Problem Set 2

Doireanna Craven

Due: October 15, 2021

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.

	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`).

The following code gave me a χ^2 test statistic of 3.7912. I then checked this in R using `chisq.test`.

[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.

```r
f0 <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, byrow=T)
                    #Set up the data in a matrix of two rows

                    #First, calculate expected values

rs<-rowSums(f0)     #This method gives the sum of each row as vector
cs<-colSums(f0)     #And sum of each column, needed for expected values
total=sum(f0)       #Also need total of all elements in the data

rsxcs<-matrix(c(rs[1]*cs[1], rs[1]*cs[2], rs[1]*cs[3],
                rs[2]*cs[1], rs[2]*cs[2], rs[2]*cs[3]),
              nrow=2, byrow=T)
                    #Set up a vector multiplying the row total by column
                    #total for each element in the data
fe=rsxcs/total      #Next step is to divide by the total of all elements

#I tried this in a for loop, but I couldn't make it to work:

   #for (i in 1:2){
    #    for (j in 1:3){
    #        fe<-matrix(c(rs[i]*cs[j]/total))
    #    }
    #}

#Now use expected values to calculate chi-squared stat:

step1 <- f0 - fe          #Difference in f0 and expected values
step2 <- (step1**2)/fe    #Squared then divided by expected values
css <- sum(step2)         #Finally summed to give chi-squared stat
css

chisq<-chisq.test(f0) #And finally use R to check if all is correct
chisq
```

(b) Now calculate the p-value from the test statistic you just created (in R). What do you conclude if $\alpha = .1$?

I used this line of code to calcuate a p-value of 0.15, which agrees with the `chisq.test` that I had already used.

```r
pv<-pchisq(css, df = 2, lower.tail=FALSE)
```

Since the p-value is greater than α, I conclude that there is not sufficient evidence to discount a hypothesis of independence at this level. That is to say, the evidence does not suggest that the reactions of the police officers depended on the driver's economic class. Using `qchisq`, I calculated that a χ^2 test statistic of 4.61 would be required to reject the null hypothesis with $\alpha = .1$.

(c) Calculate the standardized residuals for each cell and put them in the table below. The standardardised residuals are given by this formula:

