THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENTS OF ECONOMICS AND STATISTICS

AN EMPIRICAL ANALYSIS OF DEMAND IN COLLEGIATE FOOTBALL

MICHELLE ANNE PISTNER
SPRING 2015

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Economics and Statistics
with interdisciplinary honors in Economics and Statistics

Reviewed and approved* by the following:

Peter Newberry
Assistant Professor of Economics
Thesis Supervisor

Russell Chuderewicz
Senior Lecturer of Economics
Honors Adviser

Murali Haran
Professor of Statistics
Honors Adviser

*Signatures are on file in the Schreyer Honors College.

Abstract

The Pennsylvania State University is home to a strong athletic program marked by fan loyalty. However, as with many sports programs, a natural problem arises as to how to predict purchasing habits of individual buyers. This paper analyzes attendance and individual demand for the university's football program. Linear regressions were used to model attendance on a game by game basis. This analysis concluded that opponent strength and past performance were the strongest predictors of attendance. Demand was further analyzed on the individual level. Purchasing habits for the year prior was found to be the strongest predictor of an individual's purchases for the next year. However, the usefulness of this information is limited. In order to determine new buyers and returning buyers, logistic regression models and decision trees were constructed. Their results yield a simple framework relying on past behavior and account information to predict between different buying patterns.

Table of Contents

List of Figures

List of Tables

Acknowledgements

First, I would like to thank my thesis supervisor, Dr. Peter Newberry, for his guidance and support throughout the entire process. Also, Dr. James Tybout, whose feedback and advice was invaluable.

I would also like to thank Dr. Andy Wiesner for introducing me to the staff at IMG Learfield Ticket Solutions and offering insight into my analysis. Also, I would like to thank my academic adviser, Dr. Murali Haran, for all of his guidance over the past few years. And, thank you to all of those at IMG Learfield Ticket Solutions and the Athletics Department at the Pennsylvania State University, including Lowell Berg, Anthony DiFino, and Jeff Garner, for allowing me to use their data for my analysis.

Finally, I would like to thank my parents, family, and friends for their support in all of my educational endeavors.

Chapter 1

Introduction

Determinants of attendance and demand were studied for the Pennsylvania State University's football program. These models were based on a combination of previous literature and new work. Previous literature had a focus on attendance on a holistic level since this data is easily available. This research, however, primarily studies attendance on an individual level.

First, a model of attendance on a game-by-game basis was completed. This model was based upon different game metrics recorded for all home football games from 2009 to 2014. As consistent with previous results, opponent strength was found to be a strong predictor of attendance. Opponent "novelty" also had a positive influence on attendance. A discussion of these results can be found in Section 5.1.

Next, individual attendance was studied. Initial investigations revealed that the best predictor of buying habits was past purchasing history. In order to refine these results, several purchasing groups were classified, and separate analyses were run. Several applications in nonparametric statistics were used to distinguish important buying differences between groups. These results revealed that non-buyers were statistically different from those who had bought season tickets at some point. Based on these initial investigations, decision trees and logistic regressions were computed. They revealed several distinct features between buying groups. These features rely on past purchasing habits, past donation habits, and account information. A discussion of these results can be found in Section 5.2 and Section 5.3.

The outline of the remainder of the paper is as follows. First, I will discuss all relevant literature. Second, I will discuss all data cleaning procedures and limitations. Third, I will discuss sampling methodology. Finally, I will discuss analyses and conclusions.

As a final note, the data is held by certain confidentiality limitations. As such, certain restrictions on reporting will be applied when necessary.

Chapter 2

Literature Review

2.1 A Review of Determinants of Attendance in Sports

Plenty of authors have discussed potential determinants of attendance for a variety of sports. Typically, as suggested by Donihue, Findlay, and Newberry (2007) in their study of spring training games in Major League Baseballs Grapefruit League, findings are categorized into three groups. These categories include: location and stadium factors, quality of game variables, and time and weather variables (Donihue et al., 2007). A discussion of relevant literature and supporting research surrounding each category follows.

2.1.1 Quality of Game Factors

It is typically suggested that quality of game factors have an overt influence on attendance. In their paper on spring training attendance, Donihue, Findlay, and Newberry found that the quality of the opponent and past championships all have a statistically significant positive effect on game attendance. Conversely, higher ticket prices were found to have a negative effect on consumer attendance. In their study of attendance in Major League Baseball, Pan, Zhu, Gabert, and Brown (1999) also found support for the idea that performance is directly related to attendance. Their results conclude that winning percentage for the home team leads to greater attendance figures.

Baade and Tiehen (1990) also studied the determinants of attendance in Major League Baseball. Their findings, which built off of previous work by Roger Noll, suggest that the single greatest influence of attendance is a teams on-field performance. While they concluded that different cities have different definitions of performance and nonperformance, the idea that "fans will not consistently support losers" remained strong throughout (Baade and Tiehen, 1990).

Price and Sen (2003) conducted an analysis of attendance for Division 1 collegiate football. Their results indicate a strong relationship between a teams performance and attendance (Price and Sen, 2003). Also, according to their results, the prestige and tenure of the program also has a positive influence on attendance. Somewhat intuitively, they find support that a strong rivalry with an opponent leads to surges in attendance.

Similarly, Welki and Zlatoper (1990) analyzed game-day attendance for teams of the National Football League. Their result also finds support for the notion that higher quality teams are associated with higher attendance levels. They found this result to hold true for attendance of fans for both the home and visiting squads. Additionally, they concluded that expectations about the game influence consumer attendance. Games with more excitement, such as those that are closer in score, were predicted to have higher attendance levels.

Paul and Weinbach (2011) studied attendance in the Quebec Major Junior Hockey League, finding results consistent with other previous work completed on other professional and non-professional sports leagues. They found that prior results on the relationship between attendance and quality of game factors held true. They concluded that stronger opponents have a positive outcome on attendance. Finally, their results supported the intuitive increase in attendance when winning percentage for the home team increases. Similarly, in his study of the National Hockey League, Paul (2012) concluded that games involving regional rivals lead to increases in attendance. Additionally, other quality of game factors, some of which are specific to hockey, had mixed outcomes on attendance. He determined that fighting has a positive result on attendance, but an increase in scoring had a negative result (Paul, 2003).

2.1.2 Location and Stadium Factors

In their study of Major League Baseball attendance, Pan, Zhu, Gabert, and Brown (1999) point to region population as a positive predictor of attendance. They also note that competing professional teams in the vicinity have negative effects on attendance. Baade and Tiehen (1990) also found evidence that city-specific variables, such as population, can predict attendance levels. Unsurprisingly, they noted that cities with more residents typically have higher levels of attendance.

2.1.3 Time and Weather Factors

As expected, the literature offers an explanation for the interaction of weather and attendance. Donihue, Findlay, and Newberry (2007) concluded that weather affects attendance in the predicted

manner, with colder games attracting less fans. Additionally, Paul and Weinbach (2011) discovered a significant relationship between game time and attendance. Weekend games attracted more fans, as did games closer to the playoffs. Paul (2003) reached a similar conclusion to Rodney and Weinbach, noting that weekend games had higher levels of attendance.

2.2 Discussion of Past Results on Studies of Collegiate Football

In addition to the previously mentioned study by Price and Sen (2003), several other researchers have examined various facets of collegiate football. Paul, Humphreys, and Weinbach (2012) conducted a study of the interaction of uncertainty and attendance in collegiate football. Their results indicate support that attendance is positively affected when less uncertainty about the game outcome exists. In other words, attendance increased when a clear outcome was expected (Paul, Humphreys, and Weinbach).

In a survey of attendants to collegiate football games, Palanjian (2012) found support for the entertainment value of events. He determined that fans are influenced more by entertainment as they attend an increased number of games (Palanjian, 2012). He also found evidence that this valuation was strongest among undergraduate attendants; faculty, graduate students, and alumni did not present the same high predisposition to other forms of in-game entertainment besides the actual sporting event. However, his results suggest that other sources, some stemming from tradition, help form the unique environment that is a collegiate football game.

Leonard (2005) studied the importance of geography on visitor attendance to collegiate football games. He concluded that distance was a strong factor in the determination of visitor attendance. Additionally, he found support for other predictors of attendance similar to those offered in other papers concerning other sports. These variables included metrics for the strength of a fan base and performance of both teams involved in the game.

2.3 A Review of The Pennsylvania State University Athletic Department

Since the estimation of the model is so reliant on a single program, a short discussion of the programs history and functioning is warranted. The athletics department has a storied past that attracts fans in the hundreds of thousands. Over its duration, a combined 73 national championships have been earned by its various sports teams (PSU, 2014). As such, sporting events are in high demand; attendances above 100,000 are not uncommon at football games.

In terms of pricing structure, the vast majority of sports in the program follow a traditional pricing strategy. There exists, however, one major exception: football. Prior to 2014, football tickets were priced according to a traditional model. However, with the advent of more efficient technology, variable pricing was adopted for single game tickets with the beginning of the 2014 football season. Seats for more desirable games, marked by in-conference opponents or visiting teams with stronger records from the season prior, were more expensive than seats for less desirable games. This difference gave more flexibility in setting prices.

Perhaps unique to collegiate sports is the use of donations. With the exception of personal seat licenses which are used by some professional sports teams, the use of donations to determine purchasing power is exclusive to the university level. At this particular school, a donation to the athletic department is mandatory for the right to buy season tickets for football games. A minimum donation is required for the right to purchase a single seat. As expected, higher donations are welcome and are used as methods to differentiate buyers. Increased donation levels offer many perks, including more desirable seats, parking spots that are closer to the stadium, preference to purchase additional tickets for a variety of sports, and "bowl right," or partiality to purchase playoff tickets. Thus, variations in donation level can be attributed to one or more factors; these factors are unique to an individual and based upon their preferences.

