ETC 3410 - Research Project

Conor Ryan, Hui Zhang, David Kontrobarsky, Zheng Junfeng

THE IMPACT OF PHYSICAL CHARACTERISTICS IN DETERMINING TENNIS MATCH OUTCOMES

Research Questions and Motivations

Tennis is an international sport played by people all around the world. There are a huge amount of variables that define each game of tennis that range from player characteristics to point by point data. As a result, modelling and predicting the outcome of tennis matches has attracted a lot of attention in recent times. Many of the studies on the subject use massive amount of variables and complex machine learning models in order to obtain the greatest predictive power, and so these models quickly become a black box. These models are difficult to interpret for marginal effects and for determining reliable causality.

This study will instead focus on the marginal effects of certain player characteristics on the outcome of grand slam tennis matches. The aim of this study is to explore the significance of the physical characteristics such as height and age in changing the outcome of grand slam tennis matches. Height has a strong correlation with other important physical characteristics such as strength and power that play significant roles in a match of tennis. The hope is that height is able to capture these difficult to observe effects. However, height in it's own right may be expected to be beneficial due to a longer arm span that allows reaching balls easier.

Furthermore, we are interested in determining whether the magnitude, sign and significance of the marginal effects of height and age vary according to gender, court surfaces and length in time of the match. For example, different court surfaces have different characteristics that may benefit or hinder taller players. In determining whether these physical characteristics are significant, we can provide greater insight into the game of tennis and also assist players and coaches in determining what tournaments they are likely to have the most success in based on their individual characteristics and the characteristics of matches. Ultimately, we find that height that is a significant factor and the effects are more obvious when stratifying by court surface and gender.

Background Literature Review

Current research regarding the predictive ability of match statistics to predict the outcome of tennis matches quite frequently uses Markov chains as their model base. This is due to the nature of tennis that it has a great dependence on past performance. For example, if a tennis player loses a set, having that information will affect their decision making going forward in the match. A paper that aimed to develop a model which predicted the outcome of the match using Markov chains, Brown A. Klaassen and Magnus (2001) tested the use of Markov chains in tennis as a predictive model, determining the independence of points throughout a match. Their results conclude a partial relationship between set-by-set points. However, the Markov chain model is not in our best interest as it is used when predicting the outcome of the match on a point-by-point basis; whilst our model is attempting to test the causal effects of a difference in characteristics and performance to determine the winner of the match.

Barnett (2005) also uses Markov chains to use a player's statistics of the game to predict the outcome accurately. What they fail to take into consideration is the surface of the court and its relationship with performance of particular players. They mention the effects of controlling for surface and how that could omit some unobserved variables of the player's court preference from skewing the results of the data. Knottenbelt (2012) wrote another paper using Markov chains, particularly using the statistics of the game to alter the model by accounting for previous common games between the two players. The conclusion of the paper was that the statistics of previous games has effective predictive power in determining the outcome of the match.

Much like Knottenbelt (2012), other literature on predictive models of a sport such as tennis mainly focus on the application of such models with betting; using the predictive models in determining the probability of a certain player winning the match to improve the odds of winning a bet. Therefore, a large proportion of the literature of modeling a match of tennis using game statistics follows on the basis of profit; not for determining the causality of certain variables. What these papers do conclude, is that a probit or a logit model are effective in producing a regression that predicts match outcomes strongly.

One main topic that our group has considered was the greater implications of height as a correlated variable with our other variables. Determining how height affects a players

performance and its correlation with serving performance in particular would be of great interest. Frantisek V(2008) looks into the effect of height on serving performance. Our expected results when using height as a focal variable is that it does in fact correlate with serve performance. These observed correlations are possibly due to the better angle of the serve that taller players can achieve and greater kinematic impact of the ball with an increase in body height. This analysis was calculated on both the first and serve serves, which confirmed that an increase in height has a positive relationship with serve performance.

Description of the Dataset

Tennis match data was obtained from the Github of author and software developer, Jeff Sackman, who works in the field of sports statistics. He is the creator of Tennis Abstract, a website for tennis analytics and is also published on Heavy Topspin. This dataset contains match data from the start of the open era (1968) to September 2018 for both men's and women's matches. The available variables included the height, age and rank of the winner and loser as well as in-match data such as number of aces, double faults and 1st serve points one. The data chosen for analysis came from 2011 onwards and only grand slam matches were selected. We believe this is a sufficient amount of data and the computational power available would easily be able to handle it. Furthermore, since grand slams are the most important tournaments and have the most elite players of any tournament, we will be able to make more confident conclusions about elite tennis. Overall, the data contains 3895 matches, with 913 of those being women's matches.

