ivporbit:An R package to estimate the probit model with continuous endogenous regressors

Zaghdoudi, Taha

University of Jendouba, Faculty of Law Economics and Management of Jendouba

25 September 2014

ivporbit:An R package to estimate the probit model with continuous endogenous regressors

Taha Zaghdoudi*

25-10-2014

Abstract

One of the most important problem of misspecification in the probit model is the correlation between regressors and error term. To deal with this problem, some commercial software gives a solution such as Stata. For the famous R language the ivprobit gives the users the way to estimate the instrumental probit model.

1 Introduction

Many econometrics study are focused in various kinds of misspecification in the limited-dependent variable models such as correlation between regressors and error term which produce inconsistent results. To avoid this problem we apply the instrumental variable method in which we use the correlated variables as instruments. A two-stage method are used by some authors Blundell and Powell [2004] to fit the probit model but it produce non efficient result.

Otherwise, Newey [1987] expose an efficient way to estimate limited-dependent variable model by using the Amemiya's Generalized Least Squares estimators Amemiya [1978]. The idea is to include in the two-stage model a continuous endogenous regressor. This method is used in Stata when the MLE fail to estimate the model. For the ivprobit we apply the same method as used in Stata.

In the following sections we expose the Newey method to estimate the instrumental probit model then we give a simple example how to use the ivprobit Package.

2 Estimation method

Generally the model is:

$$y_{1i}^{*} = y_{2i}\beta + x_{1i}\gamma + \mu_i$$
$$y_{2i} = x_{1i}\Pi_1 + x_{2i}\Pi_2 + \upsilon_i.$$

*University of Jendouba, Faculty of Law, Economics and Management of Jendouba

Where $i = 1,..,N$, y_{2i} is a $1 \times p$ vector of endogenous variables, x_{1i} is a 1_1 vector of exogenous variable, x_{2i} is a $1 \times k_2$ vector of additional instruments, and the equation for y_{2i} is written in reduced form. By assumption, $(\upsilon_i, \upsilon_i)\tilde{}N(0,\Sigma)$, where σ_{11} is normalized to one to identify the model. β and γ are vectors of structural parameters, and Π_1 and Π_2 are matrices of reduced-form parameters. This is a recursive model: y_{2i} appears in the equation for y^*_{1i}, but y^*_{1i} does not appear in the equation for y_{2i}. We do not observe y^*_{1i}; instead, we observe:

$$y_{1i} = \{^{0\ y^*_{1i} \geq 0}_{1\ y^*_{1i} \leq 1}$$

The order condition for identification of the structural parameters requires that $k_2 \geq$. Presumably, is not block diagonal between μ_i and υ_i; otherwise, y_{2i} would not be endogenous.

To obtain the two-step estimates,Newey [1987] use minimum chi-squared estimator and the model is:

$$\text{y}^*{}_{1i}\text{=z}_i\delta\text{+}\mu_i \ \text{y}_{2i}\text{=x}_i\Pi + \upsilon_i \tag{1a}$$

(1b)

where z_i=(y$_{2i}$,x$_{1i}$),x$_i$=(x$_{1i}$, x$_{2i}$), δ=(β',δ')', $and\Pi$=(Π_1', Π_2')'.

The reduced-form equation for y^*_{1i} is:

$$y^*_{1i} = (x_i\Pi + \upsilon_i)\beta + x_{1i}\gamma + \mu_i = x_i\alpha + \upsilon_i\beta + \mu_i = x_i\alpha + \nu_i$$

where

$\nu_i = \upsilon_i\beta + \mu_i$. Because μ_i and υ_i are jointly normal,$_i$ is also normal. Note that:

$$\alpha = \{^{\Pi_1}_{\Pi_2}\}\beta + \{^{I}_{0}\}\gamma = D(\Pi)\delta$$

where $D = (\Pi) = (\Pi, I_1)$ and I_1 is defined such that x_iI_i. Letting $\widehat{z_i} = (x_i\widehat{\Pi}, x_{1i}), \widehat{z_i} = x_iD(\widehat{\Pi})\delta$, where $D(\widehat{\Pi}) = D(\widehat{\Pi}, I_1)$. Thus one estimator of α is $D(\widehat{\Pi})\delta$; donate this estimator by $\widehat{D}\delta$.

α could also be estimated directly as the solution to:

$$max_{\alpha,\lambda} = \sum_{i=1}^{N} l(y_{1i}, x_i\alpha + \widehat{\upsilon}_i\lambda)$$

where $l(.)$ is the log likelihood for probit. Denote this estimator by $\hat{\alpha}$. The inclusion of the term $\widehat{\upsilon}_i\lambda$ follows because the multivariate normality of (μ_i, υ_i) implies that, conditional on y_{2i}, the expected value of μ_i is nonzero. Because υ_i is unobservable, the least-squares residuals from fitting (1b) are used.

