An Instrumental Variables Probit Estimator using gretl

Lee C. Adkins

Professor of Economics, Oklahoma State University, Stillwater, OK 74078
lee.adkins@okstate.edu

Abstract. The most widely used commercial software to estimate endogenous probit models offers two choices: a computationally simple generalized least squares estimator and a maximum likelihood estimator. Adkins [1, 2] compares these estimators to several others in a Monte Carlo study and finds that the GLS estimator performs reasonably well in some circumstances. In this paper the small sample properties of the various estimators are reviewed and a simple routine using the gretl software is given that yields identical results to those produced by Stata 10.1. The paper includes an example estimated using data on bank holding companies.

1 Introduction

Yatchew and Griliches [19] analyze the effects of various kinds of misspecification on the probit model. Among the problems explored was that of errors-in-variables. In linear regression, a regressor measured with error causes least squares to be inconsistent and Yatchew and Griliches find similar results for probit. Rivers and Vuong [14] and Smith and Blundell [16] suggest two-stage estimators for probit and tobit, respectively. The strategy is to model a continuous endogenous regressor as a linear function of the exogenous regressors and some instruments. Predicted values from this regression are then used in the second stage probit or tobit. These two-step methods are not efficient, but are consistent. Consistent estimation of the standard errors is not specifically considered and these estimators are used mainly to test for endogeneity of the regressors–not to establish their statistical significance.

Newey [12] looked at the more generic problem of endogeneity in limited dependent variable models (which include probit and tobit). He proposed what is sometimes called Amemiya's Generalized Least Squares (AGLS) estimator as a way to efficiently estimate the parameters of probit or tobit when they include a continuous endogenous regressor. This has become one of the standard ways to estimate these models and is an option (twostep) in Stata 10.0 when the MLE is difficult to obtain. The main benefit of using this estimator is that it yields a consistent estimator of the variance-covariance matrix that and can easily be used for subsequent hypothesis tests about the parameters.

Adkins [1] compares the AGLS estimator to several alternatives, which include a maximum likelihood estimator. The AGLS estimator is simple to compute and yields significance tests that are close in size to the nominal level when samples are not very large (e.g., n=200). The other plug-in estimators are consistent for the parameters but not the standard errors, making it unlikely that they will perform satisfactorily in hypothesis testing. The latter problem is taken up by Adkins [3] who uses a Murphy and Topel [11] correction to obtain consistent standard errors with some success.

Others have explored limited dependent variable models that have discrete endogenous regressors. Nicoletti and Peracchi [13] look at binary response models with sample selection, Kan and Kao [10] consider a simulation approach to modeling discrete endogenous regressors, and Arendt and Holm [5] extends Nicoletti and Peracchi [13] to include multiple endogenous discrete variables.

Iwata [9] uses a very simple approach to dealing with errors-in-variables for probit and tobit. He shows that simple recentering and rescaling of the observed dependent variable may restore consistency of the standard IV estimator if the true dependent variable and the IV's are jointly normally distributed. His Monte Carlo simulation shows evidence that the joint normality may not be necessary to obtain improved results. However, the results for tobit were quite a bit better than those for probit. He compares this estimator to a linear instrumental variable estimator that uses a consistent estimator of standard errors. This estimator is considered by Adkins [1] in his comparison.

Blundell and Powell [6] develop and implement semiparametric methods for estimating binary dependent variable models that contain continuous endogenous regressors. Their paper "extends existing results on semiparametric estimation in single-index binary response models to the case of endogenous regressors. It develops an approach to account for endogeneity in triangular and fully simultaneous binary response models." Blundell and Powell [6], p. 655

In the following sections a linear model with continuous endogenous regressors and its estimators are considered. With respect to models having a dichotomous dependent variable, a relatively simple generalized least squares estimator discussed in Newey [12] is presented and an algorithm for its computation in gretl is given. To give the reader an idea of how this estimator compares to alternatives, including a maximum likelihood estimator (mle), some results from a simulation study conducted by Adkins [1, 2] are summarized. The results from the gretl routine and from Stata 10 are compared using an example from the banking literature.

