HW week 11

w203: Statistics for Data Science

w203 teaching team

Get familiar with the data

You receive a data set from World Bank Development Indicators.

- Load the data using `load` and see what is loaded by using `ls()`. You should see `Data` which is the data frame including data, and `Descriptions` which is a data frame that includes variable names.
- Look at the variables, read their descriptions, and take a look at their histograms. Think about the transformations that you may need to use for these variables in the section below.
- Run: `apply(!is.na(Data[,-(1:2)] ) , MARGIN= 2, mean )` and explain what it is showing.
- Can you include both `NE.IMP.GNFS.CD` and `NE.EXP.GNFS.CD` in the same OLS model? Why?
- Rename the variable named `AG.LND.FRST.ZS` to `forest.` This is going to be our dependent variable.

Decribe a model for that predicts `forest`

- Write a model with two explanatory variables.
  - Create a residuals versus fitted values plot and assess whether your coefficients are unbiased.
  - How many observations are being used in your analysis?
  - Are the countries that are dropping out dropping out by random chance? If not, what would this do to our inference?
- Now add a third variable.
- Show how you would use the regression anatomy formula to compute the coefficient on your third variable. First, regress the third variable on your first two variables and extract the residuals. Next, regress forest on the residuals from the first stage.
- Compare your two models.
  - Do you see an improvement? Explain how you can tell.

Make up a country

- Make up a country named `Mediland` which has every indicator set at the median value observed in the data.
- How much forest would this country have?

Take away

- What is the causal story, if any, that you can take away from the above analysis? Explain why.