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).

```r
# Create a matrix to store the data for Question 1
Q1_data <-matrix(c(14,6,7, 27,7,7,1, 15, 21, 13, 8, 42),nrow=3,byrow = TRUE)
rownames(Q1_data)<-c("Upper class","Lower class", "Total columns")
colnames(Q1_data)<-c("Not Stopped","Bribe Requested","Stopped/given warning","Total rows")
# Calculate the X^2 statistic using the formula: [(observed-expected)/expected]. Observed values were the values reported in the data set. Expected values are calculated using the formula: [(row total/grand total)*column total].
expected1<- (27/42)*21
expected2<- (27/42)*13
expected3<-(27/42)*8
expected4<-(15/42)*21
expected5<-(15/42)*13
expected6<-(15/42)*8
chi2_1 <-((14-expected1)^2)/expected1
chi2_2 <-((6-expected2)^2)/expected2
chi2_3<-((7-expected3)^2)/expected3
chi2_4<-((7-expected4)^2)/expected4
chi2_5<-((7-expected5)^2)/expected5
chi2_6<-((1-expected6)^2)/expected6
chi2_statistic<-sum(chi2_1,chi2_2,chi2_3,chi2_4,chi2_5,chi2_6)
chi2_statistic
# SOLUTION: The chi square statistic is 3.791168
```

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

```r
# Calculate the p-value for the test statistic. Degrees of freedom (df) is calculated using the formula: df = (rows-1)*(columns-1)
df<- (2-1)*(3-1)
pchisq(3.791168,df= df,lower.tail = FALSE)
# The p-value is 0.1502306.
# SOLUTION: The p-value is greater than the significance level (0.1502306>0.1). Therefore, fail to reject the null hypothesis (Ho= there is no association between class and bribery). There is not sufficient evidence to conclude that class and bribery are associated.
```

[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	0.3220306	-1.093382	1.182516
Lower class	-0.2014441	1.260318	-1.303106

```r
# Calculate the standardized residuals using the formula: z = [(f.observed-f.expected)/se]. Calculate the standerd error (se) using the formula: sqrt[f.expected*(1-row prop)*(1-column prop)]
residual1<- (14-expected1)/(sqrt(expected1*(1-(27/42))*(1-(21/42))))
residual2<-(6-expected2)/sqrt(expected2*(1-(27/42)*(1-(13/42))))
residual3<-(7-expected3)/sqrt(expected3*(1-(27/42)*(1-(8/42))))
residual4<-(7-expected4)/sqrt(expected4*(1-(15/42)*(1-(21/42))))
residual5<-(7-expected5)/sqrt(expected5*(1-(15/42)*(1-(13/42))))
residual6<-(1-expected6)/sqrt(expected6*(1-(15/42)*(1-(8/42))))
```

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

```r
# Standardized residuals show how far away each observed value is from the expected value. Based on the results, the"Not Stopped" individuals for both upper class and lower class individuals had observed values closest to the expected values (0.3220306 and -0.2014441, respectively). The lower class individuals who were stopped/given a warning deviated the most from the expected values (standard residual = -1.303106). The high standard residual for the lower class individuals who were stopped/given a warning might suggest that the null hypothesis (there is no association between class and bribery) is not true for this group. However, the residuals were still too low to reject the null hypothesis.
```

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 (Ho):The reservation policy has no effect on the nuber of new or repaired drinking water facilities in the villages.
# Alternative hypothesis (Ha): The reservation policy has an effect on the nuber of new or repaired drinking water facilities in the villages.
```

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

```
# Bivariate regression by hand
regressMat <- as.data.frame(matrix(c(women$reserved,women$water), nrow = 322,byrow = FALSE))
colnames(regressMat)<- c("reserved","water")
regressMat
# Calculate sums and means
mean_reserved<- mean(regressMat$reserved)
mean_water<- mean(regressMat$water)
sum_reserved<-sum(regressMat$reserved)
sum_water<- sum(regressMat$water)
# Calculate beta (slope) and alpha (y intercept)
beta <- sum((women$reserved-mean_reserved)*(women$water-mean_water))/sum((women$reserved-mean_reserved)^2)
alpha <- mean_water-(beta*mean_reserved)
# Alpha (intercept) equals 14.73832, beta (slope) equals 9.252423
# Check with lm() in R
lm(women$water~women$reserved, data=regressMat)
# SOLUTION: # Alpha (intercept) equals 14.73832, beta (slope) equals 9.252423. (when checked with the lm() function, the R output was: Intercept =  14.738  women$reserved  = 9.252. This is consistent with the results I calculated).
```

(c) Interpret the coefficient estimate for reservation policy.

