May 2006

Projection Bias in eBay Purchases: A Statistical Analysis

Chirag Mahatma
University of Pennsylvania

Projection Bias in eBay Purchases: A Statistical Analysis

Projection Bias in eBay Purchases:
A Statistical Analysis

Chirag Mahatma
Wharton Research Scholars
May 2006

1. Introduction

Numerous studies have shown that human beings deviate from rationality in their decision making processes. These studies have discovered a number of psychological biases, one of which is projection bias.

Projection bias is the systematic tendency for people to underestimate the magnitude of change in their tastes, while qualitatively understanding the direction of such shifts. Such a tendency affects not only economic decisions, but also life-related decisions.

For instance, projection bias may lead to the purchase of items, such as winter gear, or durable home goods, that are not needed or that will eventually end up being returned, due to the incorrect quantitative assessment of changes in tastes. More importantly though, this bias leads people to underestimate their ability to adapt to changes in life circumstances, affecting which job they may accept or where they choose to live in a potentially detrimental fashion. Because of such effects, it is important to understand the extent to which projection bias influences real-world decisions, big and small.

1.1 Previous Studies

A number of studies have been conducted, utilizing eBay auctions or auction data, to test for the effects of the bias in question. eBay serves as an excellent virtual marketplace, with numerous sellers and buyers and minute-to-minute data records.

One such study, conducted by Ariely and Simonson, explores projection bias in the context of value assessments of CD’s and DVD’s put up for auction. It was shown that winning bidders tend to overpay between 5 and

15% for items. While not completely attributable to projection bias, it shapes the buyer's value assessment, likely leading to a belief that the item will provide the same level of utility obtained from immediate use in the future as well.

eBay auctions, however, are not the only transactions where this bias may play a role. Conlin, O'Donoghue, and Vogelsang (2005) conducted a study utilizing catalog orders of winter weather-related items, such as coats, gloves, etc… Their primary empirical finding was an inverse relationship between the temperature on the date of order and the probability of return after order receipt. Moreover, effects such as learning were shown not to be the driving factors behind such a finding.

This paper explores the effects of projection bias via eBay auction data. In section 2, a model of the bias' effects on purchases is presented. Sources of and collection of the data are covered in Section 3. In Section 4, the results of the statistical analysis are presented. Section 5 concludes.

2. Model of Projection Bias and eBay Purchases

Under the standard economic model of rational decision-making, utility is not affected by things such as framing, anchoring, and so forth. Extending this further, prices and changes in those prices should then be determined by rational factors and changes in those factors.

The model proposed here does not take that stance. Instead, changes in prices are assumed to be affected by variables of an "irrational" nature. If the standard model is correct, these variables should have no effect on prices or price changes.

To test this hypothesis, a multiple regression analysis utilizing eBay data was conducted. If the aforementioned

variables have no effect on price, the associated coefficients should not be statistically significant.

Variation in quality, even for the same item, is a common occurrence on eBay. In order to avoid this possibly confounding factor, fairly standardized items were chosen, namely tickets to sporting events, in this case basketball games, and trading cards of certain basketball players. Such items are rather consistent across levels, as tickets are sold for certain prices by seating level, and trading cards for a player by the same company are all produced from one design. Furthermore, these items are also well-priced by markets outside of eBay, whether by arena management, or independent pricing guides. Both of these aspects assisted in ensuring proper testing of the model.

3. Data

Data for the regression analysis was collected from a number of sources. To avoid over-collection, the category of tickets was limited to those for games involving the Los Angeles Lakers or the Philadelphia 76ers; the category of trading cards was limited to those of Kobe Bryant or Allen Iverson.

3.1 eBay Data

To obtain the necessary eBay auction data, software named DeepAnalysis2.1 was utilized. Data on one month's worth of auctions was collected for both categories. In addition, data on one month's worth of auctions of the previous year was also collected. Conversion of the data to a more practical format was then done through a program called Able2Extract.

3.2 Performance Data

Team and player performance-related data had to be collected from a

number of different resources, primarily from the World Wide Web. Sites such as ESPN.com™ and Yahoo!Sports™ were used to obtain schedules, statistics, injured player lists and so forth. Betting websites focusing on the NBA were also utilized for these purposes. This data was collected so as to temporally correlate with the eBay auction data.

3.3 Compilation & Collation

While much of the performance-related data was able to be used in its raw format, the auction data had to be formatted to suit analytical purposes.

In order to form a proper sequence, the ticket auctions were organized by game date. This made certain that there were not any complications due to temporal issues. Moreover, for each game date, the associated data was averaged, producing a single data point. With regards to card auctions, an event window of two days before and after a game date was used to form a data point, similar in construction to that for tickets.

3.4 Imputation

Certain eBay data points were not able to be constructed in the above fashion, as auction data for certain game dates was not available.

In order to handle this issue, multiple data imputation utilizing stochastic regression was performed. A regression utilizing all available data points was done first. This was then used to calculate the missing values. Afterwards, a random residual from a normal distribution with the same first two moments as the regressions residuals was added to each calculated value; this introduced variation in the data. The process of adding a random residual to the calculated value was then repeated twice, for a total of three times. The obtained values were then averaged

to form the final imputed value for that data point.

(Stochastic regression was chosen as the method of imputation due to the addition of a random residual to initially calculated values. This introduced more variation into the data than methods such as “nearest-neighbor” or “hot deck” would have introduced.)

4. Regression Analysis

4.1 Covariates

A number of predictor variables were utilized in the analysis. They are discussed here briefly.

