Problem Set 2

QTM 200: Applied Regression Analysis

Vanessa Wong

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.

(a) Calculate the χ^2 test statistic by hand (even better if you can do "by hand" in `R`).

---

[1]Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multi-method 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

- The chi-squared test statistic is 3.791168.

```
1  #Calculate chi-squared test statistic by hand
2  copbribe <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, byrow=T)
3  copbribe
4  grandtotal <- sum(copbribe)
5  grandtotal
6  upperclassrowsum <- sum(copbribe[1,])
7  upperclassrowsum
8  lowerclassrowsum <- sum(copbribe[2,])
9  lowerclassrowsum
10 notstoppedcsum <- sum(copbribe[,1])
11 briberequestedcsum <- sum(copbribe[,2])
12 warningcsum <- sum(copbribe[,3])
13 e14 <- ((upperclassrowsum*notstoppedcsum)/grandtotal)
14 e6 <- ((upperclassrowsum*briberequestedcsum)/grandtotal)
15 e7a <- ((upperclassrowsum*warningcsum)/grandtotal)
16 e7b <- ((lowerclassrowsum*notstoppedcsum)/grandtotal)
17 e7c <- ((lowerclassrowsum*briberequestedcsum)/grandtotal)
18 e1 <- ((lowerclassrowsum*warningcsum)/grandtotal)
19 copbribeexpected <- matrix(c(e14, e6, e7a,e7b, e7c,e1), nrow=2, byrow=T)
20 copbribeexpected
21 unsummedchisq <- ((copbribe-copbribeexpected)^2)/copbribeexpected
22 unsummedchisq
```

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

- p-value $= 0.1502 > 0.1$, therefore we conclude that there is not enough evidence to reject (i.e. fail to reject) the null hypothesis that x and y are statistically independent.

```
1  #Calculating p-value
2  #df = (3-1)(2-1) = 2
3  pchisq(teststat, df=2, lower.tail=FALSE)
4  #p-value = 0.1502 > 0.1, therefore we conclude that there is not enough
      evidence to reject (i.e. fail to reject) the null hypothesis that x
      and y are statistically 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	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

```
1  #Calculating standardized residuals
2  r_unsquared <- (copbribe-copbribeexpected)
3  r_unsquared
4  upperrowprop <- (1-(upperclassrowsum/grandtotal))
5  lowerrowprop <- (1-(lowerclassrowsum/grandtotal))
6  notstoppedcprop <- (1-(notstoppedcsum/grandtotal))
7  bribecprop <- (1-(briberequestedcsum)/grandtotal)
8  warningcprop <- (1-(warningcsum/grandtotal))
9  sr_e14 <- r_unsquared[1,1]/sqrt(e14*upperrowprop*notstoppedcprop)
10 sr_e6 <- r_unsquared[1,2]/sqrt(e6*upperrowprop*bribecprop)
11 r_unsquared[1,2]
12 sr_e7a <- r_unsquared[1,3]/sqrt(e7a*upperrowprop*warningcprop)
13 sr_e7b <- r_unsquared[2,1]/sqrt(e7b*lowerrowprop*notstoppedcprop)
14 sr_e7c <- r_unsquared[2,2]/sqrt(e7c*lowerrowprop*bribecprop)
15 sr_e1 <- r_unsquared[2,3]/sqrt(e1*lowerrowprop*warningcprop)
16 stan_resid <- matrix(c(sr_e14, sr_e6, sr_e7a, sr_e7b, sr_e7c, sr_e1),
       nrow=2, byrow=T)
17 stan_resid
```

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

- Standardized residuals show how much each cell's observed value deviates from its respective expected value, which is the value that would be obtained if X and Y were independent 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

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

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

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

- Null: there is no relationship between the GP being reserved for female leaders and the # new/repaired drinking water facilities is 0

- Alternative: there exists some relationship between the GP being reserved for women leaders and # new/repaired drinking water facilities

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

- Linear model: y = 14.378 + 9.252x

```r
#Run bivariate regression
r <- (cov(women$water, women$reserved))/(sd(women$reserved)*sd(women$water))
r
plot(women$reserved, women$water)
# r = 0.130. There is a weak positive correlation between water and reserved.

