PLSC 503: "Multivariate Analysis for Political Research"

Exercise Three
February 16, 2022

The *subjets du jour* is (multi)collinearity (Part I) and data transformations (Part II).

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, corr$(X_1, X_2) \in (-1, 1)$ and corr$(X_1, X_3) =$ 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, show:

1. The relationship between corr(X_1, X_2) and $\widehat{\text{s.e.}(\hat{\beta}_1)}$ for $N = 10$,[1]
2. How that relationship changes as $N \to \infty$, and
3. Similarly, the relationship between corr(X_1, X_2) 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] That is, as corr(X_1, X_2) varies, what happens to $\widehat{\text{s.e.}(\hat{\beta}_1)}$?

Part II

For this exercise, we'll use some data from the *World Development Indicators* (WDI) to illustrate key course models and concepts. 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. In this assignment we'll be focusing on a subset of the whole WDI data, one containing roughly 35 variables. 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.

We'll be using these data for other exercises as well. For this exercise, we'll be using a subset of the WDI data containing the following variables:

- `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).
- `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 2010 $US).
- `GDPPerCapita` - GDP per capita (constant 2010 $US).
- `GDPPerCapGrowth` - GDP Per Capita Growth (percent annual).
- `Inflation` - Inflation (CPI, annual percent).
- `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.
- `NetAidReceived` - Net Official Development Aid Received (constant 2018 $US).
- `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` - Government expenditure on education (percent of GDP).
- `PublicHealthExpend` - Government expenditure on health (percent of GDP).
- `HIVDeaths` - Deaths due to HIV/AIDS (UNAIDS estimate).
- `WomenBusLawIndex` - Women Business & the Law Index Score.
- `PaidParentalLeave` - Paid Parental Leave (0 = No, 1 = Yes).

Of particular interest to us today is the variable `HIVDeaths` ("WBLI"), which is the number of deaths due to HIV/AIDS in each country in the data in each year. Those data are available from 1990 onward (and have some country-level missingness as well); accordingly, we'll focus only on data from 1990-2020.

For this part of this exercise, please:

1. Estimate a simple multivariate regression model of the form:

$$\begin{aligned} \texttt{HIVDeaths}_{it} &= \beta_0 + \beta_1 \texttt{Year}_{it} + \beta_2 \texttt{UrbanPopulation}_{it} + \\ &= \beta_3 \texttt{AgeDepRatioOld}_{it} + \beta_4 \texttt{GDPPerCapita}_{it} + \\ &= \beta_5 \texttt{PublicEdExpend}_{it} + \beta_6 \texttt{PublicHealthExpend}_{it} + u_{it} \end{aligned} \tag{1}$$

   In doing so,

   (a) please ignore (for now) the "panel" / time-series cross-sectional structure of the data; more important,

   (b) *pay particularly close attention to transformations to linearity*. That is, specify a model that transforms the outcome and/or covariates as you feel is appropriate, and describe and justify your decisions.

2. Present and discuss your findings in substantive terms, with particular attention to the associations between the (possibly transformed) predictors and the (possibly transformed) outcome variable.

3. Finally, assess the extent of multicollinearity among your (possibly transformed) predictors, both substantively (that is, *why* might we expect here to be problematic collinearity present?) and statistically, using whatever approaches you deem fit.

This exercise is due in electronic (PDF) form, via email, on Friday, February 25, 2022 by 11:59 p.m. ET and is worth the customary 50 points. Please be sure to include all code necessary to replicate your work with your exercise.