Predicting MLB Attendances and Prices

Rufus Petrie
Professors Magnolfi & Sullivan
Machine Learning
December 13, 2018

Abstract

In this paper, we fit an array of statistical models to predict attendance and ticket prices for Major League Baseball (MLB) games. We use data from Andrew Sweeting's 2012 paper that studies Dynamic Pricing in the secondary market for baseball tickets. Due to the highly nonlinear and discontinuous nature of MLB attendance and prices, we find that Random Forests provide the highest predictive accuracy for these tasks. For our favorite team, the Boston Red Sox, we find that sellers on the secondary market should focus on factors like row number, strength of the away team, whether the Red Sox play the Yankees, and the amount of listings on similar websites to select tickets and maximize profits.

1 Introduction

In his paper, *Dynamic Pricing Behavior in Perishable Goods Markets*, Andrew Sweeting studies the pricing dynamics of MLB tickets on secondary markets like Stubhub and Ebay. Because of the high variance in demand between games, tickets auctioned on these platforms often sell for much higher than their original prices. Also, because MLB tickets have no value after their associated game passes, sellers in this market frequently adjust their listing prices as the game date approaches to ensure that their tickets sell. For his analysis, he estimates game attendance ratios for each game by using a variety of factors like the teams playing, day of the week, team standings, etc. He then estimates the resell price ratio on the secondary market while controlling for factors like the teams playing, predicted attendance, selling location of the ticket, time of the ticket sale prior to the game, etc. He concludes that players in the secondary market for MLB tickets employ dynamic pricing strategies and reduce the prices of their listings dramatically as the gameday approaches; additionally, he claims that this behavior leads to about 16% higher profits for the ticket sellers [1].

Aside from academic purposes, information on the evolution of ticket prices could also be useful for participants in the secondary market for MLB tickets. Where Sweeting estimates expected resell prices to evaluate how market participants behave, market participants could perform the same kind of analysis to uncover better trading strategies. For example, market participants could use this type of analysis to find what teams, stadium sections, game dates, etc. have the highest prices relative to their original listings. In other words, market participants would benefit from information about which tickets to trade as well as how to adjust their prices once they obtain those tickets. However, Sweeting's empirical methods likely would not provide optimal results for these purposes.

While his models may produce correct estimates on average, market participants would likely accept some bias in their models in exchange for having closer estimates to a ticket's profitability. Therefore, we use a variety of machine learning techniques and algorithms to analyze which models provide the best predictors of ticket resale prices and what factors drive profitability. We find that random forests provide the best accuracy for predicting game attendances and prices. Additionally, we find that the timing of a ticket listing provides relatively little predictive power when compared to other information.

2 Data

Table 2.1: Variables of Interest

Variable	N	μ	σ	*Min*	*Max*
Percent Attendance	4.23×10^5	0.654	0.242	0.052	1
Resell Ratio (Panel)	2.18×10^6	1.983	1.818	0.008	150.960
Resell Ratio (Cross Section)	3.29×10^6	1.979	1.427	0.008	127.500

For this paper, we use three of the datasets from Andrew Sweeting's paper on Dynamic pricing. To predict MLB game attendances, we use a dataset describing the record of MLB attendances and the characteristics of those games spanning the years 2000 to 2007. For our independent variable, we use the proportion of maximum attendance for each game[1]. For our model covariates, we use variables like home/away team, day of week, month, year, and dummies for interleague and interdivision games. Furthermore, we also include an array of variables that describe the performance of the home and away team[2]. From Table 2.1, we see that the average MLB game has an attendance of 65.4%, and this number has significant variation. As we will see later in the paper, much of the

[1] We chose to estimate this instead of attendance because the proportion takes values in the unit interval, and this saves us some trouble when working with a few algorithms.

[2] These variables mainly include information on each team's standing in their division and wildcard race, i.e. strong indicators of likelihood to make the playoffs.

variation in our variables of interest occurs at the team level. Figure 7.1 in the appendix breaks down average attendance ratio by home team. At the extremes, we see that the Montreal Expos had an average attendance of less than 30%[3], and the Boston Red Sox and San Francisco Giants both had average attendances of over 90%.

Additionally, we use a panel and a cross section of information on Stubhub auctions to predict resell prices for MLB tickets. The panel data contains the same information as the cross section, but it contains data from sellers who have sold multiple tickets on Stubhub. Because we don't use any of the seller characteristics from the panel to make our predictions, we use the same covariates for each pricing model. For our independent variable, we use resell ratio, i.e. the ratio of a ticket's listing price divided by its original price. For our model covariates, we use information about the seats, the number of tickets being sold, the amount of time before the game each listing is made, information about competing listings, and information on the performance of each team[4].

