Project

Steven Herrera and Ethan Shen

11/09/2018

```
library(tidyverse)
library(olsrr)
library(cowplot)
library(car)
library(broom)
library(knitr)
```

Data

Using data manipulation skills in R, we shaped the dataset to show each observation as the outcome of the match for the winner and the loser of each match from 2010-2017. Below, is a glimpse of our dataset.

```
glimpse(tennis)
```

```
## Observations: 34,388
## Variables: 22
## $ tourney_id   <chr> "2017-M020", "2017-M020", "2017-M020", "2017-M020...
## $ tourney_name <chr> "Brisbane", "Brisbane", "Brisbane", "Brisbane", "...
## $ surface      <chr> "Hard", "Hard", "Hard", "Hard", "Hard", "Hard", "...
## $ tourney_date <int> 20170102, 20170102, 20170102, 20170102, 20170102,...
## $ round        <chr> "F", "SF", "SF", "QF", "QF", "QF", "QF", "R16", "...
## $ minutes      <int> 108, 87, 101, 140, 124, 61, 156, 69, 55, 89, 65, ...
## $ name         <chr> "Grigor Dimitrov", "Grigor Dimitrov", "Kei Nishik...
## $ hand         <chr> "R", "R", "R", "R", "R", "R", "R", "R", "L", "R",...
## $ ht           <int> 188, 188, 178, 196, 188, 178, 183, 196, 185, 185,...
## $ age          <dbl> 25.63450, 25.63450, 27.01164, 26.01780, 25.63450,...
## $ rank         <int> 17, 17, 5, 3, 17, 5, 4, 3, 9, 8, 17, 79, 5, 45, 4...
## $ rankpoints   <int> 2035, 2035, 4905, 5450, 2035, 4905, 5315, 5450, 3...
## $ ace          <int> 7, 4, 1, 23, 3, 3, 11, 12, 1, 11, 8, 5, 3, 6, 7, ...
## $ df           <int> 2, 1, 1, 3, 3, 0, 3, 1, 2, 1, 0, 2, 3, 1, 2, 0, 3...
## $ svpt         <int> 77, 58, 77, 97, 94, 34, 119, 53, 38, 65, 46, 103,...
## $ firsstIn     <int> 52, 36, 56, 62, 52, 19, 67, 40, 18, 44, 27, 65, 6...
## $ firsttWon    <int> 41, 27, 37, 50, 42, 18, 47, 30, 15, 36, 25, 49, 4...
## $ secondndWon  <int> 12, 18, 14, 16, 23, 10, 28, 7, 15, 15, 12, 18, 18...
## $ SvGms        <int> 13, 10, 11, 15, 14, 7, 16, 9, 7, 11, 9, 17, 15, 7...
## $ bpSaved      <int> 5, 0, 4, 6, 13, 0, 11, 2, 2, 4, 3, 4, 4, 4, 0, 0,...
## $ bpFaced      <int> 7, 0, 5, 7, 14, 0, 13, 3, 2, 4, 3, 7, 8, 4, 1, 2,...
## $ status       <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
```

Because we have 30,146 observations, we will randomly select observations to be included in a smaller dataset so that we can effectively examine exploratory data analysis. Below, is our code on how we randomly selected the observations in our new dataset.

```
set.seed(1234)
ten <- tennis %>% sample_n(1000)
```

Here is what our new dataset looks like:

```
glimpse(ten)
```

```
## Observations: 1,000
## Variables: 22
## $ tourney_id   <chr> "2016-6242", "2016-0429", "2016-O16", "2016-0328"...
## $ tourney_name <chr> "Winston-Salem", "Stockholm", "Olympics", "Basel"...
## $ surface      <chr> "Hard", "Hard", "Hard", "Hard", "Hard", "Hard", "...
## $ tourney_date <int> 20160822, 20161017, 20160808, 20161024, 20120716,...
## $ round        <chr> "R64", "R16", "R64", "R32", "QF", "F", "R32", "R3...
## $ minutes      <int> 114, 79, 68, 66, 135, NA, 66, 62, NA, 77, 109, 75...
## $ name         <chr> "Kyle Edmund", "Nicolas Almagro", "Robin Haase", ...
## $ hand         <chr> "R", "R", "R", "R", "R", "R", "R", "L", "R", "L",...
## $ ht           <int> NA, 183, 190, 183, 185, 188, NA, 188, 180, 183, 1...
## $ age          <dbl> 21.62081, 31.15674, 29.34155, 34.33265, 19.80835,...
## $ rank         <int> 85, 41, 62, 51, 326, 1, 112, 87, 37, 60, 27, 10, ...
## $ rankpoints   <int> 679, 1019, 798, 890, 136, 13755, 516, 623, 1100, ...
## $ ace          <int> 9, 3, 6, 0, 7, 5, 4, 0, 3, 2, 13, 1, 9, 3, 10, 3,...
## $ df           <int> 3, 0, 1, 0, 7, 2, 2, 0, 1, 4, 1, 4, 0, 0, 1, 2, 5...
## $ svpt         <int> 77, 58, 61, 50, 87, 106, 52, 42, 68, 57, 73, 53, ...
## $ firsstIn     <int> 41, 37, 41, 28, 45, 64, 29, 25, 39, 34, 42, 30, 5...
## $ firsttWon    <int> 33, 23, 23, 20, 31, 42, 22, 21, 23, 21, 31, 20, 3...
## $ secondndWon  <int> 18, 11, 10, 11, 23, 20, 15, 9, 20, 10, 13, 13, 21...
## $ SvGms        <int> 12, 10, 9, 9, 11, 14, 10, 7, 10, 9, 12, 10, 15, 1...
## $ bpSaved      <int> 5, 4, 4, 2, 8, 15, 0, 0, 4, 5, 0, 0, 5, 9, 3, 1, ...
## $ bpFaced      <int> 7, 7, 8, 5, 10, 19, 1, 0, 6, 9, 3, 3, 10, 14, 5, ...
## $ status       <dbl> 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1...
```

