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`).

```r
#question 1a: calculate fexpected for each cell
##calculate row total for upper class
14+6+7
##calculate row total for lower class
7+7+1
##calculate grand total, add both row totals
27+15
##calculate column total for not stopped
14+7
##calculate column total bribe requested
6+7
##calculate column total for warning
7+1

##fe for not stopped, upper class
fe1 <- (27/42)*21
##fe for bribe requested, upper class
fe2 <- (27/42)*13
##fe for warning, upper class
fe3 <- (27/42)*8
##fe for not stopped, lower class
fe4 <- (15/42)*21
##fe for bribe requested, lower class
fe5 <- (15/42)*13
##fe for warning, lower class
fe6 <- (15/42)*8

#calculate the chi-squared statistic by using sum(fo - fe)^2 / fe
x <- ((14-fe1)^2/fe1) + ((6-fe2)^2/fe2) + ((7-fe3)^2/fe3) + ((7-fe4)^2/fe4) + ((7-fe5)^2/fe5) + ((1-fe6)^2/fe6)
x
```

The chi-squared statistic is 3.791168.

(b) Now calculate the p-value (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.

```r
##df = 2 because there are 3 columns and two rows so (3-1)(2-1) = 2
pchisq(x, df=2, lower.tail = FALSE)
```

The p-value is 0.1502306. Because the p-value (0.15) is not equal to or below the 0.1 threshold, we do not find sufficient evidence to reject the null hypothesis that the variables are statistically independent.

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

```r
##calculate the standard error for each cell
##calculate se for not stopped, upper class
1-(27/42)
1-(21/42)
se1 <- sqrt(fe1*0.357*0.5)
##calculate se for bribe requested, upper class
1-(27/42)
1-(13/42)
se2 <- sqrt(fe2*0.3571429*0.6904762)
##calculate se for warning, upper class
1-(27/42)
1-(8/42)
se3 <- sqrt(fe3*0.3571429*0.8095238)
##calculate se for not stopped, lower class
1-(15/42)
1-(21/42)
se4 <- sqrt(fe4*0.6428571*0.5)
##calculate se for bribe requested, lower class
1-(15/42)
1-(13/42)
se5 <- sqrt(fe5*0.6428571*0.6904762)
##calculate se for warning, lower class
1-(15/42)
1-(8/42)
se6 <- sqrt(fe6*0.6428571*0.8095238)

##calculte standard residual for not stopped, upper class
z1 <- (14-fe1)/se1
z1
###the standardized residual of not stopped, upper class is about 0.322

##calculate standard residual for bribe requested, upper class
z2 <- (6-fe2)/se2
z2
```

```r
###the standardized residual of bribe requested, upper class is about -1.642

##calculate standard residual for warning, upper class
z3 <- (7-fe3)/se3
z3
###the standardized residual of warning, upper class is about 1.523

##calculate standard residual for not stopped, lower class
z4 <- (7-fe4)/se4
z4
###the standardized residual of not stopped, upper class is about -0.322

##calculate standard residual for bribe requested, lower class
z5 <- (7-fe5)/se5
z5
###the standardized residual of bribe requested, lower class is about 1.642

##calculate standard residual for warning, lower class
z6 <- (1-fe6)/se6
z6
###the standardized residual of warning, lower class is about -1.523
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

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

The standardized residuals they tell us how far away our observed result is from the expected result. This is helpful for interpreting our results because we can assess if there are outliers in the data (a data point that is unusually far from the expected value). If an outlier is significantly affecting our regression model, we may need to remove the data point. In this instance, the residuals are not +/- 3 so there does not seem to be any point that is unusually different from the expected value.

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.

Null hypothesis: The reservation policy had no effect on the number of new or repaired drinking water facilities in the villages. Alternative hypothesis: The reservation policy had an effect on the number of new or repaired drinking water facilities in the villages.

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

