ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Международный институт экономики и финансов

Ломасов Илья Геннадьевич

ЭФФЕКТИВНОСТЬ РЫНКА СТАВОК НА ATP ТЕННИС
(ATP TENNIS BETTING MARKET EFFICIENCY)

Выпускная квалификационная работа - БАКАЛАВРСКАЯ РАБОТА
по направлению подготовки 38.03.01 «Математика и Экономика»
образовательная программа «**Программа двух дипломов по экономике НИУ ВШЭ и Лондонского университета**»

Рецензент
PhD
В.Н. Соколов

Научный руководитель
PhD
В.Н. Соколов

Москва 2022

Table of Contents

Abstract

This paper focuses on testing the empirical relation between fundamental data available at the beginning of ATP tennis matches and the immediate pre-match odds-implied win probabilities of participants. In other words, it models the behavior and conjunctures of betting market on aggregate, depending on the information about players. The empirical findings (strong statistical evidence) suggest that there is a number of parameters that influences bettors' aggregate decision making about odds, and that this model performs well on the out-of-sample data. Odds that are given by betting companies oftentimes differ from the pre-match odds that are almost exclusively determined by supply and demand. Therefore, the present paper can serve as a basis for further research that can utilize the predictions of immediate pre-match odds given by the model in order to form profitable betting strategies that involve early betting, based on bookmaker-given odds.

Данная работа тестирует эмпирическую зависимость между техническими данными об участниках теннисных матчей под эгидой ATP и вероятностями победы того или иного участника, извлеченными из данных о коэффициентах на победу, которые были установлены на момент начала матча. Другими словами, данная работа моделирует агрегированное поведение и мнения игроков рынка ставок об исходе матчей в зависимости от информации об участниках матча. Обнаруженная зависимость предусматривает существование (подкрепленное строгими статистическими уликами) ряда параметров, которые определяют принятие решений игроками рынка ставок. Также было обнаружено, что модель хорошо показывает себя на данных, не входящих в выборку, на которой эта модель строится. Коэффициенты, которые букмекерские компании дают на момент анонса матча, часто отличаются от коэффициентов, которые устанавливаются к началу матча путём калибровки спроса-предложения. В связи с этим, данная работа может послужить базисом для будущих исследований, которые могут использовать предсказания коэффициентов на момент начала матча, данные моделью из этой работы, для формирования прибыльных стратегий, которые используют коэффициенты на момент анонса матча.

1. Introduction

Betting market is a large industry, and, needless to say, it is always profitable in the long run due to the principle that betting companies use in order to form betting odds[i]. Odds-setting process is a dynamic game with bookmaking companies and bettors as players. There is a very large number of stages in this game. At the first stage, betting companies use fundamental data to set their "starting odds" (this stage usually starts a couple hours after the match is announced). Then, player 1 (the first bettor) sees them, chooses a company and places a bet, worth some amount of money, on one of the outcomes. At the next stage, the chosen company re-adjusts its coefficients, taking into consideration the choice of the person at the previous period (both the outcome and the amount). Then the next bettor sees the changes and faces the same choice problem as bettor 1, but with updated betting odds. This part of the game goes on and on and only ends when the match starts, and the game transitions into its next part. Companies now have to take into consideration how the match is going. At the end of the game, all bettors receive the amount they bet multiplied by the coefficient at the moment of betting if they win and zero if they lose. It becomes apparent that the more people place bets, the less the updated odds depend on fundamentals and the more they depend on supply and demand. By the end of the first part of the game (by the time the match starts), odds are almost 100% determined by the market and not by fundamentals.

What makes betting companies profitable is the fact that odds that they offer are lower than the "fair" ones at each stage of this game. In the context of this paper the odds that are considered are the decimal ones (i.e., if the odd on a particular event is x, it means that if a bettor places a bet of size b then the payout is $x * b$, and net of that is $b * x - x = b * (x - 1)$. The mechanism is as follows. Denote odds on players A and B as q_A and q_B respectively. Suppose, so far the company was able to determine that market (or themselves if this is the first stage of the dynamic game) believes that probabilities of players A and B to win are p_A and p_B respectively. Then in order to break even, they need to set $q_A = 1/p_A$ and $q_B = 1/p_B$, which would imply ${}^1/_{q_A} + {}^1/_{q_B} = 1$. However, at any given stage, bookmakers set their odds in such a way that ${}^1/_{q_A} + {}^1/_{q_B} = 1 + v,\ v > 0$. And v, called vigorish or simply vig, is an indirect fee that bookmaker charges a bettor for placing a bet. It is how they make profit.

The main research questions that this paper will try to answer is what fundamental parameters bettors consider while choosing their favorite for a specific match. A person who has never watched tennis before might look at the matchup and predict the player with higher

rank to win. As del Corral & Prieto-Rodriguez (2010)[ii] found out, rank difference is a good and significant predictor for results of tennis matches of the highest level (Grand Slam) both among men (ATP) and women (WTA). However, there are countless examples where a player with higher rank is a bookmaker's underdog, and sometimes even with a considerable spread. This is where many different other factors kick in, for example, form, surface preferences, tournament level, level of tiredness, home advantage and many others. This paper is focused particularly on testing significance of a number of such parameters both individually and jointly. If the model performs well, it could potentially predict the model-consistent values of betting odds, which could be then used in order to determine if a particular bet is underpriced or overpriced.

2. Literature Review

There have been many papers related to the topic of sports prediction and betting. They can be broadly divided into two classes, the larger one concerning efficiency of betting markets, and the smaller one concerning forecasting. The first class, an extensive literature survey on which can be found in works of Sauer (1998[iii], 2005[iv]) and Vaughan Williams (1999[v], 2005[vi]), deals with efficiency in context of profitability of bets. Much attention is given to price formation and market analysis, which is looked at from a very different perspective than that in the present paper. The second class, more closely related to this paper, concerns characteristics of sport forecasts, which capture the relation between betting odds and match result (Strumbelj, 2014[vii] and Abinzano et al, 2019[viii]), within-match parameters such as number of breaks lost in tennis (Easton & Uylangco, 2010[ix]), and fundamental parameters such as ranking difference (Del Corral & Prieto-Rodriguez, 2010). The latter paper provides an extensive survey on forecasting literature, which broadly differs by the type of prediction (match winner, or point difference, or winner of tournaments and leagues), methodology used (statistical models or experts evaluation), or sports in question (team or individual, and particular sports). The two classes are related in that findings from papers about market efficiency can be used to derive results about forecasting, as argued by Stekler et al, 2010[x]. Unlike both of these classes, although related, the present paper tries to analyze the factors (fundamental sports-related variables) that bettors consider important for making bets or forecasts, rather than predicting the outcome of a particular match or tournament.

Of all sports, this paper focuses on tennis. The literature on tennis in particular contains papers connecting outcomes and difference in ranking (Boulier & Stekler, 1999[xi], Clarke & Dyte,

2000[xii], Klaasen & Magnus, 2003[xiii], Del Corral & Prieto-Rodriguez, 2010), who found that this difference is indeed a good predictor of the result. The latter factor also compares difference across genders, the issue that is side stepped in this paper by only considering male tournaments. Easton & Uylangco (2010) model the effect of within-game parameters (such as losing a break) on the overall result, and compare the predictions with the changes in betting odds during a match. Forrest & Mchale (2008)[xiv] analyze tennis in context of biases and market efficiency, paying particular attention to the so-called longshot bias, a well-known phenomenon where in extreme-odds cases too little is bet on favorites, and too much is bet on underdogs. As this papers focuses on the connection between fundamental tennis-related variables and betting rather than outcomes, the topics mentioned above are side stepped as well. The present paper is thus quite unique in the purpose of its analysis. It uses the following parameters to test their relation to betting.

3. Methodology

The main purpose of this chapter is to provide a solid theoretical foundation for the estimation and testing the main hypotheses of this paper. This chapter is divided into two subparts. The methodology for both parts is very similar, and they are arranged in the chronological order. The first part is focused on estimating the relationship between how bettors perceive player's probabilities of winning (dependent variable) and fundamental data that they have by the start of the match (explanatory variables). The second part is focused on testing if the out-of-sample data fits into the trained model with small margin of error. All the calculations are performed using *RStudio*.

