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).

```
#Chi-squared test of indepndence
tbl <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, ncol=3)
rownames(tbl) = c(”Upper class”, ”Lower class”)
colnames(tbl) = c(”Not Stopped”, ”Bribe requested”, ”Stopped/given
    warning”)
tbl
```

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

```
#Chi-squared test of indepndence
tbl <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, ncol=3)
rownames(tbl) = c(”Upper class”, ”Lower class”)
colnames(tbl) = c(”Not Stopped”, ”Bribe requested”, ”Stopped/given
    warning”)
tbl
chisq<- chisq.test(tbl)
pchisq(3.4125, df=2, lower.tail = FALSE)
```

As the p value came out to be 0.1815, which is larger than $\alpha = .1$, we fail to reject the null hypothesis which is that the variables 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	0.4369314	-1.620185	1.389297
Lower class	-0.4369314	1.620185	-1.389297

```r
se <- sqrt(13.333*(1-(27/42))*(1-(21/42)))
(14-13.33333)/se
chisq$stdres
```

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

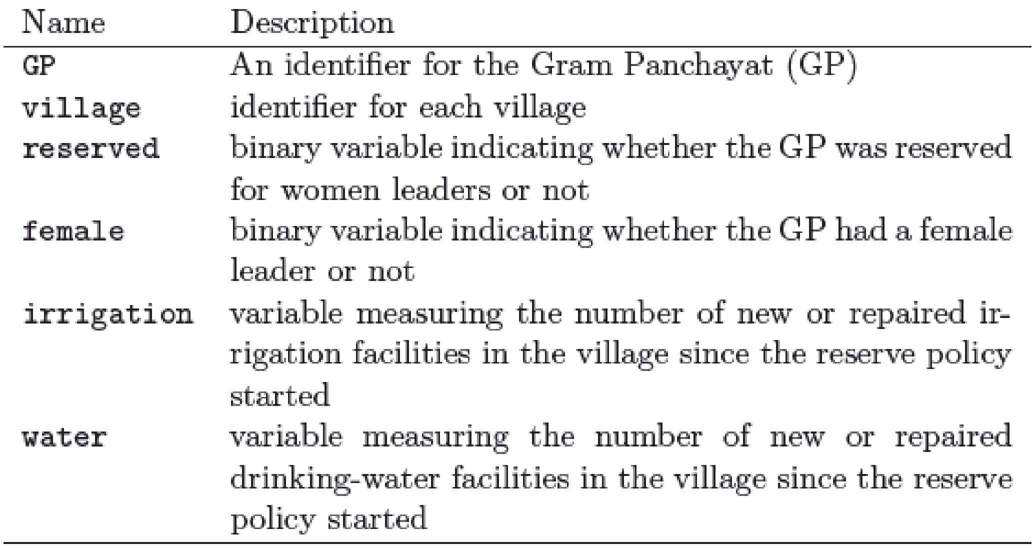
Standardized residuals help understand how far away each observed value is from ”expectation”. Since we could reject the null, and cannot disprove that the two variables are independent.

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.

null hypothesis: there is no effect of the reservation policy on the number of new or repaired drinking water facilities in the villages

$\mu = \mu 0$

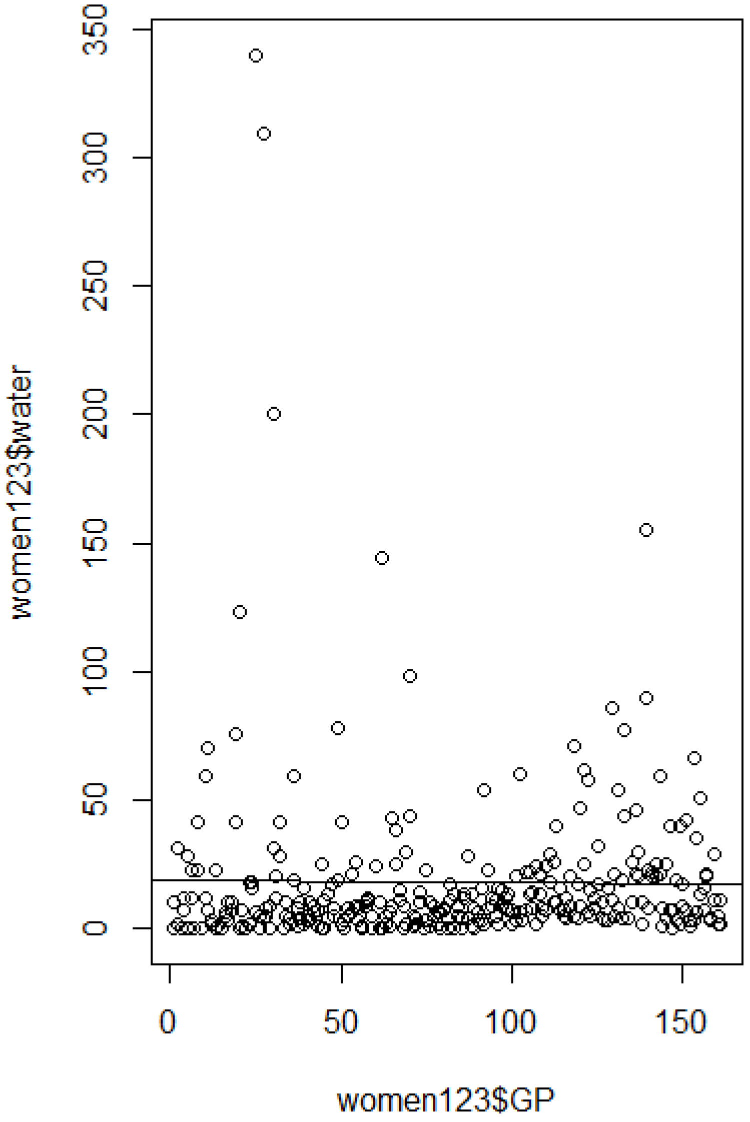
alternative hypothesis: there is an effect of the reservation policy on the number of new or repaired drinking water facilities in the villages

