2010-01-01

Identifying Influential Observations through the Intraclass Correlation Coefficient

Angel De Jesus Davalos
University of Texas at El Paso, adavalos4@miners.utep.edu

IDENTIFYING INFLUENTIAL OBSERVATIONS THROUGH

THE INTRACLASS CORRELATION COEFFICIENT

ANGEL DE JESUS DÁVALOS

Department of Mathematical Sciences

APPROVED:

Naijun Sha, Chair, Ph.D.

Amy Wagler, Ph.D.

Jorge Ibarra, M.D., M.P.H

Patricia D. Witherspoon, Ph.D.
Dean of the Graduate School

In loving memory of

Tio Rico

and

Mamá Lola

IDENTIFYING INFLUENTIAL OBSERVATIONS THROUGH THE INTRACLASS CORRELATION COEFFICIENT

by

ANGEL DE JESUS DÁVALOS

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2010

Acknowledgements

It is a great honor to thank all those who were instrumental in the completion of my thesis and Master of Science program in Statistics. With my most heartfelt thank you, I recognize the first and most important support system I had: my family. I would not be the person I am today without their love, guidance, and encouragement.

This work would not have been possible without my advisor and committee chair, Dr. Naijun Sha. He showed me a great deal about the research process and I am eternally grateful for everything he has done on my behalf. He invested a great amount of time and effort into ensuring the successful completion and defense of this thesis. It is also a great pleasure to thank the remaining members of my thesis committee: Drs. Amy Wagler and Jorge Ibarra, for their helpful suggestions and time. Additionally, I am forever indebted to Dr. Chungling Liu for having chosen me to participate as an intern during the summer of 2009 with the National Institute of Child Health and Human Development, and Dr. Aiyi Liu for providing me with the thesis topic and being an excellent mentor.

A special thank you goes to the Department of Mathematical Sciences where there were plenty of people willing to offer help and support. In particular, Dr. Joan Staniswalis was a great mentor who played a central role in encouraging me to embark on a Ph.D. program. Additionally, Dr. Ori Rosen was supportive during my course of study and I am highly appreciative of Drs. Emil Schwab and Panagis Moschopoulos for suggesting I study Statistics. Lastly, I would like to recognize my friends and colleagues for their company through it all.

Besides my family, professors, and friends, I would like to thank the LSAMP for the financial support I received through the BD fellowship and the people who administered the award: Dr. Helmut Knaust, Dr. Benjamin Flores, Ariana Arciero, and Sara Rodriguez.

Above all, I would like to give thanks and praise to God for seeing me through all of my endeavors. He provided me strength when I was weak, energy when I was fatigued,

patience when I was distressed, and wisdom when my mind was blank, thank you heavenly Father!

Abstract

In this thesis we analyze the performance of adapting the DFBETA statistic for identifying influential observations on the intraclass correlation coefficient under the assumptions of the one-way random effects model. Additionally, we introduce an approach to transforming negative intraclass correlation coefficient estimation values using the method of moments estimator. We apply this method on a data set of repeated blood pressure measurements, after which we will investigate implications of identifying influential observations.

Table of Contents

Page

Acknowledgements v

Abstract vii

Table of Contents viii

List of Tables x

List of Figures xii

Chapter

1 Introduction 1

1.1 The Intraclass Correlation Coefficient 1

1.2 Motivation and Purpose 2

2 Literature Review 3

2.1 Theoretical Background 3

2.2 Point Estimators 7

2.3 Case-Influence Procedures 10

3 Methodology 13

3.1 Distributions 13

3.2 Negative r Correction 15

3.3 Interval Estimation 22

3.4 Fisher's **Z**-Transformation 22

3.5 Diagnostics 24

4 Simulation Study and Application 26

4.1 Simulation Design 26

4.2 Results 28

4.3 Data Description 35

4.4 Systolic Analysis 36

4.5 Diastolic Analysis 44
4.6 Results 50
5 Conclusions and Discussions 52
Appendix
References 54
A Simulation Study Tables 55
Curriculum Vitae 79

List of Tables

2.1 Analysis of variance for unbalanced one-way random effects model 7

4.1 Layout for a two-period two-treatment crossover design 36
4.2 Nested Mixed Model Information 37
4.3 Nested Mixed Model Significance testing of effects 40
4.4 Analysis of variance for Systolic blood pressures under the unbalanced one-way random effects model 41
4.5 Systolic Pressure ICC estimate and confidence intervals 44
4.6 Nested Mixed Model Significance testing of effects for Diastolic Readings . 47
4.7 Analysis of variance for Diastolic blood pressures under the unbalanced one-way random effects model 47
4.8 Diastolic Pressure ICC estimate and confidence intervals 49
4.9 Raw Data of Subjects identified as influential 50

A.1 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 55
A.2 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 56
A.3 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 57
A.4 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 58
A.5 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 59
A.6 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 60
A.7 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 61
A.8 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 62
A.9 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 63
A.10 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 64
A.11 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 65

A.12 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 66
A.13 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 67
A.14 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 68
A.15 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 69
A.16 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 70
A.17 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 71
A.18 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 72
A.19 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 73
A.20 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 74
A.21 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$ 75
A.22 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$ 76
A.23 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$ 77
A.24 Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$ 78

List of Figures

3.1 PDF Plots given ρ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ 15

3.2 PDF plots given $\rho = 0.1$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ 16

3.3 Histograms of r of size 1000, with $\sigma^2 = 1$, $k = 10$, and n_i=$\{6,8,2,2,2,6,8,6,6,5\}$ 18

3.4 Histograms of r of size 1000, with $\rho = 0.1$, $\sigma^2 = 1$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ 19

3.5 Histograms of bootstrap samples from data with $r = -0.0968$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ 21

3.6 PDF plots of **Z**-Transformation with $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ 23

4.1 Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ . . . 29

4.2 Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ . . . 30

4.3 Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ . . . 31

4.4 Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$ . . . 32

4.5 Percentage $|DFZ_1| \geq 1$ with $\rho = 0.9$ and $n_i \sim DU(2, 10)$ 33

4.6 Percentage $|DFZ_1| \geq 1$ with $\rho = 0$ and $n_i \sim DU(2, 10)$ 34

4.7 Probability Plot of model 4.1 Systolic residuals 38

4.8 Plot of model 4.1 Systolic residuals versus fits 39

4.9 Diagnostic Plots of Systolic reduced model 42

4.10 Diagnostic Variance/Mean influence of Systolic readings 43

4.11 Probability Plot of model 4.1 Diastolic residuals 45

4.12 Plot of model 4.1 Diastolic residuals versus fits 46

4.13 Diagnostic Plots of Diastolic blood pressure reduced model 48

4.14 Diagnostic Variance/Mean influence of Diastolic readings 49

Chapter 1

Introduction

In this chapter, we will introduce some uses of the intraclass correlation coefficient (ICC) and describe the motivation behind the research. The primary purpose of this work is to assess how well an adaptation of the DFBETA statistic works in identifying influential observations through the ICC.

1.1 The Intraclass Correlation Coefficient

First introduced by Fisher in 1925, the intraclass correlation coefficient was presented as a slightly different view in the correlation between measurements. For example, suppose there were measurements (i.e. height, weight, blood pressure, etc.) of k pairs of brothers and we divide the brothers into two classes such as older and younger. If we proceed in this manner, the correlation between these two classes of measurements is termed an **interclass** correlation. Alternatively, we may not know which measurement belongs to the older or younger brother, or, such a distinction may be irrelevant to the purpose of the study; in these cases it is usual to use a common mean and standard deviation from all the measurements. When this is done, the correlation coefficient is an **intraclass** correlation, since we know the brothers belong to the same class (family) [3].

Since then the ICC has been commonly used in many industries for various purposes. For instance, in genetics it plays a central role in estimating the heritability of selected traits in human, animal, and plant populations. In psychology, it plays a fundamental role in reliability theory, where observations may be collected on one or more sets of judges or assessors. In sensitivity analysis, the ICC may be used to measure the effectiveness of

an experimental treatment [1]. In biometry, it is frequently used to measure the degree of intrafamily resemblance with respect to characteristics such as: blood pressure, cholesterol, weight, height, stature, lung capacity, and so forth [9]. Furthermore, in epidemiology it can be used to evaluate the association between measurements of biomarkers, outcomes, and or exposures. The ICC is often used to assess the consistency of repeated measurements of exposures on the same subject(or class) using the same instrument or different ones, either simultaneously or at several time points in reliability studies.

1.2 Motivation and Purpose

In many studies attaining exposure values can be very costly, thus the need to reduce study cost and in effect measurement error is a motivation behind accurately measuring the ICC [6]. Although there is an extensive amount of literature on the intraclass correlation coefficient, most of it is focused on point and interval estimation, and significance testing [1]. Keeping in mind the importance of identifying influential observations, as in regression, we would like to identify observations that are having an effect(improving/reducing consistency) on the estimation of the ICC. Once such an observation is identified, it allows the researcher to investigate either the correctness of the observation (i.e. check for input/instrument error) or identify factors that are improving a subject's consistency. To achieve this, we adapt the case-deletion diagnostic, DFBETA statistic used in regression, to identify influential observations on the ICC estimate.

The rest of the thesis is organized as follows: chapter 2 will lay the necessary theoretical background to describe the methodology, as well as, discuss different estimators used for the ICC and a few case-influence procedures, chapter 3 presents the methods used for assessing influential observations, chapter 4 contains design and results of the simulation study along with an application of the methodology on a data set, and lastly chapter 5 elaborates on findings and discussions for further work.

Chapter 2

Literature Review

In this chapter, the necessary theoretical background will be presented to define the ICC and present different estimators. Additionally, the only case-influence procedure for measuring influence on the ICC will be presented. Lastly, the logic behind our method for identifying influential observations will be laid out through a brief review of the DFBETAS.

2.1 Theoretical Background

The ICC can be defined under very complex models, however, we will focus on the simplest case, the unbalanced one-way random effects model, also known as the variance components model. The assumptions of this model allows us to measure the degree of coherence among repeated measurements from different subjects.

The One-Way Random Effects Model

Let Y_{ij} be the j^{th} measurement from the i^{th} subject with $(i = 1, \ldots, k,\ j = 1, \ldots, n_i)$, such that:

$$Y_{ij} = \mu + a_i + e_{ij}, \tag{2.1}$$

where μ is the grand mean, a_i is the random effect, and e_{ij} is random measurement error. The random effects, $a_i \sim N(0, \sigma_a^2)$, are independent normally distributed with mean 0 and standard deviation σ_a. The measurement errors, $e_{ij} \sim N(0, \sigma_e^2)$, are independent normally distributed with mean 0 and standard deviation σ_e. Lastly, we assume a_i, e_{ij} are

independent. Important features of the model are:

$$\begin{aligned}
&\mathbb{V}ar\{Y_{ij}\} = \sigma_a^2 + \sigma_e^2 = \sigma^2, \\
&\mathbb{C}ov\{Y_{ij}, Y_{i'j'}\} = 0, \ \textit{if } i \neq i', \\
&\mathbb{C}ov\{Y_{ij}, Y_{ij'}\} = \sigma_a^2, \ \textit{if } j \neq j'
\end{aligned}$$

from this we are able to define the intraclass correlation coefficient.

The Intraclass Correlation Coefficient

The correlation between any two measurements under the assumptions of the unbalanced one-way random effects model is,

$$\mathbb{C}orr\{Y_{ij}, Y_{i'j'}\} = \begin{cases} 0, & \textit{if } i \neq i' \\ \dfrac{\mathbb{C}ov\{Y_{ij}, Y_{ij'}\}}{\sqrt{\mathbb{V}ar\{Y_{ij}\}}\sqrt{\mathbb{V}ar\{Y_{ij'}\}}}, & \textit{if } j \neq j' \end{cases}$$

The ICC, ρ, is defined to be the correlation between any two measurements from the same subject.

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \tag{2.2}$$

Thus by definition ($\rho \geq 0$) and the closer it is to 1, the higher coherence there is among measurements from the same subject; perfect consistency occurs when $\rho = 1$ (then $\sigma_e^2 = 0$) [6].

Now that the definition of the ICC has been established, we need to show how it is estimated. Given the assumptions of the one-way random effects model, also known as the variance components model, it is necessary to estimate the variance components. At this point let us define a model that is equivalent to the one-way random effects model based on the multivariate normal distribution that will make deriving estimators easier.

The Common Correlation Model

The common correlation model is equivalent to model (2.1) if,

$$\mathbf{Y}_i \sim N_{n_i}(\mu \mathbf{1}_{n_i}, \Sigma_{n_i}) \tag{2.3}$$

where $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ is the n_i-vector of measures from the i^{th} subject, N_{n_i} is the multivariate normal distribution, μ is the grand mean, $\mathbf{1}_{n_i}$ is the n_i-vector of 1's, $\Sigma_{n_i} = \sigma^2[(1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}]$, where $\sigma_a^2 + \sigma_e^2 = \sigma^2$ is the variance of Y_{ij}, ρ (≥ 0) is the ICC, $\mathbf{I}_{n_i}$ is the $(n_i \times n_i)$ identity matrix, and $\mathbf{J}_{n_i}$ is the $(n_i \times n_i)$ matrix of 1's [4].

As a consequence of model (2.3), then:

$$\mathbf{Y} \sim N_n(\mu\mathbf{1}_n, \Sigma) \tag{2.4}$$

where $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_k')'$, $\Sigma = diag(\Sigma_{n_1}, \Sigma_{n_2}, \ldots, \Sigma_{n_k})$ is the block diagonal matrix composed of Σ_{n_i} matrices along the diagonal, and $n = \sum_{i=1}^k n_i$.

To simplify demonstrating the derivation of the unbiased variance components estimates, we will define a few matrices that will allow us to write the sum of squares in quadratic form. First, consider the variance-covariance matrix of $\mathbf{Y}$, that is, $\Sigma = \sigma_a^2\tilde{\mathbf{J}} + \sigma_e^2\mathbf{I}_n$ where $\tilde{\mathbf{J}} = diag(\mathbf{J}_{n_1}, \mathbf{J}_{n_2}, \ldots, \mathbf{J}_{n_k})$ a block diagonal matrix, and $\mathbf{I}_n$ is $(n \times n)$ identity matrix. Also, let $\mathbf{J}^\star = diag(\frac{1}{n_1}\mathbf{J}_{n_1}, \frac{1}{n_2}\mathbf{J}_{n_2}, \ldots, \frac{1}{n_k}\mathbf{J}_{n_k})$ an $(n \times n)$ block diagonal matrix.

Define the within sum of squares as $SSW = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2$ and the between sum of squares as $SSB = \sum_{i=1}^k n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$, where $\bar{Y}_{i\bullet} = \dfrac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$ and $\bar{Y}_{\bullet\bullet} = \dfrac{\sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}}{\sum_{i=1}^k n_i}$ [6]. The quadratic forms of these sum of squares are: $SSW = \mathbf{Y}'(\mathbf{I} - \mathbf{J}^\star)\mathbf{Y}$ and $SSB = \mathbf{Y}'(\mathbf{J}^\star - \frac{1}{n}\mathbf{J})\mathbf{Y}$. We will use the following quadratic form identity in the derivation of the estimators, $\mathbb{E}[\mathbf{Y}'\mathbf{A}\mathbf{Y}] = tr[\mathbf{A}\mathbb{E}(\mathbf{Y}\mathbf{Y}')] = tr[\mathbf{A}\mathbb{C}ov(\mathbf{Y})] + \mu'\mathbf{A}\mu$.

Estimation of σ_e^2

The expectation of SSW will allow us to find an appropriate unbiased estimate. So consider the following,

$$\begin{aligned}
\mathbb{E}[SSW] &= \mathbb{E}[\mathbf{Y}'(\mathbf{I}-\mathbf{J}^\star)\mathbf{Y}] \\
&= tr[(\mathbf{I}-\mathbf{J}^\star)\mathbb{E}(\mathbf{Y}\mathbf{Y}')] \\
&= tr[(\mathbf{I}-\mathbf{J}^\star)\mathbb{C}ov(\mathbf{Y})] + \mu\mathbf{1}_n'(\mathbf{I}-\mathbf{J}^\star)\mu\mathbf{1}_n \\
&= \sigma_e^2 tr[(\mathbf{I}-\mathbf{J}^\star)] + \sigma_a^2 tr[(\mathbf{I}-\mathbf{J}^\star)\tilde{\mathbf{J}}] \\
&= \sigma_e^2[tr(\mathbf{I}) - tr(\mathbf{J}^\star)] \\
&= \sigma_e^2(n-k)
\end{aligned}$$

So, the method of moments estimate is,

$$\hat{\sigma}_e^2 = \frac{SSW}{n-k} = MSW. \tag{2.5}$$

Estimation of σ_a^2

First the expectation of SSB will be demonstrated,

$$\begin{aligned}
\mathbb{E}[SSB] &= \mathbb{E}[\mathbf{Y}'(\mathbf{J}^\star - \frac{1}{n}\mathbf{J})\mathbf{Y}] \\
&= tr[(\mathbf{J}^\star - \frac{1}{n}\mathbf{J})\mathbb{E}(\mathbf{Y}\mathbf{Y}')] \\
&= tr[(\mathbf{J}^\star - \frac{1}{n}\mathbf{J})\mathbb{C}ov(\mathbf{Y})] + \mu\mathbf{1}_n'(\mathbf{J}^\star - \frac{1}{n}\mathbf{J})\mu\mathbf{1}_n \\
&= \sigma_e^2[tr(\mathbf{J}^\star) - \frac{1}{n}tr(\mathbf{J})] + \sigma_a^2[tr(\tilde{\mathbf{J}}) - \frac{1}{n}tr(\mathbf{J}\tilde{\mathbf{J}})] \\
&= (k-1)\sigma_e^2 + (n - \frac{\sum_{i=1}^k n_i^2}{n})\sigma_a^2 \\
&= (k-1)[\sigma_e^2 + n_0\sigma_a^2]
\end{aligned}$$

Now we define $\frac{SSB}{k-1} = MSB$ and $n_0 = \left(n - \frac{\sum_{i=1}^k n_i^2}{n}\right)/(k-1)$. Thus our method of moments estimate is,

$$\hat{\sigma}_a^2 = \frac{MSB - MSW}{n_0}. \tag{2.6}$$

The following section will present common point estimators used for the ICC in practice.

