Violet Chang || chan1300 || 5197617

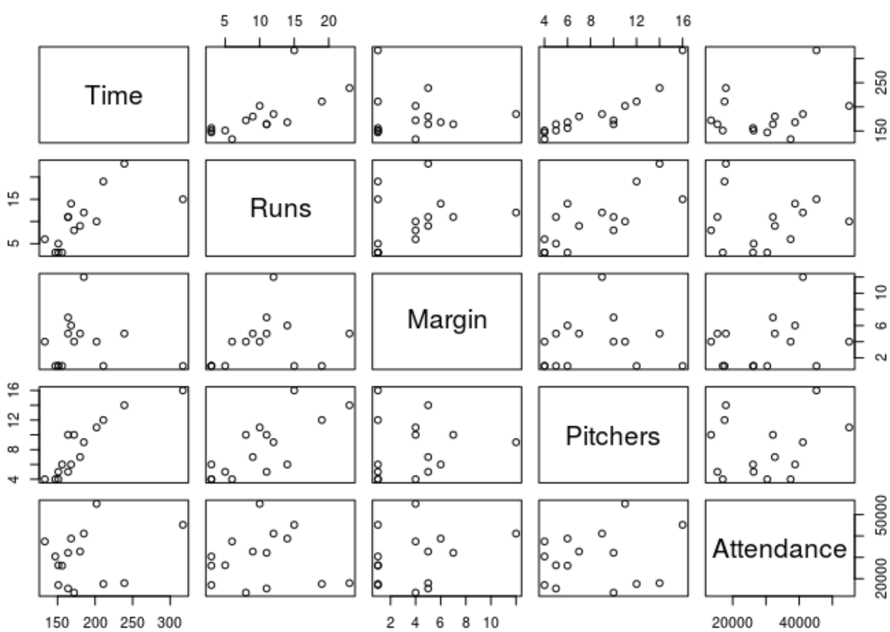
STAT 3022 Homework 4

Problem 1. In R, which diagnostic plot do we use to check for outliers and how do you tell if a point is an outlier? What statistic do we use to check for influential points and how do you tell if a point is an influential point?

- We use residual diagnostic plots, especially Residual vs. Leverage plot to find outliers. Because in the simple linear model setting, an outlier is a point where the magnitude of the residual is unusually large, if a point does not fit the general pattern in the scatterplot of residual diagnostic plot, then it is an outlier. In other word, if the absolute value of standard residual is greater than standard, then this point is a outlier.
- Cook's distance can be used to check influential points. To check if a point is influential point or not, fit the model with and without that point to see if the coefficients change very much. If an Cook's distance is greater than 1, then this point has large influence.

Problem 2. What factors can help to predict how long a Major League Baseball game will last? The datafile BaseballTimes (in the Stat2Data library) contains a sample of 15 games played on August 26, 2008. The variables are: The goal of the study is to predict Time by finding an appropriate set of quantitative predictors.

```
> library(Stat2Data)
> library(leaps)
> library(car)
> data(BaseballTimes)
> pairs(~Time+Runs+Margin+Pitchers+Attendance, data = BaseballTimes)
```



