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).

```
#Q1
#make the data given into a tble by using matrix
data<-c(14,6,7,7,7,1)
P=matrix(c(14,6,7,7,7,1), nrow=2, ncol=3,byrow=TRUE)
dimnames(P) = list(c("upper", "lower"), c("not-stop", "bribe", "stop"))
addmargins(P)#show the sum of each row and column
exp<-c(27*21/42, 27*13/42, 27*8/42, 15*21/42, 15*13/42, 15*8/42)#
    calculate the expected value
chi_square<-sum((data-exp)^2/exp)
chi_square#calculate chi-square
#we get the result of chi-square [1] 3.791168
```

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

```
#calculate p-value
p_value=pchisq(3.791168, df=2, lower.tail = F)
p_value
#we get the result of p_value [1] 0.1502306
#A p-value higher than the significance level means that we fail to
    reject the null hypothesis and conclude that whether officers were
    more likely to solicit a bribe from drivers are not independent from
    their classes.
```

[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	0.136	-0.815	0.819
Lower class	-0.183	1.094	-1.099

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

```
#calculate standardized residuals
SR=c((data-exp)/(exp^.5))
SR
rowp<-c(27/42, 27/42, 27/42, 15/42, 15/42, 15/42)
colp<-c(21/42, 13/42, 8/42, 21/42, 13/42, 8/42)
#calculte adjusted residuals
z=c((data-exp)/(exp*(1-rowp)*(1-colp))^.5)
z
#view standardized residuals
plot(SR)
mean(SR)
#we get the result [1] -0.007950638
#the residuals are very small(almost 0) and evenly distributed which does
    not indicate a significant difference between the observed value and
   expected value
```

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.

```
#Q2
#the null hypothesis is that 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
#the alternative hypothesis is that there exist 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!).

```
#the null hypothesis is that 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
#the alternative hypothesis is that there exist 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
#import data
library(readr)
women <- read_csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
lm<-lm(formula = water ~ reserved, data = women)
summary(lm)
#we get R_square as 0.01688
anova(lm)
R_square=c(1.211/(1.211+70.565))
R_square#calculate R_square
#we get R_square [1] 0.01687193 same as above
#run correlation test
cor.test(women$water, women$reserved)
#since p-value(= 0.0197) <0.05 we reject the null hypothesis and conclude that there is an significant correlation beween the two variable
```

(c) Interpret the coefficient estimate for reservation policy.

```
#the coefficient of determination is low (0.01688), indicating only 1% of
    points fall within the regression line.
#the coefficient estimate is 9.252, meaning for each unit increase in
   reserved female positions there will be 9.252 more 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.

```
#Q3
#import data
library(readr)
ff <- read_csv("fruitfly.csv")
summary(ff)#there's 25 flies, meaninf the lifespan is 57.44
#the distribution of lifespan
hist(ff$lifespan)
#we can see that it's 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?

```
#plot lifespan vs thorax graph
plot(lifespan~ thorax, main="lifespan vs thorax", xlab="thorax ", ylab="
    lifespan")
#the graph is a positive linear association
#calculate the correlation coefficient
cor(ff$lifespan, ff$thorax)
#we get the correlation coefficient [1] 0.6364835, which is close to 1,
    meaning a relatively highly positive correlation
```

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

```
#do a linear regression
lmff<-lm(lifespan~ thorax)
summary(lmff)
abline(lmff)
#the slope is -61.05, which is negetive meaning that when the one
    variable increases, the other decreases
```

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

```
#do cor.test for the significance of correlation
cor.test(ff$lifespan, ff$thorax)
#we have p_value = 1.497e-15, which can be considered highly significant
    at the 95% confidence level so we 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.
   - Now, try using the function `confint()` in `R`.

```
#90% confidence interval for the slope
#by using the formula for typical confidence intervals
error<-confint<-c(144.33-15.77*qt(0.9, df = 123), 144.33+15.77*qt(0.9, df
    = 123))
confint#we get the result [1] 124.0108 164.6492
#by using the function confint
confint(lmff, parm = "thorax", level = 0.9)
#             5 %   95 %
#thorax 118.1962 170.47
#we get the result 118.1962 170.47, which is not consistent since it's a
   95%
confint(lmff, parm = "thorax", level = 0.8)
#            10 %     90 %
#thorax 124.0133 164.6529
#we get the result 124.0133 164.6529, which is about the same as above
```

