StatR 502 Homework 7

Rebecca Hadi

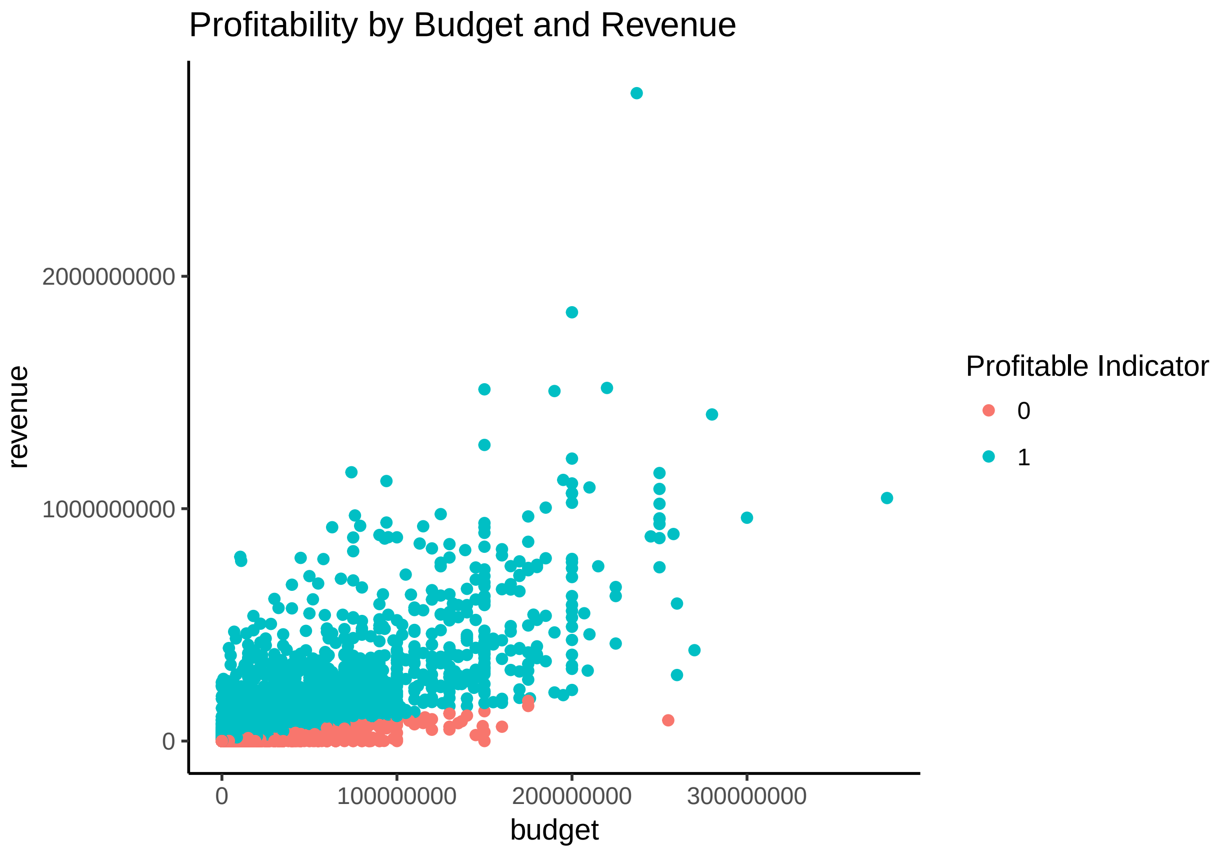
Due Thursday, Feb. 22, 2018 at 6:30 pm

Submission guidelines: please submit a PDF if possible, otherwise a self contained HTML file, and optionally your `.Rmd` file. As always, ask in the discussion forum if you're having trouble!

