BST 682 - Homework 1

Ashley I. Martinez, PharmD, MS

11 September 2018

Contents

Probability Refreshers

Problem 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?

Company	Women	Men	**Total**
Uber	900	2100	**3000**
AirBNB	675	825	**1500**
Stata	520	280	**800**
Total	**2905**	**3205**	**5300**

Using Baye's Theorem:

$$P(Uber|Woman) = \frac{P(Uber, Woman)}{P(Woman)}$$

Thus:

$$P(Uber|Woman) = \frac{P(Woman|Uber)\ P(Uber)}{[P(Woman|AirBNB)\ P(AirBNB)] + [P(Woman|Stata)\ P(Stata)] + [P(Woman|Uber)\ P(Uber)]}$$

Calculating in R:

```
> prob_UGivenW <- ((900/3000) * (3000/5300))/(((675/1500) * (1500/5300)) +
+     ((520/800) * (800/5300)) + ((900/3000) * (3000/5300)))
> prob_UGivenW
```

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

Problem 2.

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

Definition:
For a discrete random variable X with possible values $x_1, x_2, x_3, ..., x_n$, a probability mass function $f(x_i)$ is a function such that:

1. $f(x_i) \geq 0$
2. $\sum_{i=1}^{n} f(x_i) = 1$
3. $f(x_i) = P(X = x_i)$

A coin toss follows the binomial distribution. If we let p be the probability of tossing a tails, n be the number of tosses, and x be the number of tails, then the PMF is given by:

$$P(X = x) = \binom{n}{x} \; p^x \; (1-p)^{n-x}$$

Because the p in the case of a coin is 0.5 and in this problem, $n = 4$:

$$P(X = x) = \binom{4}{x} \; 0.5^x \; (0.5)^{4-x}$$

Problem 3.

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

Part A.

Progeny Group	Females	Males	**Proportion Female**
1	18	11	**0.62**
2	31	22	**0.58**
3	34	27	**0.56**
4	33	29	**0.53**
5	27	24	**0.53**
6	33	29	**0.53**
7	28	25	**0.53**
8	23	26	**0.47**
9	33	38	**0.46**
10	12	14	**0.46**
11	19	23	**0.45**
12	25	31	**0.45**
13	14	20	**0.41**
14	4	6	**0.4**
15	22	34	**0.39**
16	7	12	**0.37**
Total	**363**	**371**	**0.49**

Part B.

We know that for a function $f(y_i;\theta)$, its likelihood function is $L(\theta;y_i)$.

$$L(\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \theta^{y_1}(1-\theta)^{n_1-y_1} \ \times \ \theta^{y_2}(1-\theta)^{n_2-y_2} \ \times \ \ldots \ \times \ \theta^{y_i}(1-\theta)^{n_i-y_i}$$

$$L(\theta) = \theta^{\sum y_i}(1-\theta)^{n_i-\sum y_i}$$

To use calculus to solve for the maximum likelihood estimator, it is easier to use the log-likelihood function, $l(\theta;y_i) = logL(\theta;y_i)$.

Thus, it follows that:

$$logL(\theta) = (\sum y_i)log\theta \ + \ (n_i - \sum y_i)log(1-\theta)$$

Using the convenient properties of the natural log, we take the derivative and set it to zero.

$$\frac{\delta \ logL(\theta)}{\delta\theta} = \frac{\sum y_i}{\theta} - \frac{n_i - \sum y_i}{1-\theta} \ \equiv \ 0$$

Multiply through by $\frac{\theta}{1-\theta}$:

$$(\sum y_i)(1-\theta) \ - \ (n_i - \sum y_i)(\theta)$$

$$(\sum y_i) \ - \ n_i\theta \equiv 0$$

$$\hat{\theta} = \frac{\sum y_i}{n_i}$$

Then, to evaluate the maximum likelihood estimator, $\hat{\theta}$ for these data:

$$\hat{\theta} = \frac{(18+31+34+33+27+33+28+23+33+12+19+25+14+4+22+7)}{734} = 0.49$$

Problem 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?

For this question, we want to know the probability that for 7 consecutive years, the rainfall is ≤ 55".

