Lab 2

Psychology 310

Instructions. Work through the lab, saving the output as you go. You will be submitting your assignment as an R Markdown document.

Preamble. Today's assignment involves looking at a couple of classic issues in regression analysis.

Modern software systems like R allow one to present data visually in a variety of creative ways that can lead, in some cases, to interesting discoveries and the breaking down of barriers to understanding.

Unfortunately, one must master both the mechanics of the data analysis program and the ideas behind the analytic technique, before one can take full advantage of that technique. Often, it is a simple fact that it is not enough to present a picture of one's data. One must present the *correct* picture to obtain maximum benefit. Indeed, some times the wrong picture is worse than no picture at all.

Presenting a graphical representation with the wrong scaling can lead to views that are particularly deceptive. Consequently, one must be prepared to zoom in, zoom out, stretch and/or contract scales, transform data nonlinearly, and/or graph different subsets of data in order to explore the possibilities.

Traditional textbook exercises do not demonstrate or emphasize any of these skills. Indeed, they emphasize a narrow range of graphical representations of data, including scatterplots, histograms, and mean plots, usually with "default options" which work well for many situations, but are horribly suboptimal for others. Unfortunately, by the time most students reach the graduate level, they have adopted a number of mental sets that hinder full exploration of data.

Our first illustration of this is a simple one, which appears in the outstanding textbook *Applied Linear Regression* by Weisberg. Begin by installing the libraries `alr4` and `ggplot2` in your R system. You will need them installed to process this lab. Now, activate the `mitchell` data file with the following sequence of commands.

```
> library(alr4)
> data(Mitchell)
> attach(Mitchell)
```

You can examine the data in a variety of ways. You can ask for the names of the variables:

```
> names(Mitchell)

[1] "Month" "Temp"
```

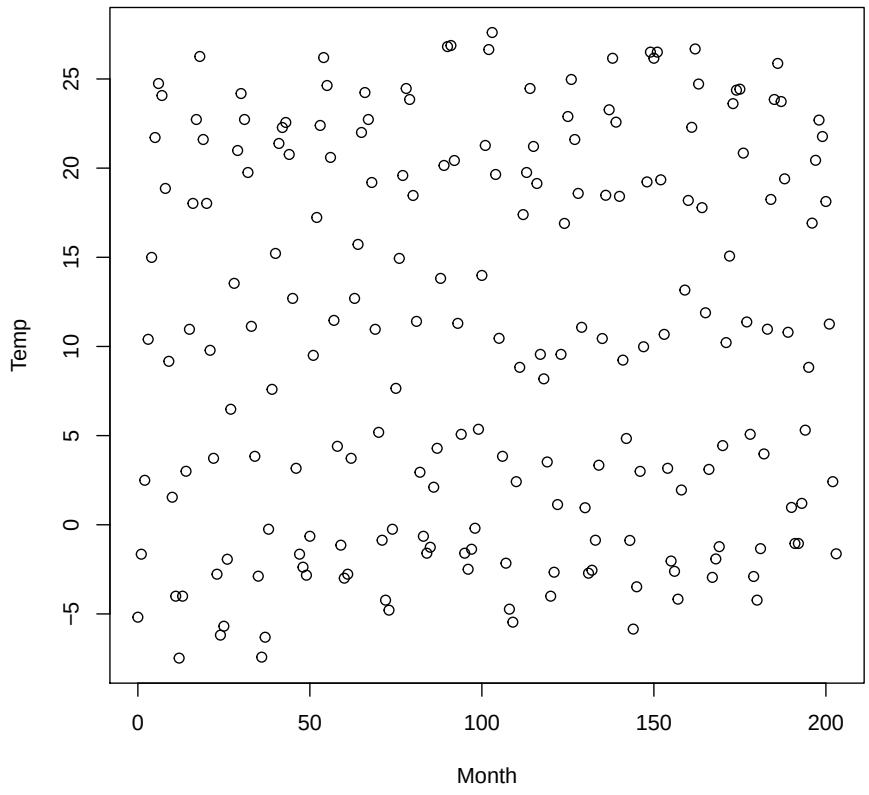
You can call for a summary of the data:

```
> summary(Mitchell)

     Month            Temp
 Min.   :  0.0   Min.   :-7.478
 1st Qu.: 50.8   1st Qu.:-0.349
 Median :101.5   Median :10.450
 Mean   :101.5   Mean   :10.313
 3rd Qu.:152.2   3rd Qu.:20.431
 Max.   :203.0   Max.   :27.606
```

The data are a plot of average temperature vs. month of the year at a particular geographic location. Take a look at the default scatterplot produced by R.

```
> plot(Month, Temp)
```



It looks as if there is a somewhat random relationship between `Month` and `Temp`.

