Problem Set 2 - Solution

Dino Wildi

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

Task 1

Calculate the χ^2 test statistic by hand (even better if you can do "by hand" in `R`).

The χ^2 test statistic is defined as

$$\sum \frac{(f_o - f_e)^2}{f_e}$$

[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.

.

In R, I am writing a function in order to calculate this formula for each cell and then add them up to return a test statistic. For the expected frequency in each cell, I need the column and row totals, as well as the grand total, which gets calculated inside the function:

```
chi.stat <- function(freqtab){
  total <- sum(colSums(freqtab))
  chisq <- 0
  for(r in 1:nrow(freqtab)){
    for(c in 1:ncol(freqtab)){
      fobs <- freqtab[r,c]
      fexp <- (sum(freqtab[r,])/total) * sum(freqtab[,c])
      c <- (fobs-fexp)^2/fexp
      chisq <- chisq + c
    }
  }
  return(chisq)
}
```

Running this function returns a Chi-squared test statistic of **3.791**.

Task 2

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

For this question I use the R function to return a p value for a given χ^2 test statistic:

```
pchisq(chi.stat(data), df=2, lower.tail = F)
```

This function returns a p-value of **0.15**. This suggests that id $\alpha = .1$, the p-value is larger than our desired significance level. We can therefore not assume a connection between the two variables.

Task 3

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

I calculated residuals with the cstest() function in R and extracting the residuals variable of the result:

```
cstest <- chisq.test(data)
cstest$residuals
```

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.136	-0.815	0.819
Lower class	-0.183	1.094	-1.099

Task 4

How might the standardized residuals help you interpret the results?

The standardized residuals show that while none of the observations are really far away from the expected frequencies, there are still sizable effects in that people from upper classes are less likely to be asked for a bribe and more likely to be given a warning than we would otherwise expect; the reverse is true for lower class drivers. However, the occurrence of drivers of both classes not being stopped at all is very close to expected frequency. I interpret this as an indication that an effect might exist, but not be significant due to a very small sample size especially for the lower class drivers (only 15 total observations in this group).

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`

Task 1

State a null and alternative (two-tailed) hypothesis.

We expect that there will be a higher amount of repaired water facilities in the subset of council heads that are reserved for women, compared to the ones where this is not the case. If the data were unassociated, the number of repaired or newly constructed water facilities would be equal. This is therefore our null hypothesis; with the alternative hypothesis being inequality. Formally speaking:

$$H_0 : \mu_{reserved} = \mu_{notreserved}$$

$$H_A : \mu_{reserved} \neq \mu_{notreserved}$$

Task 2 and 3

Run a bivariate regression to test this hypothesis in `R` (include your code!). Interpret the coefficient estimate for reservation policy.

I am using the lm() function in R to run the regression, specifying *water* as the output variable and *reserved* as the input variable. I then use the summary() command to get out the coefficient and standard errors.

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

We can see the coefficient is positive with 9.252. This means that on average, in GP's reserved to be run by women, there were 9.252 more water facilities constructed or repaired then in those not reserved for women. The

[3] Chattopadhyay and Duflo. (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica*. 72 (5), 1409-1443.

p-value is 0.02, which means that this result is significant at the 95% level. We can therefore reject the null hypothesis and conclude that there likely is a positive correlation between reserving GP leadership for women and more water facilities being constructed or repaired.

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

Task 1

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

I import the dataset and use summary() to obtain summary statistics. The statistics already tell me that the lifespan of the fruitflies has a range of 16 to 97 days and a mean and median that are close together and almost in the middle of the range (57.44 and 58.00 respectively). The distance of the 1st and 3rd quartiles from the median is exactly equal (12). This suggests that the data is roughly normally distributed. I added a bar diagram to visualise the data as well, as can be seen in Figure 1:

```
fruitfly <- read.csv("https://raw.githubusercontent.com/ASDS-TCD
    /StatsI_Fall2021/main/datasets/fruitfly.csv")
summary(fruitfly)
```

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

```
3
4 lifespanplot <- ggplot(fruitfly) + geom_bar(aes(x = lifespan)) +
      labs(x = "Lifespan in days", y = "", title = "Lifespan of
    fruitflies")
5 lifespanplot
```