Thus,

$$P(rainfall \leq 55")^7 \;=\; \Phi(\frac{55-40}{4})^7 \;=\; \Phi(3.75)^7 \;\approx\; (0.9999)^7$$

Assumptions

- Rainfall in different years are independent of each other
- Let X denote the annual rainfall in any given year
  - Let X be a normally distributed continuous random variable with parameters μ and σ^2
- $P(X \leq a) = \Phi(\frac{a-\mu}{\sigma})$, where $\Phi(x)$ is the cumulative distribution function

---

Linear Modeling Refresher

Problem 5.

> *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.*

First, we will import the data which is available on the publisher's website for download.

```
# read in the first worksheet from the workbook first row
# contains variable names
birthweight <- read.csv(file = "t2-3_birthweight.csv", header = TRUE,
    sep = ",")
```

Now, we need to make that worksheet into two matrices: one containing the regressor data and one containing the age.

```
## Create X and Y matrices for this specific regression.
X = as.matrix(cbind(1, birthweight$gestational.age))
Y = as.matrix(birthweight$birth.weight)
```

Then, we use matrix algebra to give us a matrix of estimated coefficients ($\hat{\beta}$)that minimizes the sum of squared residuals. We label this data so the output makes sense.

```
## Choose beta-hat to minimize the sum of squared residuals
## resulting in matrix of estimated coefficients:
bh = round(solve(t(X) %*% X) %*% t(X) %*% Y, digits = 2)

## Label and organize results into a data frame
beta.hat = as.data.frame(cbind(c("Intercept", "Age"), bh))
names(beta.hat) = c("Coeff.", "Est")
beta.hat
```

```
##      Coeff.      Est
## 1 Intercept -1484.98
## 2       Age   115.53
```

We can now use the built-in R function, $lm()$ to verify that we have obtained a correct effect estimate.

```
lm(birth.weight ~ gestational.age, birthweight)
```

```
##
## Call:
## lm(formula = birth.weight ~ gestational.age, data = birthweight)
##
## Coefficients:
##     (Intercept)  gestational.age
##         -1485.0            115.5
```

R Introduction

Problem 6.

> *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've definitely been using the Internet to help me with this homework assignment, as this is the first time I've ever used R. What I've been doing is entering a query into Google, like "rmarkdown insert line break," and then I'll usually click through the first few top results. I've been finding myself on StackOverflow quite a bit (https://stackoverflow.com/questions/tagged/r). I like that because I've often found that people link to more important resources in the comments.

I also like the "Quick-R" resource from StatsMethods (https://statsmethods.net). They have different topics divided into separate sections, each with detailed examples.

Problem 7.

> *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. 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?*

I downloaded the file from the publisher's website (https://www.crcpress.com/An-Introduction-to-Generalized-Linear-Models/Dobson-Barnett/p/book/9781138741515), since I couldn't find it on Canvas anywhere. It was formatted quite strangely, with multiple headers and really two tables in the first sheet. So, I used the *readxl* package to specify that I only wanted the latter 3 columns that had the actual information in them, using the following code:

```
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)

library(readxl)
gestationalage <- read_excel("/Users/ashleymartinez/Dropbox/UK/PhD/Terms/Autumn 2018/BST682/Data/t2-3_b
    range = cell_cols("G:I"))
```

But then, of course, I still had problems. The first was that it actually imported as a tibble instead of a data frame, so I had to first convert it to a data frame. Additionally, the column headers had spaces in them, which I knew would cause me trouble, so I went ahead and renamed the columns, too (along with the values for sex).

```
# Look at the imported data
head(gestationalage)
```

```
## # A tibble: 6 x 3
##   `gestational age` `birth weight`   sex
##               <dbl>          <dbl> <dbl>
## 1                40           2968     1
## 2                38           2795     1
## 3                40           3163     1
## 4                35           2925     1
## 5                36           2625     1
## 6                37           2847     1
```

