Regression Analysis in Stata

1 Purpose

The purpose of this activity is to provide you with an understanding of regression analysis (for the purposes of improving your understanding of Problem Sets 3 and 4, the quantitative reading material, and/or your final essay) and to both develop and apply that knowledge to the use of Stata as a statistical programming environment.

2 Overview

This lab can be completed during at home and everything you need to know is provided in this document.

3 Your Task

Using Stata as instructed, complete the following activities.

For this lab, we will use the UK portion of the European Social Survey (ESS) Round 7 data from 2014. You can download these data from here: `http://www.europeansocialsurvey.org/download.html?file=ESS7GB&c=GB&y=2014`. You may have to login with an email address and unzip the file. Save the file on your computer and unzip it.

3.1 Bivariate Regression

1. Use `cd` to set your working directory to wherever you saved the file. For example, on Windows, you might want to try `cd c:/users/thomas/desktop`.

2. Start a new text document in the editor of your choice (could be the Stata "do file editor" or Notepad, etc.). Write all code that you generate in that file so you have a complete record of your analysis.

3. Read the data into Stata using:

   ```
   use "ESS7GB.dta", clear
   ```

4. How many observations are there in the dataset? How many variables?

5. Examine the `polintr` variable, which measures political interest. Tabulate the variable:

   ```
   table polintr
   ```

   Then, replace the values of "Refusal", "Don't know", and "No answer" as missing values:

   ```
   replace polintr = . if polintr >= 5
   ```

   Repeat the tabulation to confirm that this work. Using what you have learned already, attempt to answer the following questions:

(a) What proportion respondents are "very interested" in politics?

(b) What is the *modal* response category for this question?

(c) What is the *median* level of interest?

(d) Using the gender variable (`gndr`), assess who is more interested in politics: men or women?

(e) Using a statistical test, assess whether the difference in political interest is statistically distinguishable from zero.

6. Now examine a different set of variables: happiness (`happy`) and religiosity (`rlgdgr`). First, remove the nonresponse categories from these variables:

```
replace rlgdgr = . if rlgdgr > 10
replace happy = . if happy > 10
```

7. How happy are people in Great Britain on average? How religious are they?

8. Draw a scatterplot of religiosity and happiness:

```
twoway scatter rlgdgr happy, jitter(4)
```

Looking at the plot, how correlated would you say these variables are? Does happiness go with religiosity?

9. Use `corr` to calculate the correlation coefficient for this relationship.

10. Recall that the correlation is simply a rescaled version of the covariation: it is the covariance of two variables divided by the product of their standard deviations. Calculate the standard deviation of each variable and — combined with the covariance you calculated above — use that to manually calculate the correlation coefficient for this relationship. Does it match the value returned by `corr rlgdgr happy, covariance`?

11. Recall that the correlation coefficient says nothing about the "size" of the relationship between two variables. Instead, it captures the degree to which that relationship is well-represented by a straight-line. If the correlation is high (close to -1 or +1), we know that the data can be well-represented by a line but we do not know how much of an increase we would expect to see in one variable given an increase in the other variable. "Regression" is a method for representing the covariance between variables in another way, which provides us with exactly this way of describing the relationship.

12. In regression, we draw a line through the scatterplot of the two variables that represents the "line of best fit" and that we can summarize by the "slope" of the line: the expected change in the y-axis variable given a one-unit increase in the x-axis variable. Use the following to see this line of best fit:

```
twoway (scatter rlgdgr happy, jitter(4)) (lfitci rlgdgr happy)
```

13. To estimate the value of the slope, all we need to know is that the slope of the line is a rescaled version of the covariance of the two variables. Unlike the correlation coefficient, however, to estimate the slope, we rescale the covariance by the variance of the x-axis variable. To do this calculation, we have to be a little careful, so let's drop the cases where religiosity and/or happiness are missing:

```
drop if rlgdgr == .
drop if happy == .
```