2.2 Point Estimators

Allan Donner has done an extensive amount of work on the ICC in the latter part of the 20^{th} century. In his 1986 paper, he summarized work done on the ICC up to that point under the assumptions of an underlying random effects model, in particular, he thoroughly explained pros and cons to the analysis of variance, unbiased, pairwise (Pearson's product-moment correlation coefficient), weighted pairwise, and maximum likelihood estimators [1].

The Analysis of Variance/Method of Moments Estimator

Given the unbiased estimators of the variance components, the method of moments (MM) estimator, otherwise known as the analysis of variance (AOV) estimator is defined using $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$,

$$r = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2} = \frac{MSB - MSW}{MSB + (n_0 - 1)MSW} \tag{2.7}$$

$$= \frac{F^\star - 1}{F^\star + (n_0 - 1)} \tag{2.8}$$

where $F^\star = \dfrac{MSB}{MSW}$. The AOV estimator is named as such because $F^\star$ is the statistic used when testing for a random effect. Table 4.4 below is the analysis of variance corresponding to model (2.1). Note, the total sum of squares be $SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$.

Table 2.1: Analysis of variance for unbalanced one-way random effects model

Source of variation	Degrees of freedom	Sum of squares	Mean square	Expected mean square
Subjects	$k-1$	SSB	MSB	$\sigma_e^2 + n_0\sigma_a^2$
Error	$n-k$	SSW	MSW	σ_e^2
Total	$n-1$	SST		

The AOV estimator is frequently adopted because of computational ease and logical construction. This estimator is consistent for ρ but counter to one's intuition it is not unbiased. A serious disadvantage of the estimator is that it may assume negative values when $MSB < MSW$. It is unreasonable to use a negative estimate for a parameter defined to be positive. Often times in practice, negative estimates will be set to 0 or the absolute value is taken [1]. In chapter 3, a method for correcting negative values will be presented. This method is based on using a truncation on the distribution of the ICC.

The Unbiased Estimator

In 1958, Olkin and Pratt derived the minimum variance unbiased estimate of the ICC. There is little reason to choose the unbiased estimate over the MM estimator since it cannot be written in closed form and the degree of bias associated with the MM estimator is very slight [1]. For the derivation of this estimator, see [7].

The Pairwise and Weighted Pairwise Estimator

The pairwise estimator is the oldest measure of intraclass correlation. It is defined as the Pearson product-moment correlation coefficient computed over all possible pairs of measurements within subjects. That is,

$$r_p = \frac{\sum_{i=1}^{k} W \sum_{j=1}^{n_i} \sum_{\substack{l=1 \\ j \neq l}}^{n_j} (Y_{ij} - \hat{\mu})(Y_{il} - \hat{\mu})}{\sum_{i=1}^{k} W(n_i - 1) \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu})^2}$$

where

$$\hat{\mu} = \sum_{i=1}^{k} W(n_i - 1) \sum_{j=1}^{n_i} Y_{ij}$$

and $W = 1/\sum_{i=1}^{k} n_i(n_i - 1)$. The main issues with this estimator are that it can assume negative values and it tends to put too much weight on those subjects (families) with many

measurements.

To overcome the disadvantage of giving high importance to subjects with many measurements, Karlin et al. (1981) proposed a weighted pairwise estimator,

$$r_{pw} = \frac{\sum_{i=1}^{k} W_i \sum_{j=1}^{n_i} \sum_{\substack{l=1 \\ j \neq l}}^{n_j} (Y_{ij} - \hat{\mu})(Y_{il} - \hat{\mu})}{\sum_{i=1}^{k} W_i (n_i - 1) \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu})^2}$$

where

$$\hat{\mu} = \sum_{i=1}^{k} W_i (n_i - 1) \sum_{j=1}^{n_i} Y_{ij}$$

and the weights W_i are constrained to $\sum_{i=1}^{k} n_i(n_i - 1)W_i = 1$. There have been several weighing methods applied to this estimator and it has been observed there is little variation in the results. A good characteristic of r_{pw} is that it does not require a specific model such as equation (2.1). However, the corresponding significance tests are problematic when the group sizes differ greatly.

The Maximum Likelihood Estimator

In 1980, Donner and Koval derived the maximum likelihood estimator, r_M. From model (2.3), r_M is obtained through the likelihood function,

$$L(\mathbf{Y}|\mu, \sigma^2, \rho) = (2\pi)^{-\frac{n}{2}} \prod_{i=1}^{k} |\Sigma_{n_i}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{k} (\mathbf{Y}_i - \mu \mathbf{1}_{n_i})' \Sigma_{n_i}^{-1} (\mathbf{Y}_i - \mu \mathbf{1}_{n_i}) \right\}.$$

The MLE estimator is obtained by finding ρ that minimizes,

$$-2 \ln L = n(1 + \ln \hat{\sigma}^2 + \ln 2\pi) + (n - k) \ln(1 - \rho) + \sum_{i=1}^{k} \ln W_i$$

where

$$\hat{\sigma}^2 = \left\{ \sum_{i=1}^{k} \frac{W_i - \rho}{W_i} \sum_{j=1}^{n_i} (Y_{ij} - \hat{\mu})^2 - \rho \sum_{i=1}^{k} \sum_{\substack{j=1 \\ j \neq l}}^{n_i} \sum_{l=1}^{n_i} \frac{(Y_{ij} - \hat{\mu})(Y_{il} - \hat{\mu})}{W_i} \right\} / \{n(1 - \rho)\},$$

$\hat{\mu} = \left(\sum_{i=1}^{k} n_i Y_{i\bullet}/W_i\right)\left(\sum_{i=1}^{k} n_i/W_i\right)^{-1}$, and $W_i = 1 + (n_i - 1)\rho$. In fact, this estimator is preferred since it will not yield negative values and performs better at extremes (values close to 0 or 1) [2]. The major drawback to this estimator is that there is no explicit expression and therefore a numerical approximation is needed.

2.3 Case-Influence Procedures

In literature, there is only one source focused on case influence on the ICC estimate.

Giraudeau et al. Method

Giraudeau et al. proposed an analytical method for measuring the subject influence on the MLE estimate of the ICC derived under the assumptions of the balanced one-way random effects model, that is ($n_i = p$) [4]. They were interested in quantifying $r_M - r_{M(i)}$ where $r_{M(i)}$ is the MLE estimate with the exclusion of the i^{th} subject, that is excluding $\mathbf{Y}_i$. The authors give the following result,

$$r_M - r_{M(i)} = (1 - r_M)\frac{\frac{k}{k-1}M_i}{k\hat{\sigma}^2 - V_i - \frac{k}{k-1}M_i} - \left(\frac{1}{p-1} + r_M\right)\frac{V_i}{k\hat{\sigma}^2 - V_i - \frac{k}{k-1}M_i},$$

where $M_i = (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$, $\bar{Y}_{i\bullet} = \frac{1}{p}\sum_{j=1}^{p} Y_{ij}$, $V_i = \frac{1}{p}\sum_{j=1}^{p}(Y_{ij} - \bar{Y}_{i\bullet})^2$, and $\hat{\sigma}^2 = \frac{1}{kp}\sum_{i=1}^{k}\sum_{j=1}^{p}(Y_{ij} - \bar{Y}_{\bullet\bullet})^2$.

Giraudeau et al.'s expression shed great light on how a subject's departure from the mean and variance it contributes to the sample. This can be analyzed from the terms, V_i the estimator of the variance of the i^{th} subject's measurements, and M_i the squared difference between the i^{th} subject's mean measures and the global mean estimator. These terms allowed them to make the following conjectures about how a subject's measurements influence the maximum likelihood estimate of ρ:

- if a subject whose measures, on average, differ greatly from the global mean then its exclusion results in a decrease in the estimate of r_M

- if a subject's variance is unusually high then its exclusion will result in an increase of r_M
- a subject's influence decreases as sample size increases
- as r_M approaches one it is difficult to make the estimate higher but it is easier to make the estimate lower
- the lower bound of r_M is $-\frac{1}{p-1}$
- as the number of measures increase the lower bound of r_M approaches 0

While their findings on how a subject influences the MLE estimate of ρ are highly important, their analytical method does not provide critical values for what should be considered a major influence (whether positive or negative) on the estimate. This leads to the question motivating this research, how much departure from the ICC estimate is enough for a subject to be considered influential?

DFBETA Statistic

Within the context of least-squares regression, the DFBETA statistic is a measure of the influence an observation has on a particular regression coefficient, $b_k\,(k = 0, 1, \ldots, p-1)$. The DFBETA statistic is,

$$(DFBETA)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{\mathbb{V}ar(b_{k(i)})}}$$

where $b_{k(i)}$ is the k^{th} estimated regression coefficient with the exclusion of the i^{th} observation and $\mathbb{V}ar(b_{k(i)})$ is the variance of the estimate excluding the i^{th} observation. Additionally, the estimated regression coefficients are assumed to be normally distributed.

The sign of the statistic determines if it's exclusion results in an increase or decrease in the coefficient estimate. For small to moderate sized data, an observation is considered influential if the $|(DFBETA)_{k(i)}| \geq 1$ and $|(DFBETA)_{k(i)}| \geq \frac{2}{\sqrt{n}}$ for large data sets [5].

This criterion for identifying influential observations is of great importance because it is the corner stone of the methodology that will be developed in chapter 3.

Chapter 3

Methodology

In this chapter, we attempt to compensate for some of the disadvantages of the estimator r. The sampling distribution will be derived for a truncated form of r as well as parametric and nonparametric approaches to dealing with negative r estimates, and interval estimation. Fisher's **Z**-transformation plays a major role in the development of our method in identifying influential observations, and therefore will be briefly discussed. Lastly, our case-influence procedure for measuring influence on the ICC estimate using an adaptation of the DFBETA statistic will be presented.

3.1 Distributions

The analysis of variance estimator, r, has several excellent appealing properties, such as ease of construction and computation, and consistency for ρ. A very severe disadvantage is that it may assume negative values. We will compensate for this issue by truncating the sampling distribution to values greater than or equal to zero.

The Cumulative Distribution Function

Recall equation (2.8), from this we construct an F-like statistic and define a transformation for the truncated F distribution. Since,

$$r = \frac{F^\star - 1}{F^\star + n_0 - 1} = \frac{c(F^\star - 1)}{c(F^\star + n_0 - 1)} = \frac{F - c}{F + c(n_0 - 1)}$$

where $c = \frac{1-\rho}{1+\rho(n_0-1)}$. From this we can see, $cF^\star = F \sim f_{u,v}(x)$ is f distributed with degrees of freedom $u = k - 1$ and $v = n - k$. Clearly we can see negative values will occur when

$F < c$. To make positive values of r, a truncated distribution can be derived in the following manner:

$$R = \frac{F - c}{F + c(n_0 - 1)} \tag{3.1}$$

where $F \sim f_{u,v}(x)$, $x \geq c$, $u = k-1$ and $v = n-k$. Furthermore, let $F(x)$ denote the CDF of the f distribution with $k-1$ and $n-k$ degrees of freedom. Then,

$$\begin{aligned} F_R(r) = P(R \leq r) &= P\left(\frac{F - c}{F + c(n_0 - 1)} \leq r\right) = P\left(F \leq \frac{c[1 + r(n_0 - 1)}{1 - r}\right) \\ &= \frac{F\left(\frac{c[1 + r(n_0 - 1)]}{1 - r}\right) - F(c)}{1 - F(c)} \end{aligned}$$

We can see this distribution depends on the parameter value ρ, so the distribution is defined below:

$$F_R(r|\rho) = \frac{F\left(\frac{c[1 + r(n_0 - 1)]}{1 - r}\right) - F(c)}{1 - F(c)}, \tag{3.2}$$

where $0 \leq r < 1$.

The Probability Density Function

By definition, the PDF of R is found by differentiating the CDF with respect to r. Thus the PDF is,

$$f_R(r|\rho) = \left(\frac{cn_0}{(1 - r)^2\,(1 - F(c))}\right) f\left(\frac{c[1 + r(n_0 - 1)]}{1 - r}\right), \tag{3.3}$$

where $0 \leq r < 1$.

Figure 3.1 displays the PDF plots for $\rho = 0.1, 0.5, 0.9$. From this plot, we can see the truncation seems to have no effect on values of ρ close to one. That is, for data that has high coherence it is unlikely to observe an ICC estimate with negative values. However, for data that lacks coherence ($\rho < 0.5$) it is possible to observe negative values for r.

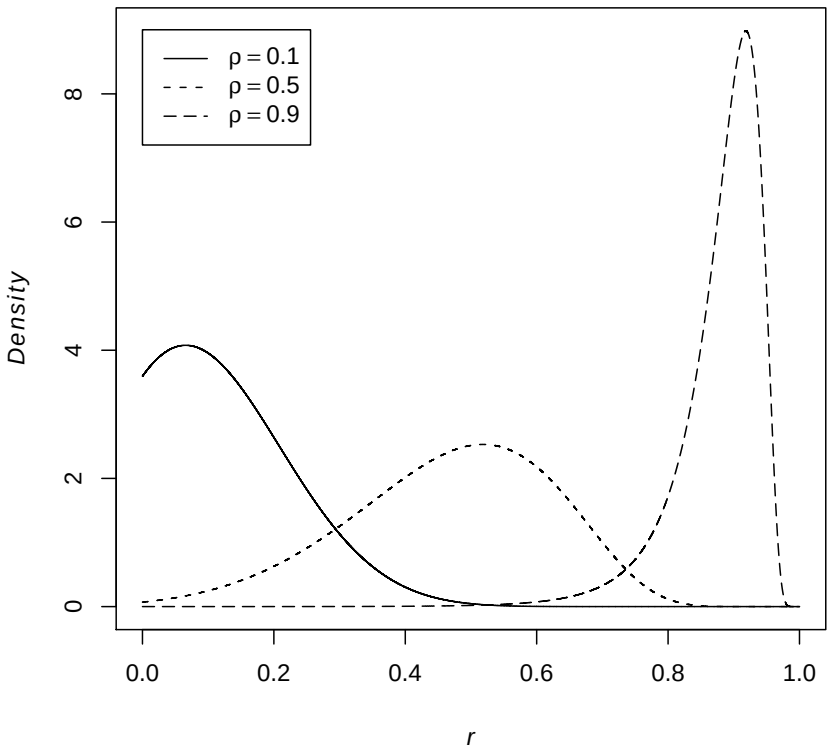

Figure 3.1: PDF Plots given ρ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Now the question remains on how to deal with negative ICC estimates given the truncated sampling distribution. As mentioned in chapter 2, in practice negative ICC values are set to zero or its absolute value is taken [6]. In the following section, the correction of values $r < 0$ based on proportion matching will be discussed. Parametric and nonparametric versions will be presented for when ρ is known or unknown.

3.2 Negative r Correction

The negative r correction procedure is motivated by the truncation. As a consequence of the truncation, observed negative values are proportionally mapped to its corresponding

positive value. To illustrate this we present the sampling distribution of the untruncated form of r. For notation purposes let,

$$\tilde{\rho} = \frac{F - c}{F + c(n_0 - 1)}$$

where F is f distributed with degrees of freedom $k - 1$ and $n - k$.

The sampling distribution of $\tilde{\rho}$ is thus,

$$f_{\tilde{\rho}}(r|\rho) = \frac{cn_0}{(1-r)^2} f\left(\frac{c[1 + r(n_0 - 1)]}{1 - r}\right),$$

where $c = \frac{1-\rho}{1+\rho(n_0-1)}$ and $f(\centerdot)$ is the PDF of the f distribution with degrees of freedom $u = k-1$ and $v = n-k$. Note, the lower bound of the untruncated distribution is observed to be $-\frac{1}{n_0-1}$.

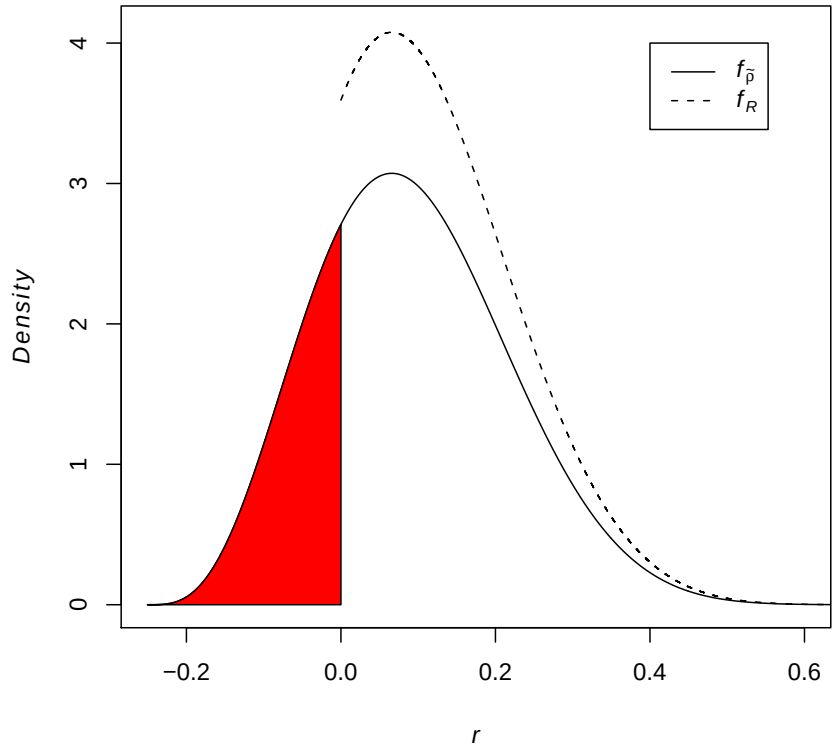

Figure 3.2: PDF plots given $\rho = 0.1$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Parametric Correction

Figure 3.2 displays densities of $\tilde{\rho}$ and r. From this, we can see negative values closer to the lower bound will be observed with less probability. When inspecting the density of r, we can see very positive values are also observed with less probability. Therefore, these two values should be matched and this provides a mechanism for the parametric correction given the value of ρ.