From Table 2.1 we see that the average MLB ticket resells for about 1.98 times its original price in both datasets. However, resell ratios in the panel data exhibit higher variance than the ones in the cross section. This may happen because repeat sellers in the panel data use more sophisticated strategies than one-off sellers and vary their price more over time. At the extremes, we see that some tickets sold for basically nothing, and one sale in the panel data went for almost 151 times the original price. As is the case with game attendance, much of the variation in resell ratios occurs at the team level. Figures 7.2 and 7.3 in the appendix show the average resell ratio by home team in the panel and cross section. The figures look similar with a few teams hanging around the 1.4 mark[5],

[3] This poor performance probably contributed to the team's move in 2004, after which point they left Montreal and became the Washington Nationals.

[4] We use the same performance variables that appear in the attendance models.

[5] Stubhub charges 25% commission, so resell ratios much lower than this would not net any profit.

the majority of teams hovering in the 1.5-2 range, and around six teams averaging a resell ratio of over 2. Figure 7.4 in the appendix shows the relationship between average resell ratios and average proportion of attendance. From this graph, we can see a positive correlation between average resell ratios and attendances, but it also appears that a few teams with very high prices help to drive this relationship[6].

3 Methodology

Unlike Sweeting's models, our models attempt to maximize predictive accuracy. Therefore, instead of worrying about the asymptotic properties of our estimators, we judge our models based on their predicted out of sample performance. To do this, we split the data into training sets for selecting models and test sets for evaluating model accuracy. For every test/train split, we allocate 80% of the data to the training set and 20% of the data to the test set. To generate these splits, we specify seeds in R and Python so that we never contaminate the test data. We then train our models using only data from the training sets and specify model characteristics that lead to the best 10-fold cross validated mean squared error in the training set[7]. Finally, we evaluate the performance of our models on the test data so that we can see how they would perform on future data.

As is the case in Sweeting's paper, we also slightly limit our usage of the data. While we use the full dataset for predicting game attendance, we only study tickets in the bottom 98% of resell ratios. According to Sweeting, this modification greatly improved the performance of his models, so we followed suit. Furthermore, this had the advantage of

[6] The Boston Red Sox dominate every other team in our variables of interest. In particular, they have an average resell ratio of around 3.3 in the panel data, almost 0.9 points higher than any other team. This likely owes to the fact that they won two championships during the span of this data.

[7] Because of the long runtimes of our algorithms, we started by training models on smaller subsections of the training set to get a vague feel for how our final parameters should look.

providing a reasonable maximum value for scaling the resell ratio variable. We found that our neural networks had the best performance when we use a min-max scaled dependent variable, and omitting observations with extremely high resell ratios made the rescaled data more coherent.

3.1 Predicting Attendance

To predict game attendances, we estimate the proportion of maximum attendance for each game by using all of the relevant covariates available in the data. In particular, these covariates included information on the teams playing, the month and day of week of the game, and the relative standings of each team. We started by estimating a linear regression including all pertinent covariates to get a baseline for model performance. Then, we estimated a linear regression using the backwards stepwise selection algorithm. We chose the best model based off of adjusted R-squared. Because this model had worse performance than the full model, we concluded that we should use more complicated specifications.

Because the simple models clearly suffered from underfitting, we wanted to perform a highly interacted regression with LASSO selection. However, we could not find an efficient enough implementation to run given the relatively large size of the data. Therefore, we decided to model random forests, as they have the ability to model highly complicated interactions and have efficient implementations in most languages. To optimize the performance of our random forests, we mainly focused on adjusting the maximum amount of covariates per sample and the number of trees. Our final model specification had a maximum of 100 covariates and grew 64 trees.

From the random forest, we moved on to neural networks, another model capable of describing complex interactions within the data. To estimate the neural networks, we had to perform a few more modifications on the data. In particular, we rescaled each of the predictors to be a standard normal random variable. Because the proportion of maximum attendance already exists in the unit interval, we didn't have to scale the dependent variable. For this same reason, we used logistic activation functions in our neural network because they also map output to the unit interval[8]. Because adjusting the learning rates and activation functions of the hidden layers led to some bizarre results, we avoided changing them too much. Therefore, we mainly focused on adjusting the number of hidden layers and number of neurons per hidden layer. Our final model used two hidden layers with 120 and 60 nodes, respectively.

