IDS 701: Final Project

Team 1: Peining Yang, Raza Lamb, Michelle Van

Introduction

The umpire calls out "quiet please": a hush falls over the crowd. A raucous afternoon in Flushing, New York, and the tension is palpable. Novak Djokovic is attempting a feat of history: the Grand Slam—winning all four Grand Slam tournaments in a single year. Ultimately, he fails: Daniil Medvedev wins his first Grand Slam, and history has to wait. In the aftermath, John McEnroe makes an offhanded comment about how tired Djokovic must be, and how long of a season he had. But, this comment logically leads the listener to a question: "How big of a role does fatigue play in determining a tennis champion?" There's a logic to both sides of the argument: obviously, fatigue is an important factor in an outcome, but doesn't experience also increase a players chance of winning? In fact, this is a conundrum that many players face in determining their schedule.

Here, we investigate this question by examining the causal relationship between the length of a tennis player's match and their probability of winning the next match. To truly understand the causal relationship at work, it would be necessary to conduct a randomized experiment in which tennis players could be randomly assigned opponents from a controlled pool. This way, we could manually limit what affects the probability of winning the next match. However, this is not a feasible experiment in the current setting, so instead we investigate two separate methods. First, a linear probability regression with fixed effects for player and opponent characteristics, where the outcome is a simple binary variable indicating win or loss. Secondly, we conducted a manual matching experiment in which we create a control and treatment group based on a tiebreak in the third set. In a tennis match, each set is usually played to six games, having to win by two games. If the players are tied at six games each, they will play a tiebreak up to seven points, with the winner having to win by two points. In this way, we treated a tiebreak as a "near-randomization" to play a third set, in which the players that won the tiebreak (and thus the match) are the control group, and the players that lost the tiebreak (but won the match in the final set) are in the treatment group.

Data

The data utilized in these analyses comes from a public GitHub Repository that contains historical match-level data for professional men's tennis (ATP) all the way back to 1968. For the purposes of this analysis, all data between 1995 and 2019 was selected. Data after 2019 was available, but was not utilized. In 2020, the ATP significantly changed their ranking system to account for tournament cancellations due to the 2020 coronavirus pandemic.

Significant data cleaning was required to enable analysis. First, the data had to be "doubled", so to speak, in order to have a row represent both the winning player and the losing player. Next, the data had to be thinned to a subset of matches that are comparable. We removed all matches from tournaments that play best-of-five matches, as those may not be comparable to best-of-three tournaments. Tournaments that included round-robin style of play were also excluded, as the rounds do not have a logical consecutive structure required for identifying the previous match. Matches where the opponent was a walkover (i.e. they forfeited the match) were not included, but matches where the opponent retired (i.e. played, but did not complete the match) were included. To extract the potential causal factor, previous match statistics were added to rows, and all first-round matches (i.e. matches where the players did not have a previous match within that tournament) were removed. Finally, during data exploration, missing and likely erroneous data were identified and removed. This included 4 matches where the match length was greater than 400 minutes, despite only playing

3 sets and no tie-breaks. There were approximately 3,000 rows with missing data in critical columns, such as minutes, previous minutes, or player rank. These observations were dropped. Finally, there were 50,320 observations with 24,250 match wins and 26,070 match losses.

Table 1 shows the breakdown of key statistics by win and loss in the data used for the first analysis. Besides opponent rank and previous match rank differential, the data is relatively balanced.

Table 1: Summary Statistics between Winners and Losers

Status	**Variables**	**Mean**	**Std.**	**Min.**	**Max.**
Winners	Match Minutes	96.34	30.79	4.00	282.00
	Age	25.98	3.68	16.72	42.79
	Height	185.94	6.69	168.00	211.00
	Player Rank	52.07	62.80	1.00	1346.00
	Opponent Rank	71.13	83.84	1.00	1890.00
	Previous Match Rank Differentials	-44.03	128.66	-2090.00	1295.00
Losers	Match Minutes	95.70	30.55	4.00	282.00
	Age	26.21	3.73	15.82	42.79
	Height	185.54	6.68	168.00	211.00
	Player Rank	77.78	86.99	1.00	1890.00
	Opponent Rank	45.32	57.54	1.00	1346.00
	Previous Match Rank Differentials	-23.12	146.22	-2125.00	1711.00

