STAT 306 Group Project:

Modelling Prize Returns to Tennis Rank

Group Members:

Soren Rajani

Yulong Peng

Sicily Xie

Date: April 2022

Introduction

In this project, we seek to understand how tennis prize winnings are affected by player rank. Our study uses observational data with the goal of explaining the nature of the relationship. Our hypothesis is that prize money has a tendency to be winner-take-all implying the relationship is exponentially increasing, however, it could be that all professional players see more linear returns to rank or even decreasing returns as there is a limited number of tournaments players can reasonably attend.

Our data[1] comes from Ultimate Tennis Statistics, an organization that records men's tennis results, rankings, prize payouts, and a host of other variables for fans taken from published tournament results. The data is cross-sectional data from 2019 and includes observations of the top 500 players in 2019. In our analysis, we utilize one response variable and up to four predictor variables. Our response variable is *prize money*, which is total lifetime earnings solely from prizes measured in USD and is continuous.

Our predictor variables are *current rank, best rank, age, and seasons played. Current rank,* and *best rank* are ordinal variables and we will treat them as continuous for this study while *age* and *seasons played* are continuous. The current rank is measured as the player rank in 2019 listed by the Association Tennis Professionals (hereafter ATP) which is how most sports organizations such as ESPN choose to rank players. The best rank is measured as the lifetime highest ATP ranking that a player had achieved by 2019. Age is measured as the age of the player in 2019. Seasons played is the total number of seasons a player had professionally competed in by 2019.

In addition to testing the relationship between best rank and prize money, we have two secondary objectives. First, we are interested in understanding whether there is an interaction between age and current rank. We believe it may be possible that players who are

[1] Ultimate Tennis Statistics. *Top 500 Player Data*. https://www.ultimatetennisstatistics.com/

both young and high ranked could receive additional returns because they are more likely to be able to attend more tournaments as younger athletes and more likely to place highly at those tournaments. Secondly, we are interested in creating a model that has some amount of predictive power. Although this study is primarily explanatory, we recognize that finding a well fit model will have implications for the interpretation of other variables that could have otherwise been overlooked.

Analysis

To begin, we visualized the data, with our predictor variable, *best rank*, on the x-axis and the total *prize money* on the y axis. The data strongly suggest a quadratic fit visually. In best rank, there is a minimum of 1, a maximum of 488, and a *standard deviation* of 136. This provides sufficient variation of our predictor variable to analyze the relationship. In terms of values of prize money, there are two particularly notable points that are far higher than any other points, these values correspond to two players who have been ranked first during their career. In our analysis, it will be important to recall these points as possible outliers.

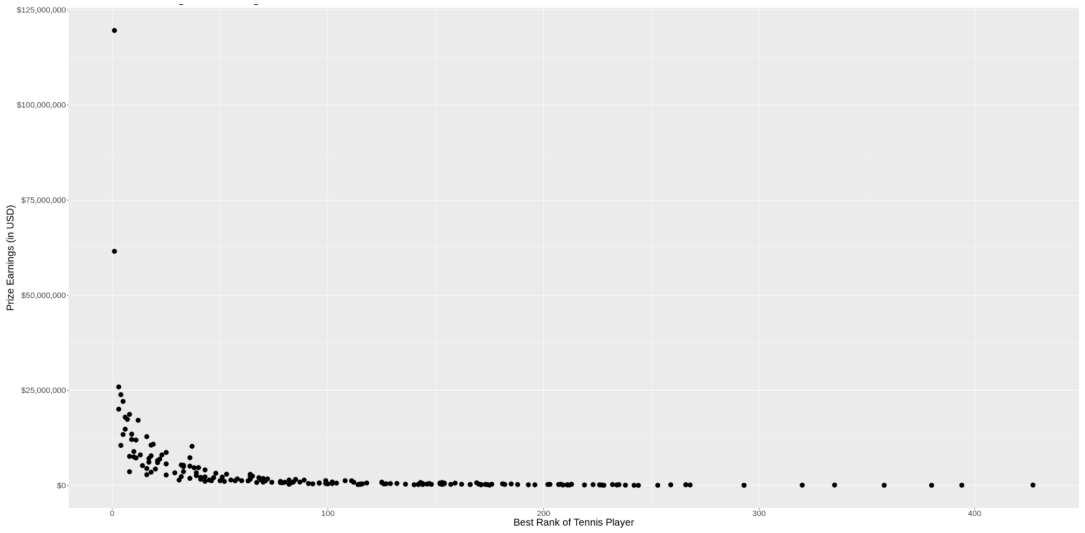

Figure 1: Player Ranking and Prize Earnings

Another concern from the data is that there is little variation in prize money for the players ranking lower. To visualize this, see figure 2.

