LinearModel

Huize Zhang

03/12/2018

```
source("../R/ExtractRecordall.R")
source("../R/importance.R")
source("../R/rest.R")
```

Extracting data

dealing with missing distance

```
# Match 1601 (Federer) and 1602 (Nadal) has missing distance of more than 40%
# (1601: 48%, 1602: not available at all) - drop the match
Nadal <- Nadal %>% filter(match_num != 1602)
Federer <- Federer %>% filter(match_num != 1601)
```

Functions to use

```
# clean the data: adding new variables created by functions in R folder
fill_in_miss <- function(dt) {
  dt %>%
    mutate(dist = ifelse(dist ==0, NA, dist)) %>%
    select(dist,match_num,SetNo, GameNo, PointNumber)
  dt_mice <- mice(dt)
  dt_mice <- complete(dt_mice)
  dt <- dt %>% mutate(dist = lag(dt_mice$dist))
}


clean_data<- function(dt, name) {
  dt %>%
    mutate(name = name) %>%
    mutate(rest = as.factor(rest(dt))) %>%
    mutate(impt = point_impt(dt)) %>%
    # filter the double fault points
    filter(Speed_KMH != 0) %>%
    # filter the point serve by the player interested
    filter(ServeIndicator == ifelse(player1 == name, 1,2)) %>%
    select(PointNumber,impt,dist,cum_dist,rest,time,MatchNo,SetNo,
           ServeNumber, name, Speed_KMH, cum_time, RallyCount, Gender)
}

# linear model
fit_lm <- function(data) lm(Speed_KMH~
                              PointNumber+ impt + time + dist + rest + MatchNo + RallyCount +
```

```
                                PointNumber * MatchNo,
                              data = data)
# point number and elasped time are highly correlated - drop one

# mixed effect model
fit_lmer <- function(data) lmer(Speed_KMH~
                                  PointNumber + impt + time + dist + rest + RallyCount+
                                  (1|MatchNo),
                                data = data)

build_linear_model <- function(dt, fit){
  by_player_fit <- dt %>%
    group_by(name) %>%
    nest() %>%
    mutate(model = map(data, fit))
  return(by_player_fit)
}

fetch_coef <- function(dt){
  player_coef_fit <- dt %>%
    unnest(model %>% map(tidy)) %>%
    dplyr::select(name, term,estimate) %>%
    spread(term, estimate)
}
```

Modelling

```
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
```

```
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
```

```
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
```

```
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
```

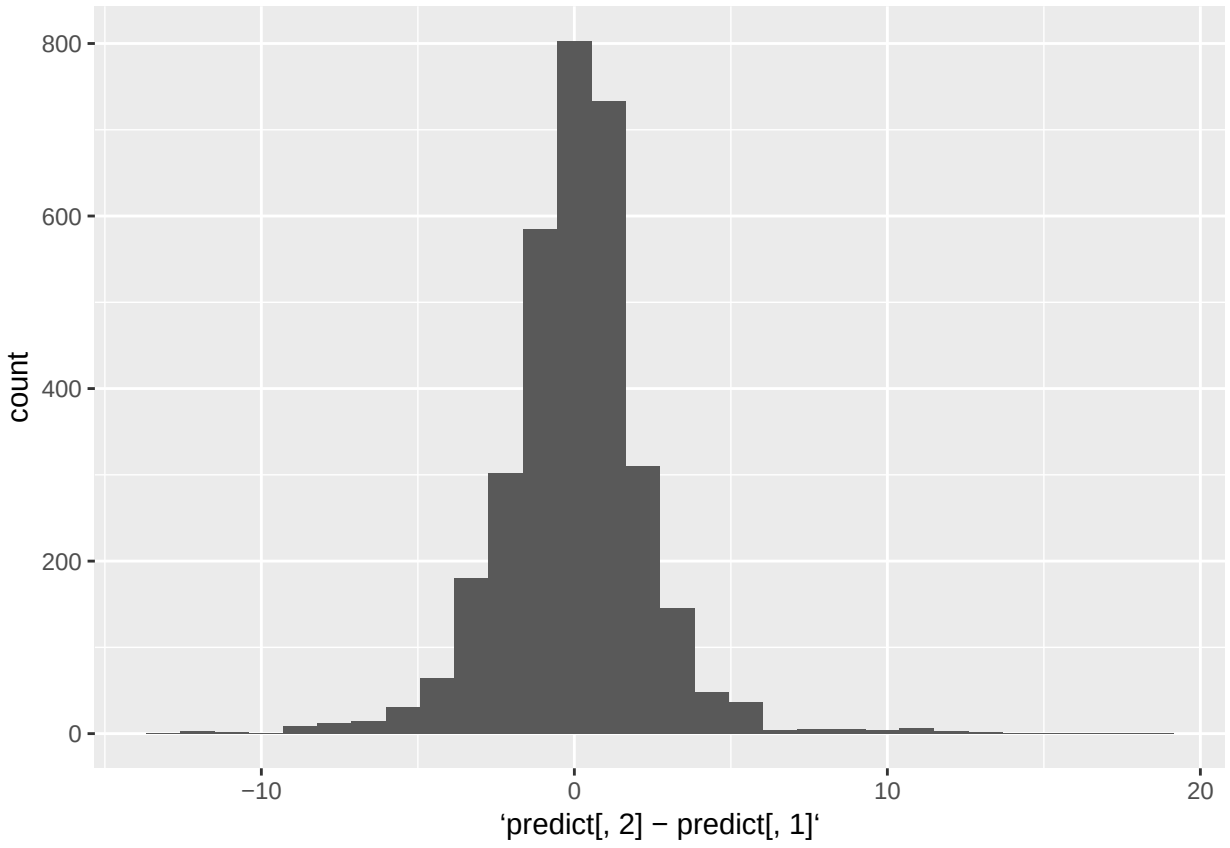
```
##   5   4
##   5   5
##
##  iter imp variable
##   1   1
##   1   2
##   1   3
##   1   4
##   1   5
##   2   1
##   2   2
##   2   3
##   2   4
##   2   5
##   3   1
##   3   2
##   3   3
##   3   4
##   3   5
##   4   1
##   4   2
##   4   3
##   4   4
##   4   5
##   5   1
##   5   2
##   5   3
##   5   4
##   5   5
```

