ANALYZING THE RELATIONSHIP BETWEEN ATTENDANCE IN HYBRID SUPPLEMENTAL INSTRUCTION AND TUTORING SESSIONS AND SUCCESS IN COURSES WITH TRADITIONALLY HIGH FAILURE RATES

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

ALLYSSA KELLEY B.S.

DENTON, TEXAS

AUGUST 2016

ACKNOWLEDGEMENTS

I would like to graciously acknowledge the many people who helped me to see this thesis to the end. First of all, thank you to my amazing support system of family and friends. Thanks to my parents, Gary and Lauria Kelley, who supported me in my pursuance of yet another degree, and who understand my passion for knowledge and education. Thank you to my siblings, Ross and Abby, for their comedic relief when it all got a little bit too stressful. A huge thank you to Audrey Parker, Cammy Boaz, and the Math Lab girls for lending an ear when I needed to vent and for always offering kind words of encouragement and support. Thank you to my advisor and good friend, Dr. Brandi Falley, for everything she has done for me, academically and personally, and to the rest of my committee members for sticking with me through the constant switching of defense dates. And finally, thank you to Mr. Paul Ingram, for informing me that I was doing everything completely wrong, crushing my world, and then helping me build it all back up again, better and stronger.

ABSTRACT

ALLYSSA KELLEY

ANALYZING THE RELATIONSHIP BETWEEN ATTENDANCE IN HYBRID SUPPLEMENTAL INSTRUCTION AND TUTORING SESSIONS AND SUCCESS IN COURSES WITH TRADITIONALLY HIGH FAILURE RATES

AUGUST 2016

Academic assistance programs have been around in some form or another since the mid-1600s, and they have grown and developed just as the demographics of students in tertiary institutions have. The supplemental instruction model is an assistance program that was built as a response to the needs of the shifting student body, and Texas Woman's University is one of many institutions that have adopted a hybrid supplemental instruction model in an effort to increase student performance in their classes. The purpose of this research was to determine which, if any, factors impact the success of a student in the TWU Supplemental Instruction and Tutoring program using logistic regression analysis to build prediction models for success. The models that were created showed that, contrary to the hypothesis of this study that the number of SIT sessions attended would provide the largest impact, whether a student was determined to be at risk by the standards set by the TWU SIT program and whether the student attended the minimum number of required SIT sessions were the variables most influential on success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Academic assistance programs have been around in some form or another since the mid-1600s, and they have grown and developed just as the demographics of students in tertiary institutions have. What started as simple tutoring for only privileged white males progressed to remedial education classes within college preparatory programs, and grew even further to include more nontraditional students of both genders and a variety of developmental education and enrichment programs (Arendale, 2002). In recent years, the higher education community has shifted its focus from just giving students an opportunity by accepting them into their institutions, to helping students succeed and prosper not only in their courses, but in their collegiate careers as well. Initially, remedial programs seemed to be sufficient in helping students to do well; however, according to a report by Complete College America, 22.5% of freshmen in 4-year colleges are enrolled in remediation, and of those, less than half (49.2%) actually complete the remediation, and less than a third (29.6%) graduate within 6 years (Remediation, 2012).

At Texas Woman's University (TWU), educators were looking to implement an academic assistance program to increase student performance in their classes, as well as in their collegiate careers in general. However, with a remedial program

already in place at TWU, and recognition of a need for a program that pulled away from remediation and instead focused on higher order thinking and a more comprehensive assistance program, the supplemental instruction (SI) model was adopted.

The purpose of this research was to determine which, if any, factors impact the success of a student in the TWU Supplemental Instruction and Tutoring (SIT) program, and the hypothesis of this study was that the number of SIT sessions attended will be the biggest influence on success. In order to do this, logistic regression analysis was used to predict success and non-success based on the factors each student faced. The variables that had the biggest impact on a student's success were identified, as well as the degree of their impact on the model.

In Chapter II, we provide a literature review of a study previously completed over the origins of the TWU SIT program, as well as look at the history of the traditional supplemental instruction program. In Chapter III, we explain the process of the data preparation, including how the data was cleaned, coded, and prepared for analysis. Chapter IV introduces the mathematical theory behind logistic regression, assumptions made by using this method, and why it was chosen for this study. Chapter V discusses the results found from the logistic regression analysis, and outlines which, if any, factors impacted success, and by how much. Lastly, Chapter VI concludes the research and explains any limitations of this study, as well as remaining questions or future research to be done.

CHAPTER II

LITERATURE REVIEW

A History of Supplemental Instruction

Supplemental instruction is a program that originated at the University of Missouri-Kansas City (UMKC) in 1973, and according to David Arendale (2002), emerged as a "response to a need at the institution created by a dramatic change in the demographics of the student body and a sudden rise in student attrition" (p. 17). Educators at UMKC recognized a need from the students for some sort of intervention that steered away from the traditional approach to developmental and remedial programs, and that would be a cost effective way to not only decrease attrition rates, but also increase student success in their collegiate career.