Of course, we know that there is indeed a well-established, cyclical relationship between time of year and temperature, so something may be seriously wrong, either with the data, our graph, or both!

Now, if you look carefully, you can see some "strands" in the plot, and this can be a clue that there is some kind of dependency. Try elongating the plot and taking another look. You can do that on your screen with the mouse. Here is what you should do.

1. First, grab the right center edge of the plot by hovering the mouse pointer over it, then left clicking. Stretch the graph so that it is as wide as you can make it.

2. Then, grab the bottom center of the plot, and *sloooowwwwly* collapse the graph upward.

What do you see? When the picture has become sufficiently revealing, save your plot.

Alternatively, you can manipulate the aspect ratio of the plot directly, using the `asp` optional parameter, which you can read about in the help file documentation on the function `plot`

Hopefully this little exercise has convinced you that a standard scatterplot, produced with some default option chosen by a computer programmer, can work well in a wide variety of circumstances, but can also lead to a very misleading picture in others.

The next part of our exercise involves an interesting data set on birth rates, death rates, and economic activity for 97 nations. The files are *poverty.dat* and *poverty.txt*. Copy the data file to your personal space, then open your personal copy.

This data file was originally distributed as part of an article in the online *Journal of Statistics Education*.

Rouncefield, M. (1995). The statistics of poverty and inequality. *Journal of Statistics Education*, 3.

The article, which you will not need to complete this exercise, may be downloaded from

http://www.amstat.org/publications/jse/v3n2/datasets.rouncefield.html

The article contains a substantial amount of background information on the meaning of the variables contained in the data file.

Read in the data file and remove the missing data with the commands:

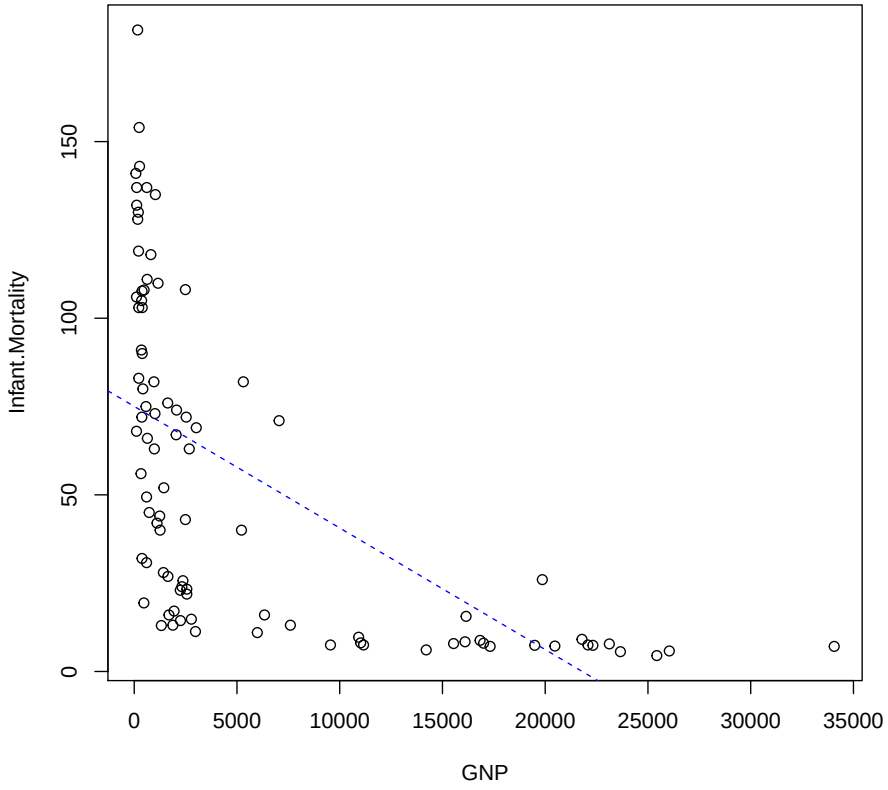
```
> library(ggplot2)
```

```
> poverty.data <- read.table("poverty.dat", header = T)
> poverty.data <- na.omit(poverty.data)
> attach(poverty.data)
> names(poverty.data)

[1] "Birth.Rate"       "Death.Rate"       "Infant.Mortality"
[4] "Male.Life.Exp"    "Female.Life.Exp"  "GNP"
[7] "Region"           "Country"
```

In this exercise, we are particularly interested in the relationship between `Infant.Mortality` and `GNP` (gross national product), often taken as an indicator of overall economic productivity.

