Analysis and Modeling of Professional Football Team Ticket Prices: A study of pricing determinants

36-707: Applied Regression Analysis

, ,
,,@stat.cmu.edu
Carnegie Mellon University
Department of Statistics

December 12, 2001

Abstract

This study analyzes the determinants of NFL ticket prices for the 2001 regular season. We have been able to construct a model which provides compelling evidence that several variables are associated with and predictive of ticket prices. Among those found with a significant, positive correlation with ticket price are whether a team is playing in a recently built stadium (opened in the last 5 years), and a team's overall popularity in the league (as measured by merchandise sales). The amount of money a team had available below the salary cap prior to the start of the 2001 season was found to be negatively correlated. A significant interaction between the 2000 season winning percentage and the presence of other professional sports teams in the same city suggests that 2001 ticket prices were more responsive to a team's 2000 season winning percentage in cities with other professional sports whose seasons overlap with the football season. Finally, a bootstrapping analysis was performed which confirmed the significance of all of the major factors in our model.

Introduction

NFL ticket prices have risen 8.7 percent since last season, raising the average ticket price for a professional football game to over $50 per seat. Clearly, football fans are willing to open their wallets to watch their favorite teams play. Prices for football tickets are generally high because of scarcity. Only 8 regular season home games are played by each team, so the supply-demand equilibrium tends to be found at a very high price level. In this paper, we study what we believe to be a more interesting question than just why do people pay so much for football tickets. We ask why they are willing to pay more to watch some teams and less for others. The design of our study is cross-sectional, but much of our data to predict 2001 season prices comes from 2000 season values for our variables.

There is a high degree of variability of average ticket prices across teams. The average price ranges from a low of $38 for the Arizona Cardinals to a high of $82 for the Washington Redskins. We develop a model which can help us think about why there exists ticket price variation across teams and what the key determinants of this variation seem to be. *A priori* we might assume that winning teams tend to have higher prices. Or, perhaps teams that play in big markets have the more exorbitant prices. Having a lot of people with a lot of money would seem to bring some healthy upward demand pressure on prices.

We would like to determine whether we can account for this variability by studying aspects of teams' past performance, characteristics of the market of their home city (population, cost of living), the financial situation of team (salary cap constraints), and any other reasonable predictors of how teams might choose to set prices. Since prices are set between seasons by the different ownership groups for NFL teams, the prices we are examining are not free market prices. When looking for significant factors to include in our model we must be cognizant of the indirect manner in which the market (i.e. the fans) actually affects the ticket prices set by the ownership.

For the purposes of constructing our model, we utilize a standard linear regression framework. Without a large quantity of data, we are unable to set aside even a small sample of our data for assessing our final model and testing its fit. Instead, we test the robustness of our specification using "case-based" bootstrapping. This essentially allows us to test the sensitivity of our coefficient estimates to different combinations (weightings) of our observation vectors.

The structure of this report is as follows. We will provide a thorough detailing of the variables we considered in our section **Description of Data**. Next, in **Analysis and Results** we provide the final specification of our explanatory model and include, where appropriate, interpretation of relevant diagnostic methods and tools. The **Discussion/Conclusion** section will give a sense of the degree to which we can generalize from our model about the relationships under study while also commenting at some length about the obstacles we faced given the type of data gathered and the analysis we have employed. The **Technical Appendices** provide the exploratory data analysis we performed as well as a greater level of detail regarding our model building and validation.

Description of Data

Our response variable is the natural logarithm of average 2001 regular season ticket prices for the 31 National Football League teams. These averages were provided by USA Today, and they are essentially a weighted average of ticket price, taking into account the different price levels for seats and the number of seats at each price level. It is customary to use a log transform for financial variables due to their propensity for right-skewing. We follow this custom here due to the right-skewing in ticket prices we detected. It has helped us achieve a more normally-distributed response variable for linear regression analysis.

There are many issues surrounding the use of average ticket price in our analysis. First, the mean ticket price may not be indicative of the price that the typical individual pays. Luxury boxes represent a severe right skew in the distribution of ticket prices for a single team and can demonstrably increase the average price while the median price remains unchanged. Also, since ticket prices are explicitly determined by team ownership prior to each season, they are an imperfect indicator of supply-demand equilibrium in the market at any point in time. The true market price is most appropriately and accurately captured in the form of prices for scalped tickets. This is where the unfettered free market truly determines the value of a ticket. Unfortunately, we were unable to compile data for scalped tickets for the purposes of examination in this report.

A notable omission from our study is the cost component to many consumers in the form of personal seat licenses (PSL's). These licenses are often a required purchase in order for fans to have the right to then purchase season tickets at their face value. USA Today did not include PSL's when they calculated the average prices used in this analysis so we have no way of accounting for its affect on the total average cost for consumers to attend a football game. We have also not collected data on the cost of parking, hotdogs, beer, and other necessary expenditures having a substantive contribution to the total cost of a Sunday football outing.

Note: Since ticket prices are determined by each team during the off-season, the prices charged reflect the team owner's perceptions of what fans are willing to pay. Those perceptions can only be based on data from the previous season. Therefore, since we are analyzing 2001 average ticket prices, we will be looking at data that was known to the team owners prior to the 2001 season. For example, we will look at 2000 winning percentage (not 2001 which is still in progress), but we will account for Pittsburgh's new stadium, which opened in 2001, as a factor that was anticipated by the organization when they set the average 2001 ticket prices 52% higher than last year.

Here are the variables we examined in this study. For additional detail and EDA, please refer to the Technical Appendices.

- **Team Variables**
  - **dummy.afc**
    Conference: 1=AFC, 0=NFC
  - **win.pct.2000**
    2000 regular season winning percentage
  - **home.win.pct.2000**
    2000 regular season home winning percentage
  - **off.td.2000**
    2000 total offensive touchdowns
  - **number.superbowl**
    Total number of Super Bowl Victories in franchise history
  - **number.probowl**
    Total number of 2000 Pro Bowl Players
- **Market Variables**
  - **city**
    Nearest major city (Sometimes the stadium is close but not technically in the same city. For instance, the New England Patriots' stadium is a 40 minute drive from Boston. Estimating relative commute times is outside the scope of this project. We will use the most reasonable home city for each team)
  - **city.pop.2000**
    City Population (from 2000 Census)
  - **cost.of.living**
    Estimates relative cost of living across differenct cities. We will construct this variable by providing a base case, living in Cincinnati on $40,000 per year. The number for all other cities is how much an individual would have to earn in those cities in order to maintain the same standard of living as they experience in Cincinnati, OH earning $40,000. The source for this data was Monster.com's moving site called Monstermoving.com which contained the salary comparison calculator used to collect cost of living data.
  - **temperature**
    Mean temperature in home city for the month of December
  - **dummy.basketball**
    1=NBA team plays in same market, 0=else
  - **dummy.hockey**
    1=NHL team plays in same market, 0=else
  - **dummy.sportstown**
    1=NBA, NHL and NFL teams share the same market, 0=else

- **Stadium Variables**
  - **capacity**
    Stadium seating capacity
  - **attendance**
    2000 average attendance
  - **capacity.utilized**
    Equal to **attendance** divided by **capacity**
    (Maximum value allowed is .999)
    Note: We calculated Denver and Pittsburgh using 2000 stadium capacity, since they opened new stadiums in 2001. Our **capacity** *variable will be relevant for above calculation for all other teams who didn't open a new stadium in 2001.*
  - **dummy.outdoor**
    1=outdoor stadium, 0=dome
  - **dummy.artificial.turf**
    1=artificial turf playing surface, 0=natural grass
  - **dummy.new.stadium**
    1=new stadium built in last 5 years, 0=else
- **Financial Variables**
  - **ticket.price.2001**
    2001 average ticket price
  - **salary.cap.room**
    As of July, 2001, the amount of money below the salary cap the team was eligible to spend.
  - **dummy.top15.merchandise**
    1=team is among top 15 in merchandise sales, 0=else

Analysis and Results

Since we have so few observations, a backward elimination technique for model-fitting is less appropriate than if we had a larger data set. Instead, we prefer to build from the ground up by examining simple linear regressions to identify any strong or moderate relationships between individual predictor variables and **log.ticket.price.2001**. Below is a table summarizing which variables had significant or near-significant results in simple linear regression models.

- Promising Predictor Variables

Variable	Status	Direction of Relationship	p-value
win.pct.2000	near-significant	positive	0.06
home.win.pct.2000	near-significant	positive	0.059
number.superbowl	near-significant	positive	0.158
number.probowl	significant	positive	0.037
attendance.2000	near-significant	positive	0.0635
capacity.utilized	near-significant	positive	0.0881
number.superbowl	near-significant	positive	0.158
dummy.outdoor	near-significant	positive	0.0878
dummy.artificial.turf	significant	negative	0.0263
dummy.new.stadium	significant	positive	0.000099
sqrt.salary.cap.room	significant	negative	0.0345
dummy.top15.merchandise	near-significant	positive	0.0928

- Final Model Specification
  We constructed multiple models, but our final specification includes 6 variables (5 variables + 1 interaction term) which capture whether a team has a new stadium, the amount of salary cap room, merchandise sales, the winning percentage from the previous season, and the presence of other sports teams who share the same market.

  Extra Sum of Squares tests were insignificant for the inclusion of any additional variables in this model. Therefore, we retain the following model as our final specification.

Model Variables

sqrt.salary.cap.room
dummy.new.stadium
dummy.top15.merchandise
win.pct.2000
dummy.sportstown

win.pct.2000:dummy.sportstown

Here are the results of our linear regression model:

```
lm(formula = log.ticket.price.2001 ~ sqrt.salary.cap.room +
    dummy.new.stadium + dummy.top15.merchandise + win.pct.2000 +
    dummy.sportstown + win.pct.2000:dummy.sportstown, data = nfl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16233 -0.05939 -0.01330  0.04520  0.21908

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.15299    0.09838  42.216  < 2e-16 ***
sqrt.salary.cap.room          -0.10372    0.03180  -3.262  0.00330 **
dummy.new.stadium              0.25798    0.04470   5.771    6e-06 ***
dummy.top15.merchandise        0.09867    0.04158   2.373  0.02601 *
win.pct.2000                  -0.32979    0.14116  -2.336  0.02815 *
dummy.sportstown              -0.40183    0.10978  -3.660  0.00124 **
win.pct.2000:dummy.sportstown  0.76472    0.21267   3.596  0.00145 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1013 on 24 degrees of freedom
Multiple R-Squared: 0.7544,     Adjusted R-squared: 0.693
F-statistic: 12.28 on 6 and 24 DF,  p-value: 2.626e-006
```

The most statistically and substantively significant variable associated with ticket prices is **dummy.new.stadium**. New stadiums have been sprouting up all over the league as of late (8 in the last five years), and it seems fairly easy to see why. Average ticket price for teams with a new stadium built in the last five years in \$65 versus \$50 for teams without a recently built stadium. Figure 1 displays the stark contrast in average prices in side-by-side boxplots. The interquartile ranges of our boxplots do not even overlap.

