ECO 395M **Spring 2022**

Upsets in Professional Tennis: Who Really Knows?

Liam Crawley and Adhi Rajaprabhakaran

May 10, 2022

1 Abstract

This paper seeks to provide insight into the probability of the underdog winning a given match on the ATP tour. We use ten years of historical match data as well as betting odds in order to explore whether a player's ranking is a significant predictor of the outcome of a given match. In particular, we use instrumental variable probit regression to determine the causal effect of a players ranking on the probability that the match will be an "upset". Our analysis features a probit regression of a ranking point ratio (favorite/underdog) on a binary "upset" variable. The initial probit regression controls for level of the tournament and round, however we suspect some endogeniety issues with the primary regressor (a ratio of ranking points). We then instrument ranking points with Vegas betting odds, as there is clear correlation between the odds Vegas places on a player to win (or lose) and his current world ranking. We suspect Vegas odds is uncorrelated with the error term from the original probit model. The first stage IV probit model uses only the favoured players Vegas odds as an instrument. We find an increase in the magnitude of the coefficient on the endogenous covariate, and an increase in its significance. The second stage includes both the favorite and the underdog's Vegas odds as instruments. We find only a slight increase in the magnitude of the coefficient on player's ATP point ratio.

2 Introduction

We are both tennis players, and are interested in statistical analysis. Unlike baseball or football, little has been done that combines tennis and statistics. Also, there is not currently a ranking system that accurately predicts the outcome of a given match, in part due to the cyclical nature of tennis matches. It is not always the case (in fact many times it is not the case) that when Player A beats Player B, and Player B beats Player C, that Player A will also beat Player C. There are certain aspects of a players games that match up well (or poorly) with other players and so a linear ranking system is unable to capture the full picture.

My original idea was to develop a cyclical ranking system and tests its causal validity, but in the interest of time we decided to instead test the causal effect of current ranking system on the outcome of the match. We define "upsets" as our dependent variable. It takes the form:

$$y_{underdog} = \begin{cases} 0 & y_{underdog} \geq 0 \\ 1 & y_{underdog} < 0 \end{cases}$$

So, any time a player with less ranking points than his opponent wins the match, we consider it an upset. This might not be robust at the highest level considering the ranking point system and things like injuries, but it is mostly true for matches within our dataset. Table 1 shows a breakdown of upset percentage by tournament. There are four major levels of tournaments on the ATP tour: ATP250 (the lowest level), ATP500, Masters 1000, and the Grand Slams. The Masters Cup is a round robin of the top eight players in the world at the end of each year. In terms of volume, there are the least amount of Grand Slams (four throughout the year), and the most amount of ATP250s. From Table 1, we see that upset percentage is negatively correlated with tournament level - the greater the level of competition in the tournament, the less

upsets there are. Also, we see a fairly high percentage of upsets across the board, which provides validity of the assumption that a players rank is somewhat unpredictive of outcome.

Table 1: Upset Percentage by Tournament Level

	Series	favored_win	underdog_win	upset_pct
1	ATP250	7760	4597	37.20
2	ATP500	2917	1556	34.79
3	GrandSlam	4166	1526	26.81
4	Masters1000	4065	2046	33.48
5	MastersCup	128	49	27.68

An explanation of the point system on the ATP tour is necessary for a full understanding of this analysis, which uses a ratio of players ranking points as its primary covariate of interest. Players on the ATP tour accumulate points by winning matches throughout the course of the ATP calender year (which runs from January to September). No points roll over past the year mark, so points are capped at the maximum available within a year. For example, if a player wins a tournament in February, the points he accumulates only stay with him until that tournament ends the following February. So, if the player loses in the first round of the same tournament he won the year before, he loses 100% of the points he gained. As such, the more wins a player has (the more points he accumulates), the more points he must "defend" the next year.

