2nd. STATA Users Group Meeting Mexico

Discussion of user-written Stata programs

Selection bias correction based on the multinomial logit: an application to the Mexican labour market.

Luis Huesca
Mario Camberos

Centro Conacyt de Investigación en Alimentación y Desarrollo, A.C.
Department of Economics.
Email: lhuesca@ciad.mx
mcamberos@ciad.mx

April 29, 2010, Universidad Iberoamericana Campus Mexico.

Goal.-

Application of the two step method ado-file –*selmlog*– explained in a robust manner by Bourguignon et al. (2004) and formally published by Bourguignon, Fourier and Gurgand (2007) -*JES*-.

Technical problem:

OLS becomes inefficient. Determination of wages generally causes a high correlation between the non-observable characteristics affecting wages and those that simultaneously determine the sector in which the individuals are currently /located/ functioning (working).

This will cause to obtain not only biased, but also inconsistent coefficients.

Evidence and facts

Bivariate selection bias
Heckman 1979

Earnings equations
Mincer 1974

Technique

Multilogit Correction (It works for nested especifications as well)

- Lee 1983 u_i and (η_j) are correlated , *iid?* $\longrightarrow (\sigma \;\; \rho_1)$
  Not true for the joint distribution.

- Dubin and McFadden 1984 : No assump. on cov. u_1 and $(\eta_j - \eta_1)$ & multicollinearity exists.

- Schmertmann 1994 u_1 and $(\eta_j - \eta_1)$ equal sign & *iid*

  ↓ Very strong hypothesis in empirical studies

- Dahl 2002

- Bourguignon, et al. (2007)
  Allows *corr* between choices $y_1 = x_1\beta_1 + \mu(P_1, \ldots, P_M) + w_1$
  w_1 is mean independent of the regresors
  $(\sigma_1\rho_1), \ldots, (\sigma_1\rho_s)$

Huesca (2005) and Zheren (2008). Recent applications BFG: Mexico and China.

Evidence and facts

As opposite to the bivariate case, when the number of events exceeds two categories, previous techniques (Lee, 1983 and so forth) present restrictions on the structure of the error terms and, generally, an inappropriate application – since those methods have been elaborated with the requirement of using an univariate transformation order–.

A correction for multivariate cases was developed in Dubin and McFadden (DM, 1984); this technique could not evaluate a model strong enough to admit maximum likelihood estimators, with complete information for the case were the number of choices were greater than two.

DM (1984) provides a model where the *J* sector must be required to establish a *J*-1 selection terms.

Bourguignon, et al. (2007) consider the case where the underlying selection process follows a polychotomous normal model, allowing correlations between alternatives.

Techinique and ado selmlog

Must be understood by the self selection of individuals in the information and self-handling of the data that identical individual exists when using samples defined with a nonrandom criterion

two step generalized methodology proposed by BFG for polytomous cases is used, allowing OLS implementation in the calculations.

$$y_s = x_s \beta'_s + u_s$$

Let's assume the information follows a Gumbel distribution G(.) *iid* for sake of normality. The following model is considered with a categorical variable *S = 1,…,M* choices based on utilities y_s^* for the individuals as follows:

$$y_s^* = z_s \gamma_s + \eta_s,$$

Where Z and η_s compose a vector of independent variables and the disturbance term which confirms the usual conditions.
The impact on the dependent variable is observed just for the case where the alternative *S* is chosen which happens when:

$$y_s^* > \max_{j \neq s}(y_j^*)$$

$$\varepsilon_s = \max_{j \neq s}(y_j^* - \eta_s); \varepsilon_s < 0$$

the vector η_s is *iid* and Gumbel distributed; thus, their respective cumulative and density functions are

$$G(\eta) = \exp(-e^{-\eta})$$

$$g(\eta) = \exp(-\eta - e^{-\eta})$$

(See McFadden, 1973). It is in this part of the model where the multimominal logit specification applies in the traditional way:

$$P(z_s \gamma_s > \varepsilon_s) = \frac{\exp(z_s \gamma_s)}{\sum_j \exp(z_j \gamma_j)}$$

$$y_1 = x_1 \beta_1 - \sigma_1 [\rho_1 m(P_1) + \sum_{s>1} \rho_s \frac{P_s}{P_s - 1} m(P_s)] + v_s$$

$\beta_{i1}^{'}$ stands for coefficients and x_{i1} as attributes of the individual. The residual term displays the usual normality statistical conditions.