3.1. Part 1 (estimation, parameter choice)

For this part the main instrument is a linear regression. It is applied to the whole dataset in order to obtain the empirical relationship between parameters as a whole. (*lm* in R)

$$y = X + \epsilon, \qquad where: \;\; y = (y_1\, y_2 \ldots y_n)^T, \qquad X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix},$$

$$\beta = (\beta_1 \ldots \beta_k), \qquad \epsilon = (\epsilon_1 \ldots \epsilon_n)$$

$$E(\epsilon_i) = 0, \;\; E(\epsilon_i^2) = \sigma^2, \;\; E(\epsilon_i) = 0$$

Assumption that has to be tested is homoskedasticity assumption:

$$Var(\epsilon) = I_n * \sigma^2, \qquad where\ I_n - n \times n \;\; identity\ matrix$$

For that reason, Breusch-Pagan test and White test are conducted.

1. Breusch-Pagan test (Breusch & Pagan, 1979)[xv]: (*lmtest::bptest* in R)

   H_0: *Homoscedasticity is present (the residuals are distributed with equal variance)*

   H_1: *Heteroscedasticity is present (the residuals are not distributed with equal variance)*

   The test is conducted in 3 steps. The first step is applying OLS model and computing squared residuals. The second step is computing Maximum Likelihood estimate of the error variance from the Step 1 regression, and dividing the residuals by it, calling this new variable g. Step 3 is applying OLS again, with g as dependent variable and the original matrix of regressors as explanatory, and then computing total sum of squares (TSS) and explained sum of squares (RSS). The resulting test statistic is:

$$LM = \frac{1}{2}(TSS - RSS) \sim \chi^2(k-1) \ \{under\ H_0\}$$

2. White test (White, 1980)[xvi]: (*skedastic::white_lm* in R)

   H_0: *Homoscedasticity is present*

   H_1: *Heteroscedasticity is present (any type)*

   This test is also conducted in 3 steps. Step 1 is applying OLS model and computing squared residuals (the same as in Breusch-Pagan). Step 2 is regressing the obtained squared residuals on the explanatory variables, their squares and their cross-products, omitting any duplicative variables (for example, if there is a dummy variable, then its square coincides with the dummy variable, and it is duplicative). The result is White auxiliary equation. Some variables and cross products may be skipped if there is perfect collinearity or insufficiency of the number of degrees of freedom. Step 3 is testing the equation for joint significance using a standard F-test for joint significance or alternatively a χ^2-test.

$$\chi^2_{st} = n * R^2 \sim \chi^2(q)\ \{under\ H_0, asymptotically\}$$

   Where $n -$ sample size, $R^2 -$ goodness of fit of White auxiliary equation, $q -$ number of regressors in White auxiliary equation.

If heteroscedasticity is detected, standard t-tests for individual significance of explanatory variables and F-test for joint significance become invalid. In this case heteroscedasticity consistent standard errors for coefficients are obtained and t-test is performed using them (*lmtest::coeftest* in RStudio).

Instead of standard F-test, Wald[xvii] test is used (*lmtest::waldtest* in R):

H_0: *all regressors are jointly insignificant* $(X\beta = 0)$

H_1: *regressors are jointly significant* $(X\beta \neq 0)$

$$WR = \left(X\hat{\beta}\right)^T [X(X^TX)^{-1}(X^Tuu^TX)(X^TX)^{-1}X^T]\left(X\hat{\beta}\right) \sim \chi^2(k)\ \{under\ H_0, asympt.\}$$

where $u -$ vector of residuals.

After that, the model with the smallest p-value of WR statistic is chosen and used for explanation. Note that even if heteroscedasticity is detected, $\hat{\beta}$ is a valid estimate for β, since it is both unbiased and consistent, and n is quite large in this paper.

3.2. Part 2 (Repeated 2-fold cross-validation)

This part works with the model that was chosen in Part 1 of this section. After that, the dataset is divided randomly into train and test sets (with equal number of observations). Then the chosen model is fitted on the train set using OLS: $y_{tr} = X_{tr}\beta_{tr} + \epsilon_{tr}$, and the coefficients $\widehat{\beta_{tr}}$ are obtained. After that, these coefficients are plugged into the test set, and prediction arror is obtained: $PE_{tr} = y_{ts} - X_{ts}\widehat{\beta_{tr}}$. Then a z-test (as the number of observations is large, asymptotically normal distribution of sample mean is assumed) is used to test the equality of its mean to 0, and p-value is obtained. The same is done, but now train and test sets are switched, i.e., $PE_{ts} = y_{tr} - X_{tr}\widehat{\beta_{ts}}$. This process (starting from a division into train & test models) is repeated 1000 times (using *set.seed* command in R), so, as a result, a vector of 2000 p-values is obtained. Then another z-test is conducted, which tests $H_0: \mu_{pvalue} = 0.1$ against $H_1: \mu_{pvalue} > 0.1$, i.e., test if Prediction Error is significantly different from 0 in this sample.

4. Data

The dataset of tennis matches (3577 observations of 50 variables) was taken from github[xviii], and it was collected by Jeff Sackmann, an author and software developer who has worked in the fields of sports statistics and test preparation. After the dataset was collected, a number of variables was generated/extracted using the already existing data from this dataset. Then the odds were taken from oddsportal.com[xix], they are average European market odds. Then, additional variables like “days since last match”, “difference in ranking points between the winner and the loser” and many others were added.

After that, the set of observations is divided into train and test sets according to the principle that has been described in Section 3. This division allows for approximately equally large samples for both train and test sets, and large samples are essential to use a large number of variables and asymptotical properties of certain distributions.

Below is the complete list of variables that were picked out for the regression analysis and their meaning:

1. **ypd** (Probability difference) – difference between odds-implied probabilities of win for the actual winner and loser. It is calculated in the following way: Let the odds on the winner and loser be q_w, q_l (European or decimal odds). Then:

$$\frac{1}{q_w}+\frac{1}{q_l}=1+v \Rightarrow \frac{1}{q_w(1+v)}+\frac{1}{q_l(1+v)}=1 \Rightarrow$$
$$p_w=\frac{1}{q_w(1+v)}=\frac{1}{q_w\left(\frac{1}{q_w}+\frac{1}{q_l}\right)}=\frac{q_l}{q_w+q_l};$$
$$p_l=\frac{1}{q_l(1+v)}=\frac{1}{q_l\left(\frac{1}{q_w}+\frac{1}{q_l}\right)}=\frac{q_w}{q_w+q_l}$$
$$ypd=p_w-p_l=\frac{q_l-q_w}{q_l+q_w}$$

   It is a numerical variable (possible values – from -1 to 1, borders excluded)
2. **xrd** (Rank difference) – difference between positions in official ATP rankings of winner and loser at the time of the start of the tournament, during which the match took place. This is a numerical variable, integer (possible values – from $-inf$ to $+inf$).
3. **xrpd** (Ranking points difference) – difference between ranking points, registered in official ATP rankings, of winner and loser at the time of the start of the tournament, during which the match took place. This is a numerical variable, integer (possible values – from $-inf$ to $+inf$).
4. **xfrm28d** (4 week form difference) – difference between 4-week forms of winner and loser by the time of the match. The formula that has been used to calculate it is
$$xfrm28d=(1.1*w\_w\_28d-w\_l\_28d)-(1.1*l\_w\_28d-l\_l\_28d)$$
   where the 1st letter indicates a person within the match (winner – w, loser – l), and the 2nd letter indicates the results their results in the specified period (wins – w, losses – l). So, w_w_28d – winner's number of wins within the last 28 days, l_w_28d – loser's number of wins within the last 28 days etc. This is a numerical variable (possible values – from $-inf$ to $+inf$).
5. **xfrm56d** (8 week form difference) – identical to **xfrm28d** by its meaning and calculation methods, but captures 8-week (56-day) forms.

