STATS707Assignment4

Syed Muhammad Adeel Ibrahim

June 5, 2019

Question 1

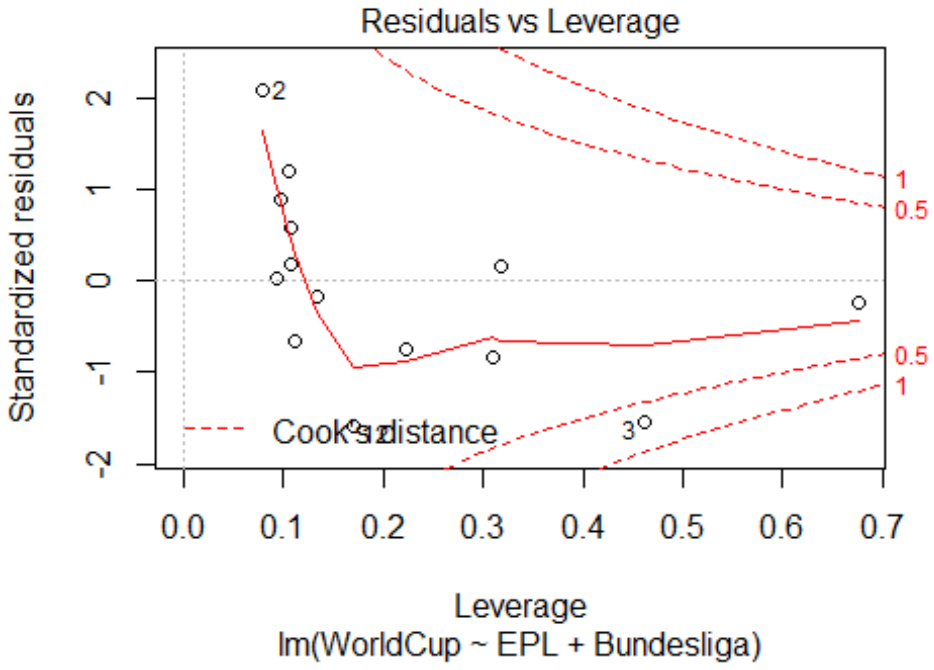
Read in the data from soccer.txt (hint, use the following command:)

read.table("soccer.txt", header=TRUE) -> soccer ####a) Our goal is to consider models that predict the average goals per game in the World Cup from the average goals per game in the largest European leagues in the preceding year. You are first asked to consider a model using the EPL and the Bundesliga. What is the p-value associated with the F-statistic? How do you interpret this?

```
soccer <- read.table("soccer.txt", header=TRUE)

eplNBundesliga.lm <- lm(formula = WorldCup ~ EPL + Bundesliga, data = soccer)
summary(eplNBundesliga.lm)

##
## Call:
## lm(formula = WorldCup ~ EPL + Bundesliga, data = soccer)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.26467 -0.11929 -0.00940  0.08374  0.36629
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3949     0.9669   0.408   0.6908
## EPL           0.2690     0.2868   0.938   0.3685
## Bundesliga    0.4836     0.1804   2.681   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1839 on 11 degrees of freedom
## Multiple R-squared:  0.416,  Adjusted R-squared:  0.3098
## F-statistic: 3.917 on 2 and 11 DF,  p-value: 0.05194

##P-value is just above .05194 which could be consider reasonable small here
which means our data doesn't suppot our null hypothesis, it is enough evident
to reject it.
```

b) Make the residuals vs leverage plot for this model. You will see one point with Cooks distance larger than 0.5. What does this mean? What are the EPL, Bundesliga and World cup goals for this point? What effect is this point having on the model?

```
plot(eplNBundesliga.lm, which = 5)
```




```
#Residual vs Leverage shows the impact of each point on the model, in the
given figure 3 is the point that heavily influencing the current model, by
theory every point should be less .5 cook's distance to qualify an ordinary
participation. over here point 3 is heavily participating which means if we
remove this point it will also change the model.
```