$m(P_s)$ are the probabilities and $(\sigma_1 \rho_1), \ldots, (\sigma_1 \rho_s)$ the coefficient terms for the polychotomous correction of selectivity bias; v_s **is an orthogonal error** parameter towards the rest of terms, having a mean expectation equal to zero. This last property is what allows using directly the OLS procedure in the estimation.

1. $P(z_s\gamma_s > \varepsilon_s) = \dfrac{\exp(z_s\gamma_s)}{\sum_j \exp(z_j\gamma_j)}$ Logit

2. $\ln y_s = \beta_s^{'} x_s + \varepsilon_s - \sigma_{\eta u}\rho_s^{'}$ Replacing terms, using a vector of Rhos

One problem that arises from this occupational selection process technique is related to the IIA as stated by Hausman and McFadden (1984). Bourguignon, *et al*. (2004; 2007) can provide fairly good correction for the outcome equation, even when the IIA hypothesis is violated in nested models.

1. exp Setting $\rho 12 = \rho 13 = \rho 23 = 0$ misspecifies param. dist.
2. exp Small corr $\rho 12 = 0.1, \rho 13 = 0.1, \rho 23 = -0.2$
3. exp Violation IIA $\rho 12 = -0.1, \rho 13 = 0.45, \rho 23 = -0.35$

$$\sigma^2 = 32, corr(u_1,\eta_1) = 0.64, corr(u_1,\eta_2) = -0.24, corr(u_1,\eta_3) = 0.14.$$

$$\begin{bmatrix} \sigma^2 & \sigma_{1u}^2 & \sigma_{2u}^2 & \sigma_{3u}^2 \\ \sigma_{1u}^2 & 1 & \rho 12 & \rho 13 \\ \sigma_{2u}^2 & \rho 12 & 1 & \rho 23 \\ \sigma_{3u}^2 & \rho 13 & \rho 23 & 1 \end{bmatrix}$$

Ensuring orthogonality so that $V[h(\eta_1,\eta_2,\eta_3)] \approx 0.16$

Empirical case

Answer the following questions: Will the differences in earnings between the formal and informal sectors of the labor market in Mexico be statistically significant? Which are the socioeconomic and occupational factors that mostly affect earnings amongst sectors?.

Logit has a practical advantage over probit when the sum of the predicted values equal to the sum of empirically observed values (Butcher and Dinardo, 1998.)

ENOE: Encuesta Nacional de Ocupación y Empleo: 2009-III.

Males and females aging from 16 to 65

Occupations = (1 ,…, 4)

Multinomial Logit

- 1: Formal self-employed
- 2: Informal self-employed
- 3: Formal wage-earners
- 4: Informal wage-earners

features for empirical application

```
To download it:
net from http://www.pse.ens.fr/gurgand/
```

To avoid endogeneity from the sample selection process we select for the objective earnings equation a vector of family background (highly recommended!).

Lee (1983), Dubin-McFadden (1984) and Dahl (2002) can be computed with selmlog as well. See help selmlog:

options **[lee dmf(#) dhl(# [all])**

`dhl` options include the order of the polynomials on the selection probabilities. With this number alone, the correction term includes only the probability to be selected on the observed outcome. If this number is followed by all, probabilities are included in polynomial form, with interactions, up to the specified order.

Syntax

1. Compute the earnings distribution using selmlog command.

```
selmlog depvar varlist [ifexp][inrange],select(depvar_m=varlist_m)
        [lee dmf(#) dhl(# [all]) showmlogit wls
        bootstrap(number_of_replications[sample_size])
        mloptions(mlogit options) gen(variable generic name)]
```

2. Computing the empirical case (Weighted Least Squares -wls- to account for heteroskedasticity present in the model due to selectivity).

