Exploratory Analysis of Tennis Data

Abstract

Linear regression of match durations using match statistics, statistical tests like Chi-Square, ANOVA, and Tukey test to test various hypotheses regarding tennis statistics. Cross-validation, Principal Component Regression are also employed.

Chapter 1: Introduction

Tennis is a popular sport with a rich history and a huge following around the world. The game of tennis is not only enjoyed by players but also by fans who closely follow the sport and enjoy watching their favorite players compete. In recent years, the popularity of tennis has grown with the advent of digital media, which has made it easier for fans to access and follow their favorite players and tournaments. As a result, there is a wealth of data available on tennis matches, players, and tournaments.

Tennis has an annual viewership of around 1 billion. Players like Roger Federer, Rafael Nadal, and Novak Djokovic are household names across the world. The industry surrounding it is worth around 2 billion USD.

In this statistics project, we will analyze a tennis dataset to gain insights into the game. The dataset we will be using contains information on tennis matches, players, and tournaments from various competitions around the world. Our goal is to use statistical analysis to explore the patterns and trends in the data which may be used for betting or other purposes.

Chapter 2: Data Description

The data has been obtained fro Jeff Sackman's work. The dataset has match statistics of tennis matches from 1968 to 2022. The dataset has a lot of missing features upto 1991. The dataset is complete from 1991. The columns are numerous and have been explained at the relevant part.

tourney_id - a unique identifier for each tournament.

tourney_name

surface

draw_size - number of players in the draw

tourney_level - For men: 'G' = Grand Slams, 'M' = Masters 1000s, 'A' = other tour-level events, 'C' = Challengers, 'S' = Satellites/ITFs, 'F' = Tour finals and other season-ending events, and 'D' = Davis Cup

tourney_date

match_num

winner_id - the player_id used in this repo for the winner of the match

winner_seed

winner_entry

winner_name

winner_hand - Right or Left

winner_ht - height in centimeters, where available

winner_ioc - country code

winner_age

loser_id

loser_seed

loser_entry

loser_name

loser_hand

loser_ht

loser_ioc

loser_age

score

best_of - '3' or '5', indicating the the number of sets for this match

round

minutes - match length, where available

w_ace - winner's number of aces

w_df - winner's number of doubles faults

w_svpt - winner's number of serve points

w_1stIn - winner's number of first serves made

w_1stWon - winner's number of first-serve points won

w_2ndWon - winner's number of second-serve points won

w_SvGms - winner's number of serve games

w_bpSaved - winner's number of break points saved

w_bpFaced - winner's number of break points faced

l_ace l_df l_svpt l_1stIn l_1stWon l_2ndWon l_SvGms l_bpSaved l_bpFaced

winner_rank- winner's ATP or WTA rank, as of the tourney_date

winner_rank_points - number of ranking points

loser_rank

loser_rank_points

Chapter 3: Methodology

ANOVA - ANOVA stands for Analysis of Variance, and it is a statistical method used to compare the means of two or more groups. ANOVA tests whether there is a significant difference between the means of the groups, taking into account the variability within each group and the sample size.

Student's t-test - Student's t-test is a statistical hypothesis test used to determine if there is a significant difference between the means of two independent groups. It assumes that the data is normally distributed

and that the variances of the two groups are equal. The t-test calculates a t-value, which is the difference between the means of the two groups divided by the standard error of the difference. If the calculated t-value is greater than the critical value, then the difference between the means is considered statistically significant.

Tukey's HSD test - Tukey's Honestly Significant Difference (HSD) test is a post-hoc test used to identify which pairs of groups have significantly different means after an ANOVA. The test uses a significance level to determine if the difference between two means is statistically significant.

Linear Regression - Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The method estimates the best-fit line that describes the relationship between the variables, which can be used to make predictions about the dependent variable based on the independent variables.

K-Fold Cross Validation - K-Fold Cross Validation is a method used to evaluate the performance of a machine learning model. The data is divided into K subsets, and the model is trained and tested K times, with each subset used once for testing and the remaining subsets used for training. This helps to reduce the risk of overfitting the model to the training data.

Chi-Square Test of Independence - The Chi-Square Test of Independence is a statistical method used to determine if there is a significant association between two categorical variables. The test compares the observed frequencies of the variables to the expected frequencies, assuming there is no association between them.

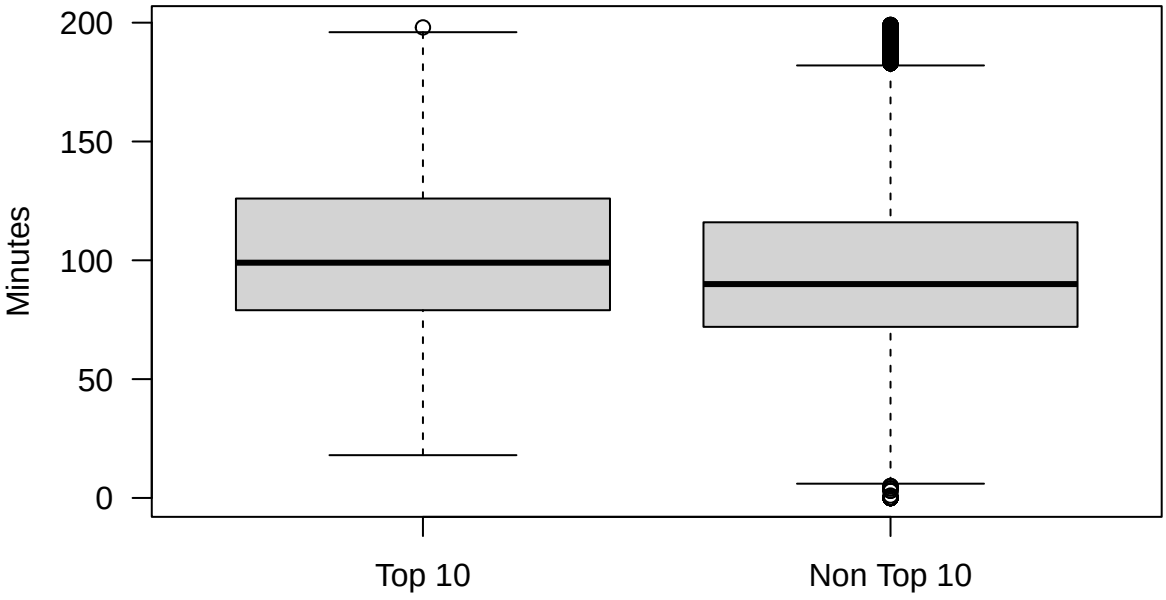
Principal Component Regression - Principal Component Regression is a statistical method used to reduce the number of variables in a regression model by combining them into a smaller number of principal components. The principal components are orthogonal, linear combinations of the original variables that explain the most variance in the data. These components can then be used in a regression model to predict the dependent variable.

