# 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 $\chi^2$ 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 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 |

- The chi-squared test statistic is 3.791168.

```
#Calculate chi-squared test statistic by hand
copbribe <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, byrow=T)
copbribe
grandtotal <- sum(copbribe)
grandtotal
upperclassrowsum <- sum(copbribe[1,])
upperclassrowsum
lowerclassrowsum <- sum(copbribe[2,])
lowerclassrowsum
notstoppedcsum <- sum(copbribe[,1])
briberequestedcsum <- sum(copbribe[,2])
warningcsum <- sum(copbribe[,3])
e14 <- ((upperclassrowsum*notstoppedcsum)/grandtotal)
e6 <- ((upperclassrowsum*briberequestedcsum)/grandtotal)
e7a <- ((upperclassrowsum*warningcsum)/grandtotal)
e7b <- ((lowerclassrowsum*notstoppedcsum)/grandtotal)
e7c <- ((lowerclassrowsum*briberequestedcsum)/grandtotal)
e1 <- ((lowerclassrowsum*warningcsum)/grandtotal)
copbribeexpected <- matrix(c(e14, e6, e7a,e7b, e7c,e1), nrow=2, byrow=T)
copbribeexpected
unsummedchisq <- ((copbribe-copbribeexpected)^2)/copbribeexpected
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.

```
#Calculating p-value
#df = (3-1)(2-1) = 2
pchisq(teststat, df=2, lower.tail=FALSE)
#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 |

```
#Calculating standardized residuals
r_unsquared <- (copbribe-copbribeexpected)
r_unsquared
upperrowprop <- (1-(upperclassrowsum/grandtotal))
lowerrowprop <- (1-(lowerclassrowsum/grandtotal))
notstoppedcprop <- (1-(notstoppedcsum/grandtotal))
bribecprop <- (1-(briberequestedcsum)/grandtotal)
warningcprop <- (1-(warningcsum/grandtotal))
sr_e14 <- r_unsquared[1,1]/sqrt(e14*upperrowprop*notstoppedcprop)
sr_e6 <- r_unsquared[1,2]/sqrt(e6*upperrowprop*bribecprop)
r_unsquared[1,2]
sr_e7a <- r_unsquared[1,3]/sqrt(e7a*upperrowprop*warningcprop)
sr_e7b <- r_unsquared[2,1]/sqrt(e7b*lowerrowprop*notstoppedcprop)
sr_e7c <- r_unsquared[2,2]/sqrt(e7c*lowerrowprop*bribecprop)
sr_e1 <- r_unsquared[2,3]/sqrt(e1*lowerrowprop*warningcprop)
stan_resid <- matrix(c(sr_e14, sr_e6, sr_e7a, sr_e7b, sr_e7c, sr_e1),
    nrow=2, byrow=T)
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

```
#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.

