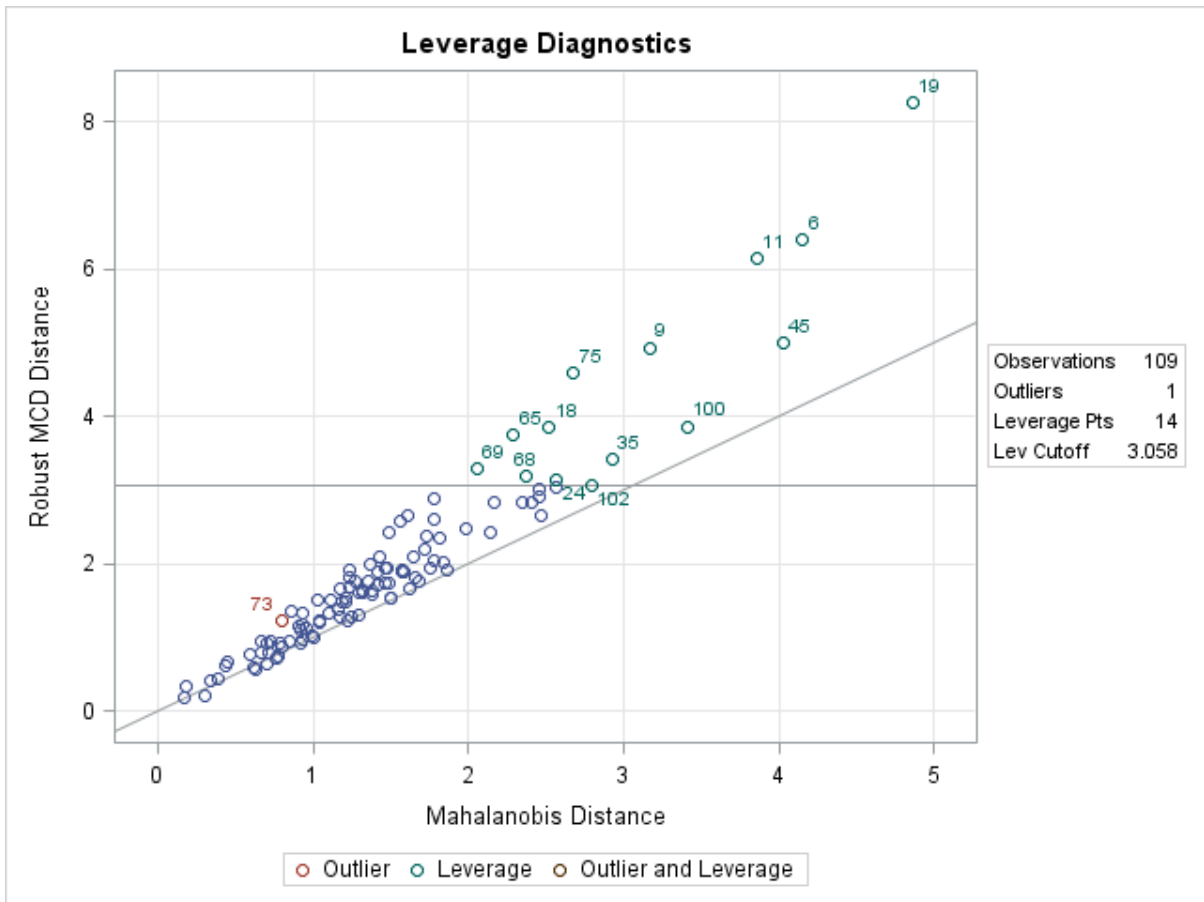


Figure F2. M-estimation DDPLLOT for Model A.

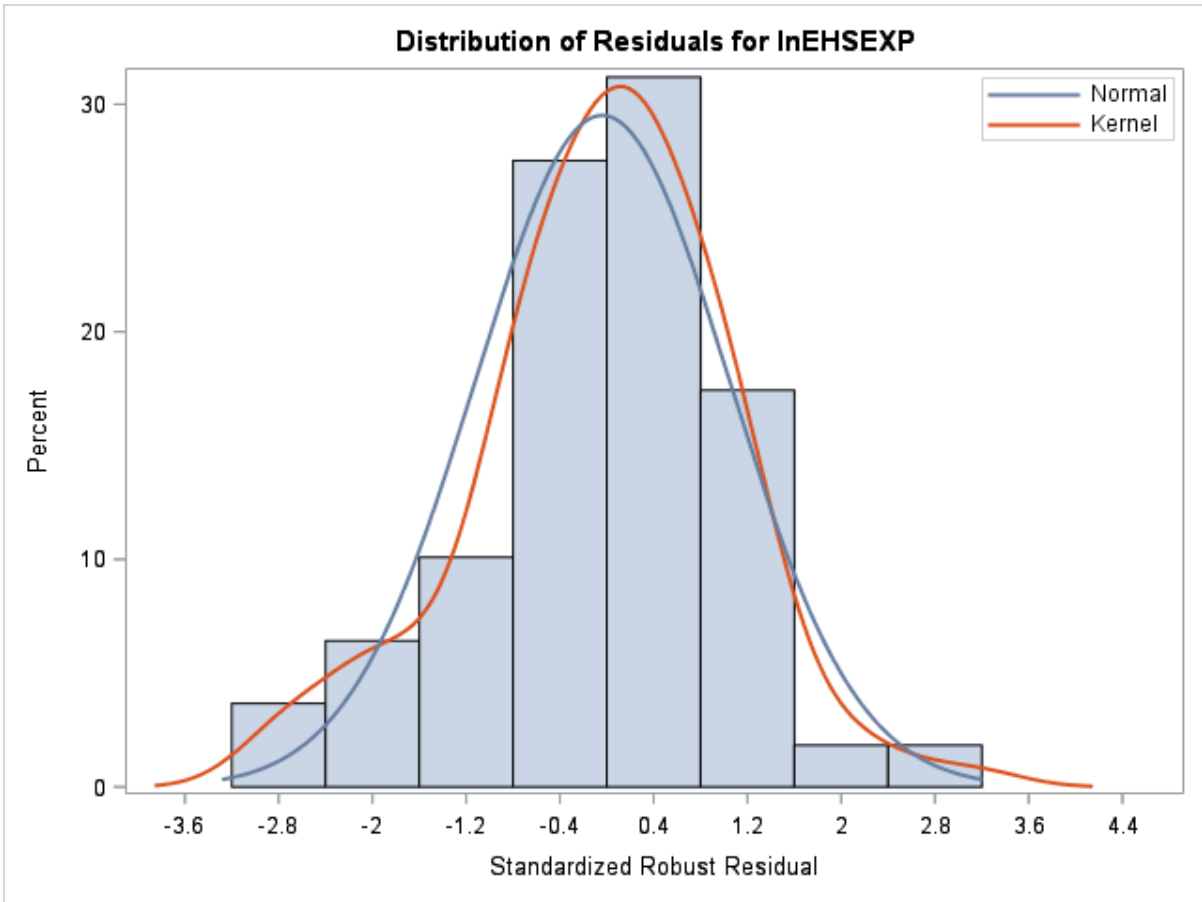


Figure F3. M-estimation Histogram of Standardized Robust Residuals for Model A.

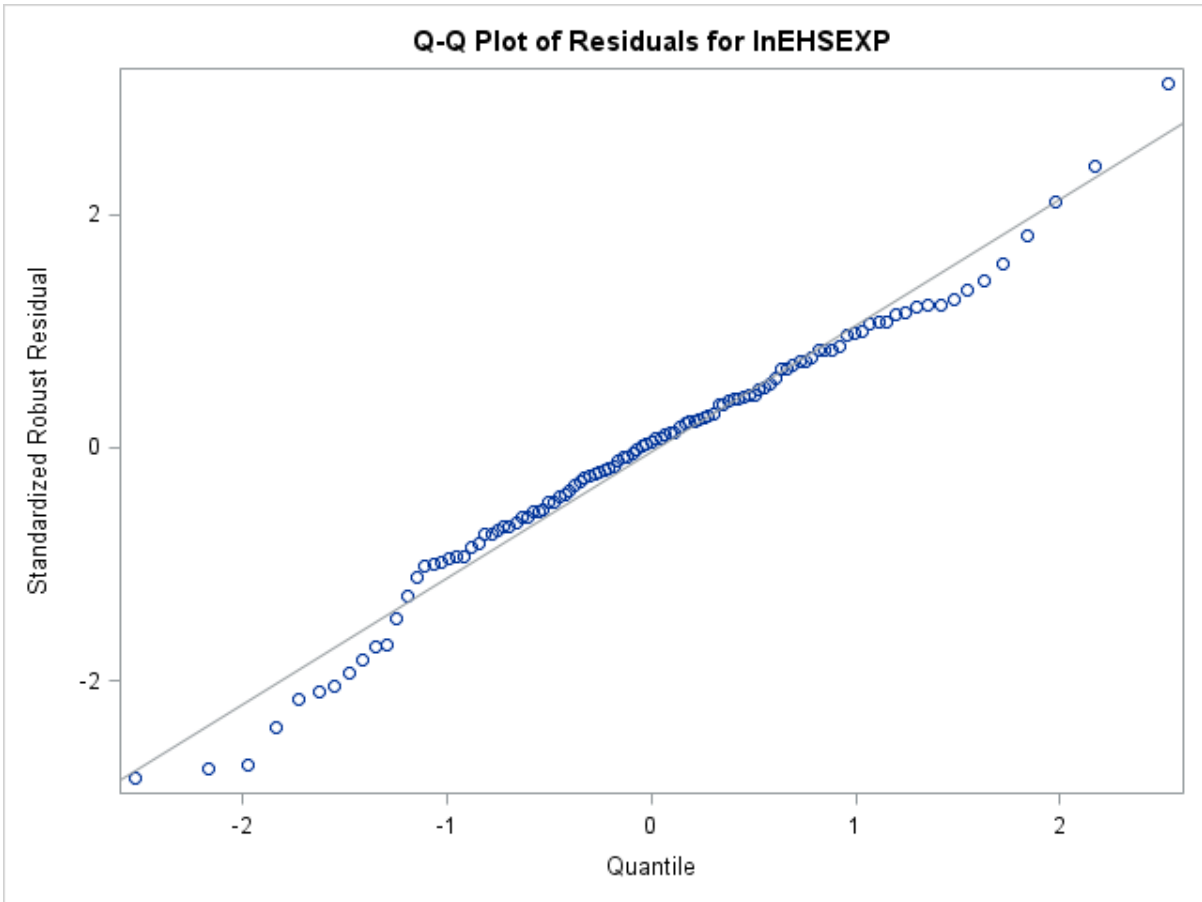


Figure F4. M-estimation Q-Q Plot for Standardized Robust Residuals for Model A.

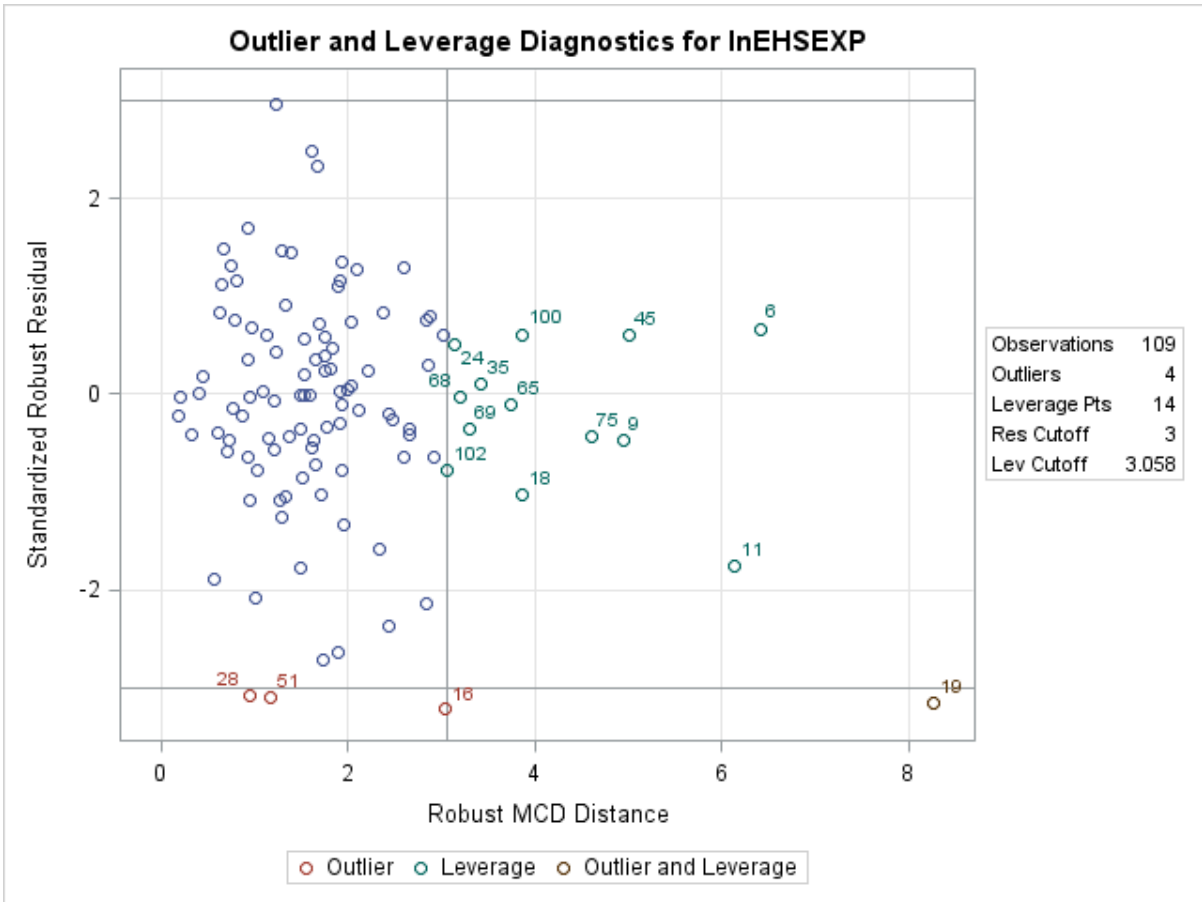


Figure F5. LTS-estimation RDPLLOT for Model A.

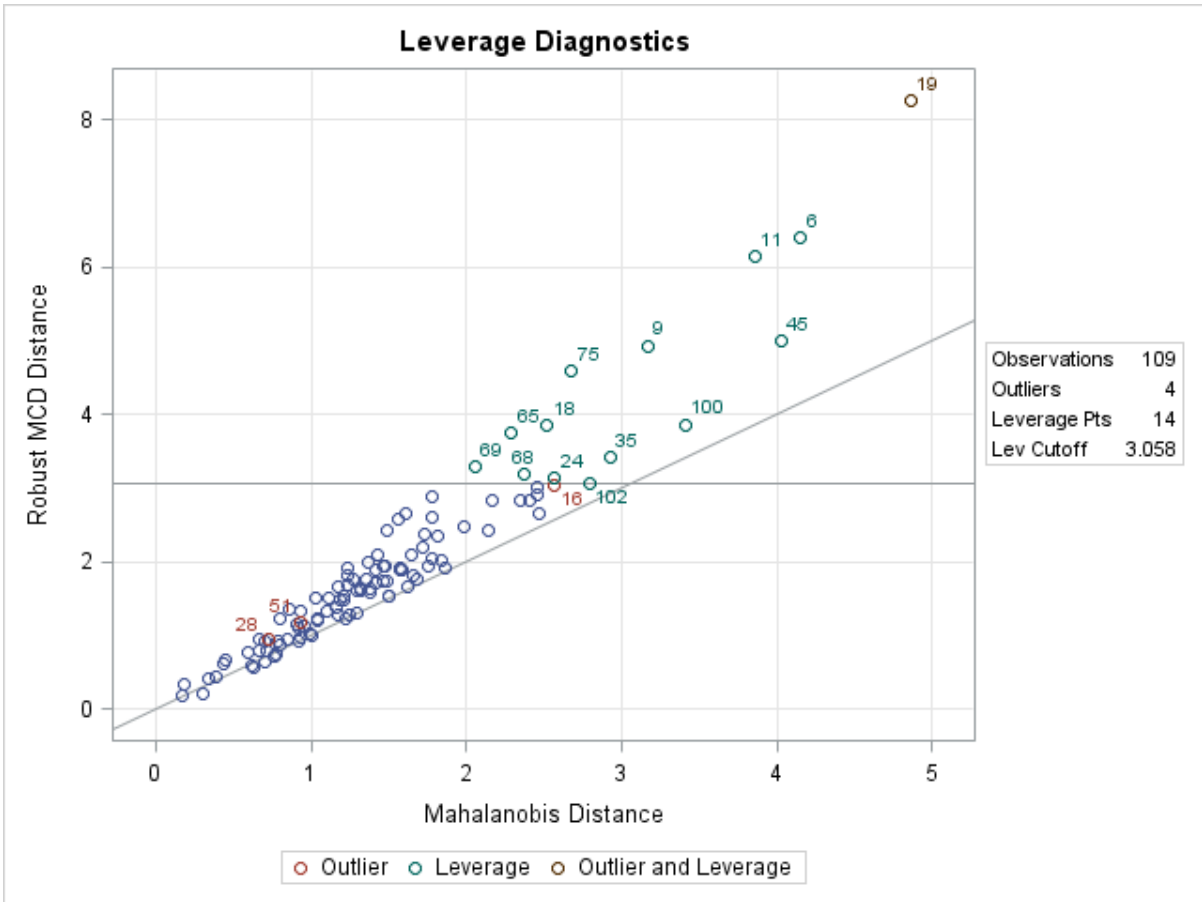


Figure F6. LTS-estimation DDPLLOT for Model A.