Chapter 4: Analysis and Result

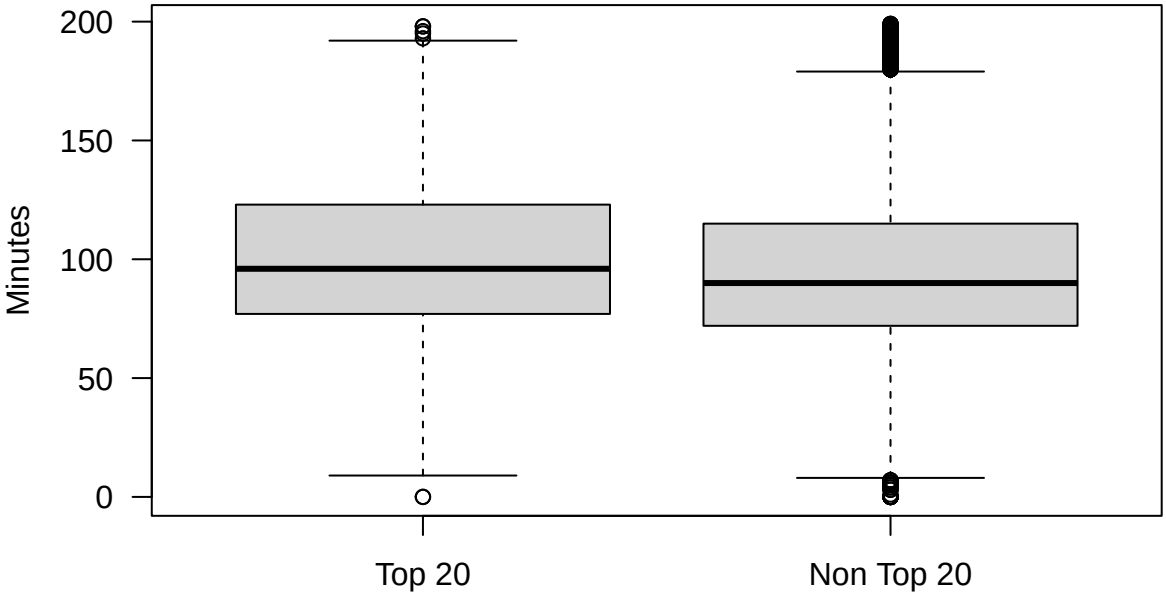
4.1 Comparing Two Samples

It is often the case that matches between closely ranked players are hard fought and therefore take longer time. We try to test this on the data.

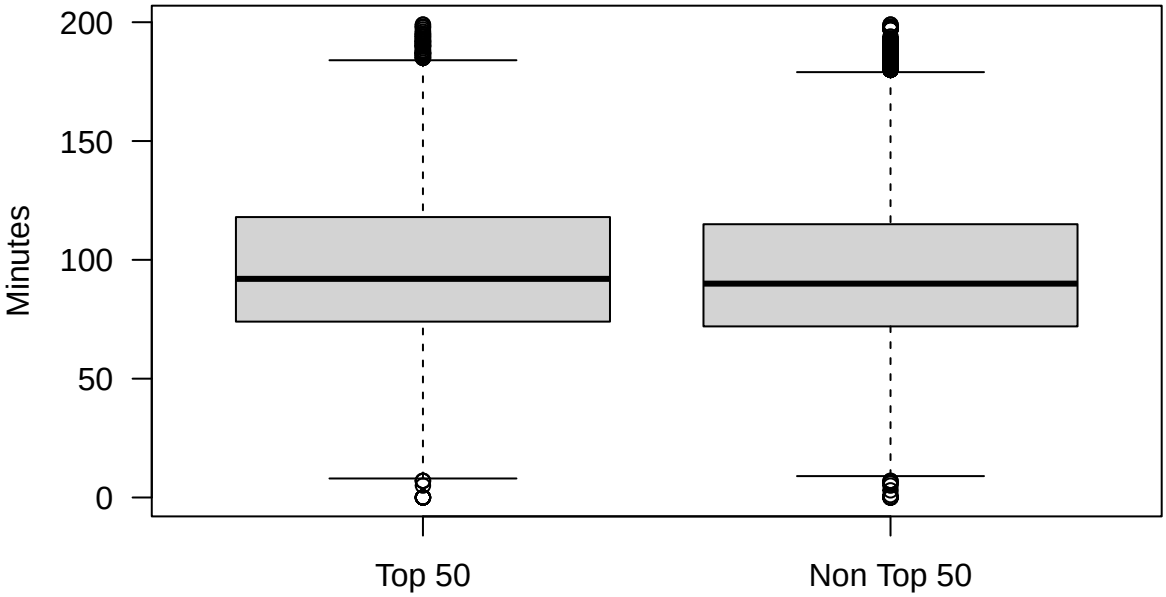
We plot the boxplots of match durations between top-10 players and matches between non-top10 players.