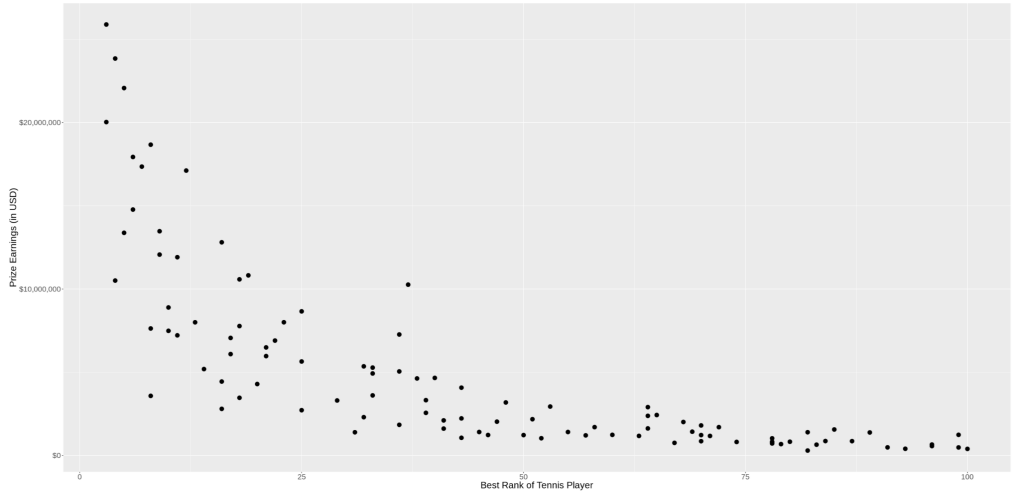

Figure 2: Player Ranking and Prize Money above 100

As we see, data above one hundred does not have the long-tailed plateau apparent in the larger dataset. For our analysis, it may be important to consider differing relationships for different sections of the data.

Table 1: Regression on Prize Money			
	Simple Model	Quadratic Model	Interaction Model
Seasons	766085****	-70807****	581883.5****
	(86001)	(81706)	(133101.1)
Age	-70807	-9040.97	150838.0
	(77156)	(73131)	(144745.4)
Current Rank	9410*	10865.99**	41768.4**
	(4946)	(4635)	(18583.6)
Best Rank	-11301****	-42350.15****	-10216.7****
	(2770)	(6436)	(2818.1)
Best Rank^2		66.76****	
		(12.67)	
Best Rank*Age			-1594.7*
			(883.2)
Adjusted R^2	.59	.64	0.59

Note: denotes significant at 10%, ** denotes significant at 5%, *** denotes significant at 1%, **** denotes significance at 0.1%*

Of the models, the quadratic model performs the best in terms of fit while the interaction model fails to improve upon the simple model enough to justify the inclusion of a new term. The coefficient on the interaction term does have a P-value of .07, however this is likely due to the diluted effect of best rank because the model performance did not improve.

Analyzing the best model, the quadratic model, we find that the coefficient on best rank is very significant at both the linear and quadratic specification suggesting the model specification is good. To explore possible issues we create the follow Q-Q plot:

