Problem Set 2

Applied Stats/Quant Methods 1

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`).

[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.

1. χ^2 test statistic = 3.791168

```
#Assign Observed Frequencies
Fo_1 <- 14
Fo_2 <- 6
Fo_3 <- 7
Fo_4 <- 7
Fo_5 <- 7
Fo_6 <- 1

#Assign Expected Frequencies
Fe_1 <- ((27/42) * 21)
Fe_2 <- ((27/42) * 13)
Fe_3 <- ((27/42) * 8)
Fe_4 <- ((15/42) * 21)
Fe_5 <- ((15/42) * 13)
Fe_6 <- ((15/42) * 8)

#Calculate chisq
chisq <- (
  ((Fo_1 - Fe_1)^2/Fe_1) + ((Fo_2 - Fe_2)^2/Fe_2) +
    ((Fo_3 - Fe_3)^2/Fe_3) +  ((Fo_4 - Fe_4)^2/Fe_4) +
    ((Fo_5 - Fe_5)^2/Fe_5) + ((Fo_6 - Fe_6)^2/Fe_6)
)
chisq
```

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

2. Ho = Class does not affect outcome of police reaction
Ha = Class does affect outcome of police reaction
P-value = 0.15
0.15 >0.1 so we fail to reject the null hypothesis

3.

```
  pchisq <- pchisq(3.791168, df = 2, lower.tail = FALSE)
    pchisq

```

[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	0.322	-1.64	1.5
Lower class	-0.322	1.64	1.5

```
tab <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE)
colnames(tab) <- c('Not Stopped','Bribed','Warned')
rownames(tab) <- c('Upper Class', 'Lower Class')
tab <- as.table(tab)
tab

chisq_test <- chisq.test(tab)
chisq_test
chisq_test$stdres

```

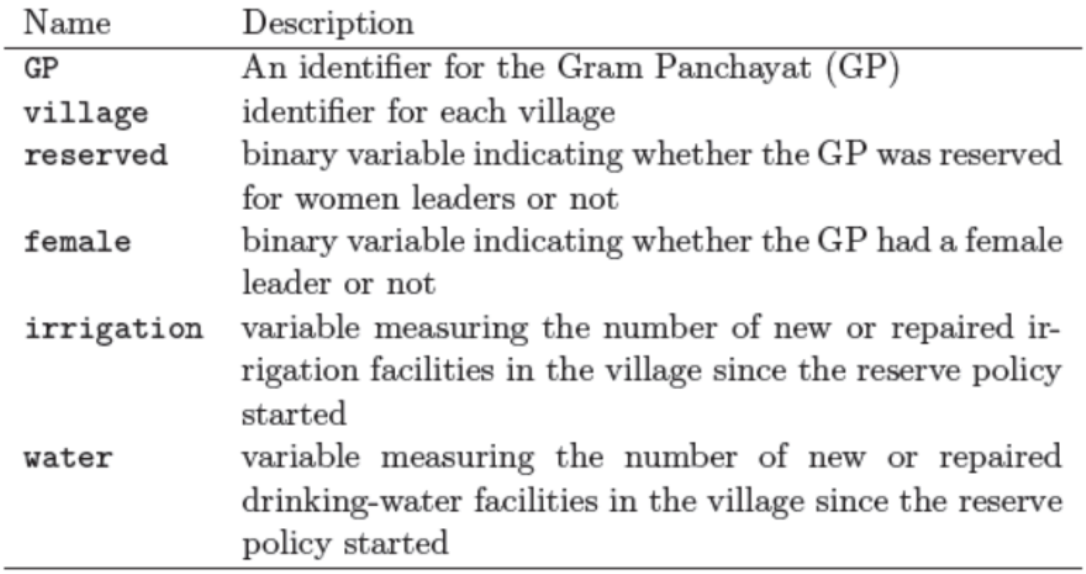
(d) How might the standardized residuals help you interpret the results?

4. The further away the standardized residuals from 0, the more the expected value differs from the observed 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.

1. Ho = Reservation policy has no impact on irrigation repair system
   Ha = Reservation policy has an impact on irrigation repair system

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

2. The p-value is 0.742 which is >than 0.05 so we fail to reject the null hypothesis.

```r
economics <- read.csv (”https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv”)
  View(economics)
  lm1 <- lm(economics$irrigation~economics$reserved)
  lm1
  summary(lm1)

```

(c) Interpret the coefficient estimate for reservation policy.

3. Co-efficient estimate is -0.3693, which is a negative weak correlation.
   This means that reservation policy has a small negative impact on irrigation repair system.