The function g performs the parametric correction given ρ is known. Given an ICC estimate,

$$r^+ = g(r|\rho) = \begin{cases} F_{\hat{\rho}}^{-1}(p_r), & r < 0 \\ r, & r \geq 0 \end{cases}$$

where

$$p_r = 1 - \frac{F_{u,v}\left(\frac{c[1+r(n_0-1)]}{1-r}\right)}{F_{u,v}(c)},$$

and $u = k - 1$ and $n - k$.

The parametric negative r correction procedure allows us to generate appropriate samples of the truncated ICC. Since the common correlation model allows for negative r values, the parametric correction is necessary to transform negative values. Figure 3.2 is a histogram of samples generated using $\rho = 0.9, 0.5$, respectively. In the sample with values generated with $\rho = 0.9$, there were no negative values observed. In the sample with values generated with $\rho = 0.5$, there were 5 negative values observed. The negative values were corrected and in both cases the histograms followed the overlaid density.

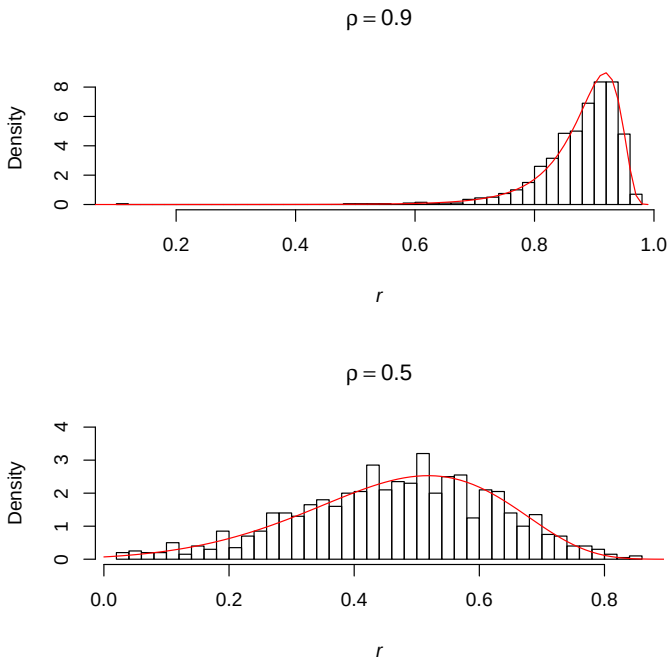

Figure 3.3: Histograms of r of size 1000, with $\sigma^2 = 1$, $k = 10$, and n_i={6,8,2,2,2,6,8,6,6,5}

When a sample is generated with ρ values close to 0, the more incidence of observed negative estimate values. To compare the negative correction procedure with what is done in practice, Figure 3.4 displays histograms of a sample with each with a different negative correction procedure applied. The point mass accumulation procedure corresponds to setting negative values to 0 and the mass accumulation on an interval procedure corresponds to taking the absolute value of negative r values.

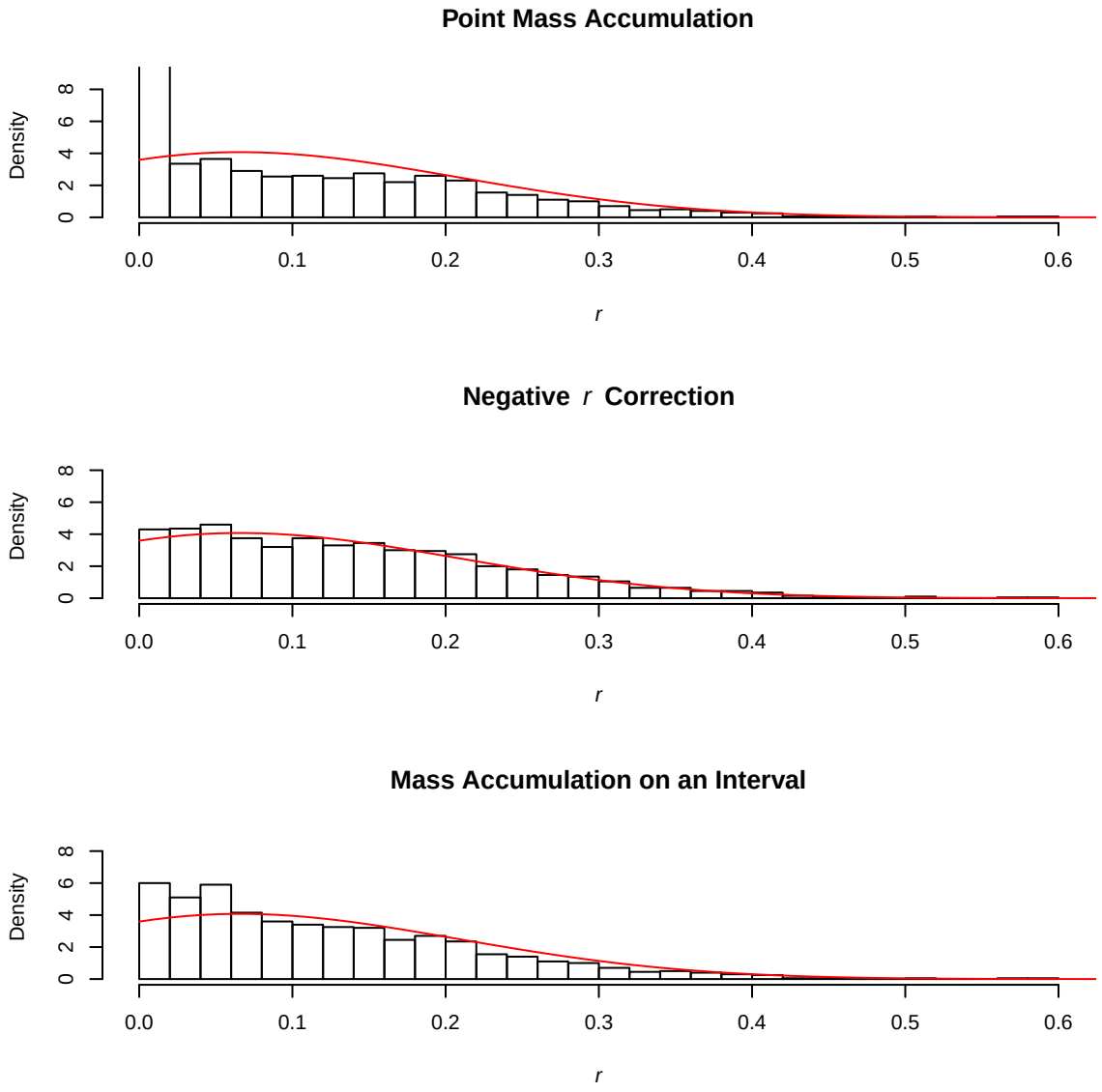

Figure 3.4: Histograms of r of size 1000, with $\rho = 0.1$, $\sigma^2 = 1$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Clearly, the negative r correction procedure is more appropriate for the density of r. To compensate for the estimator yielding negative values, the truncation was employed to address a disadvantage of the estimator. To address the the situation when ρ is unknown, a

bootstrap procedure will be presented in the following section. The nonparametric correction uses the same reasoning as the parametric procedure for correcting negative estimates.

Nonparametric Correction

The nonparametric correction procedure is dependent on our capability to produce bootstrap samples of ICCs from a particular set of data. When a data set is at hand, bootstrap samples are drawn to find the variability in the data set. Once bootstrap samples are drawn we can map negative values to positive values of the sample. The procedure is exactly the same with respect to percentile matching of negative values to positive values. The procedure is as follows:

Given a data set that satisfies the assumptions of the one-way random effects model, $\mathbf{Y}$, if the AOV estimate, $r < 0$,

1. Generate B bootstrap samples of the data to produce , r_m, $m = 1, \ldots, B$.
2. Partition r_m into $r_i^+ = r_m \geq 0$ and $r_j^- = r_m < 0$, where $i = 1, \ldots, a$, $j = 1, \ldots, b$, and $a + b = B$.
3. Find $p = \dfrac{1}{b} \displaystyle\sum_{j=1}^{b} \mathbf{1}(r_j^- \geq r)$.
4. Compute $q = p \displaystyle\sum_{i=1}^{a} \mathbf{1}(r_i^+ \geq 0) = pa$.
5. Given $r_{(q)}^+$ ordered statistics, set $r^+ = r_{(q)}^+$.

where $\mathbf{1}(\cdot)$ the indicator function. Note q is rounded to the nearest integer in the case of non-integer results.

The nonparametric correction procedure allows us to transform all of the negative values in the bootstrap sample to corresponding positive values. Figure 3.5 displays histogram plots of non-corrected and corrected bootstrap samples using data generated from the

common correlation model with $\rho = 0$, $\sigma^2 = 1$, $k = 10$, $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$, and $r = -0.0968$.

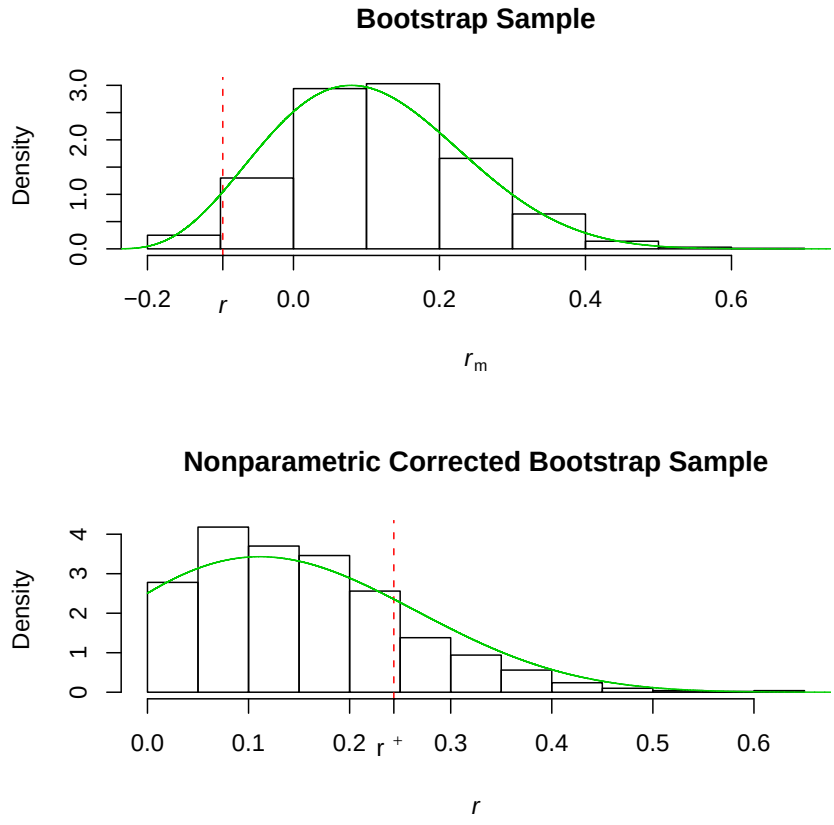

Figure 3.5: Histograms of bootstrap samples from data with $r = -0.0968$, $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

After applying the negative correction procedure on r, we found the value to be $r^+ = 0.2437$. The uncorrected and corrected bootstrap samples take the shape of the untrancated and truncated sampling densities of the ICC, shown in Figure 3.5 respectively, where the overlaid densities are $f_{\tilde{\rho}}(t|\rho = r^B_{0.5})$ and $f_R(t|\rho = r^c_{0.5})$, $r^B_{0.5} = med\{r_1, r_2, \ldots, r_B\}$ is the median of the bootstrap sample, $r^c_{0.5} = med\{(r_1^-)^+, (r_2^-)^+, \ldots, (r_b^-)^+, r_1^+, r_2^+, \ldots, r_a^+\}$ is the median of the corrected bootstrap sample, and $(r_j^-)^+$ is the corrected value of a negative

bootstrap value. As a consequence of this procedure, the bootstrap confidence intervals can be constructed.

3.3 Interval Estimation

The interval estimation for the truncated AOV estimate posed several problems. The first issue that arose was with the inability to eliminate the ρ parameter given the AOV estimator. Secondly, the restricted domain of the estimate does not allow us to achieve the nominal error rate when constructing confidence intervals for ρ. Therefore, we recommend to construct bootstrap confidence intervals for a particular data set.

Procedure

Given a data set that satisfies the assumptions of the one-way random effects model, $\mathbf{Y}$,

1. Generate M bootstrap samples of the AOV ICC estimate, r_m.
2. Using the nonparametric correction procedure, correct all negative values.
3. Using the corrected bootstrap sample, construct confidence intervals (i.e. percentile, t).

3.4 Fisher's Z-Transformation

In order for our proposed method of adapting the least-squares regression DFBETA statistic to work on the intraclass correlation coefficient, it is crucial for the coefficient to assume normality. From the derivation of the ICC distribution, we can see it a ratio of linear combinations of an F distributed statistic. In 1925, R.A. Fisher made a great contribution to statistics by deriving a transformation on the pairwise correlation coefficient and the intraclass correlation coefficient that asymptotically achieves normality.

Fisher's transformation is,

$$z_r = \frac{1}{2}\log\left(\frac{1+r(n_0-1)}{1-r}\right) \sim N(z, V_{z_r}), \tag{3.4}$$

where $z = \frac{1}{2}\log\left(\frac{1+\rho(n_0-1)}{1-\rho}\right)$ and $V_{z_r} = \frac{1}{2}\left\{\frac{1}{k-1}+\frac{1}{n-k}\right\}$ [3]. To illustrate the effectiveness of the transformation, Figure 3.6 displays pdf plots of the transformation for different ρ values.

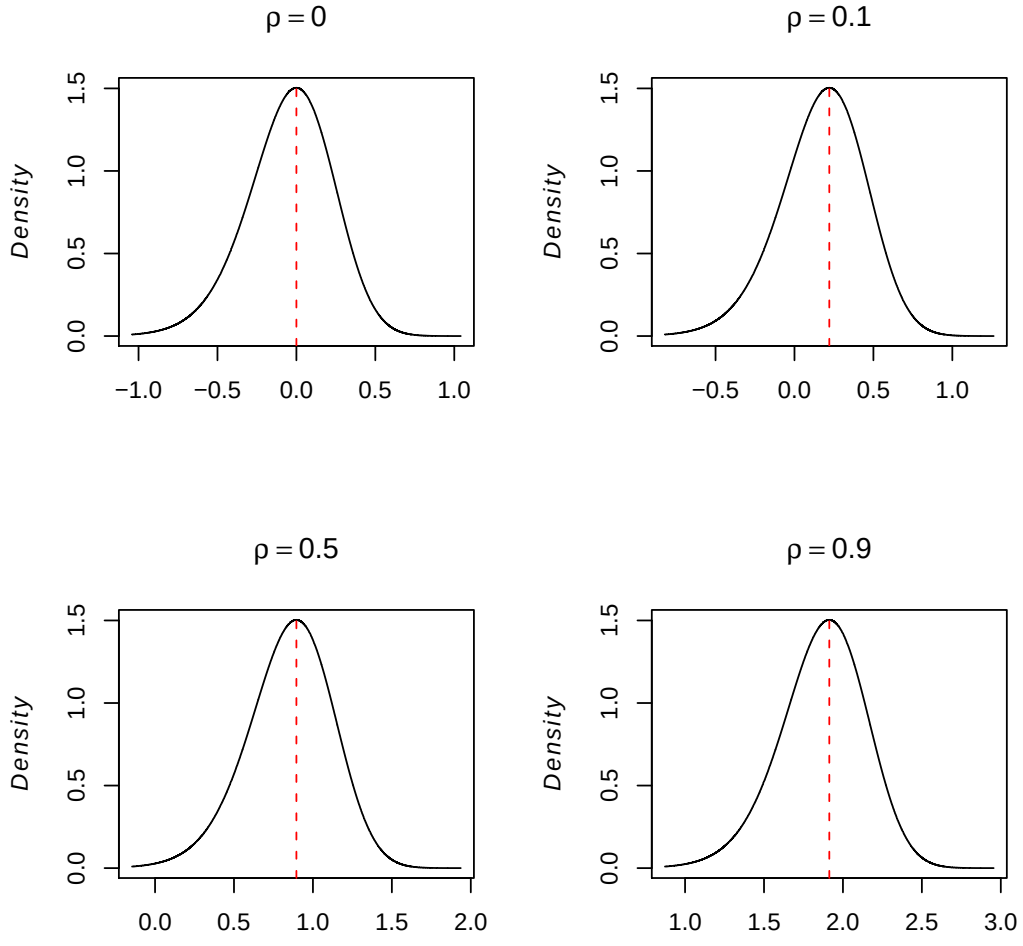

Figure 3.6: PDF plots of **Z**-Transformation with $k = 10$, and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Even with a relatively small sample size at $k = 10$, Fisher's transformation performs well in having the sampling distribution of the ICC achieve an approximate bell shaped curve of the normal distribution. When sample sizes increase to greater than 30, normality can be assumed with greater confidence. The following section will show the adaptation of the DFBETA statistic on the ICC through the use of Fisher's **Z**-transformation.

3.5 Diagnostics

In this section we present the statistic that we used to identify influential observations, the DFZ statistic. Additionally, we present a method for identifying how an observation influences the ICC estimate. This method was motivated by the work presented in Giraudeau et al.'s paper [4].

