Stat 3301: Homework 3

Due by date and time specified on Carmen

Nathan Johnson.9254

Setup:

```
library(alr4)
library(tidyverse)
```

Instructions

- Replace "FirstName LastName (name.n)" above with your information.
- Provide your solutions below in the spaces marked "Solution:".
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R. Use the option `echo = TRUE` to make sure the R code is displayed.
- Knit this document to HTML and upload both the HTML file and your completed Rmd file to Carmen
- Make sure your solutions are clean and easy-to-read by
  - formatting all plots to be appropriately sized, with appropriate axis labels.
  - only including R code that is necessary to answer the questions below.
  - only including R output that is necessary to answer the questions below (avoiding lengthy output).
  - providing short written answers explaining your work, and writing in complete sentences.
- Data files mentioned below are from the `alr4` package unless specified otherwise.

Concepts & Application In this assignment, you will

- write down the components and assumptions of a simple linear regression model
- identify the mean and variance functions of a SLR model
- estimate the parameters of a SLR model using summary statistics also using the `lm` function, and interpret the results
- plot an estimated simple linear regression line
- compare two fitted regression models
- use a fitted model to describe relationships between variables
- assess whether the SLR model is an appropriate model

Question 1 This question relates to the **NBA player** data we looked at in class (it is available on Carmen). Here we will focus on players whose position is **forward** (labeled `F` in the data set).

1. Make a plot of player weight vs. player height for all players whose position is **forward** . Use the plot to summarize the relationship between weight and height for forwards in the league.

2. We will use the normal simple linear regression model to relate weight to height for these data (weight is the response variable). Write down the general form for the model, starting out with:

$$weight_i = \beta_0 + \cdots, \quad i = 1, \ldots, n.$$

   Continue to fill out the rest of the right hand side of the equation. Your model should be expressed in terms of unknown parameters (e.g., β_1) and not specific estimated values (e.g.$\widehat{\beta}$) or numbers (e.g., 0.23). Make sure you include an error term, e_i, in the model and **be sure to specify all assumptions about its distribution**. Take a look at the Rmd files for the lecture notes/slides if you are unsure what to include or how to include it.

3. Write down the conditional mean function $E(Y \mid X)$ and the conditional variance function $\text{Var}(Y \mid X)$ for weight given height as a function of the unknown parameters. (You will estimate the parameters next.)

4. Calculate the summary statistics ($\bar{x}$, SXX, etc.) required to compute the ordinary least squares estimates of the parameters β_0 and β_1 in the mean function, and use these statistics to calculate the estimated values $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

5. Use the `lm` function in `R` to compute the least squares estimates and compare with your results above (they should be the same).

6. Calculate an unbiased estimate of the conditional variance of weight given height.

7. Provide an interpretation of all three estimated parameters. If any parameters don't have a natural interpretation, explain why.

8. Plot the data again, this time including the estimated regression line in the plot.

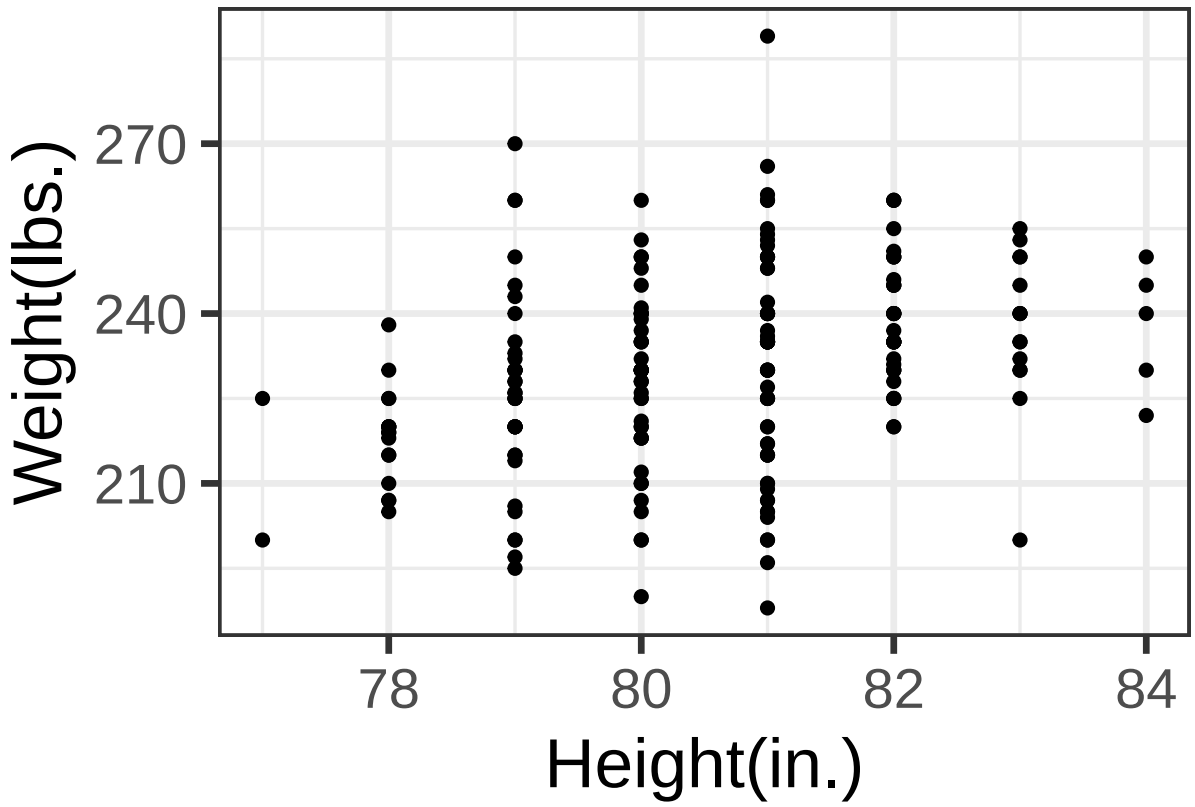
9. Compare your fitted model for `forwards` to the fitted model for `guards` we discussed in class. Summarize the differences in the estimated relationship between weight and height for players in these two positions.

