Assignment 2: The OLS Estimator

PSCI 8357, Spring 2016
January 21, 2016

This assignment must be turned in by the start of class on **Thursday, January 28**. You must follow the instructions for submitting an assignment.

Data Analysis

You will work with two datasets:

- `wdi-2010.csv`: A small selection of country-year data for 2010 from the World Bank's World Development Indicators dataset. Missing values are indicated by the string `".."`.
- `polity4.sav`: The most recent release of the Polity IV dataset. It is distributed in SPSS format, so you will need to use `read.spss` from the **foreign** package to load it into R.

Your first task is to merge the `polity2` variable (for the year 2010) from the Polity IV data into the WDI data. `polity2` is the difference between the Democracy and Autocracy scores, ranging from -10 to 10. You will find the **countrycode** package useful here. Make sure to drop all countries with a population below 500,000 before merging, since these countries are not covered by the Polity project.

Next, use the merged data to perform the following tasks:

1. Regress infant mortality on Polity score. Substantively interpret your findings.

2. Regress infant mortality on both Polity score and GDP per capita. Interpret your findings. How does including GDP per capita in the model change your conclusions about the relationship between regime type and infant mortality? What do you think accounts for the difference?

3. Now flip it over: regress Polity score on infant mortality. Interpret your findings. You will notice that, even accounting for the fact that we've

flipped the axes, the regression line is not the same as in the first part. Why not?

As always, you must follow the best practices we talked about in the first week of class. Be particularly careful about missing data here: note what is missing, how you choose to deal with missingness in your analysis, and how the missingness might limit your findings.

Simulation Study

You will draw data according to the following model. The sample size is $N = 50$. There are two covariates drawn from a joint normal distribution,

$$\begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right).$$

You can use the `mvrnorm()` function in the **MASS** package to draw from a joint normal distribution. The expected value of the response is a linear function of the covariates,

$$Y_i = 1 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where the error term is drawn independently of the covariates from a standard normal distribution, $\epsilon_i \sim N(0, 1)$.

Use simulations to answer the following questions:

1. Assume that $\beta_1 = -2$ and $\beta_2 = 0$, so only the first covariate is directly related to the response. For each value of

   $$\rho \in \{-0.75, -0.50, -0.25, 0, 0.25, 0.50, 0.75\},$$

   compare the approximate standard error of the estimated coefficient $\hat{\beta}_1$ from OLS just on X_1 and from OLS on both X_1 and X_2. Which estimator is better? How does the difference in the standard errors vary with the amount of correlation between the covariates? Give an intuitive explanation of the results.

2. Now fix $\rho = 0.75$, but assume $\beta_2 \neq 0$. We will compare the same two estimators as in the first part, but now OLS just on X_1 is biased. So instead

   of comparing the standard errors, you will compare their mean squared error:

   $$\text{MSE} = E[(\beta_1 - \hat{\beta}_1)^2] = \text{Bias}^2 + \text{Std. Error}^2.$$

   How big does β_2 have to be for it to be better, in terms of MSE, to include both variables instead of just X_1 in the regression?

3. On the basis of these results, what advice would you give a political scientist who is deciding whether to include one or both of two highly correlated variables in a regression model?

Remember to follow good coding practices in running your simulations. The less code you reuse, the better. You may find it useful to write functions, which let you run the same operations on different input values. For example, the following function computes the difference in standard deviations between two vectors:

```r
sd_diff <- function(x, y) {
    sd_x <- sd(x)
    sd_y <- sd(y)

    sd_x - sd_y
}

sd_diff(1:5, 0:10)
```

```
## [1] -1.7355
```

```r
sd_diff(1:5, 1:5)
```

```
## [1] 0
```

Weekly Visualization Challenge

How does the relationship between GDP per capita and regime type differ across continents? (Where "continents" is defined as in the `continent` variable in the **countrycode** data.)