Homework 4

Anahi Rodriguez

Due 9/28/2022

Homework Instructions

Make sure to add your name to the header of the document. When submitting the assignment on Gradescope, be sure to assign the appropriate pages of your submission to each Exercise.

For questions that require code, please create or use the code chunk directly below the question and type your code there. Your knitted pdf will then show both the code and the output, so that we can assess your understanding and award any partial credit.

For multiple choice questions, please bold your selected answer.

For written questions, please provide your answer after the indicated *Answer* prompt.

You are encouraged to knit your file as you work, to check that your coding and formatting are done so appropriately. This will also help you identify and locate any errors more easily.

Homework Setup

We'll use the following packages for this homework assignment. We'll also read in data from a csv file. To access the data, you'll want to download the dataset from Canvas, and place it in the same folder as this R Markdown document. You'll then be able to use the following code to load in the data.

```
library(ggplot2)
library(MASS)
```

Exercise 1: Mammalian Sleep [25 points]

We'll use the `msleep` dataset from the `ggplot2` package for this exercise. At first glance of the `msleep` data, you may notice some missing values encoded as NAs. For this question, we will use the sleep (`sleep_total`) and bodyweight of an animal, which have no missing values.

```
?msleep
```

part a

I wonder about how the amount of sleep required by an animal changes based on the bodyweight of an animal. For example, do animals who weigh more require more sleep. What would be the primary purpose of fitting a model like this? Bold your answer below.

(a) predicting an observation **(b) explaining a structure/system**

part b

Fit a linear model to estimate the sleep required based on the bodyweight. Interpret the slope for this model.

```
# Use this code chunk for your answer.
lm1 = lm(sleep_total ~ bodywt, data = msleep)
summary(lm1)
```

```
##
## Call:
## lm(formula = sleep_total ~ bodywt, data = msleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7008 -2.3787 -0.4268  3.2732  9.1731
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.7269205  0.4773797  22.470  < 2e-16 ***
## bodywt      -0.0017647  0.0005971  -2.956  0.00409 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.254 on 81 degrees of freedom
## Multiple R-squared:  0.09735,    Adjusted R-squared:  0.08621
## F-statistic: 8.736 on 1 and 81 DF,  p-value: 0.004085
```

Answer: estimated required amount of sleep = 10.7269205 + (-0.0017647 * body weight in kilograms). For every additional kilogram of body weight for a mammal, we expect the estimated required amount of sleep to decrease by 0.0017647 hours on average.

part c

Calculate the 90% confidence interval for the slope of this model. *Note: the multiplier (t^*) is 1.6639 for this situation.* Make sure to show your setup for the calculation. Report the 90% confidence interval below.

```
# Use this code chunk, if needed, to calculate your answer.
lb = -0.0017647 - (1.6639 * 0.0005971)
up = -0.0017647 + (1.6639 * 0.0005971)

print(c(lb, up))
```

```
## [1] -0.0027582147 -0.0007711853
```

Answer:

part d

Report and interpret the coefficient of determination for this relationship.

```
# Use this code chunk (if needed) to perform any necessary analyses.
r_2 = summary(lm1)$r.squared
r_2
```

```
## [1] 0.09735061
```

Answer: 9.735061% of the variation in the estimated required hours of sleep per mammal can be estimated by its linear relationship with body weight of the mammal in kilograms.

part e

Based on part d, calculate the correlation.

```
# Use this code chunk (if needed) to perform your analyses.
-sqrt(r_2)
```

```
## [1] -0.3120106
```

Answer: - 0.3120

part f

For the three-toed sloth, calculate the predicted total amount of sleep and the corresponding residual.

```
# Use this code chunk (if needed) to perform any necessary analyses or calculations.
msleep = msleep
sloth = subset(msleep, msleep$name == 'Three-toed sloth')
sloth[,11]
```

```
## # A tibble: 1 x 1
##   bodywt
##    <dbl>
## 1   3.85
```

```
predicted = 10.7269205 + ( -0.0017647 * 3.85)
predicted
```

```
## [1] 10.72013
```

```
observed = sloth[,6]
observed
```

```
## # A tibble: 1 x 1
##   sleep_total
##         <dbl>
## 1        14.4
```

```
observed - predicted
```

```
##   sleep_total
## 1    3.679874
```