The SI model is a peer-assisted program in which students are provided with supplemental instruction sessions outside of class. SI sessions are led by SI Leaders, who are students that have already successfully completed the course. SI Leaders are not meant to re-teach information but instead are there to "provide structure to the study session" (Arendale, 1994, p. 3). These leaders attend class, provide a model of great student attributes, and are trained at integrating "what to learn, with how to learn" (Arendale, 1994, p. 3). SI sessions are meant to be empowering to

students, and are “extensions of the classroom where students continue the learning process initiated by the professor” (Arendale, 2002, p. 22). They are a place for collaborative learning, and this supportive environment is what helps the SI model keep clear of any stigmas similar to those observed in remedial programs. According to David Arendale (1994), a former National Project Director for Supplemental Instruction,

> Assistance begins in the first week of the term. The SI leader introduces the program during the first class session and surveys the students to establish a schedule for the SI sessions. Attendance is voluntary. Students of varying abilities participate, and no effort is made to segregate students based on academic ability. Since SI is not perceived to be remediation, many unprepared students that might otherwise avoid seeking assistance will participate since there is no stigma attached. (p. 1)

Being that attendance is initially voluntary, and that even the well-prepared students are attending these sessions, underprepared students are more willing to attend and therefore benefit from the program. In a remedial program, however, the stigma attached to outside assistance may cause those same students to lose motivation and shy away from seeking outside help. Also, because assistance begins the first week of class, the SI program is looked at as a proactive model as opposed to a reactive model. Students and educators do not have to wait until there are signs of struggle (typically several weeks into the course) in order to enlist help outside of

the classroom. This allows the student to keep a consistent grade, instead of failing a test and having to work to bring his or her grade back up.

TWU CSSP Grant

The academic assistance program that was enacted at TWU in an effort to increase student performance was the Comprehensive Student Success Program, or CSSP. This grant-funded program was awarded to TWU by the Texas Higher Education Coordinating Board (THECB) and used what program officials called a hybrid SI model. The initial requirements for all students involved with the program were two-fold: "1) attend two SI sessions with a course assistant, and 2) attend tutoring twice during the semester" (Carlsen-Landy, Falley, Wheeler, and Edwards, 2014, p. 12). Next, whether a student was at risk was determined in two areas: academically and demographically through a tool called MAP-Works®. According to Carlsen-Landy et al., (2014):

> …students who were identified as at risk academically were required to attend four additional SI sessions and attend weekly tutoring. Students were considered academically at risk if their class average fell below 70%, and once they were identified as at risk academically, they were considered at risk for the rest of the semester. Students who were at risk academically or who were identified through MAP-Works® as at risk for something other than academics, such as financial, emotional, or social risks, were required to meet with the CSSP Coordinator to discuss ways to mitigate risks which may

> adversely affect the student's ability to be successful in class and at the university. (p. 12)

At risk students were also required to have contact at least twice a month with a Course Assistant (CA), who acted as a peer-mentor. However, as the CSSP program grew throughout the semesters following its creation, requirements shifted and changed slightly. First, the separate SI sessions and tutoring sessions were combined to create a supplemental instruction and tutoring session. These new SIT sessions consisted of predominately supplemental instruction, with 10-15 minutes of tutoring at the end. Another way in which the TWU CSSP strayed from the traditional SI model is that TWU CSSP called for attendance in SI (and consequently SIT) sessions to be mandatory, and in the second year of the program attendance of these sessions also became 10% of a student's grade.

As seen in the study by Carlsen-Landy et al., (2014), which examined the effectiveness of the CSSP, the hybrid SI program worked extremely well, garnering results that showed the treatment classes outperformed every one of the control classes. Although the analysis for that study only focused on the 2013 fall semester, this study continues what Carlsen-Landy et al., did and extends the analysis to both fall and spring semesters of the 2013-2014 and 2014-2015 school years.

CHAPTER III

DATA PREPARATION

This study was concerned with information from the TWU SIT program, and in an effort to use a collection of data that was a representative sample of the population, only data from the second and third year of the program was included. This is due to changes in how the program was run between the first and second year of its existence, as well as incomplete data collection for the fourth year. The time period that was used covered the 2013-2014 and 2014-2015 school years, which contained multiple sections of the four courses participating in the program: Microbiology (BACT 1003), Introductory Chemistry (CHEM 1013), College Algebra (MATH 1303), and Elementary Statistics I (MATH 1703). In total there were 2,639 students, of which 18 were excluded in the analyses because of missing information, which left 2,621 valid data entries.

In order to assemble the data for students enrolled in these four courses, the variables needed to build prediction models for success were considered. For each student, it was imperative to know the course and semester in which enrolled, the risk status of the student, the number of SIT sessions the student attended, whether the student met the minimum number of required SIT sessions, and the student's final course grade. Variables that were asked for but were not able to be included

were: when SIT sessions were attended (e.g., 6 total SIT sessions: 2 before exam one, 2 before exam two, 1 before exam three, 1 before the final exam), and after which exam a student was deemed to be at risk. (These variables will be discussed further in the Limitations chapter.) Other variables needed for this analysis included categorical or quantitative demographic information that may have impacted or influenced student success such as: age, gender, ethnicity, and class rank.