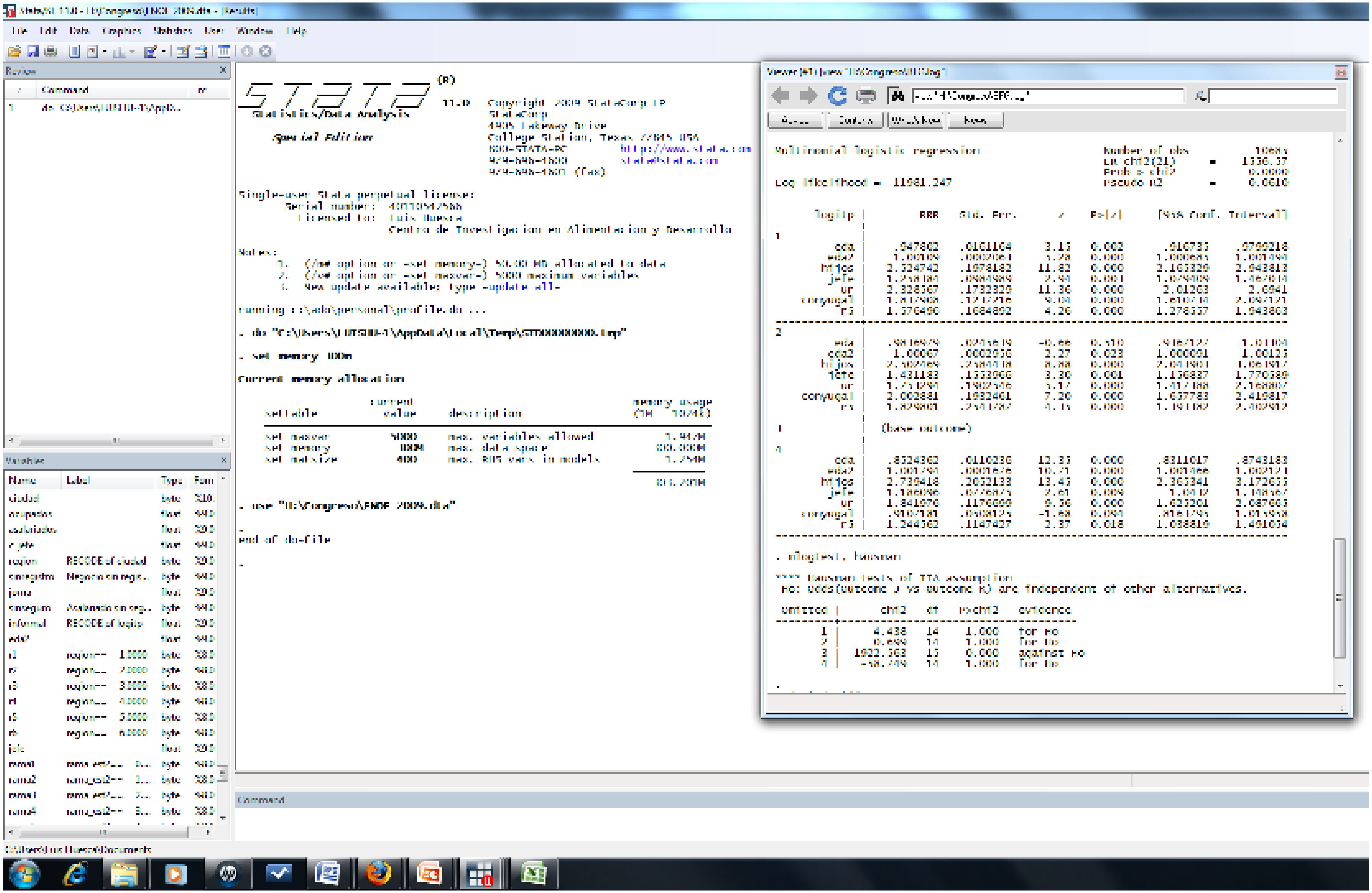
```
****Formal Self-employed:
selmlog logw1 anios_esc eda eda2 rama2 rama4 rama5 rama6 rama8  ///
if logw>0, select(logitp = eda hijos jefe ur conyugal) ///
dmf(2) wls bootstrap(100) mloptions(rrr level (95)) gen(rho_1)

****Informal Self-employed:
selmlog logw2 anios_esc eda eda2 rama2 rama4 rama5 rama6 rama8  ///
if logw>0, select(logitp = eda hijos jefe ur conyugal) ///
dmf(2) wls bootstrap(100) mloptions(rrr level (95)) gen(rho_2)

****Formal wage-earner:
selmlog logw3 anios_esc eda eda2 rama2 rama4 rama5 rama6 rama8  ///
if logw>0, select(logitp = eda hijos jefe ur conyugal) ///
dmf(2) wls bootstrap(100) mloptions(rrr level (95)) gen(rho_3)

****Informal wage-earner:
selmlog logw4 anios_esc eda eda2 rama2 rama4 rama5 rama6 rama8  ///
if logw>0, select(logitp = eda hijos jefe ur conyugal) ///
dmf(2) wls bootstrap(100) mloptions(rrr level (95)) gen(rho_4)
```