With regards to player performance, three primary covariates were used. Whether the player’s performance was above or below average was captured with a categorical variable, Player Performance, having two levels: below or above. In this case, above or below was determined by using the player’s average points per game. If the amount of points the player scored in a game was above that value, Player Performance was set to above, and vice versa.

The other two variables used were Player Points, and Alpha. Player Points is a semi-continuous variable that is simply the number of points scored by the player in a particular game. Alpha is a constructed, continuous variable that measures the proportion of the team’s final point score contributed by the player. It was found by dividing player points by the team’s final point score.

With regards to team performance, the analysis made use of a number of covariates. One category was focused on game performance, while the other on season performance (limited to the month of auction data).

With regards to game performance, three variables were used. The outcome of a game was captured

using a categorical variable, Win/Loss, having two categories, win or loss. A sports-betting inspired variable was also included, namely, Spread. This was found by subtracting the opponent's score from the team's score in the case of a win, and vice versa in the case of a loss, allowing the variable to take on negative and positive values. Finally, a semi-continuous variable, Final Score, simply the amount of points scored by the team in the game, was utilized.

Regarding season performance, two covariates were used. The team's winning or losing streak was captured in a semi-continuous variable named Streak. If the team had consecutively won the past 3 games, Streak took on a value of 3. However, if the team then lost the 4^{th} game, Streak took on a value of -1, allowing the variable to take on positive and negative values, and also provide increased precision. The team's record, i.e. 40-40, was transformed into a semi-continuous variable called Current Net Record; this was done by subtracting the number of losses from the number of wins in the case that there were more wins than losses, and vice versa for more losses than wins.

Another variable was also used; however, it crossed both categories of performance: # Players Injured. This variable captured the number of players of the team that were injured during a particular game. However, as injuries lasted for more than one game, there was a seasonal aspect to the variable as well.

4.2 Exploratory Data Analysis

Before more serious analysis was conducted, a certain amount of preliminary analysis was performed. The primary purpose of this analysis was to ascertain the normality of the various covariates, excluding those of a categorical nature however. In lieu of

examining normal quantile plots or checking the goodness-of-fit of a fitted normal distribution for each variable, a preliminary regression was conducted for each team and each player. The residuals from these regressions were then saved, and examined.

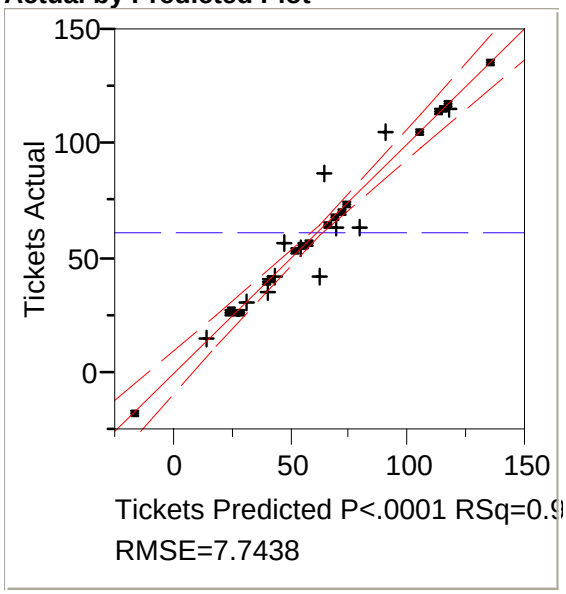
In all four regressions, the residuals were found to be normal. This was ascertained via examination of normal quantile plots for the residuals from each regression. A normal distribution was also fitted to the residuals, and a Shapiro-Wilk W goodness-of-fit test conducted. The normality of the residuals allowed the primary analysis to be conducted.

4.3 Los Angeles Lakers: Analysis

(The interpretation of the results of the analysis, so as to be given proper treatment, will be left until section 4.7. This is the case for the other three analyses presented as well.)

Initially, the ticket data points were regressed upon all of the team performance related covariates. This resulted in the full version of the linear model, including all possible categorical and continuous variable interactions. This model was then pared down through sequential elimination of most non-significant variables. In this process, the most non-significant variable is removed, and the model is refit using only the remaining variables. The most non-significant variable from that regression is then removed, and the model is refit using only the remaining variables. This process is continued until a model is reached where all coefficients on covariates are statistically significant. The final results are presented below.

Whole Model
Actual by Predicted Plot



Analysis of Variance

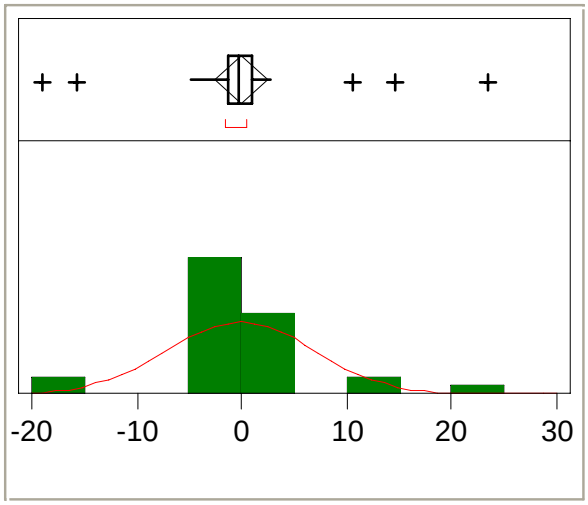
Source	DF	Sum of Squares	Mean Square	**F Ratio**
Model	5	37655.383	7531.08	**125.5897**
Error	26	1559.109	59.97	**Prob > F**
C. Total	31	39214.492		**<.0001**

