The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the web page

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Social Sciences Citation Index®

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, Mata, NetCourse, and Stata Press are registered trademarks of StataCorp LP.

The Stata Journal (2010)
10, Number 2, pp. 252–258

Computing Murphy–Topel-corrected variances in a heckprobit model with endogeneity

Juan Muro
Department of Statistics, Economic Structure,
and International Economic Organization
Universidad de Alcalá
Madrid, Spain

Cristina Suárez
Department of Statistics, Economic Structure,
and International Economic Organization
Universidad de Alcalá
Madrid, Spain

María del Mar Zamora
Department of Statistics, Economic Structure,
and International Economic Organization
Universidad de Alcalá
Madrid, Spain
mariam.zamora@uah.es

Abstract. We outline a fairly simple method to obtain in Stata Murphy–Topel-corrected variances for a two-step estimation of a heckprobit model with endogeneity in the main equation. The procedure uses `predict`'s `score` option and the powerful matrix tool `accum` in Stata and builds on previous works by Hardin (2002, *Stata Journal* 2: 253–266) and Hole (2006, *Stata Journal* 6: 521–529).

Keywords: st0191, binary choice model, heckprobit, selectivity, endogenous variables, two-step estimation, qualitative models, Murphy–Topel-corrected variances

1 Introduction

Probit models with selectivity, referred to as heckprobit models, are an important tool in empirical analysis. Estimating a heckprobit model in the presence of endogenous variables is usually achieved using a two-step method, though a full maximum likelihood (ML) method is also possible. In this article, we stress the relevance of obtaining a variance estimator (Murphy and Topel 2002; Hardin 2002) when a two-step estimation method is chosen, and we 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 illustrates the application of their approaches to a model with two index functions.

We organize the article as follows. In section 2, we describe our model. Section 3 contains the Stata procedure for computing Murphy–Topel-corrected variances and an illustration. Section 4 provides a brief summary.

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

The model considered is described by an extension of a well known result in the econometric literature first outlined by Lahiri and Schmidt (1978) and also discussed by Greene (1998, 2008).

As is well known, inefficient but consistent estimators of the parameters in the component models are given by the two-step procedure:

1. Estimate the reduced-form model for the endogenous variable by ML probit and obtain its predictions.
2. Substitute the predictions obtained in step 1 for the appropriate covariate column and estimate the heckprobit by ML.
3. Calculate appropriate corrected variance–covariance estimations; see Murphy and Topel (2002) and Greene (2008, 302–303).

We have to correct the estimated covariance matrix for the selectivity probit model in the second model. The unadjusted variance matrix is sometimes called the naïve covariance matrix because it assumes that the predicted values used as a covariate are measured without error.

A straightforward way of calculating the components of the Murphy–Topel variance expression for models with a simple index using Stata is described in detail in Hole (2006). We extend this approach in the next section for heckprobit models with endogeneity.

(*Continued on next page*)

3 Murphy–Topel-corrected variances

A sequence of commands to calculate the Murphy–Topel variance using Stata is described as follows:

```
use data
local x1 "country1-country4 aacc2-aacc6"

/* First stage: probit, save score as s0 */
probit y1 `x1´          /* `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 */
predict s0, score
```

As a result of the above Stata commands, we save the estimated covariance matrix of the probit equation, `V1`, and the predicted values of the endogenous variable to be included in the matrix of covariates of the second-stage model.

```
/*Second stage: heckprobit, save scores as s1, s2, and s3 */
local x2 "age24 age25_44 age45_64 country1-country4 aacc2-aacc6 y1hat"
local x3 "age24 age25_44 age45_64 border borderaacc"
heckprob y2 `x2´, select(y3 = `x3´)

/* `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 */
predict s1 s2 s3, score
```

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, 2009]).

