Problem Set 2

QTM 200: Applied Regression Analysis

Due: February 10, 2020

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 10, 2020. No late assignments will be accepted.
- Total available points for this homework is 100.

Question 1 (40 points): Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) Calculate the χ^2 test statistic by hand (even better if you can do "by hand" in R).

```
discrimination <- matrix(c(14, 6, 7, 7, 7, 1),ncol=3,byrow=TRUE)
colnames(discrimination) <- c("Not_Stopped","Bribe_Requested","Stopped_
    Given_Warning")
rownames(discrimination) <- c("Upper_Class","Lower_Class")
discrimination <- as.table(discrimination)


totals <- sum(discrimination)
#Sum Not Stopped
not_stopped <- sum(14 + 7)
#Sum Bribe Requested
bribe_requested <- sum(6 + 7)
#Sum Warning
warning <- (7 + 1)
#Sum Upper Class
upper <- sum(14 + 6 + 7)
#Sum Lower Class
lower <- sum(7 + 7 + 1)

#Expected Not Stopped, Upper Class
fe1 <- (upper/totals)*not_stopped
#Expected Bribe requested, Upper Class
fe2 <- (upper/totals)*bribe_requested
#Expected Warning Given, Upper Class
fe3 <- (upper/totals)*warning
#Expected Not Stopped, Lower Class
fe4 <- (lower/totals)*not_stopped
#Expected Bribe requested, Lower Class
fe5 <- (lower/totals)*bribe_requested
#Expected Warning Given, Lower Class
fe6 <- (lower/totals)*warning

##Observed-Expected squared divided by expected
#Not Stopped, Upper Class
chisquare1 <- (14-fe1)^2/fe1
#Expected Bribe requested, Upper Class
chisquare2 <- (6-fe2)^2/fe2
#Expected Warning Given, Upper Class
chisquare3 <- (7-fe3)^2/fe3
#Expected Not Stopped, Lower Class
chisquare4 <- (7-fe4)^2/fe4
#Expected Bribe requested, Lower Class
```

```r
chisquare5 <- (7-fe5)^2/fe5
#Expected Warning Given, Lower Class
chisquare6 <- (1-fe6)^2/fe6

#Test Statistic
chi_square_stat <- sum(chisquare1,chisquare2,chisquare3,chisquare4,
    chisquare5,chisquare6)
#3.791168
```

(b) Now calculate the p-value (in R).[2] What do you conclude if $\alpha = .1$?

```r
# df = (rows-1))columns-1)
df_pchisq <- (2-1)*(3-1)
pchisq(chi_square_stat, df = df_pchisq, lower.tail = FALSE)
## With a significance of alpha = 0.1 and a p-value of 0.150, I fail to
    reject the null hypothesis and conclude that
## the class and bribe solicitation are statistically independent
```

[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

(c) Calculate the standardized residuals for each cell and put them in the table below.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class 0.1651982 -1.093382 1.182516			
Lower class -0.2014441 1.260318 -1.303106			

```
z1 <- (14-fe1)/sqrt(fe1*(1-(upper/totals)*(1-(not_stopped/totals))))
z2 <- (6-fe2)/sqrt(fe2*(1-(upper/totals)*(1-(bribe_requested/totals))))
z3 <- (7-fe3)/sqrt(fe3*(1-(upper/totals)*(1-(warning/totals))))
z4 <- (7-fe4)/sqrt(fe4*(1-(lower/totals)*(1-(not_stopped/totals))))
z5 <- (7-fe5)/sqrt(fe5*(1-(lower/totals)*(1-(bribe_requested/totals))))
z6 <- (1-fe6)/sqrt(fe6*(1-(lower/totals)*(1-(warning/totals))))
matrix(c(z1, z2, z3, z4, z5, z6), ncol = 3, byrow = TRUE)
```

(d) How might the standardized residuals help you interpret the results?

```
#Standardized residuals show us how far away each observed value is from
    the "expectation"
#Thus, standardized residuals can tell us the accuracy of the expected
    values and the significance of each cell to the chi squared statistic
```

Question 2 (20 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure 1 below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).