$$\frac{f_o - f_e}{\sqrt{f_e(1 - \text{row prop.})(1 - \text{column prop.})}}$$

I calculated these in R but quite manually really. Here is an example of my code for the first standardised residual. It would have been quicker with a calculator, possibly.

```
#Standardised residual for f0[1,1]:

a <- 1-(rs[1]/total)      # 1 - row prop
b <- 1-(cs[1]/total)      # 1 - column prop
c<-(fe[1,1]*a*b)**0.5     # Denominator
stres1<- step1[1,1]/c     # f0 - fe from previous calculation
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

Standardised residuals with absolute value greater than 3 indicate considerable variation between observed and expected values within a cell. For this data, the standardised residuals are all below 2, therefore there is no evidence to support variation that cannot be explained by random chance, rather than by dependence. Within a column, the sign of the standardised residual tells us which way the data swung, so we can say that upper class drivers were marginally less likely to be stopped. Standardised residuals for outcomes once stopped were more significant though, approaching 2. This is not enough to declare that the variables are dependent, but perhaps enough to warrant a second experiment with larger n, since cells should have a frequency above 5 for a chi-squared test to be deemed reliable.

Question 2 (20 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[2] 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:

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 varables 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

[2]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.

H0: reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.

Ha: reservation policy has an effect on the number of new or repaired drinking water facilities in the villages.

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

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

The regression function returns an estimate for the slope of 9.25, and a p-value of 0.019.