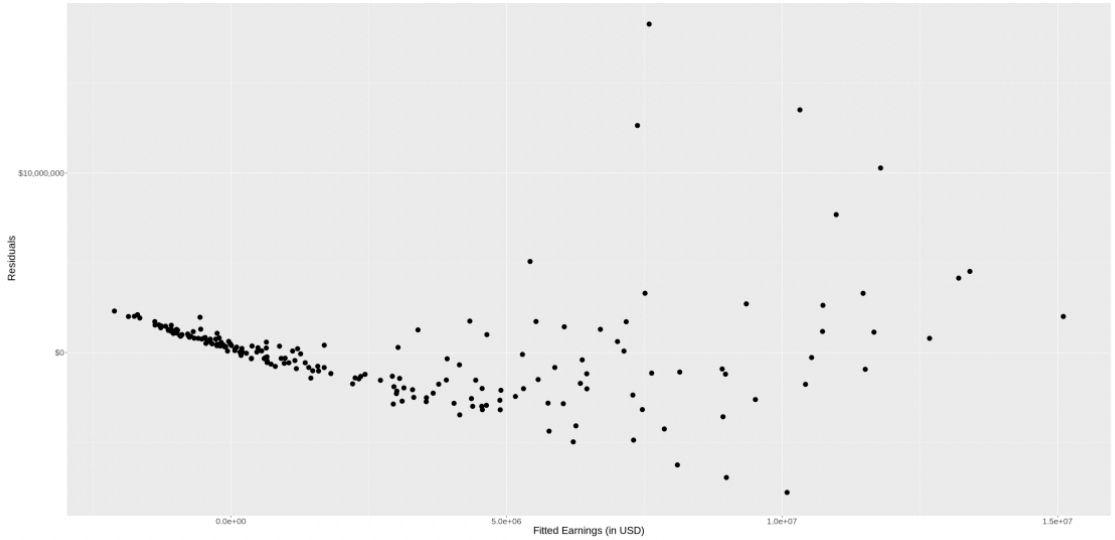

Figure 3: Residuals against Fitted Values

This graph shows a systematic problem with lower earning predictions. Regardless of specification, this relationship exists. It appears that there is some degree of plateauing or differing magnitude of relationship in the early portions of the dataset. Given this information we split the data between the players who achieved best rank above 100 (inclusively) and below 100. Going forward it will be important to test whether the relationship is quadratic at both lower and upper levels.

Although Age has demonstrated very little explanatory power in previous models, we keep it in later models as we hope to continue to explore the interaction between Age and Best Rank after splitting the data. The data split works surprisingly well in terms of generally improving goodness of fit, dramatically increasing adjusted R square, and exploring plausible explanations for the plateauing residuals for lower earning players.

After the split, residual plots of models fitted on top 100 ranked players show less plateauing effect, and residuals of the sextic model are much closer to normally distributed than before. The adjusted R squared of 0.886 of the sextic model fitted on top 100 ranked players is outstanding in all fitted models. P-values of estimates are also surprisingly all significant for that model, except for the Age variable which is still not significant. It is also worth-mentioning that seasons played seem to have less explanatory power for top 100 ranked players compared to the dominant Best Rank variable.

Initially we only fit a quadratic model whose residual plot shows curvilinearity. Then we keep increasing the power on the Best Rank term to get rid of curvilinearity. Testing models of different sizes, we find that the power on Best Rank can be increased up to 6 while all p-values associated with Best Rank remain significant. Normal plots also show increasing conformity to the 45 degree line, whereas previously they suggested insufficient fit. A model of power of 6 implies that rank has dominant explanatory power on prize money for the top 100 ranked players, which does make sense in a competitive sport. The other two models fit for top 100 ranked players improve as a result of data split, but the sextic model achieves the best fit for top 100 ranked players, while also having only significant coefficients (excluding the Age variable). Analyzing the sextic model, the adjusted R^2 improves after removing the and almost all estimates are significant at the 1% significance level (Current Rank is significant at 5% significance level) .

Overall, it seems that for the top 100 ranked players, neither Age nor interaction between Age and Best Rank have explanatory power with respect to prize money, while Best Rank is the most important when determining prize money. It also makes common sense that in popular sports, the differences in prize money between top ranks are not likely to be offset by simply playing more seasons.

Table 2: Regression on Prize Money for Top 100 Ranked Players				
	Simple Model	Sextic Model	Interaction Model	Best Model
Intercept	4635371	23142190.33453 ****	-4804984	23543929.155843 ****
Age	27226	23662.36640	438098 *	
Seasons	497540 ***	460770.91411 ****	289303	483724.674406 ****
Best Rank	-84338 ****	-3279251.51515 ****	96552	-3279208.115330 ***
Best Rank^2		191229.65378 ****		-5695.421386 ****
Best Rank^3		-5677.93130 ****		-5695.421386 ****
Best Rank^4		88.20656 ***		88.619801 ***
Best Rank^5		-0.68149 ***		-0.685807 ***
Best Rank^6		0.00206 ***		0.002081 ***
Current Rank	-10213 ***	-5441.89343 **	-7188 *	-5349.299283 **
Best Rank*Age			-7078 **	
Adjusted R^2	0.639	0.886	0.657	

