Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2021

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 GitHub in `.pdf` form.
- This problem set is due before class on Friday October 15, 2021. 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).

I did my chi square calculations by hand in R. First I created a matrix that would

resemble the table provided in the task.

```r
mat1.data <- c(14, 6, 7, 7, 7, 1)
mat1 <- matrix(mat1.data,nrow=2,ncol=3,byrow=TRUE)
mat1
rownames(mat1) <- c("Upper class","Lower class")
colnames(mat1) <-c("Not Stopped","Bribe requested","Stopped/given warning
    ")
mat1
```

After that I calculated the sums of all rows and columns in the table and I calculated my total value of all rows and columns.

```r
rowt1 <- sum(14,6,7)
rowt2 <- (7+7+1)
colt1 <- (14+7)
colt2 <- (6+7)
colt3 <- (7+1)

grad_total <- sum(rowt1, rowt2, colt1, colt2, colt3)
```

After that I calculated the expected frequencies of every observation in the table using formula: (total value of rows-grand total)/column total.

```r
f1exp <- (rowt1/grad_total)*colt1
f2exp <- (rowt1/grad_total)*colt2
f3exp <- (rowt1/grad_total)*colt3
f4exp <- (rowt2/grad_total)*colt1
f5exp <- (rowt2/grad_total)*colt2
f6exp <- (rowt2/grad_total)*colt3
```

After that I calculated the quotient of expected and observed frequencies for each number in the table, following the formula: (observed frequencies-expected frequencies)squared/expected frequencies:

```r
f1 <- (14-f1exp)^2/f1exp
f2 <- (6-f2exp)^2/f2exp
f3 <- (7-f3exp)^2/f3exp
f4 <- (7-f4exp)^2/f4exp
f5 <- (7-f5exp)^2/f5exp
f6 <- (1-f6exp)^2/f6exp
```

Sum of these quotients represents the chi square results which I got using the following formula:

```
chisq <- sum(f1,f2,f2,f4,f5,f6)
```

The final result of my chi square calculations is 21.75.

(b) Now calculate the p-value from the test statistic you just created (in `R`).[2] What do you conclude if $\alpha = .1$?
To calculate my p value I first calculated my degrees of freedom by multiplying the number of rows and columns, both subtracted with 1 and following the standard formula for calculating the p value in R:

```
df <- (2-1)*(3-1)
p.value <- pchisq(chisq, df, lower.tail=FALSE)
p.value
```

As a way to check my results I ran a chisq.test function in R to compare the results, taking into account that there are cells in the table that have the frequency ¿5.

```
chisq.results <- chisq.test(mat1, correct = FALSE)
chisq.results
```

Because the number of observations is not suited for calculations of the p value, there are inconsistencies. If calculated by hand, my p value = 1.893205e-05 and the p value got from the chi.square function is p = 0.1502. To put the meaning of alpha value in context, I assume that: H0: class and bribe are independent variables. H1: class and bribe are not independent variables. If alpha is the .1 then for my calculated p value I take that class and bribe are not independent variables, while function calculated p value is slightly bigger than alpha (I take that alpha is 0.10000), then it suggests how class and bribe are independent variables.

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	6.76	2.74	10.86
Lower class	3.70	8.61	-1.28

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

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

They are useful in helping to interpret chi-square tables by providing information about which cells contribute to a significant chi-square. If the standardized residual is beyond the range of ± 2, then that cell can be considered to be a major contributor, if it is ¿ +2, or a very weak contributor, if it is beyond -2, to the overall chi-square value. A large standardized residual provides evidence against independence in that cell. When H0 is true, there is only about a 5% chance that any particular standardized residual exceeds 2 in absolute value. From these residuals we see that the greatest contribution to the dependence of class and bribe provide the value of 10.86 for upper class and 8.61 for lower class, pointing that class and bribe are not 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).

[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. I constructed my hypothesis following the rule that H0 hypothesis indicates no relationship between the variables:

H0: There is no effect of reservation policy on the number of new or repaired drinking water facilities. H1: There is an effect of reservation policy on the number of new or repaired drinking water facilities.