2 Linear Model and Estimators

Following the notation in Newey [12], consider a linear statistical model in which the continuous dependent variable will be called y_t^* but it is not directly observed. Instead, we observe y_t in only one of two possible states. So,

$$y_t^* = Y_t\beta + X_{1t}\gamma + u_t = Z_t\delta + u_t, \quad t = 1, \ldots, N \tag{1}$$

where $Z_t = [Y_t, X_{1t}], \delta' = [\beta', \gamma']$, Y_t is the t^{th} observation on an endogenous explanatory variable, X_{1t} is a 1xs vector of exogenous explanatory variables, and δ is the $qx1$ vector of regression parameters.

The endogenous variable is related to a $1Xk$ vector of instrumental variables X_t by the equation

$$Y_t = X_{1t}\Pi_1 + X_{2t}\Pi_2 + V_t = X_t\Pi + V_t \tag{2}$$

where V_t is a disturbance. The $k - s$ variables in X_{2t} are additional exogenous explanatory variables. Equation (2) is the reduced form equation for the endogenous explanatory variable. Without loss of generality only one endogenous explanatory variable is considered below. See Newey [12] for notation extending this to additional endogenous variables.

When the continuous variable y_t^* is observed, then one could use either least squares or instrumental variable estimator to estimate δ. Collecting the n observations into matrices y^*, X, and Z of which the t^{th} row is y_t^*, X_t, and Z_t, respectively we have the least squares estimator of δ, $\hat{\delta}_{ols} = (Z^TZ)^{-1}Z^Ty^*$, which is biased and inconsistent.

The instrumental variable estimator uses the orthogonal projection of Z onto the column space of X, i.e., P_XZ where $P_X = X(X^TX)^{-1}X^T$. The IV estimator is

$$\delta_{liv} = (Z^TP_XZ)^{-1}Z^TP_Xy^*. \tag{3}$$

The (linear) instrumental variable estimator is biased in finite samples, but consistent. The heteroskedasticity robust estimator of covariance Davidson and MacKinnon [7], p. 335 is

$$\hat{\Sigma}_{HCCME} = (Z^TP_XZ)^{-1}Z^TP_X\hat{\Phi}P_XZ(Z^TP_XZ)^{-1} \tag{4}$$

where $\hat{\Phi}$ is an nxn diagonal matrix with the t^{th} diagonal element equal to $\hat{u}_t^2$, the squared IV residual.

3 Binary Choice Model and Estimators

In some cases, y_t^* is not directly observed. Instead, we observe

$$y_t = \begin{cases} 1 & y_t^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Assuming the errors of the model (1) are normally distributed leads to the probit model.

3.1 Linear, MLE, and Plug-in

There are several estimators of this model, some consistent for δ and others not. The first is least squares. The least squares estimator $\hat{\delta}_{ols} = (Z^T Z)^{-1} Z^T y^*$ is consistent if Z is exogenous. If any of the elements of Z are endogenous then it is not. Still, it is easy to compute and the degree of inconsistency may be small in certain circumstances.

The linear instrumental variable estimator (3) is also inconsistent and heteroscedastic. Iwata [9] suggests a means of rescaling and recentering (RR) the data that can bring about consistency in this case. However, in his Monte Carlo the RR versions of OLS and IV estimation don't perform particularly well for probit (although much better for tobit).

