Econ 8400: Empirical IO, Fall 2022
Problem Set 1 (Due Sept 8)

Download from CANVAS the file GMdata.RAW or GMdata.dta (in Stata format). These files contain the data from the Griliches-Mairesse paper. There are 9 variables: *index* (firm ID), *sic*3 (3 digit SIC), *yr* (year $\in \{73, 78, 83, 88\}$), *ldsal* (log of deflated sales), *lemp* (log of employment), *ldnpt* (log of deflated capital), *ldrst* (log of deflated R&D capital), *ldrnd* (log of deflated R&D), *ldinv* (log of deflated investment). See the Griliches-Mairesse paper for more details on how the data set was collected.

The equation you want to estimate is

$$ldsal_{it} = \beta_1 lemp_{it} + \beta_2 ldnpt_{it} + \beta_3 ldrst_{it} + \gamma_t + \delta_t d357 + \alpha_i + \varepsilon_{it}$$

where $d357$ is a dummy variable for computers (SIC 357) (note that the coefficient on this variable varies by time period) .

1. Report sample statistics (number of observations, mean, median, standard deviation, etc.) for the variables for both the all sample and the balanced sub-panel (i.e., those firms that are present in all years). Also report these statistics for the firms that existed at least 2 periods. Do these statistics seem different? If so what does this suggest?

2. Having gotten some initial feel for the data, let's explore firm heterogeneity.

   (a) Using only the balanced sub-panel compute (and report) the total, between, within and random effects estimators, for the above equation.

   (b) Perform a Hausman test of random effects versus fixed effects.

   (c) What have you learned about firm heterogeneity from these results?

3. In this question we will explore various "difference" estimators and what the differences can tell us. Using the balanced sub-panel compute difference estimators of the above equation. Report results from first (i.e., 5 year) differences, second (10 year) differences and "long" (15 year) differences. What does this tell you about measurement error? (See the discussion in the Griliches-Mairesse paper as well as Griliches and Hausman, "Errors in Variables in Panel Data," Journal of Econometrics, 1986, 31, 93-118.)

4. Let's explore whether entry and exit are random.

   (a) Using the full (unbalanced) panel compute the "total" and first difference estimators. Also compute an OLS estimator using only the firms that were present at least 2 periods. How do these estimates compare to the balanced panel estimates? What does this tell you?

   (b) Use a Probit model to estimate the probability that the firm exists in $t + 1$ as a function of $ldnpt_{it}$, $ldrst_{it}$, and $ldinv_{it}$. Compute the implied inverse mills ratio and include it in the above first difference regression and the OLS regression which used the firms that were present in at least two periods.

   (c) What do these results tell you about entry and exit?

Note:
(1) Working in groups on these numerical problem sets is fine, and encouraged. All members of a group should ultimately do the calculations and hand them in individually.
(2) When asked to report results present the answer in a table. Nothing fancy but don't simply attach a printout of the statistical program you used. You should attach the code you used to generate the results as an appendix.