Homework 1 - BST682

Assigned: August 31, 2018

Contents

Homework 1 Overview

This homework is intended to serve three main purposes: (1) familiarize you with our course homework submission policy (while satisfying the new Title IV regulations), (2) refresh your probability and linear modeling skills, and (3) introduce you to R. Give complete solutions, justifying your response when necessary.

This homework is due by **9:30am on Tuesday, September 11**. To complete this assignment, follow these steps:

1. Answer the questions below in a format for which you are comfortable (e.g., LaTeX, `R`, Word, paper, etc.).
2. Convert this work to a pdf and name it lastname1.pdf (replacing 'lastname' with your last name in lowercase).
3. For any questions that require programming, provide a similarly-named file that includes fully-reproducible code. (e.g., lastname1.r, lastname1.rmd or lastname1.sas)
4. Submit these files to Canvas.

Problem 1: probablity refresher 1

Uber, AirBnb and Stata have 3000, 1500, and 800 employees, respectively, and 30, 45, and 65 percent of these employees respectively are women. Resignations are equally likely among companies and genders. One woman resigns. What is the probability she worked for Uber?

```
setwd('C:/Users/court/Documents/BST682_GLM/Homework1')

#Uber  -3000 (30% women)
#AirBnb-1500 (45% women)
#Stata -800  (65% women)

denominator <- (3000*(.3)+ 1500*(.45)+800*(.65))
denominator
```

```
## [1] 2095
```

```
numerator <-3000*(.3)
numerator
```

```
## [1] 900
```

```
probability <-(numerator/denominator)
probability
```

```
## [1] 0.4295943
```

```
#The probability that the female employee that resigns will be a woman is .43
```

Problem 2: probability refresher 2

You flip four fair coins. Assuming the flips are independent, what is the pmf for the number of tails flipped?

```
#Let x equal the number of flips (x=4)
#Let k equal the number of tails (range is 1 for heads 2 for tails)
#Probability of getting tails per flip=.5.
```

$$P(X = k) = \frac{\binom{x}{k}}{2^x}$$

$$P(X = k) = \frac{\binom{4}{1}}{2^4}$$

Problem 3: probability refresher 3

Do problem 1.6 (a,b) from our text.(Page 17)

```
table16<-read_xlsx('table16.xlsx')
View(table16)
```

```
####1.6a - create the proportion
table16$prop_ftom <- (table16$females/ (table16$females+table16$males))
View(table16)
```

```
####1.6b -  MLE

#I was having issues with some of the formulas - they are  commented out below because otherwise itwoul
```

$$Y \sim Bin(n, \theta) with$$

$$\theta \in \theta(0, 1)$$

The pmf is

$$f(y; \theta)$$

$$= (nCy)\theta^y(1-\theta)^{n-y}$$

for y=0...16

$$\frac{df\theta}{d\theta} = \frac{d\theta^y(1-\theta)^{n-y}}{d\theta}$$

$$= (n-y)(1-\theta)^{n-y-1}(-1)\theta^y + y(1-\theta)^{n-y}\theta^{y-1}$$

$$= (-(n-y)\theta + y(1-\theta))(1-\theta)^{n-y-1}\theta^{y-1}$$

And 0<θ<1 when 0 < y < n

$$f(0) = 0$$

$$f(1) = 1$$

and

$$L(\frac{y}{n}) > 0$$

so that MLE= y/n Therefore, the MLE is

$$\hat{\theta} = \sum \frac{Y_i}{n_i} = 0.49$$

(Got help from Shama)

Problem 4: probability refresher 4

Assume annual rainfall in Lexington is normally distributed with a mean of 40 inches and standard deviation of 4. What is the probability that it takes more than 7 years before having a rainfall over 55 inches? What assumptions are you making?

```
#1) That the data is normally distributed
#2) Year to year, the rainfall is independent

#Let X denote the annual rainfall in any given year
```

$Z = \frac{\bar{X}-\mu}{\theta}$

```
pnormGC(55, region="below", mean=40,
        sd=4,graph=FALSE)
```

```
## [1] 0.9999116
```

```
#then we we want to take the 7 years...
```

$P(X > 7) = (0.999)^7$

```
p_x<-(0.999)^7
p_x
```

```
## [1] 0.993021
```

```
#The probability that it takes more than 7 years for the rainfall will b eover 55   inches is 99%
```