6. **xmind** (minutes difference) – difference between minutes played so far this tournament by winner and loser. This variable captures the level of tiredness of players and how hard-fought their journey to a given round has been. This is a numerical variable (possible values – from $-inf$ to $+inf$).
7. **xagemd** (age mean difference). This variable captures how far off the average age of the players on tour (proxy for career peak) are winners and losers. First, average age of all players on tour (that have played during selected period) is calculated, and is denoted as age_avg. Then absolute deviation from mean is calculated for both winner and loser. And lastly, the difference between these absolute values for winner and loser is taken.

$$xagemd = \left|age_w - age_{avg}\right| - \left|age_l - age_{avg}\right|$$

   This is a numerical variable (possible values – from $-inf$ to $+inf$).
8. **ctnm** (tournament level) – numerical representation of the level of the tournament (number of ATP points that is given for winning it). At ATP level, there are 5 possible tournament levels: 250 (ATP250), 500 (ATP500), 1000 (Masters 1000), 1300 (ATP Finals) and 2000 (Grand Slam). This is an ordinal categorical variable.
9. **crnd** (round number) – number of the round within the tournament that the match is taking place in. Round of 128 is encoded as 1, round of 64 is encoded as 2, round of 32 is encoded as 3, round of 16 is encoded as 4, quarterfinal is encoded as 5, semi-final is encoded as 6, final is encoded as 7. For the tournament which does not follow the usual knockout format (ATP Finals), relegation round (RR) is encoded as 5.5, as it is the stage that comes before the semi-final, but a win at this stage brings more ATP points than a quarterfinal of a similar level tournament. This is an ordinal categorical variable.
10. **cent** (entry status difference) – difference between entry statuses of winner and loser. Entry statuses are indicated for both winners and losers, and then the difference is taken. There are 4 types of entry statuses:
    1. Seeded – players who are in the top-25% of the draw for the tournament. The higher the ranking, the lower is the seed number, but for this variable they are all combined into a single category. This category is encoded as 4.
    2. Unseeded – players who are not top-25% of the draw but still are eligible to play in the main draw based on their rankings. The difference from seeded category is that in some tournaments seeded players skip the 1st round and start playing from the 2nd round, however, unseeded players always start from the 1st round,

as well as seeded players being protected from meeting another seeded player until later stages of tournaments. This category is encoded as 3.

3. Qualifiers – players who have won a specified number of qualification rounds. These are players that are not ranked high enough to be admitted into the main draw of the tournament. This category is encoded as 2.
4. Lucky Loser (LL), Wildcard (WC), Alternate (Alt.), Special Exempt (SE), Protected Rankings (PR). This is a category which contains 5 subcategories, with each of them being explained in detail in the appendix. This category is encoded as 1.

Therefore, range of **cent** is from -3 to 3. It is an ordinal categorical variable.

11. **cha** (home advantage/disadvantage) – a variable that is equal to 1 if the winner is playing in his home country and the loser is not, -1 if the loser is playing in his home country and the winner is not, and 0 otherwise. It is an ordinal categorical variable, and due to the fact that it can only take 3 values, it can be used both as a dummy variable and as a numerical variable in the regression.
12. **dsf** (surface) – a nominal categorical variable that takes values "grass", "clay" and "hard", which are the only 3 types of court surfaces that have been played on within the timeframe of the database.
13. **dgoatgs** – a dummy variable created for Grand Slam matches that indicates whether one of the players is Rafael Nadal or Novak Djokovic, who seem to be trusted more in Grand Slam matches. It is equal to 1 if it is a Grand Slam match, and the winner is either of the two, and the loser is not, it is equal to -1 if it is a Grand Slam match, and the loser is either of the two, and the winner is not, and it is equal to 0 otherwise.

Histograms of some of the variables (where visualization of the data is important) are included into Appendix.

5. Estimation and results

5.1. Part 1

All of the estimations were done according to the methodology that was described in Section 3 of this paper. The 'starting' model (*Model 1*) that was chosen contains all of the variables described in Section 4, with the exclusion of **frm28d**, which had lower significance than **frm56d**, and the latter includes the former by construction. Figure 1 shows the results of OLS estimation of this model. The next procedure that had to be done was testing the model for heteroscedasticity. Breusch-Pagan test statistic was equal to 172.89 with 13 degrees of freedom, which means its p-value was $< 2.2 * 10^{-16}$, so, null hypothesis of BP test is rejected at 0.1% significance level.

Figure 1. Results of OLS estimation of Model 1

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.362e-02  2.026e-02   3.634 0.000283 ***
xrpd         7.335e-05  2.217e-06  33.081  < 2e-16 ***
xrd          7.680e-04  3.899e-05  19.698  < 2e-16 ***
xfrm56d      2.185e-02  1.046e-03  20.883  < 2e-16 ***
xmind       -4.699e-04  5.842e-05  -8.043 1.19e-15 ***
xagemd       4.851e-03  1.127e-03   4.303 1.73e-05 ***
ctnm         5.467e-05  1.452e-05   3.764 0.000170 ***
crnd        -3.453e-03  5.053e-03  -0.683 0.494384
ctnmcrnd    -1.714e-05  4.780e-06  -3.586 0.000341 ***
cha          6.164e-02  8.261e-03   7.462 1.07e-13 ***
cent         1.568e-02  2.953e-03   5.309 1.17e-07 ***
dsfGrass     2.047e-02  1.590e-02   1.287 0.198134
dsfHard      8.827e-03  8.745e-03   1.009 0.312865
dgoatgs     -5.733e-02  4.014e-02  -1.428 0.153330
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2376 on 3481 degrees of freedom
Multiple R-squared:  0.5989,    Adjusted R-squared:  0.5974
F-statistic: 399.8 on 13 and 3481 DF,  p-value: < 2.2e-16
```

White test statistic was equal to 333 with 26 degrees of freedom, which means its p-value was $4.63 * 10^{-55}$, so, null hypothesis of BP test is rejected at 0.1% significance level. Given the results of these two tests, there is very strong statistical evidence of presence of heteroscedasticity of some type in this model. Hence, robust standard errors have to be used for tests. Figure 2 shows the results of t-tests using these robust standard errors. Coefficients of **crnd**, **dsfGrass, dsfHard** and **dgoatgs** are individually insignificant at 10% level, and all the other coefficients are individually significant at 0.1% level. Wald test statistic of this model has p-value of $< 2.2 * 10^{-16}$, hence, the model is jointly significant at 0.1% level. The same types of tables for the corrected model (without the individually insignificant variables and interaction variable **ctnmcrnd** from *Model 1*) are included in Appendix. Note that the p-value of Wald test statistics for the corrected model was also $< 2.2 * 10^{-16}$, while all of the explanatory variables were individually significant at 0.1% level, which means that it could also be used, however, in the context of this paper, even individually insignificant variables (which, however, do not decrease significance of the model as a whole, compared to a model with only individually significant variables) should be

Figure 2. Results of robust t-tests of Model 1

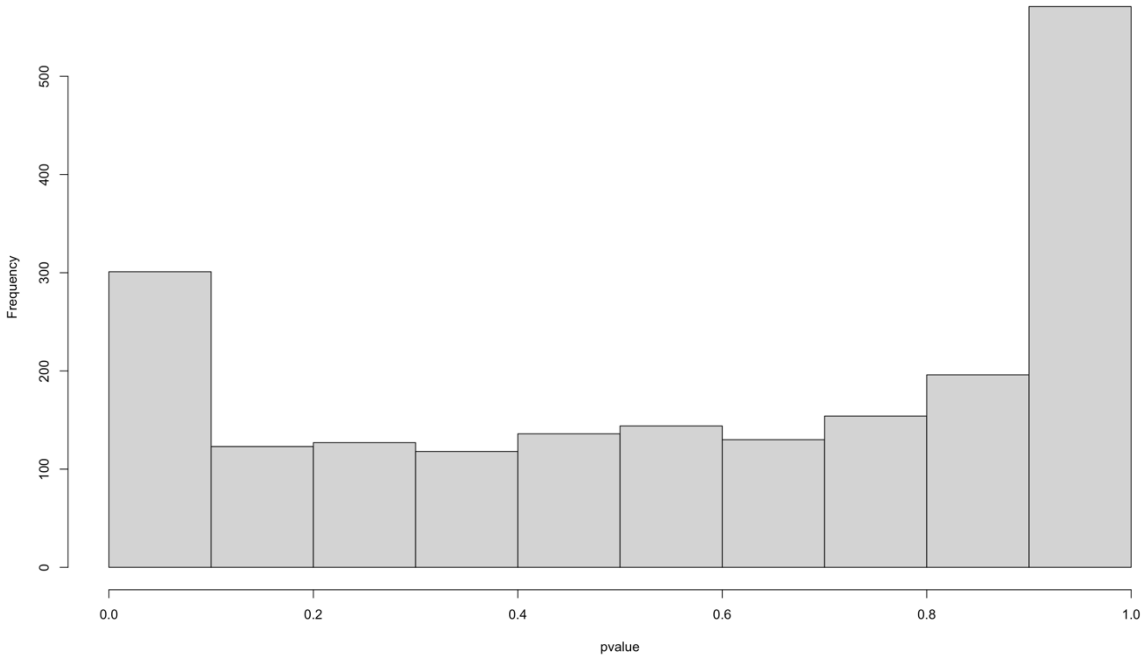
```
t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  7.3622e-02  2.0010e-02  3.6793 0.0002374 ***
xrpd         7.3350e-05  2.0377e-06 35.9964 < 2.2e-16 ***
xrd          7.6799e-04  7.7327e-05  9.9317 < 2.2e-16 ***
xfrm56d      2.1852e-02  1.0142e-03 21.5454 < 2.2e-16 ***
xmind       -4.6993e-04  5.4517e-05 -8.6198 < 2.2e-16 ***
xagemd       4.8508e-03  1.2109e-03  4.0060 6.306e-05 ***
ctnm         5.4666e-05  1.6588e-05  3.2955 0.0009924 ***
crnd        -3.4532e-03  4.7881e-03 -0.7212 0.4708291
ctnmcrnd    -1.7138e-05  5.0778e-06 -3.3750 0.0007463 ***
cha          6.1642e-02  8.2288e-03  7.4909 8.622e-14 ***
cent         1.5681e-02  3.4355e-03  4.5643 5.185e-06 ***
dsfGrass     2.0470e-02  1.6645e-02  1.2298 0.2188713
dsfHard      8.8274e-03  8.5598e-03  1.0313 0.3024883
dgoatgs     -5.7327e-02  4.2814e-02 -1.3390 0.1806652
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