# y = water
# x = reserved

ymean <- mean(women$water)
xmean <- mean(women$reserved)
ysum <- sum(women$water)
xsum <- sum(women$reserved)
yy <- (women$water) - (ymean)
xx <- (women$reserved) - (xmean)
yyxxsum <- sum(yy*xx)
yyxxsum
xxsq <- xx^2
sumxxsq <- sum(xxsq)
betawomen <- yyxxsum/sumxxsq
betawomen
# beta = 9.252
alphawomen <- ymean - (betawomen*xmean)
alphawomen
# alpha = 14.738
# linear model: y = 14.378 + 9.252x
womenreg <- lm(water~reserved, data=women)
womenreg


sd_estimate <- sqrt(sum(resid(womenreg)^2)/(dim(women)[1]-2))
sd_estimate
sigma(womenreg)
beta_se <- sd_estimate/sqrt(sum((xxsq)))
betapval <- 2*pt((betawomen-0)/beta_se, dim(women)[1]-2, lower.tail=F)
betapval
summary(womenreg)
# p = 0.0197. There is a statistically reliable relationship between the
    GP being reserved for women leaders and # new/repaired drinking water
    facilities.
```

(c) Interpret the coefficient estimate for reservation policy.

- $\alpha = 14.738$
  . When the GP is not reserved for women leaders (i.e. reserved = 0), the predicted number of new or repaired drinking-water facilities in the village is 0.3029.

- $\beta = 9.252$
  . For each additional GP that is reserved for women leaders, the number of new or repaired drinking-water facilities increases by 9.252.

# 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

---

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

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

- The distribution of overall fruitfly lifespan appears to be bimodal and approximately normal (there is a slight left skew – median is greater than the mean).

Histogram of fly$lifespan
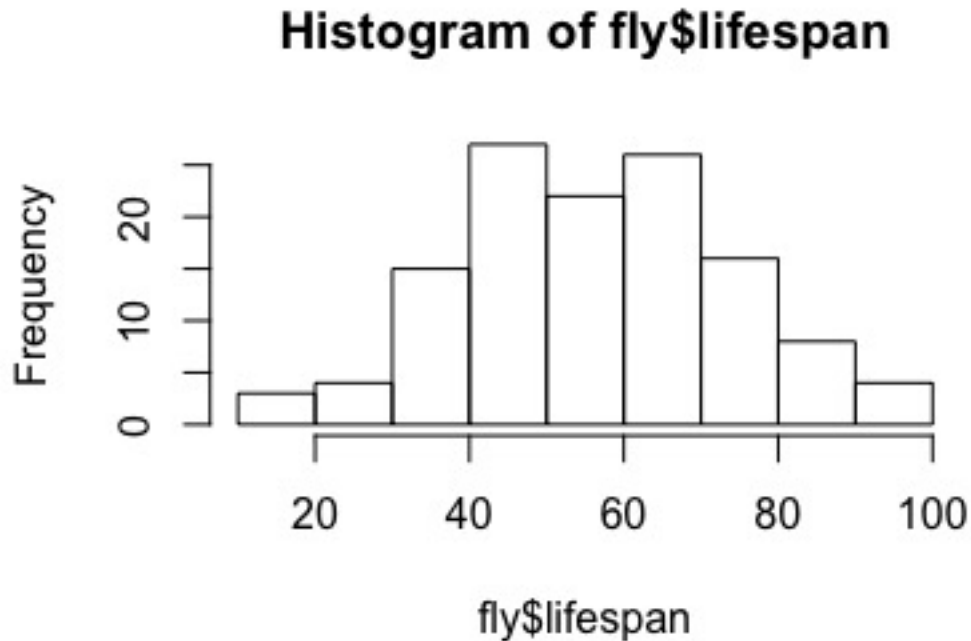
fly$lifespan

- Summary statistics for fruitfly lifespan:

|	
Min	16
Q1	46
Median	58
Mean	57.44
Q3	70
Max	97