Case-Influence DFZ

In least-squares regression, case influence is quantified by subtracting the estimate of a parameter with an estimate of the same parameter excluding an observation. For example, Cook's Distance is used to measure aggregate influence while DFBETA is used to measure influence of the i^{th} observation on each regression coefficient. Since we are seeking to measure the influence of the i^{th} subject, adopting DFBETA on the ICC is adequate. In our case deletion procedure, estimating the ICC with the exclusion of the i^{th} subject means excluding a subject's entire repeated measures from estimation. We define,

$$DFZ_i = \frac{z_{r(i)} - z_r}{\sqrt{\mathbb{V}ar_{z_{r(i)}}}}, \tag{3.5}$$

where $z_{r(i)}$ denotes a value applying Fisher's transformation on r with the exclusion of the i^{th} subject. Here as in least-squares regression, a subject is considered influential if DFZ_i in absolute value exceeds 1 for small to moderately sized data sets or $2/\sqrt{k}$ for large data sets.

Mean/Variance influence on r

If a subject is identified as influential, we would like to determine whether it was due to departure from the overall mean or whether it was due to the variance contributed by the i^{th} subjects measurements. This strategy is a direct consequence of the work done by Giraudeau [4]. Their paper decomposes $r - r_{(i)}$ into two antagonistic terms in which one is termed influence due to the departure from overall mean and the other termed influence due to the variance contributed by the i^{th} subject.

We attempt to reproduce similar terms using the AOV ICC estimate:

$$r - r_{(i)} = V_{(i)} - M_{(i)},$$

where $V_{(i)} = \left(\frac{1}{(n_{0(i)}-1)} + r\right) \frac{(n_{0(i)}-1)MSW_{(i)}}{MSB_{(i)}+(n_{0(i)}-1)MSW_{(i)}}$ and $M_{(i)} = (1-r)\frac{MSB_{(i)}}{MSB_{(i)}+(n_{0(i)}-1)MSW_{(i)}}$. Since our antagonistic terms are dependent on estimating the MSB and MSW with the exclusion of the i^{th} subject, our interpretation is not as straightforward as those derived by Giraudeau et al. Assuming there is no influence or change is very slight when excluding the i^{th} subject, we expect the two terms to be equal or close to equal. When larger positive values are observed, we may conclude the i^{th} subject's measurements on average are significantly different from the overall mean; while large negative values observed, we may conclude the i^{th} subject is contributing a significant amount of variation to the rest of the data.

When we apply our diagnostics, we will be able to identify influential observations and pinpoint what type of effect it is having on the estimation. We suspect that those measurements that contribute to an increase in variance of the total data set (i.e. $r - r_{(i)} < 0$) distort the coherence of the repeated measurements in the data. Alternatively, those subjects whose measurements on average are far from the overall mean (i.e. $r - r_{(i)} > 0$) improve coherence of repeated measurements in a data set. To test the validity of these hypotheses, we design a simulation study in which our goal is shed light on the effect a particular subject's variance has on the ICC. The design and results of such simulation study are presented in the following chapter.

Chapter 4

Simulation Study and Application

The purpose of this chapter is to study some of the properties of the DFZ statistic via a simulation study. Namely, were interested in studying the effect of altering the variance components in the percentage of times the statistic identifies an observation as influential. Secondly, a numerical example will be considered in which our diagnostics are applied to a data set of repeated blood pressure measurements.

4.1 Simulation Design

In order to study the properties of the DFZ statistic, a simulation study was designed and performed, since it is difficult to obtain results using analytic techniques. Recall, this method is applicable to the ICC under the assumptions of the unbalanced one-way random effects model. Our main interest is to gauge the effect of altering the variance components through data generated from the multivariate normal distribution using the common correlation model on the percentage of times an inconsistent observation is observed to be influential. In this section we will refer to an inconsistent observations as one that is generated using different ICC and total variance parameters, ρ and σ^2.

The main difficulty in the design of the study was controlling the infinitely many values n_i may assume. As a consequence of this, a secondary focus of the simulation study was to observe what happens under different scenarios with respect to sample size. In particular, we were interested in observing improvements in identifying the inconsistent observation as influential when sample size is increased. Thus, the data was generated using sample sizes $k = 10, 20, 50$ subjects each with n_i repeated measurements sampled from discrete

uniform distributions $DU(2, 10)$ and $DU(10, 20)$. The algorithm of the simulation study was as follows:

1. Generate 1000 data sets from the common correlation model.

$$\mathbf{Y}_m \sim N_n(\mathbf{0}_n, \Sigma),$$

   where $m = 1, \ldots, 1000$, $\mathbf{Y}_m = (\mathbf{Y}_1', \mathbf{Y}_2', \ldots, \mathbf{Y}_k')'$, $\Sigma = diag(\Sigma_{n_1}, \Sigma_{n_2}, \ldots, \Sigma_{n_k})$ is the block diagonal matrix composed of $\Sigma_{n_i} = \sigma^2[\rho J_{n_i} + (1-\rho) I_{n_i}]$ matrices along the diagonal for $i = 2, \ldots, k$ and $\Sigma_{n_1} = \sigma_1^2[\rho_1 J_{n_1} + (1 - \rho_1) I n_1]$, and $n = \sum_{i=1}^{k} n_i$.

2. Compute r_m and $r_{(1)m}$ from each $\mathbf{Y}_m$.

3. Compute DFZ_{1m}.

4. Compute the percentage of times the first observation was identified as influential:

$$P_{DFZ_1} = \frac{1}{1000} \sum_{m=1}^{1000} \mathbf{1}(|DFZ_{1m}| \geq 1)$$

The algorithm above represents computing the percentage of times the inconsistent observation (i.e. observation generated using ρ_1, σ_1^2) was observed influential. Note, for each combination of k and n_i specified above, the algorithm was applied using all combinations of parameter values such that $\sigma^2 = 1$, $\sigma_1^2 = 1, 3, 5, 10$, and $\rho, \rho_1 = 0, 0.1, \ldots, 0.9$.

When designing the simulation study, adjusting the ICC and variance was appealing over adjusting the variance components because it allows for better control and interpretation of the type of data we were generating. Recall, the definition of the ICC

$$\rho = \frac{\sigma_a^2}{\sigma^2} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

From this, we can see $\sigma_a^2 = \rho\sigma^2$ and $\sigma_e^2 = (1 - \rho)\sigma^2$, so as we adjust the ICC and variance, we are in effect adjusting the variance components. The results are tabulate in Appendix A.

4.2 Results

Before performing the simulation study, we suspected the inconsistent observation would be identified as influential at a higher rate when it is generated from a large variance population and when it is generated with a different population ρ. Below are plots of the general trends observed in the simulation study. All of the results per data set were similar and so we will highlight the most obvious findings through a series of plots from the tables in Appendix A. The trends presented below are for small data sets, however, the same trend was observed for large data with an increased percentage of inconsistent observation identified influential. When the number of subjects and repeated measurements are increased, percentage of inconsistent observation identified as influential is increased.

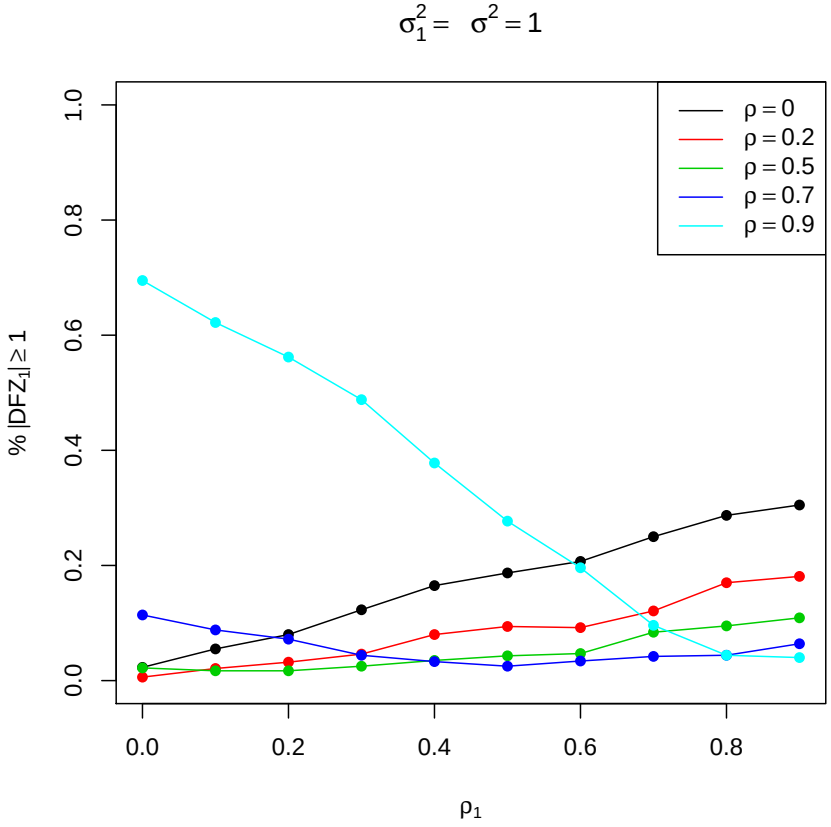

Figure 4.1: Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

We will first begin with the case where data has a small amount of subjects and repeated measurements ($k = 10$, $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$) and the inconsistent observation is generated with a population variance equivalent to the rest of the data. Each line in Figure 4.1 represents data with a different structure. For example, all of the variance of the data with $\rho = 0$ is due to random measurement error, that is $\sigma_e^2 = 1$, lack of coherence in the repeated measurements. Likewise, data with $\rho = 0.9$ has $\sigma_a^2 = 0.9$ and $\sigma_e^2 = 0.1$ which means most of the variability is due to the random effect. When data has a large random effect with respect to measurement error, there are subjects whose mean measurements deviate from the overall mean. The only two notable observations from Figure 4.1 are:

when considering a data set with high coherence ($\rho = 0.9$), inconsistent observations with large intrasubject variability will be increasingly identified; when considering a data set with lack of consistency and no random effect, inconsistent observations who deviate from the overall mean will be identified as influential at a small rate.

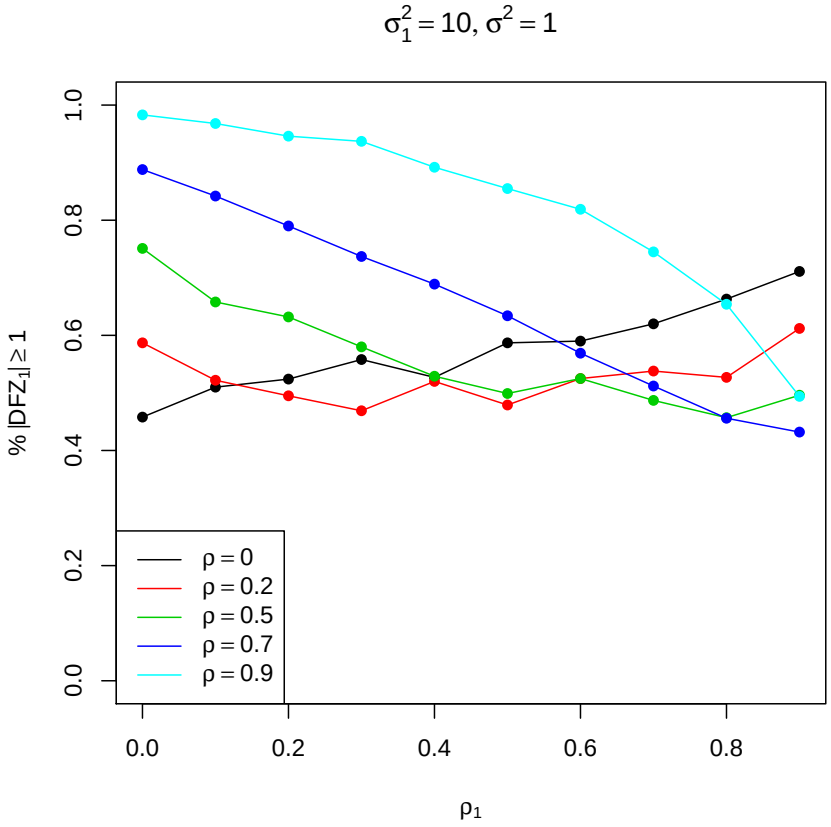

Figure 4.2: Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Figure 4.2 is an example of the effect an inconsistent observation has when it contributes a large amount of variation. In this case, it did not matter whether the inconsistent observation contributed to the variance by deviating from the overall mean or its intrasubject variation the result was an increased percentage of identification.

The following two plots were produced to observe the effect of increasing the variance

parameter of an inconsistent observation among data sets with a high and low degree of coherence.

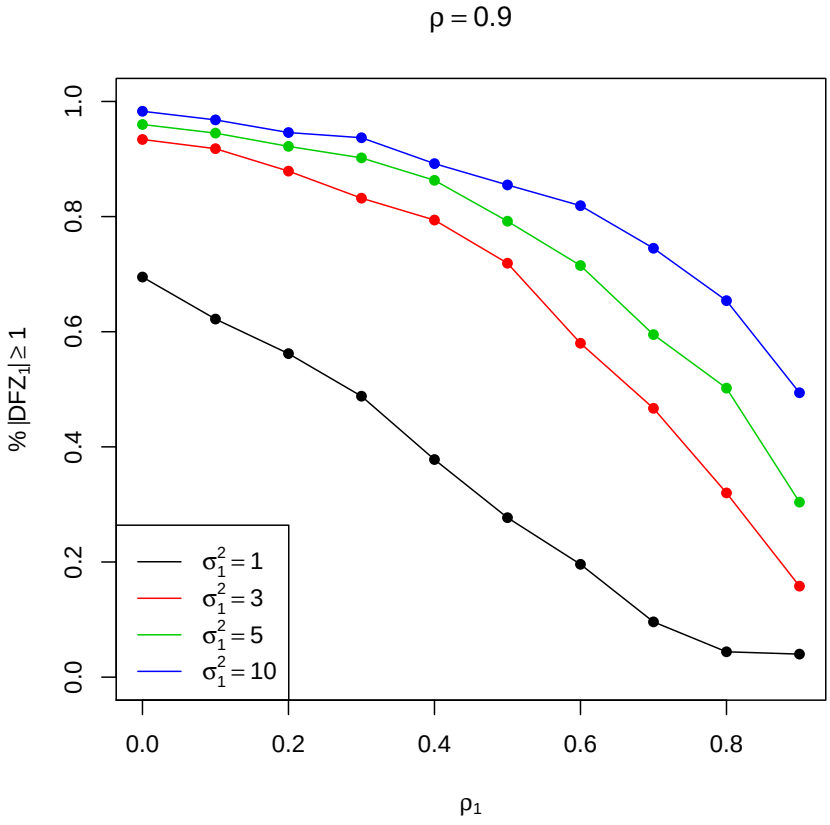

Figure 4.3: Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Figure 4.3 shows that as the inconsistent observation contributes more variation to a data set with high coherence, the more likely it is for our method to identify it as influential.

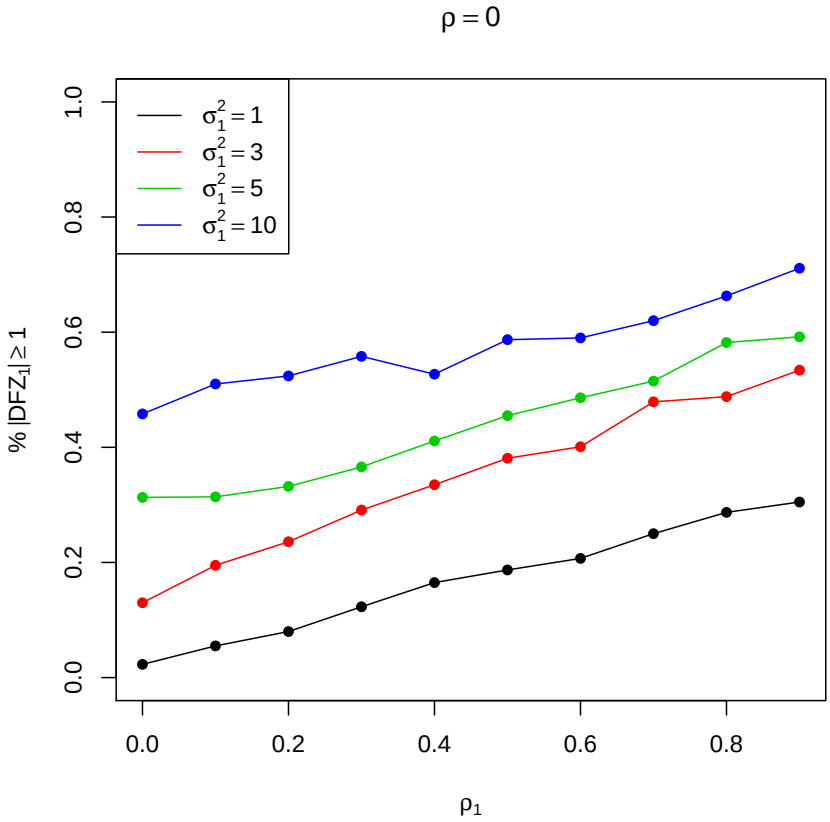

Figure 4.4: Percentage $|DFZ_1| \geq 1$ with $k = 10$ and $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$

Figure 4.4 also corroborates the observation that as an inconsistent observation contributes variance to a data set with lack of consistency in its repeated measurements, the more likely it is for it to be identified as influential.

The aim of the last two plots is to compare the effect of increasing the number of subjects in a data set. This is of interest since our method is dependent upon the asymptotic properties of Fisher's **Z**-transformation. To portray the maximum percentage of time an inconsistent observation will be identified as influential, Figures 4.5 and 4.6 display worst case scenarios; that is, inconsistent observations that have a large amount of variation within a data set with a small amount of repeated measurements.