For the following analyses, the demand of football season tickets will be discussed. While other sports will not be studied directly, their impact on consumer choices in purchasing football

season tickets will be of interest.

Chapter 3

Data Source and Description

3.1 Description of Data Source

Data used in this study was gathered from the Athletics Department at the Pennsylvania State University. This data source, which contains information about all sporting events attended and purchases for years past, encompasses information on over 500,000 individuals. At a minimum, a given individual contained in the database purchased tickets for an athletics or entertainment event hosted through the university within the past twenty years. All model estimations must be conditioned on this connection. However, the overall influence of this connection is relatively small due to the fact that many observations utilized in model estimation have had no contact with either the entertainment or athletics ticketing office within the past ten years.

A variety of information was present within the data set. Variables existed for both single game and season tickets purchases for all major sports. These variables included: purchase time, purchase cost, purchase medium, seat location, and attendance. Additionally, variables for donation levels to the athletics department and mailing address of the purchaser were available for study.

3.2 Determinants Applicable to the Data Source

While the nature of the data prevents some traditional measures of attendance from the literature to be applicable, there are several variables that retain importance. To determine a given individuals demand for collegiate football at this university, past purchasing history of football and other sports can be utilized. Additionally, the levels of donations to the athletics department can be utilized as a proxy for an individuals valuation of the athletics department. Game specific factors, such as time of game and strength of opponent, can also be used to measure the quality of a game, a factor suggested by much of the previous research as having a positive influence on attendance.

3.3 Limitations of Data Source

As with any data source, certain limitations exist. First, a minimal level of demographic data is available for study. For example, there is no direct measure of the purchaser's association with the school. This would include metrics such as whether or not a buyer previously attended the school, knows a current student, or is employed by the school. Thus, this relationship will be modeled by other proxy variables. The quantification of this relationship could potentially classify variation that is not accounted for in other metrics. However, an individuals past purchase history could serve to quantify the strength of relationship between the purchaser and the school or program.

Second, the estimated models are based on data collected from a single athletic program. This fact provides a straightforward limitation. The applicability to other programs is of question and needs to be verified. Model estimations were based on data from a large state university with a noted athletics program. It is easy to see that the resulting conclusions might not be applicable to small liberal arts colleges with smaller or newer athletics departments.

Chapter 4

Methodology

4.1 Initial Data Cleaning

In order to complete the analysis, certain observations had to be adjusted or excluded. First, the analysis was restricted to personal buyers only. This excluded accounts for recruitment, players, opponents, ground staff, game-day employees, and ticketing agencies. Removing these observations is justifiable: it is reasonable to conclude that these accounts operate on different mechanisms for purchasing tickets and attending games than the traditional buyer. Additionally, they were easily identifiable as each account has an associated type recorded with it.

The database was further restricted to addresses within the United States. A variety of observations existed for other countries and territories, including Canada, England, and Puerto Rico. These accounts were removed because the attendance at football games faced another hurdle that was not present with the rest of the accounts: international travel.

In addition to these restrictions, all accounts directly related to the university were removed. This included accounts related to university administrators, other commonwealth campuses, and other departments located within the university. This restriction did not exclude personal accounts of university faculty or staff. Again, it is reasonable to conclude that the purchasing habits of these account holders differ drastically from other buyers. They were easily to quarantine based on information for several variables, including account holder name and account type.

The final modification that was made included disregarding any accounts that involved the purchase of suite-level seats for football. These seats typically require that large sums of money be donated to the school, typically 200 times the donation amount required for season tickets for a single seat. Therefore, these purchases are clearly outliers when compared to the average buyer.

Only these restrictions on variables were used. While these restrictions eliminated only a small fraction of the overall databased, they were deemed significant to maintain the validity of the analysis.

Unless otherwise noted, distances were calculated using the Haversine equation. All distances are calculated to the zip code location of the stadium. Information on latitudes and longitutes was

obtained from a database obtained by Tom Boutell. All estimates for economic variables are at the zip code level. This data was obtained from the Economic Research Division of the United States Department of Agriculture and the Internal Revenue Service.

4.2 Overview of Analyses

First, a model of attendance was constructed on a game-by-game basis using multiple linear regression. Results for this model can be found in Section 5.1. In order to determine individual specific factors that determine attendance, a logistic model of demand was fitted using data at the individual level. This model was deemed appropriate since it can estimate which characteristics increase the likelihood of purchasing season tickets for the football program for a given year. A description of these results can be found in Section 5.2.

To further explore the results obtained in Section 5.2, decision trees and logistic regressions were used to create algorithms to identify new football season ticket holders from the remainder of the non-buying database population. A similar framework was also used to determine non-renewing football season ticket holders from the rest of the football season ticket holder population. A description of the results can be found in Section 5.3.

Chapter 5

Data Analysis and Results

5.1 A Model of Attendance

Attendance on a per game basis was studied in order to determine important determinants of demand. An outline of the analysis and results follows.

5.1.1 Sampling Schema

In order to analyze the determinants of attendance on a game by game basis, variables were collected for games from past seasons. All home games between 2009 and 2014 were included in the sample for a total of 43 games. Variables were recorded for several different dimensions of the game, including opponent strength, price, time of year, and year-by-year changes that occurred for the football program. On-field records for the previous year were collected for both the opponent and the home team. Rivalries were recorded for games against the Ohio State University and the University of Michigan. Opponent conference was also recorded. Time of day information and weather related variables were also noted. Distance between each university was obtained via Google Maps. Recruiting rankings for each year were obtained from ESPN. Proxies for head football coach were noted, and, finally, price was recorded at the amount of the minimum single game ticket price.

5.1.2 Model of Attendance

Several multiple linear regressions were calculated in order to model attendance for each individual game. All models and statistics were calculated using R using ordinary least squares estimation. A potential issue does arise since attendance has a maximum value of approximately 108,000. Several estimation techniques, such as Tobit estimation, could be used to overcome this issue. However, as long as all of the assumptions of multiple linear regression are satisfied, the estimates will still be correct. The censored nature of the data, however, does still warrant a discussion. The results for the statistically preferred model can be found in Table 5.1. The full model can be found in Table 5.2. Assumptions were verified for each of these models, and a complete

discussion can be found in Appendix A.

Table 5.1: Preferred Model of Attendance

Dependent variables	**Coefficient**	**Standard error**	**T-statistic**
Big Ten opponent	1858.3	928.0	2.00
Rival opponent	6492.6	1335.7	4.86^{**}
Penn State record for year prior	17035.1	3294.7	5.17^{**}
O'Brien coach indicator variable	-7263.6	813.1	-8.93^{**}
Power conference indicator variable	6436.1	1891.3	3.40^{**}
Distance between schools	3.73	1.7	1891.3^{*}
Constant	87995.7	2614.8	33.7^{**}

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

Table 5.2: Full Model of Attendance

Dependent variables	**Coefficient**	**Standard error**	**T-statistic**
Night game indicator variable	-2137.8	1463.4	-1.46
Winter indicator variable	-1118.7	1073.2	-1.04
Big Ten indicator	2015.5	1107.9	1.82
Rival opponent	7472.6	11733.9	4.31^{**}
Opponent record for year prior	-446.7	2075.6	-0.22
Penn State record for year prior	24293.4	5887.5	4.13^{**}
Franklin coach indicator variable	2137.1	1680.6	1.27
O'Brien coach indicator variable	-7213.3	1141.1	-6.32^{**}
Recruiting class ranking	41.8	45.9	0.91
Minimum price	24.0	69.9	0.34
Distance between schools	5.80	2.16	2.68^{*}
Power conference indicator variable	6091.1	2042.4	2.98^{**}
Constant	79529.9	7729.9	10.29^{**}

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

For the preferred model, the R^2 statistic was 81.12%. The adjusted-R^2 statistic was 77.97%. For the full model, the R^2 statistic was 83.12%. The adjusted-R^2 statistic was 76.37%.

The preferred model of attendance displays many consistencies with the analyses conducted in the literature review. Game quality appears to be a major factor in determining attendance. Rival opponents, defined as the Ohio State University and the University of Michigan, lead to significantly stronger attendance numbers. In-conference opponents also lead to slightly stronger attendance; however, this affect was not found to be statistically significant. Out-of-conference opponents that belonged to a powerhouse conference also had statistically significant increases in attendance.

Weather was found to not have as strong as an influence on attendance. The reason for this might be multi-faceted. First, fans could be immune to weather effects. Second, it could also be that stronger games may typically occur later in the season, meaning that quality of game factors outweigh weather factors.

Penn State's on-field performance seemed to be a better indicator of attendance when compared to off-season indicators of performance, such as recruiting class ranking.

Coaching variables were also included to account for personnel changes and coaching styles. When compared to Joe Paterno, Bill O'Brien brought in statistically smaller crowds whereas James Franklin was estimated to bring in larger crowds although this result was not statistically significant. I propose that this phenomenon is not a result of the individual coach, but rather is a result of the phases of the post-scandal football program. O'Brien represented the immediate fallout of the scandal, whereas Franklin represents the transition of the program back to a somewhat normal state.

Price was included to determine consumer price sensitivity. Price was found to be non-significant, and, therefore, I conclude that demand for this football program is relatively inelastic. Endogeneity might appear to be an issue here as a higher price might be issued for games with stronger predicted attendance. However, this issue is minimal since, prior to 2014, prices were equally assigned for all games of the season rather than by game-specific factors.