Answer: predicted : 10.72013, residual: 3.679874

Exercise 2: Hand Calculations [30 points]

We've used `R` to generate the summary statistics for the `msleep` dataset so far. Now let's take a moment and confirm some of these calculations "by hand" (still using `R` to perform the calculations).

part a

Calculate $\bar{x}$, $\bar{y}$, S_{xx}, and S_{xy} for the msleep dataset. Clearly label and print your results.

```
# Use this code chunk for your answer.
x_bar = mean(msleep$bodywt)
x_bar
```

```
## [1] 166.1363
```

```
y_bar = mean(msleep$sleep_total)
y_bar
```

```
## [1] 10.43373
```

```
s_xy = sum((msleep$bodywt - x_bar) * (msleep$sleep_total - y_bar))
s_xy
```

```
## [1] -89590.99
```

```
s_xx = sum((msleep$bodywt - x_bar)^2)
s_xx
```

```
## [1] 50767575
```

part b

Calculate the estimates for β_0 and β_1, using the values calculated in part a. Clearly label and print your results. Compare these estimates to what you found in Exercise 1b.

```
# Use this code chunk for your answer.

beta_1 = s_xy/s_xx
beta_1
```

```
## [1] -0.001764729
```

```
coef(lm1)[2]
```

```
##       bodywt
## -0.001764729
```

```
beta_0 = y_bar - (beta_1 * x_bar)
beta_0
```

```
## [1] 10.72692
```

```
coef(lm1)[1]
```

```
## (Intercept)
##    10.72692
```

Comparison: These values are identical.

part c

Calculate the residuals for the dataset. You can use any built-in R function to generate the fitted values of the dataset, but you should not use a built-in function to calculate the residuals. No need to print all of the residuals, but please do print the first few residuals.

```
# Use this code chunk for your answer.
fit1 = fitted(lm1)
residuals = msleep$sleep_total - fit1
residuals
```

```
##          1          2          3          4          5          6          7
##  1.4613159  6.2739266  3.6754619  4.1731130 -5.6680834  3.6798737 -1.9907612
##          8          9         10         11         12         13         14
## -3.7268411 -0.6022143 -7.7008025 -5.3678021 -1.3256358 -0.7185380  1.7738207
##         15         16         17         18         19         20         21
## -0.4268146 -2.4251558 -1.6269117  6.6792560 -5.4217146  7.2760795 -2.3321569
##         22         23         24         25         26         27         28
##  8.9731201 -6.9074969 -7.2969163 -0.6255617  0.1907268  4.1732048  1.7789031
##         29         30         31         32         33         34         35
## -0.9265676 -7.2386736 -6.6151377 -4.3769186 -4.4222881 -2.6175073 -1.2239734
##         36         37         38         39         40         41         42
##  4.3155834  8.6737324 -0.6149203  3.4731730  3.5732913  2.0731413  1.7731183
##         43         44         45         46         47         48         49
##  9.1730971  3.8735489  0.2755501 -3.0265499  3.7731289 -2.3225087 -6.8289781
##         50         51         52         53         54         55         56
## -0.9348017  5.3599608 -0.1504476  3.0580814 -1.2823876 -0.4259499  0.2750207
##         57         58         59         60         61         62         63
##  0.7731166  2.9759384 -7.0751538 -5.0330722  0.3750207  7.4789632 -5.3205675
##         64         65         66         67         68         69         70
##  2.2736442 -2.0268429 -1.1256093 -2.3267881  0.5733407 -0.1267052  5.8747030
##         71         72         73         74         75         76         77
##  3.0732577  5.1734413  2.0731642 -1.4747127 -2.1189792  5.0732771 -5.9607376
##         78         79         80         81         82         83
##  4.8746678 -1.8267370 -5.2210401 -4.4233910  1.7790443 -0.9194557
```

