Econometrics Journal (2003), volume **6**, pp. 401–407.

Standard error correction in two-stage estimation with nested samples

PINAR KARACA-MANDIC AND KENNETH TRAIN†

Department of Economics, University of California, Berkeley
E-mail: train@econ.berkeley.edu

Received: July 2003

Summary Data at different levels of aggregation are often used in two-stage estimation, with estimates obtained at the higher level of aggregation entering the estimation at the lower level of aggregation. An example is customers within markets: first-stage estimates on market data provide variables that enter the second-stage model on customers. We derive the asymptotic covariance matrix of the second-stage estimates for situations such as these. We implement the formulae in the Petrin–Train application of households' choice of TV reception and compare the calculated standard errors with those obtained without correction. In this application, ignoring the sampling variance in the first-stage estimates would be seriously misleading.

Keywords: *Standard errors, Mixed logit, Two-stage estimation.*

1. INTRODUCTION

In applications where one model is embedded within another, it is often convenient to estimate the model in two stages. One result of such estimation is that the second-stage model contains variables constructed from parameters estimated in the first stage. Thus, the covariance matrix of the second-stage estimator includes noise induced by the first-stage estimates. Amemiya (1978) derives the asymptotic covariance for two-stage estimation of multinomial logit models when both stages use the same observations. Heckman (1979) determines the correct asymptotic covariance matrix of the two step estimator that accounts for sample selection bias. In his application, the second-stage observation is a subset of the first-stage observation. Murphy and Topel (1985) give the covariance for two-stage estimation by maximum likelihood and least squares when, again, both stages use the same observations. Greene (2000) shows how Murphy and Topel's results are applied under various specific models, and Newey and McFadden (1994) and McFadden (1999) extend them to the generalized method of moments.

This paper derives the asymptotic covariance of the second-stage estimator when the two stages use different numbers of observations and the data for the second stage are nested in the observations for the first stage. For example, the second stage might model customers' demand in different markets (N_2 customers each of whom buys in one of N_1 markets). If the price in each market is correlated with the unobserved portion of customers' demand (due to, for example,

†Corresponding author.

an omitted product attribute), then some form of instrumental variables estimation is needed to account for this correlation. The first stage specifies price as a function of instruments using market-level data, and the second stage utilizes the predicted price or the residual in estimation of the customer-level model.

Kuksov and Villas-Boas (2001) derive a similar formula for a related set-up. Their samples are not nested; instead, each unit of observation in the second stage appears repeatedly in all of the first-stage observations (i.e. each customer is observed in each of several time periods.) They allow correlation over first-stage observations in the unobservables associated with each second-stage unit (i.e. correlation for the same customer over time). In contrast, our formula uses independence over first-stage observations.

We apply our formula to the Petrin–Train application of households' choice of TV reception. They are concerned that the prices of the TV reception alternatives are correlated with the unobserved attributes of the alternatives. To correct for this endogeneity, they specify a first-stage linear regression of prices on exogenous variables and instrumental variables using market level data. The second stage is a mixed logit model. The dependent variable is the TV reception choice of a customer in a given market, and the independent variables are market level prices, market level residuals estimated in the first stage and other observed attributes of the alternatives.

2. FRAMEWORK

First stage. Estimated parameters $\hat{\theta}_1$ solve moment conditions $\bar{h}(\hat{\theta}_1) = 0$ where $\bar{h}(\theta_1) = \frac{1}{N_1}\sum_m h_m(\theta_1, z_m)$, N_1 is the number of observations, and $h_m(\theta_1; z_m)$ is a vector of moments for observation m which depends on data z_m.

Second stage. Estimated parameters $\hat{\theta}_2$ solve moment conditions $\bar{g}(\hat{\theta}_2, \hat{\theta}_1) = 0$ where $\bar{g}(\theta_2, \hat{\theta}_1) = \frac{1}{N_2}\sum_n g_n(\theta_2, \hat{\theta}_1; x_n)$, N_2 is the number of observations, and $g_n(\theta_2, \hat{\theta}_1; x_n)$ is the moment for observation n based on its data and the first-stage estimates.
Each n in the second stage corresponds to an m in the first stage (e.g. n indexes customers, m indexes markets, and each customer buys in one market). Therefore, there are one or more observations in the second-stage sample from each group in the first-stage sample. Also, x_n in the second stage can include one or more elements of z_m for the m corresponding to n.

