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). The χ^2 test statistic I found was 3.79. I found this using the following code:

```
F01 <- 14
F02 <- 6
F03 <- 7
F04 <- 7
F05 <- 7
F06 <- 1


row_total_1  <- 27 #(F01 + F02 + F03)
row_total_2 <- 15 #(F04 + F05 + F06)

col_total_1 <- 21
col_total_2 <- 13
col_total_3 <- 8


Grand_total <- 42

#FeN = row total/Grand Total * Column total

Fe1 <- ((row_total_1/Grand_total)*col_total_1)

Fe1

Fe2 <- ((row_total_1/Grand_total)*col_total_2)

Fe3 <- ((row_total_1/Grand_total)*col_total_3)

Fe4 <- ((row_total_2/Grand_total)*col_total_1)

Fe5 <- ((row_total_2/Grand_total)*col_total_2)

Fe6 <- ((row_total_2/Grand_total)*col_total_3)
```

```
# xsquared <- sum((F0 - Fe)^2/Fe)

xsquared <- ((F01 - Fe1)^2/Fe1) + ((F02 - Fe2)^2/Fe2) + ((F03 - Fe3)^2/Fe3) +
((F04 - Fe4)^2/Fe4) + ((F05 - Fe5)^2/Fe5) + ((F06 - Fe6)^2/Fe6)
```

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

To answer this question I made a table for my own sense of clarity, and then calculated the degrees of freedom before using the pchisq function to get a p value of 0.1503183. If $\alpha = .1$? then we can conclude that because the p value found is 0.1503183, and this p value is slightly higher, it means that we reject the null hypothesis that there is no difference between upper and lower class drivers and the p value influences statistical significance. My code can be seen below:

```
police_experiment <-matrix(c(14, 6, 7, 7, 7, 1 ), nrow = 3, ncol = 2)
colnames(police_experiment) <- c('Not Stopped','Bribe Requested','Stopped/ Warned')
rownames(police_experiment) <- c('Upper Class','Lower Class')
police_experiment <- as.table(police_experiment)

df <- (2-1)*(3-1)


pchisq(3.79, df = 2, lower.tail = F )
```

[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			

I am struggling to put the residual values into LaTeX as I don't understand how the output of my code to calculate them corresponds to the headings for the table's rows and columns. However, I am pasting my code and the output I got from it below so you can see what I did and what sort of values I had:

```
Input:

chisq_test <- chisq.test(police_experiment, correct = FALSE)

chisq_test

chisq_test$stdres

Output received:
> chisq_test

Pearson's Chi-squared test

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

> #chisq_test$residuals
>
> chisq_test$stdres
           A          B
A  0.3220306 -0.3220306
B -1.6419565  1.6419565
C  1.5230259 -1.5230259
```

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

Because of my confusion about the output above, this is more of a guess, but as far as I understand, the lower class offered more bribes and more warnings than what our expected value was

The further away it is from zero (we don't expect any variation) the more surprised we are. The lower class having bribes requested of them and being stopped/ given a warning were quite far away from our expectations

