The most powerful tennis player: Are John Isner's wins attributed to his serve power?

Minh Nguyen

12/16/2021

Abstract

This report examines the effects of John Isner's aces on his victories in professional tennis matches. This is done by building a logistic regression model using his in-game stats as predictors. We run AIC on the model, then conduct Propensity Score Matching to determine the effects of his wins on his service game. After that, we run a linear regression on the matched dataset. We determined that **among pairs of matches that had the same probability of winning, if he aces his opponent at least once, he has an 86% higher chance of winning that same match than if he hadn't**. We conclude that achieving an ace had a significant effect on his success in winning games.

Keywords: Aces, Service Games, Break Point, Double Fault, Serve Return, Nearest Neighbour Matching

Introduction

Over the past decade and a half, professional men's tennis has been dominated by three men, colloquially known as "The Big Three". These three men are Roger Federer, Rafael Nadal, and Novak Djokovic. Among the lesser known players is John Isner. Standing at a staggering 2.08 meters tall (6 ft. 10 inches), Isner is often regarded as the best server on the tour. Affectionately referred to as a "serve bot" by his fans, Isner's playstyle is characterized by serving as hard, fast, and accurately as possible. His brute force pressures his opponents to predict the direction of the ball before he even begins service. This puts him in a position to volley should they successfully return the ball, often referred to as the "serve and volley" tactic.

Figure 1: John Isner with Roger Federer, and John Isner prior to service

Statistically speaking then, it should be no surprise that service games are Isner's primary strength, as it would seem that this is where he delivers a majority of his lethal blows. Whether this is true for all his games is what this report intends to figure out.

With that being said, we aim to **determine the effects of Isner's aces on pairs of matches where he had the same probability of winning**. To do this, we require two datasets: a dataset on all professional matches between 1991-2016, and one containing all in-game statistics for every match between 1991-2016. There are a total of 91,000 datapoints, with each datapoint representing a tennis match managed by the Association of Tennis Professionals (ATP) since 1991. The raw data comes from the ATP World Tour website, while the datasets were cleaned and prepared by Kevin Lin and Andrew J. Kuo on datahub.io [2].

To determine this effect, we construct a multivariate logistic regression model. The predictors of the model are the available match statistics of every single match in Isner's professional career up to 2016. This implies that we only need 210 of the 91,000 datapoints that we are provided. We then run AIC on the model to remove statistically insignificant predictors. Once we have the final model, we proceed to use Propensity Score Matching to determine the effect of winning on him acing his opponents per match. Finally, we construct a linear regression model to determine the correlation between his career aces and his wins.

Based on real world data, Isner holds multiple ATP records, such as the fastest recorded serve of all time, as well as the most recorded aces in a match [1]. Thus, I hypothesize that **the effect of Isner winning a match is a large ace rate**. In other words. I believe a a win from him implies acing his opponent more than once.

Data

Data Collection

There are two datasets that we will be using: a **score dataset** that contains information on all players, sets, games, and matchups since 1991, and a **match dataset** that contains specific in-game statistics. The raw data was obtained from the ATP World Tour website, with the datasets being prepared and cleaned by Kevin Lin and Andrew J. Kuo from datahub.io [2]. In total, there are approximately 91,000 observations in both the match and score datasets

The match and score datasets are connected through a primary key, where accessing specific in-game statistics requires knowing this key. The key is 15-16 characters long and alphanumeric, separated by a comma every four characters. For example, if we wanted to obtain in-game statistics for the 2006 US Open Finals between Roger Federer and Andy Roddick, the key for the match would be of the following form

2006-560-f324-r485

Each player also has a unique key as well, which is 4 characters long and alphanumeric. In the above case, Federer has the key **f324** and Roddick is **r485**. It is important to keep in mind the order in which the player ID comes within the key; **f324-r485** implies that Federer won and Roddick lost. If it was **r485-f324**, that would mean Roddick won and Federer lost. **2006-560** is the tourney ID, where 2006 represents the year and 560 represents the ID of the US Open.

To access the in-game statistics of the match, we search for **2006-560-f324-r485** in the score dataset. Here, we are given a plethora of information ranging from match duration, to aces, second serve points won, first serve points won, etc. Specific data for each of these variables is recorded for both the winners and the losers of the match. There are a total of 25 variables that are recorded for both the winner and the losers of a match. We will explain the relevant variables for the analysis in detail, without much regard for the other variables.