For the second method, the data had to be manually separated into control and treatment group. In this case, the control group is defined as a player who won their previous game in two sets, where the second set was a tiebreak. The treatment group is thus a player who won there previous game in three sets, where they lost the second set in a tiebreak. After limiting this data, there were 9,544 rows, with 7,030 in the control group and 2,514 in the treatment group. The summary statistics for these matches are listed in Table 2 below, separated by treatment and control. The treatment and control group are fairly well balanced in this case, except for the previous match rank differentials. In the treatment group, the difference is smaller, on average, indicating that the matches that the players that lost the tiebreak were closer in rank to their opponents compared to the players that won the tiebreak.

Table 2: Summary Statistics between Treatment and Control Groups

	Variables	**Mean**	**Std.**	**Min.**	**Max.**
Treatment	Match Time (mins)	95.51	30.80	12.00	282.00
	Age	26.35	3.81	16.88	40.62
	Height	186.61	7.04	168.00	211.00
	Player Rank	65.22	72.67	1.00	1078.00
	Opponent Rank	55.38	66.07	1.00	882.00
	Previous Match Rank Differentials	-24.43	124.11	-1618.00	1038.00
Control	Match Time (mins)	96.29	30.98	5.00	266.00
	Age	26.29	6.94	16.86	42.79
	Height	186.38	6.94	168.00	211.00
	Player Rank	64.63	76.97	1.00	1147.00
	Opponent Rank	57.03	73.13	1.00	1821.00
	Previous Match Rank Differentials	-31.07	131.31	-1457.00	973.00

Linear Probability Model

A regression is necessary in this context because there are several confounding features that likely cloud the true relationship between between previous match length and current winning probability. For example, we would expect very skilled players to dispatch their opponents, especially in earlier rounds, easily, and thus quickly. Therefore, previous match length could be acting as an indicator for player quality, instead of only representing exertion. In order to hopefully extract this correlation, we turned to a linear probability model.

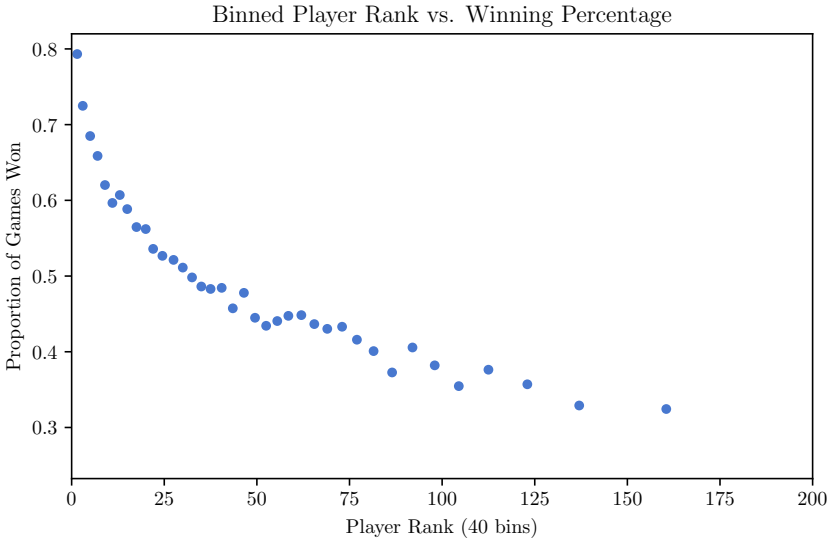

Figure 1: Proportion of Matches Won by Binned Player Rank.

Before adding controls into the model, first we investigated the relationship between various potential controls and the probability of winning. This was done by binning continuous variables into equal sized buckets and plotting the proportion of matches won within each bucket. The relationship was investigated for player rank, opponent rank, player age, previous match length, and previous match rank differential. Previous match rank differential is defined as the players rank minus the rank of their previous opponent. Figure 1 displays the binned player rank vs. the proportion of games won, and Figure 2 displays the binned previous match length vs. win percentage. Clearly, the relationship between rank and winning is not linear, while the relationship between the length of the previous match does appear roughly linear. The other plots are not shown, but did display similar trends. Opponent rank also shows a non-linear relationship, while age shows a nearly linear trend with decreasing win probability with increasing age.

