The Stata Journal

The *Stata Technical Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press, and Stata is a registered trademark of Stata Corporation.

The Stata Journal (2002)
2, Number 3, pp. 253–266

The robust variance estimator for two-stage models

James W. Hardin
Texas A&M University

Abstract. This article discusses estimates of variance for two-stage models. We present the sandwich estimate of variance as an alternative to the Murphy–Topel estimate. The sandwich estimator has a simple formula that is similar to the formula for the Murphy–Topel estimator, and the two estimators are asymptotically equal when the assumed model distributions are true. The advantages of the sandwich estimate of variance are that it may be calculated for the complete parameter vector, and that it requires estimating equations instead of fully specified log likelihoods.

Keywords: st0018, robust variance estimator, Murphy–Topel estimator, two-stage estimation, estimating equation

1 Introduction

Numerous models have been presented in the literature in which one model is embedded in another. Such models are broadly known as two-step estimation problems and are characterized by

$$\begin{aligned} \text{Model 1} &: \quad E\left\{\mathbf{y}_1|\mathbf{x}_1,\theta_1\right\} \\ \text{Model 2} &: \quad E\left\{\mathbf{y}_2|\mathbf{x}_2,\theta_2,\mathrm{E}\left(\mathbf{y}_1|\mathbf{x}_1,\theta_1\right)\right\} \end{aligned}$$

The overall model indicates that there are two parameter vectors to estimate. The first parameter vector θ_1 appears in both models, but the second parameter vector θ_2 appears only in the second model.

There are two standard approaches to estimation. The first approach is a full information maximum likelihood, FIML, model in which we specify the joint distribution $f(\mathbf{y}_1,\mathbf{y}_2|\mathbf{x}_1,\mathbf{x}_2,\theta_1,\theta_2)$ and maximize the joint log-likelihood function. Alternatively, we can adopt a limited information maximum likelihood, LIML, two-step procedure. In this approach, we estimate the first model, since it does not involve the second parameter vector. Subsequently, we estimate the second parameter vector conditional on the results of the first step estimation; we maximize the conditional log-likelihood $\mathcal{L}$ given by

$$\mathcal{L} = \sum_{i=1}^{n} \ln f\left\{y_{2i}|\mathbf{x}_{2i},\theta_2,(\mathbf{x}_{1i},\widehat{\theta}_1)\right\}$$

Here, and throughout this article, we assume that there are n observations, $\mathbf{x}_{1i}$ is the ith row of the $\mathbf{X}_1$ design matrix, $\mathbf{x}_{2i}$ is the ith row of the $\mathbf{X}_2$ design matrix, and $\widehat{\theta}_1$ is the maximum likelihood estimate obtained from the estimation of Model 1.

2 Murphy–Topel estimate of variance for two-stage models

Greene (2000) gives a concise presentation of one of the results in Murphy and Topel (1985). The presentation describes a general formula of a valid variance estimator for θ_2 in a two-stage maximum likelihood estimation model. This LIML estimation fits one model, which is then used to generate covariates for a second model of primary interest. Calculation of a variance estimate for the regressors θ_2 in the primary model of interest must address the fact that one or more of the regressors have been generated via $(\mathbf{x}_1, \widehat{\theta}_1)$.

In order to highlight the derivation and comparison to the sandwich estimate of variance, we assume that θ_1 is a $q \times 1$ vector of unknown parameters associated with an $n \times q$ matrix of covariates $\mathbf{X}$. In addition, θ_2 is a $p \times 1$ vector of unknown parameters associated with an $n \times p$ matrix of covariates $\mathbf{W}$.

Following Greene (2000), the formula for the Murphy–Topel variance estimate for θ_2 is given by

$$\mathbf{V}_2 + \mathbf{V}_2\Big(\mathbf{C}\mathbf{V}_1\mathbf{C}^{\mathrm{T}} - \mathbf{R}\mathbf{V}_1\mathbf{C}^{\mathrm{T}} - \mathbf{C}\mathbf{V}_1\mathbf{R}^{\mathrm{T}}\Big)\mathbf{V}_2 \tag{1}$$

where

$$\begin{aligned}
\mathbf{V}_1 &= (q \times q) \text{ Asymptotic variance matrix of } \widehat{\theta}_1 \text{ based on } \mathcal{L}_1(\theta_1)\\
\mathbf{V}_2 &= (p \times p) \text{ Asymptotic variance matrix of } \widehat{\theta}_2 \text{ based on } \mathcal{L}_2(\theta_2|\theta_1)\\
\mathbf{C} &= (p \times q) \text{ matrix given by } E\left\{\left(\frac{\partial \mathcal{L}_2}{\partial \theta_2}\right)\left(\frac{\partial \mathcal{L}_2}{\partial \theta_1^{\mathrm{T}}}\right)\right\}\\
\mathbf{R} &= (p \times q) \text{ matrix given by } E\left\{\left(\frac{\partial \mathcal{L}_2}{\partial \theta_2^{\mathrm{T}}}\right)\left(\frac{\partial \mathcal{L}_1}{\partial \theta_1^{\mathrm{T}}}\right)\right\}
\end{aligned} \tag{2}$$