Data Cleaning

Score Dataset

As the primary focus of our analysis is to construct a model based on data from Isner's matches, we begin by removing datapoints that do not include him (which is a majority of them). This leaves us with 210 total observations, which represent all 210 matches in his career up to 2016. We filter for his matches by searching for player ID **i186**.

After cleaning for all his relevant data, we need a binary variable that determines whether he wins or loses a match. Fortunately for us, the match dataset has two variables that determine whether a player wins or loses a match: aptly named Winner Player ID and Loser Player ID. If **i186** is in Winner Player ID, that means Isner won that match, vice versa for Loser Player ID. For our purposes, we create a new variable called Win-Loss that assigns a value of 1 if he wins and 0 if he loses. Thus, for every match he has played, we either get a 1 or a 0.

Match Dataset

We collect all the match ID's from our cleaned score dataset and find their corresponding statistics in the match dataset. We then filter the match dataset so that it contains in-game statistics for all of Isner's matches. One thing to keep in mind is that these in-game statistics are often provided for both the winner and the loser. That means in one match, we will have 25 different variables measuring the winner, and another 25 for the loser.

To find all his relevant matches, we extract the Win-Loss variable from the score dataset, and append it to the match dataset. After that, if a row has a Win-Loss value of 1, we append the winner's aces, winner's first serve points, etc. to a new dataset. Vice versa for Win-Loss value of 0. This final dataset will be the dataset containing the match statistics for all matches in his career. It will also be the one we construct our regression model with. All the relevant variables are listed below.

For convenience sake, we will only list 10 of the 25 predictors in the Description of Variables section.

Description of Variables

Score Dataset

Table 1: Description of important variables in the score dataset

Variable	**Description**
Winner Player ID	The 4 character alphanumeric ID of the winning player of a match
Loser Player ID	The 4 character alphanumeric ID of the losing player of a match
Match ID	The 16 character alphanumeric ID of a match between two players

Match Dataset

Table 2: Description of important variables in the match dataset

Variable	**Description**
Match ID	The 16 character alphanumeric ID of a match between two players
Match Duration	The length of the match in minutes
Aces	Number of serves from a server that is not touched by the receiver
Double Faults	Number of times a player serves twice and isn't able to get the ball in the receiving area (thus losing a point)
First Serve Points Won	Number of points won when the player serves for the first time and gets the ball in the receiving zone

Variable	Description
Second Serve Points Won	Number of points won when the player serves for the second time and gets the ball in the receiving zone
Break Points Saved	Number of times an opponent is one point away from winning the set, and the player serves and manages to win the point, preventing the opponent from winning the point
Service Points Won	Number of points won when the player serves
First Serve Return Won	Number of times an opponent serves and the player returns and manages to win the point
Second Serve Return Won	Number of times an opponent serves on their second try and the player returns and manages to win the point

Numerical Summaries and Plots

Below are some numerical summaries of Isner's career

Table 3: Numerical summaries of important variables

Variable	Min	First Quartile	Median	Mean	Third Quartile	Max
Aces	0.000	3.000	5.000	5.824	8.000	31.000
Match Duration	33.000	85.0	109.0	121.3	144.0	353.0
First Serve Points Won	9.00	27.00	35.00	38.19	46.00	102.00
Second Serve Points Won	3.00	11.00	15.00	16.08	19.00	43.00
Break Points Saved	0.000	1.000	3.000	3.371	5.000	19.000
Service Points Won	14.00	39.00	51.00	54.27	64.00	128.00
First Serve Return Won	0.00	12.00	16.00	17.64	21.00	49.00
Second Serve Return Won	0.00	13.00	17.00	18.38	22.00	54.00

