DataFirst

DataFirst Technical Papers

Union selection effects - some inconsistent models

by
Martin Wittenberg

Technical Paper Series
Number 25

About the Author(s) and Acknowledgments

Martin Wittenberg - Director, DataFirst and Professor, School of Economics, University of Cape Town

This is a joint DataFirst and SALDRU Paper.

Recommended citation

Wittenberg, M. (2014). Union selection effects - some inconsistent models. A DataFirst Technical Paper 25. Cape Town: DataFirst, University of Cape Town

DataFirst, University of Cape Town, Private Bag, Rondebosch, 7701, Tel: (021) 650 5708,
Email: info@data1st.org/support@data1st.org

Union selection effects – some inconsistent models

Martin Wittenberg
DataFirst, SALDRU and School of Economics
University of Cape Town

May 2014

Abstract

We show that some of the models which have been used in the South African literature to estimate union selection effects are logically inconsistent. This is a much more serious problem than a failure to identify the coefficient. It implies that the model cannot be true in any possible state of the world. Unfortunately the offending specification is becoming entrenched in the literature.

Key words: unions, selection
JEL codes: C51, J51

Union wage effects have been estimated on South African data for over twenty years (Moll 1993, Schultz and Mwabu 1998, Butcher and Rouse 2001, Hofmeyr and Lucas 2001, Casale and Posel 2010, Bhorat, Goga and van der Westhuizen 2012, Ntuli and Kwenda 2014). Many of these studies have found large union wage premiums (e.g. Schultz and Mwabu 1998, Hofmeyr and Lucas 2001). Some studies have been concerned that union membership is not an exogenously determined category and have tried to adjust the wage equations for "union selection effects". Ever since the pioneering study of Moll (1993) many of these attempts have included a variable for "other union member in the household".

Moll noted that the "other union member variable" was non-traditional in union membership studies, but defended its use:

> "This variable reflects household-specific tastes for unionization, such as the political orientation and the willingness to invest union dues and time in meetings for the sake of long-term security and wage gains. It may also reflect firm strategies of recruitment of family members by employees." (1993, p.252)

The same variable was also used by Hofmeyr and Lucas (2001) who also referred to the intuition of a common household effect:

"In anticipation that there may be some correlation among the unobserved factors leading to union membership of various household members, or that

having a member in a unionized job may make it easier for other members to obtain such a job, a further dummy variable appearing in both the [selection equations] is whether any other household member reports being a union member." (Hofmeyr and Lucas 2001, p.695)

The variable has subsequently been used *inter alia* in studies by Azam and Rospabé (2007), Bhorat et al. (2012) and most recently by Ntuli and Kwenda (2014). The justification in these studies was precedent. Before this practice becomes entrenched yet further it is important to interrogate it properly. While some of these studies have shown that the estimates of the model are sensitive to the inclusion of this variable (Hofmeyr and Lucas 2001, Casale and Posel 2010), nobody has questioned the logic more directly. Indeed Casale and Posel (2010, p.52) note that this is one of the "typical" variables used and, indeed, the "only exclusion restriction that is consistently significant (and strongly positive) in the selection equations" (p.52).

We will show that this variable is highly problematic: indeed the mathematical model underpinning it is logically inconsistent. This is a much stronger failing than failure of identification, which is another potential problem with "social spill-over" variables.

1 The model

The probit form of the union selection model can be written as

$$y_{ij}^{*} = \alpha \max\{y_{-ij}\} + \mathbf{x}_{ij}'\boldsymbol{\beta} + \varepsilon_{ij} \tag{1}$$

where $\mathbf{x}_{ij}$ is a vector of individual (and/or household) covariates, $\varepsilon_{ij} \overset{iid}{\sim} N(0,1)$ and y_{ij}^{*} is the latent propensity by individual i in household j to join a union. $\max\{y_{-ij}\}$ is a short-hand for taking the maximum over the realised union membership outcomes (i.e. dummy variables) among all individuals in household j other than i.

The problem that this model creates can be seen in the case of a two person household where the equations for the two members become:

$$\begin{aligned} y_{1j}^{*} &= \alpha y_{2j} + \mathbf{x}_{1j}'\boldsymbol{\beta} + \varepsilon_{1j} \\ y_{2j}^{*} &= \alpha y_{1j} + \mathbf{x}_{2j}'\boldsymbol{\beta} + \varepsilon_{2j} \end{aligned}$$