3.2 Predicting Prices

To predict resell prices, we once again started by estimating linear models to get a baseline idea of how models should perform on the data. As was the case with the attendance models, these severely underfit the data, so we continued to tree-based models. In particular, we fit random forests and optimized their performance by adjusting the number of variables used in each tree and the number of trees in the forest. The final random forest for the panel data used 128 trees and 100 maximum covariates, and the final random forest for the cross section used 128 trees and 110 maximum covariates.

To estimate neural networks for resell prices, we once again scaled the model covariates to have standard normal distributions. Additionally, we found that min-max scaling the resell ratio gave us the best results in terms of predictive accuracy. Because we excluded observations in the top 2% of resell ratios, this provided a better spread of values for

[8] We also tried max-min scaling our variables and using identity activation functions, but this led to significantly worse performance.

the scaled resell ratios. Furthermore, because the scaled resell ratios all exist in the unit interval, we used logistic activation functions for our hidden layers. The final neural network for the panel had two hidden layers with 120 and 60 nodes, and the final neural network for the cross section had two hidden layers with 130 and 65 nodes.

4 Results

4.1 Attendance Models

Table 4.1: Attendance Model Results

Model	Test MSE	Test R-squared
Regression	0.0218	0.6294
Neural Network	0.0026	0.9557
Random Forest	0.0009	0.9854

Table 4.1 contains the results of our attendance models. For predicting attendance, we find that random forests provide the best test MSE, followed by neural networks and then linear models. The final random forest model achieved a test MSE of 0.0009 and explained over 98% of the variation in attendance ratios in the test set. Although the random forest does not have any inferential properties or a concrete equation to interpret, we can still look at the variable importances for the model. These importances correspond to the degree to which each variable drove reductions in MSE across the trees in the random forest. Figure 7.1 in the appendix contains the top ten variable importances for our final model. For the final random forest, home record, whether a game is on the weekend, whether a game is a home opener, and whether a game was in the year 2000 all ranked highly in variable importance. Other than these, an array of home team dummies round out the top ten.

4.2 Pricing Models

Table 4.2: Pricing Model Results

Model	Test MSE	Test R-squared
Panel Data		
Regression	0.9092	0.2804
Neural Network	0.6360	0.5001
Random Forest	0.2835	0.7757
Cross Sectional Data		
Regression	0.6725	0.3238
Neural Network	0.3746	0.6243
Random Forest	0.1068	0.8926

Table 4.2 contains the results of the price models for both the panel and the cross sectional datasets. For predicting prices, we once again find that random forests provided the best test MSE, followed by neural networks and then linear models. The final random forest for the panel data achieved a test MSE of 0.2835 and explained over 77% of the variation in the test set. The final random forest for the cross section achieved a test MSE of 0.1068 and explained over 89% of the variation in the test set. The fact that our models performed better on the cross section should not surprise us, as we observed earlier that listings in both data sets have approximately the same mean, but listings in the panel have significantly higher variance. Table 7.2 in the appendix contains the top 10 variable importances for the panel and cross section models. As we can see, the attendance ratio, row number, and a couple of team dummies both provide large MSE reductions for both datasets. However, it appears that the variable importances have less concentration at the top, and information on other listings plays a larger role for the panel data. This may indicate that regular sellers both utilize a wider amount of information and focus more on demand from other sellers to set their prices.

4.3 Strategy for Trading Red Sox Tickets

Table 4.3: Red Sox Variables of Interest

Variable	N	μ	σ	*Min*	*Max*
Percent Attendance	2.79×10^4	0.834	0.207	0.193	1
Resell Ratio (Panel)	2.42×10^5	3.041	2.340	0.008	52.345
Resell Ratio (Cross Section)	3.69×10^5	2.985	1.830	0.008	28.333

To devise a trading strategy for tickets for our favorite team, the Boston Red Sox, we restricted all three datasets to Red Sox home and away games. Table 4.3 contains summary information for our variables of interest. As we can see, the average resell ratio rises by about 65% for both datasets when we keep only Red Sox games. Additionally, the standard deviation rises by about 28% for both datasets. Therefore, we can reasonably expect that Red Sox tickets have significantly different pricing behavior when compared to other tickets.

Table 4.4: Red Sox Random Forest Results

Data	Test MSE	Test R-squared
Panel	0.6009	0.6206
Cross Sectional	0.1087	0.9013

To predict resell ratios for Red Sox tickets, we fit the same models as we did for the general data. Once again, we found that random forests provide the best out of sample predictive accuracy. For both the panel data and cross section, we fit a random forest with 64 trees and a maximum of 100 features per tree. Table 4.4 contains the results of these models. As we can see, keeping only information on the Red Sox slightly improved the out of sample performance of the model in the cross sectional data, but the panel data model had significantly worse performance out of sample. Once again, this fits with the idea that regular sellers on auction websites have significantly different pricing behavior than casual sellers.

