Stat 3301: Homework 2

Jane Weissberg (weissberg.11)

Due by date and time specified on Carmen

Setup:

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

Instructions

- The html file, RMarkdown templete and data set for homework 2 can be found under the "Files/Homework/HW 2" page on Carmen.
- 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.

Question 1 Complete **Problem 1.1** from Weisberg: (Data file: `UN11`) "The data in the file `UN11`…" availa le from the alr4 package.

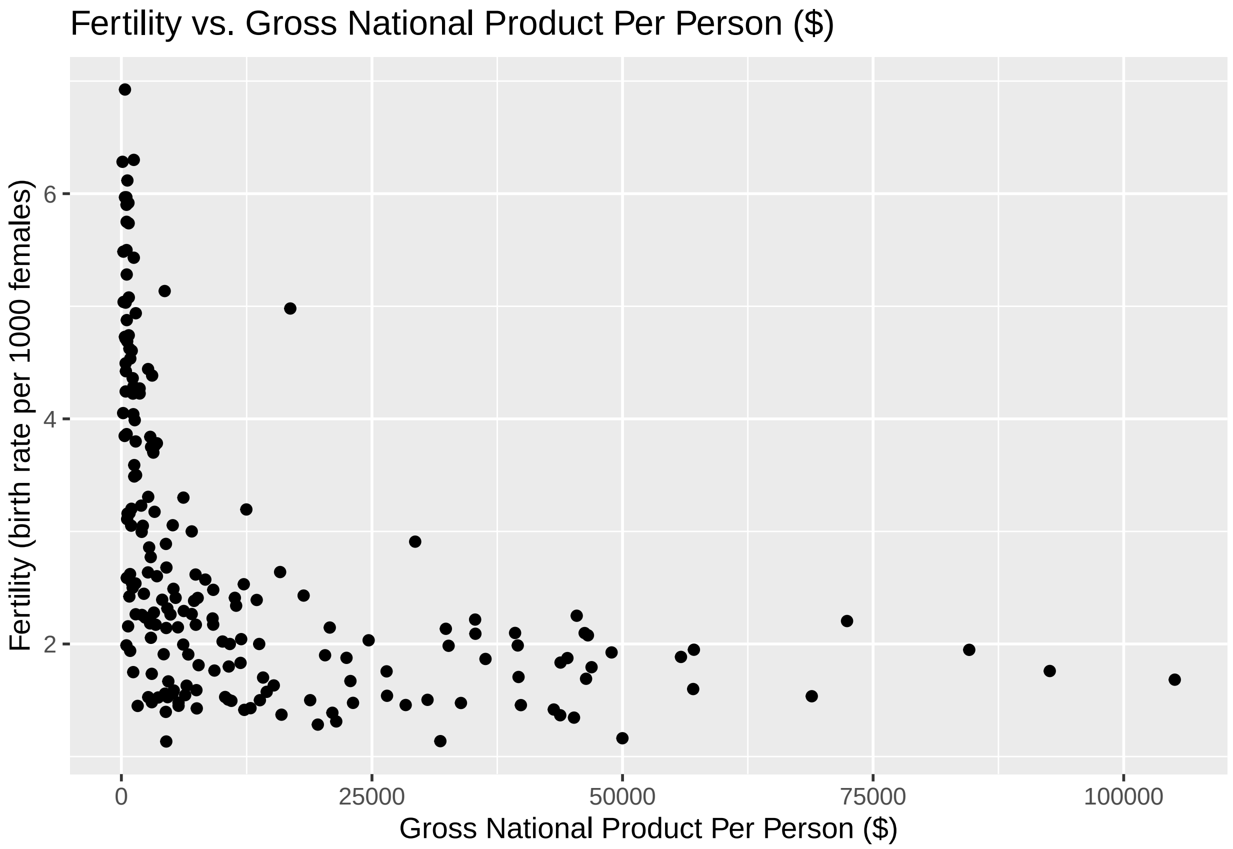
Note: for part 1.1.3, the `R` function `log` corresponds to the natural logarithm (base e). The `R` function `log10` corresponds to the base 10 logarithm. Other bases can be obtained via `log(x, base = ...)`. Note that in this class we will use the notation $\log_{10}$ to refer to the base 10 logarithm and log to refer to the base e logarithm (this is typical in the field of statistics).

Solution to Question 1 Your answers go here.