Take a Taylor expansion of both the first- and second-stage moment conditions around the true parameters θ_1^* and θ_2^* (assuming that the moment conditions are bounded and differentiable around the true parameters) and evaluate it asymptotically. For the first stage:

$$0 \stackrel{a}{=} \bar{h}(\theta_1^*) - A(\hat{\theta}_1 - \theta_1^*),$$

and for the second stage:

$$0 \stackrel{a}{=} \bar{g}(\theta_2^*, \theta_1^*) - B(\hat{\theta}_1 - \theta_1^*) - C(\hat{\theta}_2 - \theta_2^*),$$

where

$$\begin{aligned} A &= -\text{plim}\nabla_{\theta_1}\bar{h}(\theta_1^*) \\ B &= -\text{plim}\nabla_{\theta_1}\bar{g}(\theta_2^*, \theta_1^*) \\ C &= -\text{plim}\nabla_{\theta_2}\bar{g}(\theta_2^*, \theta_1^*), \end{aligned}$$

and where $\stackrel{a}{=}$ denotes equality up to asymptotically negligible remainder terms.[1] Then

$$(\hat{\theta}_1 - \theta_1^*) \stackrel{a}{=} A^{-1}\bar{h}(\theta_1^*) = A^{-1}\frac{1}{N_1}\sum_m h_m(\theta_1^*, z_m),$$

and

$$\begin{aligned}(\hat{\theta}_2 - \theta_2^*) &\stackrel{a}{=} C^{-1}(\bar{g}(\theta_2^*, \theta_1^*) - B(\hat{\theta}_1 - \theta_1^*)) \\ &\stackrel{a}{=} C^{-1}\left(\frac{1}{N_2}\sum_n g_n(\theta_2^*, \theta_1^*; x_n) - BA^{-1}\frac{1}{N_1}\sum_m h_m(\theta_1^*, z_m)\right). \qquad (1)\end{aligned}$$

Given that the N_2 second-stage observations can be grouped into N_1 first-stage observations, we can re-write

$$\sum_{n=1}^{N_2} g_n(\theta_2, \theta_1; x_n) = \sum_{m=1}^{N_1}\sum_{\ell=1}^{N^m} g_{\ell m}(\theta_2, \theta_1; x_{\ell m}),$$

where N^m is the number of observations among the N_2 second-stage observations that correspond to observation m of the first stage (e.g. number of customers in market m) and $g_{\ell m}$ is the second-stage moment for the ℓth second-stage observation that corresponds to the mth first-stage observation. Then equation (1) becomes

$$(\hat{\theta}_2 - \theta_2^*) \stackrel{a}{=} C^{-1}\left[\frac{1}{N_2}\sum_{m=1}^{N_1}\sum_{\ell=1}^{N^m} g_{\ell m}(\theta_2^*, \theta_1^*; x_{\ell m}) - BA^{-1}\frac{1}{N_1}\sum_{m=1}^{N_1} h_m(\theta_1^*, z_m)\right]$$

or, expressed more conveniently,

$$(\hat{\theta}_2 - \theta_2^*) \stackrel{a}{=} C^{-1}\left[\frac{1}{N_1}\sum_{m=1}^{N_1}\sum_{\ell=1}^{N^m} \frac{N_1}{N_2} g_{\ell m}(\theta_2^*, \theta_1^*; x_{\ell m}) - BA^{-1}\frac{1}{N_1}\sum_{m=1}^{N_1} h_m(\theta_1^*, z_m)\right].$$

Define $\tilde{g}_m = \sum_{\ell=1}^{N^m} C^{-1}\frac{N_1}{N_2} g_{\ell m}(\theta_2^*, \theta_1^*; x_{\ell m})$ and $\tilde{h}_m = C^{-1}BA^{-1}h_m(\theta_1^*, z_m)$. Then:

$$\sqrt{N_1}(\hat{\theta}_2 - \theta_2^*) \stackrel{a}{=} \frac{1}{\sqrt{N_1}}\sum_{m=1}^{N_1}(\tilde{g}_m - \tilde{h}_m).$$

Finally, letting $X_m = \tilde{g}_m - \tilde{h}_m$, we have

$$\sqrt{N_1}(\hat{\theta}_2 - \theta_2^*) \stackrel{a}{=} \frac{1}{\sqrt{N_1}}\sum_{m=1}^{N_1} X_m.$$