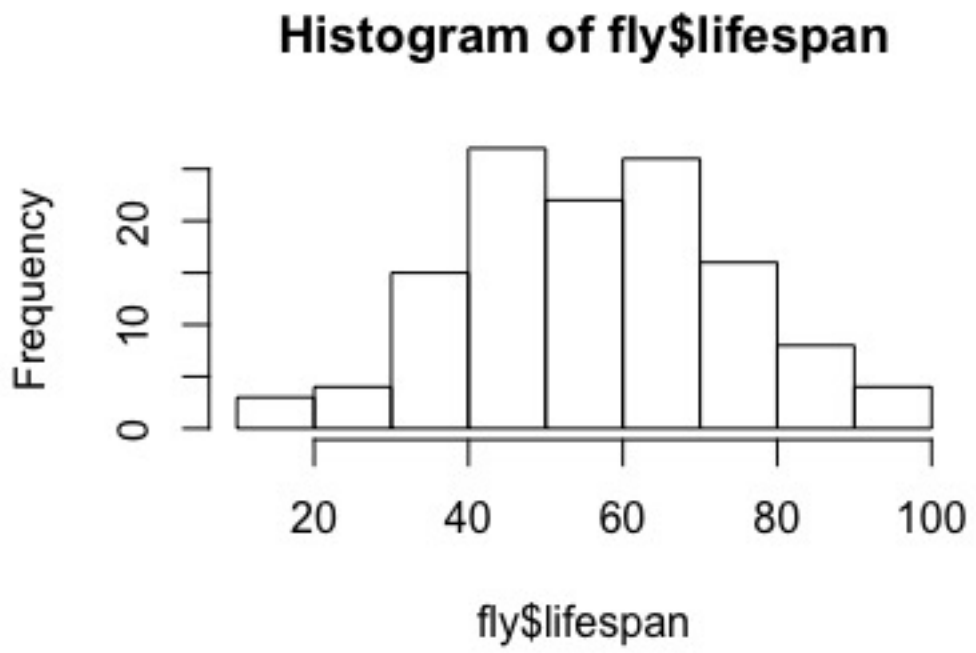
# Question 3 (40 points): Biology

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<br>1 = no females<br>2 = 1 newly pregnant female<br>3 = 8 newly pregnant females<br>4 = 1 virgin female<br>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).




–

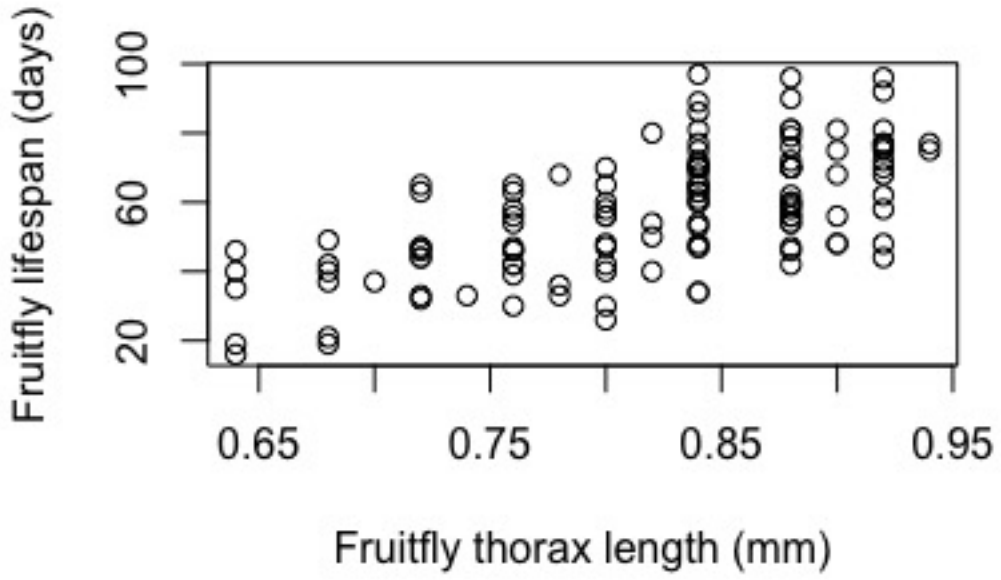
- Summary statistics for fruitfly lifespan:

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

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




*

```
#Plot thorax vs. lifespan
plot(fly$thorax, fly$lifespan,
     xlab="thorax length (mm)", ylab="lifespan (days)")
#There appears to be a linear relationship between fruitfly thorax
    and lifespan length.
#r of thorax vs. lifespan
r2 <- (cov(fly$thorax, fly$lifespan))/(sd(fly$thorax)*sd(fly$lifespan
    ))
r2
#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

