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`).

```r

# (a) Calculate the x^2 statistic

# create a table in R for the given data

data_table <- matrix(c(14, 6, 7, 7, 7, 1), ncol=3, byrow=TRUE)
colnames(data_table) <- c('NotStopped','BribeRequested','StoppedGivenWarning')
rownames(data_table) <- c('UpperClass','LowerClass')
data_table <- as.table(data_table)
data_table

barplot(height = data_table,
        beside = TRUE,
        legend.text = TRUE,
        args.legend = list(x = "topleft", cex = 0.4, box.col = "purple"))

# build the dataset for the table

df <- data.frame (first_column  = c("NotStopped"),
                  second_column = c("BribeRequested"),
                  third_column = c("StoppedGivenWarning"))

# Using the chisq.test to get the x^2 test statistic or the chi square statistic
# x^2 = 3.7912
# DF = 2 (or n-1)
# value of alpha = .1 (for part b)
```

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

```r

# (b) Now calculate thep-value from the test statistic you just created (in R).
# What do you conclude if alpha = .1

chisq <- chisq.test(data_table)
chisq
tail(df, 2)


# x^2 = 3.7912
# DF = 2
# alpha value = .01
# critical value = 1.885618
# x^2(3) = 3.7912, p < 0.1 (the probability was set at 0.1)

# Find the Critical Value
qt(p=0.1, df=2, lower.tail = FALSE)
# This gives a critical value of 1.885618

# The p-value is 0.1502 and the alpha value = 0.1
# This concludes that alpha value 0.1 is less than the p-value of 0.1502
# X^2 > critical value
# reject the Null hypothesis
```

[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			

```

# (c) Calculate the standardized residual for each cell and put them in the
# table below

chisq$residuals
# this gives
#              NotStopped  BribeRequested  StoppedGivenWarning
# UpperClass   0.1360828      -0.8153742            0.8189230
#LowerClass   -0.1825742       1.0939393           -1.0987005
```

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

```

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

# to help identify outliers
```

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).

../../../graphics/women_desc.png

(a) State a null and alternative (two-tailed) hypothesis.

```

# (a) State a null and alternative (two-tailed) hypothesis
```

```
west_bengal <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
attach(west_bengal)
west_bengal
summary(west_bengal)
str(west_bengal)

# H0 is the null hypothesis
# Ha is the alternative hypothesis

# H0: 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 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!).

```

# (b) Run a bivariate regression to test this hypothesis
# prediction - for the data = women
# water depends/extends on reserved
res_policy <- lm(water ~ reserved, data = women)
summary(res_policy)

# this gives
#Coefficients:
#               Estimate Std. Error t value Pr(>|t|)
# (Intercept)   14.738      2.286    6.446 4.22e-10 ***
# reserved       9.252      3.948    2.344   0.0197 *

# and p-value of 0.0197
# F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197

# taking an alpha of 5%
# alpha value = .05
# p-value = .0197
# the p-value is less than the alpha value that I am trying to test,
# so I can reject the null hypothesis
```

(c) Interpret the coefficient estimate for reservation policy.