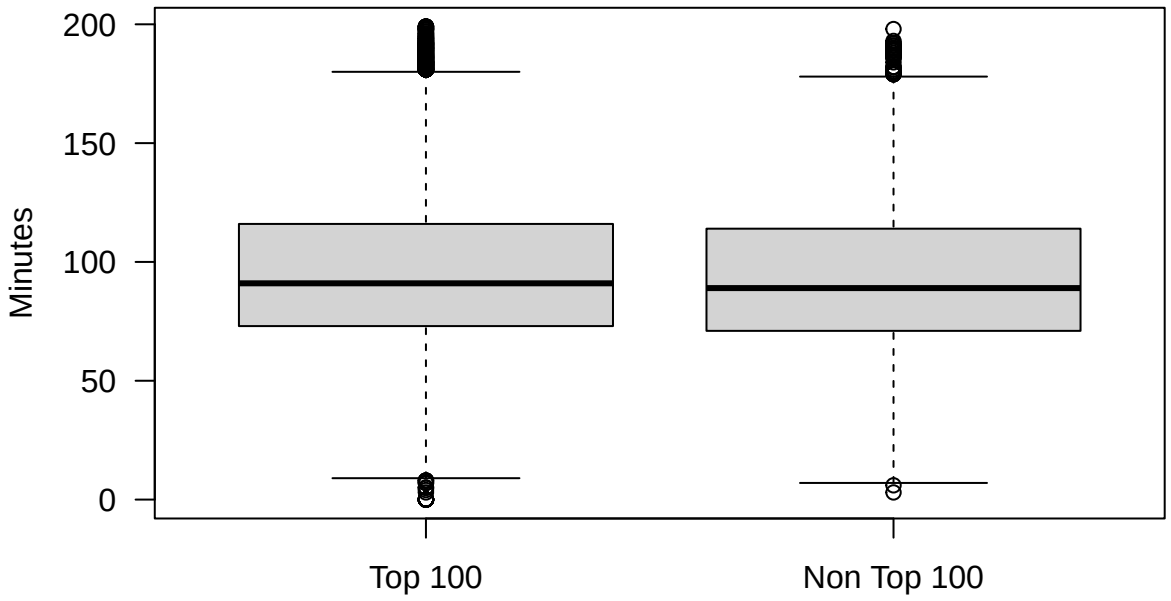
Boxplots of match durations between top-20 players and matches between non-top20 players.



Boxplots of match durations between top-50players and matches between non-top50 players.

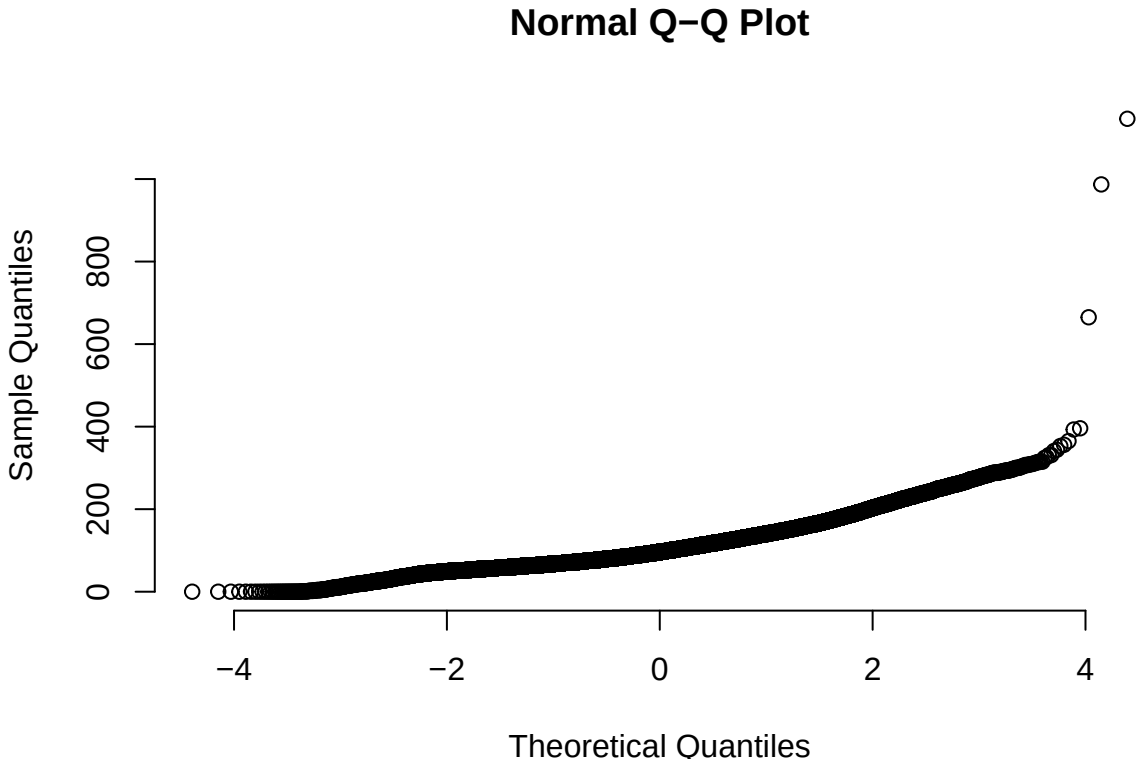


Boxplots of match durations between top-100 players and matches between non-top100 players.



We see that the boxes go from being spaced farther to being spaced closer, with the medians moving closer as well.

Before applying the tests, we check if the duration of each match follows a normal distribution. We plot it for all matches.



The sample quantiles follow the theoretical quantiles sharply. Therefore, the population is normal. We perform Student's t-test on the data.

```
##
##  Welch Two Sample t-test
##
```

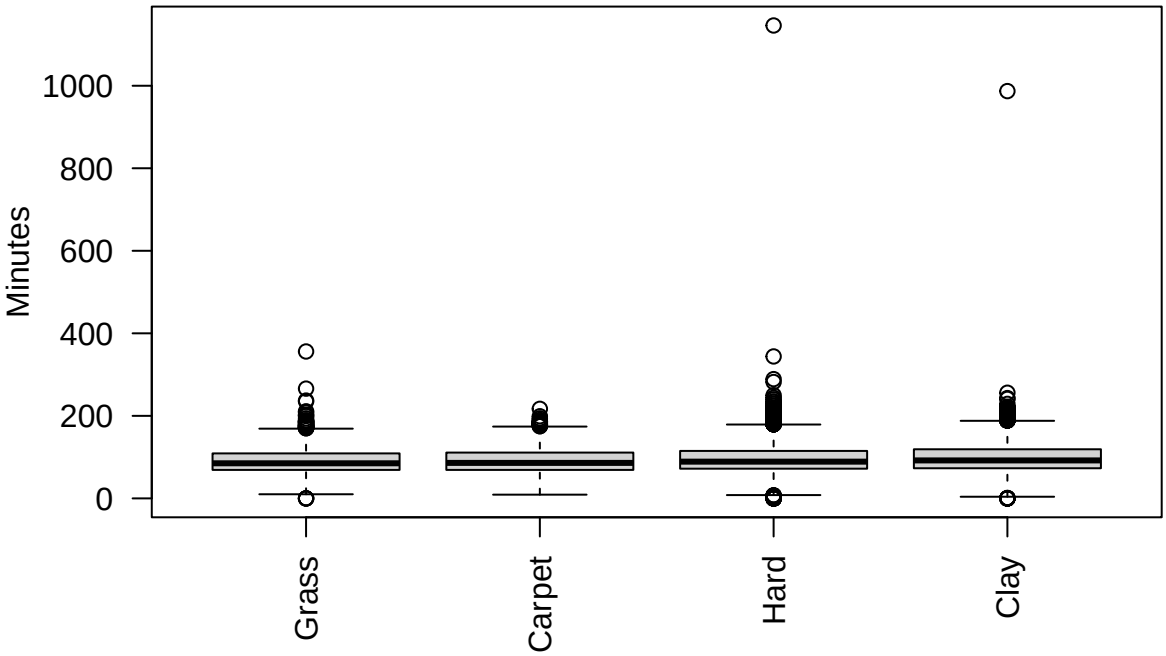
```
## data:  top10 and nontop10
## t = 9.7481, df = 1301.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   6.997358 10.523354
## sample estimates:
## mean of x mean of y
##  103.7951   95.0347

##
##  Welch Two Sample t-test
##
## data:  top20 and nontop20
## t = 11.336, df = 3918.5, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.208512 7.386900
## sample estimates:
## mean of x mean of y
## 101.24971  94.95201

##
##  Welch Two Sample t-test
##
## data:  top50 and nontop50
## t = 8.7413, df = 36521, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.057623 3.247083
## sample estimates:
## mean of x mean of y
##  97.39665  94.74430

##
##  Welch Two Sample t-test
##
## data:  top100 and nontop100
## t = 4.9485, df = 7754.1, p-value = 7.638e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.236469 2.858733
## sample estimates:
## mean of x mean of y
##  95.99243  93.94483
```