Using techniques demonstrated in class, produce a scatterplot showing `Infant.Mortality` predicted from `GNP`. Fit the data with linear regression and add the line to your scatterplot in dotted blue. The plot should look like this:



It appears that linear regression does not do a good job of describing the relationship between GNP and infant mortality. The bivariate scatterplot is L-shaped, and the straight line just doesn't fit the point pattern. Perhaps some form of polynomial regression is called for.

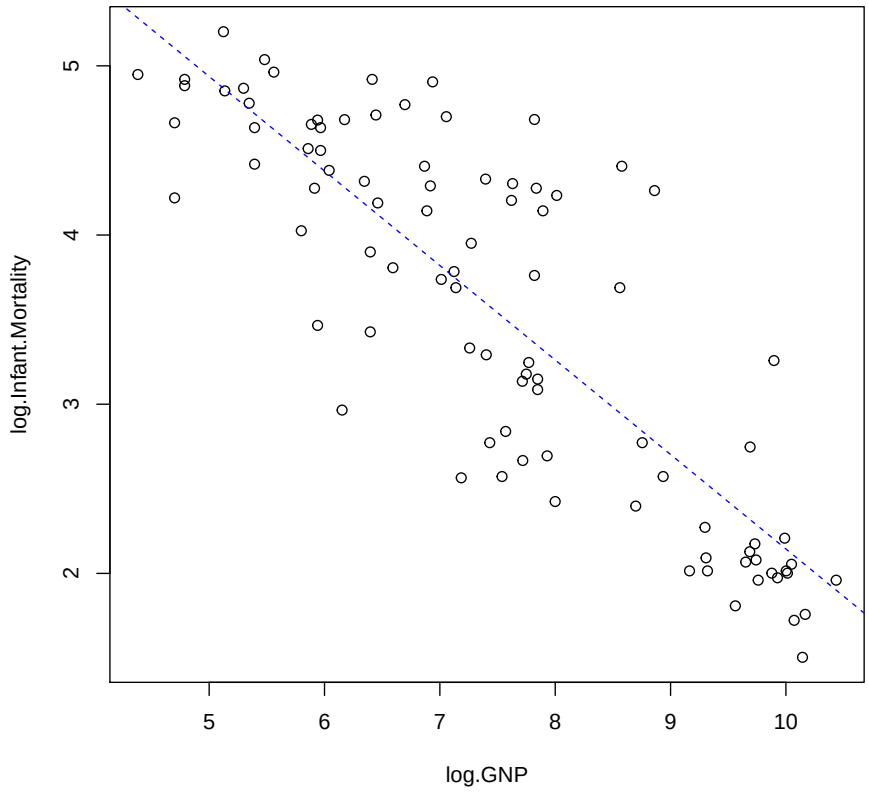
Or is it? In many cases where data are compressed along the Y-axis, log-transforming either the dependent variable, the independent variable, or both, can be very useful. In this case, we'll try transforming both variables.

Perhaps the problem is one of scaling. Let's try rescaling the plot. Create a variable called `log.GNP` and `log.Infant.Mortality` with the statements,

```
> log.GNP <- log(GNP)
> log.Infant.Mortality <- log(Infant.Mortality)
```

Now fit a linear regression predicting Infant.Mortality from log.GNP. Your plot should look like this. Notice how I extract the coefficients for further use in the code below.

```
> log.log.fit <- lm(log.Infant.Mortality ~ log.GNP)
> plot(log.GNP, log.Infant.Mortality)
> abline(log.log.fit, lty = 2, col = "blue")
```




```
> b1 <- coef(log.log.fit)[2]
> b0 <- coef(log.log.fit)[1]
```

This looks a lot better, doesn't it!