Visualisation

Comparision of lm and lmer

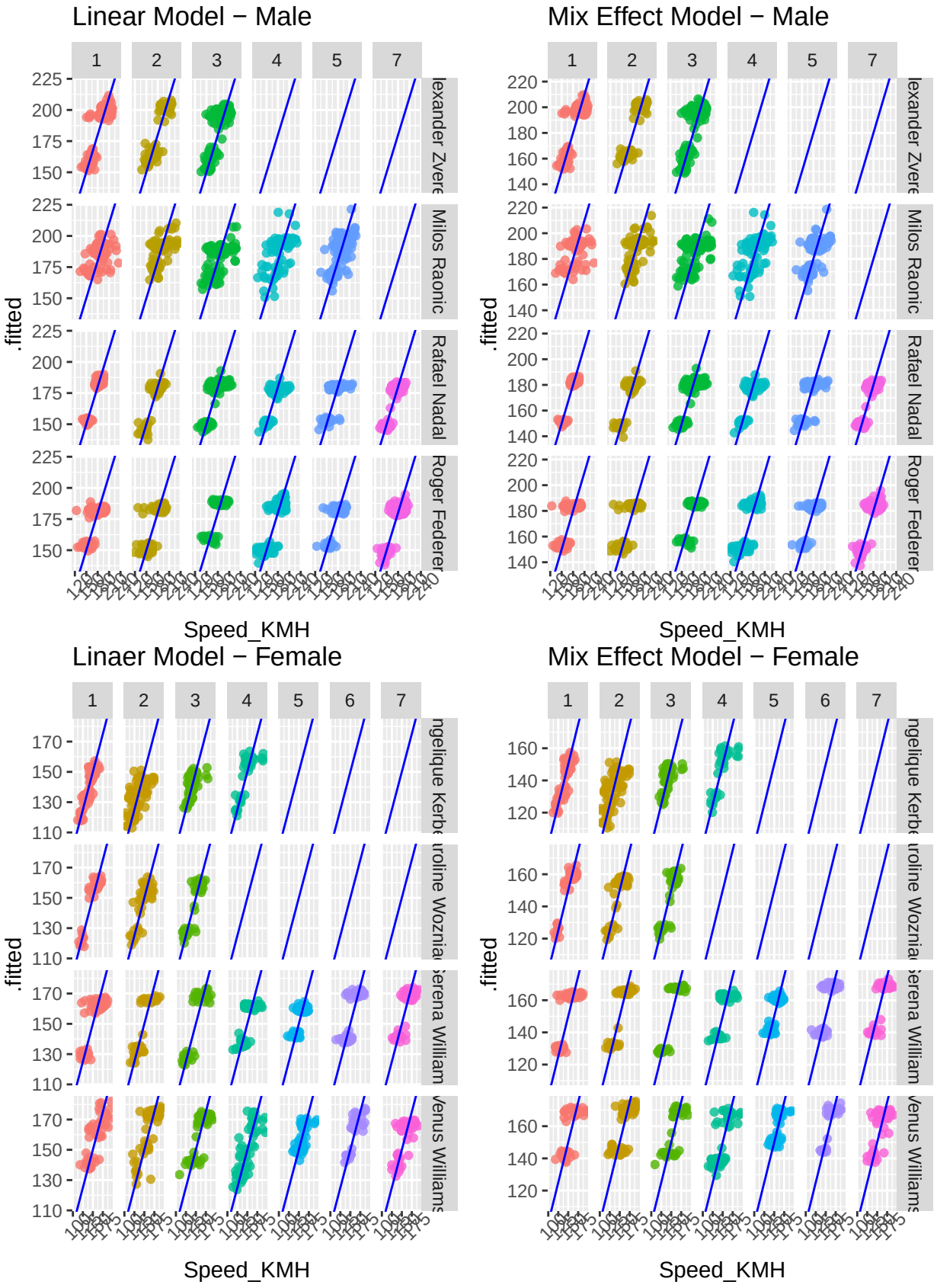
The linear model and mixed effect doesn't differentiate much in a sense that the difference between the predicted value is mostly within -5 and 5 KMH.

```
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



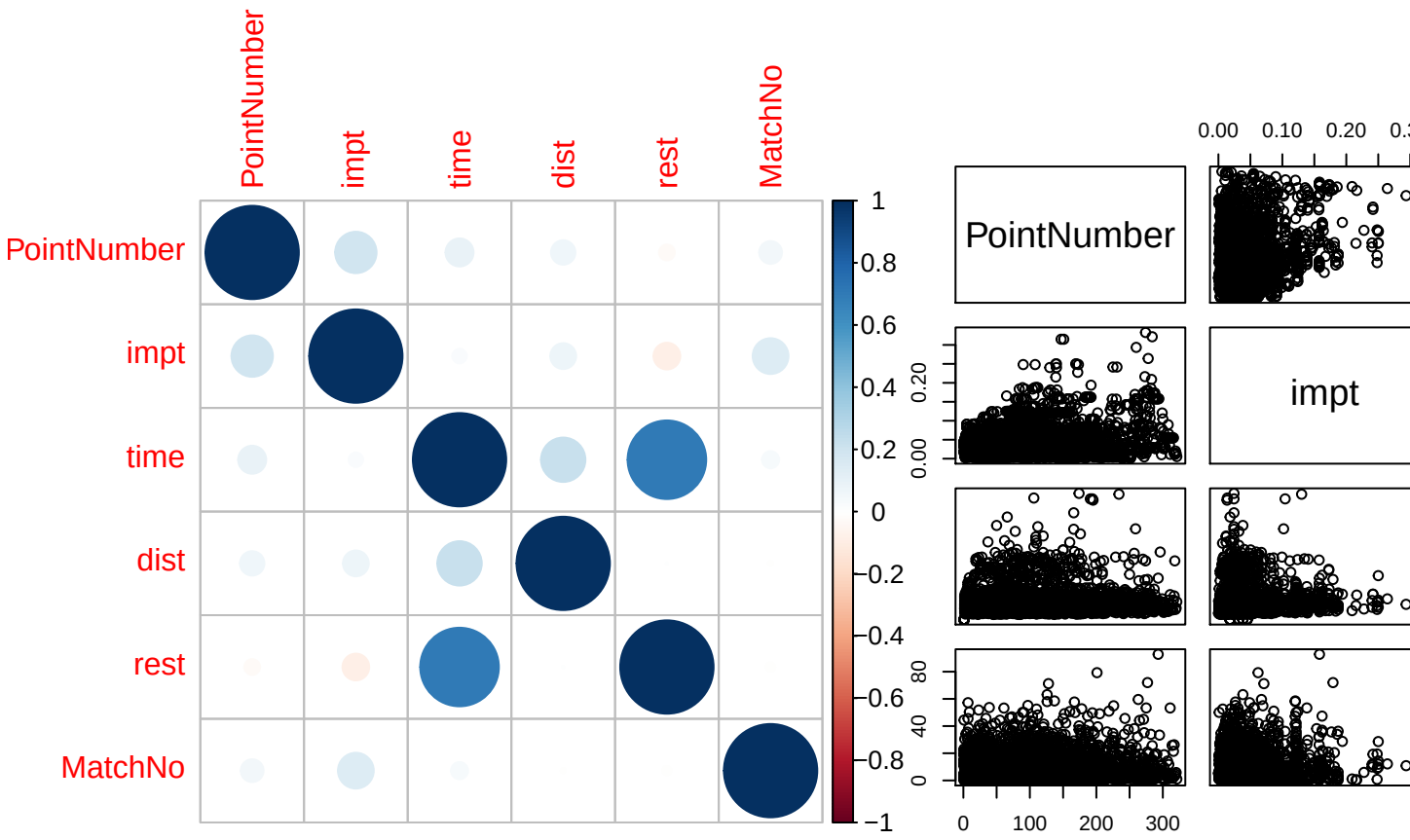
Predicted vs. Actual

This graph contrasts the player's serving speed with the predicted value. The blue line has slope of 1, which is the position where a perfect prediction lays. The two clusters indicates the effectiveness of prediction when seperating the first and second serve in modelling. For Federer, Nadal and Zeverev, the difference between first and second serve is clear as shown in mainly two clusters, while Raonic doesn't seem to have clear cut between the first and second serve speed.


Linear Model − Male
Mix Effect Model − Male
Linaer Model − Female
Mix Effect Model − Female
.fitted
Speed_KMH
1
2
3
4
5
6
7
Alexander Zverev
Milos Raonic
Rafael Nadal
Roger Federer
Angelique Kerber
Caroline Wozniacki
Serena Williams
Venus Williams

Plot explanatory variables against each other