We assume that $\mathbf{V}_1$ and $\mathbf{V}_2$ are calculated as the inverse matrix of negative second derivatives. This is not required (as indicated), and some researchers will substitute the outer product of the gradient instead. The asymptotic equivalence of these estimators is given by

$$E\left\{\left(\frac{\partial \mathcal{L}}{\partial \theta}\right)\left(\frac{\partial \mathcal{L}}{\partial \theta^{\mathrm{T}}}\right)\right\} = -E\left\{\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta^{\mathrm{T}}}\right\} \tag{3}$$

The component matrices of the Murphy–Topel estimator are estimated by evaluating the formulae at the maximum likelihood estimates $\widehat{\theta}_1$ and $\widehat{\theta}_2$. The presentation assumes the existence of a log likelihood for the first model $\mathcal{L}_1(\theta_1)$ and a conditional log-likelihood for the second (primary) model of interest $\mathcal{L}_2(\theta_2|\theta_1)$.

To gain a better appreciation and understanding of the formula in equation (1), we derive the sandwich estimate of variance for the same class of models. Our derivation assumes the first model has an estimating equation $\Psi_1(\theta_1)$, and the second model has an estimating equation $\Psi_2(\theta_2|\theta_1)$. The results we present follow from the lucid presentation of the theoretical justifications given in Stefanski and Boos (2002). Our goal is to build the sandwich estimate of variance for $\Theta = (\theta_1, \theta_2)$. We can partition the overall estimating equation as

$$[\Psi(\Theta)] = \left[\begin{array}{c} \Psi_1(\theta_1) \\ \Psi_2(\theta_2|\theta_1) \end{array}\right] = [\mathbf{0}]$$

Following Binder (1983), we know that the sandwich estimate of variance $\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathrm{T}}$ for the complete parameter vector $\Theta = (\theta_1, \theta_2)$ can be written in terms of the partitioned matrices for the complete estimating equation,

$$\begin{aligned} \mathbf{A} &= \left[\begin{array}{cc} \dfrac{\partial\Psi_1}{\partial\theta_1^{\mathrm{T}}} & \dfrac{\partial\Psi_1}{\partial\theta_2^{\mathrm{T}}} \\ \\ \dfrac{\partial\Psi_2}{\partial\theta_1^{\mathrm{T}}} & \dfrac{\partial\Psi_2}{\partial\theta_2^{\mathrm{T}}} \end{array}\right] \\ \mathbf{B} &= \left[\begin{array}{cc} \Psi_1\Psi_1^{\mathrm{T}} & \Psi_1\Psi_2^{\mathrm{T}} \\ \\ \Psi_2\Psi_1^{\mathrm{T}} & \Psi_2\Psi_2^{\mathrm{T}} \end{array}\right] \end{aligned}$$

Since the sandwich estimate of variance for estimating $\Theta = (\theta_1, \theta_2)$ is given by $\mathbf{V}_{\mathrm{S}} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathrm{T}}$, the estimated sandwich variance matrix may be partitioned to emphasize that

$$\widehat{\mathbf{V}}_{\mathrm{S}} = \left[\begin{array}{cc} \left\{\widehat{\mathbf{V}}_{\mathrm{S}}(\theta_1)\right\}_{q\times q} & \left\{\widehat{\mathbf{Cov}}_{\mathrm{S}}(\theta_1,\theta_2)\right\}_{q\times p} \\ \\ \left\{\widehat{\mathbf{Cov}}_{\mathrm{S}}^{\mathrm{T}}(\theta_1,\theta_2)\right\}_{p\times q} & \left\{\widehat{\mathbf{V}}_{\mathrm{S}}(\theta_2)\right\}_{p\times p} \end{array}\right] \qquad (4)$$

It follows that the sandwich estimate of variance for the model of interest is the lower right $p \times p$ partition of $\widehat{\mathbf{V}}_{\mathrm{S}}$.