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

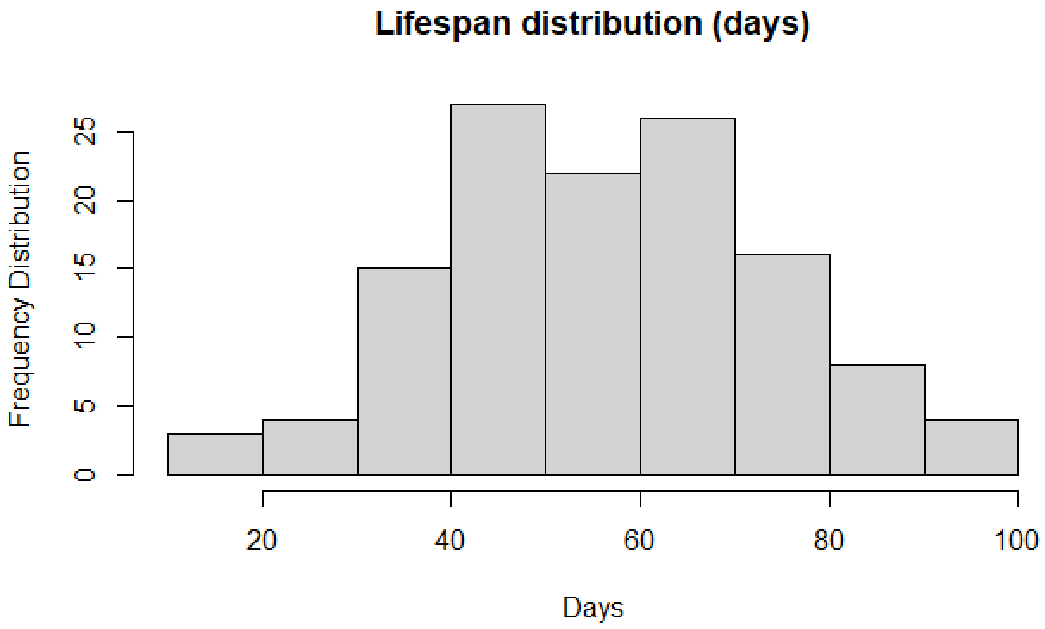
1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

```r
fruitfly<-read.csv("https://www.zoology.ubc.ca/~bio501/R/data/fruitflies.csv")
View(fruitfly)
summary(fruitfly)
hist(fruitfly$longevity.days,
     main = "Lifespan distribution (days)",
     xlab = "Days",
     ylab = "Frequency Distribution")

```

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

Figure 2: Lifespan distribution

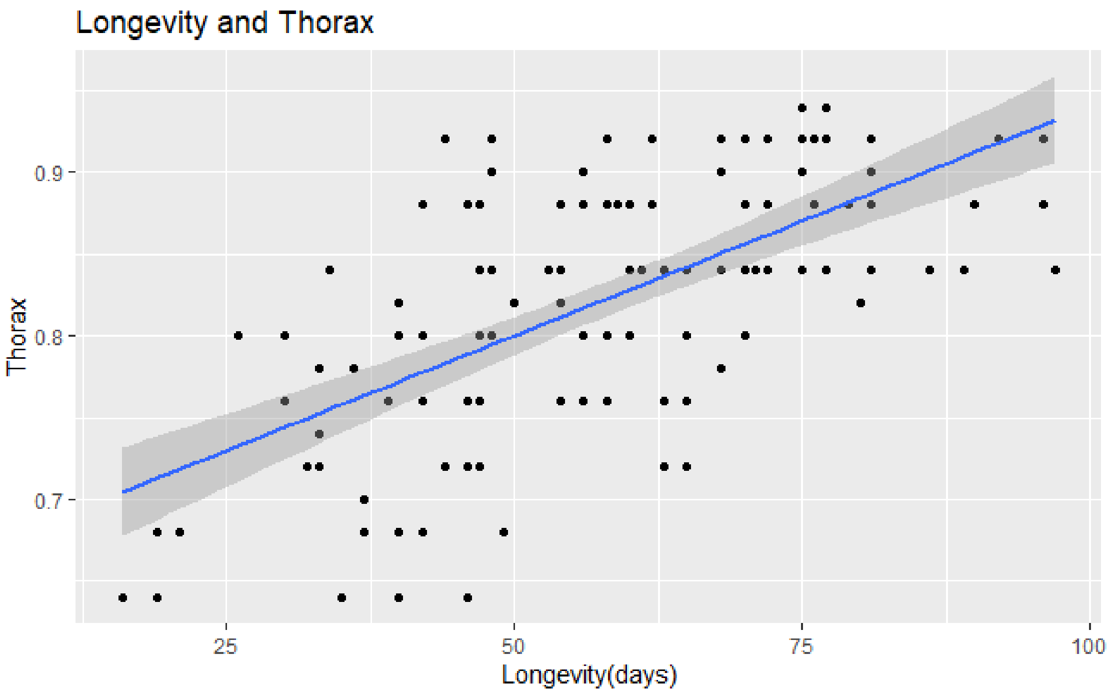


2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?
   The correlation coefficient is 0.636.
   There is a moderately strong linear relationship, the size of thorax impacts the fruit fly's lifespan. A bigger thorax results in a longer lifespan.

```r
ggplot(aes(longevity.days, thorax.mm), data = fruitfly) +
geom_point() +
geom_smooth(method = "lm", formula = y ~ x) +
ggtitle("Longevity and Thorax")+
labs(x = "Longevity(days)") +
labs(y = "Thorax")
cor.test(fruitfly$longevity.days, fruitfly$thorax)

