UNF Digital Commons

UNF Graduate Theses and Dissertations

Student Scholarship

2015

Comparing Group Means When Nonresponse Rates Differ

Gabriela M. Stegmann
University of North Florida

Comparing Group Means when Nonresponse Rates Differ

by

Gabriela Maria Stegmann

A Thesis submitted to the Department of Mathematics and Statistics in partial fulfillment of the requirements of the degree of

Master of Science in Mathematics with concentration in Statistics

UNIVERSITY OF NORTH FLORIDA

COLLEGE OF ARTS AND SCIENCE

December, 2015

This Thesis titled Comparing Group Means when Nonresponse Rates Differ is approved:

Dr. Donna Mohr

Dr. Ping Sa

Dr. Peter Wludyka

Accepted for the Department of Mathematics and Statistics:

Dr. Richard Patterson

Accepted for the College of Arts and Sciences:

Dr. Barbara Hetrick

Accepted for the University:

Dr. John Kantner
Dean of the Graduate School

DEDICATION

I dedicate this work to my husband Kris and my family, who have been unconditionally supportive of me throughout my career journey.

ACKNOWLEDGMENTS

I would like to especially thank my advisor Dr. Mohr for her constant support throughout this past year with my thesis. Her expertise and guidance made it possible for me to complete this project. I would also like to thank my committee members, Dr. Sa and Dr. Wludyka, for their valuable input. The three of them have been great mentors as well as important influences throughout my graduate studies. I would also like to thank Dr. Zhan, who encouraged me to join the statistics program, and Dr. Hochwald, who encouraged me to pursue a thesis and was so accommodating with me through the program.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

ABSTRACT

Missing data bias results if adjustments are not made accordingly. This thesis addresses this issue by exploring a scenario where data is missing at random depending on a covariate *x*. Four methods for comparing groups while adjusting for missingness are explored by conducting simulations: independent samples *t*-test with predicted mean stratification, independent samples *t*-test with response propensity stratification, independent samples *t*-test with response propensity weighting, and an analysis of covariance. Results show that independent samples *t*-test with response propensity weighting and analysis of covariance can appropriately adjust for bias. ANCOVA is the stronger method when the ANCOVA assumptions are met. When the ANCOVA assumptions are not met, a *t*-test with inverse response propensity score weighting is the superior method.

CHAPTER 1

THE PROBLEM OF MISSING DATA

1.1 Why is it Important to Treat the Missing Data?

This paper will discuss ways to handle missing data. The following example is given to illustrate how missing data can bias results. Throughout the paper, this example will be discussed in reference to different methods to analyze the data in order to reduce bias.

1.1.1 Motivational example

In a hypothetical study, a professor wants to compare the final exam scores of two sections of the same class that she teaches. At the end of the semester, she collects the scores from both sections. After collecting the scores, she notices that the mean *FinalExam* score of *Section 1* is 73% while in *Section 2* the *FinalExam* mean is 84%. After examining this using an independent samples *t*-test she found that $t = 2.1057, p = .0429$. The initial conclusion would be that the second section had higher scores than the first. However, after an in-depth look at the data, she notices that in *Section 2*, there were several students that dropped out. In *Section 1*, each student remained enrolled. She goes through her records, and now looking at the full data, she notices that students who scored low on *Exam1* in *Section 2* dropped the class before the *FinalExam*. Intuitively, it seems that the absence of weaker students may be the reason why the mean *FinalExam* score in *Section 2* is higher.

The hypothetical data set is shown below.

Table 1.1. Exams Scores for Students that Remained Enrolled

	Section 1		Section 2	
	Exam 1	Final Exam	Exam 1	Final Exam
	78	76	59	56
	51	46	69	-
	70	70	90	95
	95	86	71	-
	71	57	100	95
	95	92	94	93
	69	66	85	89
	87	82	57	47
	96	91	88	82
	72	79	87	86
	73	67	85	88
	95	86	90	-
	99	92	99	97
	87	80	71	-
	80	72	64	-
	88	84	52	-
	81	74	86	84
	62	36	91	89
	78	72	86	80
	72	64	70	-
	70	67	89	94
Mean	79.47619	73.28571	80.14286	83.92857
SD	12.65156	14.57101	14.11838	14.76724

Table 1.1 shows the scores that the students in both classes received on their *FinalExam*. As can be seen, some students in *Section 2* dropped the course and their scores are missing. Notice that in *Section 2*, the mean is significantly higher than in *Section 1*, but in *Section 2* there is a high percentage of dropouts ($t = 2.1057, p = .0429$).

Table 1.2 Exams Scores for All Students Adding the Scores Had All Students Remained Enrolled

	Section 1		Section 2	
	Exam 1	Final Exam	Exam 1	Final Exam
	78	76	59	56
	51	46	69	64
	70	70	90	95
	95	86	71	64
	71	57	100	95
	95	92	94	93
	69	66	85	89
	87	82	57	47
	96	91	88	82
	72	79	87	86
	73	67	85	88
	95	86	90	83
	99	92	99	97
	87	80	71	57
	80	72	64	62
	88	84	52	36
	81	74	86	84
	62	36	91	89
	78	72	86	80
	72	64	70	66
	70	67	89	94
Mean	79.47619	73.28571	80.14286	76.52381
SD	12.65156	14.57101	14.11838	17.7528

Table 1.2 shows all the scores, including the scores in *Section 2* as if those students had not dropped the course. As depicted, the means of both classes are similar ($t = .6461, p = .5219$).

Below is a scatterplot for the *FinalExam* grades as a function of the *Exam1* grades. “1” is for the students in *Section 1*, “2” is for students in *Section 2*, and “0” is for the grades that the students in *Section 2* who dropped the course would have received for the *FinalExam* if they had stayed.

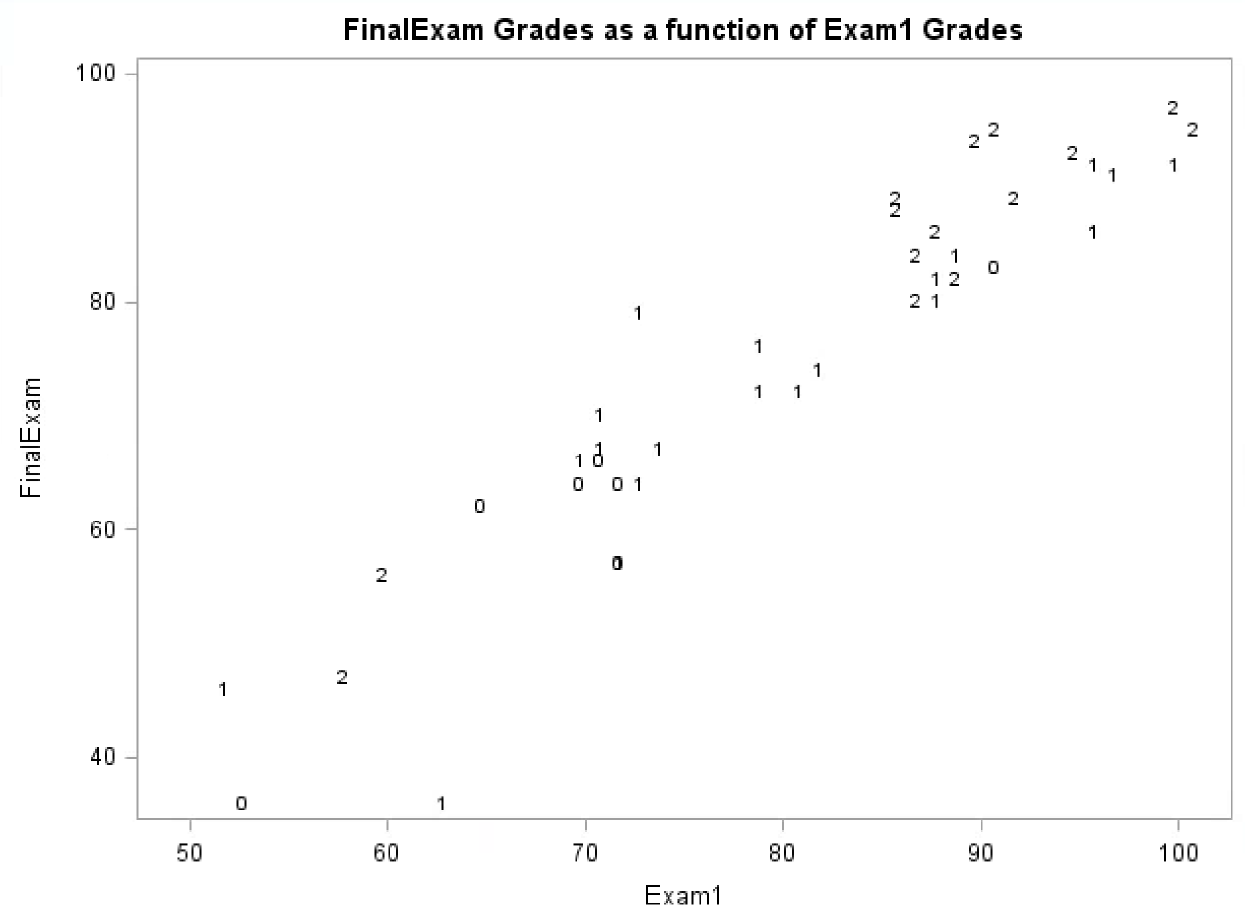

Figure 1.1 FinalExam Grades as a Function of Exam1 Grades

As can be seen in the illustration above, doing the analysis ignoring the missing observations can bias the result of a study.

