Problem Set 1

This problem set is due February 6th. Upload your write up and any code files to the Github Classroom before midnight. You may work together, but you must turn in separately written (unique) write ups and/or code.

1. Consider the random effects model under the following assumptions:

   RE.1 $y_{it} = \alpha_i + \beta' x_{it} + \gamma' z_i + \varepsilon_{it}$

   RE.2 $\mathrm{E}[\varepsilon_{it}\varepsilon_{js}|\mathbf{X}_i] = 0$, $i \neq j$ or $t \neq s$

   RE.3 $\mathrm{E}[\varepsilon_{it}|\mathbf{X}_i] = \mathrm{E}[\alpha_i|\mathbf{X}_i] = 0$

   RE.4 $\mathrm{E}[\alpha_i\varepsilon_i|\mathbf{X}_i] = 0$

   RE.5 $\mathrm{E}[\alpha_i^2|\mathbf{X}_i] = \sigma_\alpha^2$

   RE.6 $\mathrm{E}[\varepsilon_{it}^2|\mathbf{X}_i] = \sigma_\varepsilon^2$

   where $\mathbf{X}_i = [X_i, Z_i]$ and $X_i = \begin{bmatrix} x_{i1}' \\ \vdots \\ x_{iT}' \end{bmatrix}$. You may assume you have a balanced panel.

   (a) In your own words, briefly describe what restrictions assumptions RE.2–RE.6 impose on the data.

   (b) Show that the pooled OLS estimator is consistent for $\theta = (\beta', \gamma')'$ in N for above model and derive its asymptotic distribution, is it asymptotically efficient? Why or why not?

   (c) Show that applying the RE-GLS transformation to this model leads to an error term with mean 0 and variance 1. The algebra here can be a little tedious, so feel free to use a computer aid like Mathematica, sympy, Maple, or Matlab's symbolic math toolbox.[1] Include any code you use.

[1] I encourage you to use computer algebra systems for algebra and calculus steps. Sympy is free and open source, see `https://www.sympy.org/en/index.html`. Maple and Matlab freely available to students at TAMU, see `https://software.tamu.edu/`.

2. Consider the fixed effects model under the following assumptions:

FE.1 $y_{it} = \alpha_i + \beta' x_{it} + \varepsilon_{it}$

FE.2 Each unit (x_i, ε_i) is drawn from an iid process

FE.3 $\mathrm{E}[X_i' \varepsilon_{it}] = 0$

FE.4 $\mathrm{E}[X_i' X_i]$ exists and has full rank

FE.5 $\mathrm{E}[\varepsilon_{it}^2 | \mathbf{X}_i] = \sigma_\varepsilon^2$

Show the following:

(a) Show that pooled OLS is biased under these assumptions and derive the bias in terms supposing that $\mathrm{E}[\alpha_i | x_{it}] = \gamma_i' z_i$. Explain why a random effects estimator would also be biased. Can you say anything about which would be more biased?

(b) Let M_i be with within unit demeaning matrix such that $M_i = I_T - 1_T(1_T' 1_T)^{-1} 1_T'$ where I is an identity matrix and 1 is a column vector of 1s. Show that

$$X_i' X_i \geq X_i' M_i X_i.$$

Note that to show that one matrix is weakly less than other, you'll want to show that $X_i' X_i - X_i' M_i X_i$ is positive semi-definite.

HINT 1: The trick here is work it into quadratic form, which is to say something that looks like ABA and then show that B is positive semi-definite (it has all weakly positive eigenvalues).

HINT 2: M_i and $I - M_i$ are both idempotent. All the eigenvalues of an idempotent matrix are either 0 or 1.

(c) Use your analysis from the above to argue that under the above assumptions the within estimator of β has a weakly larger variance than the pooled estimator

$$\mathrm{Var}(\hat{\beta}_w | X_i) \geq \mathrm{Var}(\hat{\beta}_p | X_i).$$

What additional assumption to do you need for this inequality to hold? What does this tell us about the choice between the pooled and the within estimators?