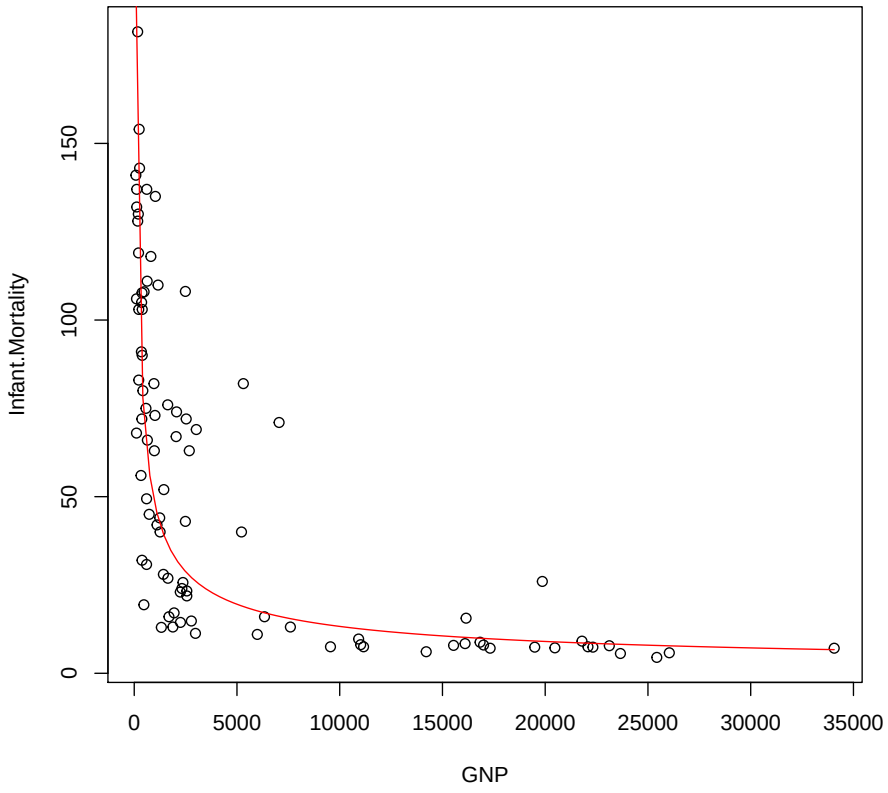
The scatterplot has been smoothed and made more linear. To begin with, we might ask how we interpret the regression coefficients of best linear fit of

this this "log-log" plot. The plot is $\log(Y) = b_1 \log(X) + b_0$. Now, suppose we exponentiate both sides. We get $\exp(\log(Y)) = \exp(b_1 \log(X) + b_0)$, or, using the laws of exponents,

$$y = X^{b_1} \exp(b_0)$$

.

We can convert the result of this log-log fitting back to the original metric by using the above equation. Here is a plot:



How do we interpret this? In absolute terms, an interpretation is difficult, but in relative terms, the meaning is clear.

Suppose we increase X by exactly 1%. What will be the proportional change in Y?

With a little algebra, we arrive at

$$\begin{aligned} X_2 &= 1.01X_1 \\ Y_2/Y_1 &= \frac{(1.01X_1)^{b_1}\exp(b_0)}{X_1^{b_1}\exp(b_0)} \\ &= 1.01^{b_1} \end{aligned} \tag{1}$$

The first few terms of the Taylor Series approximation of 1.01^{b_1} are $1 + 0.00995033b_1 + 0.0000495045b_1^2$. This is very close to $1 + .01b_1$. In other words, a proportional change of 1% in X will result in a multiple of $1 + .01b_1$, which approximately a *proportional* change of b_1% in Y.

In this case, we see that $b_1 = -0.56$, so that a 1% increase in GDP for any country corresponds to a *proportional* decrease of 0.56% in infant mortality. Because the change is proportional to where a country is, and countries with very small GDP tend to have comparatively large infant mortality rates, this means that most of the "action" or value achieved by an increase in GDP occurs at the lower levels. Countries with large GDP tend to have low mortality rates, and percentage changes in those rates are, in absolute terms, small.

If you look at the plot, you can see some points hovering above the rest. These appear to be positive outliers. They represent data points where the infant mortality rate is substantially higher than you would predict from GNP. Use the `identify` function in R to identify these points. Issue the following commands, then click on the points hovering far above the rest to see which countries are represented.

```
> plot(GNP, Infant.Mortality)
> curve(x^b1 * exp(b0), add = TRUE, col = "red")
> identify(GNP, Infant.Mortality, plot = TRUE, labels = Country)
```

Which countries are outliers? What do they have in common? Can you hypothesize why they are outliers?

There are a number of other ways we might refine this regression model, but those methods are left for another time.