The data were collected by various officials of the TWU SIT program, and was deidentified by a former program coordinator. This was done as part of the TWU CSSP Grant from which the TWU SIT program originated. The information provided was a collection of demographic and academic details of the students enrolled in the SIT supported classes. These variables included age, gender, ethnicity, class rank, course and semester in which enrolled, risk status, number of SIT sessions attended, whether the student met the minimum number of required SIT sessions, and final course grade. From final course grade, another variable called Success was created. With the focus of this study being predictors of success, all variables other than success and final course grade are independent variables, leaving success as the dependent variable.

Determining how to define success was not difficult, as it is customary to assign a label of success to a passing final course grade of A, B, or C. However, when considering how non-success is defined, the matter of whether to include withdrawals and incompletes came into question. Because the traditional SI model,

and the goal of the TWU SIT program is to increase not only academic performance, but retention as well (Arendale, 1994), a student who withdraws from the course or does not complete the course in turn does not earn a final course grade of A, B, or C and has therefore not been successful in the course. This then defines non-success as a final course grade of D, F, W, or I.

After looking over the data that were presented, it was determined that another variable needed to be created, or rather, an existing variable needed to be better defined. In order to be able to run an analysis on just the math courses (Algebra and Statistics) or just the science courses (Microbiology and Chemistry), the categorical variable course was used to create a new categorical variable called Subject. To create the subject labeled SCI, it was simply a matter of combining the microbiology and chemistry courses, and the process was repeated with the algebra and statistics courses to create the subject labeled MATH.

As previously mentioned, the data consist of 2,621 valid entries, with 1,164 students in the math courses combined, and 1,456 in the science courses. The decision to separate the analysis by subject came from the difference in requirements and nature of the courses. As shown in Table 3.1, the average number of SIT sessions attended by students in the science courses was 6.45 sessions, with a standard deviation of 5.87. Math students, however, attended an average of 4.44 sessions with a standard deviation of 3.411. With science students attending an average of two more sessions than the math students, it was hypothesized that there

may be a significant difference in the distribution of the variable SIT between the two groups. Using SPSS to run an independent samples test validated this suspicion, and showed that there was in fact a significant difference. (Table 3.2 shows the output from this test.) Therefore, all analysis was run twice, once using just math courses, and once with just science courses.

Table 3.1

Group Statistics for the Quantitative Variable SIT in Both Subjects

	Subject	N	Mean	Std. Deviation	Std. Error Mean
SIT	MATH	1173	4.44	3.411	.100
	SCI	1464	6.45	5.870	.153

Table 3.2

Independent Samples Test Showing the Significant Difference Between Subjects

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
SIT	Equal variances assumed	107.099	.000	-10.400	2635	.000	-2.009	.193	-2.388	-1.630
	Equal variances not assumed			-10.985	2419.513	.000	-2.009	.183	-2.368	-1.651

The data were presented in the form of Microsoft Excel spreadsheets, and were sent as four different documents that were separated by semester. The first hurdle in cleaning the data was to create consistent values for each variable. These inconsistencies appeared in multiple places, with one such example being Ethnicity, where in each semester "African American" was represented in a different manner (i.e., Black, B, African American, Black: Non-Hispanic). Once all of the inconsistencies had been addressed, the data was then compiled into one spreadsheet and the coding process began.

The dependent variable of success (3.1), and independent variables gender (3.2), risk status (3.3), and whether a student met the minimum required SIT session (3.4), were all variables that easily leant themselves to be coded as binary variables, being that in their raw state, they were yes/no, categorical variables. However, the multiple values for ethnicity (3.5), final course grade, (3.6) and class rank (3.7) were all enumerated with dummy code to make analysis easier. A complete list of variables, descriptions of variables, and coding can be found in the Appendix.

$$Success = \begin{cases} 0, Success \\ 1, Non - Success \end{cases} \tag{3.1}$$

$$Gender = \begin{cases} 0, Female \\ 1, Male \end{cases} \tag{3.2}$$

$$Risk = \begin{cases} 0, No \\ 1, Yes \end{cases} \tag{3.3}$$

$$MetMin = \begin{cases} 0, No \\ 1, Yes \end{cases} \tag{3.4}$$

$$Ethnicity = \begin{cases} 1, White \\ 2, Black \\ 3, Hispanic \\ 4, Asian \\ 5, Hawaiian.PacIsl \\ 6, Native\ American \\ 7, International \\ 8, Other \end{cases} \tag{3.5}$$

$$Grade = \begin{cases} 1, A \\ 2, B \\ 3, C \\ 4, D \\ 5, F \\ 6, W \\ 7, I \end{cases} \tag{3.6}$$

$$ClassRank = \begin{cases} 1, Freshman \\ 2, Sophomore \\ 3, Junior \\ 4, Senior \\ 5, PostBac \\ 6, Masters \\ 7, Doctoral \end{cases} \tag{3.7}$$

CHAPTER IV

REGRESSION ANALYSIS

As mentioned in the previous chapter, it was decided to run the analysis on math and science courses separately. With the focus of this study mainly being the effect that the number of SIT sessions has on the outcome of success, and the difference in distributions between the attendance of SIT sessions in the two subjects, the analysis would be more accurate if run separately than it would be if both subjects were included. This chapter will discuss the two models that were created, as well as the methods used to create them; however, the interpretations of these models will be examined in the next chapter.