The usual probit mle can be used to estimate the parameters. However, when any of the regressors are endogenous, then this estimator is also inconsistent (Yatchew and Griliches [19]). To develop the notation, let the probability that y_t is equal one be denoted

$$pr(y_t = 1) = \Phi(y_t, Y_t\beta + X_{1t}\gamma) = \Phi(y_t, Z_t\delta) \tag{6}$$

where Φ is the normal cumulative density, y_t is the observed binary dependent variable, and $Y_t\beta + X_{1t}\gamma$ is the (unnormalized) index function. As usual, the model is normalized assuming $\sigma^2 = 1$. Basically, this equation implies that Y_t, and X_{1t} be included as regressors in the probit model and the log likelihood function is maximized with respect to $\delta^T = [\beta^T, \gamma^T]$. Since the endogeneity of Y_t is ignored, the mle is inconsistent.

Another estimator uses predicted values of Y_t from a first stage least squares estimation of equation (2). Denote the first stage as $\hat{Y}_t = X_{1t}\hat{\Pi}_1 + X_{2t}\hat{\Pi}_2 = X_t\hat{\Pi}$ where $X_t = [X_{1t} \vdots X_{2t}]$ and $\hat{\Pi}^T = [\hat{\Pi}_1^T \vdots \hat{\Pi}_2^T]$. Then the conditional probability

$$pr(y_t = 1) = \Phi(y_t, \hat{Z}_t\delta) \tag{7}$$

with $\hat{Z}_t = [\hat{Y}_t \vdots X_{1t}]$. The parameters are found by maximizing the conditional likelihood. This is referred to here as IV probit (IVP). Although IVP is consistent for δ the standard errors estimated as the outer product of the gradient are not. This can be easily remedied using a Murphy and Topel [11] type correction.

Another estimator adds the least squares residuals from equation (2), $\hat{V}_t = Y_t - X_t\hat{\Pi}$ to (7). This brings

$$pr(y_t = 1) = \Phi(y_t, \hat{Y}_t\beta + X_{1t}\gamma + \hat{V}_t\lambda) = \Phi(y_t, \hat{Z}_t\delta + \hat{V}_t\lambda) \tag{8}$$

which is estimated by maximum likelihood, again conditional on $\hat{\Pi}$. This is similar to an estimator used by Rivers and Vuong [14] which takes the form

$$pr(y_t = 1) = \Phi(y_t, Z_t\delta + \hat{V}_t\rho) \tag{9}$$

The parameter ρ is related to λ in (8) by $\lambda = \rho + \beta$. This follows because $Z_t\delta = \hat{Z}_t\delta + \hat{V}_t\beta$. This estimator is useful for testing endogeneity, but seldom used to estimate δ.

3.2 AGLS

An efficient alternative to (8), proposed by Newey [12], and credited to Amemiya, is a generalized least squares estimator (AGLS). The AGLS estimator of the endogenous probit model is fairly easy to compute, though there are several steps–more than the two suggested by the name of its option in Stata. The basic algorithm proceeds as follows:

1. Estimate the reduced form (2), saving the estimated residuals, $\hat{V}_t$ and predicted values $\hat{Y}_t$.
2. Estimate the parameters of a reduced form equation for the probit model using the mle. In this case,

   $$pr(y_t = 1) = \Phi(y_t, X_t\alpha + \hat{V}_t\lambda) \tag{10}$$

   Note that all exogenous variables, X_{1t} and instruments X_{2t} are used in the probit reduced form and the parameters on these variables is labeled α. Let the mle be denoted $\hat{\alpha}$. Also, save the portion of the estimated covariance matrix that corresponds to $\hat{\alpha}$, calling it $\hat{J}^{-1}_{\alpha\alpha}$.
3. Another probit model is estimated by maximum likelihood. In this case it is the 2SIV estimator of equation (8). Save $\hat{\rho} = \hat{\lambda} - \hat{\beta}$ which is the coefficient of $\hat{V}_t$ minus that of $\hat{Y}_t$.
4. Multiply $\hat{\rho}Y_t$ and regress this on X_t using least squares. Save the estimated covariance matrix from this, calling it $\hat{\Sigma}$.