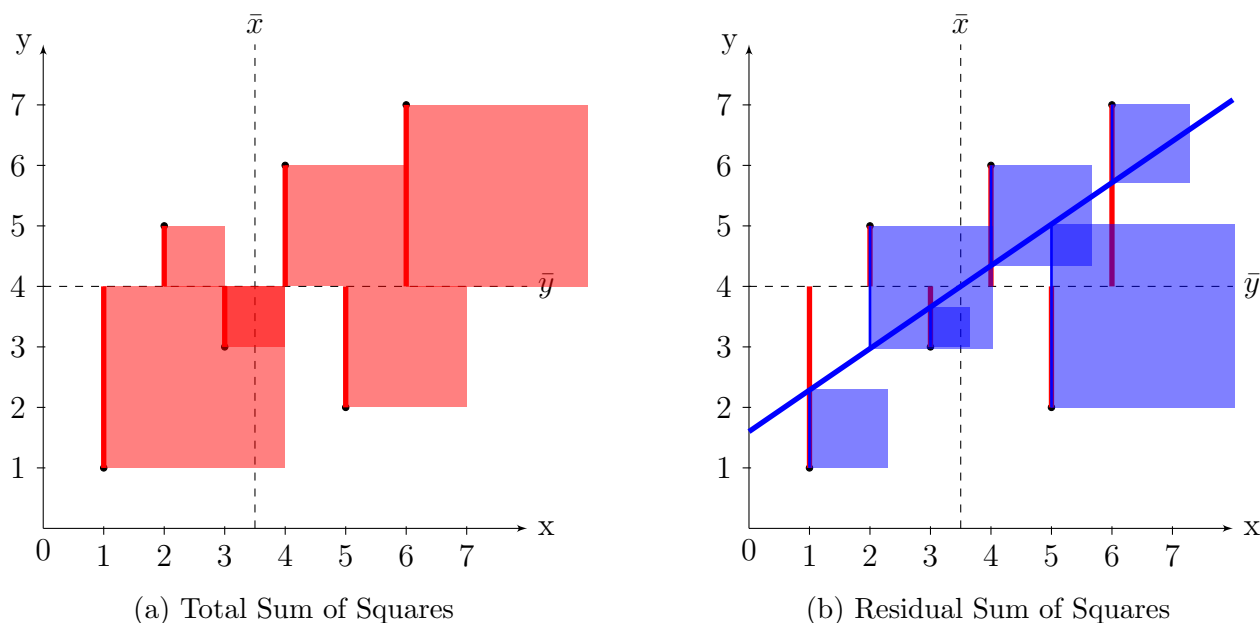(a) Total Sum of Squares   (b) Residual Sum of Squares

Figure 1: Sum of Squares for a Simple Scatterplot

How many observations did we lose by removing these missing values?

14. Now, using the reduced dataset, calculate the covariance of the two variables. Then, calculate the variance of `religiosity`. Divide the covariance by the variance of religiosity. This is the slope of the regression line. How much happier is someone if they were to increase their religiosity by one point on this question scale? Does this correspond with the relationship in your scatterplot; can you see the slope of the line matches this calculated slope?

15. We do not have to manually calculate the slope in this way. Indeed, Stata provides a command called `reg` that allows us to calculate this directly. In Stata, run the following to estimate this regression model. View the results.

    ```
    reg happy rlgdgr
    ```

    Does the slope match what you calculated by hand?

16. Another way of thinking about regression is in terms of the "sum of squares." We can think of the "sum of squares" as the total amount of variation in the y variable (happiness). You can calculate the total sum of squares for our data using:

    ```
    summ happy
    gen happy_resid = (happy - r(mean))^2
    quietly summ happy_resid
    disp r(sum)
    ```

    This quantity (the total sum of squares) is literally a sum of the area of several squares. Figure 1(a) shows this for a simple example scatterplot. The total sum of squares is simply the sum of the individual deviation of each observed value of y from the mean of y, squared.

17. In regression, we are "explaining" some of that total variation in y with the covariation between x (religiosity) and y (happiness). This explained portion of the sum of squares is called the "sum of squares, explained" or "model sum of squares"; the amount of variation in y that is not "explained" by x is called the "residual sum of squares". This residual sum of squares is literally the sum of the squared areas formed by the distance between each point and the "predicted value" (or "fitted value") of y expected by the regression line given that observation's value of x. Figure 1(b) shows the residual sum of squares for the same example scatterplot.

3

18. If all of the variation in y is explained by the variation in x, then the residual sum of squares will be *small* while the explained sum of squares will be large. You can calculate the residual sum of squares using the following:

```
reg happy rlgdgr
gen happy_resid2 = (happy - (_b[_cons] + (_b[rlgdgr] * rlgdgr)) )^2
quietly summ happy_resid2
disp r(sum)
```

How large is the residual sum of squares compared to the total sum of squares you calculated above?

19. A common way to assess the "goodness of fit" of a regression estimation is whether the explained sum of squares is large. The proportion of the total sum of squares that is "explained" by the model is called the "R-squared" statistic, sometimes called the "coefficient of determination." It is simply 1 minus the ratio of the residual sum of squares to the total sum of squares. It is interpreted as a proportion: if it is 0, then religiosity explains *none* of the variation in happiness; if it is 1, then religiosity explains *all* of the variation in happiness. What is the R-squared for this relationship?