```
## corrplot 0.84 loaded
```



Model Summary: R.squared

```
fit <- rbind(firstserve_fit_lm, secondserve_fit_lm)
fit <- fit %>% dplyr::select(name,r.squared, servenumber) %>% spread(servenumber,r.squared)
kable(fit)
```

Check significance of variables

```
p.value_1<- firstserve_lm %>%
    unnest(model %>% map(tidy)) %>%
    dplyr::select(name, term, p.value) %>%
  spread(term,p.value)
p.value_2 <- secondserve_lm %>%
  unnest(model %>% map(tidy)) %>%
    dplyr::select(name, term, p.value) %>%
  spread(term,p.value)
p.value <- rbind(p.value_1,p.value_2) %>% mutate(servenumber = c(1,1,1,1,1,1,1,1,
                                                                 2,2,2,2,2,2,2,2))
kable(p.value)
```

name	(Intercept)	dist	impt	MatchNo2	MatchNo3	MatchNo4	MatchNo5	MatchNo6
Alexander Zverev	0	0.3131287	0.2970311	0.9448075	0.0790846	NA	NA	NA
Angelique Kerber	0	0.5042417	0.1190708	0.2607037	0.7403625	0.6684184	NA	NA
Caroline Wozniacki	0	0.3062391	0.6695450	0.0044782	0.8110715	NA	NA	NA
Milos Raonic	0	0.0503439	0.9008042	0.4294504	0.5064928	0.9505076	0.1717085	NA
Rafael Nadal	0	0.6435963	0.4106366	0.1165453	0.2912795	0.0490893	0.0847120	NA
Roger Federer	0	0.1752828	0.0081017	0.5539789	0.3773872	0.6814546	0.7869452	NA
Serena Williams	0	0.9006771	0.2752173	0.7594127	0.3814218	0.2241349	0.2293765	0.7488440
Venus Williams	0	0.1133192	0.9412196	0.0134936	0.0228402	0.0068033	0.1984678	0.0120329
Alexander Zverev	0	0.1838996	0.5366250	0.1226468	0.9490110	NA	NA	NA
Angelique Kerber	0	0.0230037	0.4465877	0.0000011	0.0023410	0.7677893	NA	NA
Caroline Wozniacki	0	0.1585270	0.8789684	0.5503272	0.3600802	NA	NA	NA
Milos Raonic	0	0.5453688	0.1609601	0.4950004	0.7316810	0.5033022	0.7483451	NA
Rafael Nadal	0	0.9500617	0.8367484	0.0803482	0.8521181	0.6972094	0.9083540	NA
Roger Federer	0	0.7895568	0.0993990	0.1937457	0.8190459	0.3749221	0.9289361	NA
Serena Williams	0	0.4707976	0.3959605	0.6242775	0.3141227	0.3045262	0.0024004	0.0049559
Venus Williams	0	0.1807503	0.1546060	0.1183494	0.6895065	0.1619024	0.0075297	0.7163794

Effect of each variable in the model

```
firstserve_coef
```

```
## # A tibble: 8 x 21
##   name  `(Intercept)`    dist   impt MatchNo2 MatchNo3 MatchNo4 MatchNo5
##   <chr>         <dbl>   <dbl>  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Alex~          206.  0.0858  27.2     0.326    -6.92   NA        NA
## 2 Ange~          160. -0.0878 -50.6    -6.03     -2.13    2.88     NA
## 3 Caro~          161. -0.101  -19.7   -11.9       1.09   NA        NA
## 4 Milo~          197. -0.230    4.75   -6.12     -4.85    0.492    10.1
## 5 Rafa~          191.  0.0219  13.6    -6.56     -4.05   -8.14     -7.41
## 6 Roge~          184. -0.119   57.8    -3.10      5.03   -1.96      1.47
## 7 Sere~          169. -0.0132  25.9    -2.09      5.37   -5.59     -6.56
## 8 Venu~          158. -0.251    2.59   14.9      16.8    18.7       9.04
## # ... with 13 more variables: MatchNo6 <dbl>, MatchNo7 <dbl>,
## #   PointNumber <dbl>, `PointNumber:MatchNo2` <dbl>,
## #   `PointNumber:MatchNo3` <dbl>, `PointNumber:MatchNo4` <dbl>,
## #   `PointNumber:MatchNo5` <dbl>, `PointNumber:MatchNo6` <dbl>,
## #   `PointNumber:MatchNo7` <dbl>, RallyCount <dbl>, rest1.5 <dbl>,
## #   rest2 <dbl>, time <dbl>
```