[3]Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

(a) State a null and alternative (two-tailed) hypothesis.

```r
download.file("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", "women.csv")
women <- read.csv("women.csv")
head(women)
#2a
#Null Hypothesis: The Reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.
#Alternative Hypothesis:The Reservation policy has an effect on the number of new or repaired drinking water facilities in the villages
```

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!).

```r
#Standardize variation
mean_water_y <- mean(women$water)
mean_reserved_x <- mean(women$reserved)
sum_y <- sum(women$water)
sum_x <- sum(women$reserved)

b_hat <- sum((women$water - mean(women$water)) * (women$reserved - mean(women$reserved))) /
  sum((women$reserved - (mean_reserved_x))^2)

a_hat <- mean_water_y - (b_hat*mean_reserved_x)
#14.738

#Check
lm(women$water ~ women$reserved) #14.738
```

(c) Interpret the coefficient estimate for reservation policy.

```
#For every increase in reservation for women leaders, there was a 9.252
    increase in new or repaired drinking water facilities.
```

Question 3 (40 points): Biology

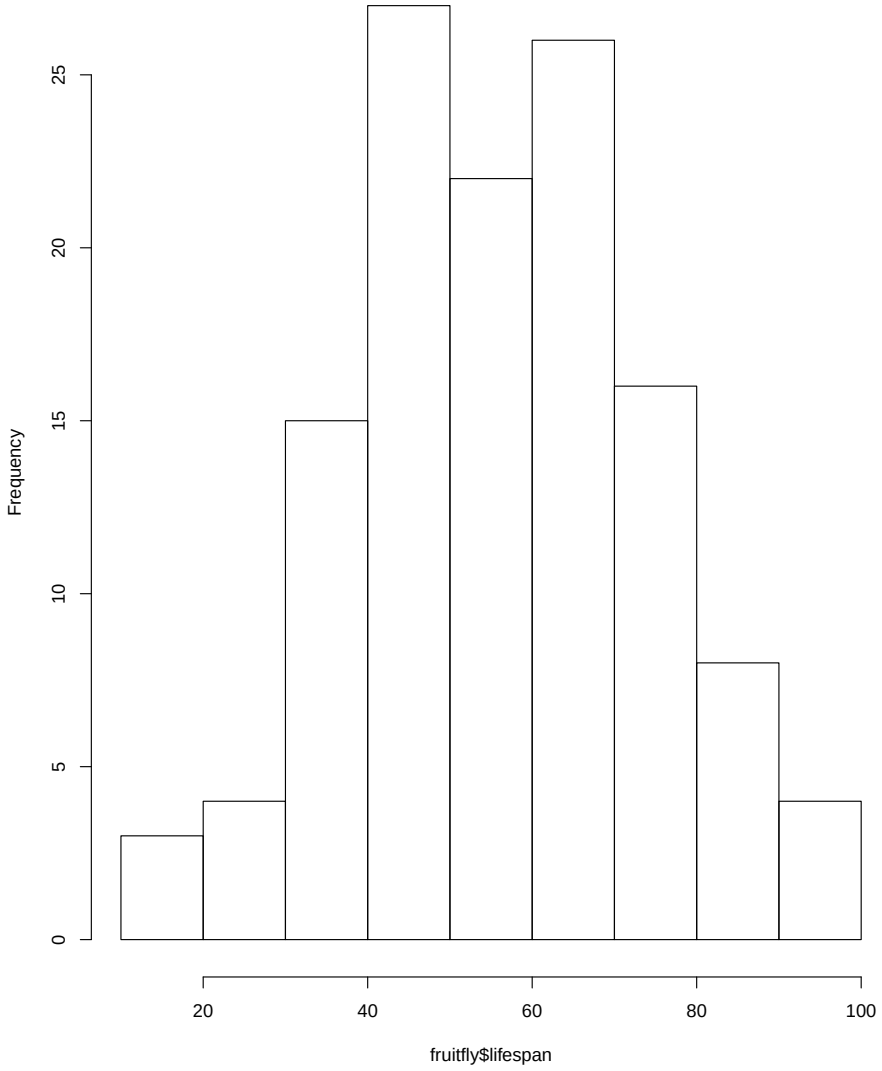
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity for male fruit flies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). The name of the data set is `fruitfly.csv`.[4]