```

# (c) Interpret the coefficient for reservation policy
# when measuring the correlation between two variables is trying to put a
# number on their association
```

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

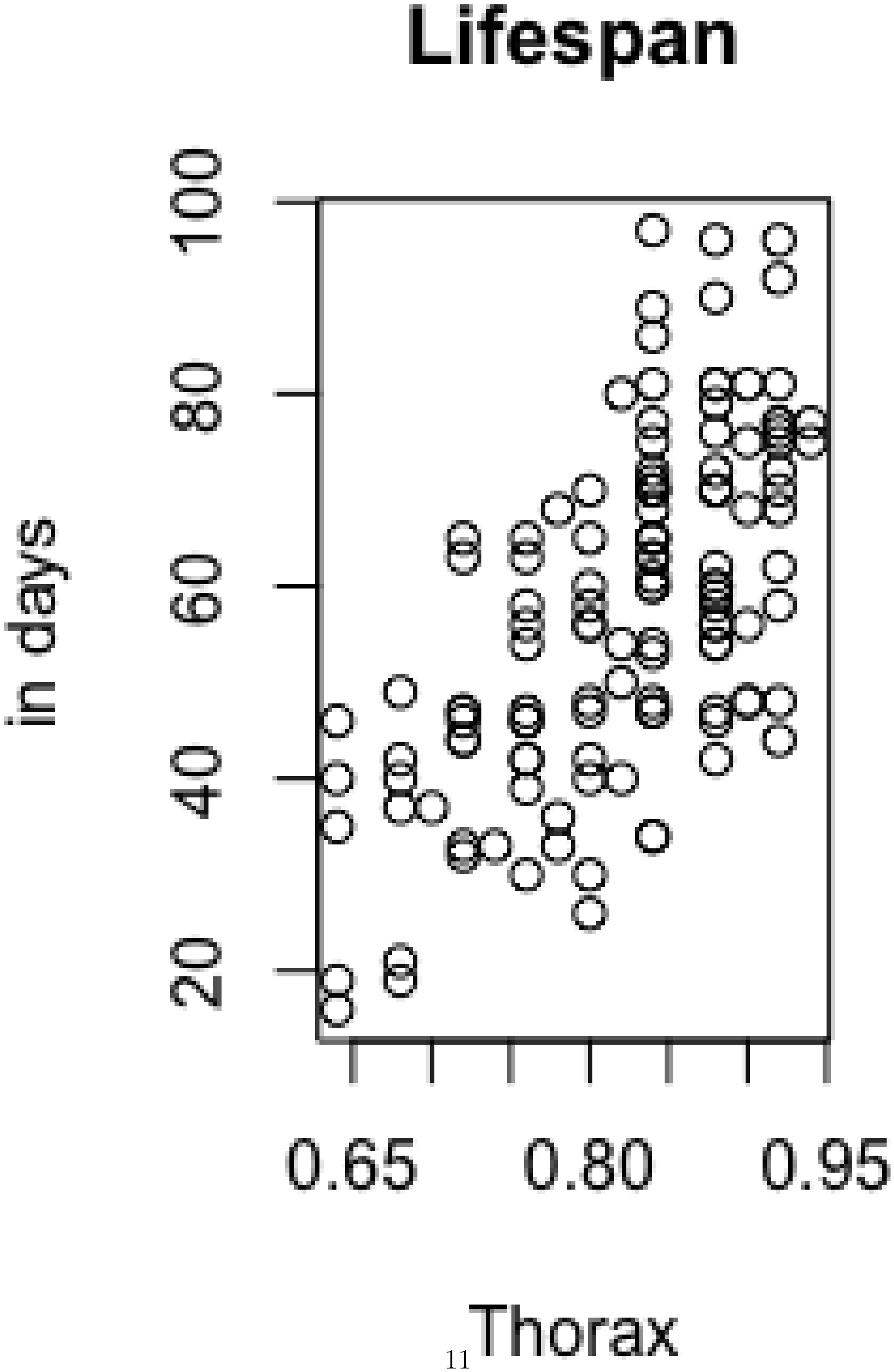
1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

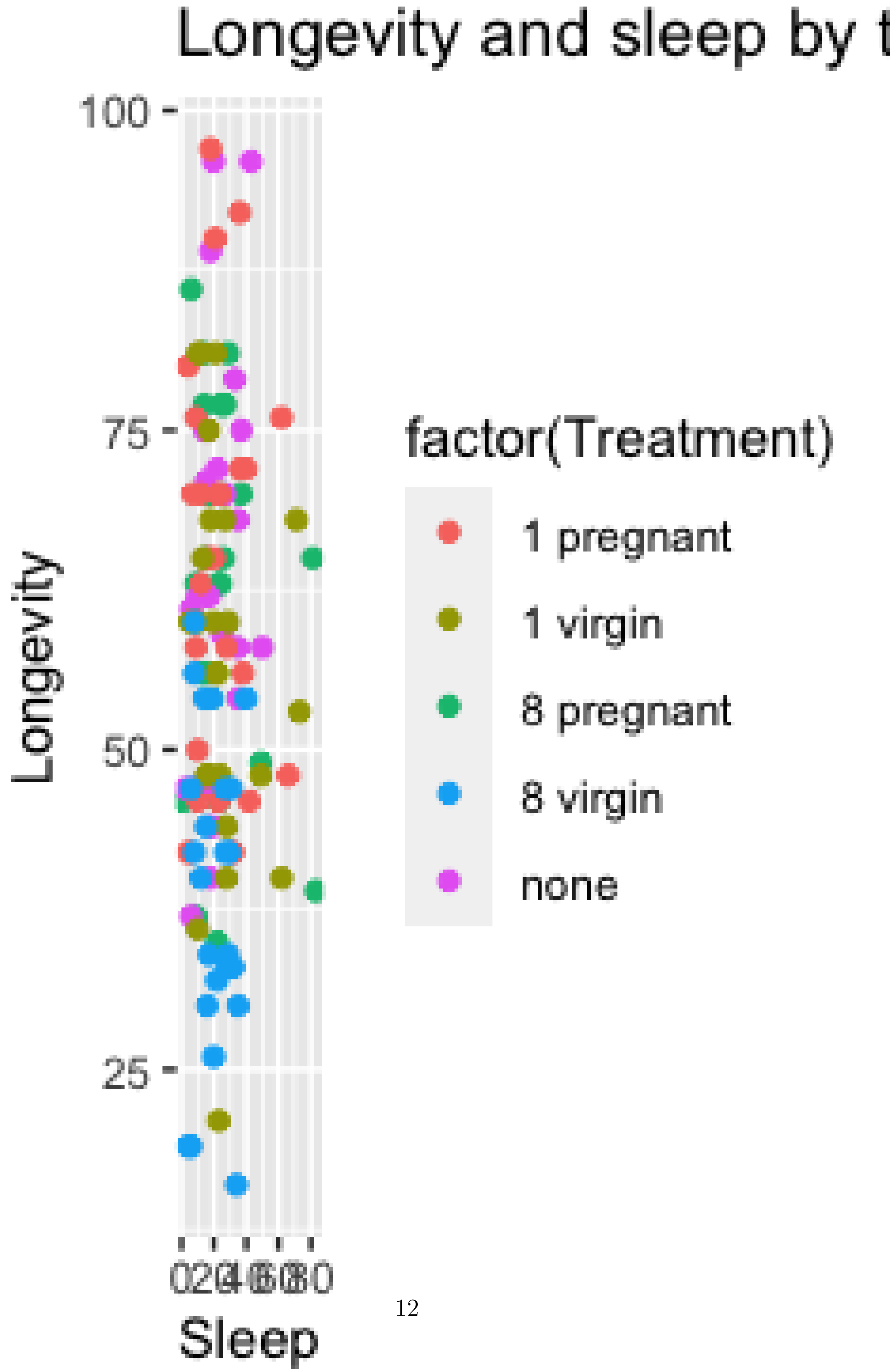
```