included into the model, as the interpretation of these variables is that betting market on aggregate does not consider them as important, which is also a valuable result of the paper. So, taking everything into consideration, this model is the optimal choice. Interpretation of coefficients can start after cross-validation.

5.2. Part 2

As it has been mentioned in Section 3, for this part, a repeating 2-fold cross validation sampling is used to obtain out-of-sample estimates of **ypd** variable, using the model specified in Part 1 of this section (1000 repetitions). Then a z-statistic (Z_{st}) is formed for testing the null hypothesis of equality to 0 of mean of the difference between actual out-of-sample values of **ypd** and the obtained estimate, with the alternative being that the mean is not equal to 0. Then p-values of these 2000 z-statistics are obtained. The following histogram shows the empirical distribution of p-value of the test statistic.

Figure 3. Histogram of P-values of Z_{st}



Its shape is not even remotely reminiscent of Normal distribution, however, as the number of observations (2000) is very large, according to Central Limit Theorem, sample mean of this 'p-value' variable has asymptotically normal distribution around its actual mean. Then the hypothesis of equality of 0.1 of the mean of 'p-value' variable is tested, with the alternative being that it is larger than one. Z-statistic of this test is 61.81537, and the p-value of this test is $< 2.2 * 10^{-16}$, which means that there is very strong statistical evidence that 'p-value' variable is on average significantly bigger than 0.1, and this in turn means that there is very strong statistical evidence of equality to 0 of mean of the difference between actual out-of-sample values of **ypd** and the obtained estimate. This means that the model performs well on the out-

of-sample data. Having confirmed that, it is now possible to start interpreting the coefficients of *Model 1*.

Figure 4. Coefficients of Model 1

(Intercept)	xrpd	xrd	xfrm56d	xmind	xagemd	ctnm
7.362160e-02	7.334996e-05	7.679918e-04	2.185158e-02	-4.699313e-04	4.850821e-03	5.466626e-05
crnd	ctnmcrnd	cha	cent	dsfGrass	dsfHard	dgoatgs
-3.453210e-03	-1.713768e-05	6.164153e-02	1.568054e-02	2.046972e-02	8.827431e-03	-5.732730e-02

1. The intercept is ≈ 0.0736, which is not interpretable in this model, as not all explanatory variables can take the value of 0, which is essential for interpreting the intercept. General concept of interpretation of a coefficient $\widehat{\beta_i}$ in a linear regression $\hat{y} = X\hat{\beta}$ is: 1 unit increase of a numerical explanatory variable x_i is associated with $\widehat{\beta_i}$ units increase in dependent variable y_i, and 1 unit decrease of an explanatory variable x_i is associated with $\widehat{\beta_i}$ units decrease in dependent variable y_i.
2. Coefficient of **xrpd** is $\approx 7.335 * 10^{-5}$, which means that keeping all the other explanatory variables fixed, 1 rank point increase in rank points difference is associated with $\approx 7.335 * 10^{-5}$ increase in the implied win probability difference. The coefficient is significant, positive, and small. The fact that it is positive is consistent with the logic that players with higher ranking points are in general more trusted in winning. The fact that it is small can be explained by the scale of ranking points: 1 point difference is extremely small for ATP rankings (histogram of **xrpd** visualizes that). The results are consistent with all the papers that concluded that outcomes of tennis matches depend positively on rank difference (Boulier & Stekler, 1999, Clarke & Dyte, 2000, Klaasen & Magnus, 2003, Del Corral & Prieto-Rodriguez, 2010)
3. Coefficient of **xrd** is $\approx 7.680 * 10^{-4}$, and it should be interpreted in the exact same way as the coefficient of **xrpd**. The coefficient is significant, positive, and small. The fact that it is small and positive has the same explanation as for **xrpd**. Unless the ranks in question are 1 and 2, a difference of 1 position in the ranking points is negligibly small.
4. Coefficient of **xfrm56d** is ≈ 0.0219. The coefficient is significant, positive, and quite large (on $(-1,1)$ scale). The fact that it is positive is consistent with the logic that keeping all other variables the same, the player in better form is trusted in winning more that the opponent, and the fact that it is quite large might be explained by the scale of **xfrm56d**, where 1-unit absolute increase is quite large in terms of relative increase.
5. Coefficient of **xmind** is $\approx -4.670 * 10^{-4}$. The coefficient is significant, negative, and small. Its negativity is consistent with the logic that **xmind** is a measure of players' tiredness, and the more tired the player is from playing longer matches in the tournament, the less bettors believe in them as opposed to a less tired player. The fact that it is small by absolute value is attributable to the scale of **xmind**, where 1-unit absolute increase is quite small in terms of relative increase.
6. Coefficient of **xagemd** is $\approx 4.851 * 10^{-3}$, which means that keeping all the other explanatory variables fixed, 1 year increase in age mean difference is associated with $\approx 4.851 * 10^{-3}$ increase in the implied win probability difference. It means that

(keeping all other variables fixed) if a player is 1 year further away from the average age of the players on tour, as opposed to the opponent, bettors tend to believe more in them. It is inconsistent with the notion that players that are close to the age of 'peak', however, it may occur due to the fact that average age of players on tour is a bad proxy for this age of 'peak'. **xagemd** is significant at 0.1% level nevertheless, hence it should be included in *Model 1*.

7. Coefficient of **ctnm** is $\approx 5.467 * 10^{-5}$, coefficient of **crnd** is $\approx -3.453 * 10^{-3}$, coefficient of **ctnmcrnd** is $\approx -1.714 * 10^{-5}$. These 3 categorical variables have to be interpreted as a group, as **crnd** is insignificant, while **ctnm** and **ctnmcrnd** are significant at 0.1% level. Coefficient of **crnd** is insignificant, which means that bettors do not consider round number separately to make any difference in win probabilities. When **ctnm** increases by 1 unit (the number of ATP points awarded for the win increases by 1), then, keeping all other variables fixed, including a fixed value of **crnd** (1 to 7), an associated increase in the implied win probability difference is approximately $5.467 * 10^{-5} - 1.714 * 10^{-5} * \boldsymbol{crnd}$, i.e., $5.467 * 10^{-5} - 1.714 * 10^{-5} * 7$ in the final, $5.467 * 10^{-5} - 1.714 * 10^{-5} * 6$ in the semi-final, and so on. This is consistent with the logic that bettors believe that the higher is the tournament level, the more effort favorites put into winning them (as the stakes are higher), however, the further the match is in the tournament, the smaller this 'motivation' proxy becomes, as players rarely reach the finals if their motivation is low, regardless of their initial status in the tournament.
8. Coefficient of **cha** is ≈ 0.0616, and it is interpreted differently from numerical variables. A person who has home advantage is on average more trusted to win than the person without home advantage (keeping all other variables fixed) by 6.16%. A person who has no home advantage is on average more trusted to win than the person with home disadvantage (keeping all other variables fixed) by 6.16%. The coefficient is significant, positive, and quite large. Its positivity is consistent with the logic that bettors believe that a person who has home advantage is more likely to win (everything else fixed). The fact that it is quite large by absolute value is attributable to the scale of **cha**, where 1-unit absolute increase is quite large in terms of relative increase.
9. Coefficient of **cent** is ≈ 0.0157, which means that 1-unit 'promotion' of a person by status of entry is associated with ≈ 0.0157 increase in the implied win probability difference. The coefficient is significant, positive, and quite large. Its positivity is explained in the same way as the positivity of **xrd** and **xrpd**. The fact that it is quite large by absolute value is attributable to the scale of **cent**, where 1-unit absolute increase is quite large in terms of relative increase.
10. Coefficients of **dsfGrass** and **dsfHard** are insignificant at 10% level, which means that if all other variables are kept fixed, surface change does not affect the beliefs of bettors.
11. Coefficient of **dgoatgs** is insignificant at 10% level, which means that even though Nadal and Djokovic have 42 Grand Slams between them (while all the other active players combined have less than 15), bettors do not consider their winning experience as the factor that gives them more (or less) chances of winning a Grand Slam match against an opponent, all of whose other fundamentals are the same.