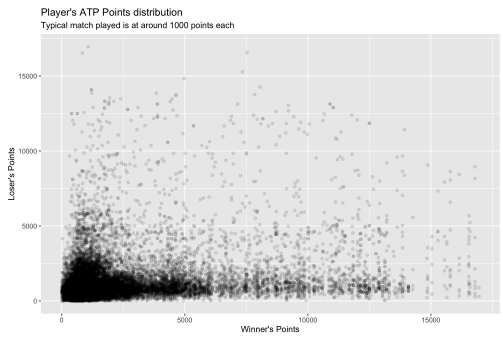

Figure 1: Distribution of ATP Points

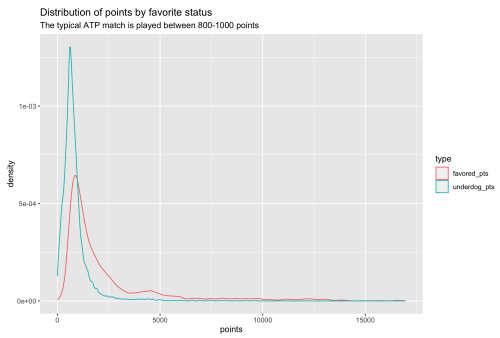

Figure 2: Density of Points

We see from Figure 1 that there is a significant clustering of points on tour towards the lower end of the spectrum. This is due mainly to to rollover effect in the sense that players have to win continuously year after year to maintain a high level of points, which is difficult. Points scale based on the level of the tournament being played, and are distributed by round. Winning an ATP250 level event would grant a player 250 ranking points, and the finalist 150. A quarterfinalist would only receive 45 points, and so forth. Winning a Grand Slam (the epitome of excellence in professional tennis) grants the player 2000 ranking points, which would propel an unranked player into about the top 20 players on the planet. There is direct linear correlation between the number of ranking points a player has and his current world ranking. Rankings are updated after every tournament.

Figure 2 shows overlapped density plots for the distribution of players who are favorites and players who are underdogs (in a given match, historically). The longer right tail on the "favorite" variable is indicative of a larger spread of players with more ranking points. The tall, narrow peak on the "underdog" variable shows that most underdogs are clustered at lower ranking points. Thus, the standard deviation of favorites is significantly larger than that of the underdogs. The average underdog has about 800 ATP ranking points, while the average favorite has around 1000.

3 Data

The data was gathered from a tennis statistics website that claims to have been gathered over the years directly from the ATP website, and compiled into excel files according to year. We use data from 2011 through 2021 in order to keep the players relatively consistent (tennis careers last anywhere from zero to twenty years, but the average is about 10). It also includes a plethora of Vegas betting odds from different casinos/bookies. Some cleaning was in order. We dropped all observations where a player retired from the match before completion or did not compete at all. Every tournament of the year at the tour level (ATP250 and up) is included, which left us with about 28,800 observations over the ten-year period.

Table 2: Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
favored_pts	28813	2317.011	2551.684	6	885	2475	16950
underdog_pts	28813	852.217	816.838	1	481	958	12005
pts_upset	28813	0.339	0.473	0	0	1	1
favored_odds	28856	1.413	0.236	1.01	1.22	1.6	5.05
underdog_odds	28856	3.965	3.115	1.54	2.3	4.21	36.44

From the table of summary statistics, we gather fairly large standard deviations on both the favorite's points and the underdog's. In our regressions, we use the ratio of favorite to underdog points. This allows us slightly simplified interpretation of the coefficient. The "favored_odds" and "underdog_odds" variables come from an average of all the available Vegas betting odds. It is worth noting that points and odds are opposite in sign - the player with higher points is the favorite but the player with lower odds is the favorite.

4 Analysis

Each of our regression models incorporate the same binary dependent variable, "pts_upset". Our most basic model is a probit GLM model of the form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{X}_i^1 + \beta_3 \mathbf{X}_i^2 + u_i$$

Where y_i is binary, x_i is our covariate of interest, favored/underdog points, $\mathbf{X}_i^1$, $\mathbf{X}_i^2$ are vectors of controls corresponding to tournament level and round, respectively, and u_i is the error term. We beleive there to be an endogeneity issue with x_i, where the error term contains more than just white noise information. This is difficult to test for a GLM errors do not follow a set distribution, such as OLS errors. But we believe that there is some unseen factor within the error term (call it ability or match-up) that is causing the predictive power of the simple probit estimate to be biased. This in turn does not allow for a causal interpretation of β_1. In order to eliminate some of this bias, we proceed to use to instrument x_i with betting odds.