```
1  #Set wd and load fruitfly dataset
2  setwd("~/GitHub/QTM200Spring2020/problem_sets/PS2") #copy/paste from
      getwd
3  getwd() #double check
4  library(readr)
5  fly <- read.csv("fruitfly.csv")
6  summary(fly$lifespan)
7  hist(fly$lifespan)
```

- 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?

– There appears to be a linear relationship between fruitfly thorax and lifespan length. r= 0.636. There is a moderate positive correlation between fruitfly thorax length and fruitfly lifespan.

*
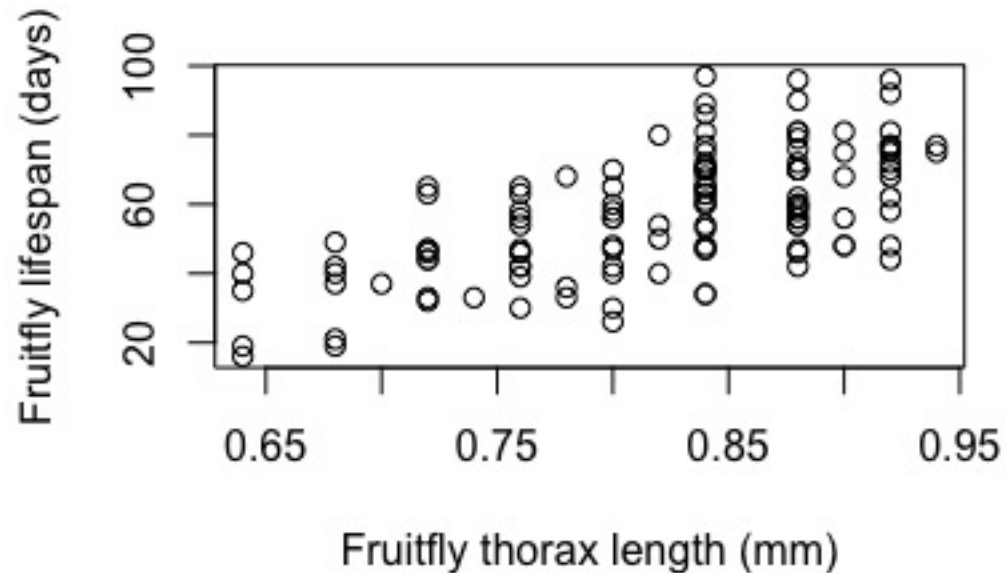
```
1 #Plot thorax vs. lifespan
2 plot(fly$thorax, fly$lifespan,
3     xlab="thorax length (mm)", ylab="lifespan (days)")
4 #There appears to be a linear relationship between fruitfly thorax
      and lifespan length.
5 #r of thorax vs. lifespan
6 r2 <- (cov(fly$thorax, fly$lifespan))/(sd(fly$thorax)*sd(fly$lifespan
      ))
7 r2
8 #r= 0.636. There is a moderate positive correlation between fruitfly
      thorax length and fruitfly lifespan.
```

- Regress **lifespan** on **thorax**. Interpret the slope of the fitted model.

  – Interpretation of beta/slope: For every one mm increase in fruitfly thorax length, there is a 144.33 day increase in fruitfly lifespan.

  – Linear model: y = -61.05 + 144.33x