6. Summary and conclusion

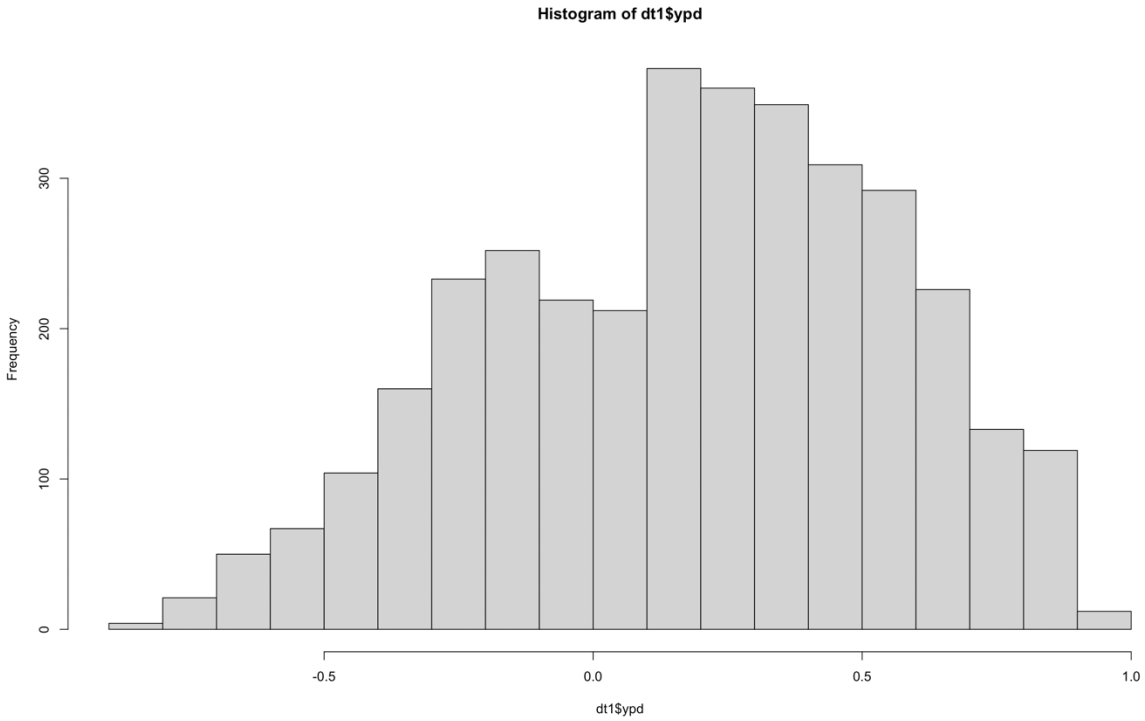
The main purpose of this paper was to find out, which fundamentals bettors (or betting market as an aggregate) consider important when they choose to place bets on one player or another. The results posted strong evidence that the difference between odds-implied probabilities of winning depend positively on rank points difference, rank difference (negative value = higher rank), 8-week form difference, tournament level (with negative correction of round within the tournament), home advantage and entry status, and negatively depend on difference between the levels of tiredness. Moreover, the developed model (including some of the variables that are not significant individually) has shown consistently accurate performance on out-of-sample data, which is a signal that it can be used to determine immediate pre-match odds even before the described dynamic game of pre-match odds setting, based on fundamentals. The paper is intended to contribute to a large variety of works focused on sports betting market efficiency, forecasting, odds determining, and studies that are focused on behavior of people who bet on sporting events frequently. The primary aim of the paper, which is finding fundamental parameters that can be used to construct odds and constructing a model of odds-setting, has been achieved.

7. Appendix

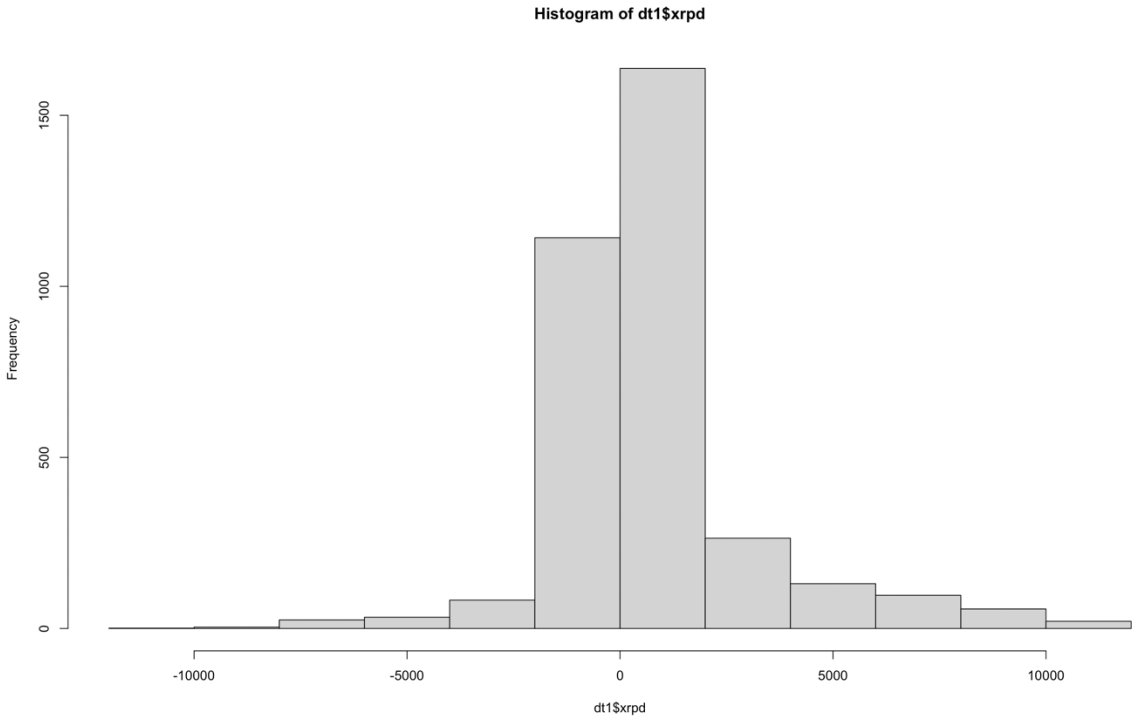
7.1. ***cent*** *variable, category 1, explanation*

1. Lucky Loser (LL) – Highest-ranked player to lose in the final round of qualifying for a tournament, but still ends up qualifying because of a sudden withdrawal by one of the players already in the main draw. In Grand Slam events, one of the four highest-ranked losers in the final qualifying round is randomly picked as the lucky loser.
2. Wildcard (WC) – Player allowed to play in a tournament, even if their rank is not adequate or they do not register in time. Typically, a few places in the draw are reserved for wild cards, which may be for local players who do not gain direct acceptance or for players who are just outside the ranking required to gain direct acceptance. Wild cards may also be given to players whose ranking has dropped due to a long-term injury.
3. Alternate (Alt.) – Player or team that gains acceptance into the main draw of a tournament when a main draw player or team withdraws. Such a player may be a lucky loser.
4. Special Exempt (SE) – Players who are unable to appear in a tournament's qualifying draw because they are still competing in the final rounds of a previous tournament can be awarded a spot in the main draw by special exempt.
5. Protected Rankings (PR) – Players injured for a minimum of six months can ask for a protected ranking, which is based on their average ranking during the first three months of their injury. The player can use their protected ranking to enter tournaments' main draws or qualifying competitions when coming back from injury (or some occurrences such as COVID-19 frozen ranking concerns in 2020–21). It is also used in the WTA for players returning from pregnancy leave.