Table 7.3 in the appendix contains the top 10 variable importances for both of the Red Sox random forests. From this table, we see that row number, attendance ratio, and the dummy "Away = Yankees" round out the top three for both models. This serves as a good sanity check because any casual baseball fan would likely name Red Sox/Yankees at Fenway if you asked them about expensive tickets. Beyond this, we see similar behavior to the variable importances for the entire dataset. In particular, team records rank higher in the cross section, and information on similar listings rank higher in the panel. Additionally, variable importances decline more slowly in the panel than they do in the cross section. Therefore, people in the secondary market for Red Sox tickets should focus on factors like row number, the strength of the opposing team, whether the Red Sox play the Yankees, and likely competition on auction websites to determine which tickets to sell. Although our models don't suggest that timing of a ticket's sale has a significant impact on its price, sellers should still probably try to sell their tickets as soon as possible.

4.4 General Observations

From the performance of these models, we can conclude that MLB game attendance and resell prices exhibit highly nonlinear behavior. Using neural networks provided a significant increase in predictive accuracy over the linear models, and the random forests provided another large increase in performance over the neural networks. This makes sense because random forests allow for both complicated interactions between variables and discontinuous predictions, whereas neural networks provide slightly smoother predictions. Therefore, future studies of MLB attendance and pricing behavior could probably benefit from evaluating how other highly flexible models perform on this type of data.

Additionally, we did not find any evidence that the timing of a ticket's listing makes a significant impact on its price. For all of our random forest models, none of the time dummies ever had particularly high variable importance. This goes against the main thrust of Andrew Sweeting's paper, i.e. that the timing of a ticket's listing has a significant relationship with its price. We likely have different results because in his paper, he uses fixed effects regressions to observe variations within sellers, whereas we build our models using variation between listings. Therefore, sellers in the secondary market can probably ignore the timing of a ticket's sale when predicting resell prices.

5 Caveats

While interpreting the results of these models, we should bear a few caveats in mind. Firstly, the models we produced don't necessarily produce the best possible out of sample MSE and R-squared as they could. Because of the size of the data and time constraints, we could not take every step possible to optimize the models. However, we have confidence that we have correctly ranked the performance of the models. In the latter stages of our testing, improvements in the neural network led to very small improvements in performance, and we doubt that it would surpass the random forest if we spent more time optimizing them. Furthermore, we would have liked to assess the performance of a highly interacted regression with regularization, but we could not find an efficient enough implementation. While we don't believe that this type of model would outperform the neural network and random forests, it would still give a more honest assessment of linear models. Finally, we also wanted to try some type of clustering algorithm, but this likely would have entailed creating a custom norm for this problem, so we didn't get around to it.

Furthermore, we should note limitations of interpreting random forest variable importances. For an example of this, we can look at Table 7.1 and see that the "Home = Expos" and "Home = Giants" variables have very similar values in terms of importance. However, as we can see from Figure 7.1, the Expos had the worst average attendance, and the Giants had the second best average attendance. Therefore, even though two variables have a similar contribution in reducing the model MSE, they can have vastly different relationships with the dependent variable. This means that while a random forest may make good suggestions about what variables to include in models, it ultimately tells us very little about the role that those variables play in explaining the data. Additionally, we observed fairly similar variable importances when we made small parameter changes in our models, but further optimizations to the models may produce different variable importance rankings.

Finally, we should bear in mind the difference between our models' results for the panel and cross sectional data. Previously, we noted that our models may perform worse on the panel data because of the higher variance in this data. However, the fact that the cross sectional data has roughly 50% more observations also probably helps to drive this discrepancy. Therefore, while we have strong reason to believe that frequent sellers behave differently, we can't precisely say how much of this difference stems from different trading strategies.

6 Conclusion

In conclusion, we find that for predicting MLB game attendance, prices, and prices for Red Sox tickets, random forests provide the highest level of predictive accuracy. We found that factors like home record, whether a game is on the weekend, whether the game is a home opener, and the particular home team had the greatest impact on attendance. We found that factors like the row number, home record, away record, and competition on auction websites had the greatest impact on resell price. For participants in the secondary market for Red Sox tickets, we find that they should focus on factors like row number, team records, whether the Red Sox play the Yankees at Fenway, and competition on other sites to maximize their profits. In general, we found that the precise timing of a ticket listing offers very little predictive power when considering prices, and casual users and frequent sellers on auction websites have significantly different behavior.