```
women <- read.csv("~/GitHub/QTM200Spring2020-master/QTM200Spring2020-
    master/problem_sets/PS2/women.csv")
View(women)
#next, calculate sums and means
mean.x <- mean(women$reserved)
mean.y <- mean(women$water)
sum(women$reserved)
sum(women$water)
numerator <- sum((women$water - mean(women$water))*(women$reserved - mean
    (women$reserved)))
denominator <- sum((women$reserved - mean(women$reserved))^2)
#calculate regression coefficients
beta.hat <- numerator/denominator
beta.hat
alpha.hat <- mean.y - (beta.hat*mean.x)
alpha.hat
#calculate p-value, start with sd, then se, then test statistic then
sd.y <- sd(women$water)
se.y <- sd.y/sqrt(sum((women$reserved - mean.x)^2))
TS <- (beta.hat - 0) / se.y
TS
p <- 2*pt(abs(TS), df= (length(women$water)), lower.tail=F)
#check
women.lm <- lm(water~reserved, data=women)
summary(women.lm)
```

Beta hat is about 9.252. Alpha hat is about 14.738. The p-value is 0.0197. Because the p-value is less than the 0.05 significance level, we reject the null hypothesis that the reservation policy had no effect on the number of new or repaired drinking water facilities in the villages.

(c) Interpret the coefficient estimate for reservation policy.

The beta coefficient represents the slope of the regression line. However, because our x-variable is categorical (0 or 1), the beta coefficient is interpreted as there being a 14.738 average increase in the number of new or repaired drinking water facilities in the village since the reserve policy started if the GP was reserved for a woman leader (as opposed to if the GP was not reserved for a woman leader).

Question 3 (40 points): Biology

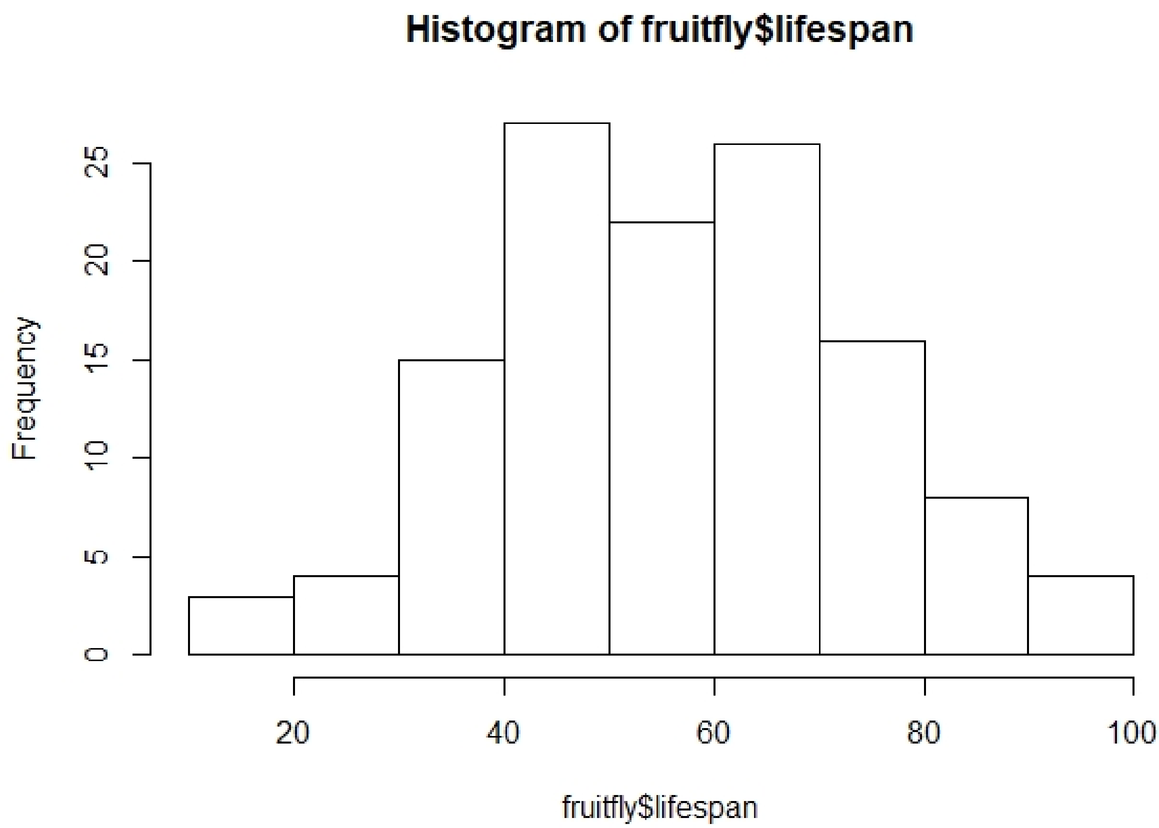
There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity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