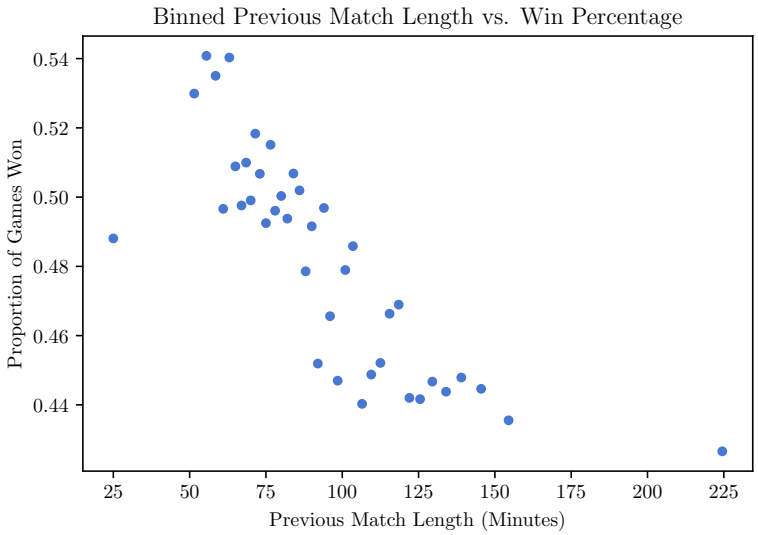

Figure 2: Proportion of Matches Won by Binned Previous Match Length.

Based on our preliminary findings, we fit three models to the data. In the first model, the only predictor included is the length of the previous match, in minutes. In the other two models, other parameters that may have an effect on win probability are included as control. This includes player age, player rank, opponent rank, and previous rank differential. In the third model, we included square terms for rank and opponent rank, based on the exploratory analysis.

Table 3: Linear Probability Models Results

	Simple Model	Model with Covariates	Model with Square Terms
Intercept	0.5647***	0.6735***	0.6733***
	(0.0073)	(0.0171)	(0.0177)
prev_minutes	-0.0009***	-0.0007***	-0.0006***
	(0.0001)	(0.0001)	(0.0001)
age		-0.0045***	-0.0047***
		(0.0006)	(0.0006)
rank		-0.0012***	-0.0022***
		(0.0000)	(0.0001)
$rank^2$			0.0000***
			(0.0000)
opp_rank		0.0013***	0.0024***
inf		(0.0000)	(0.0001)
opp_rank^2			-0.0000***
			(0.0000)
prev_rank_diff		0.0001***	0.0001***
		(0.0000)	(0.0000)
R-squared	0.0028	0.0659	0.0934
R-squared Adj.	0.0028	0.0658	0.0933

As can be seen in Table 3, in the simple model, the coefficient for previous match minutes is 0.0009, which indicates that for a player who's previous match length is 10 minutes longer, their

probability of winning is approximately 1 percentage point lower. However, once the controls are added, including the square terms, this goes down to 0.0006, which is now a 0.6 percentage point decrease in the probability of winning, given a 10 minute increase in previous match length.

Manual Matching Experiment

While the linear probability regression can control for player strength, opponent strength, age, and previous match ranking differential, it is possible that the causal estimate is confounded. After controlling for all these characteristics, previous match length could be an indicator for another factor, such as a lingering injury or simply just a bad year for a player. That is, while we are interpreting the previous match minutes and simply whether the player played more tennis, it could still be an indicator for player quality, even after controlling for player rank. To try to test this hypothesis, we designed another experiment, in which we treat a second set tie-break as a "randomization". Because a tiebreak is much shorter than a set, we can treat the winner and loser as pseudo-randomized, or something near it.

The treatment group consisted of games where the player won the first set, lost in a tiebreak for the second set and won their third set and subsequently winning the match. The control group consisted of games where the player won their first set, played a tiebreak in the second set and won the game. Similar to our previous linear probability analysis, we fit three models to the data. The first model consists of only the treatment variable as the predictor and match outcome as the response variable. In the second model, we included the same control variables as the previous models, which are age, rank, opponent's rank and the rank differential in the previous match. Building on top of the second model, we included square terms for rank and opponent's rank in the third model. Results from all three models can be seen in Table 4 below.

