Assignment 2

Last name: Type down your lastname here
First name: Type down your firstname here
Student ID: 0000000
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/Wei/TA/Teaching/STA302-Summer2017/HW/A2/FieldGoals03to06.csv",header=T)
#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
cor.test(a2$FGt, a2$FGtM1)
```

```
##
##  Pearson's product-moment correlation
##
## data:  a2$FGt and a2$FGtM1
## 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
```

From above R output, the correlation between percentage of field goals made this year and in the previous year is -0.139. This is not statistically signficantly different from 0 (p.val = 0.231). This is consistent with the claim in The New York Times article that "there is effectively no correlation between a kicker's field-goal percentage one season and his field-goal percentage the next.'

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
slope, b_1	-0.1510
estimate of σ^2	$7.723^2 = 59.64473$
P-value for $H_0 : \beta_0 = 0$	< 0.0001
P-value for $H_0 : \beta_1 = 0$	0.23

```
fit = lm(FGt~FGtM1,data=a2)
#summary(fit)
```

The slope is not statistically significantly different from 0 (p.val = 0.231) so we conclude that there is no linear relationship between the percentage of field goals kicked this year and next year, which is consistent with the conclusion about the correlation in question 1-(a) (and is, in fact, an equivalent test).

A note (not required) on the practical interpretation of this model:

The negative slope does not mean that kickers will, on average, do worse from one year to the next. We must also consider the value of the intercept. The fitted line crosses the line Y = X at 82.2. According to this fitted model, kickers who made more than 82.2% of their field goals one year will do worse the following year, on average. Kickers who made less than 82.2% of their field goals one year will, on average, do better the following year.

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:

The 0.975 quantile from a t-distribution with 74 degrees of freedom is 1.992543 (R code: *qt(0.975,74)*), so a 95% confidence interval for the slope is:

```
Low=-0.151-qt(0.975,74)*0.125
Upp =-0.151+qt(0.975,74)*0.125
c(Low,Upp) # the 95% CI
```

```
## [1] -0.40006794  0.09806794
```

```
confint(fit,level=0.95) # Method 2: the second row gives the answer
```

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

This confidence interval includes 0, we don't have evidence to support the alternative hypothesis that the slope is different from 0. So it is consistent with the conclusions that the slope and correlation are zero, and we again conclude that the data do not give evidence of a linear relationship between the percentage of field goals kicked this year and next year.

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:

The problem with this regression model is that it treats all points the as independent observations, although they come in groups of 4 points for each of 19 kickers, so this is not the right model for these data.

The fitted regression line cuts through the middle of the data. However some kickers are better than others, so the higher values of percentage of field goals scored this year will likely belong to them. For these kickers we expect the vertical deviations from their points to the regression line to be positive.

Similarly for weaker kickers, we expect negative errors. So the expectation of the error term is not 0 for all points. Also, it is reasonable to presume that errors for the four points for any one kicker are correlated (if one error is large and positive, we expect the other errors for that kicker to be large and positive). So we have violations of Gauss-Markov conditions: $E(\epsilon) = 0$ and ϵ are mutually uncorrelated.

Q3 (10 points)

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:

Player	Estimate of slope (b_1)	p-value for test with $H_0 : \beta_1 = 0$	estimate $\sigma^2(b_1)$
Mike Vanderjagt	-0.8000	0.04724	$0.1803^2 = 0.03250809$
Jay Feeley	-0.2686	0.724	$0.6606^2 = 0.4363924$

For Mike Vanderjagt there is moderate evidence that the slope is not 0 (p = 0.0472), i.e., that there is a relationship between percentage of fields goals kicked in successive years. For Jay Feely there is no evidence that the slope is different from 0 (p = 0.724), i.e., we conclude that there is no evidence of a relationship between percentage of fields goals kicked in successive years.

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

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

```
##
## 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 ?

Answer:

- Estimate the difference in their slopes

$$b_{1,MV} - b_{1,JF} = -0.8000 - (-0.2686) = -0.5314$$

- Estimate the variance of $b_{1,MV} - b_{1,JF}$

$$S^2(b_{1,MV} - b_{1,JF}) = S^2(b_{1,MV}) + S^2(b_{1,JF}) = 0.03250809 + 0.4363924 = 0.4689005$$

- The test statistic for the test with null hypothesis that $\beta_{1,MV} = \beta_{1,JF}$ is

$$t^* = \frac{b_{1,MV} - b_{1,JF}}{\sqrt{S^2(b_{1,MV}) + S^2(b_{1,JF})}} = \frac{-0.5314}{\sqrt{0.4689005}} = -0.7760348$$

- Under the null, $t^* \sim t_2$, so the p-value is 0.5189299 (R code: *2*pt(-0.7760348,2)*). So we conclude that there is no evidence at $\alpha = 5\%$ of a difference between the slopes of the regression lines for Mike Vanderjagt and for 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:

The p-value for the test with null hypothesis that the coefficient of FGtM1 is equal to 0 is $3.9 * 10^{-05} < 0.0001$. Thus we have strong evidence of a relationship between field goals made this year and percentage of field goals made the previous year, in contrast to the claim in *The New York Times* article.

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:

This regression for all players estimates a slope (-0.5037) that is closer to 0 than the slope estimate for Mike Vanderjagt (-0.8000) (so you might expect that the strength of the evidence that the slope is 0 is stronger for Mike Vanderjagt's regression). However the estimate of the standard deviation of the slope is a little smaller for the all players regression (0.1128) than for Mike Vanderjagt's regression (0.1803) resulting in similar t-statistics (-4.467 for all players and -4.447 for Mike Vanderjagt). The evidence that the slope is statistically significantly different from 0 is much stronger for the regression for all players because the error estimate is based on more observations, giving a larger degrees of freedom for the error estimate, and thus the p-value is calculated from a t-distribution with much less probability in the tails.

Jay Feely's regression is a little different from the other two as the estimated slope is closer to 0 and the standard error of the slope is larger, so the t-statistic is smaller. Jay Feely's data are too noisy to have enough power to be able to conclude that the slope for his regression is not zero.