DISCUSSION PAPER SERIES

IZA DP No. 1886

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

Axel Heitmüller

December 2005

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

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

Axel Heitmüller
London Business School
and IZA Bonn

Discussion Paper No. 1886
December 2005

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
Email: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of the institute. Research disseminated by IZA may include views on policy, but the institute itself takes no institutional policy positions.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 1886
December 2005

ABSTRACT

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

Though theoretically appealing and very popular amongst labour economists, the interpretation of the unexplained part of the Oaxaca (1973) decomposition as discrimination rather than an omitted variable problem in cross-section data has often been criticised. In this note it is shown that this problem extends also to panel data and moreover that in a fixed effects model including time invariant regressors omitted variables are a necessary and deliberate consequence. Monte Carlo simulation is used to show the extent of the bias. Special cases and practical implications are discussed.

JEL Classification: C1, C33

Keywords: decomposition, fixed effects, Monte Carlo study

Corresponding author:

Axel Heitmüller
London Business School
Regent's Park
London, NW1 4SA
United Kingdom
Email: aheitmueller@london.edu

[*] I would like to thank Pierre-Carl Michaud for very helpful comments. All remaining errors are my own.

Introduction

Though theoretically appealing and very popular amongst labour economists, the interpretation of the unexplained part of the Oaxaca (1973) decomposition as discrimination rather than an omitted variable problem in cross-section data has often been criticised. In this note it is shown that this problem extends also to panel data and moreover that in a fixed effects model including time invariant regressors omitted variables are a necessary and deliberate consequence. Monte Carlo simulation is used to show the extent of the resulting bias. Furthermore, even in the absence of time invariant characteristics the standard cross-sectional interpretation of the decomposition results may differ in longitudinal data as often no overall constant term is estimated to begin with. Several special cases and practical implications are discussed.

The Decomposition Context and the Fixed Effects Model

Suppose the model to be estimated is of the following well-known form:

$$y_{it} = \beta_0 + x_{it}^{'}\beta + z_i^{'}\gamma + \eta_i + \varepsilon_{it} \tag{1}$$

where $\eta_i = z_i^{'}\pi + \alpha_i$ so that $E(\eta_i \mid z_i) = z_i^{'}\pi$ and $\alpha_i = \eta_i - E(\eta_i \mid z_i)$. Hence (1) can be rewritten as

$$y_{it} = \beta_0 + x_{it}^{'}\beta + z_i^{'}\theta + \alpha_i + \varepsilon_{it} \tag{1'}$$

where x_{it} is an observed characteristic varying over individual *i* and time *t*; β_0 is a constant term; ε_{it} is an *iid* error term; z_i is an observed time invariant characteristic; $\theta = (\pi + \gamma)$; and α_i the unobserved fixed effect; the orthogonality condition is $E(\varepsilon_{it} \mid x_{it}, \eta_i) = 0$. Fixed effects estimation is appropriate under the assumption that $E(x_{it} \mid \alpha_i) \neq 0$ and the usual within transformation of (1') can be expressed as

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)^{'} \beta + (\varepsilon_{it} - \bar{\varepsilon}_i) \qquad (2)$$

where e.g. $\bar{y}_i = \sum_{t=1}^{T} y_{it} \Big/ T$.

Predicted outcomes from estimating (2) will in general differ from the average sample mean by a constant. Such term is of little interest within samples. However, when applying decomposition techniques such as suggested by Oaxaca (1973) to study differences in predicted conditional sample means it may well differ across groups and contain important information about the relative position of each group. For example, comparing male and female earnings, an overall constant term may contain information about structural or institutional variations in the respective labour markets such as the presence of gender anti-discrimination legislation.[1] Unless one is only interested in differences in within individual variation and there are no time-invariant variables in the decomposition of the gender earnings gap this will yield biased decomposition components.

Standard software packages such as Stata© 8.2 do often report an overall constant term. Under the constrained $\sum_{i=1}^{N} \alpha_i = 0$ and in the presence of time-invariant variables, β_0 can be recovered as $cons = \hat{\beta}_0 + \bar{z}^{'} \theta = \bar{\bar{y}} - \bar{\bar{x}}^{'} \hat{\beta}$ by estimating (2) re-adding the grand averages[2]. The main advantage of doing so is that by construction the sample average of y_{it} will equal $E(y_{it} \mid x_{it})$.[3] Also, fixed effects can be retrieved as $u_i = \alpha_i + z_i^{'} \theta = \bar{y}_i - \bar{x}_i^{'} \hat{\beta} - \hat{\beta}_0$.

[1] Clearly, it will also contain unobservable characteristics which has led to frequent criticism of the usual interpretation of the unexplained part as being due to discrimination.

[2] For example, the grand average of the dependent variable is defined as $\bar{\bar{y}} = \sum_{i=1}^{N} \sum_{t=1}^{T} y_{it} \Big/ NT$

[3] Furthermore, in two-way panel models with fixed group and time effects the above transformation is necessary to estimate a symmetric model (Greene 2000, p 564).