```r
#question 3.1, import dataset
fruitfly <- read.csv("~/GitHub/QTM200Spring2020-master/QTM200Spring2020-master/problem_sets/PS2/fruitfly.csv")
summary(fruitfly)
#Summary statistics for lifespan--min 16 days, 25th percentile: 46 days, median: 58 days, mean:57.44 days, 75th percentile: 70 days, max: 97 days. (Summary statistics for each variable can be found by executing above function).
##examine distribution of overall lifespan of fruitflies
hist(fruitfly$lifespan)
```

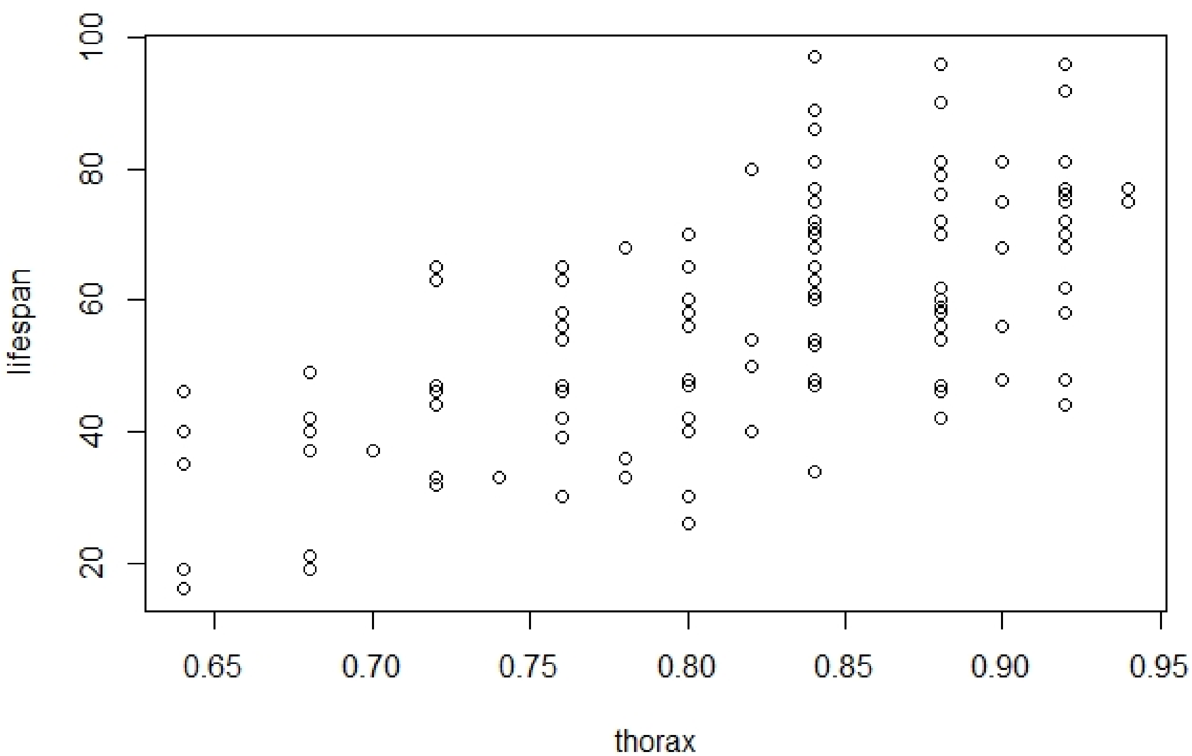
[4]Partridge and Farquhar (1981).“Sexual Activity and the Lifespan of Male Fruitflies”. *Nature.* 294, 580-581.


Histogram of fruitfly$lifespan
Frequency
0
5
10
15
20
25
20
40
60
80
100
fruitfly$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?