The main reason that a new stadium will increase averge ticket prices is the result of a large increase in corporate and luxury box seats. All of the new stadiums are designed to provide a larger number of these high priced seats than the older stadiums which were built in a different era. The average fan will also pay more to

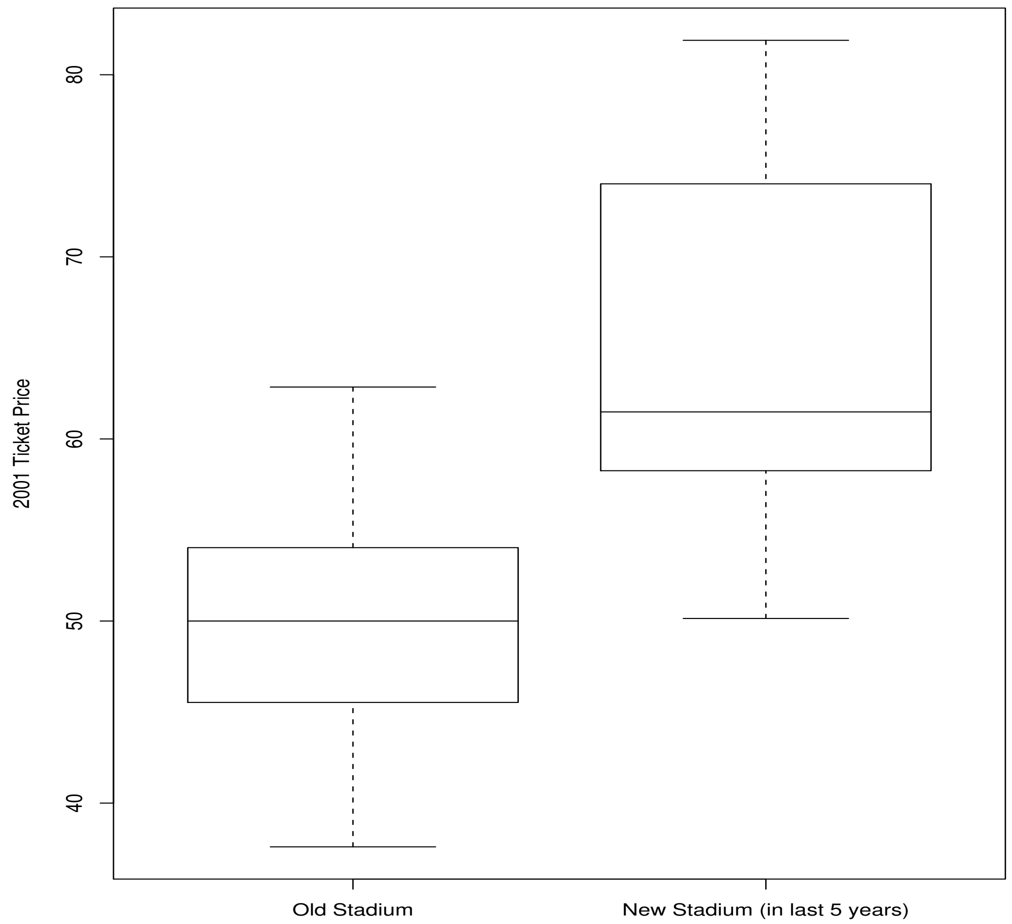

Figure 1: 2001 ticket prices by New vs. Old Stadium

attend games in a beautiful new stadium, but it is quite likely that the luxury boxes are driving a large portion of the increase in average ticket price in new stadiums.

A team whose merchandise (jerseys, mugs, pennants, etc.) sells better than average (i.e. among the top 15 teams) has an average ticket price of $56 while teams not in this category have an average price of $51. These are strictly averages to give the reader a sense of the relative importance of this variable compared to **dummy.new.stadium**. Our **dummy.top15.merchandise** variable was not found to be significant in the simple linear regression with **log.ticket.price.2001**, but its significance has surfaced in our larger model as we have been able to control for other variables which were a source of noise in the simple regression.

The amount of salary cap room is a variable which, as much as anything, is under the control of team management. The negative correlation we find between this variable and ticket prices is likely to be caused by a team trying to save some money (high cap room) when they are not generating as much revenue from ticket sales (low ticket prices).

Our interaction term **win.pct.2000:dummy.sportstown** is found to be significant suggesting that fans in a “sportstown” are more responsive to a winner (and a loser) than fans in other cities. The lower order terms of our interaction are also significant. The negative and significant coefficient on **dummy.sportstown** suggests that football, basketball and hockey are substitutes. It makes economic sense that having an additional sports team (a substitute) in the same city would tend to lessen demand pressures and consequently be associated with lower prices (on average) for football tickets.

Figure 2 displays several diagnostic plots for our regression model. Our cloud of residuals is nicely spread without a discernible pattern. Also, the normal quantile plot is reasonably straight with no obvious departure from normality. The observations which have been identified as highly influential (or possibly outliers) were removed in one calculation of our model. The model’s significance was retained and improved for all coefficients, and we achieved an increase in R^2. Figure 3 provides a scatterplot of actual versus predicted ticket price after we have converted our variables back to their regular values. This makes for an easier understanding of the strength of our model by removing the non-intuitive natural log transformed prices. It is clear that our model provides a relatively accurate prediction of ticket price. In general, the data points hug the line fairly closely, while only the Washington Redskins seem to noticeably defy the gravitational pull of our trend line.

We test our model using “case-based” bootstrapping. Normally, we would be inclined to perform this kind of analysis in situations where our normal-errors assumption was in doubt. Our diagnostic plots do not actually reveal any strong sense of non-normality. However, our small number (31) of cases motivate us to test the sensitivity of our coefficient estimates to different combinations (with repeats) of our observation vectors. In an observational study, as opposed to a designed experiment, it is appropriate to perform a naive “case-based” bootstrapping analysis in order to test the robustness of our model specification. This is the preferred version of bootstrapping for social science regression models because, as in this case, we cannot have the highest level of confidence that we have the “true” model.

Since all of the coeffients in our model are individually significant at p=.05, we know that zero is not included in their 95 percent normal confidence intervals. Bootstrap

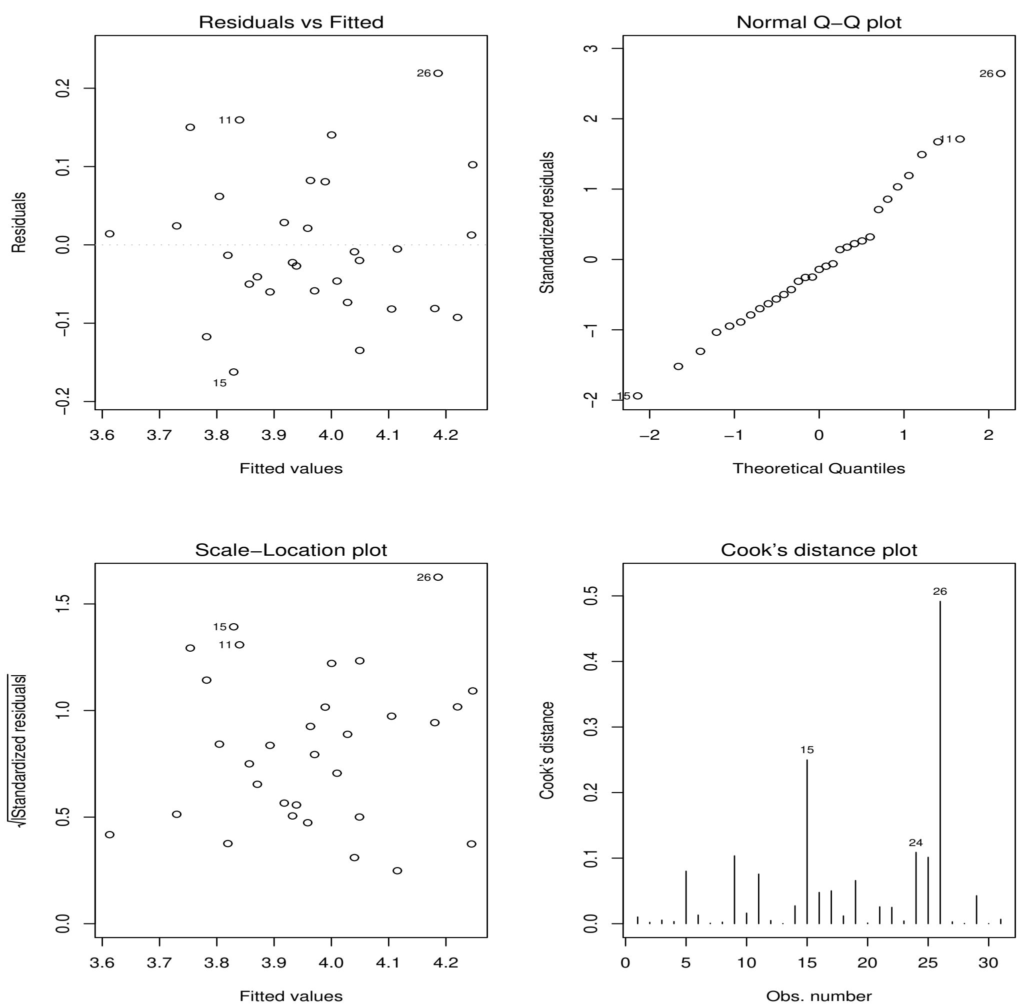

Figure 2: Diagnostic Plots of Multiple Regression

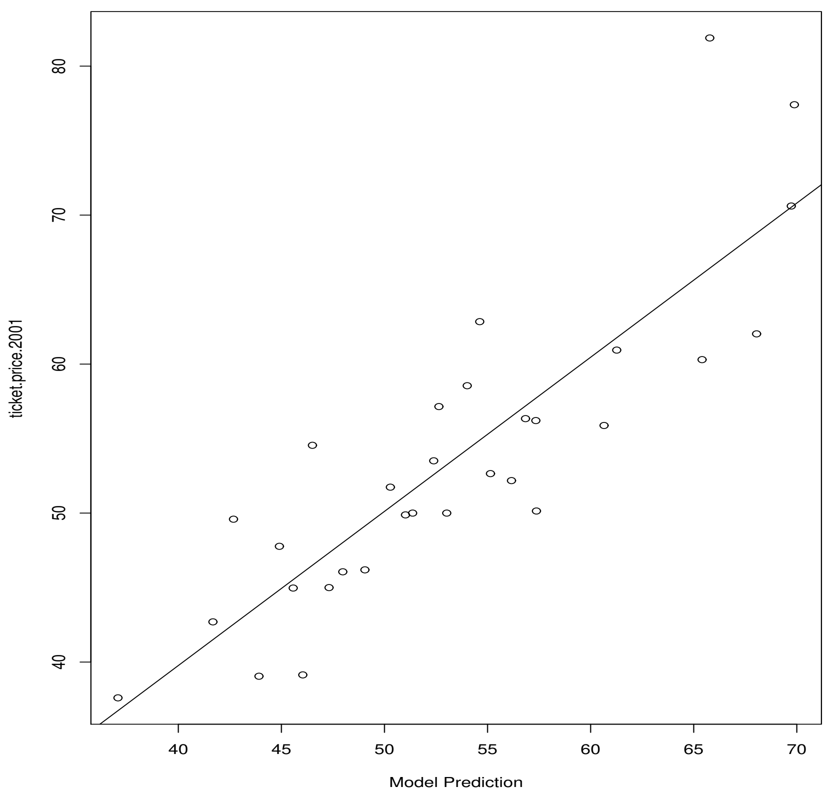

Figure 3: Scatterplot of Actual Ticket Price and Model-Predicted Ticket Price

samples give us slightly different (larger) confidence intervals to test the significance of our coefficients. Below, we provide bootstrap intervals for our coefficients.