1. Nice plot

Create a nice, polished visualization using the data for your final project. *Really* polish your figure: pick out some nice colors, make sure labels are clear, and set it up so the plot tells a little story or highlights an interesting comparison. **A good figure should be able to be understood without referring to the accompanying text.**

```
#load data
mdata <- read.csv("processed_movies.csv")
library(ggplot2)

#create plot
options(scipen=10)
ggplot(data = mdata, aes(x = budget, y = revenue, col = as.factor(profitable))) +
                     geom_point() +
                     theme_classic() +
                     labs(col = "Profitable Indicator") +
                     ggtitle("Profitability by Budget and Revenue")
```



Book problems

Do G&H Chapter 7 problems **1, 2 and 4** (pp. 152). For number 1, also do part (d) below:

G & H 1

Discrete probability simulation: suppose that a basketball player has a 60% chance of making a shot, and he keeps taking shots until he misses two in a row. Also assume his shots are independent (so that each shot has 60% probability of success, no matter what happened before).

(a) Write an R function to simulate this process.

```
i = 2 #Start with 2 shots no matter what
shot = NA  #placeholder for shot
shot[1] <- rbinom(1,1,0.6) #first shot
shot[2] <- rbinom(1,1,0.6) #second shot
while (sum(shot) >= 1) { #only enter loop if at least one shot was a success
  shot[i] <- rbinom(1,1,0.6) #next shot
  if(shot[i] == 0 & shot[i - 1] == 0) break  #if next shot and prev shot were miss, stop loop
  i <- i + 1 #if shot was sucess, go to next shot
  }
print(shot) #test output
```

```
## [1] 0 0
```

(b) Put the R function in a loop to simulate the process 1000 times. Use the simulation to estimate the mean, standard deviation, and distribution of the total number of shots that the player will take.

```r
library(dplyr)

n.sims <- 1000 #1000 sims

basketball.succ <- rep(NA,n.sims)
basketball.sim <- rep(NA, n.sims)
for (s in 1:n.sims) {
i = 2 #Start with 2 shots no matter what
shot = NA  #placeholder for shot
shot[1] <- rbinom(n.sims,1,0.6) #first shot
shot[2] <- rbinom(n.sims,1,0.6) #second shot
  while (sum(shot) >= 1) { #if first two shots were both successful, enter while loop
  shot[i] <- rbinom(1,1,0.6) #next shot
  if(shot[i] == 0 & shot[i - 1] == 0) break  #if next shot and prev shot were miss, stop loop
  i <- i + 1 #if shot was sucess, go to next shot
  }
basketball.sim[s] <- length(shot) #how many shots it took
basketball.succ[s] <- sum(shot) #how many are successes

}

#combine
sim.succ <- cbind(basketball.sim, basketball.succ)

#Calculate the mean
mean(basketball.sim)
```

```
## [1] 7.434
```

```r
#Calculate the standard deviation
sd(basketball.sim)
```

```
## [1] 6.668141
```

