Assignment 2

Last name: Du
First name: Min
Student ID: 1002602230
Course section: STA302H1F-Summer 2017

Due Date: June 3, 2017, 23:00

Q1 (20 pts) - Correlation and SLR.

Q1-(a) (6 pts): 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 ? Explain how this result supports the claim in The New York Times article.

Answer:

```
a2 = read.csv("/Users/Joy/Desktop/ASSIGNMENT2/FieldGoals03to06.csv",header=TRUE)
#str(q2data) # check the type of each column (variable) in the data set
#head(q2data,10) # have a look of the first 10 data lines

# Write your R code in the following
rp2 <- cor( a2$FGtM1, a2$FGt, method = "pearson")
rp2
```

```
## [1] -0.1391935
```

```
cor.test(a2$FGtM1, a2$FGt)
```

```
##
##  Pearson's product-moment correlation
##
## data:  a2$FGtM1 and a2$FGt
## t = -1.2092, df = 74, p-value = 0.2305
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3535538  0.0890568
## sample estimates:
##        cor
## -0.1391935
```

No, It is not significant different from zero. The correlation between percentage of field goals made and percentage of fields goals made in the previous year is -0.1391, which is close to 0, very weak and alomst neligible. According to the above code, H_0:The correlation between percentage of field goals made and percentage of fields goals made in the previous year is 0. P-value = 0.2305, which is greater than $\alpha = 0.05$, so we do not reject H_0, no evidence to show that there is correlation between percentage of field goals made and percentage of fields goals made in the previous year.

Q1-(b) (8 pts): 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.

Answer:

List of table	results
R^2	0.01937

List of table	results
intercept, b_0	94.6098
slope, b_1	-0.1510
estimate of σ^2	0.0156
P-value for $H_0 : \beta_0 = 0$	6.18e-14
P-value for $H_0 : \beta_1 = 0$	0.23

```
# code
lm_fitt = lm(a2$FGt ~ a2$FGtM1)
summary(lm_fitt)
```

```
##
## Call:
## lm(formula = a2$FGt ~ a2$FGtM1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4350  -7.0576   0.6933   5.3824  18.7047
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  94.6098    10.2525   9.228 6.18e-14 ***
## a2$FGtM1     -0.1510     0.1248  -1.209     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.723 on 74 degrees of freedom
## Multiple R-squared:  0.01937,    Adjusted R-squared:  0.006123
## F-statistic: 1.462 on 1 and 74 DF,  p-value: 0.2305
```

Using T-test, t^* = -1.209, p-value=2P[T>$|t^*|$] = 0.230512, which is greater than $\alpha = 0.05$. Therefore, we do not reject H_0, no evidence to show that the percentage of fields goals made this year and percentage of field goals made in the previous year have linear association.

Q1-(c) (6 pts): 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).

Answer:

```
# code
confint(lm_fitt, level=0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) 74.1811719 115.03840239
## a2$FGtM1    -0.3997189   0.09780225
```

95% confidence interval for the slope of the regression line is [-0.3997189, 0.09780225], 0 is contained in this confidence interval. Therefore, we do not reject H_0, no evidence to show that the percentage of fields goals made this year and percentage of field goals made in the previous year have linear association.

Q2 (5 pts)

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.

Answer: Firstly, according to Gauss-Markov assumption,var(ϵ_i)= σ^2 when it is a constant, However, in this problem var(ϵ_i) is not a constant, because the error term is between different people's percentage of fields goals made this year and percentage of field goals made in the previous years. Secondly, for cov(ϵ_i,ϵ_j) = 0, which means the correlation between two points should be equal to 0. In this problem, It can be understood as cov(Y_i,Y_j) = 0, which means the correlation between two people's percentage of field goals this year and last year should be equal to 0. However, the first 4 names from a2 are the same, there must exist relationship between the same people's percentage of field goals, the correlation between last year and this year on the same person do not equal to 0. Thirdly, according to data, for each person, we can find four different percentage of fields goals made in previous year and this year, which violate the assumption that x,y in linear model is one to one. The above three problems violate Gauss-Markov condition. So the regression model fit in Q1-(b) is not an appropriate model.

Q3 (10 pts)

Q3-(a): 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 ?