Regression

In order to predict success and determine which factors were most indicative of a favorable outcome, regression analysis was chosen because of its ability to demonstrate relationships between variables. A regression model details the relationship between a response variable and one or more explanatory variables (Utts & Heckard, 2016). A simple linear regression model assumes that the response variable, y, depends on just one explanatory variable, x, according to a linear equation like the one shown below, where β_0 is a constant, and β_1 is the coefficient of the explanatory variable.

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

(4.1)

A multiple regression model, however, will show in what way a response variable depends on multiple explanatory variables. When the explanatory variables are denoted as a linear combination, $\alpha = x_1, x_2, \cdots, x_n$, and then combined with an expansion of the simple linear regression equation (4.1), a new equation (4.2) is created to model the relationship between the one response variable and multiple explanatory variables (Moore, D. S., McCabe, G. P., & Craig, B. A., 2009).

$$\mu_{y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

(4.2)

One assumption of simple linear regression is that the response variable is of the quantitative variety, meaning that the predicted value from a model will be a quantity that can be measured or counted. Multiple regression holds the same assumption, but uses more than one explanatory variable. In this study, the response variable is dichotomous because it is either represented as "yes" or "no", and is consequently a qualitative variable. Therefore, because of the unsatisfied assumptions, using either simple linear regression or multiple regression will not be beneficial in this analysis and would result in predictive models and values that are just not valid.

Logistic Regression

Unlike simple linear regression and multiple regression, logistic regression is able to handle a response variable of a dichotomous or qualitative form. In the case of this study, success is labeled as either "yes" or "no," making it a dichotomous variable. However, similar to multiple regression, logistic regression still examines the relationship between a response and more than one explanatory variable; it just does so in a slightly different way. Where a linear regression model will predict a value of the response variable, logistic regression instead models the likelihood, or probability, of the occurrence of the response variable given the various independent variables, and can also predict the effect that a series of variables will have on the dependent variable. Although this study is interested in building a prediction model, the ability to examine the effects of independent variables on the dependent variable is also very useful.

Logistic regression analysis is one that is based on probability, odds, and an odds ratio, all of which are related ideas, but not exactly the same. Probability is defined as "a number between 0 and 1 that is assigned to a possible outcome of a random circumstance," whereas odds are defined as "the probability that an event happened compared to the probability that it did not," and an odds ratio "compares the odds of an event for two different categories" (Utts & Heckard, 2015). To further illustrate the connection between these three, it is helpful to look at how each are found. Probability is found by dividing the number of outcomes of interest by the

number of all possible outcomes. Below is the equation used to find probability, where $P(x)$ is read as "the probability of event x."

$$P(x) = \frac{outcomes\ of\ interest}{all\ possible\ outcomes} \tag{4.3}$$

Odds are then found by dividing the probability that an event will occur by the probability that the same event will not occur. It is known that the total of probabilities for all possible outcomes is equal to 1, therefore, since an event either occurs or does not occur, if the probability that x occurs is p, then the probability that x does not occurs is $1 - p$.

$$Odds = \frac{P(x\ occurs)}{P(x\ does\ not\ occur)} = \frac{p}{1-p} \tag{4.4}$$

Odds ratios are simply the odds of one event happening, in two separate categories. This is found by dividing the odds of the first category by the odds of the second category.

$$Odds\ Ratio = \frac{Odds_1}{Odds_0} = \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} \tag{4.5}$$

Probability and odds are important for logistic regression because they are the foundation for the odds ratio. The odds ratio is the tool that is able to explain the relationship between an independent and dependent variable. For an

independent variable, the odds ratio represents how the odds of the dependent variable change for a one-unit increase in the independent variable, when all other independent variables are held constant. This piece of information is useful in determining exactly how the response variable depends on the explanatory variables.

The dependent variable, Y, in logistic regression follows the Bernoulli distribution with an unknown probability, p, because it is a binary variable that can take on the value of either $\{success = 0\}$ or $\{non\ success = 1\}$. Logistic regression aims to estimate the conditional probability $P(Y = 1|x = x)$ as a function of x, but can not do so until a few issues are addressed. First, the independent variables need to be linked to essentially the Bernoulli distribution. To tie together the linear combination of variables and the Bernoulli distribution, a function is needed that maps the linear combination of variables onto the Bernoulli probability distribution with a domain from 0 to 1. The natural log of the odds ratio, the logit, is that link function.

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) \tag{4.6}$$

This link function needs to result in a domain from 0 to 1 because of the definition of probability. Probability values can only span between 0 and 1, inclusive, therefore a function that would provide values outside of this domain will not yield valid results. (This is one of the biggest reasons that linear regression

cannot be used for a dichotomous response variable.) When looking at the graph of the logit function, it is clear to see that the function is undefined at 0 and 1, and has a domain that fits the requirements that are called for.

