Problem Set 2

QTM 200: Applied Regression Analysis

Due: February 10, 2020

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 the course GitHub page in `.pdf` form.
- This problem set is due at the beginning of class on Monday, February 10, 2020. 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).

I first calculated the expected values of each cell using the formula Expected value = $\frac{\text{Row Total}}{\text{Grand Total}}$ x Column total in R.

```r
bribe_study <- matrix(c(14, 6, 7, 7, 7, 1), byrow = T, nrow = 2) # Recreating the table of data
bribe_study <- cbind(bribe_study, rowSums(bribe_study)) # Adding the total for rows as another column
bribe_study <- rbind(bribe_study, colSums(bribe_study)) # Adding the total for cols as another row
rownames(bribe_study) <- c(”Upper class”, ”Lower class”, ”Total”) # Row names
colnames(bribe_study) <- c(”Not stopped”, ”Bribe requested”, ”Stopped/given warning”, ”Total”) # Col names
bribe_study # Checking table

expected <- matrix(data = NA, nrow= 2, ncol = 3) # Matrix for expected values

# Calculating expected values and assigning expected value to each cell
for (i in 1:2)
  for (j in 1:3)
    expected[i, j] <- (bribe_study[i, 4] * bribe_study[3, j]) / bribe_study[3, 4]
expected # Checking expected values
```

Then, I calculated the χ^2 test statistic using the formula $\chi^2 = \sum \frac{(\text{Observed - Expected})^2}{\text{Expected}}$.

```r
observed <- bribe_study[-3, -4] # To make arrays comfortable
chisq <- (observed - expected)^2 / expected # Calculating each cell's contribution to TS
sum(chisq) # Chi-sq TS is 3.79

Xsq <- chisq.test(observed) # Checking answer
Xsq
```

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

I calculated the p-value in R using the pchisq() function, with input df = (rows - 1)(columns - 1). The upper tail was selected since this was a chi-square test for independence.

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

```r
rows = nrow(observed) # Number of rows
cols = ncol(observed) # Number of columns
pchisq(3.79, df = (rows-1)*(cols-1), lower.tail = F) # p-value = 0.1503
```

Since the p-value of 0.1503 is greater than the alpha value of 0.1, we do not have sufficient evidence to reject the null hypothesis that there is no association between class status and result in the study/action by police.

(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	1.52
Lower class	-0.32	1.64	-1.52

I calculated the standardized residual for each cell using this formula: adjusted residual = $\frac{\text{(Observed - Expected)}}{\sqrt{\text{Expected}(1-\text{row prop})(1-\text{col prop})}}$

```r
adjusted_resid <- matrix(data = NA, nrow= 2, ncol = 3) #Matrix for
    adjusted residuals

# Calculating adjusted residuals and assigning adjusted residual to each
    cell
for (i in 1:2)
  for (j in 1:3)
    adjusted_resid[i, j] <- (observed[i,j] - expected[i,j]) / sqrt(
   expected[i,j] * (1 - bribe_study[i, 4] / bribe_study[3,4]) * (1 -
   bribe_study[3, j] / bribe_study[3,4]))
adjusted_resid

Xsq$stdres # Checking answer
```

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

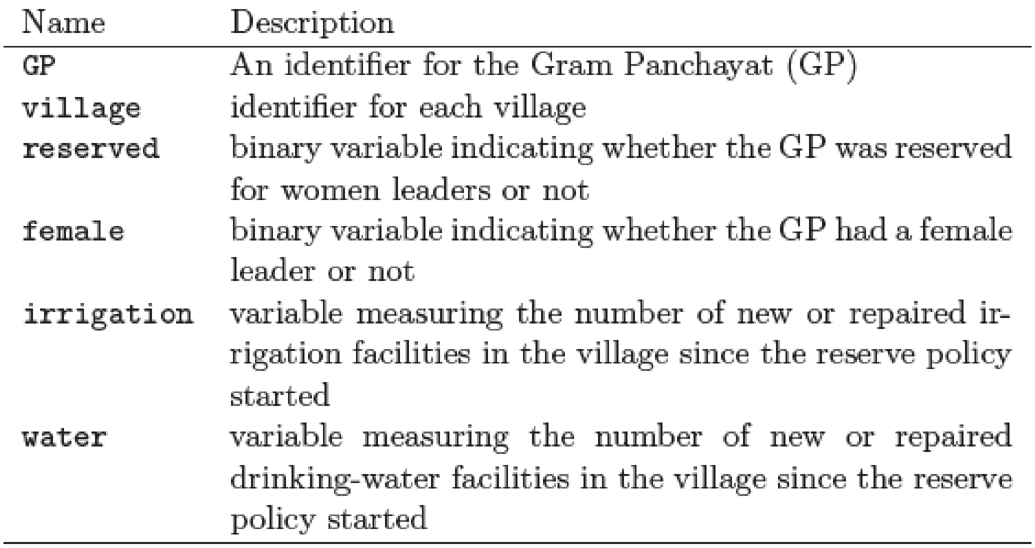
The standardized residuals lets you see how far each observed value is from the expected value. Therefore, you can pinpoint to where the deviation from independence takes place and which cell contributes most to the chi-squared test statistic. For example, I can see through the table of the residuals that requesting a bribe from the lower class had the greatest contribution to the chi-square test statistic, suggesting a stronger association between these specific levels of the explanatory/response variables.