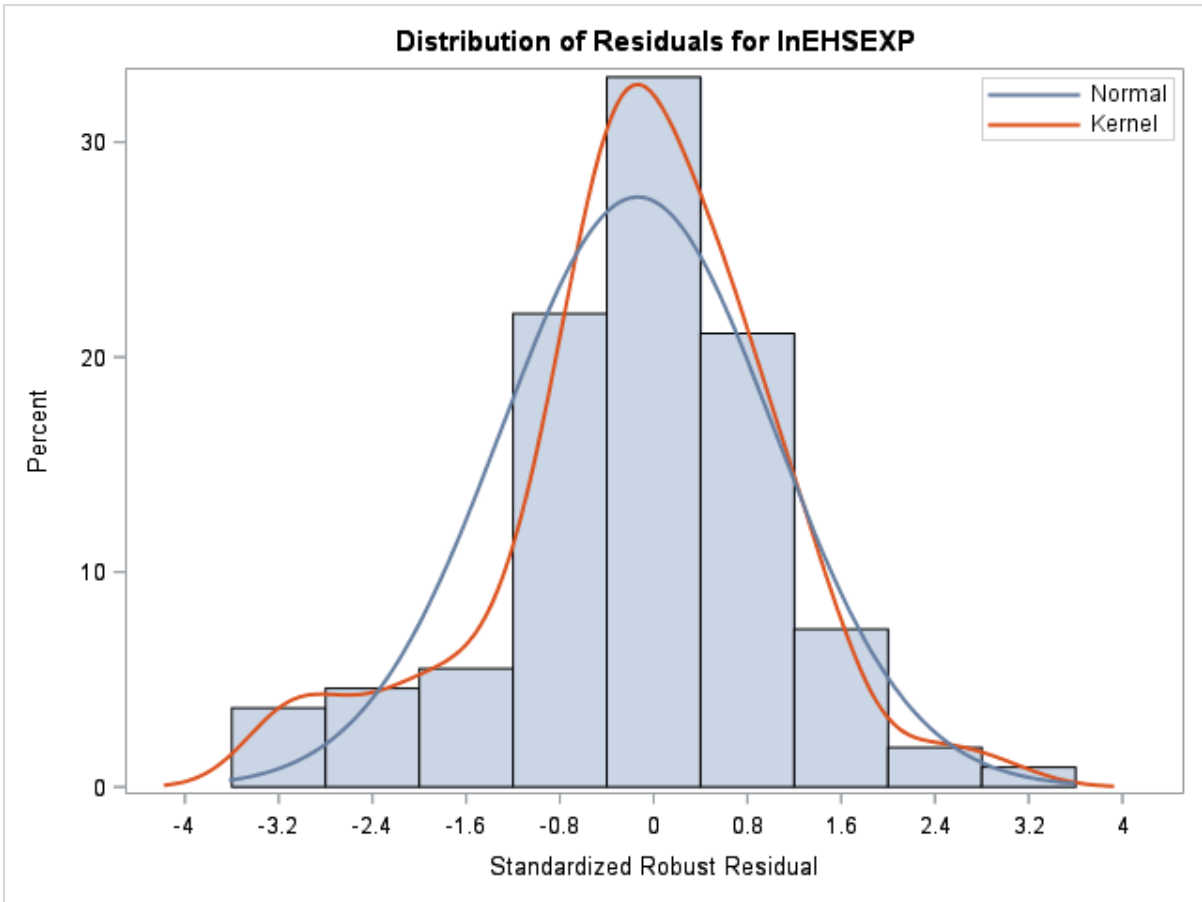


Figure F7. LTS-estimation Histogram of Standardized Robust Residuals for Model A.

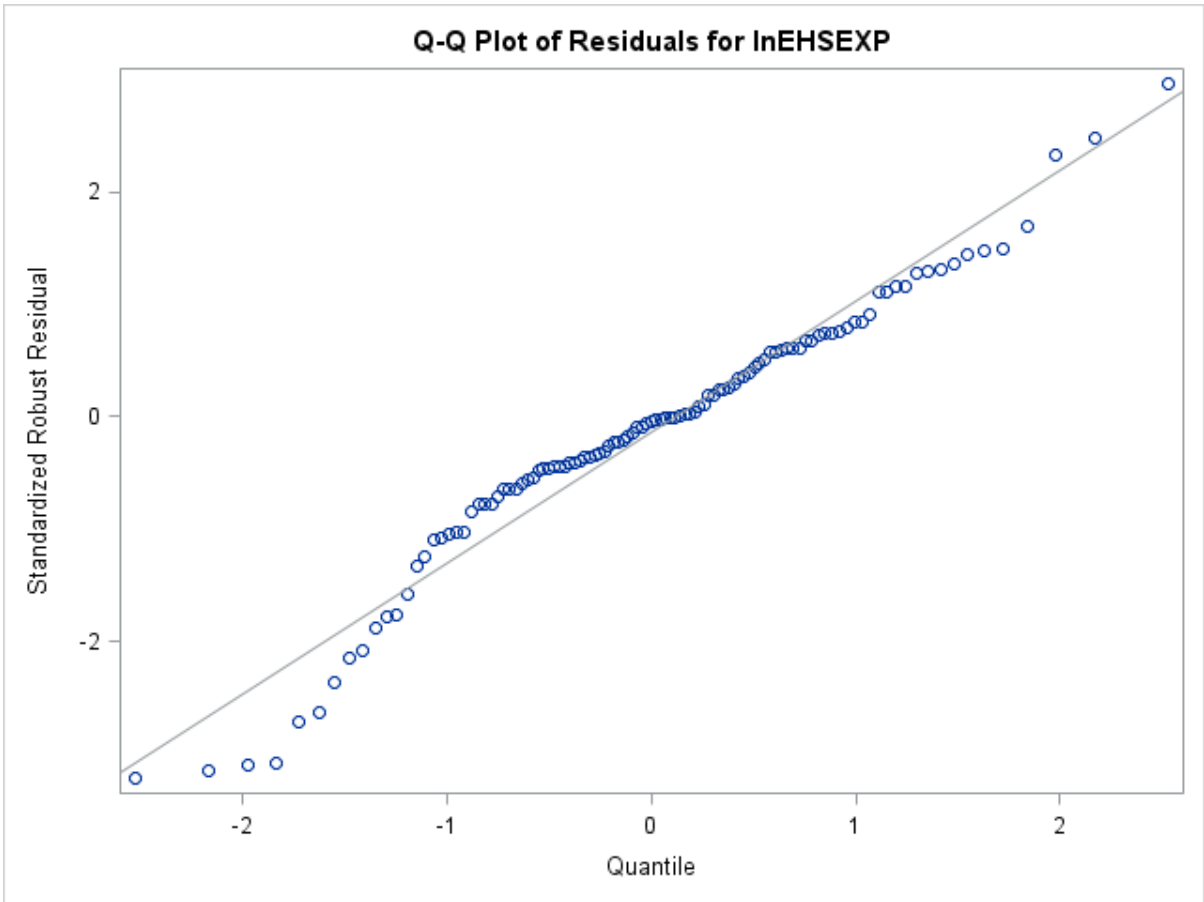


Figure F8. LTS-estimation Q-Q Plot for Standardized Robust Residuals for Model A.

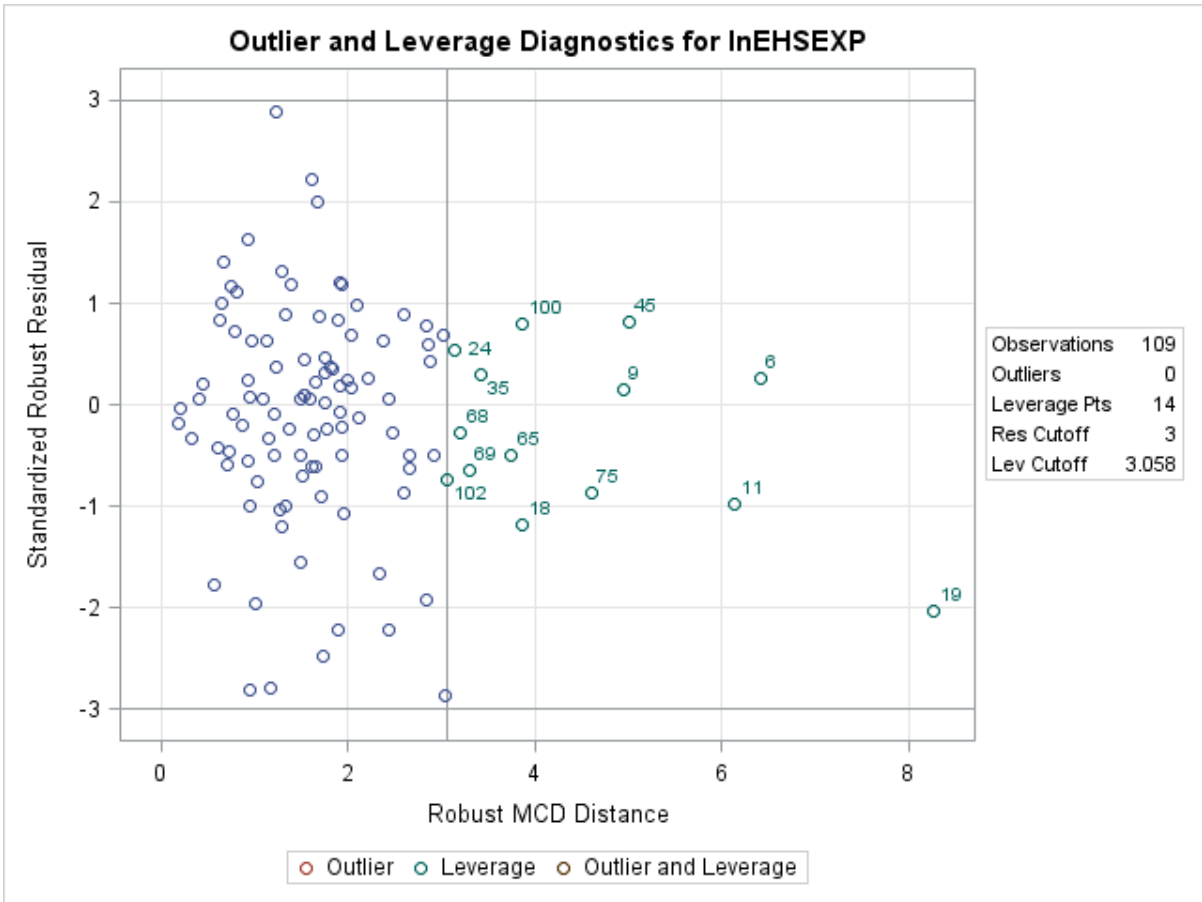


Figure F9. S-estimation RDplot for Model A.

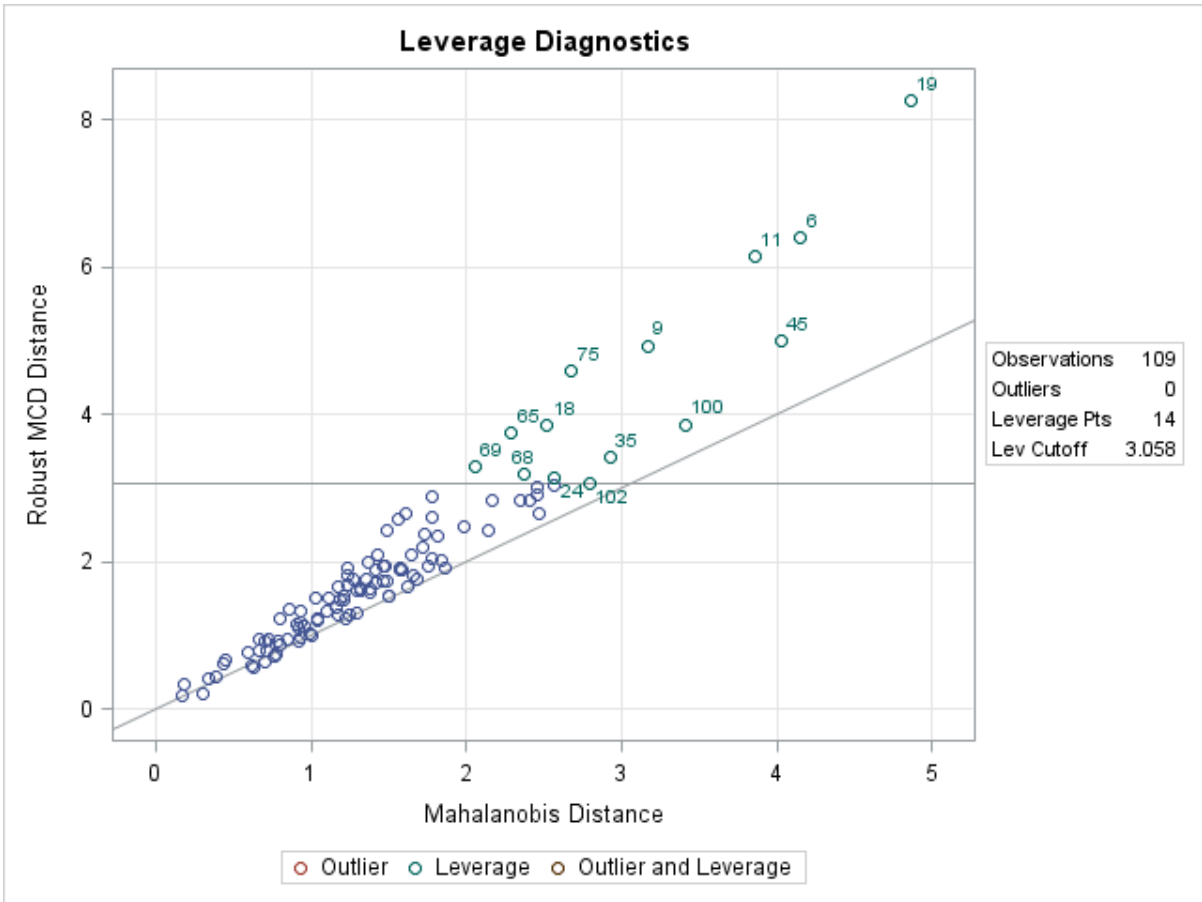


Figure F10. S-estimation DDPLLOT for Model A.

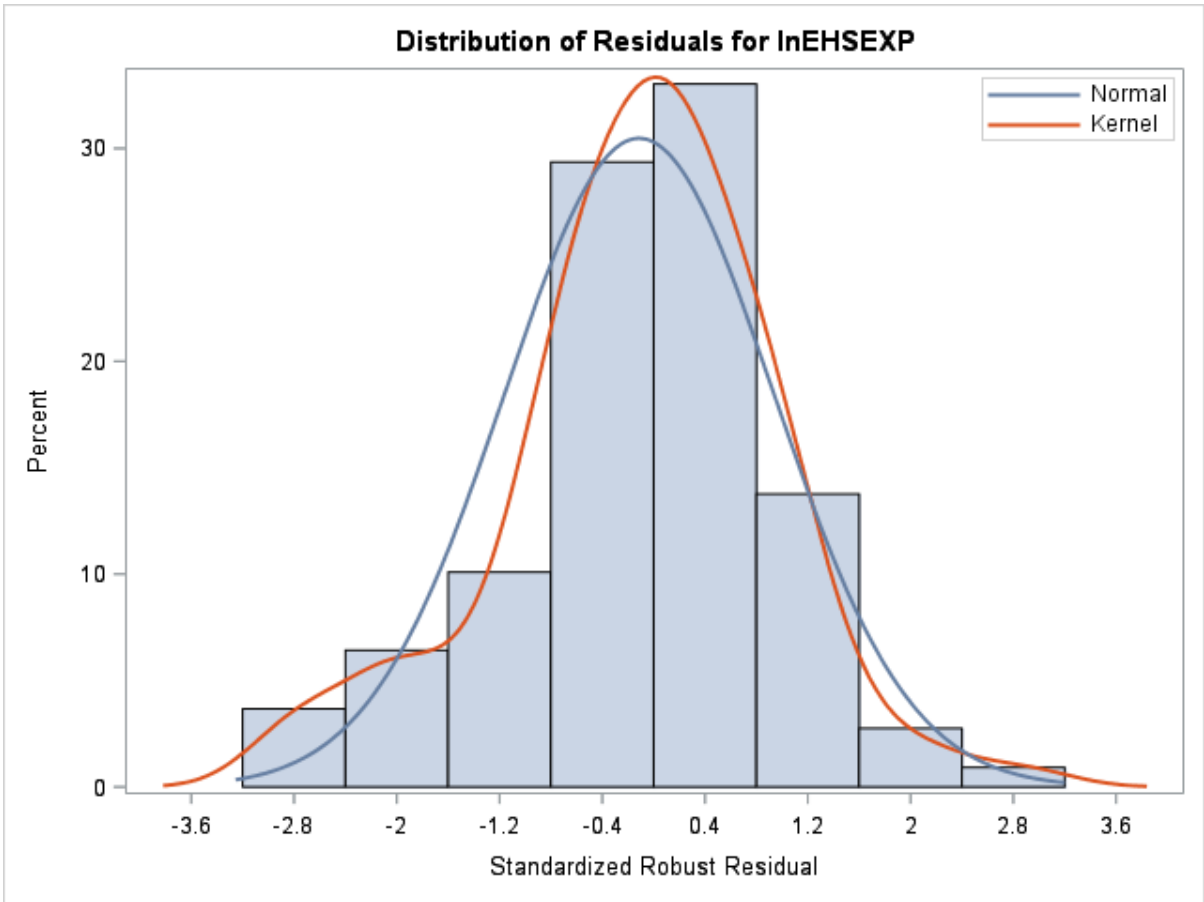


Figure F11. S-estimation Histogram of Standardized Robust Residuals for Model A.