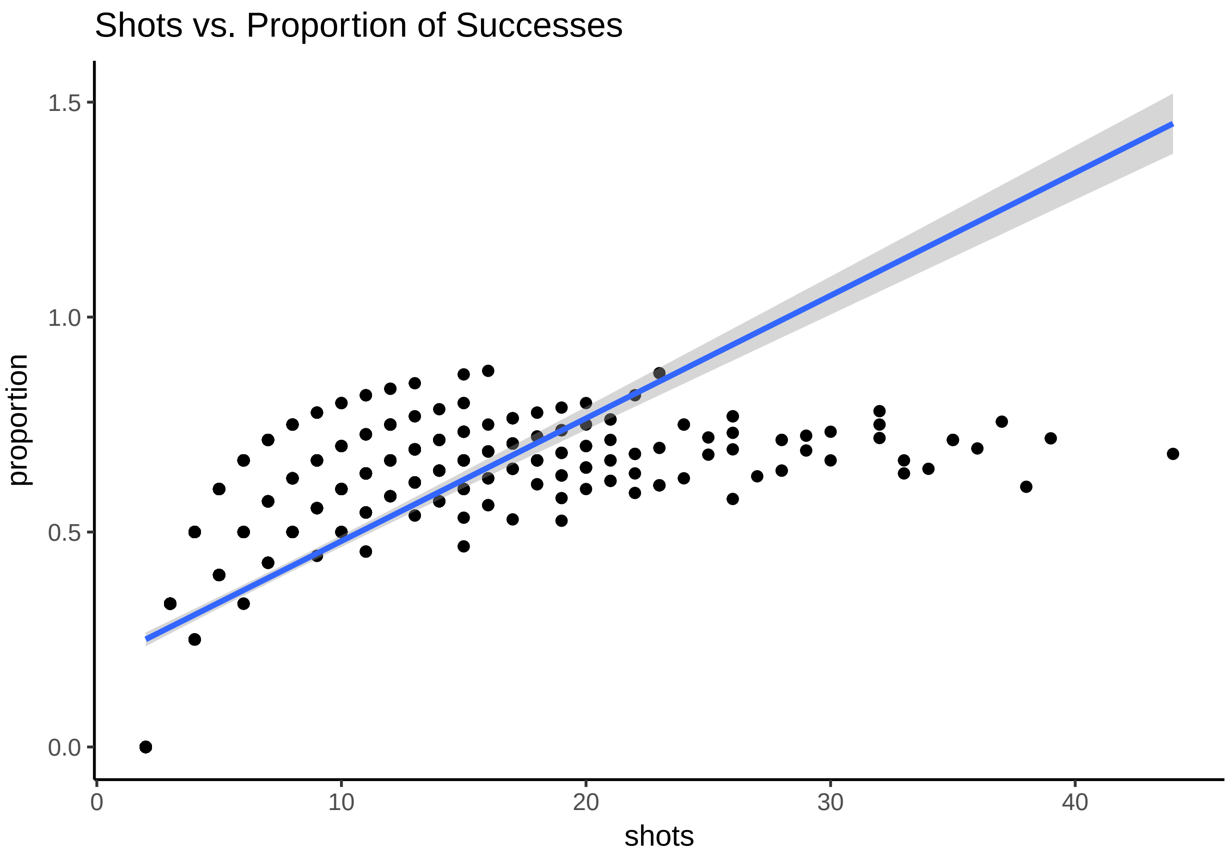
```r
#distribution of shots
summary(basketball.sim)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   2.000   5.000   7.434  10.000  44.000
```

(c) Using your simulations, make a scatterplot of the number of shots the player will take and the proportion of shots that are successes.

```r
library(ggplot2)

#convert to data frame
sim.succ <- as.data.frame(sim.succ)

#create proportion column
sim.succ$prop <- sim.succ$basketball.succ / sim.succ$basketball.sim

#Update column names
colnames(sim.succ) <- c("shots","successes","proportion")

basket.plot <- ggplot(data = sim.succ, aes(x = shots, y = proportion), ylim = c(0,1)) +
               geom_point() +
```

