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). $X^2 statiistic = (Fo–Fe)^2/Fe$

Firstly we must obtain Fo and Fe

Fo is the observed values, and can therefore be obtained from the data. Fe is the expected value and is calculated by (row total /sample total) *column total as seen below.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	Fo-14, Fe-13.5	Fo-6, Fe-8.3571	Fo-7, Fe-5.142
Lower class	Fo-7, Fe-7.5	Fo-7, Fe-4.642	Fo-1, Fe-2.857

In order to obtain the we complete the following calculation for $X^2(14–13.5)^2/13.5 + (6-8.3571)^2/8.3571 + (7–5.142)^2/5.142 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots n = 3.792844673$

```
chisq <- chisq.test(Q1data)
chisq
```

Pearson's Chi-squared test

data: Q1data $X - squared = 3.7912, df = 2, p - value = 0.1502$

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

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

$p = 0.1502306$

What we conclude if a is set at 0.1

We fail to reject the null hypothesis that "officers were more or less likely to solicit a bribe from drivers depending on their class" as p is greater than a.

It is also possible to conclude from this p value that the variables of class are independent.

[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	14, 0.32203	6, -1.64193	7, 1.5238558
Lower class	7, -0.32203	7, 1.642705238	1, -1.52294

Calculations by hand were as follows:

$= Fo–Fe/[Fe(1–rowprop) * (1–columnprop)$

$= 14–13.5/[(13.5–(21/42)) * (1 - (27/42)] = 0.32203$

$= 6–8.3571/[(8.3571–(13/42)) * (1 - (27/42)] = -1.64193$

continued for remaining 4 vlaues.

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

For us to know if a residual is large enough to indicate a departure from independence and occurs outside the scope of mere chance, we use the standard residual. This is shown as a z score.

We use this score to describe the pattern and association amongst cells.

Values of z less than -3, or z vlaues greater than 3 generally indicate that a cell has more observations than we would expect if the variables were truly independent.