Multi-Logit


Multinomial logistic regression
LR chi2(21) = 1336.37
Pseudo R2 = 0.0610
Log likelihood = 11081.247
RRR
Std. Err.
[95% Conf. Interval]
(base outcome)
Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.
chi2
df
P>chi2
evidence
for Ho
against Ho
Single-user Stata perpetual license:
Special Edition
Statistics/Data Analysis
end of do-file
Current memory allocation

Multi-Logit

Viewer (#1) [view "H:\Congreso\BFG.log"]

view "H:\Congreso\BFG.log"

Advice | Contents | What's New | News

```
Multinomial logistic regression                   Number of obs   =      10685
                                                  LR chi2(21)     =    1556.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -11981.247                       Pseudo R2       =     0.0610

------------------------------------------------------------------------------
      logitp |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
         eda |    .947802   .0161164    -3.15   0.002      .916735    .9799218
        eda2 |    1.00109   .0002063     5.28   0.000     1.000685    1.001494
       hijos |   2.524742   .1978182    11.82   0.000     2.165329    2.943813
        jefe |   1.258384   .0984989     2.94   0.003     1.079409    1.467034
          ur |   2.328567   .1732329    11.36   0.000      2.01263      2.6941
    conyugal |   1.837908   .1237216     9.04   0.000     1.610734    2.097121
          r5 |   1.576496   .1684892     4.26   0.000     1.278557    1.943863
-------------+----------------------------------------------------------------
2            |
         eda |   .9836979   .0245639    -0.66   0.510     .9367127     1.03304
        eda2 |    1.00067   .0002956     2.27   0.023     1.000091     1.00125
       hijos |   2.502469   .2584438     8.88   0.000     2.043903    3.063917
        jefe |   1.431183   .1553966     3.30   0.001     1.156837    1.770589
          ur |   1.753294   .1902546     5.17   0.000     1.417388    2.168807
    conyugal |   2.002881   .1932461     7.20   0.000     1.657783    2.419817
          r5 |   1.829801   .2543787     4.35   0.000     1.393382    2.402912
-------------+----------------------------------------------------------------
3            |  (base outcome)
-------------+----------------------------------------------------------------
4            |
         eda |   .8524362   .0110236   -12.35   0.000     .8311017    .8743183
        eda2 |   1.001794   .0001676    10.71   0.000     1.001466    1.002123
       hijos |   2.739418   .2052133    13.45   0.000     2.365341    3.172655
        jefe |   1.186096   .0776875     2.61   0.009       1.0432    1.348567
          ur |   1.841976   .1176699     9.56   0.000     1.625201    2.087665
    conyugal |   .9107181   .0508125    -1.68   0.094     .8163795    1.015958
          r5 |   1.244562   .1147427     2.37   0.018     1.038819    1.491054
------------------------------------------------------------------------------

. mlogtest, hausman

**** Hausman tests of IIA assumption
 Ho: Odds(Outcome-J vs Outcome-K) are independent of other alternatives.

 Omitted |      chi2   df   P>chi2   evidence
---------+------------------------------------
       1 |    -4.438   14    1.000   for Ho
       2 |     0.699   14    1.000   for Ho
       3 |  1922.563   15    0.000   against Ho
       4 |   -58.749   14    1.000   for Ho
----------------------------------------------

.
```

Selmlog command using BFG (Lee)

```
Selectivity correction based on multinomial logit
Second step regression
Bootstrapped standard errors (100 replications)
```

logw1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logw1						
anios_esc	-.0015537	.0045909	-0.34	0.735	-.0105516	.0074442
eda	-.0070126	.0110103	-0.64	0.524	-.0285924	.0145672
eda2	.0000231	.0001264	0.18	0.855	-.0002245	.0002708
rama2	.2603791	.1421182	1.83	0.067	-.0181675	.5389256
rama4	-.0673728	.0629758	-1.07	0.285	-.1908031	.0560575
rama5	-.6115123	.281147	-2.18	0.030	-1.16255	-.0604743
rama6	.013802	.0369808	0.37	0.709	-.058679	.086283
rama8	-.222056	.2259293	-0.98	0.326	-.6648694	.2207573
_m1	.1546043	.1118948	1.38	0.167	-.0647056	.3739141
_cons	6.215879	.334825	18.56	0.000	5.559634	6.872124
Ancillary						
sigma2	.5954783	.1259639	4.73	0.000	.3485936	.8423631
rho	.2003496	.1206472	1.66	0.097	-.0361146	.4368138

```
.
end of do-file

.
```