Task 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?*

I am producing a scatterplot between the two variables, overlaid with a regression line and a confidence interval using ggplot. I also add the code line for the correlation coefficient here:

```
1 thoraxplot <- ggplot(fruitfly) + geom_point(aes(x = thorax, y =
    lifespan)) +
2   geom_smooth(aes(x = thorax, y = lifespan), method = "lm") +
3   labs(x = "Thorax length in mm", y = "Lifespan in days", title
    = "Regression of lifespans on thorax lengths")
4 thoraxplot
5
6 r <- cor(fruitfly$thorax, fruitfly$lifespan)
```

The correlation coefficient returned is **0.636**, hinting at a strong positive correlation between the two. The plot shown in Figure 2 supports this idea; the regression line points clearly upwards with a relatively narrow confidence interval and the observated lifespans clearly increase.

Task 3

Regress ***lifespan*** *on* ***thorax****. Interpret the slope of the fitted model.*

```
1 thoraxreg <- lm(lifespan ~ thorax, data = fruitfly)
2 summary(thoraxreg)
```

The slope of the fitted model is returned as **144.33**. It should be remembered that this is the increase of the lifespan in days for an additional mm of thorax length; an increase that never occurs in real life as 2 shows: the range of thorax length lies between 0.64 and 0.94 mm. Still, this indicates a strong effect of thorax length on lifespan.

Task 4

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

In order to test a linear relationship, we first have to get a test statistic. The test statistic for a correlation is defined as $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$; with n being 125 in this case. This gives us the following R code, returning a test statistic and a p-value through the pt() function:

```
1
2 t_stat <- (r*sqrt(123))/(sqrt(1-r^2))
3 #p-value:
4 pt(t_stat, 123, lower.tail = F)
```

This yields a p-value of $p = 7.484 * 10^-16$, an extremely low number suggesting that the association is not only strong, but also extremely significant.

Task 5

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

```
1 margin.error <- qt(p = 0.1/2, df = 123, lower.tail = F)
2 upperbound <- thoraxreg$coefficients["thorax"] + (margin.error*
    15.77)
3 lowerbound <- thoraxreg$coefficients["thorax"] - (margin.error*
    15.77)
```

In order to obtain the confidence interval, we have to calculate a margin of error using the qt() function. Dividing the desired p-value of 0.1 in half as we want a two-sided interval, we get a margin of error of 1.657 standard deviations. Multiplying this with the standard error derived from the regression object and adding and substracting it from the estimated slope - 144.33 - yields a confidence interval of **[118.197, 170.469]**. Using the confint() function yields the same result:

```
1 confint(thoraxreg, level = 0.9)
```

Task 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?*

```
1 predframe <- data.frame(thorax = c(0.8))
2 predict(thoraxreg, newdata = predframe, interval = "prediction")
3 predict(thoraxreg, newdata = predframe, interval = "confidence")
```

The expected values of lifespan in both cases are **54.415**. However, the confidence intervals are significantly wider in the prediction for an individual

fruitfly, as they incorporate both the error in estimating the slope of the regression model and the error in picking an individual fruit fly. Therefore, the confidence interval for the average is **[51.91932, 56.91024]**, whereas it is **[27.37542, 81.45414]** for an individual fruit fly.

Task 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.

```
predframe_2 <- data.frame(thorax = seq(0.64, 0.94, 0.025))
confint_frame <- data.frame(predict(thoraxreg, newdata =
    predframe_2, interval = "confidence"))
predint_frame <- data.frame(predict(thoraxreg, newdata =
    predframe_2, interval = "prediction"))

predplot <- ggplot() + geom_line(aes(x = predframe_2$thorax, y =
     confint_frame$fit), data = NULL) +
  geom_line(aes(x = predframe_2$thorax, y = confint_frame$lwr),
    col = "blue", linetype = "dashed", data = NULL) +
  geom_line(aes(x = predframe_2$thorax, y = confint_frame$upr),
    col = "blue", linetype = "dashed", data = NULL) +
  geom_line(aes(x = predframe_2$thorax, y = predint_frame$lwr),
    col = "blue", linetype = "dotted", data = NULL) +
  geom_line(aes(x = predframe_2$thorax, y = predint_frame$upr),
    col = "blue", linetype = "dotted", data = NULL) +
  labs(x = "Thorax length in mm", y = "Predicted lifespan in
    days", title = "Predicted lifespan and confidence intervals
    for various thorax lengths")
predplot
```