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.

H_o: There is no association between the reservation policy and the number of drinking water facilities

H_a: There is an association between the reservation policy and the number of drinking water facilities

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

I calculated the estimators for alpha and beta using these formulas on R:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_{\mathrm{i}}-\bar{X})(Y_{\mathrm{i}}-\bar{Y})}{\sum_{i=1}^{n}(X_{\mathrm{i}}-\bar{X})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

```r
women <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header = T) # Importing dataset - can use the URL
summary(women) # Inspecting data

mean.x <- mean(women$reserved)
mean.y <- mean(women$water)
a <- sum((women$reserved - mean.x)*(women$water - mean.y))
b <- sum((women$reserved - mean(women$reserved))^2)
beta <- a / b # slope
beta # slope = 9.25

alpha <- mean.y - (beta * mean.x) # intercept
alpha # intercept = 14.74


model <- lm(water ~ reserved, data = women) # Checking answer
summary(model) # Summary of the fitted model
```

Based on the output from the summary of the linear model, we can conclude that neither the intercept or the slope are equal to 0 since the p-values are significantly less than 0.05. Therefore, there is an association between reservation policy and the number of drinking water facilities, linear or not.

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate for reservation policy was 9.25, which in this context means the mean difference in the number of drinking water facilities between reservation policies. The estimate for the intercept (which signifies the mean number of drinking water facilities in GP's with male reservation policy) was 14.74, so we can add 9.25 to this estimate in order to get the mean number of drinking water facilities in GP's reserved for females. 14.74 + 9.25 = 23.99, so there are, on average, 24 drinking water facilities in GP's with female 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.

```r
fruitfly <- read.csv(”fruitfly.csv”, header = T) # Importing dataset
summary(fruitfly) # Summary statistics

pdf(”fruitfly.pdf”) # Plot for distribution of fruitfly lifespan
hist(fruitfly$lifespan, xlab = ”Lifespan (days)”, main = NULL)
dev.off()
```

   The distribution of the overall lifespan of the fruitflies are unimodal, symmetric, and approximately normal.