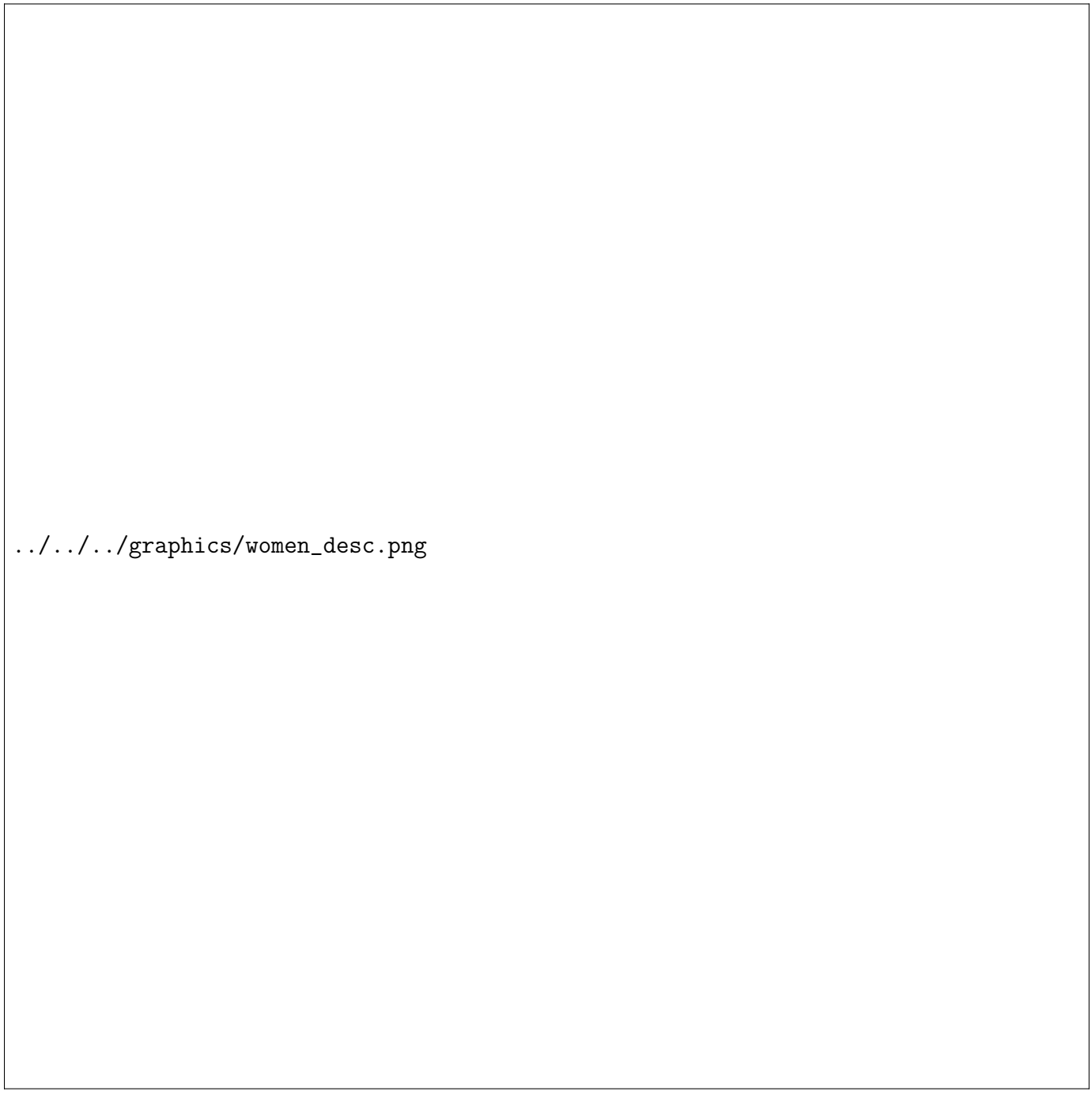
Question 2 (20 points): Economics

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.

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

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).



(a) State a null and alternative (two-tailed) hypothesis.
$H1 : u = /u0$

$H2 : u = u0$

Rejection Region for Two-Tailed Z Test (H1: u =/ 0) with a =0.05

The decision rule is: Reject H1 if $Z < -1.960 or if Z > 1.960$.

As our data is less than 30, we use the Z test

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

```
regmat <- matrix(indiadata)
r<- cov(regmat)[3,6]/sd(regmat[,3]) * sd(regmat[,6])

n <- dim(regmat)[3]
t_stat <- (r*sqrt(n-2))/sqrt(1-r^2)

2*pt(t_stat, n-2, lower.tail = FALSE)
cor(regmat[,3], regmat [,6])
cor.test(regmat[,3], regmat[,6])
```

320 degrees of freedom T-statistic: 5.493 p-value: 0.0197

We reject the null hypothesis as the p value is below our level of 0.05

(c) Interpret the coefficient estimate for reservation policy.

