Factor contributions in the presence of endogenous variables: a simulation study

Lawrence B. Dacuycuy[1] and Connie B. Dacuycuy[2]

[1]*School of Economics, De La Salle University2401 Taft Avenue, Manila, Philippines
[2]Economics Department, Ateneo de Manila University, Katipunan Avenue, Quezon City, Philippines
Corresponding Author: lawrence.dacuycuy@dlsu.edu.ph

Abstract Using the Fields inequality decomposition procedure, this paper seeks to investigate how the presence of an endogenous variable affects factor contribution estimates. In particular, we are interested in decomposing the discrepancy between Ordinary Least Squares (OLS) and Linear Instrumental Variables (IV) model based factor contribution estimates. This study adopts Monte Carlo simulation techniques to characterize small and large sample properties.

Key words Fields inequality decomposition, factor contribution, endogeneity, ordinary least squares, Monte Carlo Simulation

1. INTRODUCTION

Regression based decomposition methods have contributed immensely to the analysis of statistical discrimination and inequality. Since the seminal work of Oaxaca (1973) and Blinder (1973) on statistical discrimination, a significant number of studies on inequality decomposition analysis have been written, largely adapting to advances in the econometrics discipline.

Despite advances in methods for inequality analysis, a gap in practical applications has been noticeable. This is particularly true for a class of decomposition methods that rely on the regression framework such as the Fields factor contribution estimator which decomposes inequality (Fields, 2003).

Some papers estimating factor contributions, to a certain extent, account for specification and other classical errors that expectedly induce deviations from the true model. Dacuycuy (2009) investigated how functional assumptions on the wage—experience relationship would affect factor contributions and Dacuycuy and Dacuycuy (2012) developed a bootstrap based procedure to investigate factor contributions to changes in inequality. Dacuycuy and Dacuycuy (2014) focus on the simulation properties of the Fields contribution estimator when measurement errors are encountered.

In this short paper, we use Monte Carlo simulation methods to determine the properties of the inequality contribution estimator when endogenous variables are present. Using simple mathematical manipulations, we decompose the statistical discrepancy between IV and OLS based factor contribution estimates. A related paper by Bigotta, Krishnakumar and Rani (2012) shows asymptotic properties of the inequality factor share estimators but did not account for the possibility of having econometric problems.

This note is organized as follows. Section 2 revisits the Fields decomposition framework and discusses the role of endogenous variables by offering two propositions. Section 3 details the simulation procedure and results and the last section concludes.

2. REVISITING THE FIELDS DECOMPOSITION FRAMEWORK

One of the virtues of the framework is its computational simplicity. Consider the typical regression function $y_i = \boldsymbol{x}_i{}'\beta + \epsilon_i, i = 1,2,\dots,n.$ As discussed in Fields (2003), the contribution of the j^{th} factor to overall inequality, θ_j is conveniently given by the following expression:

$$\hat{\theta}_j^{ols} = \frac{\hat{\beta}_j^{ols}\hat{\sigma}(\boldsymbol{x}_j)\hat{\rho}_{y,x}}{\hat{\sigma}(\boldsymbol{y})}, j = 1,2,\dots.,k \qquad (1)$$

where $\hat{\rho}_{y,x} = \frac{cov(\boldsymbol{x}_j,\boldsymbol{y})}{\hat{\sigma}(y)\hat{\sigma}(\boldsymbol{x}_j)}$ pertains to the correlation coefficient of the dependent variable and the j^{th} factor; $\hat{\beta}_j^{ols}$ refers to the coefficient of the factor j and $\hat{\sigma}(\boldsymbol{x}_j)$ is just the sample standard deviation estimator.

Given the definition, we can rewrite (1) as

$$\hat{\theta}_j^{ols} = \frac{\hat{\beta}_j^{ols}cov(\boldsymbol{x}_j,\ \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2}, j = 1,2,\dots.,k \qquad (1')$$

It is obvious from formula (1) that without variable mismeasurements, omitted variables and other inconsistency – inducing problems, $\hat{\theta}_j^{ols}$converges to the true factor contribution if the true data generating process is linear. In the absence of problems inducing inconsistency, $\hat{\beta}_j^{ols}$ converges to β_j^* in probability while the other statistics, namely, the standard deviation and correlation will converge to their respective true parameters. Thus, we have

$$\hat{\theta}_j^* = \frac{\beta_j^*\sigma^*(x_j)\rho_t^*}{\sigma^*(y)}, j = 1,2,\dots.,k \qquad (2)$$

To understand how endogeneity affects factor contribution estimates, we have the following propositions:

Proposition 1 (Endogenous regressor) For simplicity, consider the linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$ for a random sample of size n. Suppose $\boldsymbol{x_1}$ is an endogenous regressor, that is $cov(\boldsymbol{x_1}, \boldsymbol{\epsilon}) > 0$. Then $\theta_1^{endo} > \theta_1^*$.