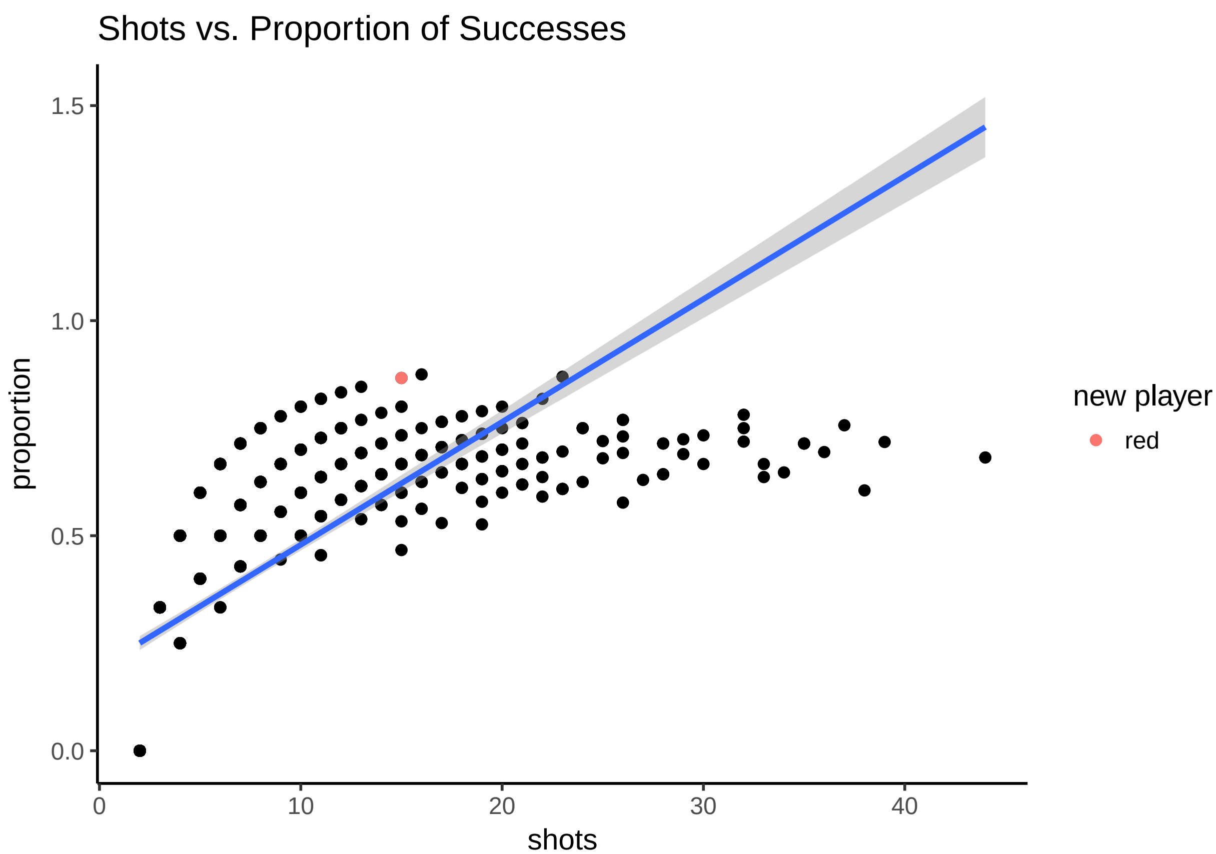
```
            theme_classic() +
            geom_smooth(method = "lm") +
            ggtitle("Shots vs. Proportion of Successes")

basket.plot
```



(d) Simulation can be used test hypotheses, even generate p values. We can consider the situation described in the problem as a *model*. Perhaps we have another basketball player and we have a null hypothesis that her shooting percentage is 60%, just like the first player. She's talking a big talk, so we have an alternative hypothesis that her shooting percentage is >60%. We test the new player, having her take shots until she misses two in a row. She takes 15 shots (i.e., 13 shots without two misses in a row, then shots 14 and 15 are both misses). Under the null model, what is the probability of taking at least 14 shots? Do you think the new player is better than the original player?

```
#bring in package
library(broom)


#add new player to plot and see how it compares to line

sim.succp2 <- as.data.frame(cbind(15,13))
colnames(sim.succp2) <- c("shots","successes")

#create proportion column
sim.succp2$prop <- sim.succp2$successes/ sim.succp2$shots

#Update column names
```

```r
colnames(sim.succp2) <- c("shots","successes","proportion")


#Add new player data to plot
basket.plot + geom_point(data = sim.succp2, aes(x = shots, y = proportion, col = "red"))  +
              labs(col = "new player")
```




```r
#Fit a linear model
basket.mod <- lm(proportion~ shots, data = sim.succ)

#predict on new data with same model
sim.succp2.predict <- predict(basket.mod, sim.succp2, se.fit = T)


#Probability of at least 14 shots   -- how many simulations had at least 14 over total simulations
sim.succ14 <- sim.succ %>%
              filter(shots >= 14)

#calculat proportion
prob <- nrow(sim.succ14) / 1000
```

The probability of taking at least 14 shots based on our simulation is 15%. Under the null model, we would predict the new player to make 60.5% with a standard error of 0.009678. Expanding the 95% CI, the player's actual proportion of shots was 86.6666667%, which is greater than the upper bound of the CI at 64.106883%.