Part 1.1.1: The predictor is ppgdp, while the response is fertility.

Part 1.1.2:

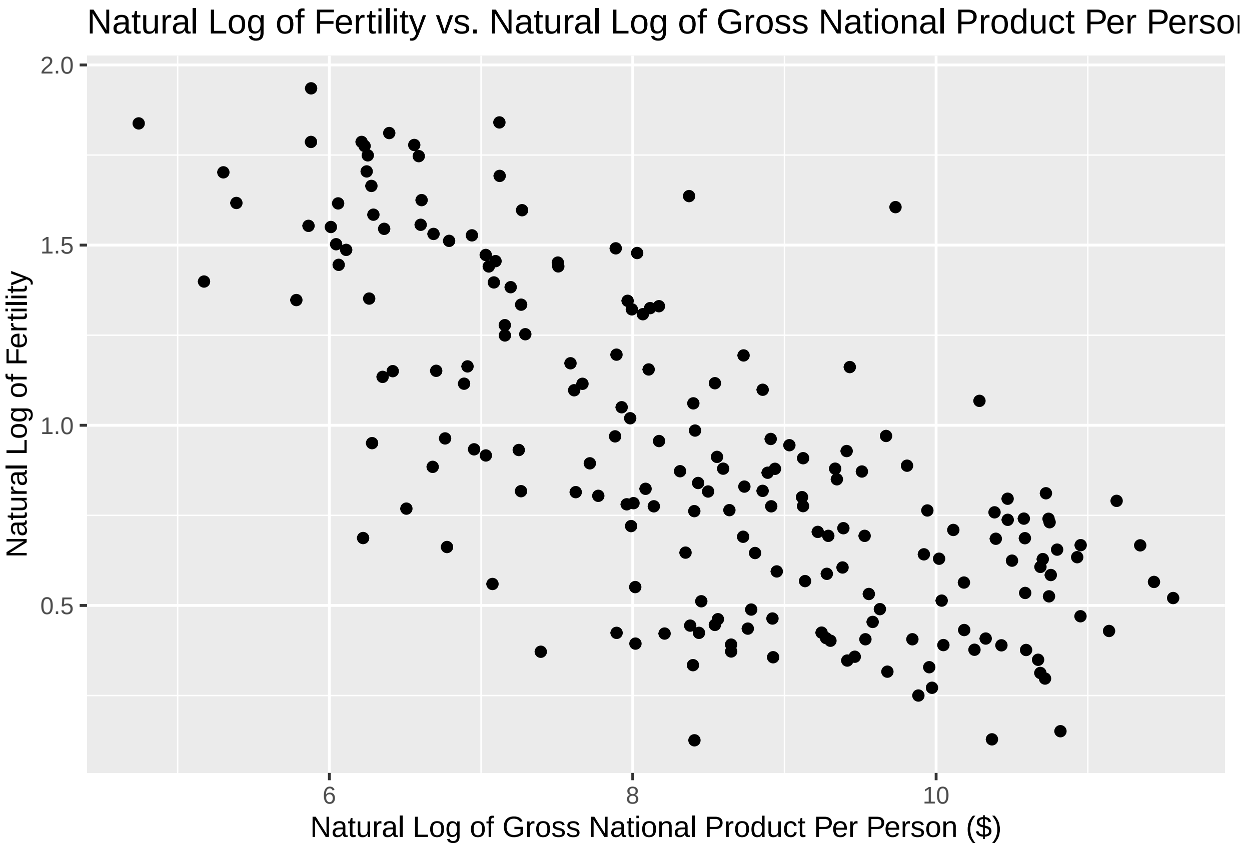
```r
data("UN11")
ggplot(data=UN11, mapping = aes(x = ppgdp, y = fertility)) + geom_point() +
  labs(title = "Fertility vs. Gross National Product Per Person ($)",
       x = "Gross National Product Per Person ($)", y = "Fertility (birth rate per 1000 females)")
```



A straight-line mean function does not seem to be plausible for a summary of this graph because the data does not follow a linear pattern.

Part 1.1.3:

```r
data("UN11")
ggplot(data=UN11, mapping = aes(x = log(ppgdp), y = log(fertility))) + geom_point() +
  labs(title = "Natural Log of Fertility vs. Natural Log of Gross National Product Per Person ($)",
       x = "Natural Log of Gross National Product Per Person ($)", y = "Natural Log of Fertility")
```



The simple linear regression model seems plausible for a summary of this graph because the data follows more of a linear pattern. ——

Question 2 Use the data file of National Basketball Association (NBA) for the 2013-2014 season (Data file: `nbahtwt`).