Proof:

The correlation between $\mathbf{y}$ and $\boldsymbol{x_1}$ is overstated since $cov(\boldsymbol{y},\ \boldsymbol{x_1}) = cov(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,\ \boldsymbol{x_1}) = cov(\beta_1 x_1,\ \boldsymbol{x_1}) + cov(\boldsymbol{\epsilon},\ \boldsymbol{x_1})$ and the estimate for β_1 is upward biased because of the assumed positive covariance between $\boldsymbol{x_1}$ and the error term $\boldsymbol{\epsilon}$. This proves the presence of endogeneity overstates the factor's contribution.

The standard approach to remedy the problem is to use instrumental variables that are assumed to be highly correlated with the endogenous variable but are not correlated with the error term. Using the second stage regression in a linear IV framework, the factor contribution estimator now becomes

$$\hat{\theta}_j^{iv} = \frac{\hat{\beta}_j^{iv} cov(\boldsymbol{z}_j,\ \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2}, j = 1,2, \ldots.., J \qquad (3)$$

where $\hat{\beta}_j^{iv}$ is the estimate based on $\boldsymbol{z}_j$ is the instrument which may be equal to the predicted value of $\boldsymbol{x}_j$. Because it comes from the first – stage regression between the endogenous variable and the instrument, it does not have an effect as great as $\boldsymbol{x}_j$. This implies that care should be exercised in determining the instrument. This concern brings us back to the fundamental requirement of having strong instruments. Using equation (1'), we have

$$\hat{\theta}_j^{ols} = \left[\frac{\hat{\beta}_j^{ols}}{\hat{\beta}_j^{iv}}\right] \frac{cov(\boldsymbol{x}_j,\ \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2} \hat{\beta}_j^{iv} \qquad (4)$$

If there is endogeneity, it is still possible to measure the discrepancy between the two factor contribution estimators. This can be shown using the following proposition:

Proposition 2 Assume that $\boldsymbol{x}_{\mathbf{j}}$ is endogenous and it is instrumented by $\boldsymbol{z}_{\mathbf{j}}$. Using factor contribution estimators (3) and (4), the discrepancy between the estimators is given by $\widehat{\Delta}_j = \left\{\frac{\widehat{\beta}_j^{ols}\left(cov(\boldsymbol{x}_j,\boldsymbol{y}) - cov(\boldsymbol{z}_j,\boldsymbol{y})\right) + (\widehat{\beta}_j^{ols} - \widehat{\beta}_j^{iv})cov(\boldsymbol{z}_j,\boldsymbol{y})}{[\widehat{\sigma}(\boldsymbol{y})]^2}\right\}$.

Proof: Define $\widehat{\Delta}_j$ as the difference between the two factor contribution estimators.

$$\widehat{\Delta}_j = \hat{\theta}_j^{ols} - \hat{\theta}_j^{iv} \qquad (5)$$

Substituting (3) and (4) into (5),

$$\widehat{\Delta}_j = \left[\frac{\hat{\beta}_j^{ols}}{\hat{\beta}_j^{iv}}\right] \frac{cov(\boldsymbol{x}_j,\ \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2} \hat{\beta}_j^{iv} - \frac{\hat{\beta}_j^{iv} cov(\boldsymbol{z}_j,\ \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2} \qquad (6)$$

Manipulating, we have

$$\widehat{\Delta}_j = \frac{\hat{\beta}_j^{iv}}{[\hat{\sigma}(\boldsymbol{y})]^2} \left\{\frac{\hat{\beta}_j^{ols} cov(\boldsymbol{x}_j,\ \boldsymbol{y}) - \hat{\beta}_j^{iv} cov(\boldsymbol{z}_j, \boldsymbol{y})}{\hat{\beta}_j^{iv}}\right\} \qquad (7)$$

$$\widehat{\Delta}_j = \left\{\frac{\hat{\beta}_j^{ols} cov(\boldsymbol{x}_j,\ \boldsymbol{y}) - \hat{\beta}_j^{iv} cov(\boldsymbol{z}_j, \boldsymbol{y})}{[\hat{\sigma}(\boldsymbol{y})]^2}\right\} \qquad (8)$$