Our presentation of the sandwich estimate of variance has thus far been under less restrictive assumptions than the presentation of the Murphy–Topel variance estimate. Thus far, we have only assumed the existence of two estimating equations. Stefanski and Boos (2002) point out that we are, in fact, building the variance estimate of an M-estimator as described in Huber (1967); the former reference gives the name *partial M-estimator* to our particular case. The two-stage models addressed by the Murphy–Topel estimator are, in fact, a special case of partial M-estimators.

For the sake of comparison of the sandwich and Murphy–Topel variance estimates, we now assume that the estimating equations are derived from models with valid log-likelihoods. Under this assumption, the estimating equations are derivatives of the model log-likelihoods,

$$\begin{aligned} \Psi_1(\theta_1) &= \frac{\partial\mathcal{L}_1(\theta_1)}{\partial\theta_1} \\ \Psi_2(\theta_2|\theta_1) &= \frac{\partial\mathcal{L}_2(\theta_2|\theta_1)}{\partial\theta_2} \end{aligned}$$

Using the log-likelihood notation $\mathcal{L}$, we calculate the sandwich estimate of variance in terms of the two log-likelihoods as

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}_1}{\partial\theta_1 \partial\theta_1^{\mathrm{T}}} & \dfrac{\partial^2 \mathcal{L}_1}{\partial\theta_1 \partial\theta_2^{\mathrm{T}}} \\ \\ \dfrac{\partial^2 \mathcal{L}_2}{\partial\theta_2 \partial\theta_1^{\mathrm{T}}} & \dfrac{\partial^2 \mathcal{L}_2}{\partial\theta_2 \partial\theta_2^{\mathrm{T}}} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 \mathcal{L}_1}{\partial\theta_1 \partial\theta_1^{\mathrm{T}}} & \mathbf{0} \\ \\ \dfrac{\partial^2 \mathcal{L}_2}{\partial\theta_2 \partial\theta_1^{\mathrm{T}}} & \dfrac{\partial^2 \mathcal{L}_2}{\partial\theta_2 \partial\theta_2^{\mathrm{T}}} \end{bmatrix} \qquad (5) \\
\mathbf{B} &= \begin{bmatrix} \left\{ \left(\dfrac{\partial \mathcal{L}_1}{\partial\theta_1}\right) \left(\dfrac{\partial \mathcal{L}_1}{\partial\theta_1^{\mathrm{T}}}\right) \right\} & \left\{ \left(\dfrac{\partial \mathcal{L}_1}{\partial\theta_1}\right) \left(\dfrac{\partial \mathcal{L}_2}{\partial\theta_2^{\mathrm{T}}}\right) \right\} \\ \\ \left\{ \left(\dfrac{\partial \mathcal{L}_2}{\partial\theta_2}\right) \left(\dfrac{\partial \mathcal{L}_1}{\partial\theta_1^{\mathrm{T}}}\right) \right\} & \left\{ \left(\dfrac{\partial \mathcal{L}_2}{\partial\theta_2}\right) \left(\dfrac{\partial \mathcal{L}_2}{\partial\theta_2^{\mathrm{T}}}\right) \right\} \end{bmatrix}
\end{aligned}$$

In equation 5, the upper right matrix entry of $\mathbf{A}$ is zero since θ_2 does not enter into $\mathcal{L}_1$. This is a common occurrence when a partial M-estimator for a model is specified from multiple estimating equations. The same result is seen in Liang and Zeger (1986) for generalized estimating equations.

Substituting the component matrices used in the calculation of the Murphy–Topel estimator, we have

$$\begin{aligned}
\mathbf{A} &= \begin{bmatrix} -\mathbf{V}_1^{-1} & \mathbf{0} \\ -\mathbf{C}^* & -\mathbf{V}_2^{-1} \end{bmatrix} \\
\mathbf{B} &= \begin{bmatrix} \mathbf{V}_1^{*-1} & \mathbf{R}^{\mathrm{T}} \\ \mathbf{R} & \mathbf{V}_2^{*-1} \end{bmatrix}
\end{aligned}$$

The inverse of $\mathbf{A}$ is then given by

$$\mathbf{A}^{-1} = \begin{bmatrix} -\mathbf{V}_1 & \mathbf{0} \\ \mathbf{V}_2\mathbf{C}^*\mathbf{V}_1 & -\mathbf{V}_2 \end{bmatrix}$$

Our use of asterisks $(*)$ as superscripts distinguishes similar matrix components. The asterisk appears when the component in the sandwich estimator differs from the corresponding component in the Murphy–Topel estimator. The difference is in the evaluation based on the two approaches described by equation (3). For example, the $\mathbf{C}$ matrix in the Murphy–Topel estimator is the outer product of the gradients, equation (2), while the $\mathbf{C}^*$ matrix in the sandwich estimator is the inverse matrix of second derivatives; see the lower left matrix of equation (5).