Finally, distance between the two competing colleges was found to be statistically significant. The inclusion of this variable was two-fold. First, games with geographically closer opponents might warrant higher visitor attendance. Second, games with more distance opponents might signify the rarity of the opponent. For example, games against Temple University are more common than the University of Minnesota. As such, there is less opportunity to attend a game with an opponent from further away. Finally, distance might also serve as a proxy for opponent strength. Opponents from further distances are more likely to be stronger, whereas opponents from closer distances might include weaker Division I football programs.

The preferred model above was used to generate predictions for individual games for the 2015. Prediction intervals for attendance was calculated for each game. A summary of the results can be found in Table 5.3.

Table 5.3: Attendance Prediction Intervals for the 2015 Season

Opponent	**Fitted Value**	**95% Prediction Interval**
University of Buffalo	97,919	[92,565, 103,274]
Rutgers University	99,875	[94,424, 105,325]]
San Diego State University	106,649	[97,643, 115,655]
United States Military Academy	104,523	[98,230, 110,817]
Indiana University	101,058	[95,786, 106,329]
University of Illinois	101,323	[96,062, 106,583]
University of Michigan	106,964	[101,292, 112,636]

There are several things to note about the following predictions. First, the total capacity of Beaver Stadium is 107,282. Upper bounds greater than this value would signify a sell-out. Second, for the specific case of San Diego State University, the model was estimated on data from 2009 to 2014, and, coincidentally, no opponent from these years were from a distance roughly greater than 1,000 miles. As such, I believe inappropriate extrapolation of distance might be occurring in this prediction and would expect attendance to be much lower.

5.2 Logistic Regression to Model Season Ticket Purchases

For the remainder of the paper, statistical modeling is achieved primarily by using logistic regression. Based upon the binomial distribution function, logistic regression offers a way to estimate a binary variable using categorical or continuous predictor variables. It is preferred over traditional linear regression when the response variable takes on one of two possible values. This model is estimated in terms of the logit function, which can be written as:

$$F(x) = \frac{1}{1+e^{-(\beta_0+\beta_1*x_1+...+\beta_n*x_n)}}$$

where the fitted value, $F(x)$, represents the probability of a "success" given a combination of predictors. The term $\beta_0 + \beta_1 * x_1 + ... + \beta_n * x_n$ represents a linear combination of predictor variables. Due to the construction of the logit function, direct interpretation of coefficients is not straightforward. Instead, these interpretations are based upon the odds ratio. For example, given a one unit increase in the value of x_n, the odds of a "success" would multiply by a factor of e^{β_n}.

For all logistic regressions, coefficients were obtained via iteratively reweighted least squares, and all coefficients represent maximum likelihood estimates. Corresponding hypothesis testing assumed coefficients were equal to zero under the null hypothesis.

Iteratively reweighted least squares is preferred to ordinary least squares in estimating coefficients for logistic regression since the estimates are more robust. The process results in better coefficient estimates, although it is computationally intensive. To start, initial estimates of the coefficients are obtained, typically through the ordinary least squares method. Then, these coefficients are used to obtain fitted values which, in turn, are used to calculate weights and a new response variable. The new response variable is then regressed on the set of predictor variables, taking the calculated weights into account (Kutner et al., 2005). This process is repeated until convergence is achieved.

Purchasing habits in relation to football season tickets were studied on an individual basis. An outline of the analysis and results follows.

5.2.1 Sampling Schema

Instead of employing a traditional random sampling scheme, stratified random sampling was employed. Stratified random sampling mitigates the potential for biases that arise by chance under a traditional simple random sampling scheme (Kutner et al., 2005). While these biases might not necessarily occur, stratification prevents the possibility of this type of bias altogether. It eliminates the potential that certain subgroups are over or under-represented in the sample as compared to the total population of interest.

For this analysis, the sample was stratified based on location. Billing zip codes were truncated to include the first three digits only. This mechanism was used since it provides a suitable amount of geographic information without the computational complications of a strategy that uses individual zip codes. For example, Figure 5.1 and Figure 5.2 display zip codes maps when considering both the three and five digit versions. Using the entire code might overcomplicate the stratification, potentially providing inconsistent trends. Perhaps, instead of showing fans by region or city, it would depict fans by socioeconomic areas within these regions.

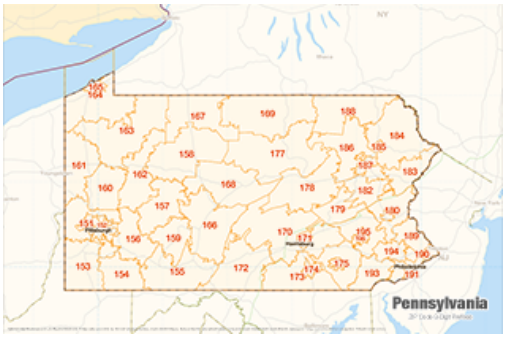

Figure 5.1: Three Digit Zip Code Boundary Map for Pennsylvania

Source: “3-Digit ZIP Code Prefix Boundary Reference Map of Pennsylvania”

Figure 5.2: Zip Code Boundary Map for Pennsylvania

Source: ”United States Zip Codes”

After restricting the data as described in Section 4.1 and truncating zip codes, incidence percentages were calculated for each region present within the database. Then, the total number of observations necessary for each region was calculated such that the sample would roughly amount to 75,000 observations. Then, using a normal distribution random number generator package in Stata, enough accounts for each three-digit region were obtained. From there, the rest of the analysis was performed.

Once restrictions were completed as outlined above, there were 72,883 observations available for analysis.

5.2.2 Exploratory Data Analysis

An initial data analysis of the database was conducted in order to determine important trends and correlations. A detailed description of the exploratory data analysis is discussed in Appendix B, and it will be discussed in brief here.

Approximately 2.4% of the accounts within the database purchased season tickets for football in 2014. This number increased to 3.3% for the 2015 season. These accounts represented the bulk of donations to the athletic department. For 2015, the median donation was the baseline amount required by the athletics department for the right to buy football season tickets. The mean donation was approximately five times that amount, suggesting the data has a positive skew.

Summary statistics for total distance between the billing zip code and the zip code of the football stadium were calculated. The results were relatively homogeneous between season ticket holders and non-season ticket holders. The mean distance from the stadium was 166 miles with a standard deviation of 377 miles.

5.2.3 Model of Season Ticket Purchases

Several logistic regressions were fit from the available data, and a suitable model was chosen based on fit statistics and included parameters. The resulting model can be found below in Table 5.4. All logistic regressions were completed using Stata.

Table 5.4: Logistic Regression of Demand Results

Dependent variables	**Coefficient**	**Standard error**	**Z-statistic**
2014 football season ticket holder	15.100	0.610	24.83^{**}
Distance (10s of miles)	0.003	0.001	2.55^{*}
Mean income (1000s of dollars)	0.002	0.001	2.66^{**}
2014 football single game buyer	8.500	0.410	20.6^{**}
Constant	-9.600	0.410	-23.2^{**}

Independent variable is season ticket holder status for the 2015 football season(0=no; 1=yes).

Football ticket holder status variables follow format of 0=no and 1=yes.

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

Several statistics were calculated to analyze the fit of the model. The model boasted a sensitivity of 99.8% and a specificity of 99.1%. Thus, the probability type I error of the model was 0.93%, and the false discovery rate was calculated as 0.28%. These statistics are based on the assumption that a fitted value greater than 50% is denoted as a "success," i.e. buying football season tickets. This guideline will be used for the remainder of the paper.

Further fit statistics suggest that this model is appropriate. The likelihood ratio chi-squared statistic is 18,357, and the associated p-value is equal to 0.000. This is suggestive that the calculated

model fits better than an empty model. The pseudo-R^2 statistic is 87.00%. Further fit statistics are discussed in Appendix C.

As expected, the single greatest predictor of purchasing football season tickets for a given year is whether or not tickets were purchased the year prior. Purchasing season tickets for the year prior increased the log odds of buying season football tickets by 15.1, all else constant. Purchasing single game tickets for the year prior increased the log odds of buying season football tickets by 8.5.

An increase in median income of the billing zip code area also increased the likelihood of buying tickets. For example, an increase in median income by $5,000 increases the probability of purchasing football season tickets by 1%, all else constant. Finally, the result for distance may prove surprising. However, this is primarily due to the collection of the database. Since the overall restricted population consists of those who came in contact with the ticketing office for a variety of reasons, it is reasonable to conclude that those who live at further distances are more likely to purchase tickets for athletic events and less likely to purchase tickets for other entertainment events.

Furthermore, it must be remembered that certain limitations exist on variables available for analysis. There is no absolute ways to capture an individuals fandom. Other variables, such as alumni status, could serve to proxy this idea, but, for the purposes of this study, the data is unavailable. However, perhaps similar to many market pricing theories, an individual's predisposition to purchase tickets is already indicated by past behavior (Beechey et al., 2015). This idea could be responsible for the strong relationship between past purchasing habits and current purchasing habits.

With such a strong connection between demand and past purchasing habits, further analyses to predict what causes consumers to buy football season tickets were conducted. The results are presented in the next section.

5.3 Determination of Potential Season Ticket Buyers

Since past purchasing habits was found to be the greatest predictor of current purchasing habits, models were constructed to determine how consumer choice influences consumer decisions. A summary of the procedure and results can be found below.

5.3.1 Sampling Schema

A separate sample was draw from the database in order to complete the analyses. Again, stratified random sampling was used. Individuals were partitioned into several distinct groups. They included: new football season ticket holders for 2015, non-renewing season ticket holders for 2015, returning season ticket holders for 2015, and non-buyers for both 2014 and 2015. The sample was stratified based on these groups so that the effects of the less-represented groups in the database would not be over-crowded.

In total, approximately 1,500 observations were randomly drawn from each group. This resulted in a total sample size of 5,998 individuals.