Adding and subtracting the counterfactual term $\hat{\beta}_j^{ols} cov(\boldsymbol{z}_j,\ \boldsymbol{y})$ to (8), we have

$$\widehat{\Delta}_j = \left\{\frac{\hat{\beta}_j^{ols} cov(\boldsymbol{x}_j,\ \boldsymbol{y}) - \hat{\beta}_j^{ols} cov(\boldsymbol{z}_j,\ \boldsymbol{y}) + \hat{\beta}_j^{ols} cov(\boldsymbol{z}_j,}{[\hat{\sigma}(\boldsymbol{y})]^2}\right. \qquad (9)$$

Grouping similar terms, we can arrive at the result.

Denoting $\widehat{\Delta}_j^1 = \frac{\widehat{\beta}_j^{ols}\left(cov(\boldsymbol{x}_j,\boldsymbol{y}) - cov(\boldsymbol{z}_j,\boldsymbol{y})\right)}{[\widehat{\sigma}(\boldsymbol{y})]^2}$ and $\widehat{\Delta}_j^2 = \frac{(\widehat{\beta}_j^{ols} - \widehat{\beta}_j^{iv})cov(\boldsymbol{z}_j,\boldsymbol{y})}{[\widehat{\sigma}(\boldsymbol{y})]^2}$, then we can now attribute the overall difference to differences in covariances and estimates.

3. DESIGN

The correlation table for the random variables $\boldsymbol{x_1}, \boldsymbol{\epsilon}, \boldsymbol{z}$ is given by the following:

	$\boldsymbol{x_1}$	$\boldsymbol{\epsilon}$	$\boldsymbol{z}$
$\boldsymbol{x_1}$	1	$\boldsymbol{\rho_{x,\epsilon}}$	$\boldsymbol{\rho_{z,x}}$
$\boldsymbol{\epsilon}$	$\boldsymbol{\rho_{x,\epsilon}}$	1	**0**
$\boldsymbol{z}$	$\boldsymbol{\rho_{z,x}}$	**0**	1

where $\boldsymbol{\rho_{x,\epsilon}}$ is the correlation between x_1 and ϵ and $\boldsymbol{\rho_{z,x}}$ is the correlation between x_1 and $\boldsymbol{z}$

To maintain simplicity of exposition, we use a simple linear regression model.[2] All variables are generated using the `drawnorm' command, employing data generating processes suitable for linear estimation using OLS. For the base model, we generate independent and identically distributed (i.i.d) standard normal variables by drawing regressors, disturbance term and measurement errors for dependent and independent variables from the multivariate standard normal distribution. We introduced endogeneity by using the correlation table below.

	$\boldsymbol{\rho_{x,z}}$		
$\boldsymbol{\rho_{x,e}}$	0.7, 0.7	0.7, 0.5	0.7, 0.1
	0.5, 0.7	0.5, 0.5	0.5,0.1
	0.1, 0.7	0.1, 0.5	0.1,0.1

We then peg the number of Monte Carlo replications at 2000 then investigate the properties of the contribution estimator using the following sample sizes: 100 and 100000.

4. RESULTS

We consider several cases. First, we assume that there is no correlation between x and ϵ which implies that OLS is consistent. Given this assumption, $\widehat{\Delta}_j$ will be equal to $\hat{\theta}_j^{ols}$. This may also mean that in large samples, the covariance between $\boldsymbol{z}$ and the independent variable is close to zero. Second, we assume a positive correlation between x and ϵ and limit correlation values to three (3), namely: high (0.7), moderate (0.5) and low (0.1).

In the base model, it is clear that the discrepancy is equal to the factor contribution from OLS when the sample size is sufficiently large. It also confirms that the factor contribution of IV is zero because $cov(\boldsymbol{z}_j,\ \boldsymbol{y}) = 0$. With endogeneity, all factor contributions of x are upward biased. This confirms the prediction of proposition 1. For a given correlation between x and $\boldsymbol{\epsilon}$, the discrepancy in terms of factor contributions is monotonically decreasing with respect to the correlation between z and x, that is, the higher is the correlation, the lower is the discrepancy. In terms of contribution, there is no doubt that a significant portion comes from the differences in covariance.

5. CONCLUDING REMARKS

This simple simulation study provides a way to assess factor contribution estimates in the presence of endogeneity. Results indicate that IV model based estimates will never be close to OLS estimates and OLS based factor contribution estimates will be upward biased if there is endogeneity. The study, however, focuses on the discrepancy between OLS and IV, with the latter being used to directly estimate factor contributions.

