Alcamentos

Departamento de Estadística, Estructura y O.E.I.

0804

A NOTE ON COMPUTING MURPHY-TOPEL CORRECTED VARIANCES IN A HECKPROBIT MODEL WITH ENDOGENEITY IN STATA

Juan Muro, Cristina Suárez y María del Mar Zamora

Universidad de Alcalá y Alcamétrica

DEPARTAMENTO DE
ESTADÍSTICA, ESTRUCTURA ECONÓMICA Y O.E.I.
Plaza de la Victoria, 2
28802 Alcalá de Henares (Madrid)
Teléfono: 91 885 42 01
http://www.uah.es/centros_departamentos/departamentos

Universidad de Alcalá

Alcamentos

Nº: 0804

A NOTE ON COMPUTING MURPHY-TOPEL CORRECTED VARIANCES IN A HECKPROBIT MODEL WITH ENDOGENEITY IN STATA

Juan Muro, Cristina Suárez y María del Mar Zamora

Universidad de Alcalá y Alcamétrica

DEPARTAMENTO DE
ESTADÍSTICA, ESTRUCTURA ECONÓMICA Y O.E.I.
Plaza de la Victoria, 2
28802 Alcalá de Henares (Madrid)
Teléfonos: 91 885 42 01
http://www.uah.es/centros_departamentos/departamentos

A NOTE ON COMPUTING MURPHY-TOPEL CORRECTED VARIANCES IN A HECKPROBIT MODEL WITH ENDOGENEITY IN STATA[1]

Juan Muro
Cristina Suárez
María del Mar Zamora*

Universidad de Alcalá and Alcamétrica

July, 2009

Abstract

We outline a fairly simple method to obtain in Stata Murphy-Topel corrected variances for a two-step estimation of a class of heckprobit models with endogeneity in the main equation. The procedure utilizes the *score* option and the powerful matrix tool *accum* in Stata and builds on previous works by Hardin (2002) and Hole (2006).

Keywords: Binary choice model with selectivity and endogenous variables, Two-step estimation in qualitative models, Murphy-Topel corrected variances, Stata program.
JEL classification: C25, L83

*Corresponding author.
Facultad de Ciencias Económicas y Empresariales
Universidad de Alcalá
Plaza de la Victoria, 2
28802 Alcalá de Henares, Madrid (SPAIN)
e-mail: mariam.zamora@uah.es

[1] The authors thank A. R. Hole and the reviewers for helpful comments and suggestion that led to improvements in the manuscript.

A NOTE ON COMPUTING MURPHY-TOPEL CORRECTED VARIANCES IN A HECKPROBIT MODEL WITH ENDOGENEITY IN STATA

Juan Muro, Cristina Suárez and María del Mar Zamora

1. Introduction.

Probit models with selectivity, heckprobit models, have become an important tool in empirical analysis. Estimating the model in presence of endogenous variables is usually made by means of a two-step method, whereas it produces consistent estimates, though a full maximum likelihood method is not discarded. In this note we stress the relevance of obtaining a corrected variance estimator, Murphy-Topel (1985), Hardin (2002), when a two-step estimation method is chosen and show a fairly simple procedure to compute Murphy-Topel corrected variances in Stata. Our procedure builds on previous work by Hardin (2002) and Hole (2006) and generalizes them to the case of a model with two index functions.

We organize the paper as follows. In section 2 we describe our model and the Murphy-Topel estimator. Section 3 contains the Stata procedure for computing Murphy-Topel corrected variances and an illustration. Section 4 concludes.

2. A Murphy-Topel estimator for a heckprobit model with endogeneity in the equation of interest.

The model used is an extension of a well known result in the econometric literature first outlined by Lahiri and Schmidt (1978) and see also Greene (1997, 1998).

In terms of log-likelihood functions our model is

$L_1(\theta_1|X_1, Y_1)$, [1]

$L_2(\theta_2, \theta_3, \rho|X_2, X_3, Y_2, Y_3)$. [2]

Being L_1, L_2 the log-likelihood functions of model (1), the reduced form equation for the endogenous variable usually a probit (X_1 is k1*T), and model (2), the heckprobit model of interest, respectively. Subindexes 2 and 3 indicate variables matrices and parameters vectors in the main equation and the selection equation respectively of the heckprobit model. Matrix X_2 (k2*T) contains the predicted value of Y_1 variable in model (1).