Throughout this paper the following notation will be used:

$$X = Exam1\ scores$$

$$Y = FinalExam\ scores$$

In each group, *X* and *Y* have a joint distribution:

$$F_j(X, Y),$$

where:

$$j = \begin{cases} 1, & \text{Section 1} \\ 2, & \text{Section 2} \end{cases}$$

The comparison of interest is:

$$E_1(Y) - E_2(Y),$$

where *FinalExam* scores between the two sections are compared.

Notice that *Y* = *FinalExam* score is being treated as a value that exists even though the student dropped the course. In this paper, the missing values will be treated as though they are true values that exist but are unknown. This is similar to how missing data is treated in various settings, such as in clinical trials and social sciences. Similarly, they consider that every subject in the study has a “true” value. If the value is missing, then it may be treated as a latent value and needs to be estimated (Schafer and Graham 2002).

1.2 Missing Data Mechanisms

Let *Y* be the variable of interest.

According to Rubin (1976), and Little and Rubin (2002), there are three missing data mechanisms:

1.2.1 Missing at Random (MAR)

Missingness is related to an observed variable other than Y, the dependent variable that is missing. For instance, a researcher may be interested in testing $Y = income\ level$ of certain families within a town. It is noted that several Hispanic families are not reporting their income level. Thus, missingness is related to the observable variable of ethnicity, which the researcher has been able to collect, rather than related to Y directly. However, this does not mean Y is not affected by MAR data. If the Hispanics in the town have a lower income level than the rest of the population, then Y is going to be over-estimated. Therefore, this needs to be accounted for when conducting the analysis of data.

1.2.2 Missing Completely at Random (MCAR)

Missingness is not related to Y in any facet and it is not related to other variables either because it is completely random. For instance, in the study described above there may be missing data due to the researcher losing a file with part of the data, or data on certain households that could not be collected due to reasons such as the family members being out on vacations, working, or in the hospital when the researcher is trying to collect the data. In these cases, the data is MCAR because it is not related to Y nor other variables. MCAR data does not bias the results.

1.2.3 Missing Not at Random (MNAR)

Missing data is related to Y, even after adjusting for other variables. In the example described above, if individuals with lower incomes failed to report it, the researcher would have MNAR data. MNAR is the most dangerous out of the three missing data mechanisms, since it biases the results of the study.

MAR and MCAR are considered ignorable missing data because the researcher has sufficient information to adjust for the bias, while MNAR is considered non-ignorable missing data because it biases the results and there is not enough information to adjust for it. Ideally the researcher prefers MCAR or MAR. When conducting a study, the researcher can prevent MNAR data by collecting information on other variables, thus converting MNAR to MAR.

In the motivational example that this thesis uses, MAR missingness is considered, since students drop out of the class based on scores received previously, which the professor has collected. It will be considered that each student has taken *Exam1*. Therefore, the professor has complete data for *Exam1* scores from each student while there are missing scores from *FinalExam*. Students that did poorly in *Exam1* had higher chances of dropping the course. Thus, missingness depends on *Exam1*, which is a variable that has been collected. Therefore, data is missing at random (MAR).

1.3 Use of *T*-test to Analyze Data without Missingness Adjustment

1.3.1 Simulation: No Missingness Adjustment

In this section, a simulation is conducted using MAR data. The data is then analyzed using an independent samples equal variance *t*-test ignoring the missing data testing whether $E_1(Y) = E_2(Y)$. All simulations in this thesis were done using SAS Version 9.2.

Let:

$$n = number\ of\ students\ in\ each\ section$$

Since in a regular lecture hall at the University of North Florida there are 140 students in the class, the simulation will have $n = 140$.

Thus:

$$X = Exam1\ Score$$

where:

$$X{\sim}Beta(4, .8) * 100$$

Notice that in both sections, the distribution of the *Exam1* score is the same.

Let:

$$Y = FinalExam\ Score,$$

where:

$$Y = X - 13 + \epsilon,$$

$$\epsilon = 5 * Z,$$

and

$$Z \sim N(0,1).$$

This means that in both sections, students' *FinalExam* scores are an average of 13 points lower than their *Exam1* scores.

Therefore, both classes have the same *Exam1* score and *FinalExam* score distributions.

The following graph shows an example of the distribution of *X*:

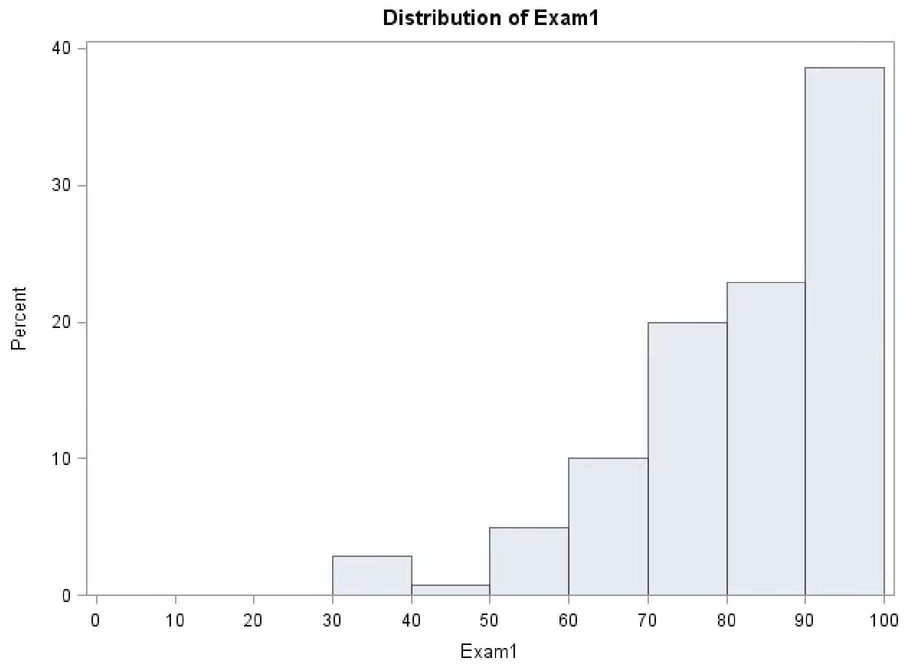

Figure 1.2 Distribution of Exam1 Grades

$$E(X) = 83.333$$

$$\sigma(X) = 15.4746$$

The drop rate is altered for each class. The probability $p_D(x)$ of dropping given the *Exam1* score is:

$$p_D(x) = \frac{e^{\beta_0+\beta_1*x}}{1+e^{\beta_0+\beta_1*x}}.$$

Various combinations of (β_0, β_1) are considered such that the probability of dropping $p_D(x)$ for a student with a score of $x = 95$ (grade 95 in *Exam1*) is 5%. Thus, both classes have the same distribution for X and Y, and the only difference is their drop rate $p_D(x)$ (with higher chances of dropping the course for students with lower values of x). Notice that since the data is being generated assuming $E_1(Y) = E_2(Y)$, and at $\alpha = .05$, 5% of the simulations should find a significant difference between the two groups.

A simulation with 1,000 replicates was conducted for each combination of (β_0, β_1) between the two sections, and the percentage of false positives (significant *p*-values) were recorded in the table below:

Table 1.3 Independent Samples with Equal Variance T-test without Missingness Adjustment for Different Drop Rates: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	4.8%								
	2	7.9%	3.9%							
	3	12.9%	6.9%	5.1%						
	4	22.7%	11.9%	6.9%	5.1%					
	5	37.4%	22.9%	13.1%	8.6%	4.5%				
	6	46.5%	31.6%	20.0%	10.1%	6.5%	4.4%			
	7	56.3%	41.5%	28.3%	15.1%	9.1%	5.8%	4.6%		
	8	63.8%	51.2%	33.0%	20.5%	11.1%	6.8%	5.3%	6.1%	
	9	67.3%	53.8%	38.9%	26.1%	16.6%	9.6%	5.6%	5.1%	5.1%

As can be seen from the table above, the greater that the dropout patterns differ between the two sections, the higher number of false positives were obtained, meaning that the *t*-test detects a difference in the mean of the *FinalExam* scores between the two sections when in reality there is none. Notice that along the diagonal, the percentage of false positives is close to 5%, which is what it is expected to be at $\alpha = .05$, given the fact that both sections have the same *FinalExam* distribution and the same dropout pattern.