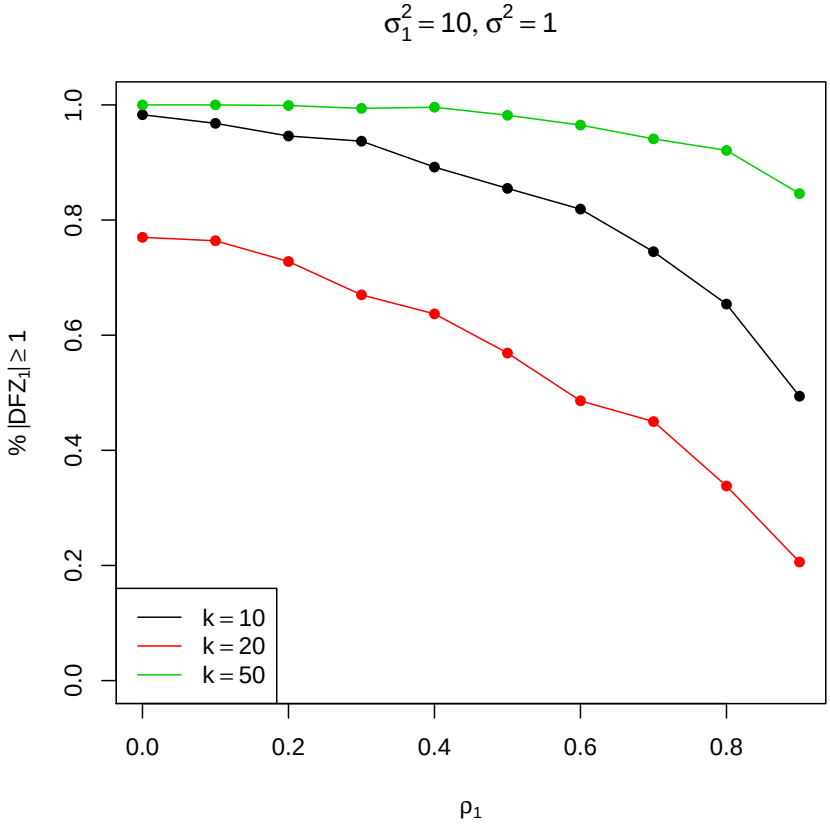

Figure 4.5: Percentage $|DFZ_1| \geq 1$ with $\rho = 0.9$ and $n_i \sim DU(2, 10)$

In Figure 4.5, we can clearly see large sample size is preferred because it almost ensures the inconsistent observation will be identified as influential. One particularly interesting observation to make from Figure 4.5 is, contrary to what is expected, the percentage of observations does not increase as the number of subjects increase. We would have expected the percentage $|DFZ_i| \geq 1$ of the data set with 20 subjects to be uniformly greater than the data set with 10 subjects. This is a major indicator that there is another component that is influencing the DFZ statistic.

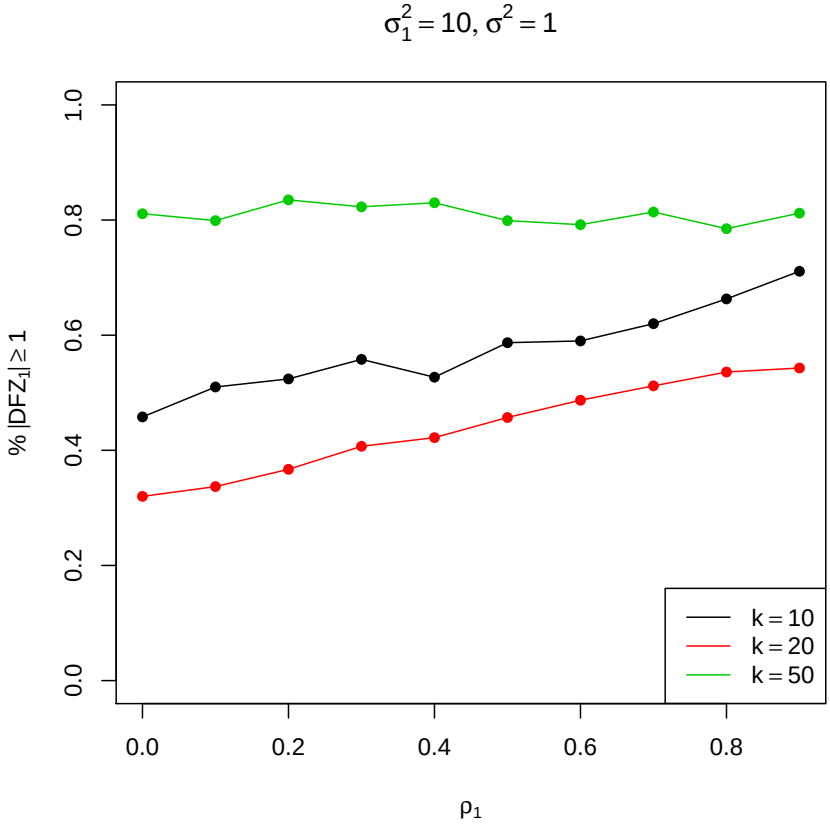

Figure 4.6: Percentage $|DFZ_1| \geq 1$ with $\rho = 0$ and $n_i \sim DU(2, 10)$

Figure 4.6, once again has similar results to those of Figure 4.5 in that a large number of subjects is preferred. When the number of subjects is large, it seems as though the percentage $|DFZ_i| \geq 1$ remains constant around 80 % regardless of the inconsistent observation's structure. Once again, in the preceding plot we observe the line corresponding to data sets with 20 subjects is lower than the line with 10 subjects. When paying particular attention to the structure of the data used for Figures 4.5 and 4.6, the number of repeated measurements used for data sets with 10, 20, and 50 subjects were 6, 3, and 8, respectively. Note, the number of repeated measurements for the data set with 20 subjects is smallest in comparison to the other two. Therefore, we may observe that DFZ is dependent on the

number of repeated measurements as well, where more weight is given to a larger number of repeated measurements.

The following section is a real example in which we apply our diagnostics. We are interested in observing those subjects identified as influential and furthermore determining how they become influential.

4.3 Data Description

The data set we will be examining is titled Estrogen. There was limited background information on the data set with respect to the type of subjects enrolled for the study, the type of tool used, and the purpose of the treatments administered. The only information known is it consists of different groups of subjects observed over 3 separate two-period crossover studies with repeated blood pressure measurements recorded. During each active treatment period, three systolic and diastolic readings were recorded per day over the course of three days. There were a total of 31 subjects in the 3 studies each of which received two treatments. Subjects received treatment for 4 weeks in each active treatment period and a 2-week washout period separated the two active treatment periods. The combinations for the crossover studies were: placebo vs. 0.625 mg, placebo vs. 1.25 mg, and 0.625 mg vs. 1.25 mg. The model used for each of the two-period crossover design experiments with repeated measurements was:

$$Y_{ijklt} = \mu + \delta_k + \gamma_{l(k)} + \alpha_i + \beta_j + \epsilon_{ijklt},$$

where δ_k is the fixed effect due to sequence k, α_i and β_j are the fixed effects due to treatment i and period j, and $\gamma_{l(k)}$ represents the l^{th} person in sequence k. The layout for such an experiment is displayed in the table below,

Table 4.1: Layout for a two-period two-treatment crossover design

	Period	
	1	2
	Time	Time
Sequence	$1\,2\,\ldots\,t$	$1\,2\,\ldots\,t$
1	A_1	A_2
2	A_2	A_1

where A_1 denotes the first treatment and A_2 denotes the second treatment [8].

The crossover design for which the data was collected is far too complex for the one-way random effects model. For this reason, we will consider only the data in the first period. This will allow us to reduce the complexity of the model and eliminate the sequence effect term. Under this circumstance, the fixed sequence and period effect terms drop out of the model because they are irrelevant. Essentially, the model reduces to a nested mixed model with a fixed effect and a random effect. The model we will use for preliminary analyses is:

$$Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \epsilon_{ijk}, \tag{4.1}$$

where α_i is the fixed effect due to treatment i and $\beta_{j(i)}$ is the random effect of the j^{th} subject in treatment i. Under this model we will assume, independent $\beta_{j(i)} \sim N(0, \sigma_a^2)$, independent $\epsilon_{ijk} \sim N(0, \sigma_e^2)$, and $\beta_{j(i)}$, ϵ_{ijk} mutually independent. We conducted two separate analyses, one for systolic and another for diastolic readings.

4.4 Systolic Analysis

The DFZ statistic was designed for use under the assumptions of the one-way random effects model. We were interested in testing for a fixed effect to determine if the model can be further reduced to meet the assumptions of the unbalanced one-way random effects model.

All subsequent modeling was performed in SAS, a summary of the model information is tabulated below.

Table 4.2: Nested Mixed Model Information

Treatment	Subjects	Repeated Measurements
1) Placebo	8	9,9,9,9,9,9,9,9
2) 0.625 mg	14	9,9,9,9,9,9,9, 9,9,9,6,9,9,9
3) 1.25 mg	9	9,9,9,9,9,9,9,9,9

The data was modelled in SAS using the proc mixed procedure with variance components and repeated measurements. Figure 4.7 is a normal-quantile plot to check the normality assumption of the model. There is no clear evidence to suggest severe departures from normality and with a Shapiro-Wilk test for normality p-value of 0.1169, there is lack of evidence to suggest the studentized residuals are not normally distributed.

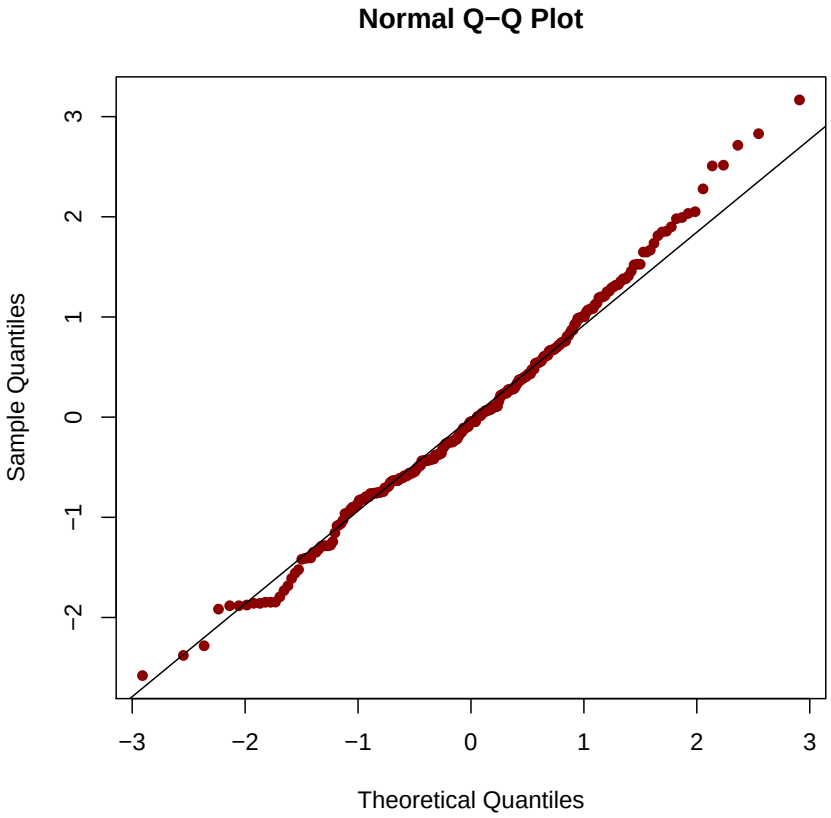

Figure 4.7: Probability Plot of model 4.1 Systolic residuals

From the variance components modeling performed in the SAS analysis, the constant variance assumption must also hold in order for the modeling to be considered adequate. Figure 4.8 is a plot of studentized residuals with respect to fits. From this, the residuals seem to be centered around zero in what seems to be a constant band.

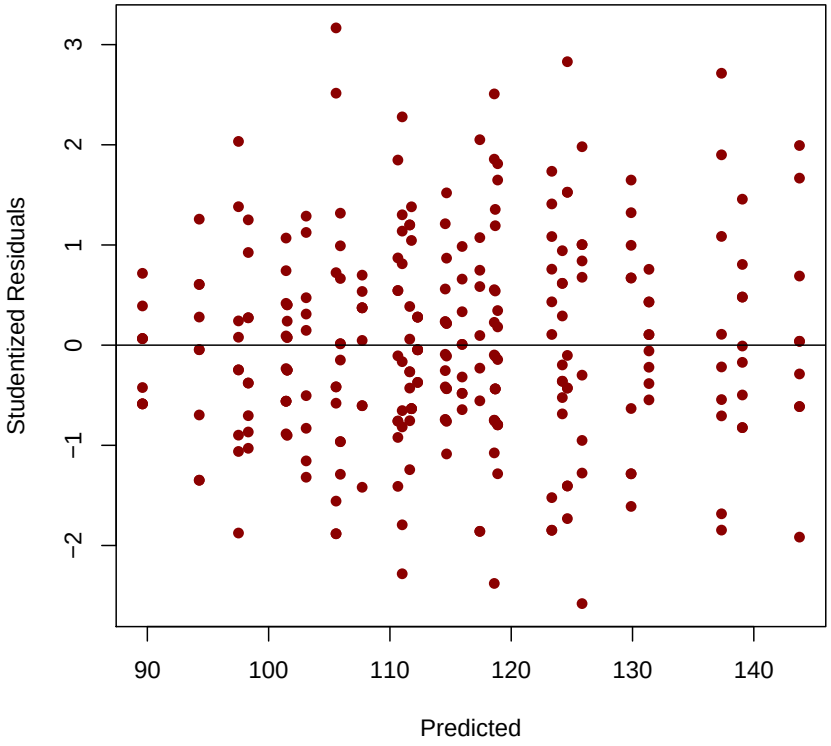

Figure 4.8: Plot of model 4.1 Systolic residuals versus fits

There is no strong evidence to suggest the data does not uphold the assumptions of model (4.1). Therefore, the subsequent testing for fixed and random effects is assumed appropriate.

Table 4.3: Nested Mixed Model Significance testing of effects

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	$p-value$
Treatment	2	28	0.04	0.9580
Covariance Parameter Estimates				
Parameter	Estimate	Standard Error	Z Value	$p-value$
Random	191.95	52.5591	3.65	0.0001
Measurement	42.2843	3.8202	11.07	$<.0001$

Table 4.3 is a summary of the SAS output from the nested mixed model. With a p-value of 0.9580, there is insufficient evidence to suggest there is a significant difference between the placebo, low, and high dose treatment effect on systolic blood pressure. This lack of significance allows us to eliminate the fixed effect term in the nested mixed model resulting in the one-way random effects model. Furthermore, the table above provides strong evidence to support the existence of variance components in the data. Thus, we may use the one-way random effects model for the systolic blood pressure measurements,

$$Y_{ij} = \mu + a_i + e_{ij},$$

where there are a total of $i = 1, \ldots, 31$ subjects with all having 9 repeated measurements except for 1 who had only 6 repeated measurements recorded. The analysis of variance table below, under the assumptions of the one-way random effects model, shows there is significant random effect and cannot be further reduced.

Table 4.4: Analysis of variance for Systolic blood pressures under the unbalanced one-way random effects model

Source of variation	DF	SS	MS	F	$P-value$
Subjects	30	49659.30	1655.31	39.15	0.000
Error	245	10360.22	42.29		
Total	275	60019.52			

Now that we know the systolic blood pressure measurements for period one achieve the assumptions of the one-way random effects model, we may apply our diagnostics to see if we may be able to identify any influential observations. Keeping in mind the observations of our simulation study and the diagnostics intrasubject/mean deviation diagnostics, we will attempt to identify those observations that are inconsistent due to large intrasubject variation or deviation from the overall mean. The repeated systolic blood pressure measurements are especially appealing for the purposes of this thesis because historically these measurements assume normality.

The intraclass correlation coefficient estimate for the systolic blood pressure measurements was 0.8108. This estimate is a strong indicator of there being good coherence among the repeated measurements in this data set. Alternatively, this can be thought of as most of the variability in this data set being explained by the random effect of the subjects in the study. Now, in an attempt to extract more information regarding the coherence of particular observations, we apply the diagnostics developed in the previous chapter.

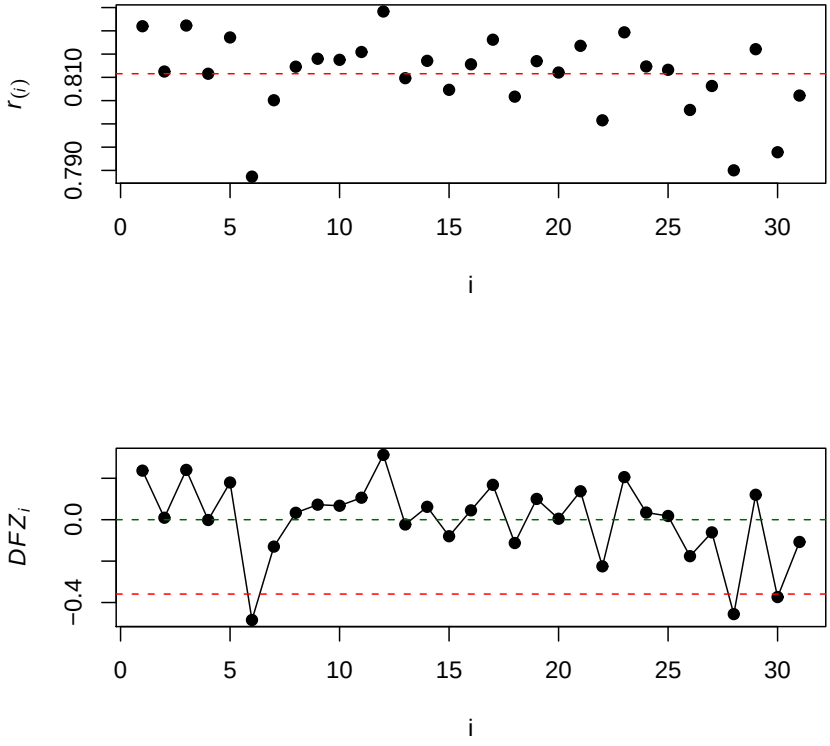

Figure 4.9: Diagnostic Plots of Systolic reduced model

There are two plots in Figure 4.9, one which is just a plot of the ICC estimate with the exclusion of the i^{th} observation and the other a plot of the DFZ_i statistic. From the plot of the ICC estimate with the exclusion of the i^{th}, it is difficult to extract any information from this with the exception of visually seeing whether the exclusion of a particular subject improves or reduces the ICC. The plot with the DFZ_i statistic shows that by our criterion, subjects 6, 28, and 30 were identified as influential on the estimation of the ICC.

