An Example of Applied Multiple Regression Using Stata

Professor M.A. Golden
Teaching Assistants Chao-yo Cheng and Zsuzsanna Blanka Magyar
Prepared for Political Science 167D, Winter 2016
Department of Political Science
University of California, Los Angeles

April 19, 2016

1 Sample problem

- Question: What is the evidence that better economic growth reduces political instability?
- Instructions: Analyze the data to answer the question and include a measure of ethnic fractionalization as a control variable.
- Data source: David Weil, *Economic Growth*, 3^{rd} edition.
- Data set: hw3 (course website)

2 Examination of the Data

- Open the dataset
- Type your equivalent path for:
  `use /mg/teaching/ps167d_win16/datasets/hw3.dta, clear`
- Scan the list of variables contained in the dataset
- What measure of political instability is included in the dataset?
- `codebook instability`
- How is this measure constructed?
- Consult the textbook [hint: in the index, look for *sociopolitical instability*]
- What years does the measure reflect? What phenomena are captured by it?
- What does the distribution of countries look like?

- Type

  ```
  summarize instability, detail
  ```

  to get:

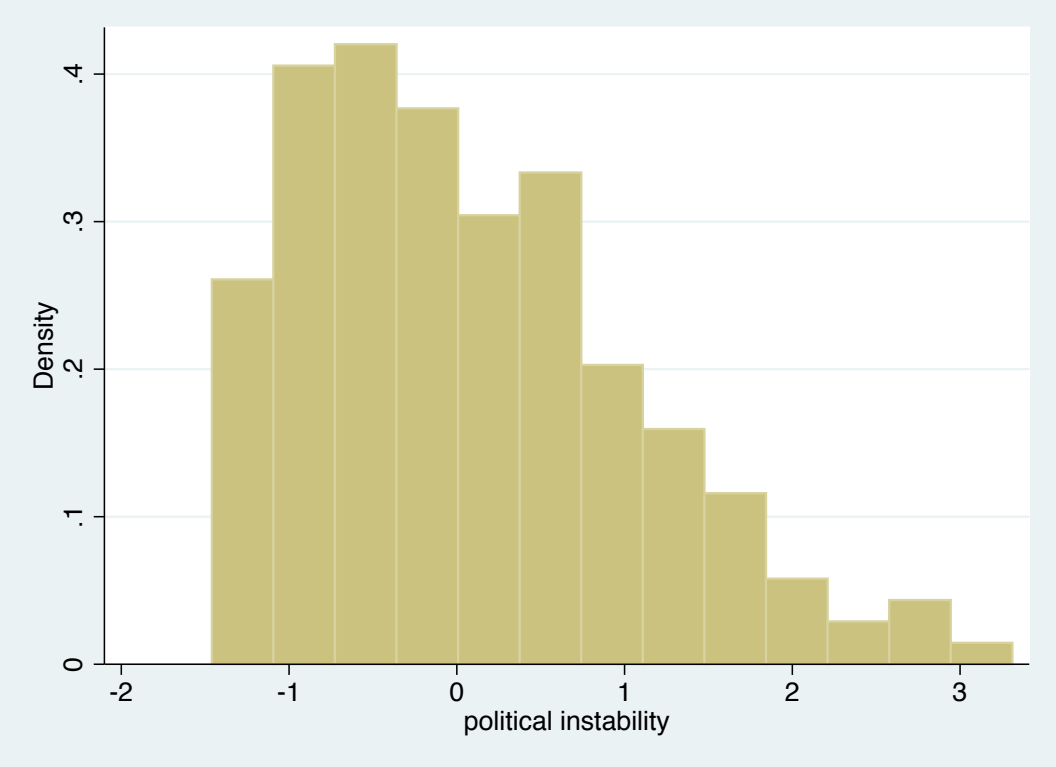
```
              political instability
-------------------------------------------------
      Percentiles      Smallest
 1%    -1.46044       -1.46045
 5%    -1.213648      -1.46044
10%    -1.062531      -1.46          Obs             188
25%    -.7153423      -1.449893      Sum of Wgt.     188

50%    -.0711304                     Mean      .0733512
                      Largest        Std. Dev.   .9854371
75%     .6765749      2.646986
90%    1.450511       2.753124       Variance       .9710862
95%    1.897404       2.7563         Skewness       .7139888
99%    2.7563         3.311713       Kurtosis       3.098159
```

  From the output, we can see that the median (that is, the 50% percentile) is -.071, which is close to the mean of .073. This probably means the variable does not have a skewed distribution that we need to worry about, but to be sure, let's look at a histogram.