The t-tests demonstrate that the difference in means is significant and that there is a decreasing trend as we expand the selection to the bottom.

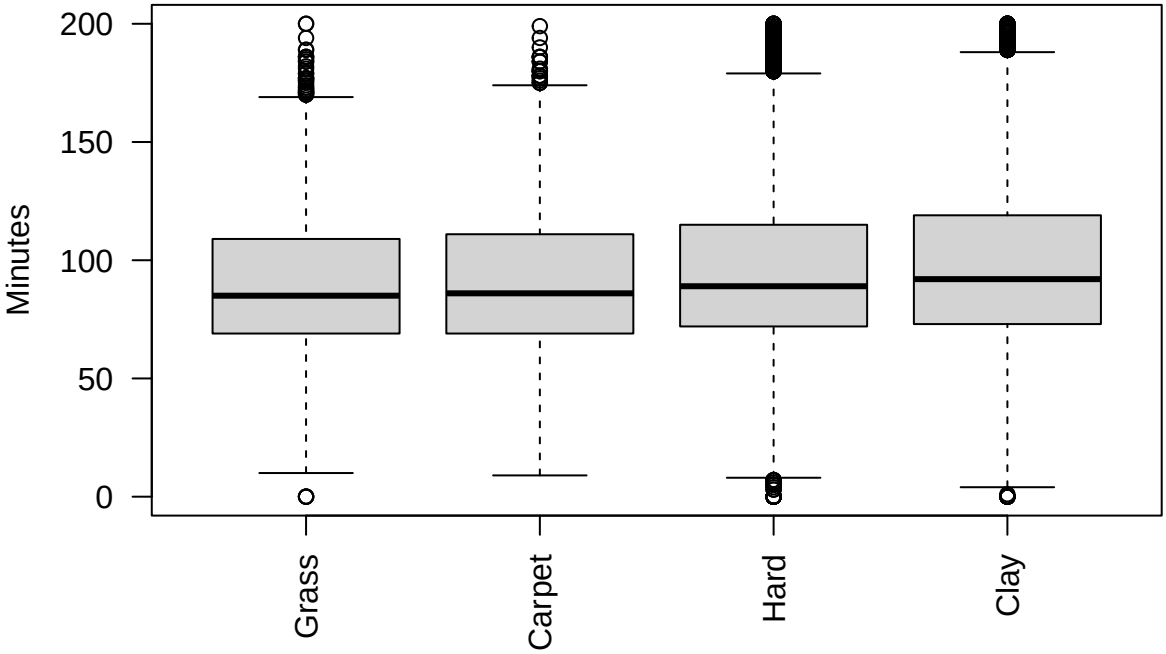
4.2 The Analysis of Variance

The surface of the court plays an important role in the game of tennis. It determines how fast the ball moves across. Grass courts are the fastest, followed by carpet and hard, with clay being the slowest. The speed of the ball dictates how long points are played, which determines the length of the game.

We look at the duration of matches (best-of-3 sets) across surfaces and test for any significant difference.

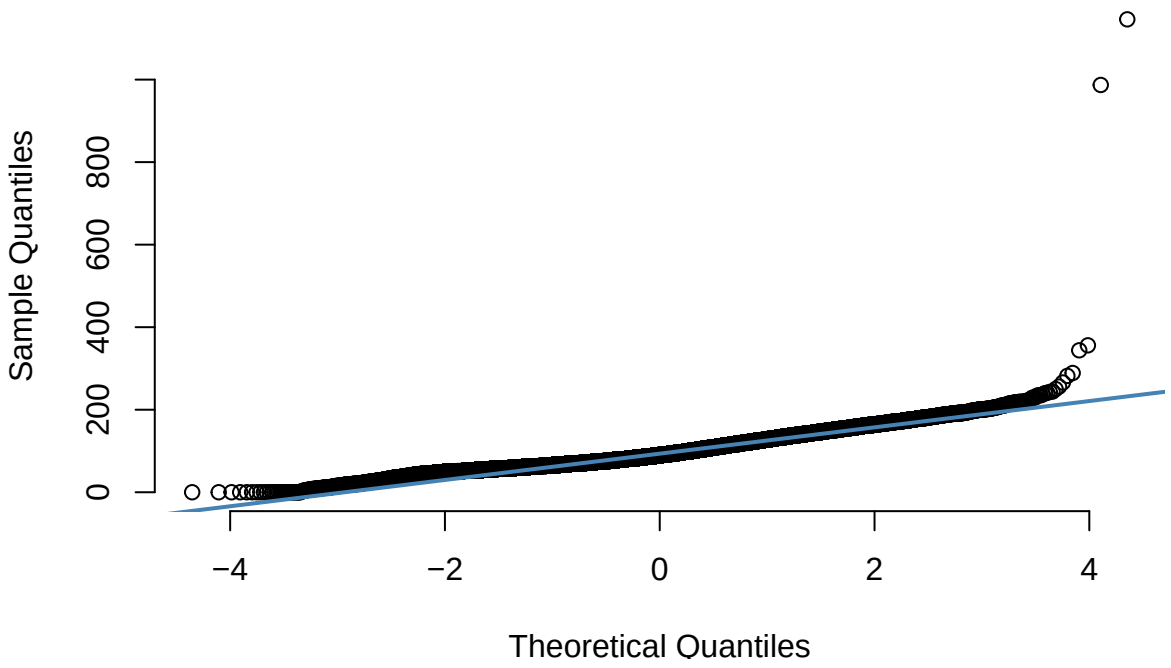


Except for some outliers, most matches end in under 200 minutes. We plot the boxplot again filtering for these outliers.