```
1  #Regress thorax on lifespan. Interpret the slope of the fitted model.
2  mean_y <- mean(fly$lifespan) #y-bar = 57.44
3  mean_y
4  mean_x <- mean(fly$thorax) #x-bar = 0.821
5  sigmalifespan_y <- sum(fly$lifespan) # = 7180
6  sigmalifespan_y
7  sigmathorax_x <- sum(fly$thorax) # = 102.62
8  sigmathorax_x
9  yybar <- (fly$lifespan-mean_y)
10 xxbar <- (fly$thorax-mean_x)
11 xxyy <- (xxbar*yybar)
12 xxyy
13 sum(xxyy) # = 107.367
14 sqxxbar <- (xxbar)^2
15 sqxxbar2 <- sum(sqxxbar)
16 sqxxbar2 # = 0.744
17 beta <- (sum(xxyy))/sqxxbar2
18 beta # = 144.33
19 alpha <- mean_y - (beta*mean_x)
20 alpha # = -61.0517
21 lm(fly$lifespan ~ fly$thorax, data=fly)
22 #linear model: y = -61.05 + 144.33x
23 #Interpretation of beta/slope: For every one mm increase in fruitfly
      thorax length, there is a 144.33 day increase in fruitfly
      lifespan.
```

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

  - p   0 , therefore there is a statistically reliable and significant linear relationship between fruitfly thorax and lifespan length.

```
1 #Test for a significant linear relationship between thorax and lifespan
2 n_fly <- dim(fly)[1]
3 fly_teststat<- (r2*sqrt(n_fly -2))/sqrt(1-(r2)^2)
4 2*pt(fly_teststat, n_fly -2, lower.tail=F)
5 #p-value = ~ 0
6 cor.test(fly$lifespan, fly$thorax)
7 # p ~ 0 , therefore there is a statistically reliable and significant
    linear relationship between fruitfly thorax and lifespan length.
```

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.

- 90% confidence interval for the slope of the fitted model is (118.3915, 170.2748). 0 is not within the interval, supporting the conclusion from the previous quesiton that there is a statistically reliable relationship between fruitfly lifespan and thorax length.

```
1  #90% CI for slope
2  #Calc critical value
3  a <- (1-(90/100))
4  critprob <- 1 - (a/2) # = 0.95
5  df <- n_fly - 2 # = 123
6  #critical value = 1.645
7  critvalue <- 1.645
8  mod = lm(fly$lifespan ~ fly$thorax, data=fly)
9  summary(mod)
10 #standard error = 15.77
11 moe2 <- critvalue*15.77
12 lower_90b <- beta - moe2
13 upper_90b <- beta + moe2
14 ci90b <- c(lower_90b, upper_90b)
15 ci90b
16 # 90% confidence interval is (118.3915, 170.2748)
17 confintfunc <- lm(lifespan ~ thorax, data=fly)
18 confint(confintfunc, level=0.90)
19 # 90% confidence interval is the same: (118, 170). 0 is not within the
       interval, supporting the conclusion from the previous quesiton that
       there is a statistically reliable relationship between fruitfly
       lifespan and thorax length.
```

- 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 range for an individual fruitfly's lifespan when thorax length = 0.8 is 27.375 to 81.454 days. The point estimate/expected value is 54.414 days.

  - The predicted range for the average lifepsan of fruitflies when thorax length = 0.8 is 51.919 to 56.910 days. The point estimate/expected value is 54.414 days.

```
1  #Prediction of (average) lifespan (y) when thorax length (x) = 0.8
2  new_fly <- data.frame(thorax=0.8)
3  predicintervals <- predict(confintfunc, newdata=new_fly, interval="
       prediction", level=0.95)
4  predicintervals
```