```r
#Regress thorax on lifespan. Interpret the slope of the fitted model.
mean_y <- mean(fly$lifespan) #y-bar = 57.44
mean_y
mean_x <- mean(fly$thorax) #x-bar = 0.821
sigmalifespan_y <- sum(fly$lifespan) # = 7180
sigmalifespan_y
sigmathorax_x <- sum(fly$thorax) # = 102.62
sigmathorax_x
yybar <- (fly$lifespan-mean_y)
xxbar <- (fly$thorax-mean_x)
xxyy <- (xxbar*yybar)
xxyy
sum(xxyy) # = 107.367
sqxxbar <- (xxbar)^2
sqxxbar2 <- sum(sqxxbar)
sqxxbar2 # = 0.744
beta <- (sum(xxyy))/sqxxbar2
beta # = 144.33
alpha <- mean_y - (beta*mean_x)
alpha # = -61.0517
lm(fly$lifespan ~ fly$thorax, data=fly)
#linear model: y = -61.05 + 144.33x
#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.

```r
#Test for a significant linear relationship between thorax and lifespan
n_fly <- dim(fly)[1]
fly_teststat<- (r2*sqrt(n_fly-2))/sqrt(1-(r2)^2)
2*pt(fly_teststat, n_fly-2, lower.tail=F)
#p-value = ~ 0
cor.test(fly$lifespan, fly$thorax)
# 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.

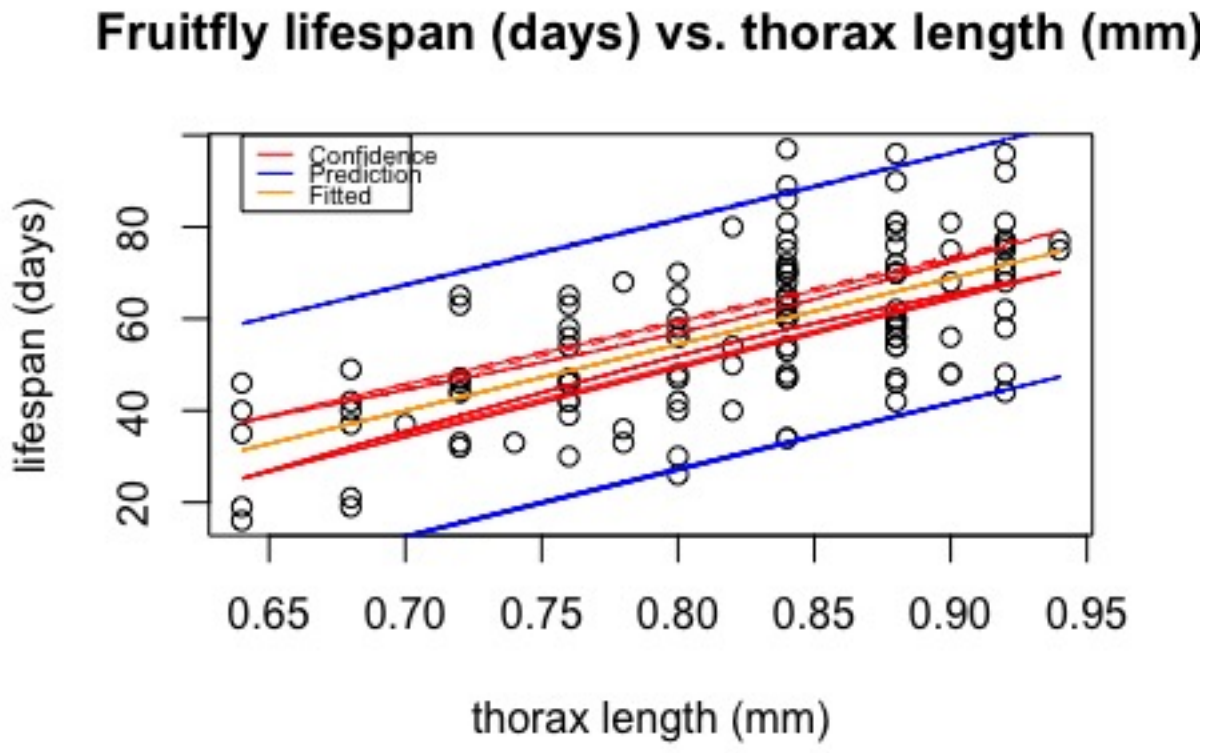
```
#90% CI for slope
#Calc critical value
a <- (1-(90/100))
critprob <- 1 - (a/2) # = 0.95
df <- n_fly - 2 # = 123
#critical value = 1.645
critvalue <- 1.645
mod = lm(fly$lifespan ~ fly$thorax, data=fly)
summary(mod)
#standard error = 15.77
moe2 <- critvalue*15.77
lower_90b <- beta - moe2
upper_90b <- beta + moe2
ci90b <- c(lower_90b, upper_90b)
ci90b
# 90% confidence interval is (118.3915, 170.2748)
confintfunc <- lm(lifespan ~ thorax, data=fly)
confint(confintfunc, level=0.90)
# 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.