- Type

  ```
  histogram instability
  ```

  to generate the following output:



This variable is distributed in kind of a weird way — it doesn't look exactly like a normal distribution, but it's probably not skewed enough to merit a logarithmic transformation. We need to look at the distribution of every variable we will use in order to make the best judgment about when to use logarithms and also to get a sense of the data we're using.

- Now review the other two variables that you will include in your analysis in identical fashion.
- What measure of economic growth is included in the dataset?
- What measure of ethnic fractionalization is included in the dataset?
- How are these variables measured? What years do they reflect? How is the data distributed? Is the distribution close enough to normal or do you need to transform either of them?
- How are all three variables ordered? What do higher and lower values mean?

3 Initial Data Analysis

Let's begin the analysis by examining the correlation of instability and growth. Correlations give us a sense of the strength of the relationship. The output of `corr instability gy` looks like this:

```
(obs=156)

             | instab~y   gy7509
-------------+------------------
 instability |   1.0000
      gy7509 |  -0.3584   1.0000
```

Note that the order of the variables does not matter in a correlation. (*Why not?*)

- Why is the sign negative?

  To answer that question, we have to think back to how our variables were measured. Higher values of growth rates mean faster growth. Higher values of instability mean more instability. So a negative correlation means that as growth increases, instability falls. Is this what we expected?

- How do we interpret the correlation coefficient?

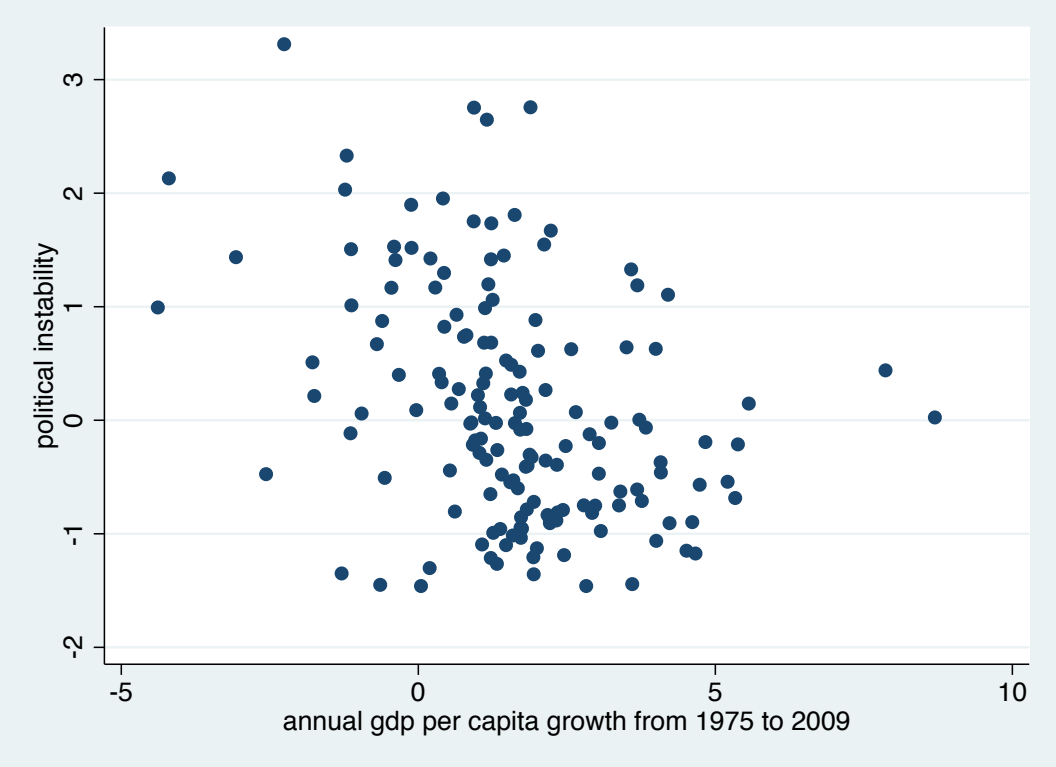
  A value of $.36$ is moderately strong. An r of $.36$ results in an r^2 of $.12$, which implies that 12 percent of the variation in the instability data is explained (linearily) by economic growth. This means that if you guess instability for a country knowing its annual average growth rate over the period 1975 to 2009, the variance of your errors will be smaller by 12 percent than if you just guess that every country has the mean amount of instability without looking at its growth rate.