Maddala (1983, p.119) notes that this set of equations is logically inconsistent. His discussion is in the context of simultaneous equation systems, but the same logic applies for the subset of two-person households. We have

$$\begin{aligned} \Pr(y_{1j}=1 \text{ and } y_{2j}=1 | \mathbf{x}_{1j}, \mathbf{x}_{2j}) &= \Phi\left(\alpha + \mathbf{x}_{1j}'\boldsymbol{\beta}\right)\Phi\left(\alpha + \mathbf{x}_{2j}'\boldsymbol{\beta}\right) \\ \Pr(y_{1j}=1 \text{ and } y_{2j}=0 | \mathbf{x}_{1j}, \mathbf{x}_{2j}) &= \Phi\left(\mathbf{x}_{1j}'\boldsymbol{\beta}\right)\left[1-\Phi\left(\alpha + \mathbf{x}_{2j}'\boldsymbol{\beta}\right)\right] \\ \Pr(y_{1j}=0 \text{ and } y_{2j}=1 | \mathbf{x}_{1j}, \mathbf{x}_{2j}) &= \left[1-\Phi\left(\alpha + \mathbf{x}_{1j}'\boldsymbol{\beta}\right)\right]\Phi\left(\mathbf{x}_{2j}'\boldsymbol{\beta}\right) \\ \Pr(y_{1j}=0 \text{ and } y_{2j}=0 | \mathbf{x}_{1j}, \mathbf{x}_{2j}) &= \left[1-\Phi\left(\mathbf{x}_{1j}'\boldsymbol{\beta}\right)\right]\left[1-\Phi\left(\mathbf{x}_{2j}'\boldsymbol{\beta}\right)\right] \end{aligned}$$

where Φ is the standard cumulative normal distribution. A quick check will verify that the four probabilities given above will add up to one **only** if $\alpha = 0$. Maddala makes the point that in truly simultaneous systems (i.e. not recursive ones) one cannot have **outcomes** on the right hand side and **propensities** on the left.

The case for three person households is on the surface a bit more complicated, (because of the "max" function) but equally inconsistent:

$$\begin{aligned} y_{1j}^{*} &= \alpha \max\{y_{2j}, y_{3j}\} + \mathbf{x}_{1j}'\boldsymbol{\beta} + \varepsilon_{1j} \\ y_{2j}^{*} &= \alpha \max\{y_{1j}, y_{3j}\} + \mathbf{x}_{2j}'\boldsymbol{\beta} + \varepsilon_{2j} \\ y_{3j}^{*} &= \alpha \max\{y_{1j}, y_{2j}\} + \mathbf{x}_{3j}'\boldsymbol{\beta} + \varepsilon_{2j} \end{aligned}$$

In this case we have eight possible outcomes, where we have omitted the covariates to make the math more transparent:

$$\begin{aligned} \Pr(y_{1j}=1, y_{2j}=1, y_{3j}=1) &= \Phi(a)\Phi(a)\Phi(a) \\ \Pr(y_{1j}=1, y_{2j}=1, y_{3j}=0) &= \Phi(a)\Phi(a)[1-\Phi(a)] \\ \Pr(y_{1j}=1, y_{2j}=0, y_{3j}=1) &= \Phi(a)[1-\Phi(a)]\Phi(a) \\ \Pr(y_{1j}=1, y_{2j}=0, y_{3j}=0) &= \Phi(0)[1-\Phi(a)][1-\Phi(a)] \\ \Pr(y_{1j}=0, y_{2j}=1, y_{3j}=1) &= [1-\Phi(a)]\Phi(a)\Phi(a) \\ \Pr(y_{1j}=0, y_{2j}=1, y_{3j}=0) &= [1-\Phi(a)]\Phi(0)[1-\Phi(a)] \\ \Pr(y_{1j}=0, y_{2j}=0, y_{3j}=1) &= [1-\Phi(a)][1-\Phi(a)]\Phi(0) \\ \Pr(y_{1j}=0, y_{2j}=0, y_{3j}=0) &= [1-\Phi(0)][1-\Phi(0)][1-\Phi(0)] \end{aligned}$$

Again it is evident that these probabilities will sum to one only if $\alpha = 0$.

What are the implications of these findings? The estimated coefficients for these models **cannot** produce probabilities that would add up to one for households other than one person ones (which are obviously uninteresting). That means they cannot correspond to the "Data Generating Process" in any conceivable state of the world. That is a much stronger failing than a failure of identification – which typically means that there are multiple possible states of the world which could all generate the observable data.

2 Could one rescue the "social spill over" intuition?

The ideas that access to "union jobs" may run through social networks or that there may be "household tastes" for unionisation are attractive. Is there any way to reformulate the model in ways that would allow this to be estimated?

2.1 A recursive system

One way of rewriting the model to make it logically consistent would be to remove the causal arrow "pointing back" from other members of the household.

