Pol Sci 630: Problem Set 6: Dummy Variables and Interactions

Prepared by: Anh Le (anh.le@duke.edu)

Due Date: Wednesday, Oct 12, 2016 (Beginning of Class)

Note 1: It is absolutely essential that you show all your work, including intermediary steps, and comment on your R code to earn full credit.

Note 2: Please use a *single* PDF file created through knitr to submit your answers. knitr allows you to combine R code and LaTeX code in one document, meaning that you can include both the answers to R programming and math problems. Also submit the source code that generates the PDF file (i.e. .Rnw file)

Note 3: Make sure that the PDF files you submit do not include any references to your identity. The grading will happen anonymously. You can submit your answer at the following website: `http://ps630-f15.herokuapp.com/`

1 Merging data (8 points)

The most common merging task in political science is to merge datasets based on country-year. The biggest obstacle is that country codes can come in many forms (country name, World Bank code, COW code, ISO2, ISO3, etc.)

This exercise will let you dip your toes in the sea of pain that is merging real world data. You're expected to Google and read help files to figure out two packages: 1) `countrycode`, which converts between different types of country codes, and 2) `psData`, a package that automates the downloading of many common Political Science dataset.

Note: In the past `psData` didn't work for some students. Let me know early if you encounter problems you can't solve.

1.1 Download WDI data

Download GDP per capita ('NY.GDP.PCAP.CD') and FDI ('BX.KLT.DINV.CD.WD') from WDI, 2007-2009, `extra = FALSE`. What country indicators are there?

Note: There should be 792 rows

1.2 Download Polity data

Use `PolityGet()` in package `psData` to download Polity data. Download the 'polity2' variable (*not* the entire dataset). Use 'iso3c' as the format for the country code.

What country indicators are there?

Note: There should be 16351 rows

1.3 Convert country code

To merge WDI and Polity data we must first create a common country ID. (We can't use country name, because there's no guarantee they will be the same). Use package `countrycode` to convert the country code in WDI data from 'iso2c' to 'iso3c'. Store this newly created country code in the WDI data frame.

1.4 Merge

Merge the WDI and the Polity data based on 'iso3c' and 'year' (Note: There should be 492 rows).

There are two variables showing country names in the merged dataset. Why? Clean them up so we only have 1 country name variable in the merged dataset.

1.5 Check merged result

(Optional) Figure out which country years appear in WDI data but not in Polity data. Note: There should be 300 unmatched records.

In real research, this is useful to check that you are not throwing away data erroneously. There are more than one way to do this and should require some Googling.

2 Factors and Regression with Factors (8 points)

2.1 Dichotomize a continuous variable

Create a new factor variable in your merged dataset, called `polity2_binary` that is 1 (labeled 'democracy') when `polity2` ≥ 0, and 0 (labeled 'dictatorship') otherwise.

2.2 Regression with one binary variable

Regress FDI on the binary variable `polity2_binary`. From the regression result, report the average amount of FDI that democracy and dictatorship gets.

Note: You should know this from the regression result, not from running `mean()`

2.3 Regression with interaction and interpretation

Regress FDI against `polity2_binary`, gdppc, and their interaction term.

Let's say that I want to plot FDI against gdppc with two lines, one representing democracy, the other representing dictatorship (similar to the last plot in the lab). What would be the intercept and slope of these two lines? (

Note 1: No need to draw to plot; I'm only interested in the values of the intercept and the slope

Note 2: Don't look at the regression results and calculate the values by hand / calculator. Instead, use R code to extract the regression coefficients and calculate the values in R. We do things this way so that if your model changes (and it always does), all of the calculations change accordingly and you don't have to manually hunt down each calcultation. This leads to better reproducibility.

2.4 Demonstrating substative meaning of coefficients

In research, we usually have to demonstrate the substantive meaning of our regression result. As emphasized in class, when we have an interaction term we can't interpret the size of the coefficients by themselves anymore.

So what to do? A common solution is to give the estimated outcome for a "typical" country, varying one important factor. For example, imagine that we have a country with median gdppc. What would be its FDI if it were a 1) dictatorship and 2) democracy, holding gdppc at the median value?

Hint: You could either calculate using regression formula, or feed `newdata` to `predict`