- The correlation coefficient suggests that we will need control variables in a regression model, but it also documents that there is an obvious relationship between growth rates and political instability.

- Now let's generate a scatterplot of the relationship. In a scatterplot, as in regression, the order of the variables is important. We always put our dependent variable before our independent variable, so that the dependent variable appears on the y axis. Let's type

  ```
  scatter instability gy
  ```

  Here is the output that Stata produces:



There is an obvious relationship that appears in the data: we can visualize the negative line from high instability/low growth to low instability/high growth. But the relationship does not appear very strong, and there is clearly a lot of variance in political instability that is not connected to growth rates. We know this because the country points are very diffused.

4 Bivariate Regression Analysis

- Following our preliminary data analysis, we move to formal regression analysis. Let's begin with a bivariate regression that recapitulates the scatterplot but in a more precise way. We tell Stata

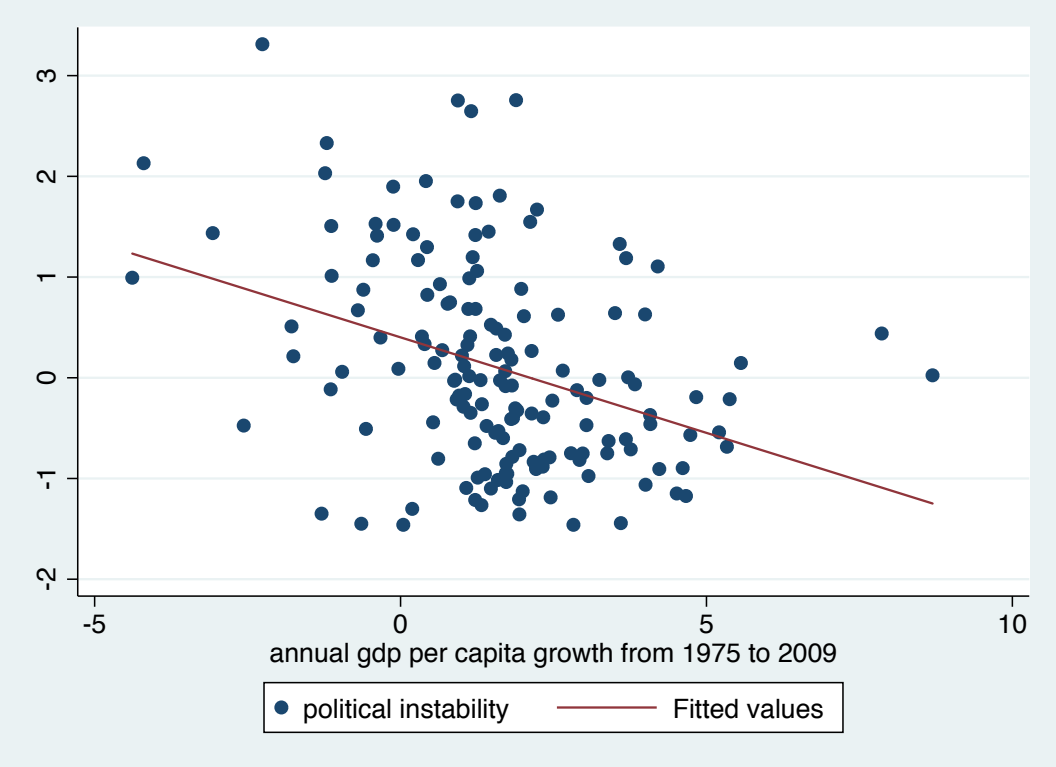
  ```
  reg instability gy
  ```

  and here is the output:

```
. reg instability gy

      Source |       SS           df       MS      Number of obs   =       156
-------------+----------------------------------   F(1, 154)       =     22.69
       Model |  20.5722244         1  20.5722244   Prob > F        =    0.0000
    Residual |  139.608517       154  .906548813   R-squared       =    0.1284
-------------+----------------------------------   Adj R-squared   =    0.1228
       Total |  160.180741       155  1.03342414   Root MSE        =    .95213

------------------------------------------------------------------------------
 instability |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gy7509 |  -.1896942   .0398207    -4.76   0.000    -.2683595   -.1110288
       _cons |   .4013789   .0996595     4.03   0.000     .2045026    .5982551
------------------------------------------------------------------------------
```

A bivariate regression estimates the linear relationship between two variables: one cause and one effect. The given slope is the slope of the regression line (here it's negative, and its value is $-.1896942$) and the given intercept (in this case, $.4013789$) where $x = 0$. The intercept is given in the regression results by the value of the constant.

