Julia Hofmann and Colleen Werkheiser

Economics 312

Final Project

Efficiency of Fixed and Random Effects Estimators: A Monte Carlo Analysis

Introduction

This project investigates the differences in efficiency between fixed and random effects models for estimating relationships in panel data using Monte Carlo simulation methods. Small sample behaviors of the estimators are also investigated. We are curious as to under what circumstances random effects is more efficient, and how much more efficient it is in these cases. We also want to confirm that the fixed effects estimator is an accurate estimator even under the random effects assumptions as long as the number of time observations is small. It is common that fixed effects is used in this case, so we want to confirm that this is a valid method to use. Our results matched our expectations, though we were unable to make strong conclusions about how much more efficient the random effects estimator is. Both estimators were found to be accurate for our data, suggesting the convention of using fixed rather than random effects does not decrease the accuracy of the estimation by a large amount.

Theory

There are two main models used in estimation with panel data. Those models are fixed and random effects. Due to the two-dimensional nature of panel data, there exist both unit and time fixed effects models, the first of which assumes the differences in data occur in a fixed manner across individuals and not at all across time and the second of which assumes fixed differences across time and no differences across individuals. We most often assume, especially in cases where the number of years measured in the data is small, that more of the differences occur across individuals, so unit fixed effects is a much more common estimator to use. Because of this convention, for the remainder of this paper, the term fixed effects refers to the unit fixed effects model.

The unit fixed effects model is given by the following equation:

$$y_{it} = \beta_{1i} + \beta_2 x_{2it} + \beta_3 x_{3it} + e_{it},$$

where e_{it} is a random error term. The fixed effects model assumes that variation between individuals is fixed with respect to time. This model includes an individual-specific, time-invariant intercept such that each individual has their own intercept. This intercept is the β_{1i} term in the model, and these intercepts represent the fixed effects that the model is named after.

The fixed effects model is estimated by de-meaning the data. This process first averages the data across time to get the following equation:

$$\bar{y}_{it} = \beta_{1i} + \beta_2 \bar{x}_{2it} + \beta_3 \bar{x}_{3it} + \bar{e}_i .$$

Then, by subtracting this equation from the underlying model, the transformed model is obtained with the de-meaned y depending on the de-meaned x's and the de-meaned error term. The process of de-meaning the data also serves to eliminate the individual intercepts β_{1i} from the model being estimated, thus decreasing the number of degrees of freedom used. This transformed equation can be estimated using OLS and the β_2 and β_3 in the transformed model are the same as in the underlying model.

The fixed effects model is appropriate when there is some factor that makes individuals different from one another, but this factor is constant over time. Such time-invariant factors include innate ability for individuals or historical and institutional factors for countries.

The random effects model used for panel data assumes that the differences between individuals are random as opposed to fixed. This is modeled by including a fixed intercept $\overline{\beta}_1$ and a random variable u_i which varies across individuals in place of the individual intercept β_{1i} in the fixed effects model. The u_i term is assumed to have constant variance and mean zero, similar to a random error term. It is also assumed to be uncorrelated with the x's in the regression, as the e_{it} term is. The random effects model for each individual i at each unit of time t is thus:

$$y_{it} = \overline{\beta}_1 + u_i + \beta_2 x_{2,it} + \beta_3 x_{3,it} + e_{it} .$$

The random effects estimator allows us to look at variables that vary over time as well as those that do not. For example, characteristics of individuals in the sample such as gender or race, which do not vary over time, can be factored into random effects while they cannot in fixed effects. Additionally, random effects is estimated using GLS while fixed effects is estimated using OLS and as such, random

effects estimates will generally have smaller variances. As a result, the random effects model is more efficient.

While random effects is more efficient than fixed effects, problems often arise that make it not applicable as a model. Most often, the random effects themselves, u_i, are correlated with the *x*'s, simply because the random variation across individuals is often related to other observations of the individuals. This violates our assumptions about u_i, and makes random effects an invalid estimator. Although it is possible to correct for this endogeneity by using instruments, it is often better to simply use fixed effects instead, given how difficult strong instruments are to find.

Therefore, one would expect to see better estimates for the β coefficients using the fixed effects estimator if the random effects assumptions were not met. The fixed effects estimator is also appropriate if the variance in the data is much larger across individuals than across time. In this case, one can treat the variance across individuals as fixed over time.

Not a great deal of econometric literature has investigated the use of fixed versus random effects models. In particular, the differences in efficiency, although acknowledged, are generally not measured. Most studies concerned with fixed and random effects are concerned with their application in meta-analysis contexts.