Since oversampling could represent biases in model estimation, appropriate corrections were used. As outlined by King and Zeng (2001), oversampling biases the estimates of the intercept for logistic regressions while leaving stable estimates for the coefficients of independent variables. In order to correct for these biases, they suggest the following:

$$\hat{\beta}_0^{\,corrected} = \hat{\beta}_0 - \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]$$

where τ represents the true population proportion of the oversampled event and $\bar{y}$ represents the sample proportion of the oversampled event.

5.3.2 Model Analyses

As discussed in the previous section, the most important predictor in purchasing habits was purchasing habits in the previous year. While this result has some practical applicability, it is not useful in determining new season ticket holders from the total population of non-buyers within the database. Also, it is not useful in determining which season ticket holders will not renew from the population of total season ticket holders.

In order to determine these two groups from the rest of the database population, decision trees were constructed. These trees utilize conditional probabilities in order to make informed decisions on consumer behaviors. Due to oversampling techniques, these resulting probabilities are biased. However, the structure of the decision tree remains valid as research suggests that oversampling improves the overall accuracy of the results (Lavery, 2015). For example, consider a population where 5% of all cases are deemed a success, and 95% of all cases are non-successful. In order to maximize the computing efficiency of decision tree algorithms, a simple oversampling technique is used such that 50% of all sample observations are a success. Each non-success observation in the sample is representative of approximately 95 non-success observations in the population As a result, the node probabilities are biased.

In this context, they were preferred to more traditional techniques due to their robustness to outliers and practicable applicability. Decision trees were constructed using SAS Enterprise Miner, a data mining suite that is used to supplement the base SAS software. The program can be used for a variety of data mining applications, such as decision trees, neural networks, and component analysis.

All trees included in this paper were constructed manually using a combination of statistical validation, such as the LogWorth statistic, and past economic theory on consumer purchases. In SAS Enterprise Miner, LogWorth is defined as the negative logarithm of the p-value that is computed from the Chi-Squared test (SAS, 2015). As such, higher values of LogWorth are associated with better predictors for the given dependent variable.

In some instances, decision trees were calculated automatically; this was used to develop marketing recommendations related to a single variable. They were created using the Classification and Regression Tree algorithm (CART), a method that was pioneered by Leo Breiman (Breiman et al., 1984). As with most decision tree algorithms, this method works by partitioning the data into several rectangular spaces. In SAS Enterprise Miner, these partitions are determined via the Gini impurity index, a measure of node impurity. After a tree is constructed, cross-validation, or "pruning", is automatically conducted by SAS Enterprise Miner. This step removes unneccesary tree nodes in order to minimize misclassification. For this analysis, 20% of each data set was used to validate "pruning" decisions.

Based on the information obtained from the decision trees, logistic regressions were computed. All logistic regressions were calculated using Stata.

Several groups were defined in the analysis. They include: new season ticket holders for 2015, returning season ticket holders for 2015, non-renewed season ticket holders for 2015, and non-season ticket holders for either 2014 or 2015. In order to guide variable selection, several statistical analyses were performed to determine which variables contained potentially useful predictor information between the groups. Kruskal-Wallis tests were preferred to a traditional analysis of variance since the assumptions of normality appeared to be violated upon initial inspection. A complete discussion of these tests can be found in Appendix D, and the results will be discussed in brief here.

Several variables were studied, including consumer distance from the stadium, money spent on other athletic events, total athletic donations paid for the 2014 season, total spent on single game football tickets in 2014, and athletic account length. With the exception of athletic account length, each test yielded a significant result, indicative of differences between at least one of the groups. Multiple comparisons were calculated for each test, and the results can be found in Appendix D.

A decision tree model was constructed in order to predict new season ticket holders from the database population of non-buyers. The results can be found in Figure 5.3.

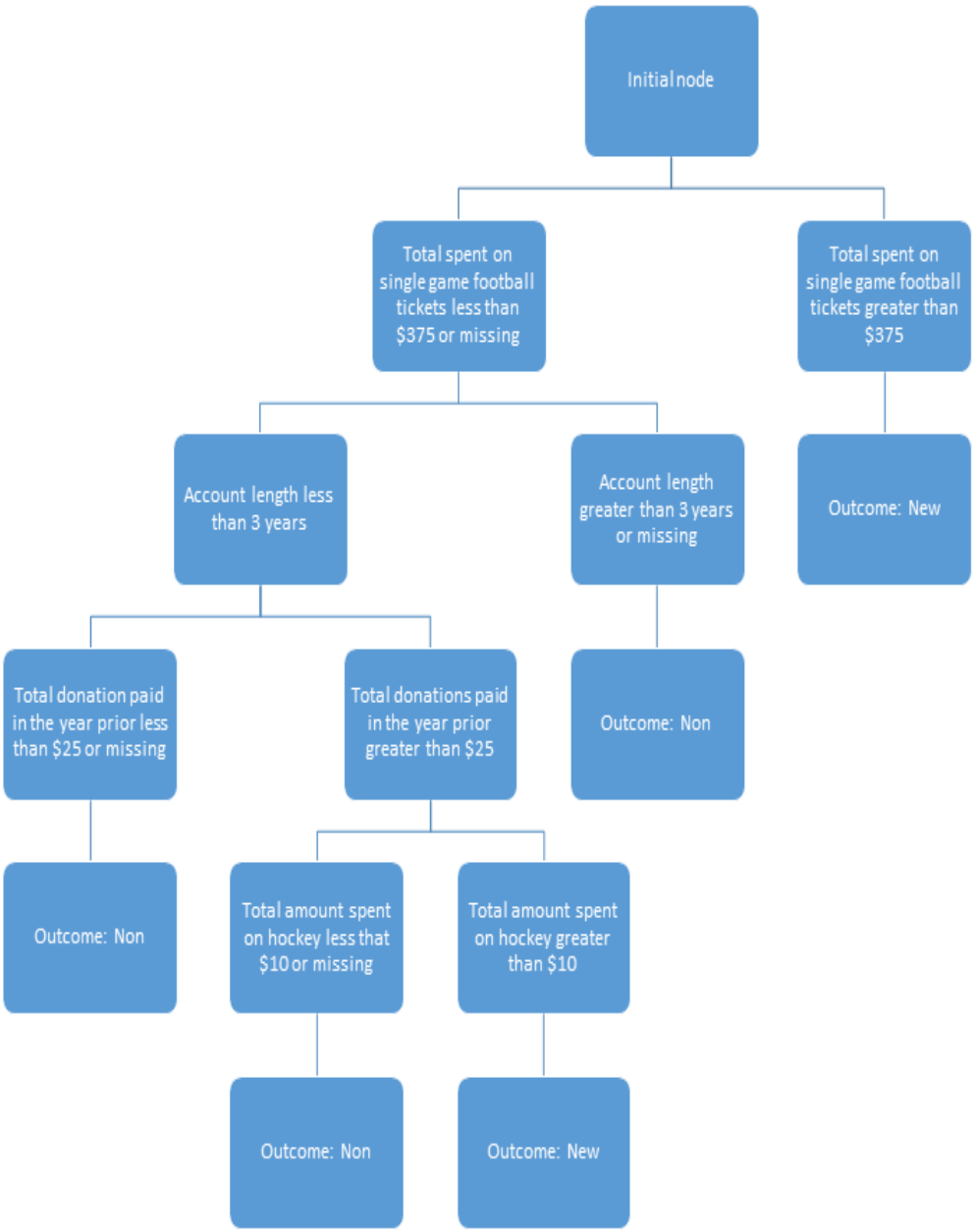

Figure 5.3: Decision Tree to Predict New Season Ticket Holders

This decision tree presents excellent predictive power between the new season ticket holders and non-buyers. While still constructed on over-sampled data, this technique does not require any corrections. Rather, it must be noted that percentages in each node are not representative of the actual population. The model's predictive power, nonetheless, remains valid.

The tree reveals somewhat intuitive results. First, those who spend more money on single game football tickets are more likely to become new season ticket holders. As such, it is possible

that some new season ticket holders devote substantial money to the program before making the long-term commitment. Second, account length also seems to play a role in purchasing decisions. Again, it offers a simple explanation. Those who have newer accounts had more recent contact with the department, and, are more likely to purchase tickets. On the other hand, those with longer accounts typically had a longer time to express interest in purchasing. Finally, donations for the year prior also offers some predictive power. This could be due to the fact that donations increase the total number of "points" which are responsible for determining seat locations, among other things.

Lastly, a logistic regression was created to determine new season ticket buyers from the database population of non-buyers. The statistically preferred model can be found in Table 5.5. The full model can be found in Table 5.6. Further analyses were conducted to determine if each model was appropriate; a discussion of these results can be found in Appendix E.

Table 5.5: Preferred Logistic Regression to Predict New Season Ticket Holders

Dependent variables	**Coefficient**	**Standard error**	**Z-statistic**
Total dollars paid for donations in 2014	0.0151	0.0041	3.68^{**}
Total dollars paid for single game tickets in 2014	0.0013	0.0006	2.44^{*}
Account length in years	-0.2458	0.0271	-9.09^{**}
Constant	-1.1181	0.1453	-7.703^{**}

Independent variable is season ticket holder status for the 2015 football season(0=no; 1=yes).

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

For the preferred model, the pseudo-R^2 value was 22.80%. The sensitivity of the model was 69.23%, and the specificity of the model was 99.73%. The probability type I error of the model was 30.77%, and the false discovery rate was 2.00%.