```
#question 3.2, plot lifespan vs thorax, calculate correlation coefficient
plot(lifespan~thorax, data=fruitfly)
mean.lifespan <- mean(fruitfly$lifespan)
mean.thorax <- mean(fruitfly$thorax)
sd.lifespan <- sd(fruitfly$lifespan)
sd.thorax <- sd(fruitfly$thorax)
r.fruitfly <- (1/(length(fruitfly$lifespan)-1))*sum(((fruitfly$lifespan - mean.lifespan)/sd.lifespan)*(fruitfly$thorax - mean.thorax)/sd.thorax)
r.fruitfly
#check
cor(fruitfly$lifespan, fruitfly$thorax)
```

Yes, it looks like there is a linear relationship. The correlation coefficient is about 0.636.



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

```
#question 3.3 regress lifespan on thorax, interpret the slope of the
    fitted model
##step 1: calculate sums and means (already had means from the previous
    problem)
mean.thorax
mean.lifespan
thorax.sum <- sum(fruitfly$thorax)
lifespan.sum <- sum(fruitfly$lifespan)

big.sum <- sum((fruitfly$lifespan - mean(fruitfly$lifespan))*(fruitfly$
    thorax-mean(fruitfly$thorax)))
big.sum

small.sum <- sum((fruitfly$thorax-mean(fruitfly$thorax))^2)
small.sum
beta.hat.flies <- big.sum/small.sum
beta.hat.flies
alpha.hat.flies <- mean.lifespan - (beta.hat.flies*mean.thorax)
alpha.hat.flies
#check
flies.lm <- lm(lifespan~thorax, data=fruitfly)
summary(flies.lm)
```

The slope of the fitted model is beta hat which is 144.3331. This means that for each millimeter (mm) increase in thorax length, there is a 144.3331 increase in lifespan (days).

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

```
##question 3.4 test for a significant linear relationship betwen lifespan
     and thorax
TS.flies <- (r.fruitfly*(sqrt(length(fruitfly$lifespan)-2))/(sqrt(1-(r.
    fruitfly)^2)))
TS.flies
p.flies <- 2*pt(TS.flies, length(fruitfly$lifespan)-2, lower.tail=F)
p.flies
```

The null hypothesis is that there is no relationship between lifespan and thorax. However, our p-value (1.496761e-15) is less than the significance level of 0.05. Thus, we have sufficient evidence to reject the null hypothesis that there is no relationship between lifespan and thorax.

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.

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