Note: denotes significant at 10%, ** denotes significant at 5%, *** denotes significant at 1%,*

*****denotes significant at 0.1%*

For players who have not achieved top 100, all three models have a higher adjusted R^2 squared (around 0.73) compared with before the data partition. All three residual plots are more acceptable as normally distributed than residual plots of top 100 ranked players. One notable outcome of the split is that, in all three fitted models, estimates of the Age variable are significant at 1% significance level. The Age variable demonstrates explanatory power in all three models. Meanwhile, the estimate of interaction term is also significant at 1% significance level in the interacted model.

Seasons appear to have a somewhat dominant explanatory power for prize money (all three large positive estimates are significant at 0.1% significance level). Comparing the quadratic model and the interaction model, we prefer the interaction model because low-ranked players tend to earn more money from accumulated competition experience since they do not receive as many one-time large payouts as top-ranked players. We tried to increase the power of Best Rank in the quadratic model similar to the top 100 ranked quadratic model, but the outcome is not satisfying at all. One explanation for a lower power of Best Rank in the lower ranked dataset may be that players do not see as quick of an increase in pay outside of the top 100 because there are no large payout tournaments not dominated by the highest ranking players.

Table 3 presents the findings of the regression results discussed above:

Table 3: Regression on Prize Money for the Rest Players			
	Simple Model	Quadratic Model	Interaction Model
Intercept	-83437.7	336567.174 *	-909276.75 ***
Age	13256.8 ***	14815.511 ***	51650.83 ****
Seasons	102018.7 ****	86154.244 ****	87448.73 ****
Best Rank	-431.8	-4352.098 ****	4019.24 ***
Best Rank^2		8.020 ****	
Current Rank	-391.7 **	-397.529 **	-305.80 *
Best Rank * Age			-208.35 ***
Adjusted R^2	0.7122	0.7485	0.7435

Note: denotes significant at 10%, ** denotes significant at 5%, *** denotes significant at 1%, ****denotes significant at 0.1%*

Conclusion

The initial model appears quadratic with the explanatory variable “Best Rank” of tennis players and the response variable “Prize Money.” We notice a large variation in prize money for the players ranking around 0 to 100. We use three different models to analyze and study these data.

- Simple Model: PrizeMoney = Age + Best Rank + Seasons + Current Rank
- Quadratic Model: PrizeMoney = Age + Best Rank + (Best Rank)^2 + Seasons + Current Rank（BEST MODEL）
- Interaction Model: PrizeMoney = Age + Best tRank + Age*Best Rank + Seasons + Current Rank

Of the models, the quadratic model performs the best on the metric of adjusted R^2 returning a value of 0.64. (Note: a higher adjusted R-squared indicated that the new term improves the model more than expected by chance). Due to analysis of the residual plot, we decided to split our initial model into two parts, the players who achieved the best rank above 100 (inclusive) and those below 100. The best performing model on the Top 100 Ranked Players dataset was the sextic model returning an adjusted R^2 of 0.886 while still maintaining significant P-values. For the below 100, the relationship was found to be less strongly influenced by a quadratic best rank term, and more dependent on season's player.

The interpretation of these results is twofold. First, players at the top see a higher return to rank compared with players at the bottom, where total number of seasons played is a better determiner of prize payouts. This suggests that players who earn high prize money must be doing so consistently enough to achieve a high rank, players at the bottom do not enjoy nearly as successful careers despite still being world-class athletes.

Secondly, because the best performing model for the top 100 was sextic, whereas below 100 did not see explanatory power increase beyond quadratic, we may want to be wary of overfitting. The relationship is certainly very sharply exponential with best rank, implying that the highest ranking top players dominate the prize pools. This, however, does not provide reasoning for a sextic model. We see that the top 100 dataset has more outliers, thus rewarding a model that may be overfitting. This may imply that there is an increased "luck factor" at top levels where even a 50th ranked player may win a large payout but not achieve a lifetime high rank (whereas the data suggests this is very unlikely for a lower ranked player). This result may be useful for future work understanding sports payouts and dominance in athletics.