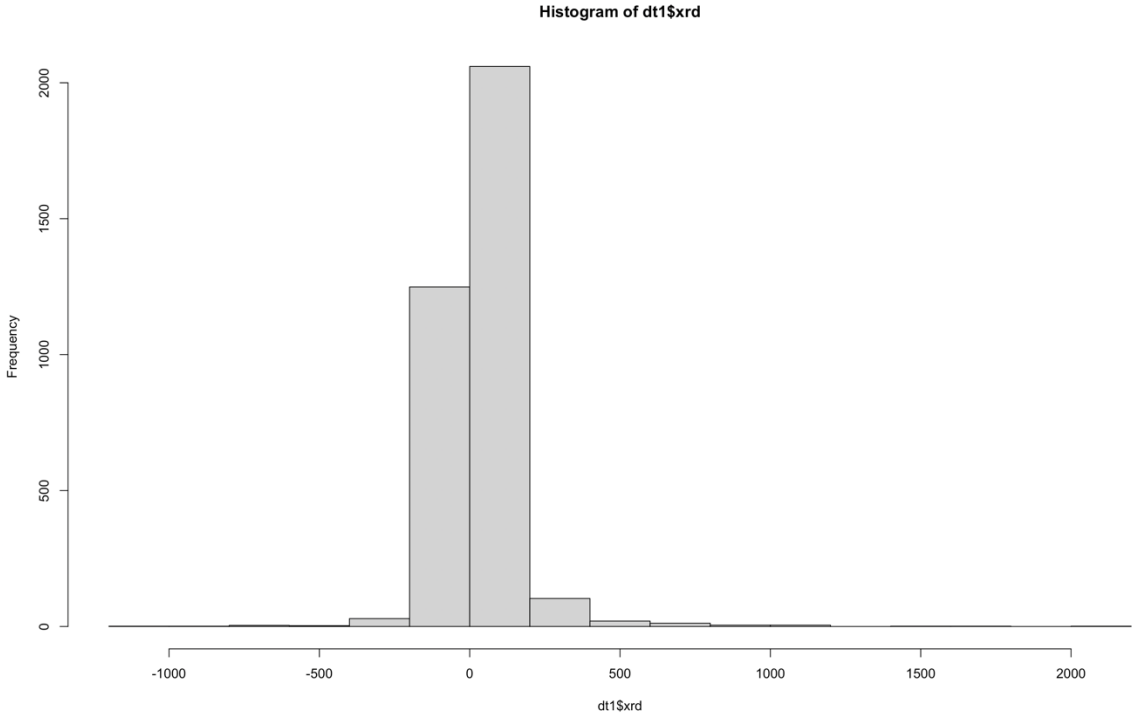
7.2. Histograms of some of the variables from the dataset