We can carry out the matrix multiplication $\mathbf{V}_{\mathrm{S}}(\Theta) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-\mathrm{T}}$ to see that the matrix elements of equation (4) are given by

$$\begin{aligned}
\mathbf{V}_{\mathrm{S}}(\theta_1) &= \mathbf{V}_1\mathbf{V}_1^{*-1}\mathbf{V}_1 \\
&= \mathbf{V}_{\mathrm{S}1} \\
\mathbf{Cov}_{\mathrm{S}}(\theta_1,\theta_2) &= \mathbf{V}_1\mathbf{R}^{\mathrm{T}}\mathbf{V}_2 - \mathbf{V}_1\mathbf{V}_1^{*-1}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}}\mathbf{V}_2 \\
&= \mathbf{V}_1\mathbf{R}^{\mathrm{T}}\mathbf{V}_2 - \mathbf{V}_{\mathrm{S}1}\mathbf{C}^{*\mathrm{T}}\mathbf{V}_2 \\
\mathbf{V}_{\mathrm{S}}(\theta_2) &= \mathbf{V}_2\mathbf{C}^{*}\mathbf{V}_1\mathbf{V}_1^{*-1}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}}\mathbf{V}_2 - \mathbf{V}_2\mathbf{R}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}}\mathbf{V}_2 - \mathbf{V}_2\mathbf{C}^{*}\mathbf{V}_1\mathbf{R}^{\mathrm{T}}\mathbf{V}_2 \\
&\quad + \mathbf{V}_2\mathbf{V}_2^{*-1}\mathbf{V}_2 \\
&= \mathbf{V}_2\mathbf{V}_2^{*-1}\mathbf{V}_2 + \mathbf{V}_2\left(\mathbf{C}^{*}\mathbf{V}_1\mathbf{V}_1^{*-1}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}} - \mathbf{R}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}} - \mathbf{C}^{*}\mathbf{V}_1\mathbf{R}^{\mathrm{T}}\right)\mathbf{V}_2 \\
&= \mathbf{V}_{\mathrm{S}2} + \mathbf{V}_2\left(\mathbf{C}^{*}\mathbf{V}_{\mathrm{S}1}\mathbf{C}^{*\mathrm{T}} - \mathbf{R}\mathbf{V}_1\mathbf{C}^{*\mathrm{T}} - \mathbf{C}^{*}\mathbf{V}_1\mathbf{R}^{\mathrm{T}}\right)\mathbf{V}_2 \qquad (6)
\end{aligned}$$

The sandwich estimate of variance for θ_1 is the usual result for a single model. This is the expected result since the first model does not involve the second parameter vector.

As mentioned previously, this is the same result obtained for the case of generalized estimating equations (GEE). Looking at the variance estimate in this way highlights why the sandwich estimate of variance for the regression coefficients in GEE is said to be robust to misspecification of the assumed correlation structure—because the correlation parameters do not enter the calculation of the variance matrix for the regression coefficients (they only affect the efficiency of the coefficient estimates).

The sandwich estimate of variance for θ_2 given in equation (6) has a form that is similar to the Murphy–Topel variance estimate in equation (1). The differences are in the use of the sandwich estimators, $\mathbf{V}_{\mathrm{S}1}$ and $\mathbf{V}_{\mathrm{S}2}$ from the individual models, and the specification of the matrix of second derivatives estimator $\mathbf{C}^*$ over the outer product of the gradient estimator $\mathbf{C}$.

3 Example

Greene (2000) provides a model of consumer behavior. The dependent variable of interest is the number of derogatory reports for a sample of people applying for a credit card. This variable is a nonnegative integer that is zero for the majority of applicants, but values up to ten are not unusual. We address this dependent variable via a Poisson regression model. The original study included a secondary model for the outcome of the application. This outcome is binary and modelled using logistic regression. The predicted probability of the logistic model is used as one of the covariates in the Poisson model of interest.

From the original study, the author makes 100 observations available for our use. The initial logistic model is in terms of `z`, an indicator of whether the application is accepted. This outcome is a function of `age`, the applicant's age in years; `income`, the annual income; `ownrent`, an indicator of whether the applicant owns their home;

and `selfemp`, an indicator of whether the applicant is self employed. The model also includes a constant term.

