A SAS macro to perform Dong and Lewbel's "Simple Estimator for Binary Choice Models" (2015)

Nicolas Moreau

DOCUMENTS DE TRAVAIL CEMOI / CEMOI WORKING PAPERS

A SAS macro to perform Dong and Lewbel's "Simple Estimator for Binary Choice Models" (2015)

Nicolas Moreau[1]

http://cemoi.univ-reunion.fr

Centre d'Economie et de Management de l'Océan Indien
Université de La Réunion

January 2017

Abstract

This paper presents a SAS macro to estimate Dong and Lewbel's "Simple Estimator for Binary Choice Models with Endogenous Regressors" (2015) on cross-section data.

JEL codes: C25, C26.

[1] E-mail: nicolas.moreau@univ-reunion.fr

Introduction

We introduce the SAS macro *specreg* written to perform Dong and Lewbel's "Simple Estimator for Binary Choice Models with Endogenous Regressors" (2015) on cross-section data. The computation requires several steps, all of which are written in *specreg* in SAS Interactive Matrix Language (IML) without resorting to existent procedures. We hope that this helps the user to gain a clear understanding of the "mechanics" behind the estimator.

Although it is usually advisable to use matrix calculation rather than "do loops" in SAS/IML, *specreg* relies heavily on looping, because otherwise it is unlikely to obtain estimates in a reasonable amount of time for very large datasets. The complete SAS/IML code is available at http://cemoi.univ-reunion.fr.

A brief presentation of Dong and Lewbel's (2015) simple estimator for binary choice models

The binary choice model is:

$$D = I(X\beta + V + \varepsilon \geq 0),$$

where*D* is the dependent variable, *V* the special regressor, and *X* the matrix of all other regressors in the model. The special regressor is supposed to have a positive impact on the binary outcome; its impact is normalized to 1. Some or all of the elements of *X* can be endogenous. Let *Z* be the matrix of excluded instruments from the model. Let *S* be the matrix of all the elements of *X* and *Z*.

To find the vector of estimates $\hat{\beta}$, it is necessary to:
- run an OLS regression of *V* on *S* and save the residuals $\hat{u}$ from this regression.
- perform White's test to assess whether the error term u from the preceding regression is heteroscedastic. This involves running an OLS regression of $\hat{u}^2$on $\tilde{S}$, where $\tilde{S}$ includes all the elements of *S* and all the squares and cross products of all the elements of *S*.Let $\tilde{S}\hat{c}$ be the fitted values from this regression.
- estimate for each observation *i* the nonparametric density $\hat{f}(\hat{u}_i)$ if the error term u is homoscedastic or otherwise the nonparametric density $\hat{f}(\frac{\hat{u}_i}{\sqrt{\tilde{S}_i\hat{c}}})$.
- construct for each observation *i* the variable $\hat{T}_i$ defined as $\hat{T}_i = [D_i - I(V_i \geq 0)]/\hat{f}_i$ if the error term u is homoscedastic or otherwise as $\hat{T}_i = [D_i - I(V_i \geq 0)][\sqrt{\tilde{S}_i\hat{c}}]/\hat{f}_i$.
- run a IV regression of $\hat{T}$ on *X* using excluded instruments *Z* to obtain $\hat{\beta}$.

The estimated mean marginal effects of *X* on choice probabilities are calculated as $\overline{m}\hat{\beta} = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_i\,\hat{\beta}$, with $\hat{m}_i$ being the first derivative of the average index function with respect to *X* evaluated at X_i (see Lewbel, Dong and Yang, 2012).

Finally, statistical inference for the parameters of interest and marginal effects at the mean are provided with bootstrapping.

Syntax of the SAS macro *specreg*

The syntax is %macro
specreg(data=,depvar=,specreg=,regendo=,regexo=,instr=,density=,trim=,het=,hetvar=,ilpm=,nboot=);

Data specifies the data set. *depvar* is the binary outcome variable. *specreg* is the special regressor. *regexo* specifies the list of exogenous regressors; this list can be empty. *regendo* specifies the list of endogenous regressors; if this list is empty, IV regressions reduce to OLS estimations. *instr* specifies the list of excluded instruments from the regression; *instr* can be empty if there are no endogenous regressors in *regendo*. All variables in *depvar*, *specreg*, *regexo*, *regendo*, and *instr* must be numeric.