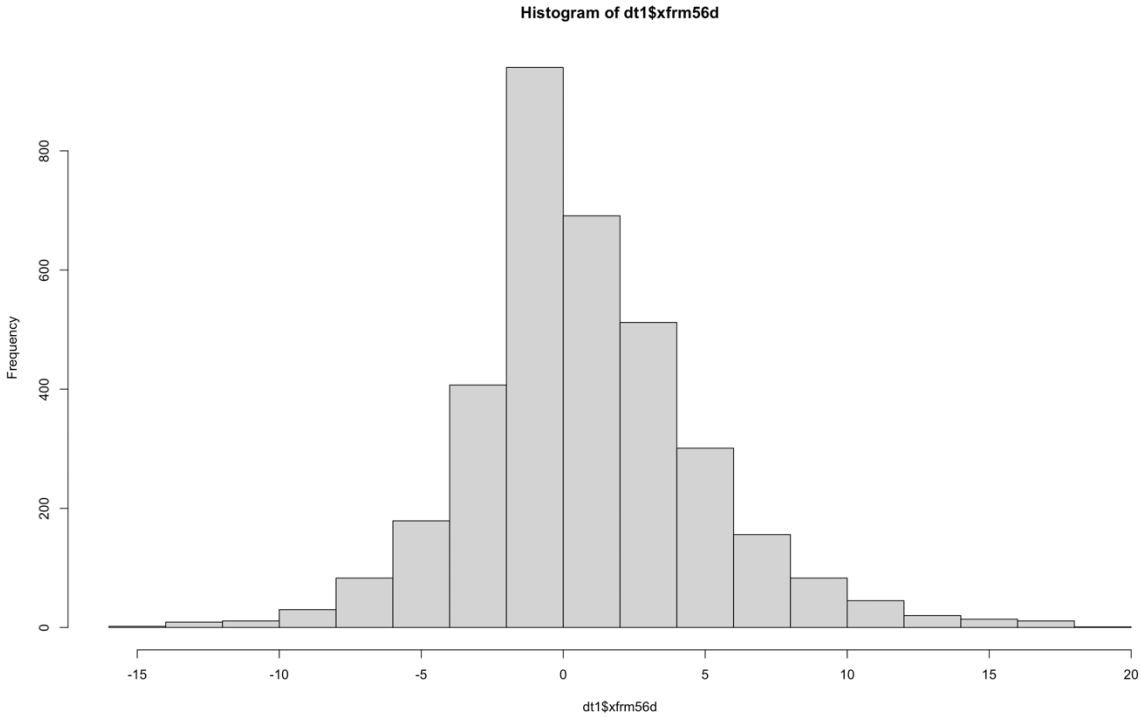
*Graph 1. Histogram of **ypd***



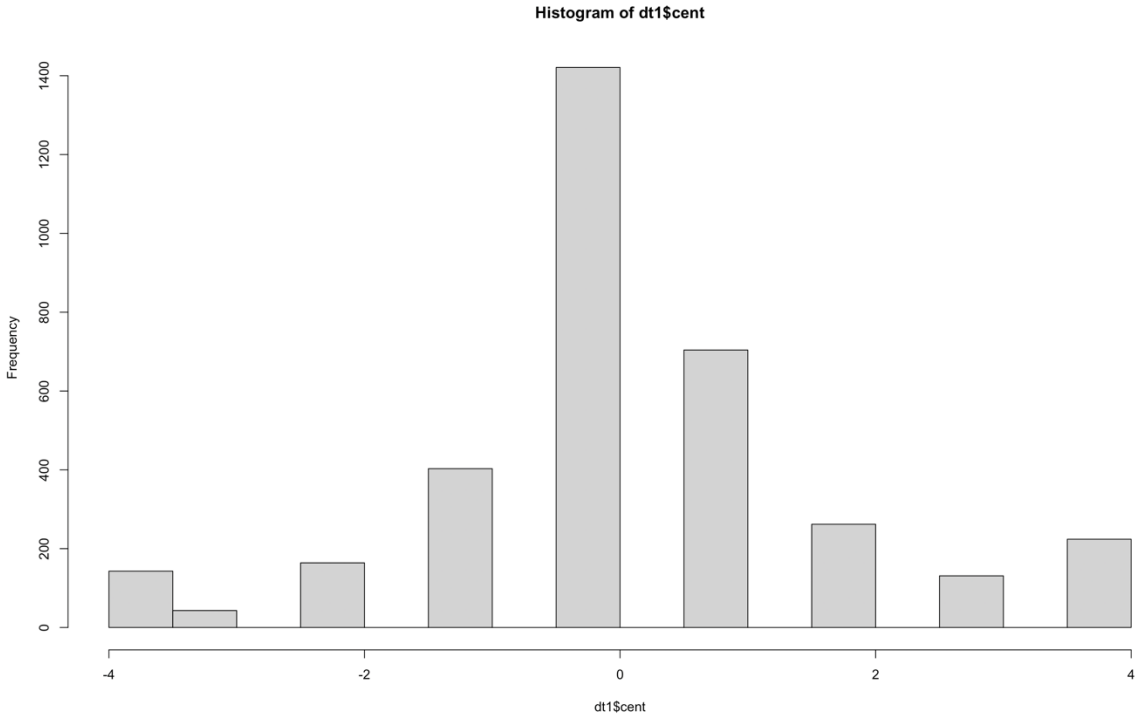
*Graph 2. Histogram of **xrpd***



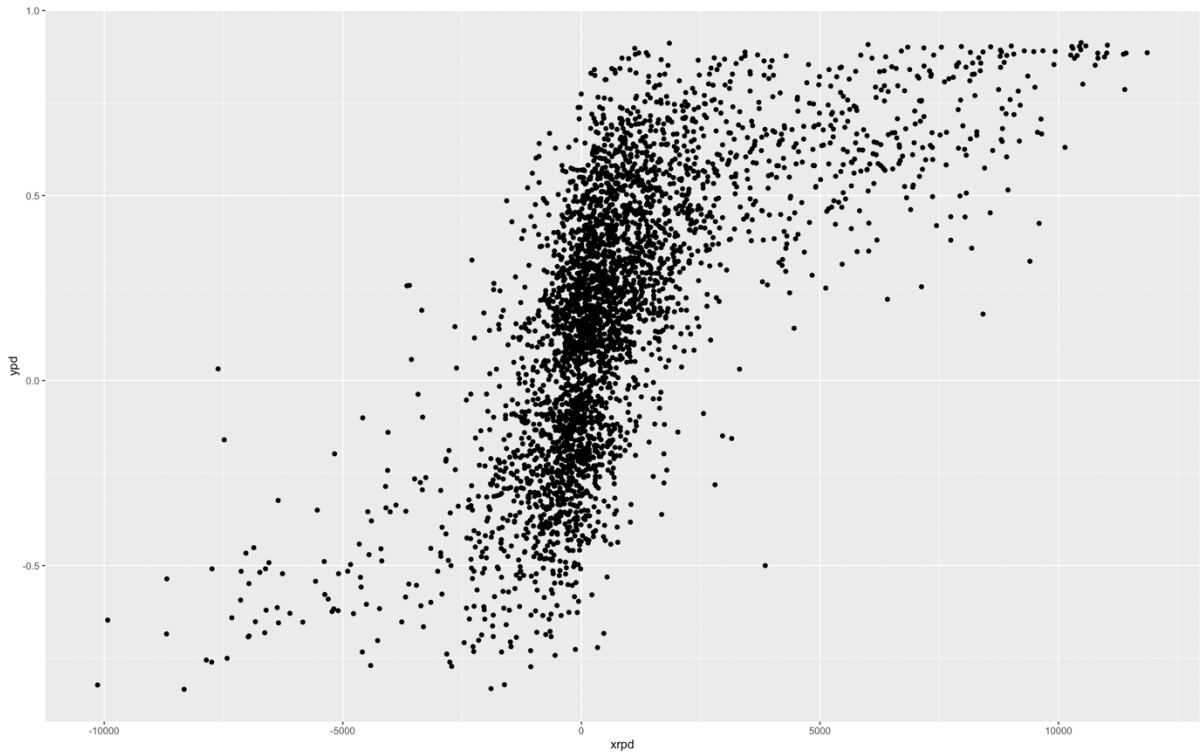
*Graph 3. Histogram of **xrd***



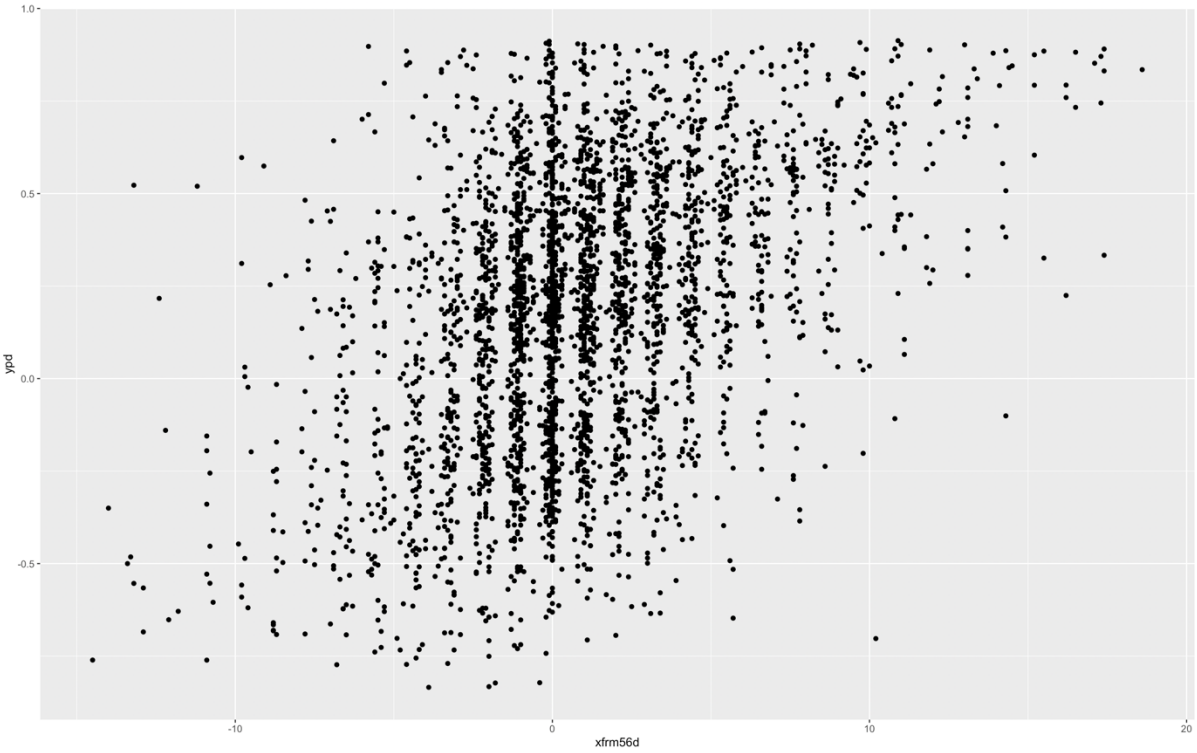
*Graph 4. Histogram of **xfrm56d***



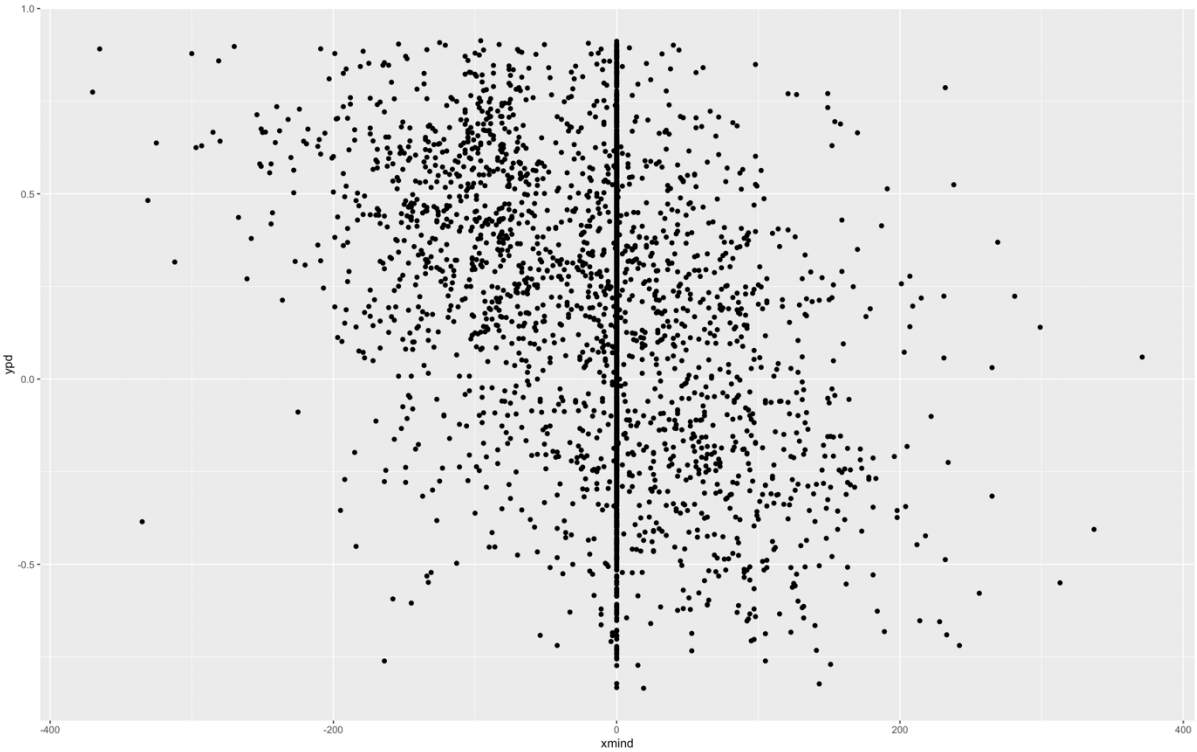
*Graph 5. Histogram of **cent***



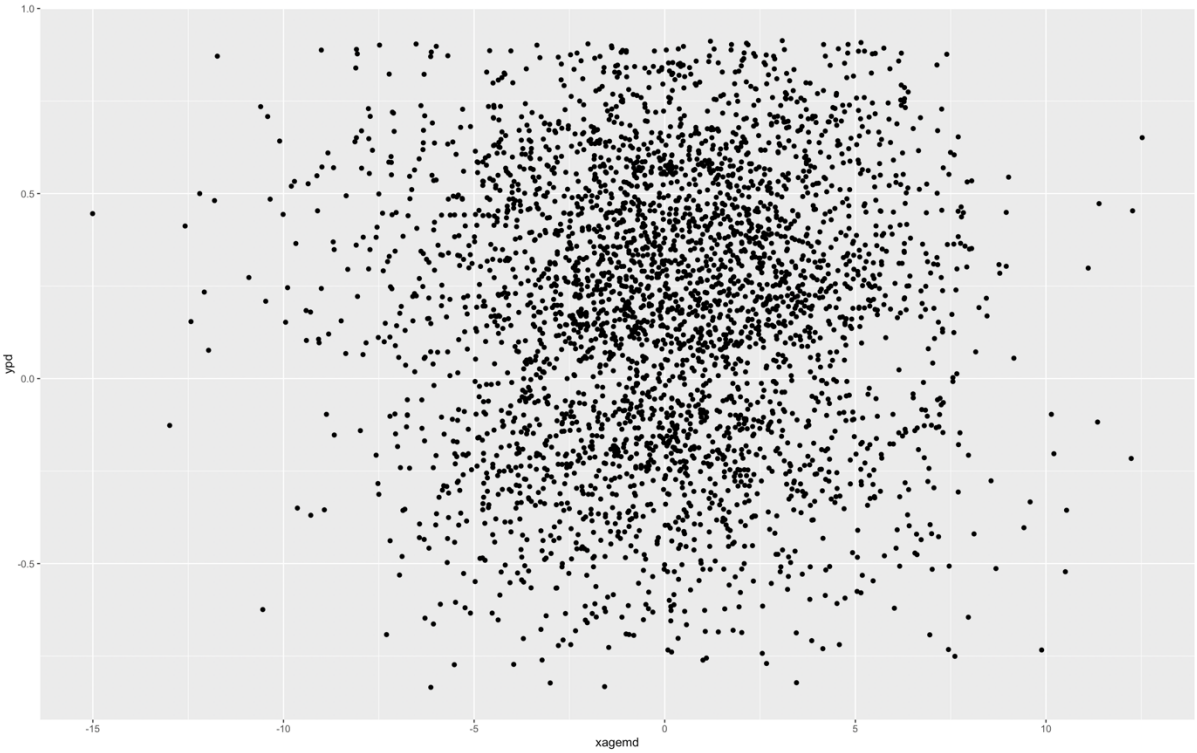
*Graph 6. Plot of **ypd** (y axis) and **xrpd** (x axis)*



*Graph 7. Plot of **ypd** (y axis) and **xfrm56d** (x axis)*



*Graph 8. Plot of **ypd** (y axis) and **xmind** (x axis)*
Note that large concentration of points along the y-axis is attributed to the 1st round matches, when none of the players have had time to play any matches in the tournament yet.



*Graph 9. Plot of **ypd** (y axis) and **xagemd** (x axis)*

7.3. Results of OLS estimation, B-P test, White test, robust t-test for the corrected model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.338e-02  6.448e-03   8.279  < 2e-16 ***
xrpd         7.006e-05  2.104e-06  33.302  < 2e-16 ***
xrd          7.908e-04  3.907e-05  20.237  < 2e-16 ***
xfrm56d      2.195e-02  1.050e-03  20.902  < 2e-16 ***
xmind       -4.686e-04  5.834e-05  -8.032 1.30e-15 ***
xagemd       4.371e-03  1.128e-03   3.876 0.000108 ***
ctnm         3.127e-05  6.470e-06   4.834 1.40e-06 ***
cha          6.279e-02  8.310e-03   7.556 5.28e-14 ***
cent         1.653e-02  2.962e-03   5.581 2.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2391 on 3486 degrees of freedom
Multiple R-squared:  0.5933,    Adjusted R-squared:  0.5924
F-statistic: 635.6 on 8 and 3486 DF,  p-value: < 2.2e-16
```

Figure 5. Results of OLS estimation of corrected model

```
        studentized Breusch-Pagan test

data:  lm11pdc
BP = 147.76, df = 8, p-value < 2.2e-16
```

Figure 6. Results of BP test for corrected model

```
> whc
# A tibble: 1 × 5
  statistic  p.value parameter method       alternative
      <dbl>    <dbl>     <dbl> <chr>        <chr>
1      313. 4.73e-57        16 White's Test greater
```

Figure 7. Results of White test for corrected model

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  5.3380e-02  6.3610e-03  8.3918 < 2.2e-16 ***
xrpd         7.0059e-05  2.0334e-06 34.4545 < 2.2e-16 ***
xrd          7.9075e-04  7.9130e-05  9.9931 < 2.2e-16 ***
xfrm56d      2.1954e-02  1.0222e-03 21.4763 < 2.2e-16 ***
xmind       -4.6862e-04  5.5817e-05 -8.3956 < 2.2e-16 ***
xagemd       4.3706e-03  1.2160e-03  3.5941 0.0003301 ***
ctnm         3.1272e-05  7.4526e-06  4.1961 2.784e-05 ***
cha          6.2787e-02  8.2651e-03  7.5966 3.882e-14 ***