3. Consider the following model

$$
\begin{aligned}
y_{it} &= \alpha_i + \beta' x_{it} + \gamma' z_i + \varepsilon_{it} \\
x_{it} &= [x^1_{it}, x^2_{it}] & z_i &= [z^1_i, z^2_i] \\
\alpha_i &\sim U[-1, 1] & \theta &= (\beta', \gamma')' = (1, 4, -3, 5)' \\
x^1_{it} &= 15 - 3a_i - 2z^2_i + 0.25x^1_{it-1} + x^{1*}_{it} & x^{1*}_{it} &\sim N(0, 1/4) \\
x^2_{it} &\sim \text{Poisson}(5) \\
z^1_i &\sim N(0, 4) & z^2_i &\sim N(0, 1/4) \\
\varepsilon_{it} &= 0.8\varepsilon_{it-1} + u_{it} & u_{it} &\sim N(0, 7) \\
x^1_{i0} &= 15 - 3a_i - 2z^2_i & \varepsilon_{i0} &= 0 \\
i &= 1, \ldots, N & t &= 1, \ldots, T_i.
\end{aligned}
$$

Here α_i and z_2 are assumed to be unobserved to the analyst. Note that the above normal random variables are specified as $N(\mu, \sigma^2)$.

(a) Which of the various A assumptions in the course notes are satisfied in the this model? Which are not?

(b) Based on your answer to part (a), if you were interested in the effect of x^1 on y, what would be an appropriate estimator and specification? What would be an appropriate variance estimator?

(c) What if you were only interested in the effect of x_2 on y? How would you specify the model? What estimator would you use for β^2? What would be an appropriate variance estimator?

(d) Find the "true" values of σ^2_ε and σ^2_α under this data generating process. Note that in this case σ^2_α should include the unobserved z_2. What do they tell you about how similar/different the RE estimates will be from the pooled and within estimates?

HINT 1: you can rewrite ε_{it} to be

$$\varepsilon_{it} = \sum_{s=0}^{t} (\varepsilon_{i0} + u_{it-s})(0.8)^s, \quad u_{i0} = 0$$

which will make it a lot easier to work with expectations and variances. Apply the expectation/variance operator to both sides, and then take $t \to \infty$ (since it's stationary and ergodic we can treat any slice we observe as part of an infinite chain)

HINT 2: Recall that $\sum_{t=0}^{\infty} ar^t = \frac{a}{1-r}$ for $|r| < 1$ and constant a.

(e) Find the correlation between x_{it}^1 and the unobserved heterogeneity $\alpha_i + 5z_i^2$.

(f) Conduct a Monte Carlo experiment with this data generating process and

i. $N \in \{50, 200\}$ (the number of U.S. states and roughly the number of countries in the world)

ii. $T_i - 1 \sim \text{Poisson}(24)$

Code and implement your own versions of the RE-GLS, RE-MLE (with gradients), the within, and the Mundlak estimators. By your own version, I mean that `lm` and `optim` are fine, but I don't want to see `lmer` or `feols` here. The only packages I'd like to see would be for data manipulation, graphics, or tabling (e..g, `dplyr`, `data.table`, `ggplot`, `xtable`, `matrixStats`), but you can complete this assignment without loading any packages. If you decide to parallelize your simulations, you may use any packages you like for that.

When conducting your simulation, use a 25 period burn-in in generating x^1 and ε to ensure they are not too tied to the initial conditions. Use the specification(s) you described in part (b). Build your simulations to answer the following questions:

i. Is the pooled estimator biased for β_1? Does it roughly match what you would expect for omitted variable bias based on your answer to question 1?

ii. Is the RE-GLS estimator biased for estimating β_1? Does the RE-MLE do any better?[2] Do you notice any differences in the RE estimates of β_1? If so, what might explain it?

iii. Is the within estimator biased for estimating β_1?

iv. For the RE-GLS, pooled, and within estimators, calculate the standard error on $\hat{\beta}_1$ using both classical and clustered variance estimator. For the clustered variance you can either estimate the asymptotic variance or conduct an appropriate bootstrap. Which one better reflects the observed uncertainty in your estimates across simulations? Why?

[2]Use the MLE estimator we described in class that assumes that the unobserved heterogeneity is normal.

v. Conduct the Hausman test for comparing RE-GLS with the within estimator. Does a rejection reflect statistical size or power in this simulation? Also analyze the performance of the Wald test we described on the Mundlak-specified model fit with OLS and clustered standard errors.

vi. Conduct a second simulation that changes the DGP to satisfy the random effects assumptions. In this simulation record only the Hausman test results and the Wald test on the Mundlak model results. Does a rejection here reflect size or power?

vii. Using the above results on size and power, compare the Hausman test to a Wald test. Which has better properties in these simulations? Is it surprising?