6. REFERENCES

Blinder, A. (1973). Wage discrimination: reduced form and structural estimates. *Journal of Human Resources*, 436-455.

Brigotta, M., Krishnakuman, J., & Rani, U. (2012). Fuurther theoretical results on the regression-based approach to inequality decomposition and application to India. *Researh Papers by the Institute of Economics and Econometrics*.

Dacuycuy, L. (2009). The functional specification of the wage-experience relationship and male wage inequality in the Philippines: a decomposition analysis. *DLSU Business & Economics Review, 18*(1).

Dacuycuy, L., & Dacuycuy, C. (2012). Decomposing temporal changes in covariate contributions to wage inequality. *Applied Economics Letters*, 1279-1283.

Dacuycuy, L., & Dacuycuy, C. (2014). Decompositions behaving badly: interesting lessons from simulations. *Manuscript*.

[2] We compute the factor contribution statistics using the `gfields.ado' program written by S. Kolenikov in STATA.

Fields, G. (2003). Accounting for income inequality and its change: a new method with application to the distribution of earnings in the United States. In S. Polacheck (Ed.), *Research in Labor Economics* (Vol. 22, pp. 1-38). Amsterdam: Elsevier.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 693-709.

APPENDIX

Table 1Decomposition results for N=100

		Factor contributions			Components		Components (%)		Covariance	
Degree of Correlation		$\hat{\theta}_j^{ols}$	$\hat{\theta}_j^{iv}$	$\hat{\Delta}_j$	$\hat{\Delta}_j^1$	$\hat{\Delta}_j^2$	$\hat{\Delta}_j^1$	$\hat{\Delta}_j^2$	y and x	y and z
x and ϵ	x and z									
Zero	Zero	0.144	0.010	0.134	0.142	-0.008	106.165	-6.165	0.400	0.004
High	High	0.703	0.056	0.647	0.572	0.076	88.287	11.713	1.104	0.209
High	Moderate	0.703	0.033	0.670	0.633	0.037	94.467	5.533	1.104	0.111
High	Low	0.703	0.010	0.693	0.694	-0.001	100.166	-0.166	1.104	0.014
Moderate	High	0.520	0.059	0.461	0.402	0.058	87.360	12.640	0.902	0.204
Moderate	Moderate	0.520	0.034	0.485	0.458	0.027	94.365	5.635	0.902	0.108
Moderate	low	0.520	0.010	0.509	0.512	-0.003	100.638	-0.638	0.902	0.012
Low	High	0.206	0.071	0.135	0.124	0.011	91.581	8.419	0.500	0.198
Low	Moderate	0.206	0.041	0.165	0.163	0.002	98.746	1.254	0.500	0.103
Low	Low	0.206	0.011	0.195	0.202	-0.007	103.694	-3.694	0.500	0.008

Table 2Decomposition results for N =100000

		Factor contributions			Components		Components (%)		Covariance	
Degree of Correlation		$\hat{\theta}_j^{ols}$	$\hat{\theta}_j^{iv}$	$\hat{\Delta}_j$	$\hat{\Delta}_j^1$	$\hat{\Delta}_j^2$	$\hat{\Delta}_j^1$	$\hat{\Delta}_j^2$	y and x	y and z
x and ϵ	x and z									
Zero	Zero	0.138	0.000	0.138	0.138	0.000	100.006	-0.006	0.400	0.000
High	High	0.703	0.046	0.658	0.578	0.080	87.874	12.126	1.100	0.196
High	Moderate	0.703	0.023	0.680	0.639	0.041	94.015	5.985	1.100	0.100
High	Low	0.703	0.001	0.703	0.701	0.002	99.768	0.232	1.100	0.004
Moderate	High	0.519	0.050	0.469	0.406	0.063	86.603	13.397	0.900	0.196
Moderate	Moderate	0.519	0.026	0.494	0.461	0.032	93.506	6.494	0.900	0.100
Moderate	low	0.519	0.001	0.518	0.517	0.001	99.753	0.247	0.900	0.004
Low	High	0.202	0.063	0.138	0.123	0.016	88.586	11.414	0.500	0.196
Low	Moderate	0.202	0.032	0.169	0.161	0.008	95.246	4.754	0.500	0.100

Low	Low	0.202	0.001	0.200	0.200	0.000	99.844	0.156	0.500	0.004