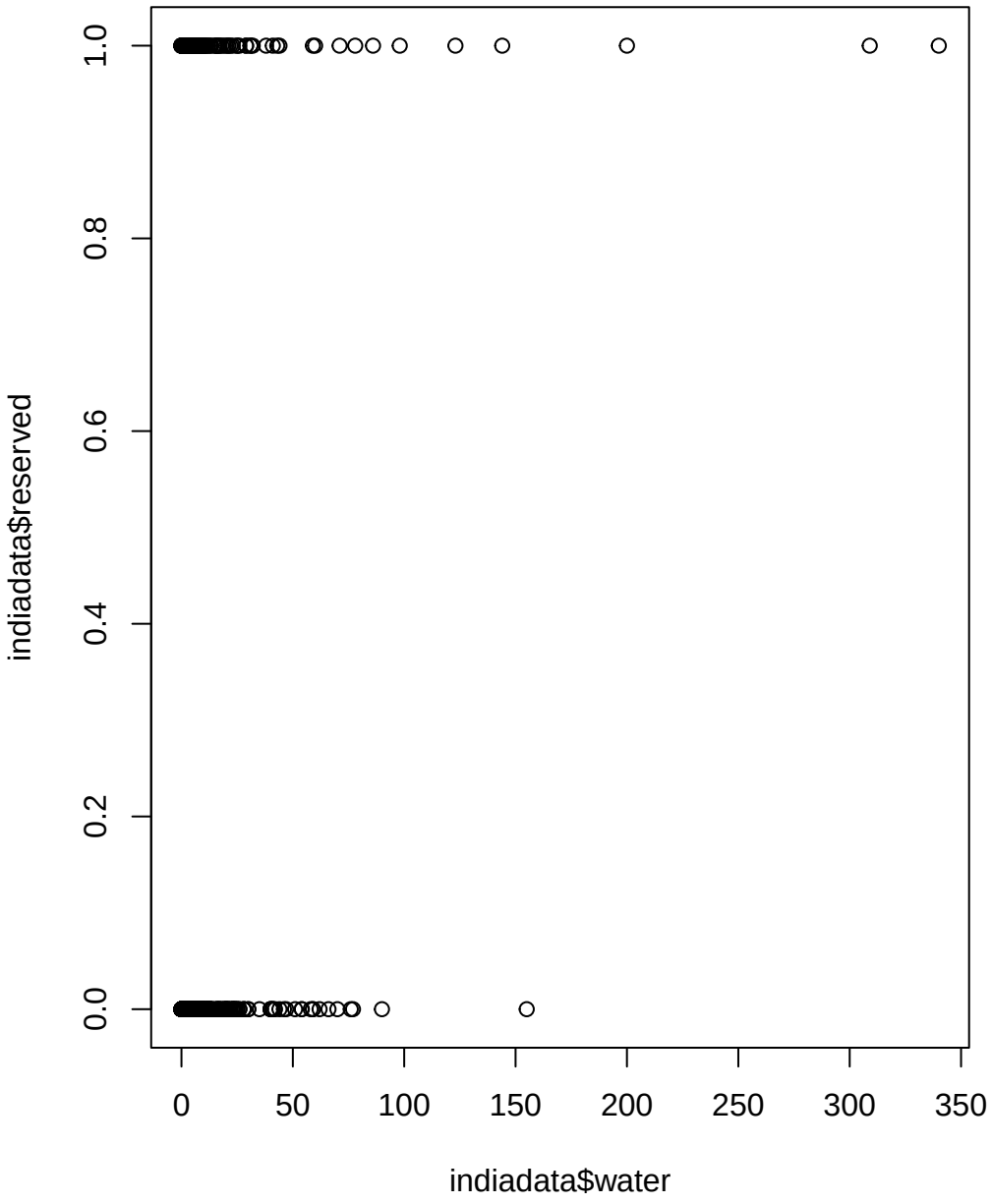


Figure 2: Relationship on water in india

```
lm(formula = water ~ reserved, data = indiadata)

summary(reg)
```

Coefficients: (Intercept) 14.738 reserved 9.252

From our coefficent estimates one can say that the there is a posotive relationship between reservation policy and water output in villages that have female heads.

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.

```
install.packages("faraway")
library(faraway)

flydata <- read.csv("http://stat2.org/datasets/FruitFlies.csv")

summary(flydata)
str(flydata)
```