The parameter *density* is used to select a non-parametric estimator for the density of the residuals, with the possibility of five choices. The default is the sorted data estimator of Lewbel (2000) and Lewbel and Schennach (2007) if *density* is not specified by the user.

If *density*=1, the standard normal kernel (mean=0, standard deviation=1) with Silverman's rule of thumb for bandwidth selection is computed. If *density*=2, the Epanechnikov kernel with Silverman's rule of thumb for bandwidth selection is used instead.

If *density*=1 or *density*=2, direct evaluation of the kernel density occurs at any point *i*. This involves *N* kernel evaluations per observation requiring a total of *N*×*N* kernel evaluations, where *N* is the sample size.
If *density*=3 or *density*=4, direct evaluation of the kernel density at any point *i* only involves 401 evaluations from a grid, requiring a total of *N*×401 kernel evaluations. This significantly decreases calculation time for a large *N*. If *density*=3 (or 4), the standard normal (Epanechnikov) kernel is used.

The parameter *trim* specifies the percentage applied to remove the top and bottom values of $\hat{T}$. For instance, when *trim*=5, the left-most and right-most 5% of observations are removed. When *trim*=0.05, the bottom and top 0.05% are excluded. The default is no trimming if *trim* is not specified by the user.

White's test for heteroscedasticity of the special regressor is automatically supplied. If it is heteroscedastic (at the 5% level), *specreg* switches to the estimator robust to heteroscedasticity. The user can still request the primary estimator to be computed by specifying the parameter *het*. If *het*=no, *het*=0, or *het*=none (or whatever character/value except a blank), the primary estimator is computed regardless of whether the homoscedasticity assumption is rejected or not.

To perform White's test, all the elements of *S* and all the squares and cross products of all the elements of *S* are automatically used (see Dong and Lewbel, 2015). Some of these variables may be redundant if *S* includes dummies or squared variables, for instance. Duplicate variables are then removed.

The user may not want to perform the pure form of White's test if the corresponding regression involves a large number of covariates. With the parameter *hetvar*, users can specify their own list of squares and cross products to add to *S*. The default is the pure form of White's test if *hetvar* is not specified by the user.

Nboot is the number of bootstrapped samples drawn to provide bootstrap standard errors for the estimated parameters and marginal effects at the mean. No statistical inference is produced if *nboot* is not specified be the user.

Note that *specreg* will stop executing if:
- the dependent variable or the special regressor are not specified
- the dependent variable is not binary 0 and 1
- the number of excluded instruments is less than the number of endogenous regressors
- not any regressors are specified
- some variables have missing values.

Results presentation and output data files

A first table provides descriptive statistics for the dependent variable, regressors, and excluded instruments.

A second table includes preliminary estimates from an instrumental linear probability model. This preliminary estimation procedure is used to verify the positive effect of the special regressor on the dependent variable. As Dong and Lewbel (2015) point out, a positive effect is requested for the estimator to be valid. If the estimated sign appears to be negative, the negative of the special regressor is automatically used afterwards. However, if *ilpm*=no, *ilpm*=0, or *ilpm*=none (or whatever

character/value except a blank), this step is ignored: the linear probability model is not estimated and the possible change of sign of the special regressor not made.

White's test for heteroscedasticity with the special regressor is then presented.

Two validity checks are shown, both of which relate to the support of the special regressor. The first evaluates whether the empirical distribution of the special regressor exhibits excess kurtosis. A negative value indicates a light-tailed distribution relative to the standard normal, which may weaken the results. The second compares the spread of observations of the special regressor to the spread of the fitted values. Measures of the empirical spread of the special regressor have to be at least comparable to those of $X\hat{\beta}$.