Bootstrap percentile-based 95% confidence intervals for independent variable coefficients:

Intercept

```
    2.5%    97.5%
3.857473 4.370667
```

sqrt.salary.cap.room

```
     2.5%     97.5%
-0.18335735 -0.02362473
```

dummy.new.stadium

```
     2.5%     97.5%
0.1305388 0.3612133
```

dummy.top15.merchandise

```
       2.5%       97.5%
0.009822568 0.180039609
```

win.pct.2000

```
      2.5%      97.5%
-0.6189132  0.0940375
```

dummy.sportstown

```
      2.5%      97.5%
-0.5964222 -0.1676532
```

win.pct.2000:dummy.sportstown

```
     2.5%     97.5%
0.2786191 1.1676652
```

Clearly, all but one of our coefficients are significantly different from zero (i.e. zero is not in the 95 percent bootstrap interval). The coefficient on **win.pct.2000** includes the value zero in its 95 percent bootstrap confidence interval. This causes us to give pause when attempting to interpret the relationship being measured by this variable. Since **win.pct.2000** is a lower order term for the interaction with **dummy.sportstown**, its coefficient represents the relationship between winning percentage and ticket prices for those teams in cities without an NBA or NHL team, where **dummy.sportstown**=0. The significant, negative relationship found in our regression t-test was not upheld with the application of our bootstrapping 95% confidence interval analysis. We will, however, keep the term in our model, since it is customary to always retain lower order terms when a significant interaction is to be kept in the model.

Overall, the takeaway from our bootstrapping analysis is that our model specification appears to be appropriate for assigning directionality to the relationships between our independent variables and the response, **log.ticket.price.2001**. In other words, we feel confident that we can conclude which variables have a positive association and which have a negative association with ticket prices.

Interpreting Our Coefficients

For the purpose of providing the reader with a sense of how ticket prices tend to change with changes in our independent variables, we will examine a single team, the Miami Dolphins, and watch how their ticket price might be predicted to change. The 2001 season average ticket price for a Miami Dolphins game is $56.34. Our model predicted an average price of $56.83. We got pretty close to the actual price with our prediction. Now, we we see how that price can be expected to change.

The values for the Dolphins on our independent variables are listed in the table below along with a demonstration of how predicted price changes for new values of the variables. *Note: Estimated price differences are for changes in the particular variable holding all other variables constant at their actual value.*

- Miami Dolphins

Variable	**Actual Value**	**New Value**	**Price Difference**
sqrt.salary.cap.room	1.1	2	-$5.04
dummy.new.stadium	0	1	+$16.64
dummy.top15.merchandise	1	0	-$5.31
win.pct.2000	.688	.200	-$10.81
dummy.sportstown	1	0	-$6.61

The Miami Dolphins have a relatively average ticket price (slightly above average). Our model predicts that they could increase average ticket price by $16.64 if they would build a new stadium. That increase is on par with the average price difference between new-stadium teams and old-stadium teams across the league. This adds to our confidence of the reasonableness of our prediction model. If the Dolphins were to have had a perfect record last year and had opened a new stadium this season, then they would be predicted to have an average ticket price of $83.82 for the 2001 season. This would give them the highest ticket price in the league, close to $2 over the Washington Redskins. Since the 1972 Dolphins are the only team in the modern era to go an entire season undefeated, this is not too far-fetched to consider.

Discussion/Conclusion

When performing any analysis it is sometimes more interesting to discover variables that are not significant than those which are found to have statistical significance. Cost-of-living was a variable that we were expecting to be highly correlated with ticket prices. It failed to reveal a significant relationship in both the simple and multiple regression models. Also, city population failed to demonstrate the positive relationship with ticket price that we anticipated.

Our temperature variable was included in our analysis as an attempt to identify any relationship between the regional climate and how this may affect the willingness of fans to endure cold temperatures in the many outdoor stadiums around the league. We expected a positive relationship between December temperature and ticket prices, since, this could have an influence on the margins. No significant relationship emerged in the analysis, however. Low temperatures were clearly not associated with lower ticket prices. It seems that football fans are willing to tough it out in cold temperatures to watch their teams.

Offensive touchdowns as an indicator of on-the-field excitement, number of Super Bowl victories as a proxy for fan loyalty, and number of Pro Bowl players as a measure of star attraction were all found to be less predictive of ticket prices than the other variables in our model. All three would seem to have the potential for a positive relationship with ticket price, but they were not able to provide any additional explanatory capability.

This report has attempted to understand many of the underlying factors which determine ticket prices in the National Football League. If future work is done, it would be interesting to examine the extremes of ticket pricing for each team by looking at the lowest priced tickets and highest priced tickets that are available. From the least expensive ticket for each team one might get a better sense of the minimum cost requirements, the barriers to entry. For all price levels of tickets, we could compare the scalped price to face value to get a sense of the true market price. The percent markup would also provide an indication of relative demand pressures for the tickets to different teams. Unfortunately, our use of an average ticket price for each team prevented a more in depth examination of these important issues.

An additional layer of complexity is introduced if we were to dig deeper and look at the scalping prices of tickets to a particular team for all 16 of its games, not just its home games. If a popular team like the St. Louis Rams comes to play in Pittsburgh, presumably scalping prices will be higher than for the game when a bad team like the Cincinnati Bengals comes to play the Steelers. Tracking market prices for 16 games rather than just 8 home games would increase the amount of data we have, thus enriching the analysis with the introduction of a higher degree of precision.

Future work might also introduce additional years of data. While this was not feasible for this report, it would be interesting to analyze intra-team changes in independent variables and examine their impact on ticket price. For instance, if the Steelers have successive years of low winning percentage, do their ticket prices rise more slowly than the average NFL team over the same period. By performing a time series, we could control for team effects to get a clearer understanding of the true relationships between ticket price and the many covariates.

At the conclusion of this analysis, however, we feel that we have followed a reasonable path to identifying ticket price determinants given the data at our disposal.

Technical Appendices

1. ### Response Variable

   We selected a log transformation of raw average 2001 season ticket price as our response variable. There was virtually no correlation between ticket price and cost-of-living, which might have served to provide an adjusted version of ticket prices. Here are the results of different correlations which were all statistically insignificant at any reasonable p-values:

```
> cor(log.ticket.price.2001,cost.of.living)
[1] 0.09728868
> cor(ticket.price.2001,cost.of.living)
[1] 0.09155404
> cor(log.ticket.price.2001,log.cost.of.living)
[1] 0.08290283
> cor(ticket.price.2001,log.cost.of.living)
[1] 0.08915233
```

   - **ticket.price.2001**

     2001 average ticket price

```
> stem(ticket.price.2001)

  The decimal point is 1 digit(s) to the right of the |

  3 | 899
  4 | 3
  4 | 55668
  5 | 000002234
  5 | 566679
  6 | 0123
  6 |
  7 | 1
  7 | 7
  8 | 2
```

     We see some right-skew to ticket prices. Perhaps a log transformation would be appropriate. We create a new variable **log.ticket.price.2001** below.

   - **log.ticket.price.2001**

     Natural logarithm of 2001 average ticket price

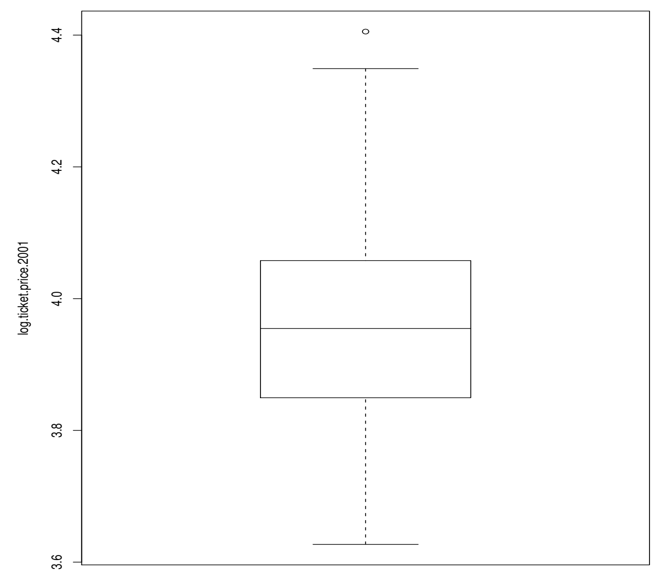

Figure 4: Boxplot of the Natural Logarithm of 2001 ticket prices

```
> stem(log(ticket.price.2001))

  The decimal point is 1 digit(s) to the left of the |

  36 | 367
  37 | 5
  38 | 11337
  39 | 011115568
  40 | 023357
  41 | 0134
  42 | 6
  43 | 5
  44 | 1
```

The log transform has pulled the positive outliers closer to the main body of the data, but the boxplot of **log.ticket.price.2001** in Figure 4 identifies the Redskins ticket price as an outlier even after the transformation.

2. **Team Variables**

– **dummy.afc**

Conference: 1=AFC, 0=NFC

```
> table(dummy.afc)
dummy.afc
 0  1
15 16
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.afc))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.afc)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30046 -0.11501 -0.01545  0.08196  0.47791

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.92747    0.04698  83.602   <2e-16 ***
dummy.afc    0.07424    0.06539   1.135    0.266
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1819 on 29 degrees of freedom
Multiple R-Squared: 0.04255,    Adjusted R-squared: 0.009538
F-statistic: 1.289 on 1 and 29 DF,  p-value: 0.2655
```

As we might expect, conference does not seem significantly associated with ticket prices.

