**Lab Assignment 11**

**I. Objectives**: To learn the key components of multiple regression analysis…

**II. Datasets**: “UN-data.csv”

**III. Packages**: none

**IV. Preparation**

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear any data in memory by typing rm(list=ls())

3) Download dataset “UN-data.csv” and place it in your working directory

4) Download R script “POLS 6480 Lab 11.R” and place it in your working directory.

5) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

**V. Instructions for Lab 11**

The dataset you will use a small dataset from the 2005 United Nations Human Development report (all data is for 2003). After listing the country’s name, the first column is the Human Development Index (HDI), which combines female life expectancy at birth (Life), educational attainment, and income per capita. The second column is the fertility rate (Fert), measured as births per adult female, and the third column is the percentage of women using contraception (Cont). Technological adoption is measured using the share of the population using cell phones (Cell) and the share of the population using the internet (Inter), both measured as subscribers per 1000 people. Wealth is measured using per capita gross domestic product (GDP) in US dollars. The dataset also contains carbon dioxide emissions per capita (CO2) measured in metric tons. Finally, to measure gender equality, the dataset contains female adult literacy rate (Liter) and adult women in the labor force per 100 men (FemEc), a.k.a. the female economic activity rate.

1. Load the dataset, typing the following line, changing the directory if needed:

> UN <- read.csv("C:/UN-data.csv")

Your goal is to explore how the percent of women using contraceptives (Cont) affects the birthrate (Fert). Before you run a simple regression model, however, you should inspect the data, because R is interpreting some variables (Cont, Cell, Inter, Liter) as factors even though they are integer-valued. If you were to run the model the obvious way, defining the model as lm(Fert ~ Cont, data = UN), then R would estimate a separate coefficient for each of the 21 levels that the variable Cont takes on!

When you run into such a problem, you might estimate the simple regression by typing:

> simple <- lm(Fert ~ as.numeric(Cont), data = UN)

but this model uses a command that simply assigns values of 1 to *L* for the *L* different levels that the variable inside the () takes on; in this case, values from 1 to 21. This is not a proper solution, but it does at least provide an single slope coefficient (type simple$coefficients); this would be appropriate if you believe Cont should be treated as an ordinal variable rather than interval.

To code the variable properly, you must tell R first to recognize the characters and then to use those to code the data numerically. Here is such a process for creating a new variable in the UN data frame: > UN$Contracept <- as.numeric(as.character(UN$Cont))

Use the new variable to estimate the model:

> correct <- lm(Fert ~ Contracept, data = UN)

> correct$coefficients

The latter line displays the intercept and slope. Recall that the coefficient indicates the effect that a one-unit increase in the explanatory variable has on the response variable’s value. In this case, we can say that increasing the percent of women using contraception by 1 unit has the effect of \_\_\_\_\_\_\_\_\_\_ (increasing/decreasing) the number of babies born by \_\_\_\_\_ per adult female.

2. Before we move on to multivariate regressions, it is worth examining the intercept. Recall that the intercept tells you the expected value of the response variable when the value of the explanatory variable equals zero. So, if 0% of women were using contraception, how many babies would each adult woman be expected to bear?

Is 0% in the range of the values of Contracept? What is the actual range of values that the Contracept variable takes on? Minimum = \_\_\_\_\_; Maximum = \_\_\_\_\_. What is the average value of Contracept? Mean = \_\_\_\_\_.

In lab 10, you predicted the expected revenues of a Tom Cruise movie for three different Freshness ratings. You could also predict the fertility rate given contraceptive use. The following code will predict fertility rates at minimum, median, and maximum values of contraceptive use:

> predict(correct, data.frame(Contracept = c(13, 74, 84)))

3. Obviously regression analysis tells us a lot more than just the values of the intercept and slope. Two important pieces of information are the degree to which the model fits the data and the statistical significance of each explanatory variable. To assess model fit, we usually look at the R2, which tells us what percentage of the variance is explained by the model; a high percentage is perceived as good. To assess statistical significance, we usually look at the *t* statistic and *p* value. The easiest ways to see these quantities is to type the following:

> summary(correct)

The R2 for the model is shown as \_\_\_\_\_\_; the *t* statistic is \_\_\_\_\_\_\_, which – according to the asterisks shown – is statistically significant at better than the .001 level.