In order for betting odds to be a valid instrument, it must be correlated with x_i, which embodies a players rank, and uncorrelated with the error term. We find strong correlation between Vegas odds and a players rank, and no evidence to suggest that odds are correlated with u_i. The instrumental variable probit model is then as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 X_i^1 + \beta_3 X_i^2 + u_i$$
$$x_i = z_i^1 \Pi_1 + z_i^2 \Pi_2 + \mathbf{v}_i$$

Where x_i is the endogenous covariate, z_i^1 is a vector of exogenous variables (i.e. $X_i^1 + X_i^2$), z_i^2 is a vector of instruments (in this case only the favorite's average Vegas odds), Π_1, Π_2 are matrices of reduced form parameters, and $\mathbf{v}_i$ is the error. By assumption and as mentioned before, $(u_i, \mathbf{v}_i) \sim N(0, \Sigma)$. The final model is nearly identical except we add underdog average odds as an additional instrument.

5 Results

The results from each model are summarized in the following tables. We will focus on the coefficient on "fu_pt_ratio", our variable that encompasses players' rank. Table 3 is the initial probit model. We find a negative, significant coefficient on "fu_pt_ratio". Interpreting the effect of a single covariate in a GLM, all else fixed, is not intuitive. However, we can conclude from this model that as the ratio of favorite to underdog ranking points increases (in other words the favorite gains more points or the underdog loses points), the chances of an upset decrease. This makes sense intuitively. The more highly ranked the favorite is, or the larger the ranking spread between two players, the less of a chance there is of an upset occurring.

After running the our initial probit regression, we began testing our instrumental variable models. The first stage IV model, which used only average betting odds of the favorite as an instrument (in addition to the vector of exogenous controls), finds a significant coefficient on "fu_pt_ratio". The results from the first stage IV model populate Table 4. The magnitude of the effect has increased relative to the initial GLM model. This result implies that when instrumented with Vegas odds, the predictive power of a players ranking on determining the probability of an upset in a given match between two players has increased. Given the access to information that Vegas has, this is not a surprising result. A player's ranking alone is not extremely predictive of the outcome of a given match but Vegas uses more information than just a player's ranking when calculating the odds for a given match. So, when we instrument a player's rank with the favorite's odds, the predictive power of rank improves. It is worth noting that this could be due to a faulty instrument. We discuss this at greater length below.

The final model uses two formal instruments on "fu_pt_ratio" as well as the vector of exogenous controls. We add the underdog's average Vegas odds (in addition to the favorite's) as an instrument, and find that the coefficient on "fu_pt_ratio" decreases slightly in magnitude as compared to the first-stage IV probit model. See Table 5 for detailed results of the second stage IV probit model. A more robust analysis would require weak instrument tests, however there are no post-estimation commands for IV probit models in R. We believe that both instruments have a strong enough effect on the endogenous variable of interest to permit identification. However, the second condition necessary for the instrument to pass the Anderson-Rueben weak instrument test is that the instrument cannot have a direct explanatory effect on the outcome variable.

Vegas odds are almost certainly predictive of upsets, however there have been studies (Conley, Hansen, Rossi (2012)) that have explored the validity of linear models that relax the aforementioned "excursion principle". There has been little to no research as to whether relaxing this principle in a non-linear model still allows for causal interpretation. One could also argue that Vegas odds are not predictive of the outcome of tennis matches, but bettors would beg to differ I believe. We ran a simple probit model with a ratio of Vegas odds in addition to the same controls used in our model against the binary upset outcome variable and find a significant coefficient on Vegas odds, which could unravel the validity of our model. The assumptions on GLM models are more relaxed than linear models as well, so it might be that our instrument remains valid in a probit model despite being invalid in a linear model. For the sake of our analysis, we maintain that Vegas odds to not affect the outcome of a tennis match, despite having some predictive power.