A third table presents the parameter estimates along with bootstrap standard errors, corresponding t-statistics, and p-values. The fourth table provides the 99%, 95%, and 90% confidence intervals for the parameters of interest, respectively. The confidence intervals are computed using the percentile method.

A fifth table shows the estimates for the marginal effects at the mean along with bootstrap standard errors, corresponding t-statistics, and p-values. The sixth and final table includes the 99%, 95%, and 90% confidence intervals for the marginal effects at the mean, respectively.

Four temporary output data files are created. *discarded obs* includes the internal identification number of observations that are removed from the estimation procedure when trimming. *bootselec* includes the internal identification number of observations contained in each bootstrap sample. *bviboot* includes the parameter estimates obtained on each bootstrap sample, whereas *mean_meffects_boot* includes the corresponding estimated mean marginal effects.

Some examples

The dependent variable is y_1, the special regressor is x_0, the endogenous regressor is x_1. All variables from x_2 to x_{12} are exogenous regressors. The excluded instruments are x_{13} and x_{14}. All variables except x_2 and x_3 are binary. The number of observations is 19957.

Macro call (1):

%***specreg***(data=file1,depvar=y1,specreg=X0,regendo=X1,regexo=X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12, instr=X13 X14,density=**2**,trim=**5**,het=no,hetvar=,ilpm=,nboot=**99**),

performs the estimator with the Epanechnikov kernel to estimate the nonparametric density of the residuals. The Epanechnikov kernel is also used to estimate the marginal effects at the mean. The left-most and right-most 5% of observations are removed, while 99 bootstrap replications are used for statistical inference.

All the elements of *regendo*, *regexo*, and *instr* as well as all the non-redundant squares and cross products of all of these elements are used to perform White's test.

The simple estimator that does not account for heteroscedasticity is computed, even if the assumption of homoscedasticity is rejected.

In contrast, the macro call:

%***specreg***(data=file1,depvar=y1,specreg=X0,regendo=X1,regexo=X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12, instr=X13 X14,density=**2**,trim=**5**,het=,hetvar=,ilpm=,nboot=**99**),

performs the simple estimator that accounts for heteroscedasticity if the assumption of homoscedasticity is rejected. All the elements of *regendo*, *regexo*, and *instr* as well as all the non-redundant squares and cross products of all of these elements are used to perform the estimator.

The macro call:

%***specreg***(data=file1,depvar=y1,specreg=X0,regendo=X1,regexo=X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12, instr=X13 X14,density=**2**,trim=**5**,het=,hetvar=X22 X33 X23,ilpm=,nboot=**99**),

only uses the variables x_{22}, x_{33}, and x_{23} as cross products and squares to perform White's test and the estimator robust to heteroscedasticity, with $x_{22}= x_2x_2$, $x_{33}= x_3x_3$, and $x_{23}= x_2x_3$.

Computation takes 1 h 1 min with macro call (1); the corresponding outputs are presented below. Computation time decreases to 31 min with the nonparametric estimation of the density of the residuals with 401 evaluations per observation (*density*=4); it is reduced to 32 min when the sorted density estimator (default, when *density*=) is used. In comparison, computation takes 1 h 52 min with the standard normal kernel (*density*=1) and 54 min with the standard normal kernel and 401 grid point evaluations (*density*=3).

Desccriptive statistics

Variable	N	N Miss	Minimum	10th Pctl	Lower Quartile	Median	Mean
y1	19957	0	0.00	0.00	0.00	1.00	0.70
x0	19957	0	21.00	27.00	31.00	35.00	34.54
x1	19957	0	0.00	0.00	0.00	0.00	0.39
x2	19957	0	14.00	17.00	19.00	22.00	22.43
x3	19957	0	0.00	2.00	2.00	4.00	4.25
x4	19957	0	0.00	0.00	0.00	0.00	0.40
x5	19957	0	0.00	0.00	0.00	0.00	0.27
x6	19957	0	0.00	0.00	0.00	0.00	0.09
x7	19957	0	0.00	0.00	0.00	0.00	0.08
x8	19957	0	0.00	0.00	0.00	0.00	0.19
x9	19957	0	0.00	0.00	0.00	0.00	0.21
x10	19957	0	0.00	0.00	0.00	0.00	0.20
x11	19957	0	0.00	0.00	0.00	0.00	0.19
x12	19957	0	0.00	0.00	0.00	0.00	0.04
x13	19957	0	0.00	0.00	0.00	1.00	0.51
x14	19957	0	0.00	0.00	0.00	0.00	0.01

