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).

```
#Create the table
data1 <- c(14, 6, 7, 7, 7, 1)
bribe <- as.table(matrix(data1, nrow = 2, ncol = 3, byrow = T))
rownames(bribe) <- c("Upper class", "Lower class")
colnames(bribe) <- c("not-stopped","bribe requested", "stopped/given
    warnings")
bribe
#Calculate expected values
addmargins(bribe)
Expected_Value <- c(27*21/42, 27*13/42, 27*8/42, 15*21/42, 15*13/42, 15*8
    /42)
Expected_Value
#Calculate chi-square statistics
```

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

```
#Calculate the p-value
p_value <- pchisq(3.791168, df = 2, lower.tail = FALSE)
p_value#[1] 0.1502306 > alpha = 0.1
#We therefore cannot reject the null hypothesis and conclude that whether
     officers were more likely to solicit a bribe from drivers are not
    independent from their classes

#Check with R
chisq.test(bribe, correct = TRUE)#same result
```

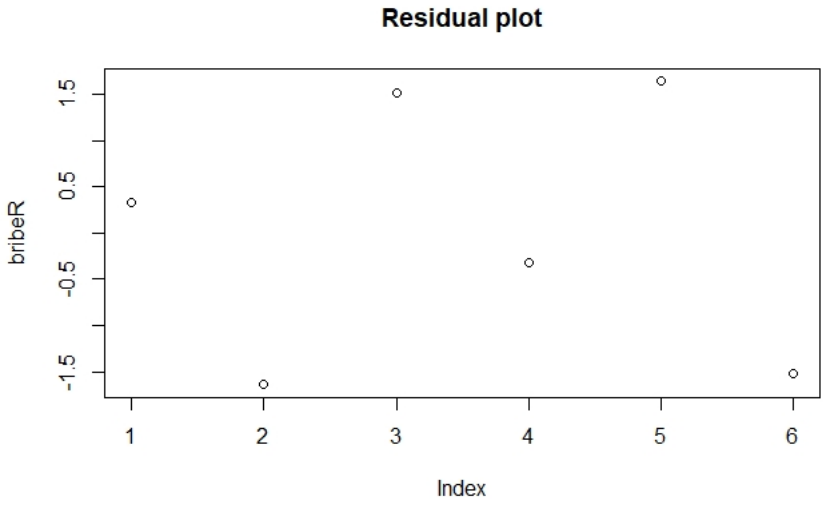
[2]Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.326	-1.642	1.523
Lower class	-0.322	1.642	-1.523

```
#Calculate standardized(adjusted) residuals
row.prop <- c(27/42, 27/42, 27/42, 15/42, 15/42, 15/42)
col.prop <- c(21/41, 13/42, 8/42, 21/42, 13/42, 8/42)
briberR <- c((data1-Expected_Value)/sqrt(Expected_Value*(row.prop-1)*(col.
    prop-1)))
#View standardized residuals
plot(briberR, main = "Residual plot")
mean(briberR)#[1] 0.0006666667
#reported it in a table
briberR <- as.table(matrix(round(briberR, digits = 3), nrow = 2, ncol = 3,
    byrow = T))
rownames(briberR) <- c("Upper class", "Lower class")
colnames(briberR) <- c("not-stopped","bribe requested", "stopped/given
    warnings")
briberR
```

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



Since our residuals are relatively large and evenly distributed, we cannot decern a significant difference between our observed value and expected value, however to make sure of the lack of dependency we have to do a chi-sqr test.

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).

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

[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.

Our null hypothesis is:
There's no association between whether a village have female leaders position reserved or not and the number of new or repaired drinking water facilities in that village
Our alternative hypothesis is:
There's an association between whether a village have female leaders position reserved or not and the number of new or repaired drinking water facilities in that village

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

```
#Import dataset
data2 <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header = T)
summary(data2)
#correlation testing on regression model between reserved and water
lm1 <- lm(data2$water~data2$reserved)
summary(lm1)
cor.test(data2$water, data2$reserved)
#Test shows that we can reject the null hypothesis and conclude that there is an significant correlation between whether a village have female leaders position reserved or not and the number of new or repaired drinking water facilities in that village
#Coefficient estimate is 9.252, which means that for each unit increase in reserved female positions, there will be 9.252 more repaired or new drinking water facilities
```

(c) Interpret the coefficient estimate for reservation policy.