– **win.pct.2000**
2000 regular season winning percentage

```
> summary(win.pct.2000)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0630  0.3440  0.5630  0.5003  0.6565  0.8130
```

```
> stem(win.pct.2000)

  The decimal point is 1 digit(s) to the left of the |

  0 | 6
  1 | 99
  2 | 55
  3 | 11188
  4 | 444
  5 | 006666
  6 | 33339999
  7 | 555
  8 | 1
```

There seems to be some minor left skew to the distribution, but there are no outliers.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~win.pct.2000))

Call:
lm(formula = log.ticket.price.2001 ~ win.pct.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32083 -0.11080 -0.01403  0.10098  0.43968

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.80715    0.08692  43.801   <2e-16 ***
win.pct.2000  0.31709    0.16201   1.957     0.06 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1748 on 29 degrees of freedom
Multiple R-Squared: 0.1167,     Adjusted R-squared: 0.08622
F-statistic: 3.831 on 1 and 29 DF,  p-value: 0.06002
```

Our simple regression shows a nearly significant correlation between ticket price and winning percentage of the previous year.

– **home.win.pct.2000**

2000 regular season home winning percentage

```
> table(home.win.pct.2000)
home.win.pct.2000
0.125  0.25 0.375   0.5 0.625  0.75 0.875
    1     1     8     5     8     5     3
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~home.win.pct.2000))

Call:
lm(formula = log.ticket.price.2001 ~ home.win.pct.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.28237 -0.11077 -0.03363  0.08417  0.45816

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.7828     0.0983  38.482   <2e-16 ***
home.win.pct.2000   0.3289     0.1674   1.965   0.0591 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1747 on 29 degrees of freedom
Multiple R-Squared: 0.1175,     Adjusted R-squared: 0.08705
F-statistic:  3.86 on 1 and 29 DF,  p-value: 0.05908
```

Home winning percentage of the previous year seems to have approximately the same correlation with ticket price and regular winning percentage.

– **off.td.2000**

Total offensive touchdowns in 2000 season

```
> summary(off.td.2000)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   6.00   15.50   20.00   20.45   22.50   37.00

> stem(off.td.2000)

  The decimal point is 1 digit(s) to the right of the |

  0 | 69
  1 | 22444
  1 | 56888899
  2 | 001112223
  2 | 889
  3 | 233
  3 | 7
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~off.td.2000))

Call:
lm(formula = log.ticket.price.2001 ~ off.td.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32149 -0.12045 -0.04428  0.11458  0.44911

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.886368   0.099273  39.148   <2e-16 ***
off.td.2000 0.003883   0.004578   0.848    0.403
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1837 on 29 degrees of freedom
Multiple R-Squared: 0.02421,    Adjusted R-squared: -0.009441
F-statistic: 0.7194 on 1 and 29 DF,  p-value: 0.4033
```

There seems to be a very weak correlation between ticket prices and offensive touchdowns.

– **number.superbowl**
Total number of Super Bowl Victories in franchise history

```
> table(number.superbowl)
number.superbowl
 0  1  2  3  4  5
16  6  3  3  1  2
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~number.superbowl))

Call:
lm(formula = log.ticket.price.2001 ~ number.superbowl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30400 -0.11270 -0.02138  0.09098  0.38195

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.93100    0.04018  97.827   <2e-16 ***
number.superbowl  0.03081    0.02124   1.451    0.158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1795 on 29 degrees of freedom
Multiple R-Squared: 0.06767,    Adjusted R-squared: 0.03553
F-statistic: 2.105 on 1 and 29 DF,  p-value: 0.1575
```

The number of Super Bowl Championships in franchise history is within striking distance of a significant correlation if we can remove some noise by controlling for other factors of variability.

- **number.probowl**
  Total number of 2000 Pro Bowl Players

```
> table(number.probowl)
number.probowl
0 1 2 3 4 5 7 9
4 7 4 4 4 4 2 2
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~number.probowl))

Call:
lm(formula = log.ticket.price.2001 ~ number.probowl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.29919 -0.10898 -0.01732  0.10292  0.41415

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.88244    0.04909  79.081   <2e-16 ***
number.probowl  0.02720    0.01244   2.187    0.037 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1723 on 29 degrees of freedom
Multiple R-Squared: 0.1415,     Adjusted R-squared: 0.1119
F-statistic: 4.781 on 1 and 29 DF,  p-value: 0.03698
```

Clearly there is a significant positive correlation between number of Pro Bowl players on a team's roster and ticket price.

3. **Market Variables**

– **city**

Nearest major city (sometimes stadium is close but not technically in the same city. For instance, New England's stadium is a 40 minute drive from Boston. Estimating relative commute times is outside the scope of this project.)

NFL Cities

```
1   San.Francisco
2         Chicago
3     Cincinnati
4         Buffalo
5          Denver
6       Cleveland
7           Tampa
8         Phoenix
9       San.Diego
10    Kansas.City
11   Indianapolis
12         Dallas
13          Miami
14   Philadelphia
15        Atlanta
16       New.York
17   Jacksonville
18       New.York
19        Detroit
20      Green.Bay
21      Charlotte
22         Boston
23        Oakland
24       St.Louis
25      Baltimore
26     Washington
27    New.Orleans
28        Seattle
29     Pittsburgh
30      Nashville
31    Minneapolis
```

– **city.pop.2000**
City Population (from 2000 Census)

```
> summary(city.pop.2000)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 102300  391100  563400 1166000  871600 8008000

> stem(city.pop.2000)

  The decimal point is 6 digit(s) to the right of the |

  0 | 133333444445555666667788
  1 | 02235
  2 | 9
  3 |
  4 |
  5 |
  6 |
  7 |
  8 | 00
```

Clearly the two teams from New York (Giants and Jets) are outliers in this category.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~city.pop.2000))

Call:
lm(formula = log.ticket.price.2001 ~ city.pop.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33885 -0.11615 -0.01075  0.09037  0.43987

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.965e+00  3.936e-02 100.745   <2e-16 ***
city.pop.2000 4.697e-10  1.787e-08   0.026     0.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.1859 on 29 degrees of freedom
Multiple R-Squared: 2.383e-005, Adjusted R-squared: -0.03446
F-statistic: 0.0006911 on 1 and 29 DF,  p-value: 0.9792
```

City population does not appear to have any correlation with ticket prices. Natural logarithm and square root transformations of population failed to provide a better fit. There is no compelling reason to transform this variable.

– **cost.of.living**

Cost of living variable which estimates relative cost of living across different cities. We will construct this variable by providing a base case, living in Cincinnati on \$40,000 per year. The number for all other cities is how much an individual would have to earn in those cities in order to maintain the same standard of living as they experience in Cincinnati, OH earning \$40,000.

```
> summary(cost.of.living)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  28450   34850   40000   52460   55670  150100

> stem(cost.of.living)

  The decimal point is 4 digit(s) to the right of the |

   2 | 812233445666778
   4 | 0045772666
   6 | 85
   8 | 99
  10 |
  12 |
  14 | 00
```

Again, New York is a clear outlier. We will use a log transform to create **log.cost.of.living**. But first...

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~cost.of.living))

Call:
lm(formula = log.ticket.price.2001 ~ cost.of.living)

Residuals:
       Min         1Q     Median         3Q        Max
-0.3292938 -0.1249895  0.0002882  0.0883311  0.4269097

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.936e+00  6.598e-02  59.648   <2e-16 ***
cost.of.living 5.719e-07  1.086e-06   0.526    0.603
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1851 on 29 degrees of freedom
Multiple R-Squared: 0.009465,   Adjusted R-squared: -0.02469
F-statistic: 0.2771 on 1 and 29 DF,  p-value: 0.6026
```

Clearly, there is not a significant correlation between ticket price and cost of living in the cities.

– **log.cost.of.living**
Natural logarithm of **cost.of.living**

```
> summary(log.cost.of.living)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.26   10.46   10.60   10.75   10.93   11.92

> stem(log(cost.of.living))

  The decimal point is 1 digit(s) to the left of the |

  102 | 657799
  104 | 448990115
  106 | 019057
  108 | 7333
  110 | 3
  112 | 2
  114 | 00
  116 |
  118 | 22
```

The natural log only lessened the right skew to a small degree. We will check the fit of our simple regression to determine if we prefer the transformed variable.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~log.cost.of.living))

Call:
lm(formula = log.ticket.price.2001 ~ log.cost.of.living)

Residuals:
      Min        1Q    Median        3Q       Max
-0.329607 -0.123788  0.001405  0.085959  0.423573

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.59595    0.82622   4.352 0.000153 ***
log.cost.of.living  0.03439    0.07676   0.448 0.657491
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1853 on 29 degrees of freedom
Multiple R-Squared: 0.006873,   Adjusted R-squared: -0.02737
F-statistic: 0.2007 on 1 and 29 DF,  p-value: 0.6575
```

Our fit is not much improved, so we would be indifferent between these two variables for cost of living. Another possible transformation of our cost of living variable is explored below.