Command

Selmlog command using (dmf(0)) Dubin-McFadden [1]

Selectivity correction based on multinomial logit
Second step regression
Bootstrapped standard errors (100 replications)

logw1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logw1						
anios_esc	-.0017757	.0039499	-0.45	0.653	-.0095173	.005966
eda	-.0090108	.0114329	-0.79	0.431	-.0314188	.0133972
eda2	.0000463	.0001292	0.36	0.720	-.000207	.0002995
rama2	.2633093	.1395896	1.89	0.059	-.0102812	.5368998
rama4	-.0640174	.0540833	-1.18	0.237	-.1700188	.041984
rama5	-.6069665	.3037591	-2.00	0.046	-1.202323	-.0116096
rama6	.0148896	.0354548	0.42	0.675	-.0546005	.0843797
rama8	-.2273555	.2071633	-1.10	0.272	-.6333882	.1786772
_m2	.0868744	.5257237	0.17	0.869	-.9435251	1.117274
_m3	.0084358	.2283435	0.04	0.971	-.4391092	.4559808
_m4	.0013133	.2748971	0.00	0.996	-.5374752	.5401018
_cons	6.225136	.35089	17.74	0.000	5.537404	6.912868
Ancillary						
sigma2	.5230322	.3951511	1.32	0.186	-.2514498	1.297514
rho2	.1540642	.6441352	0.24	0.811	-1.108418	1.416546
rho3	.0149602	.3021221	0.05	0.961	-.5771881	.6071086
rho4	.002329	.355908	0.01	0.995	-.6952379	.6998959

.
end of do-file

.

Command

Selmlog command using (dmf(1)) Dubin-McFadden [2]

-all correlation coefficients sum-up to zero-

```
Selectivity correction based on multinomial logit
Second step regression
Bootstrapped standard errors (100 replications)
```

logw1	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
logw1						
anios_esc	-.001937	.0066486	-0.29	0.771	-.014968	.011094
eda	-.0047785	.0148917	-0.32	0.748	-.0339657	.0244087
eda2	-1.55e-06	.0001717	-0.01	0.993	-.0003381	.000335
rama2	.250801	.1822725	1.38	0.169	-.1064466	.6080486
rama4	-.0605817	.0633696	-0.96	0.339	-.1847838	.0636205
rama5	-.6005671	.2961673	-2.03	0.043	-1.181044	-.0200898
rama6	.0184748	.0405116	0.46	0.648	-.0609265	.0978762
rama8	-.2428746	.2314719	-1.05	0.294	-.6965512	.210802
_m1	-.0623649	.1937439	-0.32	0.748	-.442096	.3173662
_m2	.772115	2.668067	0.29	0.772	-4.4572	6.00143
_m3	.4767232	1.60533	0.30	0.766	-2.669665	3.623111
_m4	.4728991	1.584504	0.30	0.765	-2.632671	3.578469
_cons	6.750295	2.153209	3.13	0.002	2.530084	10.97051
Ancillary						
sigma2	1.25447	3.443743	0.36	0.716	-5.495144	8.004083
rho1	-.0714142	.1494277	-0.48	0.633	-.3642872	.2214588
rho2	.8841504	1.330742	0.66	0.506	-1.724055	3.492356
rho3	.5458967	.7634907	0.72	0.475	-.9505175	2.042311
rho4	.5415177	.7639026	0.71	0.478	-.9557039	2.038739

```
.
end of do-file

.
```