Coefficient estimate is 9.252, which means that for each unit increase in reserved female positions, there will be 9.252 more repaired or new drinking water facilities

Question 3 (40 points): Biology

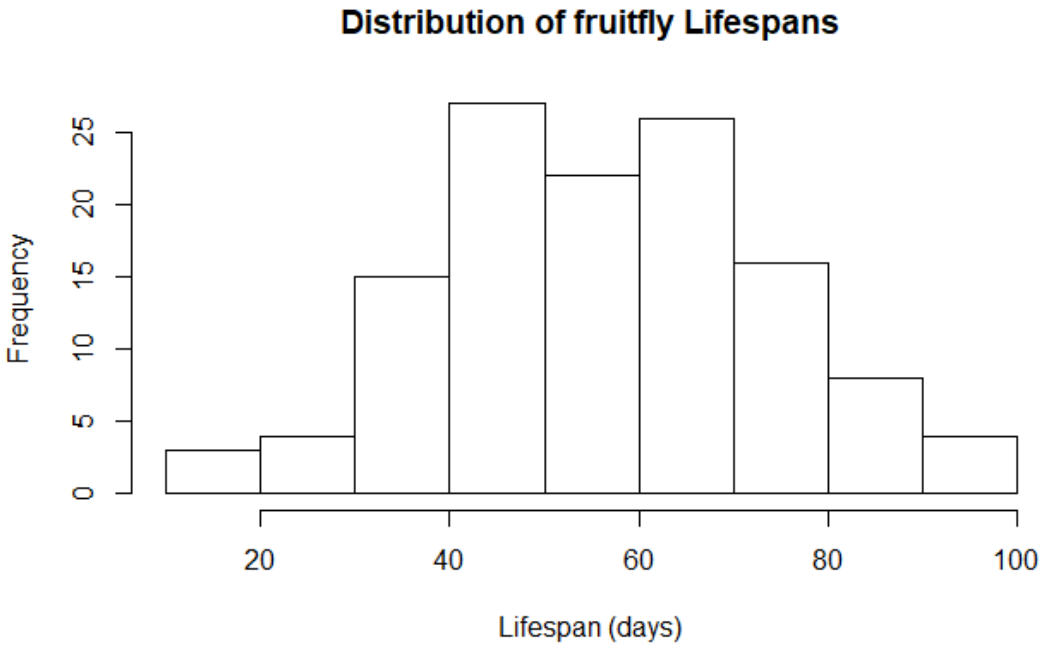
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity 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.

```
#Import dataset
data3 <- read.csv("fruitfly.csv", header = T)
summary(data3) #25 flies, mean lifespan = 57.44
#View the distribution of lifespan
hist(data3$lifespan, main = "Distribution of fruitfly Lifespans", xlab =
   "Lifespan (days)") #Approximately normal
```

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



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?

```r
#Plot lifespan and thorax and calculate the correlation coefficient
plot(data3$lifespan~data3$thorax, main="Lifespan vs. thorax", xlab ="
    length of thorax (mm)", ylab ="lifespan (days)" )
#The distribution can be discribed as a positive linear association
cor(data3$lifespan, data3$thorax)#[1] 0.6364835
#correlation coefficient is closer to 1, which means it is a fairly high
    positive correlation
```

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

```r
#Linear regression
y <- data3$lifespan
x <- data3$thorax
lm2 <- lm(y~x)
summary(lm2)
abline(lm2)
```

```
Coefficients:
Estimate      Std. Error       t value      Pr(>|t|)
(Intercept)    -61.05      13.00  -4.695  7.0e-06 ***
data3$thorax   144.33      15.77   9.152  1.5e-15 ***

Residual standard error: 13.6 on 123 degrees of freedom
Multiple R-squared:  0.4051,        Adjusted R-squared:  0.4003
F-statistic: 83.76 on 1 and 123 DF,  p-value: 1.497e-15
```

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

```
Pearson's product-moment correlation

data:  data3$lifespan and data3$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
```

```r
#test for the significance of the correlation
cor.test(y, x)
```

```
3 #p_value = 1.497e-15, which is highly significant at a 95% confidence
    level, we can therefore reject the null hypothesis and conclude that
    the correlation between lifespan and thorax is significant
```

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.

```
#90% confidence interval for the slope of the model
#Method 1
confint <- c(144.33-15.77*qt(0.90, df = 123), 144.33+15.77*qt(0.90,
    df = 123))
confint
```

     The resulting confidence interval is (124.0108, 164.6492)