c) Find a point estimate, confidence and predition interval for average World Cup goals if the EPL and Bundesliga both average 4 goals per game. Is it sensible to use this model for this sort of prediction? Explain.

```
newdata = data.frame(EPL=4, Bundesliga=4)

predict(eplNBundesliga.lm, newdata)

##        1
## 3.405318

predict(eplNBundesliga.lm, newdata, interval = "confidence")

##        fit      lwr      upr
## 1 3.405318 2.462119 4.348518

predict(eplNBundesliga.lm, newdata, interval = "prediction")
```

```
##        fit     lwr      upr
## 1 3.405318 2.37892 4.431716

#its not a good idea to predict result through such models, since we are
unable to identify importance and irrelevance of multiple variables,
therefore it is wise to use recommend method to identity relevance of
variables for prediction
```

d) It would also be possible to also include the La Liga and Serie A competitions in the model (please ignore the incomplete total and Champions league variables. Using stepwise selection and your chosen information criteria, try to come up with a better model. Report your findings, including diagnostics and any concerns.

```
allLeague.lm <- lm(formula = WorldCup ~ EPL + Bundesliga + LaLiga + SerieA,
data = soccer)

# backward Model
allLeague.modBackward <- step(allLeague.lm)

## Start:  AIC=-40.84
## WorldCup ~ EPL + Bundesliga + LaLiga + SerieA
##
##              Df Sum of Sq     RSS     AIC
## - SerieA      1  0.000343 0.37108 -42.825
## - LaLiga      1  0.001353 0.37209 -42.787
## - EPL         1  0.022849 0.39359 -42.001
## <none>                    0.37074 -40.838
## - Bundesliga  1  0.124683 0.49542 -38.780
##
## Step:  AIC=-42.83
## WorldCup ~ EPL + Bundesliga + LaLiga
##
##              Df Sum of Sq     RSS     AIC
## - LaLiga      1  0.001011 0.37209 -44.787
## - EPL         1  0.023896 0.39498 -43.952
## <none>                    0.37108 -42.825
## - Bundesliga  1  0.234190 0.60527 -37.976
##
## Step:  AIC=-44.79
## WorldCup ~ EPL + Bundesliga
##
##              Df Sum of Sq     RSS     AIC
## - EPL         1  0.029748 0.40184 -45.711
## <none>                    0.37209 -44.787
## - Bundesliga  1  0.243153 0.61525 -39.747
##
## Step:  AIC=-45.71
## WorldCup ~ Bundesliga
##
##              Df Sum of Sq     RSS     AIC
```

```
## <none>                     0.40184 -45.711
## - Bundesliga  1   0.23525 0.63709 -41.259

allLeague.startmod <- lm(WorldCup ~ 1, data = soccer)

# Forward Model
allLeague.modForward <- step(allLeague.startmod, scope= list(upper = WorldCup
~ EPL + Bundesliga + LaLiga + SerieA, lower= WorldCup ~ 1),
direction="forward")

## Start:  AIC=-41.26
## WorldCup ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + Bundesliga  1  0.235252 0.40184 -45.711
## <none>                     0.63709 -41.259
## + SerieA      1  0.076574 0.56052 -41.051
## + EPL         1  0.021847 0.61525 -39.747
## + LaLiga      1  0.019373 0.61772 -39.691
##
## Step:  AIC=-45.71
## WorldCup ~ Bundesliga
##
##          Df Sum of Sq     RSS     AIC
## <none>                 0.40184 -45.711
## + EPL     1 0.0297482 0.37209 -44.787
## + LaLiga  1 0.0068640 0.39498 -43.952
## + SerieA  1 0.0000252 0.40182 -43.711

# Both Model
allLeague.modBoth <- step(allLeague.startmod, scope= list(upper = WorldCup ~
EPL + Bundesliga + LaLiga + SerieA, lower= WorldCup ~ 1), direction="both")

## Start:  AIC=-41.26
## WorldCup ~ 1
##
##              Df Sum of Sq     RSS     AIC
## + Bundesliga  1  0.235252 0.40184 -45.711
## <none>                     0.63709 -41.259
## + SerieA      1  0.076574 0.56052 -41.051
## + EPL         1  0.021847 0.61525 -39.747
## + LaLiga      1  0.019373 0.61772 -39.691
##
## Step:  AIC=-45.71
## WorldCup ~ Bundesliga
##
##              Df Sum of Sq     RSS     AIC
## <none>                     0.40184 -45.711
## + EPL         1  0.029748 0.37209 -44.787
## + LaLiga      1  0.006864 0.39498 -43.952
```