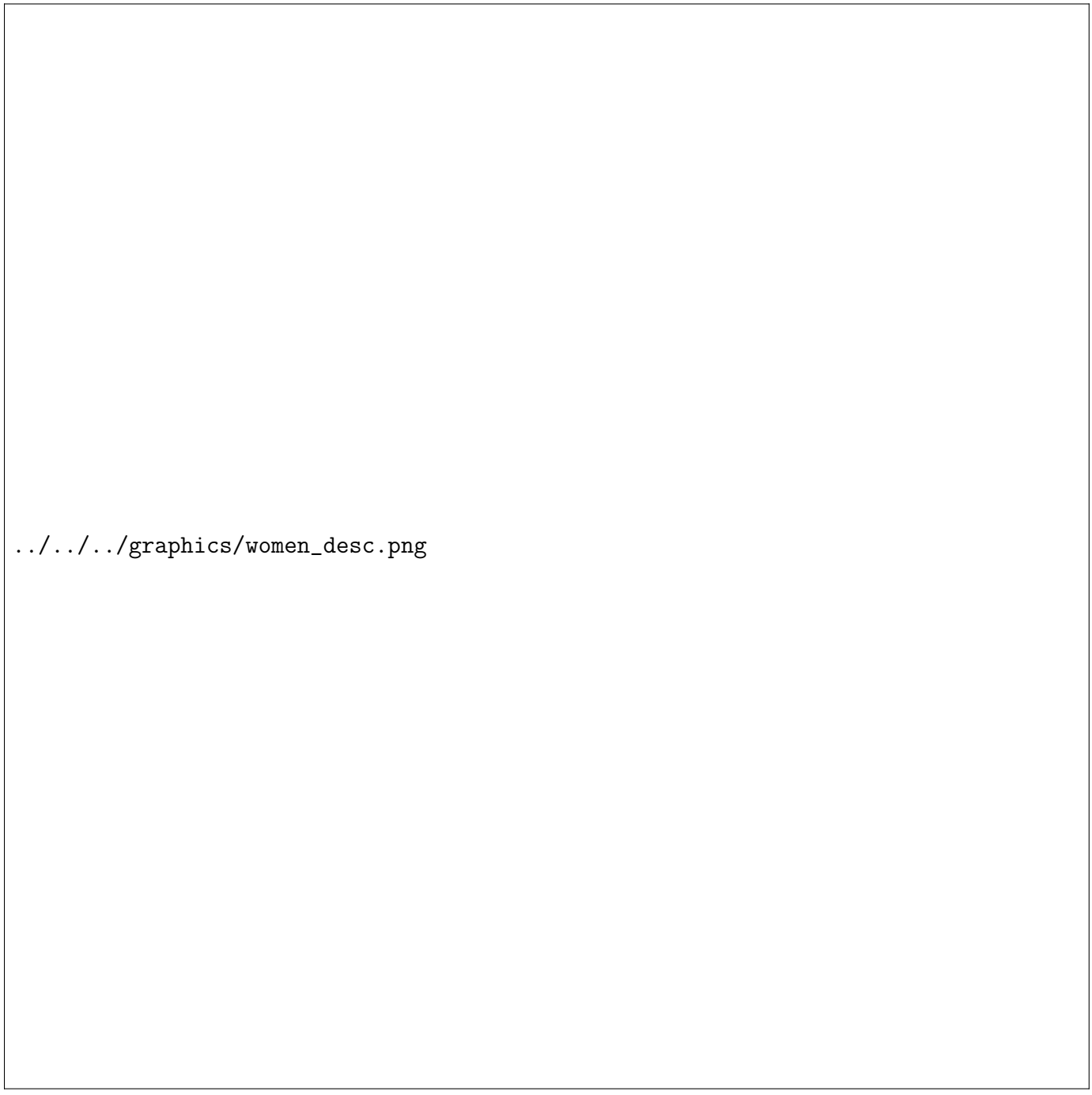
Question 2 (20 points): Economics

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.

Ho = the reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.

Ha = the reservation policy does have an impact on the number of new and repaired drinking water facilities

(b) Run a bivariate regression to test this hypothesis in R (include your code!). The bivariate regression returned a p value of 0.7422. Based on this we would reject the null hypothesis.

Here is my code:

```
policy_data <-
read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.

summary(policy_data)

x <- policy_data$reserved

y <- policy_data$irrigation

lm1 <- lm(policy_data$irrigation~policy_data$reserved)

summary(lm1)
```

(c) Interpret the coefficient estimate for reservation policy.

The estimate coeeficient outputted is -0.36. This is negative so the reservation policy has a negative impact on the irrigation repair systems.

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.

   My code:

```
FruitFlies <- read.csv("https://www.zoology.ubc.ca/~bio501/R/data/fruitflies.csv")

summary(FruitFlies)


hist(FruitFlies$longevity.days,
     main = "Scatter Plot of the overall lifespan of FruitFlies",
     xlab = "Number of Days",
     ylab = "Frequency Distribution")
```

   Here is the histogram graph:

   It looks like a normal distribution.