$\mu \neq \mu 0$

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

```r
#plotting water versus GP
plot(women123$water~women123$GP)
#adding the linear regression model
temp.model <- lm(women123$water~women123$GP)
temp.model
abline(temp.model)
```

The $\beta 0 = 18.380474$, which is the equivalent of an intercept The $\beta 1 = -0.006653$, which is the equivalent of the slope.


women123$water
women123$GP
0
50
100
150
200
250
300
350
0
50
100
150

(c) Interpret the coefficient estimate for reservation policy.

```r
c(round(mean(women123$water), 2), round(sd(women123$water),2)) #the mean and standard deviation of water
standardized.x <- ((women123$water - mean(women123$water)))/sd(women123$water) #create standardized distance for each observation of water
round(standardized.x, 2) #vecot of standaridized values
r <- (1/(322-1))*sum(standardized.x) #computing correlation coefficient
r
```

The coefficient estimate for reservation policy came out to be $1.197 * 10^{-17}$, which illustrates that value is close to 0, illustrating no correlation between the reservation policy on the number of new or repaired drinking water facilities in the villages.

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.

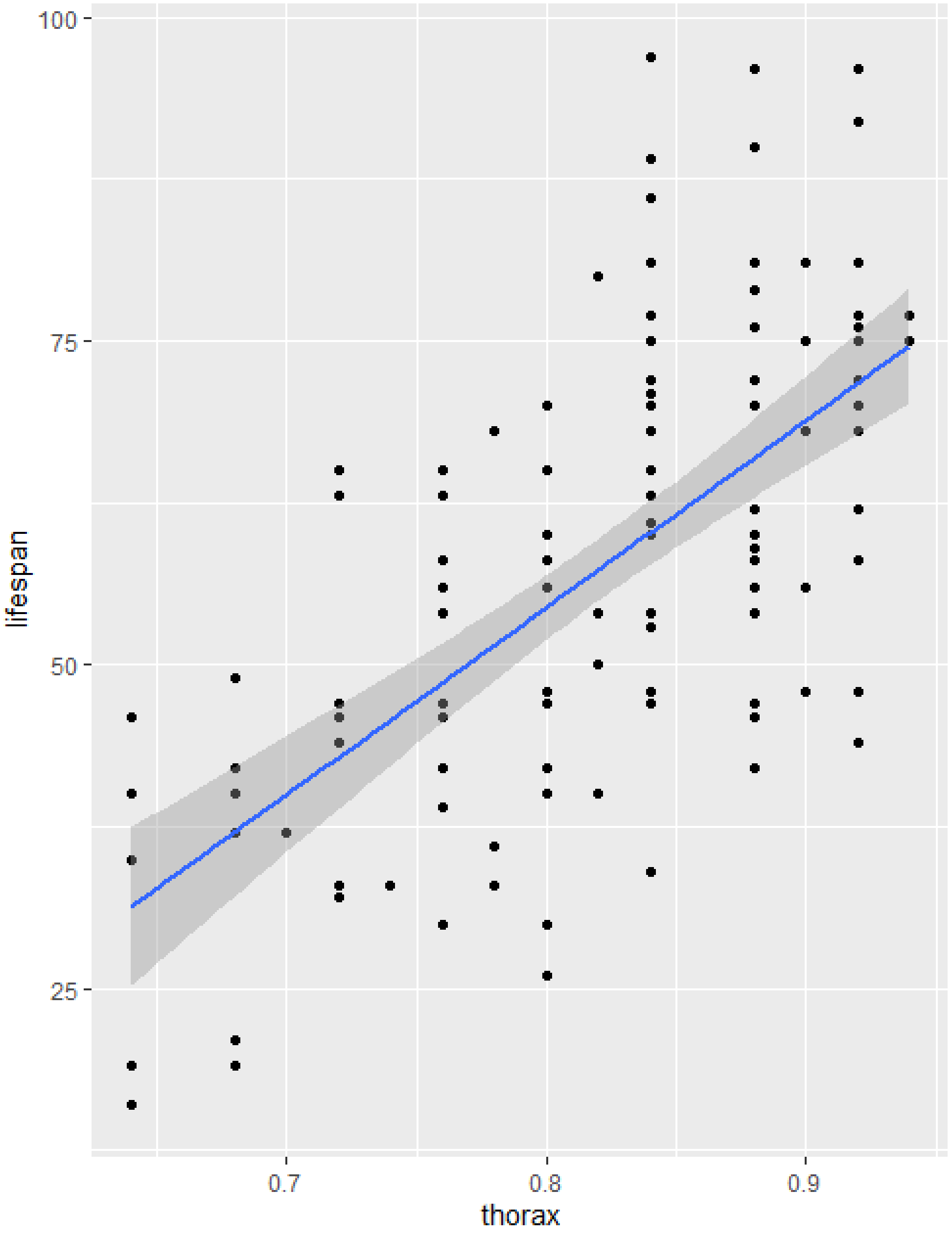
```r
#importing data
library(readr)
fruitfly <- read_csv("~/GitHub/QTM200Spring2020/problem_sets/PS2/fruitfly.csv")
View(fruitfly)
summary(fruitfly)
```