Table 5.6: Full Logistic Regression to Predict New Season Ticket Holders

Dependent variables	Coefficient	Standard error	Z-statistic
Total dollars paid for hockey in 2014	-0.0042	0.0079	-0.53
Total dollars paid for wrestling in 2014	0.0029	0.0040	0.72
Total dollars paid for single game tickets in 2014	0.0014	0.0005	2.58^{**}
Total donations paid in 2014	0.0159	0.00411	3.86^{**}
Account length in years	-0.2507	0.0271	-9.25^{**}
PA indicator variable	0.3162	0.2535	1.25
Median household income for buyer zip code	-4.87e-6	7.54e-6	-0.65
Constant	-1.0456	0.5547	-1.89

Independent variable is season ticket holder status for the 2015 football season(0=no; 1=yes).

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

For the full model, the pseudo-R^2 value was 23.26%. The sensitivity of the model was 69.23%, and the specificity of the model was 99.73%. The type I error was 7.04%, and the false discovery rate was 30.77%.

The preferred model yields an identical pattern of results as the decision tree. As such, the reasoning behind each coefficient remains identical to those discussed before.

For the full model, many variables appeared to have limited predictive power. The same relationships between account length, donations paid, and single game purchase amount remains consistent with the previous models. Total dollars spent on both hockey and wrestling seem to have negligible effects. There is no evidence of a difference between state of residence for the buyers. Additionally, there is no evidence that income has predictive power. This could be suggestive that those who purchase tickets view them more as a necessity. Also, it could be a by-product of the fact that income data is on the zip code level and not the individual level.

Further inspection revealed information for maximizing marketing strategies for obtaining new season ticket holders. Donations to the athletic department in excess of $40 in the year prior to football purchase were indicative of an increase in the probability of buying for each consumer

choice by 69.68% compared to those who donated less. The increase in log odds is 3.42, and the 95% confidence interval for the log odds is [2.60, 4.23]. A potential explanation for this is the points structure of season ticket purchases in the athletic department. Points are accrued through donation history, and they dictate much of an individual's purchasing capability, including seat location within the stadium and purchases for other sports.

Purchasing habits for single game football sales for the year prior to purchase also led to differentiation among consumers. Single game purchases of football tickets in excess of $40 were indicative of an increase in the probability of buying for each consumer choice by 18.00% compared to those who spent less. The increase in log odds is 1.63, and the 95% confidence interval for the log odds is [0.99, 2.28]. There are several potential explanations for this phenomenon. First, those deciding whether or not to commit to season tickets might be more likely to purchase football tickets to several games first. Second, it could capture the effect of those who purchased mini-game plans and later converted those plans to traditional season tickets.

Money spent on other sports also further differentiated new season ticket holders. By spending more than $5 on other athletic department events, the total probability of buying for each consumer choice increased by 17.55%. The increase in log odds is 0.98, and the 95% confidence interval for the log odds is [0.42, 1.54].

Distance also yielded some differentiation. Consumers who lived more than 70 miles from the football stadium were indicative of an increase in the probability of buying for each choice by 5.63%. The increase in log odds is 0.665, and the 95% confidence interval for the log odds is [0.25, 1.09]. While this result may seem non-intuitive, it warrants a simple explanation. The population of study includes only the members of the database. As location from the stadium increases, there are less reasons other than athletic events to be included in the database. Conversely, those who live closer may be included in the population due to interest in other entertainment events, not just athletic events.

Next, a decision tree model was constructed in order to determine returning football season ticket holders from the population of all season ticket holders. The results can be found in Fig-

ure 5.4.

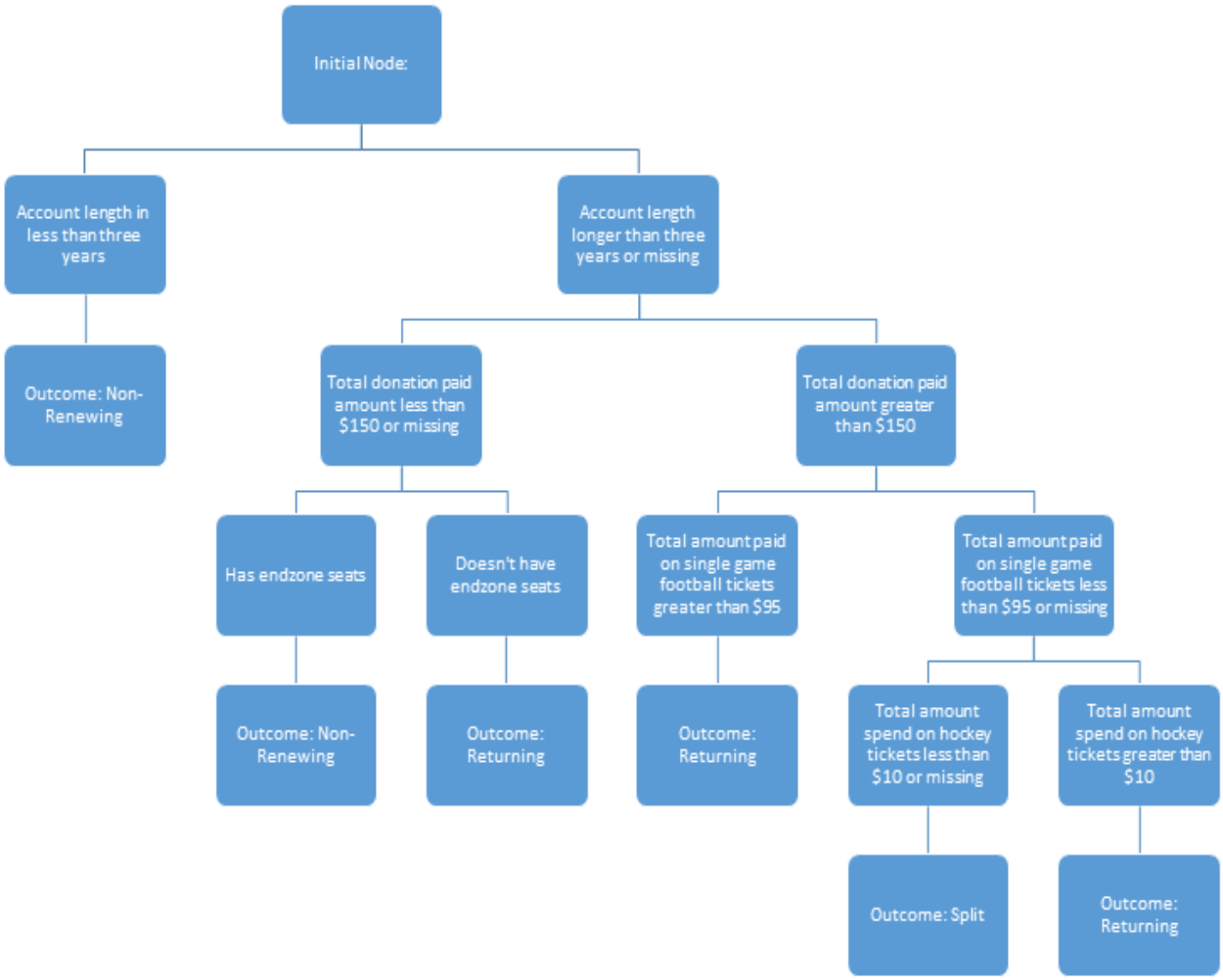

Figure 5.4: Decision Tree to Predict Returning Season Ticket Holders

The decision tree resulted in the best predictive model to predict non-returning season ticket holders. Again, as discussed with the previous model, no bias corrections are needed due to over-sampled data. However, the probabilities in each node have no direct use. This model is suggestive that donations, money spent on single game football tickets, account length, and money spent on other sports can be used to determine non-renewing accounts. It is suggestive that those with longer

accounts are more likely to return. Furthermore, within these long-standing accounts, those with lower donations are less likely to return. Finally, those who spend less on any athletic sport are less likely to renew when compared to their higher spending counterparts.

Lastly, a logistic regression was created to determine non-renewing season ticket buyers from the database population of season ticket buyers. The results can be found in Table 5.7. Further analyses were conducted to determine if the model was appropriate; a discussion of these results can be found in Appendix E.

Table 5.7: Full Logistic Regression to Predict Returning Season Ticket Holders

Dependent variables	**Coefficient**	**Standard error**	**Z-statistic**
Total dollars paid for hockey in 2014	0.0005	0.0002	1.71
Total dollars paid for wrestling in 2014	0.0010	0.0007	1.35
Total dollars paid for single game tickets in 2014	0.00015	0.00005	5.82^{**}
Total donations paid in 2014	0.0003	0.00005	5.53^{**}
Account length in years	0.0317	0.0033	9.61^{**}
PA indicator variable	0.0298	0.1137	0.26
Median household income for buyer zip code	-4.85e-6	3.02e-6	-1.61
Endzone seats dummy variable	0.0284	0.0866	0.33
Upper level seats indicator variable	-0.3259	0.0824	-3.96^{**}
Constant	-4.00	0.5270	-7.59^{**}

Independent variable is season ticket holder status for the 2015 football season(0=no; 1=yes).

* indicates significance at 95% confidence.

** indicates significance at 99% confidence.

For this model, the pseudo-R^2 value was 8.09%. The sensitivity of the model was 68.27%, and the specificity of the model was 59.72%. The type I error was 40.28%, and the false discovery rate was 31.73%.

This full model reveals that it is extremely difficult to determine non-returning season ticket holders from the rest of the group. Variables similar to the decision tree noted statistical significance, albeit at an uninformative level. These results support information obtained in primary

data analyses, such as the Kruskal-Wallis tests, which discovered that the data presents very little differences between non-returning, new, and returning season ticket holders.

The decision tree, however, still appears informative. This could be due to the relatively robust construction of this measure or its nature of conditional expectancies.