Cursorily, we see an increasing order in the duration.

Before applying the tests, we check if the duration of each match follows a normal distribution.



The sample quantiles follow the theoretical quantiles sharply. Therefore, the population is normal.

We perform ANOVA on the data.

```
##                Df   Sum Sq Mean Sq F value Pr(>F)
## surface         3   372315  124105   128.3 <2e-16 ***
## Residuals   74651 72208668     967
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also perform Tukey's HSD test on the data.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = minutes2 ~ surface)
##
## $surface
##                   diff       lwr        upr     p adj
## Clay-Carpet   6.274172  5.111599  7.4367449 0.0000000
## Grass-Carpet -1.071745 -2.585524  0.4420335 0.2643362
## Hard-Carpet   3.507524  2.378675  4.6363734 0.0000000
## Grass-Clay   -7.345917 -8.543299 -6.1485352 0.0000000
## Hard-Clay    -2.766648 -3.412014 -2.1212812 0.0000000
## Hard-Grass    4.579270  3.414603  5.7439362 0.0000000
```

The differences in durations are in the expected order of grass, carpet, hard, and clay. The differences are highly significant. We see that there is not a significant difference between grass and carpet courts. Both grass and carpet are considered quite similar by players as well.

4.3 The Analysis of Categorical Data

The height of a player is an important factor in tennis. Being taller allows better services, which may allow one to hit more aces and to win more points on serve. There is also a notion that the advantage taller players have had is increasing over the years.

We perform categorical data analysis to test these hypotheses. We extract the results of matches in which the taller player has won or whether the shorter player has won. The results are categorized by surface.

```
##         Short  Tall
## Carpet   7419  7041
## Clay     8088  7937
## Grass   23055 28009
## Hard    23674 30552
```

As the samples are quite large, we perform the Chi-Square test of independence on the data.

```
##
##  Pearson's Chi-squared test
##
## data:  table4
## X-squared = 426.32, df = 3, p-value < 2.2e-16
```

The p-value indicates that there is a significant effect of the surface on taller or shorter players winning.

4.4 Linear Regression

We regress the duration of the match in minutes against all available variables in the dataset.

```
lg = dataset[,c(3,13,15,21,23,25:45,46,48,50),]
lg$year = as.factor(lg$year)
lg = lg[minutes<=200&best_of == 3]
lg = na.omit(lg)
model = lm(minutes~., data = lg)
summary(model)
```

```
##
## Call:
## lm(formula = minutes ~ ., data = lg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.451   -5.781   -0.535    5.093  123.938
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.449e+01  7.034e+00   2.060  0.03937 *
## surfaceClay   3.790e+00  1.554e-01  24.391  < 2e-16 ***
## surfaceGrass -3.522e+00  1.928e-01 -18.262  < 2e-16 ***
## surfaceHard   1.406e+00  1.473e-01   9.546  < 2e-16 ***
## winner_ht    -2.450e-02  6.187e-03  -3.960 7.51e-05 ***
## winner_age    1.406e-01  1.016e-02  13.833  < 2e-16 ***
```

```
## loser_ht     -2.665e-02  6.080e-03  -4.382 1.18e-05 ***
## loser_age     1.015e-01  9.860e-03  10.293  < 2e-16 ***
## best_of              NA         NA      NA       NA
## roundER      -1.746e+01  7.056e+00  -2.474  0.01335 *
## roundF       -9.357e+00  6.846e+00  -1.367  0.17173
## roundQF      -1.257e+01  6.844e+00  -1.837  0.06617 .
## roundR128    -1.491e+01  6.847e+00  -2.177  0.02945 *
## roundR16     -1.354e+01  6.843e+00  -1.979  0.04779 *
## roundR32     -1.485e+01  6.843e+00  -2.171  0.02996 *
## roundR64     -1.486e+01  6.844e+00  -2.171  0.02996 *
## roundRR      -1.362e+01  6.847e+00  -1.989  0.04667 *
## roundSF      -1.134e+01  6.844e+00  -1.656  0.09764 .
## w_ace        -3.022e-01  1.056e-02 -28.633  < 2e-16 ***
## w_df         -3.825e-01  2.106e-02 -18.166  < 2e-16 ***
## w_svpt        7.570e-01  1.376e-02  55.000  < 2e-16 ***
## w_1stIn      -7.741e-02  1.314e-02  -5.890 3.88e-09 ***
## w_1stWon     -6.681e-02  2.595e-02  -2.575  0.01003 *
## w_2ndWon      2.581e-02  2.659e-02   0.970  0.33181
## w_SvGms       3.602e-01  7.233e-02   4.979 6.40e-07 ***
## w_bpSaved    -1.004e+00  7.339e-02 -13.682  < 2e-16 ***
## w_bpFaced     9.242e-01  6.995e-02  13.213  < 2e-16 ***
## l_ace        -2.815e-01  1.233e-02 -22.839  < 2e-16 ***
## l_df         -3.760e-01  1.816e-02 -20.704  < 2e-16 ***
## l_svpt        7.178e-01  1.398e-02  51.334  < 2e-16 ***
## l_1stIn      -1.092e-01  1.105e-02  -9.879  < 2e-16 ***
## l_1stWon     -3.787e-02  2.658e-02  -1.425  0.15419
## l_2ndWon      1.732e-02  2.739e-02   0.632  0.52722
## l_SvGms       1.930e-01  7.396e-02   2.609  0.00907 **
## l_bpSaved     2.865e-01  7.520e-02   3.809  0.00014 ***
## l_bpFaced    -3.201e-01  7.096e-02  -4.511 6.46e-06 ***
## winner_rank  -4.575e-03  4.895e-04  -9.345  < 2e-16 ***
## loser_rank   -2.705e-03  3.264e-04  -8.286  < 2e-16 ***
## year1992      2.218e+00  2.612e-01   8.489  < 2e-16 ***
## year1993      2.948e+00  2.578e-01  11.434  < 2e-16 ***
## year1994     -1.223e+00  2.576e-01  -4.747 2.07e-06 ***
## year1995     -5.593e+00  2.612e-01 -21.410  < 2e-16 ***
## year1996     -6.720e+00  2.630e-01 -25.547  < 2e-16 ***
## year1997     -6.433e+00  2.658e-01 -24.199  < 2e-16 ***
## year1998     -6.568e+00  2.669e-01 -24.605  < 2e-16 ***
## year1999     -5.504e+00  2.733e-01 -20.144  < 2e-16 ***
## year2000     -3.865e+00  2.745e-01 -14.079  < 2e-16 ***
## year2001     -2.429e+00  2.728e-01  -8.903  < 2e-16 ***
## year2002     -1.980e+00  2.766e-01  -7.157 8.30e-13 ***
## year2003     -1.816e+00  2.771e-01  -6.555 5.60e-11 ***
## year2004     -1.681e+00  2.753e-01  -6.105 1.03e-09 ***
## year2005     -9.862e-02  2.740e-01  -0.360  0.71889
## year2006      7.948e-01  2.742e-01   2.898  0.00375 **
## year2007      1.735e+00  2.793e-01   6.213 5.21e-10 ***
## year2008      2.815e+00  2.803e-01  10.043  < 2e-16 ***
## year2009      3.945e+00  2.813e-01  14.021  < 2e-16 ***
## year2010      3.912e+00  2.831e-01  13.817  < 2e-16 ***
## year2011      5.894e+00  2.834e-01  20.799  < 2e-16 ***
## year2012      5.796e+00  2.853e-01  20.317  < 2e-16 ***
## year2013     -6.482e-01  2.883e-01  -2.249  0.02452 *
```

```
## year2014      1.058e-01  2.906e-01   0.364  0.71577
## year2015      2.286e-01  3.525e-01   0.648  0.51677
## year2016      1.329e+00  2.857e-01   4.653 3.28e-06 ***
## year2017      1.556e+00  2.880e-01   5.405 6.49e-08 ***
## year2018      2.934e+00  2.849e-01  10.298  < 2e-16 ***
## year2019      4.693e+00  2.868e-01  16.361  < 2e-16 ***
## year2020      8.337e+00  3.633e-01  22.951  < 2e-16 ***
## year2021      8.273e+00  2.886e-01  28.667  < 2e-16 ***
## year2022      1.205e+01  2.879e-01  41.852  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.673 on 72871 degrees of freedom
## Multiple R-squared:  0.8987, Adjusted R-squared:  0.8986
## F-statistic:  9651 on 67 and 72871 DF,  p-value: < 2.2e-16
```