G & H 2

Continuous probability simulation: the logarithms of weights (in pounds) of men in the United States are approximately normally distributed with mean 5.13 and standard deviation 0.17; women with mean 4.96 and standard deviation 0.20. Suppose 10 adults selected at random step on an elevator with a capacity of 1750 pounds. What is the probability that the elevator cable breaks?

```r
library(magrittr)

#create 1000 variables each
log.weight.male <- rnorm(1000,5.13,0.17)
log.weight.female <- rnorm(1000,4.96,0.20)

#convert to data frame and add col names
#female
log.weight.female %<>% as.data.frame(log.weight.female)
colnames(log.weight.female) = c("log.weight")
#male
log.weight.male %<>% as.data.frame(log.weight.male)
colnames(log.weight.male) = c("log.weight")

#add gender
log.weight.female <- log.weight.female  %>%
                      mutate(gender = "female")
log.weight.male <-   log.weight.male %>%
                     mutate(gender = "male")


#Combine into one data set
log.weight <- rbind(log.weight.female, log.weight.male)

#Simulate 1000 times

#Sample 10 adults randomly (can be any gender)
total.weight <- rep(NA,n.sims)

for (s in 1:n.sims) {
  log.weight.sample <- sample_n(log.weight,10, replace = TRUE) #take sample
   #get the total weight and convert back to original scale
   total.weight[s] <- sum(exp(log.weight.sample$log.weight))
}

#convert to data frame
total.weight %<>% as.data.frame(total.weight)
colnames(total.weight) = c("weight")

#create subset where the total weight would break the elevator
break.sims <- total.weight %>%
              filter(weight > 1750)

prob.break <- nrow(break.sims) / nrow(total.weight)
```

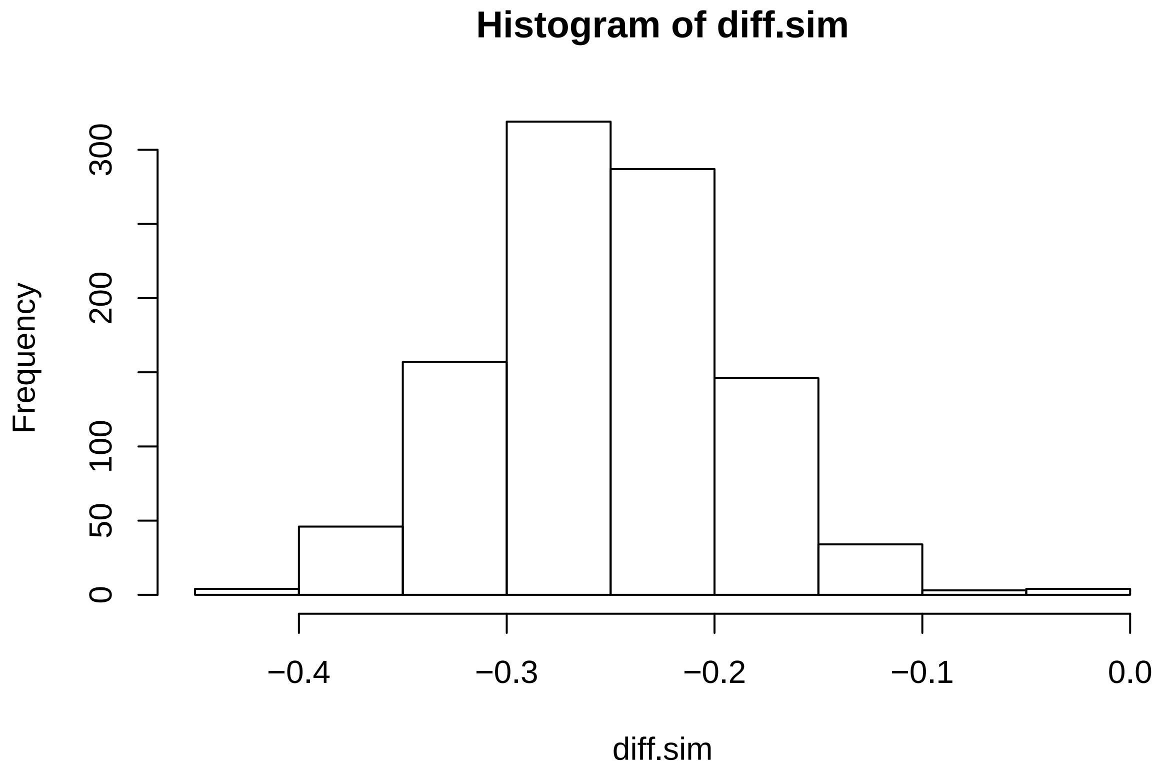
The probability that the elevator will break is 4.8 %.