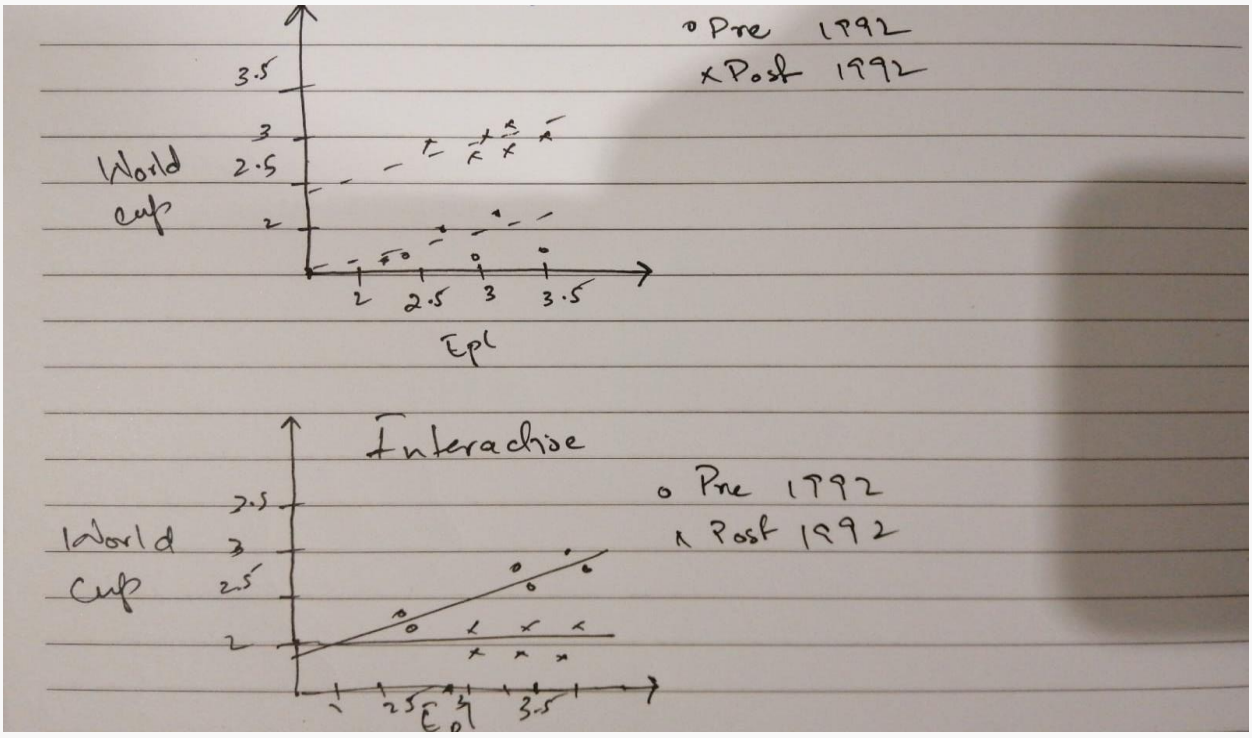
```
## + SerieA      1  0.000025 0.40182 -43.711
## - Bundesliga  1  0.235252 0.63709 -41.259

# While looking into all the stepwise types it seems like predictive models
are same with AIC of -45.71
```