As is well known a inefficient but consistent estimation of the model in (1) and (2), see conditions that guarantee consistency for example in Maddala(1983), Chapter 5, 122-147 and 246-247, a two-step corrected estimation, is

1. Estimate (1) by ML probit and obtain Y_i predictions, i.e., $\Phi(\theta_1' X_{1i})$.
2. Substitute predictions obtained in previous step in place of observed Y_i in (2) and estimate the heckprobit by ML.
3. Calculate appropriate corrected variance-covariance estimations; Murphy-Topel (1985), see also Greene (1997, pp.141-142), Hardin (2002) and Hole (2006).

We have to correct the estimated covariance matrix for the selectivity probit model in (2), sometimes named the naïve covariance matrix, due to its conditional nature. As is well known, Murphy-Topel (1985), the estimate of the variance for a two-step model is

$$\hat{V}_2 + \hat{V}_2\left(\hat{C}\hat{V}_1\hat{C}' - \hat{R}\hat{V}_1\hat{C}' - \hat{C}\hat{V}_1\hat{R}'\right)\hat{V}_2. \qquad [3]$$

Where V_i, i= 1, 2, stands for the covariance matrices of models in step 1 and step 2. In addition,

$$\hat{C} = \sum_{i=1}^{N}\left(\frac{\partial L_{i2}}{\partial \hat{\theta}}\right)\left(\frac{\partial L_{i2}}{\partial \hat{\theta}_1'}\right). \qquad [4]$$

$$\hat{R} = \sum_{i=1}^{N}\left(\frac{\partial L_{i2}}{\partial \hat{\theta}}\right)\left(\frac{\partial L_{i1}}{\partial \hat{\theta}_1'}\right). \qquad [5]$$

Where L_{ij} j=1, 2 stands for each observation i´s contribution to the likelihood function of the respective model; $\theta = [\theta_2\ \theta_3\ \rho]$.

A fairly easy way of calculating expressions in (4) and (5) for models with a simple index using Stata is described in detail in Hole (2006). We extend in the next section the method for heckprobit models with endogeneity.

3. A Stata program to calculate Murphy-Topel corrected variances.

A program to calculate (3) in Stata is described as follows:

```
/*Fist stage: probit, save score as s0 */
probit Y1 X1, score(s0)                    /* X1 contains k1 variables (included in k1 the constant)*/
```

```
matrix V1=e(V)                  /*Variance estimate, matrix dimension (k1,k1)*/
predict double Y1hat            /*Generate prediction of endogenous variable for second stage*/
```

As a result of the above Stata sentences we get covariance matrix of model (1), V1, and the predicted values of the endogenous variable, Y1hat, included in matrix X_2 of model (2).

```
/*Second stage: heckprobit, save scores as s1, s2 and s3 */
heckprob Y2 X2 Y1hat , sel(Y3= X3) score(s1 s2 s3)
/* X2 and X3 contain (k2-1) and k3 variables, respectively (included in k2-1 and k3 the constant) */
matrix V2=e(V)                  /*matrix dimension (k2+k3+1, k2+k3+1)*/
scalar TP=_b[Y1hat]             /*Coef. of endogenous variable in main equation*/
matrix coef=e(b)                /*vector dimension: k2+k3+1*/
```

In the second stage we obtain heckprobit ML estimates and the naïve covariance matrix. Table 1 shows two step heckprobit ML estimation results where standard errors, z-statistics, probabilities and confidence intervals derive from the naïve covariance matrix (the data and model come from Muro, Suárez and Zamora (2006, 2008))

[Insert Table 1]

Given the initial estimates, we calculate in turn $\hat{C}$ and $\hat{R}$ matrices. For the sake of clarity we remind that in a heckprobit model we have censored and uncensored observations. Only uncensored observations, those who satisfy Y_3=1, enter in the main equation. So, we can split summations in (4) and (5) in two parts: uncensored and censored. S1 and S3 scores computed in Stata are vectors with null values for censored observations whilst S2 has no null values in the whole sample.