We first look the dummy variables of surface and their impact on duration. As discussed in the previous section, surfaceGrass has a negative coefficient, indicating the shorter duration of matches. surfaceClay has a smaller coefficient than surfaceHard though. This may be explained by the fact that there are more hard-court matches in a year than those of clay.

The round a match is played in, whether a final or quarterfinal, does not seem to have an impact. Both winner_rank and loser_rank are negative indicating that matches between higher ranked players are contested closely and therefore take longer time.

w_bpFaced is the number of break points that the winner has faced, and it has a relatively large positive value. This is cosistent with expectation and can be explained in the following manner. The higher the break points one faces, the more is the chance of losing and the match ending quickly. Therefore, if a player who has faced many break points were to win, it would indicate a 'comeback' and would lengthen the match.

The dummy variables for year indicate that the matches played in the late-2010s were longer than the ones played in the 1990s.

For further analysis, we restrict ourselves to matches played on 'hard' courts and played after the year 2010, due to computational issues.

4.5 Resampling Methods

We select 20% of the observations for testing and calculate the Mean Square Error.

```
lg = dataset[,c(3,13,15,21,23,25:45,46,48,50),]
lg = lg[minutes<=200&best_of == 3]
lg = na.omit(lg)
lg2 = lg[year>2010&surface=='Hard']

lg2 = lg2[,c(2:5, 8:28)]
```

```
set.seed(1)
train <- sample(13898, 2800)
```

```
lm.fit <- lm(minutes ~ w_svpt + w_ace + l_svpt + l_ace , data = lg2, subset = train)
```

```
mean((lg2$minutes - predict(lm.fit, lg2))[-train]^2)
```

```
## [1] 93.30913
```

4.6 Linear Model Selection and Regularization

We use best subset selection from the module 'leaps'.

```
library(leaps)
lg2 = lg[year>2010&surface=='Hard']
lg2 = lg2[,c(2:5, 8:28)]

regfit.full <- regsubsets(minutes ~ ., data = lg2, nvmax = 15)
reg.summary<- summary(regfit.full)
reg.summary
```

```
## Subset selection object
## Call: regsubsets.formula(minutes ~ ., data = lg2, nvmax = 15)
## 24 Variables  (and intercept)
##             Forced in Forced out
## winner_ht       FALSE      FALSE
## winner_age      FALSE      FALSE
## loser_ht        FALSE      FALSE
## loser_age       FALSE      FALSE
## w_ace           FALSE      FALSE
## w_df            FALSE      FALSE
## w_svpt          FALSE      FALSE
## w_1stIn         FALSE      FALSE
## w_1stWon        FALSE      FALSE
## w_2ndWon        FALSE      FALSE
## w_SvGms         FALSE      FALSE
## w_bpSaved       FALSE      FALSE
## w_bpFaced       FALSE      FALSE
## l_ace           FALSE      FALSE
## l_df            FALSE      FALSE
## l_svpt          FALSE      FALSE
## l_1stIn         FALSE      FALSE
## l_1stWon        FALSE      FALSE
## l_2ndWon        FALSE      FALSE
## l_SvGms         FALSE      FALSE
## l_bpSaved       FALSE      FALSE
## l_bpFaced       FALSE      FALSE
## winner_rank     FALSE      FALSE
## loser_rank      FALSE      FALSE
## 1 subsets of each size up to 15
## Selection Algorithm: exhaustive
##          winner_ht winner_age loser_ht loser_age w_ace w_df w_svpt w_1stIn
## 1  ( 1 ) " "       " "        " "      " "       " "   " "  "*"    " "
## 2  ( 1 ) " "       " "        " "      " "       " "   " "  "*"    " "
## 3  ( 1 ) " "       " "        " "      " "       "*"   " "  "*"    " "
## 4  ( 1 ) " "       " "        " "      " "       "*"   " "  "*"    " "
## 5  ( 1 ) " "       " "        " "      " "       "*"   " "  "*"    " "
```