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

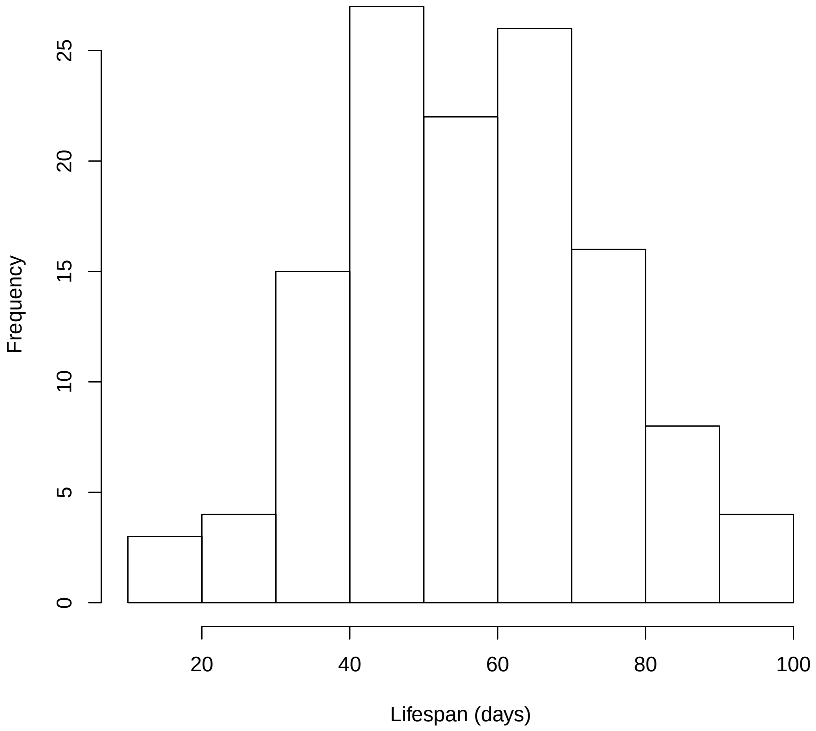


Figure 2: Distribution of Fruitfly Lifespan

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?

```r
pdf("thorax_lifespan.pdf") # Plot for lifespan vs. thorax
plot(fruitfly$thorax, fruitfly$lifespan, main = NULL, xlab = "Thorax (mm)", ylab = "Lifespan (days)") # Moderately linear
dev.off()
cor(fruitfly$thorax, fruitfly$lifespan) # Correlation coefficient = 0.64
```

The correlation coefficient betwen the two variables is 0.64, so fruitfly lifespan and thorax length has a moderate, linear relationship.

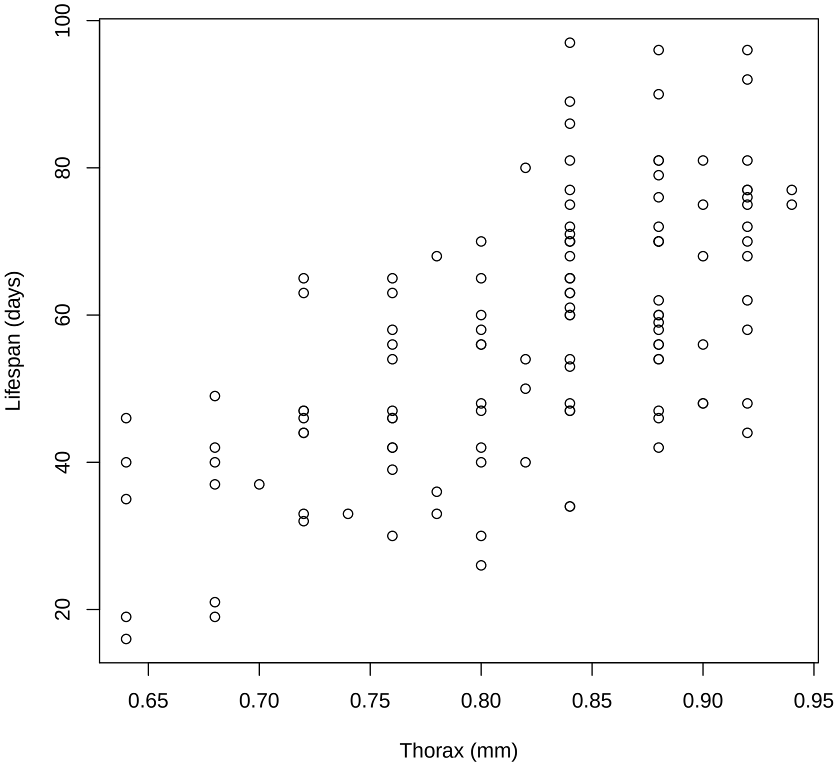


Figure 3: Plot of Fruitfly Lifespan vs. Thorax Length

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

```r
# Regress lifespan on thorax
mean.x <- mean(fruitfly$thorax) # Mean of x
mean.y <- mean(fruitfly$lifespan) # Mean of y
a <- sum((fruitfly$thorax - mean.x)*(fruitfly$lifespan - mean.y))
b <- sum((fruitfly$thorax - mean(fruitfly$thorax))^2)
 beta <- a / b # slope
 beta # slope = 144.33

 alpha <- mean.y - (beta * mean.x) # intercept
 alpha # intercept = -61.05

 model2 <- lm(lifespan ~ thorax, data = fruitfly) # checking answer
summary(model2)
```

The slope of the fitted model is 144.33, which means that for every 1mm in increase in thorax length, there is an increase of about 144.33 days in fruitfly lifespan.

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and

interpret your results of your test.