```
5 # The predicted range for an individual fruitfly 's lifespan when thorax
      length = 0.8 is 27.375 to 81.454 days. The point estimate/expected
      value is 54.414 days.
6 confintervals <- predict(confintfunc, newdata=new_fly, interval="
      confidence", level=0.95)
7 confintervals
8 # The predicted range for the average lifepsan of fruitflies when thorax
      length = 0.8 is 51.919 to 56.910 days. The point estimate/expected
      value is 54.414 days.
```

- For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.

Fruitfly lifespan (days) vs. thorax length (mm)

  —
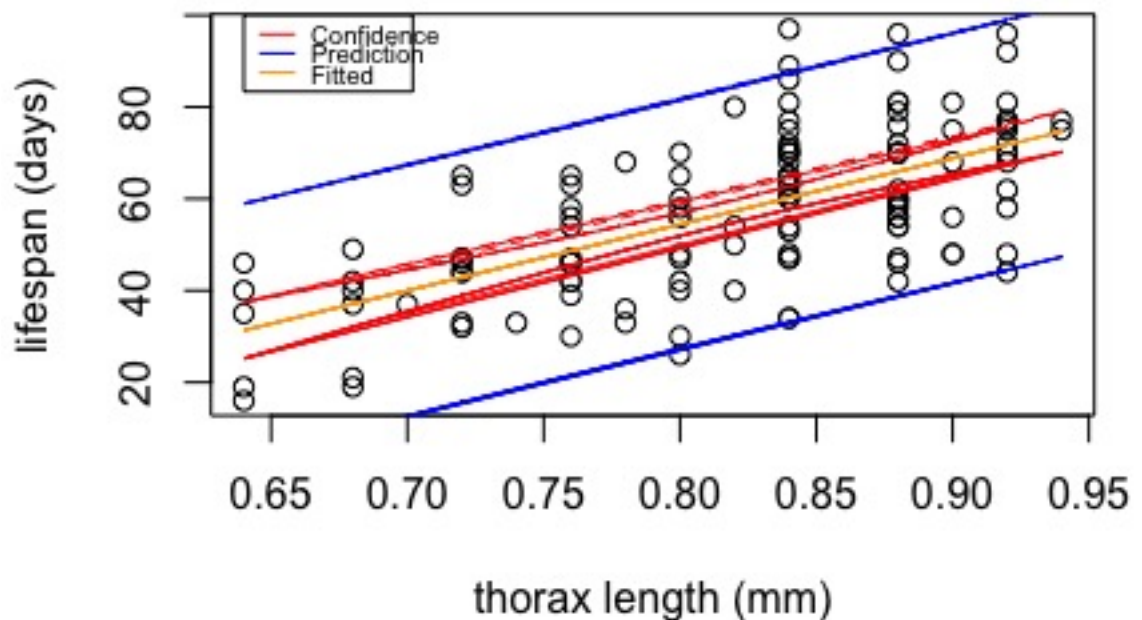
```
1 confintfunc
2 confseq <- predict(confintfunc, newdata=fly, interval = "confidence")
3 confseq
4 predicseq <- predict(confintfunc, newdata=fly, interval = "prediction")
5 predicseq
6 plot(fly$thorax, fly$lifespan,
7      xlab="thorax length (mm)", ylab="lifespan (days)", main="Fruitfly
      lifespan (days) vs. thorax (mm)")
```

```r
8  Confidence <- matlines(fly$thorax, confseq[,c("lwr","upr")], col="red")
9  Prediction <- matlines(fly$thorax, predicseq[,c("lwr","upr")], col="blue"
       )
10 fitted.values = predicseq[1:125]
11 Fitted <- lines(fly$thorax[1:125], fitted.values[1:125], col="orange",
       lwd=1)
12 legend(x=0.64, y=100, legend=c("Confidence", "Prediction", "Fitted"), cex
       =0.6, col=c("red", "blue", "orange"), lty=1)
```

- R Code:

```r
1  #######################
2  # load libraries
3  # set wd
4  # clear global .envir
5  #######################
6
7  # remove objects
8  rm(list=ls())
9  # detach all libraries
10 detachAllPackages <- function() {
11   basic.packages <- c("package:stats", "package:graphics", "package:
       grDevices", "package:utils", "package:datasets", "package:methods", "
       package:base")
12   package.list <- search()[ifelse(unlist(gregexpr("package:", search()))
       ==1, TRUE, FALSE)]
13   package.list <- setdiff(package.list, basic.packages)
14   if (length(package.list)>0)  for (package in package.list) detach(
       package,  character.only=TRUE)
15 }
16 detachAllPackages()
17
18 # load libraries
19 pkgTest <- function(pkg){
20   new.pkg <- pkg[!(pkg %in% installed.packages()[,  "Package"])]
21   if (length(new.pkg))
22     install.packages(new.pkg,  dependencies = TRUE)
23   sapply(pkg,  require,  character.only = TRUE)
24 }
25
26 # here is where you load any necessary packages
27 # ex: stringr
28 # lapply(c("stringr"),  pkgTest)
29
30 lapply(c("stringr"),  pkgTest)
31
32 # set working directory
33 setwd("~/GitHub/QTM200Spring2020/problem_sets/PS1")
34
35
36 #######################
```