e) Someone suggests that the start up of the Champions League format in 1992 affected the goals scored in the individual countries leagues, and therefore the relationship with the world cup. Consider predicting the world cup goals with a single league-sketch data that would indicate a main effect of the champions league, but no interaction. Make another sketch of data that would indicate a significant interaction. (Do not worry about making your sketch agree with the observed data for a particular league.) You may either make an electronic drawing, or take a picture of each sketch to embed in your assignment.

```
championsnoNointeract.lm <- lm(WorldCup~EPL+Champions, data=soccer)
summary(championsnoNointeract.lm)

##
## Call:
## lm(formula = WorldCup ~ EPL + Champions, data = soccer)
##
## Residuals:
##        8        9       10       11       12
##  0.27135 -0.01925  0.02071 -0.17893 -0.09389
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1929     2.7724   1.152    0.369
## EPL          -0.9995     1.2629  -0.791    0.512
## Champions     0.7740     0.7184   1.077    0.394
##
## Residual standard error: 0.2401 on 2 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.378,  Adjusted R-squared:  -0.2439
## F-statistic: 0.6078 on 2 and 2 DF,  p-value: 0.622

championsInteract.lm <- lm(WorldCup~EPL*Champions, data=soccer)
summary(championsInteract.lm)

##
## Call:
## lm(formula = WorldCup ~ EPL * Champions, data = soccer)
##
## Residuals:
##        8        9       10       11       12
##  0.15952  0.03564 -0.14060 -0.04300 -0.01157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -90.58      79.92  -1.133    0.460
## EPL              34.40      30.18   1.140    0.458
```

```
## Champions         40.22      33.61   1.197    0.443
## EPL:Champions    -14.87      12.67  -1.174    0.449
##
## Residual standard error: 0.2202 on 1 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.7385, Adjusted R-squared:  -0.04616
## F-statistic: 0.9412 on 3 and 1 DF,  p-value: 0.6215
```

```
# it is visible that AIC now dropped to 21.02 for entries with champions Leage
```



Question 2

The following table gives the number of full time and part time academic employees of different ranks at the University of Michigan.

| Full Time | Part Time

——————————————————

Assistant Professor | 716 | 83

Associate Professor | 727 | 86

Full Professor | 1226 | 330

a) Is Full Time/Part Time status independent of rank? Give your computer output and state your conclusion.

```
entries <- c(716 , 83, 727, 86, 1226, 330)

x <- matrix(entries, ncol=2, byrow=TRUE)
chisq.test(x)
```

```
##
##  Pearson's Chi-squared test
##
## data:  x
## X-squared = 68.631, df = 2, p-value = 1.25e-15
```

```
# since pvalue < 0.0001 therefore Rank and fulltime/parttime roles are not independent
```

b) Show how the degrees of freedom for the test is computed.

```
#df = (Colunm - 1) X (Row - 1)
(2 - 1) * (3 - 1)
```

```
## [1] 2
```

c) Show how the expected value for full time Associate Professors is computed.

```
total <- sum(entries)
PFullTime <- (716 + 727 + 1226)/total
PAssosiate <- (727 + 86)/total

PFullTimeAssosiate <- PFullTime * PAssosiate

PFullTimeAssosiate * total
```

```
## [1] 684.9422
```

d) A newspaper article suggests Assistant Professors are more likely that other ranks to be part time because of parenting duties. Is this a correct interpretation of this data? Explain.

```
x - chisq.test(x)$expected
```

```
##           [,1]      [,2]
## [1,]  42.85259 -42.85259
## [2,]  42.05777 -42.05777
## [3,] -84.91035  84.91035
```

```
# Thats not true, In fact Full Professors are the ones that more likely be part timer.
```