To further illustrate the effect of the difference in drop rates, the following graph shows the drop rate for *Section 1* ($\beta_0 = 6.055561, \beta_1 = -.09$) in grey and *Section 2* ($\beta_0 = 4.055561, \beta_1 = -.07$) in black. As shown in the above table, 12.90% of the simulations detected a difference between the means when in reality there was not one.

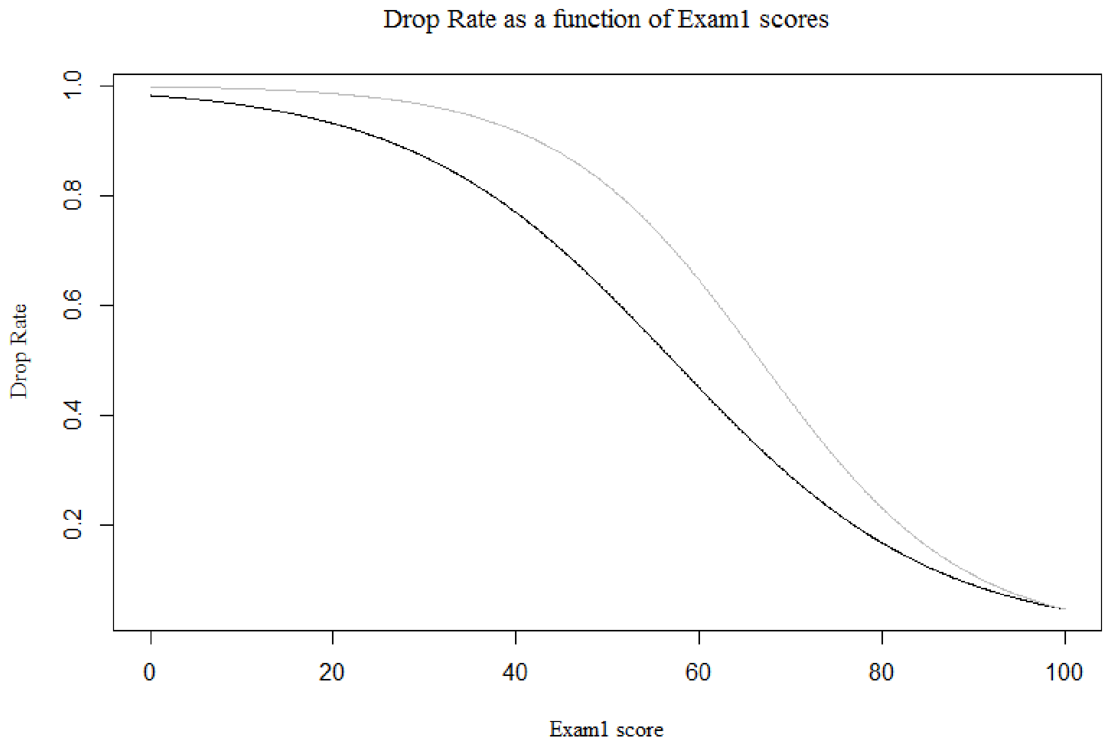

Figure 1.3 Drop Rate as a Function of Exam1 Grades

As can be seen in the graph above, the grey line (representing *Section 1*) is above the black line (representing *Section 2*) through the graph, meaning that students in *Section 1* have a higher chance of dropping than in *Section 2*. Even though the drop rates are only slightly different, the Type I error rate is inflated.

1.4 Conclusion

Without missingness adjustment, the analysis of data is biased for MAR data. The following chapters will explore three methods discussed by Little (1986) and the use of analysis of covariance. Conducting simulations, I will first explore the Type I error rate of each method. If the Type I error rate is close to 5% at $\alpha = .05$, then I will proceed to conduct further simulations to test the power of the method.

CHAPTER 2

LITTLE'S METHODS FOR MISSINGNESS ADJUSTMENT

2.1 Overview of Little's Methods for Adjusting for Missingness

In his article "Survey Nonresponse Adjustments for Estimates of Means," Little (1986) discusses the risk of analyzing data when observations are missing and addresses methods to adjust for missingness.

Little defines the respondents' mean as the mean of the observed responses without making any adjustments for missingness:

$$\bar{y}_R = \sum_{i=1}^{n_R} \frac{y_i}{n_R}$$

where

$$i = 1,2, \dots n_R$$

$$R = respondents$$

$$n_R = number\ of\ respondents.$$

In the motivational example given above, the number of respondents is the number of students who remained enrolled and took the final exam. The respondents' mean for the *FinalExam* scores in *Section 2* is the mean that the professor initially calculated where she did not account for the missing observations:

$$\bar{y}_R = 83.9286$$

As was seen in the motivational example, the respondents' mean has a potential for bias.

Little proposes a method to adjust for missingness. He recommends that observations be classified into *C* adjustment cells defined by a covariate *x*. Then, the adjusted mean is calculated by:

$$\bar{y}_A = \sum_{c=1}^{C} p_c * \bar{y}_{cR},$$

where

$$c = 1,2, \dots, C$$

$$p_c = \frac{n_c}{n}.$$

In the motivational example, for *Section 2*, three adjustment cells can be created by grouping the *Exam1* scores into terciles. Let,

$$c = \begin{cases} 1, & \text{if } x \leq 70 \\ 2, & \text{if } 70 < x \leq 88 \\ 3, & \text{if } 88 < x \end{cases}$$

Then,

Table 2.1 Information Needed for Adjusted Mean

	c = 1	c = 2	c = 3
y_{cR}	47+56	89+88+84+80+86+82	94+95+89+93+97+95
$\bar{y}_{cR}$	51.5	84.83	93.83
n_c	6	8	7
n	21	21	21
p_c	6/21	8/21	7/21

$$\bar{y}_A = \sum_{c=1}^{3} p_c * \bar{y}_{cR} = 78.3095$$

Even though this result is not the true value of $\bar{y}$, it is closer than $\bar{y}_R$ is. Thus, if the adjustment cells are appropriately created, this method can reduce the bias when calculating the mean.

The variance may be calculated by:

$$Var(\bar{y}_A) = \sum_{c=1}^{C} \frac{p_c^2 S_{cR}^2}{n_{cR}}.$$

The bias for $\bar{y}_A$ is given by:

$$Bias(\bar{y}_A) = \sum_{c=1}^{C} \left(p_c - \pi_c\right) * E(y_{cR}) + \sum_{c=1}^{C} \pi_c * \left(E\left(y_{cR}\right) - E\left(y_c\right)\right).$$

Notice that for the first term, as the sample size increases, $p_c \to \pi_c$. Thus,

$$\lim_{n \to \infty} \sum_{c=1}^{C} \left(p_c - \pi_c\right) * E(y_{cR}) = 0$$

However, the second term does not go to 0 as n increases, since respondents' and nonrespondents' distribution of *Y* may differ. Therefore, as $n \to \infty$, $E(y_{cR}) \neq E(y_c)$. The adjusted mean $\bar{y}_A$ has zero bias if $E(y_{cR}) = E(y_c)$, that is, if the mean of *Y* is the same for respondents and nonrespondents given stratifier *c*.

Let:

$$r = \begin{cases} 1, & \text{if the i}^{\text{th}}\text{observation was collected (student remained enrolled)} \\ 0, & \text{if the i}^{\text{th}}\text{observation was not collected (student dropped)} \end{cases}$$

Then, cells should be created such that *y* and *r* are independent given *c*. Two potential stratifiers are:

Predicted Mean Stratification: The adjustment cells *c* are created based on the predicted value of *y*. This can be done by modeling the distribution of *y* given *x*. Then, $\hat{y}(x)$ is the predicted mean of *y* given *x*. Therefore, $\hat{y}(x)$ is grouped into *C* intervals. Even though it is unlikely that *y* and *r* are completely independent within each cell, in each interval their relationship will be weaker, and therefore bias will be reduced. In the motivational example, as was shown above, these adjustment cells may be formed according to the students' *Exam1* scores.

Response Propensity Stratification: The adjustment cells *c* are created based on the response propensity. Let $p_R(x) = pr(r = 1|x)$. Then, $p_R(x)$ can be estimated by $\hat{p}_R(x)$ from the logistic regression of the response indicator *r* on *x*, and adjustment cells are formed by grouping $\hat{p}_R(x)$. Similarly to the groups created by the predicted mean stratification, in each adjustment cell the relationship between *y* and *r* will be weaker, thus reducing the bias.

Weighting by the Inverse of the Response Propensity Score: A third method that Little mentions in his paper is using the response propensity score, but rather than forming adjustment cells, each y for the respondents is weighted by the inverse of $\hat{p}_R(x)$. Little reports that he prefers the previous two methods over this one, because for extremely low values of $\hat{p}_R(x)$, the variance becomes inflated.