Asymptotically, $\sqrt{N_1}(\hat{\theta}_2 - \theta_2^*) \stackrel{a}{\sim} N(0, \mathrm{Var}(X_m))$, such that $\hat{\theta}_2 \stackrel{a}{\sim} N(\theta_2^*, \mathrm{Var}(X_m)/N_1)$. $\mathrm{Var}(X_m)$ is approximated by its sample analog, $\frac{1}{N_1}\sum_m \hat{X}_m\hat{X}_m'$, with $\hat{X}_m$ calculated at the estimated parameters (and A, B, and C replaced by sample analogs). This is a re-expression of the Murphy

[1] A and C are square matrices in this notation, such that the parameters are exactly identified by the moment conditions. The notation can of course be generalized.

and Topel (1985) result for an application in which the correlation in the random components of the first and second steps take a particular form due to the nested structure of the data. Note that if there is no error in the first stage (i.e. $h_m(\theta_1^*) = 0 \quad \forall m$), then this formula for the asymptotic covariance becomes the robust covariance estimator of $\hat{\theta}_2$ that allows for correlation over observations within each market.[2] Also note that these formulae assume that $N_1 \to \infty$, though $\mathrm{Var}(X_m)$ will decline as N_2 increases as well.

3. APPLICATION

We adapt and apply the formula to the model of Petrin and Train (2002). Their model is a mixed logit of choice at the consumer level that uses an explanatory variable that is estimated by regression on market level data. They examine customers' choice of TV reception, with the alternatives being: antenna only, cable with basic or extended service, cable with premium packages, or satellite. Each customer lives in a franchise area, called a market, and the price and other attributes of the alternatives vary over markets. Some attributes of the alternatives, such as quality of programming, are not measurable and hence not included as explanatory variables in the customer choice model. Since the omitted attributes are expected to be related to price, their omission induces correlation between the unobserved portion of utility and price. One of the ways they address this correlation is through a control function approach; in particular, they regress market price against market-level instruments and then enter the residual of this price regression as an explanatory variable in the customer-level choice model. Unobserved utility conditional on this price residual need not be correlated with price; the density of this conditional unobserved utility determines the form of the choice model.

The model is specified as follows. The first stage consists of OLS applied to linear regressions:

$$p_{mj} = \beta_j' z_m + \mu_{mj},$$

where p_{mj} is the price of alternative j in market m and z_m are instruments. The estimated residuals $\hat{\mu}_{mj} = p_{mj} - \hat{\beta}_j' z_m$ are calculated. The utility that customer n who resides in market m obtains from alternative j is specified as

$$U_{nmj} = \alpha' w_{nmj} + \lambda_j \mu_{mj} + \varepsilon_{nmj}.$$

The explanatory variables w_{nmj} include the price and other observed attributes of alternative j in market m, interacted in some cases with demographics of the customer. In estimation, the $\hat{\mu}_{mj}$ is used in lieu of the true μ_{mj}. Petrin and Train assume that the error ε_{nmj} contains a component that is normally distributed and common to all non-antenna alternatives, plus an i.i.d. extreme value term. Their choice model is therefore a mixed logit, with mixing over the normal error component.

The price of antenna only is zero for all customers, and the price of satellite does not vary over markets. Price regressions are therefore estimated only for the two cable alternatives, and only the utility for these two alternatives includes price residuals. The two cable alternatives are denoted $j = 2, 3$.

[2]The robust covariance estimator is the 'cluster' estimator implemented by Rogers (1993) and marketed by Stata.

With this specification, we have $\theta_1' = \langle \beta_2', \beta_3' \rangle$ and $\theta_2' = \langle \alpha', \lambda_2, \lambda_3 \rangle$. The first-stage moments are $h_m' = \langle z_m \mu_{m2}, z_m \mu_{m2} \rangle$ and the second-stage moments are $g_{nm} = \nabla_{\theta_2} L_{nm}(\theta_2, \hat{\theta}_1)$, where L_{nm} is the log-likelihood for customer n in market m. We need to calculate A, B and C.

Matrix A is the expected moment matrix of the instruments. Let Z be the matrix whose mth row is z_m'. Then A is approximated by its sample analog,

$$\left(\begin{array}{c|c} Z'Z/N_1 & 0 \\ \hline 0 & Z'Z/N_1 \end{array}\right).$$

Matrix C is the negative of the expected Hessian in the second-stage model evaluated at the true parameters: $C = -\nabla_{\theta_2\theta_2} L(\theta_2^*, \theta_1^*)$, where L is the expected log-likelihood of an observation. Using the information identity, C can be approximated by the sample variance of the scores at the estimated parameters,

$$\frac{1}{N_2} \sum_{m=1}^{N_1} \sum_{\ell=1}^{N^m} \nabla_{\theta_2} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1) \nabla_{\theta_2} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1)',$$

where $L_{\ell m}$ is the log-likelihood for the ℓth household in market m.