The code above produces a prediction for the values of 0.64 to 0.94, which equals the range of thorax lengths in the dataset. It does so in steps of 0.025 mm, resulting in 13 distinct predictions. Figure 3 then shows the line of predictions in a solid black line, the confidence interval for averages in dashed blue lines, and the confidence interval for a single prediction in dotted blue lines.

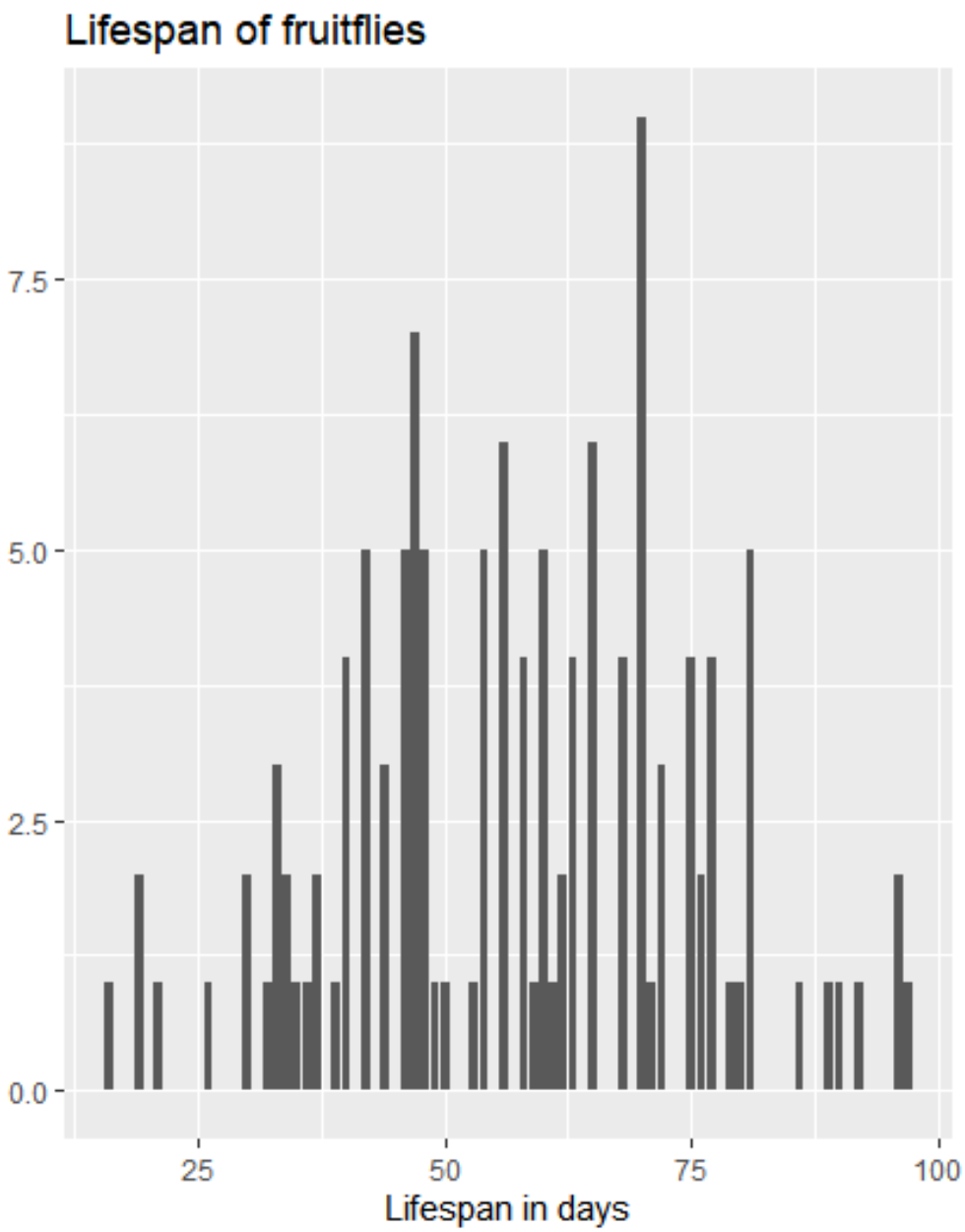