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	534.12024	21.83774	24.46	<.0001
Win/Loss[L]	-31.09613	3.299066	-9.43	<.0001
Spread	0.7669848	0.359797	2.13	0.0426
Current Net Record	8.0383399	1.025859	7.84	<.0001
# Players Injured	-40.11256	3.327156	-12.06	<.0001
Final Score	-3.914449	0.189044	-20.71	<.0001

After this model was finalized, regression diagnostics were performed. The residuals, hats and Cook's D Influence values were saved and examined. Residuals were examined for normality, while the hats and Cook's D values were examined to determine if high leverage or high influence points existed within the data. The results of these examinations follow.

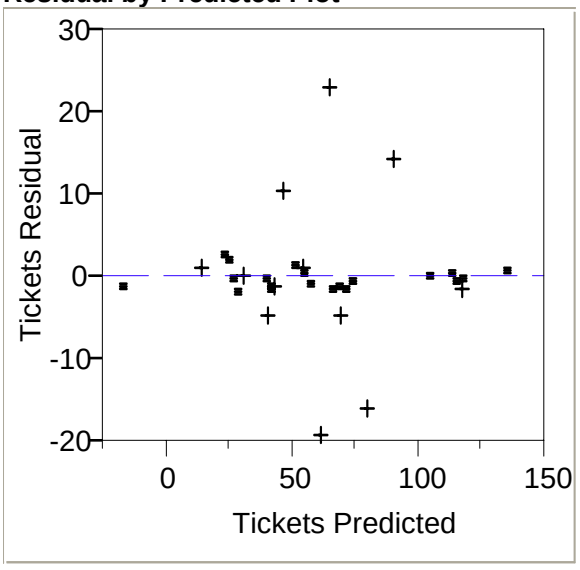
Residual Tickets



Fitted Normal
Parameter Estimates

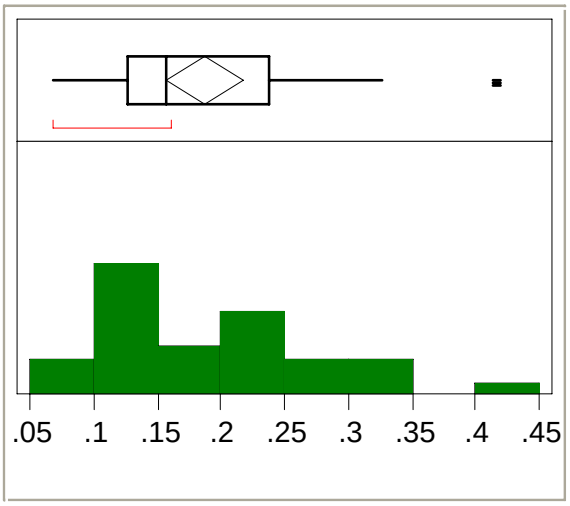
Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	**Mu**	**0.000000**	-2.55687	2.556872
Dispersion	Sigma	7.091814	5.68553	9.428422

Residual by Predicted Plot



As can be seen, the residuals showed no problems with regards to regression assumptions. The distribution and residual plot were both within the requirements.

h Tickets



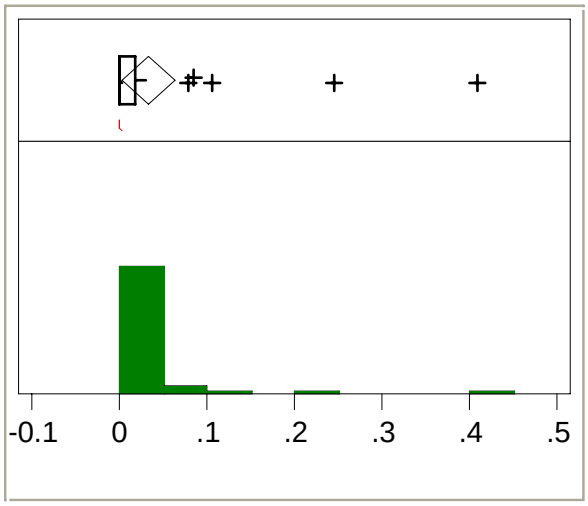
Quantiles

100.0%	maximum	0.41589
99.5%		0.41589
97.5%		0.41589
90.0%		0.31500
75.0%	quartile	0.23720
50.0%	median	0.15678
25.0%	quartile	0.12728
10.0%		0.10045
2.5%		0.06817
0.5%		0.06817
0.0%	minimum	0.06817

In this case, a point had high leverage if it had a hat value greater than or equal to 0.5625. As the largest hat value was 0.41580, it was concluded that there were no high leverage points.

The Cook's D values were examined next.

Cook's D Influence Tickets



Quantiles

100.0%	maximum	0.40640
99.5%		0.40640
97.5%		0.40640
90.0%		0.09826
75.0%	quartile	0.01789
50.0%	median	0.00113
25.0%	quartile	0.00013
10.0%		0.00002
2.5%		0.00000
0.5%		0.00000
0.0%	minimum	0.00000

In this case, a point had high influence if it had a Cook's D value greater than or equal to 1. As the largest value was 0.40640, it was concluded that there were no high leverage points.

Finally, the Durbin-Watson test was conducted to test for autocorrelation, as the data had an associated temporal sequence. The results, presented below, showed no significant signs of such an effect.