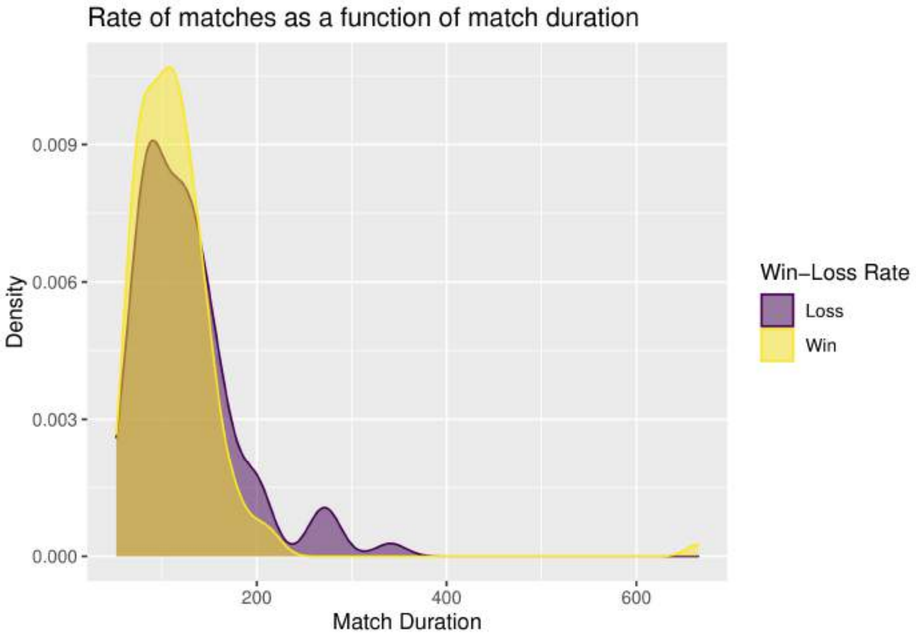

Figure 2: Isner's career match duration

Figure 2 shows Isner's match duration among his wins and losses. It would seem that a majority of his matches, regardless of win or loss, are concentrated at 90~100 minutes. This is the standard for most professional tennis matches, so is not of interest. But what is significant is the slight bump at 230 minutes onwards. This could perhaps be due to fatigue, as John Isner's large frame makes it difficult to move across the court efficiently, thus losing him more games.

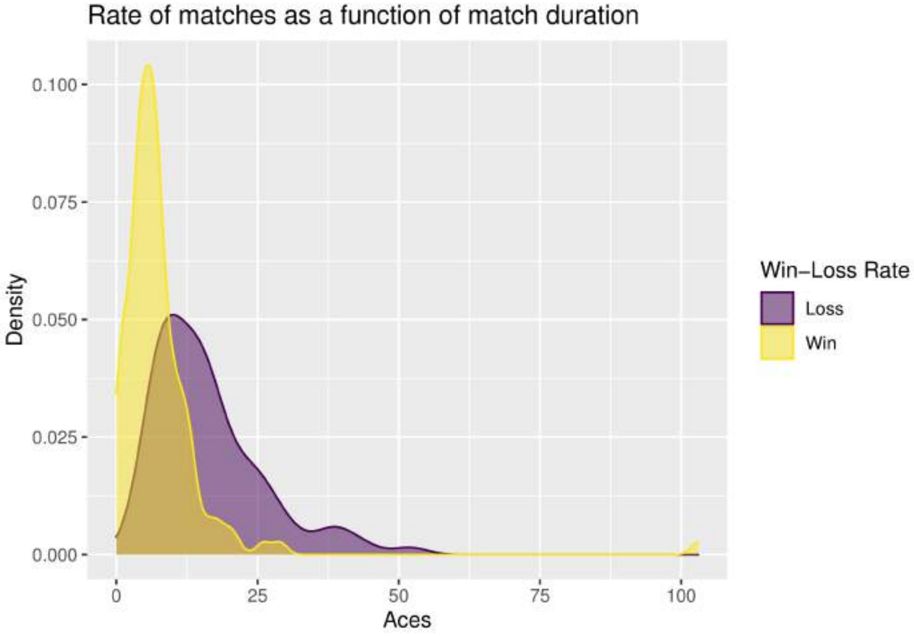

Figure 3: Isner's career match duration

Figure 2 reveals something surprising to us, namely that a significant proportion of his won matches had him acing his opponent less than <12 times. In contrast, him acing >12 times usually indicates a loss. While correlation does not equal causation, this adds evidence against the myth that Isner wins matches primarily through brute force serving. Isner specialists claim that his primary tactics are a powerful forehand and serve-volley. So perhaps an in-court metric to measure shot speeds would yield results that align more with the prevailing stereotype around his game.