Another transformation

An additional way to reduce right skew in **cost.of.living** is to use a $-x^{-p}$ transformation. This did, in fact, remove virtually all of our right skew (see Figure 5), but our simple linear regression (presented below) still does not seem to uncover any significant correlation between **log.ticket.price.2001** and our new transformed variable, **transform2.cost.of.living**. The Adjusted-R^2 is negative.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~transform2.cost.of.living))

Call:
lm(formula = log.ticket.price.2001 ~ transform2.cost.of.living)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.336820 -0.118073 -0.007385  0.087687  0.435266 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                3.972e+00  7.001e-02  56.736   <2e-16 ***
transform2.cost.of.living -1.052e+07  1.042e+08  -0.101     0.92    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1859 on 29 degrees of freedom
Multiple R-Squared: 0.0003516,  Adjusted R-squared: -0.03412
F-statistic: 0.0102 on 1 and 29 DF,  p-value: 0.9203
```

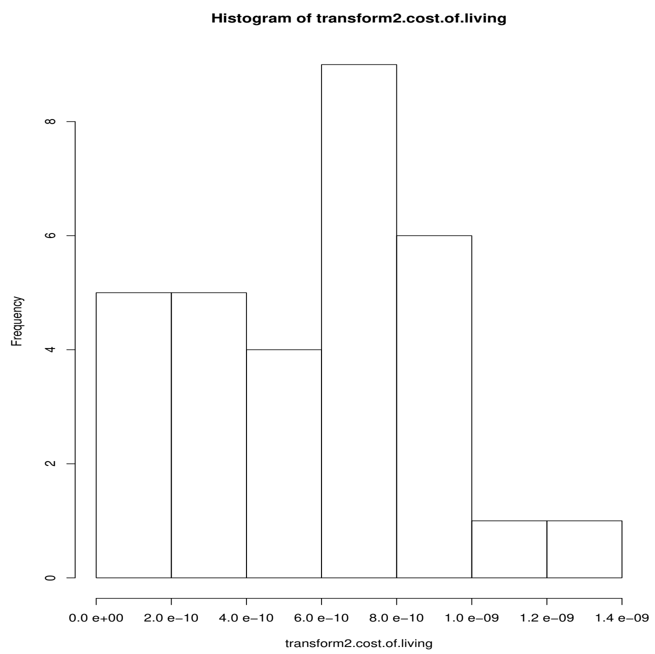

Figure 5:

– **temperature**
Mean temperature for the month of December

```
> summary(temperature)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   31.00   36.00   39.87   48.50   69.00

> stem(temperature)

  The decimal point is 1 digit(s) to the right of the |

  1 | 8
  2 | 09
  3 | 000111124566699
  4 | 1135899
  5 | 4557
  6 | 29
```

There does not appear to be any major skewing that necessitates a transformation.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~temperature))

Call:
lm(formula = log.ticket.price.2001 ~ temperature)

Residuals:
      Min        1Q    Median        3Q       Max
-0.366059 -0.100113  0.007971  0.080041  0.441273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.888809   0.116620  33.346   <2e-16 ***
temperature 0.001931   0.002804   0.688    0.497
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1844 on 29 degrees of freedom
Multiple R-Squared: 0.01608,    Adjusted R-squared: -0.01785
F-statistic: 0.4739 on 1 and 29 DF,  p-value: 0.4967
```

The correlation between average December temperature and ticket price is not close to significance.

– **dummy.basketball**

1=NBA team plays in same market, 0=else

```
> table(dummy.basketball)
dummy.basketball
 0  1
14 17
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.basketball))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.basketball)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30558 -0.10107 -0.02623  0.10131  0.47279

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.00610    0.04866  82.334   <2e-16 ***
dummy.basketball -0.07352    0.06571  -1.119    0.272
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1821 on 29 degrees of freedom
Multiple R-Squared: 0.04138,    Adjusted R-squared: 0.008326
F-statistic: 1.252 on 1 and 29 DF,  p-value: 0.2724
```

There is little or no correlation between ticket price and the presence of a local professional basketball team.

– **dummy.hockey**
1=NHL team plays in same market, 0=else

```
> table(dummy.hockey)
dummy.hockey
 0  1
13 18
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.hockey))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.hockey)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33806 -0.11549 -0.01208  0.09186  0.44031

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.966777   0.051571  76.918   <2e-16 ***
dummy.hockey -0.001708   0.067679  -0.025     0.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1859 on 29 degrees of freedom
Multiple R-Squared: 2.196e-005, Adjusted R-squared: -0.03446
F-statistic: 0.0006367 on 1 and 29 DF,  p-value:  0.98
```

There is little or no correlation between ticket price and the presence of a local professional hockey team. In fact, the Adjusted-R^2 is negative.

– **dummy.sportstown**
1=NBA, NHL and NFL teams all play in same market, 0=else

```
> table(dummy.sportstown)
dummy.sportstown
 0  1
18 13
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.sportstown))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.sportstown)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30706 -0.09310 -0.02204  0.10399  0.47131

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.98869    0.04333  92.049   <2e-16 ***
dummy.sportstown -0.05463    0.06691  -0.816    0.421
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1838 on 29 degrees of freedom
Multiple R-Squared: 0.02247,    Adjusted R-squared: -0.01124
F-statistic: 0.6666 on 1 and 29 DF,  p-value: 0.4209
```

Again, there is little or no correlation between ticket price and the presence of three professional sports teams sharing the same market.

4. **Stadium Variables**

   - **capacity**

     Stadium seating capacity

```
> summary(capacity)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  60270   65530   70270   70300   73560   85410

> stem(capacity)

  The decimal point is 4 digit(s) to the right of the |

  6 | 00134
  6 | 5556666789
  7 | 011333334
  7 | 56999
  8 | 0
  8 | 5
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~capacity))

Call:
lm(formula = log.ticket.price.2001 ~ capacity)

Residuals:
     Min       1Q   Median       3Q      Max
-0.35269 -0.08189  0.01391  0.09353  0.36894

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.637e+00  3.674e-01   9.900 8.28e-11 ***
capacity    4.676e-06  5.205e-06   0.898    0.376
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1834 on 29 degrees of freedom
Multiple R-Squared: 0.02708,    Adjusted R-squared: -0.006468
F-statistic: 0.8072 on 1 and 29 DF,  p-value: 0.3763
```

As we might expect, there is absolutely no correlation between capacity and ticket price.

- **attendance.2000**
  2000 average attendance

```
> summary(attendance.2000)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  44960   60070   65600   65820   72810   80930

> stem(attendance.2000)

  The decimal point is 4 digit(s) to the right of the |

  4 |
  4 | 5
  5 | 34
  5 | 5789
  6 | 0003344
  6 | 5667889
  7 | 0334
  7 | 66889
  8 | 1
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~attendance.2000))

Call:
lm(formula = log.ticket.price.2001 ~ attendance.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37239 -0.11720 -0.02912  0.10550  0.33233

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.498e+00  2.443e-01  14.320 1.11e-14 ***
attendance.2000 7.101e-06  3.681e-06   1.929   0.0635 .
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.1751 on 29 degrees of freedom
Multiple R-Squared: 0.1137,     Adjusted R-squared: 0.08318
F-statistic: 3.722 on 1 and 29 DF,  p-value: 0.06355
```

The correlation between attendance and ticket price is on the verge of significance with $p = .0635$.

– **capacity.utilized**
Equal to **attendance** divided by **capacity**
(Maximum value allowed is .999)
Note: Must calculate Denver and Pittsburgh using 2000 stadium capacity, since they opened new stadiums in 2001. Our **capacity** *variable will be relevant for above calculation for all other teams.*

```
> summary(capacity.utilized)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.6140  0.9330  0.9810  0.9392  0.9970  0.9990

> stem(capacity.utilized)

  The decimal point is 1 digit(s) to the left of the |

  6 | 1
  6 |
  7 | 4
  7 | 6
  8 | 3
  8 | 8
  9 | 0224
  9 | 5556668888999
  10 | 000000000
```

Attempts to transform this left-skewed variable did not meet with much success.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~capacity.utilized))

Call:
lm(formula = log.ticket.price.2001 ~ capacity.utilized)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30461 -0.12668 -0.03988  0.09350  0.43403

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)         3.3728     0.3375   9.994 6.69e-11 ***
capacity.utilized   0.6314     0.3577   1.765   0.0881 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1767 on 29 degrees of freedom
Multiple R-Squared: 0.09699,    Adjusted R-squared: 0.06585
F-statistic: 3.115 on 1 and 29 DF,  p-value: 0.08812
```

The correlation between **capacity.utilized** and ticket prices is very close to significance.

– **pop.per.seat**
Equal to **city.pop.2000** divided by **capacity**

```
> summary(pop.per.seat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.680   5.702   7.383  15.880  12.490 100.800

> stem(pop.per.seat)

  The decimal point is 1 digit(s) to the right of the |

   0 | 2455555666677777889
   1 | 00123788
   2 | 3
   3 |
   4 | 3
   5 |
   6 |
   7 |
   8 |
   9 |
  10 | 11
```

New York and Chicago are the home cities for 3 outliers in the distribution.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~pop.per.seat))

Call:
lm(formula = log.ticket.price.2001 ~ pop.per.seat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33855 -0.11614 -0.01219  0.09661  0.43861

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.9674831  0.0402610  98.544   <2e-16 ***
pop.per.seat -0.0001069  0.0014162  -0.076     0.94
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1859 on 29 degrees of freedom
Multiple R-Squared: 0.0001965,  Adjusted R-squared: -0.03428
F-statistic: 0.005701 on 1 and 29 DF,  p-value: 0.9403
```

We achieve an insignificant correlation between ticket prices and population per stadium seat. We apply a log transform to this variable to help reduce right-skew. That may provide a better fit.

– **log.pop.per.seat**
Natural logarithm of **pop.per.seat**

```
> summary(log.pop.per.seat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.519   1.741   1.999   2.259   2.524   4.613

> stem(log.pop.per.seat)

  The decimal point is at the |

  0 | 5
  1 | 4
  1 | 56667788899
  2 | 000112334
  2 | 56899
  3 | 1
  3 | 8
  4 |
  4 | 66
```

The outliers have been pulled much closer to the rest of the data. Now, let's see if we have improved the correlation with ticket prices.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~log.pop.per.seat))

Call:
lm(formula = log.ticket.price.2001 ~ log.pop.per.seat)

Residuals:
     Min       1Q   Median       3Q      Max
-0.32036 -0.10301 -0.03159  0.12595  0.42921

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.03148    0.09300  43.350   <2e-16 ***
log.pop.per.seat -0.02909    0.03848  -0.756    0.456
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.1841 on 29 degrees of freedom
Multiple R-Squared: 0.01932,    Adjusted R-squared: -0.0145
F-statistic: 0.5713 on 1 and 29 DF,  p-value: 0.4558
```

We are still left with a correlation that is not significant, even after the transformation to reduce right skew. This is surprising since it might make sense that a high concentration of people would increase market demand, thus producing upward pressure on prices. We have not found evidence to support this hypothesis.

– **dummy.outdoor**
1=outdoor stadium, 0=dome

```
> table(dummy.outdoor)
dummy.outdoor
 0  1
 6 25
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.outdoor))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.outdoor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.36624 -0.14367  0.02996  0.10919  0.41213

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.85136    0.07213  53.395   <2e-16 ***
dummy.outdoor  0.14188    0.08032   1.766   0.0878 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1767 on 29 degrees of freedom
Multiple R-Squared: 0.09715,    Adjusted R-squared: 0.06601
F-statistic:  3.12 on 1 and 29 DF,  p-value: 0.08785
```

It looks like we're close to seeing a significant positive correlation between outdoor stadiums and ticket prices. My suspicion, however, is that this may be a reflection of the fact that very few domes have been built recently. Also, of the 31 stadiums in the league, only 6 are indoor dome facilities. So our sample of indoor stadiums is also very small.

– **dummy.artificial.turf**
1=artificial turf playing surface, 0=natural grass

```
> table(dummy.artificial.turf)
dummy.artificial.turf
 0  1
22  9
```

Just 5 years ago, half of the league was playing on artificial surfaces. Now, fewer than one-third play on those surfaces. Let's check out the correlation with ticket prices.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.artificial.turf))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.artificial.turf)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38465 -0.09823  0.01744  0.09296  0.39372

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.01166    0.03636 110.346   <2e-16 ***
dummy.artificial.turf -0.15800    0.06747  -2.342   0.0263 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1705 on 29 degrees of freedom
Multiple R-Squared: 0.159,      Adjusted R-squared:  0.13
F-statistic: 5.483 on 1 and 29 DF,  p-value: 0.02628
```

We have a significant correlation with $p = .0263$, but again we have a variable that is highly associated with the age of the stadium. All of the new stadiums are built with natural grass as the playing surface.