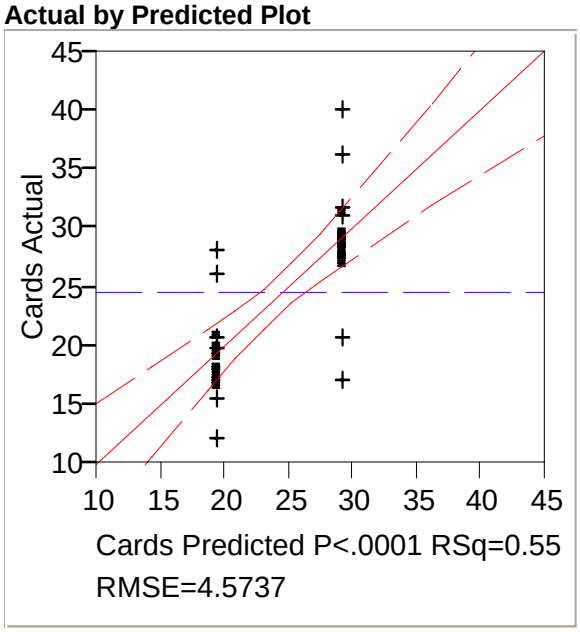
Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	**Prob<DW**
2.7741341	32	-0.3908	**0.9751**

4.4 Kobe Bryant: Analysis

To begin, the card data points were regressed on all of the player-related covariates. This produced the full linear model, including all possible interactions between categorical and continuous variables. The model was then pared down using sequential elimination of most non-significant variables. (For more detail on this process, refer back to section 4.3) The final results are shown below.

Whole Model
Actual by Predicted Plot



Analysis of Variance

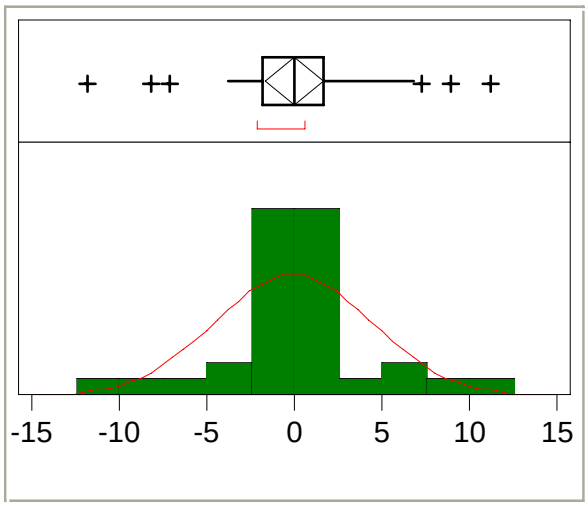
Source	DF	Sum of Squares	Mean Square	**F Ratio**
Model	1	769.5688	769.569	**36.7885**
Error	30	627.5624	20.919	**Prob > F**
C. Total	31	1397.1312		**<.0001**

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	24.213917	0.810108	29.89	<.0001
Player Performance [Above]	4.9135868	0.810108	6.07	<.0001

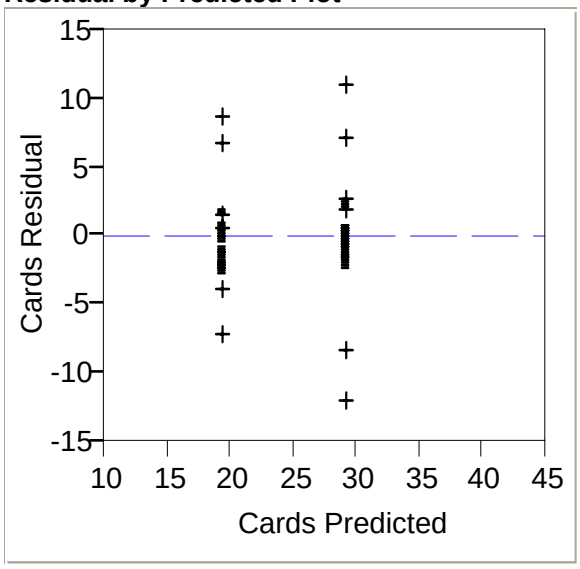
After this model was finalized, regression diagnostics were performed. The residuals, hats and Cook's D Influence values were saved and examined. Residuals were examined for normality, while the hats and Cook's D were examined to determine if there were points with high leverage or high influence. The results of these examinations are presented below.

Residual Cards



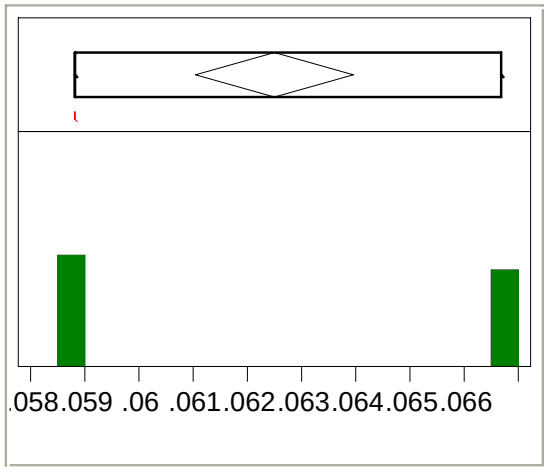
Normal(1.4e-15,4.49933)

Residual by Predicted Plot



As is shown, the residuals showed no problems with regards to regression assumptions. The distribution was within requirements. The residual plot appeared not to be within requirements; however, as this was a regression with only a categorical covariate, this was to be expected.