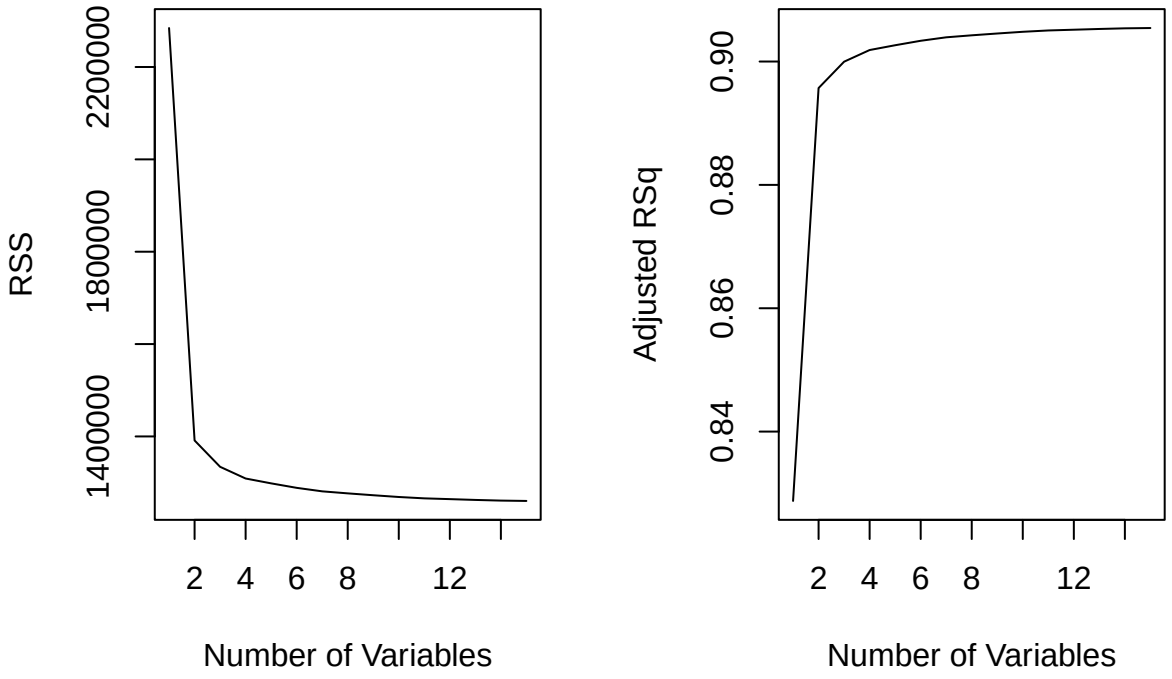
```
## 6  ( 1 )  " "       " "         " "       " "        "*"    " "  "*"     " "
## 7  ( 1 )  " "       " "         " "       " "        "*"    " "  "*"     " "
## 8  ( 1 )  " "       " "         " "       " "        "*"    "*"  "*"     " "
## 9  ( 1 )  " "       " "         " "       " "        "*"    "*"  "*"     " "
## 10  ( 1 ) " "       " "         " "       " "        "*"    "*"  "*"     " "
## 11  ( 1 ) " "       " "         " "       " "        "*"    "*"  "*"     " "
## 12  ( 1 ) " "       " "         " "       " "        "*"    "*"  "*"     "*"
## 13  ( 1 ) "*"       " "         " "       " "        "*"    "*"  "*"     "*"
## 14  ( 1 ) "*"       " "         " "       " "        "*"    "*"  "*"     "*"
## 15  ( 1 ) "*"       " "         " "       " "        "*"    "*"  "*"     "*"
##           w_1stWon w_2ndWon w_SvGms w_bpSaved w_bpFaced l_ace l_df l_svpt
## 1  ( 1 )  " "      " "      " "     " "       " "        " "   " "  " "
## 2  ( 1 )  " "      " "      " "     " "       " "        " "   " "  "*"
## 3  ( 1 )  " "      " "      " "     " "       " "        " "   " "  "*"
## 4  ( 1 )  " "      " "      " "     " "       " "        "*"   " "  "*"
## 5  ( 1 )  " "      " "      " "     " "       " "        "*"   "*"  "*"
## 6  ( 1 )  " "      " "      " "     " "       " "        "*"   "*"  "*"
## 7  ( 1 )  " "      " "      "*"     " "       " "        "*"   "*"  "*"
## 8  ( 1 )  " "      " "      "*"     " "       " "        "*"   "*"  "*"
## 9  ( 1 )  " "      " "      "*"     " "       " "        "*"   "*"  "*"
## 10  ( 1 ) " "      " "      " "     "*"       "*"        "*"   "*"  "*"
## 11  ( 1 ) " "      " "      "*"     "*"       "*"        "*"   "*"  "*"
## 12  ( 1 ) " "      " "      "*"     "*"       "*"        "*"   "*"  "*"
## 13  ( 1 ) " "      " "      "*"     "*"       "*"        "*"   "*"  "*"
## 14  ( 1 ) " "      " "      " "     "*"       "*"        "*"   "*"  "*"
## 15  ( 1 ) " "      " "      "*"     "*"       "*"        "*"   "*"  "*"
##           l_1stIn l_1stWon l_2ndWon l_SvGms l_bpSaved l_bpFaced winner_rank
## 1  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 2  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 3  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 4  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 5  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 6  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 7  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 8  ( 1 )  " "     " "      " "      " "     " "       " "        " "
## 9  ( 1 )  " "     " "      " "      " "     " "       " "        "*"
## 10  ( 1 ) " "     " "      " "      " "     " "       " "        "*"
## 11  ( 1 ) " "     " "      " "      " "     " "       " "        "*"
## 12  ( 1 ) " "     " "      " "      " "     " "       " "        "*"
## 13  ( 1 ) " "     " "      " "      " "     " "       " "        "*"
## 14  ( 1 ) " "     "*"      " "      "*"     " "       " "        "*"
## 15  ( 1 ) " "     "*"      " "      " "     "*"       " "        "*"
##           loser_rank
## 1  ( 1 )  " "
## 2  ( 1 )  " "
## 3  ( 1 )  " "
## 4  ( 1 )  " "
## 5  ( 1 )  " "
## 6  ( 1 )  "*"
## 7  ( 1 )  "*"
## 8  ( 1 )  "*"
## 9  ( 1 )  "*"
## 10  ( 1 ) "*"
## 11  ( 1 ) "*"
```

```
## 12  ( 1 ) "*"
## 13  ( 1 ) "*"
## 14  ( 1 ) "*"
## 15  ( 1 ) "*"
```

We see that upto 4 variables, the variables selected are winner_aces, loser_aces, winner_servicepoints, and loser_servicepoints.