The variables in the original data were labelled for the winner and loser, so for each match player 1 and player 2 were randomly assigned to each player. This made it possible to create a binary dependent variable indicating whether player 1 has won or lost. Since there was a winner statistic and loser statistic for many of the metrics, these were combined into one variable as the difference between player 1 and player 2 for each statistic. This gives us an idea of how players match up against each other and also vastly reduces the amount of variables we need to work with. A description of the variables can be found in table 10 in the appendix.

The match win proportion was calculated for each player in the data and the included variable was the difference between these proportions. Finally, the log difference in rank is calculated in order to account for the non linearity of the relationship between ranking and player quality. In terms of data cleaning, matches with player heights missing and retirements were removed as

retirements are often due to injury and are likely to bias the results. A brief description of the variables is in Table 1.

Table 11 shows the correlation matrix between the variables. The difference in log rank appears to be highly correlated with multiple variables in the data so it will not be included in the analysis. Otherwise, there does not appear to be a significant amount of multicollinearity in the data as shown by the Variance Inflation Factors (VIF) all of which are less than 2 while the difference in log rank had a VIF of 11, which is considered quite high. Due to the players randomly being assigned to player 1 and player 2, this greatly reduces the likelihood of endogeneity in the variables.

Summary statistics for the variables in our data can be found in figure 1 in the appendix. The calculated values are the average for the winners over the entire dataset. For example, the mean of -32 for rank difference means that on average the winner of the match was ranked 32 places higher than their opponent which is not surprising.

Methodology

In this project, the logit model will be estimated using maximum likelihood under the assumption that the error term, v_i, follows a logistic distribution. Under the logit model, the predicted probability that player 1 wins the i^{th} match conditional on the predictor variables is:

$$E(Player1win_i) = \Lambda(\beta_0 + \beta_1 Heightdiff_i + \beta_2 rankdiff_i + \beta_3 agediff_i + \beta_4 rankpointsdiff_i + \beta_5 diffwinproportion_i + \beta_6 1stservewondiff_i + \beta_7 diffaces_i + \beta_8 2ndservewondiff_i + \beta_9 1stserveindiff_i) \quad (1)$$

The dependent variable is a binary variable so the predicted value is a probability between 0 and 1. This makes OLS inappropriate for the problem at hand since it's predictions are not bounded between 0 and 1. Although the coefficients of the probit and logit models will be different, the marginal effects will be almost identical when evaluated at the mean or median of the variables. However, differences would be present in the tails of these distributions. Since the focus of this study is marginal effects of each of the variables, the choice of model is of little importance. Therefore, we have chosen to use the logit model due to it's more common use in sport research

We do not use any quadratic or log transformations in our models. The reason for this is that

all the variables have negative values so squaring these variables would make them positive and taking the log of these variables is not possible mathematically speaking.

Our baseline model involves estimating equation (1) with all the available data and all the variables (except for difference in log rank due to reasons explained in the data section). Several of the variables in our data; court surface, gender and length of match will be used for stratification. That is, the data will be subset and separate logit regressions will be fit in order to see if the significance and marginal effects of the difference variables e.g height difference, age difference etc change according to gender, court surface and the length of the match. The large sample size of our data ensures hypothesis tests will have sufficient power even when the data is subset. We also include multiple control variables such as difference in ranking, difference in rank points and difference in win proportion in order to account for differences in skill among players. In this way, we can make more reliable causal statements about the marginal effect of height and age. These hypotheses will be tested using the z test for coefficients and Wald tests.

Results

The results for the logit model when estimating equation (1) can be found in the appendix (table 1).

From first glance at the sign of these coefficients, it seems as though most of the included variables have a positive effect in increasing a player's chances of winning the game. However, the difference in 1st serves in, difference in height and rank difference all have negative signs. These results suggest that overall, being taller than your opponent could actually decrease a players chances of winning a match. Furthermore, the coefficient for ranking difference suggests that a lower ranked player has a lower chance of winning which of course is expected.. We will need to investigate whether this effect is actually significant and to also test the individual significance of the other coefficients using a z test.

The z test statistic is standard normally distributed and will test the null hypothesis that each regression coefficient is equal to zero against the alternative that they are not equal to zero at a 5% significance level. The critical value is 1.96 for a two tailed test and we reject the null

hypothesis if the t statistic is greater than 1.96. Summary of the results can be found in the appendix (table 2).