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

```r
# 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.
confintervals <- predict(confintfunc, newdata=new_fly, interval="confidence", level=0.95)
confintervals
# 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.




–

```r
confintfunc
confseq <- predict(confintfunc, newdata=fly, interval = "confidence")
confseq
predicseq <- predict(confintfunc, newdata=fly, interval = "prediction")
predicseq
plot(fly$thorax, fly$lifespan,
     xlab="thorax length (mm)", ylab="lifespan (days)", main="Fruitfly lifespan (days) vs. thorax (mm)")
```

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

- R Code:

```
#########################
# load libraries
# set wd
# clear global .envir
#########################

# remove objects
rm(list=ls())
# detach all libraries
detachAllPackages <- function() {
  basic.packages <- c("package:stats", "package:graphics", "package:grDevices", "package:utils", "package:datasets", "package:methods", "package:base")
  package.list <- search()[ifelse(unlist(gregexpr("package:", search()))==1, TRUE, FALSE)]
  package.list <- setdiff(package.list, basic.packages)
  if (length(package.list)>0)  for (package in package.list) detach(package, character.only=TRUE)
}
detachAllPackages()

# load libraries
pkgTest <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[, "Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}

# here is where you load any necessary packages
# ex: stringr
# lapply(c("stringr"), pkgTest)

lapply(c("stringr"), pkgTest)

# set working directory
setwd("~/GitHub/QTM200Spring2020/problem_sets/PS1")


#########################
```

```
# Problem 1
######################

#Calculate chi-squared test statistic by hand
copbribe <- matrix(c(14, 6, 7, 7, 7, 1), nrow=2, byrow=T)
copbribe
grandtotal <- sum(copbribe)
grandtotal
upperclassrowsum <- sum(copbribe[1,])
upperclassrowsum
lowerclassrowsum <- sum(copbribe[2,])
lowerclassrowsum
notstoppedcsum <- sum(copbribe[,1])
briberequestedcsum <- sum(copbribe[,2])
warningcsum <- sum(copbribe[,3])
e14 <- ((upperclassrowsum*notstoppedcsum)/grandtotal)
e6 <- ((upperclassrowsum*briberequestedcsum)/grandtotal)
e7a <- ((upperclassrowsum*warningcsum)/grandtotal)
e7b <- ((lowerclassrowsum*notstoppedcsum)/grandtotal)
e7c <- ((lowerclassrowsum*briberequestedcsum)/grandtotal)
e1 <- ((lowerclassrowsum*warningcsum)/grandtotal)
copbribeexpected <- matrix(c(e14, e6, e7a,e7b, e7c,e1), nrow=2, byrow=T)
copbribeexpected
unsummedchisq <- ((copbribe-copbribeexpected)^2)/copbribeexpected
unsummedchisq
teststat <- sum(unsummedchisq)
teststat
#chi-squared test statistic is 3.791168

#Calculating p-value
#df = (3-1)(2-1) = 2
pchisq(teststat, df=2, lower.tail=FALSE)
#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.

#Calculating standardized residuals
r_unsquared <- (copbribe-copbribeexpected)
r_unsquared
upperrowprop <- (1-(upperclassrowsum/grandtotal))
lowerrowprop <- (1-(lowerclassrowsum/grandtotal))
notstoppedcprop <- (1-(notstoppedcsum/grandtotal))
bribecprop <- (1-(briberequestedcsum)/grandtotal)
warningcprop <- (1-(warningcsum/grandtotal))
sr_e14 <- r_unsquared[1,1]/sqrt(e14*upperrowprop*notstoppedcprop)
sr_e6 <- r_unsquared[1,2]/sqrt(e6*upperrowprop*bribecprop)
r_unsquared[1,2]
sr_e7a <- r_unsquared[1,3]/sqrt(e7a*upperrowprop*warningcprop)
sr_e7b <- r_unsquared[2,1]/sqrt(e7b*lowerrowprop*notstoppedcprop)
sr_e7c <- r_unsquared[2,2]/sqrt(e7c*lowerrowprop*bribecprop)
sr_e1 <- r_unsquared[2,3]/sqrt(e1*lowerrowprop*warningcprop)
```

```
stan_resid <- matrix(c(sr_e14, sr_e6, sr_e7a, sr_e7b, sr_e7c, sr_e1),
    nrow=2, byrow=T)
stan_resid

#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.