7 Appendix

Table 7.1: Attendance Random Forest Variable Importances

Variable	Importance
Home Record	0.0793
Home = Expos	0.0495
Home = Giants	0.0437
Saturday Game	0.0402
Home = Marlins	0.0394
Home = Red Sox	0.0372
Year = 2000	0.0301
Home Opener	0.0297
Home = Cubs	0.0290
Home = Rays	0.0262

Table 7.2: Pricing Random Forest Variable Importances

Panel Data		*Cross Sectional Data*	
Variable	Importance	Variable	Importance
Attendance Ratio	0.1429	Attendance Ratio	0.1561
Row Number	0.1047	Row Number	0.1335
Section Listings Sq.	0.0424	Away Game Record	0.0493
Section Listings	0.0424	Home Game Record	0.0488
Away = Yankees	0.0373	Away = Yankees	0.0473
Home Game Record	0.0330	Home = Cubs	0.0356
Similar Listings	0.0309	Section Listings Sq.	0.0350
Similar Listings Sq.	0.0308	Away = Boston	0.0347
Away Game Record	0.0252	Section Listings	0.0342
Home = Cubs	0.0209	Similar Listings	0.0272

Table 7.3: Red Sox Pricing Random Forest Variable Importances

Panel Data		*Cross Sectional Data*	
Variable	Importance	Variable	Importance
Away = Yankees	0.1047	Row Number	0.1464
Row Number	0.0999	Away = Yankees	0.1109
Attendance Ratio	0.0700	Attendance Ratio	0.0834
Section Listings	0.0593	Away Game Record	0.0720
Section Listings Sq.	0.0580	Section Listings	0.0605
Similar Listings Sq.	0.0381	Home Game Record	0.0589
Similar Listings	0.0367	Section Listings Sq.	0.0514
EB Section Listings	0.0340	Similar Listings	0.0313
EB Section Listings Sq.	0.0339	Similar Listings Sq.	0.0295
Home Game Record	0.0331	Home = Yankees	0.0285

References

[1] Andrew Sweeting. Dynamic pricing behavior in perishable goods markets: Evidence from secondary markets for major league baseball tickets. *Journal of Political Economy*, 120(6):1133–1172, 2012.