6 Conclusion

This paper seeks to determine the causal effect of a professional tennis ranking on the probability of an upset in a given match between two players on the ATP tour. It is the first step in developing a ranking system that is more predictive of outcome. We conclude that there is some causation between two player's rankings and the outcome of a given match, however the predictive power is very small. Much of our causal analysis relies on the assertion that Vegas odds are a valid instrument. If they are not, we cannot confirm a causal link between ranking and upset probability. However, the root of the motivation for the analysis was to test whether or not the current ranking system is a good predictor of the outcome of a match. We find that Vegas odds seem to have more predictive power than a player's simple ranking, which might imply that there exists a more complete way to rank players.

Perhaps the most useful extension of this project would be to select better instruments. Post-estimation on the ivprobit models also must be incorporated - we were surprised that an R package did not exist for this. The most complete version would require Vegas' formula for calculating a given players' odds in a given match and using that to create a more complete ranking system, then testing the strength of the new system against the old.

Regression Tables

Table 3: Probit Model (no instruments)

	Dependent variable:
	pts_upset
fu_pt_ratio	-0.029^{***}
	(0.002)
SeriesATP500	-0.047^{**}
	(0.023)
SeriesGrandSlam	-0.259^{***}
	(0.022)
SeriesMasters1000	-0.081^{***}
	(0.021)
SeriesMastersCup	-0.164
	(0.225)
Round2ndRound	-0.134^{***}
	(0.019)
Round3rdRound	-0.138^{***}
	(0.033)
Round4thRound	-0.214^{***}
	(0.062)
RoundFinals	-0.087^{*}
	(0.049)
RoundQuarterfinals	-0.093^{***}
	(0.027)
RoundRoundRobin	-0.231
	(0.252)
RoundSemifinals	-0.058
	(0.036)
Constant	-0.175^{***}
	(0.015)
Observations	28,813
Log Likelihood	$-18{,}101.960$
Akaike Inf. Crit.	36,229.920
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: First Stage Instrumental Variable Regression

	Coef	S.E.	t-stat	p-val
Intercep	-0.09	0.52	-0.18	0.86
Series_ATP500	-0.08	0.05	-1.79	0.07
Series_Masters1000	-0.20	0.04	-4.55	0.00
Series_GrandSlam	-0.26	0.05	-5.58	0.00
Series_MastersCup	-0.33	0.47	-0.70	0.48
Round_1stRound	0.41	0.52	0.77	0.44
Round_2ndRound	0.29	0.52	0.55	0.58
Round_3rdRound	0.23	0.53	0.44	0.66
Round_4thRound	0.15	0.54	0.28	0.78
Round_Quarterfinals	0.16	0.53	0.31	0.76
Round_Semifinals	0.19	0.52	0.36	0.72
Round_Finals	0.12	0.53	0.24	0.81
fu_pt_ratio	-0.13	0.01	-23.24	0.00

Table 5: Second Stage Instrumental Variable Regression

	Coef	S.E.	t-stat	p-val
Intercep	-0.14	0.55	-0.25	0.80
Series_ATP500	-0.11	0.05	-2.13	0.03
Series_Masters1000	-0.21	0.05	-4.43	0.00
Series_GrandSlam	-0.38	0.05	-7.74	0.00
Series_MastersCup	-0.32	0.50	-0.64	0.52
Round_1stRound	0.39	0.56	0.70	0.48
Round_2ndRound	0.24	0.56	0.43	0.66
Round_3rdRound	0.18	0.56	0.33	0.74
Round_4thRound	0.10	0.57	0.18	0.86
Round_Quarterfinals	0.16	0.56	0.28	0.78
Round_Semifinals	0.19	0.55	0.34	0.73
Round_Finals	0.14	0.56	0.25	0.80
fu_pt_ratio	-0.11	0.00	-23.45	0.00

Note that data was gather exclusively from: http://tennis-data.co.uk/alldata.php