STAT 511: HW #3 Q:1 & 2

Rumil Legaspi

15 February 2021

```r
setwd("C:/Users/RUMIL/Desktop/APU/STAT 511 - Millie Mao (Applied Regression Analysis)/week 4/Week 4")

gpa_data = read.table(file = "GPA.txt", header = FALSE, sep = "")

#Adding headers
names(gpa_data) <- c("GPA", "ACT")

#Defining dependent and independent vars
ACT = gpa_data$ACT #X
GPA = gpa_data$GPA #Y

gpa_lm = lm(GPA ~ ACT, data = gpa_data)
summary(gpa_lm)
```

```
##
## Call:
## lm(formula = GPA ~ ACT, data = gpa_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.74004 -0.33827  0.04062  0.44064  1.22737
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
## ACT          0.03883    0.01277   3.040  0.00292 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 118 degrees of freedom
## Multiple R-squared:  0.07262,    Adjusted R-squared:  0.06476
## F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

size.lm = lm(hours~size, data = lotdata) summary(size.lm)

#ANOVA decomp[osition aov(size.lm)

#ANOVA with F-test anova(size.lm)

#General linear test red.lm = lm(hours~1, data = lotdata) #hours~1 is reduced model anova(red.lm, size.lm)

1. Refer to the GPA problem in HW#1

(a). Setting up the ANOVA table

```r
anova(gpa_lm)
```

```
## Analysis of Variance Table
##
## Response: GPA
##            Df Sum Sq Mean Sq F value   Pr(>F)
## ACT         1  3.588  3.5878  9.2402 0.002917 **
## Residuals 118 45.818  0.3883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b). What does MSR measure in your ANOVA table? What does MSE measure? Under what condition do MSR and MSE estimate the same quantity?

To begin with, the MSR (Mean Squared of Regression) comes from the sum of squares (***of our X values) divided by the degrees of freedom from our regression model. In this case MSR measures our regression model's variability (separate from the variability of our error(MSE)).

MSE is the mean squared error and measures how close our regression line is to the data points. It measures the distances from the data points to our regression line. Those distances are the errors and the MSE is squaring those errors. **Ultimately** it measures the variability/spread of those errors.

Both the MSR and MSE estimate the same quantity when the slope β_1 is zero or not. (*** from the slide 8/39, the same quantity error is the same quantity measured when slope is zero or not)

(c). At $\alpha = 0.05$, conduct an F-test of whether or not $\beta_1 = 0$. State the null and alternative hypotheses, decision rule, and conclusion.

Null hypothesis: H0: $\beta_1 = 0$ (slope is horizontal/ no relationship)

Alternative hypothesis: H1: β_1 /= 0 (slope exists/ relationship exists)

Given our F-value is 9.2402

___note: 1. If the f statistic is larger than the critical value

2. or the p value is less than alpha then it is significant and we reject the null___

```r
#Using qt() to find our critical value we get
qt(0.95, 118)
```

```
## [1] 1.65787
```

```r
#*** why do we use 0.995 and not 0.99 here?
```

The F-value is greater than our critical value (1.65787) **AND** our p-value (0.002917) is less than our $\alpha = 0.05$ we reject our null and therefore a non-zero slope exists, as well as a significant relationship between ACT(X) and GPA(Y) scores.

(d). Obtain the R-squared from your regression. Interpret this number

---

2. Refer to the Muscle Mass problem in HW#1.

```
muscle_data = read.table(file = "Muscle.txt", header = FALSE, sep = "")

#Adding headers
names(muscle_data) <- c("Muscle", "Age")

#Defining dependent and independent vars
Age = muscle_data$Age #X
Muscle = muscle_data$Muscle #Y

#creating our linear model
muscle_lm = lm(Muscle ~ Age, data = muscle_data)
summary(muscle_lm)
```

```
##
## Call:
## lm(formula = Muscle ~ Age, data = muscle_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1368  -6.1968  -0.5969   6.7607  23.4731
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.3466     5.5123   28.36   <2e-16 ***
## Age          -1.1900     0.0902  -13.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.173 on 58 degrees of freedom
## Multiple R-squared:  0.7501, Adjusted R-squared:  0.7458
## F-statistic: 174.1 on 1 and 58 DF,  p-value: < 2.2e-16
```

(a). Setting up ANOVA Table

```
anova(muscle_lm)
```

```
## Analysis of Variance Table
##
## Response: Muscle
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## Age        1 11627.5 11627.5  174.06 < 2.2e-16 ***
## Residuals 58  3874.4    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(b). At $\alpha = 0.05$, conduct an F-test of whether or not $\beta_1 = 0$. State the null and alternative hypotheses, decision rule, and conclusion.

Null hypothesis: H0: $\beta_1 = 0$ (slope is horizontal/ no relationship between X and Y)

Alternative hypothesis: H1: β_1 /= 0 (slope is non zero and X is linearly related to Y)

```r
#Using qt() to find our critical value we get
qt(0.95, 58)
```

```
## [1] 1.671553
```

Our F-statistic = 174.06

The F-value (174.06) is greater than our critical value (1.671553) **AND** our p-value (2.2e-16) is less than our $\alpha = 0.05$, we reject our null and therefore a non-zero slope exists, as well as a significant relationship between AGE(X) and MUSCLE(Y).

(c). What proportion of the total variation in Muscle Mass remains "unexplained" in the regression with Age? Is this proportion relatively small or large?

The unexplained variation is the error component of the regression equation. It is the mean squared error (MSE) which is the sum of squares divided by the degrees of freedom.

The ratio 11627.5 (MSR)/ 66.8 (MSE) shows a ratio that is fairly relatively large. We can definitely see that our model is contributing far more to the variance than the error is. Which is also true when we compare our F statistic (174.06) with the critical value (1.671553) and see that our model is a significant in predicting the variance between age and muscle mass.

(d). Obtain the R-squared from your regression. Interpret this number