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

```
#Method 2
confint(lm2, "x", level = 0.90)
```

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?

```
#predict a lifespan of a individual fruit fly
predicted_lifespan <- 0.8*144.33-61.05
predicted_lifespan #[1] 54.414
#predict the average lifespan for thorax = 0.8 and the respective
    confidence interval
predict(lm2, newdata=data.frame(x=0.8), df = 123, interval = "confidence"
    , level = 0.90)
```

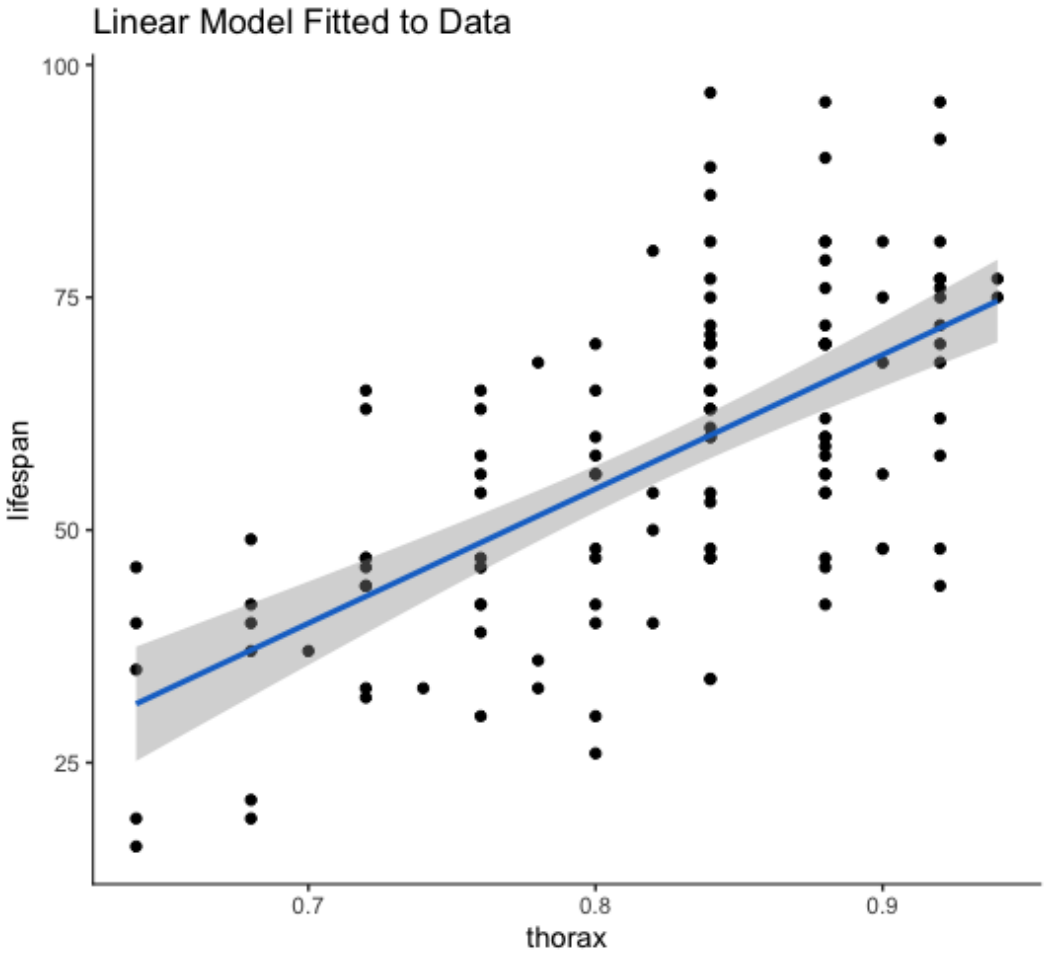
     Result:

```
       fit      lwr      upr
54.41478 52.32539 56.50416
```

     The expected lifeline is around 54 days and the confidence interval is (52.34, 56.50)

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.

```
#plot the fitted lifespan for a sequence of thorax values
install.packages("ggplot2")
library(ggplot2)
```




```
ggplot(data = data3, aes(x = thorax, y = lifespan)) +
  geom_point() +
  stat_smooth(method = "lm", col = "dodgerblue3") +
  theme(panel.background = element_rect(fill = "white"),
        axis.line.x=element_line(),
        axis.line.y=element_line()) +
  ggtitle("Linear Model Fitted to Data")
```