Amemiya [1978] shows that the estimator of δ:

$$\max_{\delta}(\widetilde{\alpha} - \hat{D}\delta)'\hat{\Omega}^{-1}(\widetilde{\alpha} - \hat{D}\delta) \tag{2}$$

where$\hat{\Omega}$ is a consistent estimator of the covariance of $\sqrt{N}(\widetilde{\alpha} - \hat{D}\delta)$, is asymptotically efficient relative to all other estimators that minimize the distance between $\widetilde{\alpha}$ ans D($\hat{\Pi}$) δ. Thus an efficient estimator of δ is :

$$\hat{\delta}\text{=}(\hat{D}'\hat{\Omega}^{-1}\hat{D})\ ^{-1}(\hat{D}'\hat{\Omega}^{-1}\widetilde{\alpha}) \tag{3}$$

and

$$\text{Var}(\hat{\delta})\text{=}(\hat{D}'\hat{\Omega}^{-1}\hat{D})\ ^{-1} \tag{4}$$

To implement this estimator, we need $\hat{\Omega}^{-1}$. Consider the two-step maximum likelihood estimator that results from first fitting (1b) by OLS and computing the residuals $\hat{\upsilon}_i = y_{2i} - x_i\,\hat{\Pi}$. The estimator is then obtained by solving:

$$\underset{\delta,\lambda}{max} = \sum_{i=1}^{N} l(y_{1i}, z_i\delta + \hat{\upsilon}_i\lambda)$$

This is the two-step instrumental variables (2SIV) estimator proposed by Rivers and Vuong [1988], and its role will become apparent shortly.

From Proposition 5 of [Newey, 1987],$\sqrt{N}(\widetilde{\alpha}\text{-}\hat{D}\delta) \underset{\rightarrow}{d} N(0,\Omega)$, where

$$\Omega = J_{\alpha\alpha}^{-1} + (\lambda - \beta)'\Sigma_{22}(\lambda - \beta)Q^{-1}$$

and $\Sigma_{22} = E\{\upsilon_i'\upsilon_i\}.J_{\alpha\alpha}^{-1}$ is simply the covariance matrix of $\widetilde{\alpha}$, ignoring that $\hat{\Pi}$ is an estimated parameter matrix. Moreover, Newey shows that the covariance matrix from an OLS regression of $y_{i2}(\hat{\lambda} - \hat{\beta})$ on x_i is a consistent estimator of the second term.$\hat{\lambda}$ can be obtained from solving (2), and the 2SIV estimator yields a consistent estimate, $\hat{\beta}$.

The basic algorithm proceeds as follows:

1. Each of the endogenous right-hand-side variables is regressed on all the exogenous variables, and the fitted values and residuals are calculated. The matrix $\hat{D} = D(\hat{\Pi})$ is assembled from the estimated coefficients.
2. probit is used to solve (2) and obtain $\widetilde{\alpha}$ and $\widetilde{\lambda}$. The portion of the covariance matrix corresponding to α, $J_{\alpha\alpha}^{-1}$, is also saved.
3. The 2SIV estimator is evaluated, and the parameters $\hat{\beta}$ corresponding to y_{2i} are collected.
4. $y_{i2}(\hat{\lambda}\text{-}\hat{\beta})$ is regressed on x_i. The covariance matrix of the parameters from this regression is added to $J_{\alpha\alpha}^{-1}$, yielding $\hat{\Omega}$.
5. Evaluating (3) and (4) yields the estimates $\hat{\delta}$ and Var($\hat{\delta}$).
6. A Wald test of the null hypothesis H 0 : λ=0, using the 2SIV estimates, serves as our test of exogeneity.

3 Example

To dwonload the package you can just type install.packages(”ivprobit”) in R or Rstudio.

```
# load the package
library(ivprobit)
```

The data we use in this example represents the Foreign-Exchange Derivatives Use By Large U.S. Bank Holding Companies from 1996 to 2000.

```
# load the database
dat<-system.file("data","eco.RData",package="ivprobit")
load(dat)
pro<-library("ivprobit", lib.loc="C:/Program Files/R/R-3.0.3/library")
#load data
dat<-system.file("data","eco.RData",package="ivprobit")
load(dat)
```

The function is ivprobit(y x,y2 z, data):

- y: the dichotomous l.h.s vector
- x: the r.h.s. exogenous variables matrix.
- y2: the r.h.s. endogenous variables vector or matrix
- z: the complete set of instruments (a matrix)
- data: the dataframe

In this example we have d2 the dichotomus l.h.s vector, the r.h.s exogenous variables are (ltass,roe and div), the r.h.s. endogenous variables matrix is (eqrat,bonus) and the instrumental variables are (ltass,roe,div,gap,cfa).

```
# fit the instrumental probit model
 pro<-ivprobit(d2~ltass+roe+div,cbind(eqrat,bonus)~ltass+roe+div+gap+cfa,mydata)
 # the results summary
 summary(pro)

##                     Coef         S.E.  t-stat   p-val
## (Intercept) -1.6862e+01  9.7317e+00 -1.7327 0.08355 .
## X1ltass      9.4760e-01  6.2070e-01  1.5267 0.12724
## X1roe        6.6492e-02  1.1312e-01  0.5878 0.55684
## X1div        2.0378e-06  4.2745e-06  0.4767 0.63368
## yhateqrat    9.4371e+00  1.2937e+01  0.7295 0.46593
## yhatbonus   -1.4173e-06  2.8163e-06 -0.5032 0.61493
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4 Conclustion

In this paper we introduced the ivprobit to estimate the parameters of an dichotomous choice model that contains endogenous regressors. The routine is simple and yields the same results as the two-step option in the commercially available Stata software.

References

Takeshi Amemiya. The estimation of a simultaneous equation generalized probit model. *Econometrica: Journal of the Econometric Society*, 46:1193–1205, 1978.

Richard W Blundell and James L Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

Whitney K Newey. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, 36(3):231–250, 1987.

Douglas Rivers and Quang H Vuong. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3):347–366, 1988.