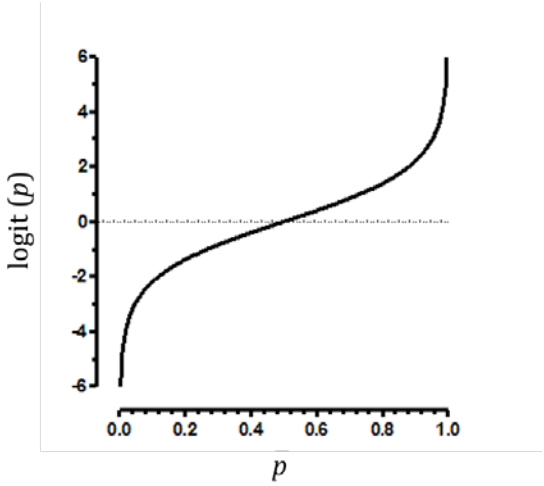

Figure 4.1: Graph of the logit(p) Function

When using this logit link function, probabilities run along the x-axis, however, because we are predicting probabilities, it is more useful for probabilities to lie on the y-axis. This transformation is accomplished by taking the inverse of the logit function in equation (4.6), and using α as the linear combination of independent variables.

$$logit^{-1}(\alpha) = p = \frac{e^{\alpha}}{1 + e^{\alpha}} \tag{4.7}$$

Once this transformation is done, you can see from the graph of $logit^{-1}(\alpha) = p$ that the probability is on the y-axis, yet still is bounded by 0 and 1, and $logit\ (p)$ now runs along the x-axis. This graph forms what is called a sigmoid "S" curve, which is a key property of fitting the data for logistic regression.

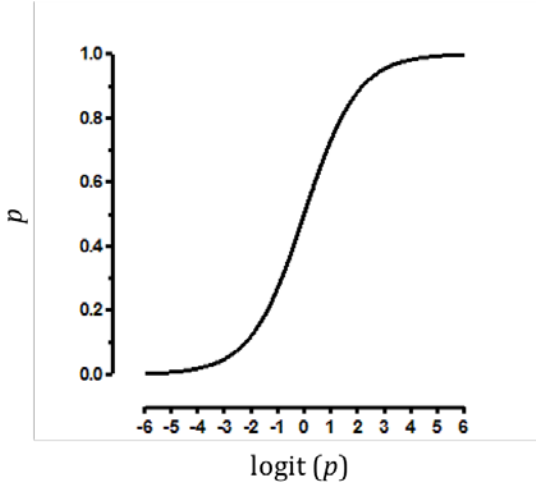

Figure 4.2: Graph of the logit^{-1}(p) Function

Next, we will discuss the estimated regression equation, or ERE. It is important to remember that one of the goals of this logistic regression analysis is to find the likelihood, or probability, of success. It is known that the natural log of the odds ratio is equivalent to a linear combination of the independent variables (Foltz, 2015), and from equation (4.1) that combination is $\beta_0 + \beta_1 x_1$. Therefore, by using the inverse of the logit function, or the antilog, it is possible to solve for p, the probability of success.

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1$$

$$\frac{p}{1-p} = \exp(\beta_0 + \beta_1 x_1)$$

$$p = \exp(\beta_0 + \beta_1 x_1)\,(1-p)$$

$$p = \exp(\beta_0 + \beta_1 x_1) - \exp(\beta_0 + \beta_1 x_1)\,(p)$$

$$p - \exp(\beta_0 + \beta_1 x_1)\,(p) = \exp(\beta_0 + \beta_1 x_1)$$

$$p(1 - \exp(\beta_0 + \beta_1 x_1)) = \exp(\beta_0 + \beta_1 x_1)$$

$$p = \frac{\exp(\beta_0 + \beta_1 x_1)}{1 + \exp(\beta_0 + \beta_1 x_1)} = \hat{p} \qquad (4.8)$$

Now that the estimated regression equation has been found, once the data are run through SPSS and coefficients are calculated, it is possible to plug those coefficients into the ERE and produce a predicted probability of success given the linear combination of independent variables.

Variable Selection Methods

When running a regression analysis, there are many different ways to select the variables that are entered into the regression models. Enter is a selection method in which all variables are entered in a single step. Stepwise selection creates a type of multiple regression model in which the predictive variables are chosen by an automatic procedure based on criteria decided on before hand. This process can take form in one of three ways. Forward selection will determine the most significant indicator of the response variable, and will keep building

increasingly better models by adding in one explanatory variable at time. This continues until adding explanatory variables no longer betters the model. Backwards elimination, however, selects all possible explanatory variables initially, and then discards those that are insignificant, and again will continue until doing so no longer produces a stronger model. Stepwise selection progresses just as forward selection, yet at each step the procedure allows for deletions that again would better the model.