With a p-value of 0.019, the probability of observing the estimated slope would be very small if the null hypothesis were true. I therefore reject the null hypothesis at the $\alpha = 0.05$ level, and accept that reservation policy does have an effect on the number of new or repaired drinking water facilities in the villages.

(c) Interpret the coefficient estimate for reservation policy.

The slope of 9.25 suggests that adopting a policy of reserving $\frac{1}{3}$ of village council heads for women typically leads to 9.25 additional new or repaired drinking water facilities in the villages.

Question 3 (40 points): Biology

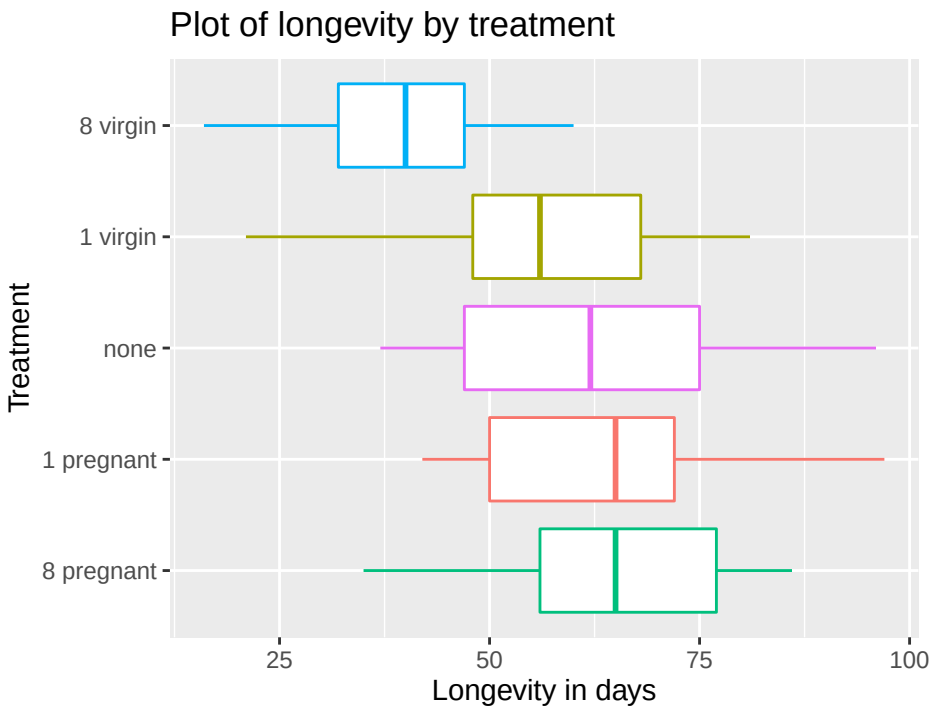
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity 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`.[3]

`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

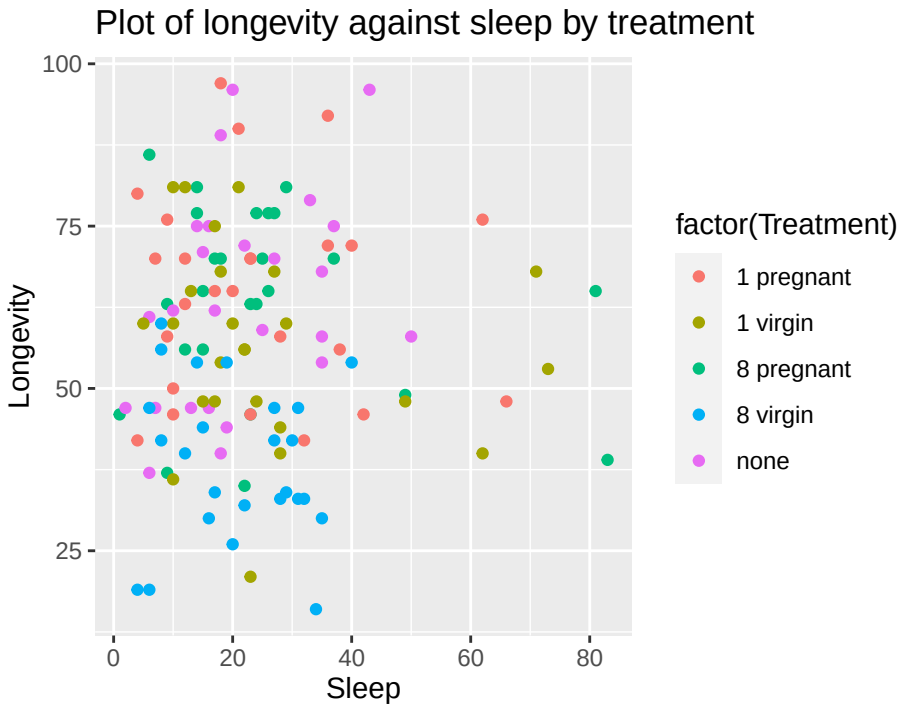
1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

   The boxplots on the next page compare the distribution of the lifespans of male fruitflies in the various treatment groups.
   Compared to the treatment where no females were introduced, we can see that the lifespan of males in treatment groups where virgin females were introduced is shorter, and where pregnant females were introduced the lifespan of males is marginally longer.

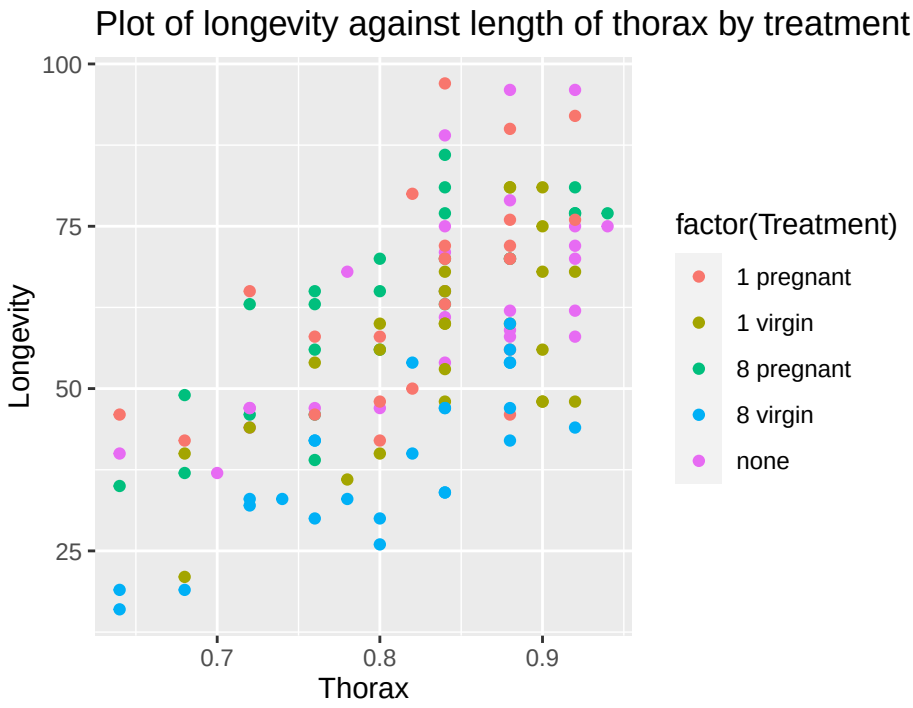
[3]Partridge and Farquhar (1981)."Sexual Activity and the Lifespan of Male Fruitflies". *Nature*. 294, 580-581.



Below, a scatter plot of males' lifespan against percentage of each day spent sleeping shows no correlation.



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?

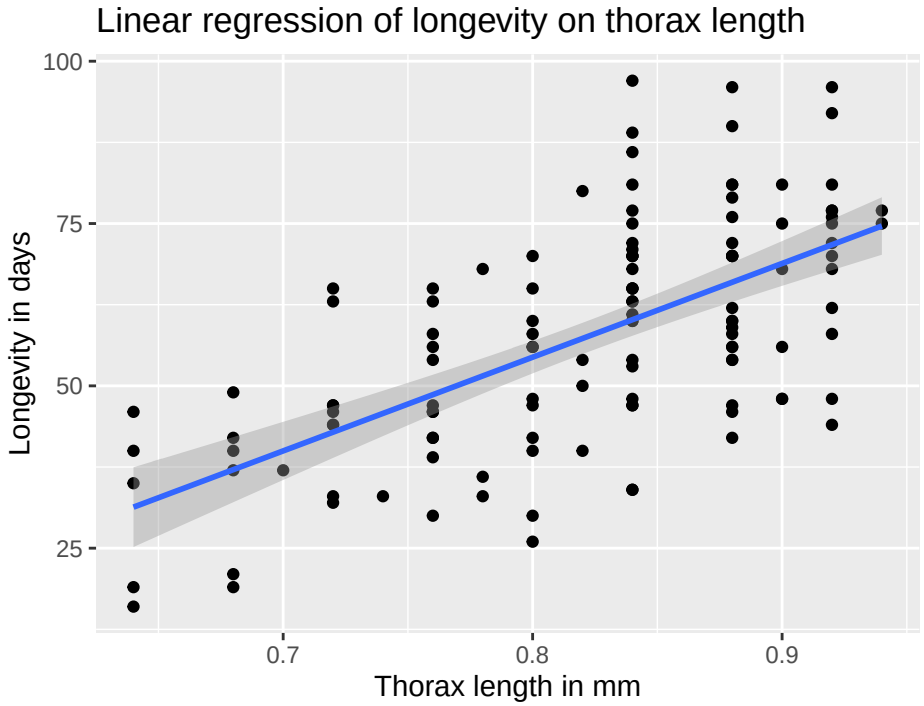


Lifespan and thorax appear moderately positively correlated.

The correlation coefficient between longevity and thorax length is equal to 0.64. The p-value for this test is significantly less than 0.001, with $\alpha = 0.5$. The code below calcuates the correlation coefficient and p-value by hand in R, which I have then checked using `cor.test`.