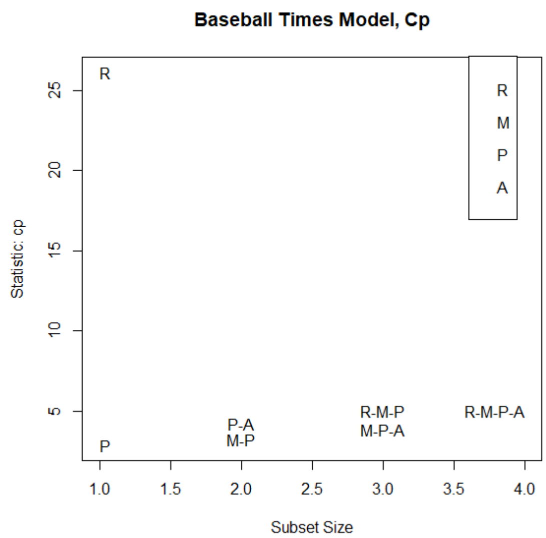
#2 (a): From the plot we know that the relationship between Runs and the Pitchers with Time is more like positive linear, while Margin and Attendance shows non-linear relationship with Time. There seems exists a linear like relationship between Runs and Pitchers.

```
> lm3=lm(Time~Runs+Margin+Pitchers+Attendance, data=BaseballTimes)
> lm4=lm(Time~Runs+Pitchers, data=BaseballTimes)
> summary(lm3)
Call:
lm(formula = Time ~ Runs + Margin + Pitchers + Attendance, data = BaseballTimes)

Residuals:
    Min      1Q  Median      3Q     Max 
-25.755 -11.163  -1.571  13.090  36.716 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 88.0151955 17.7733306   4.952 0.000577 ***
Runs         1.5613428  1.6764324   0.931 0.373612    
Margin      -3.7278867  2.0794572  -1.793 0.103269    
Pitchers     8.7322001  2.4849861   3.514 0.005594 ** 
Attendance   0.0007269  0.0005105   1.424 0.184889    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.77 on 10 degrees of freedom
Multiple R-squared:  0.8557,	Adjusted R-squared:  0.798 
F-statistic: 14.83 on 4 and 10 DF,  p-value: 0.0003299
> summary(lm4)
Call:
lm(formula = Time ~ Runs + Pitchers, data = BaseballTimes)

Residuals:
    Min      1Q  Median      3Q     Max 
-38.025  -8.525  -3.397   9.757  50.518 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 94.87600   13.95551   6.798 1.91e-05 ***
Runs        -0.06436    1.56861  -0.041 0.967948    
Pitchers    10.78571    2.40462   4.485 0.000745 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.33 on 12 degrees of freedom
Multiple R-squared:  0.7998,	Adjusted R-squared:  0.7665 
F-statistic: 23.97 on 2 and 12 DF,  p-value: 6.436e-05
> anova(lm4, lm3)
Analysis of Variance Table

Model 1: Time ~ Runs + Pitchers
Model 2: Time ~ Runs + Margin + Pitchers + Attendance
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     12 5983.4                           
2     10 4312.2  2    1671.2 1.9377 0.1944
```

#2 (b): From the nested anova table, we find that the p-value is 0.19 which is larger than 0.05, so we can’t reject hypothesis H0 that the coefficients for Margin and Attendance are zero (Beta1 = Beta2 = 0). In other words, we don’t have enough evidence to say we can remove the Margin and Attendance simultaneously.

```
> forwardModel = regsubsets(Time~Runs+Margin+Pitchers+Attendance,
```

```
+                    data=BaseballTimes, nbest=2,method="forward")
> par(mfrow=c(1,1))
> subsets(forwardModel, statistic="cp", names=c("R","M","P","A"),
+        main="Baseball Times Model, Cp")
```

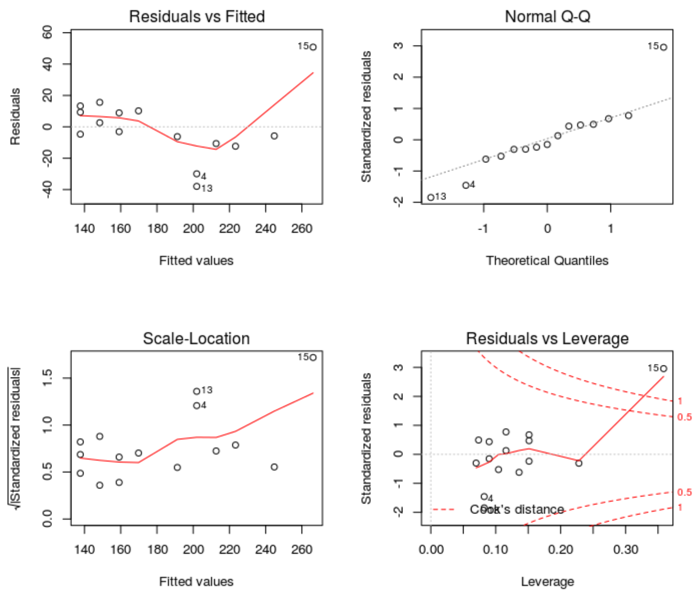



```
> summary(forwardModel)$cp
[1]  2.877398 26.138773  3.222566  4.237234  3.867410  5.027822  5.000000
```