There are some who argue that a stepwise selection method is less than adequate because "it seems unwise to let an automatic algorithm determine the questions we do and do not ask about our data" (Judd & McClelland, 1989). In a frequently asked questions page from Stata discussing these issues, it is noted that there are in fact many ways that stepwise selection can tarnish the results of a study (Sribney, B., Harrell, F., & Conroy, R., 1998). The many issues related to this particular selection process made it easy to choose the Enter selection method as the one to be used in this research.

Models for Math Courses

The following shows the logit function, as well as the estimated regression equation that was produced by SPSS when running a logistic regression analysis on just cases in which the subject was math. To model the relationships between success and the chosen explanatory variables, it is necessary to remember back to equation (4.2) from earlier in this chapter.

$$logit(p) = 2.687 + 1.653MetMin - 3.025Risk + 0.188SIT - 0.068Age$$

(4.9)

$$ERE = \hat{p} = \frac{\exp(2.687 + 1.653MetMin - 3.025Risk + 0.188SIT - 0.068Age)}{1 + \exp(2.687 + 1.653MetMin - 3.025Risk + 0.188SIT - 0.068Age)}$$

(4.10)

Table 4.1

SPSS Coefficients Table for Models Built From the Math Courses

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	-.068	.023	9.037	1	.003	.934	.894	.977
	Ethnicity	.018	.075	.056	1	.813	1.018	.879	1.178
	Gender	.163	.273	.359	1	.549	1.177	.690	2.009
	ClassRank	-.104	.108	.938	1	.333	.901	.729	1.113
	SIT	.188	.045	17.425	1	.000	1.207	1.105	1.318
	Risk	-3.025	.208	210.938	1	.000	.049	.032	.073
	MetMin	1.653	.278	35.230	1	.000	5.221	3.025	9.011
	Constant	2.687	.482	31.147	1	.000	14.691		

a. Variable(s) entered on step 1: Age, Ethnicity, Gender, ClassRank, SIT, Risk, MetMin.

Models for Science Courses

The models used to represent the science courses were constructed in the same way as explained for math. Data values in which subject was equal to science were run through a logistic regression analysis in SPSS, using the enter variable selection method, and Table 4.3 shows the coefficients table that was then produced. The following models were built from that table.

$$\text{logit}(p) = 1.479 + 2.225MetMin - 3.229Risk + 0.062SIT - 0.156Ethnicity$$

(4.11)

$$ERE = \hat{p} = \frac{\exp(1.479 + 2.225MetMin - 3.229Risk + 0.062SIT - 0.156Ethnicity)}{1 + \exp(1.479 + 2.225MetMin - 3.229Risk + 0.062SIT - 0.156Ethnicity)}$$

(4.12)

Table 4.2

SPSS Coefficients Table for Models Built From the Science Courses

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Age	-.035	.019	3.221	1	.073	.966	.930	1.003
	Ethnicity	-.156	.063	6.144	1	.013	.855	.756	.968
	Gender	.433	.326	1.767	1	.184	1.542	.814	2.921
	ClassRank	-.014	.093	.022	1	.883	.986	.822	1.184
	SIT	.062	.021	8.777	1	.003	1.064	1.021	1.109
	Risk	-3.229	.178	328.102	1	.000	.040	.028	.056
	MetMin	2.225	.227	96.225	1	.000	9.249	5.930	14.426
	Constant	1.479	.428	11.942	1	.001	4.390		

a. Variable(s) entered on step 1: Age, Ethnicity, Gender, ClassRank, SIT, Risk, MetMin.

CHAPTER V

ASSESMENT OF PREDICTIVE MODELS

Although the study by Carlsen-Landy et al. (2014) looked at the effectiveness of the TWU CSSP by comparing treatment and control classes, this study has only taken data from within the TWU SIT program, and is looking within the courses to find its results. This chapter will look further into the models created in the last chapter, and will explain and interpret the findings for both the models built for the math courses as well as the models built for the science courses.

Models for Math Courses

The predictive models that were built for the math courses in the last chapter provided some interesting information. Initially, the hypothesis of this study was that the more SIT sessions a student attended, the more likely he or she was to receive a successful ending course grade. In our model (4.9), $logit(p)$ is the odds that a student will succeed in the course, and as mentioned in the previous chapter, the *log odds* are modeled as a linear combination of the explanatory variables, where positive coefficients are associate with a higher probability that the student will succeed. We see from the regression equation (4.9) that although the independent variable SIT provides a positive coefficient, $\beta = 0.188$, it is not the

factor that has the largest impact on success. The independent variable MetMin, which describes if the student met the minimum required SIT sessions, has the biggest positive impact on the likelihood of success with a coefficient of $\beta = 1.653$. However the independent variable Risk, which describes if a student was ever at risk, has the largest impact of all of the factors on success, albeit negative, with a coefficient of $\beta = -3.025$. We can interpret this to mean that although attending SIT sessions does positively influence a students likelihood of succeeding in the course, the more important factor to consider is a student being labeled at risk. If a student is labeled at risk, his or her predicted *log odds* of a successful ending course grade is decreased by about 3 times what it would be if he or she wasn't labeled at risk, holding all other variables constant.