4. The basic idea underlying multivariate regression is that there may be multiple explanatory variables that each affect the response variable simultaneously. More importantly, the explanatory variables might be related to each other. For example, countries that experience higher use of contraception might also experience higher female labor force participation, and those factors reinforce each other to determine the birth rate. To figure out whether and to what extent contraceptive use affects the fertility rate, we would want to hold constant the value of labor force participation. (It is worth considering further whether both factors are determined by GDP!)

Before estimating a multivariate model, it is informative to inspect the correlations. To calculate correlations, however, we need to have numerical data. To see whether R perceives female labor force participation (FemEc) as a factor, you can type is.factor(UN$FemEc), which will give either a TRUE or FALSE response. If it is numeric rather than a factor, you can move on.

Examine the correlations between contraceptive use and female labor force participation:

> cor(UN$FemEc, UN$Contracept , use="complete.obs")

The reason for the extra code inside the parentheses is that there are missing values of the Contracept variable! R will not calculate the correlation coefficient without this code.

What is the bivariate correlation coefficient? *r* = \_\_\_\_\_\_\_.

You can also examine whether each of these variables is correlated to GDP per capita:

> cor(UN$GDP, UN$FemEc , use="complete.obs")

> cor(UN$GDP, UN$Contracept , use="complete.obs")

A visual way to see the relationships between three variables is the pairs command, which shows scatterplots for all the variables in a data frame or for selected variables. For just the three variables whose correlations we were examining, you can type:

> pairs(~ GDP + FemEc + Contracept, data = UN, pch=19)

5. Recall that the basic code for a regression model allows you to add as many explanatory variables as you like, using the + symbol, for example:

> model <- lm(dataset$yvar ~ dataset$xvar + dataset$control)

> model <- lm(yvar ~ xvar + control, data = dataset)

These two lines will estimate the same model; you can also eliminate any reference to dataset by typing attach(dataset) before running the models.

To estimate a simple regression model in which FemEc is the sole explanatory variable and Fert is the response variable, you would type:

> labor.model <- lm(Fert ~ FemEc, data = UN)

> summary(labor.model)

The intercept, \_\_\_\_\_\_\_, tells us that the expected fertility rate if female labor force participation was equal to 0 is roughly \_\_\_\_\_ babies per adult female. However, the range of female labor force participation rates is from 29 to 91, with an average of about 65. To predict the fertility rate at the mean value of female labor force participation, you could type the following:

> predict(labor.model, data.frame(FemEc = mean(FemEc)))

The slope coefficient, \_\_\_\_\_\_\_, tells us that for each unit increase in female labor force participation, we predict the fertility rate will \_\_\_\_\_\_\_ (increase/decrease) by roughly \_\_\_\_\_ babies per adult female. Therefore, it would take a \_\_\_\_\_\_\_ unit increase in female labor force participation to reduce the number of babies per adult female by one.

The *t* statistic and/or *p* value indicate that the relationship \_\_\_\_\_\_\_ (is/is not) significant at the .05 significance level.

6. We already saw above that the use of contraceptives affects the fertility rate. To estimate the multiple regression model in which the variable Contracept has been added to the specification, you would type:

> multi.model <- lm(Fert ~ FemEc + Contracept, data=UN)

> summary(multi.model)

Does contraceptive use have a statistically significant effect on fertility rate, holding female economic participation rate constant?

Does the female economic participation rate have a statistically significant effect on the fertility rate, holding contraceptive use constant?

Compare the R2 values of three models – just FemEc, just Contracept, and both FemEc and Contracept. By how much does the multivariate model improve over either bivariate model?