- Let's examine the regression line, which is the best fit line between values of x (in this case, growth rates) and values of y (in this case, political instability) for our dataset.



As you can see, the graphic is identical to the scatterplot already produced except that we have added the line of best fit. The addition of that line helps us visualize the strength of the relationship between the independent and dependent variables. You can see that, although the line is clearly sloping downwards, lots of country-dots are far from the line. This suggests a lot of unexplained variance in the dependent variable.

- Now let's interpret the bivariate results. In this model, the annual average rate of economic growth between 1975 and 2009 is negatively related to political instability, and the relationship is statistically significant at the $p = .01$ or smaller level. We know this from the negative sign on the coefficient for growth, and from the the p-value.
- The overall amount of variance explained by growth rates is 12.8 percent. This is the r^2 in the regression. Is this a lot or a little? It's not a lot, but is an acceptable amount to move forward with an interpretation; that is, it's not so little that you think you are entirely on the wrong track.
- Recall, by the way, that the r^2 is simply the square of the correlation coefficient (i.e. the r), which we discussed above.
- Let's add more substance to our interpretation. A one-percent change in the growth rate is associated with a $.18969$ unit reduction in instability, on average. We talk about the growth rate in percentages because those are the units in which we measure growth rates. We talk

about instability as "units" because there is no natural interpretation to the units used to measure this variable. In general, we have to talk about each variable with whatever metric is used to measure it.

- Substantively, we want to understand the magnitude of the effect. We want to know whether the effect is substantively large or small, separately from whether it is statistically significant. Even a statistically significant result may have a very trivial substantive effect.

- To unpack this, we are going to use the range of observed values of our independent and dependent variables to calculate the maximum effect that observed growth could have on instability in our dataset. The regression results provide information about the average effect of growth on instability for every x value. But we might also want to know how large the change in y is from the minimum to the maximum observed values as a function of the minimum and maximum observed values of the change of x in the dataset. That is, we want to know how much of the range of y is statistically explained by the range of x in our dataset.

- Here's how to do the calculations. We begin by considering the range of each variable. Our dependent variable, instability, ranges from -1.46 to 3.31, and our independent variable, growth rates, ranges from -4.38 to 8.70.

  Now let's recall the basic regression equation:

  $$\hat{y} = \alpha + \beta * x$$

  The bivariate regression output includes values for both unknowns in the equation, so we can predict the maximum change that growth can generate in political instability. If growth is at its maximum value of 8.70, then:

  $$\hat{y} = .40 + (-.19) * 8.7$$

  and therefore, $\hat{y}$, or the level of predicted instability, is equal to -1.25.

  If growth is at its minimum value of -4.38, then:

  $$\hat{y} = .40 + (-.19) * -4.38$$

  and therefore, $\hat{y}$, or the level of predicted instability, is equal to 1.23.

  This means that the maximum change in instability generated by growth is 1.23 - (-1.25), or the difference between the minimum and maximum values of instability that are associated with the observed range of growth rates. This difference in instability is equal to 2.48, which is large; for instance, it is equal to $2.48/4.77$ — the former is the range of *predicted* instability and the latter is the range on *observed* instability. This means that 52 percent (which is what you get if you calculate $2.48/4.77$) of the observed range in instability is explained when we move the growth rate from the minimum to the maximum. One way to think about this is that going from the maximum to the minimum in the growth rate is associated with a 52 percent reduction in the observed range of instability.

- Let's also investigate a few countries. We sort our data on instability, using

  ```
  sort instability
  ```

  We then list our countries by instability, using

  ```
  list country instability
  ```

  We observe that Switzerland is measured on instability at about the minimum of -1.25 and that Kenya is close to the maximum instability measure of 1.23. This suggests that increasing growth from the minimum to the maximum would make Kenya as stable a polity as Switzerland! That would be a huge improvement in political stability for Kenya. (Likewise, if the reverse were to occur, it would be a massive deterioration of political stability for Switzerland.)