A logistic model where $\mathbf{z}$ is the outcome (whether the application is successful) and $\mathbf{X}$ is the matrix of covariates has log likelihood given by

$$\mathcal{L}_z = \sum_{i=1}^{n} [z_i \mathbf{x}_i \theta_1 - \ln\{1 + \exp(\mathbf{x}_i \theta_1)\}]$$

The results for fitting the logistic regression model are given by

```
Logit estimates                                   Number of obs   =        100
                                                  LR chi2(4)      =       8.80
                                                  Prob > chi2     =     0.0662
Log likelihood = -53.924625                       Pseudo R2       =     0.0755

------------------------------------------------------------------------------
           z |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0732769   .0316192    -2.32   0.020    -.1352493   -.0113045
      income |   .2192029   .1800238     1.22   0.223    -.1336373    .5720431
    ownrent |    .189368   .5417589     0.35   0.727    -.8724599    1.251196
     selfemp |  -1.943879   1.037069    -1.87   0.061    -3.976497    .0887385
       _cons |   2.723656   1.055066     2.58   0.010     .6557644    4.791547
------------------------------------------------------------------------------
```

The second stage is the Poisson model of interest. We model `y`, the number of derogatory reports, as a function of `age`, the applicant's age in years; `income`, the applicant's annual income; `expend`, the monthly average expenditures of the applicant; and $\widehat{z}$, the predicted probability that the application for a credit card is accepted stored in `zhat`. The predicted probabilities are calculated from the fitted first stage logistic model. The logistic model also includes a constant term.

A Poisson model where $\mathbf{y}$ is the outcome (number of derogatory reports) and $\mathbf{W}$ is the matrix of covariates has log-likelihood given by

$$\mathcal{L}_y = \sum_{i=1}^{n} \{y_i \mathbf{w}_i \theta_2 - \exp(\mathbf{w}_i \theta_2) - \ln\Gamma(y_i + 1)\}$$

The results for fitting the second stage Poisson model are given by

```
Poisson regression                                Number of obs   =        100
                                                  LR chi2(4)      =      27.21
                                                  Prob > chi2     =     0.0000
Log likelihood = -78.330992                       Pseudo R2       =     0.1480

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0731059   .0542458     1.35   0.178    -.0332139    .1794258
      income |   .0452336   .1741114     0.26   0.795    -.2960184    .3864856
      expend |  -.0068969     .00202    -3.41   0.001    -.0108561   -.0029378
        zhat |   4.632355   3.661774     1.27   0.206     -2.54459     11.8093
       _cons |  -6.319947   3.930768    -1.61   0.108    -14.02411    1.384217
------------------------------------------------------------------------------
```

This output includes the naive standard errors, which assume that there is no error in the generation of the $\widehat{z}$ predictor in the first-stage logit regression model.

4 Obtaining estimates in Stata

Stata makes it relatively easy to obtain the Murphy–Topel variance estimates for a two-stage model. The powerful commands that we have at our disposal are the `matrix accum` and `matrix vecaccum` commands.

A do-file for generating the results of the two stage estimation proceeds as follows. In this construction, we will include the details for building the naive, Murphy–Topel, and robust variance estimates. We begin with the specification of the two models.

```
                        /* Assumption: the data is already loaded */
logit z age income ownrent selfemp          /* First stage: logit */
matrix V1 = e(V)                            /* First stage variance estimate */
logit z age income ownrent selfemp, robust
matrix V1s = .99 * e(V)                     /* Undo the 100/(100-1) adjustment */
predict double zhat                         /* Covariate for second stage */

poisson y age income expend zhat            /* Second stage: poisson */
matrix V2 = e(V)
poisson y age income expend zhat, robust
matrix V2s = .99 * e(V)                     /* Undo the 100/(100-1) adjustment */
predict double yhat
scalar zz = _b[zhat]                        /* Coeff on generated variable */
```

With this much of the two-stage estimation specified, we obtain several of the pieces that we need to construct the desired estimates of variance. Thus far, we have `V1`, the variance of the first stage model; `V1s`, the robust variance of the first stage model; `V2`, the variance of the second stage model; and `V2s`, the robust variance of the second stage model. We note that `V2` is the naive variance estimate of the two-stage estimation and `V2s` is the naive robust variance estimate. Both naive estimators assume that the `zhat`$= \widehat{z}$ predictor from the first stage is without error.