Therefore, all coefficients appear to be significant at the 5% level except for the age difference and rank difference. Interestingly, difference in rank does not seem to be significant in determining the probability that a player will win. This is perhaps due to the fact that this metric assumes that equal difference in rank indicates equal difference in player ability, which is not the case. For example, the difference in ability between the 1st ranked player and the 10th ranked player is much greater than the difference in ability between the 50th ranked player and the 60th ranked player. Also, difference in ranking does seem to be significant.

Now that the significance of the variables has been ascertained, the marginal effects which is the main focus of this study can be investigated. In logit models, the marginal effects of regressors are not directly given by the coefficients, this is because the model is non linear which means that the gradient is dependant on which point you are on the curve. The preferred method is to assume the median values for the discrete variables and the mean for the continuous variables. Theoretically speaking, we would expect these values to be zero or very close to zero due to each player having been randomly assigned player 1 or player 2.

We can calculate the marginal effects on regressors with the response variable from this equation. The marginal effect of the difference in height (assumed to be a continuous regressor) between two players can be estimated as:

$$\frac{\partial\, \widehat{Player1win}}{\partial\, heightdiff}$$

$$= -0.0233388 \times \lambda(0.2716804 - 0.0233388\, \widehat{heightdiff}$$
$$- 0.0006561\, \widehat{rankdiff} + 0.0020341\, \overline{agediff}$$
$$+ 0.000067\, \widehat{rankpointsdiff} + 5.083102\, \overline{diffwinproprtion}$$
$$+ 0.1931885\, \widehat{1stservewondiff} + 0.0450903\, \widehat{diffaces}$$
$$+ 0.2051032\, \widehat{2ndservewondiff} - 0.0070844\, \widehat{1stserveindiff})$$

$$\frac{\partial \widehat{Player1win}}{\partial\, heightdiff}$$

$$= -0.0233388 \times \lambda(0.2716804 - 0.0233388 \times (0) - 0.0006561 \times (2) + 0.0020341 \times (0.065234) + 0.000067 \times (-29) + 5.083102 \times (-0.0062137) + 0.1931885 \times (0) + 0.0450903 \times (-0) + 0.2051032 \times (0) - 0.0070844 \times (38))$$

$$\frac{\partial \widehat{Player1win}}{\partial\, heightdiff} = -0.0233388 \times \lambda(-0.03223417842) = -0.0233388 \times 0.2499350711$$
$$= -0.005833184637$$

The marginal effect of height difference shows that controlling all other independent variables, for a unit value increase in the height difference between player 1 and player 2, the probability that player 1 wins the game is estimated to decrease by about 0.006, for a match with characteristics as described above. This prediction is valid since height difference is significant at 5% significance level. As we discussed before, taller players are less agile and generally move around the court at a slower pace even though they have a better angle for serving and more power to hit the ball. This would be the reason for a negative effect of height difference.

We also calculate the marginal effect of the difference in win proportion between two players as:

$$\frac{\partial \widehat{Player1win}}{\partial diffwinproportion} = 1.27$$

The marginal effect of the difference in win proportion shows that for a 1 percentage point increase in the difference in win proportion between player 1 and player 2, the probability that player 1 wins the game is estimated to increase by about 1.27 percentage points where the match otherwise assumes the mean and median values, as described previously.

We also estimate the marginal effects of the difference in first serve points won between player 1 and player 2, the difference in second serve points won between player 1 and player 2, the difference in rank points between player 1 and player 2, the difference in the number of aces between player 1 and player 2 and the difference of the amount of first serves in between player 1 and player 2. We will only estimate the marginal effects of significant regressors.

Let P_1 be the probability of player 1 winning the match when they have won 1 more first serve than their opponent and all other variables are at their median or mean.

$$P^1 = \Lambda(0.2716804 - 0.0233388 \times (0) - 0.0006561 \times 2 + 0.0020341 \times 0.065234 + 0.000067 \times (-29) + 5.083102 \times (-0.0062137) + 0.1931885 \times 1 + 0.0450903 \times (0) + 0.2051032 \times (0) - 0.0070844 \times 38) = \Lambda(0.1609543216) = 0.5401519355$$

Let P_2 be the probability of player 1 winning the match when they have won 2 more first serve than their opponent and all other variables are at their median or mean.

$$P^2 = \Lambda(0.2716804 - 0.0233388 \times (0) - 0.0006561 \times 2 + 0.0020341 \times 0.065234 + 0.000067 \times (-29) + 5.083102 \times (-0.0062137) + 0.1931885 \times 2 + 0.0450903 \times (0) + 0.2051032 \times (0) - 0.0070844 \times 38) = \Lambda(0.3541428216) = 0.5876218406$$

Thus, the marginal effect of an additional point difference in first serve points won between player 1 and player 2 is estimated as:

$$ME_1^2(1) = P^2 - P^1 = 0.5876218406 - 0.5401519355 = 0.0474493$$

This means that winning two additional first serve points than your opponent is expected to increase the chance of winning by 0.047 as compared to winning 1 more first serve points than your opponent. The choice of 1 and 2 is arbitrary as their no particularly meaningful value for number of first serves won.