Methods

Logistic Regression

Since we aim to model Isner's win-loss, we utilize multivariate logistic regression as a means of doing so. We choose a frequentist approach simply because of its clarity and simplicity of use. Our initial regression model is defined as follows:

$$l_{WinLoss} = log(\frac{p}{1-p}) = \beta_0 + \sum_{i=1}^{25} \beta_i x_i + \epsilon$$

In the above model $l_{WinLoss}$ is the log-odds of Isner winning or losing. The β_0 is the y-intercept, while $\sum_{i=1}^{25} \beta_i x_i$ are all 25 parameters of the model. We have not listed all 25 parameters of the model due to time constraints.

Akaike Information Criterion (AIC)

AIC is a method of variable selection, where the aim is to determine variables that do not contribute a statistically significant amount to the prediction. In our specific case, we run backwards AIC on our logistic regression model. Backwards AIC is essentially a combination of backwards selection and the entire AIC process.

The way the process works, is that we refer to the step() function offered in the Base R package. When we run step(model, direction = "backwards"), the function loops through our model and determines the effect of removing each predictor independently on the accuracy of the model. A lower AIC score usually implies a less cumbersome model that is still almost exactly as accurate as the original. The output of the function is the final model with most statistically insignificant predictors removed.

Propensity Score Matching.

General Idea

The primary method of interest for this report is Propensity Score Matching (PSM), which we will utilize to determine the effect of winning on the amount of aces that John Isner serves. The technique begins by assigning a probability to every observation we have; that's 210 observations corresponding to 210 games played by Isner. We construct the probabilities by basing it on the 11 predictors that we have in our final model. This probabability is called the **propensity score**. The assigned probabilities will be our best guess in determining the probability of a match being a win, regardless of whether Isner actually won or lost the match.

After that is done, we split these matches into a treatment and control group. The general goal of PSM is to determine the effect of administering something (usually a drug, a coupon, etc) to our designated treatment group. Depending on the goals of PSM, the result that we want to measure will usually depend on this treatment. In our case, our "treatment" group is the set of John Isner's wins.

Once we have these probabilities, we begin the matching process. The idea behind matching can be explained as follows. Suppose we take two games, one from our treatment group, and one from our control group:

Game 1: 110 minutes long and Isner won 30 second-serve points. He won this game

Game 2: 100 minutes long and Isner won 40 second-serve points, He lost this game

The matching algorithm we run would dictate that the assigned probability for both games is similar to each other, as there is not much difference in its variables. In this case, there isn't much difference between 100 and 110 minutes, as well as 30 and 40 second-serve points. There are many other algorithms that match probabilities with each other, but the above example illustrates **Nearest Neighbour Matching**. This is the algorithm we utilized for this report, and can be found in the *arm* package.

Assumptions

One important assumption with PSM is that it relies entirely on observational data. We also assume that it is possible to match pairs from both treatment and control groups. In fact, in the case of the Nearest Neighbour approach, our matches are constructed primarily of pairs from both groups.

Results

Variable Selection

Given the cumbersome nature of having 25 variables, we run backwards AIC on our model to remove insignificant variables. The final model is given below:

$$
\begin{aligned}
l_{WinLoss} &= log(\frac{p}{1-p}) \\
&= \beta_0 + \beta_{DoubleFaults}x_{DoubleFaults} \\
&+ \beta_{FirstServeIn}x_{FirstServeIn} \\
&+ \beta_{FirstServeTotal}x_{FirstServeTotal} \\
&+ \beta_{FirstServePointsWon}x_{FirstServePointsWon} \\
&+ \beta_{SecondServePointsWon}x_{SecondServePointsWon} \\
&+ \beta_{FirstServeReturnWon}x_{FirstServeReturnWon} \\
&+ \beta_{FirstServeReturnTotal}x_{FirstServeReturnTotal} \\
&+ \beta_{SecondServeReturnWon}x_{SecondServeReturnWon} \\
&+ \beta_{SecondServeReturnTotal}x_{SecondServeReturnTotal} \\
&+ \beta_{ServiceGamesPlayed}x_{ServiceGamesPlayed} \\
&+ \beta_{ReturnGamesPlayed}x_{ReturnGamesPlayed} \\
&+ \beta_{IsnerWinLoss}x_{IsnerWinLoss} + \epsilon
\end{aligned}
$$

Propensity Score Matching

The results of this model will be our propensity score. We then append this forecast to our match dataset. Our goal is to use this forecast to create matches. For every match that Isner won, we want a lost match that was considered statistically similar to the won match based on our predictors (this assigns a propensity score to every observation).