Chapter 6

Discussion of Results

The Pennsylvania State University boasts a storied athletic program that has documented success in many sports. This paper focused primarily on the football program. I analyzed various different facets of the program, including attendance and individual purchasing habits.

First, a general model of attendance was created. The models are discussed in detail in Section 5.1. Data was collected for every home football game between 2009 and 2014. Although the overall sample size is small, it is important to remember that the overall population size is relatively small when considering all comparable games. It is not accurate, for example, to utilize games that occurred in a home stadium with much smaller capacity.

Several important consistencies were revealed by these models. First, attendance is greatly influenced by the strength of the opponent. In-conference games and "rival" opponents all naturally lead to higher attendance figures. Second, weather expectations appear to have minimal influence on attendance. However, this might be due to the fact that stronger teams are typically played towards the end of the season, and opponent strength is a stronger determinant of attendance the weather. Next, distance between the two schools seemed to have an effect, although in somewhat a counter-intuitive manner. However, a simple explanation exists. I propose that distance serves as another measure of game quality. Closer opponents are more likely to be lesser-known Division I football programs that those from across the country. Also, price seems to have no effect on attendance, suggesting that demand for football is relatively inelastic. Endogeneity issues might be suggested; however, this seems to be of minimal effect since, prior to 2014, prices for a given seat remained consistent throughout the season. Lastly, these models indicate that the scandal's effect on football attendance is on the decline. While the scandal did decrease attendance, the new coaching era has seemed to reverse the trend. In order to make this association stronger, another season of attendance data would need to be collected and incorporated into the model.

After examining demand on a per-game basis, it was analyzed on the individual level. Logistic regressions revealed that the most important factor in buying habits in a given year was the previous years actions. As such, individuals have internalized their decisions to purchase tickets and made decisions accordingly. This results is, however, not useful in determining new season ticket holders

or marketing strategies.

The last part of this paper focused on determining new football season ticket holders and returning football season ticket holders. Logistic regressions and decision trees were used to analyze the buying patterns. These groups were defined as new season ticket holders, returning season ticket holders, non-renewal season ticket holders, and non-season ticket holders.

Based on a simple decision tree algorithm, new buyers could be accurately predicted from the entire database of non-buyers. Using metrics on total dollars spend on donations, football single game tickets, other sports tickets, and account length, new buyers could be predicted with a high level of accuracy. Logistic regressions further support these claims.

Next, the same set of analyses were computed in order to predict returning season ticket holders. These analyses proved to be less successful, although a decision tree algorithm provided minimal success. This model used a combination of purchase history, donation history, and account length to differentiate between the two groups.

A final note on these analyses includes a discussion of their limitations. First, they were constructed using a single year's data since a high level of detail was not available for other years. However, this could be mitigated by re-examining the analysis after another year of purchases. Second, these predictions will always involve a level of uncertainty since individual behavior is unpredictable. However, they can provide some recommendations on differentiating between different buying groups.

6.1 Recommendations for Marketing

First, on a holistic level, demand is most directly influenced by opponent quality. Other factors appear to have minimal influence. Second, higher quality opponents from farther distances would yield the greatest return of attendance outside of the main rivals.

On an individual level, there appears to be several metrics that can be used to differentiate between different buying groups. To predict new buyers, purchasing habits in football and other

sports appears to be of importance. Those who spend more money on other sports should be targeted. As such, marketing at athletic events seems a likely way to influence new would-be buyers.

Also, out of the database population, those with shorter account lengths are more likely to buy. More specifically, those who were in the database for less than three years are the most likely to purchase season tickets.

As for predicting those who aren't going to return, the task is much harder. However, several recommendations can be made. First, season ticket holders who have held their seats for three years or less are susceptible to be lost. Second, those with minimal additional dollars spent on the football program are also less likely to renew. These dollars can be spent on either donations or single game football tickets. As a final note, those who spend less money on other sports are less likely to renew; however, this is not as strong of a predictor as other variables.

6.2 Recommendations to Increase Statistical Accuracy

Several recommendations to improve statistical modeling are warranted. First, and perhaps the easiest to implement, would be the recommendation to obtain alumni information from the Alumni Association. This information could be used to better model buyer relationship to the school could be done with minimal assumptions and costs. Basic merges on buyer name and address could be made between the two databases. As opposed to all other recommendations made in this section, this process could be accomplished in a mere week whereas the rest would take several years to build enough information for statistical analyses.

Secondly, buyer relationship to the school on different levels should be coded into the database. This would include variables such as whether or not the buyer has a child enrolled in the school or whether they are related to any current players. This information is typically asked by the sales staff already, and formally quantifying it would increase the accuracy of buying predictions.

Another important metric to record would be who buyers attend games with, such as family,

friends, or work acquaintances. It is reasonable to assume that each group operates on different buying mechanisms, and retaining such variables could yield important trends.

The final recommendation centers on statistical recording procedures. Currently, database coding requires that any current season ticket holder who sold tickets on the exchange be recorded as a single game buyer. Separate variables should be created for these, if possible, in order to quarantine the separate groups.

6.3 Concluding Remarks

This investigation of attendance and demand yielded interesting results. Overall, game-by-game attendance was influenced by many factors. However, the strongest influence was exerted by strength of opponent characteristics. On an individual level, past buying characteristics, past donation levels, and account characteristics were the best predictors of buying habits.

Several considerations must be accounted for when understanding these results. First, they are based on a single year of information since complete information could not be obtained for other years. Second, individual decisions, at their very nature, will induce some uncertainty into the results. Finally, generalizations need to be refined as more data becomes available. Nonetheless, these inferences still appear valid and offer important information.

Appendix A

Assumptions Analysis for Model of Attendance

As with most statistical procedures, linear regression is based upon a variety of assumptions that must be upheld in order to maintain the validity of the model. These assumptions include: linearity, independence, normality of errors, and errors with mean of zero and constant variance.

For the preferred model, assumptions of the regression model were analyzed. A normal probability plot was constructed. It can be found in Figure A.1. Further, the Anderson-Darling test for normality was computed, and the test statistic was calculated as 0.21 with a p-value of 0.84. Thus, there is no significant evidence that the residuals appear to depart from normality. Independence was deemed reasonable, as both game-specific and season-specific factors were studied.

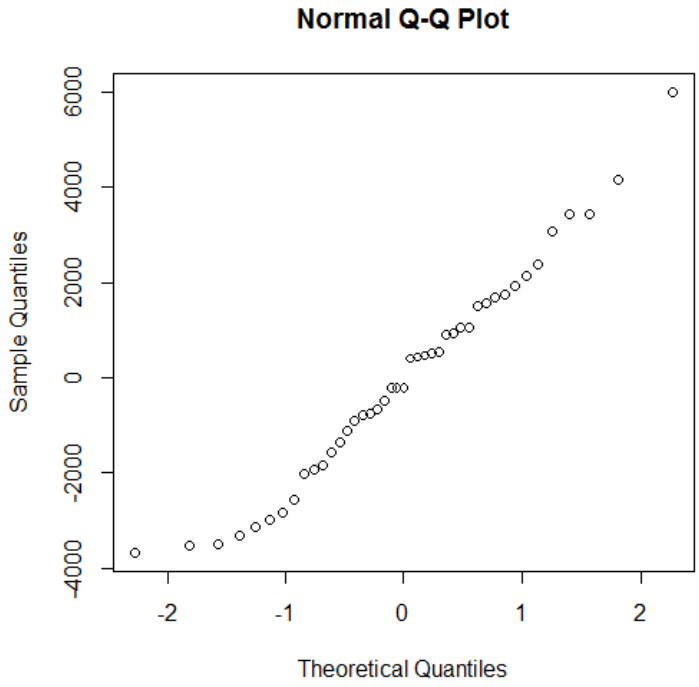

Figure A.1: Normal Probability Plot of Residuals for the Preferred Model of Attendance

Variance inflation factors were examined, and their sum was calculated as 7.38. While this value is still high, it is still within acceptable bounds, such as those suggested by Kutner.

For the full model, assumptions of the regression model were analyzed. A normal probability plot was constructed. It can be found in Figure A.2. Further, the Anderson-Darling test for normality was completed, and the test statistic was calculated as 0.44 with a corresponding p-value of 0.280. The resulting conclusion of this test is that there is no strong evidence supporting non-normality of the residuals. Independence was determined to not be an issue since attendance for each game is driven by many game-specific factors and few season-specific factors.

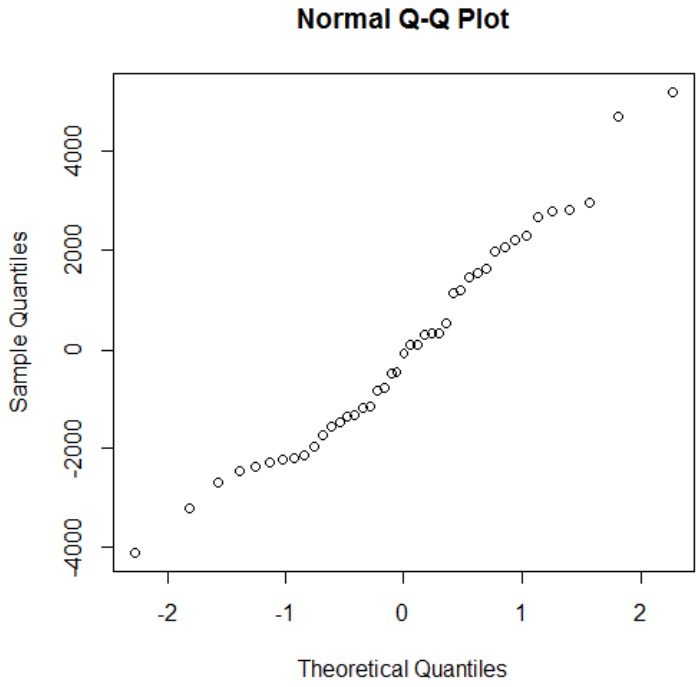

Figure A.2: Normal Probability Plot of Residuals for the Full Model of Attendance

Again, variance inflation factors were also examined for the full model, and their sum was calculated as 24.43. Over-inflation appears to be an issue in this full model. So, the preferred model was developed in order to improve statistical accuracy.