Problem 5: linear models refresher

Using the data from `Table 2.3 Birthweight and gestational age.xls`, calculate by matrix algebra the effect estimate resulting from regressing birth weight on gestational age.

```
table23<-read_xls('table23.xls')
View(table23)
```

```
ols <- function(XX, yy) {
missing.data <- apply(is.na(XX), 1, any) | is.na(yy)
X <- cbind(Intercept=1, as.matrix(XX[!missing.data,]))
y <- yy[!missing.data]
solve(crossprod(X)) %*% t(X) %*% y
```

```
}
ols(table23[,-2], table23$bweight)
```

```
##                  [,1]
## Intercept -1447.2432
## gestage     120.8943
## sex        -163.0393
```

```
#The Intercept is -1.447.2 and the beta coefficient for gestage is 120.9
```

Problem 6: R intro 1

You will inevitably use the Google to problem solve with programming in R – many of you already do. Having go to resources for answering your questions and/or developing new skills can be quite helpful. Search around for what might be (or already is) a resource you will turn to as you improve your R skills. Give the site and url. What, in particular, makes this suitable for you?

```
#I have used this a ton to get used to the nomenclature in r as well as practice a bit -
#  http://homerhanumat.github.io/tigerstats/instructorNotes.html#shiny_apps

#I also really enjoy Andy Field -
 # I've used this book for class (and I actually enjoy reading it)
#https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-r/book236067
```

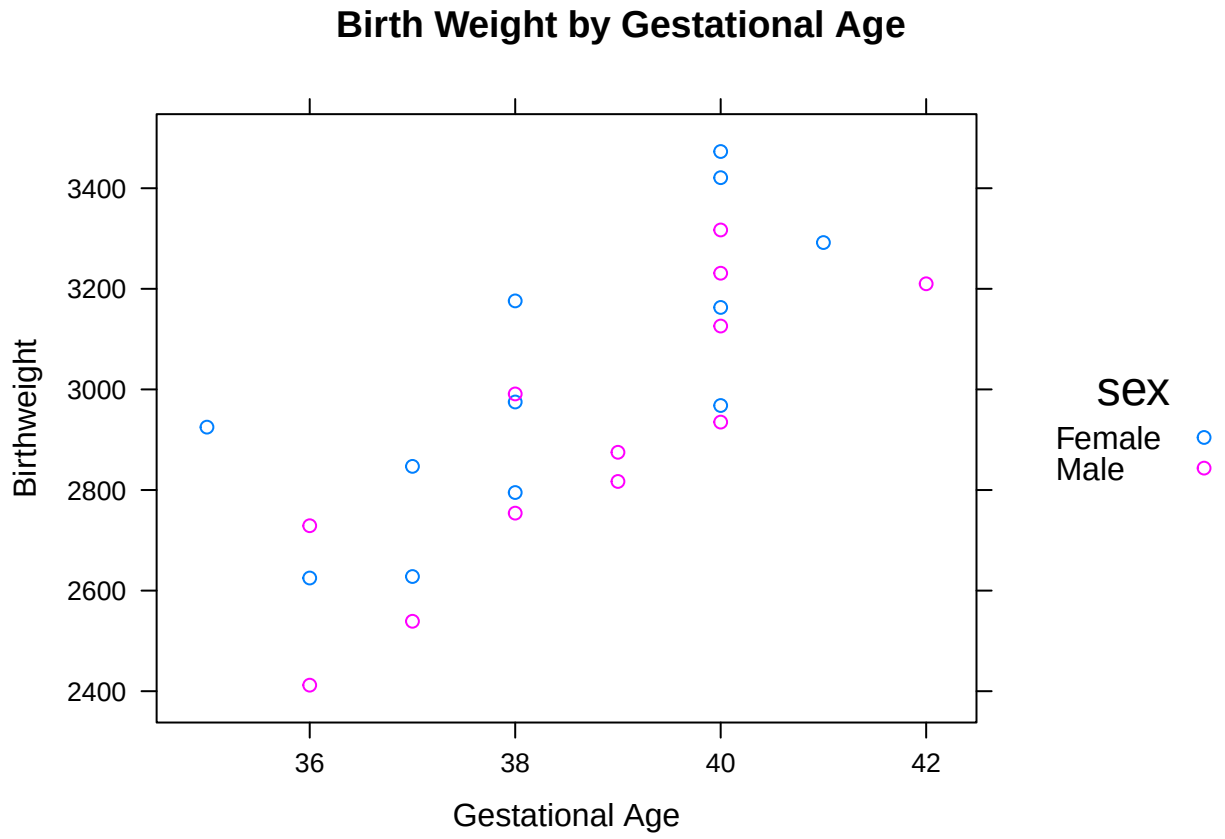
Problem 7: R intro 2

Import the data from `Table 2.3 Birthweight and gestational age.xls` into R. Each observation should be a single row. *Tip*: I added a second sheet to make this easier if you prefer. Use the `Import Dataset` functionality in RStudio's `Environment` tab and select Sheet 2. This simple example shows why some abhor Excel... *Tip 2*: Use the `readxl` package.