Since, the difference in rank between player 1 and player 2 is insignificant at the 5% significance level, there is no reason to calculate its marginal effect on the probability for player 1 winning. This is due to the fact that the variable does not directly affect the probability of player 1 winning and its marginal effect also would not be significant. The remainder of the marginal effects can be found in the appendix (table 8).

As discussed earlier, an area of interest is to see if these results are consistent when considering the court surfaces individually, that is, hard court, grass court and clay court. This was done by subsetting the data to each court surface and conducting multiple logit regressions. These results are available in the appendix of the report.

Table 3 suggests the hard court data has similarities to the overall regression. Height remains as a significant factor when determining a players likelihood of winning a game (t stat = -4.1)

and all other regressors that were significant in the initial regression are still significant. The negative coefficient on height difference suggests that being taller your opponent is potentially still a disadvantage on hard court. Interestingly, the difference in rank between the two players now becomes a significant regressor (t stat = 3).

For the other two court surfaces; grass and clay, the coefficient for height difference becomes insignificant which suggests height becomes less of an important factor on these two court surfaces as compared to hard court. For clay court, this makes sense. Clay is the slowest court surface of the three and so the advantage of being tall and having a powerful serve is largely taken away from players. However, for grass court the results are somewhat unexpected since grass traditionally favours a strong serve and a serve-and-volley game style. Also, although age becomes a significant regressor for the clay court regression, it is only at the 10% level and so could very well be up to random chance.

We are also interested in determining if there are any differences between male and female games in regards to the variables we are considering. The data is a subset for male and female games and two logit models are fit. The results can be found in the appendix.

For male games, the difference in height between players is still highly significant (t stat = -3.86) however for women's game it is only significant at the 10% significance level (t-stat = -1.75). It is important to mention that there are far less female games than male games which would potentially alter the significance of variables (2982 for males and 913 for females). Despite this, the sample sizes are large in both cases which means the results can probably be relied upon. These results may be due to the fact that women are generally less powerful than men and so being a taller women may not necessarily lead to being able to serve significantly more aces. This theory is backed up the scatter plots shown in figure 1 and figure 2. For mens matches, the positive correlation between aces and difference in height is clear while the womens data appears to be essentially random.

The final regression that will be conducted will involve adding an interaction term between height difference and the length of the match. The hope is that this regressor will be able to capture the lack of stamina that taller players may have in longer matches. The results for this regression are located in table 9.

The coefficient is significant at the 1% level (t-stat = 3) and it's sign is negative. The interpretation is that as the length in time of the match of the match increases, the players that are taller than their opponents are generally at a disadvantage. This is possibly due to reasons discussed earlier that taller players often have less stamina due to being heavier and this struggle to keep up in matches that last for long periods of time.

Limitations

One of the most prevalent limitations for our current regression of the dataset, is our lack of accountability for the effects of the unobserved variables within a match of tennis. Not every single match of tennis will record all variables associated with the physical characteristics in a game of tennis. One extremely important variable which isn't accounted for throughout our data is the advantage of home court and weather conditions and its effects on performance.

There are four major grand slam tournaments in the world, one being housed in Melbourne, Australia. Since the Australian open occurs during the summer months, the weather throughout the main duration of the tournament is fairly dry and hot. As Australia is in the Southern Hemisphere, when players from all over the globe come to compete at the Australian open, majority of the players are commuting from the Northern Hemisphere, where it would be winter during the same time period. Therefore, if a British player went from their home country conditions compared to the much hotter, drier conditions in Melbourne during January, that particular player would be at a disadvantage. Therefore, negatively affecting their performance in relation to other players; compared to players from Australia, who would have acclimatized to the weather conditions.

Other such unobserved variables that are omitted in our dataset are the difference in a players stamina and the difference in their experience throughout tournaments and clutch moments. Obviously these variables would have an effect on the probability of players winning the game, but are not taken into consideration in our regression.