The following two chapters will evaluate these three methods for adjusting for missingness based on predicted mean stratification, response propensity stratification, and weighting by the inverse of the response propensity score.

CHAPTER 3

MISSINGNESS ADJUSTMENT BASED ON PREDICTED MEAN

3.1 Use of *T*-Test to Analyze Data Using Predicted Mean Stratification Missingness Adjustment

3.1.1 Generation of Data

In the following set of simulations, data is generated similarly to the data in Chapter 1 (Section 1.3). Therefore, both classes have the same *Exam1* score distribution and *FinalExam* score distribution. The *FinalExam* score means are compared between the two sections using an independent samples *t*-test assuming unequal variances using $\bar{y}_A$ and $Var(\bar{y}_A)$ as described by Little, and using Satterthwaite's approximation for degrees of freedom:

$$df = \frac{\left(Var(\bar{y}_1) + Var(\bar{y}_2)\right)^2}{\sum_{j=1}^{2}\sum_{c=1}^{3}\frac{\left(\frac{p_{jc}^2 * s_{jcR}^2}{n_{jcR}}\right)}{n_{jcR} - 1}}$$

The null hypothesis is $E_1(Y) = E_2(Y)$.

3.1.2 Simulation 1. Predicted Mean Stratification: Type I Error Rate

In the first set of simulations, I stratified according to the predicted mean. The *FinalExam* scores are related to *Exam1* scores. Therefore, I used *x* as the covariate for stratifying by the predicted mean.

The data was split into adjustment cells such that the relationship between the drop rate and the *FinalExam* grades is minimally dependent given *Exam1* grades. The way this was conducted was by splitting the data into terciles according to *Exam1*. Thus, students whose *Exam1* scores fell in the first tercile were in adjustment cell 1, students in the middle tercile were in adjustment cell 2, and students in the upper tercile were in adjustment cell 3.

I ran simulations using Little's formula and comparing the group means with an independent samples unequal variance *t*-test at different drop rates. If any adjustment cell had one or fewer observations, then the mean and variance of this adjustment cell could not be calculated; therefore, this replicate had to be discarded. The simulations were created such that there is no difference between *Section 1* and *Section 2* with respect to either *Exam1* or *FinalExam*. Therefore, 5% of the *p*-values should be significant at $\alpha = .05$, regardless of the drop rate. Below are the results with the percentage of significant *p*-values out of 1,000 replicates. If replicates were discarded due to having one or fewer observations in a cell, the number of discarded replicates is reported after the percentage of significant *p*-values:

Table 3.1 Independent Samples with Unequal Variance T-test using Predicted Mean Stratification for Missingness Adjustment: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	23.6%								
	2	26.1%	23.2%							
	3	28.2%	24.4%	24.3%						
	4	31.7%	27.9%	25.9%	24.0%					
	5	35.3%	31.0%	28.6%	27.6%	26.2%				
	6	37.3%	34.2%	31.0%	25.4%	26.8%	26.0%			
	7	39.5%	36.8%	33.2%	28.4%	28.9%	23.4%	27.0%		
	8	44.5%	40.6%	34.3%	33.0%	26.7%	25.8%	25.9%	27.6%	
	9	42.9%	39.4%	34.6%	33.9%	28.5%	26.5%	25.9%	23.7%	24.7%

As can be seen from the simulations, the Type I error rate is still noticeably high. Notice, however, that for extremely different drop patterns the Type I error rate is slightly lower than when conducting the simulations without any adjustments for missing data. This means that when the drop rates are drastically different, using this method gives slightly improved results over not adjusting for missingness at all. Something surprising was that along the diagonal, where the dropout patterns are similar for both sections, the Type I error rate is higher than in the previous set of simulations where no adjustments were done. I decided to investigate further, to see the origin of this.

3.1.3 Simulation 2. Predicted Mean Stratification: Normally Distributed and Fixed Cells.

Since an assumption for a *t*-test is that the data has to be normally distributed, I decided to investigate whether the lack of normality within adjustment cells is what is causing this high Type I error rate along the diagonal. Therefore, I created an unrealistic scenario where each adjustment cell has a normal distribution. Let,

$$X|c = 1 \sim N(50, 10)$$

$$X|c = 2 \sim N(75, 10)$$

$$X|c = 3 \sim N(85, 10)$$

The groups were fixed in advance such that each group was normally distributed. However, *y* and *r* are more strongly associated than in the previous simulation, since the observations are not being ranked and divided into terciles.

Below are the results. Notice that on the left of the slash bar are the percent of significant *p*-values and on the right of the bar are the number of replicates that had to be ignored due to having at least one cell with one or fewer observations:

Table 3.2 Independent Samples with Unequal Variance T-test using Predicted Mean Stratification for Missingness Adjustment: Normally Distributed and Fixed Cells

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	45.10	40.10	34.40	28.40	22.62	17.24	12.65	9.25	6.73
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	5.2%								
	2	6.3%	4.6%							
	3	10.9%	6.7%	5.1%						
	4	19.0% / 1	10.3%	5.7%	4.7%					
	5	27.3%	17.6%	10.3%	6.0%	4.2%				
	6	33.5%	28.9%	16.2%	9.5%	5.7%	5.9%			
	7	43.3%	31.1%	23.7%	13.2%	7.6%	4.5%	5.0%		
	8	47.9%	37.6%	25.1%	14.8%	10.0%	7.0%	4.3%	4.2%	
	9	49.6%	38.4%	27.1%	17.6%	9.9%	6.9%	4.9%	6.9%	5.3%

Now it can be seen that along the diagonal, the Type I error rate is as expected (around 5%). This indicates that a possible explanation for the extremely high Type I error rate in the previous simulation along the diagonal was due to the lack of normality within the adjustment cells or the way the cells were formed based on the observed covariate. Therefore, if each adjustment cell has a normal distribution, when the two sections have similar expected grades and similar drop rates, the *t*-test is able to correctly detect that the two groups are not different. However, as the drop rates differ, the Type I error becomes higher. Additionally, one replicate in one of the simulations had to be dropped due to having an adjustment cell with one or fewer observations. Thus, having a normal distribution within each cell only gives better results if the drop rate patterns are the same between the two sections.

3.1.4 Simulation 3. Predicted Mean Stratification: Normally Distributed and Random Cells.

I decided to re-do the unrealistic simulation above generating random data coming from Normal distributions with means 50, 75, and 85, and standard deviations 10, but instead of fixing the groups, I decided to rank the data and group the data in terciles of the covariate *x*.

Below are the results of the simulations:

Table 3.3 Independent Samples with Unequal Variance T-test using Predicted Mean Stratification for Missingness Adjustment: Normally Distributed and Random Cells

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	45.10	40.10	34.40	28.40	22.62	17.24	12.65	9.25	6.73
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	9.6% / 1								
	2	9.2% / 2	10.9%							
	3	11.5% / 3	10.1%	9.8%						
	4	14.7% / 1	12.8%	10.6%	11.4%					
	5	17.8% / 1	15.7%	12.7%	11.9%	11.5%				
	6	23.4% / 2	20.1%	15.8%	13.7%	13.8%	13.6%			
	7	28.1% / 1	22.9%	21.6%	17.6%	13.4%	10.7%	14.0%		
	8	29.3% / 1	26.0%	21.4%	16.4%	16.7%	14.0%	13.0%	12.2%	
	9	29.7%	25.4%	22.9%	17.1%	14.7%	13.8%	11.7%	14.5%	13.5%

The Type I error rate is still high, but it is lower than it was in previous simulations. Along the diagonal the Type I error rate was increased in comparison to the previous set of simulations with fixed normal groups. This is possibly due to the fact that now that the groups are not fixed with a normal distribution, the assumption of normality in the *t*-test is being violated again.

3.1.5 Simulation 4. Predicted Mean Stratification: Beta Distribution and Fixed Cells.

I conducted simulations in which each of the three groups had a non-normal distribution and compared the Type I error rates.

In the following set of simulations, the adjustment cells were fixed such that:

$$X|c = 1 \sim Beta(1, 4) * 40$$

$$X|c = 2 \sim 70 + Beta(4, 1) * 30$$

$$X|c = 3 \sim 40 + Beta(10, 10) * 30$$

Below are the results:

Table 3.4 Independent Samples with Unequal Variance T-test using Predicted Mean Stratification for Missingness Adjustment: Beta Distribution and Fixed Cells

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	67.76	63.40	57.87	51.00	42.49	32.58	22.18	13.80	8.25
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	7.6% / 631								
	2	6.9% / 523	5.8% / 314							
	3	16.2% / 396	9.7% / 186	5.9% / 64						
	4	26.3% / 398	17.0% / 182	6.3% / 31	4.2%					
	5	40.5% / 403	28.9% / 159	17.6% / 39	10.0% / 1	4.3%				
	6	45.3% / 395	38.8% / 150	30.2% / 33	19.6% / 1	8.8%	4.9%			
	7	53.1% / 392	43.9% / 164	42.6% / 31	36.3% / 1	20.9%	7.9%	5.9%		
	8	52.2% / 397	51.7% / 178	52.8% / 32	45.8% / 1	31.4%	14.9%	5.8%	5.3%	
	9	57.8% / 413	56.8% / 180	54.3% / 36	50.6% / 4	37.0%	16.7%	8.3%	5.4%	4.7%

Not only is the Type I error rate high, but also has numerous replicates that had to be discarded due to one of the adjustment cells having one or fewer observations. When adjustment cells have non-normal distribution and when there is an extreme drop rate, this method detects a difference between groups when there is none, or cannot analyze the data due to having an empty cell.