h Cards

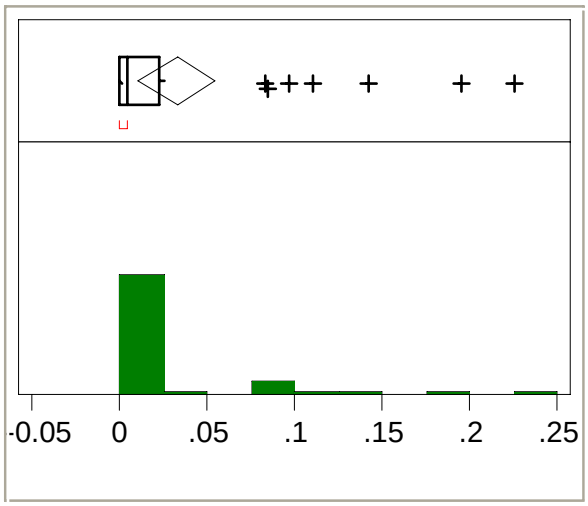


Quantiles

100.0%	maximum	0.06667
99.5%		0.06667
97.5%		0.06667
90.0%		0.06667
75.0%	quartile	0.06667
50.0%	median	0.05882
25.0%	quartile	0.05882
10.0%		0.05882
2.5%		0.05882
0.5%		0.05882
0.0%	minimum	0.05882

In this regression, a point had high leverage if it had a hat value greater than or equal to 0.1875. As the largest value was 0.06667, it was concluded that there were no points with high leverage.

Cook's D Influence Cards



Quantiles

100.0%	maximum	0.22545
99.5%		0.22545
97.5%		0.22545
90.0%		0.13210
75.0%	quartile	0.02293
50.0%	median	0.00466
25.0%	quartile	0.00069
10.0%		0.00011
2.5%		0.00000
0.5%		0.00000
0.0%	minimum	0.00000

In this analysis, a point had high influence if it had a Cook's D value

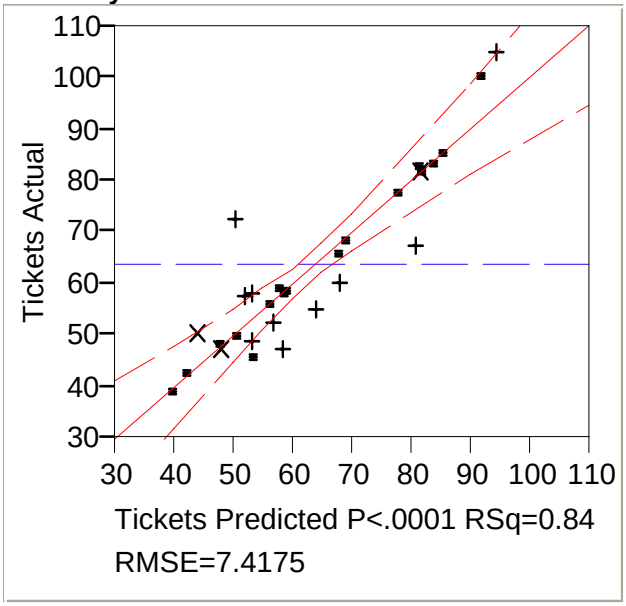
greater than or equal to 1. As the highest value was 0.22545, it was concluded that there were no points of high influence.

A Durbin-Watson test was not conducted in this case due to the categorical nature of the only predictor variable.

4.5 Philadelphia 76ers: Analysis

As for the Los Angeles Lakers, the ticket data points were regressed onto all of the covariates, producing the full linear model, including all possible interactions between categorical and continuous variables. Sequential elimination of most non-significant variables was utilized next to pare down this model to the final version. (For more detail on this process, please refer back to section 4.3) The results of this process are presented below.

Whole Model
Actual by Predicted Plot



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	**F Ratio**
Model	5	7081.3287	1416.27	**25.7412**
Error	24	1320.4672	55.02	**Prob > F**
C. Total	29	8401.7960		**<.0001**

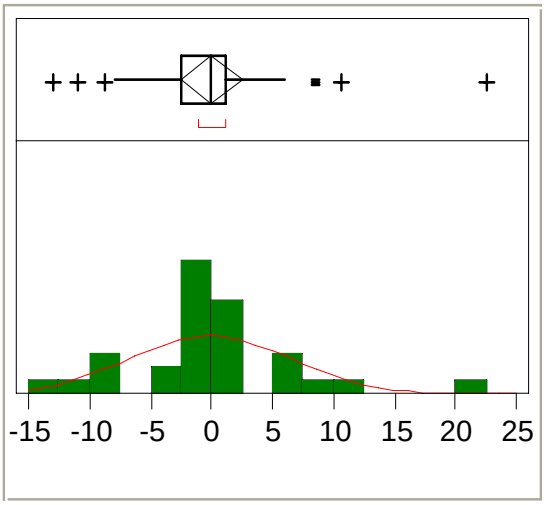
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	15.522953	17.19045	0.90	0.3755
Win/Loss[L]	36.078716	4.05347	8.90	<.0001
Streak	9.2255705	1.87054	4.93	<.0001
Spread	0.6862011	0.208932	3.28	0.0031
Current Net Record	1.8142903	0.865209	2.10	0.0467
Final Score	0.5210698	0.171414	3.04	0.0056

As with the previous two analyses presented, after the regression analysis was completed, regression diagnostics were conducted. The residuals, hats, and Cook's D Influence values were saved and examined. Residuals were again examined for normality, with Cook's D and hat values

being examined to determine the existence of high influence or high leverage points within the data. The results of these examinations follow.