#2 (c): From the cp value and the plot we say that only Pitchers is selected by the forward selection technique, for it has the smallest cp value.

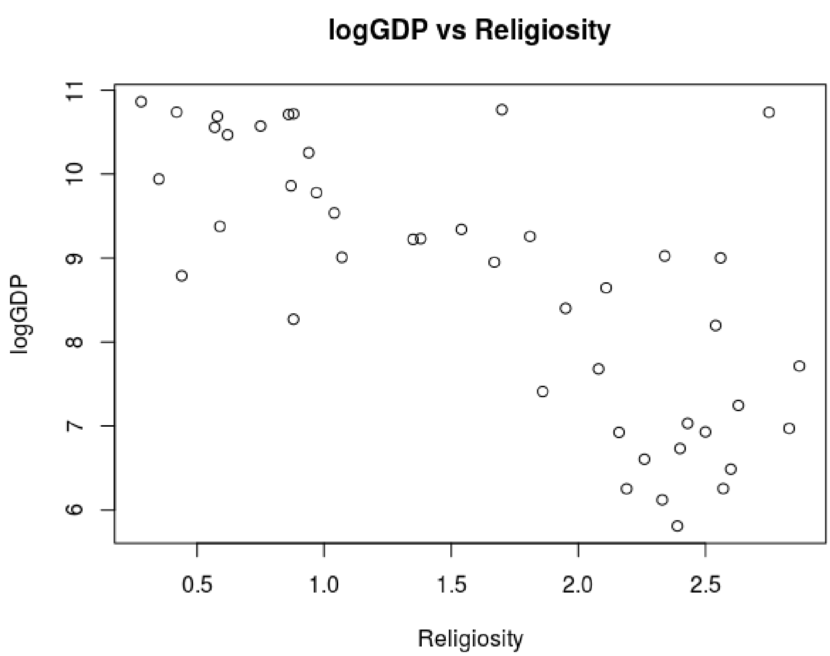
```
> seleModel=lm(Time~Pitchers, data=BaseballTimes)
> summary(seleModel)
Call:
lm(formula = Time ~ Pitchers, data = BaseballTimes)

Residuals:
    Min      1Q  Median      3Q     Max 
-37.945  -8.445  -3.104   9.751  50.794 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  94.843     13.387   7.085 8.24e-06 ***
Pitchers     10.710      1.486   7.206 6.88e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.46 on 13 degrees of freedom
Multiple R-squared:  0.7998,	Adjusted R-squared:  0.7844 
F-statistic: 51.93 on 1 and 13 DF,  p-value: 6.884e-06
> par(mfrow=c(2,2))
> plot(seleModel)
```

#2 (d): In part a, we find out the there is a linear relationship between Pitchers and Time. From the plots, the zero-mean assumption is possibly violated, and the normal q-q plot is not that linearly for some points are outside the diagonal line. The constant variance seems hold. Based on these, I think there is no enough evidence to say that this is violated or not for this is just a small sample.



Problem 4.10

> data(ReligionGDP)
> ReligionGDP$logGDP=log(ReligionGDP$GDP)
> plot(logGDP~Religiosity, data=ReligionGDP, main="logGDP vs Religiosity")



#4.10 (a): From the plot we can find a negative relationship between logGDP and Religiosity. In other words, when Religiosity goes higher, the logGDP will possibly get lower.