Another considered limitation with our dataset, is the fact that our data only analyses the effects of this particular physical characteristics on a game of tennis in a particular era of tennis; our dataset is specifically from 2011 onwards. For a greater in depth analysis, one could introduce a panel analysis on the different eras/years to definitively determine the importance of each

physical characteristic. This greater pool of information would also increase the validity of our conclusion; as some characteristics may have a larger impact historically then what our data suggests. To further expand our dataset, more tournament types (less important tournaments) could be integrated into the existing dataset in order to be able to make conclusions about tennis in general rather than purely at the elite level. . Since the larger the sample size, the more reliable the results of the regression become due to more powerful tests.

An additional limitation of the data is the fact that only the difference in height of the two players was considered instead of the heights themselves. It seems likely that the effect of difference in height would be greater in some ranges of heights as compared to others which our model has not captured. Also, a comparison of singles and doubles matches in regards to important winning factors would have provided greater insight into the game.

Conclusion

The aim of this report was to test the causal effects of physical characteristics and other factors in determining the winner of a tennis match. After analysing the data initially, we had expectations that the rank difference would be a highly significant predictor as it would act as a proxy for the level of skill for each player. Since, that the rank difference is not linear in parameters and that the difference in skill between the 50th rank player and the 60th rank player will be less than the skill difference between the 11th best player and the 1st best player; it may have resulted in our testing concluding that rank difference is not a significant regressor, which was surprising.

It was also interesting to observe that there was a significant difference in the importance of height between surfaces, and how taller player would perform worse on average on the hard court, but were not all too significant on grass and clay courts. The majority of the other variables in the regression had all been as significant as expected; height difference, ranking points, win proportion, aces, first serves in, first serves won and second serves won. Age was determined to be insignificant, as there must have been a great spread of ages throughout different skill level; which makes sense since age isn’t too much of a factor until you reach the end of their career. The extreme significance of the difference in win proportion was expected, indicating that past performance is a large indicator for future performance in tennis.

Team contributions

Conor:

Writing up the literature review, by researching previous studies performed by professional analysts to assist in increasing our depth of the field we were analysing. Wrote up most of the limitations section of the final report, detailing the limitations of our motivation questions and dataset and how that would have skewed the data. Wrote the conclusion to the report, to summarise our findings after running the regression. Assisted in the formatting of the results section of the report.

Hui:

Selecting the econometric model (logit model), and explaining this econometric method. Writing some of the results, including the calculation and explanation of marginal effects.

David:

Preparing the data to be analysed by joining datasets from different years and gender, creating new variables (The difference variables and win proportion variable), deleting irrelevant variables. Writing the data description section and some of the results (Hypothesis tests, regressions) and motivations/research questions. Creating the regression tables and scatter plots that are in the appendix.

Jevon:

Researching the background of the topic and find out the research questions and motivations.

References

Frantisek Vaverka, (2008), The influence of Body height on the serve in tennis, University of Ostrava: Research gate, taken from: www.researchgate.net/Frantisek_vaverka/publication/251351599-THE-INFLUENCE-OF-THE-BODY-HEIGHT-ON-THE-SERVE-IN-TENNIS.pdf?origin=publication_detail

William J Knottenbelt, (2012), A common-opponent stochastic model for predicting the outcome of professional tennis matches, Computer and Mathematics with Applications, pages

3820-3827, taken from: www.sciencedirect.com/science/article/pii/S0898122112002106

Tristan Barnett and Stephen Clarke, (2005), Combining player statistics to predict outcomes of tennis matches, IMA Journal of Management Mathematics 16, 113-120, doi:10.1093/imnman/dpi001

Barnett T. Brown A. Clarke S., Developing a model that reflects outcomes of tennis matches, Faculty of Life and Social Sciences, Swinburne University, Melbourne, Taken From: www.strageticgames.com.au/8mcs.pdf

Michael Summers, (2011), Clay Vs. Grass: A Statistical Comparison of the French Open and Wimbledon, Pepperdine University Malibu, 9026, CA, Taken From: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.911.4217&rep=rep1&type=pdf

Appendix

Summary statistics

	Winner mean
Rank difference	-32
difference win proportion	0.168
Height difference	1cm
Age difference	-0.0562
Aces difference	2.77
Second serves won difference	2.07
First serves won difference	4.95
First serves in	0.29