`No`	serial number (1-25) within each group of 25
`type`	Type of experimental assignment 1 = no females 2 = 1 newly pregnant female 3 = 8 newly pregnant females 4 = 1 virgin female 5 = 8 virgin females
`lifespan`	lifespan (days)
`thorax`	length of thorax (mm)
`sleep`	percentage of each day spent sleeping

1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

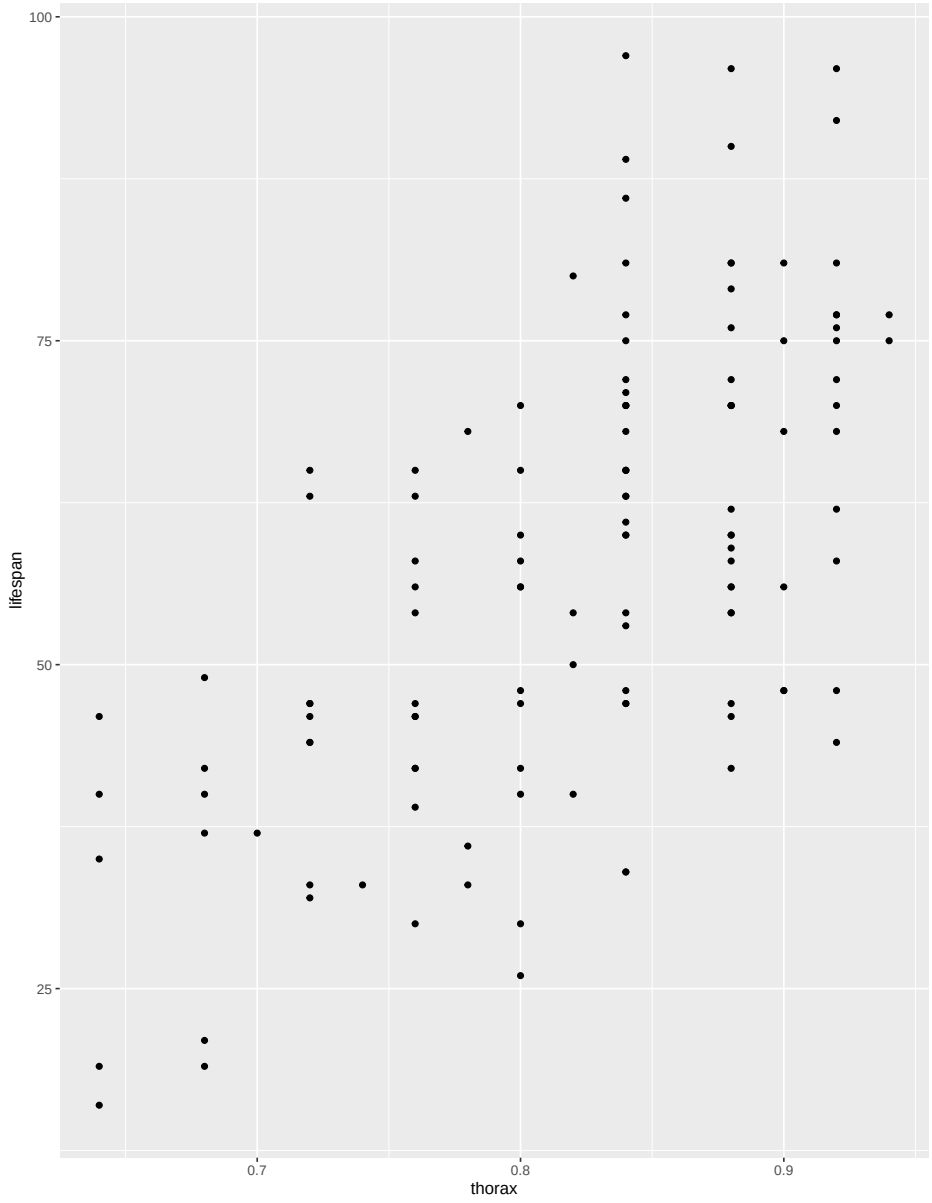
```
fruitfly <- read.csv("fruitfly.csv")

#3.1
summary(fruitfly)
# No          type       lifespan        thorax          sleep
# Min.   : 1   Min.   :1   Min.   :16.00   Min.   :0.640   Min.   : 1.00
# 1st Qu.: 7   1st Qu.:2   1st Qu.:46.00   1st Qu.:0.760   1st Qu.:13.00
# Median :13   Median :3   Median :58.00   Median :0.840   Median :20.00
# Mean   :13   Mean   :3   Mean   :57.44   Mean   :0.821   Mean   :23.46
# 3rd Qu.:19   3rd Qu.:4   3rd Qu.:70.00   3rd Qu.:0.880   3rd Qu.:29.00
# Max.   :25   Max.   :5   Max.   :97.00   Max.   :0.940   Max.   :83.00
hist(fruitfly$lifespan)
```