20. Recall from earlier that the regression slope is simply a scaled version of the covariance of the two variables. The correlation coefficient is also a scaled version of the covariance. Now, you can also note that there is a direct correspondence between the correlation coefficient and the R-squared value you just calculated. R-squared is simply the bivariate correlation coefficient, squared. Try it and confirm this is the case.

21. This focus on the residuals allows us to demonstrate a few Stata programming features that have useful practical applications. After running a regression using `reg`, several commands will generate useful information:

    (a) `predict FITTED` will create a new variable `FITTED` containing the "fitted value" of y for each observation (given its value of x)

    (b) `predict RESIDUALS, residuals` will create a variable `RESIDUALS` containing the residuals for each observation in your dataset. These residuals will be equal to `happy - FITTED`

22. Returning to our main focus, another common measure of model fit is called the residual mean squared error, or sometimes simply $\hat{\sigma}$. This is simply the standard deviation of the residuals (the unexplained variation in y). As religiosity explains more and more of the variation in happiness, then the residuals will become small and have a lower variance (and lower standard deviation).[1] There are three ways to calculate sigma: (a) by manually calculating the residuals and taking the standard deviation thereof, (b) extracting the residuals from the regression object (returned by `reg`), or (c) using Stata's calculation of sigma in the raw regression output:

```
gen happy_resid3 = (happy - (_b[_cons] + (_b[rlgdgr] * rlgdgr)) )^2
quietly summ happy_resid2
disp sqrt(r(sum)/(_N-))

predict RESIDUALS, residuals
summ RESIDUALS
```

The reason this quantity is useful as a measure of model fit is that it is on the scale of the y-axis variable (happiness) and can be directly compared to the standard deviation of y. If religiosity explained a lot of the variation in happiness, then $\hat{\sigma}$ would be small compared to the standard deviation of `happy` (returned by `summ happy`). Is it?

---

[1]Recall the standard deviation is simply the square root of the variance.

23. Returning to our estimated regression model, use `reg happy rlgdgr` to see the regression results again. Can you make sense of all of the information that is displayed? What are the values of the following quantities:

   (a) Regression slope

   (b) $\hat{\sigma}$

   (c) R-squared

   (d) y-intercept or constant (the value of happiness when religiosity = 0)

24. You will note that the table includes a number of other statistics that we haven't discussed, including the standard errors of the regression slope and y-intercept, and t-statistics and p-values for each of these. We are not going to cover the calculation of the standard errors, but can you see how the t-statistics are calculated given what we covered in Lab 2?

25. Is the estimated regression slope substantively large? Is the slope significantly different from zero?

26. In order to ensure your understanding, you may want to redo some of these exercises using other combinations of variables (e.g., political interest, etc.). One particularly illustrative example would be to use gender as an x-axis variable. (You will want to redefine `gender` as:

```
gen gender = gndr - 1
```

In that case, you will hopefully note some similarities between the conditional mean of an outcome variable given gender:

```
tabstat polintr, by(gender)
```

and the results of a corresponding regression analysis:

```
reg polintr gender
```

27. Recall that the p-value in a statistical significance test represents the probability of observing a t-statistic as large or larger than the one observed in a sample of the data from a population where the true t-statistic is zero (in this case, where the slope of the relationship between religiosity and happiness is non-existent).

As an **advanced exercise**, you emulate that process of repeated sampling through a process known as "bootstrapping". In bootstrapping, you treat your sample data as if they were a population and draw repeated samples (with replacement) from your sample and apply an estimator (in this case, a regression estimator) to each resampled dataset. In essence, bootstrapping creates an approximate sampling distribution for your statistic.

If you extract the estimated slope from each bootstrap sample, this bootstrap sampling distribution mimics the hypothetical distribution of slopes you would expect to observe in a population where the true slope was equal to that observed in your actual sample data. The proportion of slopes less than zero in this bootstrapped distribution can be interpreted as the p-value (the probability of observing a slope as large as the one you observed) from a population where the true slope was zero.