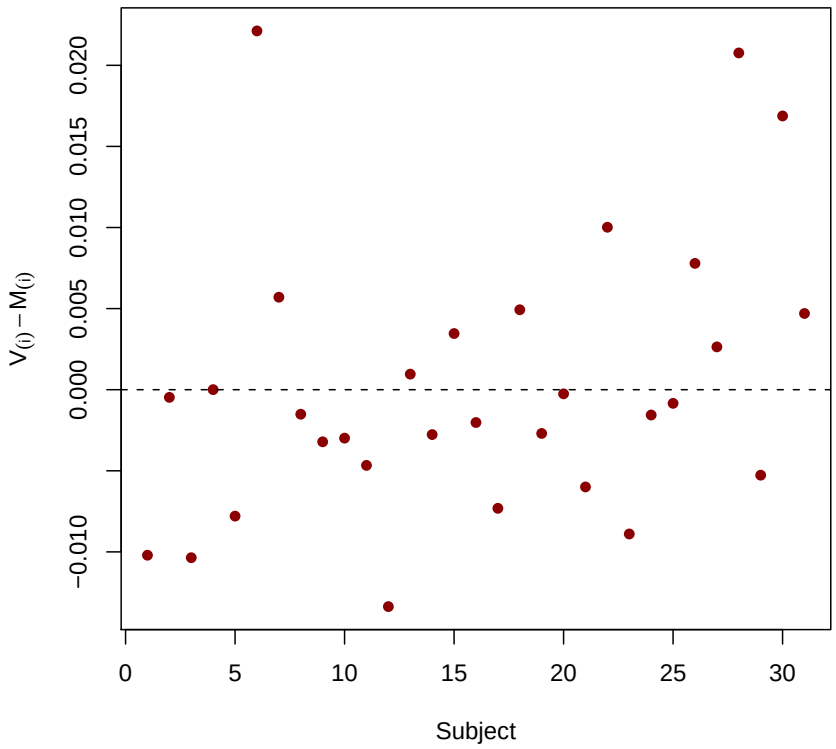

Figure 4.10: Diagnostic Variance/Mean influence of Systolic readings

In chapter 3, we showed $r - r_{(i)}$ can be interpreted in such a way to determine whether a subject is influential due to its intrasubject variation or severe deviation from the overall mean. The $r - r_{(i)}$ formula led us to identifying the two antagonistic terms that reveal how a subject is influential. In Figure 4.10, we can see subjects 6, 28, and 30 are positive and furthermore visibly further from the rest of the points. This indicates the $M_{(i)}$ terms is smaller than the $V_{(i)}$ term, in effect, implying the exclusion of these subjects, one at a time, from the data set results in a decrease of MSB. In other words, these subjects on average deviate from the overall mean of the data set.

Table 4.5: Systolic Pressure ICC estimate and confidence intervals

Omitting subject	$r_{(i)}$	Confidence Interval
none	0.8108	(0.802,0.861)
6	0.7887	(0.783,0.845)
28	0.7900	(0.781,0.844)
30	0.7939	(0.783,0.852)
6, 28	0.7633	(0.751,0.829)
6, 30	0.7632	(0.752,0.826)
28, 30	0.7702	(0.759,0.832)
6, 28, 30	0.7326	(0.718,0.801)

As with any diagnostic procedure, it is important to report estimates with and with the exclusion of those observations identified as influential. Table 4.5 displays the ICC estimates and 95% bootstrap confidence intervals with the exclusion of one, two, and all three influential observations. One particularly interesting observation to make is: when all three observations are excluded the confidence interval no longer overlaps with the one with all observations included and the ICC estimate is considerably reduced.

4.5 Diastolic Analysis

The same modeling procedure was performed on the diastolic blood pressure readings of the same 31 subjects. First we modeled the nested mixed model to determine if there is a significant treatment effect on the diastolic blood pressure readings of these subjects.

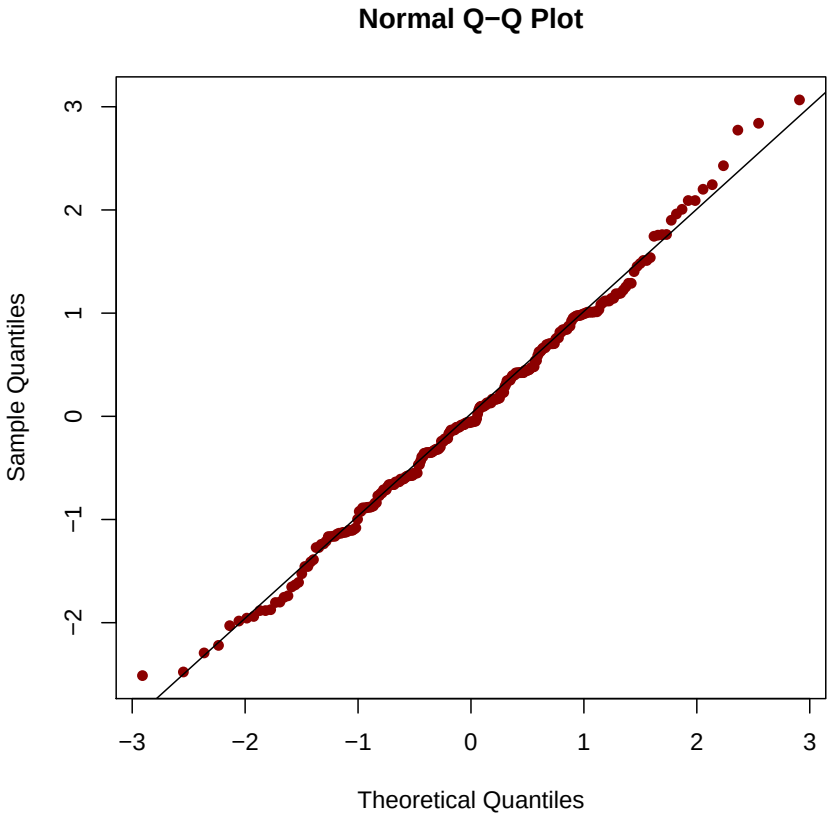

Figure 4.11: Probability Plot of model 4.1 Diastolic residuals

The normal probability plots of the studentized residuals shows there are no severe deviations from normality. After performing a Shapiro-Wilk test for normality, the p-value observed was 0.5091 indicating the diastolic data studentized residuals may be assumed to be normally distributed.

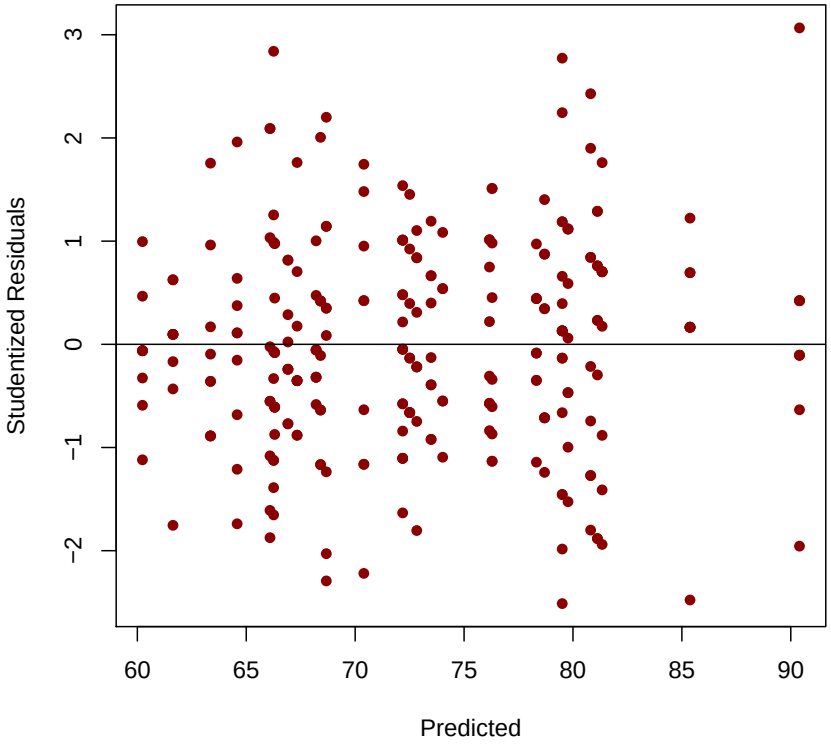

Figure 4.12: Plot of model 4.1 Diastolic residuals versus fits

Figure 4.12 is a plot of the diastolic data studentized residuals versus fits. From this, we can see the residuals may be assumed to be centered around zero. However, it is difficult to make the assumption of constant variance. The hour-glass shape of the plot shows there is a clear trend in the residuals and constant variance is difficult to assume. For our purposes we will assume the constant variance assumption holds and proceed with the significance testing.

Table 4.6: Nested Mixed Model Significance testing of effects for Diastolic Readings

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	$p-value$
Treatment	2	28	0.49	0.6202
Covariance Parameter Estimates				
Parameter	Estimate	Standard Error	Z Value	$p-value$
Random	56.8645	15.6751	3.63	0.0001
Measurement	16.0480	1.4499	11.07	$< .0001$

The nested mixed model significance testing performed in SAS, is summarized in Table 4.6. The significance test for the fixed effect determined there is no significant treatment effect on the diastolic blood pressure readings. Similar to the results of the systolic readings, the diastolic readings have significant covariance parameters due to random effect and measurement error. This allows us to model the data using the one-way random effect model and perform subsequent testing for presence of a random effect. Table 4.7 shows that under the assumptions of the one-way random effects model, the diastolic data has a significant random effect present.

Table 4.7: Analysis of variance for Diastolic blood pressures under the unbalanced one-way random effects model

Source of variation	DF	SS	MS	F	$P-value$
Subjects	30	15281.94	509.40	31.74	0.000
Error	245	3932.00	16.05		
Total	275	19213.94			

The ICC estimate for the diastolic data was observed to be 0.7754. This indicates there is a fair amount of coherence among the repeated measurements of the subjects. Up to this

point, the analysis has been identical to that of the systolic readings. Since the readings correspond to the same subjects, these results are not surprising.

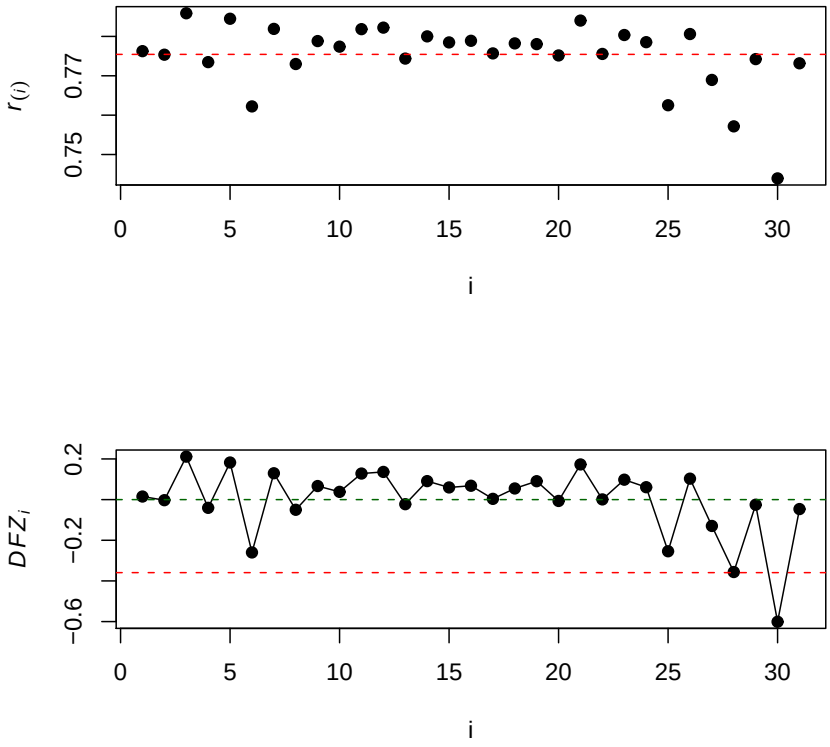

Figure 4.13: Diagnostic Plots of Diastolic blood pressure reduced model

After performing the diagnostics on the diastolic blood pressure data set, there were only two observations identified as influential. Subjects 28 and 30 were once again identified as influential, however, inconsistent with the systolic blood pressure measurements the subject number 6 was not identified as influential.

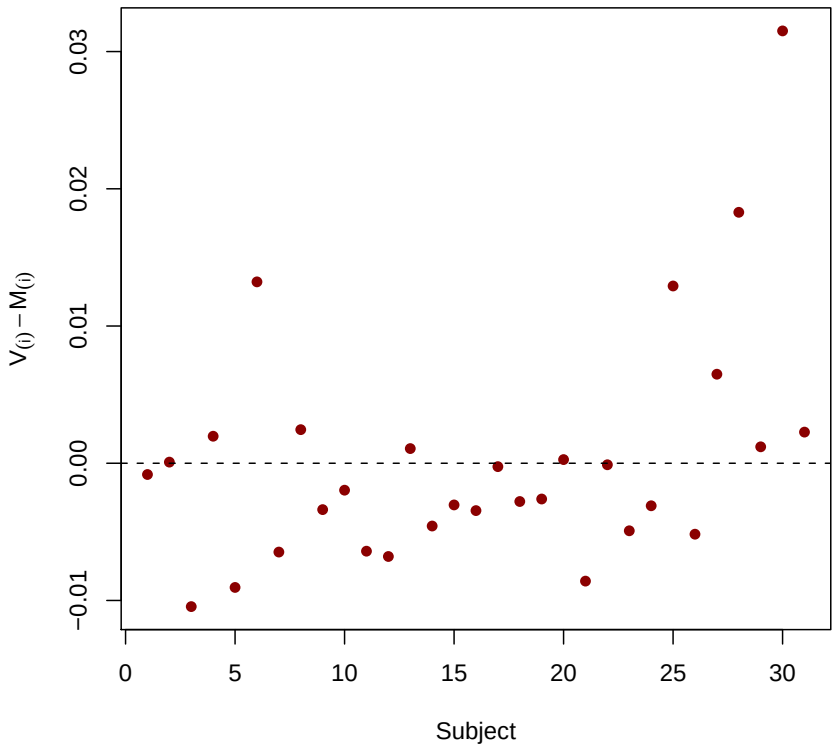

Figure 4.14: Diagnostic Variance/Mean influence of Diastolic readings

Once again, subjects 28 and 30 were identified as influential because their measurements on average deviate from the overall mean of the diastolic blood pressure measurements.

Table 4.8: Diastolic Pressure ICC estimate and confidence intervals

Omitting subject	$r_{(i)}$	Confidence Interval
none	0.7754	(0.763,0.836)
28	0.7572	(0.742,0.823)
30	0.7439	(0.726,0.816)
28, 30	0.7207	(0.702,0.795)

For the diastolic blood pressures, there were only two subjects identified as influential and so their exclusion resulted in a reduction of the ICC estimate, but, the reduction is not as severe as was the case with the systolic blood pressure data. This is due in large part because there were three subjects excluded from the analysis.

4.6 Results

For systolic blood pressure readings, our method identified three subjects as influential on the intraclass correlation coefficient. In all cases, the exclusion of each subject's measurements resulted in a reduction of the ICC estimate. For the diastolic blood pressure readings, only two subjects were identified as influential. Using the diagnostic tools presented in Chapter 3, we were able to determine the subjects that were identified as influential were identified as such because on average their measurements deviated from the overall mean. In both data sets, subjects 28 and 30 were identified as influential for the same reason. Table 4.9 displays the raw data of those subjects identified as influential.

Table 4.9: Raw Data of Subjects identified as influential

		Day								
		1			2			3		
		Reading			Reading			Reading		
Subject	Type	1	2	3	1	2	3	1	2	3
6	Systolic	156	154	144	148	140	142	144	132	140
28	Systolic	90	94	86	90	92	86	86	87	90
30	Systolic	144	139	142	134	138	134	136	148	142
6	Diastolic	86	76	86	88	88	86	90	86	86
28	Diastolic	62	64	60	60	60	56	58	59	60
30	Diastolic	92	83	90	92	88	90	90	102	92

Not surprisingly, the subjects that were identified as influential were those patients

whose blood pressure measurements were at extremes. That is, the repeated measurements of these subjects were on average either hyper or hypotensive. Two of the three subjects identified as influential using the systolic readings had readings consistently around the hypertensive level. One of the subjects using the systolic readings had measurements in the hypotensive level. For the diastolic readings one subject had measurements in the hypotensive level while the other had measurements in the hypertensive level. One particular inconsistency to note from the data is, subject 6 had systolic readings that were clearly hypertensive, however, the diastolic readings fell within the normal to pre-hypertensive levels.

These findings corroborate those found by Giraudeau et al. with respect to case-influence on the ICC. The reason why these subjects were identified as influential is that they did not contribute much to the total variance of the data set and their measurements were on average far from the global mean of the entire data set. When it is of interest to measure consistency of a blood pressure tool, these results indicate it is not advisable to exclude people whose blood pressure measurements are hypo/hypertensive.

Chapter 5

Conclusions and Discussions

This work set out to develop a method for identifying influential observations through the intraclass correlation coefficient. This problem reduced to measuring case influence on the ICC, using an adaptation of the DFBETA statistic used in least-squares regression. The use of Fishers **Z**-transformation was highly important to satisfy the normality assumption required. By adapting DFBETA statistic to the ICC, we are able to benefit from the criterion used to identify observations as influential, and its ability to adjust as sample size increases. As a consequence of the asymptotic properties of the transformation, the percentage of inconsistent observations the DFZ statistic identifies as influential increases as sample size increases. In literature, there was only one attempt to measure case influence.