5 Multivariate Regression Analysis

- Usually we need to include control variables in a regression equation. We need proper control variables to avoid drawing inaccurate conclusions. If we omit a necessary control variable, even the sign on the variable of interest could be wrong!
- Usually, there are *many* causes of important outcomes, and control variables also allow us to separate out the marginal effect of each individual variable.
- In this example, we are going to re-estimate the effect of growth on political instability controlling for ethnic fractionalization. The regression we run is therefore:

  ```
  reg instability gy ethnic
  ```

  and the output looks like this:

```
. reg instabil gy eth

      Source |       SS           df       MS      Number of obs   =       119
-------------+----------------------------------   F(2, 116)       =     15.76
       Model |  23.5917089         2  11.7958544   Prob > F        =    0.0000
    Residual |  86.8148792       116  .748404131   R-squared       =    0.2137
-------------+----------------------------------   Adj R-squared   =    0.2001
       Total |  110.406588       118  .935649052   Root MSE        =     .8651

------------------------------------------------------------------------------
           instability |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------------+------------------------------------------------------------------
                gy7509 |   -.116276   .0525486    -2.21   0.029    -.2203551    -.012197
ethnicfractionalization |   1.231711   .3307438     3.72   0.000     .5766313    1.886791
                 _cons |  -.1955526    .213331    -0.92   0.361    -.6180816    .2269763
------------------------------------------------------------------------------
```

- Our interpretation proceeds as in the bivariate example, except that now each of the two independent variables is estimated holding the value of the other constant. Let's think of it

this way: ethnic fractionalization is a confounding variable that may affect *both* growth and instability. We want to know the true impact of growth on instability after we purge both variables of any impact of ethnic fractionalization.

- We begin by examining the coefficient on growth. It's still negative, and now it's $-.1163$, and it's still statistically significant. This is the size of the effect of growth on instability holding ethnic fractionalization constant. We call that the marginal effect of growth.

- Our interpretation of the constant ($x = 0$) is now slightly changed. The constant is now the average value of y when $x = 0$, for both x_1 and for x_2; that is, both regressors are equal to 0.

- The overall amount of variance explained by growth rates and ethnic fractionalization is 21 percent. We have improved the r^2 with the addition of ethnic fractionalization in the regression. Adding more variables almost always improves the r^2 just because the model can find more ways of predicting the right outcome. Sometimes this does not represent a genuine improvement in the model. In this case, we have theoretical reasons to believe that ethnic fractionalization probably affects growth rates and political instability, so the addition of the variable can be justified on those grounds.

- Now we want to repeat the exercise already performed above and see how the range on x affects y, but in this context we need to figure out what to do with our control variable. In this example, we are going to hold it to its average value. This is a convenient practice. In the particular case of linear models, it also doesn't matter what value we hold the control variable to: the predicted change in the outcome due to a change in x_1 is the same regardless of what value we chose for x_2, because all of the "effect" of our x on y is capture by β.

- The observed ranges on our variables of interest of course don't change because we are still working with the same dataset. Our dependent variable, instability, ranges from -1.46 to 3.31, and our independent variable, growth rates, ranges from -4.38 to 8.70. Our regression equation is slightly different, however, because it includes a second regressor:

  $$\hat{y} = \alpha + \beta_1 * x_1 + \beta_2 * x_2$$

  The regression output includes values for all unknowns in the equation, so we can predict the maximum change that growth can generate in political instability at the average value of ethnic fractionalization (which we know when we ask Stata to summarize the variable is $.4468698$). If growth is at its maximum value of 8.70, then:

  $$\hat{y} = -.20 + (-.12) * 8.7 + 1.2 * .45$$

  and therefore, $\hat{y}$, or the level of predicted instability, is equal to $-.70$.

  If growth is at its minimum value of -4.38, then:

  $$\hat{y} = -.20 + (-.12) * -4.38 + 1.2 * .45$$

  and therefore, $\hat{y}$, or the level of predicted instability, is equal to $.87$.

This means that the maximum change in instability generated by growth is .87 - (-70), or the difference between the minimum and maximum values of instability that are associated with the observed range of growth rates holding ethnic fractionalization at the mean. This difference in instability is equal to 1.57, which is large; for instance, it is equal to $1.57/4.77$ — the former is the range of predicted instability and the latter is the range on observed instability. This means that 33 percent (which is what you get if you calculate $1.57/4.77$) of the observed range in instability is explained when we move the growth rate from the minimum to the maximum, holding ethnic fractionalization at the mean.

- Note that now that we are controlling for ethnic fractionalization, the substantive magnitude of the effect of growth on instability is smaller. This is because previously some of the variation in instability that is caused by ethnic fractionalization was wrongly attributed to growth.
- Again, we have to ask ourselves if the impact is large or small. If we can explain 33 percent of the range in instability moving growth from the minimum to the maximum (holding ethnic fractionalization at its mean), is this a lot? It seems reasonable to argue that it is. But keep in mind that you have to explicitly offer this interpretation. You could again look up specific countries to make it easier for your reader to grasp the importance of your finding.