```r
37 # Problem 1
38 ######################
39
40 #Calculate chi−squared test statistic by hand
41 copbribe <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, byrow=T)
42 copbribe
43 grandtotal <- sum(copbribe)
44 grandtotal
45 upperclassrowsum <- sum(copbribe[1,])
46 upperclassrowsum
47 lowerclassrowsum <- sum(copbribe[2,])
48 lowerclassrowsum
49 notstoppedcsum <- sum(copbribe[,1])
50 briberequestedcsum <- sum(copbribe[,2])
51 warningcsum <- sum(copbribe[,3])
52 e14 <- ((upperclassrowsum*notstoppedcsum)/grandtotal)
53 e6 <- ((upperclassrowsum*briberequestedcsum)/grandtotal)
54 e7a <- ((upperclassrowsum*warningcsum)/grandtotal)
55 e7b <- ((lowerclassrowsum*notstoppedcsum)/grandtotal)
56 e7c <- ((lowerclassrowsum*briberequestedcsum)/grandtotal)
57 e1 <- ((lowerclassrowsum*warningcsum)/grandtotal)
58 copbribeexpected <- matrix(c(e14, e6, e7a,e7b, e7c,e1), nrow=2, byrow=T)
59 copbribeexpected
60 unsummedchisq <- ((copbribe−copbribeexpected)^2)/copbribeexpected
61 unsummedchisq
62 teststat <- sum(unsummedchisq)
63 teststat
64 #chi−squared test statistic is 3.791168
65
66 #Calculating p−value
67 #df = (3−1)(2−1) = 2
68 pchisq(teststat, df=2, lower.tail=FALSE)
69 #p−value = 0.1502 > 0.1, therefore we conclude that there is not enough
      evidence to reject (i.e. fail to reject) the null hypothesis that x
      and y are statistically independent.
70
71 #Calculating standardized residuals
72 r_unsquared <- (copbribe−copbribeexpected)
73 r_unsquared
74 upperrowprop <- (1−(upperclassrowsum/grandtotal))
75 lowerrowprop <- (1−(lowerclassrowsum/grandtotal))
76 notstoppedcprop <- (1−(notstoppedcsum/grandtotal))
77 bribecprop <- (1−(briberequestedcsum)/grandtotal)
78 warningcprop <- (1−(warningcsum/grandtotal))
79 sr_e14 <- r_unsquared[1,1]/sqrt(e14*upperrowprop*notstoppedcprop)
80 sr_e6 <- r_unsquared[1,2]/sqrt(e6*upperrowprop*bribecprop)
81 r_unsquared[1,2]
82 sr_e7a <- r_unsquared[1,3]/sqrt(e7a*upperrowprop*warningcprop)
83 sr_e7b <- r_unsquared[2,1]/sqrt(e7b*lowerrowprop*notstoppedcprop)
84 sr_e7c <- r_unsquared[2,2]/sqrt(e7c*lowerrowprop*bribecprop)
85 sr_e1 <- r_unsquared[2,3]/sqrt(e1*lowerrowprop*warningcprop)
```

