Problem Set 2

Paula Montano/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`). Complete question 1 was done by hand in an additional sheet attached at the end of this document.

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

[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			
Lower class			

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

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.

westBengal ¡- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/

```
str(westBengal)
head(westBengal)
summary(westBengal)
```

Ho: When GP was reserved for women leaders the number of new or repaired drinking-water facilities decreased in the village.

Ha: When GP was reserved for women leaders the number of new or repaired drinking-water facilities increased in the village.

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

(c) Interpret the coefficient estimate for reservation policy.

Question 3 (40 points): Biology

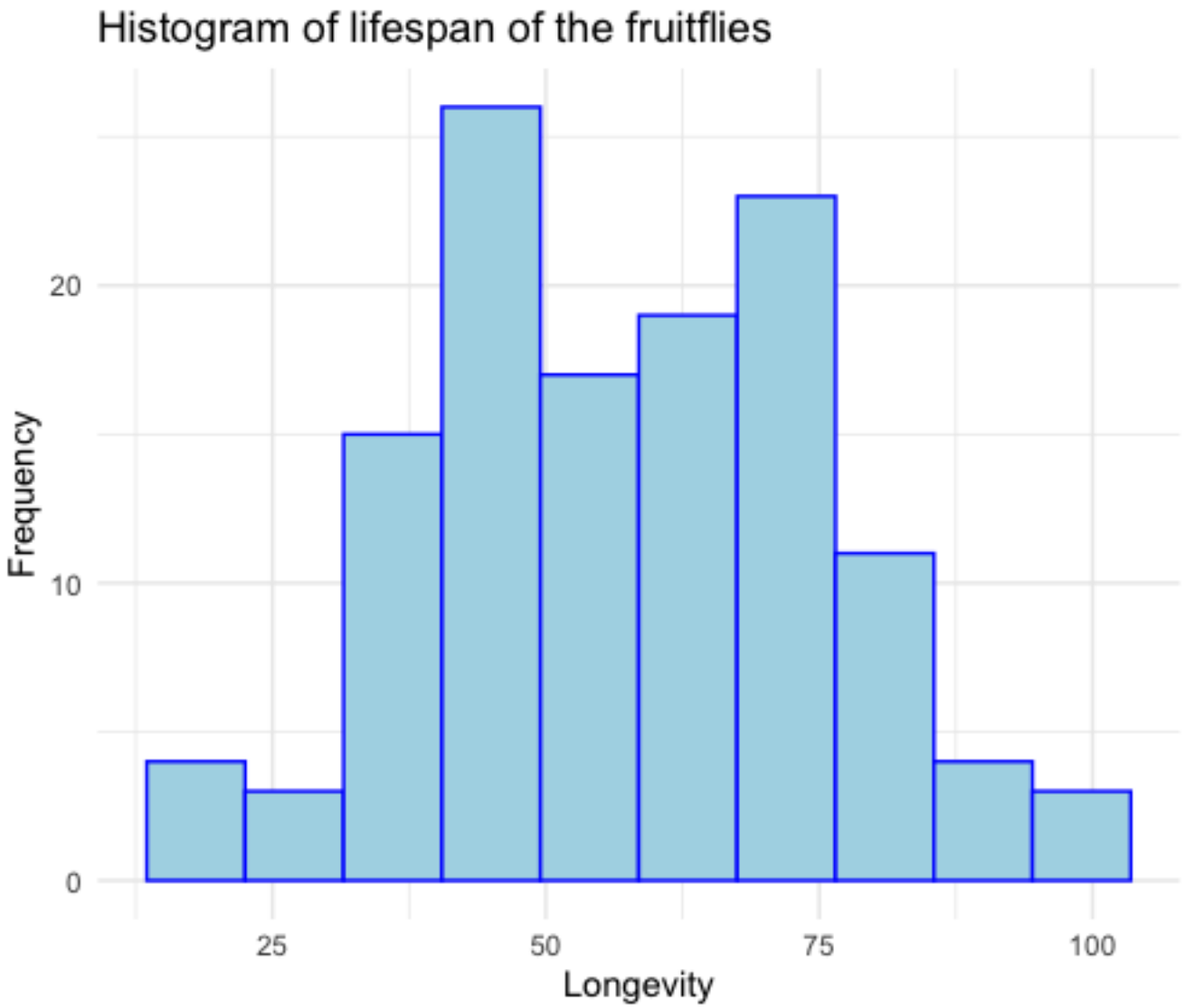
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity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