> lm1=lm(logGDP~Religiosity, data=ReligionGDP)
> summary(lm1)
Call:

```
lm(formula = logGDP ~ Religiosity, data = ReligionGDP)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8387 -0.8108  0.1272  0.5833  3.5923 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  10.9961     0.3656  30.079  < 2e-16 ***
Religiosity  -1.4013     0.2001  -7.005 1.43e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.085 on 42 degrees of freedom
Multiple R-squared:  0.5388,	Adjusted R-squared:  0.5278 
F-statistic: 49.06 on 1 and 42 DF,  p-value: 1.432e-08
```

#4.10 (b): The summary of this linear model shows that the percentage of the variability is 53.88% based on the R-squared value., which means there is 53.88% of logGDP data can be explained by this model.

#4.10 (c): While religiosity increases 1, the logGDP will decrease 1.4 on the average.

```
> ReligionGDP$EastEurope=as.factor(ReligionGDP$EastEurope)
> ReligionGDP$MiddleEast=as.factor(ReligionGDP$MiddleEast)
> ReligionGDP$Asia=as.factor(ReligionGDP$Asia)
> ReligionGDP$WestEurope=as.factor(ReligionGDP$WestEurope)
> ReligionGDP$Americas=as.factor(ReligionGDP$Americas)
> lm2=lm(logGDP~Religiosity+EastEurope+MiddleEast+Asia+WestEurope+Americas, data=ReligionGDP)
> summary(lm2)
Call:
lm(formula = logGDP ~ Religiosity + EastEurope + MiddleEast + 
    Asia + WestEurope + Americas, data = ReligionGDP)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5274 -0.5720 -0.0760  0.5457  2.3395 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.2019     0.7452  12.348 1.09e-14 ***
Religiosity  -0.9979     0.2852  -3.498  0.00124 ** 
EastEurope1   0.7901     0.6709   1.178  0.24639    
MiddleEast1   1.9374     0.4797   4.039  0.00026 ***
Asia1         0.9856     0.4556   2.163  0.03706 *  
WestEurope1   2.0538     0.6975   2.944  0.00556 ** 
Americas1     1.5937     0.4778   3.336  0.00195 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8947 on 37 degrees of freedom
Multiple R-squared:  0.7235,	Adjusted R-squared:  0.6787 
F-statistic: 16.14 on 6 and 37 DF,  p-value: 5.095e-09
```

#4.10 (e): The summary of this linear model shows that the percentage of the variability is 72.35% based on the R-squared value., which means there is 72.35% of logGDP data can be explained by this model.

#4.10 (f): While religiosity increases 1, the logGDP will decrease 0.9979 on the average.

```
> anova (lm2, lm1)
Analysis of Variance Table

Model 1: logGDP ~ Religiosity + EastEurope + MiddleEast + Asia + WestEurope +
    Americas
Model 2: logGDP ~ Religiosity
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     37 29.615
2     42 49.405 -5    -19.79 4.9449 0.001448 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#4.10 (g): The nested F-test statistic value is 4.94 and the p-value is 0.0014 < 0.05. So we have the strong evidence to say that we reject the hypothesis H0 that Beta2 = Beta3 = Beta4 = Beta5 = Beta6 = 0. We can conclude that the five religions are important factor for predicting logGDP and the new model is better than before.

Problem 5.12:

a) Because an explanatory variable is one that explains changes in that variable, for here, the response variable here is the final score, and the explanatory variable here are the four different fonts.

b) This is a randomized experiment as we assign 40 students to 4 fonts groups randomly.

c)The one-way analysis of variance (ANOVA) is used to determine whether there are any statistically significant differences between the means of three or more independent groups. Because there are four different fonts here, we will use ANOVA to compare and analysis the data.

Problem 5.20:

a)

Source	DF	SS	MS	F
Type	K-1=2	37.51	37.51/2=18.76	18.76/3.68=5.10
Error	N-K=9	33.09	33.09/9=3.68	
Total	N-1=11	70.60		

b) The MS for county type tells me the variability of the mean types, so I can compare the MS with the variability of error to find out if there is more variability.

c)