(b) Run a bivariate regression to test this hypothesis in `R` (include your code!). First I read the .csv file and for a pre-check I ran a linear model using the lm function in R:

```r
women <- read.csv ("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv", fill= TRUE)
```

```r
class(women)
typeof (women)

influence <- lm(women$water ~ women$reserved, data = women)
summary(influence)
```

After that I did my regression by hand:

```r

# BETA 1 #
beta <- sum((women$water - mean(women$water)) * (women$reserved - mean(
    women$reserved)))/
  sum((women$reserved - mean(women$reserved))^2)
beta

# Alpha #
alpha <- mean(women$water)-beta*mean(women$reserved)
alpha

reg_DF <- as.data.frame(cbind(women$reserved,women$water))
reg_DF

# STANDARD EROR #
sig <- sigma(influence)
sig
```

```

# SE FOR BETA 1 #

beta_se <- sig/sqrt(sum((reg_DF$V1)-mean(reg_DF$V1))^2)
beta_se

# SE FOR ALPHA #


alpha_se <- sig * sqrt((1/dim(reg_DF) [1]) + (mean(reg_DF$V1)^2)/sum((reg_DF$V1-mean(reg_DF$V1))^2))
alpha_se

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

#alpha
alphaC <- 2*pt((alpha-0)/alpha_se, dim(reg_DF) [1]-2, lower.tail = F)
```

(c) Interpret the coefficient estimate for reservation policy.

In the model generated using the function my estimate beta coefficient is 9.25, which indicates that one more reserved space for women increases the number of new and repaired water facilities. My calculated beta is the same. Standard error in the model and by my calculations are the same (std. eror = 2.28) small levels of standard deviation or that means in the sample could differ from the estimated mean of the population for two standard deviations on average. My p value is the same as in the model calculated by function and suggests that there is statistical significance in the relationship between seats reserved for women and number of repaired water facilities (.01 ¡ .05).

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. I inspected my dataset using the summary function in R:

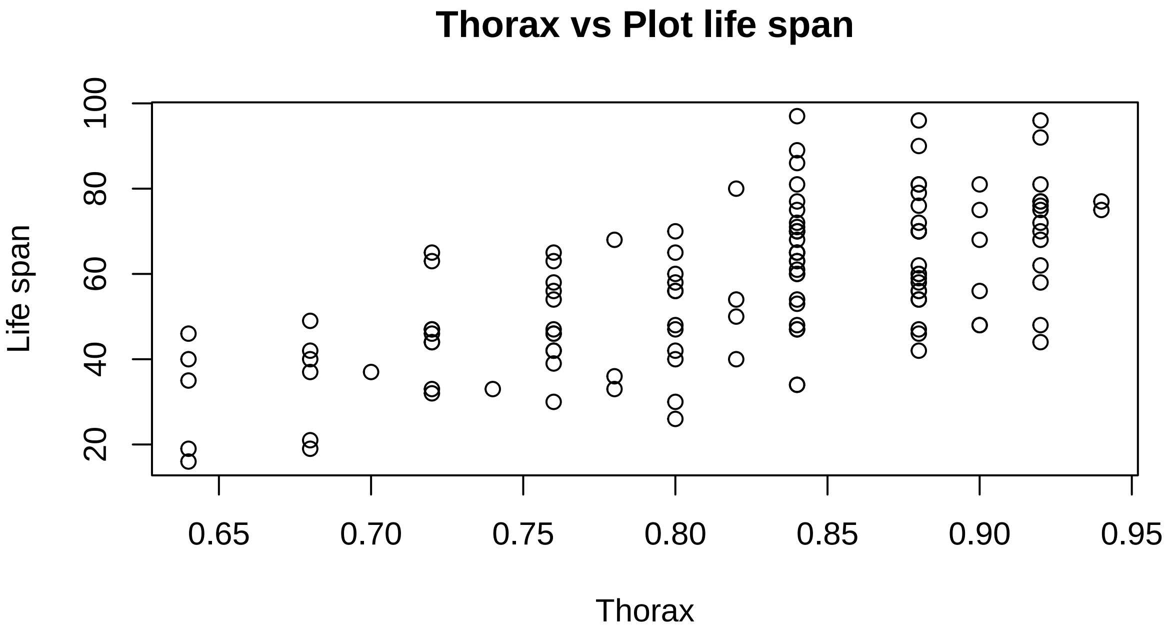
```
fruitfly <-read.csv ("http://mldata.org/repository/data/download/csv/uci-20070111-fruitfly/")
typeof(fruitfly)
summary(fruitfly)

fruitfly
```

   From the summary, I read that the min. longevity of a fruit fly is 16 and max. 97. (not sure about the units, so I report numbers only). Mean is 57.44.

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?

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



I rendered the graph using the following formula:

```
plot(fruitfly$Thorax, fruitfly$Longevity,
     main = ”Thorax vs Plot life span”,
     xlab = ”Thorax”,
     ylab = ”Life span”)
```

My correlation result is 0.63 which indicates medium strong level of association.