Residual Tickets



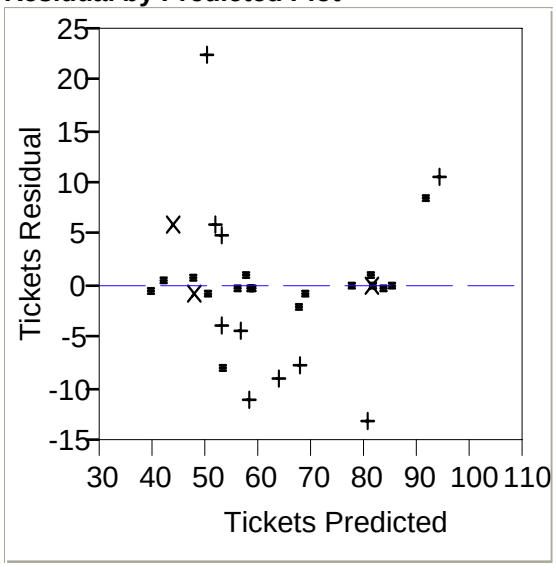
Normal(3.2e-14,6.74784)

Fitted Normal

Parameter Estimates

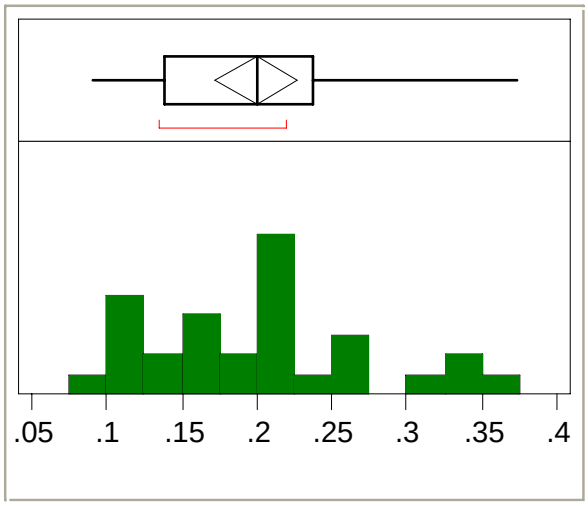
Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	**Mu**	**0.00000**	-2.51969	2.519685
Dispersion	Sigma	6.747841	5.37403	9.071225

Residual by Predicted Plot



As can be seen, the residuals showed no problems with regards to regression assumptions. Both the distribution and the residual plot were within requirements.

h Tickets

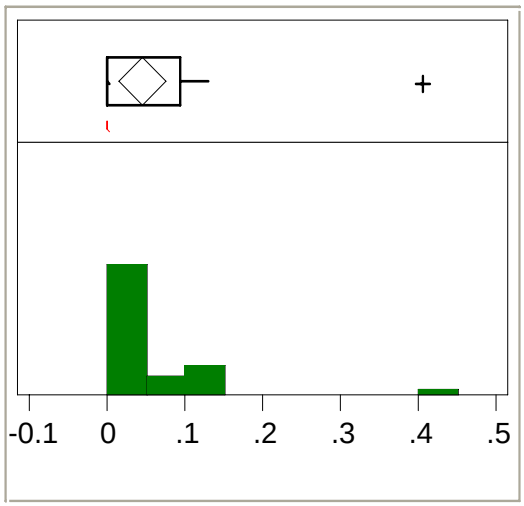


Quantiles

100.0%	maximum	0.37284
99.5%		0.37284
97.5%		0.37284
90.0%		0.32725
75.0%	quartile	0.23688
50.0%	median	0.20018
25.0%	quartile	0.13896
10.0%		0.11567
2.5%		0.09112
0.5%		0.09112
0.0%	minimum	0.09112

In this regression, a point had high leverage if it had a hat value greater than or equal to 0.6. As the highest hat value was 0.37284, it was concluded that there were no high leverage points in the data.

Cook's D Influence Tickets



Quantiles

100.0%	maximum	0.40410
99.5%		0.40410
97.5%		0.40410
90.0%		0.12272
75.0%	quartile	0.09513
50.0%	median	0.00095
25.0%	quartile	0.00008
10.0%		0.00001
2.5%		0.00000
0.5%		0.00000
0.0%	minimum	0.00000

A point had high influence in this case if it had a Cook's D value greater than or equal to 1. As the highest value was 0.40410, it was concluded that no points of high influence were present in the data.

Finally, a Durbin-Watson test was conducted to test for autocorrelation as the data had a related temporal sequence. The results, presented below, showed no significant signs of such an effect.

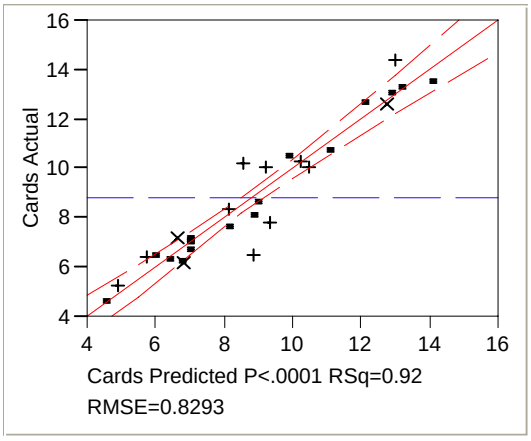
Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
2.3957521	30	-0.2103	0.8034

4.6 Allen Iverson: Analysis

Initially, the card data points were regressed on all of the player performance-related variables. This produced the full linear model, including all possible interactions between variables of a categorical or continuous nature. This model was then pared down utilizing the sequential elimination of most non-significant variables, a process explained in section 4.3. The final results follow.