Partial derivatives of log-likelihood L_2 with respect to the parameter vector of model (2) have two components: the first one is the derivative with respect to the index; the second one the

derivative of the index with respect to the parameter. The first component is the scores vector calculated in Stata's *heckprob* option: S1 for θ_2 , S2 for θ_3 , and S3 for ρ. The second component is a matrix with X_2, X_3 and a vector of ones.

Partial derivatives of log-likelihood L_2 with respect to the parameter vector of model in (1) have similarly two components. The first component is the S1 score vector, which has null values for censored observations. The second component is matrix X1 times the estimated parameter of Y1hat in model in (2) times the derivative of Y1hat with respect to the index function of model in (1). The formula is

$$\hat{C} = \breve{X}' diag\left[S2 * S1 \frac{\partial Y1hat}{\partial (X_1 \theta_1)} \hat{\gamma} \right] X_1.$$

Where X (with the strange hat) has as components X_2 times S1/S2, X_3, and S3/S2; the derivatives in our probit model are N (0, 1) pdf; γ hat is the estimated parameter of Y1hat in model in (2).

For matrix R in (5) a similar reasoning leads us to the formula

$$\hat{R} = \breve{X}' diag\left[S2 * S0 \right] X_1.$$

With the equivalences noted above.

The Stata program continues as follows

```
gen const = 1                  /* needed for the program */
gen X2_s = X2 * s1 / s2        /* s1 is the true score */
gen Y1hat_s= Y1hat*s1/s2
```

```
gen a_s = s1/ s2
gen s3_s = s3 / s2              /* auxiliary parameter */
/*For main  and selection equations*/
matrix accum C = X1 const X2_s Y1hat_s a_s X3 const s3_s          /*
            */          [iw=s2*s1*(s0*((1-Y1) + (2*Y1-1)*Y1hat)* (2*Y1-1))²*TP], nocons
/*For main and selection equations*/
matrix accum R = X1 const X2_s Y1hat_s a_s X3 const s3_s          /*
            */          [iw=s2*s0], nocons
/* Get only the desired partition */
matrix C= C [k1+1..k1+k2+k3+1,1..k1]                    /* see Hole (2006)
matrix R= R [k1+1..k1+k2+k3+1,1..k1]

matrix M =V2 + (V2 * (C*V1*C' -R*V1*C' -C*V1*R') *V2)

capture program drop doit
matrix b=e(b)
program define doit, eclass
            ereturn post b M
            ereturn local vcetype "Mtopel"
            ereturn display
end
doit
```

Derivation of similar Stata sentences for the case in which model (2) is a regression equation, with continuous dependent variable, is straightforward.

Table 2 shows two step heckprobit ML estimation results where standard errors, z-statistics, probabilities and confidence intervals derive from the Murphy-Topel corrected covariance matrix.

[Insert Table 2]

4. Conclusions.

In this paper we show how to compute in a fairly simple way Murphy-Topel corrected variances in Stata in the context of a selectivity probit model with endogeneity. We use the *score*

[2] Term in brackets is equivalent to normalden(xb).

option and the powerful matrix tool *accum* in Stata to make a program that compute the corrected covariance matrix and allows a quick change of alternative specifications.

Our illustration shows the importance of constructing Murphy-Topel covariance matrices in order to properly asses the significance of covariates in presence of endogenous variables in econometric models with dependent qualitative variables. In particular, in our case endogenous variable significance is altered when Murphy-Topel variances are considered.

References.

Greene, W. (1997), *Econometric Analysis*. 3rd ed. Prentice Hall. Englewood Cliffs, NJ.

Greene, W. (1998), "Gender Economics Courses in Liberal Arts Colleges: Further Results*",* *Journal of Economic Education,* 29, pp. 291-300.

Hardin, J.W. (2002), "The robust variance estimator for two stage models. *The Stata Journal*, 2, pp. 253-266.

Hole, A.R. (2006), "Calculating Murphy-Topel Variance Estimates in Stata: A Simplified Procedure", *The Stata Journal*, 6, pp. 521-529.

Lahiri, K. and P. Schmidt (1978), "On the estimation of Triangular Structural Systems", Econometrica, 46, pp. 1217-1221.