Matrix B is the expected derivative of the second-stage scores with respect to the first-stage parameters: $B = -\nabla_{\theta_2\theta_1} L(\theta_2^*, \theta_1^*)$. Analogous to the proof for the information identity, $B = \nabla_{\theta_2} L(\theta_2^*, \theta_1^*) \nabla_{\theta_1} L(\theta_2^*, \theta_1^*)'$. Its empirical analog is therefore $\frac{1}{N_2} \sum_{m=1}^{N_1} \sum_{\ell=1}^{N^m} \nabla_{\theta_2} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1) \nabla_{\theta_1} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1)'$. The first term is the score. The second term takes a particularly convenient form in this application. Note that

$$\frac{\partial L_{\ell m}}{\partial \lambda_j} = \frac{1}{P_{\ell m}} \frac{\partial P_{\ell m}}{\partial (\lambda_j \mu_{mj})} \mu_{mj},$$

where $P_{\ell m}$ is the probability of the chosen alternative of the ℓth customer in market m. Since, by definition, $\mu_{mj} = p_{mj} - \beta_j' z_m$, we have

$$\frac{\partial L_{\ell m}}{\partial \beta_j^k} = \frac{1}{P_{\ell m}} \frac{\partial P_{\ell m}}{\partial (\lambda_j \mu_{mj})} \frac{\partial \lambda_j \mu_{mj}}{\partial \beta_j^k} = \frac{1}{P_{\ell m}} \frac{\partial P_{\ell m}}{\partial (\lambda_j \mu_{mj})} (-\lambda_j z_m^k) = \frac{\partial L_{\ell m}}{\partial \lambda_j} \left(-\frac{\lambda_j}{\mu_m j} z_m^k \right),$$

where superscript k refers to the kth element of β_j and z_m. Then, collecting elements, we have:

$$\nabla_{\theta_1} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1) = -\begin{pmatrix} \nabla_{\lambda_2} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1) \frac{\hat{\lambda}_2}{\hat{\mu}_{m2}} z_m \\ \nabla_{\lambda_3} L_{\ell m}(\hat{\theta}_2, \hat{\theta}_1) \frac{\hat{\lambda}_3}{\hat{\mu}_{m3}} z_m \end{pmatrix}.$$

Using these quantities, X_m is calculated as

$$X_m = \sum_{\ell=1}^{N^m} C^{-1} \frac{N_1}{N_2} g_{\ell m}(\hat{\theta}_2, \hat{\theta}_1; x_{\ell m}) - C^{-1} B A^{-1} h_m(\hat{\theta}_1, z_m),$$

where the terms A, B, C, $g_{\ell m}$, and h_m are defined as above. The covariance of the second-stage estimator is then calculated as the empirical covariance of X_m over the markets in the sample, divided by N_1.

Table 1 gives the estimated model from Petrin and Train (2002). The first column of standard errors is calculated from the formulae just described. The second column gives the standard

Table 1. Mixed logit model of TV reception choice.