The null hypothesis for this test is that the true population correlation coefficient is equal to 0, while the alternate hypothesis is that the population correlation coefficient is not equal to 0.

```r
# Pearson correlation
pear_cor <- (cov(fruitfly$thorax, fruitfly$lifespan)) / (sd(fruitfly$thorax) * sd(fruitfly$lifespan))
pear_cor

cor(fruitfly$thorax, fruitfly$lifespan) # Checking answer

t_stat <- (pear_cor * sqrt(125 - 2)) / sqrt(1 - pear_cor^2) # Calculating the t-test statistic
t_stat

p_value <- 2 * pt(t_stat, 125 - 2, lower.tail = FALSE) # Calculating the p-value
p_value

cor.test(fruitfly$thorax, fruitfly$lifespan) # Checking answer
```

Since the p-value of 1.5e-15 is significantly less than the alpha value of 0.05, we reject the null hypothesis that the true population correlation coefficient is equal to 0. While we cannot conclude that there is a linear relationship between the two variables, there is a statistically significant relationship.

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

   - Use the formula for typical confidence intervals to find the 90% confidence interval around the point estimate.

     I calculated the 90% confidence intervals around the slope of the fitted model using this formula:

     $$90\% \text{ confidence interval: } \beta_1 \pm t_{90} se_{\hat{\beta}_1}$$

     ```r
     slope <- 144.33 # Slope
     se <- 15.77 # Standard error (from R output in summary)
     sample_size <- nrow(fruitfly)
     t90 <- qt((1 - 0.9) / 2, lower.tail = F, df = sample_size - 2) # t-score for 90% conf interval
     lower_90 <- slope - t90 * se # Lower bound
     upper_90 <- slope + t90 * se # Upper bound
     c90 <- c(lower_90, upper_90)
     c90 # (118.2, 170.5)
     ```

   - Now, try using the function `confint()` in `R`.

     ```r
     confint(model2, level = 0.9) # Checking answer
     ```

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 predicted individual fruitfly's lifespan when thorax = 0.8mm is 54.41 days, and the average lifespan of fruitflies when thorax = 0.8 is also 54.41 days. The prediction interval is wider (27.38, 81.45) than the confidence interval (51.92, 56.91) because it is for an individual response rather than for a mean.

```r
new_fruitfly <- data.frame(thorax = 0.8) # Creating new data
predict(model2, new_fruitfly, se.fit = T) # Running prediction: 54.51 days
prediction <- predict(model2, new_fruitfly, interval="prediction", level=0.95) # Prediction intervals
prediction # (27.38, 81.45)
confidence <- predict(model2, new_fruitfly, interval="confidence", level=0.95) # Confidence intervals
confidence # (51.92, 56.91)
```

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.

```r
# First plot attempt
new_df <- cbind(fruitfly, prediction, row.names = NULL)
pdf("conf_interval2.pdf")
ggplot(new_df, aes(thorax, lifespan)) + 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)
dev.off()

# Second plot attempt
newww_fruitfly <- fruitfly
newww_fruitfly$thorax <- 0.8
pred <- predict(lm(fruitfly$lifespan~fruitfly$thorax), newdata = newww_fruitfly, interval = "prediction", level = 0.95)
pred

pdf("conf_interval.pdf")
#ggplot(fruitfly, aes(x=thorax, y=lifespan)) + geom_point() + geom_smooth(method=lm, se=TRUE)
new_df2 <- cbind(fruitfly, pred, row.names = NULL) # New dataframe
ggplot(new_df2, aes(thorax, lifespan)) + 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)
dev.off()
```

