GSERM 2017: “Regression III”

Exercise Two

Exercise

The topic of the homework is leverage, discrepancy, influence, and outlier detection, and the application of the bootstrap. The substantive question is from public health, and the data in question are 2004 statistics on a number of country-level variables ($N = 134$, after accounting for missing data). Specifically, they are:

- `country` is the country name.
- `isocode` is the three-letter ISO code for that country, and
- `ccode` is the three-digit Correlates of War (COW) identifier.
- `HALE` is country’s *health-adjusted life expectancy*. This WHO measure reflects both a country’s lifespan and its health performance, and is calculated as the “average number of years that a person can expect to live in ‘full health,’ excluding years lived in less than full health due to disease and/or injury.”
- `GDPPerCap` is Gross Domestic Product (GDP) per capita, in constant U.S. dollars.
- `Openness` is a measure of trade openness, defined as $\frac{\text{Imports + Exports}}{\text{GDP}} \times 100$; that is, total trade as a percentage of GDP.
- `UNEducation` is UN’s education index, which reflects a weighted combination of literacy and educational enrollments.
- `BattleDeathsLag` is the lagged (i.e., 2003) number of battle deaths per 100,000 members of the population; it captures the potential association between war and health performance.
- `RefugeesLag` is the lagged (2003) number of refugees housed in the country, per 1000 members of the native population. It is designed to reflect any influence that refugees might have on health performance.

The dependent variable of interest is HALE; for the time being, we’ll assume that all five of the other variables belong on the right-hand side of the model, untransformed and in a linear/additive fashion.

Part I

1. Estimate the model discussed above, and (very) briefly discuss your "findings."
2. Address the question of whether – and, if so which of, and to what extent – the findings are being driven by a small number of particularly influential observations. It is probably wise to start with / rely upon the discussion from the Day 3 class for this, though you should also use your own judgement as to what kinds of things can and should be considered.
3. Finally, estimate and provide a brief discussion/justification of a "final" model – that is, one that deals with outliers, if any. Please note: *Your "final" model need not necessarily be any different from your initial one.* What I do ask, however, is that you (briefly) justify your decisions about your final model in light of whatever you find (or do not find) in your analysis of influence and outliers.

Part II

Reestimate the model with the data you settled on in Part I, using a simple nonparametric bootstrap. Present the revised estimates of the $\hat{\beta}$s and their associated standard errors, and briefly discuss (in whatever manner you feel is appropriate) the differences between the conventional estimates, standard errors, and inferences and those from the bootstrap.

This exercise is worth 50 possible points, and is due electronically (submitted as a .PDF file to `zorn@psu.edu`) on or before 5: p.m. St. Gallen time on Monday, June 26, 2017.