Solution to Question 1 Your answers go here.

```
library(readr)
nba = read.csv('nbahtwt.csv')
```

Part 1.1:

```
nba %>% filter(pos=="F") %>% ggplot(aes(x=ht,y=wt)) + geom_point(size=2) + theme_bw(26) +
  xlab("Height(in.)") + ylab("Weight(lbs.)")
```



As seen in the plot, as height increases in forwards, the mean weight also increases. The variance also seems to be relatively normal for each height.

Part 1.2:

$$weight_i = \beta_0 + \beta_1 height_i + e_i, \quad i = 1, \ldots, n.$$

Part 1.3:

$$E(weight_i \mid height_i) = \frac{P(weight_i \cap height_i)}{P(height_i)}$$

$$Var(weight_i \mid height_i) = E(weight_i^2 \mid height_i) - E(weight_i \mid height_i)^2$$

Part 1.4:

```
nbastats = nba %>% filter(pos=="F") %>% summarize(xbar = mean(ht), ybar = mean(wt), SXX = sum((ht - xba
xbar = nbastats$xbar
ybar = nbastats$ybar
SXY = nbastats$SXY
SXX = nbastats$SXX

B_1_hat = SXY/SXX
B_0_hat = ybar - (B_1_hat*xbar)

B_1_hat
```

```
## [1] 3.777441
```

```
B_0_hat
```

```
## [1] -73.90178
```

The parameters B_0, B_1 can be represented as statistics: $\hat{B}_0, \hat{B}_1$. These statistics are calculated in the code above. $\hat{B}_0 = -73.90178$ pounds. $\hat{B}_1 = 3.777441$ pounds/inch.

Part 1.5:

```
forwards.lm = lm(wt ~ ht, data = nba, subset = (pos=="F"))

forwards.lm
```

```
##
## Call:
## lm(formula = wt ~ ht, data = nba, subset = (pos == "F"))
##
## Coefficients:
## (Intercept)           ht
##     -73.902        3.777
```

They are in fact the same coefficients.

Part 1.6: $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i-(B_0+B_1x_i))}{n-2}$

```
n = sum(nba$pos == "F")
s2_hat = nba %>% filter(pos=="F") %>% summarize(s2_hat = sum( (wt - (B_0_hat + (B_1_hat * ht)))^2 )/(n-

s2_hat
```

```
##     s2_hat
## 1 253.2537
```

```
s_hat = sqrt(s2_hat)
s_hat
```

```
##     s2_hat
## 1 15.91395
```

