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 Multi-method 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`).

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

```
#calculate p value
pchisq(3.791,df=2,lower.tail=FALSE) #p=.150

#If alpha =.1, we fail to reject that null hypothesis that a driver's
    class has no bearing on whether or not they are asked for a bribe
```

[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			
Lower class			

```r
(14-13.5)/sqrt(13.5*(1-27/42)*(1-21/42)) #=.322
(6-8.357)/sqrt(8.357*(1-27/42)*(1-13/42)) #=-1.642
(7-5.143)/sqrt(5.143*(1-27/42)*(1-8/42)) #=1.523
(7-7.5)/sqrt(7.5*(1-15/42)*(1-21/42)) #=-.322
(7-4.643)/sqrt(4.643*(1-15/42)*(1-13/42)) #=1.642
(1-2.857)/sqrt(2.857*(1-15/42)*(1-8/42)) #=-1.523

#create table

table1results<-matrix(c(.322,-1.642,1.523,-.322,1.642,-1.523),ncol=3,
    byrow=TRUE)
rownames(table1results)<-c("Upper Class","Lower Class")
colnames(table1results)<-c("Not Stopped","Bribe Requested","Stopped/Given
     Warning")
table1<-as.table(table1results)
View(table1)
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	.322	-1.642	1.523
Lower class	-.322	1.642	-1.523

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

```
#They provide a sense of how far each estimate is from the center of the
    chi squared distribution. Since none of them are that large, we can
    infer that they are fairly close to this center and thus these results
     are perfectly reasonable assuming H0 is true.
```

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.

```
#Load data
library(readr)
women <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
View(women)

#Part a

#H0: There is no association between the variables reserved and water
#HA: There is an association between the variables reserved and water
```

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

```
meanx<-mean(women$reserved)
meany<-mean(women$water)
sumx<-sum(women$reserved)
sumy<-sum(women$water)

#Estimate beta
beta.numerator<-sum((women$water-meany)*(women$reserved-meanx))
beta.denominator<-sum((women$reserved-meanx)^2)
beta.hat<-beta.numerator/beta.denominator
#beta=9.252

#Solve for alpha
#meany=a + beta(meanx)
alpha<-meany-beta*meanx
#alpha = 14.738
```

(c) Interpret the coefficient estimate for reservation policy.

```
#Reserving GP for women leaders on average results in an increase of
    9.252 new or repaired drinking water facilities in the village
```

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

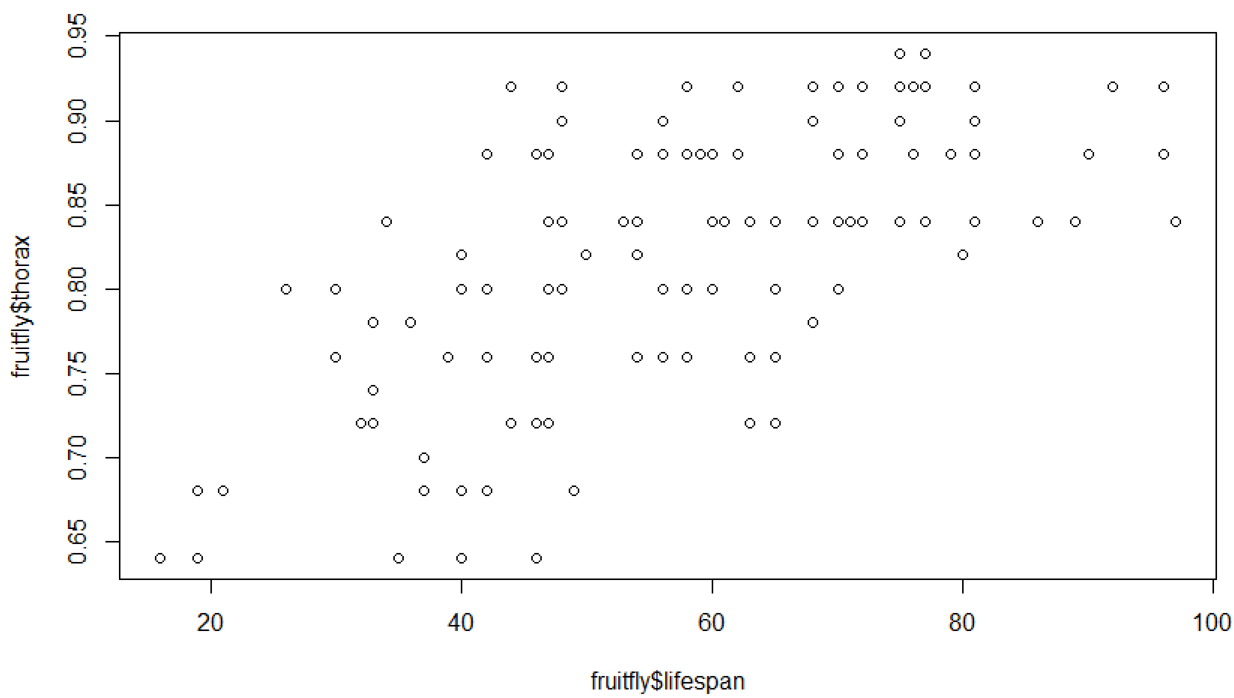
1. Import the data set and obtain summary statisticsc and examine the distribution of the overall lifespan of the fruitflies.

```
#import data
fruitfly<-read.csv("GitHub/QTM200Spring2020/problem_sets/PS2/fruitfly.csv")
summary(fruitfly$lifespan)
```

[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(fruitfly$lifespan, fruitfly$thorax)
cor(fruitfly$lifespan, fruitfly$thorax) # r=.636 -- there is a relatively
    weak positive linear relationship
```



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

```r
model1<-lm(fruitfly$thorax~fruitfly$lifespan)
model1
#intercept = .660, #coefficient=.003
#for every increase in lifespan of 1 day, there is on average an increase
    in the length of the thorax by .003 mm
alpha1<-.660
beta1<-.003
```

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

```r
sd_estimate<-sigma(model1)

#calculate standard error
beta_se<-sd_estimate/sqrt(sum((fruitfly$lifespan-mean(fruitfly$lifespan))
    ^2))
beta_se
alpha_se<-sd_estimate*sqrt((1/dim(fruitfly)[1])+(mean(fruitfly$lifespan)
    ^2/sum((fruitfly$lifespan-mean(fruitfly$lifespan))^2)))
alpha_se

#find p values
2*pt((beta-0)/beta_se,dim(fruitfly)[1]-2,lower.tail=FALSE) #=~0
2*pt((alpha-0)/alpha_se,dim(fruitfly)[1]-2,lower.tail=FALSE) #=~0
```

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`.

```
confint(model1, level=.9)
#in repeated sampling, 90% of the sample slopes will fall between .002
   and .003
```

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?

```
#fitted model
Fit<-predict(lm(fruitfly$lifespan~fruitfly$thorax),newdata=fruitfly,se.
   fit=TRUE)
Fit
#the model predicts a lifespan of 48.641 days

#confidence interval
CI<-predict(lm(fruitfly$lifespan~fruitfly$thorax),newdata=fruitfly,
   interval="confidence",level=.95)
CI
#the confidence interval is (45.572,51.711)

#prediction interval
PI<-predict(lm(fruitfly$lifespan~fruitfly$thorax),newdata=fruitfly,
   interval="prediction",level=.95)
PI
#the prediction interval is (21.543,75.740)
```

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(Fit$fit)
lines(CI)
lines(PI)
```