For our purposes, we're going to use a matching function from the *arm* package. This function finds pairs of matches with the closest propensity score between won and lost matches. Once this process if complete, we just need to reduce the match dataset to pairs of observations that are matched with each other.

After the matching process, we run a linear regression model on our new matched dataset. The purpose of this model is to observe the difference between having serving at least one ace, and not serving any aces on Win-Loss rate. The results of the model are below:

y-intercept	Double Faults	First-Serves In	First-Serves Total
-1.137	0.312	-0.172	-0.091

First-Serve Points Won	Second-Serve Points Won	First-Serve Return Won	First-Serve Return Total
0.522	0.211	-0.003	-0.019

Second-Serve Return Won	Second-Serve Return Total	Service Games Played	Return Games Played
-0.078	0.129	-0.467	-0.068

Win-Loss Rate
0.862

Please note that this process was a modification of the code provided by Rohan Alexander [4].

Recall that we interpret the above results as individual coefficients for our linear regression model. An interesting thing to note is that if we an 86.2% chance of winning a match, then we would have landed at least one ace. In other words, if we didn't land an ace, we would not have an 86.2% chance of winning. The realistic scenario, is that Isner would have much less than an 86.2% chance at winning, as not landing aces implies a weak service game in general.

Conclusion

There are several key results in our analysis that we will reiterate. We hypothesized that the effect of Isner winning a match is usually a high number of aces. What we've discovered using PSM was a little bit more complicated.

Before we got to that, we had to clean the dataset of relevant variables. This involved finding Isner's player ID in the score dataset, which is i186. After we got his ID, we ran through the score dataset to determine which matches he won and lost, and compiled this as a binary variable. We then appended this variable to our match dataset, and found all of the 25 in-game statistics for every match he played since the beginning of his career.

Once we had these predictors, we constructed a logistic regression model with them, with Win-Loss as output. We then ran AIC on the model, which left us with a final model with 11 predictors. With that final model, we began Propensity Score Matching. We explain this process theoretically in the Methods section and practically in the Results section. The resulting dataset contains data on pairs of observations that have a similar propensity to win. After we have all of the matched observations in one dataset, we construct a regression model on this dataset.

Our results indicate that among pairs of matches that had the same probability of winning, if he aces his opponent at least once, he has an 86% higher chance of winning that same match than if he hadn't. This implies that achieving an ace had a significant effect on his success in winning games.

Weaknesses and Next Steps

A paper by Gary King and Richard Nielsen claim that PSM increases model "imbalances, inefficiency, model dependance, and bias" [3]. Indeed, King argues that one of the disadvantages of PSM is that it requires a very large sample population, with a lot of overlapping covariates between the treatment and control group [3]. While a significant amount of our observations had overlap, our sample population is quite small at 212 observations. It would seem that we suffer from this issue that King claims.

An important factor that was left out of the analysis was the serve and volley. It is normally the case that if Isner plays against a high level opponent, they are capable of returning Isner's serve. This return usually forces the ball into a high arc. As a result, his natural counterplay is to serve hard, then run to the net to volley or smash that return. Our model does not account for this as a technique for Isner to win his games. It is recommended that future studies attempt to find a correlation between First Serve Points Won and winning a match.

