Exercise 1:

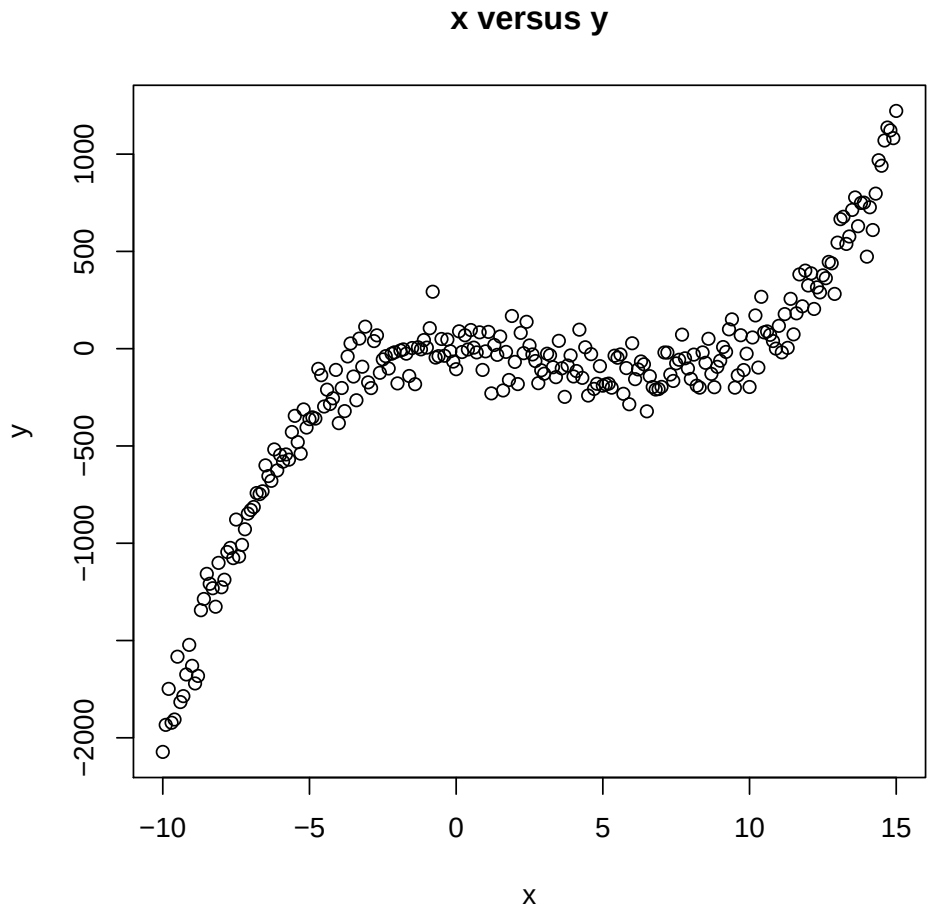
(a) Create 2 fake variables where (1) X is a sequence of numbers from -10 to 15 in steps of 0.1, and (2) Y equals $x^3 - 10x^2$ plus a random normal deviate with mean 0 and standard deviation 100. Show all of your R code for this problem.

(b) Plot X versus Y in a scatterplot. Be sure to label your axes and provide a plot title.

Answer:

```
> x = seq(-10, 15, 0.1)
> y = x^3 - 10*x^2 + rnorm(length(x), 0, 100)

> plot(x, y, main = 'x versus y')
```



Exercise 2:

(a) Calculate correlation for: $X < 0$, $0 < X < 8$, $X > 8$, and the entire range of X.

(b) Using about 250 words...

- ...explain why the 4 correlation figures you calculated differ from each other. Be sure to talk about the sign and magnitude of each number and relate them to those of each other.
- ...describe major limitations of using correlation to describe the relationship we see between X and Y (hint: what types of relationships does correlation (not) work for? Do X and Y share that kind of relationship? How do you know?