```
##question 3.5, run a 90% confidence interval for the slope of the fitted model
confint(flies.lm, level = 0.90)
```

   The 90 percent confidence interval for the slope of the fitted model is (118.196, 170.470). This means that if we took take 100 trials/samples, we would expect 90 of the CIs calculated to contain the true slope and we would expect 10 of the CIs to not include the true slope.

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?

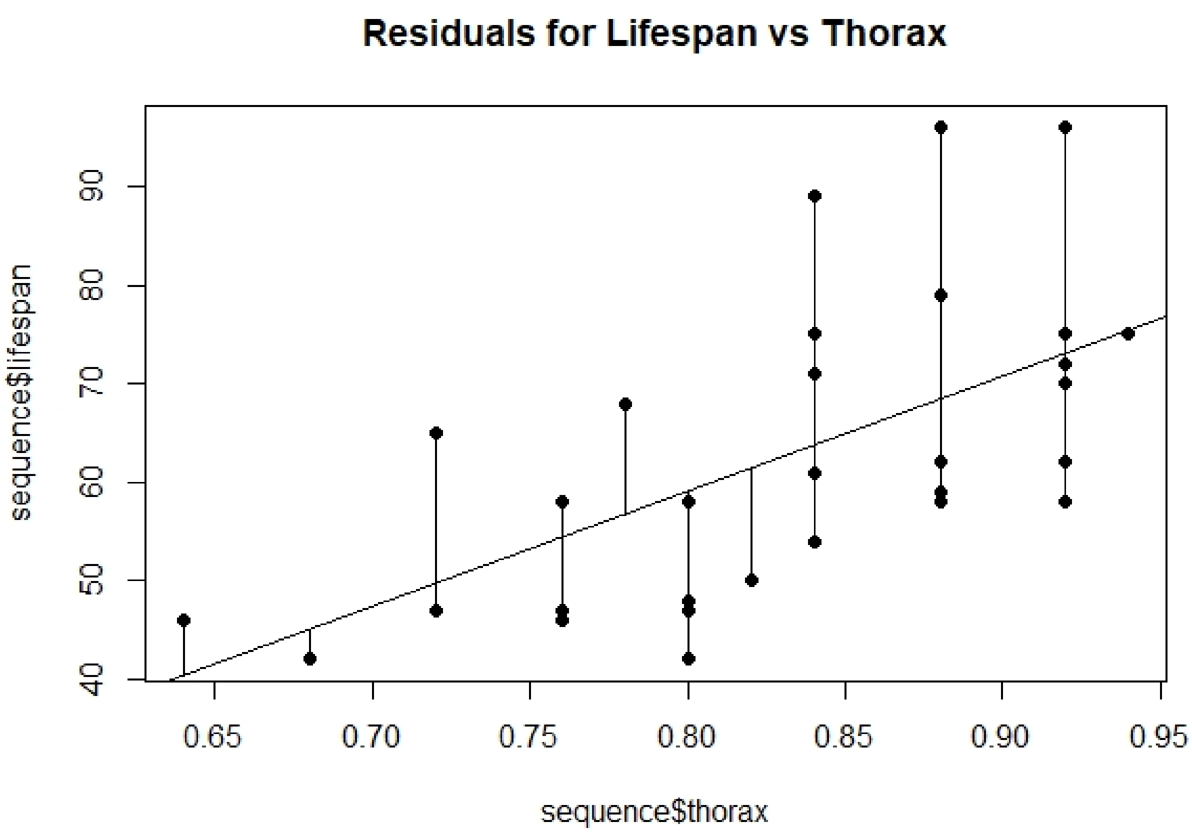
```
##question 3.6 Use the predict() function in R to (1) predict an individual fruit???y's lifespan when thorax=0.8
##this is going to be a prediction interval
new_DF <- fruitfly; new_DF$thorax <- 0.8
predict(lm(new_DF$lifespan~new_DF$thorax), newdata=new_DF, interval ="prediction", level=0.95)

##question 3.6 cont.and (2) the average lifespan of fruit???ies when thorax=0.8 by the ???tted model.
##this is going to be a confidence interval
predict(lm(new_DF$lifespan~new_DF$thorax), newdata=new_DF, interval ="confidence", level=0.95)
```

   The expected value of the lifespan for both individual and average is 57.44 days. For an individual, the prediction interval is (22.53736, 92.34264). For the average, the confidence interval is (54.33063, 60.54937).

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.

```
##question 3.7
sequence <- fruitfly[30:59,]
#prediction interval
predict(lm(sequence$lifespan~sequence$thorax), newdata=sequence, interval
    ="prediction", level=0.95)
#confidence interval
predict(lm(sequence$lifespan~sequence$thorax), newdata=sequence, interval
    ="confidence", level=0.95)
#create the fitted plot
plot(sequence$thorax, sequence$lifespan, pch=19, main = "Residuals for
    Lifespan vs Thorax")
sequence.fit <- lm(lifespan~thorax, data=sequence)
abline(sequence.fit)
preds <- predict(sequence.fit)
segments(sequence$thorax, sequence$lifespan, sequence$thorax, preds)
```



I chose to analyze observations 30 through 59. The fitted plot is provided above. I was unsure if I should list each prediction and confidence interval because there are 20 different intervals for each of the 20 observations, however, the intervals can be created by running the code.