5. Combine the last two steps into a matrix, $\Omega = \hat{J}_{\alpha\alpha}^{-1} + \hat{\Sigma}$.
6. Then, the AGLS estimator is

$$\delta_A = [D(\hat{\Pi})^T \Omega^{-1} D(\hat{\Pi})]^{-1} D(\hat{\Pi})^T \Omega^{-1} \hat{\alpha} \tag{11}$$

The estimated variance covariance is $[D(\hat{\Pi})^T \Omega^{-1} D(\hat{\Pi})]^{-1}$ and $D(\hat{\Pi}) = [\hat{\Pi} \vdots I_1]$ where I_1 is a kxs selection matrix such that $X_{1t} = X_t I_1$.

The AGLS estimator is one of the options available in Stata 10 (the other is an mle). Adkins [2, 1] conducts a Monte Carlo simulation to compare the bias of each of these estimators as well as the size of nominal 10% significance test of model parameter. He finds that in some circumstances the AGLS estimator performs reasonably well and can be used successfully to test for significance, especially if the sample is small and the instruments not very strong. The main findings of Adkins [1] are reproduced in the next section.

4 Summary of Simulation Results from Adkins (2008)

The main results of Adkins [1] can be summarized as follows:

1. When there is no endogeneity, OLS and Probit work well (as expected). Bias is very small and tests have the desired size.
2. It is clear that OLS and Probit should be avoided when you have an endogenous regressor. Both estimators are significantly biased and significance tests do not have the desired size.
3. Weak instruments increases the bias of AGLS. The bias worsens as the correlation between the endogenous regressor and the equation's error increases.
4. The actual size of a parameter significance test based on the instrumental variable probit is reasonably close to the nominal level in nearly every instance. This is surprising for at least two reasons. 1) The bias of IVP is substantial when instruments are weak. 2) The test statistic is based on an inconsistent estimator of the standard error. No attempt was made to estimate the covariance of this estimator consistently, as is done in Limdep 9 Greene [8]. This is explored further in Adkins [3] who uses a Murphy and Topel [11] correction to obtain consistent standard errors.
5. The size of the significance tests based on the AGLS estimator is also reasonable, but the actual size is larger than the nominal size–a situation that gets worse as severity of the endogeneity problem increases. When instruments are very weak, the actual test rejects a true null hypothesis twice as often as it should.

6. Linear instrumental variables estimators that use consistent estimators of standard errors can be used for this purpose (significance testing) though their performance is not quite up to that of the AGLS estimator. The Linear IV estimator performs better when the model is just identified.
7. There is an improvement in bias and the size of the significance test when samples are larger (n=1000). Mainly, smaller samples (n=200) require stronger instruments in order for bias to be small and tests to work properly (other than IVP, which as mentioned above, works fairly well all the time).
8. There is little to be gained by pretesting for endogeneity. When instruments are extremely weak it is outperformed by the other estimators considered, except when the no endogeneity hypothesis is true (and probit should be used). Bias is reduced by small amounts, but it is uncertain what one would use as an estimator of standard errors for a subsequent t-test.
9. When instruments are weak, t-tests based on ML are no better than ones based on AGLS (in fact, one could argue that they are worse). Significance testing based on the ML estimator is much more reliable in large samples.

The picture that emerges from this is that the AGLS estimator may be useful when the sample is relatively small and the instruments not very strong. It is also useful when the mle cannot be computed–a situation which limited the simulations conducted by Adkins [1, 2]. Given the usefulness of the AGLS estimator, a gretl script is provided to compute is and its standard errors. The script is provided below in section 6. In the next section, a brief example is given the results from Stata 10 and the gretl script are compared.

5 Example

In this section a brief example based on Adkins et al. [4] is presented and the results from Stata and gretl compared.