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	-.1119905	.0202599
age25_44	-.1537932	.0282394	-5.45	0.000	-.2091414	-.098445
age45_64	-.0720658	.0262045	-2.75	0.006	-.1234256	-.020706
country1	-.6337745	.0913744	-6.94	0.000	-.812865	-.4546841
country2	.1011763	.0213756	4.73	0.000	.0592809	.1430717
country3	.3173457	.0176745	17.96	0.000	.2827044	.3519871
country4	-.2298831	.0296122	-7.76	0.000	-.287922	-.1718443
aacc2	.6911638	.0330778	20.90	0.000	.6263325	.755995
aacc3	.9601613	.2879347	3.33	0.001	.3958196	1.524503
aacc4	.8319112	.4004438	2.08	0.038	.0470557	1.616767
aacc5	.5350787	.0405529	13.19	0.000	.4555965	.6145609
aacc6	.6227361	.0574185	10.85	0.000	.5101979	.7352743
y1hat	-1.523668	.5041929	-3.02	0.003	-2.511868	-.5354684
_cons	-.5805467	.0394076	-14.73	0.000	-.6577843	-.5033092
y3						
age24	.9653056	.0344481	28.02	0.000	.8977886	1.032823
age25_44	.912114	.0241071	37.84	0.000	.8648649	.9593631
age45_64	.4015542	.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	-.6547092	.0639639	-10.24	0.000	-.7800761	-.5293423
rho	-.5748316	.0428282			-.6527504	-.4848782

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

Given the initial estimates, we calculate the $\widehat{\mathbf{C}}$ and $\widehat{\mathbf{R}}$ matrices described in Hardin (2002) and Hole (2006). For the sake of clarity, we remind readers that in a heckprobit model, we have censored and uncensored observations. Only uncensored observations enter into the main equation. Thus we can split summations into two parts: uncensored and censored. `s1` and `s3` scores computed in Stata are vectors with null values for censored observations, while `s2` has no null values in the whole sample.

Partial derivatives of the log likelihood of the second stage with respect to the parameter vector in the second stage have two components: the first one is the derivative with respect to the index, and the second one is the derivative of the index with respect to the parameter. The first component is the score vector calculated in Stata's `heckprob` command: `s1` for the parameters of the main equation, `s2` for the parameters of the selection equation, and `s3` for the correlation term ρ. The second component is a matrix with `‘x2’`, `‘x3’`, and a vector of 1s.

Partial derivatives of the log likelihood of the second stage with respect to the parameter vector in the first stage also have 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 the heckprobit model times the derivative of `y1hat` with respect to the index function of the probit model. The formula is

$$\widehat{\mathbf{C}} = \breve{X}'\mathrm{diag}\left(\texttt{s2}\times\texttt{s1}\frac{\partial\texttt{y1hat}}{\partial\,\text{`}\texttt{x1}\text{'}\theta_1}\widehat{\gamma}\right)\text{`}\texttt{x1}\text{'}$$

where $\breve{X}$ has as components `‘x2’` times `s1`/`s2`, `‘x3’`, and `s3`/`s2`; the derivatives in our probit model are $N(0,1)$ probability density function; and $\widehat{\gamma}$ is the estimated parameter of `y1hat` in the heckprobit model.

For matrix $\widehat{\mathbf{R}}$, a similar reasoning leads us to the formula

$$\widehat{\mathbf{R}} = \breve{X}'\mathrm{diag}\left(\texttt{s2}\times\texttt{s0}\right)\text{`}\texttt{x1}\text{'}$$

with the equivalences noted above.