I used the following formula for the correlation:

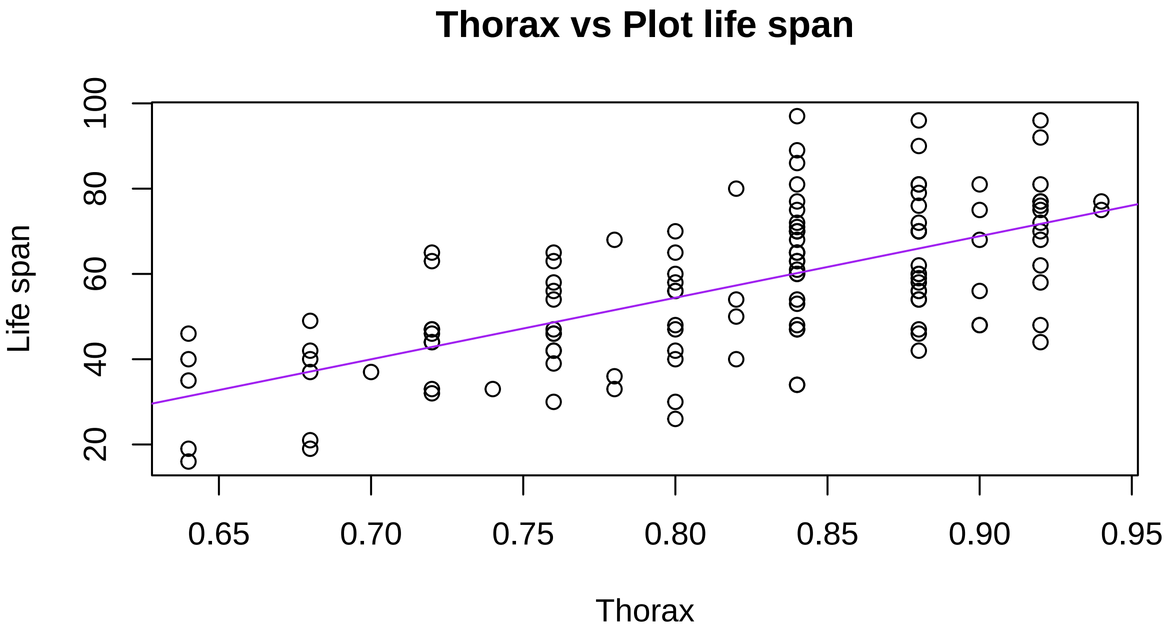
```
correlation.results <- cor(fruitfly$Longevity,fruitfly$Thorax)
correlation.results
```

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

I regressed the results and rendered the slope using the following formula:

```
reg.results <- lm(fruitfly$Longevity ~ fruitfly$Thorax)
summary(reg.results)

abline(lm(fruitfly$Longevity ~ fruitfly$Thorax), col = ”purple”)
```



The slope of the fitted model indicates the linear relationship in the data and that the size of thorax is positively associated with the lifespan of a fruit fly.

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

   I tested the significance of linear relationship using the chi square test.

   ```
   fruitfly_chisq <- chisq.test(table(fruitfly$Longevity, fruitfly$Thorax))
   fruitfly_chisq
   ```

   I got that the association between the size of thorax and the lifespan is statistically significant the p value of .05. And the association according to the chi square is strong (=642).

5. Provide the 90% confidence interval for the slope of the fitted model.

    - Use the formula of confidence interval.

    - Use the function `confint()` in `R` .

    My work gets really sloppy at this point and I am not sure why my results are not proper at this point. As per task, I calculated the CI by hand:

    ```
    # calculating by hand:

    tvalue <- (0.0028068-0)/0.0003067
    tvalue
    # same sa that from the table

    # confidence interval for beta 1

    #beta 1+t score*se
    #beta 1-tscore*se

    upperCI <- 0.0028068 + (9.151614*0.0003067)
    lowerCI <- 0.0028068 - (9.151614*0.0003067)

    upperCI
    lowerCI
    ```

    I got as my lower CI = -1.38e-11 and as my upper CI = 0.0056136. Which I am not sure how to interpret. Using formula confint I did not manage to get proper results and my calculations rendered NA as a result. I used the following formula:

    ```
    confint(reg.results, 'fruitfly$Longevity', 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?

    I tried doing this one and following one. You can see some of mu R code, but I started getting the error message which I did not know how to amend. The message says that the data in my formula is the wrong, i.e. it is an lm object which I tried turning into data frame or list, but I was not successful.

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.