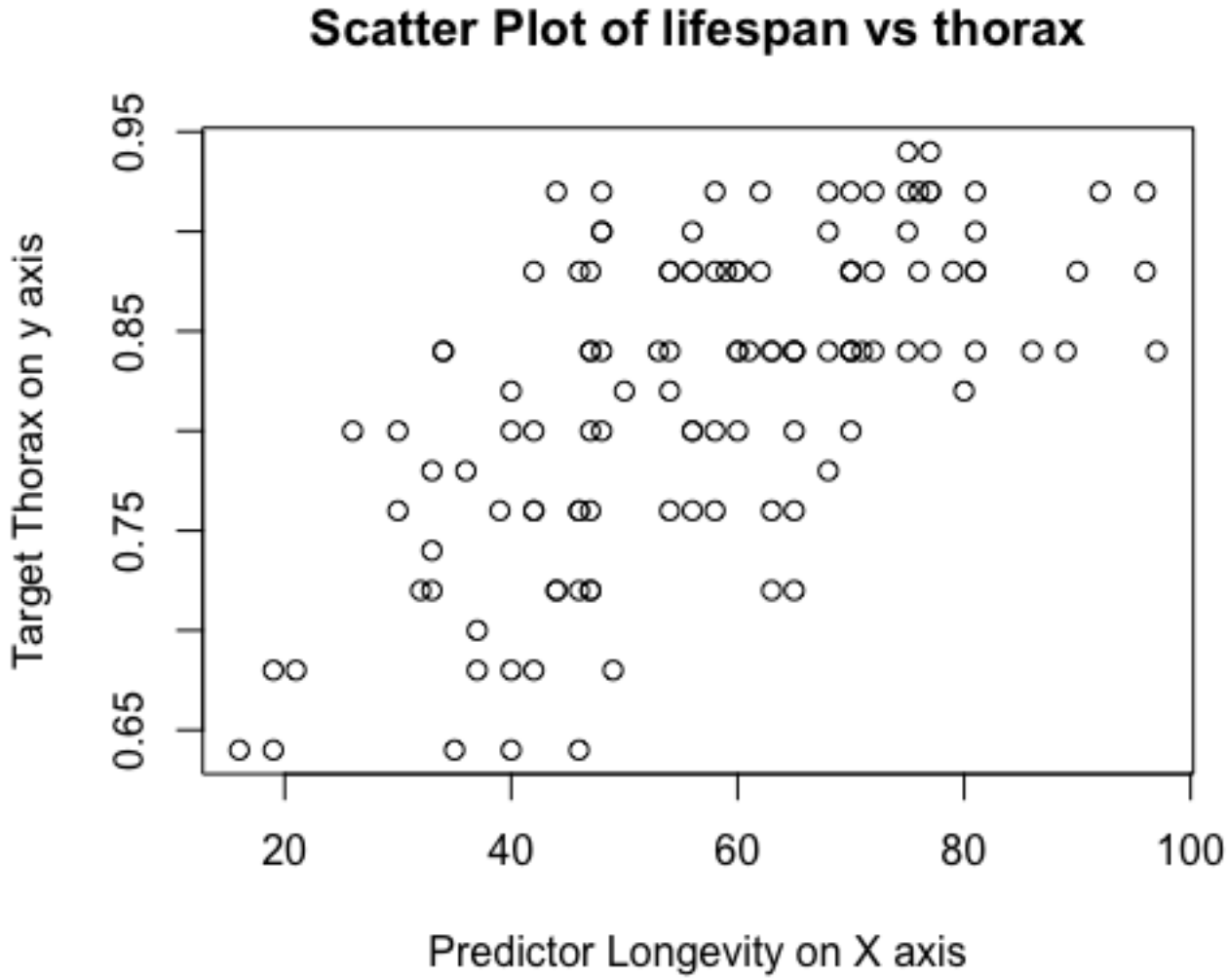
```
str(fruitfly)
head(fruitfly)
summary(fruitfly)

lifespan_histogram <- ggplot(fruitfly, aes(x = Longevity)) +
  geom_histogram(bins = 10, color = "blue", fill = "lightblue") +
  labs(x = "Longevity", y = "Frequency",
       title = "Histogram of lifespan of the fruitflies") +
  theme_minimal()
```