```
secondserve_coef
```

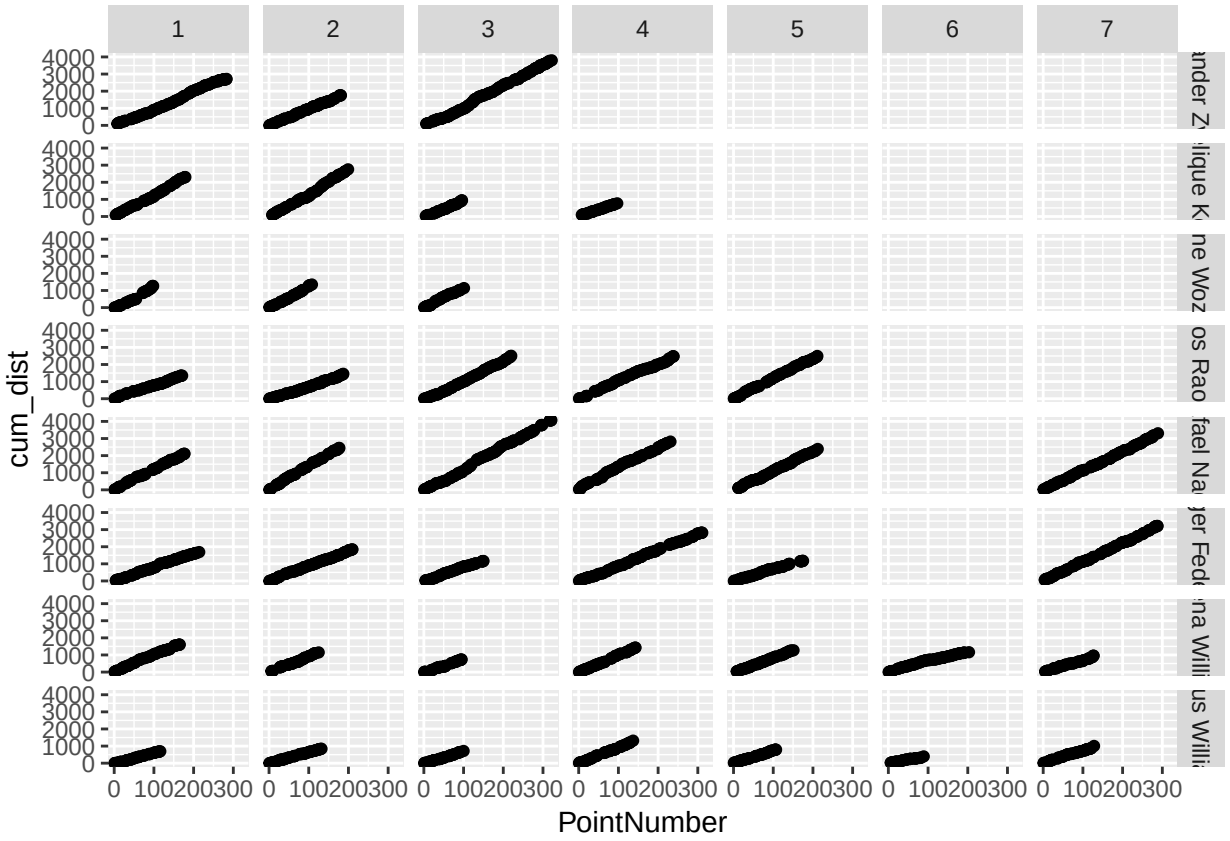
```
## # A tibble: 8 x 21
##   name  `(Intercept)`     dist   impt MatchNo2 MatchNo3 MatchNo4 MatchNo5
##   <chr>         <dbl>    <dbl>  <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 Alex~          167. -0.129   -16.4      7.04    0.249    NA       NA
## 2 Ange~          141. -0.191   -13.9    -16.5   -14.4      -1.28    NA
## 3 Caro~          121.  0.189    -9.80     6.25    7.69     NA       NA
## 4 Milo~          174. -0.113   -69.3      7.26    3.83      7.40     3.37
## 5 Rafa~          155. -0.00477   4.77   -10.8    -1.03     -2.18     0.670
## 6 Roge~          158. -0.0323  -50.4     -7.45    1.49     -4.23    -0.520
## 7 Sere~          125.  0.0699   24.1     -2.74    6.49      5.44    15.4
## 8 Venu~          143.  0.178   -53.5    -11.6     2.92      9.58    17.8
```

```
## # ... with 13 more variables: MatchNo6 <dbl>, MatchNo7 <dbl>,
## #   PointNumber <dbl>, `PointNumber:MatchNo2` <dbl>,
## #   `PointNumber:MatchNo3` <dbl>, `PointNumber:MatchNo4` <dbl>,
## #   `PointNumber:MatchNo5` <dbl>, `PointNumber:MatchNo6` <dbl>,
## #   `PointNumber:MatchNo7` <dbl>, RallyCount <dbl>, rest1.5 <dbl>,
## #   rest2 <dbl>, time <dbl>
```

```
coefficient <- rbind(firstserve_coef,secondserve_coef)%>%
  mutate(servenumber = c(1,1,1,1,1,1,1,1,
                         2,2,2,2,2,2,2,2))
```

correlation

The correlation between Point Number and cummulated match time is high and the plot shows the relationship is linear. Thus it doesn't matter much to use point number or match time as x variable when plotting the models.