Appendix

Supplementary Materials

Below is a glimpse of the raw match and score datasets

```
## # A tibble: 6 x 54
##    tourney_order match_id  match_stats_url~ match_time match_duration winner_aces
##            <dbl> <chr>     <chr>            <time>              <dbl>       <dbl>
## 1              0 1991-730~ /en/scores/1991~ 01:20                  80           3
## 2             0 1991-730~ /en/scores/1991~ 01:29                  89           1
## 3             0 1991-730~ /en/scores/1991~ 00:55                  55           0
## 4             0 1991-730~ /en/scores/1991~ 01:09                  69           2
## 5             0 1991-730~ /en/scores/1991~ 01:29                  89           6
## 6             0 1991-730~ /en/scores/1991~ 01:49                 109          10
## # ... with 48 more variables: winner_double_faults <dbl>,
## #   winner_first_serves_in <dbl>, winner_first_serves_total <dbl>,
## #   winner_first_serve_points_won <dbl>, winner_first_serve_points_total <dbl>,
## #   winner_second_serve_points_won <dbl>,
## #   winner_second_serve_points_total <dbl>, winner_break_points_saved <dbl>,
## #   winner_break_points_serve_total <dbl>, winner_service_points_won <dbl>,
## #   winner_service_points_total <dbl>, winner_first_serve_return_won <dbl>, ...
```

```
## # A tibble: 6 x 24
##   tourney_year_id tourney_order tourney_slug tourney_url_suffix tourney_round_n~
##   <chr>                   <dbl> <chr>        <chr>              <chr>
## 1 1991-7308                   1 adelaide     /en/scores/archiv~ Finals
## 2 1991-7308                   1 adelaide     /en/scores/archiv~ Semi-Finals
## 3 1991-7308                   1 adelaide     /en/scores/archiv~ Semi-Finals
## 4 1991-7308                   1 adelaide     /en/scores/archiv~ Quarter-Finals
## 5 1991-7308                   1 adelaide     /en/scores/archiv~ Quarter-Finals
## 6 1991-7308                   1 adelaide     /en/scores/archiv~ Quarter-Finals
## # ... with 19 more variables: round_order <dbl>, match_order <dbl>,
## #   winner_name <chr>, winner_player_id <chr>, winner_slug <chr>,
## #   loser_name <chr>, loser_player_id <chr>, loser_slug <chr>,
## #   winner_seed <chr>, loser_seed <chr>, match_score_tiebreaks <chr>,
## #   winner_sets_won <dbl>, loser_sets_won <dbl>, winner_games_won <dbl>,
## #   loser_games_won <dbl>, winner_tiebreaks_won <dbl>,
## #   loser_tiebreaks_won <dbl>, match_id <chr>, match_stats_url_suffix <chr>
```

Here is a glimpse of the final dataset after Propensity Score Matching

```
##   isner_win_loss aces double_faults first_serves_in first_serves_total
## 1              0   14             2              61                 89
## 2              0    9             1              67                 91
## 3              0   17             4              71                108
## 4              0    7             1              54                 86
## 5              0   17             1              91                133
## 6              0   17             5              88                121
##   first_serve_points_won second_serve_points_won first_serve_return_won
## 1                     43                       8                     12
## 2                     37                       7                      2
## 3                     42                      19                     12
## 4                     38                      16                      9
## 5                     63                      22                     11
## 6                     60                      15                     14
##   first_serve_return_total second_serve_return_won second_serve_return_total
## 1                       65                       8                        25
## 2                       30                       8                        25
## 3                       59                      11                        28
## 4                       53                      11                        30
## 5                       65                      13                        46
## 6                       71                      10                        36
##   service_games_played return_games_played      .fitted cnts
## 1                   14                  15 8.904585e-08    1
## 2                  12                  11 1.241874e-06    1
## 3                  15                  14 3.139980e-05    1
## 4                  15                  15 5.950273e-05    1
## 5                  18                  18 6.166897e-05    1
## 6                  19                  19 6.997469e-05    1
```

Bibliography

1. Association of Tennis Professionals, *John Isner: Overview*, ATP World Tour, Retrieved December 19, 2021, from https://www.atptour.com/en/players/john-isner/i186/overview
2. Lin, K.; Kuo, A.J.; *ATP World Tour tennis data*, datahub.io, Retrieved December 16, 2021, from https://datahub.io/sports-data/atp-world-tour-tennis-data#readme
3. King, G.; Nielsen, R. (2019), *Why Propensity Scores Should Not Be Used for Matching*, Political Analysis, Retrieved December 19, 2021, from https://gking.harvard.edu/files/gking/files/psnot.pdf
4. Rohan, A. (2020), *Running Through a Propensity Score Matching Example*, STA304 Lectures, Retrieved December 16, 2021, from https://q.utoronto.ca/courses/236142/files/17304472?wrap=1