Stat 3301: Homework 2

Nathan Johnson.9254

Due by date and time specified on Carmen

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.

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

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).

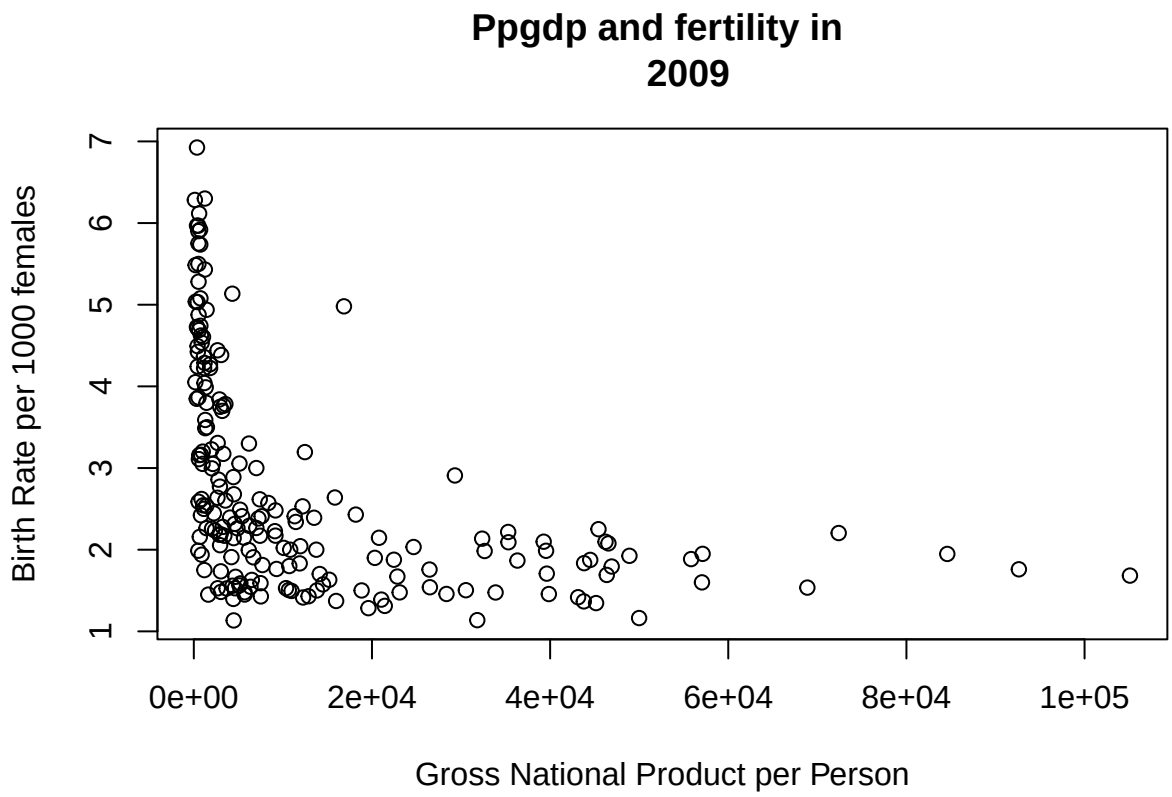
Solution to Question 1 Your answers go here.

```
library(readr)
data(UN11)
```

Part 1.1.1: The predictor variable in this problem is the ppdgp, the gross national product per person in U.S. dollars from 2009. The response variable in this problem is fertility, the birth rate per 1000 females from 2009.

Part 1.1.2:

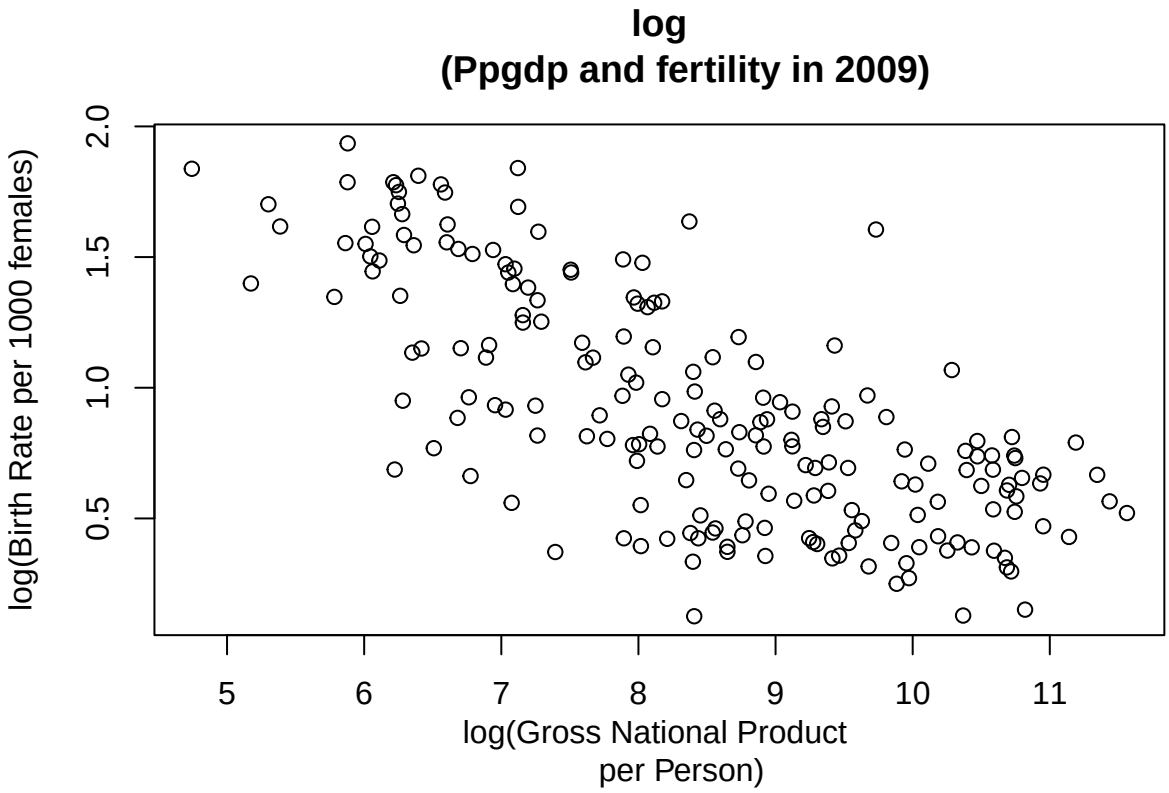
```
plot(UN11$ppgdp, UN11$fertility, xlab = "Gross National Product per Person",
       ylab = "Birth Rate per 1000 females", main = "Ppgdp and fertility in
     2009")
```