3.1.6 Simulation 5. Predicted Mean Stratification: Beta Distribution and Random Cells

In the next set of simulations I used the same distributions as above, but rather than fixing the groups, I ranked the data and grouped it based on terciles of *x*.

Below are the results:

Table 3.5 Independent Samples with Unequal Variance T-test using Predicted Mean Stratification for Missingness Adjustment: Beta Distribution and Random Cells

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop	67.76	63.40	57.87	51.00	42.49	32.58	22.18	13.80	8.25
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	13.3% / 955								
	2	7.3% / 918	5.9% / 778							
	3	16.0% / 844	10.2% / 606	9.1% / 293						
	4	19.6% / 811	13.5% / 542	9.7% / 153	9.7% / 19					
	5	28.3% / 813	18.8% / 515	16.1% / 174	13.7% / 18	10.2%				
	6	31.7% / 817	27.4% / 504	19.1% / 141	18.4% / 10	15.8%	14.1%			
	7	37.4% / 802	30.6% / 533	24.6% / 175	21.1% / 17	18.6%	14.3%	15.8%		
	8	38.4% / 797	31.2% / 529	28.0% / 156	24.7% / 9	20.5%	19.0%	15.8%	13.2%	
	9	35.9% / 816	36.1% / 504	28.3% / 170	26.0% / 8	23.2%	18.5%	16.6%	17.0%	14.8%

There was a high Type I error rate and excessive discarded replicates due to adjustment cells with one or fewer observations. Thus, this method is not appropriate when the cells have a non-normal distribution or when there is an extreme dropout rate.

3.2 Conclusion

In this chapter I conducted simulations where I generated data for two sections with similar *Y* (*FinalExam*) distributions but different drop rates. I compared the two sections with an independent samples unequal variance *t*-test adjusting for missingness using Little's predicted mean stratification for missing data adjustment using *X* (*Exam1*) as the covariate. Based on the simulations conducted above, it appears that if the dropout patterns are different between the two groups, this method may be an improvement over ignoring missing data as long as there are no cells with one or fewer observations. However, the Type I error rate is still excessively high, making this method inappropriate.

In the next chapter, I will explore missingness adjustment based on the response propensity.

CHAPTER 4

MISSINGNESS ADJUSTMENT BASED ON RESPONSE PROPENSITY

4.1 Use of *T*-Test to Analyze Data Using the Response Propensity to Adjust for Missingness

4.1.1 Simulation 1. Response Propensity Stratification: Type I Error Rate

In the next set of simulations, realistic random data is generated in the same manner as in the original simulations (Chapter 1, Section 1.3). In order to create groups based on the response propensity, let:

$$r = \begin{cases} 1, & \text{if student remains enrolled} \\ 0, & \text{if student drops out} \end{cases}$$

A logistic regression is used to model the probability $\hat{p}_R(x)$ of a student remaining enrolled as a function of their *Exam1* score *x*. Students are ranked based on their response propensity score $\hat{p}_R(x)$ for each section and grouped into terciles. Thus, students whose $\hat{p}_R(x)$ scores fell in the first tercile were in adjustment cell 1, students in the middle tercile were in adjustment cell 2, and students in the upper tercile were in adjustment cell 3. This simulation tested whether $E_1(Y) = E_2(Y)$.

Using Little's formula for means and standard deviations, simulations are conducted comparing the group means with an independent samples unequal variance *t*-test at different drop rates. If any adjustment cell had one or fewer observations, then the mean and variance of this adjustment cell could not be calculated and this replicate had to be discarded. Since the simulations were created such that there was no difference between *Section 1* and *Section 2*, 5% of the *p*-values should be significant at $\alpha = .05$, regardless of the drop rate. Below are the results with the percentage of significant *p*-values out of 1,000 replicates. If any replicate was discarded due to having one or fewer observations in a cell, the number of discarded replicates are reported after the percentage of significant *p*-values:

Table 4.1 Independent Samples with Unequal Variance T-test using Response Propensity Stratification for Missingness Adjustment: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	23.6%								
	2	26.1%	23.2%							
	3	28.2%	24.4%	24.3%						
	4	31.7%	27.9%	25.9%	24.0%					
	5	35.3%	31.0%	28.6%	27.6%	26.2%				
	6	37.3%	34.2%	31.0%	25.4%	26.8%	26.0%			
	7	39.5%	36.8%	33.2%	28.4%	28.9%	23.4%	27.0%		
	8	44.5%	40.6%	34.3%	33.0%	26.7%	25.83% / 1	25.9%	27.6%	
	9	42.9%	39.4%	34.6%	33.9%	28.5%	26.5%	25.9%	23.7%	24.7%

Even though the results seem better than analyzing the data without making adjustments for missing observations, the Type I error rate remains excessively high. Along the diagonal where even though the dropout patterns are the same between the two classes the Type I error rate is higher than in the simulations where no missingness adjustments were made. Thus, this method seems to only show an improvement over a *t*-test without adjusting for missingness if the dropout patterns are substantially different between the two sections, but not if the dropout patterns are similar.

4.1.2 Simulation 2. Inverse Response Propensity Weighting without Stratification: Type I Error Rate

The following set of simulations uses the same data that was used in previous simulations, and an independent samples *t*-test is conducted based on the method that Little describes where observations are weighted by the inverse of the response propensity score. Below are the results:

Table 4.2 Independent Samples with Unequal Variance T-test using Inverse Response Propensity without Stratification for Missingness Adjustment: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	18.6%								
	2	15.9%	13.4%							
	3	16.3%	11.4%	9.0%						
	4	13.3%	9.1%	7.7%	5.7%					
	5	13.4%	11.1%	8.0%	7.6%	4.8%				
	6	13.0%	11.3%	7.2%	6.4%	5.6%	4.4%			
	7	13.7%	9.4%	8.1%	5.2%	5.1%	3.9%	4.7%		
	8	12.4%	9.0%	6.3%	5.5%	4.8%	4.1%	3.8%	5.8%	
	9	13.1%	8.9%	9.0%	6.4%	4.2%	4.1%	4.1%	3.2%	4.6%

In these simulations, the Type I error rate seems to have diminished in comparison with the other methods. It appears that the largest number of false positives is when β_1 is furthest away from zero. This makes intuitive sense, since β_1 is the part of the dropout that depends on *x*. Therefore, for extremely low values of *x*, the dropout is extremely high, and it is possible that the lowest values of *x* are not being represented by observed values of *y*.

It appears that this method is giving improved results in comparison to previous methods. The Type I error rate is higher than 5%, but it is not as excessively high as when using previous methods. The next step is to examine the power for this method. It is possible, as Little explains in his paper, that the reason why the *t*-test is not detecting a difference between means is due to extremely large variance rather than appropriately correcting for bias. Therefore, the next simulations will show whether this method has an appropriate power in detecting a difference when there is a real difference between the two sections.

4.1.3 Simulation 3. Inverse Response Propensity Weighting without Stratification with Different E(Y) between the Sections: Power

In the following simulation, the *Exam1* scores have the same distribution as in previous simulations. However, the *FinalExam* scores will be different between the two sections:

Section 1: $Y = X - 13 + \epsilon$

Section 2: $Y = X - 8 + \epsilon$

where:

$$\epsilon \sim Z * 5, and$$

$$Z \sim N(0,1).$$

Thus,

$$E_1(Y) - E_2(Y) = 5.$$

Simulations were conducted to see whether conducting a *t*-test weighting by the inverse of the response propensity score is able to detect this difference. First, a simulation with 1,000 replicates was run with no drop rate, and an independent samples *t*-test with equal variances was conducted. In this simulation, 74.70% of the times the *t*-test was able to detect the difference between the two sections. Therefore, 74.70% power is an appropriate goal.

The following table shows the percentage of the times that the *t*-test with the inverse $\hat{p}_R(x)$ weighting adjustment discussed detects the difference between the two sections.