   The distribution of the overall lifespan of fruit files have shown that the minimum lifespan is 16 days, 1st quartile is 46 days, while the median is 58 days. The 3rd quartile is 70 days and the maximum lifespan of the data was 97 days. The average of the overall lifespan is 57 days.

[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?

```
#lifespan vs thorax
ggplot(fruitfly , aes(x= thorax, y=lifespan))+
  geom_point()+
  geom_smooth(method=lm)
```



The plot illustrates a positive linear relationship, illustrating that as the length of thorax increases, the lifespan of the fruit fly seems to increase.

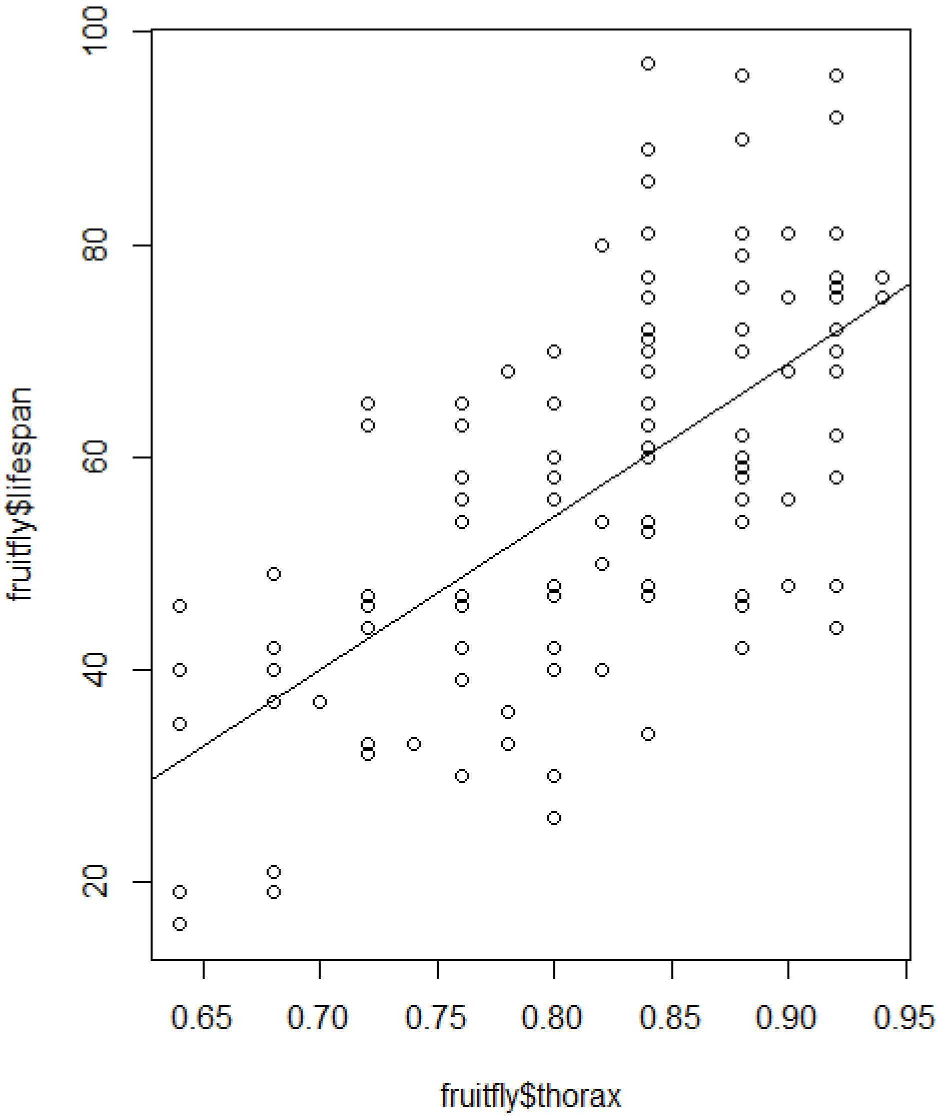
```r
#correlation coefficient
#for y value
c(round(mean(fruitfly$lifespan), 2), round(sd(fruitfly$lifespan),2))
#for x value
c(round(mean(fruitfly$thorax), 2), round(sd(fruitfly$thorax),2))

standardized.lifespan <- (fruitfly$lifespan - mean(fruitfly$lifespan))/sd(fruitfly$lifespan)
standardized.thorax <- (fruitfly$thorax- mean(fruitfly$thorax))/sd(fruitfly$thorax)
r_fruitfly <- (1/(125-1))*sum(standardized.lifespan*standardized.thorax)
r_fruitfly
```

