Implementing the Oaxaca-Choe decomposition method in Stata

Alfonso Miranda (CIDE)

(alfonso.miranda@cide.edu)

Introduction

- Oaxaca, R. (1973) and Blinder, A. S. (1973) describe methods that the aim is to uncover what proportion of the log-wage gap between two groups, say men and women, is explained by differences in observable characteristics across groups (also known as the 'E' part) and what proportion of the gap is left 'unexplained' once the effect of observables is netted out via regression analysis (also known as the 'U' part).

- he work of Oaxaca and Choe (2016) extends the usual toolkit in two important directions:
  (a) To take into account that the two groups may have different degrees of labour market attachment that contribute to the observed wage gap;
  (b) To take into account the role of unobserved heterogeneity at the panel level.

Some detail

- Oaxaca-Choe decomposition involves fitting Wooldridge (1995)'s correlated random effects (Heckman) sample selection estimator for each compared group, v.g. men and women, to get coefficients on:
  (a) time-varying controls;
  (b) time-fixed controls;
  (iii) inverse Mills ratio terms.

  for decomposing the wage-gap into its Explained, Unexplained, and Selection components.

Wooldridge's CRE (Heckman) sample selection estimator

Consider fitting the following system for pooled cross-section data with $i = 1, \ldots, N$ individuals and $t = 1, \ldots, T$ periods

$$logw^*_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{w}_i\gamma + \delta_t + c_i + u_{it} \tag{A.1}$$

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

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

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

conditional on c_i, all control variables are exogenous and $\epsilon^s_{it} = c_i + v_{it}$, with $\epsilon^s_{it} \sim \mathcal{N}(0, 1)$. Define $\epsilon^{logw}_{it} = c_i + u_{it}$. Sample selection bias arises whenever $E(\epsilon^{logw}_{it} | \epsilon^s_{it}) \neq 0$.

- 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 $S_{im} = \{S_{im1}, S_{im2}, \dots, S_{imT}\}$. This is an important complication.
- Use a CRE approach as a way of dealing with the dependency of ϵ^s_{imt} on the whole history of selection.
- Fitt equation S by probit for each t to get a predicted inverse Mills ratio $\widehat{\lambda}_{imt}$. Then, in a second step, fit the regression of

$$logw_{it} \text{ on } \mathbf{x}_{it}, \bar{\mathbf{x}}_{it}, \mathbf{w}_i, d2_t\mathbf{w}_i, \dots, dT_t\mathbf{w}_i, \widehat{\lambda}_{it}, d2_t\widehat{\lambda}_{it}, \dots, dT_t\widehat{\lambda}_{it}$$

by POLS in the selected sample.

- 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. Bootstrapping the standard errors is a popular choice.

Defining E, U, and S in the panel context

Method 1 The 'explained part' is anything due to differences in characteristics and the 'unexplained part' is anything due to differences in parameters. Differences in c_i and selection are split into their E and U components.

Method 2 Consider differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage as Explained or non discriminatory. That is, given observed characteristics and coefficients in the logit model for $\widehat{\lambda}_{it}$, the correlation between S and $logw$ is considered as explained. Differences in c_i and $\widehat{\lambda}_{it}$ are split into their E and U components.

Method 3 Define the selection component S as containing only differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage. Differences in c_i and $\widehat{\lambda}_{it}$ are split into their E and U components.

Method 4 Define S as anything affecting differences in selection:

(i) differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage,
(ii) differences in characteristics that enter the probit model for $\widehat{\lambda}_{it}$,
(iii) differences in coefficients in the probit model for $\widehat{\lambda}_{it}$.

- The E part contains differences in time-varying and time-fixed characteristics that affects log-wage (including those affecting c_i).
- The U part contains differences in coefficients on time-varying and time-fixed characteristics that affects log-wage (including those affecting c_i).

Method 5 Define E as:

(i) differences in time-varying variables,
(ii) differences in time-fixed variables (including differences time fixed vars that affect c_i),
(iii) differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage,
(iv) differences in characteristics that enter the probit model for $\widehat{\lambda}_{it}$,

- U contains differences in coefficients in time-varying variables, differences in coefficients in the probit model for $\widehat{\lambda}_{it}$.