We plot the RSS and adjusted R^2.

```
reg.summary<- summary(regfit.full)
par(mfrow = c(1, 2))
plot(reg.summary$rss, xlab = "Number of Variables",
     ylab = "RSS", type = "l")
plot(reg.summary$adjr2, xlab = "Number of Variables",
     ylab = "Adjusted RSq", type = "l")
```



We see that with these 4 variables, our adjusted r^2 moves upto 0.9. Considering that a majority of a match of tennis involves winning points on serve, the result is unsurprising. In fact, a tennis match can be won only by breaking the opponent's serve, or rather losing on serve rather than winning.

4.7 Moving Beyond Linearity

We perform Principal Component Regression and analyze the results.

```
library(pls)
```

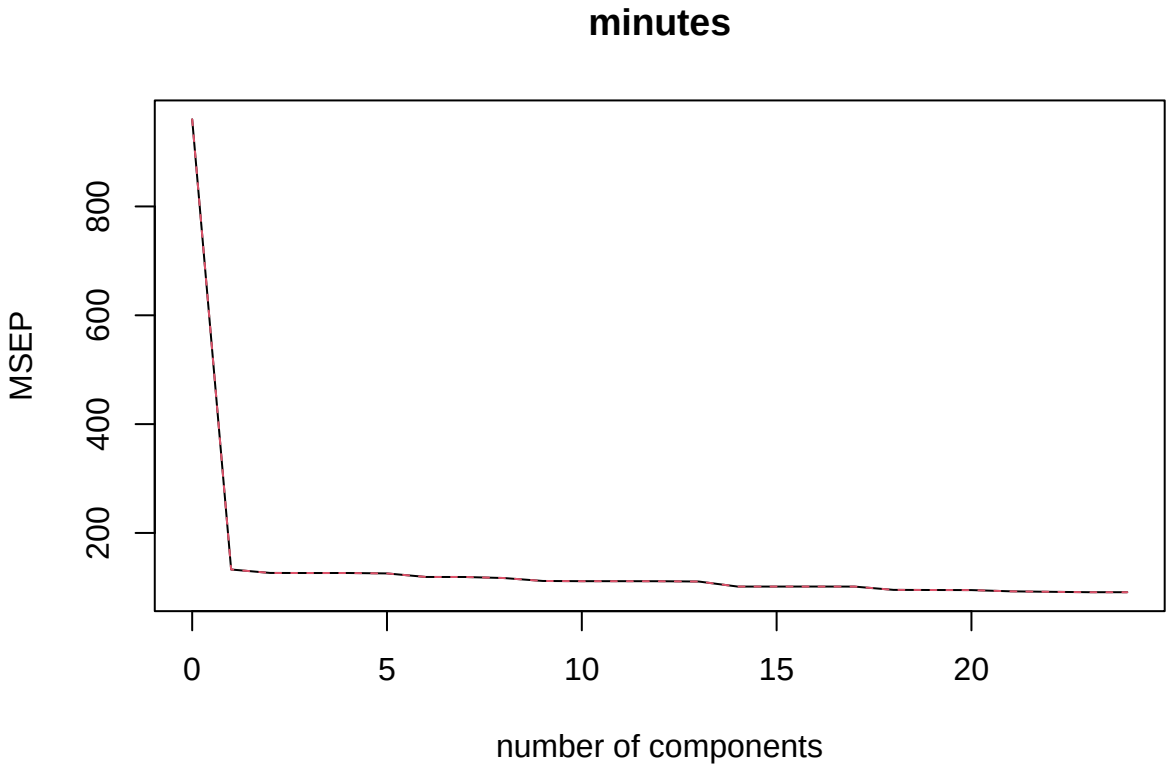
```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
set.seed(2)
pcr.fit <- pcr(minutes ~ ., data = lg2, scale = TRUE,
validation = "CV")
summary(pcr.fit)
```

```
## Data:    X dimension: 13898 24
##  Y dimension: 13898 1
## Fit method: svdpc
## Number of components considered: 24
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           30.98    11.53    11.25    11.24    11.24    11.21    10.91
## adjCV        30.98    11.52    11.25    11.24    11.24    11.21    10.91
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       10.91    10.82    10.57     10.55     10.55     10.55     10.52
## adjCV    10.91    10.82    10.57     10.55     10.55     10.55     10.52
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## CV        10.07     10.07     10.07     10.07     9.767     9.754     9.747
## adjCV     10.07     10.07     10.07     10.07     9.766     9.753     9.746
##        21 comps  22 comps  23 comps  24 comps
## CV        9.630     9.585     9.545     9.542
## adjCV     9.629     9.583     9.544     9.541
##
## TRAINING: % variance explained
##          1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X          38.45    48.24    55.43    61.39    66.63    71.17    75.40    79.55
## minutes    86.17    86.83    86.84    86.84    86.92    87.60    87.61    87.81
##          9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X          83.26     86.53     89.51     92.06     94.25     96.12     97.75
## minutes    88.38     88.42     88.42     88.44     88.49     89.46     89.46
##          16 comps  17 comps  18 comps  19 comps  20 comps  21 comps  22 comps
## X            98.62     99.30     99.56     99.70     99.80     99.89     99.95
## minutes      89.46     89.46     90.10     90.12     90.14     90.38     90.47
##          23 comps  24 comps
## X            99.98    100.00
## minutes      90.55     90.56
```

```
validationplot(pcr.fit, val.type = "MSEP")
```



The prinicipal component analysis shows us that the use of one component explains around 86% of the variance.

Conclusion

We have demonstrated the effect of surface on match duration. We have tested if the height of a player determines victory differently across surfaces. With regression we have predicted the duration of a match with ingame statistics. We can further develop our predictions using other machine-learning models, or by engineering many extra features.

References