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

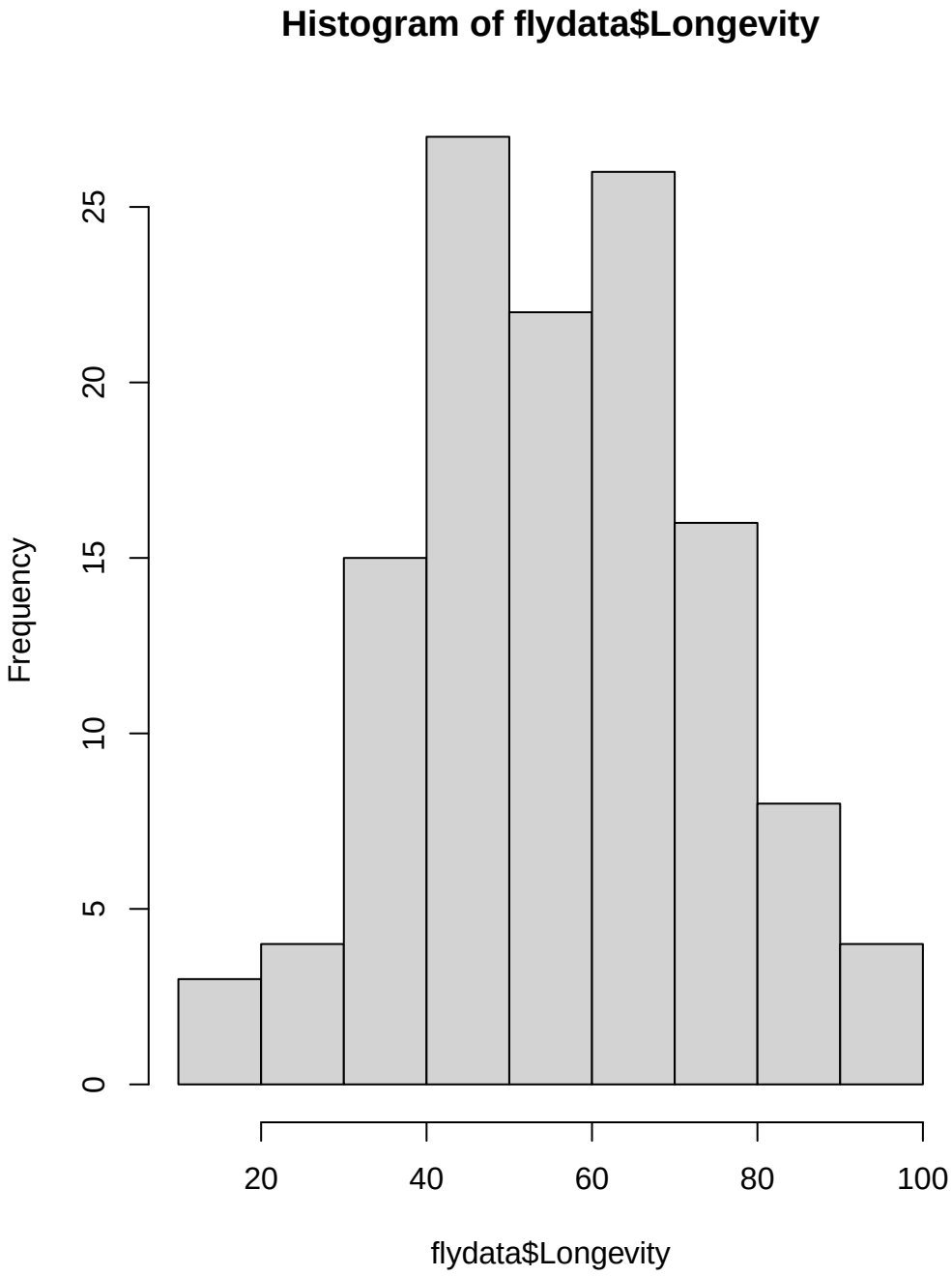


Figure 3: Fly lifespan distribution

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?

```
install.packages("ggpubr")
library("ggpubr")
ggscatter(flydata, x = "Longevity", y = "Thorax",
          add = "reg.line", conf.int = TRUE,
          cor.coef = TRUE, cor.method = "pearson",
          xlab = "days", ylab = "mm")
```

It would appear there is a linear relationship and that it is posotive as the correlation coefficent is 0.64.

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

```
ggplot(flydata, aes(x = Longevity, y = Thorax)) +
     geom_point() +
     stat_smooth()

model <- lm(Longevity ~ Thorax, data = flydata)
model

ggplot(flydata, aes(x = Longevity, y = Thorax)) +
  geom_point() +
```