It is also of worth to mention that the independent variable Age was a significant factor in predicting the odds of success as well. With a coefficient of $\beta = -0.068$ it does have a negative impact, but a small one. Additionally, three independent variables were not entered into the model because they did not provide significant results. Ethnicity, Gender, and Class Rank, were all not significant at the $\alpha = 0.05$ level of significance, providing p-values of $p = 0.0813, 0.549, \& 0.333$ respectively, whereas SIT, Risk, MetMin, and Age were significant with p-values of $p = 0.000, 0.000, 0.000, \& 0.003$ respectively.

We can see from the tables below that the Cox & Snell R^2 indicates that 37.4% of the variation in the response variable is explained by the logistic model, and

Nagelkerke's R^2 of .544 posits a moderate relationship between the explanatory variables and response variable. Lastly, in examining the χ^2 and p-values of the model, we can see that with a p-value of $p = 0.000$, we reject the null hypothesis that states that this model does not do any better of a job at predicting the likelihood of success compared to a model with no predictors or independent variables.

Table 5.1

Omnibus Tests of Math Courses Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	545.038	7	.000
	Block	545.038	7	.000
	Model	545.038	7	.000

Table 5.2

Math Courses Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	808.231[a]	.374	.544

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Models for Science Courses

The predictive model built for the science courses was once again surprising when considering the initial hypothesis of this study, but the results followed closely

with the results from the math courses. SIT, Risk, and MetMin were once again the most influential predictors of the likelihood of success, with coefficients of $\beta = 0.062, -3.229, \& \ 2.225$ respectively. These coefficients also mimic the strength of the impacts that the same factors had on the math courses. SIT provided a positive, but small relationship to predicting success. However, Risk and MetMin both showed a much stronger relationship, with Risk negatively affecting success, and MetMin having a positive impact. Again, these results can be interpreted to say that if a student is at risk while taking a SIT supported course, it will have a strong negative impact on his or her likelihood of success, whereas if the student attends the minimum number of required of SIT sessions, the likelihood of succeeding in the course will increase.

One difference in the model of the science courses as compared to the model of the math courses is that age was not a significant predictor of success at the $\alpha = 0.05$ level of significance. Age, Ethnicity, Gender, and Class Rank all provided *p*-values greater than 0.05, and were therefore not significant, whereas SIT, Risk, and MetMin were less than 0.05 and were significant.

Lastly, we can once again examine the Omnibus and Model Summary table to observe the R^2 and χ^2 values. From the tables below, we see that the Cox & Snell R^2 provides a value of 0.439, which is a bit higher than that of the math courses, and tells us that 43.9% of the variation in the response variable is explained by the logistic model. Similarly, Nagelkerke's R^2 of 0.600 is just a tad higher than that of

the math courses. Finally, in examining the χ^2 and p-values of the model, we can see that with a p-value of $p = 0.000$, we once again reject the null hypothesis that the intercept and all coefficients are zero.

Table 5.3

Omnibus Tests of Science Courses Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	841.741	7	.000
	Block	841.741	7	.000
	Model	841.741	7	.000

Table 5.4

Science Courses Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1076.370	.439	.600

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

CHAPTER VI

CONCLUSION

Limitations

A few limitations of this study involve the variables or factors that cannot be (or were not) measured. For example, a student's motivation to do well in the course will have a large impact on his or her success; however, other than a survey (which would most likely provide biased results) motivation is difficult to measure. Next, a student's attendance and participation in class lectures could be useful. Although attendance can easily be collected, whether a student is paying attention and/or actively participating is much harder to measure. Lastly, a student's natural ability to comprehend the subject matter is another factor that could affect success, but again, is difficult to measure and control for.

It is of worth to note that the models built in this study were built as training models, and were not tested for their validity. This was done because the models themselves are not being used to predict the probability of success. Predicting a specific student's actual likelihood of succeeding in the course does not have a practical use, as data must be collected after the student has finished the course. The purpose of the logistic regression models built in this study was to analyze the

relationships between the independent variables and dependent variable, which was found by examining the coefficients of the models

Next, it is important to talk about the independent variable MetMin. A student in the TWU SIT program is required to attend one SIT session per exam (4 per semester); however, once labeled at risk, a student must begin attended SIT sessions once per week. Therefore, if a student was deemed at risk after the first exam, the required number of SIT sessions will increase from 4 to 9 (typically, this depends on which week of the semester the exam was given.) Similarly, if a student is not at risk after the first exam, but becomes at risk after the second, his or her number of required SIT sessions will increase from 4 to about 6. The test after which a student became at risk was not information that was available for this research; therefore, this study had to make do with what was available, and defined MetMin to mean 4 SIT sessions.

Results

Although the results from both the model for the math courses and the model for the science courses were not exactly aligned with the hypothesis that the more SIT sessions a student attended, the better his or her chance at succeeding in the course, the information provided by the logistic regression analysis is still useful and interesting. Knowing that the biggest (negative) indicator of success is if a student is at risk is useful for TWU SIT program officials. Also, the results of this study tell instructors and program officials that meeting the minimum number of

required SIT sessions is essential for a student to do well in the course. Lastly, it also shows that gender, ethnicity, class rank, and even (sometimes) age, have no effect on whether a student will end the course successfully.