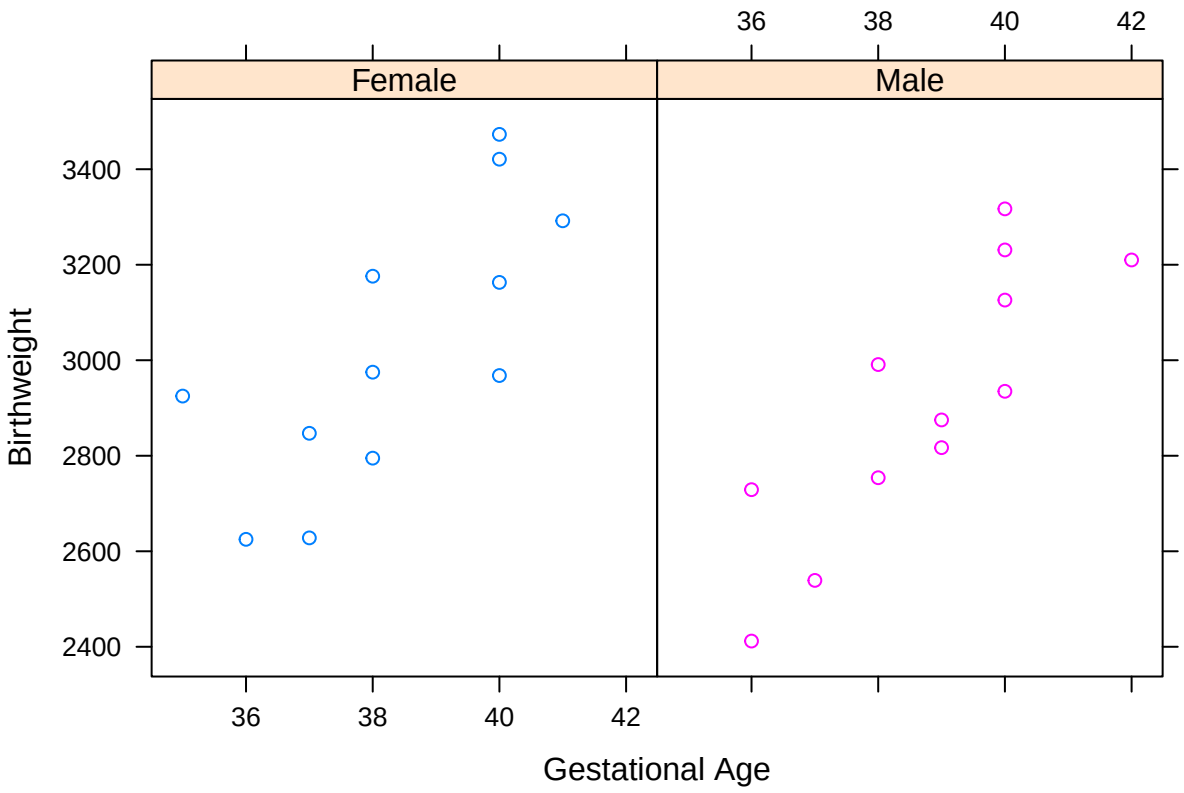
```
#Used the import function to get table 2.3 in.
View(table23)
attach(table23)
table23$sex[sex=='1']<-"Female"
table23$sex[sex=='2']<-"Male"
```

Plot birthweight by age and give each gender a different color on the same plot. Now, do the same plot stratified by gender (Tip: look at the Introduction to R notes). What observations do you have?

```
xyplot(bweight ~ gestage,
    data = table23,
    group=sex,
    auto.key = list(
        space = "right",
        title = "sex"),
    main = "Birth Weight by Gestational Age",
    xlab = "Gestational Age",
    ylab = "Birthweight")
```




```
xyplot(bweight ~ gestage |sex,
    data = table23,
    group=sex,
    layout = c(2,1),
    xlab = "Gestational Age",
    ylab = "Birthweight")
```



Problem 8: R intro 3

Using `R` and `lm`, confirm your regression parameter estimate in Problem 5.

```
lm(bweight ~ gestage, data=table23)
```

```
##
## Call:
## lm(formula = bweight ~ gestage, data = table23)
##
## Coefficients:
## (Intercept)      gestage
##     -1485.0        115.5
```