Appendix B

Exploratory Data Analysis

Below are the results of the exploratory data analysis. Such analysis is necessary in order to better understand the relationships present within the data. Table B.1 displays the amount of accounts in the sample that purchased different groups of football tickets. Roughly 2.4% of the sample purchased season tickets for the 2014 season. For the 2015 season, 3.3% of those in the sample purchased tickets. These percentages meet the standard minimum requirements for the implementation of a logistic regression model (Kutner et al., 2005).

Table B.1: Purchasing Habits of Stratified Sample

Number of 2014 season ticket holders in sample	1,758
Number of 2014 single game buyers in sample	2,496
Number of 2015 season ticket holders in sample	2,413

Descriptive analyses were completed for distance variables. Since the data was stratified by location, some cohesiveness of the data is expected. However, simple analyses can still prove

informative.

Distances were calculated using the Haversine equation. This equation relates the distance between two points as a function of latitude and longitude. For this analysis, distance was calculated between the billing zip code and University Park, PA, the location of the football stadium. The distance calculated is as the crow flies. Therefore, it is not the individuals driving distance, but, nonetheless, it still provides an accurate gauge of relative location compared to the school.

Table B.2 displays the summary statistics for these calculated differences. As seen below, both groups are relatively homogenous. The overall mean is 166, and the median is 96. This is suggestive of a negative skew in the data; large outlier accounts with extreme distances are causing the inflation in the median.

Table B.2: Summary Statistics for Distance in Miles

Group	**Mean**	**Median**	**Standard deviation**
Overall	167	97	336
2015 season ticket holders	161	107	274
2015 non-season ticket holders	167	97	338

To further represent the stratified sample, Figure B.1 contains a map of Pennsylvania and surrounding states as represented in the sample population. As expected, darker colors indicate higher frequency among the sample. This is also indicative of the overall trend of the data that is available. White areas represent zip codes that are not represented in the sample; this could be due to random chance or lack of entries in the database.

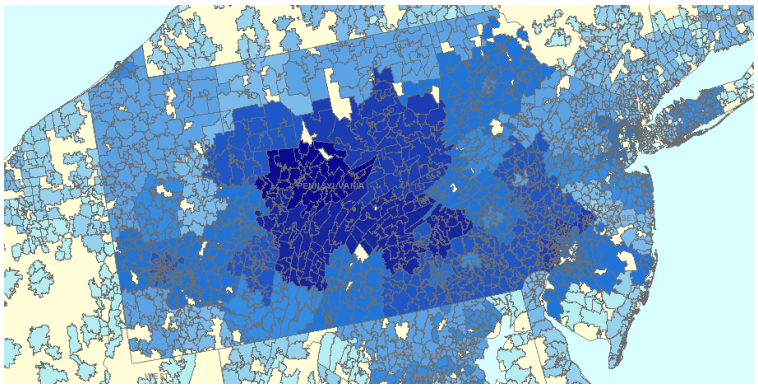

Figure B.1: Pennsylvania Map of Stratified Sample

Appendix C

Assumptions Analysis for Model of Individual Demand

Although less restrictive than a traditional linear regression model, there are still several assumptions that must be explored for a logistic regression (Kutner et al., 2005). A review of model fit reveals that both meaningful predictors have been chosen (p-value: 0.000), and no misspecification blatantly exists (p-value: 0.992). Additionally, the Hosmer-Lemeshow goodness-of-fit test reveals that the model also accurately fits the data (test statistic: 7.04; p-value: 0.53).

Further diagnostics reveal that multicolinearity does not appear to influence the model. The maximum variance inflation factor is 1.03, meeting established guidelines that indicate multicolinearity is low (Kutner et al., 2005). Furthermore, for observations on which the model was based, its specifications correctly predict 99% of all accounts.

Appendix D

Testing for Predictors between Different Buying Groups

In order to determine which predictors would be useful in distinguishing between different buying groups, statistical tests were performed. Initial inspection between the groups revealed probable violations of the assumption of normality, rendering traditional methods of analysis of variance inappropriate. Instead, its nonparametric alternative was utilized. The Kruskal-Wallis test was deemed appropriate since it is mathematically equivalent to one-way analysis of variance and is robust to extreme values. Test statistics were calculated in R, and p-values were approximated via the normal approximation.

After initial testing, multiple comparisons were conducted where appropriate. The Wilcoxon Rank-Sum test, the nonparametric equivalent to a t-test, was used. In order to control for type I error inflation, a simple Bonferroni correction for multiple comparisons was used. Although somewhat conservative, this correction controls for the escalation of type I error which can lead

to spurious conclusions. For example, if 20 separate hypothesis tests are conducted at 95% confidence, statistical probabilities suggest one significant conclusion will be the result of type I error.

First, distances between different buying groups were studied. Initial inspection was completed, and a boxplot of the results can be found in Figure D.1. The Kruskal-Wallis test was performed, resulting in a test statistic of 1805.9 and a p-value of approximately 0.000. Multiple comparisons were then conducted to determine which groups were statistically different. These comparisons were conducted at a family-wide error rate of 0.05. A summary of the multiple comparisons can be found in Table D.1.

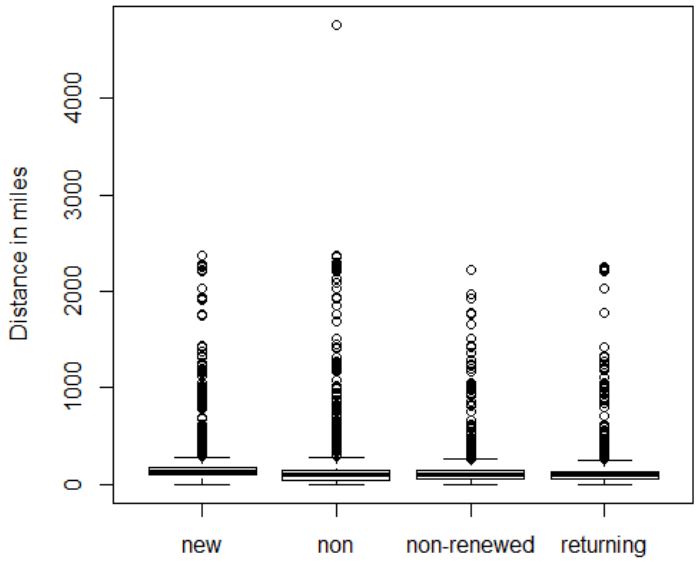

Figure D.1: Boxplot of Distance by Buying Group

Table D.1: Multiple Comparisons for Distance by Buying Groups

Compared Groups	**Observed Distance**	**Statistical Significance at** $\alpha = 0.0083$
New versus non	834.7	True
New versus non-renewing	879.4	True
New versus returning	908.6	True
Non versus non-renewing	44.7	False
Non versus returning	73.9	False
Non-renewing versus returning	29.1	False

Dollars spent on other sports was also compared between groups. Initial inspection, as seen in Figure D.2, also revealed the need for a nonparametric test. The Kruskal-Wallis test was computed, and the test statistic was calculated at 190.7 with a p-value of 0.001. Multiple comparisons were again conducted at a family-wide error rate of 0.05. The results can be found in Table D.2.

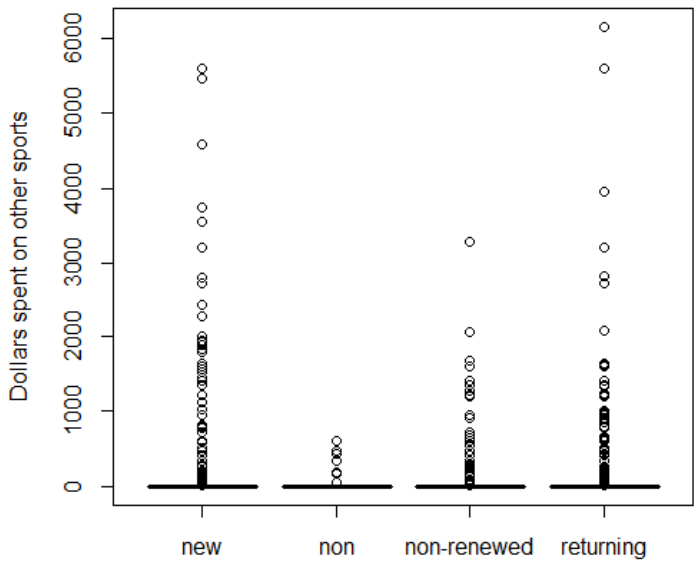

Figure D.2: Boxplot of Dollars Spent on Other Sports by Buying Groups

Table D.2: Multiple Comparisons for Dollars Spent on Other Sports by Buying Groups

Compared Groups	**Observed Distance**	**Statistical Significance at** $\alpha = 0.0083$
New versus non	244.1	True
New versus non-renewing	129.1	False
New versus returning	17.7	False
Non versus non-renewing	115.0	False
Non versus returning	261.7	True
Non-renewing versus returning	146.7	False

Dollars spent on single game football tickets in 2014 was also studied. Again, normality appeared to be violated, as seen in the boxplot in Figure D.3. The Kruskal-Wallis test was performed, and the resulting test statistic was 2861.3 with a p-value of approximately 0.000. Again, multiple comparisons were conducted at a family-wide error rate of 0.05. The results of these comparisons can be found in Table D.3.