The main goal of Adkins et al. [4] was to determine whether managerial incentives affect the use of foreign exchange derivatives by bank holding companies (BHC). There was some speculation that several of the variables in the model were endogenous. The dependent variable of interest is an indicator variable that takes the value 1 if the BHC uses foreign exchange derivative. The independent variables are as follows:

Ownership by Insiders When managers have a higher ownership position in the bank, their incentives are more closely aligned with shareholders so they have an incentive to take risk to increase the value of the call option associated with equity ownership. This suggests that a higher ownership position by

insiders (officers and directors) results in less hedging. The natural logarithm of the percentage of the total shares outstanding that are owned by officers and directors is used as the independent variable.

Ownership by Institutional Blockholders Institutional blockholders have incentive to monitor the firm's management due to the large ownership stake they have in the firm (Shleifer and Vishny [15]). Whidbee and Wohar [18] argue that these investors will have imperfect information and will most likely be concerned about the bottom line performance of the firm. The natural logarithm of the percentage of the total shares outstanding that are owned by all institutional investors is included as an independent variable and predict that the sign will be positive, with respect to the likelihood of hedging.

CEO Compensation CEO compensation also provides its own incentives with respect to risk management. In particular, compensation with more option-like features induces management to take on more risk to increase the value of the option (Smith and Blundell [16]; Tufano [17]). Thus, higher options compensation for managers results in less hedging. Two measures of CEO compensation are used: 1) annual cash bonus and 2) value of option awards.

There is a possibility that CEO compensation is endogenous in that successful hedging activity could in turn lead to higher executive compensation. The instruments used for the compensation variables are based on the executive's human capital (age and experience), and the size and scope of the firm (number of employees, number of offices and subsidiaries). These are expected to be correlated with the CEOs compensation and be predetermined with respect to the BHCs foreign exchange hedging activities.

BHC Size The natural logarithm of total assets is used to control for the size of the BHC.

Capital The ratio of equity capital to total assets is included as a control variable. The variable for dividends paid measures the amount of earnings that are paid out to shareholders. The higher the variable, the lower the capital position of the BHC. The dividends paid variable is expected to have a sign opposite that of the leverage ratio.

Like the compensation variables, leverage should be endogenously determined. Firms that hedge can take on more debt and thus have higher leverage, other things equal.

Foreign Exchange Risk A bank's use of currency derivatives should be related to its exposure to foreign exchange rate fluctuations. The ratio of interest income from foreign sources to total interest income measures foreign exchange exposure. Greater exposure, as represented by a larger proportion of income being derived from foreign sources, should be positively related to both the likelihood and extent of currency derivative use.

Profitability The return on equity is included to represent the profitability of the BHCs. It is used as a control.

5.1 Results

In this section the results of estimation are reported. Table 1 contains some important results from the reduced form equations. Due to the endogeneity of leverage and the CEO compensation variables, instrumental variables estimation is used to estimate the probability equations. Table 2 reports the coefficient estimates for the instrumental variable estimation of the probability that a BHC will use foreign exchange derivatives for hedging. The first column of results correspond to the Stata two-step estimator and the second column, gretl.

In Table 1 summary results from the reduced form are presented. The columns contain p-values associated with the null hypothesis that the indicated instrument's coefficient is zero in each of the four reduced form equations. The instruments include number of employees, number of subsidiaries, number of offices, CEO's age–which proxies for his or her experience, the 12 month maturity mismatch, and the ratio of cash flows to total assets (CFA). The p-values associated with the other variables have been suppressed to conserve space.

Each of the instruments appears to be relevant in that each is significantly different from zero at the 10% (p-value < 0.1) in at least one equation; the number of employees, number of subsidiaries, and CEO age and CFA are significant in one equation; the number of offices, employees, subsidiaries are significant in two equations.

The overall strength of the instruments can be roughly gauged by looking at the overall fit of the equations. The R^2 in the leverage equation is the smallest (0.29), but is still high relative to the results of the Monte Carlo simulation. The instruments, other than the 12 month maturity mismatch, appear to be strong and we have no reason to expect poor performance from either the AGLS or the mle in terms of bias.