Here's the code:

```
reg happy rlgdgr, vce(boot)
```

What is p-value of the regression slope, based on the bootstrap sampling distribution?
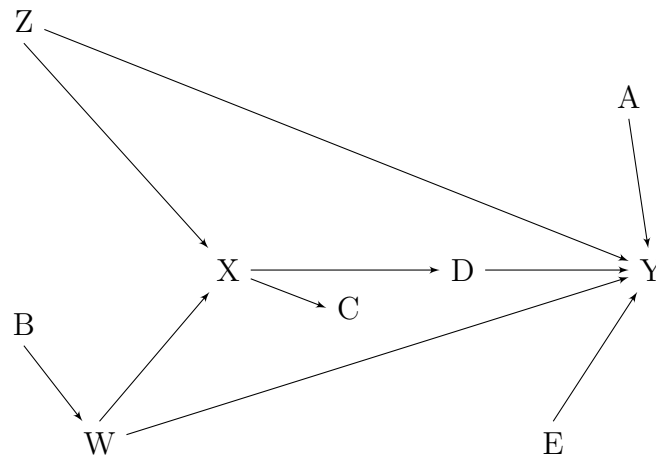
Figure 2: A Causal Graph

3.2 Multivariate Regression

1. Think again about political interest. How correlated are interest, happiness, gender, and religiosity? If they are correlated, can we interpret those correlations as causal relationships? What criteria for causal inference to we have to satisfy in order to interpret this correlation as a causal relationship?

2. One major barrier to causal inference here is confounding: the relationship between these two factors may be explained by some other variable or variables. We might be concerned that there is some earlier variable, that precedes both religiosity and (current) happiness in time that causes both variables, and in turn makes it appear that the two variables are related due to a causal relationship between religiosity and happiness.

   To think about this possibility, it can be useful to draw a "causal diagram" in the form of a "directed, acyclic graph" (DAG). Figure 2 shows an example DAG. We can imagine that religiosity is **X** and happiness is **Y**. We are concerned bout variables like **Z** and **W** that cause both. Draw a possible DAG, labeling possible confounding variables. (You might also want to include other variables in your graph, like **A** and **E** that cause happiness but that you do not think cause religiosity.)

3. Once you have a reasonable DAG showing possibly causal relationships, think about whether you can observe any of the variables in your DAG. Are all of the potentially confounding variables you named available in the ESS dataset? You can find a complete listing of variables by typing `codebook` or looking in the complete documentation of the ESS questionnaire here: `http://www.europeansocialsurvey.org/docs/round7/fieldwork/united_kingdom/ESS7_questionnaires_GB.pdf`.

4. If you have access to potentially confounding variables, then we can use an elaboration of the bivariate regression models we just discussed to help to control for this confounding. If potentially confounding variables in your DAG are unobserved (either because they are not included in the ESS or because they are perhaps "unobservable" factors), there is nothing we can do to rule out the potential for confounding. We are, in that situation, completely unable to draw a causal inference from the observational data in front of us. We would need more data!

5. How can multiple regression help us to address confounding? The mathematics here is a bit complicated, but the following exercises are meant to provide some intuition. Before we begin, make sure you have a firm grasp on the variables you have identified. Use `summ` or other commands to know the key statistics of each variable and `twoway scatter` to produce scatterplots or other graphics comparing these variables to one another.

6. Multiple regression works by simultaneously accounting for the influence of multiple "right hand side" (RHS) variables on a single "outcome" variable ($\mathbf{Y}$ in the example DAG in Figure 2). How does it do this? Let's start with a simple example (which you can expand using the variables in your DAG). In the below model we are trying to see whether there is still an association between religiosity and happiness once we "control for" interest in politics:

```
drop if happy == .
drop if rlgdgr == .
drop if polintr == .
drop if gndr == .
reg happy rlgdgr polintr
```

Basically, this model says that we think happiness might be a function of religiosity, political interest, and gender, and we want to know how much each contributes to happiness. By including all variable in one regression equation, we estimate the influence of each variable on the outcome accounting for the simultaneous influence of the other RHS variables. Is there still a relationship between religiosity and happiness once we account for these other factors?

7. Try estimating bivariate relationships between each of these RHS variables and the outcome. How do the estimated slopes in the bivariate models compare to the estimated slopes in the multivariate specification?

```
reg happy rlgdgr
reg happy polintr
```

8. When we regress an outcome on a variable, we aim to *break* the correlation between the residuals (the unexplained parts of y) and the RHS variable. TO see this, look at the correlations between the residuals from each of the bivariate models and the RHS variable we included:

```
quietly reg happy rlgdgr
predict residuals1, residuals
quietly reg happy polintr
predict residuals2, residuals
corr residuals1 rlgdgr
corr residuals2 polintr
```

This correlation is basically zero in each case.

