A Note on Decompositions in Fixed Effects Models in the Presence of Time-Invariant Characteristics

Axel Heitmueller
London Business School and IZA

The issue

Decomposition techniques are very popular among economists. The idea is to disaggregate the difference in outcomes across two groups in an *explained* and *unexplained* part. The former is due to differences in the Xs (observable variables) and the latter due to differences in the Betas (coefficients).

Question: Can this technique also be applied to panel data?

Problem

Very often a fixed effects model is the appropriate specification. However, time-invariant variables such as sex and country of birth are wiped out.

No coefficients for these variables are estimated using the standard `xtreg,fe` command

Hence, these variables can not be used for the decomposition in the classical sense. Yet, as it turns out they become part of the overall constant term and bias the decomposition result.

More formally

Stata© estimates the following equations:

$$y_{it} = a \ + \beta \ x_{it} + \theta \ z_i \ + \alpha_i \ + \varepsilon_{it}$$

$$y_{it} - \bar{y}_i + \bar{\bar{y}} = a \ + \beta \ (x_{it} - \bar{x}_i + \bar{\bar{x}}) + \theta \ \bar{z} + \varepsilon_{it} - \bar{\varepsilon}_i + \bar{\bar{\varepsilon}}$$

where

$$\bar{\bar{y}} = \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \Big/ NT$$

Key restriction

The key identification restriction is

$$\sum_{i=1}^{N} \alpha_i = 0$$

i.e. the sum of fixed effects over all N individuals is equal to zero

Yet, no such assumption is made for the time-invariant variables i.e.

$$\sum_{i=1}^{N} z_i \neq 0$$

Consequence

As a result, the overall constant term that is estimated by Stata© will also contain $\theta\ \bar{z}$ (the time-invariant variable and its coefficient) so that $c = a + \theta\ \bar{z}$ and the individual terms are not identifiable.

Hence, the decomposition terms will be

$$\bar{\bar{y}}^1 - \bar{\bar{y}}^2 = \hat{\beta}^1(\bar{\bar{x}}^1 - \bar{\bar{x}}^2) + (\hat{\beta}^1 - \hat{\beta}^2)\bar{\bar{x}}^2 + (\hat{c}^1 - \hat{c}^2)$$

When they actually should be

$$\bar{\bar{y}}^1 - \bar{\bar{y}}^2 = \hat{\beta}^1(\bar{\bar{x}}^1 - \bar{\bar{x}}^2) + \hat{\theta}^1(\bar{z}^1 - \bar{z}^2) + (\hat{\beta}^1 - \hat{\beta}^1)\bar{\bar{x}}^2 + (\hat{\theta}^1 - \hat{\theta}^2)\bar{z}^2 + (\hat{a}^1 - \hat{a}^2)$$

Does this matter?

One way to test the magnitude of the bias is using Monte Carlo studies

Data created had the following features:

- Panel data, two groups, each N=1000 and T=20
- Time-variant variables are correlated with fixed effects
- Time-invariant variables are orthogonal

Table 1: Monte Carlo results

Component	True	Expected	Estimated	MSE	Relative MSE
Model I					
Total gap	4.1462	4.1462	4.1450	0.0020	
Explained part	-1.5500	0.7000	0.7000	5.0627	
Unexplained excluding constant	3.7056	0.7982	0.7985	8.4600	
Unexplained including constant	5.7056	3.4462	3.4450	5.1126	
Constant	2.0000	2.6480	2.6465	0.4311	

Note: Monte Carlo simulation 5,000 replications. Each sample has 1,000 observations and each individual is observed over 20 time periods. The true value refers to equation (6) where all parameters can be estimated including the time-invariant ones. In contrast, the expected value refers to a fixed effects estimation where time-invariant variables are swiped out from the estimation and consequently not part of the decomposition. The Mean Square Error refers to the deviation from the true model. The relative MSE is 3 SLS MSE derived from an auxiliary regression relative to the fixed effects MSE in column 5.

What can be done?

Testing for mean equality in time-invariant variables in the two groups (more difficult if more than one)

Testing for zero mean in time-invariant variables (again more complicated if more than one variable)

Auxiliary regression to recover coefficient for time-invariant variable: (1) retrieve fixed effects using `predict,u` (2) regress fixed effects on time-invariant variables (3) use coefficients for decomposition (key assumption: time-invariant variables are orthogonal)

Table 1: Monte Carlo results

Component	True	Expected	Estimated	MSE	Relative MSE
Model I					
Total gap	4.1462	4.1462	4.1450	0.0020	1.0000
Explained part	-1.5500	0.7000	0.7000	5.0627	0.0124
Unexplained excluding constant	3.7056	0.7982	0.7985	8.4600	0.0022
Unexplained including constant	5.7056	3.4462	3.4450	5.1126	0.0137
Constant	2.0000	2.6480	2.6465	0.4311	0.1062
Model II					
Total gap	6.3962	6.3962	6.3950	0.0020	1.0000
Explained part	0.7000	0.7000	0.7000	0.0003	1.0000
Unexplained excluding constant	3.7056	0.7982	0.7985	8.4600	0.0022
Unexplained including constant	5.7056	5.6962	5.6950	0.0023	1.0000
Constant	2.0000	4.8980	4.8965	8.4029	0.0025
Model III					
Total gap	3.0962	3.0962	3.0950	0.0020	-
Explained part	0.7000	0.7000	0.7000	0.0003	-
Unexplained excluding constant	0.4056	0.7982	0.7985	0.1630	-
Unexplained including constant	2.4056	2.3962	2.3950	0.0023	-
Constant	2.0000	1.5980	1.5965	0.1759	-

Note: Monte Carlo simulation 5,000 replications. Each sample has 1,000 observations and each individual is observed over 20 time periods. The true value refers to equation (6) where all parameters can be estimated including the time-invariant ones. In contrast, the expected value refers to a fixed effects estimation where time-invariant variables are swiped out from the estimation and consequently not part of the decomposition. The Mean Square Error refers to the deviation from the true model. The relative MSE is 3 SLS MSE derived from an auxiliary regression relative to the fixed effects MSE in column 5.

Conclusion

Time-invariant variables pose a problem for decomposition techniques in fixed effects models.

This is not surprising and very similar to the implicit problems of decomposition techniques in cross-section data.

However, ignoring ***observable*** time-invariant variables in panel data is different from omitting ***unobservable*** variables in cross-section data and may be overcome.