Command

Selmlog command using BFG (dmf(2))

```
Selectivity correction based on multinomial logit
Second step regression
Bootstrapped standard errors (100 replications)
------------------------------------------------------------------------------
       logw1 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logw1        |
   anios_esc |   .0491248   .0048183    10.20   0.001     .0106928    .0819471
       rama2 |   .2426938   .0806902     3.01   0.029     .1811145    .2968432
       rama4 |  -.0559203   .0610801    -0.92   0.361    -.1756351    .0637945
       rama5 |  -.6100826   .3005651    -2.03   0.042    -1.199179   -.0209857
       rama6 |   .0225234   .0045593     4.94   0.022    -.0668373    .0311884
       rama8 |  -.2639496   .2446007    -1.08   0.081    -.7433581    .0154593
          r1 |   .1154463   .0096481    11.97   0.000     .0736529    .1545455
          r2 |   .0558856   .0088745     6.30   0.019     .0280532    .0824401
          r3 |   .0481392   .0892543     0.54   0.596    -.1267961    .2230745
          r4 |   .1249463   .0517734     2.41   0.039     .0949263    .1848189
          r6 |   .1829422   .0986251     1.85   0.064     .1035946    .2262438
         _m1 |  -.0545378   .3358121    -0.16   0.871    -.1271664    .6036427
         _m2 |   1.162403   1.005957     1.16   0.068     .6983051    2.023111
         _m3 |   .3831932   .0771588     4.97   0.032     .2909283    .4954781
         _m4 |   .2251297   .0696599     3.23   0.047     .1901813    .2904432
       _cons |   6.162259   .5968858    10.32   0.000    4.9923841    7.332133
-------------+----------------------------------------------------------------
>
Anciliary    |
      sigma2 |   .8620013   .4056217    2.13    0.054    -1.892966    3.61696
        rho1 |  -.0587404   .0250426   -2.35    0.051    -0.549567    .432086
        rho2 |   1.251995   .0948975   13.19    0.000     1.227962    1.28111
        rho3 |   .4127279   .0538643    7.66    0.004     .3993362    .468449
        rho4 |   .2424817   .0531211    4.56    0.031     .2198673    .263537
------------------------------------------------------------------------------

.
. ****Cuenta propia informal:
. selmlog logw2 anios_esc rama2 rama4 rama5 rama6 rama8 r1 r2 r3 r4 r6 ///
> if logw>0, select(logitp = eda eda2 hijos jefe ur conyugal r5) ///
> dmf(2) wls bootstrap(100) mloptions(rrr level (95)) gen(rho_2)

Selectivity correction based on multinomial logit
Second step regression
Bootstrapped standard errors (100 replications)
------------------------------------------------------------------------------
       logw2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logw2        |
   anios_esc |   .0186409   .0087613     2.13   0.033     .0014691    .0358127
```

Conclusions :

`selmlog` command is a useful tool to correct selection bias in polytomous cases (From Lee to BFG).

The empirical application confirms for the Mexican case, that choices are selected in a non-randomly process: Individuals decide where to work!

An advantage is not to depend on the IIA-Hausman-Mc Fadden’s test for nested models.

Our suggestion is not to specify models with a great number of covariates when computing the ado.

In earnings equations use familiar background as variables for selection.

The inference with a great number of reps is time consuming, 100 reps is recommended.

References

Bourguignon Francois, Fournier M. and Gurgand Marc (2004). Selection Bias Corrections Based on the Multinomial Logit Model: Monte-Carlo comparisons, mimeo Delta, (download from http\\:www.pse.ens.fr\senior\gurgand\selmlog13.htm).

Bourguignon, François, M. Fournier and M. Gurgand (2007) "Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons.", *Journal of Economic surveys*, 21, pp. 174-205.

Butcher, K. F. and John Dinardo (1998), "The immigrant and native-born wage distributions: Evidence from united states census", *NBER Working paper No. 6630.*

Dahl G. B., "Mobility and the Returns to Education: Testing a Roy Model with Multiple Markets", *Econometrica*, vol. 70, 2367-2420, 2003.

Dubin, J. A. and D. L. McFadden. (1984) "An Econometric Analysis of Residential Appliance Holdings and Consumption." *Econometrica*, 52 (March), pp. 345-62.

Hausman, J. and D. McFadden (1984) "Specification tests for the multinomial logit model." *Econometrica 52* (5), pp. 1219-40.

Heckman, James (1979) "Sample selection bias as a specification error", *Econometrica* Vol. 47(1), pp. 153-61.

Huesca Luis (2005) "La Distribución salarial del mercado de trabajo en México: Un análisis de la Informalidad", PhD thesis, Department of Applied Economics, Universitat Autónoma d'Barcelona.

Lee L.F., "Generalized Econometric Models with Selectivity", *Econometrica*, vol. 51, 507-512, 1983.

McFadden, D. L. (1973) "Conditonal Logit Analysis of Qualitative Choice Behavior." *Frontiers in Econometrics*, Academic Press.

Mincer, J. (1974) Schooling, experience and earnings. Columbia University Press.

Schmertmann, C. (1994) "Selectivity Bias Correction Methods in Polychotomous Sample Selection Models." *Journal of Econometrics*, 60 (January-February), pp. 101-32.

Zheren, Wu (2008) "Self-selection and earnings of migrants: Evidence from rural China", Discussion paper 08-25, Graduate School of Economics and Osaka School of International Public Policy (OSIPP), Japan.