```
## [1] 0.9600482
```

Point Number

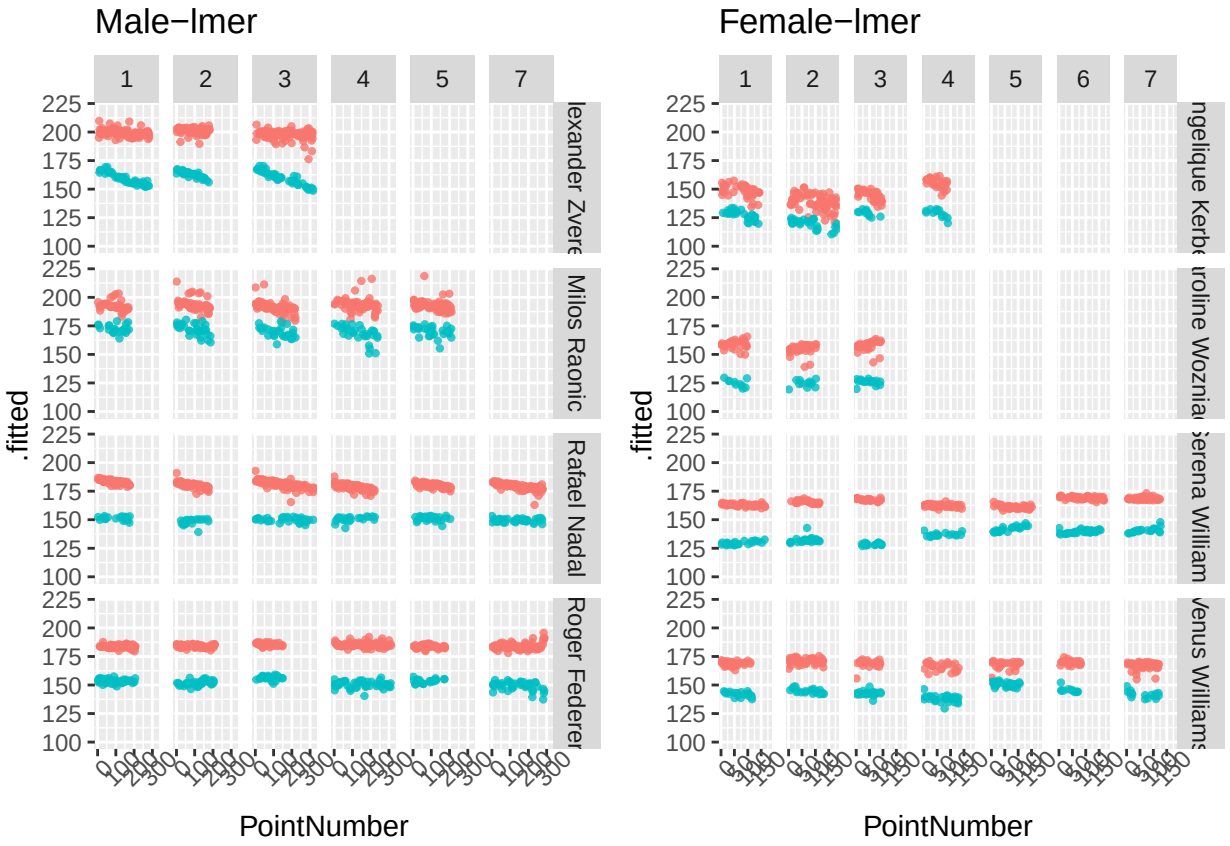
For all players, there's significant difference between first and second serve speed. We can also find that on average, Male's serving is significantly faster than female's.

For the first serve, Zeverev and Raonic have their serving speeds mostly above 175 KMH, which are much higher than those of Nadal or Federer's. Notice that Nadal does have a few fast serve at around 200 KMH.

For the second serve, Zeverev shows clear evidence of reducing of speed for each match and this could be due to the fact of his young age, thus lack of experience or fatigue. For Raonic, we can find some evidence of reduce of serving speed but the variation of serve varies a lot for each match. For Nadal and Federer, whose serving speed is relatively consistent across game, fatigue can be captured by the variation of the serving speed.

Looking at female's data, we could see that Serena and Venus Williams, who played the final game shows a relatively stable serving speed like Federer and Nadal. While Kerber has a similar high variation of serving speed like Raonic and Wozniakic's first serve seems to increase as the game proceeds.

Based on these, we could capture the fatigue through the reduce of the serving speed (slope) as well as the variation of the serving speed (variance). Attention need to be paid to players like Raonic, whose serving speed naturally variates a lot in each match and Wozniakic, whose first serve seems to go against our hypothesis that serving speed will decrease as the game proceeds.



