Assignment 3: Gravity Estimation
Fall 2022
Due date: November 8

This assignment looks at the estimation of "gravity" regressions for commuting flows between Detroit neighborhoods. For this assignment, please produce the following five outputs that are described in detail below:

i. A table reporting a log-linear gravity regression. The table should have four columns, each employing a different Stata estimator to estimate the same specification: 'reg', 'xtreg', 'areg', 'reghdfe'. The table should have two panels (distance vs time). Report the computation time associated with each regression.

ii. A table with eight columns reporting the results of regressing commuting flows on the same covariate when including observations for which commuting flows are zero. Use the Stata estimators specified below.

iii. A table reporting computation times in a race between 'reghdfe', 'FixedEffectModels.jl', and 'fixest' for the non-zero commuting flows.

iv. A PDF that contains the tables listed above and responses to the questions about them that appear below.

v. Code that produces all of the above.

Submit all your code, the TeX output, and the compiled PDFs in one ZIP file via Canvas before the deadline. Please allow us to run your code on our machines by minimizing your use of machine-specific filepaths. Define a global variable for the filepath at the top of your scripts or, preferably, write everything using "relative filepaths."

Download data. Here is a link to the 'data' folder for this assignment contains a 19 MB ZIP file. 'Detroit.csv' contains commuting flows between Detroit census tracts in 2014.

- The keys that define a row are two Census tracts: 'home_ID' and 'work_ID'

- 'flows' is the integer-valued number of people who reside in 'home_ID' and work in 'work_ID'
- 'distance_straight_miles' is the straight-line distance between the two tracts
- 'distance_Google_miles' is the driving distance reported by Google Maps
- 'duration_minutes' is the driving time reported by Google Maps

Stata estimation. The 'reg', 'xtreg', 'areg', and 'glm' packages are included in Stata by default. Install the 'reghdfe', 'ppml', 'poi2hdfe', and 'ppmlhdfe' on your machine using the 'ssc install' command. Use the 'timer' function in Stata.

Warning: 'outreg2' can be quite slow if you have many fixed effects, so we suggest using the 'estout' package to produce TeX tables for this assignment.

Table 1: Log-linear estimation. Estimate a regression of log (non-zero) commuters on log bilateral (Google driving) distance, origin fixed effects, and destination fixed effects. When using 'reg', you will have to create the fixed-effect dummies. When using the 'xtreg', 'areg', and 'reghdfe' commands, use the fixed-effect options to absorb them. You should find that your coefficient on log distance when using 'reg' is -.40716; the standard error should be .00120.

Your table should have two panels. In the top panel, use the Google-driving-distance covariate. In the bottom panel, use the Google-driving-time covariate. Report the computation time for each of the eight regressions, and include answers to the following questions about Table 1 in your TeX file.

i. Are the point estimates and standard errors numerically identical across the different estimators? Should they be?

ii. Are the number of observations and R-squared statistics identical? Should they be?

iii. Compare the relative computation times of these estimators.

iv. Are the coefficients on the distance and time covariates the same? Should they be?

Table 2: Zeros. Now we examine the roles of zeros in the commuting matrix. Run a regression using the Google-driving-distance covariate. Use Stata estimators and the following specifications to produce a table containing 8 columns:

i. A log-linear regression that omits observations in which flow equals zero.

ii. A log-linear regression that omits observations in which flow equals zero. In addition, set the dependent variable to log of flow plus one.

iii. A log-linear regression in which the dependent variable is log of flow plus one.

iv. A log-linear regression in which the dependent variable is log of flow plus 0.01.

v. A log-linear regression in which the dependent variable is log flow, but flows that are zero in the data ($X_{ij} = 0$) are replaced by $X_{ij} = 1e^{-10}X_{jj}$ as the dependent variable.

vi. An estimate of the same constant-elasticity specification that uses [the 'poi2hdfe' command] to implement the [PPML estimator] of Guimaraes, Figueirdo, and Woodward (REStat 2003) and Silva and Tenreyro (REStat 2006). Use all observations, including zeros, in this and the next column.

vii. An estimate of the same constant-elasticity specification that uses [the 'ppmlhdfe' command] to implement the PPML estimator.

viii. An estimate of the constant-elasticity specification that uses [the 'ppmlhdfe' command] to implement the PPML estimator. Omit observations in which flow equals zero.

Include answers to the following questions about Table 2 in your TeX file:

i. Are your results sensitive to the omission of zeros?

ii. How well does transforming the dependent variable to be $log(x+1)$, $log(x+0.01)$, or $log(1e^{-10}X_{jj})$ if zero work? Is the result sensitive to the choice of transformation?

iii. Examine the residuals from your log-linear regression. Are they heteroskedastic? Report a Breusch–Pagan test statistic and a scatterplot of the residuals that addresses this question.

iv. How do the computation times compare?

Table 3: Comparing Stata's reghdfe, Julia's FixedEffectModels, R's fixest. Now estimate the log-linear specification of Table 1 using non-zero commuting flows using 'reghdfe' again. Compare the speed of this calculation to the speed of estimating it in Julia using the 'FixedEffectModels' package and in R using the 'fixest' package. Use heteroskedastic-robust standard errors in all cases.

- Julia:
  - Use the [StatFiles package] to load a DTA file if need be.
  - Use the [FixedEffectModels package] to estimate.
  - Use the [RegressionTables package] to produce tables.
  - Use the ['@time' macro] in Julia to track performance.
  - Note that Julia functions are compiled the first time they are run, so the first run will be slow and you should not use the full-size data the first time you call your function.

  If you have considerable difficulty getting started in Julia, after consulting your classmates, you should ask the TA for help. We also recommend that you read the QuantEcon.org lectures on [Programming in Julia] to get started. To install Julia on your own machine, download Julia 1.6 from this link. Consider installing the packages that are [dependencies for the QuantEcon package]. Package installation can take a few minutes. Don't panic.

- R:
  - Use the [foreign package] to load the DTA file.
  - Use the [fixest package] to estimate ([short intro to fixest], [fixest at GitHub].
  - You likely know better than us how to export a pretty table from an R regression. R-blogger's [5 ways to measure running time of R code]

- Include answers to the following questions in your submission:

  i. Verify that 'reghdfe', 'FixedEffectModels', and 'fixest' return identical point estimates. Are the standard errors identical?

  ii. Which estimator is faster? By what magnitude?