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


Histogram of lifespan of the fruitflies
Frequency
0
10
20
25
50
75
100
Longevity

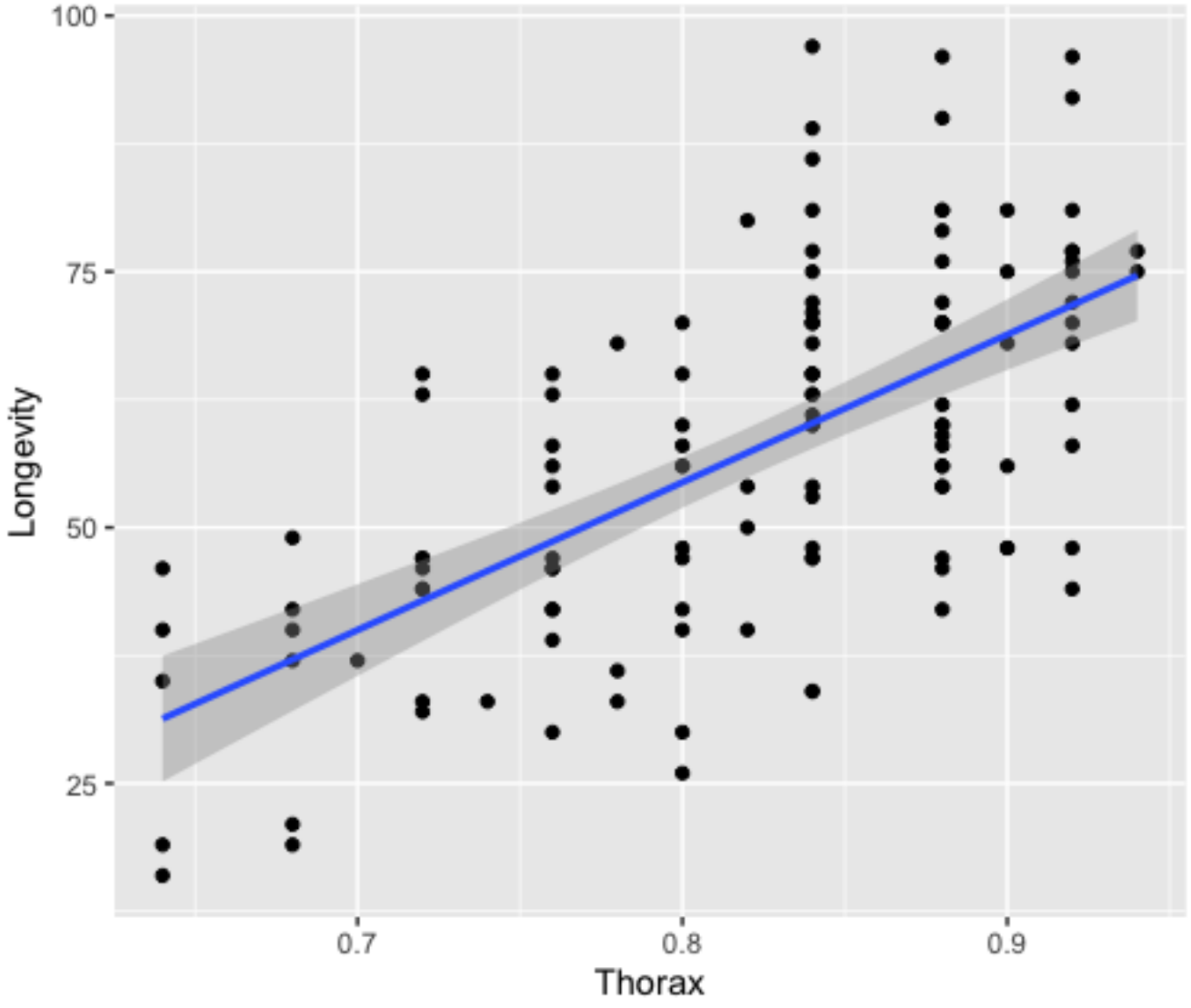
2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?

3.
```r
Plot_lifespan_vs_thorax <- plot(fruitfly$Longevity, fruitfly$Thorax,
                                main = "Scatter Plot of lifespan vs thorax",
                                xlab = "Predictor Longevity on X axis",
                                ylab = "Target Thorax on y axis")


lifespan_thorax <- lm(Longevity ~ Thorax, data = fruitfly)
summary(lifespan_thorax)
class(lifespan_thorax)

ggplot(aes(Thorax, Longevity), data = fruitfly) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x)
str(lifespan_thorax)

##The variables lifespan and thorax have a positive relation. The plot depicts a positive linear relationship when a value in X increases it also increses in Y. The correlation coefficient shows a strong association between liespan and thorax variable.
```