Part 1: Draw a scatterplot of height on the horizontal axis versus weight on the vertical axis for the NBA players who are playing in Center position.

Hint: The center players are denoted as "C" in the data set.

Part 2: What can you say about the relationship by looking at the plot produced in part 1?

Part 3: What can you say about the the relationship between average (mean) weight and height by looking at the plot produced in part 1.

Part 4: What can you say about the the relationship between the variance (variability) of weight around its conditional mean at different values of height by looking at the plot produced in part 1.

Hint: Refer the Lecture notes "Simple Linear Regression" from Week 2.

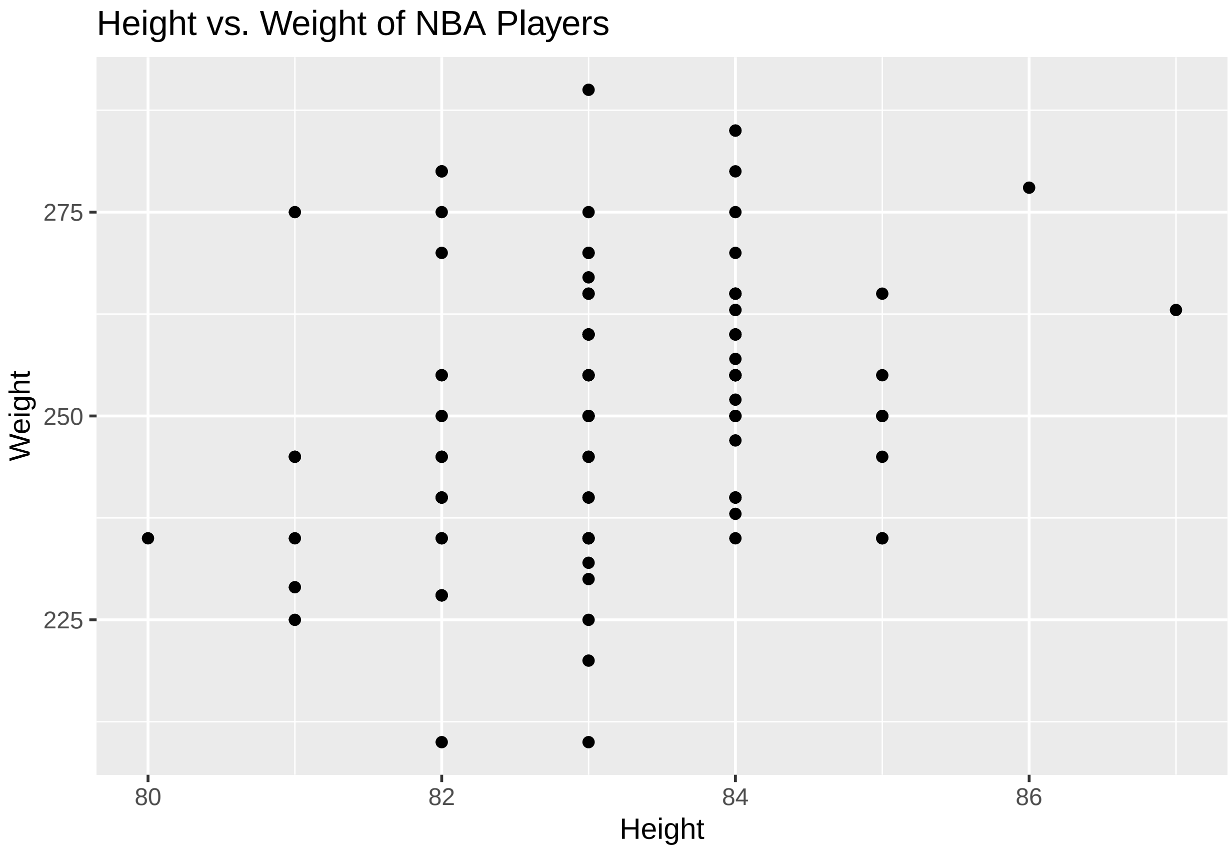
Solution to Question 2 Your answers go here.