cent         1.6531e-02  3.4890e-03  4.7380 2.245e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 8. Results of robust t-tests for corrected model

8. Bibliography and references

Website links are underlined

[i] https://www.quora.com/How-are-sports-betting-odds-calculated
ii del Corral, J & Prieto-Rodrıguez, J. Are differences in ranks good predictors for Grand Slam tennis matches? International Journal of Forecasting 26 (2010) 551–563.
iii Sauer, R. D. (1998). The economics of wagering markets. Journal of Economic Literature, 36, 2021–2064.
iv Sauer, R. D. (2005). The state of research on markets for sports betting and suggested future directions. Journal of Economics and Finance, 29, 416–426.
v Vaughan Williams, L. (1999). Information efficiency in betting markets: A survey. Bulletin of Economic Research, 51, 1–30.
vi Vaughan Williams, L. (2005). Information efficiency in financial and betting markets. Cambridge: Cambridge University Press.
vii Štrumbelj, E. On determining probability forecasts from betting odds. International Journal of Forecasting. 30 (2014) 934–943.
viii Abinzano, I. & Muga, L. & Santamaria, R. Hidden Power of Trading Activity: The FLB in Tennis Betting. Exchanges Journal of Sports Economics 2019, Vol. 20(2) 261-285.
ix Easton, S. & Uylangco, K. Forecasting outcomes in tennis matches using within-match betting markets. International Journal of Forecasting 26 (2010) 564–575.
x Stekler, H.O. & Sendor, D. & Verlander, R. Issues in sports forecasting. International Journal of Forecasting 26 (2010) 606–621.
xi Boulier, B., & Stekler, H. (1999). Are sports seedings good predictors? An evaluation. International Journal of Forecasting, 15, 83–91.
xii Clarke, S., & Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. International Transactions in Operational Research, 7, 585–594.
xiii Klaassen, F., & Magnus, J. (2001). Are points in tennis independent and identically distributed? Evidence from a dynamic binary panel data model. Journal of the American Statistical Association, 96, 500–509.
xiv Forrest, D. & Mchale, I. Anyone for Tennis (Betting)? The European Journal of Finance Vol. 13, No. 8, 751–768, December 2007.
xv Breusch, T.C. & Pagan, A.R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. Econometrica. Vol. 47, No. 5 (Sep., 1979), pp. 1287-1294.
xvi White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. Econometrica. Vol. 48, No. 4 (May, 1980), pp. 817-838.
xvii https://warwick.ac.uk/fac/soc/economics/staff/academic/corradi/teaching-ec976/msfe-week6.pdf
xviii https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_matches_2021.csv
https://raw.githubusercontent.com/JeffSackmann/tennis_atp/master/atp_matches_2022.csv
xix https://www.oddsportal.com/results/#tennis