As seen in the scatterplot, a straight-line mean function does not seem to be plausible for the relation between ppgdp on the x-axis and fertility on the y-axis.

Part 1.1.3:

```
plot(log(UN11$ppgdp), log(UN11$fertility), xlab = "log(Gross National Product
     per Person)", ylab = "log(Birth Rate per 1000 females)", main = "log
     (Ppgdp and fertility in 2009)")
```



Yes, a simple linear regression model does seem plausible for a summary of this graph. The graph has a lot of variance, but more importantly seems to have a linear trend. As the mean gross national product per person increases, the mean birth rate per 1,000 females decreases. The variance of the birth rate per 1,000 females remains relatively the same as the gross national product per person increases. There is some more variance at the very lowest gross national product per person though.

Question 2 Problem 1.3 from Weisberg: (Data file: `Mitchell`) The data shown in Figure 1.12. . . "

Hint: Part 1.3.2 asks you to make a plot where the length of the horizontal axis is at least 4 times the length of the vertical axis. This can be done by specifying the R chunk option `fig.asp = ?` for an appropriate value of the question mark, like this:

You might try Googling "fig.asp RMarkdown" or something along those lines if you can't find help on the `fig.asp` option in R. If you get an error that says `figure margins too large` when you try to make the plot, just try adjusting the value of `?` until you get something close a 4 to 1 ratio.

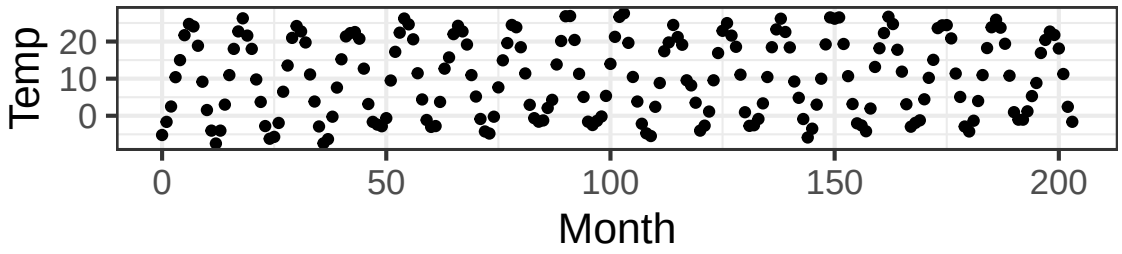
Solution to Question 2 Your answers go here.

Part 1.3.1: If we use a linear regression model for this graph the temperature of the soil does not seem to depend on the month number. It's linear relationship is a flat line so the average soil temperature does not seem to correlate with months after January 1976.

Part 1.3.2:

```
data(Mitchell)
```

```
Mitchell %>% ggplot(aes(x=Month, y=Temp)) + geom_point() + theme_bw(16)
```



Since this data is not linear, we can't use a linear regression model to describe it. That said, the temperature of the soil does follow a sin wave pattern, likely following the seasons of the year. If the soil has a low temperature, it is a winter month, if the soil has a high temperature, it is a summer month.

Question 3 Complete **Problem 1.6** from Weisberg: (Data file `Rateprof`) "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)
```

Solution to Question 3 Quality:

-Clarity: There is a strong positive correlation between quality and clarity with almost a perfect.

-Helpfulness: There is a strong positive correlation between quality and helpfulness.

-Easiness: There is a weak to moderate positive correlation between quality and easiness.

-Rater-Interest: There is a moderately strong positive correlation between quality and rater-interest.

Clarity:

-Helpfulness: There is a moderately-strong positive correlation between clarity and helpfulness.

-Easiness: There is a moderate positive correlation between clarity and easiness.

-Rater-Interest: There is a moderate positive correlation between clarity and rater-interest.

Helpfulness:

-Easiness: There is a moderate positive correlation ebtween helpfulness and easiness.

-Rater-Interest: There is a moderate positive correlation between helpfulness and rater-interest

Easiness:

-Rater-Interest: There is a weak to moderate correlation between easiness and rater-interest.

```
data(Rateprof)
```

```
pairs(Rateprof[8:12], pch = '.')
```