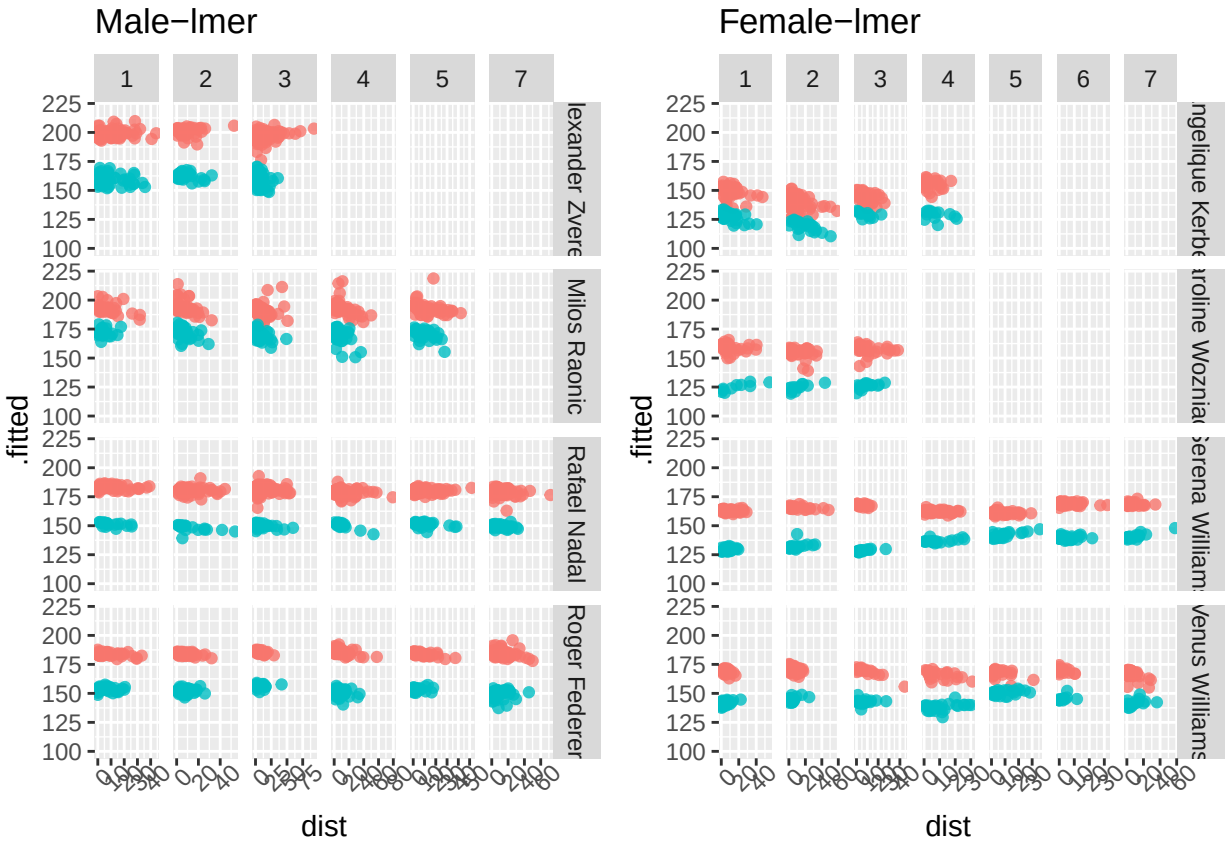
name	1	2
Alexander Zverev	-0.0237823	-0.0507146
Angelique Kerber	-0.0017940	-0.1175797
Caroline Wozniacki	0.0148459	-0.0211741
Milos Raonic	-0.0605693	0.0168979
Rafael Nadal	-0.0460939	0.0149624
Roger Federer	-0.0265033	-0.0090077
Serena Williams	-0.0593490	0.0227047
Venus Williams	0.2033241	0.0339929

Running Distance

Running distance in general doesn't seem to affect much on the male's serving speed, although little evidence (Nadal's second serve) supports that it may reduce the second serving speed.

While for female, running distance seems to be an increasing factor of the second serving speed. We can see that Wozniaki's "seemingly increasing serving speed" in the previous graph is due to the increase of second serve speed. Serena also exhibit this pattern in the second serve.

Another thing to notice is that due to the nature of the female's game (3 games a match rather than 5 games as male's). We observe less data for female than male, thus we would expect female data to have higher variation (i.e. Kerber and Venus)



name	1	2
Alexander Zverev	0.0857788	-0.1293555
Angelique Kerber	-0.0878480	-0.1910633
Caroline Wozniacki	-0.1013958	0.1894563
Milos Raonic	-0.2295072	-0.1126488
Rafael Nadal	0.0218750	-0.0047737
Roger Federer	-0.1192071	-0.0323338
Serena Williams	-0.0132218	0.0698717
Venus Williams	-0.2506983	0.1777874

Point Importance

- In general, point importance doesn't affect the serving speed very much.