Variable	Std Dev	Upper Quartile	90th Pctl	Maximum
y1	0.46	1.00	1.00	1.00
x0	4.87	39.00	40.00	41.00
x1	0.49	1.00	1.00	1.00
x2	4.17	25.00	28.00	38.00
x3	2.68	6.00	8.00	23.00
x4	0.49	1.00	1.00	1.00
x5	0.44	1.00	1.00	1.00
x6	0.28	0.00	0.00	1.00
x7	0.27	0.00	0.00	1.00
x8	0.39	0.00	1.00	1.00
x9	0.41	0.00	1.00	1.00
x10	0.40	0.00	1.00	1.00
x11	0.39	0.00	1.00	1.00
x12	0.19	0.00	0.00	1.00
x13	0.50	1.00	1.00	1.00
x14	0.12	0.00	0.00	1.00

Preliminary estimation with the instrumental variables linear probability model

Parameter Estimate	Standard Error	T-statistic	P-value

INTERCEPT	0.605	0.035	17.361	0
X0	0.015	0.002	7.935	0
X1	-0.148	0.041	-3.622	0
X2	-0.011	0.003	-4.415	0
X3	-0.005	0.003	-1.723	0.085
X4	-0.185	0.01	-18.504	0
X5	-0.08	0.01	-8.36	0
X6	0.067	0.011	6.086	0
X7	0.131	0.011	12.31	0
X8	-0.02	0.01	-2.044	0.041
X9	-0.002	0.01	-0.166	0.868
X10	0.011	0.01	1.17	0.242
X11	0.031	0.01	3.173	0.002
X12	-0.057	0.018	-3.228	0.001

Message: duplicate variables in Stild were found and deleted
The number of distinct columns in stild (not including the constant term) is 94

White's test for heteroscedasticity of the special regressor X0

White's test	
Test statistic	P-value
3398.7612	0

Trimming applied: the left-most and right-most 5% of observations are excluded, 1798 observations are discarded

Excess kurtosis for special regressor X0 is -0.353887
Caution! Negative value for excess kurtosis may weaken the results

Comparing measures of spread of the special regressor to those of the fitted values Xbhat

	Variance	1%	5%	10	25%
Special regressor	20.252	-11.544	-8.544	-6.544	-2.544
Xbhat	6.266	-4.712	-2.587	-1.43	0.328

Comparing measures of spread of the special regressor to those of the fitted values Xbhat

	50%	75%	90%	95%	99%
Special regressor	0.456	3.456	5.456	6.456	6.456
Xbhat	2.059	3.61	4.914	5.699	7.137

Caution! The spread of the special regressor is requested to be comparable or larger

Estimates and Statistical Inference with bootstrap standard errors

	Parameter estimate	Standard error	T-statistic	P-value
INTERCEPT	20.015214	0.8711846	22.974712	0
X1	-1.31235	0.6952769	-1.887521	0.0590903
X2	-0.613811	0.0244926	-25.06108	0
X3	-0.447161	0.037092	-12.05544	0
X4	-3.907761	0.210291	-18.58263	0
X5	-1.483964	0.1637462	-9.062588	0
X6	0.6238221	0.2032307	3.0695272	0.002144
X7	0.8090402	0.2253164	3.5906852	0.0003298

X8	-0.289195	0.1800046	-1.6066	0.1081422
X9	0.0719056	0.1458722	0.4929357	0.622058
X10	0.1116195	0.1609223	0.6936235	0.4879184
X11	0.0753156	0.1589741	0.4737598	0.6356712
X12	-0.194633	0.337562	-0.576584	0.5642204