Table 4: Manual Matching Model Results

	Simple Model	Model with Covariates	Model with Square Terms
Intercept	0.4921***	0.6170***	0.6304***
	(0.0060)	(0.0369)	(0.0370)
treat	-0.0510***	-0.0482***	-0.0465***
	(0.0116)	(0.0112)	(0.0110)
age		-0.0049***	-0.0047***
		(0.0013)	(0.0013)
rank		-0.0011***	-0.0024***
		(0.0001)	(0.0001)
$rank^2$			0.0000***
			(0.0000)
opp_rank		0.0014***	0.0024***
		(0.0001)	(0.0002)
opp_rank^2			-0.0000***
			(0.0000)
prev_rank_diff		0.0000	0.0000
		(0.0000)	(0.0000)
R-squared	0.0020	0.0653	0.0952
R-squared Adj.	0.0019	0.0648	0.0945

Results show that in the simple model, players in the treatment group have a model coefficient of -0.0510, which indicates that if a player loses the tiebreak in the second set and wins the match

in the third set, their probability of winning is 5.10 percentage points lower compared to a player who won the tiebreak in the second set. After including the control variables and square terms, the model coefficient for the treatment group increases to -0.0465, which is now a 4.65 percentage point decrease in win probability compared to players in the control group. The coefficients for all three models showed a p-value of less than 0.05, indicating that the results are statistically significant.

In addition, we also explored interaction terms between the player's age and whether they are in the treatment group. Intuitively, we would expect the increase in age to affect a player's probability of winning given that they have previously played a longer match. We created six bins for the age variable, spanning from the minimum of 22.48 and maximum of 40.624. Results from the model is shown in Table 5 below.

Table 5: Model Results of Manual Matching Model with Interaction Terms

	Model with Interaction Terms
Intercept	0.5361***
	(0.0145)
treat	-0.0876***
	(0.0286)
age(22.48, 24.501)	-0.0357*
	(0.0206)
age(24.501, 26.185)	-0.0309
	(0.0207)
age(26.185, 27.86)	-0.0500**
	(0.0205)
age(27.86, 30.073)	-0.0626***
	(0.0206)
age(30.073, 40.624)	-0.0862***
	(0.0207)
treat:age(22.48, 24.501)	0.0439
	(0.0404)
treat:age(24.501, 26.185)	0.0255
	(0.0401)
treat:age(26.185, 27.86)	0.0518
	(0.0407)
treat:age(27.86, 30.073)	0.0572
	(0.0404)
treat:age(30.073, 40.624)	0.0453
	(0.0398)
R-squared	0.0044
R-squared Adj.	0.0033

As we can see from the results above, the interaction terms between the treatment variable and age were in fact not statistically significant. It is interesting to note that players in the treatment group (lost in second tiebreak and won in the third set) aged 24.501 to 26.185 have the lowest probability of winning compared to players of all other ages.

Discussion

The results of both analyses appear to confirm what may initially seem obvious—a longer previous match decreases the probability of winning, holding all else equal. Controlling for other factors significantly decreases the impact of previous match length, but it still has a significant impact.

Based on our results, an aggressive match strategy is an ideal one—clearly there is value in not just winning, but winning as fast as possible.

Limitations

Although an in-depth analysis and inferences were drawn from the data, limitations were met throughout our analysis.

When dealing with the linear probability model that analyzed the players probability of winning a match, we only included his opponent's ranking and opponent's previous ranking differential. Although these are good indicators in representing the winning player's probability, we did not take into consideration factors that could have affected the losing player's loss such as the player's physical conditions.

The data was collected by a public open source, as such, there were certain areas of the data that were not as reliable as we expected them to be. Due to changing regulations in tennis tournaments, the number of sets required to win a game, number of points needed to win in a tiebreak, and the number of allowed minutes in a game changed over the past decades. Therefore, comparing the winning player's probability back in the early 1900's to 2010's is not a straightforward comparison and thus is a limitation in our analysis.