Since Stata applies a small sample adjustment $n/(n-1)$ to the robust variance estimates, we undo that adjustment in defining those matrices. It is still left to calculate $\mathbf{R}$, $\mathbf{C}$, and $\mathbf{C}^*$. The first two matrices are relatively easy to calculate using the `predict` command to get intermediate results, and we apply the `matrix accum` command to generate the desired matrix. Continuing our development, the do-file is augmented with the following steps. The only trick we need to apply is that there is a constant in each of the two stages of estimation. The `matrix accum` command will (by default) add a constant to the end of the variable list. Instead, we generate our own constant variable, include it twice in the list, and specify the `nocons` option to prevent adding another constant to the list.

To highlight the calculation of these matrices with the `matrix accum` command, note that

$$
\begin{aligned}
\frac{\partial \mathcal{L}_z}{\partial \theta_1} &= \sum_{i=1}^{n} \mathbf{x}_i (z_i - \widehat{z}_i)\mathbf{x}_i^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}} \ \mathrm{Diag}(z_i - \widehat{z}_i) \ \mathbf{X} \\
\frac{\partial \mathcal{L}_z}{\partial \theta_2} &= \mathbf{0} \\
\frac{\partial \mathcal{L}_y}{\partial \theta_2} &= \sum_{i=1}^{n} \mathbf{w}_i (y_i - \widehat{y}_i)\mathbf{w}_i^{\mathrm{T}} = \mathbf{W}^{\mathrm{T}} \ \mathrm{Diag}(y_i - \widehat{y}_i) \ \mathbf{W} \\
\frac{\partial \mathcal{L}_y}{\partial \theta_1} &= \sum_{i=1}^{n} \mathbf{x}_i (y_i - \widehat{y}_i)(\widehat{z})(1-\widehat{z}_i)\widehat{\theta}_{2\widehat{z}}\mathbf{x}_i^{\mathrm{T}} = \mathbf{X}^{\mathrm{T}} \ \mathrm{Diag}\{(y_i - \widehat{y}_i)\widehat{z}_i(1-\widehat{z})\widehat{\theta}_{2\widehat{z}}\} \ \mathbf{X}
\end{aligned}
$$

where $\widehat{\theta}_{2\widehat{z}}$ is the estimated coefficient in the second stage model for the generated predictor $\widehat{z}$.

```
gen byte cons = 1

matrix accum C = age income ownrent selfemp cons age income expend zhat cons /*
                          */ [iw=(y-yhat)*(y-yhat)*zhat*(1-zhat)*zz], nocons
matrix accum R = age income ownrent selfemp cons age income expend zhat cons /*
                          */ [iw=(y-yhat)*(z-zhat)], nocons

matrix C = C[6..10,1..5]                    /* Get only the desired partition */
matrix R = R[6..10,1..5]                    /* Get only the desired partition */
```

At this point, we have all of the necessary information for building the Murphy–Topel variance estimate. However, we still need an estimate of $\mathbf{C}^*$ for the sandwich variance estimate. This second derivative is complicated by the dependence on the fitted values from the first stage.

$$
\begin{aligned}
\frac{\partial^2 \mathcal{L}_y}{\partial \theta_{2j} \partial \theta_{1k}} &= \sum_{i=1}^{n} \mathbf{w}_{ij}\{-\widehat{y}_i \widehat{z}_i (1-\widehat{z}_i)\widehat{\theta}_{2\widehat{z}}\}\mathbf{x}_{ik}^{\mathrm{T}} + \sum_{i=1}^{n} \mathcal{I}(\widehat{z})_i (y_i - \widehat{y}_i)\widehat{z}_i(1-\widehat{z}_i)\mathbf{x}_{ik}^{\mathrm{T}} \\
&= \mathbf{W}^{\mathrm{T}} \ \mathrm{Diag}\{-\widehat{y}_i\widehat{z}_i(1-\widehat{z}_i)\widehat{\theta}_{2\widehat{z}}\} \ \mathbf{X} \ + \ \mathcal{I}(\widehat{z})^{\mathrm{T}} \ \mathrm{Diag}\{(y_i - \widehat{y}_i)\widehat{z}_i(1-\widehat{z}_i)\} \ \mathbf{X}
\end{aligned}
$$

where $\mathcal{I}(\widehat{z})$ is an $(n \times p)$ matrix; the column associated with the generated covariate $\widehat{z}$ from the first stage is equal to one, and all other columns are zero. We can form this matrix using Stata's accumulation commands, but we address the two matrix products separately.

```
matrix accum Cs1 = age income ownrent selfemp cons age income expend zhat cons /*
                                        */ [iw=-yhat*zz*zhat*(1-zhat)], nocons
matrix Cs1 = Cs1[6..10,1..5]                 /* Get only the desired partition */

gen dd = (y-yhat)*zhat*(1-zhat)
matrix vecaccum Cs2 = dd age income ownrent selfemp cons, nocons
matrix Cs2 = J(5,3,0) , Cs2’ , J(5,1,0)     /* Plug into the relevant column */

matrix Cs  = -(Cs1 + Cs2’)
```