G& H 4

Predictive simulation for linear regression: take one of the models from Exercise 3.5 or 4.8 that predicts course evaluations from beauty and other input variables. You will do some simulations. **(a)** Instructor A is a 50 year old woman who is a native English speaker and has a beauty score of -1. Instructor B is a 60 year old man who is a native English speaker and has a beauty score of -0.5. Simulate 1000 random draws of the course evaluation rating of these two instructors. In your simulation, account for the uncertainty in the regression parameters (that is, use the sim() function) as well as the predictive uncertainty. **(b)** Make a histogram of the difference between the course evaluations for A and B. What is the probability that A will have a higher evaluation?

```
#load data
library(AER)
library(arm)
data(TeachingRatings)

#Fit linear model
beauty.mod <- lm(eval ~ beauty + factor(gender) + factor(native) + age, data = TeachingRatings)

#Simulate model
beauty.sim <- sim(beauty.mod,1000)

#extract matrix from sim object
beauty.sim.coef <- as.data.frame(beauty.sim@coef)

#extract standard error
beauty.sim.se  <- as.data.frame(beauty.sim@sigma)
colnames(beauty.sim.se) <- c("error")



#Simulate evaluations for A & B

#Instructor A - inferential and predictive uncertainty
#50 year old woman who is a native English speaker and has a beauty score of -1
a.sim <- beauty.sim.coef$`(Intercept)` +  #intercept
                    (beauty.sim.coef$beauty * -1) +  #beauty
                    (beauty.sim.coef$`factor(gender)female` * 1) + #gender
                    (beauty.sim.coef$age * 50) +    #age
                    beauty.sim.se$error #error


#Instructor B - inferential and predictive uncertainty
#Instructor B is a 60 year old man who is a native English speaker and has a beauty score of -0.5.
b.sim <- beauty.sim.coef$`(Intercept)` +  #intercept
                    (beauty.sim.coef$beauty * -0.5) +  #beauty
                    (beauty.sim.coef$age * 60) +    #age
                    beauty.sim.se$error #error


#difference between a & B

diff.sim <- a.sim - b.sim
```

```r
#compare A & B
hist(diff.sim)
```




```r
#What's the probability A has a higher evaluation?
#proprotion of diff.sim that are greater than 0 over total diff.sim

diff.sim %<>% as.data.frame(diff.sim)
colnames(diff.sim) <- c("diff")

diff.sim.a <- diff.sim %>%
              filter(diff > 0)

prob.eval <- nrow(diff.sim.a) / nrow(diff.sim)
```

In the simulation, there are no outcomes where A has a higher evaluation than B.