PLSC 503: "Multivariate Analysis for Political Research"

Exercise Three

February 12, 2024

The *subjets du jour* is (multi)collinearity.

Part I

Consider a model like:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

where Y, $\mathbf{X}$, and u meet all the usual assumptions of the classical linear regression model, $r_{X_1 X_2} = \text{corr}(X_1, X_2) \in (-1, 1)$ and $\text{corr}(X_1, X_3) = \text{corr}(X_2, X_3) = 0$. (In other words, there is a non-zero correlation between X_1 and X_2, but no correlation among the predictors otherwise).

Using simulations,[1] show:

1. The relationship between $r_{X_1 X_2}$ and $\widehat{\text{s.e.}(\hat{\beta}_1)}$ for $N = 10$,[2]
2. How that relationship changes as $N \to \infty$, and
3. Similarly, the relationship (if any) among $r_{X_1 X_2}$, N, and $\widehat{\text{s.e.}(\hat{\beta}_3)}$.

Hint: To generate two normal variates that are correlated to a particular degree using R, check out the various `norm2d` commands in the `fMultivar` package, or the `mvtnorm` package.

[1] Note that, in doing so, it's wise to hold other things – like the means and variances of the Xs and the us – constant at some level.

[2] That is, with $N = 10$, as $r_{X_1 X_2}$ varies, what happens to $\widehat{\text{s.e.}(\hat{\beta}_1)}$?

Part II

For this exercise, we'll again use some data from the *World Development Indicators* (WDI) to illustrate key course models and concepts.[3] In this assignment we'll be focusing on a subset of the whole WDI data, one that *only includes the year 2018* and which containing roughly 36 variables. Specifically:

- `ISO3` - The country's International Standards Organization (ISO) three-letter identification code (e.g., `CHE` for Switzerland).
- `Year` - The year that row of data applies to. The combination of `ISO3` and `Year` uniquely identifies every observation in the data.
- `Region` - The geographical region of the country. There are seven regions specified: East Asia & the Pacific, Europe & Central Asia, Latin America & the Caribbean, the Middle East & North Africa, North America, South Asia, and Sub-Saharan Africa.
- `country` - The name of the country (useful for labeling, etc.).
- `LandArea` - Land area (sq. km).
- `ArablePercent` Arable Land (percent of total land area).
- `Population` - Popluation.
- `PopGrowth` - Population Growth (percent).
- `RuralPopulation` - Rural Population (percent of total).
- `UrbanPopulation` - Urban Population (percent of total).
- `BirthRatePer1K` - Birth Rate (births per 1K people).
- `FertilityRate` - Fertility Rate (births per woman).
- `PrimarySchoolAge` - Primary school starting age (years).

[3]The WDI are data collected on around 1,400 demographic, political, social, and economic indicators collected annually by the World Bank. Data are collected at the national level, for each of around 215 countries in the international system. The length of time for which each variable is collected varies significantly; the longest series extend back to 1960, while others are only available for very recent years or at discrete time points. (The World Bank also gathers and publishes WDI data on regional groupings, but for our purposes we'll focus on data at the national level.) Detailed information on the WDI is available at the WDI website. The WDI data is available for bulk download (and on-line analysis) at the World Bank's website. The code for obtaining and creating the WDI data for this exercise is available on the course Github repository, and makes use of the very useful `WDI` and `countrycode` packages, created by Vincent Arel-Bundock and his collaborators.

- `LifeExpectancy` - Life Expectancy at birth (years).
- `AgeDepRatioOld` - Age Dependency Ratio (old), percent of the working age population.
- `CO2Emissions` - CO2 Emissions (metric tons per capita).
- `GDP` - GDP (constant 2015 $US).
- `GDPPerCapita` - GDP Per Capita (constant 2015 $US).
- `GDPPerCapGrowth` - GDP Per Capita Growth (percent annual).
- `GNIPerCapita` - GNI Per Capita (constant 2015 $US).
- `AdjNetNatlIncome` - Adjusted Net National Income (constant 2015 $US).
- `Inflation` - Inflation (CPI, annual percent).
- `Unemployment` - Unemployment (percent of the total labor force).
- `TotalTrade` - Total Trade (percent of GDP).
- `Exports` - Exports (percent of GDP).
- `Imports` - Imports (percent of GDP).
- `FDIIn` - Inward Foreign Direct Investment (FDI) (percent of GDP).
- `AgriEmployment` - Percent of total employment in agriculture.
- `MobileCellSubscriptions` - Mobile / cellular subscriptions per 100 people.
- `NaturalResourceRents` - Total natural resource rents (percent of GDP).
- `MilitaryExpenditures` - Military expenditures (percent of GDP).
- `GovtExpenditures` - Government Expenditures (percent of GDP).
- `PublicEdExpend` - Public expenditure on education (percent of GDP).
- `PublicHealthExpend` - Public expenditure on health (percent of GDP).
- `WomenBusLawIndex` - Women Business & the Law Index Score.
- `PaidParentalLeave` - Paid Parental Leave (0 = No, 1 = Yes).

Your assignment is:

1. Estimate a regression model to examine the hypothesis that economic performance is related to (drives?) government expenditures. More specifically, choose one of `GovtExpenditures`, `MilitaryExpenditures`, `PublicEdExpend`, or `PublicHealthExpend` as your main outcome / variable of interest, and choose from among the other variables for your predictors, with a focus on measures that (positively or negatively) reflect the health and strength of the economy.

2. Assess the extent of multicollinearity, both substantively (that is, *why* might we expect here to be problematic collinearity present?) and statistically, using whatever approaches you deem fit.

3. Discuss what, if any, measures you might take to ameliorate any problematic collinearity among the covariates. Address the potential problems/pitfalls of each approach. If you decide to undertake any of these, show and briefly discuss your work.

This homework is worth 50 possible points, and will be due via email *in PDF format* to Christopher Zorn (`zorn@psu.edu`) and Nathan Morse (`nam@psu.edu`) at or before **11:59 p.m. EST** on **Wednesday, February 21, 2024**.