The simulations from Adkins [1] suggest discarding extra instruments, and this would be recommended here. Which to drop, other than the mismatch variable is unclear. CFA, Age, and subsidiaries are all strongly correlated with lever-

age; office and employees with options; and, employees, subsidiaries, and offices with bonuses. The fit in the leverage equation is weakest, yet the p-values for each individual variable is relatively high. For illustrative purposes, I'll plow forward with the current specification.

Table 1. Summary Results from Reduced-form Equations. The table contains p-values for the instruments and R^2 for each reduced form regression. The data are taken from the Federal Reserve System's Consolidated Financial Statements for Bank Holding Companies (FR Y-9C), the *SNL Executive Compensation Review*, and the *SNL Quarterly Bank Digest*, compiled by SNL Securities.

	Reduced Form Equation		
	Leverage	Options	Bonus
Instruments	Coefficient P-values		
Number of Employees	0.182	0.000	0.000
Number of Subsidiaries	0.000	0.164	0.008
Number of Offices	0.248	0.000	0.000
CEO Age	0.026	0.764	0.572
12 Month Maturity Mismatch	0.353	0.280	0.575
CFA	0.000	0.826	0.368
R-Square	0.296	0.698	0.606

Table 2: **IV Probit Estimates of the Probability of Foreign-Exchange Derivatives Use By Large U.S. Bank Holding Companies (1996-2000)**. This table contains estimates for the probability of foreign-exchange derivative use by U.S. bank holding companies over the period of 1996-2000. To control for endogeneity with respect to compensation and leverage, we use an instrumental variable probit estimation procedure. The dependent variable in the probit estimations (i.e., probability of use) is coded as 1 if the bank reports the use of foreign-exchange derivatives for purposes other than trading. The data are taken from the Federal Reserve System's Consolidated Financial Statements for Bank Holding Companies (FR Y-9C), the *SNL Executive Compensation Review*, and the *SNL Quarterly Bank Digest*, compiled by SNL Securities. Approximate p-values based on the asymptotic distribution of the estimators are reported in parentheses beneath the parameter estimates.

	Instrumental Variables Probit	
	Stata (twostep)	gretl
Leverage	21.775	21.775
	(13.386)	(13.386)
Option Awards	-8.79E-08	-8.79E-08
	(5.31E-08)	(5.31E-08)
Bonus	1.76E-06	1.76E-06

Continued from preceding page

	Instrumental Variables Probit	
	Stata	gretl
	(8.88E-07)	(8.88E-07)
Total Assets	0.36453	0.36453
	(0.17011)	(0.17011)
Insider Ownership %	0.25882	0.25882
	(0.11623)	(0.11623)
Institutional Ownership %	0.36981	0.36981
	(0.13477)	(0.13477)
Return on Equity	-0.033852	-0.033852
	(0.028188)	(0.028188)
Market-to-Book ratio	-0.0018722	-0.0018722
	(0.0012422	(0.0012422)
Foreign to Total Interest Income Ratio	-3.5469	-3.546958
	(3.8414)	(3.8414)
Derivative Dealer Activity Dummy	-0.2799	-0.2799
	(0.24675)	(0.24675)
Dividends Paid	-8.43E-07	-8.43E-07
	(5.62E-07)	(5.62E-07)
D=1 if 1997	-0.024098	-0.024098
	(0.27259)	(0.27259)
D=1 if 1998	-0.24365	-0.24365
	(0.26195)	(0.26195)
D=1 if 1999	-0.24156	-0.24156
	(0.28171)	(0.28171)
D=1 if 2000	-0.128	-0.127999
	(0.27656)	(0.27656)
Constant	-9.673	-9.673
	(2.5351)	(2.5351)
Sample size	794	794