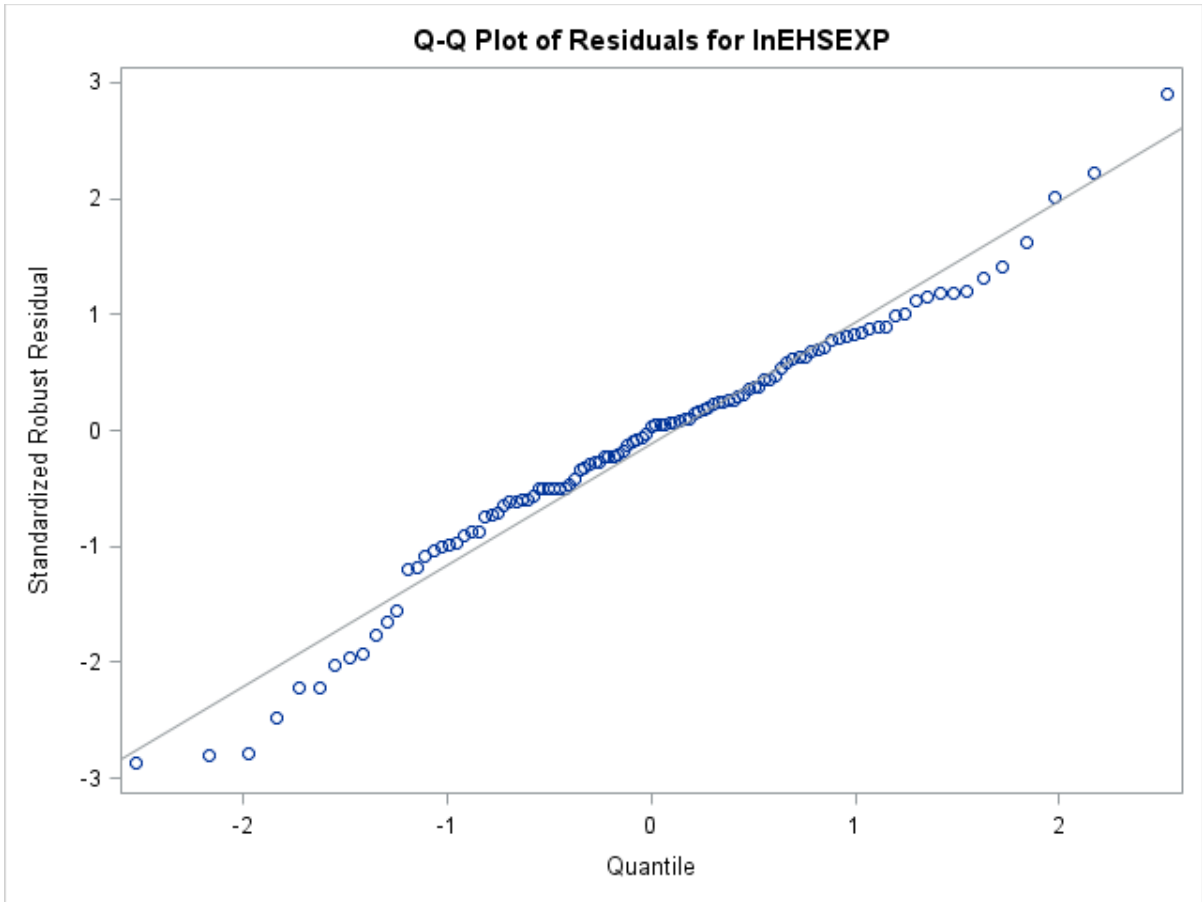


Figure F12. S-estimation Q-Q Plot for Standardized Robust Residuals for Model A.

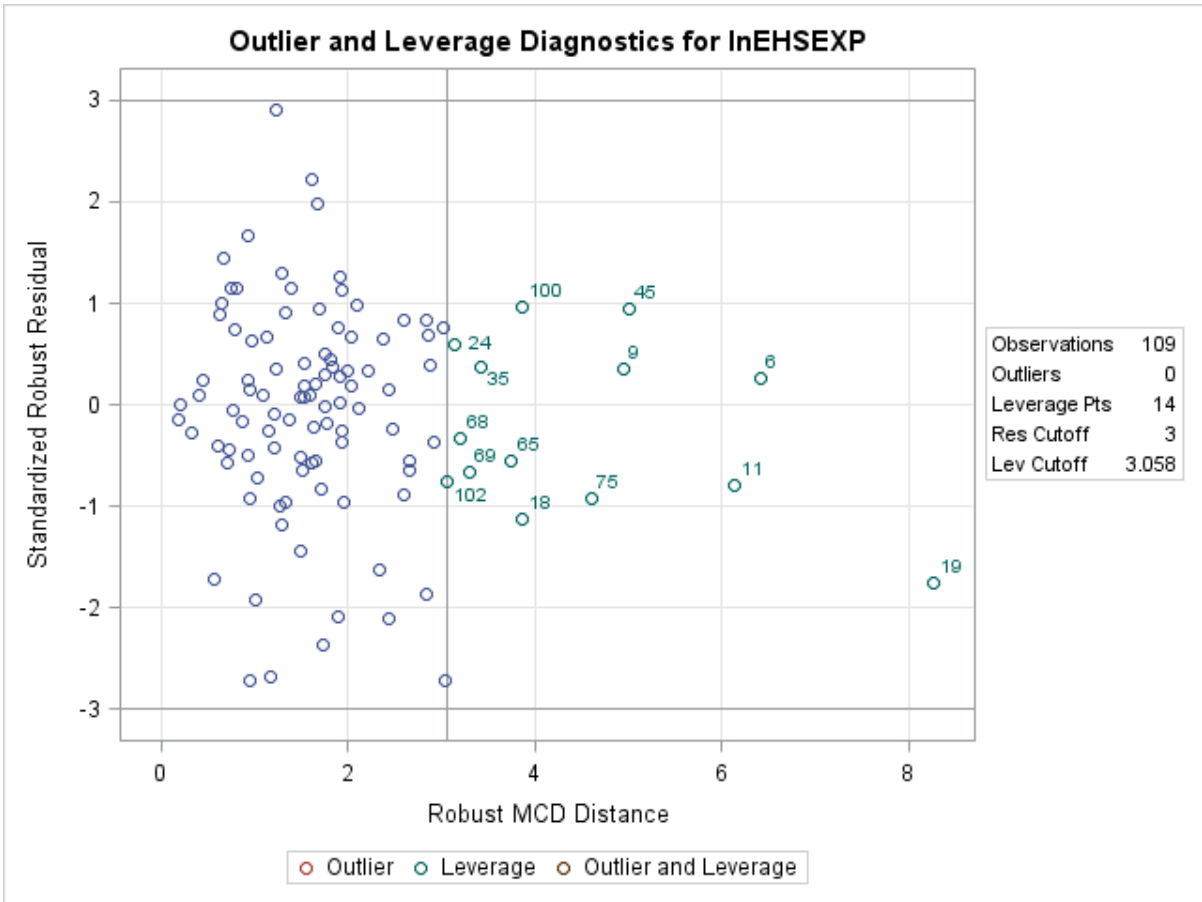


Figure F13. MM-estimation RDPLLOT for Model A.

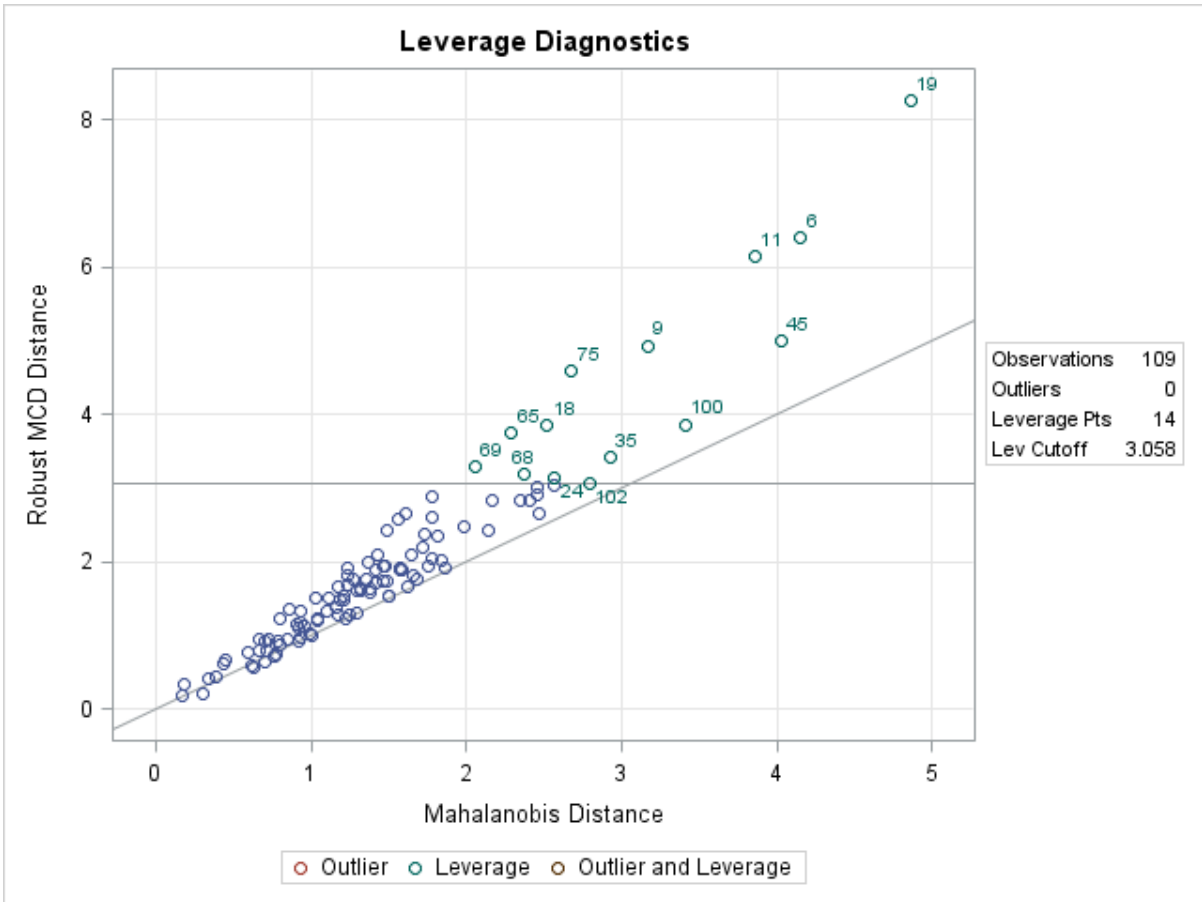


Figure F14. MM-estimation DDPLLOT for Model A.

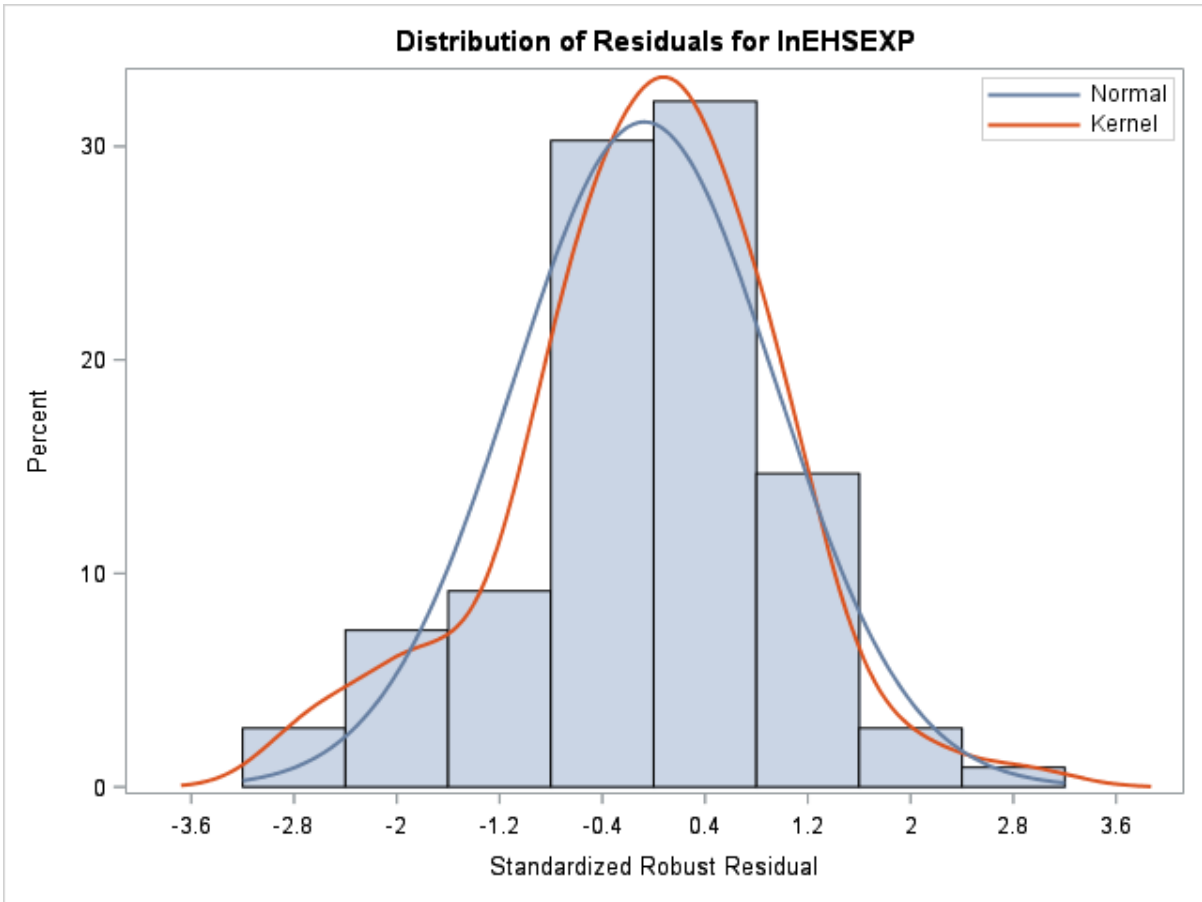


Figure F15. MM-estimation Histogram of Standardized Robust Residuals for Model A.

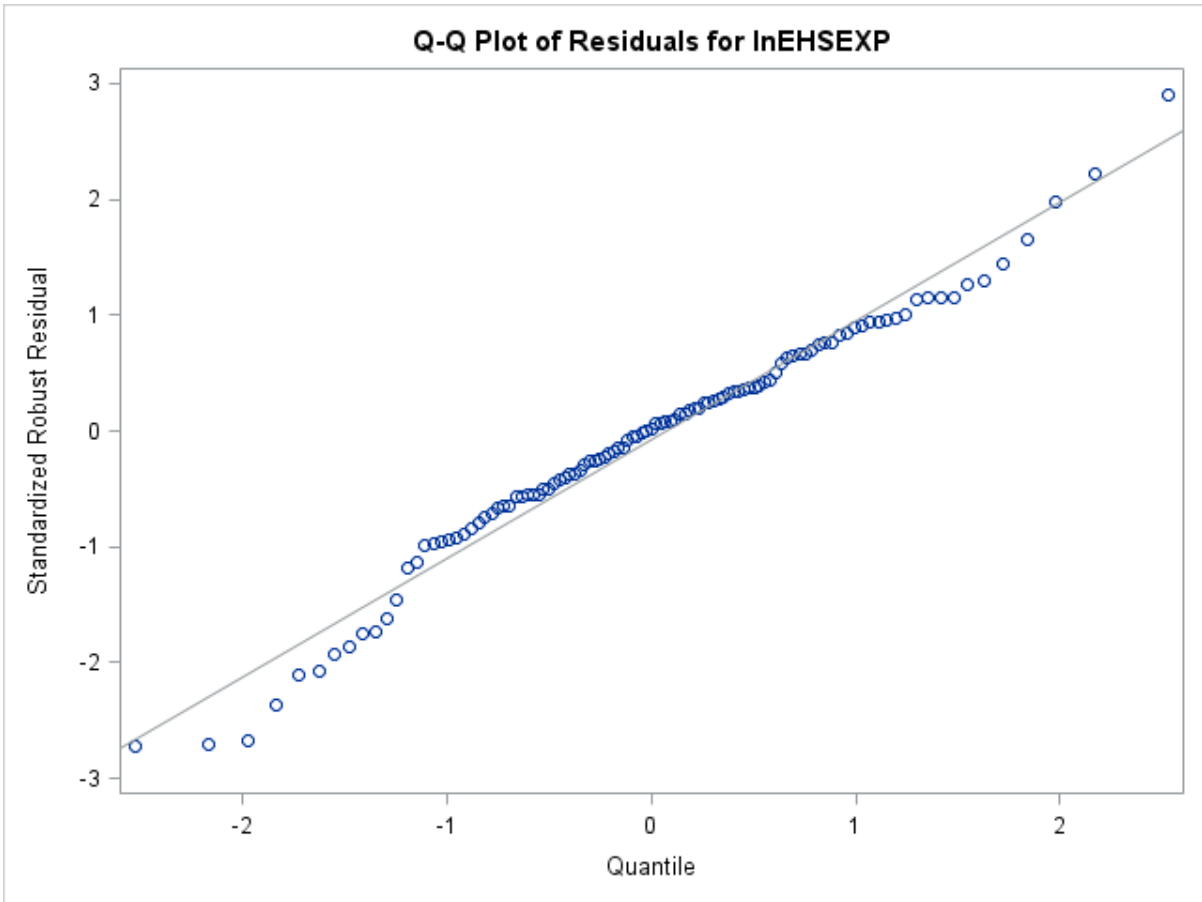


Figure F16. MM-estimation Q-Q Plot for Standardized Robust Residuals for Model A.

APPENDIX G: FIT DIAGNOSTICS FOR THE SQUARE ROOT OF ENVIRONMENTAL HEALTH AND SAFETY FULL-TIME EMPLOYEES.