```
1
2 # r = correlation coefficient
3 # r = (covariance of x and y) divided by (sd of x * sd of y)
4
5 r <- cov(fruitflies$Longevity, fruitflies$Thorax)/
6     (sd(fruitflies$Longevity)*sd(fruitflies$Thorax))
7
8 n <- 125                                    # Number of observations
9
10 test_stat <- (r*sqrt(n-2))/sqrt(1-r^2)     # Formula for test statistic
11
12 p_value <- 2*pt(test_stat, n-2, lower.tail = FALSE)
13                                            # 2-sided test, t-distribution
14
15                                            # And now the quick way:
16 cor.test(fruitflies$Longevity, fruitflies$Thorax)
```

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




```
1
2 # Regression on Longevity and Thorax
3
4 longevity_thorax <- lm(Longevity ~ Thorax, data = fruitflies)
5 summary(longevity_thorax)
6
7 ggplot(aes(Thorax, Longevity), data = fruitflies) +
8     geom_point() +
9     geom_smooth(method = "lm", formula = y ~ x)+
10    labs(title ="Linear regression of longevity on thorax length",
11         x = "Thorax length in mm",
12         y = "Longevity in days")
```

The slope of the line is 144.33. Using an appropriate scale, this means that a 0.1mm increase in thorax length is associated with approximately 14 days longevity. The direction of association is not implied.

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.
   The null hypothesis is that there is no relationship between lifespan and thorax length, therefore the slope would be zero.
   H0: $\beta 1 = 0$
   Ha: $\beta 1$!= 0

```r
# Test statistic = (estimate of slope - zero)/standard error of slope

p_value_f <- 2*pt((144.33/15.77), df = 125-2 , lower.tail = FALSE)
# 2-sided test, t-distribution
```

   The p-value is in the order of 10exp(-15). The probability of observing the estimated slope would be virutally nil if the null hypothesis were true. Therefore we accept the alternative hypothesis that there is a relationship between thorax length and lifespan.

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` .

```r
# CI = estimated slope +/- tscore x standard error
# Calculate t-90 for df = 123

t90 <- qt(0.05, 123, lower.tail=FALSE)      # t90 = 1.66

lower_90 <- 144.33 - (t90*15.77)            # 1.66 se's below mean
upper_90 <- 144.33 + (t90*15.77)            # 1.66 se's above mean

# Check using confint:

confint(longevity_thorax, level = 0.9)
```

   The 90% confidence interval for the estimated slope is [118, 170]. A 0.1mm increase in thorax length is associated with an increase in lifespan of between 12 to 17 days.

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?

```
#The newdata argument in predict() requires data a dataframe
nd <- data.frame(Thorax = 0.8)

#Find the mean of lifespans that the model associates with Thorax = 0.8mm
predict(longevity_thorax, newdata= nd, se.fit=TRUE)

pred_int <- predict(longevity_thorax, newdata=nd, interval="prediction",
    level = 0.9)

conf_int <- predict(longevity_thorax, newdata=nd, interval="confidence",
    level = 0.9)
```

The model predicts that a thorax of 0.8mm will be associated with a lifespan of approximately 54 days. A pen and paper calculation using $\beta 0$ and $\beta 1$ agress with this. The prediction interval is [32, 77], and the 90% confidence interval is [52, 57].
The prediction interval indicates the range within which the lifespan of a fruitfly with thorax length = 0.8mm is likely to fall. The confidence interval gives the range in which the *average* lifespan of *many* fruitflies with thorax = 0.8mm is likely to fall.

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.

```
# Generate thorax lengths and corresponding longevity, and combine

Thorax <- runif(125, 0.65, 0.94)
Longevity <- -61.05 + Thorax*144.33
nd2 <- as.data.frame(cbind(Thorax, Longevity))

# Generate predictions from nd2

pred_int_2 <- predict(longevity_thorax, newdata=nd2,
                      interval="prediction", level = 0.9)

# Will need one data frame to plot from, including a column for new
# thorax values. It cannot have the same column title though:

Thorax2 = Thorax
mydata <- cbind(fruitflies, pred_int_2, Thorax2)

# Regression line + confidence intervals from original data:

```

```r
p <- ggplot(mydata, aes(Thorax, Longevity)) +
    geom_point() +
    stat_smooth(method = lm)

# Add prediction intervals based on new data:

p + geom_line(aes(x=Thorax2, y = lwr), color = "red", linetype = "dashed")+
    geom_line(aes(x =Thorax2, y = upr), color = "red", linetype = "dashed")+
    labs(title = "Confidence and prediction intervals",
         x = "Thorax length in mm", y = "Longevity in days")
```