Question 3

A university has as its goal to have 82% domestic students with full government funding, 14.5% international students, and 3.5% other students. A sample of 1000 students shows 807 fully funded domestic students, 167 international students, and 26 other students.

a) Are the universitys students distributed in the desired way? Perform the relevant hypothesis test and give your conclusion.

```
x <- c(807, 167, 26)
(test <- chisq.test(x, p = c(.82, .145, .035)))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  x
## X-squared = 5.8583, df = 2, p-value = 0.05344
```

```
test$expected
```

```
## [1] 820 145  35
```

```
# Based on p-value 0.05344 which is greater than .0001 we cannot reject null Hypothesis
```

b) Someone suggests using a sample of 100 students rather than 1000 students, so each can be interviewed personally and asked a more extensive set of questions. Discuss the pros and cons of this idea with respect to the test performed above.

```
# sampling 100 students out of another sample makes no sense and it would result abnormalities in distribution. Theoritically it will have a similar result as above. but practically it could cause serious distribution change since we are taking 1/10th of the sample and could change p-value drastically.
```

Question 4

Consider the data on kg of tomatoes produced by plots with differing soil salinity in the file â??salinity.txt.â??

a) Treating salinity as a factor, perform a one way anova. What is your conclusion? Explan what parts of the output you are basing this conclusion on.

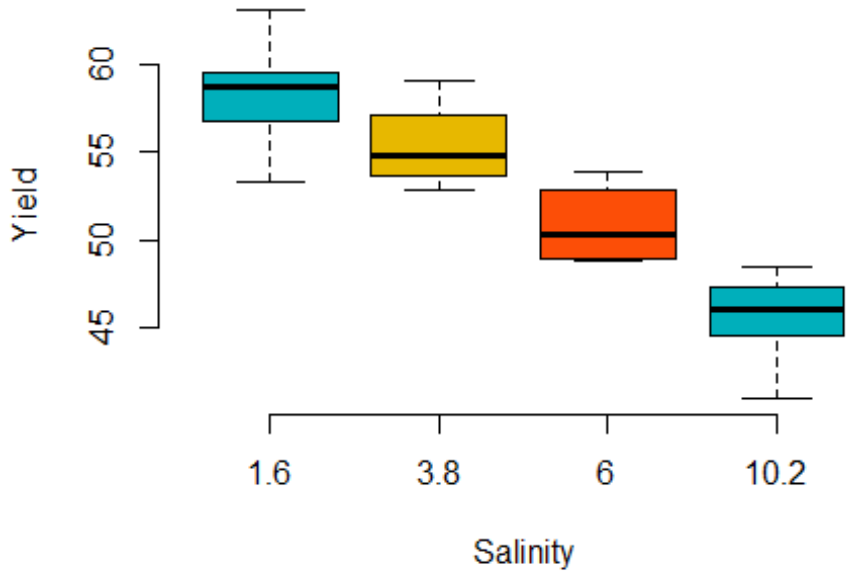
```
salinity <- read.table("salinity.txt", header=TRUE)

sal.anov <- aov(yield~factor(salinity), salinity)

summary(sal.anov)
```

```
##                  Df Sum Sq Mean Sq F value   Pr(>F)
## factor(salinity)  3  456.5  152.17   17.11 5.87e-05 ***
## Residuals        14  124.5    8.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

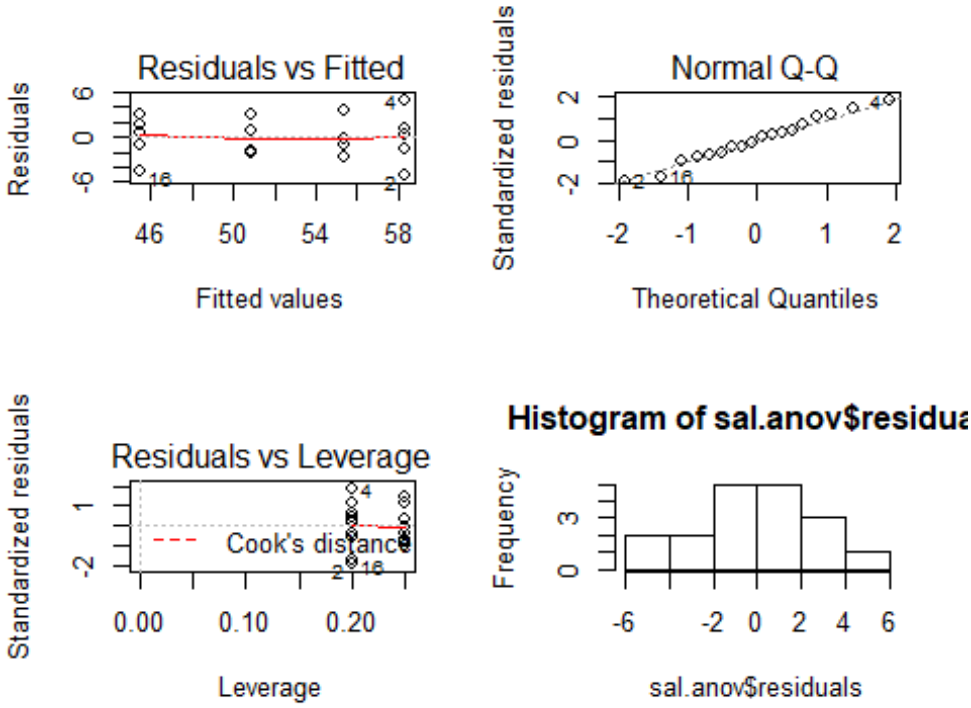
```
# according to P value it could be seen that there is significant difference
between groups. since the value is much lesser than .001
```

b) Produce relevant diagnostic plots and outline any concerns.