[4]Partridge and Farquhar (1981)."Sexual Activity and the Lifespan of Male Fruitflies". *Nature*. 294, 580-581.


Histogram of fruitfly$lifespan
Frequency
0
5
10
15
20
25
20
40
60
80
100
fruitfly$lifespan

2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?

```
library(tidyverse)
qplot(x = thorax, y = lifespan, data = fruitfly)
#Yes, there seems to be a linear correlation

cor(fruitfly$thorax, fruitfly$lifespan, method="pearson") #0.6364835
#Check if null p = 0
cor.test(fruitfly$thorax, fruitfly$lifespan)
#0.6364835
```



3. Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.

```
lm(fruitfly$lifespan~fruitfly$thorax)
#For every 1 mm increase in thorax size, lifespan increases by 144.33
    days
```

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.

```
cor.test(fruitfly$lifespan,fruitfly$thorax) #P-value = 1.497e-15

#Pearson's product-moment correlation

#data:  fruitfly$lifespan and fruitfly$thorax
#t = 9.1521, df = 123, p-value = 1.497e-15
#alternative hypothesis: true correlation is not equal to 0
#95 percent confidence interval:
#0.5188709 0.7304479
#sample estimates:
#cor
#0.6364835

## With a p-value of 1.497e-15, I reject the null hypothesis that there
    is no correlation in the population between thorax and lifespan.
```

5. Provide the 90% confidence interval for the slope of the fitted model.

   - Use the formula for typical confidence intervals to find the 90% confidence interval around the point estimate.

   - Now, try using the function `confint()` in `R`.

