STAT 511: HW #3 Q:1 & 2

Rumil Legaspi

21 February 2021

Contents

1. Refer to the GPA problem in HW#1

Workspace Setup

```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
```

(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.

(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: $\beta_1 /= 0$ (slope exists/ relationship exists)

Given our F-value is 9.2402

note: null rejection rules

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*

```
#Using qt() to find our critical value we get


qt(0.95, 118)
```

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

The F-value is greater than our critical value (1.980272) **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

- R squared = 1 - unexplained variation/total variation
- R squared = variability in Y explained by X/total variability
- R squared = SSR(sum of squares of regression)/SST(sum of squares of total variation(SSR+SSE))
- R squared = 1 - SSE(sum of squares of error)/SST(sum of squares of total variation(SSR+SSE))

```
#Using R functions to find R squared
summary(gpa_lm)$r.squared
```

```
## [1] 0.07262044
```

```
#Reading from ANOVA table we can calculate manually
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
```

```
#SSR(sum of squares of regression)
SSR_gpa <- 3.58

#SSE(sum of squares of errors)
SSE_gpa <- 45.818

#SST(sum of squares of total variation)

SST_gpa <- SSR_gpa + SSE_gpa

#R squared
rsquared_gpa <- SSR_gpa/ SST_gpa
rsquared_gpa
```

```
## [1] 0.07247257
```

The R squared we found [1] 0.07247257 indicates that our ACT scores (input variable) explains close 7% of the variability in our dependent variable GPA. **The relationship between our model and the dependent variable GPA is very weak.**

---

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

Workspace Setup

```
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
```

```r
#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

```r
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: $\beta_1 /= 0$ (slope is non zero and X is linearly related to Y)

```r
#Using qt() to find our critical value we get
#Since alternative is not equal we're looking at both tails
# .025 on both sides
qt(0.975, 58)
```

```
## [1] 2.001717
```

Our F-statistic = 174.06

The F-value (174.06) is greater than our critical value (2.001717) **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.

```
#explained variation, From summary
rsquared.muscle <- 0.7501
#Unexplained variation
1 - rsquared.muscle
```

```
## [1] 0.2499
```

~25% is unexplained, this number is small and therefore our model is strong since our explained variability is ~75%.

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 (2.001717) 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

```
#Using R functions to find R squared
summary(muscle_lm)$r.squared
```

```
## [1] 0.7500668
```

```
#Reading from ANOVA table we can calculate manually
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
```

```
#SSR(sum of squares of regression)
SSR_muscle <- 11627.5

#SSE(sum of squares of errors)
```

```r
SSE_muscle <- 3874.4

#SST(sum of squares of total variation)

SST_muscle <- SSR_muscle + SSE_muscle

#R squared
rsquared_muscle <- SSR_muscle / SST_muscle
rsquared_muscle
```

```
## [1] 0.7500693
```

The R squared we found [1] 0.7500693 indicates that Age (input variable) explains helps explain close to 75% of the variability in our dependent variable Muscle. In other words, **the relationship between our model and the dependent variable Muscle is strong.**

3. Refer to the GPA problem in HW #1

(a). Compute the Pearson correlation coefficient and attach the appropriate sign.

```r
#Computing the sample correlation of two variables
#Choose any continuous variable, order does not matter
cor.test(ACT, GPA, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  ACT and GPA
## t = 3.0398, df = 118, p-value = 0.002917
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.09482051 0.42804747
## sample estimates:
##       cor
## 0.2694818
```

The result of our Pearson correlation coefficient is a positive 0.269.

$r = +0.27$

(b). Obtain Spearman rank correlation coefficient.

```r
#spearman correlation
cor.test(ACT, GPA, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  ACT and GPA
## S = 197904, p-value = 0.0005048
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3127847
```

The result of our Spearman correlation coefficient is 0.31.

$r_s = 0.31$

(c). Which correlation coefficient is stronger? Why?