```
boxplot(yield ~ factor(salinity), data = salinity,
        xlab = "Salinity", ylab = "Yield",
        frame = FALSE, col = c("#00AFBB", "#E7B800", "#FC4E07"))
```



```r
par(mfrow=c(2,2))
plot(sal.anov, which=c(1,2,5))
hist(sal.anov$residuals)
```




```r
# while looking at the residual histogram it seems that distribution is not fully uniform, but Normal Q-Q shows that it is dributed well with few outliners, further based on only one factor Residual vs Factor graph is not formed with meaningful data
```

c) Compute intervals for pairwise comparisons based on Tukeys honest significant differences. Describe the optimal salinity level(s).

```r
TukeyHSD(sal.anov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = yield ~ factor(salinity), data = salinity)
##
## $`factor(salinity)`
##             diff        lwr        upr     p adj
## 3.8-1.6   -2.88  -8.694385  2.9343846 0.4969233
## 6-1.6     -7.43 -13.244385 -1.6156154 0.0109691
## 10.2-1.6 -12.78 -18.261854 -7.2981457 0.0000472
## 6-3.8     -4.55 -10.678899  1.5788995 0.1830783
## 10.2-3.8  -9.90 -15.714385 -4.0856154 0.0010845
## 10.2-6    -5.35 -11.164385  0.4643846 0.0760885
```

```
# from the answer it can be observed that 6-1.6, 10.2-1.6, 10.2-3.8 are the
significants.
# most optimal salinity level is 10.2-1.6
```

d) Describe the advantages and disadvantages of using an ANOVA rather than regression in this situation.

```
# Main advantage of using Anova over Regression is that it could use single
dependent variable to analyze. but when you want more granuality in your
result Regression would be a better choice, over here we had only on
dependent variable that used for analysing the outcome of grouped data. While
regression couldn't do well in group data. However in above example
regression would do better since the factor/group are numeric data.
```