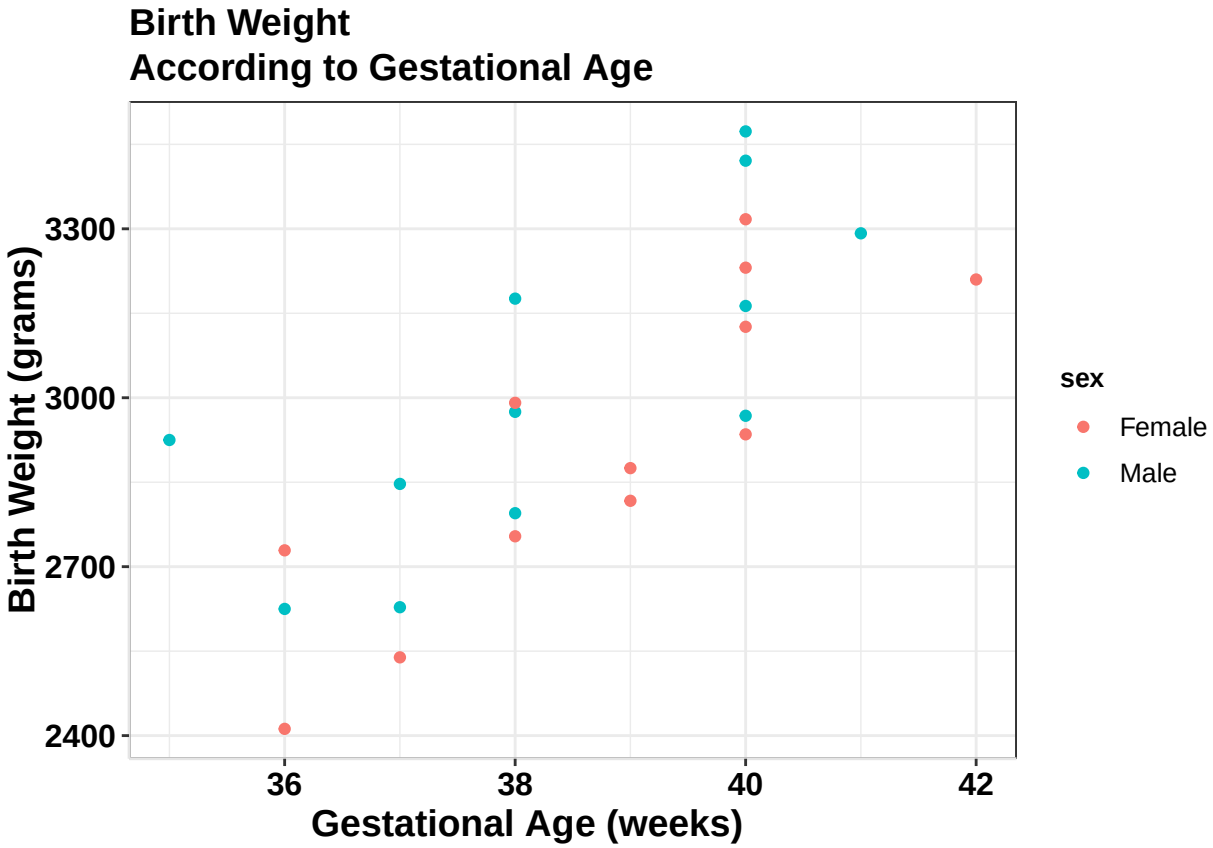
```
# Change it into a dataframe
df <- as.data.frame(gestationalage)

# Rename the columns
names(df) = c("age", "weight", "sex")

# Replace the numbers in sex with logical values
df$sex[df$sex == 1] <- "Male"
df$sex[df$sex == 2] <- "Female"
```

Then, I use the function *ggplot*2 to create a plot that would show the relationship between gestational age and weight by sex.

```
library(easyGgplot2)
```

```
## Loading required package: ggplot2
```

```
weight_age <- ggplot2.scatterplot(data = df, xName = "age", yName = "weight",
    groupName = "sex")
ggplot2.customize(weight_age, backgroundColor = "white", mainTitle = "Birth Weight\nAccording to Gestati
    xtitle = "Gestational Age (weeks)", ytitle = "Birth Weight (grams)")
```



I can also stratify by sex in two different plots.

```
# First, subset my data into males and female
attach(df)
```