Figure 1

Table 1: Initial Logit Regression Results

	Dependent variable:
	Player 1 Win Probability
Height difference	-0.023^{***}
	(0.006)
Rank difference	-0.001
	(0.001)
Age difference	0.002
	(0.011)
Ranking points difference	0.0001^{***}
	(0.00002)
Difference in win proportion	5.083^{***}
	(0.324)
Difference in aces	0.045^{***}
	(0.007)
Difference in first serves in	-0.007^{***}
	(0.002)
Difference in 1st serves won	0.193^{***}
	(0.008)
Difference in 2nd serves won	0.205^{***}
	(0.009)
Intercept	0.272^{***}
	(0.089)
Observations	3,895
Log Likelihood	$-1,201.335$
Akaike Inf. Crit.	2,422.671
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Initial Logit Z statistics

	Dependent variable:
	Player 1 Win Probability
Height difference	−0.023
	t = −3.992***
Rank difference	−0.001
	t = −0.977
Age difference	0.002
	t = 0.185
Ranking points difference	0.0001
	t = 2.754***
Difference in win proportion	5.083
	t = 15.688***
Difference in aces	0.045
	t = 6.094***
Difference in first serves in	−0.007
	t = −3.708***
Difference in 1st serves won	0.193
	t = 24.238***
Difference in 2nd serves won	0.205
	t = 21.660***
Constant	0.272
	t = 3.047***
Observations	3,895
Log Likelihood	−1,201.335
Akaike Inf. Crit.	2,422.671
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3: Logit regression for hard court data

	Dependent variable:
	Player 1 Win Probability
Height difference	−0.037***
	(0.009)
Rank difference	−0.003**
	(0.001)
Age difference	−0.019
	(0.016)
Ranking points difference	0.0001
	(0.00003)
Difference in win proportion	5.335***
	(0.476)
Difference in aces	0.039***
	(0.011)
Difference in first serves in	−0.009***
	(0.003)
Difference in 1st serves won	0.193***
	(0.012)
Difference in 2nd serves won	0.201***
	(0.013)
Constant	0.227*
	(0.131)
Observations	1,940
Log Likelihood	−579.594
Akaike Inf. Crit.	1,179.188
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Logit regression for clay court data

	Dependent variable:
	Player 1 Win Probability
Height difference	−0.012
	(0.012)
Rank difference	0.001
	(0.001)
Age difference	0.043*
	(0.023)
Ranking points difference	0.0002***
	(0.0001)
Difference in win proportion	4.835***
	(0.688)
Difference in aces	0.030
	(0.020)
Difference in first serves in	−0.006
	(0.004)
Difference in 1st serves won	0.208***
	(0.017)
Difference in 2nd serves won	0.210***
	(0.019)
Constant	0.403**
	(0.182)
Observations	967
Log Likelihood	−279.247
Akaike Inf. Crit.	578.493
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Logit regression for grass court data

	Dependent variable:
	Player 1 Win Probability
Height difference	−0.007
	(0.011)
Rank difference	0.0001
	(0.001)
Age difference	0.005
	(0.021)
Ranking points difference	0.00004
	(0.00004)
Difference in win proportion	5.059***
	(0.607)
Difference in aces	0.066***
	(0.013)
Difference in first serves in	−0.005
	(0.003)
Difference in 1st serves won	0.185***
	(0.015)
Difference in 2nd serves won	0.224***
	(0.020)
Constant	0.292*
	(0.170)
Observations	988
Log Likelihood	−325.100
Akaike Inf. Crit.	670.200
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Logit regression for male games

	Dependent variable:
	Player 1 Win Probability
Height difference	-0.027^{***}
	(0.007)
Rank difference	-0.001
	(0.001)
Age difference	0.009
	(0.013)
Ranking points difference	0.0001^{***}
	(0.00003)
Difference in win proportion	5.175^{***}
	(0.399)
Difference in aces	0.041^{***}
	(0.008)
Difference in first serves in	-0.020^{***}
	(0.003)
Difference in 1st serves won	0.195^{***}
	(0.009)
Difference in 2nd serves won	0.201^{***}
	(0.011)
Constant	1.073^{***}
	(0.189)
Observations	2,982
Log Likelihood	-887.762
Akaike Inf. Crit.	1,795.524
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Logit regression for female games

	Dependent variable:
	Player 1 Win Probability
Height difference	-0.028^{*}
	(0.016)
Rank difference	0.0004
	(0.002)
Age difference	0.001
	(0.027)
Ranking points difference	-0.0001
	(0.0001)
Difference in win proportion	5.001^{***}
	(0.819)
Difference in aces	-0.078^{*}
	(0.040)
Difference in first serves in	-0.296^{***}
	(0.028)
Difference in 1st serves won	0.652^{***}
	(0.053)
Difference in 2nd serves won	0.278^{***}
	(0.030)
Constant	-0.095
	(0.139)
Observations	913
Log Likelihood	-177.987
Akaike Inf. Crit.	375.975
Note:	*p<0.1; **p<0.05; ***p<0.01