Bootstrap Confidence intervals for the model parameters with the percentile method

	0.5%	2.5%	5%	95%	97.5%	99.5%
INTERCEPT	18.257	18.599	18.856	21.642	22.062	22.554
X1	-3.762	-2.934	-2.704	-0.236	-0.211	-0.05
X2	-0.674	-0.668	-0.658	-0.578	-0.577	-0.576
X3	-0.555	-0.536	-0.52	-0.393	-0.385	-0.36
X4	-4.32	-4.282	-4.274	-3.57	-3.489	-3.463
X5	-1.924	-1.857	-1.805	-1.261	-1.212	-1.206
X6	0.096	0.13	0.317	0.981	1.031	1.108
X7	0.352	0.408	0.482	1.233	1.239	1.454
X8	-0.643	-0.597	-0.548	0.042	0.089	0.149
X9	-0.149	-0.144	-0.112	0.376	0.42	0.488
X10	-0.187	-0.148	-0.133	0.41	0.456	0.649
X11	-0.224	-0.167	-0.128	0.416	0.443	0.483
X12	-0.867	-0.735	-0.668	0.419	0.537	0.588

Marginal effects at the mean and Statistical Inference with bootstrap standard errors

	Parameter estimate	Standard error	T-statistic	P-value
INTERCEPT	0.8414446	0.085513	9.8399577	0
X1	-0.055172	0.0334514	-1.649303	0.0990856
X2	-0.025805	0.0024717	-10.44031	0
X3	-0.018799	0.0026378	-7.126726	1.028E-12
X4	-0.164283	0.0076241	-21.54802	0
X5	-0.062386	0.0065495	-9.52536	0
X6	0.0262256	0.0090043	2.9125739	0.0035846
X7	0.0340123	0.010051	3.3839835	0.0007144
X8	-0.012158	0.0076573	-1.587745	0.1123441
X9	0.0030229	0.0061319	0.4929806	0.6220263
X10	0.0046925	0.0067074	0.699598	0.4841784
X11	0.0031663	0.0066026	0.4795481	0.6315488
X12	-0.008182	0.014412	-0.567749	0.5702055

Bootstrap Confidence Intervals for marginal effects at the mean with the percentile method

	0.5%	2.5%	5%	95%	97.5%	99.5%
INTERCEPT	0.684	0.694	0.711	1.007	1.036	1.118
X1	-0.186	-0.136	-0.126	-0.009	-0.008	-0.002
X2	-0.033	-0.031	-0.03	-0.022	-0.022	-0.021
X3	-0.028	-0.025	-0.024	-0.015	-0.014	-0.014
X4	-0.186	-0.183	-0.181	-0.152	-0.151	-0.149
X5	-0.082	-0.078	-0.076	-0.054	-0.052	-0.047
X6	0.004	0.006	0.013	0.042	0.045	0.049
X7	0.014	0.016	0.02	0.054	0.055	0.062
X8	-0.028	-0.025	-0.024	0.002	0.004	0.006
X9	-0.006	-0.006	-0.004	0.016	0.018	0.019
X10	-0.008	-0.007	-0.006	0.018	0.019	0.024
X11	-0.009	-0.006	-0.006	0.017	0.018	0.019
X12	-0.036	-0.032	-0.031	0.018	0.023	0.026

References

Dong, Y. and A. Lewbel (2015), "A Simple Estimator for Binary Choice Models With Endogenous Regressors", Econometric Reviews, 34, 82-105.

Lewbel, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables", Journal of Econometrics, 97, 145-177.

Lewbel, A. Dong, Y., and T. Yang (2012), "Comparing Features of Convenient Estimators for Binary Choice Models With Endogenous Regressors", Canadian Journal of Economics, 45, 809-829.

Lewbel, A. and S. Schennach (2007), "A Simple Ordered Data Estimator for Inverse Density Weighted Functions", Journal of Econometrics, 186, 189-211.