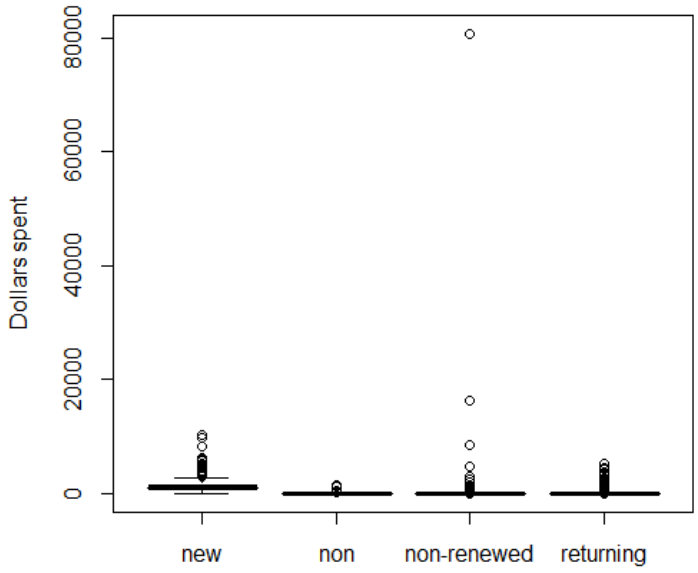

Figure D.3: Dollars Spent on Single Game Football Tickets by Buying Groups

Table D.3: Multiple Comparisons for Dollars Spent on Single Game Football Tickets by Buying Groups

Compared Groups	**Observed Distance**	**Statistical Significance at** $\alpha = 0.0083$
New versus non	2775.5	True
New versus non-renewing	2574.0	True
New versus returning	2361.8	True
Non versus non-renewing	201.5	True
Non versus returning	413.7	True
Non-renewing versus returning	212.2	True

A final test was done for differences in account length between the buying groups. Again, the nonparametric test was preferred, as seen in Figure D.4. The Kruskal-Wallis test was computed, and a test statistic of 2567.2 with a p-value of 0.401 was calculated. Thus, no differences between the groups were found to be statistically significant.

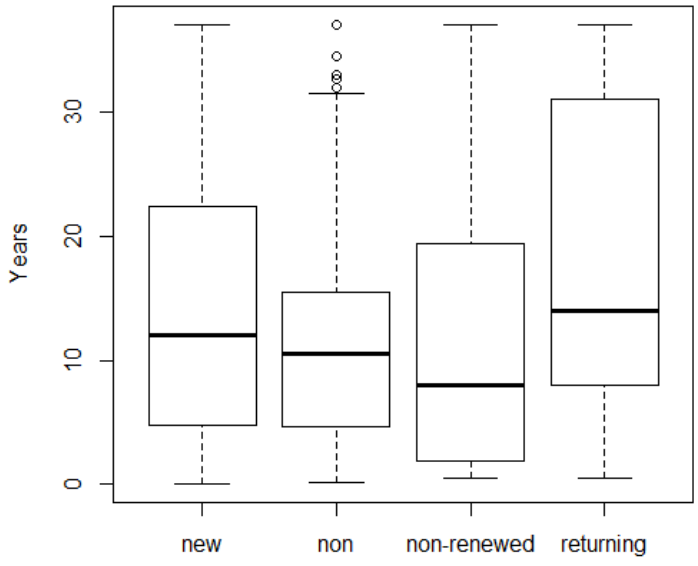

Figure D.4: Account Length by Buying Groups

Appendix E

Analysis of Assumptions for Logistic Models to Predict between Groups

For the full model of predicting new season ticket holders, meaningful predictors have been chosen (p-value: 0.000). The maximum variance inflation factor is 1.29, and the mean of all variation inflation factors is 1.10. While this is still within acceptable limits, it still was improved upon in the preferred model.

For the preferred model of predicting new season ticket holders, meaningful predictors have also been chosen (p-value: 0.000). The maximum variance inflation factor is 1.02, and the mean of all variance inflation factors is 1.01.

For the full model of predicting returning season ticket holders from the population of all season ticket holders, the model fits better than an empty alternative (p-value: 0.000). Additionally, the maximum variance inflation factor is 1.31, and the mean of all variance inflation factors is 1.11.

Bibliography

Baade, R. A. and Tiehen, L. J. (1990). An analysis of Major League Baseball attendance, 1969-1987. *Journal of Sports and Social Issues*, 14:36–61.

Beechey, M., Gruen, D., and Vickery, J. (2015). The efficient markey hypothesis: A survey. *Economic Research Development*.

Boutell, T. (2014). Zip code latitude and longitude database.

Breiman, L., Friedman, J., Stone, C.J., and R.A. Olshen (1984). *Classification and Regression Trees*. Chapman and Hall/CRC, London.

Donihue, M., Findlay, D., and Newberry, P. (2007). An analysis of attendance at Major League Baseball spring training games. *Journal of Sports Economics*, 8:14–32.

Higgins, J. (2004). *Introduction to Modern Nonparametric Statistics*. Brooks/Cole Cengage Learning, Belmont, CA.

Humphreys, B. and Ruseski, J. (2008). The size and scope of the sports industry in the United States. *IASE/NAASE Working Paper Series, No. 08-11*.

IRS (2014). Individual income tax statistics - 2012 ZIP code data (SOI).

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *FJJ/Shraban*.

Kutner, M., Nachtsheim, C., and Neter, J. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin, Boston.

Lavery, R. (2015). An animated guide: Regression trees in JMP and SAS Enterprise Miner. *North East SAS Users Group*.

Leonard, J. (2005). The geography of visitor attendance at college football games. *Journal of Sports Behavior*, 28:231–52.

Palanjian, S. (2012). Factors influencing student and employee attendance at college football games. Master's thesis, The University of North Carolina at Chapel Hill.

Pan, D., Zhu, Z., Gabert, T., and Brown, J. (1999). Team performance, market characteristics, and attendance of Major League Baseball: A panel data analysis. *The Mid-Atlantic Journal of Business*, 35:77–92.

Paul, R. (2003). Variations in NHL attendance: The impact of violence, scoring, and regional rivalries. *The American Journal of Economics and Sociology*, 62:345–364.

Paul, R., Humphrey, B., and Weinbach, A. (2012). Uncertainty of outcome and attendance in college football: Evidence from four conferences. *The Economic Labour Relations Review*, 23:69–81.

Paul, R. and Weinbach, A. (2011). Determinants of attendance in the Quebec Major Junior Hockey League: Role of winning, scoring, and fighting. *International Atlantic Economic Society*.

Penn State athletics (2014). GoPSUSports.com. CBS Interactive, Inc.

Price, D. and Sen, K. (2003). The demand for game day attendance in college football: An analysis of the 1997 Division 1-A season. *Managerial and Decision Economics*, 24:35–46.

SAS (2015). Getting started with SAS (R) Enterprise Miner (TM). *SAS: The Power to Know*.

Stock, J. and Watson, M. (2007). *Introduction to Econometrics*. Pearson/Addison Wesley, Boston.

Stringer Sites (2015). 3-digit ZIP code prefix map of Pennsylvania. Stringer Sites.

USDA (2015). Unemployment and median household income for the U.S., states, and counties, 2000-2013.

United States ZIP Code Maps (2015). United States ZIP codes. *UnitedStatesZipCodes.org*.

Welki, A. and Zlatoper, T. (1998). U.S. professional football game-day attendance. *Journal of Sports Economics*, 27:285–98.

Whitehead, J. (2015). An introduction to logistic regression. *Appalachian State University*.

ACADEMIC VITA

Michelle Pistner

949 Falcon Road
Saint Marys, PA 15857
map5672@psu.edu

EDUCATION

The Pennsylvania State University **University Park, PA**
Schreyer Honors College *Class of 2015*

- Bachelor of Science in Statistics, Bachelor of Science in Economics
- Honors in Statistics and Economics

EXPERIENCE

IMG Learfield Ticket Solutions **University Park, PA**
Data Analyst Sept. 2014 – Current

- Analyzed sales trends in order to determine marketing initiatives
- Conducted a variety of statistical analyses on databases consisting of over three million observations

University Libraries at the Pennsylvania State University **University Park, PA**
Bednar Internship Sept. 2014 – Current

- Provided tutoring services to faculty and students at the Pennsylvania State University
- Constructed online resource library for faculty and students that consisted of information on different statistical methodologies and procedures

Bates White, LLC **Washington, D.C.**
Summer Consultant June 2014-Aug. 2014

- Conducted nonparametric analysis, including local regressions and smoothing estimations, in order to estimate a client's potential insurance liability in relation to allegations proposed in a court case
- Wrote programming scripts to compile and manage several databases that consisted of over 75,000 observations

INVOLVMENT

Professor's Aide, Introductory Macroeconomics **The Pennsylvania State University**
Spring 2014

- Performed data analysis, including regressions and ANOVA, to assist the professor in classroom decisions

Learning Assistant, Introductory Macroeconomics **The Pennsylvania State University**
Fall 2013-Spring 2013

- Led peer discussions on classroom questions relating to course materials
- Graded student homework and exams

Phi Eta Sigma, THON Committee **The Pennsylvania State University**
Spring 2012-Spring 2014

HONORS AND AWARDS

- Evan Pugh Senior Scholar Award, The Pennsylvania State University
- Evan Pugh Junior Scholar Award, The Pennsylvania State University
- The President's Freshman Award, The Pennsylvania State University
- Dean's List, all semesters
- Member: Mu Sigma Rho

SKILLS

R, SAS, Stata, Minitab, C++, Visual Basics, LaTex, ArcGIS, and Microsoft Office