7. We saw earlier that female labor force participation is correlated to contraceptive use and that each of these explanatory variables *on its own* affects the fertility rate. Suppose one developed a causal model in which the level of labor force participation determines the level of contraceptive use, and that the effect of female labor force participation on fertility is entirely indirect, filtered through its effect on contraceptive use. Once we account for contraceptive use, the direct effect of labor force participation (evident in the model labor.model) disappears. This would be an example of what is often called a mediating relationship. You can examine the premise of this model by looking at the regression of contraceptive use on female labor force participation:

> med.model <- lm(Contracept ~ FemEc, data = UN)

> summary(med.model)

Does female labor force participation have a statistically significant effect on contraceptive use?

What share of the variance in contraceptive use is explained by female labor force participation?

8. If you subtract the R2 of med.model from 1, this tells you the share of the variance in contraceptive use that is not explained by female labor force participation. The deviations of the true value of Contracept from the values predicted by a linear function of FemEc (i.e., the regression line based on med.model) are referred to as *residuals*. You can think of these residuals as the portion of Contracept that is independent of FemEc.

You can create a new variable from the residuals using the code shown in lines 24–25. Because there are missing values of Contracept, residuals exist for 35 of the 39 countries in the UN dataset. Therefore, we start by creating a new dataset, named postest, that contains 35 cases. Next, we create a new variable named Indep that is the residual of Contracept after we account for the effect of FemEc.

Line 26 respecifies the regression model in 6 using this new variable in place of Contracept:

> alt.model <- lm(Fert ~ FemEc + Indep, data=postest)

> summary(alt.model)

Examine the coefficient for Indep; is it different from the coefficient for Contracept?

Examine the R2 of alt.model and multi.model; are they different?

Examine the coefficient for Fert in alt.model; is it different from the coefficient for Fert in multi.model?

The process you just went through “works” by replacing the variable for contraceptive use with a new variable that is uncorrelated to female labor market participation. (Check the correlation!) Consequently FemEc gets “credit” for its total effect on fertility rates (whereas previously its effect was hidden once contraceptive use was included). Note that this method is only appropriate if you have an unambiguous causal path, whereby female labor market participation affects contraceptive use and not the reverse.

9. On your own, re-run the model in 7, but add the variable HDI to the specification. Recall that the Human Development Index combines life expectancy, educational attainment and income. Then, re-run the model in 6, with the variable HDI added to the specification. Once we control for the country’s value on the Human Development Index, does Contracept have an independent effect on the fertility rate?

10. One advantage of linear regression is the ease of making predictions. In a simple regression, the intercept and slope define a line, so one can enter a value for the explanatory variable into the equation and generate a predicted value of the response variable. In a multiple regression, the intercept and the slopes define a plane (or a hyperplane for more than two explanatory variables). If you enter a value for each explanatory variable, you can generate a predicted value for the response variable.

Suppose we wished to predict the fertility rate. Lines 29–31 in the R script create a new data frame that has four hypothetical data points, created by setting FemEc equal to its first and third quartiles (51 and 80, respectively) and setting Contracept equal to its first and third quartiles (56 and 77.5, respectively).

Line 32 uses the predict() command to generate the predicted values. The first entry is the name of the regression model that you are using to generate the prediction. The second entry is the name of the data frame that you are using, having told R to use new data. The code is:

> predict(multi.model, newdata = twobytwo)

The impact of changing setting FemEc from its first to its third quartile can be assessed by comparing the first and second predictions (and/or comparing the third and fourth predictions, but because the model is linear, these deltas will be equal).The impact of changing Contracept from its first quartile to its third quartile can be assessed by comparing the first and third predictions (and/or comparing the second and fourth predictions). Based on these comparisons, which variable has a larger effect on predicted values of fertility rates?

11. To clear the **Environment**, type rm(list=ls()) or click on the broom icon.

To clear the **Console** window, type Ctrl-*l*