Part 1:

```r
nba_data <- read_csv('nbahtwt.csv')
```

```
## Rows: 505 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): name, pos
## dbl (3): ht, wt, age
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
center_players <- subset(nba_data, pos == 'C')
ggplot(data = center_players, aes(x = ht, y = wt)) + geom_point() +
labs(title = "Height vs. Weight of NBA Players", x = "Height", y = "Weight")
```



Part 2: There seems to be a positive correlation between the height and weight of centers in the NBA.

Part 3: There is a positive, somewhat linear relationship between average weight and height for NBA centers.

Part 4: Some of the variances of weight around the mean are greater than others. For example, the variance of weight for heights of 80, 86, and 87 are very small, while the variance for a height of 83 is much greater.

Question 3 Complete **Problem 1.6** from our class text book, Weisberg: (Data file `Rateprof` from the alr4 package) "In the website and online forum `RateMyProfessors.com`…"

Note: In addition to summarizing the relationships, **also** reproduce the scatterplot matrix.

Recall that you can create a subset of the whole data set that corresponds to the variables you want to plot using:

```
subset(Rateprof, select = c(quality, clarity, helpfulness, easiness, raterInterest))
```

or

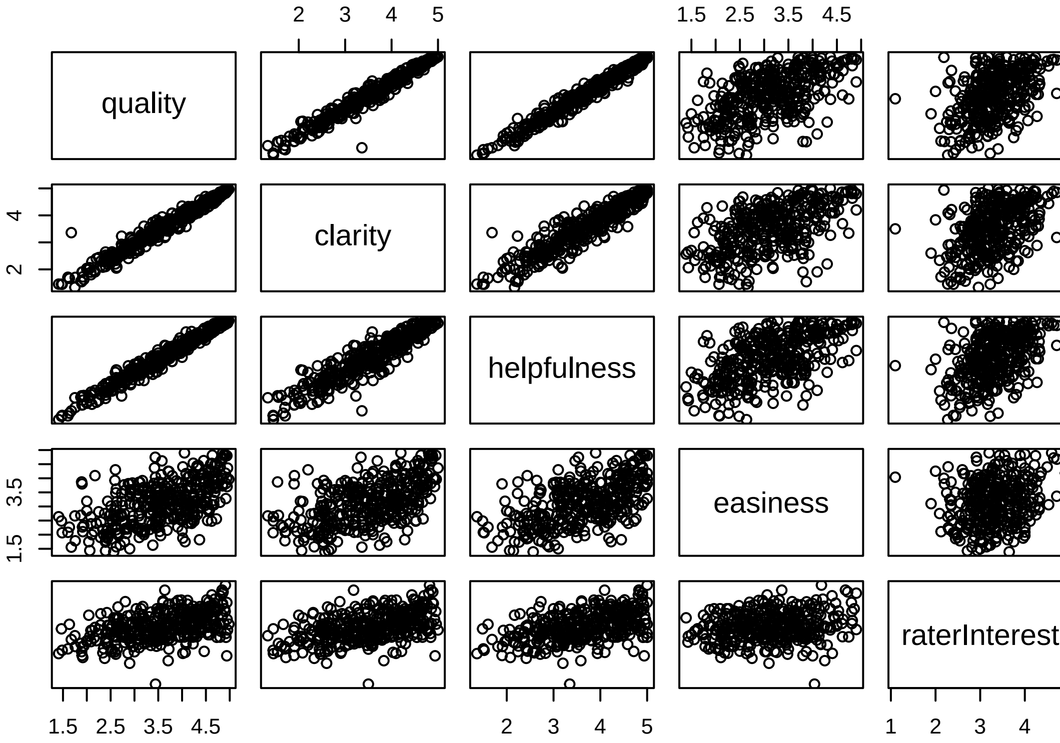
```
Rateprof %>% select(quality, clarity, helpfulness, easiness, raterInterest)
```

```


data("Rateprof")
subset_data <- subset(Rateprof, select = c(quality, clarity, helpfulness, easiness, raterInterest))

data("Rateprof")

pairs(subset_data)
```



Solution to Question 3

Clarity and helpfulness both have a strong positive relationship with quality. Helpfulness and clarity also both have strong positive relationships. The relationship between easiness and quality, helpfulness, and clarity is not as strong. Rater interest has a slightly positive relationship with quality, helpfulness, and clarity. However, its' relationship with easiness is not as strong.