– **dummy.new.stadium**

1=new stadium built in last five years, 0=else

```
> table(dummy.new.stadium)
dummy.new.stadium
 0  1
23  8
```

8 teams (26 percent) are playing in stadiums which opened in the last five years. The following regression suggests that we have identified a very important factor for predicting ticket prices.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.new.stadium))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.new.stadium)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27069 -0.07939  0.01193  0.09852  0.24382

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.89769    0.02973 131.113  < 2e-16 ***
dummy.new.stadium  0.26386    0.05852   4.509  9.9e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1426 on 29 degrees of freedom
Multiple R-Squared: 0.4121,     Adjusted R-squared: 0.3919
F-statistic: 20.33 on 1 and 29 DF,  p-value: 9.905e-005
```

Our correlation is highly significant and positive, as we might have expected. What is most surprising about this variable is the incredibly large amount of variance in ticket prices that it explains. We have achieved an R^2 value of 0.4121.

Figure 6 clearly shows that untransformed ticket prices for the 2001 season are distinctly higher for teams with a recently built stadium. The Interquartile ranges of our two boxplots do not even overlap.

The following regression output calculates the untransformed average ticket

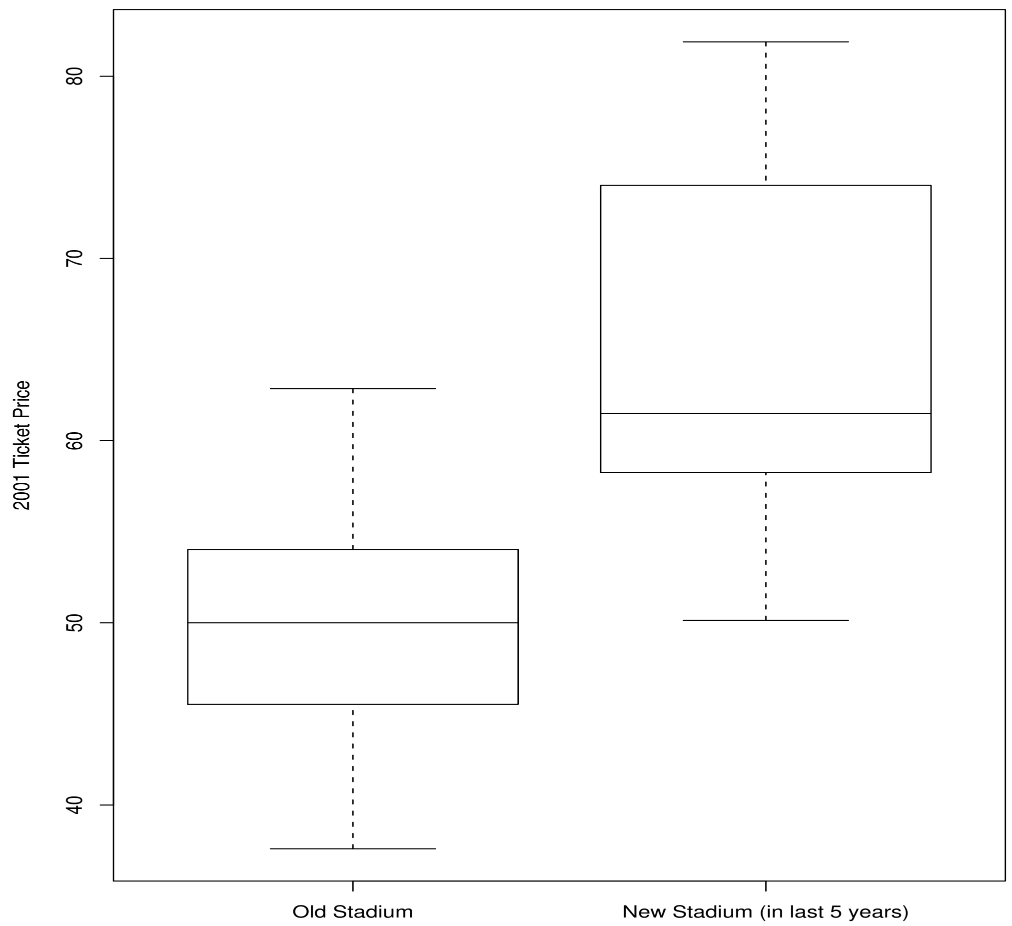

Figure 6: 2001 ticket prices by New vs. Old Stadium

prices for teams with and without a new stadium. A recently built stadium looks to net an extra \$15, on average, for tickets.

```
> summary(lm(ticket.price.2001~dummy.new.stadium+dummy.no.new.stadium-1))

Call:
lm(formula = ticket.price.2001 ~ dummy.new.stadium +
dummy.no.new.stadium -
    1)

Residuals:
     Min       1Q   Median       3Q      Max
-14.8012  -4.6761   0.1691   5.2539  16.9488

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
dummy.new.stadium      64.941      2.752   23.60   <2e-16 ***
dummy.no.new.stadium   49.711      1.623   30.63   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 7.783 on 29 degrees of freedom
Multiple R-Squared: 0.981,      Adjusted R-squared: 0.9797
F-statistic: 747.7 on 2 and 29 DF,  p-value:     0
```

5. **Financial Variables**

– **salary.cap.room**

As of July, 2001, the amount of money below the salary cap the team was eligible to spend.

```
> stem(salary.cap.room)

  The decimal point is at the |

  0 | 0112235677
  1 | 0111123689
  2 | 334
  3 | 15
  4 | 135
  5 | 1
  6 | 1
  7 | 3
```

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~salary.cap.room))

Call:
lm(formula = log.ticket.price.2001 ~ salary.cap.room)

Residuals:
      Min        1Q    Median        3Q       Max
-0.369616 -0.079788 -0.008668  0.078845  0.398809

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      4.03999    0.04480  90.179   <2e-16 ***
salary.cap.room -0.03759    0.01650  -2.278   0.0303 *
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.1713 on 29 degrees of freedom
Multiple R-Squared: 0.1518,     Adjusted R-squared: 0.1226
F-statistic:  5.19 on 1 and 29 DF,  p-value: 0.03026
```

Figure 7 and Figure 8 show the residuals for the two regression models

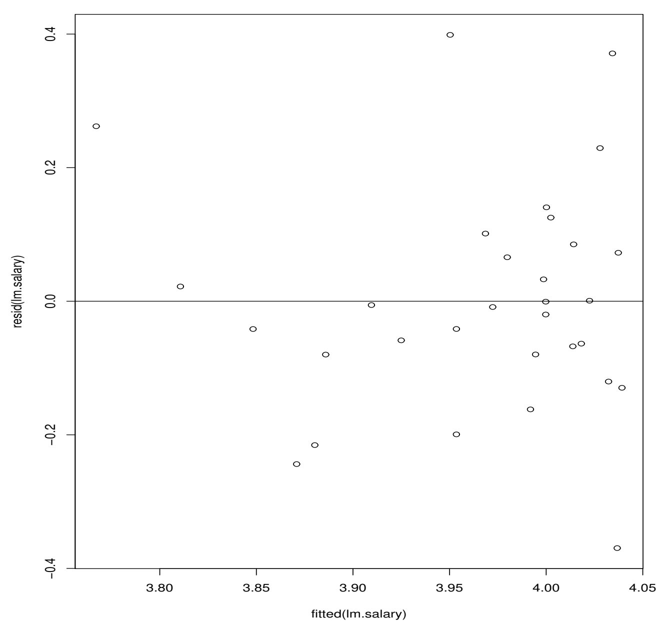

Figure 7: Residuals of **salary.cap.room** regression

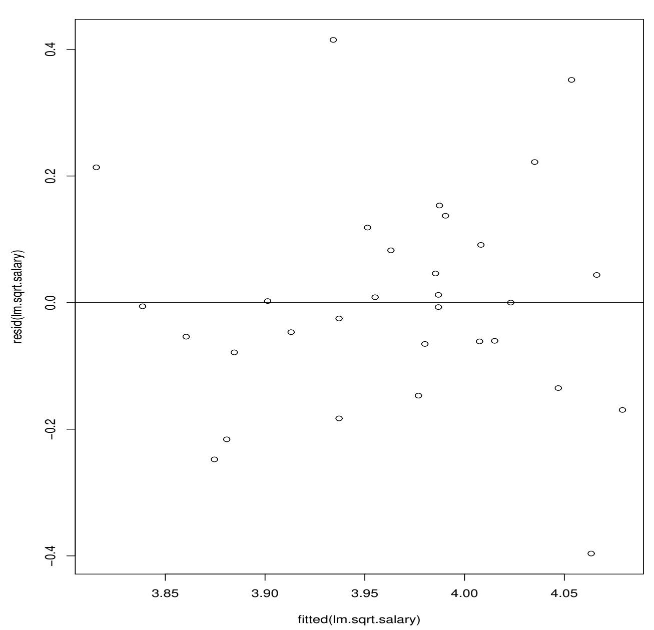

Figure 8: Residuals of **sqrt.salary.cap.room** regression

over **salary.cap.room** and **sqrt.salary.cap.room**. Figure (7) displays a definite positive trend in the residuals as the fitted values increase. While Figure 8 does not entirely remove this trend, it suggests that the square root transformation lessens the degree to which the pattern violates our goal of random normal residuals.

We will therefore introduce this transformation **sqrt.salary.cap.room** as our preferred variable for salary cap room.