Maddala, G.S. (1983), *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge University Press. New York.

Muro, J., C. Suárez and M.M. Zamora (2006), "The Demand for Low-cost Carriers: An Empirical Micro Analysis". Unpublished paper.

Muro, J., C. Suárez and M.M. Zamora (2008), "The Impact of E-Commerce on the Tourist Purchase Decision: An Empirical Analysis". Alcamentos 0801.

Murphy, K.M. and R.H. Topel (1985), "Estimation and Inference in Two-step Econometric Models", *Journal of Business and Economic Statistics*, 3, pp. 88-97.

Table 1. Two step Heckprobit estimation results (uncorrected covariance matrix).

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y2						
age24	-.0458653	.033738	-1.36	0.174	-.1119904	.0202599
age25_44	-.1537932	.0282394	-5.45	0.000	-.2091414	-.098445
age45_64	-.0720658	.0262045	-2.75	0.006	-.1234256	-.020706
country1	-.6337746	.0913743	-6.94	0.000	-.812865	-.4546842
country2	.1011763	.0213756	4.73	0.000	.0592809	.1430717
country3	.3173458	.0176745	17.96	0.000	.2827044	.3519871
country4	-.2298831	.0296122	-7.76	0.000	-.287922	-.1718443
AACC2	.6911637	.0330778	20.90	0.000	.6263325	.755995
AACC3	.9601613	.2879347	3.33	0.001	.3958197	1.524503
AACC4	.8319111	.4004437	2.08	0.038	.0470559	1.616766
AACC5	.5350787	.0405529	13.19	0.000	.4555965	.6145609
AACC6	.6227361	.0574185	10.85	0.000	.5101979	.7352743
Y1hat	-1.523668	.5041928	-3.02	0.003	-2.511868	-.5354686
_cons	-.5805468	.0394076	-14.73	0.000	-.6577843	-.5033093
Y3						
age24	.9653056	.0344481	28.02	0.000	.8977886	1.032823
age25_44	.912114	.0241071	37.84	0.000	.8648649	.9593631
age45_64	.4015543	.0243098	16.52	0.000	.3539079	.4492006
border	-1.654292	.0162945	-101.52	0.000	-1.686229	-1.622356
borderAACC	-.8970486	.0150943	-59.43	0.000	-.9266328	-.8674644
_cons	1.139033	.0220057	51.76	0.000	1.095902	1.182163
/athrho	-.6547091	.0639639	-10.24	0.000	-.780076	-.5293422
rho	-.5748316	.0428282			-.6527504	-.4848781

LR test of indep. eqns. (rho = 0): chi2(1) = 97.60 Prob > chi2 = 0.0000

Table 2. Two step Heckprobit estimation results (Murphy-Topel corrected covariance matrix).

	Mtopel				
	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Y2						
age24	-.0458653	.0337474	-1.36	0.174	-.1120089	.0202784
age25_44	-.1537932	.0282431	-5.45	0.000	-.2091486	-.0984378
age45_64	-.0720658	.026204	-2.75	0.006	-.1234248	-.0207068
country1	-.6337746	.0916261	-6.92	0.000	-.8133585	-.4541907
country2	.1011763	.0223539	4.53	0.000	.0573636	.1449891
country3	.3173458	.0179434	17.69	0.000	.2821774	.3525141
country4	-.2298831	.0299708	-7.67	0.000	-.2886249	-.1711414
AACC2	.6911637	.0338813	20.40	0.000	.6247577	.7575698
AACC3	.9601613	.2992126	3.21	0.001	.3737153	1.546607
AACC4	.8319111	.4155639	2.00	0.045	.0174209	1.646401
AACC5	.5350787	.0414957	12.89	0.000	.4537486	.6164088
AACC6	.6227361	.0590565	10.54	0.000	.5069876	.7384847
Y1hat	-1.523668	.5229309	-2.91	0.004	-2.548594	-.4987426
_cons	-.5805468	.0395836	-14.67	0.000	-.6581292	-.5029643
Y3						
age24	.9653056	.0344495	28.02	0.000	.8977859	1.032825
age25_44	.912114	.0241066	37.84	0.000	.8648659	.959362
age45_64	.4015543	.0243096	16.52	0.000	.3539084	.4492001
border	-1.654292	.0163054	-101.46	0.000	-1.68625	-1.622334
borderAACC	-.8970486	.015094	-59.43	0.000	-.9266324	-.8674649
_cons	1.139033	.022006	51.76	0.000	1.095902	1.182164
athrho						
_cons	-.6547091	.064002	-10.23	0.000	-.7801507	-.5292675