Method 6 Define E as:

(i) differences in time-varying variables,

- U contains differences in coefficients in time-varying variables,
- S contains differences in time-fixed variables, differences in coefficients on time-fixed variables, differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage, differences in characteristics that enter the probit model for $\widehat{\lambda}_{it}$, differences in coefficients in the probit model for $\widehat{\lambda}_{it}$.

Example with data from the MXFLS Mexican Family Life Survey Home (ENNViH)

```
. de lincome age female $educat sel nchild

                  storage   display    value
variable name     type      format     label      variable label
-------------------------------------------------------------------------------
lincome           float     %9.0g                 log of income per month
age               float     %9.0g                 age
female            float     %9.0g                 female
noschool          float     %9.0g                 No formal schooling
preschool         float     %9.0g                 Preschool or kinder
jrhigh            float     %9.0g                 Jr High
ojrhigh           float     %9.0g                 Open Jr High
highsch           float     %9.0g                 High School
ohighsch          float     %9.0g                 Open High School
tradesch          float     %9.0g                 Trade school
college           float     %9.0g                 College
graduate          float     %9.0g                 Graduate
dksch             float     %9.0g                 Don't know
sel               float     %9.0g                 Positive income
nchild            float     %9.0g                 Number of children<6 years old
```

```
. bysort female: su lincome age female $educat sel nchild
-----------------------------------------------------------------------------------------------------
-> female = 0
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     lincome |      5,852      10.374    .7293295   8.188689   11.69525
         age |      8,746    44.16305    10.66742         20         65
      female |      8,746           0           0          0          0
    noschool |      8,746    .0695175    .2543466          0          1
   preschool |      8,746    .0018294    .0427349          0          1
-------------+---------------------------------------------------------
      jrhigh |      8,746    .2492568    .4326075          0          1
     ojrhigh |      8,746    .0102904    .1009242          0          1
     highsch |      8,746    .1001601    .3002305          0          1
    ohighsch |      8,746    .0052595    .0723359          0          1
    tradesch |      8,746    .0096044    .0975358          0          1
-------------+---------------------------------------------------------
     college |      8,746     .098788    .2983942          0          1
    graduate |      8,746    .0059456    .0768824          0          1
       dksch |      8,746    .0080037    .0891095          0          1
         sel |      8,746    .6691059    .4705619          0          1
      nchild |      8,746    .1808827    .4638866          0          4

-----------------------------------------------------------------------------------------------------
-> female = 1
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     lincome |      2,514    10.10162    .8269909   8.188689   11.69525
         age |     10,618    43.17282    10.63039         20         65
      female |     10,618           1           0          1          1
    noschool |     10,618    .0928612    .2902515          0          1
   preschool |     10,618    .0013185    .0362891          0          1
-------------+---------------------------------------------------------
      jrhigh |     10,618    .2395931    .4268553          0          1
     ojrhigh |     10,618    .0158222    .1247931          0          1
     highsch |     10,618    .0806178    .2722601          0          1
    ohighsch |     10,618    .0030138    .0548174          0          1
    tradesch |     10,618     .014127    .1180199          0          1
-------------+---------------------------------------------------------
     college |     10,618    .0589565    .2355543          0          1
    graduate |     10,618     .002637    .0512867          0          1
       dksch |     10,618    .0065926    .0809304          0          1
         sel |     10,618    .2367678    .4251186          0          1
      nchild |     10,618    .1692409    .4509338          0          4
```