Whole Model
Actual by Predicted Plot



Analysis of Variance

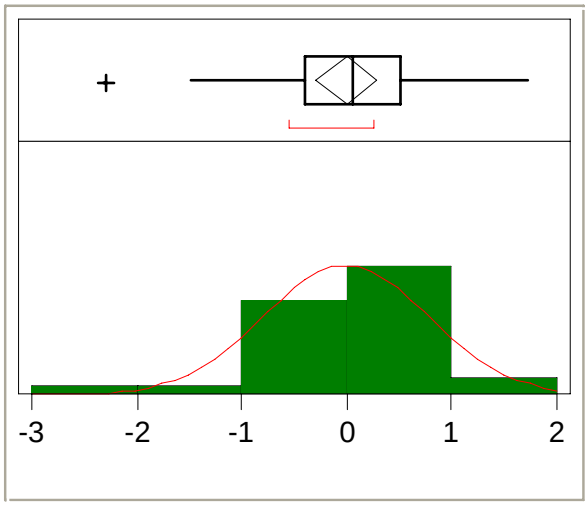
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	201.94945	67.3165	**97.8761**
Error	26	17.88209	0.6878	**Prob > F**
C. Total	29	219.83154		**<.0001**

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	3.9843295	0.491752	8.10	<.0001
Player Performance [Above]	-3.066261	0.214827	-14.27	<.0001
Player Points	0.6236971	0.047547	13.12	<.0001
Alpha	-43.64081	4.711154	-9.26	<.0001

After the model was finalized, regression diagnostics were performed. The residuals, hats, and Cook's D Influence values were saved and examined. While the hats and Cook's D values were being examined to determine if there were any points with high leverage or high influence, residuals were examined for normality. The findings of these examinations are shown below.

Residual Cards



Normal(-2e-15,0.78525)

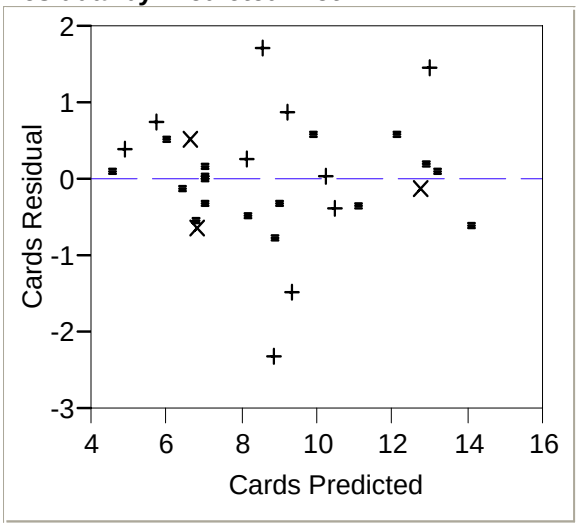
Quantiles

100.0%	maximum	1.723
99.5%		1.723
97.5%		1.723
90.0%		0.867
75.0%	quartile	0.504
50.0%	median	0.051
25.0%	quartile	-0.394
10.0%		-0.762
2.5%		-2.302
0.5%		-2.302
0.0%	minimum	-2.302

Fitted Normal Parameter Estimates

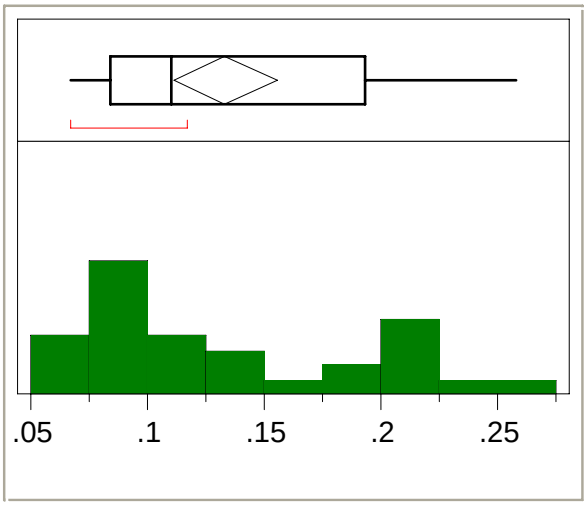
Type	Parameter	Estimate	Lower 95%	Upper 95%
Location	**Mu**	**-0.000000**	-0.293219	0.293219
Dispersion	Sigma	0.785254	0.625382	1.055629

Residual by Predicted Plot



As can been seen, the residuals showed no problems with regards to regression assumptions. Both the distribution and the residual plot are well within requirements.