```
# The alpha (9.252) gives us the slope of the linear relationship. For every increase of 1 in regards to the GP being reserved for women, the number of new or repaired drinking-water facilities in the village increases by 9.252. If the GP being reserved for women is 0 (x=0), the y-intercept (beta) is 14.738. This is the number of new or repaired drinking-water facilities in the village if there were no reservations for women on the GP.
```

Question 3 (40 points): Biology

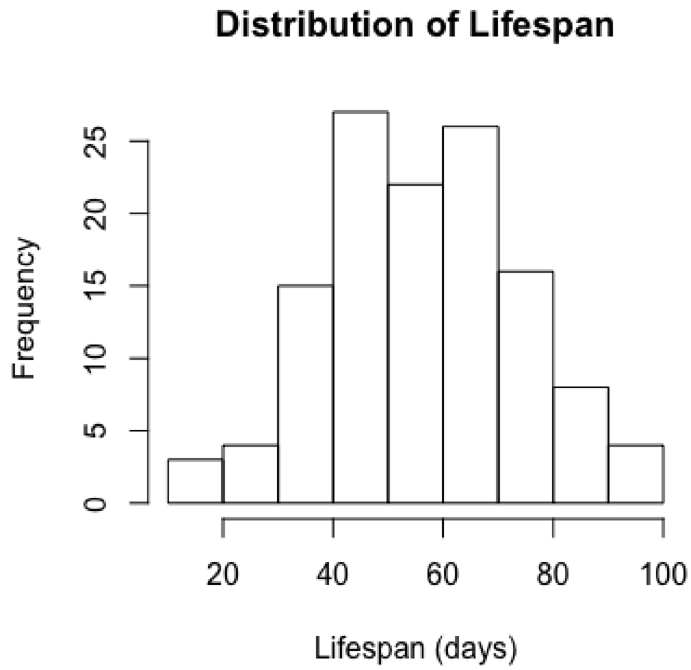
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity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

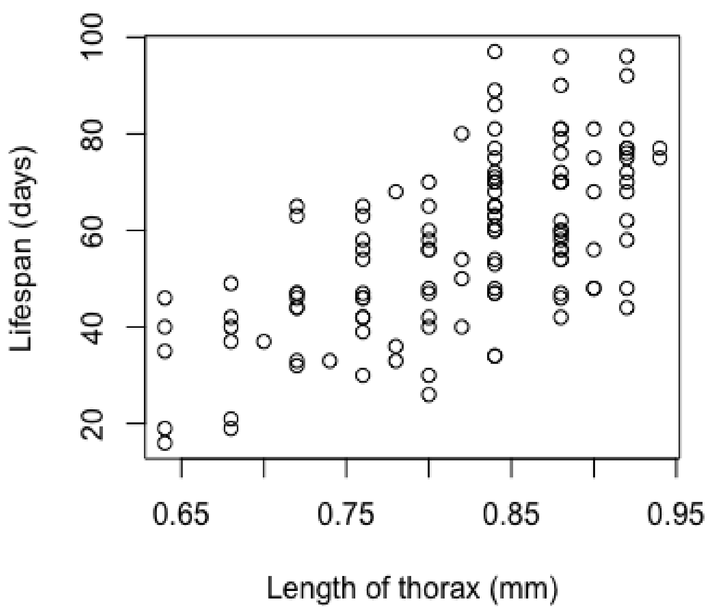
```r
# Summarize the fruitfly data, and use a histogram to examine the distribution
summary(fruitfly)
No            type        lifespan          thorax            sleep
Min.   : 1    Min.   :1   Min.   :16.00   Min.   :0.640   Min.   : 1.00
1st Qu.: 7    1st Qu.:2   1st Qu.:46.00   1st Qu.:0.760   1st 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, main = "Distribution of Lifespan", xlab = "Lifespan (days)")
# There is an approximately normal distribution of fruitflies based on their lifespan. The distribution is centered at the mean value of 57.44 days.
```

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


Distribution of Lifespan
Frequency
0
5
10
15
20
25
20
40
60
80
100
Lifespan (days)

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 and calculate the correlation coefficient
plot(fruitfly$thorax,fruitfly$lifespan, xlab = "Length of thorax (mm)",
    ylab = "Lifespan (days)")
cor(fruitfly$thorax,fruitfly$lifespan,method = "pearson")
# The correlation coeffieient is  0.6364835, and there appears to be a
    linear relationship. There is a moderate, positive linear association
    between the two variables.
```



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

```
# Run a regression for the variables thorax and lifespan
lm1<- lm(fruitfly$lifespan~fruitfly$thorax)
lm1
# The slope is 144.33. This meand that for every 1mm increase in length
    of thorax, the lifespan of fruitflies increases by 144.33 days.
```

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

```
# Use cor.test to test for the significance of the linear relationship
    between the two variables.
cor.test(fruitfly$lifespan,fruitfly$thorax)
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

# The p-value is 1.497e-15, which is less than a significance level of a =0.05 (1.497e-15<0.05). Since there is such a small p-value, reject the null hypothesis (Ho: there is no correlation between the length of the thorax and the lifespan of fruitflies).
```

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.

```
# Calculate the confidence interval for the slope at 0.90
    significance using the formula: slope +/- zscore*standard error.
summary(lm1)
slope = 144.33
zscore = 1.645
standard error = 15.77
upper <- 144.33 + (1.645*15.77)
lower <-144.33 - (1.645*15.77)
# SOLUTION: The 90% confidence interval for the slope is
    (118.3884,170.2717)
```

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

```
# Use the command confint() to calculate the confidence interval for the
    slope at 0.90
confint(lm1, level = 0.90)
# SOLUTION: The 90% confidence interval for the slope is (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?

```
# Calculate the prediction interval and confidence interval
new_fruitfly <- fruitfly
new_fruitfly$thorax<- 0.8
prediction_interval <- predict(lm(fruitfly$lifespan~fruitfly$thorax),
    newdata = new_fruitfly, se.fit = T, interval = "prediction", level=
    0.90)
confidence_interval <- predict(lm(fruitfly$lifespan~fruitfly$thorax),
    newdata = new_fruitfly, se.fit = T, interval = "confidence", level=
    0.90)

#Expected value of lifespan for an individual with thorax = 0.8
# from prediction interval: fit = 54.41478 lwr = 31.775371 upr = 77.05419
# from confidence interval: fit = 54.41478 lwr = 52.32539 upr = 56.50416

# Graph the confidence interval
```

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.

```
# Graph the confidence interval
plot(fruitfly$thorax,fruitfly$lifespan, xlab="Thorax Length (mm)", ylab="
    Lifespan (days)")
lines(fruitfly$thorax, fitted(lm1), col="blue")

#The blue line represents the fitted regression. I was unable to figure
    out how to graph the confidence intervals.
```