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

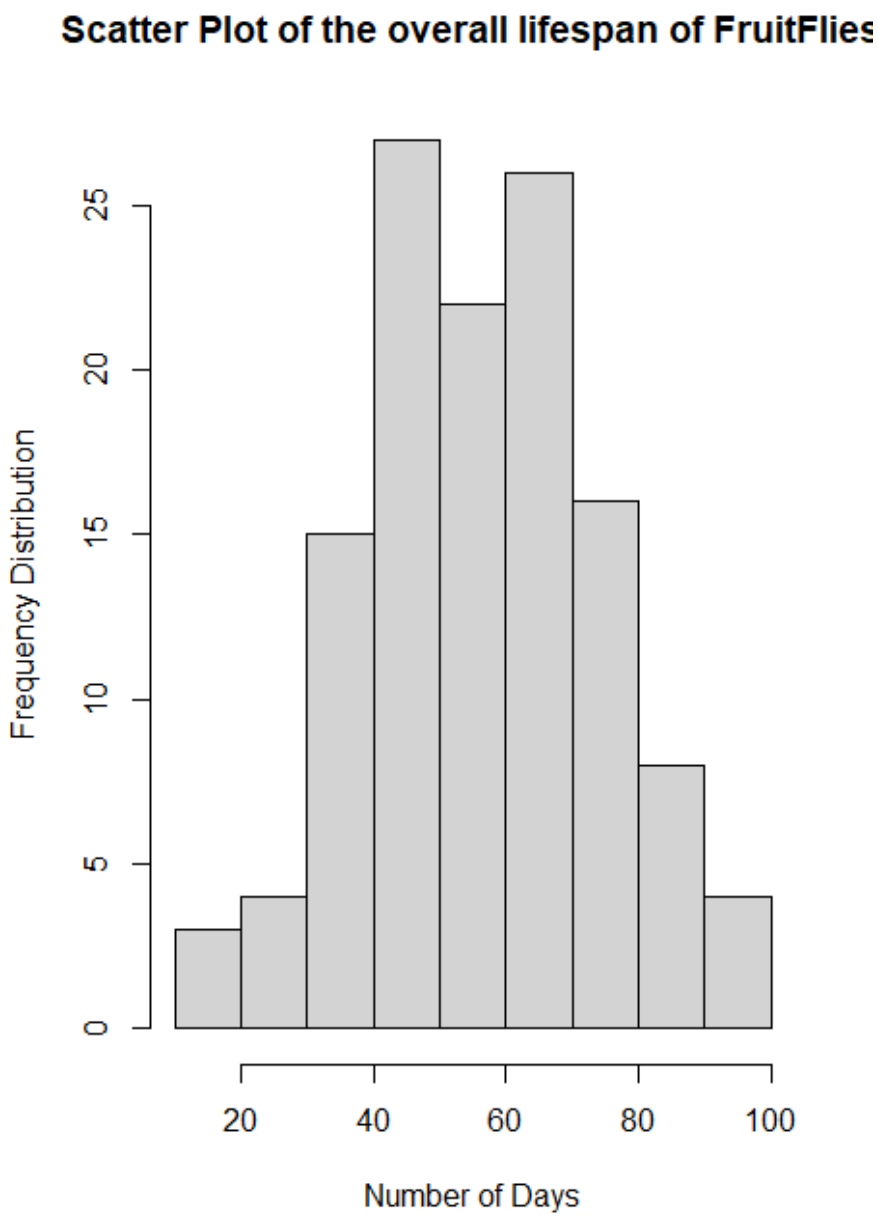

Scatter Plot of the overall lifespan of FruitFlies
Frequency Distribution
0
5
10
15
20
25
20
40
60
80
100
Number of Days

Figure 2:

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 plot can be observed at Figure 3.

   It does look like a linear relationship. The correlation coefficient of 0.63 implies a strong positive correlation.

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

   The slope of the fitted model is 144.33 –¿ for every 0.1 increase in Thorax, Longevity increases with one increase in thorax there is a 14,43 increase in days.

   The code used for this was as follows:

```
lm <- lm(FruitFlies$longevity.days ~ FruitFlies$thorax.mm, data = FruitFlies)
lm
summary(lm)
```