h Cards

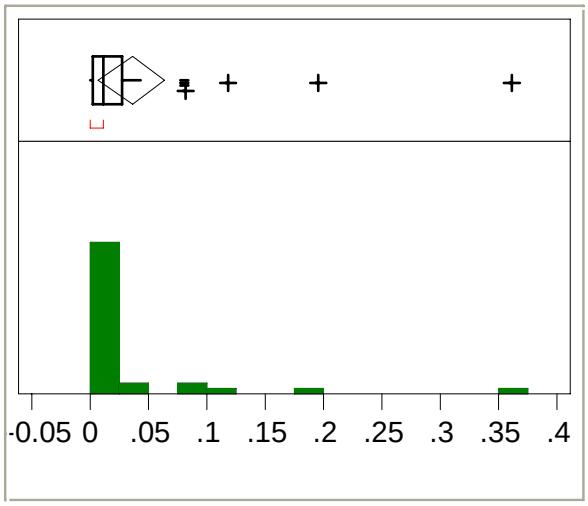


Quantiles

100.0%	maximum	0.25790
99.5%		0.25790
97.5%		0.25790
90.0%		0.22150
75.0%	quartile	0.19295
50.0%	median	0.11010
25.0%	quartile	0.08464
10.0%		0.06883
2.5%		0.06669
0.5%		0.06669
0.0%	minimum	0.06669

With regards to this regression, a point had high leverage if it had a hat value greater than or equal to 0.4. As the highest hat value was 0.25790, it was concluded that there were no points with high leverage.

Cook's D Influence Cards



Quantiles

100.0%	maximum	0.36010
99.5%		0.36010
97.5%		0.36010
90.0%		0.11391
75.0%	quartile	0.02635
50.0%	median	0.01039
25.0%	quartile	0.00261
10.0%		0.00038
2.5%		0.00002
0.5%		0.00002
0.0%	minimum	0.00002

Here, a point had high influence if it had a Cook's D value greater than or equal to 1. As the highest value was 0.36010, it was concluded that there were no high influence points.

Finally, a Durbin-Watson test was conducted to test for autocorrelation due to the temporal sequence related to the data. The results of this test, shown below, showed no signs of this effect.

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.9766419	30	0.0041	0.4213

4.7 Interpretation of Results

The results of each of the regression analyses conducted showed that variables of an "irrational" nature have explanatory power with regards to price or price changes.

Regarding the Los Angeles Lakers, the regression analysis performed showed that Win/Loss, Spread, Current Net Record, # Players Injured and Final Score all had an effect on the price of game tickets. The respective coefficients generally went in the expected direction. Losing a game or having many injured players led to a drop in price, while having a large positive spread or a favorable current net record led to an increase in price. The coefficient on Final Score however did not go in the direction expected. A higher final score was shown to lead a drop in price. There may have been a correlation between one or more of the other variables and Final Score that was not able to be seen in the initial data exploration.

With regards to Kobe Bryant, the analysis completed showed that Player Performance had an effect on the price of his trading cards. The coefficient on this variable went in the direction expected; an above average performance led to an increase in price, while a below average performance led to a decrease in price.

For the Philadelphia 76ers, the regression conducted showed that Win/Loss, Streak, Spread, Current Net Record and Final Score all had an effect on the price of tickets to their games. The associated betas generally went in the expected direction. Winning games consecutively, having a positive spread, having a favorable net record, and scoring more points in a game all led to an increase in price. The coefficient on

Win/Loss was a bit counterintuitive however. Losing a game was shown to lead to an increase in price. It is possible that there was a correlation between this variable and one or more of the other covariates that was not able to be detected in the preliminary exploration of the data.

Finally, for Allen Iverson, the regression analysis performed demonstrated that Player Performance, Player Points and Alpha all had an effect on the price of his trading cards. However, in this case, only one of the coefficients went in the expected direction, namely that associated with Player Points. Scoring more points led to an increase in price. However, the other coefficients were counterintuitive. Above average performance and a higher alpha were shown to lead to a decrease in price. As before, there may have been some correlation that was not able to be detected via the exploratory data analysis.

5. Discussion

Numerous studies have proven projection bias to be a cognitive bias in the area of decision-making. Due to its effects on both small and large decisions, studying these influences is of paramount importance.

This paper attempted to analyze the influences of projection bias on decision making in the arena of eBay auctions. A number of “irrational” variables were shown to have an effect on prices, contrary to the decision making model put forth by economists. However, due to the imputation of certain data points and the small size of the datasets, these results cannot be taken as definitive. Further work on this topic is a necessary task for future economists, psychologists, and statisticians.

References:

Roth, Alvin E and Ockenfels, Axel. "Last Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet" The American Economic Review September (2002): 1093-1102

Nelson, Leif D. and Morrison, Evan L. "The Symptoms of Resource Scarcity: Judgments of Food and Finances Influence Preferences for Potential Partners" Psychological Science 16-2 (2005): 167-173

Loewenstein, George; O'Donoghue, Ted and Rabin, Matthew. "Projection Bias in Predicting Future Utility" The Quarterly Journal of Economics November (2003): 1209-1245

Gilbert, Daniel T.; Gill, Michael J. and Wilson, Timothy D. "The Future Is Now: Temporal Correction in Affective Forecasting" Organizational Behavior and Human Decision Processes 88-1 (2002): 430-442

Loewenstein, George. "Out of Control: Visceral Influences on Behavior" Organization Behavior and Human Decision Processes 65-3 (1996): 272-289

Bajari, Patrick, and Hortascu, Ali. "Economic Insights from Internet Auctions" Journal of Economic Literature 42 (June 2004): 457-484

Conlin, Michael; O'Donoghue, Ted and Vogelsang, Timothy J. "Projection Bias in Catalog Orders" Syracuse University, Cornell University (2005): 1-28

Ariely, Dan and Simonson, Itamar. "Buying, Bidding, Playing or Competing? Value Assessment and Decision Dynamics in Online Auctions" 1-19