Table 4.3 Independent Samples with Unequal Variance T-test using Inverse Response Propensity Weighting without Stratification for Missingness Adjustment with Different E(Y) between the Sections: Power

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	65.1%								
	2	56.5%	66.1%							
	3	58.2%	61.2%	63.0%						
	4	55.8%	59.6%	62.1%	63.1%					
	5	52.3%	58.5%	60.6%	64.4%	67.1%				
	6	54.5%	60.7%	62.4%	67.6%	67.1%	67.9%			
	7	55.0%	60.5%	62.1%	66.8%	67.6%	66.5%	70.9%		
	8	54.8%	57.6%	65.0%	65.9%	67.5%	67.7%	67.9%	70.1%	
	9	56.5%	60.4%	65.6%	65.4%	67.5%	68.5%	69.5%	68.6%	70.8%

The results above show an acceptable level of power, between 52% and 70%, depending on the level of drop rate. Thus, it would seem that this method has an appropriate power and an improved Type I error rate in comparison to previous methods or not adjusting for missingness. As shown on the table, the power increases as β_1 gets closer to zero. This is consistent with results from previous simulations, where a β_1 that is closer to zero brings less biased results. Therefore, for extremely low values of *x*, the dropout is extremely high, and it is possible that the lowest values of *x* are not being represented by observed values of *y*.

4.1.4 Simulation 4. Inverse Response Propensity Weighting without Stratification with Different E(Y) between the Sections: Power

In the following simulation, the *Exam1* scores have the same distribution as in previous simulations. However, the *FinalExam* scores are different between the two classes:

Section 1: $Y = X - 13 + \epsilon$

Section 2: $Y = .85 * X + 4.5 + \epsilon$

where:

$$\epsilon \sim Z * 5$$

$$Z \sim N(0,1)$$

Then:

$$E_1(Y) - E_2(Y) = 5$$

I now conduct simulations to see whether conducting a t-test weighting by the inverse of the response propensity score is able to detect this difference. A simulation with 1,000 replicates was done without missing observations, and an independent samples with equal variances t-test was conducted. 81.00% of the times the t-test was able to detect the difference between the two sections.

The following table shows the percentage of the times that the t-test with inverse response propensity weighting detects the difference between the two sections according to each drop rate.

Table 4.4 Independent Samples with Unequal Variance T-test using Inverse Response Propensity Weighting without Stratification for Missingness Adjustment with Different E(Y) between the Sections: Power

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	67.9%								
	2	62.2%	69.6%							
	3	63.6%	66.1%	68.5%						
	4	60.5%	66.0%	67.3%	68.0%					
	5	57.3%	63.7%	64.9%	70.0%	72.9%				
	6	60.6%	64.8%	69.2%	73.9%	72.2%	74.3%			
	7	61.3%	66.5%	67.5%	74.6%	73.1%	73.5%	76.5%		
	8	61.0%	64.3%	70.7%	71.8%	74.4%	75.3%	74.9%	76.4%	
	9	62.1%	66.4%	71.3%	71.6%	72.1%	75.7%	76.6%	74.2%	77.3%

As can be seen in the above table, this method is showing an appropriate amount of power. Between 57% and 77% of the simulations detected a difference between the two sections when a difference was present.

4.2 Conclusion

In this chapter, I first conducted simulations where I generated data for two sections with similar *Y* (*FinalExam*) distributions but differing drop rates. I compared the two sections with an independent samples unequal variance *t*-test adjusting for missingness using Little's inverse response propensity stratification for missing data adjustment, using $\hat{p}_R(x)$ as the covariate. To estimate $\hat{p}_R(x)$, I used a logistic regression to calculate the probability of *r* (whether the student took the final exam or not) based on *x* (the *Exam1* score). I created adjustment cells based on terciles of $\hat{p}_R(x)$, and implemented Little's method this way using different drop rates for each section. This method showed an excessive amount of Type I error rate.

In the second part of this chapter, I used the estimated response propensity to adjust for missingness, but I weighted the observations by the inverse of $\hat{p}_R(x)$ rather than using stratification. The simulations showed a reduced amount of Type I error rates. Therefore, I proceeded to evaluate whether this method has an appropriate level of power. I generated new data where the distribution of *Y* is different for each section, and I conducted simulations where I compared the two sections for different drop rates. The analysis using this method was able to detect the difference between the groups in 52% to 77% of the simulations depending on the dropout patterns. A *t*-test without missing data was able to detect the difference between the two sections in 74% to 81% of the simulations. This indicates that this method can accurately compare the difference between the two sections when the drop rates differ. This method seems considerably more accurate when β_1 (the dropout associated with the covariate *x*) is closer to zero. This is expected, since when *x* is extremely low and β_1 is further from zero, there is an excessive number of missing observations; therefore, there are no *y* values observed when *x* is extremely low.

Out of the methods discussed, it appears that using an independent samples *t*-test with equal variances weighting by the inverse of the response propensity is giving the best results. It has the lowest percentage of Type I error, and it seems to have an appropriate amount of power to detect a difference between the two sections when a difference is present.

CHAPTER 5

USE OF ANALYSIS OF COVARIANCE TO ADJUST FOR MISSINGNESS

5.1 Use of Analysis of Covariance to Estimate Group Differences

5.1.1 Analysis of Covariance

In clinical trials, researchers use an analysis of covariance sometimes to compare the difference in response to treatments between the groups. In the context of this paper, notice that rather than comparing the *FinalExam* score means between the two sections, ANCOVA would compare the *FinalExam* score mean given *Exam1* score:

$$E_1(Y|X = x_0) = E_2(Y|X = x_0)$$

ANCOVA assumes that the slopes the regression lines for both sections are parallel and both sections have the same expected baseline (Kutner et al. 2005). Then:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \epsilon_{ij}$$

where

$$\beta_{11} = \beta_{12},$$

$$E_1(X) = E_2(X),$$

and

$$\epsilon_{ij} \sim N(0, \sigma^2).$$

Thus,

$$E_1(Y) - E_2(Y) = E_1(E_1(Y|X)) - E_2(E_2(Y|X))$$

$$= E_1(\beta_{01} + \beta_{11}X) - E_2(\beta_{02} + \beta_{12}X)$$

$$= \beta_{01} + \beta_{11}E_1(X) - \beta_{02} - \beta_{12}E_2(X)$$

$$= \beta_{01} - \beta_{02}$$

In his paper “On Efficiency of Constrained Longitudinal Data Analysis versus Longitudinal Analysis of Covariance”, Lu (2010) discusses randomized clinical trials where subjects may drop out after a few visits, and he compares constrained longitudinal data analysis (cLDA) and longitudinal analysis of covariance when analyzing these data sets. However, since this thesis is concerned with ANCOVA only, cLDA will not be discussed.

In his Web Appendix, Lu shows that if the probability of drop-out depends on the baseline value, then ANCOVA gives an unbiased estimate of the between-group differences. He assumes that the data is missing at random and the probability of missingness depends on the observed baseline and treatment group.

The estimated postbaseline mean in group *j* is:

$$\hat{Y}_j = \hat{\beta}_{0j} + \hat{\beta}_1\bar{X},$$

where $\bar{X}$ is the observed baseline mean for subjects included in the analysis from both groups. Thus, the estimated mean difference at postbaseline time points between groups is:

$$\hat{Y}_1 - \hat{Y}_2 = \hat{\beta}_{01} + \hat{\beta}_1\bar{X} - \hat{\beta}_{02} - \hat{\beta}_1\bar{X} = \hat{\beta}_{01} - \hat{\beta}_{02}.$$

which is an unbiased estimate of $\beta_{01} - \beta_{02}$.

Lu conducts simulations comparing two treatment groups with three models of missing data: in the first model, missingness is MAR and depends on the previously observed value. In the second model, missingness depends also on the treatment group. In the third model, intermittent MCAR missingness is generated. He furthermore generated two other scenarios where the third model is combined with the first and the second models. The simulations showed a Type I error rate between 5.0% and 5.3% and a power between 73.1% and 91.2%.

In the following section, I will examine the Type I error rate and power of ANCOVA using this paper's scenario and compare it to the *t*-test using inverse response propensity weighting.

5.1.2 Simulation 1. ANCOVA: Type I Error Rate

Data was generated similarly to previous simulations with:

$$Y = X - 13 + \epsilon$$

for both sections.

Then, $E_1(Y|X = x_0\ ) = E_2(Y|X = x_0)$ was tested using a regular ANCOVA in 1,000 replicates at each drop rates. Since both sections have the same expected conditional outcome, 5% of the *p*-values should be significant at $\alpha = .05$. The table below shows the Type I error rate for simulations with different drop rates.

Table 5.1. ANCOVA Assuming Equal E(Y) for both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	4.90%								
	2	4.30%	5.90%							
	3	5.40%	5.60%	5.60%						
	4	4.50%	5.60%	5.20%	6.10%					
	5	4.90%	6.60%	5.90%	6.30%	5.50%				
	6	5.10%	6.40%	4.90%	6.40%	5.20%	4.50%			
	7	5.30%	5.40%	5.90%	4.80%	5.90%	3.70%	4.70%		
	8	4.70%	5.20%	5.60%	5.20%	5.80%	4.70%	4.60%	4.50%	
	9	6.60%	5.50%	4.70%	5.20%	4.50%	4.70%	5.10%	4.90%	4.90%

The Type I error rate is around 5%, which is what is expected. Therefore, in the following set of simulations I will test the power of using ANCOVA to adjust for missing data.

