STA302/1001HF: Methods of Data Analysis I **Summer 2017**

Assignment 2

Out: May 22, 2017 *Due: June 3, 2017*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 11 :00pm, June 3, 2017. Submit your solution produced from Rmarkdown as instructed by Crowdmark : one pdf file for one question.

Most problems on this assignment require using R. Your turned in solutions should not include all of the R output and graphs that you will produce. Write your solutions and include only sparingly R output or graphs when necessary to support a point you are making in response to the problem question.

Late assignments will be subject to a deduction of 10% of the total marks for the assignment for each day late. Any late assignment after the day I post the solution will get zero mark.

Presentation of solutions is very important. A Rmarkdown template for the solution is provided. Solution will be accepted only by Rmarkdown. Mark will be deducted if the instructions herein are not followed.

Data

This data for this assignment were taken from *A modern Approach to Regression with R* by Simon J. Sheather. The data set is posted at the course website.

The article *N.F.L Kickers Are Judged on the Wrong Criteria* appeared in *The New York Times* on November 12, 2006. The article states that "there is effectively no correlation between a kicker's fied-goal percentage one season and his field-goal percentage the next". Sheather collected the data to examine this claim.

The variables in the dataset are :

- Name : name of kicker.
- Yeart : year of data collected.
- Teamt : Team of kicker during this year
- FGAt : Number of fields goals attempted during this year.
- FGt : Percentage of fields goals made during this year.
- Team.t.1 : Team in previous year.
- FGAtM1 : Number of field goals attempted in previous year.
- FGtM1 : Percentage of fields goals made in previous year.
- FGAtM2 : Number of field goals attempted two years ago.
- FGtM2 : Percentage of fields goals made two years ago

Use R to do the analysis for the questions in next page.

Questions

Q1 (20 points) Correlation analysis and simple linear regression.

(a) (6 points) Find the correlation between percentage of field goals made and percentage of fields goals made in the previous year. Is this estimated correlation significant different from zero ? (Use *cor.test()* in R). Explain how this result supports the claim in *The New York Times* article.

(b) (8 points) Carry out a simple linear regression using the variables percentage of fields goals made this year and percentage of field goals made in the previous year. In a table, give the values of

- R^2
- the intercept, b_0
- the slope, b_1
- the estimate of the variance of the error term ($\hat{\sigma}^2$)
- the p-value for the test with null hypothesis that the intercept is 0
- the p-value for the test with null hypothesis that the slope is 0

Explain how this analysis is consistent with your answer to Q1-(a).

(c) (6 points) Give a 95% confidence interval for the slope of the regression line in Q1-(b). Explain how the confidence interval is consistent with the conclusions of Q1-(a) and Q1-(b).

Q2 (5 points) Conclusions from regression analysis are valid only if the right model was fit to the data. Why is the regression model fit in Q1-(b) not an appropriate model ? In particular, you should consider how it violates the Gauss-Markov conditions. You do not need to look at plots of the residuals for this question. Instead comment on the Gauss-Markov conditions in the context of the data being considered.

Q3 (10 points) Answer the following two parts.

(a) (5 points)In 2003, Mike Vanderjagt had the highest percentage of field goals made (100%) and Jay Feely had the lowest percentage (70.3%). For each of these two players, carry out a regression to examine the relationship between the percentage of fields goals made in a year and the percentage of field goals made in the previous year. (Note that this is 2 regressions, each using only 4 data points.) What do you conclude ?

(b) (5 points) We can test for a difference between the slopes of the regressions for Mike Vanderjagt and Jay Feely using a t-test, similar to the two-sample t-test for the difference between two means. We can estimate the difference in their slopes by $b_{1,MV} - b_{1,JF}$ where $b_{1,MV}$ and $b_{1,JF}$ are the estimated slopes for Mike Vanderjagt and Jay Feely, respectively. You also need to find an estimate of the standard deviation of $b_{1,MV} - b_{1,JF}$. Under the regression model assumptions and assuming that there is no difference in the slopes, the estimate of the difference in slopes divided by the estimate of the standard deviation of the differences will have approximately a t-distribution with 2 degrees of freedom (using Satterthwaite's approximation). What do you conclude from this t-test ?

Q4 (10 points) R output from a multiple regression is given next page. This regression uses all the data, but fits 19 separate lines, one for each player. In this regression, the lines were forced to be parallel so the coefficient of FGtM1, the percentage of field goals made in the previous year, is the same for all players.

(a) (5 points) Find the p-value for the test with null hypothesis that the coefficient of FGtM1 is equal to 0. What do you conclude about the relationship between field goals made this year and percentage of field goals made the previous year ?

(b) (5 points) Explain, in words, why the test considered in part Q4-(a) is more powerful than the tests about the slopes considered in Q3-(a).

```
> fit=lm(FGt~FGtM1+Name,data=a2)
> summary(fit)

Call:
lm(formula = FGt ~ FGtM1 + Name, data = a2)

Residuals:
     Min       1Q   Median       3Q      Max
-11.1808  -4.0045  -0.5093   4.3053  13.3134

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              126.6872    10.0057  12.661  < 2e-16 ***
FGtM1                     -0.5037     0.1128  -4.467  3.9e-05 ***
NameDavid Akers           -4.6463     4.4007  -1.056  0.29559
NameJason Elam            -3.0167     4.4217  -0.682  0.49790
NameJason Hanson           2.1172     4.3949   0.482  0.63186
NameJay Feely            -10.3737     4.4514  -2.330  0.02341 *
NameJeff Reed             -8.2955     4.3994  -1.886  0.06454 .
NameJeff Wilkins           2.3102     4.3931   0.526  0.60106
NameJohn Carney           -5.9774     4.4159  -1.354  0.18130
NameJohn Hall             -8.4865     4.4528  -1.906  0.06180 .
NameKris Brown           -13.3598     4.5186  -2.957  0.00455 **
NameMatt Stover            8.7363     4.4060   1.983  0.05230 .
NameMike Vanderjagt        4.8955     4.3994   1.113  0.27055
NameNeil Rackers          -6.6200     4.3985  -1.505  0.13793
NameOlindo Mare          -13.0365     4.4528  -2.928  0.00493 **
NamePhil Dawson            3.5524     4.3931   0.809  0.42215
NameRian Lindell          -4.8674     4.4244  -1.100  0.27598
NameRyan Longwell         -2.2315     4.3970  -0.508  0.61379
NameSebastian Janikowski  -3.9763     4.4126  -0.901  0.37138
NameShayne Graham          2.1350     4.3932   0.486  0.62888
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 6.212 on 56 degrees of freedom
Multiple R-squared:  0.5199,Adjusted R-squared:  0.3569
F-statistic: 3.191 on 19 and 56 DF,  p-value: 0.0003849
```