The correlation variable is 0.636, illustrating a positive correlation between the two variables. This means that as lifespan of the fruit fly increases, the length of the thorax increases as well.

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

```r
#slope
plot(fruitfly$lifespan~fruitfly$thorax)
lst.model <- lm(fruitfly$lifespan~fruitfly$thorax)
lst.model
abline(lst.model)
```



The slope came out to be 144.33, illustrating that as the thorax length increases by 1mm, there is a 144.33 day increase in lifespan.

The y-intercept was shown to be -61.05, illustrating that when the thorax is 0mm, the fruitfly could not live.

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

```r
#significant linear relationship between lifespan and thorax
#beta_1
beta <- sum((fruitfly$lifespan - mean(fruitfly$lifespan))* (fruitfly$thorax - mean(fruitfly$thorax)))/sum((fruitfly$thorax - mean(fruitfly$thorax))^2)
beta
reg1 <- lm(lifespan~thorax, data = fruitfly) # checking regression to check if our estimates are correct
reg1

sd_estimate <- sqrt(sum(resid(reg1)^2/(dim(fruitfly)[1]-2)))
sd_estimate

sigma(reg1)

# SE for beta_1
beta_se <- sd_estimate/sqrt(sum((fruitfly$thorax - mean (fruitfly$thorax))^2))
beta_se

2*pt((beta-0)/beta_se, dim(fruitfly)[1]-2, lower.tail = F)

#to check the right p-value
summary(lm(fruitfly$thorax~fruitfly$lifespan))
```

As the p value is $1.5 * 10^{-15}$, the p value is very close to 0, leading us to reject the null hypothesis of $\beta = 0$, and thus we reject the hypothesis that lifespan and thorax are independent of each other. This illustrates that there is a relationship between lifespan and thorax in the linear regression model of the data set.

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.

     ```r
     #90% confidence interval for the slope
     summary(lst.model)
     #slope
     b1 <- 144.33
     #standard error of slope value
     s <- 15.77
     # size
     n <- 125
     # t-value
     t <- abs(qt(0.1/2, df = n-2)) #because the test is a two-tailed test
     t
     #left and right confidence intervals
     left <- b1 - s*t
     right <- b1 + s*t
     left
     right
     ```

     The confidnece interval for the slope of the fitted model is (118.1938, 170.4662)

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

   ```r
   #using confint
   confint(lst.model, level = 0.9)
   ```

   The confidence interval came out to be (118.19616, 170.4700) when using the confint function in R.

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?

   ```r
   #individual fruitfly's lifespan when thorax=0.8
   new_fruitfly<- fruitfly;
   new_fruitfly$thorax <- 0.8
   ```

```
p <- as.data.frame(predict(lst.model, newdata = new_fruitfly, interval = ”prediction”))
mean(p$fit)

#average lifespan of fruitflies
c<- as.data.frame(predict(lst.model, newdata = new_fruitfly, interval = ”confidence”))
mean(c$upr)
mean(c$lwr)
```

When predicting an individual fruit fly's lifespan when thorax = 0.8, it was found that the predicted lifespan would be 57.44, which is around 57 days.

When predicting the average lifespan of fruit flies when thorax = 0.8 was found through the confidence intervals of (54.15853, 60.72147), which is around 54 to 61 days.

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.