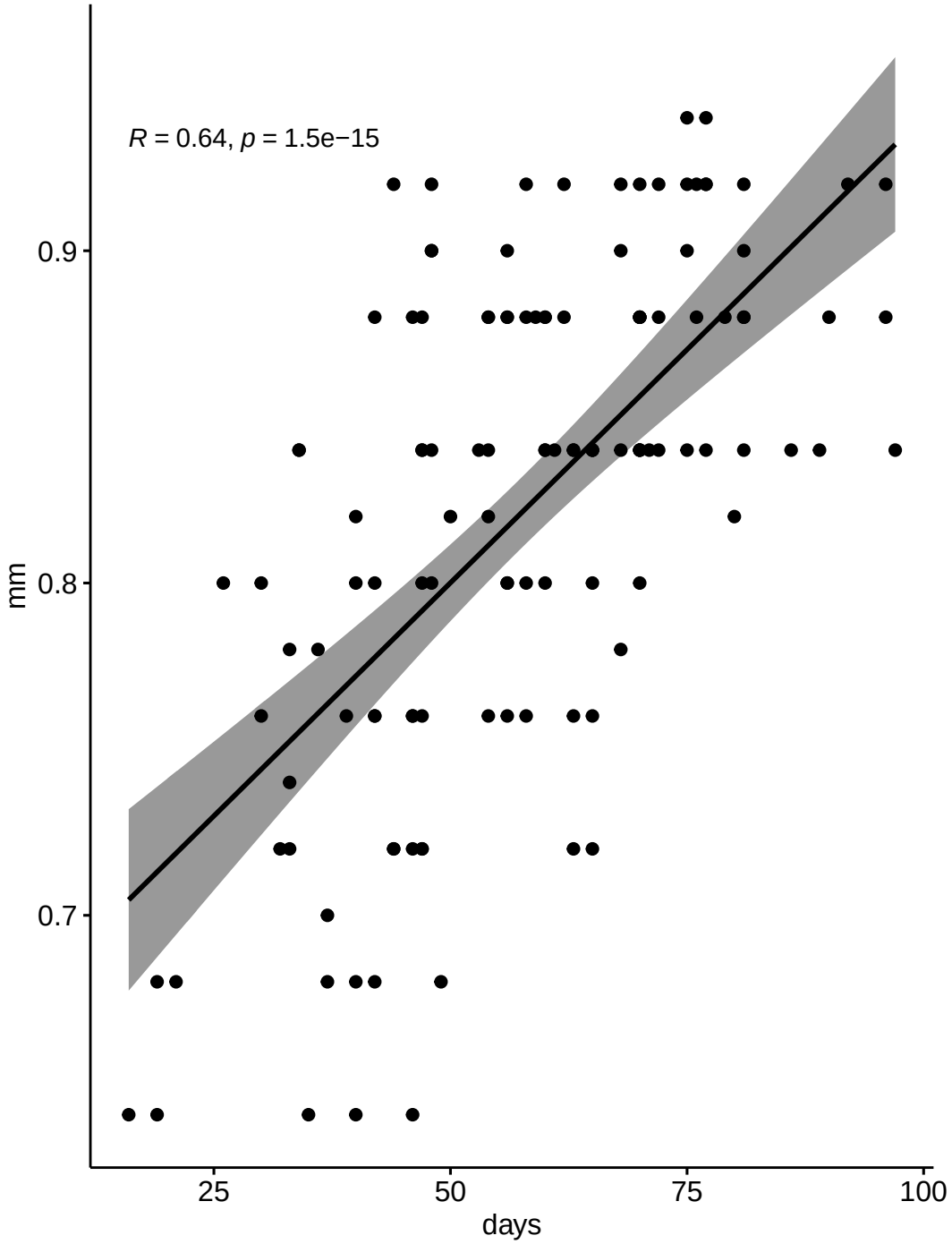


Figure 4: Fly lifespan distribution

```
  stat_smooth(method = lm)

summary(model)


confint(model)
```

The slope is 144.3

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

```
summary(model)
```

Call: lm(formula = Longevity ˜ Thorax, data = flydata)

Residuals: Min 1Q Median 3Q Max -28.415 -9.961 1.132 9.265 36.812

Coefficients: Estimate Std. Error t value Pr(¿—t—) (Intercept) -61.05 13.00 -4.695 7.0e-06 *** Thorax 144.33 15.77 9.152 1.5e-15 *** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.6 on 123 degrees of freedom