Men are relatively older and have higher qualifications than women

```
. bysort female: su lincome age female $educat sel nchild if sel==1
----------------------------------------------------------------------------------------------------
-> female = 0
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lincome |     5,852      10.374    .7293295   8.188689   11.69525
         age |     5,852    42.83903    10.27546         20         65
      female |     5,852           0           0          0          0
    noschool |     5,852    .0615174    .2402975          0          1
   preschool |     5,852    .0013671    .0369516          0          1
-------------+--------------------------------------------------------
      jrhigh |     5,852     .265892    .4418448          0          1
     ojrhigh |     5,852    .0109364    .1040129          0          1
     highsch |     5,852    .1074846    .3097549          0          1
    ohighsch |     5,852    .0064935    .0803271          0          1
    tradesch |     5,852    .0093985    .0964974          0          1
-------------+--------------------------------------------------------
     college |     5,852    .0849282    .2787987          0          1
    graduate |     5,852    .0032468    .0568926          0          1
       dksch |     5,852    .0046138    .0677739          0          1
         sel |     5,852           1           0          1          1
      nchild |     5,852    .1954887    .4818286          0          4

----------------------------------------------------------------------------------------------------
-> female = 1
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lincome |     2,514    10.10162    .8269909   8.188689   11.69525
         age |     2,514    42.31424    9.365033         20         65
      female |     2,514           1           0          1          1
    noschool |     2,514    .0640414    .2448753          0          1
   preschool |     2,514    .0019889    .0445612          0          1
-------------+--------------------------------------------------------
      jrhigh |     2,514    .2728719    .4455242          0          1
     ojrhigh |     2,514    .0190931    .1368795          0          1
     highsch |     2,514    .1165473     .320944          0          1
    ohighsch |     2,514    .0067621      .08197          0          1
    tradesch |     2,514    .0286396    .1668246          0          1
-------------+--------------------------------------------------------
     college |     2,514    .1077963    .3101847          0          1
    graduate |     2,514    .0031822    .0563322          0          1
       dksch |     2,514    .0043755    .0660158          0          1
         sel |     2,514           1           0          1          1
      nchild |     2,514    .1372315    .4027636          0          4
```