Armed with the accumulated information, the Murphy–Topel estimate may now be calculated as

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

and the Sandwich estimate may be calculated as

```
matrix Ms = V2s + (V2 * (Cs*V1s*Cs’ - R*V1*Cs’ - Cs*V1*R’) * V2)
```

Once the final estimates are formed, we can post them to the estimation areas and list them in the usual manner so that the variance estimates are available for testing. To post the Murphy–Topel estimates, we can append the following code to our do-file.

```
matrix b = e(b)

capture program drop doit
program define doit, eclass
        est post b M                  /* For sandwich results: est post b Ms */
        est local vcetype "Mtopel"    /* For sandwich results:
                                       * est local vcetype "Robust" */
        est display
end
doit
```

Alternatively, we can make the obvious adjustments to the do-file to list the sandwich estimate of variance results. If we run the above do-file, the final results with the Murphy–Topel variance estimates are listed as

```
             |              MTopel
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y            |
         age |   .0731059   .1096293     0.67   0.505    -.1417636    .2879755
      income |   .0452336   .4375397     0.10   0.918    -.8123285    .9027957
      expend |  -.0068969    .004265    -1.62   0.106    -.0152561    .0014623
        zhat |   4.632355   10.82669     0.43   0.669    -16.58757    25.85228
       _cons |  -6.319947   9.661564    -0.65   0.513    -25.25626    12.61637
```

5 Results

The results of the two-stage estimation problem along with standard errors are listed in Table 1. Results for the Murphy–Topel and sandwich estimators are similar for this model, and both estimators are approximately double the size of the naive results.

Interested readers will note differences in the results for the Murphy–Topel standard errors listed here and the ones listed in the cited text (current results for the text are listed in the errata on the author’s web site). The difference between the calculation used here and the calculation used in the text is for $\mathbf{V}_1$. The Murphy–Topel variance estimate specifies only that a valid variance estimate from the model should be used. We use the inverse matrix of negative second derivatives $\mathbf{V}_1$, while the text uses the outer product of the gradient $\mathbf{V}_1^*$.

Table 1: Coefficients and standard errors from the second stage Poisson model. Naive standard errors are calculated for the Poisson regression model, assuming the values of $\widehat{z}$ are true (without error).

	Coefficient	**Naive SE**	**Murphy–Topel SE**	**Sandwich SE**
age	.0731059	.0542458	.10962933	.09863122
income	.0452336	.1741114	.43753973	.36183127
expend	-.0068969	.0020200	.00426497	.00300891
$\widehat{z}$	4.632355	3.661774	10.826693	8.2048782
constant	-6.319947	3.930768	9.6615637	7.9570337

6 Simulation

A simulation study is not possible for every type of two-stage model that we may encounter. Here, we simulate data for two models similar to the previously illustrated example. The initial model is a logistic regression described by $y_i^* = \text{Logit}(\gamma_0 + \gamma_1 \mathbf{x}_{1i} + \gamma_2 \mathbf{x}_{2i} + \gamma_3 \mathbf{x}_{3i} + \gamma_4 \mathbf{x}_{4i})$. We simulate the covariates such that $\mathbf{X}_1 \sim \text{Uniform}(-.5, .5)$, $\mathbf{X}_2 \sim \text{Normal}(0, 1)$, $\mathbf{X}_3 \sim \text{Discrete Uniform}\{-1, 0, 1\}$, and $\mathbf{X}_4 \sim \text{Exponential}(1) - 1$. Logistic error is added to the calculation of the continuous outcome, and a binary outcome y_i is then generated.

The second model is a linear regression model for which data are generated such that $z_i = \beta_0 + \beta_1 \mathbf{w}_{1i} + \beta_2 \mathbf{x}_{2i} + \beta_3 \mathbf{x}_{3i} + \beta_4 y_i$ where $\mathbf{W}_1 \sim \text{Uniform}(-.5, .5)$. Normally, distributed error is added to the outcome z_i. We estimate the regression model using the fitted values $\widehat{y}_i$ from the first-stage logistic regression.