The model is overidentified, the sample is large (700+), and the instruments are very strong. Compared to maximum likelihood (ML) estimation, a few differences were found (see Adkins [2]). Leverage is significant in ML at the 10% level, but not with AGLS. Similarly, return-on-equity, market-to-book, and dividends paid are all significant in the ML regression but not AGLS. This divergence of results is a little troubling. In terms of the small sample properties documented by Adkins [1], ML p-values tend to be too small when instruments were mildly strong and correlation low. If the endogeneity problem is not severe, then the ML estimation and AGLS results tend to diverge. In this case, then AGLS estimator appears to be more reliable for testing significance. In the case of very strong instruments, the AGLS estimator tended to be insignificant too often. In the banking example, the empirical model falls between these two extremes and a strong recommendation can not be made for one over the other.

However, for the purposes of this paper, the news is excellent: the Stata results (column 1) and those from the simple gretl script (column 2) are basically identical. In situations where the AGLS is called for, one can confidently use the gretl script provided below to estimate the parameters of probit model that contains continuous endogenous regressors.

6 gretl Script

The following script was used with gretl 1.7.8 to produce the results in column 2 of Table 2.

```
# Variable definitions
# y2 = r.h.s. endogenous variables
# x  = the complete set of instruments
# x1 = r.h.s. exogenous variables
# y1 = dichotomous l.h.s. variable

list y2 = eqrat bonus optval
list x = const ltass linsown linstown roe mktbk perfor \
     dealdum div dum97 dum98 dum99 dum00 no_emp no_subs \
     no_off ceo_age gap cfa
list x1 = const ltass linsown linstown roe mktbk perfor \
     dealdum div dum97 dum98 dum99 dum00
list y1 = d2

matrix X = { x }
matrix Y = { y2 }
matrix Y1 = { y1 }
matrix X1 = { x1 }
matrix Z = X1~Y

matrix b = invpd(X'*X)*X'*Y
matrix d = invpd(X'X)*X'Z

scalar kx = cols(X)
scalar ky = cols(Y)
scalar s = cols(Y)

loop foreach i y2
    ols $i x --quiet
    genr uhat$i = $uhat
    genr yhat$i = $yhat
endloop

matrix d = invpd(X'X)*X'Z

# step 2  RF probit
```

```
probit y1 x uhat* --quiet
genr J = $vcv
matrix alph = $coeff
matrix alpha = alph[1:kx]
matrix lam = alph[kx+1:kx+ky]
matrix Jinv=J[1:kx,1:kx]

# Step 3 2siv
probit y1 x1 uhat* yhat* --quiet
matrix beta = $coeff
matrix beta = beta[rows(beta)-ky+1:rows(beta)]
matrix rho = lam - beta

# step 4 v2*inv(x'x)
matrix rhoY=Y*rho
series ry =  rhoY
ols ry x --quiet
matrix v2 = $vcv

matrix omega = (v2+Jinv)

# Step 5
matrix cov = invpd(d'*invpd(omega)*d)
matrix se = sqrt(diag(cov))
matrix delt = cov*d'*invpd(omega)*alpha
print delt se
```

This code could be used as the basis for a more elegant gretl function that could be used to estimate this model. Basically, one just needs to load the data and replace the variable names to be used in the list statements. This version of the code illustrates just how easy it is to perform matrix computations in gretl in that the code mimics the steps listed in section 3.2.

One of the very useful properties of gretl is the way in which matrix computations and native gretl results can be intermixed. In this case, the usual probit mle can be estimated using native gretl routines and the resulting variance covariance matrix can be saved, converted to a matrix and used in subsequent computations. The –quiet option reduces the amount of output to a manageable level.

7 Conclusion

In this paper a simple gretl script is used 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 10 software.

The next step is to duplicate the maximum likelihood estimator, a considerably more challenging undertaking given the multitude of ways the mle can be computed. It should be noted that the only other commericial software that estimates this model via mle is Limdep; Limdep and Stata use different algorithms and yield different results.