######################
# Problem 2
######################
women <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/
    master/PREDICTION/women.csv")
women
#Null and alternative hypotheses
#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

#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.

#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.




#######################
# Problem 3
#######################
#Set wd and load fruitfly dataset
setwd("~/GitHub/QTM200Spring2020/problem_sets/PS2") #copy/paste from
    getwd
getwd() #double check
library(readr)
fly <- read.csv("fruitfly.csv")
summary(fly$lifespan)
hist(fly$lifespan)
#The distribution of overall fruitfly lifespan appears to be bimodal and
    approximately normal.

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

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

```
mean_y <- mean(fly$lifespan) #y-bar = 57.44
mean_y
mean_x <- mean(fly$thorax) #x-bar = 0.821
sigmalifespan_y <- sum(fly$lifespan) # = 7180
sigmalifespan_y
sigmathorax_x <- sum(fly$thorax) # = 102.62
sigmathorax_x
yybar <- (fly$lifespan-mean_y)
xxbar <- (fly$thorax-mean_x)
xxyy <- (xxbar*yybar)
xxyy
sum(xxyy) # = 107.367
sqxxbar <- (xxbar)^2
sqxxbar2 <- sum(sqxxbar)
sqxxbar2 # = 0.744
beta <- (sum(xxyy))/sqxxbar2
beta # = 144.33
alpha <- mean_y - (beta*mean_x)
alpha # = -61.0517
lm(fly$lifespan ~ fly$thorax, data=fly)
#linear model: y = -61.05 + 144.33x
#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 thorax and lifespan
n_fly <- dim(fly)[1]
fly_teststat<- (r2*sqrt(n_fly-2))/sqrt(1-(r2)^2)
2*pt(fly_teststat, n_fly-2, lower.tail=F)
#p-value = ~ 0
cor.test(fly$lifespan, fly$thorax)
# p ~ 0 , therefore there is a statistically reliable and significant
    linear relationship between fruitfly thorax and lifespan length.


#90% CI for slope
#Calc critical value
a <- (1-(90/100))
critprob <- 1 - (a/2) # = 0.95
df <- n_fly - 2 # = 123
#critical value = 1.645
critvalue <- 1.645
mod = lm(fly$lifespan ~ fly$thorax, data=fly)
summary(mod)
#standard error = 15.77
moe2 <- critvalue*15.77
lower_90b <- beta - moe2
upper_90b <- beta + moe2
ci90b <- c(lower_90b, upper_90b)
ci90b
# 90% confidence interval is (118.3915, 170.2748)
confintfunc <- lm(lifespan ~ thorax, data=fly)
```

```
confint(confintfunc, level=0.90)
# 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.

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




confintfunc
confseq <- predict(confintfunc, newdata=fly, interval = "confidence")
confseq
predicseq <- predict(confintfunc, newdata=fly, interval = "prediction")
predicseq
plot(fly$thorax, fly$lifespan,
     xlab="thorax length (mm)", ylab="lifespan (days)", main="Fruitfly lifespan (days) vs. thorax (mm)")
Confidence <- matlines(fly$thorax, confseq[,c("lwr","upr")], col="red")
Prediction <- matlines(fly$thorax, predicseq[,c("lwr","upr")], col="blue")
fitted.values = predicseq[1:125]
Fitted <- lines(fly$thorax[1:125], fitted.values[1:125], col="orange", lwd=1)
legend(x=0.64, y=100, legend=c("Confidence", "Prediction", "Fitted"), cex=0.6, col=c("red", "blue", "orange"), lty=1)



#Questions:

#Q2: can binary variables be used for bivariate regression? It seems like we need to use
#the female variable to address the researchers' hyoptheses
#Q3: running predict function --> matrix, not a specific expected value or prediction interval
#Q3: difference between predicting average value vs value for an
```

```
    individual?
#
#
#
#
#
```