Figure 7.1: Average Proportion of Maximum Attendance by Home Team

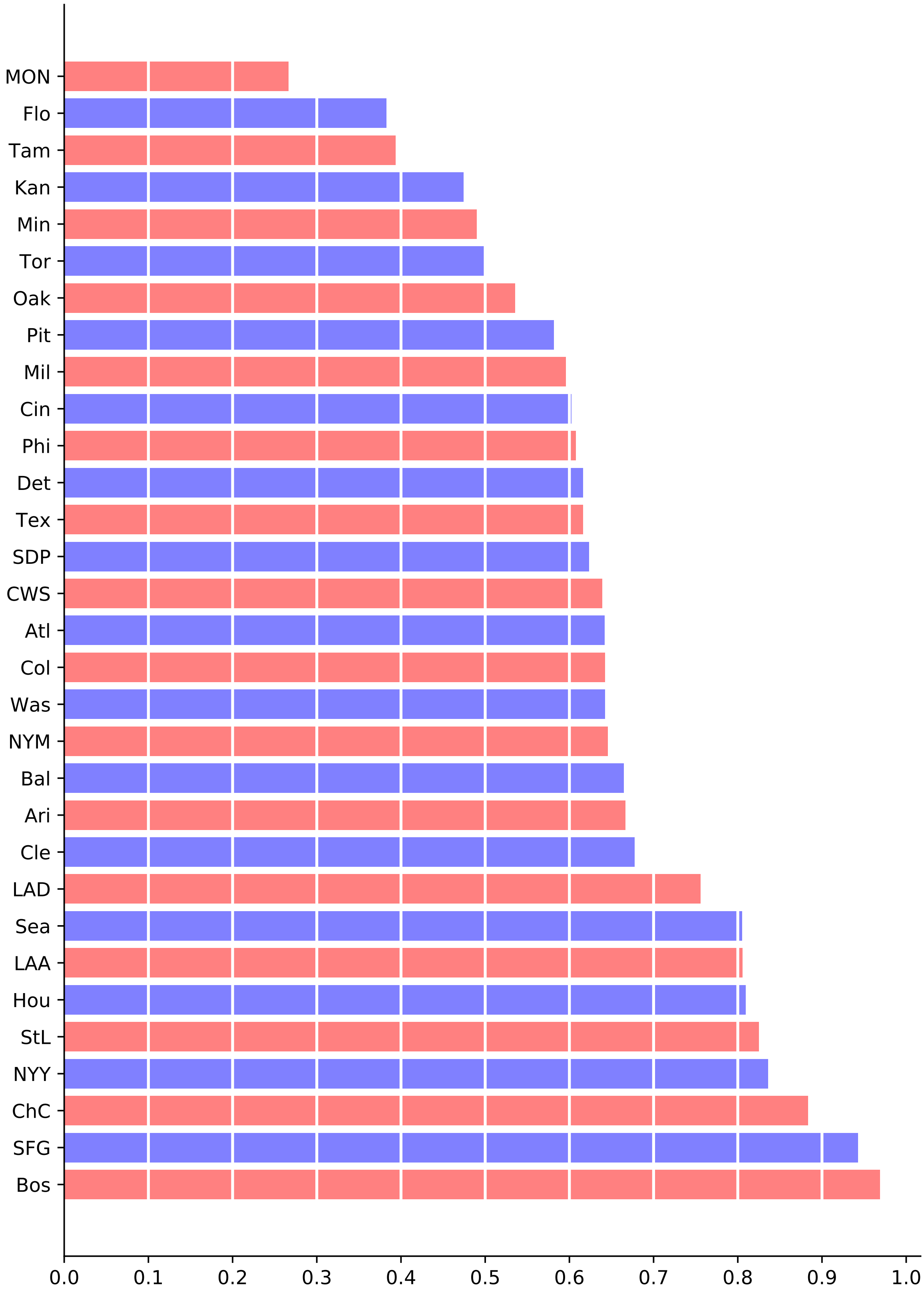


Figure 7.2: Average Resell Ratio by Home Team (Panel)