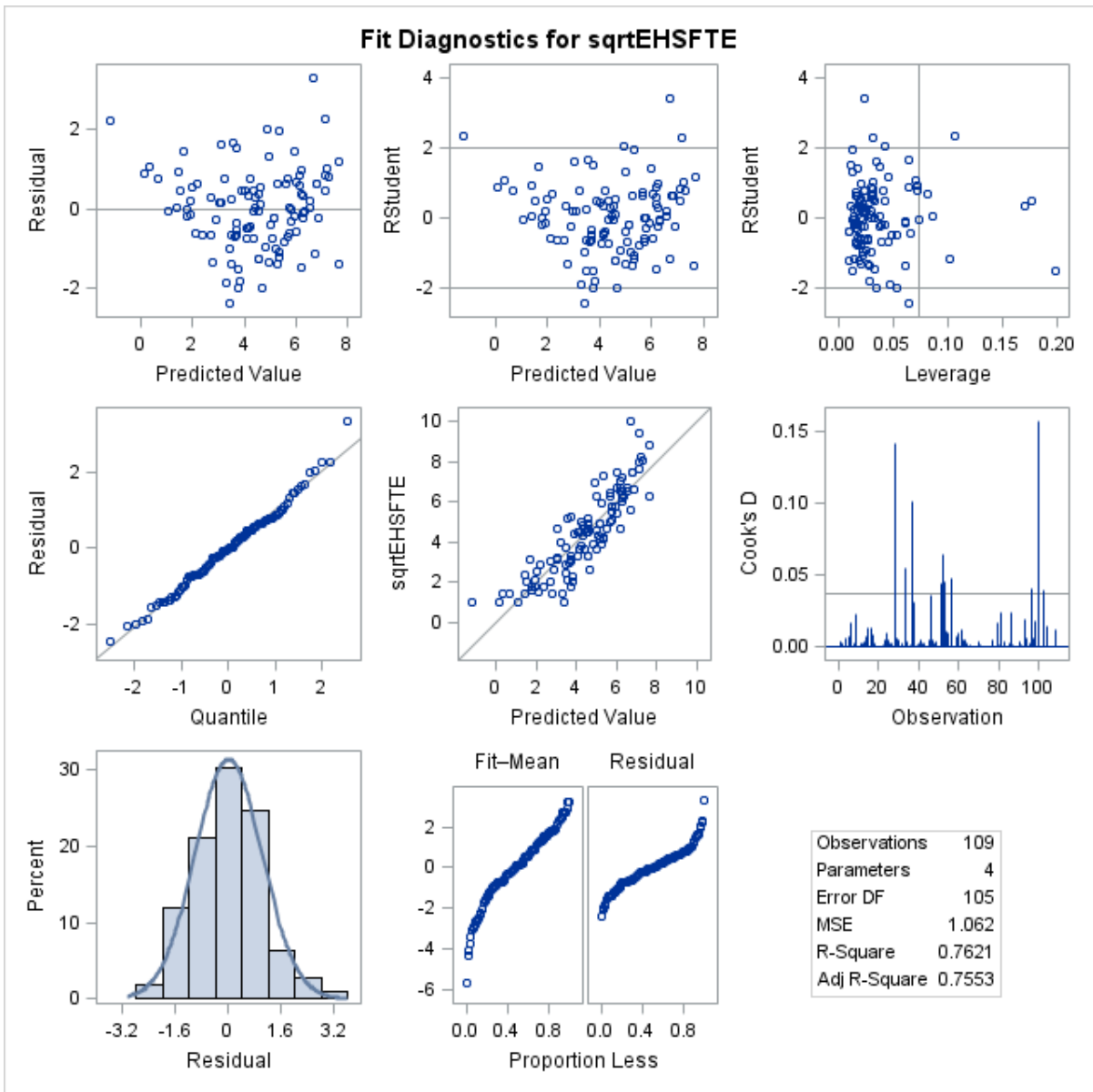


Figure G1. Fit Diagnostics for Model D.

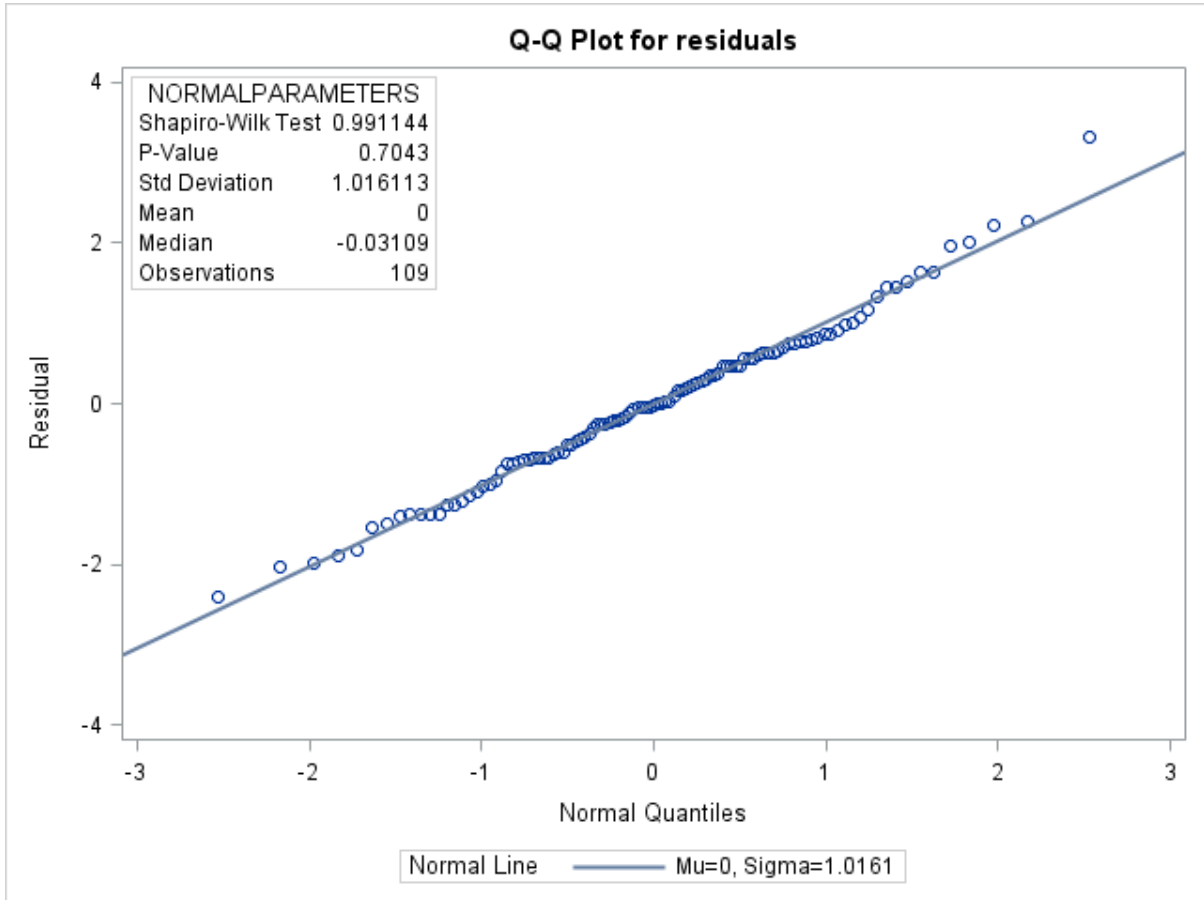


Figure G2. Q-Q Plot of Residuals with Shapiro-Wilk's Test for Model D.

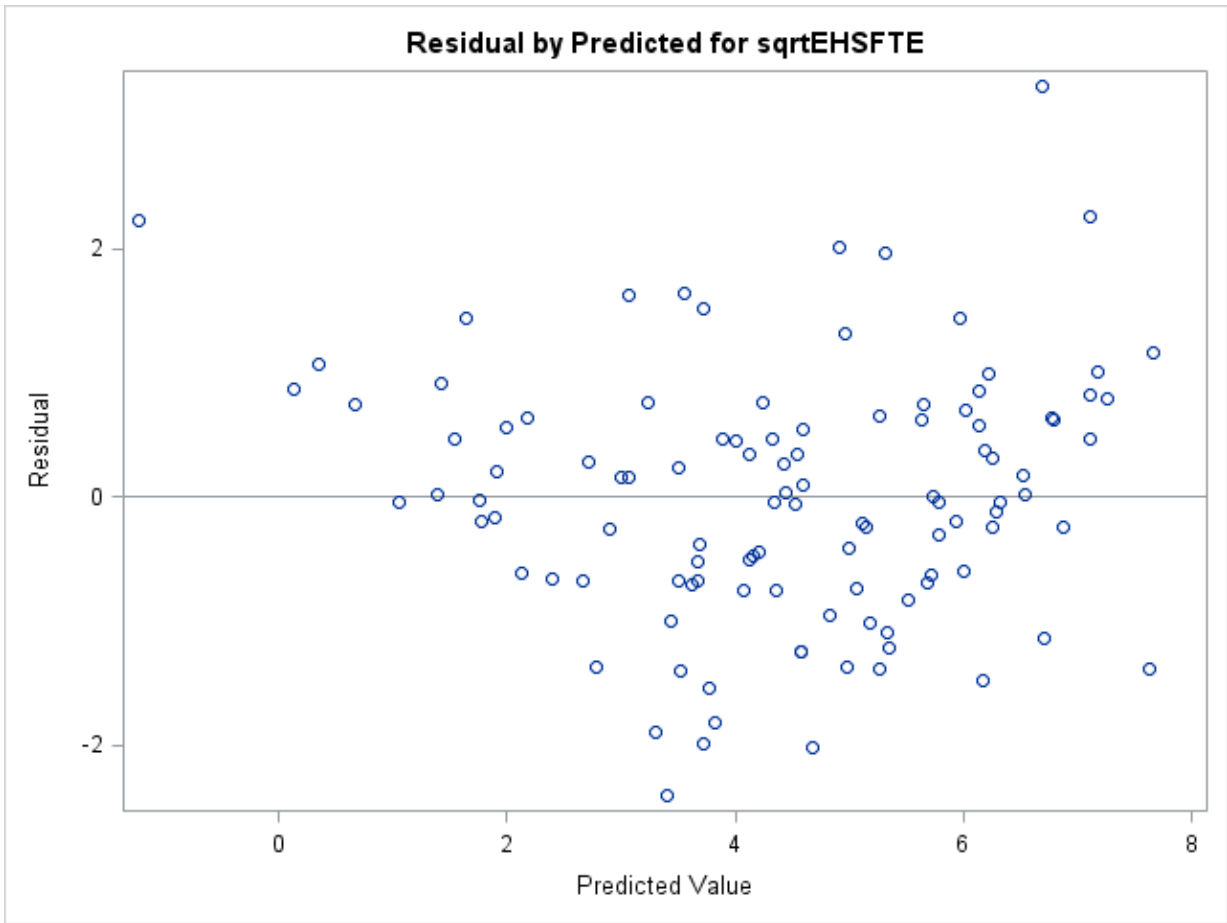


Figure G3. Residuals by Predicted for Model D.

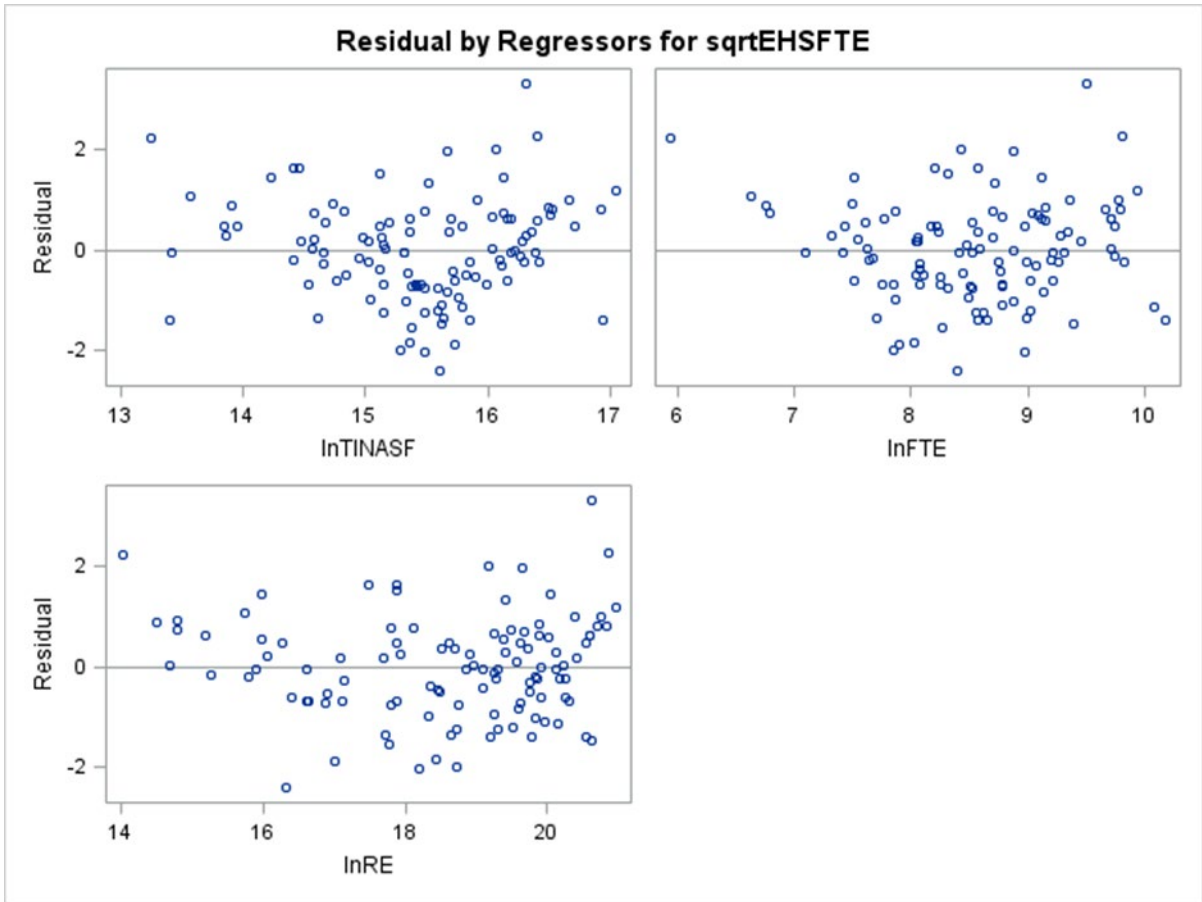


Figure G4. Residuals by Regressor for Model D.

APPENDIX H: OUTLIER ANALYSIS FOR MODEL D.

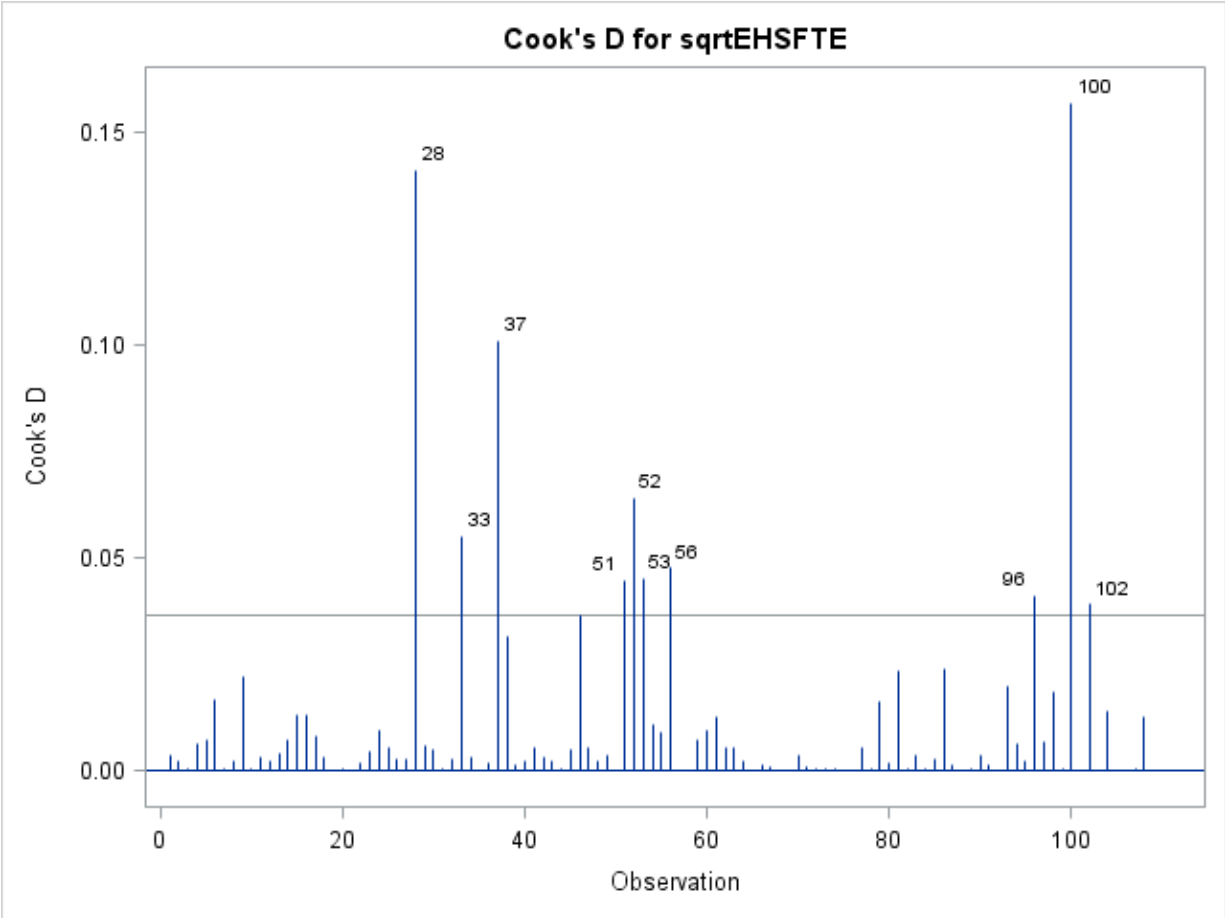


Figure H1. Cook's D for Model D.