– **sqrt.salary.cap.room**
The square root of **salary.cap.room**

```
> stem(sqrt(salary.cap.room))

  The decimal point is at the |

  0 | 1334
  0 | 567888
  1 | 0000011334
  1 | 55579
  2 | 0113
  2 | 57
```

We have eliminated virtually all of the right skew using the square root transform. The following model produced the residuals plot in Figure 8.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~sqrt(salary.cap.room)))

Call:
lm(formula = log.ticket.price.2001 ~ sqrt(salary.cap.room))

Residuals:
      Min        1Q    Median        3Q       Max
-0.396309 -0.071944 -0.005856  0.086869  0.414930

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            4.09375    0.06543  62.563   <2e-16 ***
sqrt(salary.cap.room) -0.10330    0.04657  -2.218   0.0345 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1719 on 29 degrees of freedom
Multiple R-Squared: 0.1451,     Adjusted R-squared: 0.1156
F-statistic:  4.92 on 1 and 29 DF,  p-value: 0.03453
```

– **dummy.top15.merchandise**
1=team is among top 15 in merchandise sales, 0=else

```
> table(dummy.top15.merchandise)
dummy.top15.merchandise
 0  1
16 15
```

The coding should be self-explanatory. We would have preferred to have quantitative data for merchandise sales, but unfortunately only a ranking was available to us. We view this variable as an indicator of team popularity more than a significant component of a team's financial health because we suspect that the teams do not receive all of the proceeds from merchanise sales related to their franchise. It is likely that the National Football League absorbs a fair portion of the revenues. During the course of this analysis this hypothesis could not be proven or disproven.

Simple linear regression:

```
> summary(lm(log.ticket.price.2001~dummy.top15.merchandise))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.top15.merchandise)

Residuals:
      Min        1Q    Median        3Q       Max
-0.285291 -0.108559 -0.008505  0.110788  0.493082

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.91229    0.04424  88.436   <2e-16 ***
dummy.top15.merchandise  0.11055    0.06360   1.738   0.0928 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.177 on 29 degrees of freedom
Multiple R-Squared: 0.09436,    Adjusted R-squared: 0.06313
F-statistic: 3.021 on 1 and 29 DF,  p-value: 0.09278
```

This correlation is on the precipice of significance in the expected positive direction.

6. Multivariate Model: A first stab

```
>summary(lm(log.ticket.price.2001~dummy.new.stadium+sqrt.salary.cap.room+
                                                dummy.top15.merchandise))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.new.stadium +
sqrt.salary.cap.room +
    dummy.top15.merchandise)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22439 -0.05924 -0.01029  0.04552  0.25235

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.90948    0.05765  67.813  < 2e-16 ***
dummy.new.stadium        0.26733    0.05040   5.304 1.35e-05 ***
sqrt.salary.cap.room    -0.06121    0.03324  -1.841  0.07657 .
dummy.top15.merchandise  0.13049    0.04384   2.977  0.00608 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1196 on 27 degrees of freedom
Multiple R-Squared: 0.6149,     Adjusted R-squared: 0.5721
F-statistic: 14.37 on 3 and 27 DF,  p-value: 8.678e-006
```

7. Interactions: A first stab

```
> summary(lm(log.ticket.price.2001~dummy.basketball*win.pct.2000))

Call:
lm(formula = log.ticket.price.2001 ~ dummy.basketball * win.pct.2000)

Residuals:
     Min       1Q   Median       3Q      Max
-0.31866 -0.09394 -0.02723  0.07049  0.45975

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.02526    0.12654  31.810   <2e-16 ***
dummy.basketball              -0.38016    0.16767  -2.267   0.0316 *
win.pct.2000                  -0.03635    0.22489  -0.162   0.8728
dummy.basketball:win.pct.2000  0.63743    0.31046   2.053   0.0499 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1659 on 27 degrees of freedom
Multiple R-Squared: 0.2587,     Adjusted R-squared: 0.1763
F-statistic:  3.14 on 3 and 27 DF,  p-value: 0.04159
```

The above interaction of **dummy.basketball** and **win.pct.2000** has significance at $p = .05$ in the main interaction term as well as the lower-order term **dummy.basketball**. This significance provides a fairly intuitive interpretation.

The negative and significant coefficient on **dummy.basketball** suggests that football and basketball are substitutes. It makes economic sense that having an additional sports team (a substitute) in the same city would tend to lessen demand pressures and consequently be associated with lower prices for football tickets. The presence of a professional basketball team, however, might at the same time be considered evidence that the city is a “sports town”. A city which falls into this category might be more responsive when its team starts having success. That is the effect we may have discovered in our main interaction above. The positive and significant coefficient on our main interaction supports the hypothesis that winning percentage matters in a sports town.

A rise in demand produces the following economic reality: *In the wake of* *a winning season, the owners raise prices to meet the town's excitement and match the higher demand for tickets to watch its winning football team the following season.*

8. Multivariate Model: A second stab

```
>summary(lm(log.ticket.price.2001~sqrt.salary.cap.room+dummy.new.stadium+dummy.top15.merchandise+win.pct.2000*dummy.basketball))

Call:
lm(formula = log.ticket.price.2001 ~ sqrt.salary.cap.room +
dummy.new.stadium +
    dummy.top15.merchandise + win.pct.2000 * dummy.basketball)

Residuals:
      Min        1Q    Median        3Q       Max
-0.187137 -0.064023 -0.002681  0.055091  0.203673

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.16117    0.10263  40.545  < 2e-16 ***
sqrt.salary.cap.room          -0.08645    0.03064  -2.821  0.00945 **
dummy.new.stadium              0.26698    0.04516   5.911 4.24e-06 ***
dummy.top15.merchandise        0.08976    0.04125   2.176  0.03961 *
win.pct.2000                  -0.38084    0.15123  -2.518  0.01886 *
dummy.basketball              -0.37022    0.10816  -3.423  0.00223 **
win.pct.2000:dummy.basketball  0.73556    0.20411   3.604  0.00142 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1021 on 24 degrees of freedom
Multiple R-Squared: 0.7503,     Adjusted R-squared: 0.6878
F-statistic: 12.02 on 6 and 24 DF,  p-value: 3.175e-006
```

Clearly, we have a model with several significant variables. The R^2 of .75 is also impressive to achieve on so few observations (31).

Figure 9 displays several diagnostic plots for our regression model. The residual cloud in “Residuals vs. Fitted” is very nicely spread without any discernible pattern. The normal quantile plot also forms a nice straight line, revealing no obvious departure from normality. Observations 6, 15, and 26 are identified as having large values in our Cook’s Distance plot.

Below we run the model again, excluding those three observations that stand out in our Cook’s Distance plot.

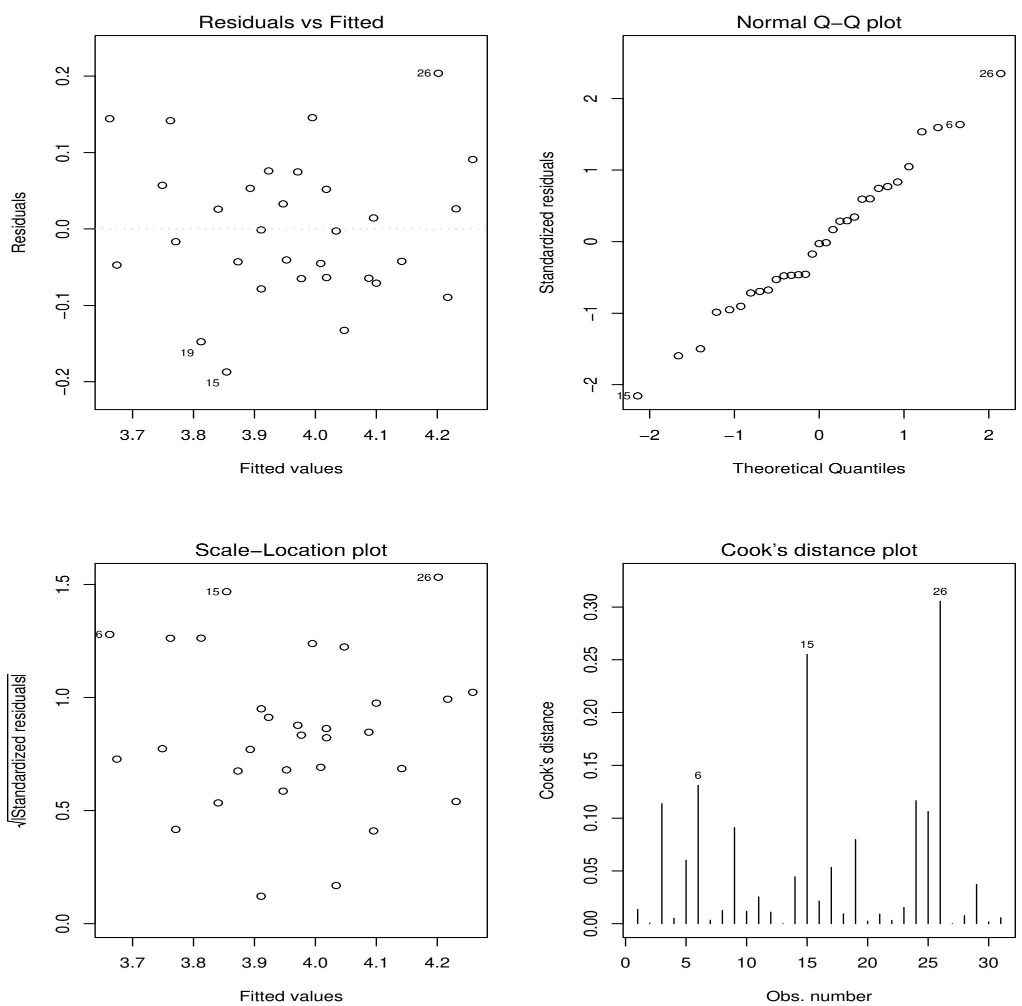

Figure 9: Diagnostic Plots of Multiple Regression

```
>summary(lm(log.ticket.price.2001~sqrt.salary.cap.room+dummy.new.stadium+dummy.top15.merchandise+win.pct.2000*dummy.basketball,data=nfl[-c(6,15,26),]))

Call:
lm(formula = log.ticket.price.2001 ~ sqrt.salary.cap.room +
dummy.new.stadium +
    dummy.top15.merchandise + win.pct.2000 * dummy.basketball,
    data = nfl[-c(6, 15, 26), ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.12975 -0.05457 -0.00478  0.04173  0.14071

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.18145    0.08996  46.481  < 2e-16 ***
sqrt.salary.cap.room          -0.09792    0.02929  -3.343 0.003087 **
dummy.new.stadium              0.23034    0.03863   5.962 6.44e-06 ***
dummy.top15.merchandise        0.09701    0.03426   2.832 0.009996 **
win.pct.2000                  -0.37755    0.12622  -2.991 0.006961 **
dummy.basketball              -0.38123    0.09462  -4.029 0.000607 ***
win.pct.2000:dummy.basketball  0.72611    0.17746   4.092 0.000522 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08241 on 21 degrees of freedom
Multiple R-Squared: 0.7948,     Adjusted R-squared: 0.7361
F-statistic: 13.55 on 6 and 21 DF,  p-value: 2.85e-006
```

A remarkable increase in both R^2 and Adjusted-R^2 has resulted from the removal of those observations. Each of our coefficients has remained significant.

9. Multivariate Model: A third stab