(Table 8)	Marginal Effects
The difference in second serve points won	$ME_1^2(2) = P^2 - P^1 = 0.0502739827$
The difference in rank points	$ME_1^2(2) = P^2 - P^1 = 0.0000167467$
The difference in aces	$ME_1^2(2) = P^2 - P^1 = 0.011268567$
The difference in first serves in	$ME_1^2(2) = P^2 - P^1 = -0.0017486059$

Table 9 Logit regression With minutes interaction variable

	Dependent variable:
	Player 1 Win Probability
Height difference	0.029
	(0.018)
Rank difference	−0.0004
	(0.001)
Age difference	−0.001
	(0.012)
Ranking points difference	0.0001**
	(0.00003)
Difference in win proportion	5.351***
	(0.374)
Difference in aces	0.043***
	(0.008)
Added minutes interaction variable	−0.022***
	(0.003)
Difference in 1st serves won	0.200***
	(0.009)
Difference in 2nd serves won	0.203***
	(0.010)
minutes	0.009***
	(0.002)
Height_diff:minutes	−0.0003***
	(0.0001)
Constant	−0.450**
	(0.193)
Observations	3,234
Log Likelihood	−968.715
Akaike Inf. Crit.	1,961.429
Note:	*p<0.1; **p<0.05; ***p<0.01