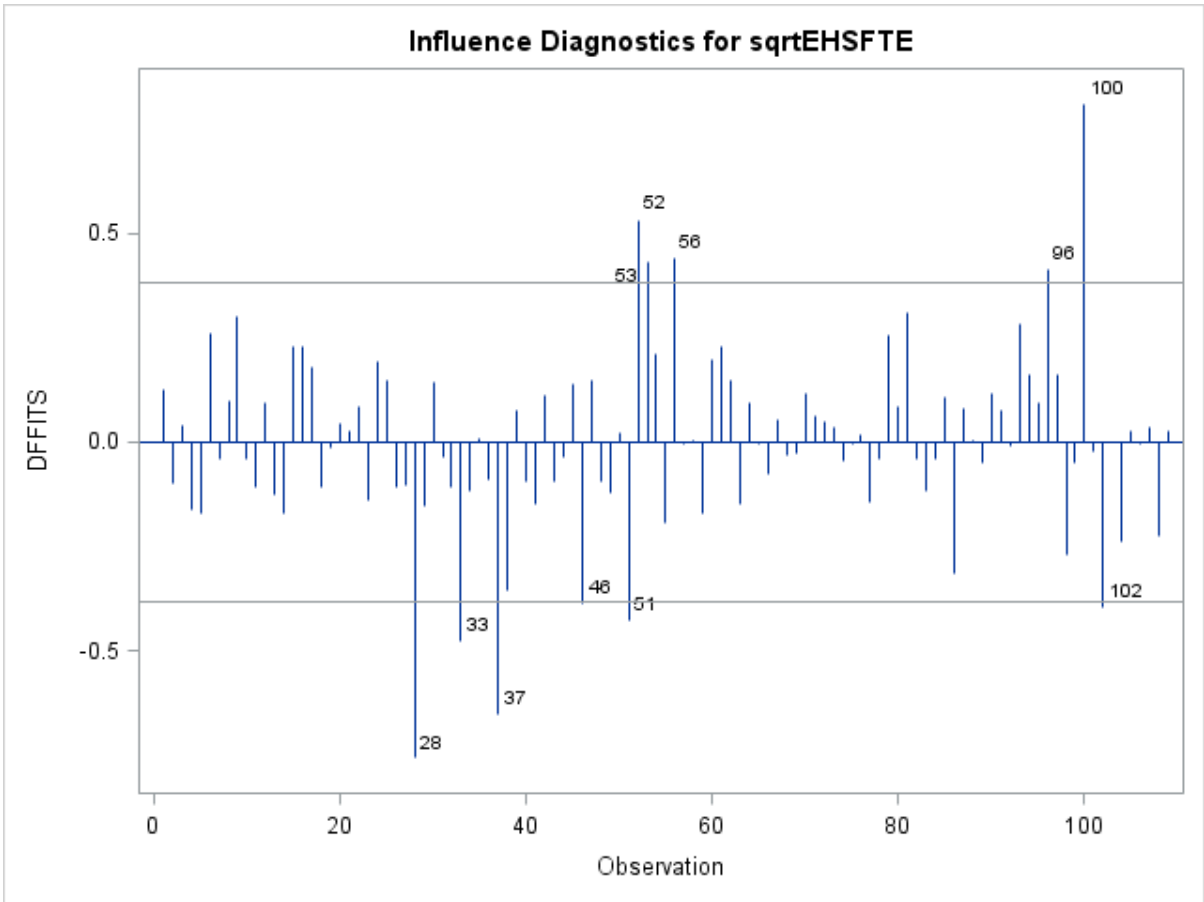


Figure H2. Difference in Fits (DFFITS) Influence Diagnostics for Model D.

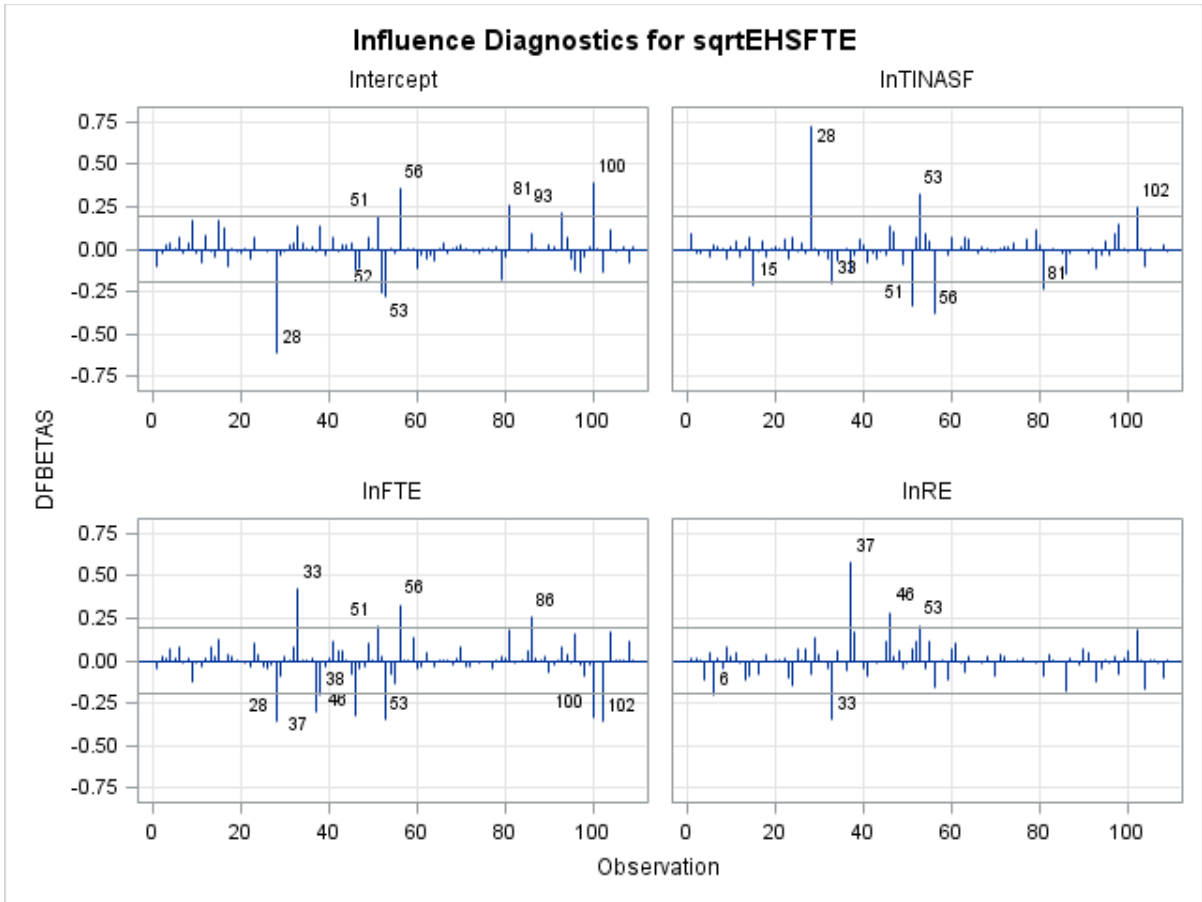


Figure H3. Difference in Betas (DFBETAS) Influence Diagnostics for Model D.

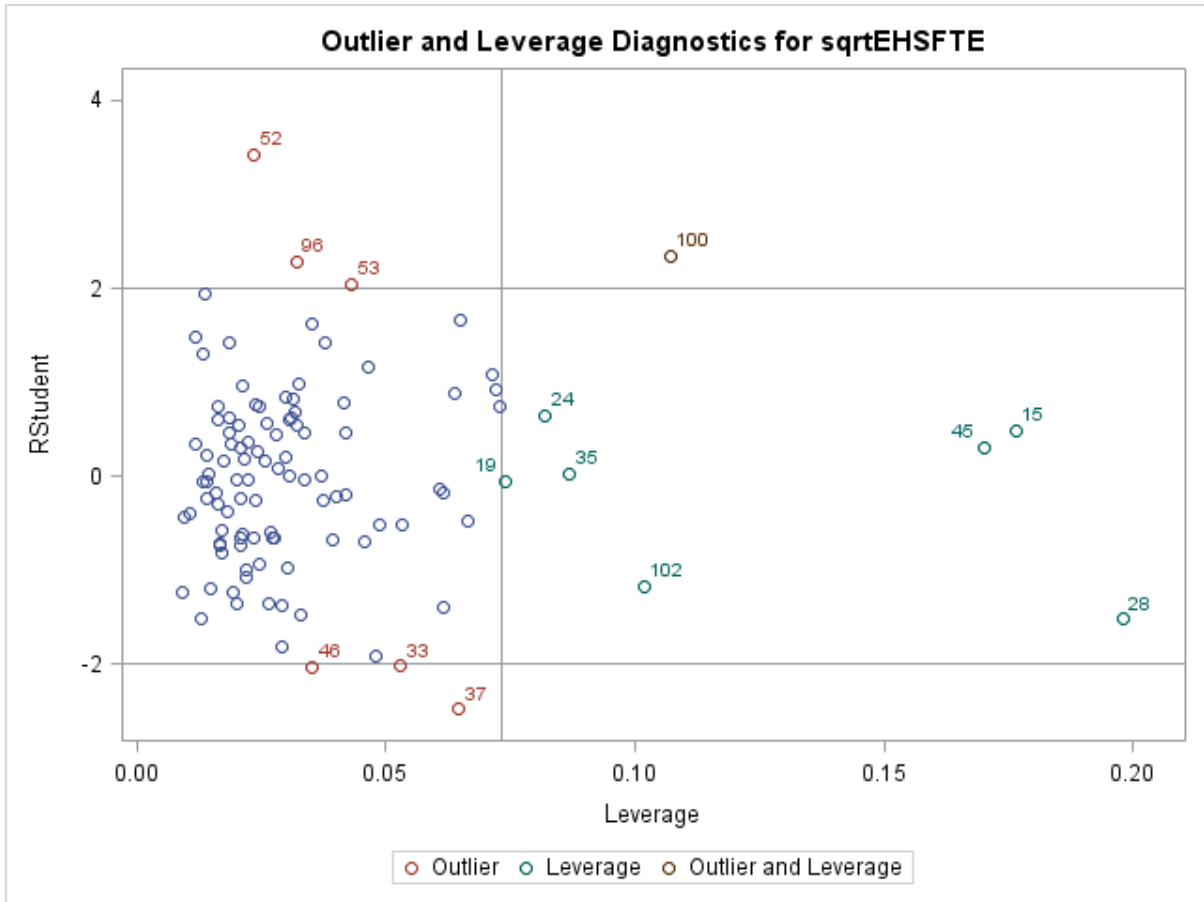


Figure H4. Outlier and Leverage Diagnostics using OLS for Model D.

APPENDIX I: ROBUST REGRESSION ESTIMATION FOR MODEL D.

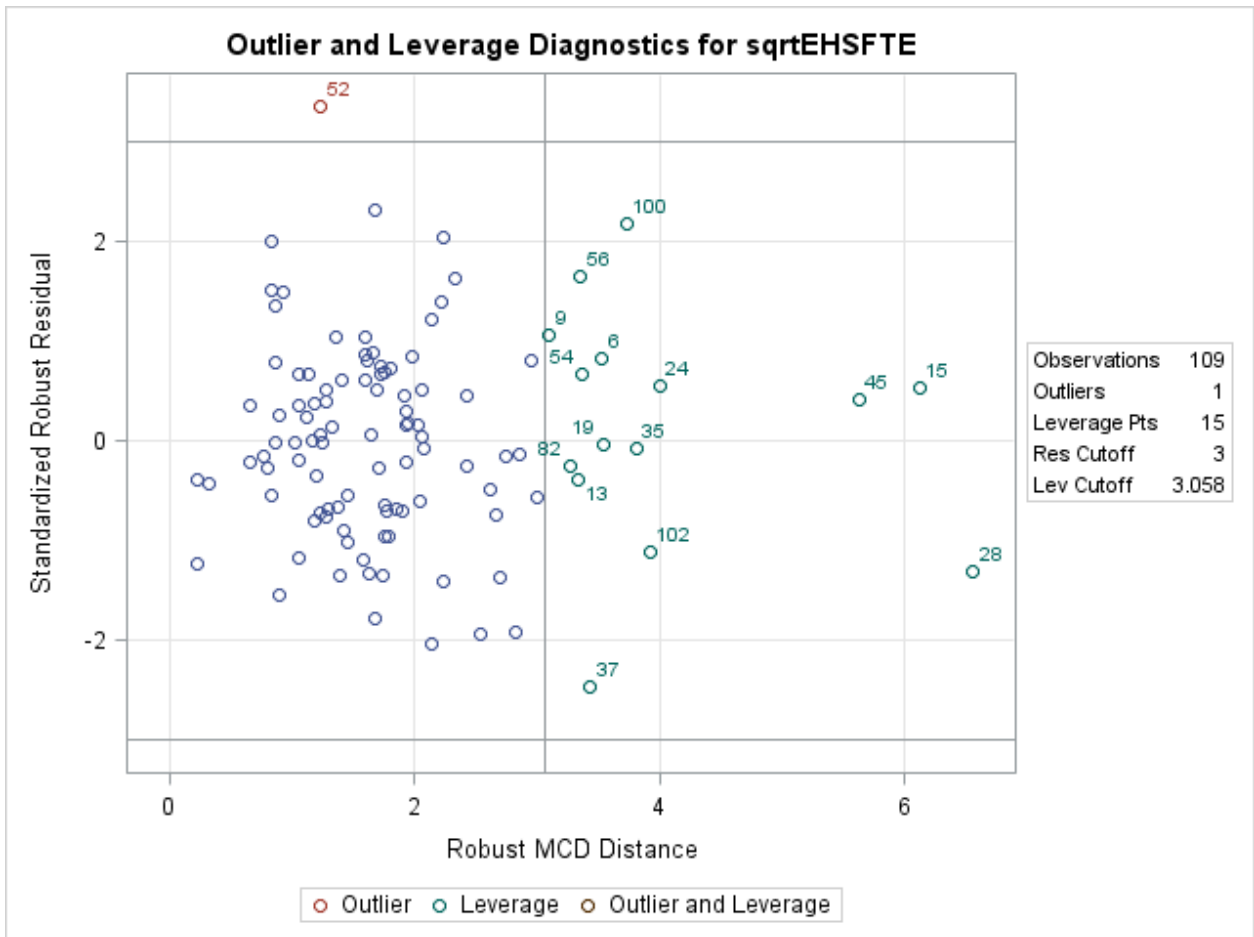


Figure I1. M-estimation RDPLLOT for Model D.

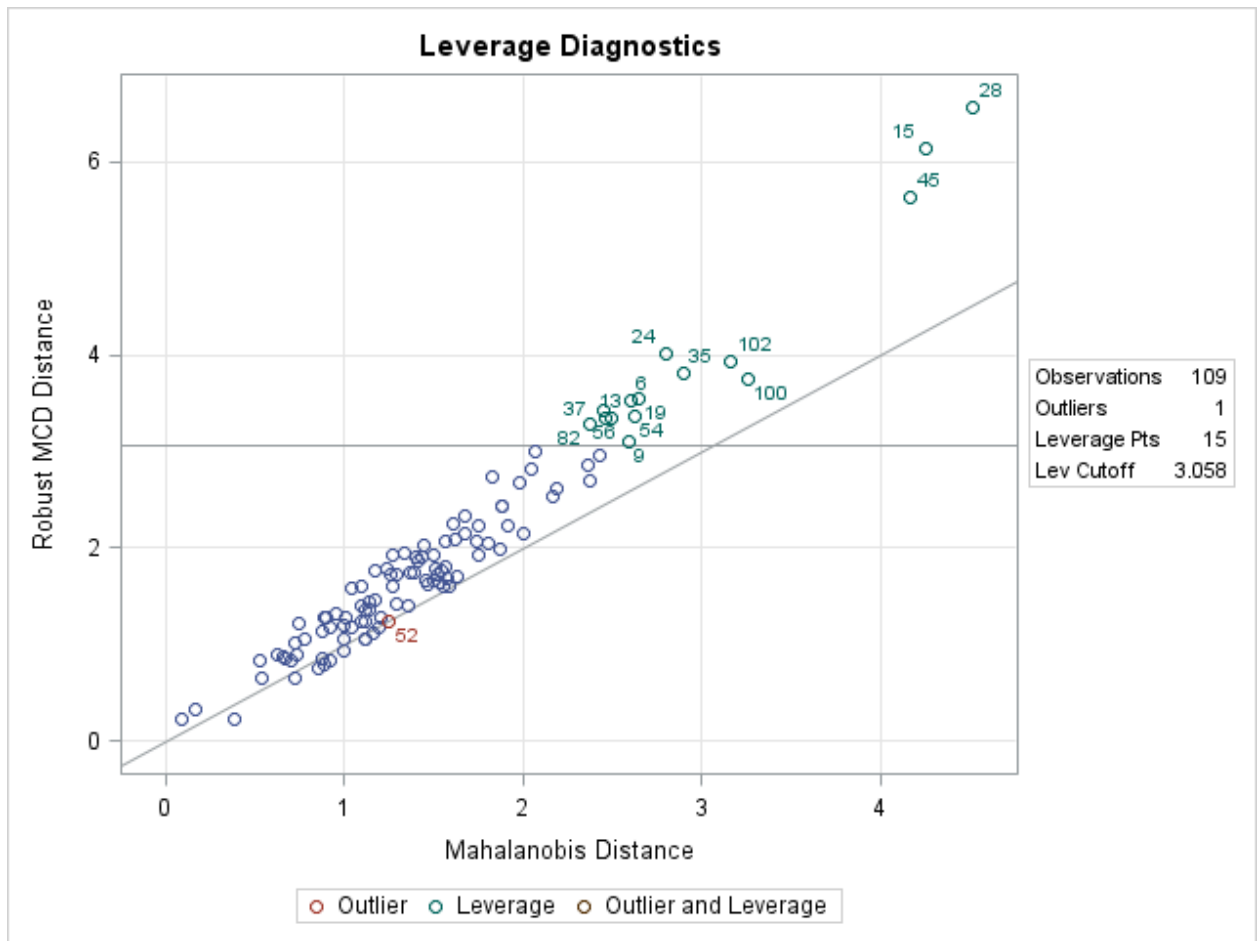


Figure I2. M-estimation DDPLLOT for Model D.

