STAT 3032 - Homework 2

Saksham Goel

February 9, 2018

Question 2 - Real Estate Markets

The following snippet of R Code loads the desired dataset and applies a simple linear model to it.

```
file = read.table("indicators.txt", header=T)
xCol = file$LoanPaymentsOverdue
yCol = file$PriceChange
mod = lm(yCol ~ xCol)
mod
```

```
##
## Call:
## lm(formula = yCol ~ xCol)
##
## Coefficients:
## (Intercept)         xCol
##       4.514       -2.249
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = yCol ~ xCol)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.5145     3.3240   1.358   0.1933
## xCol         -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

Question 2 - Part a: Generate a 95% confidence interval for β_1

The following snippet of R Code uses the inbuilt function confint() which outputs the values of the confidence interval for all the coefficicents (in case of liner model - β_0 and β_1).

```
confint(mod, level = 0.95)
```

```
##                 2.5 %     97.5 %
## (Intercept) -2.532112 11.5611000
```

```
## xCol        -4.163454 -0.3335853
```

The 95% confidence interval can thus be given as follows:

$$-4.163454 \leq \beta_1 \leq -0.3335853$$

The above confidence interval suggests that there is a significant negative linear correlation because the confident interval first of lies in the region below zero, which prooves that the relation is negative. Also the relation is quite significant because the magnitude of confidence interval is rather big.

Question 2 - Part b: Generate a 95% confidence interval for E(Y|X=4)

First we find the estimate value for E(Y|X=4) using the fitted regression model as follows:

```
predict(mod, data.frame(xCol = c(4)))
```

```
##         1
## -4.479585
```

We see that the predicted value that is the expected value is given as follows: E(Y|X=4) = -4.479585.

Now finding the Confidence Interval using the following snippet of R Code:

```
x_curr = 4 #the value at which we need to find the confidence interval
y_hat = predict(mod, data.frame(xCol = c(4)))
x_bar = mean(xCol)
sxx = sum((xCol-x_bar)^2)
s = 3.954 #found through the summary of the linear model
n = length(xCol)
t_mult = qt(0.975, df = n-2) # found using the T Table, equal to t0.25 for 16 degrees of freedom
yPILower = y_hat - (t_mult * s * sqrt((1/n) + (((x_curr - x_bar)^2)/sxx) ))
yPIUpper = y_hat + (t_mult * s * sqrt((1/n) + (((x_curr - x_bar)^2)/sxx) ))
yPILower
```

```
##        1
## -6.64885
```

```
yPIUpper
```

```
##         1
## -2.310321
```

The 95% confidence interval can thus be given as follows:

$$-6.64885 \leq E(Y|X = 4) \leq -2.310321$$

Because the 95% confidence interval does not contain the value 0 in it, we can say that 0 is not a feasible value for it.

Question 7 - Explaining Confidence Intervals

By definition of confidence interval, confidence interval represents the expected value of the response variable at a particular value of the independent variable which makes it an interval for the parameters β_0 and β_1 and does not make it able to create an interval for the random variable that is the actual value that we recorded (Y). This fact shows that the interval is not supposed to cover the estimates of the actual observed value but just the expectation of the response variable, which is why it doesn't necessarily contain the 95% of points (which are the actual observed values). A prediction interval is actually an interval that tries to cover the observed value because it tries to estimate the value of the response value, which is why it covers 95% of the

observations (value of the response variable for a particular value of independent variable for a particular trial) as it is supposed to. Another argument is that when we calculate the confidence interval the standard error is calculated only using the variablility because of the parameters β_0 and β_1, while the observed value contains variability due to the error term e_i. This is why the confidence interval does not contain the 95% of the observations. While when we calculate the prediction interval, we always account for the variability due to the error term e_i, because of which the prediction interval contains the 95% of the observed values.

Question InText - Inheritance of Height

Question Intext - Part a:

The response variable is daughter's height. I was able to give this answer because seeing the summary table I can see that there is a value of the slope for the variable mheight which makes it the independent variable. Because mheight is an independent variable, it directly infers that the variable dheight is response variable.

Question Intext - Part b:

Calculating the test statistic for β_0 can be done through the following piece of R code, which uses values from the summary table and just basic math operations.

```
beta0 = 29.91744
stdErr = 1.62247
testStatistic = (beta0 - 0) / stdErr
testStatistic
```

```
## [1] 18.43944
```

The test statistic we find is given as follows: testStatistic = 18.4394

Question Intext - Part c:

We know thet the t-Statistic for β_1 follows a relation with the f-statistic. The relation is given as follows:

$$t - statistic^2_{\beta_1} = f - statistic$$

So the t-statistic is given as follows, using the R Code

```
fStatistic = 435.5
tStatistic = sqrt(fStatistic)
tStatistic
```

```
## [1] 20.86864
```

Hence, the t-statistic we find is given as follows: tStatistic = 20.86864

Question Intext - Part d:

To find RSS, we can use the value of Residual Standard error from the summary table. We know the relation between RSS and Residual Standard error is given as follows:

$$\frac{RSS}{n-2} = s^2$$

So using the following snippet of R Code we find:

```r
s = 2.266
n = 1375
RSS = (s^2) * (n-2)
RSS
```

```
## [1] 7050.02
```

Hence, the value of RSS we find is given as follows: RSS = 7050.02