Extra : A brief summary of R on SLR

In this section, I am using a simple data to give you a summary of the R relevant function and code regarding the simple linear regression (SLR) model.

```r
# make up data
x=c(18,23,25,35,65,54,34,56,72,19,23,42,18,39,37)
y=c(202,186,187,180,156,169,174,172,153,199,193,174,198,183,178)
```

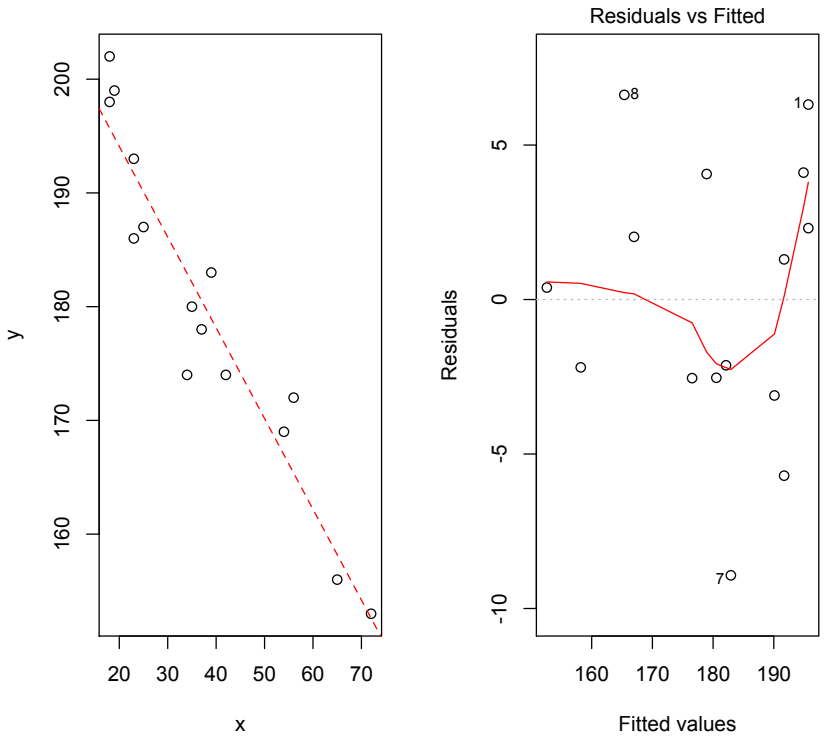
Listing 1 – R example

A scatter plot of the data and the regression line are shown in the next graph. The graph is obtained using the commands :

```r
#make a scatter plot of the data
plot(x,y)

#adding the regression line (with red colour and line type is 2) in the scatter plot
m = lm(y~x)
abline(m,col="red",lty=2)

# want to get the residual plot vs fitted value
plot(m,which=1) # which=2: for the Normal QQ-plot

# Put two plots in one panel
par(mfrow=c(1,2))
plot(x,y)
abline(m,col="red",lty=2)
plot(m,which=1) # which=2: for the Normal QQ-plot
```

Listing 2 – R example

Here is the plot produced from running the last 4 lines



Functions of **lm()**, I summarize here a few functions of interest that you can explore. Each of the following functions acts on **m (m=lm(y∼ x))**, defined by

- **coefficients(m)** - model coefficients.
- **coef(m)** - the same as **coefficients(m)**.
- **confint(m,level=0.99)** - confidence intervals for the regression coefficients.
- **deviance(m)** - residual sum of squares (SSE).
- **fitted(m)**- vector of fitted y values.
- **residuals(m)** -vector of model residuals.
- **resid(m)** - the same as **residuals(m)**.
- **summary(m)** - the summary function already described.
- **vcov(m)** - variance-covariance matrix of the main parameters.
- **df.residual(m)** - the degree of freedom of MSE.
- **plot(m,which=1)** - diagnostic plot.
  which=1 : residuals vs fitted.
  which=2 : Normal QQ-plot.
  which=3 : scale-location.
  which=4 : Cook's distance.
  which=5 : residuals vs leverage.
  which=6 : Cook's distance vs leverage

Plotting the two bands requires some care. Here is a way to do it :

```
pred.xframe = data.frame(x=18:72)
pi = predict(m,interval="prediction",newdata=pred.xframe)
pci = predict(m,interval="confidence",newdata=pred.xframe)
plot(x,y,ylim=range(y,pi,na.rm=TRUE))
pred.x=pred.xframe$x
matlines(pred.x,pci,lty=c(1,2,2),col="blue")
matlines(pred.x,pi,lty=c(1,3,3),col="red")
legend("topright", c("Fitted line", "CIs for E(Y)", "PI for Y"), lty=c(1,2,3), col=c("red
    ","blue","red"))
grid()
```

Listing 3 – R example