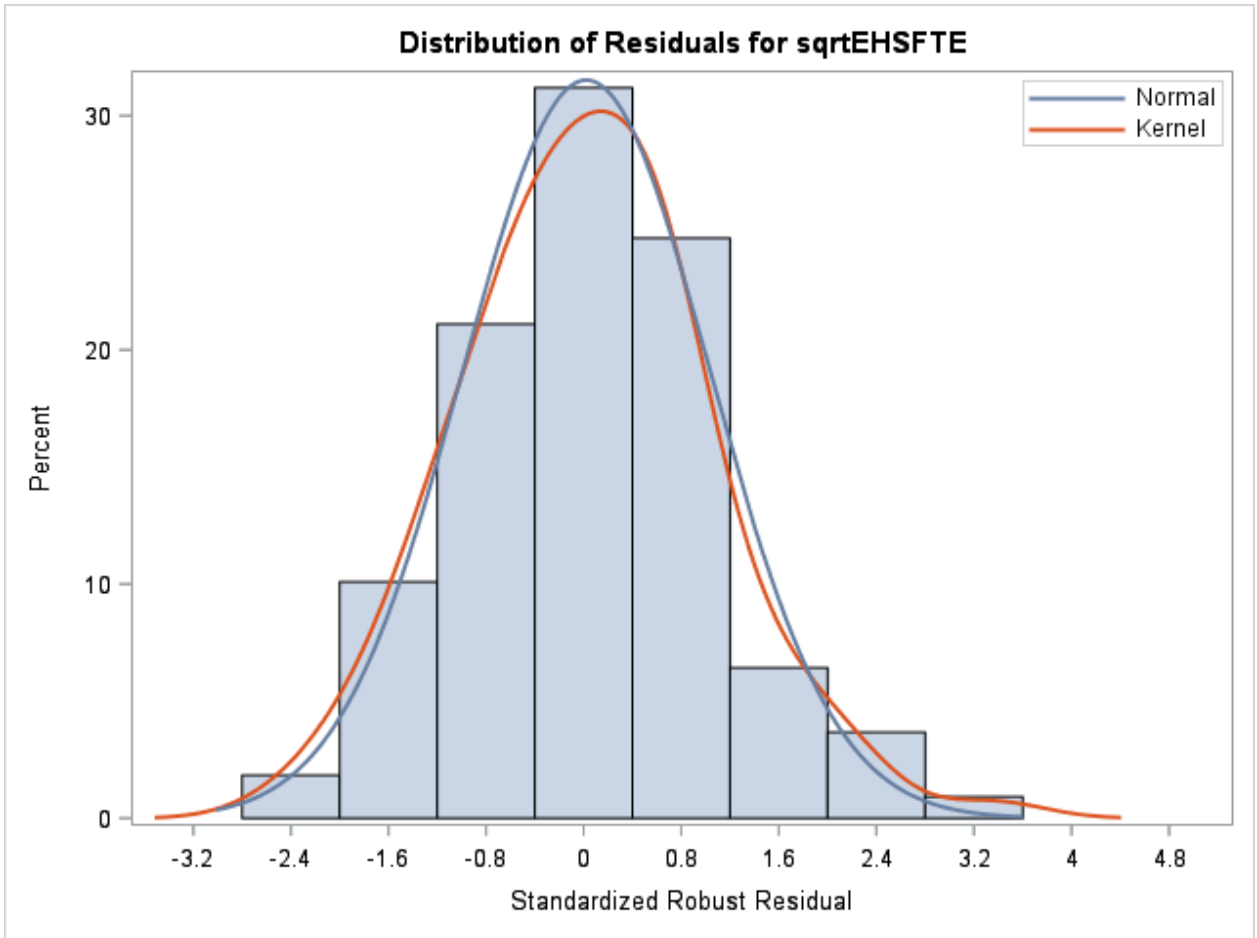


Figure I3. M-estimation Histogram of Standardized Robust Residuals for Model D.

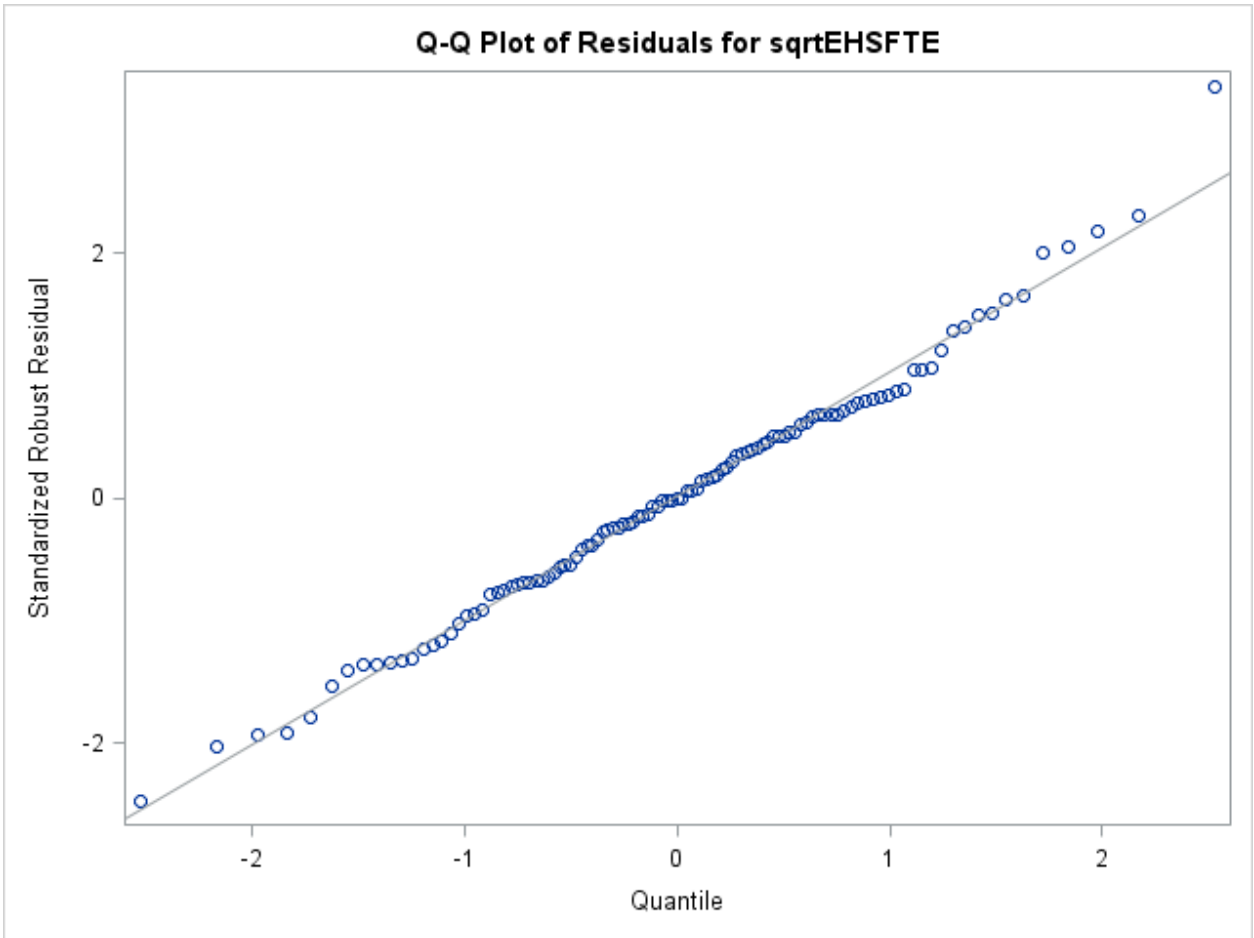


Figure I4. M-estimation Q-Q Plot for Standardized Robust Residuals for Model D.

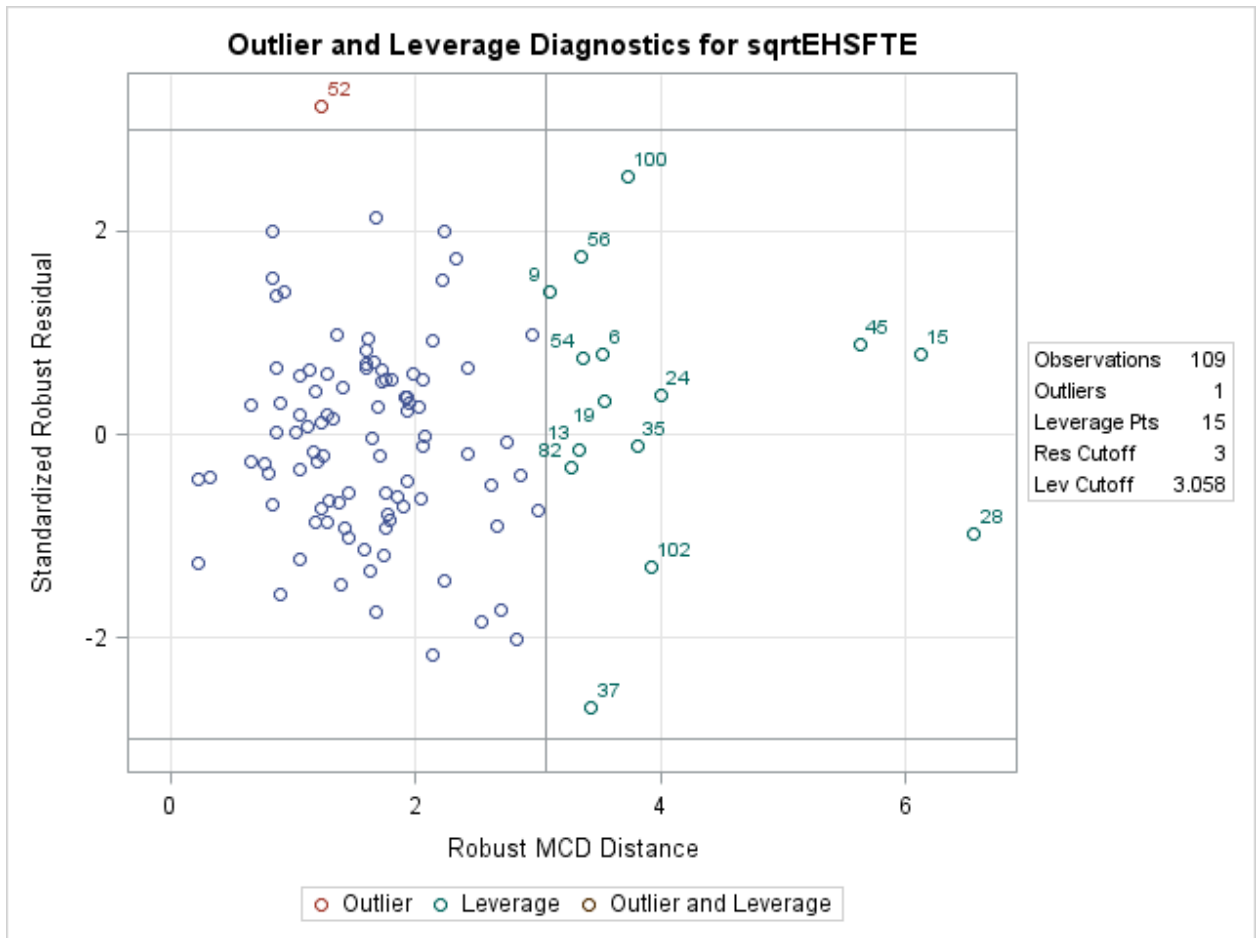


Figure I5. LTS-estimation RDPLLOT Model D.

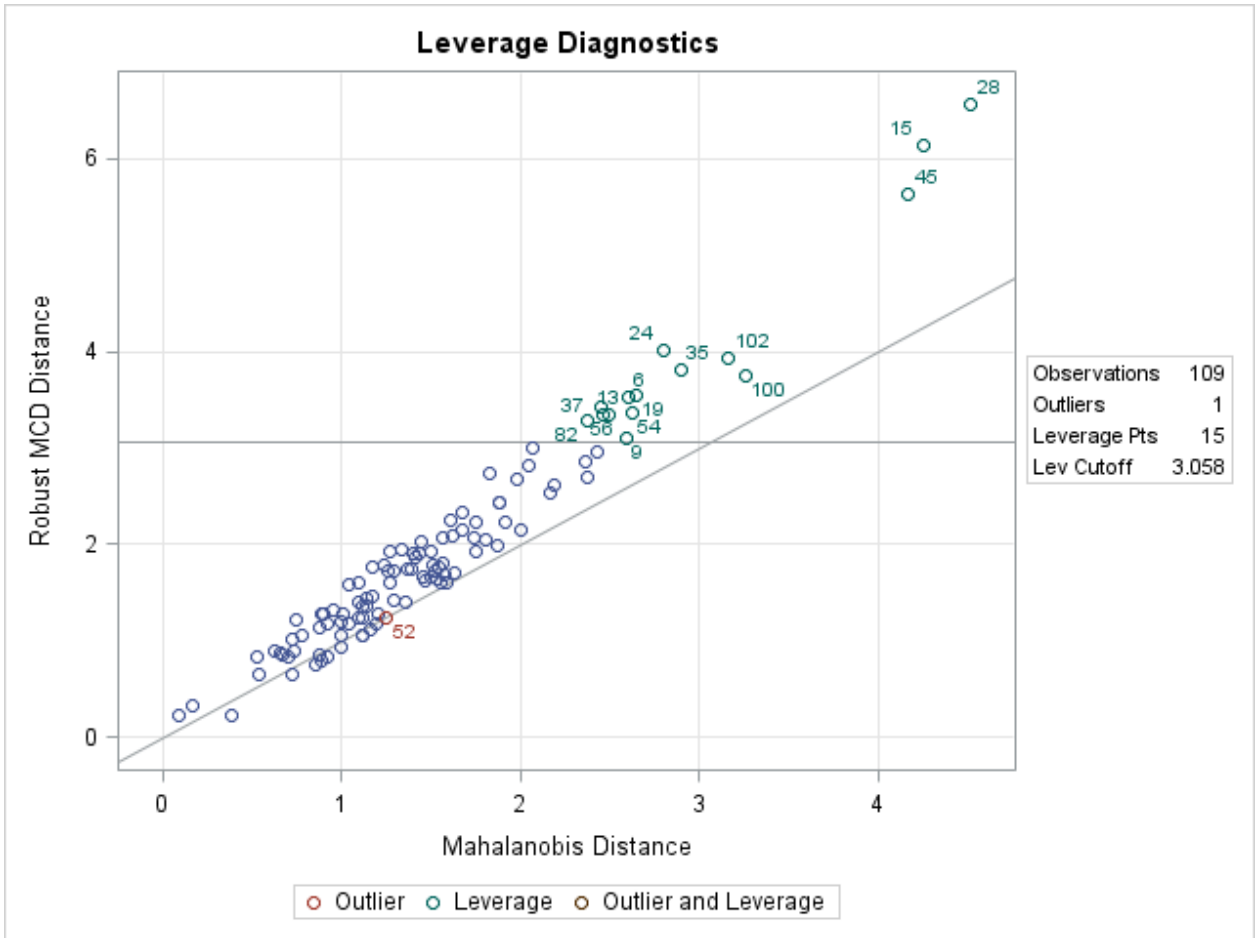


Figure I6. LTS-estimation DDplot for Model D.

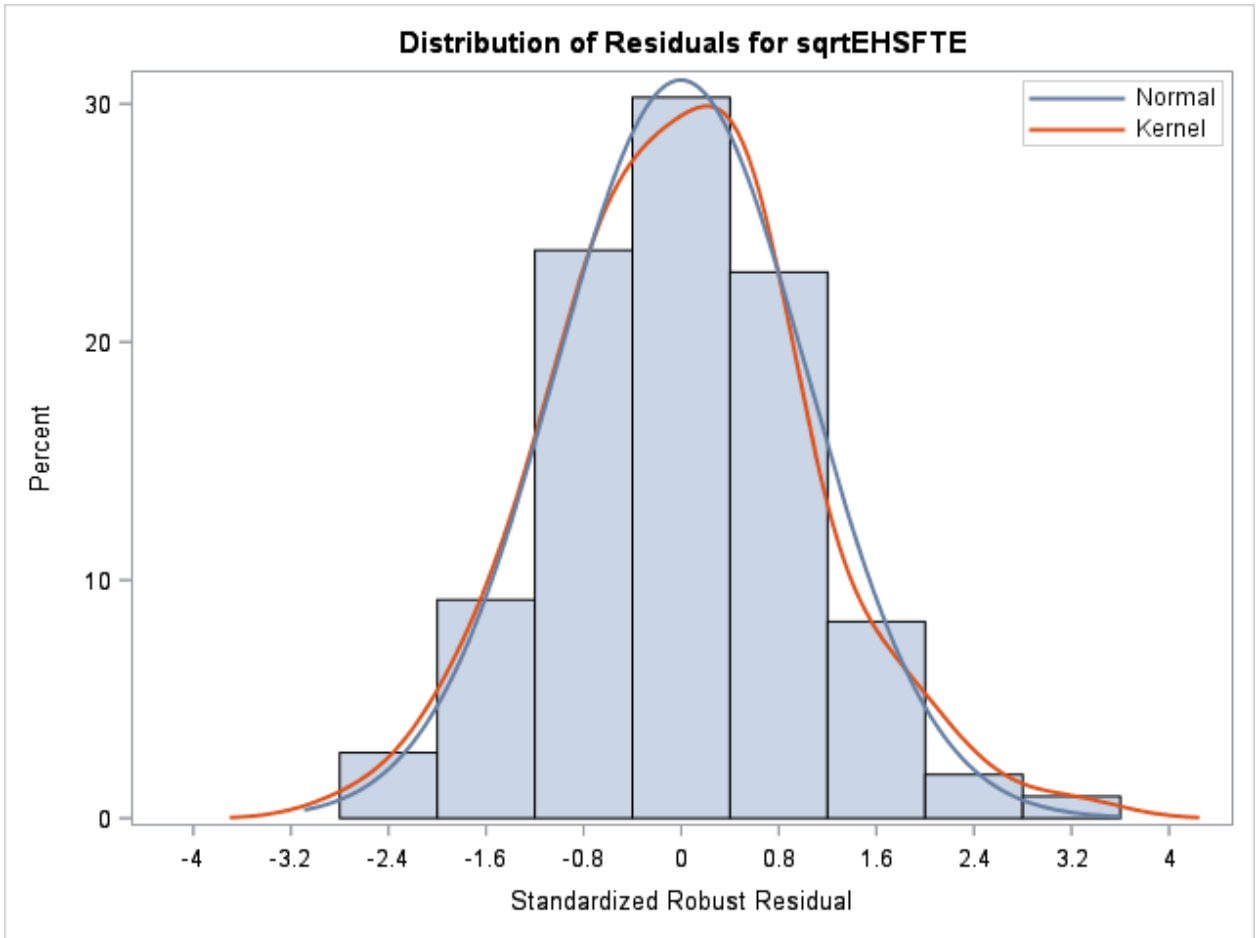


Figure I7. LTS-estimation Histogram of Standardized Robust Residuals for Model D.