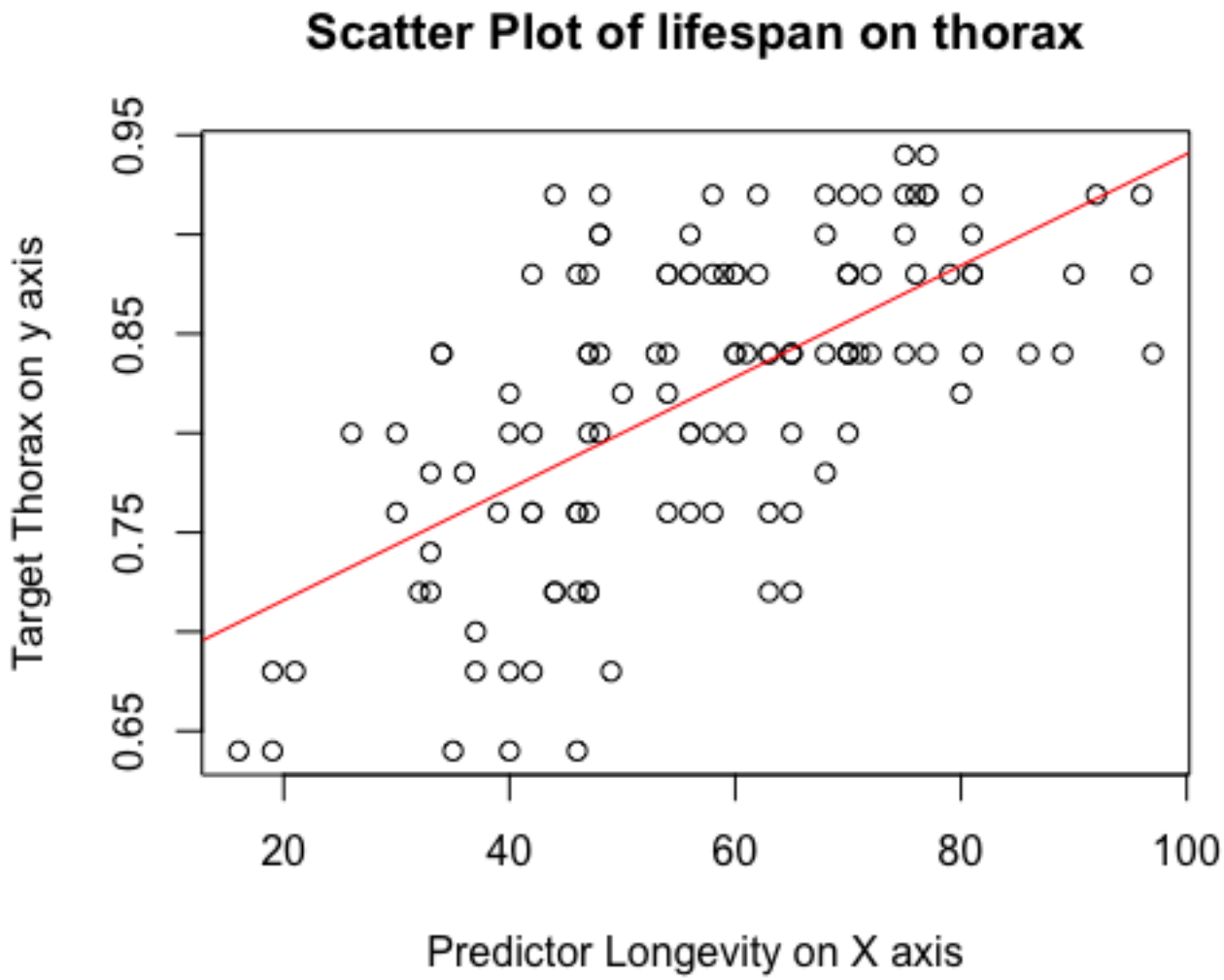
Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.