```
86  stan_resid <- matrix(c(sr_e14, sr_e6, sr_e7a, sr_e7b, sr_e7c, sr_e1),
        nrow=2, byrow=T)
87  stan_resid
88
89  #How might the standardized residuals help you interpret the results?
90  #Standardized residuals show how much each cell's observed value deviates
        from its respective expected value, which is the value that would be
        obtained if X and Y were independent variables.
91
92  #######################
93  # Problem 2
94  #######################
95  women <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/
        master/PREDICTION/women.csv")
96  women
97  #Null and alternative hypotheses
98  #Null: there is no relationship between the GP being reserved for female
        leaders and the # new/repaired drinking water facilities is 0
99  #Alternative: there exists some relationship between the GP being
        reserved for women leaders and # new/repaired drinking water
        facilities
100
101 #Run bivariate regression
102 r <- (cov(women$water, women$reserved))/(sd(women$reserved)*sd(
        women$water))
103 r
104 plot(women$reserved, women$water)
105 # r = 0.130. There is a weak positive correlation between water and
        reserved.
106
107 # y = water
108 # x = reserved
109
110 ymean <- mean(women$water)
111 xmean <- mean(women$reserved)
112 ysum <- sum(women$water)
113 xsum <- sum(women$reserved)
114 yy <- (women$water) - (ymean)
115 xx <- (women$reserved) - (xmean)
116 yyxxsum <- sum(yy*xx)
117 yyxxsum
118 xxsq <- xx^2
119 sumxxsq <- sum(xxsq)
120 betawomen <- yyxxsum/sumxxsq
121 betawomen
122 # beta = 9.252
123 alphawomen <- ymean - (betawomen*xmean)
124 alphawomen
125 # alpha = 14.738
126 # linear model: y = 14.378 + 9.252x
127 womenreg <- lm(water~reserved, data=women)
```

```
128  womenreg
129

130
131  sd_estimate <- sqrt(sum(resid(womenreg)^2)/(dim(women)[1]-2))
132  sd_estimate
133  sigma(womenreg)
134  beta_se <- sd_estimate/sqrt(sum((xxsq)))
135  betapval <- 2*pt((betawomen-0)/beta_se, dim(women)[1]-2, lower.tail=F)
136  betapval
137  summary(womenreg)
138  # p = 0.0197. There is a statistically reliable relationship between the
         GP being reserved for women leaders and # new/repaired drinking water
         facilities.
139
140  #Interpret the coefficient estimate for reservation policy
141  # alpha = 14.738. When the GP is not reserved for women leaders (i.e.
         reserved = 0), the predicted number of new or repaired drinking-water
         facilities in the village is 0.3029.
142  # beta = 9.252. For each additional GP that is reserved for women leaders
         , the number of new or repaired drinking-water facilities increases by
         9.252.
143

144

145

146
147  ########################
148  # Problem 3
149  ########################
150  #Set wd and load fruitfly dataset
151  setwd("~/GitHub/QTM200Spring2020/problem_sets/PS2") #copy/paste from
         getwd
152  getwd() #double check
153  library(readr)
154  fly <- read.csv("fruitfly.csv")
155  summary(fly$lifespan)
156  hist(fly$lifespan)
157  #The distribution of overall fruitfly lifespan appears to be bimodal and
         approximately normal.
158
159  #Plot thorax vs. lifespan
160  plot(fly$thorax, fly$lifespan,
161      xlab="thorax length (mm)", ylab="lifespan (days)")
162  #There appears to be a linear relationship between fruitfly thorax and
         lifespan length.
163  #r of thorax vs. lifespan
164  r2 <- (cov(fly$thorax, fly$lifespan))/(sd(fly$thorax)*sd(fly$lifespan))
165  r2
166  #r= 0.636. There is a moderate positive correlation between fruitfly
         thorax length and fruitfly lifespan.
167
168  #Regress thorax on lifespan. Interpret the slope of the fitted model.
```