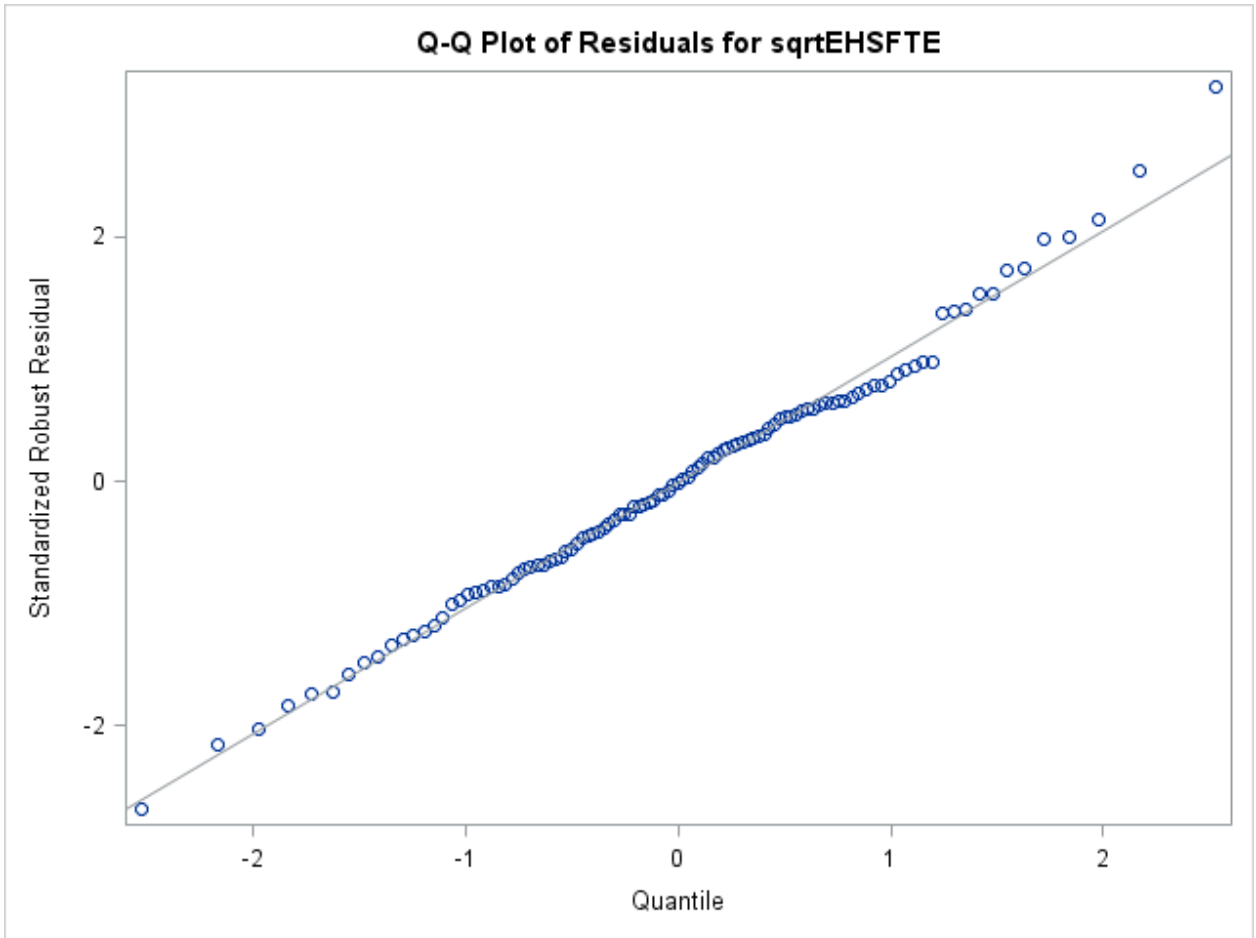


Figure I8. LTS-estimation Q-Q Plot for Standardized Robust Residuals for Model D.

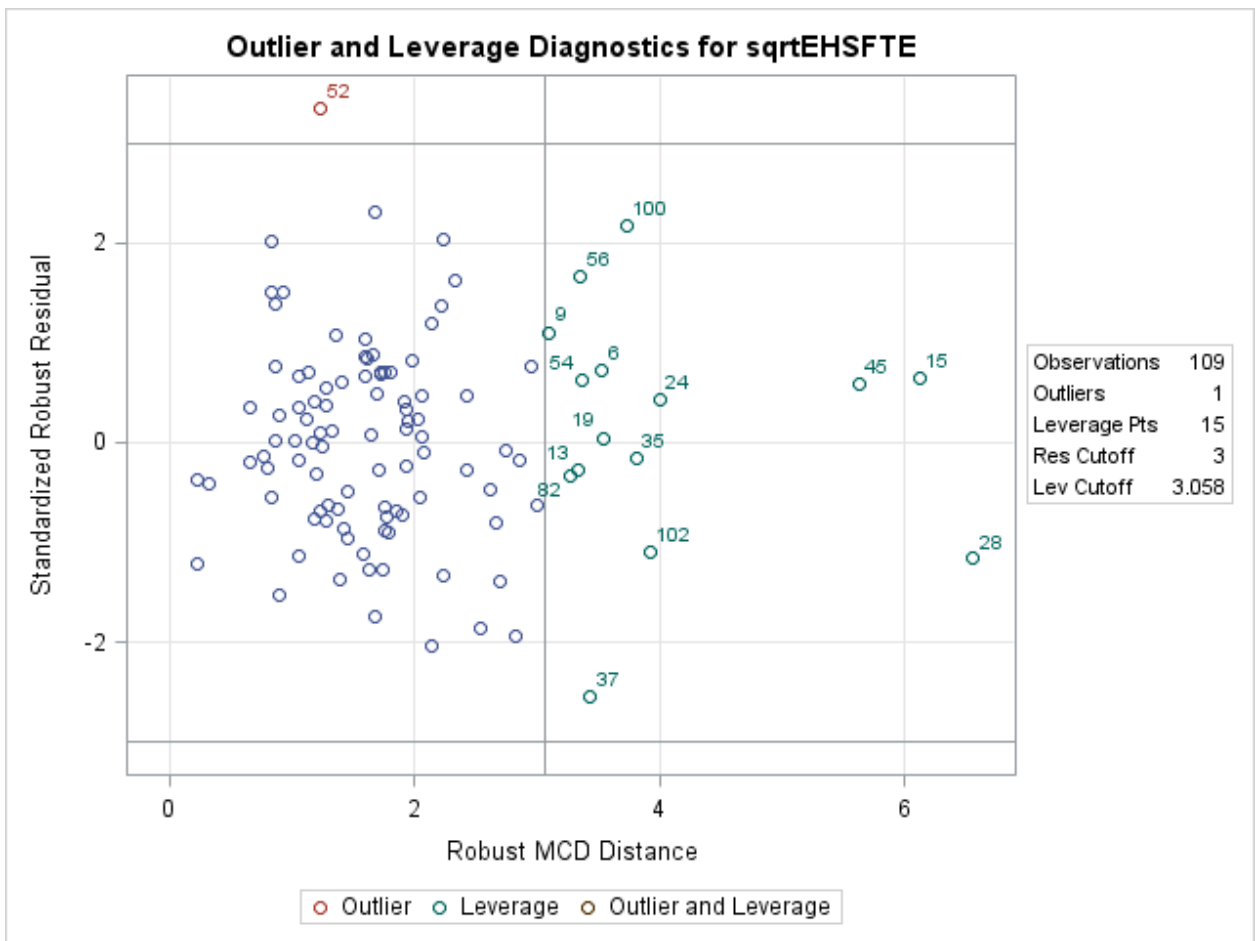


Figure I9. S-estimation RDPLT for Model D.

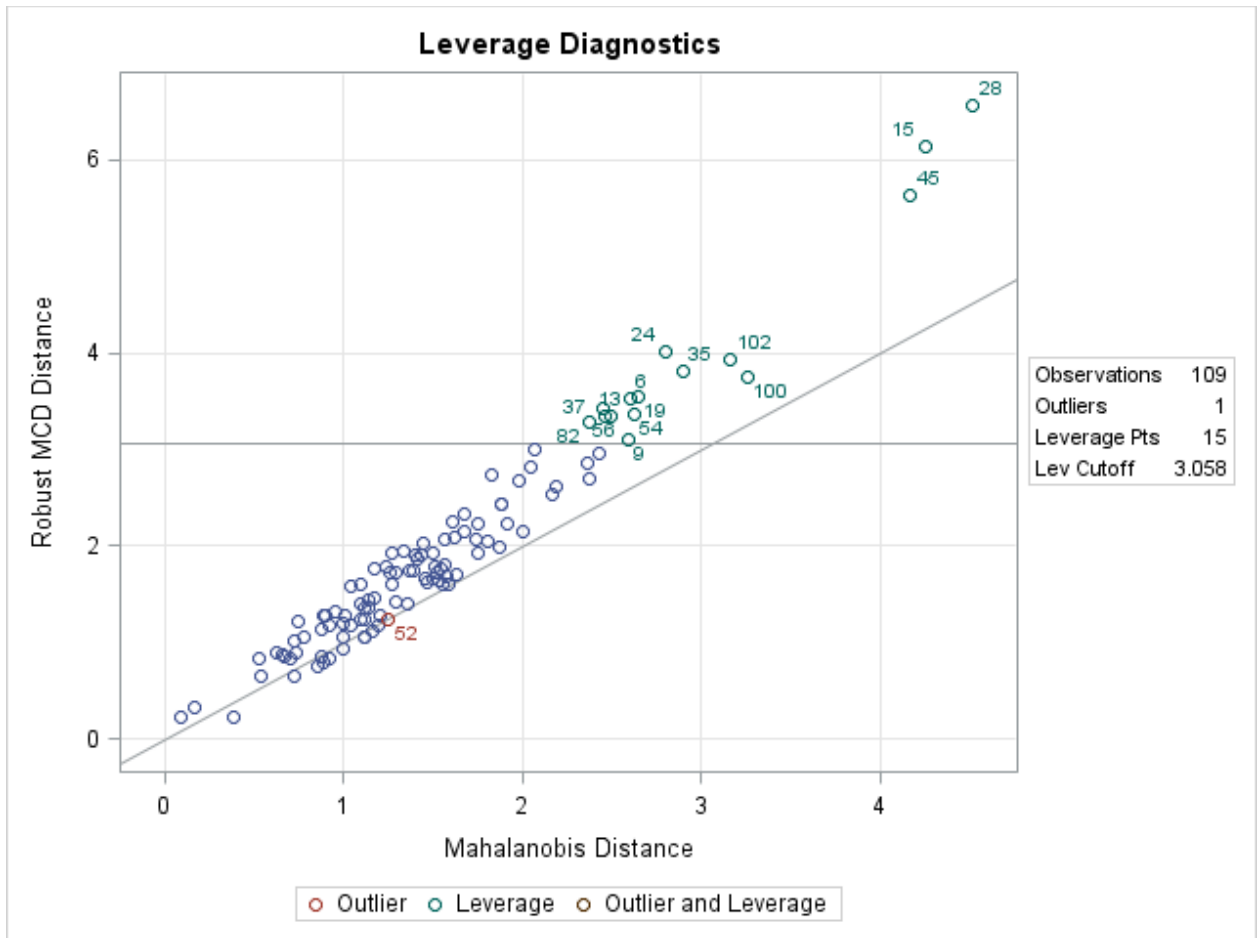


Figure I10. S-estimation DDPlot for Model D.

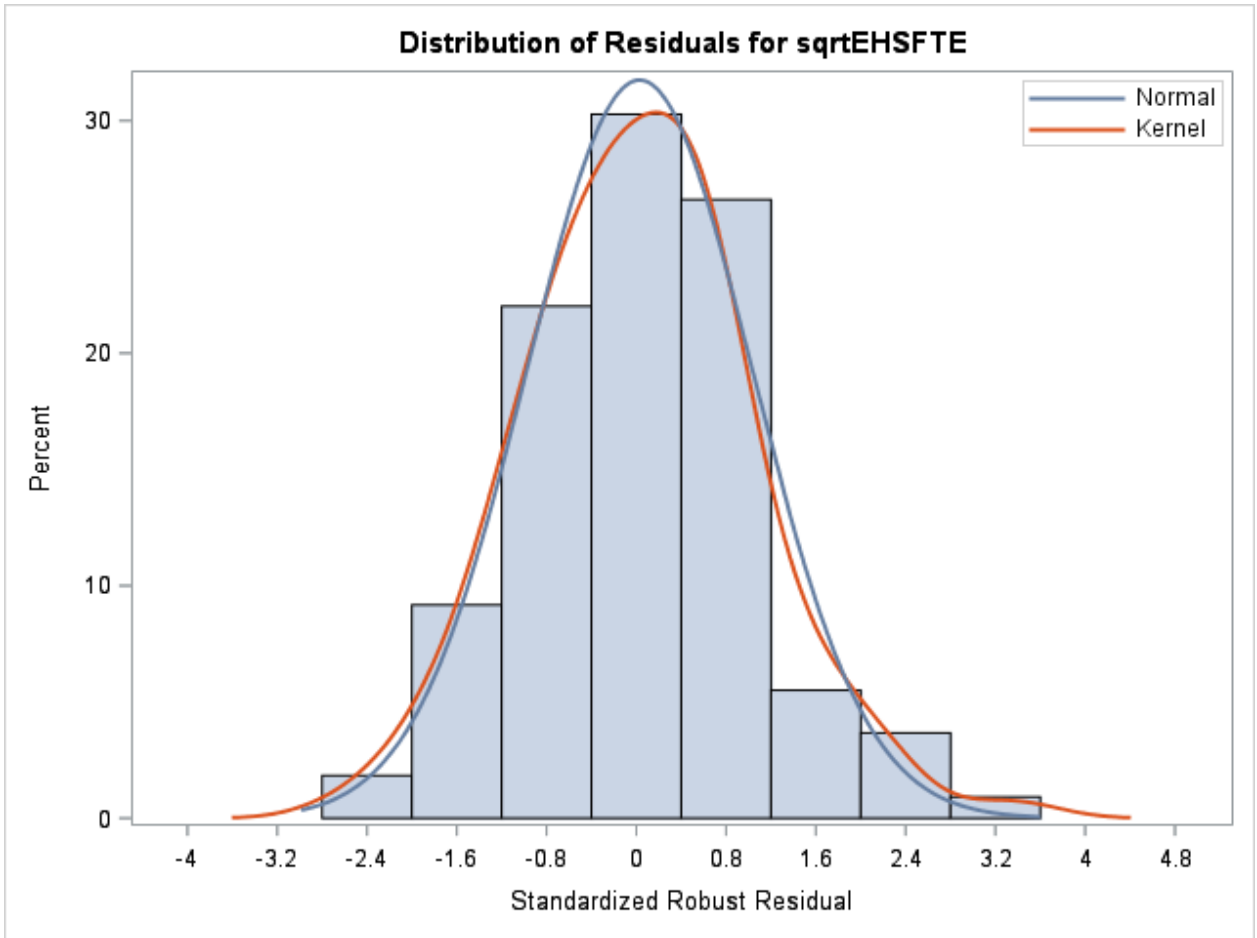


Figure I11. S-estimation Histogram of Standardized Robust Residuals Model D.