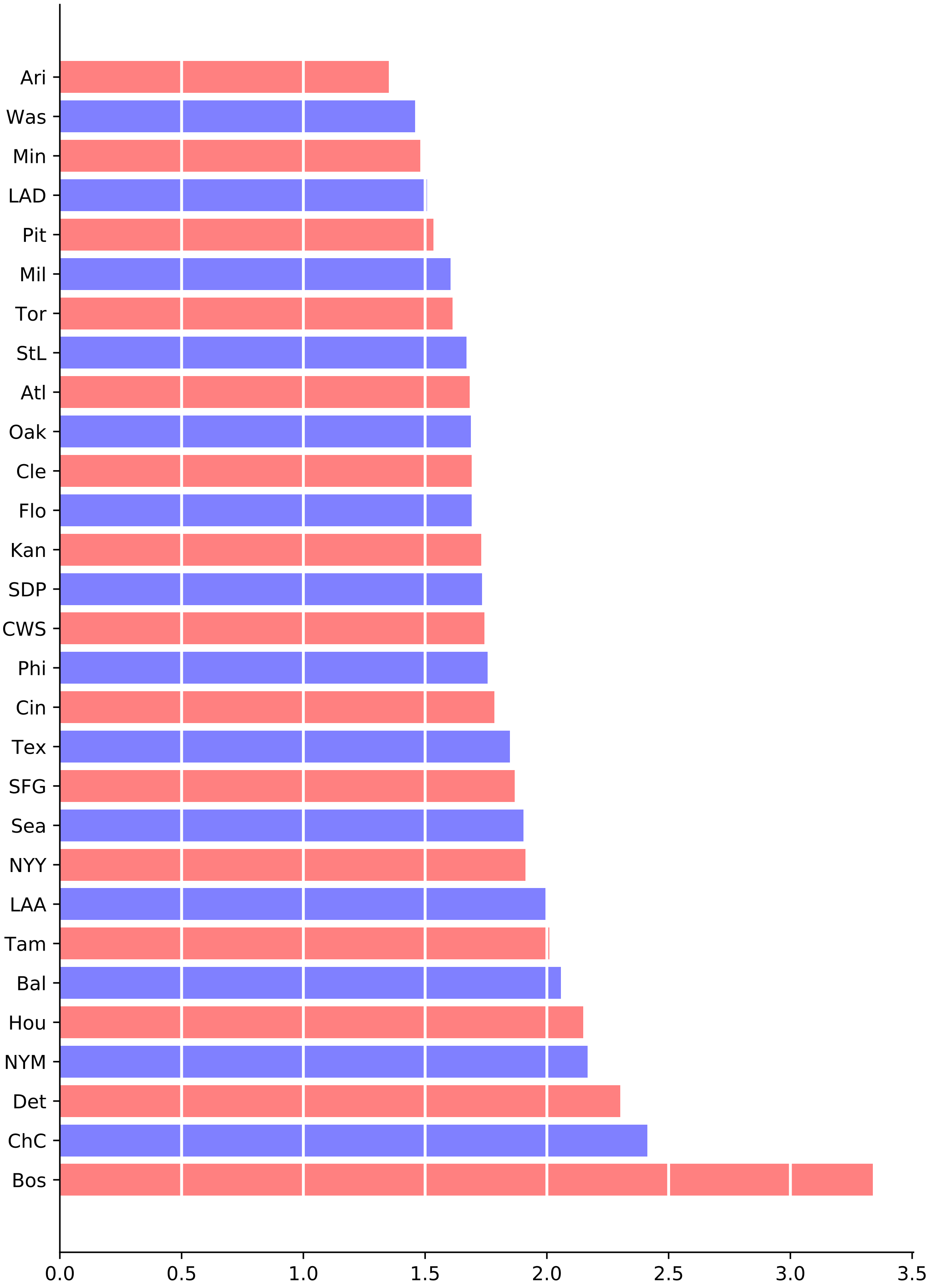


Figure 7.3: Average Resell Ratio by Home Team (Cross Section)

