Assignment 4: Going Fishing

The Principled Way

PSCI 8357, Spring 2016
February 4, 2016

This assignment must be turned in by the start of class on **Thursday, February 18**. That's right: *two* weeks from today. You must follow the instructions for submitting an assignment.

Main Task

You will be working with the dataset `neumayer.dta`, the replication data for Neumayer (2005). Each observation is a country-year. The response variable is `lneuplusasyl`, the natural logarithm of the number of people seeking asylum in Western Europe from the given country in the given year. The covariates for each country of origin include the following. I have tried my best to match variable names to the descriptions in the text of Neumayer (2005), but there may be some errors below.

- `lngdp`: logged GDP per capita, in 1997 USD (ln *GDP*)
- `growthrate3years`: average annual economic growth over the previous three years (*GROWTH*)
- `lnpop`: logged population (ln *POPULATION*)
- `ecdis`: a Freedom House index of discrimination against ethnic minorities, measured on a 0-4 scale (*ECONDISCRIMINATION*)
- `free`: sum of the two Freedom House indices of political rights and civil liberties, each ranging from 1 (most free) to 7 (least free) (*AUTOCRACY*)
- `pts`: average of two Political Terror Scales measuring human rights violations, ranging from 1 (best) to 5 (worst) (*RIGHTSVIOLATION*)
- `sfallmax`: magnitude score for civil war, ethnic war, or collapse of state authority (*DOMWAR/STATEFAIL*)
- `genpoliticidemag`: magnitude score for number of deaths from genocide and politicide (*GEN/POLITICIDE*)
- `uppsalaexternalintensity`: intensity of external conflicts, on an ordered categorical scale (0 = no conflict, 1 = 25-999 deaths that year in

minor conflict, 2 = 25-999 deaths that year in conflict totaling 1000+ deaths across years, 3 = 1000+ deaths that year) (*EXTERNALWAR*)

- `urban`: urban population as a percentage of total population (*%URBAN*)
- `sharepop1564`: ages 15–64 as a percentage of total population (*%POP15-64*)
- `food`: net per capita food production (*FOOD*)
- `sumdead`: aggregate deaths from natural disasters (*NATURALDISASTER*)
- `bnksum`: total number of guerrilla and riot events (*DISSIDENTVIOLENCE*)
- `distmineurope`: logged minimum air distance between the country's capital and the closest Western European capital (ln *DISTANCE*)
- `christ`: percentage of Christians (*%CHRIST*)
- `colsec`: number of years between 1900 and 1960 as a colony of any Western European destination country (*COLONY*)
- `aidipolated`: aid as percentage of GDP (*AID*)
- `tradeipolated`: trade as percentage of GDP (*TRADE*)
- `arrivalsipolated`: number of tourist arrivals (*TOURISTS*)

Like Neumayer, you should restrict your sample to developing countries, as indicated by the `developing` variable.

You will take the following steps:

1. You will randomly split the sample 50-50 into a *training set* and a *validation set*.

2. Using only the training set, you will run various models in search of an interesting finding. By "interesting" and "finding" I mean:

    - The model must include at least one higher-order term.
    - The higher-order term itself must be statistically significant, according to the nominal *p*-value.
    - The test of the relevant composite hypothesis must also be statistically significant, again according to the nominal *p*-value.

3. Select a model from the training set. Present the estimates and nominal *p*-values from the model you fished for.

4. Run the exact same model on the validation set. How do your estimates and inferences change? **Note:** You should only do this after you have made a final decision about what model you want to present. In other words, unlike the training set, you should only use the validation set

once.

5. Answer the following questions.
    - From a statistical and scientific standpoint, what are the advantages and disadvantages of this sample-splitting exercise?
    - Should we trust the results of the hypothesis tests you report from the training set? What about the ones from the validation set?
    - Would it be ethical to “fish” for a finding in this way, but only report the validation set results? Why or why not?

Your model need not include all of the covariates. You should be thoughtful about what you choose to control for.

You may, for the purposes of this exercise, assume the errors are independent and identically distributed across observations.

Weekly Visualization Challenge

Compare the shape and strength of the estimated relationship of interest between the training and validation sets. Make sure to include any relevant measures of uncertainty.

Hints

Sampling

You can use the `sample()` function to randomly sample from a vector.

```r
sample(1:10, size = 5)
```

```
## [1] 9 3 2 6 8
```

Seeding

When you randomly sample, your results will be different every time. Because, you know, randomness.

```
sample(1:10, size = 5)
```

```
## [1] 10  7  6  4  9
```

```
## Same code, different result
sample(1:10, size = 5)
```

```
## [1]  5 10  2  7  4
```

But sometimes you want to draw something at random *once* and then keep it the same, even when you re-run your code in the future. For example, at the outset of this assignment, you are going to split your sample in half. You don't want the sample split to be different every time you recompile your R Markdown document.

To ensure that a random draw will give you the same results every time, you can use the `set.seed()` function. Pick any number you like, and set it as the "random seed" using `set.seed()`. This will let you replicate your results from any function that uses random numbers.

```
set.seed(97)  # I use the jersey numbers of my favorite Bengals
sample(1:10, size = 5)
```

```
## [1] 1 6 8 2 9
```

```
## Do it again
set.seed(97)
sample(1:10, size = 5)
```

```
## [1] 1 6 8 2 9
```

References

Neumayer, Eric. 2005. "Bogus Refugees? The Determinants of Asylum Migration to Western Europe." *International Studies Quarterly* 49 (3): 389–410. http://dx.doi.org/10.1111/j.1468-2478.2005.00370.x.