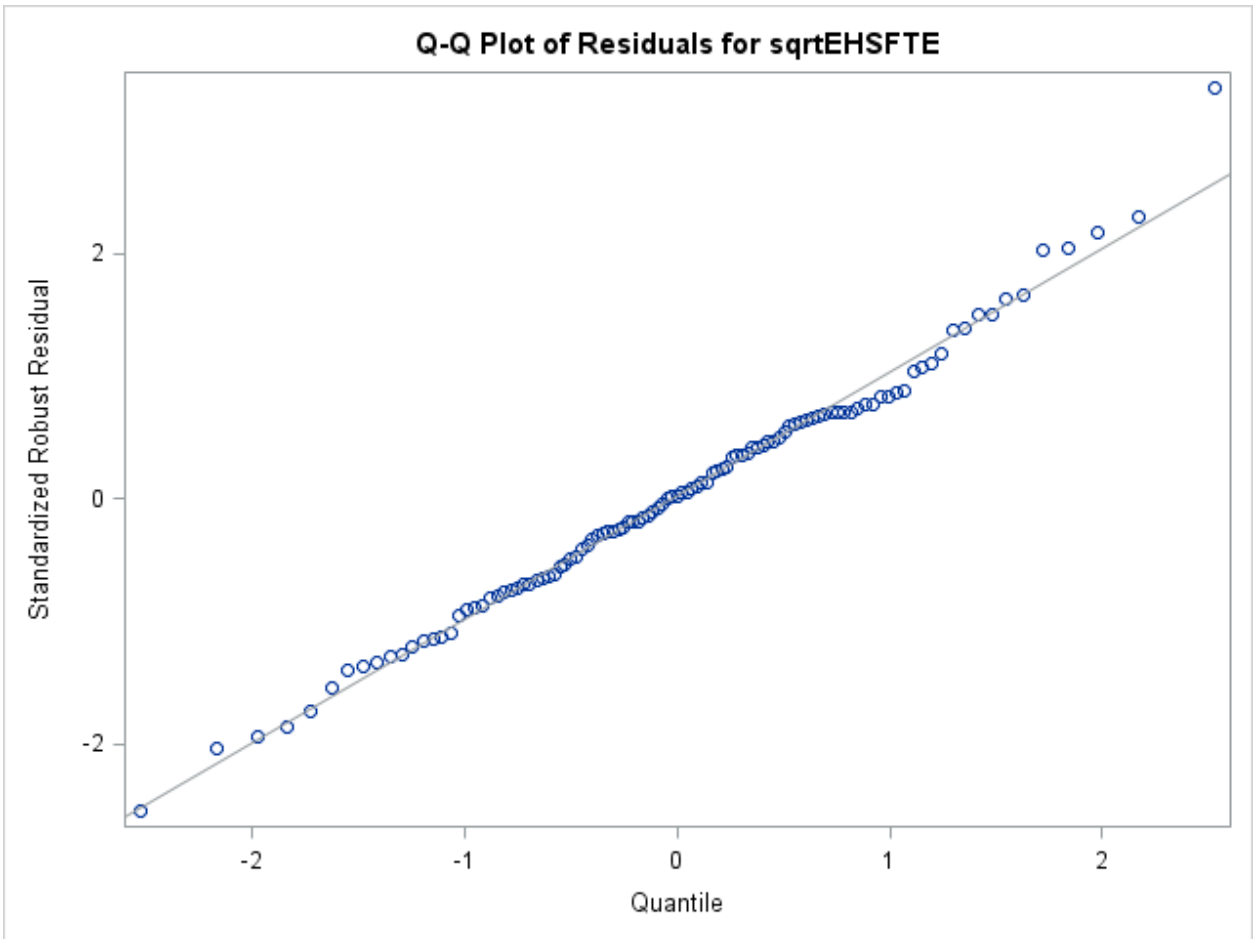


Figure I12. S-estimation Q-Q plot for Standardized Robust Residuals for Model D.

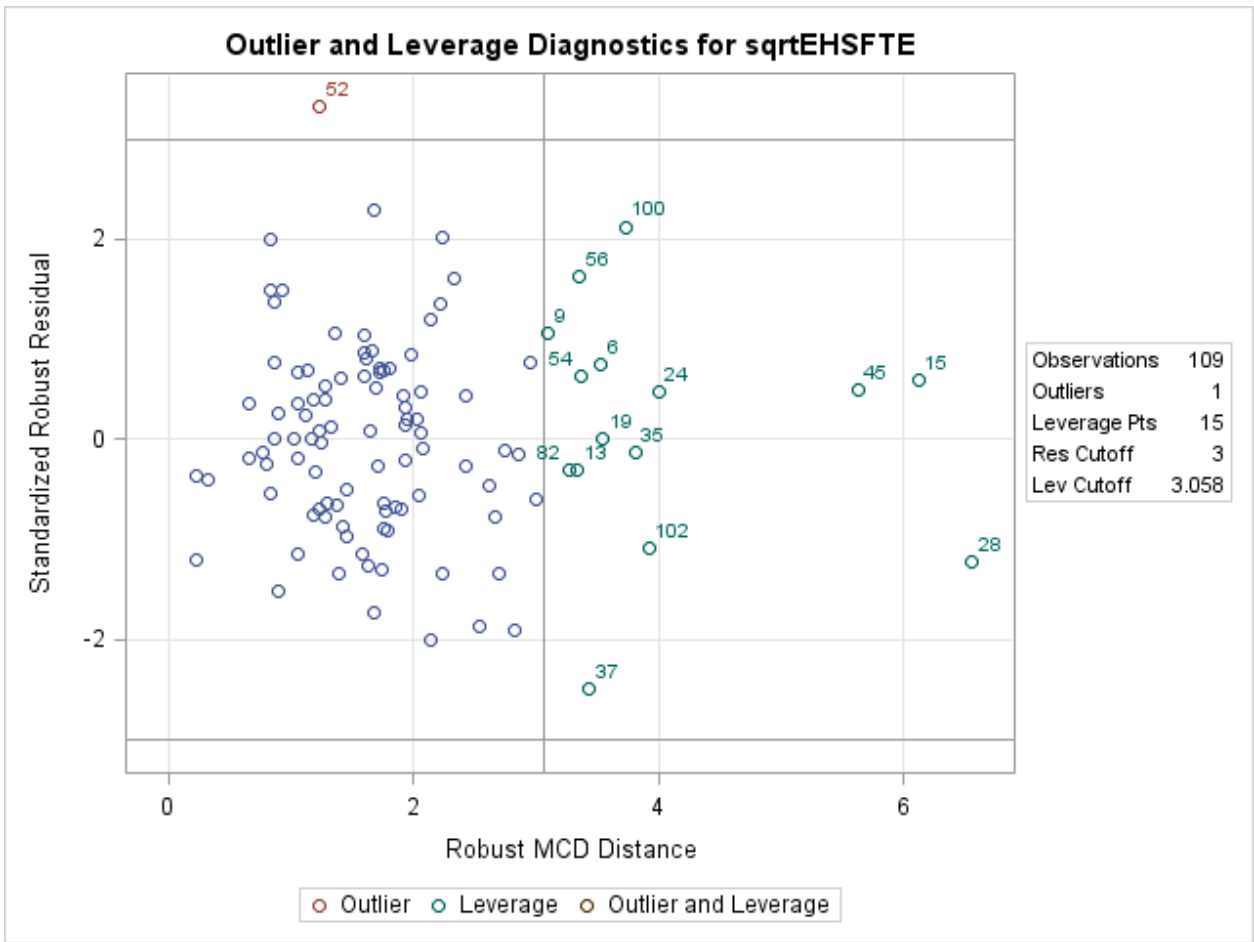


Figure I13. MM-estimation RDPLT for Model D.

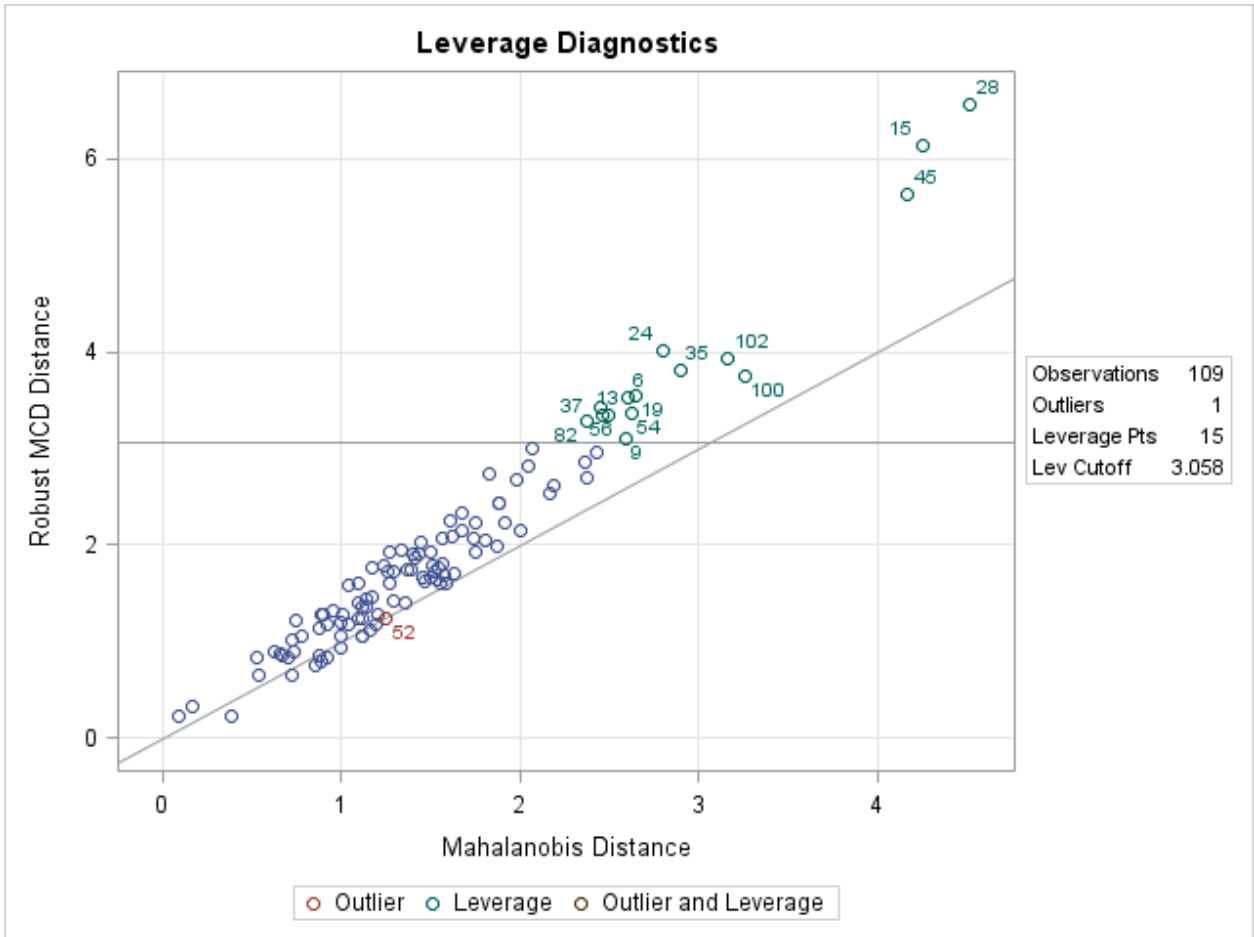


Figure I14. MM-estimation DDPLLOT for Model D.

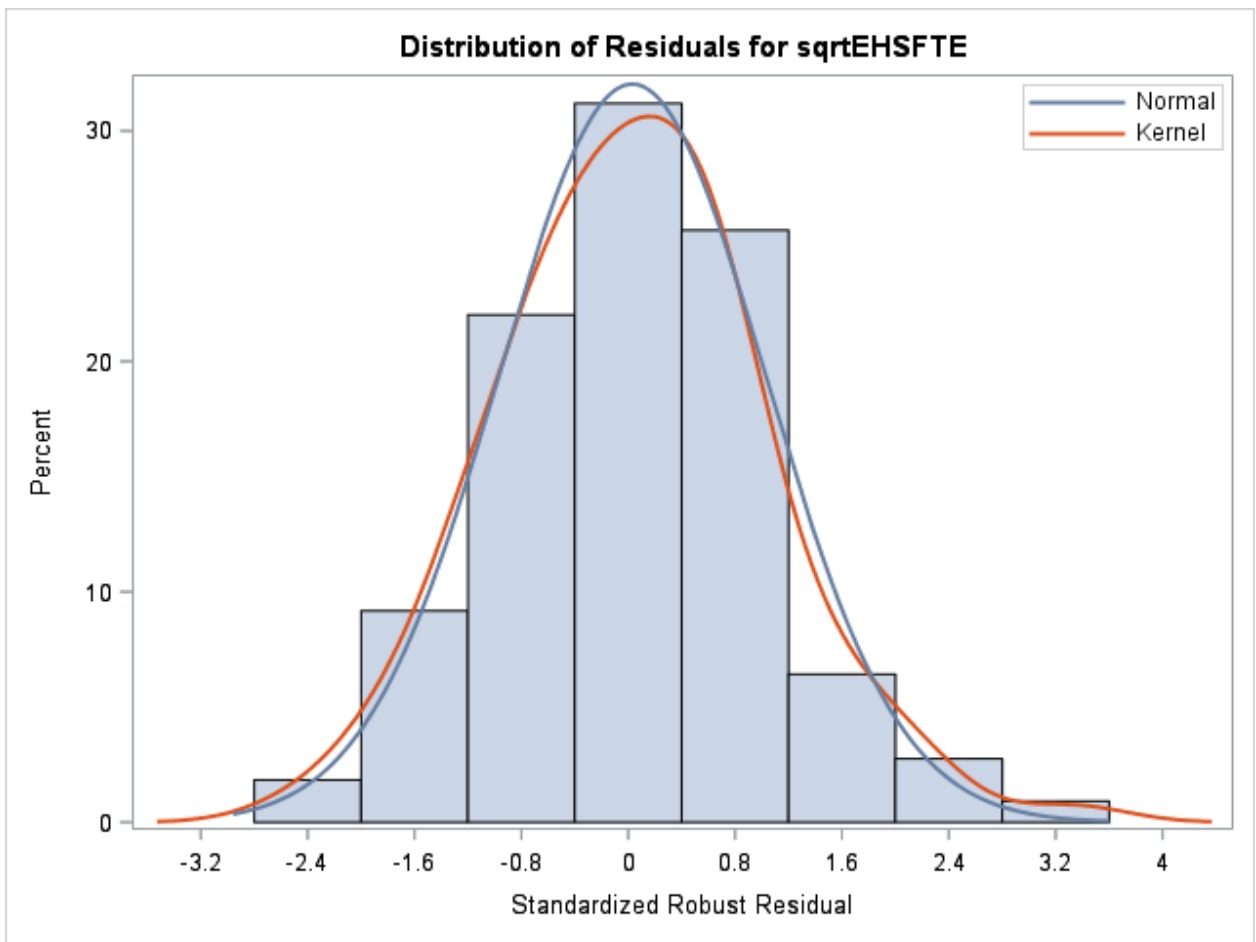


Figure I15. MM-estimation Histogram of Standardized Robust Residuals for Model D.

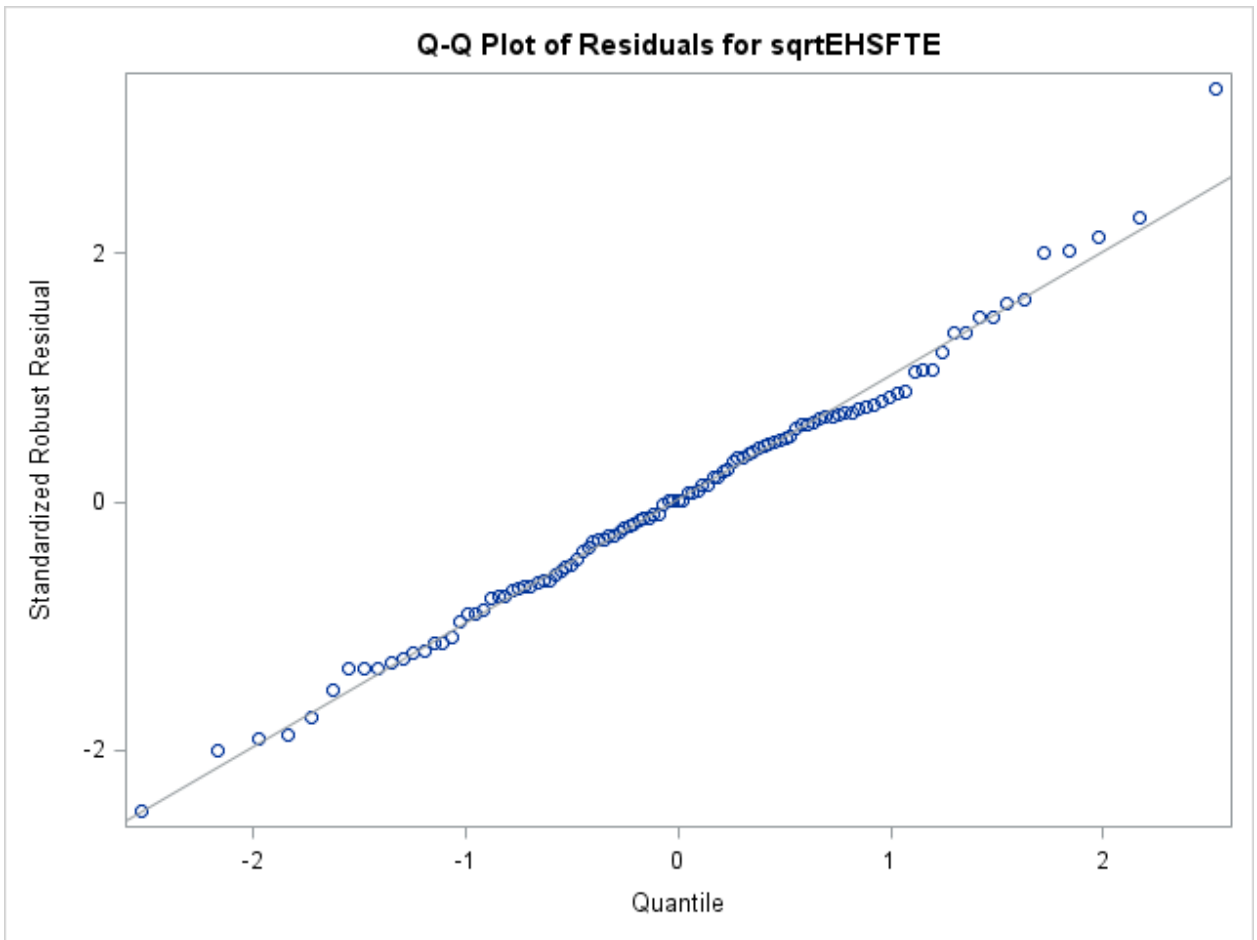


Figure I16. MM-estimation Q-Q Plot for Standardized Robust Residuals for Model D.