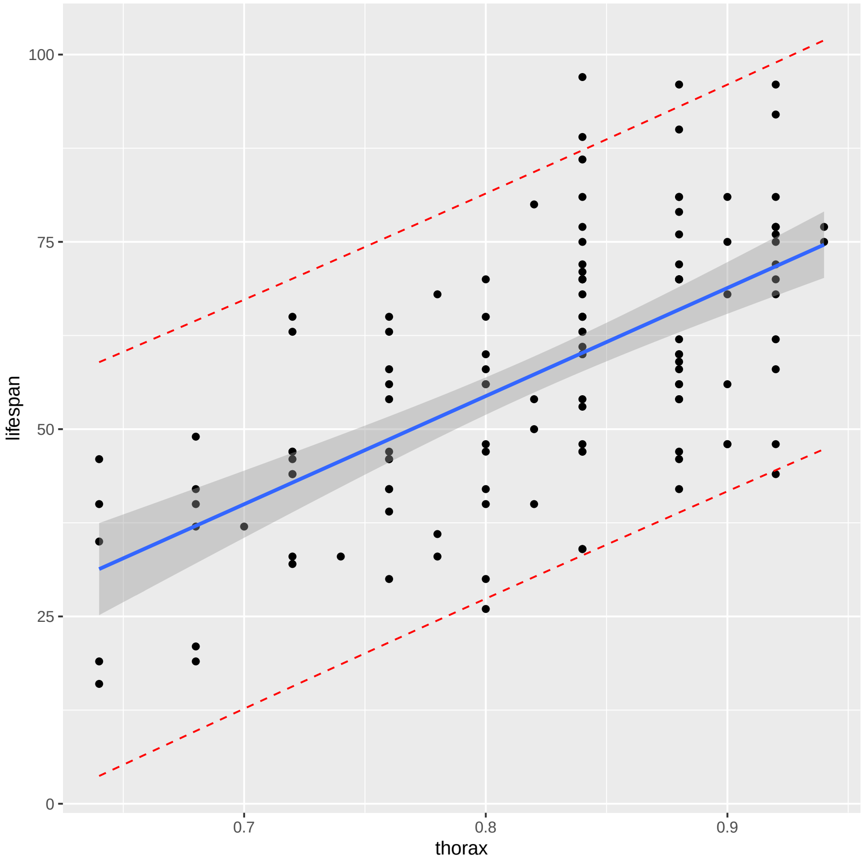


Figure 4: Plot of Fruitfly Lifespan vs. Thorax Length with Confidence Intervals and Prediction Intervals

The shaded region around the blue line is the confidence intervals, and the red dotted lines signify the prediction intervals.

(I wasn't sure which one was right so I just put both...)

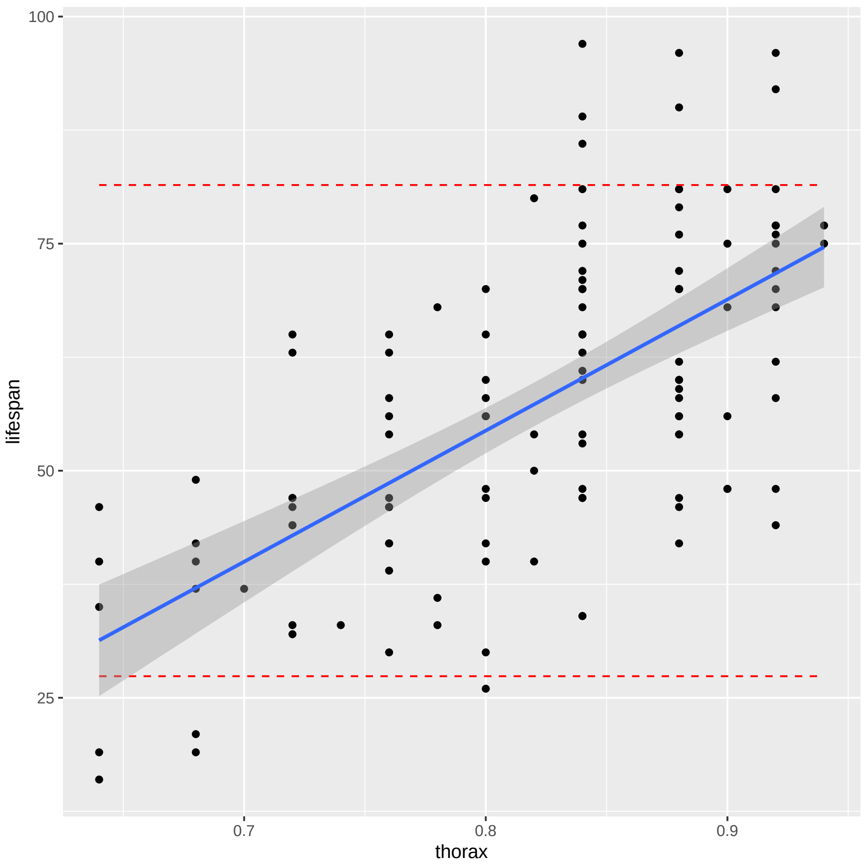


Figure 5: Plot of Fruitfly Lifespan vs. Thorax Length with Confidence Intervals and Prediction Intervals