Appendix

Asymmetries of information and risk aversion play a relevant role in tourism demand. Insufficiently informed and high risk aversion tourists are very prone to organize their travel through a tourist package supplied by a tourist agency. Yet informed and low risk aversion tourists tend to organize their travel by their own using more and more the new communication and information technologies. This translates to tourism economic models a problem that has rarely been addressed, the endogeneity of the travel organization mode variables. In this note we illustrate our procedure for calculating Murphy-Topel corrected variances in a Heckprobit model with endogeneity by means of the estimation of a model of low cost carrier (LCC) demand in which endogeneity of the travel organization mode variable is controlled for in a fairly simple way.

Our model is

$$Y_{1i}^{*} = \theta_1' X_{1i} + \upsilon_i \,, \qquad [a.1]$$
$$Y_{2i}^{*} = \theta_2' X_{2i} + u_i \,, \qquad [a.2]$$
$$Y_{3i}^{*} = \theta_3' X_{3i} + v_i, \qquad [a.3]$$

Where Y_{2i}=1 stands for a tourist who travels in a LCC versus full service carrier; Y_{3i} =1 for a tourist who travels by air carrier versus road; Y_{1i} =1 for a tourist who travels with a tourist package, and matrix X_2 contains the predicted value of Y_{1i}. Also, $Y_{3i} = 1[\theta_3' X_{3i} + v_i, > 0]$; $Y_{2i} = 1[\theta_2' X_{2i} + u_i > 0]$ Y_{3i} ,$\forall$ Y_{3i} = 1 ; Y_{2i} is unobserved ,$\forall$ Y_{3i} = 0; $Y_{1i} = 1[\theta_1' X_{1i} + \upsilon_i > 0]$. Under joint normality υ_i, u_i, v_i, are distributed as a trivariate normal variable (TVN).

Variables in our model are:

Y1: Dichotomous. 1= Tourist visits Spain with a package tour.
Y2: Dichotomous. 1= Tourist travels in a low cost carrier (versus full service airline).
Y3: Dichotomous. 1= Tourist travels by air (versus road).
Age: Four discrete categories. Under 24 (*age_24*), between 24 and 44(*age25_44*), between 45 and 64 (*age45_64*) and over 64.
Country of residence: It captures five different origins: France (*country1*), Germany (*country2*), United Kingdom (*country3*), Italy (*country4*), and the rest of the world. We also use this information to distinguish whether a country of residence has border with Spain (*border*).
Tourist main destination: Six dummy variables capture the main tourism destinations in Spain: Andalusia (*AACC2*), Canary Islands (*AACC3*), Balearic Island (*AACC4*), Catalonia (*AACC5*), Community of Valencia (*AACC6*), and other destinations. We also use this information to distinguish whether a Community has border (*borderAACC*).

Our data come from the 2004 wave of EGATUR the Spanish Foreign Tourism Expenditure Survey. It is an annual survey on non-resident visitors coming to Spain expenditures. The survey is conducted on a monthly basis in the frontiers. The EGATUR sample provides a very rich data set on tourists' behaviour, socioeconomic categories, attributes of the trip and other relevant variables.

Relación de títulos publicados en la colección ALCAMENTOS.

Nº	Autor/es	Título
0801	Juan Muro Cristina Suárez Maria del Mar Zamora	The impact of e-commerce on the tourist purchase decision: An empirical micro analysis
0802	Jhon James Mora Juan Muro	Diploma earning differences by gender in Colombia
0803	José Mª Arranz Ana I. Gil	Alcoholic beverages as determinants of traffic fatalities
0804	Juan Muro Cristina Suárez Maria del Mar Zamora	A note on computing Murphy-Topel corrected variances in a heckprobit model with endogeneity in Stata