```
> 1-pf(5.1,2,9)
[1] 0.03305489
```

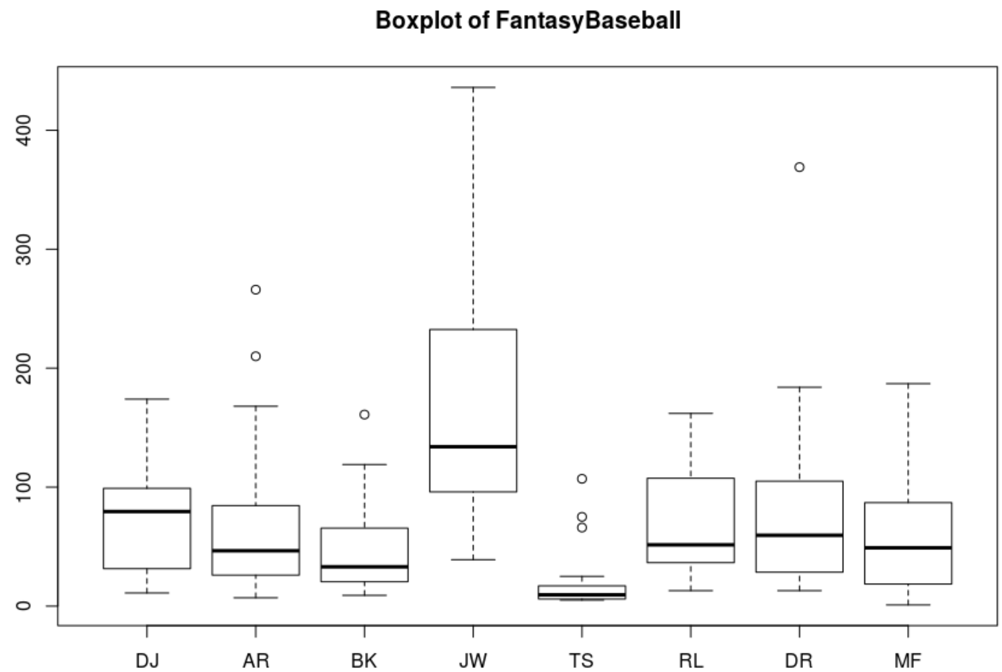
So, the p value 0.033

d)H_0: Beta1 = Beta2 = Beta3 = 0; H_1: Beta1 != Beta2 != Beta3 != 0, I.e, at least one of county’s size is different from others.

Since the p value is 0.033 < 0.05, we say that we reject H0, which means there is at least one county’s sizes is different from others.

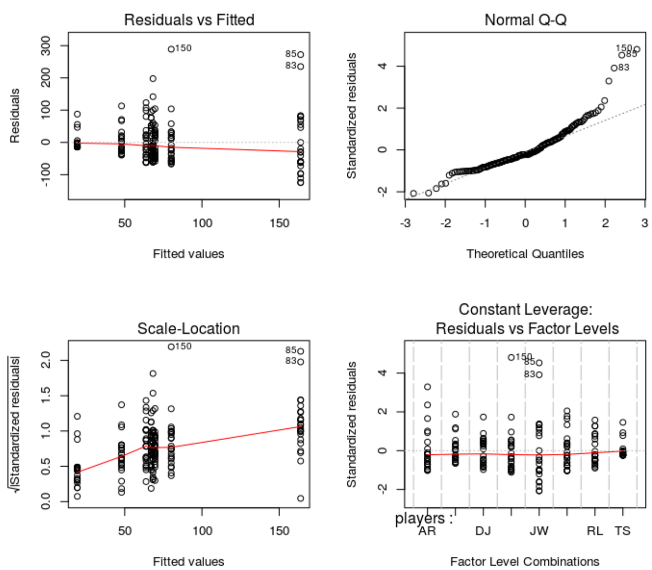
Problem 5.24:

```
> data(FantasyBaseball)
> boxplot(FantasyBaseball[,2:9], main="Boxplot of FantasyBaseball")
```




```
#5.24 (a): This boxplot shows that the most of the distribution is skewed to the right. IT is
easy to find out that there are some decisions are harder to make which will cost more time.
Generally, the time participants take are same while JW is the slowest and TS is the fastest in
making decisions.