```
Call:
lm(formula = log.ticket.price.2001 ~ sqrt.salary.cap.room +
             dummy.new.stadium + dummy.top15.merchandise +
             win.pct.2000 + dummy.sportstown +
             win.pct.2000:dummy.sportstown, data = nfl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16233 -0.05939 -0.01330  0.04520  0.21908

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.15299    0.09838  42.216  < 2e-16 ***
sqrt.salary.cap.room          -0.10372    0.03180  -3.262  0.00330 **
dummy.new.stadium              0.25798    0.04470   5.771    6e-06 ***
dummy.top15.merchandise        0.09867    0.04158   2.373  0.02601 *
win.pct.2000                  -0.32979    0.14116  -2.336  0.02815 *
dummy.sportstown              -0.40183    0.10978  -3.660  0.00124 **
win.pct.2000:dummy.sportstown  0.76472    0.21267   3.596  0.00145 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1013 on 24 degrees of freedom
Multiple R-Squared: 0.7544,     Adjusted R-squared: 0.693
F-statistic: 12.28 on 6 and 24 DF,  p-value: 2.626e-006
```

This model has greater overall significance, as measured by the F-statistic, than the regression utilizing the interaction between basketball and winning percentage. The variable **dummy.sportstown** makes for a more significant set of coefficients for both the interaction and the lower order dummy variable. The significant negative correlation between winning percentage and ticket price among teams who do not share their market with hockey and basketball teams does not have an intuitive interpretation. Why would there ever be a *negative* correlation between winning percentage and and ticket price among any subset of teams?

Figure 10 displays several diagnostic plots for our regression model. Observation numbers 11,15,24 and 26 are identified in our plots as potential outliers. We will run our model again without those observations to check if our vari-

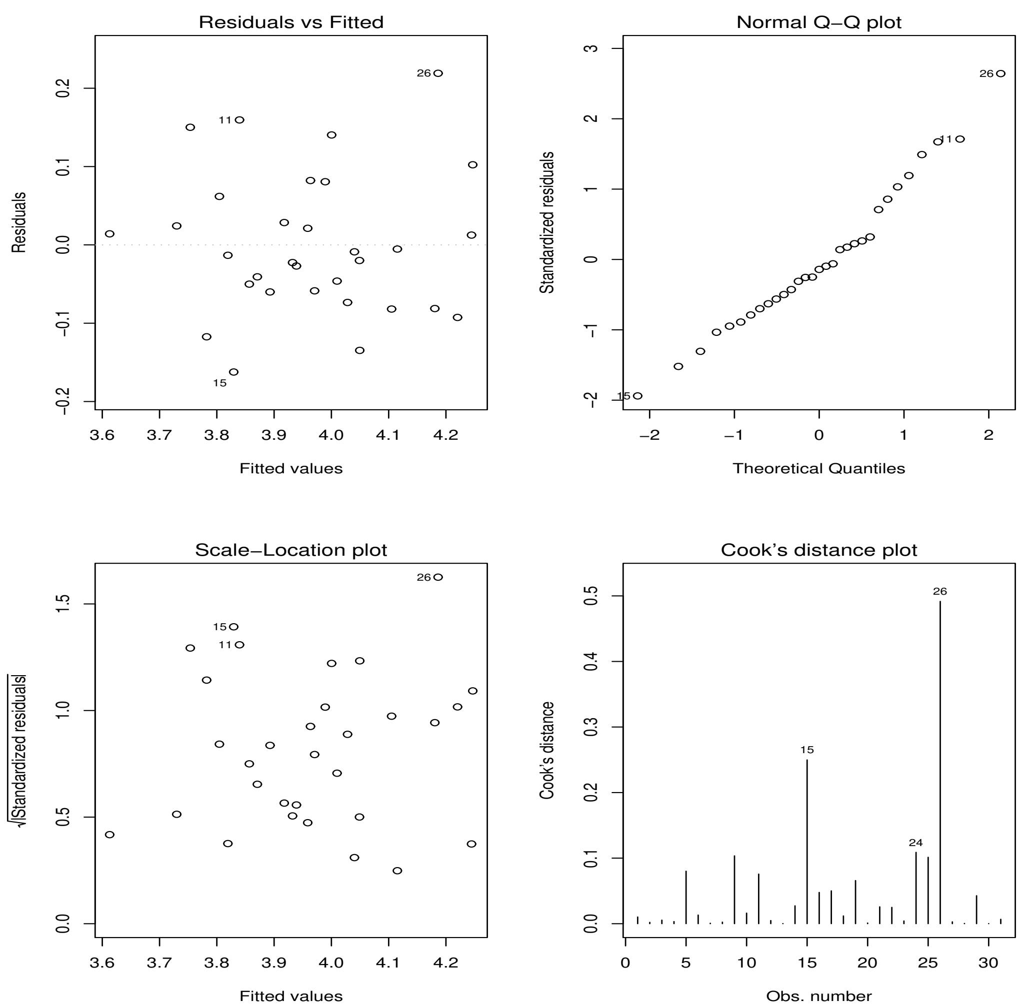

Figure 10: Diagnostic Plots of Multiple Regression

ables are still significant.

```
Call:
lm(formula = log.ticket.price.2001 ~ sqrt.salary.cap.room +
dummy.new.stadium +
    dummy.top15.merchandise + win.pct.2000 * dummy.sportstown,
    data = nfl[-c(11, 15, 24, 26), ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.090472 -0.057495  0.006619  0.027448  0.128216

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    4.18488    0.07827  53.465  < 2e-16 ***
sqrt.salary.cap.room          -0.11524    0.02711  -4.250 0.000392 ***
dummy.new.stadium              0.25807    0.03519   7.333 4.35e-07 ***
dummy.top15.merchandise        0.14225    0.03192   4.456 0.000242 ***
win.pct.2000                  -0.44736    0.11260  -3.973 0.000749 ***
dummy.sportstown              -0.37277    0.07835  -4.758 0.000120 ***
win.pct.2000:dummy.sportstown  0.72785    0.15410   4.723 0.000130 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07012 on 20 degrees of freedom
Multiple R-Squared: 0.8624,     Adjusted R-squared: 0.8211
F-statistic: 20.89 on 6 and 20 DF,  p-value: 1.228e-007
```

Removing those outliers brings all of our coefficients to a level of significance of .001. Also, our R^2 has risen to .86. We therefore cannot conclude that our results are being driven by a select few highly influential observations.

Extra Sum of Squares tests were insignificant for the inclusion of any additional variables in this model. Therefore, we retain this model as our final specification.

10. "Case-based" Bootstrapping

We will test our model using "case-based" bootstrapping. Normally, we would be inclined to perform this kind of analysis in situations where our normal-errors assumption was in doubt. Our diagnostic plots do not actually reveal any strong sense of non-normality. However, our small number (31) of cases motivate us to test the sensitivity of our coefficient estimates to different combinations (with repeats) of our observation vectors.

Since all of the coeffients in our model are individually significant at p=.05, we know that zero is not included in their 95 percent Normal confidence intervals. Bootstrap samples give us slightly different (larger) confidence intervals to test the significance of our coefficients. Below, we provide the 95 percent, percentile-based bootstrap intervals for our coefficients.

Bootstrap percentile-based 95% confidence intervals

Intercept

```
> # 95% percentile-based bootstrap interval for b0
> quantile(boot.data$b0, c(0.025,0.975))

    2.5%    97.5%
3.857473 4.370667
```

sqrt.salary.cap.room

```
> # 95% percentile-based bootstrap interval for b1
> quantile(boot.data$b1, c(0.025,0.975))

       2.5%       97.5%
-0.18335735 -0.02362473
```

dummy.new.stadium

```
> # 95% percentile-based bootstrap interval for b2
> quantile(boot.data$b2, c(0.025,0.975))

     2.5%     97.5%
0.1305388 0.3612133
```

dummy.top15.merchandise

```
> # 95% percentile-based bootstrap interval for b3
```

```
> quantile(boot.data$b3, c(0.025,0.975))

       2.5%       97.5%
0.009822568 0.180039609
```

win.pct.2000

```
> # 95% percentile-based bootstrap interval for b4
> quantile(boot.data$b4, c(0.025,0.975))

      2.5%      97.5%
-0.6189132  0.0940375
```

dummy.sportstown

```
> # 95% percentile-based bootstrap interval for b5
> quantile(boot.data$b5, c(0.025,0.975))

      2.5%      97.5%
-0.5964222 -0.1676532
```

win.pct.2000:dummy.sportstown

```
> # 95% percentile-based bootstrap interval for b6
> quantile(boot.data$b6, c(0.025,0.975))

     2.5%     97.5%
0.2786191 1.1676652
```

Clearly, all but one of our coefficients are significantly different from zero (i.e. zero is not in the 95 percent interval). The coefficient on **win.pct.2000** includes the value zero in its 95 percent bootstrap confidence interval. This causes us to give pause when attempting to interpret the relationship being measured by this variable.

Since **win.pct.2000** is a lower order term for the interaction with **dummy.sportstown**, its coefficient represents the relationship between winning percentage and ticket prices for those teams in cities without an NBA or NHL team, where **dummy.sportstown**=0. The significant, negative relationship found in our regression t-test was not upheld with the application of our bootstrapping 95% confidence interval analysis. We will, however, keep the term in our model, since it is customary to always retain lower order terms when a significant interaction is to be kept in the model.

Bibliography and Credits

I conferred with Brian Junker on the analysis found in this report and used his class handouts and notes extensively. The data was collected entirely from internet sources.

Sources

- CoachBox.com
  `http://www.coachbox.com/cap.htm`
- ESPN Webpage
  `http://sports.espn.go.com/nfl/attendance?year=2000`
- Milwaukee Journal Sentinel Web Page
  `http://www.jsonline.com/sports/sday2/sday121698.asp`
- Monstermoving.com Salary Comparison Calculator
  `http://www.monstermoving.com/Find_A_Place/Calculators/SalaryCalc/index.asp`
- NFL.com
  `http://www.nfl.com/probowl2001/afc.html`
- StatFox.com
  `http://www.statfox.com/nfl/weather/`
- SuperBowl.com
  `http://www.superbowl.com/u/xxxv/history/recaps/`
- United States Census Bureau Homepage
  `http://www.census.gov/Press-Release/www/2001/tables`
- USA Today Webpage
  `http://www.usatoday.com/sports/nfl/stories/2001-09-05-ticket-prices.htm`
- Weather.com
  `http://www.weather.com/`