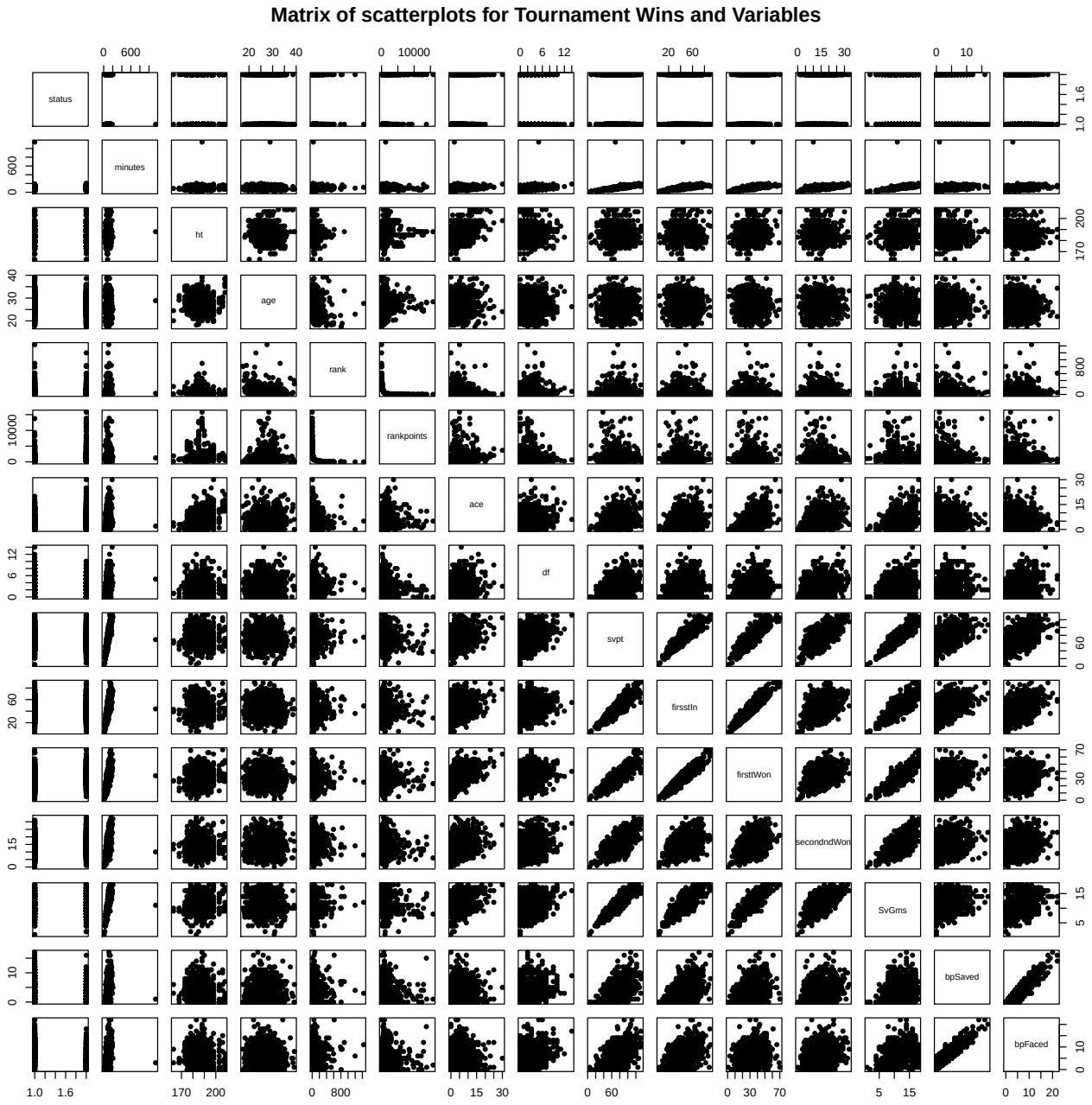
Exploratory Data Analysis

To begin our exploratory data analysis, we will examine a matrix plot of the variables in our dataset to consider multicollinearity.

Matrix Plot

```
ten <- ten %>%
  mutate(status = as.factor(status))

pairs(status ~ minutes + ht + age + rank + rankpoints + ace +
        df + svpt + firsstIn + firsttWon + secondndWon +
        SvGms + bpSaved + bpFaced, data=ten, pch = 16,
      main = "Matrix of scatterplots for Tournament Wins and Variables")
```



Logistic Regression Model

To begin our regression models, we will

```
#backwards <- ols_step_backward_aic(fullmodel)
```

Model Assessment

Prediction

Conclusion