The Spearman correlation coefficient is stronger compared to the Pearson correlation between the variables ACT(X) and GPA(Y). This is because Spearman measures the data **points by ranks**. In other words, ACT and GPA are better suited because they are both **ordinal** types of data indicating a kind of ranking system, whereas Pearson does not.

---

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

(a). Compute the Pearson product-moment correlation coefficient.

```r
#Age is x, Muscle is Y. But still placement order does not matter
cor.test(Age, Muscle, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Age and Muscle
## t = -13.193, df = 58, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9180874 -0.7847085
## sample estimates:
##       cor
## -0.866064
```

The result of our Pearson correlation coefficient is a negative 0.86 indicating a fairly strong negative linear relationship between our variables, age and muscle.

$r = -0.86$

(b). Based on Part (a), test whether muscle mass and age are significantly correlated in the population at alpha = 0.05. State the null and alternative hypotheses, decision rule, and conclusion.

- 1. **Pearson Correlation Hypothesis Testing Stating Our Hypotheses:**
  - Null hypothesis: H_0: $\rho = 0$ (There is no statistically significant linear correlation between Age and Muscle)
  - Alternative hypothesis: H_1: $\rho \neq 0$ (There exists a statistically significant linear correlation between Age and Muscle)

```
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
```

```
#Helps us find our T value
cor.test(Age, Muscle, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  Age and Muscle
## t = -13.193, df = 58, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9180874 -0.7847085
## sample estimates:
##       cor
## -0.866064
```

```
#Gives us our critical value so we can compare with T statistic
qt(0.975, 58)
```

```
## [1] 2.001717
```

- 2. Calculating t statistic:
  - $t = -13.193$
  - Critical value = 2.00
- 3. Comparing with critical value and p value
  - $|t = -13.193| > 2.00$
  - p value 2.2e-16 $< \alpha$ (0.05)
- 4. Draw conclusion
  - Also, since the p value is very small and we can reject the NULL hypothesis that there is no linear correlation between age and muscle.
  - Therefore we can conclude with the alternative hypothesis that age and muscle are linearly correlated.

(c). Compute the Spearman rank correlation coefficient.

```
#Age is x, Muscle is Y. But still placement order does not matter
cor.test(Age, Muscle, method = "spearman")
```

```
## Warning in cor.test.default(Age, Muscle, method = "spearman"): Cannot compute
## exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Age and Muscle
## S = 67147, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.8657217
```

(d). Repeat the test in Part (b) using the Spearman rank correlation from Part (c).

- 1. **Spearman Correlation Hypothesis Testing Stating our Hypotheses:**
  - Null hypothesis: H_0: $\rho = 0$ (There is no statistically significant rank correlation between Age and Muscle)
  - Alternative hypothesis: H_1: $\rho \neq 0$ (There exists a statistically significant rank correlation between Age and Muscle)

```
#Helps us find our T value
cor.test(Age, Muscle, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  Age and Muscle
## S = 67147, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.8657217
```

```
#Gives us our critical value so we can compare with T statistic
qt(0.975, 58)
```

```
## [1] 2.001717
```

- 2. Calculating t statistic:
  - $t = -13.193$
  - Critical value = 2.00
- 3. Comparing with critical value and p value
  - $|t = -13.193| > 2.00$
  - p value 2.2e-16 $< \alpha$ (0.05)
- 4. Draw conclusion
  - Similarly, since the p value is very small we can reject the NULL hypothesis that there is no significant rank correlation between age and muscle.
  - Therefore we can conclude with the alternative hypothesis that there is a significant rank correlation between age and muscle.

(e). How do your correlation coefficient estimates and test conclusions in Parts (a) and (b) compare to those obtained in Parts (c) and (d)?

Based on the results from our hypothesis tests on both Pearson and Spearman, both age and muscle hold a significant rank and linear correlation relationship. These two hypothesis help verify our correlation coefficient estimates and allow us to safely conclude that a correlation does exist.