Figure 1: Lifespan of fruitflies

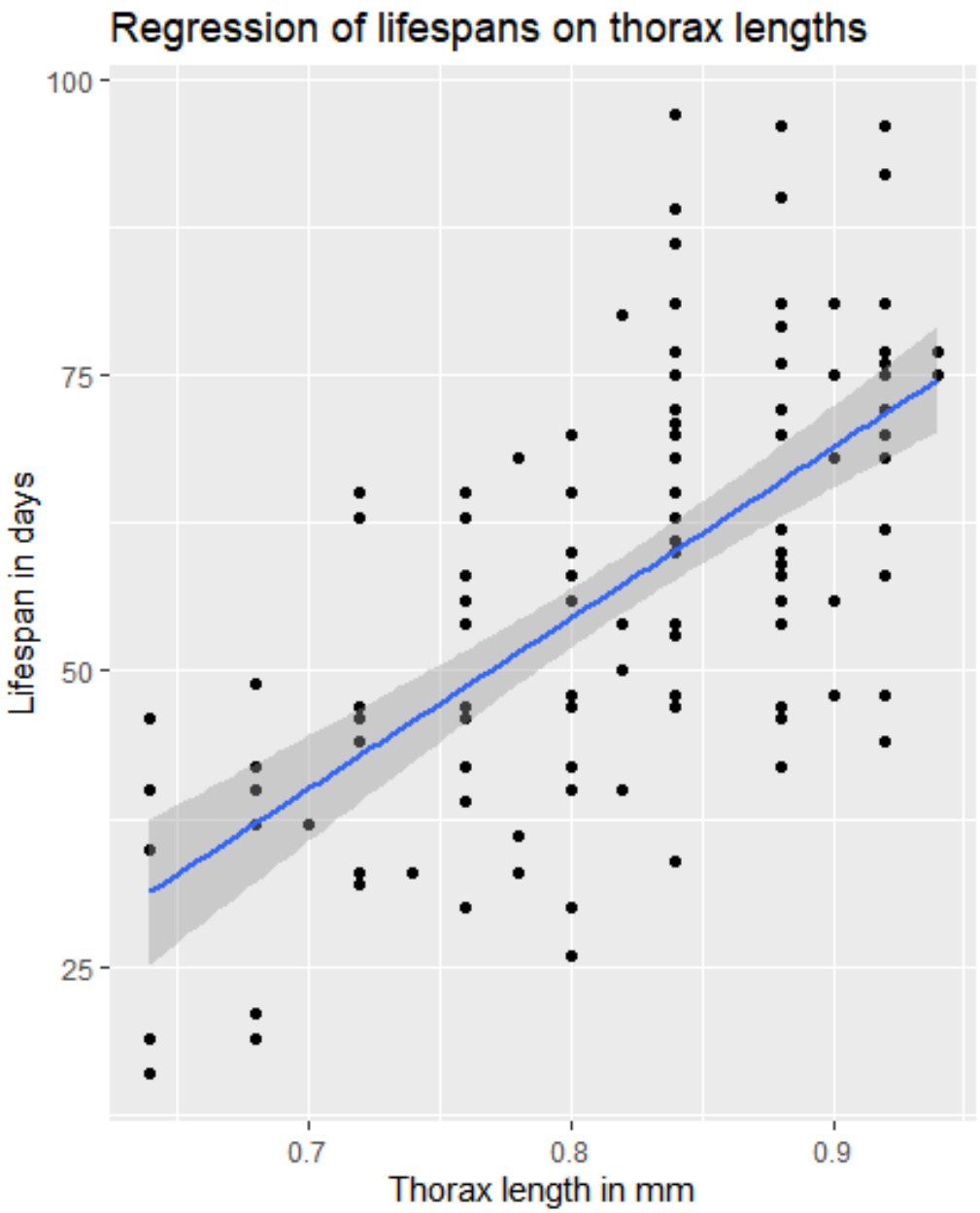

Figure 2: Lifespan of fruitflies and thorax length

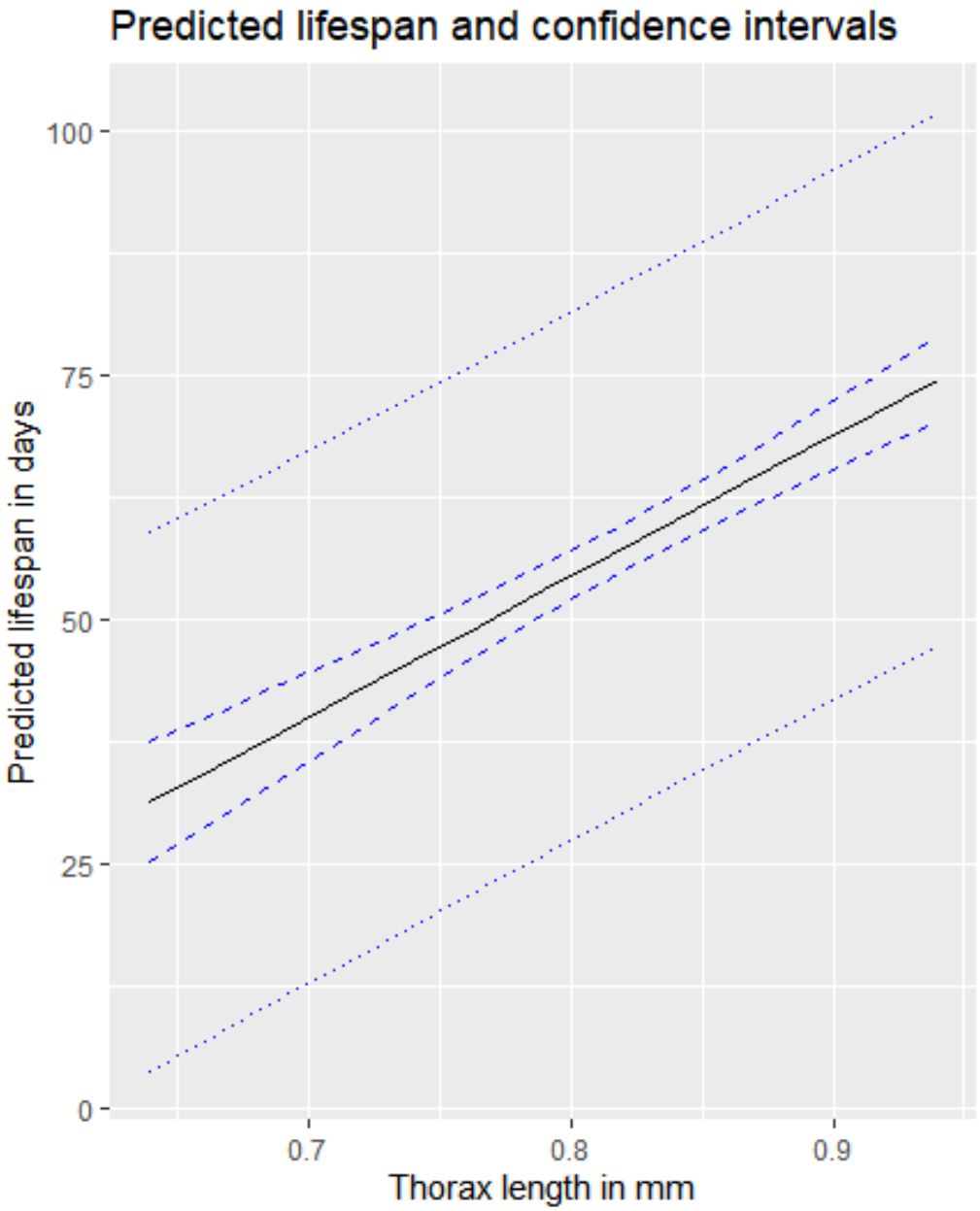

Figure 3: Predicted lifespan of fruitflies and thorax length