5.1.3 Simulations 2. ANCOVA: Power

In this set of simulations, in order to test the power of ANCOVA, data was generated such that:

$$Y_1 = X - 13 + \epsilon$$

$$Y_2 = X - 8 + \epsilon$$

Simulations with 1,000 replicates using an ANCOVA to compare the two sections at each drop rate were conducted. Since both sections have a different $E_j(Y|X = x_0)$, it is expected that there will be a high percentage of significant p-values. Below are the results.

Table 5.2. ANCOVA Assuming Different E(Y) between the Sections: Power

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	100.0%								
	2	100.0%	100.0%							
	3	100.0%	100.0%	100.0%						
	4	100.0%	100.0%	100.0%	100.0%					
	5	100.0%	100.0%	100.0%	100.0%	100.0%				
	6	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%			
	7	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%		
	8	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
	9	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

The ANCOVA detected the difference in conditional means in 100% of the simulations, regardless of the drop rate. This method is showing appropriate Type I error rate and strong power.

5.1.4 Simulations 3. ANCOVA Using Inverse Propensity Weighting: Type I Error Rate

In the next set of simulations, a combination of Little’s inverse propensity weighting method and ANCOVA were used to see whether this combination would improve upon these two methods individually. An ANCOVA using the inverted propensity weighting method was used with:

$$Y = X - 13 + \epsilon$$

for both sections.

Below are the results.

Table 5.3. ANCOVA with Inverse Propensity Weighting Assuming Equal E(Y) for both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	7.80%								
	2	7.80%	8.50%							
	3	8.00%	5.70%	6.80%						
	4	6.10%	7.20%	6.10%	6.70%					
	5	7.40%	7.60%	6.50%	6.40%	5.60%				
	6	7.70%	7.80%	5.00%	7.00%	5.40%	4.40%			
	7	6.80%	7.00%	7.20%	5.00%	6.10%	3.80%	4.50%		
	8	5.90%	7.30%	5.80%	6.40%	5.90%	4.50%	4.70%	4.60%	
	9	8.50%	6.40%	5.50%	5.00%	4.50%	5.00%	5.40%	5.10%	4.80%

In using ANCOVA, the Type I error rate was slightly higher using the inverted propensity weighting than without it.

5.2 Conclusion

In this chapter, I first conducted simulations generating data for the two sections with similar *Y* (*FinalExam*) distributions but different drop rates. I compared the two sections using an Analysis of Covariance. The Type I error rate was close to 5%; in fact, the Type I error was lower using ANCOVA than using an independent samples *t*-test with inverse propensity weighting. I then proceeded to investigate the power of using ANCOVA to adjust for missingness, and this method had a higher power than the independent samples *t*-test with inverse propensity weighting. Using a combination of ANCOVA and inverse propensity weighting had a slightly higher Type I error rate; therefore, I did not proceed to investigate the power of this method, since it appears that ANCOVA without any weighting is a better method.

Analysis of covariance is based on the assumptions that β_1 and *E(X)* are equal in both sections. The next chapter will compare the performances of the ANCOVA and the independent samples *t*-test with inverse response propensity score weighting.

CHAPTER 6

ANCOVA LIMITATIONS AND INVERSE RESPONSE PROPENSITY SCORE WEIGHTING SOLUTIONS

6.1 Limitations of ANCOVA

The previous chapter showed that ANCOVA appears to be a more powerful test than independent samples *t*-test with inverse response propensity weighting for missingness adjustment. However, in the previous simulations it was assumed that the ANCOVA assumptions were met. In this chapter, the focus is on exploring what happens when the ANCOVA assumptions are not met, but ANCOVA is mistakenly used regardless.

ANCOVA is based on the following two assumptions:

(1) $\beta_{11} = \beta_{12}$, and

(2) $E_1(X) = E_2(X)$.

Thus,

$$E_1(Y) - E_2(Y) = (\beta_{01} - \beta_{11}E_1(X)) - (\beta_{02} - \beta_{12}E_2(X))$$

$$= (\beta_{01} - \beta_{02}) + (\beta_{11}E_1(X) - \beta_{12}E_2(X))$$

$$= \beta_{01} - \beta_{02},$$

and this is how it is possible to compare group means for Y using ANCOVA.

However, if either of the above assumptions are violated, then this comparison cannot take place.

(1) Assume that $\beta_{11} \neq \beta_{12}$.

Then,

$$E_1(Y|X = x_0) - E_2(Y|X = x_0) = (\beta_{01} - \beta_{02}) + (\beta_{11} - \beta_{12})x_0$$

is sensitive to the choice of x_0.

(2) Assume $E_1(X) \neq E_2(X)$.

Then,

$$E_1(Y) - E_2(Y) = (\beta_{01} - \beta_{02}) + \left(\beta_{11}E_1(X) - \beta_{12}E_2(X)\right),$$

and $E_1(Y) - E_2(Y)$ cannot be estimated as a simple linear combination of parameters.

This chapter explores the use inverse response propensity weighting and ANCOVA for situations when ANCOVA assumptions are violated.

6.1.1 Different E(X) and Equal E(Y) between the Sections

In the following simulations, random data was generated so that the two sections have a different *Exam1* score mean but the same *FinalExam* mean:

$$X_1 \sim Beta(4, .8)$$

$$Y_1 = X - 13 + \epsilon$$

and

$$X_2 \sim Beta(4, 1.13)$$

$$Y_2 = X - 8 + \epsilon.$$

Then,

$$E_1(X) = 83$$

$$E_1(Y) = 70$$

and

$$E_2(X) = 78$$

$$E_2(Y) = 70.$$

Thus, both sections have a different mean *Exam1* score and the same *FinalExam* score. The interest is in the difference between *FinalExam* scores between the two classes. Therefore, since both classes have the same expected *FinalExam* score, when conducting the analysis, a significant difference between the sections should not be found.

6.1.1.1 Use of ANCOVA when E(X) is Different but E(Y) is Equal between the Sections: Type I Error Rate

In this simulation, an Analysis of Covariance is used fitting $Y_{ij} = \beta_{0j} + \beta_1 X_{ij}$ and testing $\beta_{01} = \beta_{02}$. Since $E(Y)$ is equal for both sections, the test conducted should not detect a difference between the two sections. However, since $E(Y|X = x_0)$ is different for both sections, ANCOVA is probably going to find a significant difference.

Table 6.1. ANCOVA with Different E(X) and Equal E(Y) for Both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	31.98	28.24	24.29	20.16	16.75	13.01	10.14	7.85	6.13
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	100.0%								
	2	100.0%	100.0%							
	3	100.0%	100.0%	100.0%						
	4	100.0%	100.0%	100.0%	100.0%					
	5	100.0%	100.0%	100.0%	100.0%	100.0%				
	6	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%			
	7	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%		
	8	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
	9	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

As it was expected, there was a 100% Type I error. Both sections had the same *FinalExam* score mean, but due to the difference in baseline the ANCOVA found a significant difference between both sections. The reason for this is because ANCOVA is detecting $E_1(Y|X = x_0) \neq E_2(Y|X = x_0)$, rather than just comparing $E_1(Y) - E_2(Y)$.

6.1.1.2 Use of Independent Samples T-Test with Inverse Response Propensity Score Weighting when E(X) is Different but E(Y) is Equal between the Sections: Type I Error Rate

In the following simulations, the same scenario as above is simulated and an independent samples *t*-test with inverse response propensity score weighting is used. Below are the results:

Table 6.2. Independent Samples T-Test with Inverse Response Propensity Score Weighting with Different E(X) and Equal E(Y) for Both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	18.10%								
	2	15.90%	15.70%							
	3	11.90%	13.20%	10.60%						
	4	13.60%	10.10%	8.10%	6.90%					
	5	14.20%	10.10%	8.60%	6.10%	5.80%				
	6	14.00%	10.90%	8.90%	6.10%	5.20%	4.50%			
	7	15.50%	10.20%	7.90%	5.70%	4.10%	3.70%	3.60%		
	8	14.50%	11.50%	7.50%	5.20%	4.00%	4.40%	4.60%	4.60%	
	9	14.10%	10.20%	7.60%	5.00%	3.20%	4.30%	4.80%	4.70%	3.20%

Even though for high drop rates the Type I error rate was 18.10%, this method was able to give improved estimates compared to ANCOVA when both sections have a different *Exam1* mean and the same *FinalExam* mean.

6.1.2 Equal E(X), Different X and Y Relationship, Equal E(Y)

In the following simulations, random data was generated so that the two sections have the same *Exam1* score mean, the same *FinalExam* mean, but a different relationship between *Exam1* and *FinalExam* scores:

$$X \sim Beta(4, .8)$$

$$Y_1 = X - 13 + \epsilon$$

$$E_1(Y) = 70,$$

and

$$Y_2 = 1.15X - 25 + \epsilon$$

$$E_2(Y) = 70.$$

It is clear that the ANCOVA assumption of equal slopes is being violated.

