STAT 308 – Homework 4 Solutions

For the problems in which calculations are needed, please include your R code with your answers, otherwise you will not be given full credit. Please upload your assignment by Thursday, September 29, 11:59 pm in a pdf file to Sakai.

- 1. Suppose we perform a simple linear regression where

$$n = 50, \bar{x} = -0.208, \bar{y} = 1.516, s_x = 2.354, s_y = 3.185$$

$$\hat{\beta}_0 = 1.745, \hat{\beta}_1 = 1.102, s_{Y|X} = 1.868, s_{\hat{\beta}_0} = 0.265, s_{\hat{\beta}_1} = 0.113$$

– a. Calculate the sample correlation coefficient, r, and the r^2.

```
n <- 50
b1 <- 1.102
sx <- 2.354
sy <- 3.185
r <- b1*sx/sy
r
```

```
## [1] 0.8144766
```

```
r^2
```

```
## [1] 0.6633721
```

$r = 0.814$, $r^2 = 0.663$.

– b. Calculate the estimate of the regression variance $s^2_{y|x}$. (Hint: Intuitively, this is the variance of Y that is not explained through the linear model with X.)

```
SST <- (n-1)*sy^2
SSE <- SST*(1-r^2)
syx2 <- SSE/(n-2)
syx2
```

```
## [1] 3.485971
```

$s^2_{y|x} = 3.486$ - 2. Use the following incomplete ANOVA table to answer the following questions.

– a. What is the mean squares for the model (MSM)?

```
df_total <- 34
df_model <- 1
SSM <- 1.47
f <- 0.18
MSM <- SSM/df_model
MSM
```

	df	Sum Sq	Mean Sq	F value	Pr(>F)
Model	1	1.47		0.18	
Error					
Total	34				

```
## [1] 1.47
```

$MSM = 1.47$.

– b. What is the mean squared error (MSE)?

```
MSE <- MSM/f
MSE
```

```
## [1] 8.166667
```

$MSE = 8.1667$

– c. What are the error degrees of freedom?

```
df_error <- df_total - df_model
df_error
```

```
## [1] 33
```

$df_error = 33$.

– d. What is the sum of squared errors (SSE)?

```
SSE <- MSE*df_error
SSE
```

```
## [1] 269.5
```

$SSE = 269.5$

– e. What is the p-value used to test for a significant linear relationship between X and Y?

```
p <- 1 - pf(f,df_model,df_error)
p
```

```
## [1] 0.6741264
```

p-value = 0.674

- 3. Reconsider the dataset `AdRevenue.csv` as well as our simple linear regression model of ad revenue (in millions of dollars) based on circulation (in millions).

– a. Obtain an ANOVA table for this model.

```
adrev <- read.csv("../data/AdRevenue.csv")
mod <- lm(AdRevenue ~ Circulation,adrev)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: AdRevenue
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Circulation  1 1027759 1027759  576.52 < 2.2e-16 ***
## Residuals   68  121223    1783
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

– b. Using the ANOVA table, perform a hypothesis test for a linear relationship between ad revenue and circulation. Be sure to properly state your hypotheses, your test statistic, p-value, and a decision and conclusion at $\alpha = 0.05$.

$H_0 : \beta_1 = 0$, $H_a : \beta_1 \neq 0$

$f - stat = 576.52$, $p - value = 2.2e - 16$

We reject H_0 and say that there is a significant linear relationship between magazine circulation and ad revenue.

– c. What is the distribution the test statistic follows under H_0? In other words, what is the distribution we use to calculate the p-value?

$f - stat \sim F_{1,68}$

– d. Using the ANOVA table, calculate the value of r^2. Interpret this value in the context of the problem.

```
SSM <- anova(mod)[1,2]
SSE <- anova(mod)[2,2]
SST <- SSM + SSE
r2 <- SSM/SST
r2
```

```
## [1] 0.894495
```

89.45% of the variation in ad revenue can be explained by its linear relationship with magazine circulation.