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 fruitfly
pl<-0.8*144.33-61.05
pl #we get result [1] 54.414
#predict the average lifespan for thorax = 0.8 and confidence intervals
plmff <- predict.lm(lmff, thorax = 0.8, df = 123, interval = "confidence"
    )
#calculate the average
mean(plmff) #we get the average[1] 57.44
#Find the confidence interval at 95%
summary(plmff)  #we get the mean of lwr as 54.16 and the mean of upr as
    60.72
#so the confident interval of the lwr and upr is (54.16, 60.72)
```

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 with prediction
    intervals and confidence intervals
#add predictions
pred.int<-predict(lmff, interval = "prediction")
ffp<-cbind(ff, pred.int)
#plot regression line and confidence intervals
install.packages("ggplot2")
library(ggplot2)
p <- ggplot(ffp, aes(x = thorax, y = lifespan)) +
  geom_point() +
  stat_smooth(method = "lm", col = "blue") +
  theme(panel.background = element_rect(fill = "white"),
        axis.line.x=element_line(),
        axis.line.y=element_line()) +
  ggtitle("Fitted Linear Model of lifespan vs thorax")
#add prediction intervals
p + geom_line(aes(y = lwr), color = "red", linetype = "dashed")+
    geom_line(aes(y = upr), color = "red", linetype = "dashed")
p#plot with prediction intervals in red dash and confidence intervals in
   gray
```