9. But when we exclude a RHS variable (leaving it out of the regression equation), the residuals of y might still be associated with that variable:

```
corr residuals1 rlgdgr polintr
corr residuals2 rlgdgr polintr
```

10. We can see that the excluded variables appear to still be associated with the outcome (i.e., there is some residual some of squares that is still associated with the excluded RHS variables). If we include these, we can account for the independent influences of each of these factors. But how does this work? To understand multiple regression we have to understand a simpler concept: the *partial correlation.*

A partial correlation is the correlation between two variables controlling for the influence of a third variable. If we want to know the partial correlation between `happy` and `rlgdgr`, controlling for `polintr`, we regress each variable on `polintr` and correlate their residuals:

```
quietly reg happy polintr
predict residuals3_happy, residuals
quietly reg rlgdgr polintr
predict residuals3_rlgdgr, residuals
corr residuals3_happy residuals3_rlgdgr
```

This removes the common influence of the third variable from the original correlation:

```
corr happy rlgdgr
```

which in this case is small.

11. Multiple regression is simply a series of partial correlations. If we want to know the slope of the relationship between religiosity and happiness, we simply regress the residuals of the regression of `happy` on `polintr` on the residuals from the regression of `rlgdgr` on `polintr`:

```
reg residuals3_happy residuals3_rlgdgr
```

Similarly, if we want to know the influence of political interest controlling for religiosity, we do something analogous:

```
quietly reg happy rlgdgr
predict residuals4_happy, residuals
quietly reg polintr rlgdgr
predict residuals4_polintr, residuals
reg residuals4_happy residuals4_polintr
```

Note how the slope for `polintr` and the slope for `rlgdgr` in the previous two partial correlation models are identical to the slopes from a multivariate regression of `happy` on both `polintr` and `rlgdgr`:

```
reg happy polintr rlgdgr
```

All a multiple regression model is doing is engaging a series of simultaneous regressions of the outcome on each RHS variable and each RHS variable on all the other RHS variables.

12. Everything that we learned above about bivariate regressions — residuals, fitted/predicted values, measures of model fit, statistical significance, etc. — applies in a multiple regression context. Explore some of these aspects of the multivariate regression models we have just estimated.

13. The only thing that is different about bivariate and multivariate regression is the substantive interpretation of estimated regression slopes. Rather than talk about these slopes as the "influence of X on Y" we instead interpret them as the "influence of X on Y controlling for Z" or the "influence of X on Y all else constant". This is a subtle change but it's important to remember that in a multivariate regression we are not estimating a *raw* relationship but rather than relationship accounting for the variance in X and Y explained by the other RHS variables in the model. For this reason, the slopes in a multivariate regression are sometimes called "marginal effects" due to their representing the change of influence of one variable while the other variables remain constant.

3.3 Survey-Weighted Analyses

14. Now that you have a sense of how to estimate and interpret regression models, we need to handle a very important final caveat to all of the above. When we work with survey data, we need to acknowledge that such data are often *weighted*. That is, they do not represent a simple random sample from a population but instead come from a more complex survey design and the *representativeness* of the sample data hinges on the estimates being re-weighted to make them match the target population.

15. The easiest way to do this is to use the `svyset` command to instruct Stata that the dataset should be weighted. The specific of this command are pretty complex but for our purposes we need to tell stata, via this command, that the `pspwght` weight variable contains the weights that make our sample representative.

    A simple example of this is as follows:

    ```
    * set survey weighting globally
    svyset [pweight=pspwght]

    * calculate unweighted means
    mean happy polintr rlgdgr

    * calculate weighted means
    svy: mean happy polintr rlgdgr
    ```

    How do the unweighted and weighted means differ?

16. The difference is that in the unweighted analysis, every row in the dataset counts equally whereas in the weighted analysis, some rows count for more than others. Individuals that are underrepresented in the sample relative to the population are given weights larger than 1 and individuals that are overrepresented in the sample relative to the population are given weights smaller than 1. You might want to examine the distribution of weights using some of the following:

    ```
    hist pspwght
    summ pspwght

    tabstat pspwght, by(gndr)
    tabstat pspwght, by(region)
    tabstat pspwght, by(yrbrn)
    ```

17. For regression analysis, we can either use `svyset` to globally specify that the data are supposed to be weighted. Or we can the `[pweight=...]` option to `reg`:

    ```
    ** unweighted command behaves as normal
    reg happy polintr rlgdgr

    * survey-weighted model
    reg happy polintr rlgdgr [pweight=pspwght]

    ** 'svy:'-prefixed command provides weighted results
    svy: reg happy polintr rlgdgr
    ```

    The last two sets of results should be basically identical, but both should differ from the unweighted regression.