Data

The data used in our Monte Carlo simulation were pulled from two panel data sets given by Hill, Griffiths, and Lim, nlspanel.dta and crime.dta[1]. We chose one variable from each dataset, specifically looking for one that varied more across individuals and one that varied more across time, respectively. We then made the data's variance more extreme in the already more varied component by adding in random variance across individuals or across time. The variable corresponding to our *b2* coefficient was made to vary more across individuals and *b3* made to vary more across time. We therefore expect fixed effects to be better at estimating *b2* and random effects to be better at estimating *b3*. One limitation of

[1] Retrieved from http://principlesofeconometrics.com/poe4/poe4.htm

our data is that we only have four years observed in our time variable. However, this is common of panel data, so a simulation with this specification is reasonable.

Results and Analysis

The underlying model that we chose to estimate was the following:

$$y_{it} = 4 + u_i + 10x_{2,it} + 28x_{3,it} + e_{it},$$

with the u_i having a variance of 5 and the e_{it} having a variance of 1. Our x_2 variable had much larger variance across individuals than time, while our x_3 variable had much more variance across time than individuals. Therefore, we expected fixed effects to give a better estimate for β_2 and random effects to give a better estimate for β_3.

We first used data with 90 individuals and 4 years. We ran a simulation 5000 times using random effects and then the same simulation using fixed effects, with the results summarized below:

```
Random effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    9.999996    .0003437   9.998648   10.00117
          b3 |      5000          28    .0000377   27.99986   28.00015

Fixed effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    10.00001    .0003546    9.99881   10.00174
          b3 |      5000          28    .0000379   27.99983   28.00013
```

Fixed effects gives a mean *b2* estimate closer to the true β_2 than random effects, as expected. The standard deviation of mean values of *b3* is smaller for random effects than fixed effects with both giving the true β_3 value as the mean estimate. This suggests that random effects produces a better estimate of β_3, although we are unable to determine how much more accurate this estimate is due to rounding in Stata.

It is somewhat surprising that both fixed and random effects give such accurate estimates of *b3*, considering most of the variance in x_3 is across time and not across individuals, suggesting that fixed effects are not occurring. However, considering that there are only four years observed, even a large variation across years does not make much of a difference in the models' efficiencies. In order to account for variation over time in the fixed effects model we would have to perform a time fixed effects estimate. However this is not necessary when the time variable only takes on a few values. Therefore, both random and fixed effects are appropriate for data with a small number of years regardless of whether most variance is across individuals or time.

Another important aspect of our results to note is that the standard deviation of both the *b2* and *b3* values was smaller for random effects than for fixed effects, even though *b2* was more accurate for fixed effects. This suggests that the random effects estimator is more precise, though it is difficult to tell how much more so than fixed effects. A separate Monte Carlo study would have to be done in order to determine the difference in precision for these two estimators.

We then decreased our sample size to investigate how well both fixed and random effects models did at estimating both coefficients.
Restricting our sample to fifty individuals, the results were as follows:

```
Random Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    9.999995    .0008581   9.997337   10.00324
          b3 |      5000          28    .0000501   27.99984    28.0002

Fixed Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    10.00001    .0008729   9.996778   10.00371
          b3 |      5000          28    .0000506    27.9998   28.00017
```

As can be observed from the data, the standard deviations of both coefficient estimates increased, as was to be expected. In addition, we can see that random effects gives a worse estimate of *b2* with 50

individuals than with our original 90 individuals. Although both standard deviations increased for the estimates of *b3*, the means stayed constant and remained very close to the underlying value of β_2.

Decreasing sample size again to twenty-five individuals yielded the following results:

```
Random Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    10.00002    .0027255   9.990821   10.01029
          b3 |      5000          28    .0000731   27.99972   28.00031

Fixed Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    10.00007    .0027985   9.989896   10.01078
          b3 |      5000          28    .0000723    27.9997   28.00028
```

It is apparent that both estimates of *b2* became less accurate. In fact, fixed effects is worse at estimating *b2* than random effects is at this sample size. The exact reasons for this are unclear, but this suggests more uncertainty in the estimation of *b2* as sample size is decreased. It also shows that the random effects estimator is more efficient than the fixed effects estimator, even when fixed effects is an appropriate model. Yet again, however, both estimates of *b3* are accurate, despite their higher standard deviations due to the smaller sample size.