The Stata program continues as follows:

```
generate const = 1      /* Needed for the program */
local x2 "age24 age25_44 age45_64 country1 country2 country3 country4 aacc2 /*
        */      aacc3 aacc4 aacc5 aacc6 y1hat"
foreach a1 of local x2 {
        generate `a1´_s = `a1´ * s1/s2
}

/* s2 is the true score */
generate a_s = s1/s2
generate s3_s = s3/s2   /* Auxiliary parameter */
local x2_s "age24_s age25_44_s age45_64_s country1_s country2_s country3_s  /*
        */      country4_s aacc2_s aacc3_s aacc4_s aacc5_s aacc6_s y1hat_s"

/* For main and selection equations */
matrix accum C = `x1´ const `x2_s´  a_s  `x3´ const s3_s        /*
        */      [iw=s1*s2*(s0*((1-y1)+(2*y1-1)*y1hat)*(2*y1-1))*TP], noconstant

/*For main and selection equations*/
matrix accum R = `x1´ const `x2_s´  a_s  `x3´ const s3_s        /*
        */      [iw=s2*s0], noconstant

/* Get only the desired partition; see Hole (2006) */
matrix C = C[11..31,1..10]
matrix R = R[11..31,1..10]

/* For Murphy-Topel matrix */
matrix M = V2 + (V2 * (C*V1*C´ -R*V1*C´ -C*V1*R´) * V2)

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

In the first `matrix accum` command, the term in brackets is equivalent to `normalden(xb)`.

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.

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

	Coef.	Mtopel 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	-.6337745	.0916261	-6.92	0.000	-.8133584	-.4541906
country2	.1011763	.0223539	4.53	0.000	.0573636	.1449891
country3	.3173457	.0179434	17.69	0.000	.2821774	.3525141
country4	-.2298831	.0299708	-7.67	0.000	-.2886249	-.1711414
aacc2	.6911638	.0338813	20.40	0.000	.6247577	.7575698
aacc3	.9601613	.2992127	3.21	0.001	.3737152	1.546607
aacc4	.8319112	.415564	2.00	0.045	.0174207	1.646402
aacc5	.5350787	.0414957	12.89	0.000	.4537486	.6164088
aacc6	.6227361	.0590565	10.54	0.000	.5069875	.7384847
y1hat	-1.523668	.522931	-2.91	0.004	-2.548594	-.4987424
_cons	-.5805467	.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	.4015542	.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	-.6547092	.064002	-10.23	0.000	-.7801508	-.5292676

4 Summary

In this article, we demonstrate how the Murphy–Topel variance estimator for a heckprobit second-stage model can be estimated following the sequence of commands given by Hardin (2002) and Hole (2006).

5 Acknowledgments

The authors would like to thank A. R. Hole and the reviewers for helpful comments and suggestions that led to improvements in this article.

6 References

Greene, W. H. 1998. Gender economic courses in liberal arts colleges: Further results. *Journal of Economic Education* 29: 291–300.

———. 2008. *Econometric Analysis*. 6th ed. Upper Saddle River, NJ: Prentice Hall.

Hardin, J. W. 2002. The robust variance estimator for two-stage models. *Stata Journal* 2: 253–266.

Hole, A. R. 2006. Calculating Murphy–Topel variance estimates in Stata: A simplified procedure. *Stata Journal* 6: 521–529.

Lahiri, K., and P. Schmidt. 1978. On the estimation of triangular structural systems. *Econometrica* 46: 1217–1221.

Muro, J., C. Suárez, and M. M. Zamora. 2006. The demand for low-cost carriers: An empirical micro analysis. Unpublished manuscript.

———. 2009. Access and use of e-commerce in the Spanish tourism market. In *Advances in Tourism Destination Marketing*, ed. M. Kozak, J. Gnoth, and L. L. A. Andreu, 170–182. London: Routledge.

Murphy, K. M., and R. H. Topel. 2002. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics* 20: 88–97.

About the authors

Juan Muro is a professor of economics at the Universidad de Alcalá, Spain. His research interests include microeconometrics, labor economics, duration models, treatment effects, and efficiency measures and production frontiers.

Cristina Suárez is a lecturer of econometrics in the Department of Statistics, Economic Structure, and International Economic Organization at the Universidad de Alcalá, Spain. Her main research interests are microeconometrics, applied industrial economics, and service markets.

María del Mar Zamora is a lecturer of econometrics in the Department of Statistics, Economic Structure, and International Economic Organization at the Universidad de Alcalá, Spain. Her current research projects focus on applied microeconometric models and tourism behavior.