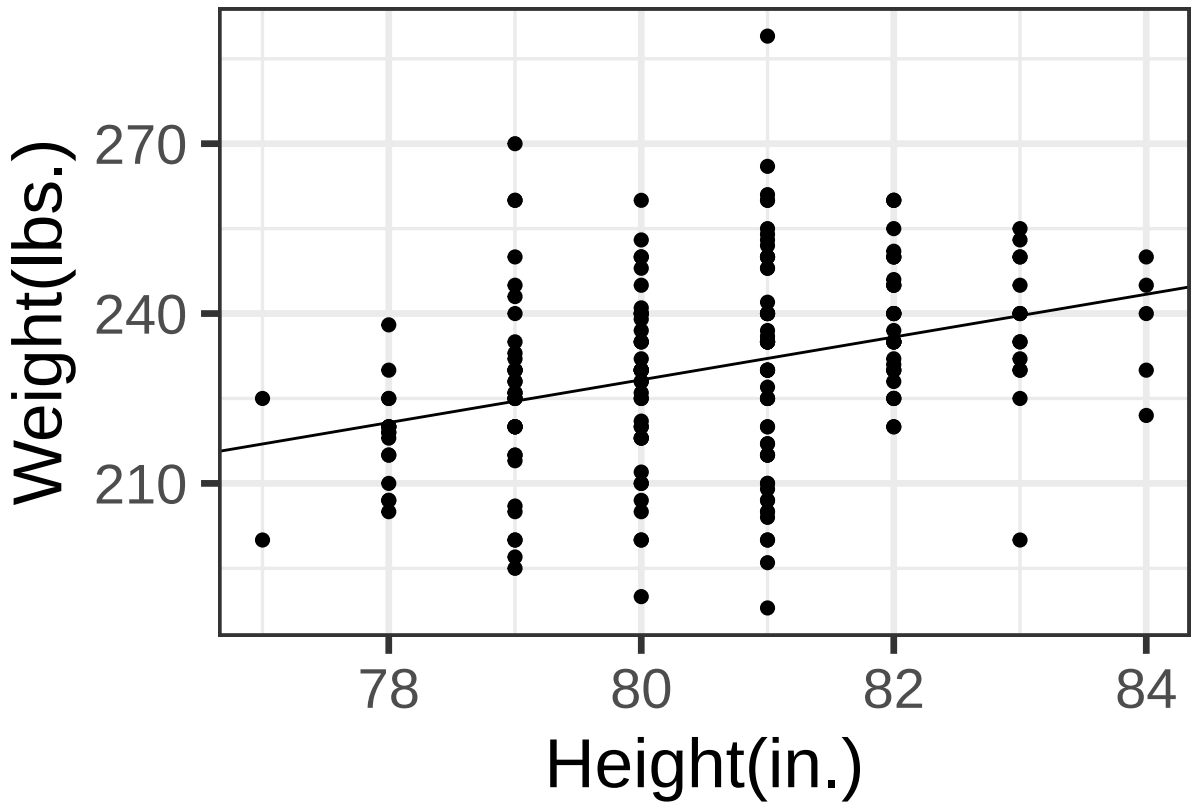
Part 1.7: So when a players height is 0, they are estimated to be -73.90178 pounds. Of course, a player will never be 0 inches tall or even close to it so it's not interpretable.

For every inch a player is, they are estimated to be 3.777441 pounds heavier.

Lastly, $\hat{\sigma^2}$ (253.2537 pounds) is hard to interpret in this scenario, but $\hat{\sigma}$ is easier so that will be interpreted instead. The standard deviation is roughly 15.9134 pounds so at any given height, most players will be within 3 standard deviations of the least squares regression for that given height.

Part 1.8:

```
nba %>% filter(pos=="F") %>% ggplot(aes(x=ht,y=wt)) + geom_point(size=2) + geom_abline(intercept = B_0_
  xlab("Height(in.)") + ylab("Weight(lbs.)")
```



Part 1.9: In terms of relation between weight and height, the guards have a greater slope. The forward's slope is about 3.777 pounds per inch, but the guard's slope is 4.338 pounds per inch.

The guards also have a lower y intercept at -131.54. One part of this is that guards are already generally shorter than forwards. On top of that, the slope is steeper which means that an extrapolation of that slope to height 0 is going end up lower than forwards are.

The variance of guards was 145.33 while the variance of the forwards was 253.2357 which shows that guards were often times closer to the same weight at any given height than the forwards.

Question 2 Problem 2.2 from Weisberg: (Data file: `UBSprices`) The international bank UBS...

Solution to Question 2 Your answers go here.

Part 2.2.1: Points above the $y = x$ line show that the price of rice from 2003 to 2009 has inflated while the points below the $y = x$ line show that the price of rice has deflated.

Part 2.2.2: The city with the greatest increase in rice price would be Vilnius. A way to calculate this would be to subtract the 2003 price from the 2009 price and take the greatest number, but since Vilinius is such a great outlier, the plot suffices for this.

The city with the greatest decrease in rice price would be Mumbai. Once again, you could subtract 2003 price from 2009 price and take the lowest number, but the plot clearly shows Mumbai as the greatest outlier.

Part 2.2.3: No, it does not. If we remove clear outliers (Vilnius, Budapest, Nairobi, Mumbai, Seoul), then the data clearly trends above y=x. These outliers, specifically Seoul, Nairobi, and Mumbai, skew the data below the line $y = x$.

Part 2.2.4: The first glaring problem is the variance of 2009 price given any 2003 price does not appear to be constant since at ~10 on the x-axis, the variance is narrower than the variance at ~20.

The distribution of values of the Y does not appear to be normal at each value of X either. At ~10, the data may be roughly normal, but by ~20, it does not appear normal since the values appear to be evenly distributed.