Explanatory variable	Estimates	Standard errors	
		Asympt. formula	Uncorrected
Price, in dollars per month (1–4)	−0.0969	0.0407	0.0174
Price for income group 2 (1–4)	0.0150	0.0023	0.0024
Price for income group 3 (1–4)	0.0247	0.0033	0.0030
Price for income group 4 (1–4)	0.0269	0.0035	0.0033
Price for income group 5 (1–4)	0.0308	0.0034	0.0036
Number of cable channels (2,3)	0.0026	0.0035	0.0015
Number of premium channels (3)	0.0448	0.0243	0.0162
Number of over-the-air channels (1)	0.0222	0.0151	0.0089
Whether pay per view is offered (2,3)	0.5813	0.1741	0.0761
Indicator: ATT is cable company (2)	−0.1949	0.2388	0.1060
Indicator: ATT is cable company (3)	−0.2370	0.2345	0.1199
Indicator: Adelphia Comm is cable company (2)	0.3425	0.2932	0.1224
Indicator: Adelphia Comm is cable company (3)	0.2392	0.3030	0.1491
Indicator: Cablevision is cable company (2)	0.1342	0.3608	0.2402
Indicator: Cablevision is cable company (3)	0.7350	0.3838	0.2516
Indicator: Charter Comm is cable company (2)	−0.0580	0.2311	0.1006
Indicator: Charter Comm is cable company (3)	−0.1757	0.1856	0.1270
Indicator: Comcast is cable company (2)	−0.0938	0.3682	0.1190
Indicator: Comcast is cable company (3)	0.1656	0.2723	0.1316
Indicator: Cox Comm is cable company (2)	−0.0577	0.3267	0.1475
Indicator: Cox Comm is cable company (3)	0.0874	0.4386	0.1691
Indicator: Time-Warner is cable company (2)	−0.0817	0.2261	0.0995
Indicator: Time-Warner is cable company (3)	−0.0689	0.2017	0.1203
Education level of household (2)	−0.0619	0.0267	0.0220
Education level of household (3)	−0.1123	0.0329	0.0278
Education level of household (4)	−0.1967	0.0367	0.0368
Household size (2)	−0.0518	0.0290	0.0240
Household size (3)	0.0134	0.0291	0.0287
Household size (4)	0.0050	0.0447	0.0358
Household rents dwelling (2–3)	−0.2436	0.0913	0.0863
Household rents dwelling (4)	−0.2149	0.1327	0.1562
Single family dwelling (4)	0.7649	0.2022	0.1521
Residual for extended-basic cable price (2)	0.0805	0.0422	0.0177
Residual for premium cable price (4)	0.0873	0.0423	0.0178
Alternative specific constant (2)	2.972	0.8984	0.5012
Alternative specific constant (3)	2.903	1.379	0.6904
Alternative specific constant (4)	4.218	2.319	1.087
Error components, standard deviation (2–4)	0.5553	0.6826	0.6410
Log-likelihood at convergence: −14635.47			
Number of observations: 11810			

Alternatives: (1) Antenna only, (2) basic and extended cable, (3) premium cable, (4) satellite. Variables enter alternatives in parentheses and zero in other alts.

errors that are produced by the standard maximum likelihood estimation routine, which treats the estimated price residuals as true. In this application, ignoring the sampling variance in the first-stage estimates would be seriously misleading. For example, the standard error on the base price coefficient rises from 0.0174 without correction to 0.0407 when the first-stage sampling variance is considered.[3] As expected, the standard error correction has the greatest impact for the base price coefficient and the coefficients of the price residuals. The impact is minimal for the demographic variables.

ACKNOWLEDGEMENT

We gratefully acknowledge the assistance of James Powell.

REFERENCES

Amemiya, T. (1978). On two-step estimation of multivariate logit models. *Journal of Econometrics 8*, 13–21.

Greene, W. (2000). *Econometric Analysis*, 4th edn. Upper Saddle River, NJ: Prentice-Hall.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–62.

Kuksov, D. and M. Villas-Boas (2001). Endogeneity and individual customer choice. Working Paper, Haas School of Business, University of California, Berkeley, `http://groups.haas.berkeley.edu/marketing/PAPERS/VILLAS/Paper.pdf`.

McFadden, D. (1999). *Economics 240b*, Lecture Notes. Department of Economics, University of California, Berkeley, `http://emlab.berkeley.edu/users/mcfadden/e240b\_f01/ch3.pdf`.

Murphy, K. and R. Topel (1985). Estimation and inference in two step econometric models. *Journal of Business and Economic Statistics 3*, 370–9.

Newey, W. and D. McFadden (1994). Large sample estimation and hypothesis testing. In Z. Griliches and M. Intriligator (eds), *The Handbook of Econometrics*, vol. 4, pp. 2111–245. Amsterdam: Elsevier Science Publishers.

Petrin, A. and K. Train (2002). Omitted product attributes in discrete choice models. Working Paper, Department of Economics, University of California, Berkeley, `http://elsa.berkeley.edu/wp/train1202.pdf`.

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin 13*, 19–23. (Reprinted in Stata Technical Bulletin Reprints, vol. 3, 88–94).

[3]For the second-stage standard errors, Petrin and Train (2002) use a bootstrap method that can be applied with canned software, unlike the matrix manipulations required by the formula in this paper. Details of the bootstrap procedure can be found in Petrin and Train (2002). The standard errors based on the asymptotic formula and the bootstrap are similar. For example, the price coefficient receives a standard error of 0.0407 using the asymptotic formula and 0.0400 using the bootstrap procedure. The standard error on the number of cable channels is 0.0035 using the asymptotic formula and 0.0029 using the bootstrap procedure. This similarity suggests that bootstrapping on the first-stage regression provides reasonable standard errors and can perhaps serve as a useable method for researchers who do not want to program the asymptotic formula.