part d

Calculate the SSR, SSE, and SST for the model. You may use anything that you have calculated in the earlier parts of this question. Clearly label and print these three values.

```
# Use this code chunk for your answer.
sst = sum((msleep$sleep_total - y_bar)^2)
sst
```

```
## [1] 1624.066
```

```
sse = sum((fit1 - msleep$sleep_total)^2)
sse
```

```
## [1] 1465.962
```

```
ssr = sum((fit1 - y_bar)^2)
ssr
```

```
## [1] 158.1038
```

part e

How are SSR, SSE, and SST related to each other?

Answer: SST can be broken down into SSE and SSR. SST is a measure of the total variance. SSR is the variance that can be explained by the linear relationship between x and y. SSE is the variance that cannot be measured by the linear relationship between x and y.

part f

Calculate the coefficient of determination from the values calculated in part d. How does this compare to what you found in Exercise 1d (based on Exercise 1b)?

```
# Use this code chunk for your answer.
ssr/sst
```

```
## [1] 0.09735061
```

```
summary(lm1)$r.squared
```

```
## [1] 0.09735061
```

Comparison: These values are the same because R^2 measures the variance in estimated y that is explained by its linear relationship with x and similarly, ssr/sst measured the variance in the linear relationship that we can explain.

Exercise 3: Professor Beauty [20 Points]

In the article, "Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity" in the journal *Economics of Education Review* (http://www.sciencedirect.com/science/article/pii/S0272775704001165), Hamermesh and Parker explored how course evaluations may be associated with a professor's physical appearance. The data were gathered from end of semester student evaluations for a large sample of professors from the University of Texas at Austin. Later, six undergraduate students were asked to look at each professor's photo and rate their physical beauty from 1 to 10.

We will focus on two variables:

- `score` is the average student rating that each professor received.
- `bty_avg` is the average score provided by the six student raters.

part a

Load the data Prof_Evals.csv into your markdown file as `prof_evals`. To do this, you'll want to download the csv file into the same Folder as your Homework4.Rmd file, and then use a line of code to call in the csv file. Print the number of professors included in the dataset and the number of variables recorded for each professor.

```
# Use this code chunk to answer the question.
setwd("~/Desktop/data")
prof_evals = read.csv("Prof_Evals.csv")
dim(prof_evals)[1] # number of profs
```

```
## [1] 463
```

```
dim(prof_evals)[2] # number of variables recorded for each professor
```

```
## [1] 19
```

part b

Create a simple linear model with these two variables, named `prof_model`. Run the summary of your model.

```
# Use this code chunk to answer the question.
prof_model = lm(score ~ bty_avg, data = prof_evals)
summary(prof_model)
```

```
##
## Call:
## lm(formula = score ~ bty_avg, data = prof_evals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9246 -0.3690  0.1420  0.3977  0.9309
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.88033    0.07614   50.96  < 2e-16 ***
## bty_avg      0.06664    0.01629    4.09 5.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5348 on 461 degrees of freedom
## Multiple R-squared:  0.03502,    Adjusted R-squared:  0.03293
## F-statistic: 16.73 on 1 and 461 DF,  p-value: 5.082e-05
```

part c

In RMarkdown, you can directly input code into your text space (with a certain formatting) and then have the knitted file convert that code into the number produced by that code!

I can extract a particular value from this model into my text space by doing the following:

- Type r, a space, and then your code
- Then surround this all with backtick marks (the character that you use to create a code chunk, typically found on the top left of your keyboard).

An example follows. For this to run, remove the two # signs, one in the R chunk and the other in the line of text immediately following the chunk:

```
age_model = lm(score ~ age, data = prof_evals)
```

$\hat{\beta}_0 = 4.4619324$

Knit the pdf, and you'll notice that this is printed as a value! The mathematical symbol before is surrounded by dollar signs, since that is typed using TeX style. The part after the equals sign is the code transformed to its value!