- However, as point becomes more importance, player's second serve speed tends to decrease i.e. Raonic and Kerber, which can be viewed as a conservative approach taken by players.

- Also, its interesting to notice that Federer's first serve, serving speeds increase as points become more importance, which shows a more aggressive approach taken by him trying to win the point via an ace serve. This pattern also shows in Serena's second serve

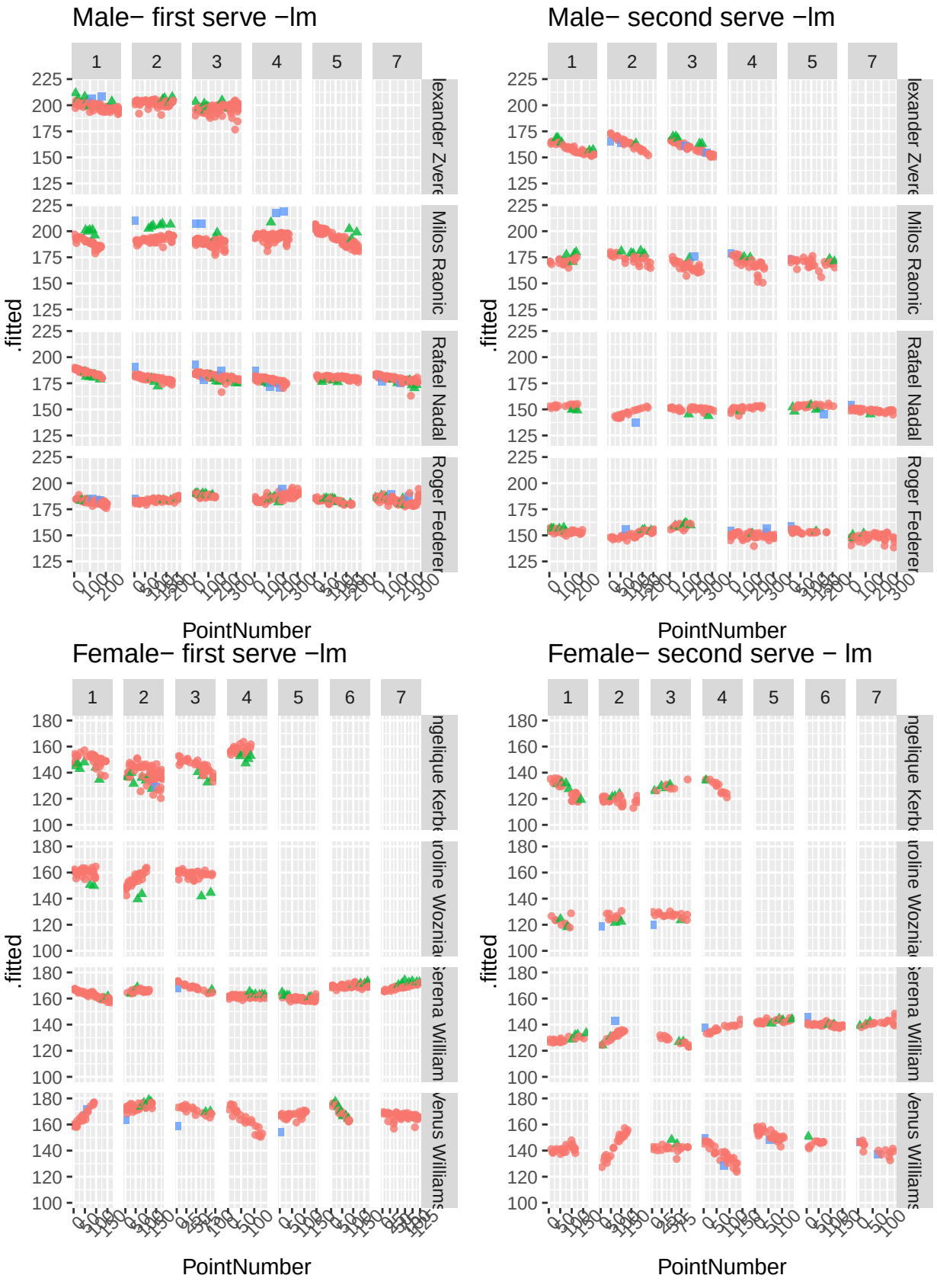
name	1	2
Alexander Zverev	27.237862	-16.375615
Angelique Kerber	-50.609753	-13.944933
Caroline Wozniacki	-19.691040	-9.801391
Milos Raonic	4.749611	-69.320415
Rafael Nadal	13.587819	4.773605
Roger Federer	57.766386	-50.371347
Serena Williams	25.867572	24.147334
Venus Williams	2.587014	-53.511695

Rest

In general, after having the game break, players tends to have higher serving speed, which indicates less fatigue. The improvement for Nadal is marginal while it is more obvious in Zverev's first serve and Federer's Second serve. Raonic's behaviour is interesting in a sense that after each game break, his serving usually drop (green dots) while after the set break, he would have faster serving (blue dotss)

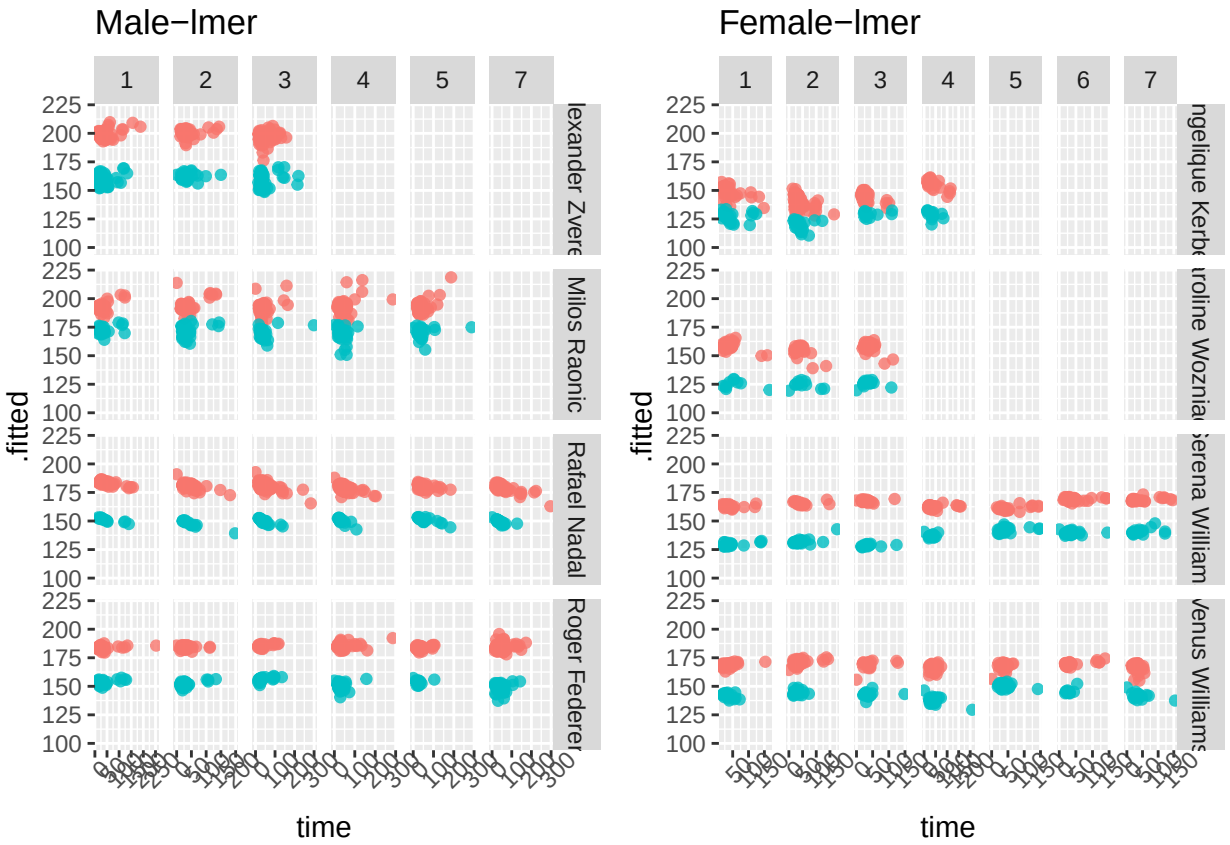
In female players, Kerber and Wozniaki behave similarly with a decrease of serving speed after the scheduled break. While for Serena Williams, she has similar behaviour to top male players with marginal increase of

serving speed after the breaks



Time