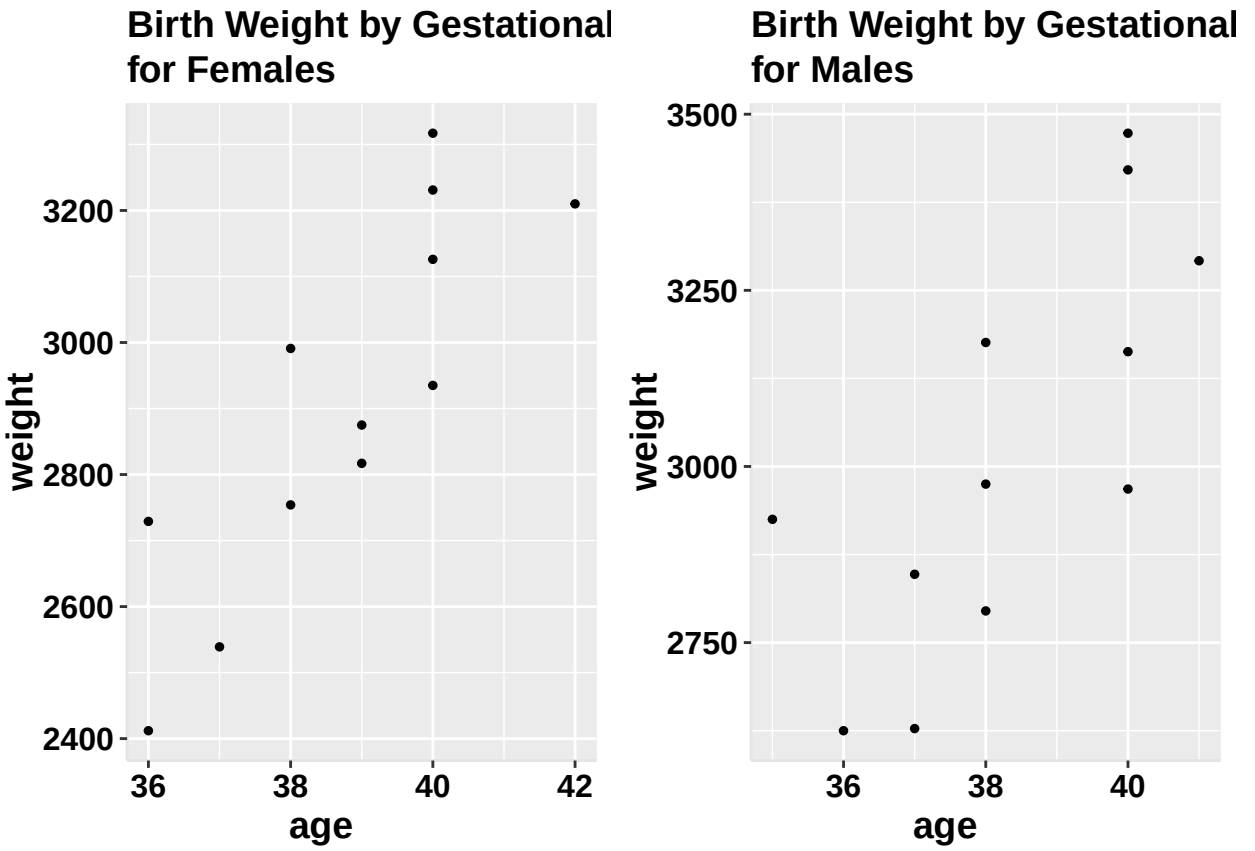
```
## The following object is masked from package:easyGgplot2:
##
##     weight
```

```
df_female <- df[which(sex == "Female"), ]
detach(df)

attach(df)
```

```
## The following object is masked from package:easyGgplot2:
##
##     weight
```

```
df_male <- df[which(sex == "Male"), ]
detach(df)

# Now plot them on separate plots using the multiplot
# function from ggplot2
library(easyGgplot2)
weight_age_female <- ggplot2.scatterplot(data = df_female, xName = "age",
    yName = "weight", mainTitle = "Birth Weight by Gestational Age\nfor Females")
weight_age_male <- ggplot2.scatterplot(data = df_male, xName = "age",
    yName = "weight", mainTitle = "Birth Weight by Gestational Age\nfor Males")
weight_age_stratify <- ggplot2.multiplot(weight_age_female, weight_age_male,
```

```
    cols = 2)
```



Problem 8.

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

I already did this; see problem #5.