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

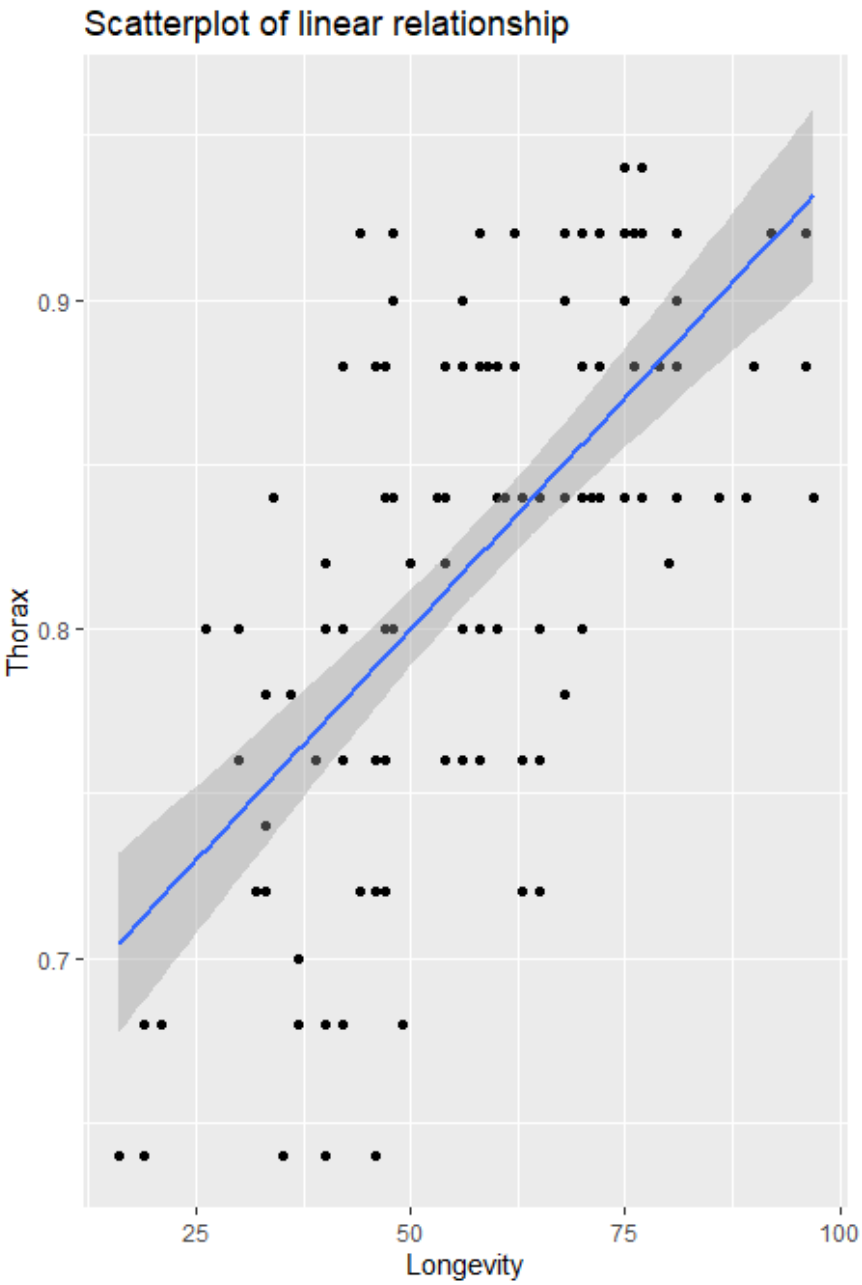

Scatterplot of linear relationship
Thorax
0.9
0.8
0.7
25
50
75
100
Longevity

Figure 3:

We reject the null hypothesis – ¿ 1.497 e -15 –¿ Our null Hypothesis is Ho= no linear relationship between lifespan and thorax –¿ one variable influences variance in the other, therefore we reject it

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

   - Use the formula of confidence interval.

     ```
     lower_interval <- (144.33 - ((15.77*1.645)))
     lower_interval
     upper_interval <- (144.33 + ((15.77*1.645)))
     upper_interval
     ```

     The confidence interval I found was between 118.3884 and 170.2717

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

```
confint(lm)
confint(lm, level = 0.90)
```

My 90 per cent confidence interval output looked like this:

```
     confint(lm)
                          2.5 %   97.5 %
(Intercept)          -86.79221 -35.3112
FruitFlies$thorax.mm 113.11646 175.5497
> confint(lm, level = 0.90)
                            5 %     95 %
(Intercept)          -82.60361 -39.4998
FruitFlies$thorax.mm 118.19616 170.4700
```

¿

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?

```
# lm is the regression line --> we want to use the regression line to
#make prediction off
FruitFlies_ <- lm(longevity.days ~ thorax.mm, data = FruitFlies)
new_df <- data.frame (thorax.mm = 0.8)

predict(FruitFlies_, newdata = new_df)
# It would live 54.4 days

predict(FruitFlies_, newdata = new_df, interval = 'confidence')

The above code produces the following output:

fit      lwr      upr
1 54.41478 51.91932 56.91024
```

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.

```
model1 <- lm(longevity.days ~ thorax.mm, data=FruitFlies)
summary(model1)

plot(FruitFlies$longevity.days, FruitFlies$longevity.days, ylim=c(100, 200), xlab="
abline(model1, col="lightblue")
```

```
lower_interval <- (144.33 - ((15.77*1.645)))
lower_interval
upper_interval <- (144.33 + ((15.77*1.645)))
upper_interval

summary(FruitFlies$thorax.mm)
newx <- seq(0.760, 0.880, by=0.05)
plot(FruitFlies$longevity.days, FruitFlies$thorax.mm, ylim=c(100, 200), xlab="QUET"
abline(model1, col="lightblue")
```

I struggled with the graphing part of this and couldn't get it to work for me.