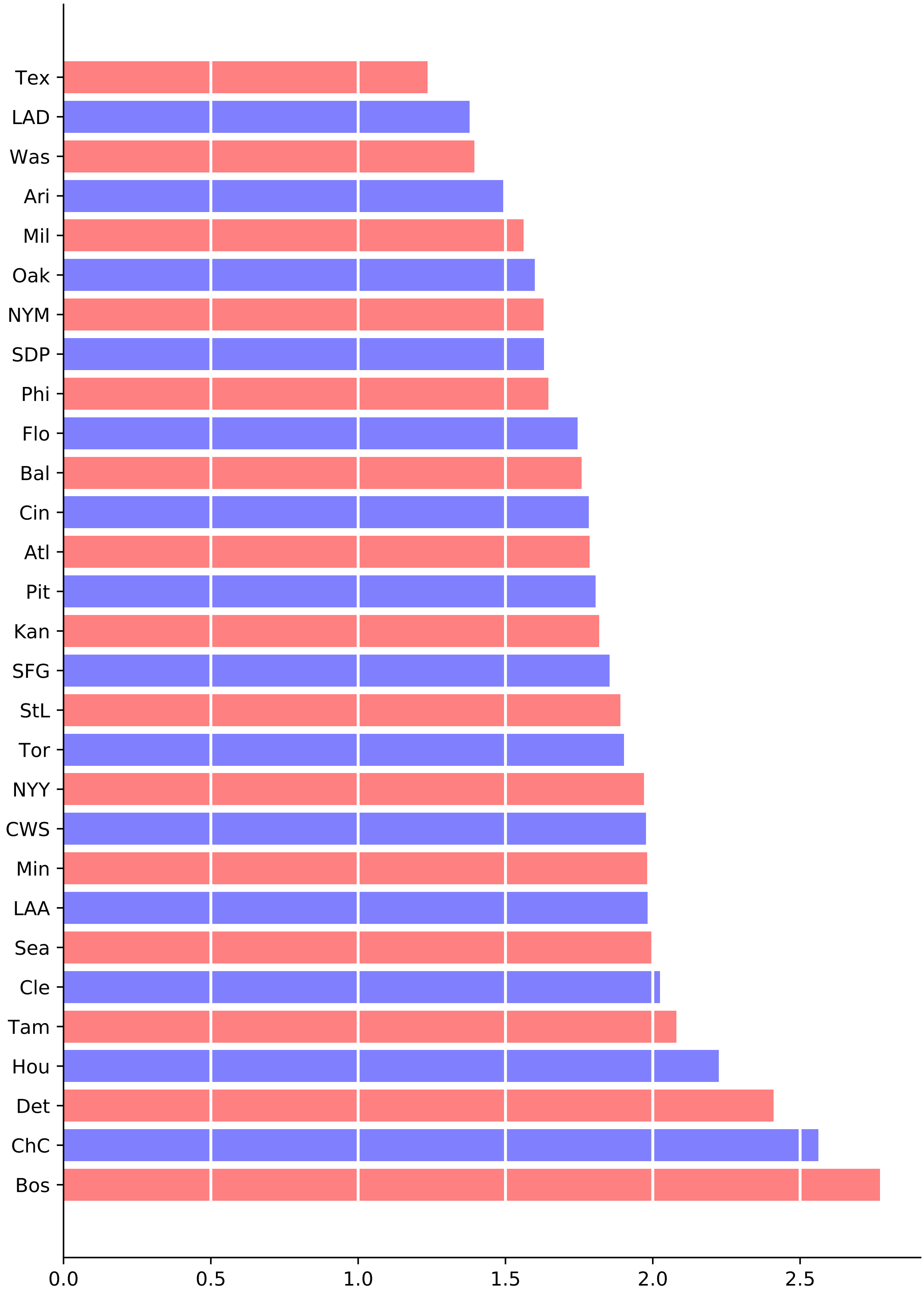
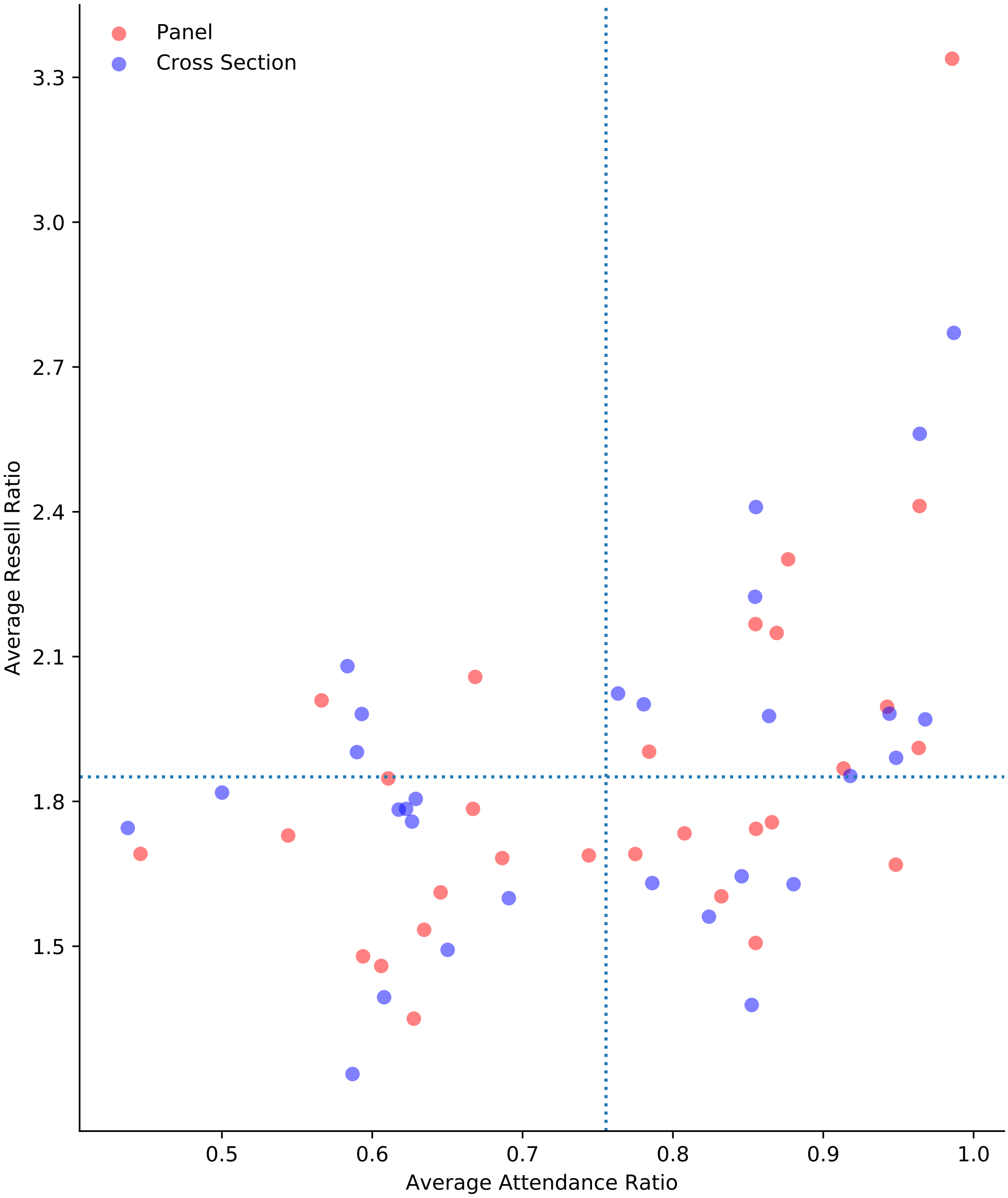


Figure 7.4: Resell Ratio vs. Attendance by Home Team


● Panel
● Cross Section
Average Resell Ratio
Average Attendance Ratio