```r
169  mean_y <- mean(fly$lifespan) #y-bar = 57.44
170  mean_y
171  mean_x <- mean(fly$thorax) #x-bar = 0.821
172  sigmalifespan_y <- sum(fly$lifespan) # = 7180
173  sigmalifespan_y
174  sigmathorax_x <- sum(fly$thorax) # = 102.62
175  sigmathorax_x
176  yybar <- (fly$lifespan-mean_y)
177  xxbar <- (fly$thorax-mean_x)
178  xxyy <- (xxbar*yybar)
179  xxyy
180  sum(xxyy) # = 107.367
181  sqxxbar <- (xxbar)^2
182  sqxxbar2 <- sum(sqxxbar)
183  sqxxbar2 # = 0.744
184  beta <- (sum(xxyy))/sqxxbar2
185  beta # = 144.33
186  alpha <- mean_y - (beta*mean_x)
187  alpha # = -61.0517
188  lm(fly$lifespan ~ fly$thorax, data=fly)
189  #linear model: y = -61.05 + 144.33x
190  #Interpretation of beta/slope: For every one mm increase in fruitfly
        thorax length, there is a 144.33 day increase in fruitfly lifespan.
191
192  #Test for a significant linear relationship between thorax and lifespan
193  n_fly <- dim(fly)[1]
194  fly_teststat <- (r2*sqrt(n_fly-2))/sqrt(1-(r2)^2)
195  2*pt(fly_teststat, n_fly-2, lower.tail=F)
196  #p-value = ~ 0
197  cor.test(fly$lifespan, fly$thorax)
198  # p ~ 0 , therefore there is a statistically reliable and significant
        linear relationship between fruitfly thorax and lifespan length.
199
200
201  #90% CI for slope
202  #Calc critical value
203  a <- (1-(90/100))
204  critprob <- 1 - (a/2) # = 0.95
205  df <- n_fly - 2 # = 123
206  #critical value = 1.645
207  critvalue <- 1.645
208  mod = lm(fly$lifespan ~ fly$thorax, data=fly)
209  summary(mod)
210  #standard error = 15.77
211  moe2 <- critvalue*15.77
212  lower_90b <- beta - moe2
213  upper_90b <- beta + moe2
214  ci90b <- c(lower_90b, upper_90b)
215  ci90b
216  # 90% confidence interval is (118.3915, 170.2748)
217  confintfunc <- lm(lifespan ~ thorax, data=fly)
```

```
218  confint ( confintfunc , level =0.90)
219  # 90% confidence interval is the same: (118, 170). 0 is not within the
          interval , supporting the conclusion from the previous quesiton that
          there is a statistically reliable relationship between fruitfly
          lifespan and thorax length.
220
221  #Prediction of (average) lifespan (y) when thorax length (x) = 0.8
222  new_fly <- data.frame(thorax =0.8)
223  predicintervals <- predict(confintfunc , newdata=new_fly , interval="
          prediction", level =0.95)
224  predicintervals
225  # The predicted range for an individual fruitfly's lifespan when thorax
          length = 0.8 is 27.375 to 81.454 days. The point estimate/expected
          value is 54.414 days.
226  confintervals <- predict(confintfunc , newdata=new_fly , interval="
          confidence", level =0.95)
227  confintervals
228  # The predicted range for the average lifepsan of fruitflies when thorax
          length = 0.8 is 51.919 to 56.910 days. The point estimate/expected
          value is 54.414 days.
229
230
231
232
233  confintfunc
234  confseq <- predict(confintfunc , newdata=fly , interval = "confidence")
235  confseq
236  predicseq <- predict(confintfunc , newdata=fly , interval = "prediction")
237  predicseq
238  plot(fly$thorax , fly$lifespan ,
239       xlab="thorax length (mm)", ylab="lifespan (days)", main="Fruitfly
          lifespan (days) vs. thorax (mm)")
240  Confidence <- matlines(fly$thorax , confseq[,c("lwr","upr")], col="red")
241  Prediction <- matlines(fly$thorax , predicseq[,c("lwr","upr")], col="blue
          ")
242  fitted.values = predicseq[1:125]
243  Fitted <- lines(fly$thorax[1:125], fitted.values[1:125], col="orange",
          lwd=1)
244  legend(x=0.64, y=100, legend=c("Confidence", "Prediction", "Fitted"), cex
          =0.6, col=c("red", "blue", "orange"), lty=1)
245
246
247
248  #Questions:
249
250  #Q2: can binary variables be used for bivariate regression? It seems like
          we need to use
251  #the female variable to address the researchers' hyoptheses
252  #Q3: running predict function --> matrix, not a specific expected value
          or prediction interval
253  #Q3: difference between predicting average value vs value for an
```

```
      individual?
254 #
255 #
256 #
257 #
258 #
```