In 1996, Giraudeau et al. published a paper in which they studied the properties of case-influence on the ICC under the assumptions of the balanced one-way random effect model using the maximum likelihood estimator. They developed a mathematical formula to show $r - r_{(i)}$ is dependent upon intrasubject variation and a subject's deviation from the overall mean. However, their method does not provide strict criterion for determining whether an observation should be considered influential and influence decreases as sample size increases. From this, our appeal for the DFZ is obvious since it has a criterion and adjusts for large sample sizes.

The diagnostics we have developed seemed to perform adequately at identifying inconsistent observations as influential under certain circumstances. In the large sample case, the DFZ statistic performed exceptionally well regardless of the structure of the data. In the small to moderate sample size case, it was a bit difficult for inconsistent observations to be identified as influential unless the inconsistent observation contributed a large amount

of variation or the structure of the inconsistent observation must be dramatically different from the rest of the data. That is, if the inconsistent observation lacks coherence among its repeated measurements it must be among a data set whose repeated measurements have strong coherence and vice versa. One major drawback to this method is that, it is sensitive to the unbalanced repeated measurement. Those observations that have less repeated measurements will be less likely to be identified as influential.

Our real example of repeated blood pressure measurements was a large sample case where subjects who were identified as influential, were identified as such due to their deviation from the overall mean. When dealing with blood pressure measurements, patients who are hypo/hypertensive deviate from the normal blood pressure range. As expected, those patients whose measurements on average were in the hypo/hypertensive range were identified as influential. Given the methods we have developed, we were only able to tell they were influential due to deviating from the mean.

Our work was able to help with some of the drawbacks of the analysis of variance intraclass correlation coefficient estimator under the assumptions of the unbalanced one-way random effects model. One major drawback that we addressed was correcting for it being able to assume negative values by developing our proportion correction procedure. Additionally, through the use of DFZ we have strict criteria for determining whether an observation is influential. However, the major drawbacks are due to the unbalanced repeated measurements, small sample size, lack of a strong $r - r_{(i)}$ diagnostic, and the simplicity of the one-way random effects model. Some future work as a consequence of these drawbacks is to address extending DFZ to more complicated models and developing diagnostics that can address a comparison of the coherence among individual subjects in comparison to others.

References

[1] A. Donner. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, 54(1):67–68, 1986.

[2] A. Donner and J.J. Koval. The estimation of intraclass correlation in the analysis of family data. *Biometrics*, 36(1):19–25, March 1980.

[3] R. A. Fisher. *Statistical methods for research workers*. Hafner Publishing Company Inc., NY, 1958.

[4] B. Giraudeau, A. Mallet, and C. Chastang. Case influence on the intraclass correlation coefficient estimate. *Biometrics*, 52:1492–1497, December 1996.

[5] M. Kutner, C.J. Nachtsheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill Irwin, NY, 5th edition, 2005.

[6] A. Liu, E. F. Schisterman, and C. Wu. Multistage evaluation of measurement error in a reliability study. *Biometrics*, 62:1191, December 2006.

[7] I. Olkin and J. Pratt. Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, 29(1):201–211, March 1958.

[8] R. L. Ott and M. Longnecker. *An Introduction to Statistical Methods and Data Analysis*. Duxbury, CA, 2001.

[9] D. J. Young and M. Bhandary. Test for equality of intraclass correlation coefficient under unequal family sizes. *Biometrics*, 54:1363, December 1998.

Appendix A

Simulation Study Tables

Table A.1: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=10 $\sigma^2 = 1$ $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$										
	ρ_1 ($\%\|DFZ_1\| \geq 1$)									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.494	0.654	0.745	0.819	0.855	0.892	0.937	0.946	0.968	0.983
0.8	0.422	0.537	0.592	0.628	0.719	0.792	0.820	0.890	0.907	0.952
0.7	0.432	0.456	0.512	0.569	0.634	0.689	0.737	0.790	0.842	0.888
0.6	0.447	0.418	0.476	0.522	0.544	0.572	0.644	0.689	0.765	0.822
0.5	0.496	0.457	0.487	0.525	0.499	0.529	0.580	0.632	0.658	0.751
0.4	0.514	0.431	0.443	0.477	0.512	0.484	0.538	0.57	0.625	0.681
0.3	0.555	0.531	0.499	0.483	0.492	0.487	0.509	0.550	0.564	0.580
0.2	0.612	0.527	0.538	0.525	0.479	0.520	0.469	0.495	0.522	0.587
0.1	0.651	0.619	0.577	0.538	0.521	0.516	0.498	0.485	0.512	0.504
0	0.711	0.663	0.62	0.59	0.587	0.527	0.558	0.524	0.510	0.458

Table A.2: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=10 $\sigma^2 = 1$ $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$										
	ρ_1 ($\%\|DFZ_1\| \geq 1$)									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.304	0.502	0.595	0.715	0.792	0.863	0.902	0.922	0.945	0.960
0.8	0.245	0.326	0.408	0.490	0.610	0.665	0.728	0.795	0.854	0.892
0.7	0.273	0.257	0.282	0.335	0.422	0.498	0.587	0.652	0.740	0.769
0.6	0.323	0.290	0.266	0.298	0.350	0.399	0.453	0.492	0.594	0.670
0.5	0.356	0.313	0.257	0.286	0.301	0.298	0.386	0.452	0.473	0.566
0.4	0.400	0.329	0.306	0.303	0.282	0.322	0.326	0.342	0.398	0.452
0.3	0.468	0.370	0.343	0.323	0.282	0.313	0.289	0.345	0.353	0.385
0.2	0.499	0.425	0.388	0.347	0.286	0.289	0.323	0.298	0.297	0.328
0.1	0.53	0.496	0.427	0.393	0.375	0.360	0.321	0.297	0.283	0.30
0	0.592	0.582	0.515	0.486	0.455	0.411	0.366	0.332	0.314	0.313

Table A.3: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=10 $\sigma^2 = 1$ $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.158	0.320	0.467	0.580	0.719	0.794	0.832	0.879	0.918	0.934
0.8	0.168	0.180	0.256	0.314	0.390	0.501	0.602	0.65	0.727	0.795
0.7	0.219	0.173	0.181	0.181	0.256	0.300	0.405	0.477	0.544	0.635
0.6	0.241	0.208	0.175	0.154	0.182	0.230	0.288	0.326	0.387	0.454
0.5	0.300	0.225	0.195	0.166	0.164	0.197	0.212	0.210	0.313	0.384
0.4	0.322	0.283	0.224	0.204	0.174	0.147	0.158	0.181	0.188	0.284
0.3	0.370	0.311	0.257	0.225	0.214	0.171	0.170	0.161	0.171	0.187
0.2	0.378	0.370	0.313	0.263	0.257	0.202	0.181	0.130	0.160	0.151
0.1	0.451	0.401	0.352	0.340	0.288	0.250	0.221	0.172	0.172	0.156
0	0.534	0.488	0.479	0.401	0.381	0.335	0.291	0.236	0.195	0.130

Table A.4: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=10 $\sigma^2 = 1$ $n_i = \{6, 8, 2, 2, 2, 6, 8, 6, 6, 5\}$										
	ρ_1 ($\%\|DFZ_1\| \geq 1$)									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.040	0.044	0.096	0.196	0.277	0.378	0.488	0.562	0.622	0.695
0.8	0.067	0.039	0.036	0.053	0.071	0.101	0.122	0.196	0.249	0.347
0.7	0.064	0.044	0.042	0.034	0.025	0.033	0.044	0.072	0.088	0.114
0.6	0.082	0.066	0.058	0.033	0.027	0.024	0.026	0.022	0.039	0.055
0.5	0.109	0.095	0.084	0.047	0.043	0.035	0.025	0.017	0.017	0.022
0.4	0.12	0.115	0.089	0.058	0.060	0.036	0.034	0.016	0.014	0.016
0.3	0.150	0.116	0.124	0.092	0.071	0.063	0.033	0.019	0.017	0.006
0.2	0.181	0.170	0.121	0.092	0.094	0.080	0.046	0.032	0.021	0.006
0.1	0.236	0.206	0.194	0.155	0.136	0.113	0.078	0.051	0.027	0.017
0	0.305	0.287	0.250	0.207	0.187	0.165	0.123	0.080	0.055	0.023

Table A.5: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=10 $\sigma^2 = 1$ $n_i = \{17, 11, 13, 20, 15, 20, 15, 17, 12, 12\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.553	0.693	0.815	0.858	0.908	0.944	0.985	0.992	0.998	1.000
0.8	0.411	0.539	0.589	0.726	0.787	0.853	0.911	0.951	0.984	0.998
0.7	0.342	0.462	0.542	0.623	0.666	0.747	0.815	0.887	0.948	0.991
0.6	0.463	0.408	0.499	0.544	0.594	0.643	0.715	0.807	0.904	0.979
0.5	0.491	0.417	0.433	0.465	0.510	0.590	0.629	0.739	0.835	0.943
0.4	0.549	0.479	0.443	0.460	0.538	0.529	0.571	0.656	0.743	0.901
0.3	0.618	0.576	0.512	0.479	0.523	0.510	0.542	0.586	0.660	0.838
0.2	0.674	0.616	0.551	0.527	0.514	0.530	0.542	0.547	0.579	0.706
0.1	0.732	0.658	0.668	0.631	0.593	0.589	0.569	0.563	0.515	0.635
0	0.828	0.810	0.775	0.735	0.724	0.710	0.671	0.635	0.572	0.531

Table A.6: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=10 $\sigma^2 = 1$ $n_i = \{17, 11, 13, 20, 15, 20, 15, 17, 12, 12\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.278	0.560	0.741	0.839	0.883	0.940	0.977	0.985	0.998	0.999
0.8	0.245	0.275	0.423	0.534	0.689	0.746	0.830	0.932	0.969	0.995
0.7	0.31	0.258	0.269	0.363	0.448	0.589	0.687	0.798	0.896	0.974
0.6	0.353	0.263	0.234	0.273	0.298	0.433	0.515	0.620	0.760	0.897
0.5	0.404	0.353	0.274	0.245	0.264	0.351	0.374	0.467	0.640	0.793
0.4	0.468	0.385	0.317	0.302	0.257	0.275	0.30	0.350	0.454	0.657
0.3	0.488	0.469	0.376	0.346	0.284	0.243	0.272	0.276	0.367	0.517
0.2	0.553	0.501	0.483	0.429	0.333	0.294	0.270	0.255	0.266	0.374
0.1	0.661	0.624	0.542	0.491	0.472	0.408	0.341	0.305	0.251	0.256
0	0.781	0.737	0.715	0.679	0.663	0.590	0.568	0.469	0.376	0.256

Table A.7: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=10 $\sigma^2 = 1$ $n_i = \{17, 11, 13, 20, 15, 20, 15, 17, 12, 12\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.107	0.318	0.546	0.707	0.819	0.918	0.943	0.972	0.994	0.999
0.8	0.191	0.132	0.170	0.334	0.459	0.629	0.719	0.847	0.921	0.967
0.7	0.228	0.173	0.133	0.145	0.208	0.310	0.454	0.586	0.711	0.868
0.6	0.263	0.214	0.16	0.121	0.105	0.151	0.229	0.321	0.503	0.627
0.5	0.300	0.244	0.192	0.143	0.128	0.112	0.143	0.177	0.267	0.473
0.4	0.347	0.316	0.266	0.202	0.144	0.111	0.093	0.121	0.148	0.293
0.3	0.408	0.360	0.308	0.271	0.211	0.158	0.106	0.085	0.094	0.151
0.2	0.435	0.436	0.375	0.341	0.287	0.221	0.177	0.112	0.091	0.076
0.1	0.582	0.553	0.497	0.456	0.424	0.339	0.253	0.211	0.106	0.051
0	0.732	0.696	0.687	0.611	0.601	0.527	0.488	0.397	0.295	0.100

Table A.8: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=10 $\sigma^2 = 1$ $n_i = \{17, 11, 13, 20, 15, 20, 15, 17, 12, 12\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.035	0.019	0.035	0.109	0.241	0.425	0.576	0.733	0.799	0.901
0.8	0.052	0.035	0.016	0.021	0.01	0.032	0.071	0.123	0.196	0.335
0.7	0.063	0.053	0.036	0.035	0.009	0.01	0.007	0.016	0.019	0.057
0.6	0.106	0.077	0.052	0.040	0.020	0.019	0.005	0.004	0.005	0.006
0.5	0.101	0.100	0.079	0.060	0.038	0.011	0.012	0.008	0.002	0.002
0.4	0.155	0.119	0.106	0.076	0.053	0.035	0.025	0.007	0.005	0.001
0.3	0.184	0.159	0.149	0.112	0.100	0.063	0.037	0.014	0.004	0.002
0.2	0.219	0.263	0.188	0.160	0.122	0.093	0.060	0.053	0.021	0.003
0.1	0.377	0.319	0.280	0.251	0.230	0.169	0.124	0.068	0.034	0.003
0	0.555	0.509	0.486	0.467	0.370	0.359	0.294	0.213	0.119	0.024

Table A.9: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=20 $\sigma^2 = 1$ $n_i = \{3,3,6,7,6,5,7,7,5,9,2,5,3,4,4,3,8,9,7,10\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.206	0.338	0.45	0.486	0.569	0.637	0.670	0.728	0.764	0.770
0.8	0.183	0.204	0.277	0.326	0.384	0.441	0.496	0.564	0.591	0.610
0.7	0.217	0.206	0.213	0.270	0.296	0.319	0.367	0.418	0.456	0.495
0.6	0.241	0.204	0.221	0.223	0.231	0.243	0.283	0.32	0.362	0.391
0.5	0.290	0.263	0.251	0.226	0.238	0.245	0.253	0.279	0.305	0.346
0.4	0.300	0.292	0.254	0.246	0.245	0.239	0.264	0.257	0.242	0.261
0.3	0.367	0.347	0.329	0.297	0.255	0.247	0.230	0.241	0.242	0.245
0.2	0.462	0.395	0.359	0.346	0.314	0.308	0.259	0.247	0.226	0.220
0.1	0.483	0.467	0.442	0.393	0.392	0.335	0.309	0.306	0.276	0.246
0	0.543	0.536	0.512	0.487	0.457	0.422	0.407	0.367	0.337	0.320

Table A.10: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=20 $\sigma^2 = 1$ $n_i = \{3,3,6,7,6,5,7,7,5,9,2,5,3,4,4,3,8,9,7,10\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.071	0.129	0.239	0.330	0.426	0.464	0.519	0.571	0.610	0.641
0.8	0.077	0.086	0.087	0.109	0.214	0.245	0.289	0.345	0.380	0.430
0.7	0.096	0.072	0.081	0.090	0.115	0.142	0.160	0.189	0.228	0.306
0.6	0.130	0.095	0.088	0.081	0.080	0.106	0.121	0.124	0.163	0.204
0.5	0.150	0.119	0.108	0.094	0.084	0.077	0.090	0.095	0.11	0.131
0.4	0.159	0.165	0.153	0.102	0.105	0.096	0.074	0.058	0.087	0.083
0.3	0.251	0.175	0.156	0.132	0.134	0.10	0.077	0.079	0.073	0.082
0.2	0.259	0.255	0.216	0.184	0.169	0.141	0.141	0.106	0.084	0.082
0.1	0.347	0.299	0.276	0.264	0.229	0.196	0.189	0.154	0.126	0.097
0	0.430	0.425	0.342	0.346	0.317	0.286	0.262	0.226	0.197	0.168

Table A.11: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=20 $\sigma^2 = 1$ $n_i = \{3,3,6,7,6,5,7,7,5,9,2,5,3,4,4,3,8,9,7,10\}$										
	ρ_1 ($\%\|DFZ_1\| \geq 1$)									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.03	0.051	0.119	0.183	0.267	0.323	0.369	0.439	0.451	0.498
0.8	0.035	0.020	0.040	0.061	0.080	0.103	0.141	0.178	0.228	0.256
0.7	0.050	0.034	0.023	0.028	0.039	0.035	0.060	0.097	0.123	0.137
0.6	0.060	0.051	0.024	0.022	0.016	0.029	0.032	0.044	0.052	0.077
0.5	0.084	0.072	0.048	0.043	0.035	0.025	0.030	0.023	0.026	0.042
0.4	0.084	0.079	0.063	0.057	0.049	0.024	0.020	0.028	0.016	0.017
0.3	0.141	0.108	0.087	0.064	0.059	0.045	0.030	0.028	0.026	0.025
0.2	0.162	0.123	0.139	0.131	0.080	0.066	0.056	0.044	0.030	0.019
0.1	0.229	0.176	0.155	0.160	0.133	0.133	0.098	0.075	0.059	0.037
0	0.311	0.305	0.275	0.229	0.233	0.190	0.147	0.122	0.106	0.096

Table A.12: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=20 $\sigma^2 = 1$ $n_i = \{3,3,6,7,6,5,7,7,5,9,2,5,3,4,4,3,8,9,7,10\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.001	0.0	0.004	0.009	0.029	0.046	0.076	0.113	0.147	0.157
0.8	0.0	0.0	0.001	0.0	0.002	0.003	0.007	0.014	0.020	0.032
0.7	0.003	0.001	0.0	0.0	0.001	0.001	0.0	0.0	0.004	0.006
0.6	0.003	0.001	0.005	0.001	0.001	0.0	0.0	0.0	0.0	0.0
0.5	0.002	0.002	0.001	0.002	0.001	0.001	0.0	0.001	0.0	0.0
0.4	0.006	0.005	0.002	0.002	0.001	0.002	0.0	0.0	0.0	0.0
0.3	0.010	0.013	0.012	0.008	0.007	0.001	0.001	0.0	0.0	0.0
0.2	0.025	0.020	0.012	0.019	0.007	0.006	0.007	0.001	0.002	0.0
0.1	0.053	0.045	0.035	0.020	0.021	0.012	0.009	0.005	0.004	0.002
0	0.097	0.060	0.075	0.055	0.041	0.030	0.031	0.016	0.010	0.001