But, among those who work, women have higher qualifications than men

```
Bootstrap results                               Number of obs     =     19,364
                                                Replications      =         20

                         (Replications based on 9,682 clusters in pid_link)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
gap	.2723789	.0185867	14.65	0.000	.2359496	.3088083
E1	-.03733	.0079794	-4.68	0.000	-.0529692	-.0216907
U1	.3097089	.0199107	15.55	0.000	.2706846	.3487332
S1	0	(omitted)				
E2	.1991702	.5650059	0.35	0.724	-.9082209	1.306561
U2	.0732087	.5753357	0.13	0.899	-1.054429	1.200846
S2	0	(omitted)				
E3	-.03733	.0079794	-4.68	0.000	-.0529692	-.0216907
U3	.0732087	.5753357	0.13	0.899	-1.054429	1.200846
S3	.2365002	.5661459	0.42	0.676	-.8731254	1.346126
E4	-.0344269	.00832	-4.14	0.000	-.0507337	-.01812
U4	-.0715769	.4013026	-0.18	0.858	-.8581155	.7149616
S4	.3783828	.3935239	0.96	0.336	-.3929099	1.149675

Most of the wage-gap is due to differences in selection. And most of the difference in selection is due to differences in coefficients on $\widehat{\lambda}_{it}$ in the second stage (that is, given observed characteristics and coefficients in the logit model for $\widehat{\lambda}_{it}$, correlation between S and $logw$).

The end, thanks!!


Plan de Estudios
1.º Semestre
Políticas Públicas
Procesos Políticos y Políticas Públicas
Matemáticas Aplicadas
Estadística
Diseño de Investigación: la lógica de la inferencia
2.º Semestre
Econometría
Inferencia Causal y Métodos Cualitativos
Microeconomía
Métodos Experimentales para el Análisis de Políticas Públicas
Taller: Manejo de Base de Datos y Programación en Stata
3.º Semestre
Seminario de Investigación I
Elegir tres materias optativas
Análisis Espacial
Teoría de Redes Sociales
Análisis Logitudinal
Series de Tiempo
Métodos Cualitativos en Campo: etnografía, historia oral, observación participante
Machine Learning
Big Data
Análisis de Redes Neuronales
4.º Semestre
Evaluación de Impacto: enfoques cuantitativo y cualitativo
Métodos Mixtos Aplicados al Análisis de Políticas
Elegir un curso
Diseño de Encuestas: del muestro al análisis
Métodos Cualitativos y Problemas Públicos Contemporáneos
Seminario de Investigación II
El plan de estudios puede tener modificaciones.
CIDE
CIDE
CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS
Maestría en Métodos para el Análisis de Políticas Públicas
Informes
Lunes a viernes de 9:00 a 17:00 horas
CIDE Santa Fe, Ciudad de México
Oficina de Promoción y Admisiones
Teléfono directo (55) 5727-9811, lada sin costo 01 800 021 24 33
decide@cide.edu
Carretera México-Toluca 3655, col. Lomas de Santa Fe, 01210, Ciudad de México.
CIDE Región Centro, Ciudad de Aguascalientes
Oficina de Admisiones y Control Escolar
Teléfono (449) 2514821 / 9945150 exts. 5255 y 5192
decideregioncentro@cide.edu
Circuito Tecnopolo Nte 117, Fracc. Tecnopolo Pocitos II, C.P. 20313, Aguascalientes, Ags.
http://cide.edu/programa/metpol/
www.cide.edu

El Centro de Investigación y Docencia Económicas (**CIDE**) es una institución especializada en investigación y educación superior en el área de Ciencias Sociales. El CIDE forma parte de la red de Centros Públicos del Consejo Nacional de Ciencia y Tecnología (CONACYT).

El **CIDE** ofrece diez posgrados, todos registrados en el Programa Nacional de Posgrados de Calidad (PNPC):

Programas de tiempo completo:
- Doctorados: Políticas Públicas; Ciencia Política

- Maestrías: **Métodos para el Análisis de Políticas Públicas** (*en proceso de registro al PNPC*); Historia Internacional; Economía; Economía Ambiental; Ciencia Política; Administración y Políticas Públicas

Programas profesionalizantes:
- Maestría en Gerencia Pública

Programas vinculados con la industria:
- Maestría en Periodismo sobre Políticas Públicas

Las y los alumnos de la **Métodos para el Análisis de Políticas Públicas** no pagan colegiatura y son becarios de CONACYT, por lo que reciben una beca de manutención de aproximadamente $11,000 pesos (monto estimado 2018).

La fortaleza de nuestros programas académicos se basa en:
• Estudiantes de alto rendimiento que pasaron por un exigente y riguroso proceso de selección.
• Planta académica de alto nivel, experta en los temas que imparte; 95% de la planta cuenta con doctorado y 75% pertenece al Sistema Nacional de Investigadores.

Maestría en Métodos para el Análisis de Políticas Públicas

Es un programa de nueva creación, único en América Latina, que forma profesionistas en el uso riguroso de métodos cuantitativos y cualitativos para el análisis de políticas públicas.

Los estudiantes conocerán el estado del arte en las principales teorías sociales, métodos de análisis y estrategias de recolección de datos en experimentos, estudios observacionales y de caso.

Perfil de ingreso
Son bienvenidos profesionistas de cualquier área con interés en aproximarse a los problemas públicos de México y el mundo desde distintas perspectivas y disciplinas. Buscamos aspirantes que deseen desarrollar su capacidad analítica, lógica matemática y pensamiento crítico.

Perfil de egreso
Los egresados serán capaces de formular y evaluar políticas públicas basadas en evidencia, un perfil que tiene alta demanda en el sector público y en áreas del sector privado vinculadas a la economía de la información. Asimismo, nuestro perfil tiene alta demanda en los mejores programas de doctorado del mundo.

Profesorado
Los profesores son egresados de los mejores programas doctorales internacionales en ciencias sociales y contribuyen a la investigación de frontera en temas como salud, política de drogas, educación y medio ambiente.
Los estudiantes tendrán la oportunidad de colaborar en proyectos de investigación y aprovechar la riqueza de enfoques metodológicos.
En la sede Región Centro encontrará una comunidad vibrante que dialoga sobre los grandes problemas nacionales desde diferentes disciplinas y métodos de análisis.

References

- Aguilar-Rodriguez, A., Miranda, A., Zhu, Yu. (2018). Decomposing the language pay gap among the indigenous ethnic minorities of Mexico: is it all down to observables? Economics Bulletin 38 (2): 689-695.
- Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates, The Journal of Human Resources 8,436–455.
- Heckman, J. J. (1979). Sample Selection Bias as a Specification Error, Econometrica, 47,153–161.
- Jann, B. (2008). The Blinder–Oaxaca decomposition for linear regression models, The Stata Journal 8(4): 453-479.
- Oaxaca, R. (1973). Male-female wage differentials in urban labor markets, International Economic Review 14,693–709.
- Oaxaca, R. and Choe, C. (2016). Wage decompositions using panel data sample selection correction, Korean Economic Review 32, 201–218.
- Wooldridge, J. M. (1995). Selection corrections for panel data models under conditional mean independence assumptions, Journal of Econometrics 68: 115–132.
- Wooldridge, J.M. (2010). Econometric Analysis of Cross Section and Panel Data (2nd edition). The MIT Press.