6.1.2.1 Use of ANCOVA when E(X) and E(Y) are Equal but Assumption of Equal Slopes is Violated: Type I Error Rate

In the following set of simulations, and ANCOVA is being used to analyze the data.

Table 6.3. ANCOVA with Equal E(X), Equal E(Y), and Different Relationship between X and Y in Both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	38.40%								
	2	39.10%	37.20%							
	3	36.20%	35.30%	34.30%						
	4	35.60%	34.10%	30.20%	26.10%					
	5	34.10%	31.20%	29.10%	23.80%	21.80%				
	6	31.70%	31.70%	26.50%	24.80%	21.00%	19.00%			
	7	33.40%	30.90%	25.90%	23.20%	20.10%	16.50%	16.30%		
	8	33.50%	27.40%	26.50%	22.20%	20.30%	16.20%	16.10%	14.50%	
	9	32.90%	29.60%	24.30%	20.50%	16.80%	15.40%	16.00%	13.80%	12.60%

As shown in the above table, the Type I error rate is high. The ANCOVA detected a large number of significant differences between the groups when in fact there are no differences in mean *FinalExam* score means. Additionally, notice that the larger the dropout, the higher the Type I error rate becomes. As in the previous simulations, this is likely due to the high dropout among students with low scores.

6.1.2.2 Use of Independent Samples T-Test with Inverse Response Propensity Weighting when E(X) and E(Y) are Equal but ANCOVA Assumption of Equal Slopes is Violated: Type I Error Rate

In the following simulations, the same scenario as above is represented, but rather than analyzing the data with an ANCOVA, the independent samples *t*-test with inverse response propensity weighting is used.

Table 6.4. Independent Samples T-Test with Inverse Response Propensity Score Weighting with Equal E(X), Equal E(Y), and Different Relationship between X and Y in Both Sections: Type I Error Rate

		Section 1								
	Type	1	2	3	4	5	6	7	8	9
	Drop%	24.46	21.97	19.06	15.98	13.33	11.09	9.21	7.23	5.96
	β_0	6.06	5.06	4.06	3.06	2.06	1.06	0.06	-0.94	-1.94
	β_1	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01
Section 2	1	19.70%								
	2	15.80%	14.20%							
	3	14.70%	11.60%	9.80%						
	4	11.40%	8.60%	7.10%	6.60%					
	5	11.70%	10.10%	8.20%	7.00%	4.70%				
	6	10.60%	9.90%	6.50%	6.30%	5.70%	6.10%			
	7	11.00%	8.30%	6.80%	6.00%	6.00%	4.30%	5.40%		
	8	10.80%	8.90%	7.90%	5.00%	4.90%	4.50%	5.00%	5.90%	
	9	10.40%	8.30%	8.30%	7.10%	3.80%	5.10%	4.80%	4.80%	5.00%

As shown in the table above, even though when the dropout is high the Type I error is also high, it is still more reasonable than when an ANCOVA was used. The higher Type I error rate with the higher dropout is possibly due to those extreme observations that are not being represented.

6.2 Conclusion

This chapter explored scenarios where the ANCOVA assumptions are violated. It was found that when $E(X)$ or β_1 are different between the groups, ANCOVA is an inappropriate tool and a *t*-test with inverse response propensity score weighting can control Type I error rates.

It was found that when the *Exam1* score means are different between the two classes, the ANCOVA detected differences in *FinalExam* score means 100% of the times when the simulations were set such that there were no differences in *FinalExam* score means. However, independent *t*-tests with inverse response propensity weighting had a Type I error rate between 3.20% and 18.10%, showing that this method is superior to ANCOVA when the baselines are different between the two groups.

Additionally, it was found that when $E_1(X) = E_2(X)$ and $E_1(Y) = E_2(Y)$ but the relationships between *X* and *Y* were different (meaning the two sections had different slopes connecting *Exam1* and *FinalExam* scores), the Type I error rate for ANCOVA was excessively high (12.60% to 39.10%), while the Type I error rate for independent samples *t*-test with inverse response propensity weighting was again between 3.80% and 19.70%.

CHAPTER 7

DISCUSSION

This thesis addressed the importance of missing data and explored four methods for comparing group means while adjusting for missingness: independent samples with unequal variances *t*-test with predicted mean stratification, independent samples with unequal variances *t*-test with response propensity stratification, independent samples with equal variances *t*-test with inverse response propensity score weighting, and analysis of covariance.

Before testing the missingness adjustment methods, a set of simulations using an independent samples *t*-test assuming equal variances with no missingness adjustment was conducted. It was found that when the drop patterns are equal between the two sections, the Type I error rate was around 5%, indicating that in these situations a *t*-test is an appropriate tool to compare two group means. However, when the drop patterns differed between the two sections, the Type I error rate increased, making this tool inappropriate for group comparisons.

The two stratification methods suggested by Little were evaluated. It was found that when the dropout patterns differed between the two sections, these two methods offered a slight improvement over a *t*-test without any missingness adjustment. However, they still had an excessively high Type I error rate. Furthermore, when the drop patterns were similar between the two sections, the Type I error rate was increased in comparison to *t*-test without missingness adjustment. This appeared to be due to the lack of normality within the cells.

It was found that independent samples t-test with inverse response propensity weighting and ANCOVA can both control bias caused by dropout. When $E_1(X) = E_2(X)$ and $\beta_{11} = \beta_{12}$, ANCOVA was found to be a stronger tool than inverse response propensity weighting: it had the least amount of Type I error rate and the highest power. However, the use of ANCOVA is not appropriate if the assumptions are not met. In this case, using an independent samples equal variances t-test with inverse response propensity score weighting is more appropriate, and simulations supported this by showing that it had a lower level of Type I error rate and appropriate power when the dropout is not extreme.

Therefore, an independent samples t-test assuming equal variances may be used when the drop patterns are equal between the two sections. Otherwise, if ANCOVA assumptions are met, the group means may be compared using this tool. If the drop patterns are different the two sections and the ANCOVA assumptions are not met, then a t-test with inverse response propensity weighting is an appropriate tool as long as the dropout is not extremely high.

These methods are relatively simple to implement, and they are applicable in studies aimed to compare groups. Examples of potential research settings that may use these two methods are educational settings, clinical trials, survey research, and other studies involving group comparisons where it is possible to collect data that may be used as a covariate.

This thesis illustrates the importance of collecting information that may be used as a covariate. Data predictive of outcome as well as data related to dropout may be used in the study if missingness is an issue.

A potential limitation of this study is the fact that the drop rates considered were not extremely high. The Type I error rate and power of ANCOVA did not appear to be affected by the drop rate; however, if extreme drop rates were considered, it is possible that ANCOVA would not have enough power to detect a difference between groups. Similarly, in the simulations involving *t*-test weighted by the inverse response propensity score, the Type I error rate started becoming inflated as the drop rate increased. It appears that in these situations it is possible that the extreme values are missing and not represented by any observations.

In future research involving inverse response propensity weighting, it would be interesting to find a way of estimating the lowest values of Y that are not being represented due to excessive missingness. It is possible that using regression imputation to estimate those values might be a possible substitute. Additionally, since this study only considered moderate drop rates (below 31%), scenarios with more extreme drop rates should be considered in future research to see whether this method can handle the amount of missing data. Finally, further research could explore the application of inverse response propensity weighting in other settings where the goal may not be comparisons of two groups, but other forms of estimation.

REFERENCES

Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W. (2005). *Applied Linear Statistical Models, 5th Edition*. McGraw-Hill: New York.

Little, R. (1986). “Survey Nonresponse Adjustments for Estimates of Means.” I*nternational Statistical Review,* 52, 139-157.

Little, R., and Rubin, D. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.: Hoboken.

Lu, K. (2010). “On Efficiency of Constrained Longitudinal Data Analysis versus Longitudinal Analysis of Covariance.” *Biometrics*, 66, 891-896.

Rubin, D. (1976). “Inference on Missing Data.” *Biometrica,* 63, 581-592.

Schafer, J. and Graham, J. (2002). “Missing Data: Our View of the State of the Art.” *Statistical Methods*, 7, 147-177.

VITA

Gabriela Maria Stegmann was born and moved to the United States Gabriela graduated and continued to pursue her education at the University of North Florida. After completing her Bachelor of Arts in Psychology with minor in Mathematics in December, 2007, she stayed at the university, obtaining her Master of Arts in Counseling Psychology in April, 2010. She worked in the mental health field at Riverpoint Behavioral Health and with children with developmental delays at the Early Steps program at the University of Florida. In 2013, Gabriela returned to the UNF for her Master of Science in Mathematics with concentration in Statistics. Gabriela is expected to graduate in December, 2015.