```

Figure 3: Lifespan and Thorax



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

```
lm2 <- lm(fruitfly$longevity.days~fruitfly$thorax)
lm2
summary(lm2)

```

For every 0.1mm increase in thorax, longevity increases 14.433 days

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.
Ho = Thorax has no effect on longevity
Ha = Thorax has effect on longevity
P-value 1.497e-15 >than 0.05 so we reject the null hypothesis (obtained from summary statistics).
There is a significant relationship between thorax and longevity.

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

We are 90percent confident that the true slope lies between 118.19 and 170.47.

- Use the formula of confidence interval.

```r
slope <- 144.33
critical_value <- 1.645
residual_se <- 15.77
margin_of_error <- critical_value * residual_se
confidence_upper90 <- slope + margin_of_error
confidence_lower90 <- slope - margin_of_error
confidence_upper90
confidence_lower90
confint(lm2, level = 0.9, lower.tail = FALSE)
```

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

```r
  confint(lm2, level = 0.9)

```

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?

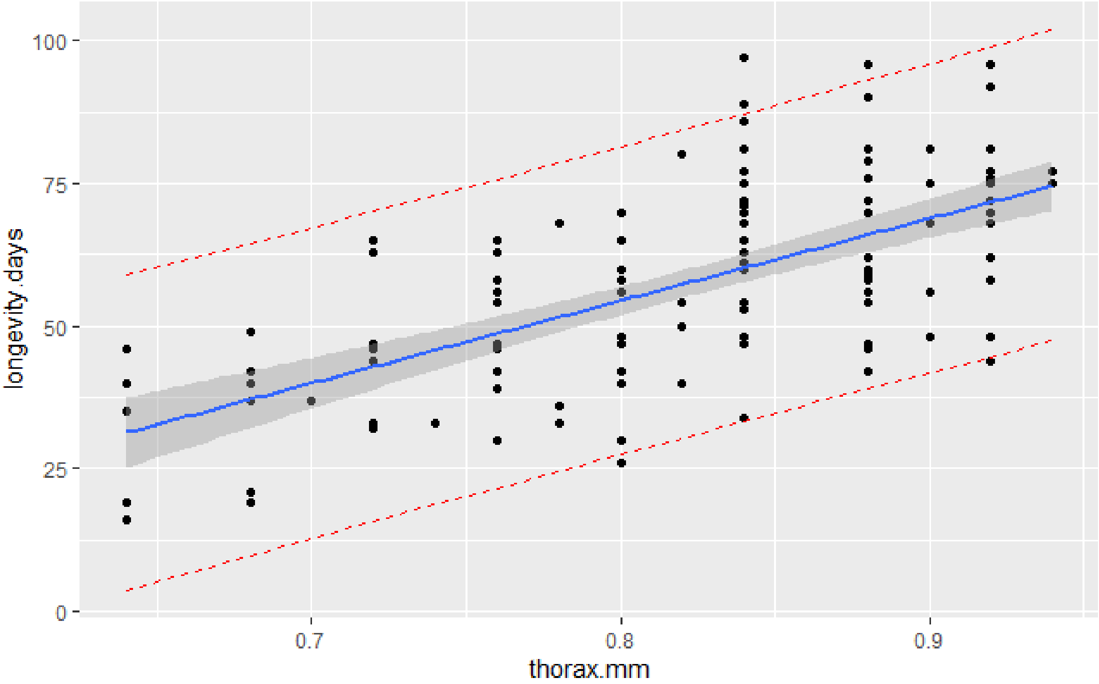
Predicted individual's lifespan = 54.4 days when thorax = 0.8
Prediction interval is between 27.3 and 81.4 days.
Confidence interval is between 51.9 and 56.9 days.

```r
fruitfly_lm <- lm(longevity.days~thorax.mm, data = fruitfly)
new_df <- data.frame(thorax.mm =c(0.8))
predict(fruitfly_lm, newdata = new_df)
predict(fruitfly_lm, newdata = new_df, interval=”prediction”)
predict(fruitfly_lm, newdata = new_df, interval =”confidence”)

```

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.

Figure 4: Prediction of lifespan and thorax



```
my_conf <- ggplot(data, aes(x=fruitfly$thorax.mm, y=fruitfly$longevity.days)) +
geom_point() +
geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE)

my_predict <- predict(fruitfly_lm, interval = "prediction")
my_df <- cbind(fruitfly, my_predict)

ggplot(my_df, aes(thorax.mm, longevity.days)) +
geom_point() +
geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
geom_line(aes(y=upr), color = "red", linetype = "dashed")+
geom_smooth(method=lm, se=TRUE)

```