> dim(FantasyBaseball)
[1] 24  9
> times=with(FantasyBaseball, c(DJ,AR,BK,JW,TS,RL,DR,MF))
> players=rep(colnames(FantasyBaseball)[2:9], each=24)
> newData=data.frame("players"=as.factor(players), "times"=times)
> lm5=lm(times~players, data=newData)
> summary(lm5)
Call:
lm(formula = times ~ players, data = newData)

Residuals:
    Min      1Q  Median      3Q     Max
-124.88  -36.91  -13.33   24.47  288.88

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    68.292     12.528   5.451 1.59e-07 ***
playersBK     -20.333     17.718  -1.148   0.2526
playersDJ       1.333     17.718   0.075   0.9401
playersDR      11.833     17.718   0.668   0.5050
playersJW      95.583     17.718   5.395 2.09e-07 ***
playersMF      -4.458     17.718  -0.252   0.8016
playersRL      -1.167     17.718  -0.066   0.9476
playersTS     -48.958     17.718  -2.763   0.0063 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

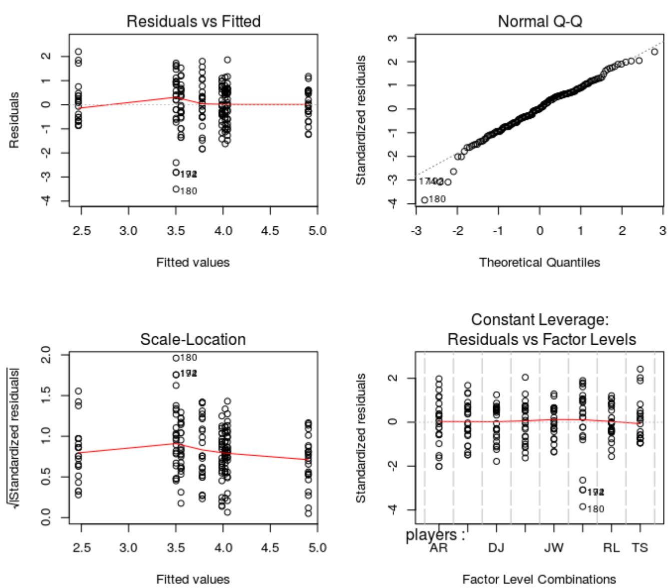
```
Residual standard error: 61.38 on 184 degrees of freedom
Multiple R-squared:  0.293,     Adjusted R-squared:  0.2661
F-statistic: 10.89 on 7 and 184 DF,  p-value: 1.788e-11
> anova(lm5)
Analysis of Variance Table

Response: times
           Df Sum Sq Mean Sq F value    Pr(>F)
players     7 287196   41028  10.891 1.788e-11 ***
Residuals 184 693126    3767
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> an1=aov(times~players)
> summary(an1)
             Df Sum Sq Mean Sq F value   Pr(>F)
players       7 287196   41028   10.89 1.79e-11 ***
Residuals   184 693126    3767
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> an1=aov(times~players, data=newData)
> par(mfrow=c(2,2))
> plot(an1)
```