Answer: For Mike Vanderjagt, using T-test to find the relationship between the percentage of fields goals made in a year and the percentage of field goals made in the previous year. From the code shown below, t^* = -4.437, p-value=2P[T>$|t^*|$] = 0.047226, which is smaller than α = 0.05, so we reject H_0, we have evidence to show that the Mike Vanderjagt 's percentage of fields goals made in a year and the percentage of field goals made in the previous year have linear association. For Jay Feely, we also use T-test to find whether there exists relationship between the percentage of fields goals made in a year and the percentage of field goals made in the previous year. From the code shown below, t^* = -0.407, p-value=2P[T>$|t^*|$] = 0.723433, which is significantly greater than α = 0.05, so we do not reject H_0, we have no evidence to show that the Mike Vanderjagt 's percentage of fields goals made in a year and the percentage of field goals made in the previous year have linear association.

Player	Estimate of slope (b_1)	p-value for test with $H_0 : \beta_1 = 0$	estimate $\sigma^2(b_1)$
Mike Vanderjagt	-0.8	0.04724	0.0325
Jay Feeley	-0.2686	0.724	0.4364

```
# Example: run a SLR for name="David Akers"
DA =lm(FGt~FGtM1, data=a2[a2$Name=="David Akers",])
summary(DA)
```

```
##
## Call:
## lm(formula = FGt ~ FGtM1, data = a2[a2$Name == "David Akers",
##     ])
##
## Residuals:
##       5       6       7       8
##  3.2671  3.7034 -7.5581  0.5876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  98.0942    45.9835   2.133    0.167
## FGtM1        -0.2116     0.5596  -0.378    0.742
##
## Residual standard error: 6.398 on 2 degrees of freedom
## Multiple R-squared:  0.06671,    Adjusted R-squared:  -0.3999
## F-statistic: 0.143 on 1 and 2 DF,  p-value: 0.7417
```

```
# For MV = Mike Vanderjagt# S{b0} =1
MV =lm(FGt~FGtM1, data=a2[a2$Name=="Mike Vanderjagt",])
summary(MV)
```

```
##
## Call:
## lm(formula = FGt ~ FGtM1, data = a2[a2$Name == "Mike Vanderjagt",
##     ])
```

```
##
## Residuals:
##    45    46    47    48
##  2.06  2.78 -1.22 -3.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 157.2187    15.7080  10.009  0.00984 **
## FGtM1        -0.8000     0.1803  -4.437  0.04724 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.645 on 2 degrees of freedom
## Multiple R-squared:  0.9078, Adjusted R-squared:  0.8616
## F-statistic: 19.68 on 1 and 2 DF,  p-value: 0.04724
```

```
# For JF = Jay Feely
JF =lm(FGt~FGtM1, data=a2[a2$Name=="Jay Feely",])
summary(JF)
```

```
##
## Call:
## lm(formula = FGt ~ FGtM1, data = a2[a2$Name == "Jay Feely", ])
##
## Residuals:
##      17      18      19      20
## -6.1994 -0.9046  6.3171  0.7869
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  97.9850    51.5890   1.899    0.198
## FGtM1        -0.2686     0.6606  -0.407    0.724
##
## Residual standard error: 6.316 on 2 degrees of freedom
## Multiple R-squared:  0.07634,    Adjusted R-squared:  -0.3855
## F-statistic: 0.1653 on 1 and 2 DF,  p-value: 0.7237
```

Q3-(b): 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 ? (To estimate the p-value, you can use a t-table.)

Answer: We want to test the difference between the slopes of the regressions for Mike Vanderjagt and Jay Feely, so we set H_0:$\beta_{1,MV}$=$\beta_{1,JF}$, H_1:$\beta_{1,MV}$-$\beta_{1,JF}$!=0. From the above code, $b_{1,MV}$=-0.8000, $b_{1,JF}$= -0.2686, so the difference in their slopes is $b_{1,MV} - b_{1,JF}$= -0.5314. The standard deviation of $b_{1,MV} - b_{1,JF}$ = sqrt(var$(b_{1,MV}) + var(b_{1,JF})$)= sqrt($s^2 b_{1,MV} + s^2 b_{1,JF}$)= 0.6848, t = $b_{1,MV} - b_{1,JF} - 0/sqrt(var(b_{1,MV} - b_{1,JF}))$ = -0.776. According to T-table, p-value= 2P[T>$|t^*|$]=0.5189, which is greater than α=0.05. Therefore, we do not reject H_0, we have no evidence to show that there are difference between the slopes of the regressions for Mike Vanderjagt and Jay Feely.

Q4 (10 pts)

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.

Q4-(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 ?

Answer: P-value for the test with null hypothesis that the coefficient of FGtM1 is equal to 0 is 0.0003849, which is smaller than α=0.05, so we reject H_0, we have evidence to show that the coefficient of FGtM1 does not equal to 0.

Q4-(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).

Answer: Q3-(a) put 2 groups of data together, Q4-(a) uses indicated variable to foucus on each players' data, so Q4-(a) is more accurate and more powerful.