Another possibility is to use the plug-in IVP estimator with Murphy-Topel standard errors. In very preliminary research Adkins [3] finds that this estimator compares favorably to AGLS and ML estimation in approximating the nominal size of 10% tests of parameter significance. Like the AGLS estimator, this should also be a relatively simple computation in gretl.

Bibliography

[1] Adkins, Lee C. [2008*a*], Small sample performance of instrumental variables probit estimators: A monte carlo investigation.

[2] Adkins, Lee C. [2008*b*], 'Small sample performance of instrumental variables probit estimators: A monte carlo investigation', Department of Economics, Oklahoma State University, Stillwater OK 74078. available at http://www.learneconometrics.com/pdf/JSM2008.pdf.

[3] Adkins, Lee C. [2009], A comparison of two-step and ml estimators of the instrumental variables probit estimators: A monte carlo investigation.

[4] Adkins, Lee C., David A. Carter and W. Gary Simpson [2007], 'Managerial incentives and the use of foreign-exchange derivatives by banks', *Journal of Financial Research* **15**, 399–413.

[5] Arendt, Jacob Nielsen and Anders Holm [2006], Probit models with binary endogenous regressors, Discussion Papers on Business and Economics 4/2006, Department of Business and Economics Faculty of Social Sciences University of Southern Denmark.

[6] Blundell, Richard W. and James L. Powell [2004], 'Endogeneity in semiparametric binary response models', *Review of Economic Studies* **71**, 655–679. available at http://ideas.repec.org/a/bla/restud/v71y2004ip655-679.html.

[7] Davidson, Russell and James G. MacKinnon [2004], *Econometric Theory and Methods*, Oxford University Press, Inc., New York.

[8] Greene, William H. [2007], *LIMDEP Version 9.0 Econometric Modeling Guide, Volume 1*, Econometrics Software, Inc., 15 Gloria Place, Plainview, NY.

[9] Iwata, Shigeru [2001], 'Recentered and rescaled instrumental variable estimation of tobit and probit models with errors in variables', *Econometric Reviews* **24**(3), 319–335.

[10] Kan, Kamhon and Chihwa Kao [2005], Simulation-based two-step estimation with endogenous regressors, Center for Policy Research Working Papers 76, Center for Policy Research, Maxwell School, Syracuse University. available at http://ideas.repec.org/p/max/cprwps/76.html.

[11] Murphy, Kevin M. and Robert H. Topel [1985], 'Estimation and inference in two-step econometric models', *Journal of Business and Economic Statistics* **3**(4), 370–379.

[12] Newey, Whitney [1987], 'Efficient estimation of limited dependent variable models with endogenous explanatory variables', *Journal of Econometrics* **36**, 231–250.

[13] Nicoletti, Cheti and Franco Peracchi [2001], Two-step estimation of binary response models with sample selection, Technical report, Faculty of Economics, Tor Vergata University, I-00133 Rome, Italy. Please do not quote.

[14] Rivers, D. and Q. H. Vuong [1988], 'Limited information estimators and exogeneity tests for simultaneous probit models', *Journal of Econometrics* **39**(3), 347–366.

[15] Shleifer, A. and R. W. Vishny [1986], 'Large shareholders and corporate control', *Journal of Political Economy* **94**, 461–488.

[16] Smith, Richard J. and Richard W. Blundell [1985], 'An exogeneity test for a simultaneous equation tobit model with an application to labor supply', *Econometrica* **54**(3), 679–685.

[17] Tufano, P. [1996], 'Who manages risk? an empirical examination of risk management practices in the gold mining industry', *Journal of Finance* **51**, 1097–1137.

[18] Whidbee, D. A. and M. Wohar [1999], 'Derivative activities and managerial incentives in the banking industry', *Journal of Corporate Finance* **5**, 251–276.

[19] Yatchew, Adonis and Zvi Griliches [1985], 'Specification error in probit models', *The Review of Economics and Statistics* **67**(1), 134–139.