#5.24 (b): The plots shows that this model is not that acceptable for from normal q-q plot we see a obvious curve pattern and the zero mean assumption doesn’t not hold according to the residuals vs fitted value plot. So we need to correct the model and can’t reject any hypothesis because the anova and summary table can’t be trusted.

```
> newData$log=log(newData$times)
> lm6=lm(log~players, data=newData)
> summary(lm6)
Call:
lm(formula = log ~ players, data = newData)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5026 -0.5643  0.0110  0.5941  2.2040

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.7798     0.1900  19.896  < 2e-16 ***
playersBK     -0.2254     0.2687  -0.839    0.403
playersDJ      0.2427     0.2687   0.904    0.367
```

```
playersDR     0.2672     0.2687   0.995    0.321
playersJW     1.1233     0.2687   4.181 4.48e-05 ***
playersMF    -0.2772     0.2687  -1.032    0.304
playersRL     0.2078     0.2687   0.773    0.440
playersTS    -1.3109     0.2687  -4.879 2.30e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9307 on 184 degrees of freedom
Multiple R-squared:  0.3307,    Adjusted R-squared:  0.3053
F-statistic: 12.99 on 7 and 184 DF,  p-value: 1.538e-13
> anova(lm6)
Analysis of Variance Table

Response: log
           Df Sum Sq Mean Sq F value    Pr(>F)
players     7  78.75 11.2500  12.989 1.538e-13 ***
Residuals 184 159.37  0.8661
---
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> an2=aov(log~players, data=newData)
> summary(an2)
             Df Sum Sq Mean Sq F value   Pr(>F)
players       7  78.75  11.250   12.99 1.54e-13 ***
Residuals   184 159.37   0.866
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow=c(2,2))
> plot(an2)
```



#5.24 (c): The plots shows that this model is better than before because the normal q-q plot shows a more similar linear pattern than previous one and the other three plots also shows better pattern. This time, the assumption can stand and the anova table is more believeable.

#5.24 (d): The F-statistic value is 12.99, p-value is approximately 0, which indicate that we reject the hypothesis H0 that all coefficient values (beta) are equal to zero. In other words, we have the strong evidence to show that there is at least one participant’s mean selection time is different from others.

```
> source("LSDtest.R")
> lsd=LSDtest(newData$log, newData$players, alpha=0.05)
```

```
> lsd
     pair         est         lwr         upr
1  AR~BK  0.22537843 -0.30467454  0.75543139
2  AR~DJ -0.24274982 -0.77280279  0.28730315
3  AR~DR -0.26720058 -0.79725355  0.26285239
4  AR~JW -1.12326875 -1.65332172 -0.59321578
5  AR~MF  0.27720147 -0.25285150  0.80725443
6  AR~RL -0.20778389 -0.73783686  0.32226908
7  AR~TS  1.31090586  0.78085289  1.84095883
8  BK~DJ -0.46812824 -0.99818121  0.06192473
9  BK~DR -0.49257901 -1.02263198  0.03747396
10 BK~JW -1.34864717 -1.87870014 -0.81859420
11 BK~MF  0.05182304 -0.47822993  0.58187601
12 BK~RL -0.43316231 -0.96321528  0.09689066
13 BK~TS  1.08552744  0.55547447  1.61558041
14 DJ~DR -0.02445076 -0.55450373  0.50560221
15 DJ~JW -0.88051893 -1.41057190 -0.35046596
16 DJ~MF  0.51995128 -0.01010169  1.05000425
17 DJ~RL  0.03496593 -0.49508704  0.56501890
18 DJ~TS  1.55365568  1.02360271  2.08370865
19 DR~JW -0.85606817 -1.38612114 -0.32601520
20 DR~MF  0.54440205  0.01434908  1.07445502
21 DR~RL  0.05941669 -0.47063627  0.58946966
22 DR~TS  1.57810645  1.04805348  2.10815942
23 JW~MF  1.40047021  0.87041724  1.93052318
24 JW~RL  0.91548486  0.38543189  1.44553783
25 JW~TS  2.43417461  1.90412164  2.96422758
26 MF~RL -0.48498535 -1.01503832  0.04506762
27 MF~TS  1.03370440  0.50365143  1.56375737
28 RL~TS  1.51868975  0.98863678  2.04874272
#5.24 (e): From the Fisher’s LSD procedure, we know that:
JW’s and TS’s average selection times is significantly different from any of other player,
DR’s is different from MF.
```