If we could identify the "original" (first) union member (and number this person as 1 within the household) the following model would be consistent:

$$\begin{aligned} y^*_{1j} &= \mathbf{x}'_{1j}\boldsymbol{\beta}+\varepsilon_{1j} \\ y^*_{2j} &= \alpha\, y_{1j} + \mathbf{x}'_{2j}\boldsymbol{\beta}+\varepsilon_{2j} \\ &\dots \\ y^*_{kj} &= \alpha\, y_{1j} + \mathbf{x}'_{kj}\boldsymbol{\beta}+\varepsilon_{kj} \end{aligned}$$

It might be tempting to simply **impose** this structure, e.g. enter the Head of Household's union status as explanatory variable in the union membership equation for other household members. The problem, of course, is that the model has to be true to the underlying data generating process, and if it turns out that the influence within the household works in different ways we end up with a misspecified model, albeit one that is logically coherent.

2.2 Putting propensities on the right hand side

Another logically consistent model would be

$$y^*_{ij} = \alpha\, \overline{y}^*_{-ij} + \mathbf{x}'_{ij}\boldsymbol{\beta}+\varepsilon_{ij}$$

In this case the average **propensity** of other household members appears on the right hand side. In the two person household case this becomes the simultaneous equation model

$$\begin{aligned} y^*_{1j} &= \alpha y^*_{2j} + \mathbf{x}'_{1j}\boldsymbol{\beta}+\varepsilon_{1j} \\ y^*_{2j} &= \alpha y^*_{1j} + \mathbf{x}'_{2j}\boldsymbol{\beta}+\varepsilon_{2j} \end{aligned}$$

which, as Maddala notes, raises estimation issues (Maddala 1983, pp.246-7). Obviously the latent variables on the right hand side are not observed, so at best one can estimate reduced form equations which, in this case, would imply that union member status of an individual is a function not only of that person's explanatory variables, but also of all of the explanatory variables of every other member of the household.

A second point to note is that the errors in the equation which can be estimated (i.e. the reduced form) are no longer independent of each other. Instead the joint outcome for the household would have to be estimated – which would be by a multivariate probit. An additional issue raised by Maddala is that the normalisation implicit in the structural equation (i.e. that the errors are standard normal) would not carry over to the estimation of the reduced form, so that typically one would only be able to identify the signs of the coefficients (Maddala 1983, pp.246-7).

3 Conclusion

We have argued that the way in which union selection effects have been estimated in the South African literature is based on a model that is logically

inconsistent. Any "correction" made on the basis of such a model will also be fatally flawed. We have also suggested that the intuition underpinning the use of that model is likelly to create serious estimation issues. Perhaps the choice of Casale and Posel (2010) not to "correct" for selection is in the current situation the more defensible option.

References

Azam, J.-P. and Rospabé, S.: 2007, Trade unions vs. statistical discrimination: Theory and application to post-Apartheid South Africa, *Journal of Development Economics* **84**, 417–444.

Bhorat, H., Goga, S. and van der Westhuizen, C.: 2012, Institutional wage effects: Revisiting union and bargaining council wage premia in South Africa, *South African Journal of Economics* **80**(3), 400–414.

Butcher, K. and Rouse, C.: 2001, Wage effects of unions and industrial councils in South Africa, *Industrial and Labor Relations Review* **54**(2), 349–374.

Casale, D. and Posel, D.: 2010, Unions and the gender wage gap in South Africa, *Journal of African Economies* **20**(1), 27–59.

Hofmeyr, J. and Lucas, R. E.: 2001, The rise in union wage premiums in South Africa, *Labour* **15**(4), 685–719.

Maddala, G.: 1983, *Limited-dependent and qualitative variables in econometrics*, Econometric Society Monographs, Cambridge University Press, Cambridge.

Moll, P.: 1993, Black South African unions: Relative wage effects in international perspective, *Industrial and Labor Relations Review* **46**(2), 245–261.

Ntuli, M. and Kwenda, P.: 2014, Labour unions and wage inequality among African men in South Africa, *Development Southern Africa* **31**(2), 322–346.

Schultz, T. P. and Mwabu, G.: 1998, Labor unions and the distribution of wages and employment in South Africa, *Industrial and Labor Relations Review* **51**(4), 680–703.

About DataFirst

DataFirst is a Research Unit and Data Service based at the University of Cape Town, South Africa. We give researchers online access to survey and administrative microdata (data at unit record level) from South Africa and other African countries. We assist researchers to use the data via our online helpdesk and offer formal training courses in microdata analysis.

DataFirst also trains African data managers in microdata curation. We conduct research on the quality and usability of South African microdata, and we work with African microdata producers to improve the quality of their data products."

We aim for a data rich research-policy interface in South Africa, where data reuse by policy analysts in academia serves to refine inputs to government planning.

DataFirst

www.datafirst.uct.ac.za
Level 3, School of Economics Building, Middle Campus, University of Cape Town
Private Bag, Rondebosch 7701, Cape Town, South Africa
Tel: +27 (0)21 650 5708
info@data1st.org / support@data1st.org