Finally, decreasing our sample size once more to fifteen individuals results in the following:

```
Random Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    9.999929    .0047292   9.983551   10.01625
          b3 |      5000          28    .0000964    27.9997    28.0004

Fixed Effects:

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
          b2 |      5000    10.00008    .0047464   9.984583   10.01589
          b3 |      5000          28    .0000954   27.99962   28.00031
```

Both estimates of *b2* are again worse than our ninety-individual sample case and the fifty and twenty-five individual cases, with larger standard deviations than any of the larger sample sizes. This trend of worse estimates for *b2* as sample size decreases is interesting to note. Nevertheless, the estimates of *b3* are again exactly 28, the true value, though the standard deviations of *b3* are also larger than those of the larger samples.

This suggests something about the behavior of data with different variance structures over both fixed and random effects. The data being represented by *b3* have a much higher variance across time than across individuals, while those represented by *b2* have a higher variance across individuals than across time. It makes sense, therefore, that both fixed and random effects are much better at estimating models whose main variance is across time, as limiting the number of individuals does not change the variance structure of the data nearly as much as it does in cases where the main variance is across individuals. This also explains why the estimates of *b2* get so much worse as sample size is decreased, as removing observations changes the very structure of the data.

Conclusion and Assessment of Validity

Overall, the differences in efficiency between fixed and random effects for our data do not seem to be that large. In both the case where we expect random effects to be more accurate and that where we expect fixed effects to be more accurate, random effects still produces estimates with a smaller standard deviation, suggesting it is more precise. This continues to be the case as number of individuals is decreased.

When more of the variance in the data is across individuals than across time, fixed effects produces more accurate estimates, as we expect. When more of the variance is across time, both fixed and random effects models estimate the coefficients equally well, although, as noted before, random effects is more precise.

Because most panel data sets include far more observations of individuals than of time periods, both models as we used them, unit fixed effects and random effects, produce good estimates. If it were the case that panels included larger ranges of time, our fixed effects estimator would presumably not produce as accurate estimates of β_2. Instead, random effects, which accounts for both unit effects and time effects, would likely be the better estimator.

Nonetheless, we recognize the difficulties in employing random effects, regardless of its increased efficiency as an estimator. The likelihood of u_i being correlated with x_{it} is very high, given the types of data often collected in panels. In this case, employing instrumental variables to correct for this issue, while plausible, is more often than not very difficult simply because finding strong instruments for the endogenous variables is challenging in practice.

As sample size decreases, the standard deviation of the coefficient estimates increases. This is exactly what should occur, according to OLS and GLS assumptions.

We structured our underlying model so that e_{it} and u_{it} were uncorrelated with x_{2it} and x_{3it}. We also took our x_{2it} and x_{3it} from unrelated data sets so that they would be uncorrelated with one another. We also checked to make sure that they were uncorrelated and found that they were. When increasing the respective variances in the data, we used randomly generated numbers so that we would not introduce any additional correlation or trends in our data that were not there originally. These procedures in structuring our data and model meant that all of the assumptions for random effects were met.

However, there were limitations due to the data we had. The conclusions that can be drawn for the β_3 estimates are not as valid as those for the β_2 estimates because our data had a small number of time observations, especially compared to the number of individuals. This meant that our estimates for β_3 were so precise that it was difficult to compare the two estimation methods.

The external validity of our model was examined in the previous conclusions section in that random effects is more efficient when the correct assumptions are met, but those assumptions are often not met, meaning that fixed effects is used more often in estimating panel data whether variance is greatest across individuals or time.

There are a few ways that one might want to expand our study. One is to re-run the simulation with data that include a larger number of observations for the time variable. This would mean that fixed effects should no longer be as accurate an estimator, which would allow one to see how much more accurate random effects is when it is most accurate.

Our simulations suggested that the random effects estimator is more precise. Therefore, another investigation that could be done is that one could change the amount of variance greatly in order to attempt to determine how much more precise random effects is than fixed effects.

References

Field, Andy P. "Meta-Analysis of Correlation Coefficients: A Monte Carlo Comparison of Fixed- and Random-Effects Methods." *Psychological Methods* 6 (2001): 161-180. doi: 10.1037//1082-989X.6.2.161.

Hunter, John E. and Frank L. Schmidt. "Fixed Effects vs. Random Effects Meta-Analysis Models: Implications for Cumulative Research Knowledge." *International Journal of Selection and Assessment* 8 (2000): 275-292. doi: 10.1111/1468-2389.00156/.

Hill, Carter R., William E. Griffiths, and Guay C. Lim. *Principles of Econometrics, Fourth Edition*. New York: John Wiley & Sons, 2012.