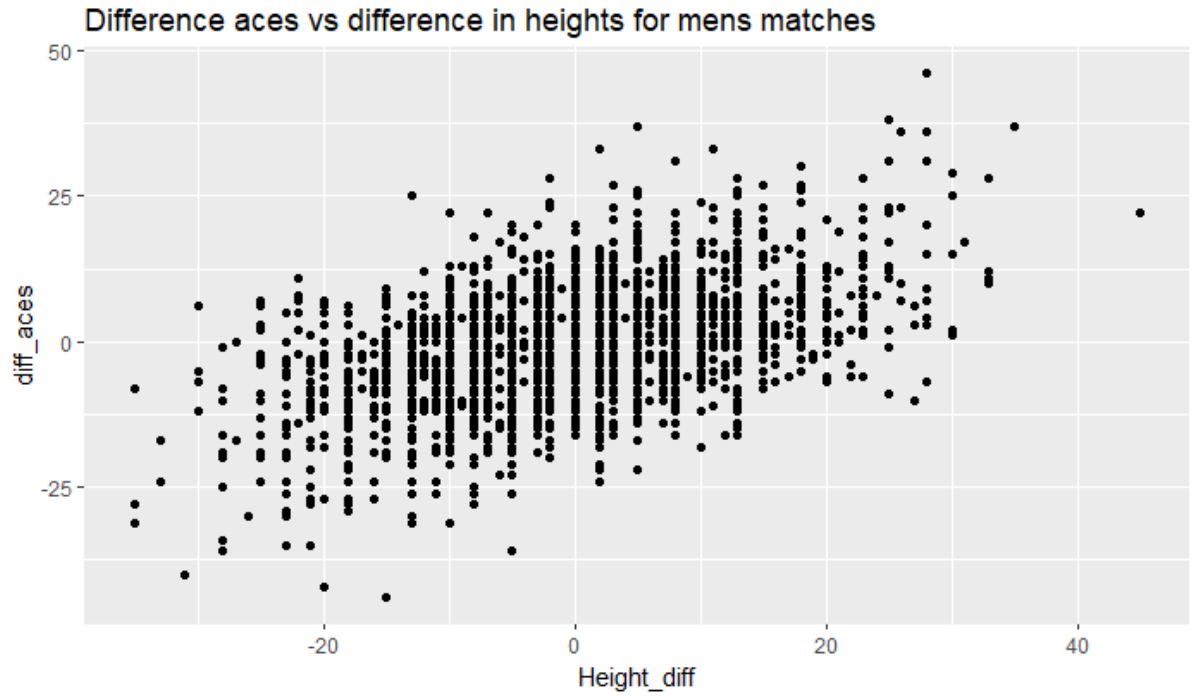

Figure 1

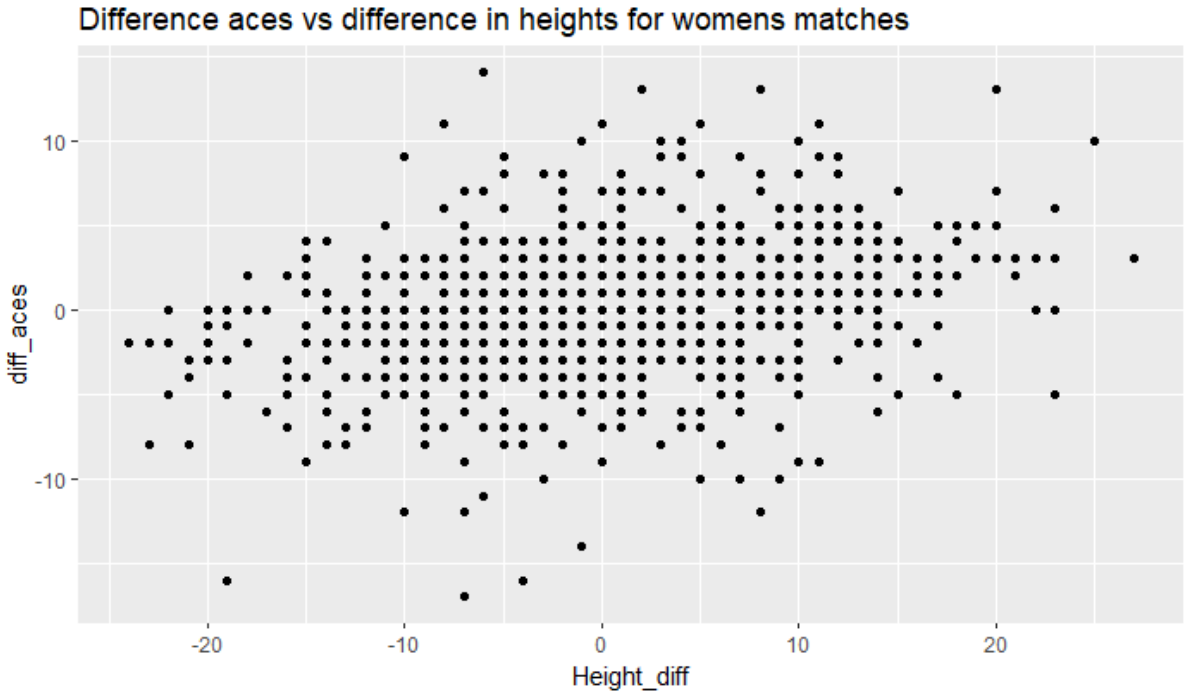

Figure 2

Variable	Description
tourney_name	Name of the tournament the match was played in. One of US open, Australian Open, Wimbledon or Roland Garros.
surface	The court surface that the match is played on
minutes	The length of the match in minutes
Player1	Name of the 1st player
Player2	Name of the 2nd player
round	The round that the match is played in (R128, R64 etc)
Player1_win	A binary variable indicating whether or not the 1st player won the match
Height_diff	The difference in height between player 1 and player 2
rank_diff	The difference in rank between player 1 and player 2
age_diff	The difference in age between player 1 and player 2
Rank_points_diff	The difference in ranking points between player 1 and player 2
diff_win_proportion	The difference in win proportion between player 1 and player 2 for matches in the data
1stservewon_diff	The difference in 1st serve points won between player 1 and player 2
diff_aces	The difference in number of aces between player 1 and player 2
2ndservewon_diff	The difference in 2nd serve points won between player 1 and player 2
1stservein_diff	The difference in number of first serves in between player 1 and player 2
Diff_logrank	The difference in log rank between player 1 and player 2
Gender	Men's or women's match

Table 10

	Height_diff	rank_diff	age_diff	rank_points_diff	diff_win_proprtion	firstservewon_diff	diff_aces	secondservewon_diff	firstservein_diff	diff_logrank
Height_diff	1									
rank_diff	-0.09	1								
age_diff	-0.12	0.03	1							
rank_points_diff	0.05	-0.41	0.02	1						
diff_win_proprtion	0.18	-0.47	-0.01	0.66	1					
firstservewon_diff	0.16	-0.19	0	0.28	0.35	1				
diff_aces	0.47	-0.22	-0.08	0.16	0.34	0.28	1			
secondservewon_diff	-0.07	-0.09	-0.04	0.12	0.16	-0.26	0.01	1		
firstservein_diff	-0.05	0	0	0.02	-0.02	0.19	-0.05	0.02	1	
diff_logrank	-0.11	0.71	0	-0.87	-0.76	-0.31	-0.27	-0.14	0	1

Table 11