Table A.13: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=20 $\sigma^2 = 1$										
$n_i = \{15, 14, 15, 16, 19, 16, 14, 11, 13, 20, 17, 14, 18, 14, 11, 17, 20, 16, 16, 13\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.404	0.687	0.802	0.885	0.929	0.961	0.990	0.992	0.998	1.000
0.8	0.277	0.399	0.558	0.654	0.776	0.870	0.935	0.950	0.992	0.999
0.7	0.316	0.29	0.398	0.512	0.645	0.722	0.814	0.867	0.943	0.992
0.6	0.376	0.308	0.337	0.410	0.501	0.556	0.677	0.755	0.886	0.966
0.5	0.445	0.341	0.340	0.329	0.389	0.503	0.538	0.642	0.775	0.922
0.4	0.480	0.415	0.367	0.366	0.351	0.422	0.491	0.521	0.630	0.833
0.3	0.520	0.506	0.439	0.387	0.374	0.373	0.410	0.428	0.558	0.701
0.2	0.608	0.558	0.522	0.447	0.423	0.414	0.412	0.423	0.429	0.558
0.1	0.686	0.614	0.599	0.559	0.518	0.489	0.463	0.424	0.430	0.435
0	0.819	0.769	0.739	0.689	0.682	0.639	0.618	0.572	0.502	0.413

Table A.14: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=20 $\sigma^2 = 1$										
$n_i = \{15, 14, 15, 16, 19, 16, 14, 11, 13, 20, 17, 14, 18, 14, 11, 17, 20, 16, 16, 13\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.141	0.349	0.629	0.806	0.878	0.938	0.973	0.990	0.999	1.000
0.8	0.177	0.134	0.213	0.421	0.552	0.709	0.806	0.882	0.954	0.976
0.7	0.225	0.152	0.124	0.180	0.271	0.394	0.532	0.694	0.820	0.921
0.6	0.258	0.201	0.160	0.124	0.144	0.204	0.34	0.453	0.626	0.748
0.5	0.306	0.234	0.195	0.156	0.111	0.120	0.175	0.263	0.399	0.553
0.4	0.383	0.297	0.253	0.201	0.162	0.139	0.120	0.142	0.222	0.406
0.3	0.408	0.385	0.318	0.272	0.201	0.124	0.119	0.120	0.128	0.235
0.2	0.475	0.440	0.382	0.359	0.302	0.232	0.170	0.120	0.093	0.124
0.1	0.575	0.547	0.514	0.436	0.391	0.349	0.268	0.225	0.128	0.073
0	0.753	0.702	0.654	0.653	0.603	0.544	0.499	0.421	0.281	0.131

Table A.15: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=20 $\sigma^2 = 1$										
$n_i = \{15, 14, 15, 16, 19, 16, 14, 11, 13, 20, 17, 14, 18, 14, 11, 17, 20, 16, 16, 13\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.053	0.109	0.352	0.623	0.762	0.855	0.932	0.970	0.987	0.997
0.8	0.110	0.065	0.054	0.079	0.236	0.365	0.557	0.682	0.782	0.889
0.7	0.138	0.102	0.047	0.051	0.059	0.11	0.222	0.312	0.46	0.619
0.6	0.145	0.116	0.116	0.069	0.04	0.042	0.065	0.134	0.207	0.338
0.5	0.211	0.167	0.130	0.102	0.066	0.049	0.027	0.046	0.084	0.168
0.4	0.246	0.204	0.164	0.124	0.106	0.056	0.035	0.017	0.024	0.062
0.3	0.294	0.257	0.235	0.184	0.142	0.086	0.068	0.022	0.014	0.029
0.2	0.381	0.341	0.299	0.250	0.197	0.170	0.109	0.070	0.019	0.008
0.1	0.491	0.447	0.406	0.406	0.321	0.274	0.214	0.123	0.050	0.008
0	0.653	0.65	0.620	0.556	0.502	0.467	0.399	0.316	0.209	0.071

Table A.16: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=20 $\sigma^2 = 1$										
$n_i = \{15, 14, 15, 16, 19, 16, 14, 11, 13, 20, 17, 14, 18, 14, 11, 17, 20, 16, 16, 13\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.006	0.005	0.003	0.01	0.055	0.13	0.266	0.395	0.524	0.642
0.8	0.012	0.008	0.005	0.002	0.002	0.001	0.004	0.018	0.042	0.067
0.7	0.015	0.014	0.006	0.003	0.003	0.0	0.0	0.001	0.002	0.005
0.6	0.029	0.021	0.011	0.008	0.007	0.001	0.0	0.0	0.0	0.0
0.5	0.043	0.026	0.026	0.012	0.009	0.004	0.001	0.0	0.0	0.0
0.4	0.058	0.053	0.039	0.025	0.01	0.007	0.006	0.0	0.0	0.0
0.3	0.096	0.075	0.071	0.038	0.034	0.025	0.011	0.003	0.0	0.0
0.2	0.153	0.108	0.106	0.082	0.064	0.035	0.022	0.009	0.003	0.0
0.1	0.243	0.225	0.195	0.164	0.128	0.114	0.059	0.032	0.007	0.0
0	0.439	0.393	0.408	0.360	0.296	0.266	0.229	0.138	0.075	0.01

Table A.17: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=50 $\sigma^2 = 1$										
$n_i = \{8, 10, 7, 6, 10, 5, 2, 8, 6, 2, 9, 7, 7, 4, 2, 7, 4, 6, 2, 9, 7, 6, 5, 6, 10,$ $8, 10, 2, 3, 4, 6, 7, 6, 7, 7, 8, 6, 10, 10, 4, 10, 3, 8, 4, 5, 5, 6, 7, 8, 2\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.846	0.921	0.941	0.965	0.982	0.996	0.994	0.999	1.000	1.000
0.8	0.711	0.809	0.867	0.897	0.936	0.962	0.975	0.986	0.992	0.999
0.7	0.648	0.772	0.820	0.858	0.908	0.921	0.929	0.964	0.971	0.991
0.6	0.635	0.743	0.792	0.843	0.832	0.865	0.914	0.931	0.954	0.980
0.5	0.652	0.707	0.778	0.789	0.834	0.849	0.884	0.894	0.924	0.969
0.4	0.671	0.695	0.751	0.752	0.784	0.791	0.859	0.867	0.897	0.921
0.3	0.697	0.715	0.731	0.759	0.79	0.804	0.816	0.849	0.861	0.910
0.2	0.706	0.725	0.770	0.766	0.766	0.820	0.818	0.836	0.847	0.881
0.1	0.771	0.750	0.779	0.795	0.798	0.811	0.820	0.815	0.811	0.849
0	0.812	0.785	0.814	0.792	0.799	0.830	0.823	0.835	0.799	0.811

Table A.18: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=50 $\sigma^2 = 1$										
$n_i = \{8, 10, 7, 6, 10, 5, 2, 8, 6, 2, 9, 7, 7, 4, 2, 7, 4, 6, 2, 9, 7, 6, 5, 6, 10,$ $8, 10, 2, 3, 4, 6, 7, 6, 7, 7, 8, 6, 10, 10, 4, 10, 3, 8, 4, 5, 5, 6, 7, 8, 2\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.629	0.822	0.898	0.95	0.959	0.983	0.99	0.995	0.999	1
0.8	0.492	0.656	0.745	0.826	0.894	0.924	0.952	0.968	0.986	0.987
0.7	0.454	0.544	0.67	0.709	0.778	0.84	0.871	0.912	0.945	0.974
0.6	0.466	0.51	0.551	0.645	0.677	0.759	0.796	0.848	0.92	0.941
0.5	0.526	0.524	0.539	0.599	0.649	0.704	0.736	0.825	0.855	0.903
0.4	0.568	0.525	0.544	0.558	0.627	0.648	0.703	0.732	0.791	0.84
0.3	0.576	0.565	0.525	0.565	0.573	0.611	0.653	0.675	0.704	0.784
0.2	0.643	0.584	0.579	0.583	0.603	0.589	0.638	0.629	0.663	0.734
0.1	0.716	0.667	0.62	0.626	0.626	0.615	0.614	0.589	0.645	0.641
0	0.748	0.726	0.702	0.668	0.655	0.668	0.661	0.676	0.625	0.645

Table A.19: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=50 $\sigma^2 = 1$										
$n_i = \{8, 10, 7, 6, 10, 5, 2, 8, 6, 2, 9, 7, 7, 4, 2, 7, 4, 6, 2, 9, 7, 6, 5, 6, 10,$ $8, 10, 2, 3, 4, 6, 7, 6, 7, 7, 8, 6, 10, 10, 4, 10, 3, 8, 4, 5, 5, 6, 7, 8, 2\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.446	0.675	0.837	0.908	0.945	0.97	0.986	0.992	0.997	0.998
0.8	0.294	0.443	0.564	0.702	0.794	0.848	0.892	0.93	0.971	0.98
0.7	0.342	0.353	0.465	0.535	0.622	0.722	0.781	0.847	0.891	0.932
0.6	0.424	0.39	0.385	0.439	0.522	0.595	0.684	0.726	0.796	0.894
0.5	0.433	0.377	0.381	0.404	0.412	0.499	0.58	0.622	0.713	0.801
0.4	0.489	0.418	0.396	0.38	0.397	0.427	0.492	0.551	0.607	0.73
0.3	0.538	0.462	0.425	0.42	0.413	0.423	0.411	0.449	0.517	0.564
0.2	0.582	0.538	0.488	0.463	0.45	0.405	0.421	0.426	0.441	0.498
0.1	0.62	0.603	0.555	0.548	0.477	0.454	0.452	0.406	0.393	0.44
0	0.706	0.666	0.614	0.574	0.51	0.531	0.537	0.471	0.436	0.431

Table A.20: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=50 $\sigma^2 = 1$										
$n_i = \{8, 10, 7, 6, 10, 5, 2, 8, 6, 2, 9, 7, 7, 4, 2, 7, 4, 6, 2, 9, 7, 6, 5, 6, 10,$ $8, 10, 2, 3, 4, 6, 7, 6, 7, 7, 8, 6, 10, 10, 4, 10, 3, 8, 4, 5, 5, 6, 7, 8, 2\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.081	0.207	0.397	0.546	0.708	0.796	0.896	0.914	0.941	0.955
0.8	0.13	0.079	0.095	0.174	0.285	0.404	0.529	0.623	0.736	0.792
0.7	0.161	0.117	0.094	0.09	0.122	0.178	0.239	0.328	0.456	0.545
0.6	0.177	0.128	0.115	0.071	0.083	0.085	0.112	0.17	0.263	0.333
0.5	0.22	0.185	0.149	0.111	0.068	0.068	0.065	0.09	0.14	0.203
0.4	0.295	0.213	0.182	0.151	0.102	0.09	0.076	0.069	0.069	0.12
0.3	0.293	0.269	0.234	0.19	0.146	0.104	0.07	0.051	0.041	0.057
0.2	0.37	0.3	0.308	0.27	0.221	0.168	0.128	0.06	0.041	0.034
0.1	0.446	0.405	0.355	0.296	0.268	0.228	0.181	0.12	0.071	0.035
0	0.518	0.503	0.466	0.428	0.376	0.33	0.259	0.199	0.11	0.057

Table A.21: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 10$

k=50 $\sigma^2 = 1$										
$n_i = \{10, 13, 16, 16, 17, 14, 19, 12, 15, 10, 20, 10, 11, 20, 14, 14,$ 19,14,16,18,13,10,12,14,17,11,12,19,17,12,17,15,12,15, $12, 16, 16, 12, 18, 15, 19, 14, 19, 11, 18, 17, 18, 17, 13, 11\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.71	0.847	0.915	0.95	0.982	0.988	0.995	0.999	1	1
0.8	0.526	0.682	0.79	0.833	0.921	0.943	0.963	0.983	0.995	1
0.7	0.464	0.615	0.692	0.753	0.83	0.862	0.92	0.948	0.979	0.997
0.6	0.486	0.563	0.613	0.701	0.752	0.797	0.844	0.902	0.948	0.976
0.5	0.527	0.514	0.6	0.662	0.711	0.745	0.778	0.831	0.898	0.947
0.4	0.577	0.558	0.617	0.638	0.681	0.732	0.757	0.8	0.835	0.895
0.3	0.596	0.593	0.624	0.65	0.672	0.699	0.716	0.743	0.798	0.853
0.2	0.704	0.658	0.644	0.677	0.671	0.681	0.701	0.723	0.727	0.777
0.1	0.739	0.698	0.688	0.689	0.728	0.714	0.752	0.745	0.703	0.734
0	0.846	0.821	0.805	0.773	0.796	0.808	0.801	0.801	0.753	0.73

Table A.22: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 5$

k=50 $\sigma^2 = 1$										
$n_i = \{10, 13, 16, 16, 17, 14, 19, 12, 15, 10, 20, 10, 11, 20, 14, 14,$ 19,14,16,18,13,10,12,14,17,11,12,19,17,12,17,15,12,15, $12, 16, 16, 12, 18, 15, 19, 14, 19, 11, 18, 17, 18, 17, 13, 11\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.439	0.723	0.831	0.897	0.951	0.981	0.991	0.999	1	1
0.8	0.26	0.419	0.61	0.696	0.785	0.866	0.919	0.956	0.981	0.993
0.7	0.321	0.313	0.437	0.522	0.652	0.759	0.815	0.872	0.947	0.973
0.6	0.361	0.328	0.352	0.456	0.498	0.614	0.683	0.767	0.85	0.923
0.5	0.375	0.353	0.345	0.367	0.418	0.495	0.544	0.662	0.751	0.851
0.4	0.475	0.417	0.345	0.369	0.379	0.451	0.456	0.564	0.641	0.756
0.3	0.496	0.448	0.416	0.398	0.394	0.394	0.433	0.511	0.54	0.659
0.2	0.576	0.525	0.49	0.469	0.438	0.443	0.431	0.454	0.492	0.533
0.1	0.631	0.639	0.588	0.556	0.515	0.53	0.482	0.448	0.436	0.445
0	0.792	0.761	0.735	0.708	0.647	0.675	0.647	0.597	0.529	0.469

Table A.23: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 3$

k=50 $\sigma^2 = 1$										
$n_i = \{10, 13, 16, 16, 17, 14, 19, 12, 15, 10, 20, 10, 11, 20, 14, 14,$ 19,14,16,18,13,10,12,14,17,11,12,19,17,12,17,15,12,15, $12, 16, 16, 12, 18, 15, 19, 14, 19, 11, 18, 17, 18, 17, 13, 11\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.191	0.502	0.684	0.841	0.902	0.95	0.967	0.982	0.995	0.997
0.8	0.175	0.194	0.356	0.495	0.638	0.771	0.827	0.904	0.937	0.973
0.7	0.219	0.167	0.197	0.257	0.378	0.496	0.633	0.743	0.838	0.9
0.6	0.231	0.21	0.175	0.203	0.246	0.35	0.445	0.552	0.653	0.762
0.5	0.285	0.249	0.206	0.182	0.199	0.275	0.296	0.393	0.5	0.616
0.4	0.35	0.283	0.238	0.222	0.186	0.196	0.257	0.3	0.364	0.523
0.3	0.401	0.365	0.327	0.286	0.215	0.209	0.193	0.217	0.247	0.345
0.2	0.494	0.448	0.404	0.37	0.314	0.277	0.246	0.198	0.219	0.241
0.1	0.582	0.573	0.499	0.488	0.425	0.372	0.345	0.267	0.22	0.167
0	0.728	0.699	0.68	0.654	0.596	0.587	0.496	0.47	0.392	0.255

Table A.24: Percentage Subject 1 Identified Influential with $\sigma_1^2 = 1$

k=50 $\sigma^2 = 1$										
$n_i = \{10, 13, 16, 16, 17, 14, 19, 12, 15, 10, 20, 10, 11, 20, 14, 14,$ 19,14,16,18,13,10,12,14,17,11,12,19,17,12,17,15,12,15, $12, 16, 16, 12, 18, 15, 19, 14, 19, 11, 18, 17, 18, 17, 13, 11\}$										
	ρ_1 $(\%\|DFZ_1\| \geq 1)$									
ρ	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
0.9	0.018	0.024	0.118	0.301	0.483	0.649	0.77	0.822	0.876	0.923
0.8	0.029	0.019	0.011	0.033	0.07	0.135	0.25	0.317	0.476	0.568
0.7	0.044	0.026	0.026	0.015	0.011	0.026	0.065	0.101	0.182	0.235
0.6	0.071	0.034	0.026	0.017	0.007	0.006	0.016	0.022	0.04	0.093
0.5	0.098	0.078	0.059	0.041	0.014	0.011	0.007	0.009	0.011	0.015
0.4	0.136	0.087	0.072	0.06	0.027	0.034	0.015	0.003	0.002	0.008
0.3	0.164	0.126	0.127	0.087	0.058	0.046	0.026	0.009	0.001	0.002
0.2	0.245	0.182	0.174	0.158	0.103	0.086	0.05	0.027	0.007	0
0.1	0.345	0.338	0.272	0.249	0.221	0.199	0.123	0.071	0.022	0.004
0	0.528	0.506	0.501	0.453	0.431	0.37	0.327	0.226	0.189	0.058

Curriculum Vitae

Angel de Jesus Dávalos was born and raised in El Paso, Texas. He is the third child born to Cristina Dávalos. He attended local area public schools, graduated from Jefferson High School in June 2004, and entered The University of Texas at El Paso in the fall of 2004. In May 2008, he received his Bachelor's degree in Mathematics.

In the fall of 2008, he entered the Master of Science in Statistics program. During his first year, he served as a teaching and research assistant. He worked as an intern for Dr. Aiyi Liu of the National Institute of Child Health and Human Development of the National Institutes of Health in the summer of 2009. The summer work led to his master's thesis topic. In his second year of graduate studies, he was awarded The University of Texas System Louis Stokes Association for Minority Participation's Bridge to the Doctorate Fellowship. Upon completion of the Master of Science in Statistics program, he plans to pursue a Ph.D. degree in Biostatistics at The University of North Carolina at Chapel Hill.

Current address: 4506 Blanco Avenue Apt. 4
El Paso, Texas 79905