4.
```r
Plot_lifespan_thorax <- ggplot(aes(Longevity, Thorax), data = fruitfly) +
  geom_point(alpha = 0.4) +
  labs(x = "Lifespan", y = "Thorax",
       title = "Scatter Plot of lifespan vs thorax")

plot(fruitfly$Longevity, fruitfly$Thorax,
     main = "Scatter Plot of lifespan on thorax",
     xlab = "Predictor Longevity on X axis",
     ylab = "Target Thorax on y axis")
lm(fruitfly$Thorax ~ fruitfly$Longevity)
abline(lm(fruitfly$Thorax ~ fruitfly$Longevity), col = "red")

##The red slope shows the positive relation between lifespan and thorax
   in the fruitflies. The slope is steep and shows the strong association
    between the input and output variables.
```



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

```
lm(Longevity~Thorax, data = fruitfly)
```



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

```
z90 <-  qnorm((1-.90) / 2, lower.tail = FALSE)
n <- length(na.omit(fruitfly$lifespan_thorax))
fruitfly_mean <- mean(fruitfly$lifespan_thorax, na.rm = TRUE)
fruitfly_sd <- sd(fruitfly$lifespan_thorax, na.rm = TRUE)
lower_90 <- fruitfly_mean - (z90 * (fruitfly_sd  / sqrt(n)))
upper_90 <- fruitfly_mean + (z90 * (fruitfly_sd  / sqrt(n)))
confint90 <- c(lower_90, upper_90)

##Function confint()
confint(lifespan_thorax, parm = 0.90, level = 0.90)
```

- Use the formula of confidence interval.
- Use the function `confint()` in `R` .

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?

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.

(a) Calculate the X^2test statistic by hand

H0: The variables are statistically independent
Ha: The variables are statistically dependent

If H0 is true, then we would expect f observed = f expected

f observed = fo = observed frequency = the raw count

f expected = fe = what we would expect for independent samples

$f1e = \frac{row\ total}{grand\ total} * column\ total$

$\frac{27}{42} * 21 = 13.5$

$\frac{27}{42} * 13 = 8.36$

$\frac{27}{42} * 8 = 5.14$

$\frac{15}{42} * 21 = 7.5$

$\frac{15}{42} * 13 = 4.64$

$\frac{15}{42} * 8 = 2.85$

First, we calculate CHI-SQUARE TEST

	Not Stopped	Bribe requested	Stopped/ given warning	**Total**
Upper class	fo= 14 fe= 13.5	fo= 6 fe= 8.36	fo= 7 fe= 5.14	27
Lower class	fo= 7 fe= 7.5	fo= 7 fe= 4.64	fo= 1 fe= 2.85	15
Total	21	13	8	42

Then we calculate the X^2test statistic by hand

$$x^2 = \sum \frac{(fo - fe)^2}{fe}$$

$$= \frac{(14-13.5)^2}{13.5} + \frac{(6-8.36)^2}{8.36} + \frac{(7-5.14)^2}{5.14} + \frac{(7-7.5)^2}{7.5} + \frac{(7-4.64)^2}{4.64} + \frac{(1-2.85)^2}{2.85}$$

Answer:

$x^2 = 3.79$

(b) Calculate the p-value from the test statistic

df = (rows -1) (columns -1)

df = (3-1) (2-1)

p-value = pchisq(3.79, df = 2, lower.tail=FALSE)

p-value = 0.1503183

If p ≤ α we conclude that the evidence supports the alternative hypothesis Ha.
If p > α we cannot reject the null hypothesis H0.

What we conclude from α = .1
Our p-value is greater than α, therefore we cannot reject our null hypothesis. The variables are not statistically dependent.

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

	Not Stopped	Bribe requested	Stopped/ given warning	
Upper class	0.32	-1.64	1.24	
Lower class	-0.32	1.65	-1.52	

$$z11 = \frac{14 - 13.5}{\sqrt{13.5\left(1 - \frac{27}{42}\right)(1 - \frac{21}{42})}} = 0.32$$

$$z12 = \frac{6 - 8.36}{\sqrt{8.36\left(1 - \frac{27}{42}\right)(1 - \frac{13}{42})}} = -1.64$$

$$z13 = \frac{7 - 5.14}{\sqrt{5.14\left(1 - \frac{27}{42}\right)(1 - \frac{8}{42})}} = 1.24$$

$$z14 = \frac{7 - 7.5}{\sqrt{7.5\left(1 - \frac{15}{42}\right)(1 - \frac{21}{42})}} = -0.32$$

$$z15 = \frac{7 - 4.64}{\sqrt{4.64\left(1 - \frac{15}{42}\right)\left(1 - \frac{13}{42}\right)}} = 1.65$$

$$z16 = \frac{1 - 2.85}{\sqrt{2.85\left(1 - \frac{15}{42}\right)\left(1 - \frac{8}{42}\right)}} = -1.52$$

(d) The standardized residuals help us to identify how far away is each observed value from the predicted value (fo from fe). The standardized residuals are small and denote that the expected values are not so distant from the observed values.