Project

Steven Herrera and Ethan Shen

11/09/2018

```r
library(tidyverse)
library(olsrr)
```

Exploratory Data Analysis

```r
atp <- read_csv("files/atp.csv")
atp2016 <- read_csv("files/atp2016.csv")
atp2015 <- read_csv("files/atp2015.csv")
atp2014 <- read_csv("files/atp2014.csv")
atp2013 <- read_csv("files/atp2013.csv")
atp2012 <- read_csv("files/atp2012.csv")
atp2011 <- read_csv("files/atp2011.csv")
atp2010 <- read_csv("files/atp2010.csv")

atp1 <- atp %>%
  filter(tourney_date < 20171113)

winners2017 <- atp1 %>%
  filter(round == "F")

winners2016 <- atp2016 %>%
  filter(tourney_date < 20161114) %>%
  filter(round == "F")

winners2015 <- atp2015 %>%
  filter(tourney_date < 20151115) %>%
  filter(round == "F")

winners2014 <- atp2014 %>%
  filter(tourney_date < 20141109) %>%
  filter(round == "F")

winners2013 <- atp2013 %>%
  filter(tourney_date < 20131104) %>%
  filter(round == "F")

winners2012 <- atp2012 %>%
  filter(tourney_date < 20121105) %>%
  filter(round == "F")

winners2011 <- atp2011 %>%
  filter(tourney_date < 20111114) %>%
  filter(round == "F")
```

```r
winners2010 <- atp2010 %>%
  filter(tourney_date < 20101121) %>%
  filter(round == "F")
```

```r
winners <- rbind(winners2017, winners2016, winners2015, winners2014, winners2013,
                 winners2012, winners2011, winners2010)
```

```r
hard <- winners %>%
  filter(surface == "Hard",
         best_of == 3,
         !is.na(w_ace),
         !is.na(w_df),
         !is.na(w_svpt),
         !is.na(w_1stIn),
         !is.na(w_1stWon),
         !is.na(w_2ndWon),
         !is.na(w_SvGms),
         !is.na(w_bpSaved),
         !is.na(w_bpFaced)) %>%
  group_by(winner_name) %>%
  mutate(mean_w_ace = mean(w_ace),
         mean_w_df = mean(w_df),
         mean_w_svpt = mean(w_svpt),
         mean_w_1stIn = mean(w_1stIn),
         mean_w_1stWon = mean(w_1stWon),
         mean_w_2ndWon = mean(w_2ndWon),
         mean_w_SvGms = mean(w_SvGms),
         mean_w_bpSaved = mean(w_bpSaved),
         mean_w_bpFaced = mean(w_bpFaced),
         mean_minutes = mean(minutes),
         num = n())

myvars <- names(hard) %in% c("loser_id", "loser_seed", "loser_entry", "loser_name",
                             "loser_hand", "loser_ht", "loser_ioc", "loser_age",
                             "loser_rank", "loser_rank_points", "l_ace", "l_df",
                             "l_svpt", "l_1stIn", "l_1stWon", "l_2ndWon", "l_SvGms",
                             "l_bpSaved", "l_bpFaced")
hard <- hard[!myvars]
```

```r
clay <- winners %>%
  filter(surface == "Clay",
         best_of == 3,
         !is.na(w_ace),
         !is.na(w_df),
         !is.na(w_svpt),
         !is.na(w_1stIn),
         !is.na(w_1stWon),
         !is.na(w_2ndWon),
         !is.na(w_SvGms),
         !is.na(w_bpSaved),
         !is.na(w_bpFaced)) %>%
  group_by(winner_name) %>%
  mutate(mean_w_ace = mean(w_ace),
         mean_w_df = mean(w_df),
         mean_w_svpt = mean(w_svpt),
         mean_w_1stIn = mean(w_1stIn),
         mean_w_1stWon = mean(w_1stWon),
         mean_w_2ndWon = mean(w_2ndWon),
         mean_w_SvGms = mean(w_SvGms),
         mean_w_bpSaved = mean(w_bpSaved),
         mean_w_bpFaced = mean(w_bpFaced),
         mean_minutes = mean(minutes),
         num = n())

myvars <- names(clay) %in% c("loser_id", "loser_seed", "loser_entry", "loser_name",
                             "loser_hand", "loser_ht", "loser_ioc", "loser_age",
                             "loser_rank", "loser_rank_points", "l_ace", "l_df",
                             "l_svpt", "l_1stIn", "l_1stWon", "l_2ndWon", "l_SvGms",
                             "l_bpSaved", "l_bpFaced")
clay <- clay[!myvars]
```

```
grass <- winners %>%
  filter(surface == "Grass",
         best_of == 3,
         !is.na(w_ace),
         !is.na(w_df),
         !is.na(w_svpt),
         !is.na(w_1stIn),
         !is.na(w_1stWon),
         !is.na(w_2ndWon),
         !is.na(w_SvGms),
         !is.na(w_bpSaved),
         !is.na(w_bpFaced)) %>%
  group_by(winner_name) %>%
  mutate(mean_w_ace = mean(w_ace),
         mean_w_df = mean(w_df),
         mean_w_svpt = mean(w_svpt),
         mean_w_1stIn = mean(w_1stIn),
         mean_w_1stWon = mean(w_1stWon),
         mean_w_2ndWon = mean(w_2ndWon),
         mean_w_SvGms = mean(w_SvGms),
         mean_w_bpSaved = mean(w_bpSaved),
         mean_w_bpFaced = mean(w_bpFaced),
         mean_minutes = mean(minutes),
         num = n())

myvars <- names(grass) %in% c("loser_id", "loser_seed", "loser_entry", "loser_name",
                              "loser_hand", "loser_ht", "loser_ioc", "loser_age",
                              "loser_rank", "loser_rank_points", "l_ace", "l_df",
                              "l_svpt", "l_1stIn", "l_1stWon", "l_2ndWon", "l_SvGms",
                              "l_bpSaved", "l_bpFaced")
grass <- grass[!myvars]
```

```
testmodel <- lm(num ~ mean_w_ace + mean_w_df + mean_w_svpt + mean_w_1stIn + mean_w_1stWon + mean_w_2ndW
```

```
backward <- ols_step_backward_aic(testmodel)
```

```
## Backward Elimination Method
## ---------------------------
##
## Candidate Terms:
##
## 1 . mean_w_ace
## 2 . mean_w_df
## 3 . mean_w_svpt
## 4 . mean_w_1stIn
## 5 . mean_w_1stWon
## 6 . mean_w_2ndWon
## 7 . mean_w_SvGms
## 8 . mean_w_bpSaved
## 9 . mean_w_bpFaced
## 10 . mean_minutes
##
##
## Variables Removed:
##
## - mean_w_df
## - mean_w_bpSaved
## - mean_w_1stIn
## - mean_w_bpFaced
##
## No more variables to be removed.
```

Linear Regression

Linear Regression Assumptions

Hypothesis Tests + Confidence Intervals

Conclusion