Economics Bulletin

Appendix and Supplemental material not intended for publication-Round 1

Submission Number:EB-18-00123

PANEL is the Program for Longitudinal Studies, Experiments and Surveys of CIDE. Acknowledgement: We would like to thank Ronald Oaxaca and Chung Choe for providing us with a copy of their Stata code for implementing their panel decomposition method. We are also grateful to Quetzali Ramirez-Guillen, Valeria Gracia-Olvera, Cristina Alvarez-Venzor and Javier Valbuena for their excellent research assistance. Funding: This work was supported by CONACYT repatriation program for Alfonso Miranda [grant number 191523]. Conflict of Interest: We declare not to have any conflict of interest.
Submitted: Feb 06 2018. **Revised:** March 30, 2018.

Decomposing the language pay gap among the indigenous ethnic minorities of Mexico: Is it all down to observables? (online appendix)

Adriana Aguilar-Rodriguez[a,c], Alfonso Miranda[b,c,d,*], Yu Zhu[e,**]

[a]Center for Research in Geospacial Information Science (CentroGeo), Mexico.
[b]Economics Division, Centre for Economic Research and Teaching (CIDE), Mexico.
[c]Program for Longitudinal Studies, Experiments and Surveys (PANEL), CIDE, Mexico.
[d]Institute for the Study of Labor (IZA), 53113 Bonn, Germany
[e]Economic Studies, University of Dundee, Dundee, DD1 4HN, UK.

1. Wooldridge's correlated random effects (Heckman) sample selection estimator

Here we breifly summarize the methods of Wooldridge (1995) and Wooldridge (2009), p. 834-835. Consider fitting the following system for pooled cross-section data with $i = 1,\ldots,N$ individuals, $m = 1,\ldots,M$ municipalities, and $t = 1,\ldots,T$ periods

$$logw^*_{imt} = \mathbf{x}_{imt}\beta + \theta BIL_{imt} + \mathbf{w}_{mt}\gamma + \delta_t + c_m + u_{imt} \tag{A.1}$$

$$S^*_{imt} = \mathbf{z}_{it}\pi_1 + \mathbf{w}_{mt}\pi_2 + \alpha_t + c_m + v_{imt} \tag{A.2}$$

$$S_{imt} = 1\,(S^*_{it} > 0) \tag{A.3}$$

$$logw_{imt} = \begin{cases} logw^*_{imt} \text{ if } S_{imt} = 1 \\ \text{missing otherwise.} \end{cases} \tag{A.4}$$

We suppose that, conditional on the municipal fixed-effect c_m, all control variables are exogenous and $\varepsilon^s_{imt} = c_m + v_{imt}$, with $\varepsilon^s_{imt} \sim \mathcal{N}(0,1)$. Define $\varepsilon^{logw}_{imt} = c_m + u_{imt}$. Sample selection bias arises whenever $E(\varepsilon^{logw}_{imt}|\varepsilon^s_{imt}) \neq 0$.

Under this model a straightforward extension of the two-step Heckman model is not available because ε^s_{imt} depends on the whole history of selection $S_{im} = \{S_{im1}, S_{im2}, \ldots, S_{imT}\}$ — as opposed

*Corresponding author: Alfonso Miranda, Economics Division, Center for Research and Teaching in Economics (CIDE), Region Centro campus. Circuito Tecnopolo 117, Aguascalientes 20313, Aguascalientes, Mexico. Telephone: +524499945150 ext. 5152.

**Acknowledgement: We would like to thank Ronald Oaxaca and Chung Choe for providing us with a copy of their Stata code for implementing their panel decomposition method. We are also grateful to Quetzali Ramirez-Guillen, Valeria Gracia-Olvera, Cristina Alvarez-Venzor and Javier Valbuena for their excellent research assistance. Funding: This work was supported by CONACYT repatriation program for Alfonso Miranda [grant number 191523]. Conflict of interest: We declare not to have any conflict of interest.

Email addresses: aaguilar@centrogeo.org.mx, alfonso.miranda@cide.edu, yuzhu@dundee.ac.uk
URL: www.cide-panel.org (PANEL)

to being function of S_{imt} only. This is an important complication that requires careful consideration. In this context, Wooldridge suggests an estimator that follows a strategy similar to Chamberlain (1980)'s correlated random effects approach as a way of dealing with the dependency of ε^{s}_{imt} on the whole history of selection. Namely, Wooldridge suggests fitting equation A.2 by probit for each t to get a predicted inverse Mills ratio $\widehat{\lambda}_{imt}$. Then, in a second step, the regression of

$$logw_{imt} \text{ on } BIL_{imt}, \mathbf{x}_{imt}, \bar{\mathbf{x}}_{im}, \mathbf{w}_{mt}, d2_t\mathbf{w}_{mt}, \ldots, dT_t\mathbf{w}_{mt}, \widehat{\lambda}_{imt}, d2_t\widehat{\lambda}_{imt}, \ldots, dT_t\widehat{\lambda}_{imt}$$

is fitted by POLS in the selected sample, where $d2_t, \ldots, dT_t$ are time dummy indicators and $\bar{\mathbf{x}}_m$ is the time average of individual level control variables over time for the m-th municipality. Standard errors are suitably clustered at the municipal level to allow for arbitrary heteroskedasticity or serial correlation. Because we have a two-step estimator, to get valid standard errors it is important to take into account the variation of first stage parameters. In this context, bootstrapping the standard errors is a popular choice.

References

Chamberlain, G. (1980) "Analysis of covariance with qualitative data" *The Review of Economic Studies* **47**, 225–238.

Wooldridge, J. M. (1995) "Selection corrections for panel data models under conditional mean independence assumptions" *Journal of Econometrics* **68**, 115–132.

Wooldridge, J. M. (2009) *"Econometric analysis of cross section and panel data"* MIT press.