Jake Snitily, Calvin Chapman, Katharine Franklin, Brandon Poblette, Xinxin Tong
MATH 425: Applied Stat Models
Professor Mamun
09/30/2024

Assignment 2

Github: https://github.com/Jsnit/MATH425-Assignment-2

Guidelines: *Remember that neatness is important! Please do not hand in any unlabeled or unedited R Output. Include in your write-up only those results that are necessary to present a complete solution. In particular, questions must be answered in order (including graphs), and all graphs must be fully labeled (e.g., x and y axis, title). Include your R code for all questions at the very end of your homework. The code won't be graded, but can help me to figure out what you may have done wrong, if your answer is not correct. You will often be asked to continue problems on successive homework assignments, so please save all your R code.*

1. Grade Point Average (KNNL Problem #2.23 on page 93): The director of admissions of a small college selected 120 students at random from the new freshman class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year (Y) can be predicted from the ACT test score (X). Assume that simple linear regression is appropriate.

   The data can be found on the website: "http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR19.txt".

   Make sure you understand which column is X and which is Y.

(a) Set up the ANOVA table.

```
Analysis of Variance Table

Response: V1
           Df Sum Sq Mean Sq F value   Pr(>F)
V2          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) Conduct an F test of whether or not $\beta 1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

Since the p-value (0.002917) is less than 0.01, we reject the null hypothesis and conclude that there is a correlation between ACT score and GPA. However, the p-value isn't significantly smaller than the α risk, so we want to run other tests as well.

(c) Obtain correlation coefficient *r* and attach the appropriate sign.

The correlation coefficient is 0.2694.

2. Muscle Mass (KNNL Problem #2.27 on page 93): A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass. Assume that first-order regression model is appropriate. Muscle Mass (KNNL Problem #2.27 on page 93): A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass. Assume that first-order regression model is appropriate.
   The data can be found on the website "http://users.stat.ufl.edu/~rrandles/sta4210/Rclassnotes/data/textdatasets/KutnerData/Chapter%20%201%20Data%20Sets/CH01PR27.txt".
   Make sure you understand which column is X and which is Y.

(a) Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

Null hypothesis (H0): There is no linear relationship between age and muscle mass (the slope of the regression is zero).

Alternative hypothesis (H1): There is a negative linear relationship between age and muscle mass (the slope of the regression is negative).

If the p-value associated with the slope (X coefficient) is less than 0.05 (the significance level), reject the null hypothesis in favor of the alternative hypothesis.

P-val: `4.12398690167249e-19`

(b) Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year. Why is it not necessary to know the specific ages to make this estimate?

`-1.18999551413858`

Since the slope is constant in a linear model, the age difference itself is sufficient for estimating the change in muscle mass. Specific ages do not affect this estimate because the relationship is linear and uniform across the entire range of ages.

(c) Obtain a 95 percent confidence interval for the mean muscle mass for women of age 60. Interpret your confidence interval. (Kat)

(82.78266, 87.08213)

Based on the data, we can estimate that the average muscle mass for women aged 60 is likely between 82.78 and 87.08 with 95% confidence.

(d) Obtain a 95 percent prediction interval for the muscle mass of a woman whose age is 60. Is the prediction interval relatively precise? (Kat)

(68.28508, 101.5797)

While the prediction interval provides useful information, its broad range suggests a high level of uncertainty in predicting the exact muscle mass for a single 60-year-old woman.

(e) Set up the ANOVA table.

```
> anova(model)
Analysis of Variance Table

Response: mmm_y
          Df  Sum Sq Mean Sq F value    Pr(>F)
age_x      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
```

(f) Test whether or not $\beta 1 = 0$ using an F test with $\alpha = .05$. State the alternatives, decision rule, and conclusion.

Reject the null hypothesis, $\beta 1$ not equal 0. The alternative that we tested is H0 : $\beta 1 = 0$ and Ha : $\beta 1$ not equal 0. From ANOVA table, we can conclude that p-value we have from f-test is less than 2.2*10^-16, which is less than 0.05. So we reject the null hypothesis.

(g) What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?

Total variation that remains unexplained:

SSE / SSTO = 3874.4 / 15501.93 = 0.2499332 This is relatively small when predicting the human population

(h) Obtain R^2 & R

R^2 = SSR/SSTO = 11627.5/15501.93 = 0.7501

$$R = -\sqrt{R^2} = -\sqrt{0.7501} = -0.866$$

Code for question 1:

```
data=read.table("/Users/calvinchapman/Downloads/GPA_data.txt")
data

model=lm(V1~V2,data=data)
anova(model)
#For the F test we look at the P-value of V2. Since it is less than 0.05, we reject
```

```
#the null hypothesis and declare that there is a correlation

summary(model)
r.squared=summary(model)$r.squared
correlation.coefficient=sqrt(r.squared)
r.squared
#0.0726
correlation.coefficient
cor(data)
#0.2694
```

Code for Question 2:

Code for Question 2 (ab):

```
setwd("C:/Users/brand/Documents/Math 425/")
muscle_age_data = read.csv("musclemass_age_data.csv", header=T)

# x = AGE
x = muscle_age_data$age
# y = MUSCLEMASS
y = muscle_age_data$musclemass

# Set graph to 1x1
#par(mfrow = c(1, 1))

model = lm(y~x)
summary(model)
```

```
p_value <- summary(model)$coefficients[2, 4]

print(paste("P-value:", p_value))

# 95% confidence interval for the slope
confint(model, level = 0.95)
# Extract the slope
slope <- coef(model)[2]

# Print slope
print(paste("Estimated decrease in muscle mass for a one-year increase in age:", slope))
if(p_value < 0.05) {
  print("Reject the null hypothesis. There is significant evidence of a negative linear relationship between age and muscle mass.")
} else {
  print("Fail to reject the null hypothesis. There is not enough evidence to conclude a significant negative relationship between age and muscle mass.")
}

(e,f)
### Set up the ANOVA table.
anova(model)
```