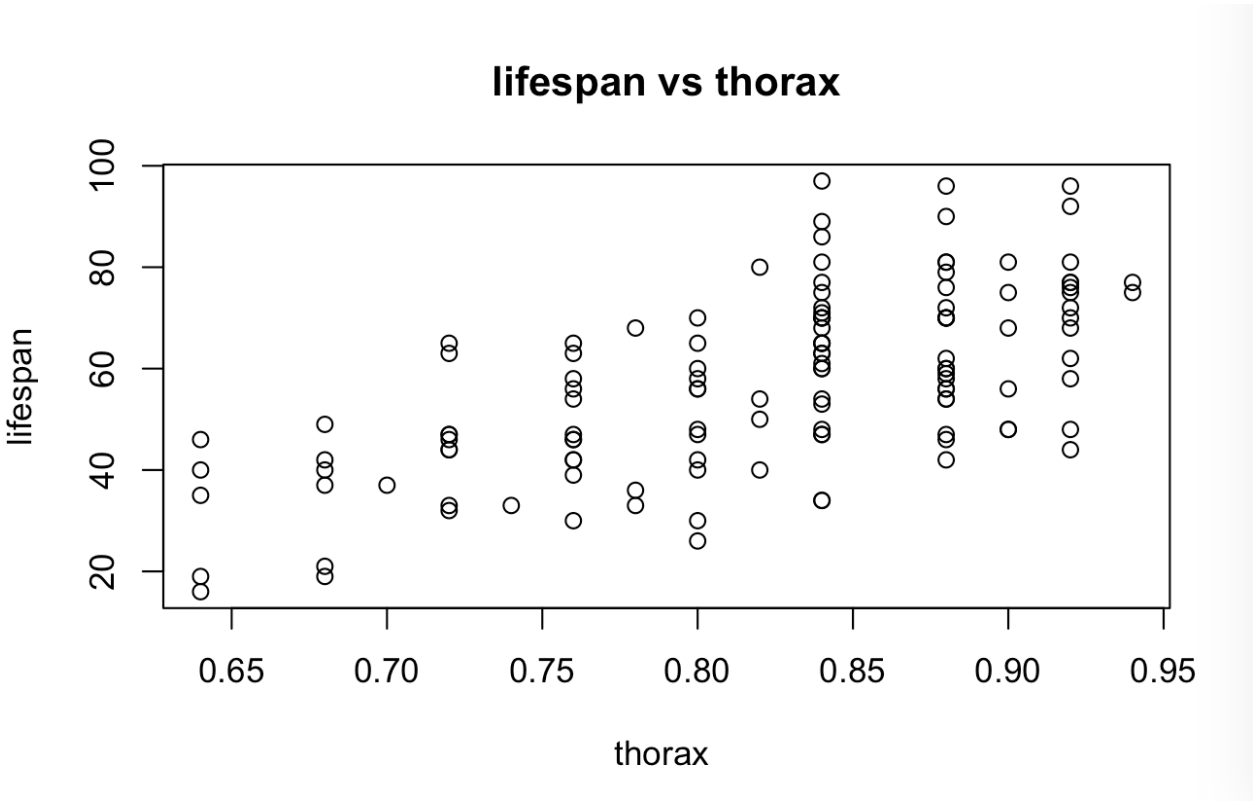

Figure 2: Lifespan vs Thorax.

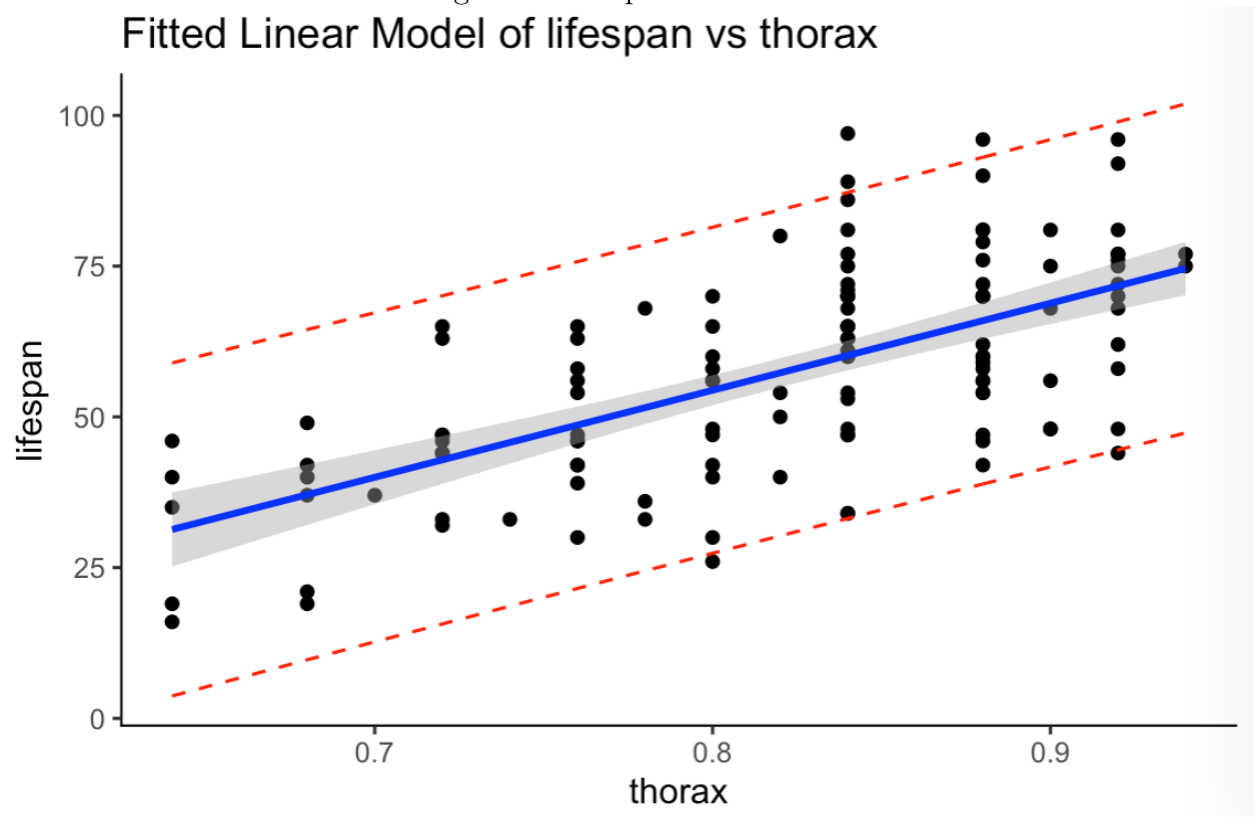

Figure 3: Fitted Linear Model of lifespan vs thorax.