PLSC 503: "Multivariate Analysis for Political Research"

Exercise Three
February 10, 2025

Part I

We're going to investigate what effect "dummying out" a continuous variable has on our conclusions in a linear regression model. Consider a simple model like this:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

where $u_i \sim N(0, 1)$ and

$$X_i \sim \begin{cases} \phi_0 \text{ with probability } P_i \\ \phi_1 \text{ with probability } (1 - P_i) \end{cases}$$

where $P_i \in [0, 1]$ and ϕ_0 and ϕ_1 are normal densities with means of 0 and 1, respectively, and variances $\sigma^2_{0,1}$ as discussed below.[1] You'll be considering models where X_i is replaced by D_i,

$$Y_i = \gamma_0 + \gamma_1 D_i + e_i \tag{2}$$

where

$$D_i = \begin{cases} 0 \text{ if } X_i \leq 0.5 \\ 1 \text{ otherwise.} \end{cases}$$

Your assignment:

1. First, demonstrate (trivially) that for any value of P, when $\sigma^2_0 = \sigma^2_1 = 0$ the estimates from (1) and (2) are identical.

2. For $P = 0.5$, discuss the effect on parameter estimates ($\hat{\beta}_1$ vs. $\hat{\gamma}_1$, and, if relevant, $\hat{\beta}_0$ vs. $\hat{\gamma}_0$), standard errors, model fit, etc. of symmetrically increasing σ^2_0 and σ^2_1 (say, from $\sigma^2_{0,1} = 0.0001$ to $\sigma^2_{0,1} = 1$).

3. Similar to part 2, discuss what happens if, for a given value of $\sigma^2_{0,1}$ (say, $\sigma^2_{0,1} = 0.1$), P is varied from 0 to 1 with D defined as above.

[1] This is known as a *mixture distribution*, specifically a mixture of two normals. You'll know you got it right if the histogram of the resulting X is bimodal for small values of $\sigma^2_{0,1}$ and values of P not close to 0 or 1. ***Hint***: `rbinom()` is a great way to draw binary values with $\Pr(P_i) = P$.

Part II

We're once again dipping into the World Development Indicators (WDI) data well, this time to model nations' environmental impact. More specifically, the data are from the year 2018, and the assigned outcome of choice is each country's CO2 emissions, measured in metric tons per capita (`CO2Emissions`). You'll be asked to regress those emissions data on several variables, also drawn from the WDI data:

- `PopDensity` is the country's population density, measured as people per square kilometer of land area;
- `MilitaryExpenditures` are exactly that: the country's expenditures on the military, as a percentage of GDP;
- `GDPPerCapita` is the by-now-probably-familiar measure of wealth, gross domestic product per capita, measured in constant (2015) U.S. dollars;
- `RuleOfLawIndex` is an index of the rule of law, drawn from the World Bank's World Governance Indicators;
- `FemalePopPct` is (unsurprisingly) the percentage of the country's population that is female;
- `FertilityRate` is the country's fertility rate, specifically the "the number of children that would be born to a woman if she were to live to the end of her childbearing years and bear children in accordance with age-specific fertility rates of the specified year;"
- `FemalePctLegis` is the percentage of the members of the country's national legislature / parliament that are female.

I'll leave it to you to think about what – if any – relationships we might find between these predictors and CO2 emissions, and why. For purposes of this exercise, please:

1. Estimate a linear regression model to examine the relationship(s) between the variables above and CO2 emissions, and discuss the results of that regression in substantive terms. In doing so, consider carefully the value of transforming the dependent and/or any of the independent variables in the model. Justify those decisions, and be sure that your interpretation of the resulting regression results accounts for them.
2. Re-estimate your model, dichotomizing `FemalePopPct` into those countries that have $\geq$ 50 percent of their population female vs. those that do not. Briefly discuss what, if any, effect this change has on your "findings."

This homework is worth 50 possible points, and will be due via email *in PDF format* to Christopher Zorn and Morrgan Herlihy at or before **11:59 p.m. EST** on **Wednesday, February 19, 2025**.