Multiple R-squared: 0.4051, Adjusted R-squared: 0.4003

F-statistic: 83.76 on 1 and 123 DF, p-value: 1.497e-15

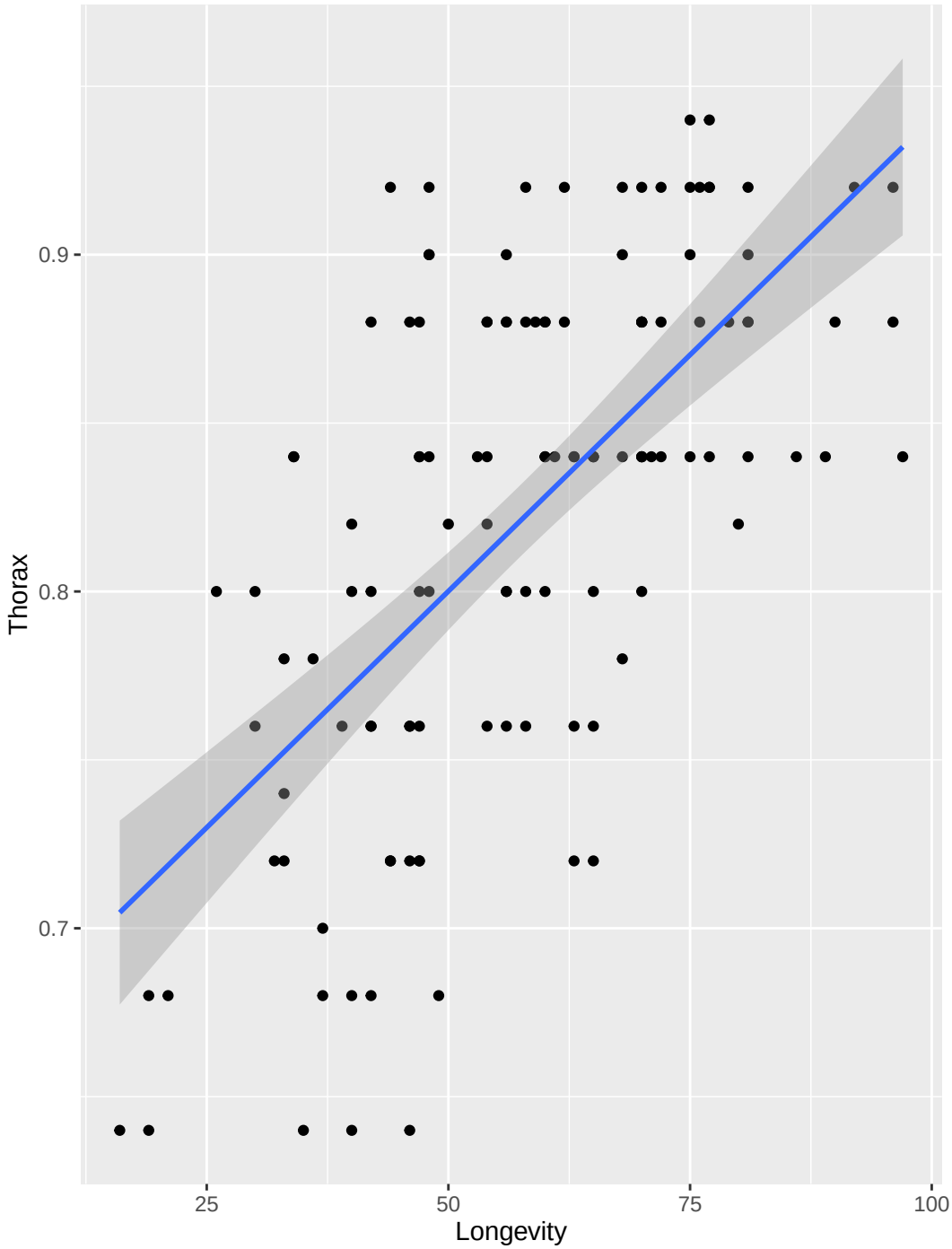


Figure 5: Fly lifespan distribution

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` .

   Hypotheses: $H0 : B0 = 0$

   $Ha : B0a = /0$

   Test statistic:t=B00/seB or -61.05/13 = -4.696

   P = 7.06

   B0:Bˆ0±t×se or-61.05(+/-)-4.696*13 = (-122.098, -1/500)

```r
confint(model)
```

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?

```r
class(model)
new_df <- data.frame(Thorax = c(runif(0.8)))
predict(model, newdata = new_df)
?predict()

newDF1 <- model ;newDF1$Thorex <- 0.8
predict(lm(newDF1$Thorex ~ newDF1$Longevity), newdata=newDF1, se.fit = T)
```

   The code would not provide me an output

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.

```
ggplot(aes(newDF1$Thorax ~ Longevity), data = flydata) +
      geom_point() +
      geom_smooth(method = lm)
```

Figure 6: fitted values