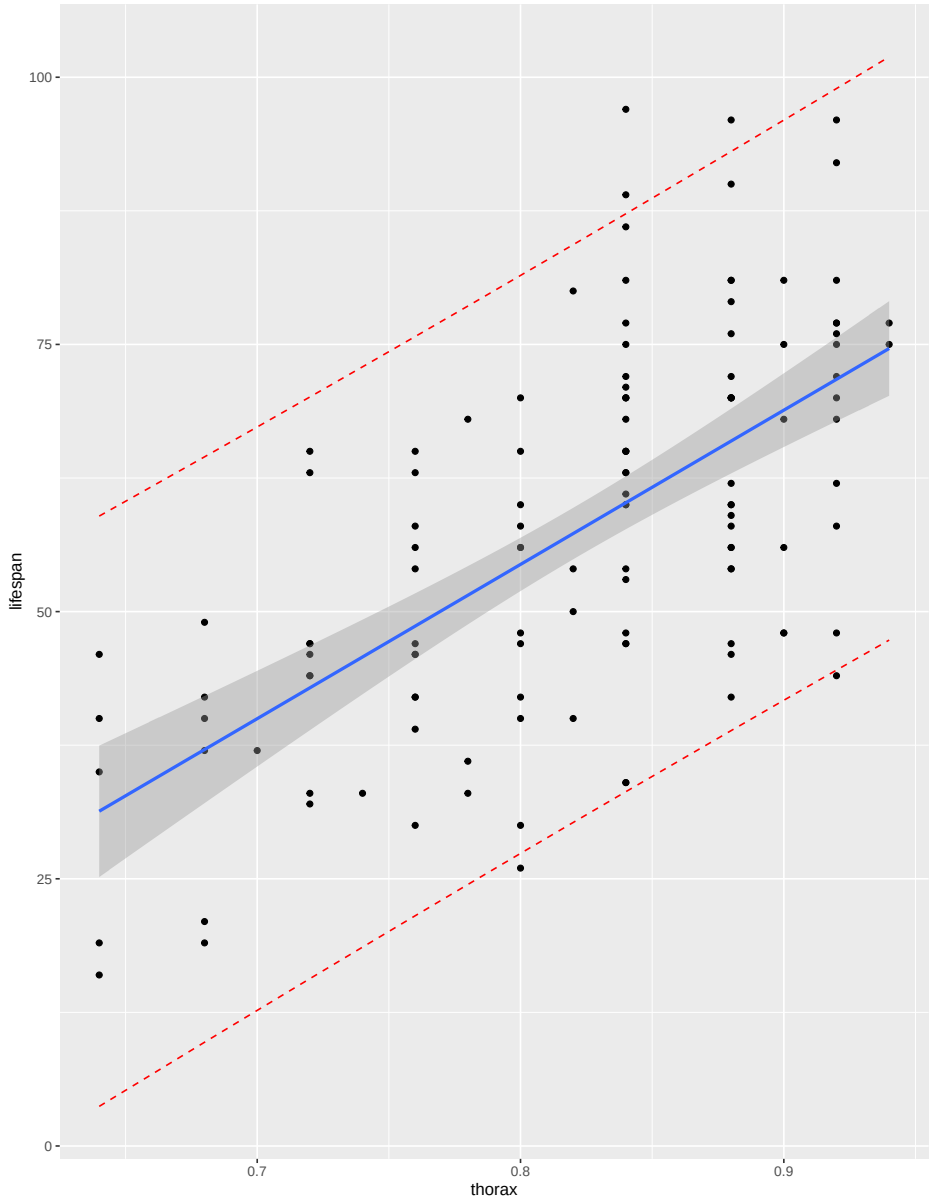
```
##p-value = 1.497e-15 (see 3.4)
t.test(lm(fruitfly$lifespan~fruitfly$thorax), level = 0.90)
#Formula based on t values:
confint(lm(fruitfly$lifespan~fruitfly$thorax), level = 0.90)
#[118.19616,170.4700]
```

6. Use the `predict()` function in `R` to (1) predict an individual fruitfly's lifespan when `thorax`=0.8 and (2) the average `lifespan` of fruitflies when `thorax`=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What are the prediction and confidence intervals around the expected values?

```
new_fruitfly <- fruitfly
new_fruitfly$thorax <- 0.8
#Condifence intervals around expected values of lifespan
pred_interval <- predict(lm(fruitfly$lifespan~fruitfly$thorax), newdata=
    new_fruitfly , interval = "prediction", level = 0.95)
conf_interval <- predict(lm(fruitfly$lifespan~fruitfly$thorax), newdata=
    new_fruitfly , interval = "confidence", level = 0.95)

#3.6.2
#Expected Values of Lifespan:
predict(lm(fruitfly$lifespan~fruitfly$thorax), newdata=new_fruitfly , se.
    fit = TRUE)
```

7. For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.

```
fitted_lifespan <- predict(lm(fruitfly$lifespan~fruitfly$thorax), newdata
    =new_fruitfly , se.fit = TRUE)
reg <- lm(fruitfly$lifespan~fruitfly$thorax)

new_df <- cbind(fruitfly , pred_interval , conf_interval)
names(new_df)[9] <- "fit_conf"
names(new_df)[10] <- "lwr_conf"
names(new_df)[11] <- "upr_conf"

ggplot(new_df, aes(x=thorax, y=lifespan))+
  geom_point()+
  geom_smooth(method=lm, se=TRUE)+
  geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y=upr), color = "red", linetype = "dashed")
```


100
75
50
25
0
lifespan
0.7
0.8
0.9
thorax