# 1. Import the data set and obtain summary statistics and examine the distrubtion of
# the overall lifespan of the fruitflies.

# import data set
dat<-read.csv(”http://stat2.org/datasets/FruitFlies.csv”)
# or
dat<-read.table(”http://stat2.org/datasets/FruitFlies.csv”, sep=’,’, header=TRUE)
attach(dat)
dat
# can also get the descriptive statistics in R using Hmisec package
install.packages(”Hmisc”)
library(Hmisc)
describe(dat)
```

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

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

# 2. Plot lifespan vs thorax.
plot(Longevity~Thorax, xlab="Thorax",
     main="Lifespan", ylab= "in days")

boxplot(Longevity~Treatment, data=dat, frame =FALSE)
summary(dat)

qplot(Sleep, Longevity, data=dat, color=factor(Treatment))+
  labs(title="Longevity and sleep by treatment")

# The correlation coefficient between the two variables Longevity (or lifespan) and thorax
# using the Pearson correlation
cor(dat$Longevity, dat$Thorax)
# the correlation deficit between the two variables lifespan vs thorax is 0.6364835

# this gives a summary of rhe statistics of the data
summary(dat)
# examines the distrubution of the overall lifespan of the fruitflies
summary(dat$Longevity)
```


Lifespan
in days
100
80
60
40
20
0.65
0.80
0.95
Thorax


Longevity and sleep by t
100
75
50
25
Longevity
0
20
40
60
80
Sleep
factor(Treatment)
● 1 pregnant
● 1 virgin
● 8 pregnant
● 8 virgin
● none

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

```

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


lm(dat$Longevity ~ dat$Thorax, data=dat)
# this gives:
# Call:
# lm(formula = dat$Longevity ~ dat$Thorax)

# Coefficients:
# (Intercept)   dat$Thorax
# -61.05        144.33

# 144.3 gives the slope of the line
```

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

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` .

```

# 5. Provide the 90 per cent confidence interval for the slope of the fitted model
# DF 123
qt(.05, 123, lower.tail = FALSE)

#Coefficients:
#  (Intercept)    dat$Thorax
#   -61.05         144.33
```

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.