As can easily be seen, unless $\sum_{i=1}^{N} z_i \neq 0$ the fixed effects will contain $z_i^{'}\theta$ and more importantly the constant term will contain $\bar{z}^{'}\theta$. In such a case the explained and unexplained components are biased. The gap in predicted outcomes y_{it} between two groups is usually expressed as

$$\bar{\bar{y}}^1 - \bar{\bar{y}}^2 = (\bar{\bar{x}}^1 - \bar{\bar{x}}^2)^{'}\hat{\beta}^1 + \bar{\bar{x}}^{2^{'}}(\hat{\beta}^1 - \hat{\beta}^2) + (\hat{\beta}_0{}^1 - \hat{\beta}_0{}^2) \qquad (5)$$

where *j=1,2* indicates the group and the double bar represents the sample average. As shown above, the constant terms contain $\bar{z}^{'}\theta$. An appropriate decomposition expression would therefore be

$$\bar{\bar{y}}^1 - \bar{\bar{y}}^2 = (\bar{\bar{x}}^1 - \bar{\bar{x}}^2)^{'}\hat{\beta}^1 + (\bar{z}^1 - \bar{z}^2)^{'}\hat{\theta}^1 + \bar{\bar{x}}^{2^{'}}(\hat{\beta}^1 - \hat{\beta}^2) + \bar{z}^{2^{'}}(\hat{\theta}^1 - \hat{\theta}^2) + (\hat{\beta}_0{}^1 - \hat{\beta}_0{}^2) \quad (6)$$

Three special cases can be distinguished: (a) $\bar{z}^1 = \bar{z}^2 \neq 0$, i.e. the explained component will be unbiased while the unexplained component will still be biased; (b) $\bar{z}^1 = \bar{z}^2 = 0$, i.e. both the explained and unexplained component is unbiased; (c) alternatively it becomes possible to retrieve θ from an auxiliary regression of u_i on z_i enabling the decomposition of (6) directly.

Monte Carlo Results

Monte Carlo simulation is used to demonstrate the magnitude of the bias for three different specifications of the form (1). Specification I employs the usual Oaxaca decomposition to fixed effects results. Specification II and III are equivalent to cases (a) and (b).

Each of the two samples contains 1,000 observations and each individual is observed over 20 time periods. The respective parameters are:

$\beta_0^1 = 1; \beta_0^2 = 3; \beta^1 = 1; \beta^2 = 1.4; \theta^1 = 0.3; \theta^2 = 2.5$.[4] Furthermore, $E(x_{it} \mid \alpha_i) = 0.5$ and $E(z_i \mid \alpha_i) = 0$.[5]

Table 1 reports the estimated decomposition components alongside their respective MSE for 5,000 replications. The *true* values refer to equation (6), the *expected* values are those expected from equation (5) and reported for comparison purposes only. The Mean Square Error (MSE) refers to the average of differences between estimated and true values for each repetition. All three models are constraint such as to derive predictions equal to the sample mean as described above. Two different unexplained components are reported, one including one excluding the constant term.

Clearly, the bias in the total gap estimate is very small and by construction equal in all three specifications. However, as expected the bias is substantial in both the explained and unexplained parts in model I. The bias in the explained part virtually disappears in case the group-means of the time-invariant variables are equal (model II). Furthermore, all components are estimated with much reduced bias if the group-means in time-constant characteristics equal zero (model III).

Are there ways to overcome the bias? The literature suggests at least four methods to deal with time-invariant variables in panel models which are random effects, the Hausman-Taylor (1981) instrument variable approach, pooled OLS ignoring individual effects[6], and an auxiliary regression to recover θ as discussed in cases (c). Column six reports the *relative* MSE between the easy-to-implement latter method and the standard fixed effects approach. Clearly, the bias is significantly reduced.

[4] Other parameter values as well as different values for T and N have been tried without fundamentally changing the main results.

[5] The exact means and standard deviations can be obtained from the author upon request.

[6] See Oaxaca and Geisler 2003 for a two stage GLS estimator which yields equivalent coefficients for the time-invariant variables as the pooled OLS.

Note however that the remaining *absolute* bias can still be large particularly so if the magnitude of the time-invariant variable or its parameters is relatively large.

Conclusion

Panel data is becoming ever more widely available and it is tempting to use methods originally developed for cross-section data in a longitudinal environment such as the popular Oaxaca (1973) decomposition. Yet, this note shows that the interpretation and estimation of longitudinal decompositions is cumbersome when employing fixed effects methods particularly so in the presence of time-invariant characteristics. Unless the group-means for these variables obey certain properties decomposition or time invariant variables are orthogonal to the fixed effects, results are subject to a substantial bias. These results are very similar to the cross-section case with omitted variables. However, the important distinction is that in the presence of time invariant variables in panel data the omitted regressors are observable and knowingly omitted.

References

Greene, W.H., 2000, Econometric Analysis, Fourth Edition, Prentice Hall International, New Jersey.

Hausman, J.A. and W.E. Taylor, 1981, Panel Data and Unobservable Individual Effects, Econometrica 49(6): 1377-1398.

Oaxaca, R.L., 1973, Male-Female Wage Differentials in Urban Labor Markets, International Economic Review 14(1): 693-709.

Oaxaca, R.L. and I. Geisler, 2003, Fixed Effects Models with Time-Invariant Variables. A Theoretical Note, Economics Letters 80: 373-377.

Appendix: Tables

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 are not part of the decomposition. The Mean Square Error refers to the deviation from the true model. The relative MSE is 2 SLS MSE derived from an auxiliary regression relative to the fixed effects MSE in column 5. For model specification see text.