Now... for this exercise, report the values of the intercept and slope for your model in the space below using the automated process.

Answer: $\hat{\beta}_0 = 3.8803326$ $\hat{\beta}_1 = 0.066637$

part d

Interpret $\hat{\beta}_0$, $\hat{\beta}_1$, and R^2 in context with the appropriate value plugged in (you don't have to use `R` code for the interpretation statements, but you can!)

Answer:

- $\hat{\beta}_0$: We estimate that when a professor is rated as a 0 on the beauty scale, that their evaluation rating by students will equal 3.88033 on average.
- $\hat{\beta}_1$: For every additional unit increase in beauty rating of a professor, we estimate that the evaluation rating by their students will increase by 0.06664 on average.
- R^2: 3.502% of the variation in estimated professor evaluation by students can be explained by its linear relationship with beauty rating of that professor.

Exercise 4: Professor Beauty Inference [20 points]

part a

Calculate the expected evaluation score values for professors with average beauty scores of 6.25 and 9.5.

```r
# Use this code chunk for your answer.
as.numeric(coef(prof_model)[1] + (coef(prof_model)[2] * 6.25))
```

```
## [1] 4.296814
```

```r
as.numeric(coef(prof_model)[1] + (coef(prof_model)[2] * 9.5))
```

```
## [1] 4.513384
```

Answer:

part b

Calculate an 95% confidence interval for the true slope.

Print both the lower and upper bound. Make sure these are clearly labeled.

```r
# Use this code chunk for your answer.
lb = 0.06664 - (2 * 0.01629)
ub = 0.06664 + (2 * 0.01629)
print(c(lb,up))
```

```
## [1]  0.0340600000 -0.0007711853
```

part c

For the default hypothesis test for the intercept, report the following:

- The null and alternative hypotheses for the slope (in symbols)
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.05$

No need to extract these specific values in your code. Simply observe the values in the output and report this information in text below.

```r
# Use this code chunk for your answer, if needed.
```

Answer: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, t-statistic: 4.09, p_value: .0000508, We would reject the null: $H_0 : \beta_1 = 0$ at the 5% level and say that their is evidence to suggest the true slope is different from 0

part d

Suppose that the University of Illinois had previously found that for each increase in the student perceived average beauty rating of a professor of 1 on a scale from 1 to 10, that the estimated average student rating score of that professor would increase by 0.1, on average.

Might this same relationship be true for the University of Texas at Austin, where our data came from? Or is there evidence that this relationship for the University of Texas is different from the University of Illinois?

In other words, use a t-test to test:

- $H_0 : \beta_1 = 0.1$
- $H_1 : \beta_1 \neq 0.1$

Calculate and report the value of your test statistic.

```r
# Use this code chunk for your answer.
t_stat = (0.06664 - 0.1)/0.01629
t_stat
```

```
## [1] -2.047882
```

Answer: -2.047882

part e

If a news article used this data to make the claim “Professors get a bump in their evaluation scores if they’re viewed as attractive by students!” would you agree with that conclusion? Why or why not?

There is not an objectively right answer for this question. Consider all of the information that we’ve compiled and calculations that we’ve performed to help guide your response.

Answer: I would say that if professors are viewed as attractive by their students, their is evidence that it would increase their average evaluation scores but only slightly. Our linear model tells us that their is a positive relationship between these two variables and that the value of β_1 is different from 0. Our model predicts this value to be 0.06664 and rejects that this value is 0.1, telling us that this effect exists but is **small**. For example, if a professor was rated as 10 in attractiveness, their evaluation score would only increase on average by 0.6664 – a small change for being rated as high in attractiveness as possible on the scale.

Exercise 5: Formatting [5 points]

The last five points of the assignment will be earned for properly formatting your final document. Check that you have:

- included your name on the document
- properly assigned pages to exercises on Gradescope
- select **page 1 (with your name)** and this page for this exercise (Exercise 5)
- all code is printed and readable for each question
- generated a pdf file