REFERENCES

Annotated SPSS Output Logistic Regression. UCLA: Statistical Consulting Group. from http://www.ats.ucla.edu/stat/spss/output/logistic.htm

Arendale, D. (1994). Understanding the supplemental instruction model. In D. Martin & D. Arendale (Eds.), Supplemental instruction: Increasing achievement and retention (Vol. 60, pp. 11-21). San Francisco, CA: Jossey-Bass.

Arendale, D. (2002). History of supplemental instruction (SI): Mainstreaming of developmental education. In D. B. Lundell & J. Higbee (Eds.) *Histories of developmental education* (pp. 15-27). Minneapolis: Denter for Research on Developmental Education and Urban Literacy, General College, University of Minnesota.

Carlsen-Landy, B., Falley, B., Wheeler, A., & Edwards, D. (2014). Adaptations of supplemental instruction: Our course assistants wear many hats. Supplemental Instruction Journal, 1(1), 7-21.

Foltz, B. [Brandon Foltz]. (2015, March 8). *Statistics 101: Logistic Regression, An Introduction*. [Video File]. Retrieved from https://www.youtube.com/watch?v=zAULhNrnuL4&list=PLIeGtxpvyG-JmBQ9XoFD4rs-b3hkcX7Uu

Foltz, B. [Brandon Foltz]. (2015, March 9). *Statistics 101: Logistic Regression, Logit and Regression Equation*. [Video File]. Retrieved from https://www.youtube.com/watch?v=NmjT1_nClzg&list=PLIeGtxpvyG-JmBQ9XoFD4rs-b3hkcX7Uu&index=3

Foltz, B. [Brandon Foltz]. (2015, March 8). *Statistics 101: Logistic Regression Probability, Odds, and Odds Ratio*. [Video File]. Retrieved from https://www.youtube.com/watch?v=ckkiG-SDuV8&list=PLIeGtxpvyG-JmBQ9XoFD4rs-b3hkcX7Uu&index=2

Judd, C. M., & McClelland, G. H. (1989). Data analysis: A model-comparison approach. London: Harcourt Brace Jovanovich.

Logistic Regression Variable Selection Methods. (2012). Retrieved from https://www.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics.help/logistic_regression_methods.htm

Moore, D. S., McCabe, G. P., & Craig, B. A. (2009). Introduction to the practice of statistics. New York: W.H. Freeman.

Remediation: Higher Education's Bridge to Nowhere. (2012, April). Retrieved from http://www.completecollege.org/

Sribney, B., Harrell, F., & Conroy, R. (1998, May). Problems With Stepwise Regression. Retrieved from http://www.stata.com/support/faqs/statistics/stepwise-regression-problems/

Utts, J. M., & Heckard, R. F. (2015). Mind on statistics. Stamford, CT: Cengage Learning.

APPENDIX A

List of Variables

VARIABLE	DEFINITION	CODING
Dependent Variables		
Success	A passing grade of A, B, or C	(0) No (1) Yes
Grade	Final course grade	(1) A (2) B (3) C (4) D (5) F (6) W (7) I
Independent Variables		
Ethnicity	Student's Ethnicity	(1) White (2) Black (3) Hispanic (4) Asian (5) Hawaiian Pacific Islander (6) Native American (7) International (8) Other
Gender	Student's Gender	(1) Female (2) Male
Age	Student's Age	No coding, Student's age in *n* whole number years
Class Rank	Student's Class Rank	(1) Freshman (2) Sophomore (3) Junior (4) Senior (5) Post Baccalaureate (6) Master's (7) Doctoral
Risk	Was the student ever declared at risk	(0) No (1) Yes

SIT	How many SIT sessions attended	No coding necessary, number of SIT sessions attended in *n* whole numbers
MetMin	Was the minimum requirement of SIT sessions met	(0) No (1) Yes
Semester	Semester in which student was enrolled	(F13) Fall 2013 (S14) Spring 2014 (F14) Fall 2014 (S15) Spring 2015
Section	Course and section number in which the student was enrolled	(B1) BACT 1003.01 (B2) BACT 1003.02 (C1) CHEM 1013.01 (C3) CHEM 1013.03 (A1) MATH 1303.01 (A2) MATH 1303.02 (A3) MATH 1303.03 (A4) MATH 1303.04 (A5) MATH 1303.05 (S1) MATH 1703.01 (S2) MATH 1703.02 (S3) MATH 1703.03 (S4) MATH 1703.04 (S5) MATH 1703.05 (S9) MATH 1703.09 (S16) MATH 1703.16
Course	Course in which the student was enrolled	(B) Microbiology (C) Introductory Chemistry (A) College Algebra (S) Elementary Statistics I
Subject	Subject in which the student was enrolled	(SCI) Science – Microbiology and Introductory Chemistry (MATH) Math – College Algebra and Elementary Statistics I