

Download from the course home page the file GMdata (in Stata or ASCII format). It contains the data from the Griliches-Mairesse paper. There are 9 variables (which appear in this order in the ASCII file): *index* (firm ID), *sic3* (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), *ldmd* (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

$$ldsal_{it} = \beta_1 * lemp_{it} + \beta_2 * ldnpt_{it} + \beta_3 * ldrst_{it} + d_t + d_t * d357 + a_i + \epsilon_{it}$$

where d_t are time dummy variables and $d357$ is a dummy variable for computers (SIC 357).

1. Report sample statistics (number of observations, mean, median, standard deviation, etc.) for the variables for both the all sample and the balanced sun-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. (i) Using only the balanced sub-panel compute (and report) the total, between, within and random effects estimators, for the above equation.
(ii) Perform a Hausman test of random effects versus fixed effects.
(iii) What have you learned about firm heterogeneity from these results?
3. (i) 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?
(ii) 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 1st difference regression and the OLS regression which used the firms that were present in at least 2 periods.

4. (i) Estimate the following model

$$l\text{dsal}_{it} = \beta_1 * l\text{emp}_{it} + \beta_2 * l\text{dnpt}_{it} + \beta_3 * l\text{drst}_{it} + d_t + d_t * d357 + \alpha_i + \omega_{it} + \epsilon_{it}$$

where ω_{it} is not serially correlated but is “transmitted” (i.e., correlated with the current choice of labor and future choice of the other inputs), and α_i is a firm “fixed” effect. First-difference the data (to get rid of the fixed effect) and use lagged values of the inputs as instruments.

- (ii) Estimate the same model without the fixed effect, i.e.

$$l\text{dsal}_{it} = \beta_1 * l\text{emp}_{it} + \beta_2 * l\text{dnpt}_{it} + \beta_3 * l\text{drst}_{it} + d_t + d_t * d357 + \omega_{it} + \epsilon_{it}$$

but now let $\omega_{it} = \rho\omega_{i,t-1} + v_{it}$ and continue to assume that it is “transmitted”. Quasi-difference the data and used lagged values of the inputs and output as instruments. Note, that you will need to do some of your own programming here (the STATA Arellano-Bond command does not allow for serial correlation). You can either program this outside of STATA or you can write a loop that for each value of ρ computes an IV regression, takes the residual and interacts it with the lagged output. You then choose the ρ that sets this last moment to zero (or at least as close as possible).

- (iii) Add a fixed effects to the above model

$$l\text{dsal}_{it} = \beta_1 * l\text{emp}_{it} + \beta_2 * l\text{dnpt}_{it} + \beta_3 * l\text{drst}_{it} + d_t + d_t * d357 + \alpha_i + \omega_{it} + \epsilon_{it}$$

where $\omega_{it} = \rho\omega_{i,t-1} + v_{it}$. Difference the quasi-differences and then use lagged values of the inputs as instruments.

5. Compute an Olley-Pakes like estimator by computing the following steps.

- (i) Regress $l\text{dsal}_{it}$ on $l\text{emp}_{it}$, the dummy variables and a 2nd order polynomial in

$l\text{dnpt}_{it}$, $l\text{drst}_{it}$, and $l\text{div}_{it}$. Report the coefficients on $l\text{emp}_{it}$ and the dummy variables.

- (ii) To compute the remaining coefficients define:

$$y_{it}^* = l\text{dsal}_{i,t+1} - (\hat{\beta}_1 * l\text{emp}_{i,t+1} + \hat{\theta}_1 * d_{t+1} + \hat{\theta}_2 * d_{t+1} * d357)$$

where the “hats” denote coefficients estimated in (i). Use NLLS to minimize the sum of squares of the following residuals:

$$\xi_{it} = y_{it}^* - \beta_2 * ldnpt_{it+1} - \beta_3 * ldrst_{it+1} - (\hat{\phi}_{it} - \beta_2 * ldnpt_{it} - \beta_3 * ldrst_{it}) - (\hat{\phi}_{it} - \beta_2 * ldnpt_{it} - \beta_3 * ldrst_{it})^2$$

where $\hat{\phi}_{it}$ denotes the value of the polynomial computed in (i). Note that, β_2 and β_3 are the coefficients to be estimated in this non-linear regression). Report the estimates of β_2 and β_3 .

(iii) Use a Probit model to estimate the probability that the firm exists in $t+1$ as a function of $ldnpt_{it}$, $ldrst_{it}$, and $ldinv_{it}$. Denote by \hat{P}_{it} the predicted probability from this model. Repeat (ii) but now instead of $\hat{h}_{it} [\equiv \hat{\phi}_{it} - \beta_2 * ldnpt_{it} - \beta_3 * ldrst_{it}]$ and \hat{h}_{it}^2 include \hat{P}_{it} and \hat{P}_{it}^2 .

(iv) Repeat (ii) but include a 2nd order polynomial in both \hat{P}_{it} and \hat{h}_{it} .

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.