In general, as the point is played longer, player's serving speed will decrease marginally (i.e. Nadal). However, for Raonic, it seems to have a positive effect on the serving speed

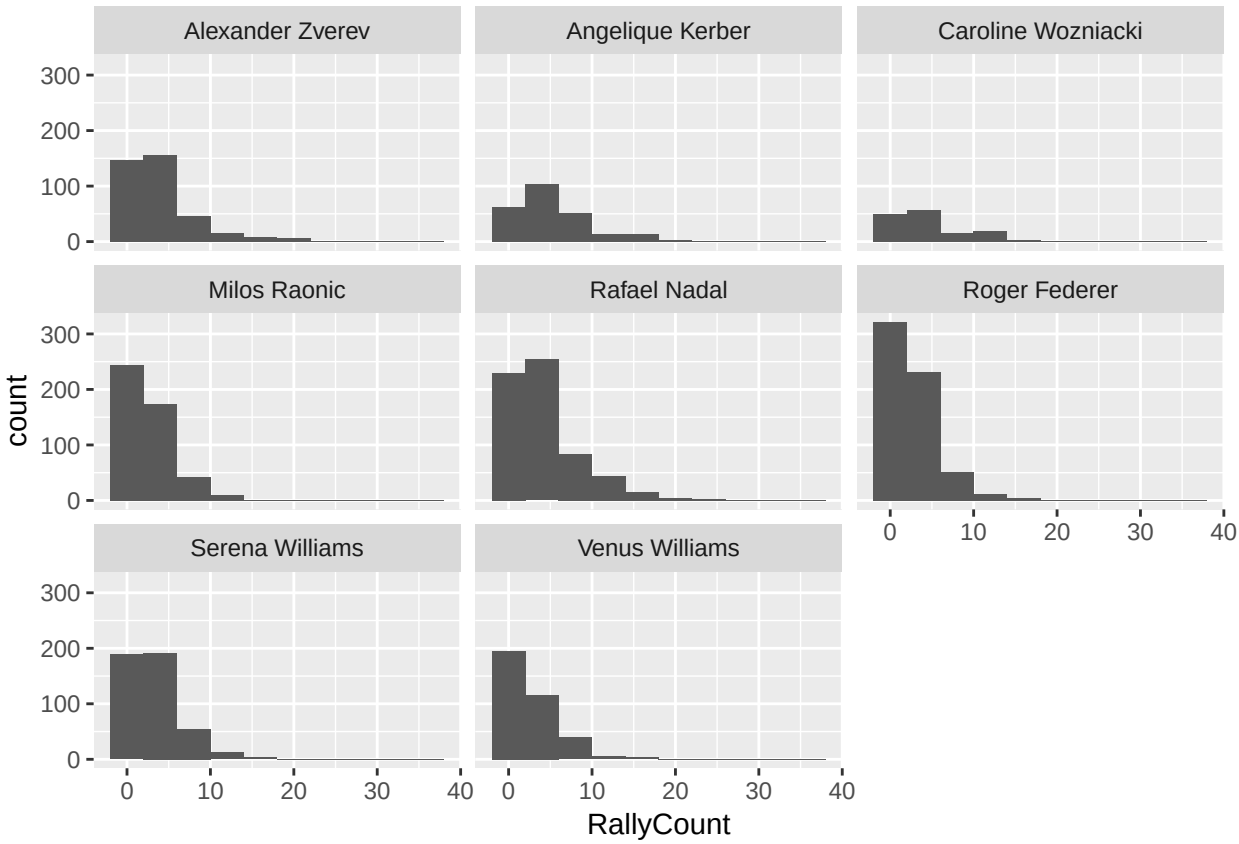


name	1	2
Alexander Zverev	-0.0789898	0.0386050
Angelique Kerber	-0.0443874	0.0112923
Caroline Wozniacki	0.0678250	-0.0033280
Milos Raonic	0.0318582	0.0277208
Rafael Nadal	-0.0570707	-0.0734474
Roger Federer	0.0160707	0.0277107
Serena Williams	-0.0456531	0.0332924
Venus Williams	0.0549028	-0.0275053

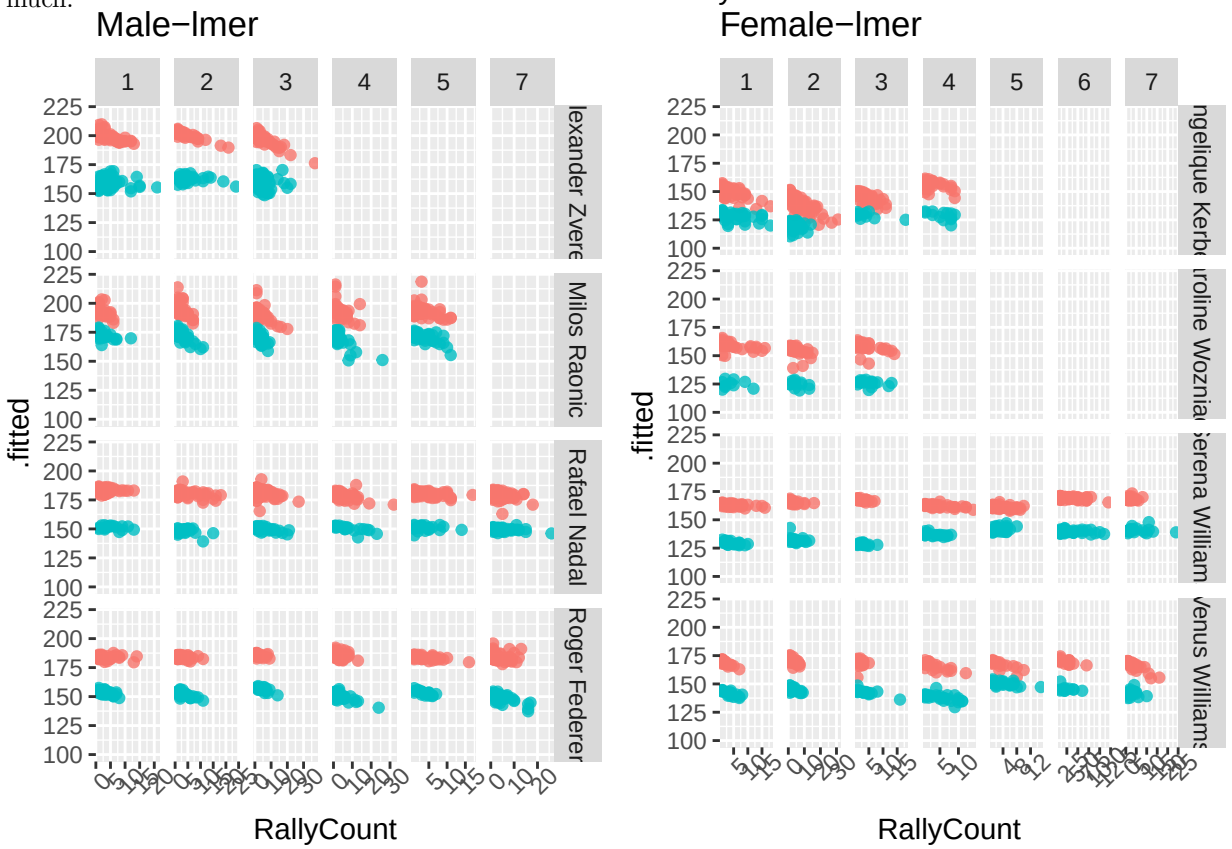
Rally Count

From the first plot, Raonic and Federer play relatively fewer long rally points, while Zverev and Nadal has more long rally points, which may help to understand if rally count would have an effect on fatigue (Serving speed)

The number of rally played in each game is also a factor that would decrease the serving speed and the effect is obvious for the first serve of Zeverev, Kerber, Venus Williams . It is interesting to know that although Nadal and Serena Williams have played a relative number of long rally game, it doesnt seem to affect his serving speed


Alexander Zverev
Angelique Kerber
Caroline Wozniacki
Milos Raonic
Rafael Nadal
Roger Federer
Serena Williams
Venus Williams
count
RallyCount

much.


Male−lmer
Female−lmer
.fitted
RallyCount

name	1	2
Alexander Zverev	-0.7209449	-0.1292412
Angelique Kerber	-0.7502675	-0.2400600
Caroline Wozniacki	-0.4591091	-0.1281165
Milos Raonic	-0.7383053	-0.6922816
Rafael Nadal	-0.2305573	-0.1109862
Roger Federer	-0.2626039	-0.5187073
Serena Williams	-0.2073328	-0.1916860
Venus Williams	-0.6522186	-0.5561305