Answer:

```
> cor(x[x<0], y[x<0])

[1] 0.9321448

> cor(x[x>0 & x<8], y[x>0 & x<8])

[1] -0.4094365

> cor(x[x>8], y[x>8])

[1] 0.9207798

> cor(x, y)

[1] 0.7989372
```

The four numbers are different because the relationship between x and y is different depending on the domain of x. For example, between 0 and 8, the relationship is negative, while it is highly positive outside that domain. The correlation for the entire domain of x is lower in magnitude because x and y are non-linearly related.

Exercise 3:

(a) Use 'read.csv' to load the WB dataset into the variable 'wb'. Then, use the command 'na.omit' on 'wb' to drop all observations with NAs in them (hint: use '?na.omit' to research how to use the command. Pay particular attention to the examples section.). Save this new subset into the variable 'wb1' (hint: 'wb1' should contain 191 observations).

(b) Using 'wb1', calculate the correlation between (1) GDP and population, and (2) GDP and life expectancy WITHOUT using the correlation command (hint: use the formula for calculating correlation from covariance). Show all of your R code for this problem.

Answer:

```
> wb = read.csv('c:/users/daniel/dropbox/ps6/data/wb2k.csv')
> wb1 = na.omit(wb)
> cov(wb1$gdp, wb1$pop)/(sd(wb1$gdp)*sd(wb1$pop))

[1] 0.2755039

> cov(wb1$gdp, wb1$life)/(sd(wb1$gdp)*sd(wb1$life))

[1] 0.1863872
```

Student must show the above formulas/code in order to get credit.

Exercise 4:

(a) Run 2 linear regressions: GDP on population (i.e. GDP is the DV, population is the IV) and life expectancy on GDP. Show regression tables for both regressions.

(b) In about 250 words, describe what these regression tables are telling us. At a minimum, you should talk about the meaning of the stars and what the estimated effects mean in real world terms. Keep this discussion purely to the numbers presented in the regression tables (i.e. don't talk about causes or bring in outside information).

Answer:

```
> summary(m1 <- lm(gdp ~ pop, data=wb1))

Call:
lm(formula = gdp ~ pop, data = wb1)

Residuals:
     Min       1Q   Median       3Q      Max
-1568491  -118185  -109462   -94838  9265890

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.094e+05  5.921e+04   1.848 0.066218 .
pop         1.855e+00  4.709e-01   3.940 0.000115 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 792500 on 189 degrees of freedom
Multiple R-squared:  0.0759,        Adjusted R-squared:  0.07101
F-statistic: 15.52 on 1 and 189 DF,  p-value: 0.0001146

> summary(m2 <- lm(life ~ gdp, data=wb1))

Call:
lm(formula = life ~ gdp, data = wb1)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-26.756  -6.819   3.810   7.532  14.132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.649e+01  7.588e-01  87.624  < 2e-16 ***
gdp         2.364e-06  9.065e-07   2.608  0.00983 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.27 on 189 degrees of freedom
Multiple R-squared:  0.03474,        Adjusted R-squared:  0.02963
F-statistic: 6.802 on 1 and 189 DF,  p-value: 0.009832
```

For part A, code is unnecessary. Both models must be shown, but regression tables can be shown in pretty format or raw R output. For part B, student must interpret the coefficients in substantive terms (e.g. coefficient β means that for every unit change in X, there is a β change in Y.). They must also talk in some capacity about statistical signifiance. They can use *either* the p-value or signifiance stars to do this. Deduct points if they talk about causes or speculative stuff.

Exercise 5:

(a) Now, using about 250 words, speculate on some of the reasons why we see the results in the regression tables in problem 4. Why does does population have a positive effect on GDP? Likewise for GDP and life expectancy. In essence, come up with *theories* to explain the data.

(b) Further, using another 250 words (separate from the previous count), talk about whether you think GDP causes life expectancy. Is the answer to this question a strict yes-no? Are there other factors that might cause conflation or a common response situation?

Answer:

For part A, give full credit so long as both regressions are addressed and the explanations make a modicum of sense. For part B, give full credit if the explanation makes sense, and if they address the issue of conflation/common response. They do not have to give an example of how conflation/common response might occur but they do need to provide a rationale for their answer.