We consider six different sample sizes, $\{20, 40, 60, 80, 100, 1000\}$, and we expect similar coverage probabilities for the two estimators when data are generated from the fitted models. In addition, we simulate error for the regression model that depends on the value of $\mathbf{W}_1$. Here, we wish to investigate the robustness properties in terms of the coverage probabilities for the variance estimators; especially that for the β_1 coefficient on $\mathbf{W}_1$. Results of the simulations are listed in Table 2 and Table 3.

The coverage probabilities estimated from the simulations indicate that the Murphy–Topel and sandwich estimates of variance have similar coverage probabilities when the models are correct. For the heteroskedastic regression model, the sandwich estimate of variance has coverage probability that is closer to the nominal level than the Murphy–Topel estimate. This is especially true for large sample sizes where the Murphy–Topel underestimates the variance of the covariate on which the errors depend.

(*Continued on next page*)

Table 2: Coverage probabilities from 10,000 replications for simulation of correct model. Both estimators exhibit coverage probabilities close to the nominal level for all covariates.

	Murphy–Topel		Sandwich	
Coefficient	$p = 0.900$	$p = 0.950$	$p = 0.900$	$p = 0.950$
	$n = 20$		$n = 20$	
β_0	0.933	0.968	0.938	0.967
β_1	0.885	0.936	0.869	0.923
β_2	0.908	0.947	0.897	0.937
β_3	0.896	0.943	0.895	0.939
β_4	0.923	0.962	0.928	0.961
	$n = 40$		$n = 40$	
β_0	0.941	0.977	0.949	0.978
β_1	0.887	0.941	0.880	0.935
β_2	0.913	0.956	0.907	0.951
β_3	0.917	0.973	0.919	0.959
β_4	0.934	0.978	0.937	0.973
	$n = 60$		$n = 60$	
β_0	0.940	0.977	0.942	0.976
β_1	0.895	0.947	0.887	0.941
β_2	0.913	0.958	0.911	0.954
β_3	0.917	0.960	0.917	0.958
β_4	0.931	0.971	0.931	0.970
	$n = 80$		$n = 80$	
β_0	0.937	0.976	0.934	0.975
β_1	0.896	0.947	0.890	0.943
β_2	0.913	0.959	0.901	0.956
β_3	0.915	0.959	0.915	0.958
β_4	0.930	0.971	0.928	0.963
	$n = 100$		$n = 100$	
β_0	0.928	0.969	0.930	0.968
β_1	0.892	0.944	0.888	0.940
β_2	0.912	0.960	0.908	0.959
β_3	0.907	0.955	0.908	0.953
β_4	0.921	0.962	0.921	0.962
	$n = 1000$		$n = 1000$	
β_0	0.912	0.956	0.911	0.957
β_1	0.898	0.947	0.899	0.948
β_2	0.904	0.955	0.904	0.954
β_3	0.905	0.954	0.904	0.953
β_4	0.911	0.953	0.906	0.953

Table 3: Coverage probabilities from 10,000 replications for simulation of heteroskedastic model. The second stage regression model includes heteroskedastic error, depending on the values of the covariate associated with β_1. Note the discrepancy in results for the rows associated with this estimator.

	Murphy–Topel		Sandwich	
Coefficient	$p = 0.900$	$p = 0.950$	$p = 0.900$	$p = 0.950$
	$n = 20$		$n = 20$	
β_0	0.917	0.946	0.941	0.970
β_1	0.790	0.854	0.838	0.897
β_2	0.905	0.941	0.907	0.948
β_3	0.896	0.937	0.910	0.952
β_4	0.918	0.950	0.933	0.963
	$n = 40$		$n = 40$	
β_0	0.935	0.966	0.956	0.982
β_1	0.804	0.873	0.867	0.921
β_2	0.918	0.955	0.920	0.960
β_3	0.913	0.951	0.922	0.963
β_4	0.933	0.969	0.948	0.977
	$n = 60$		$n = 60$	
β_0	0.938	0.972	0.949	0.976
β_1	0.805	0.874	0.872	0.925
β_2	0.918	0.958	0.915	0.961
β_3	0.916	0.956	0.917	0.960
β_4	0.935	0.974	0.931	0.970
	$n = 80$		$n = 80$	
β_0	0.938	0.973	0.943	0.981
β_1	0.890	0.877	0.876	0.935
β_2	0.924	0.959	0.915	0.959
β_3	0.917	0.958	0.911	0.957
β_4	0.935	0.970	0.935	0.963
	$n = 100$		$n = 100$	
β_0	0.934	0.974	0.938	0.977
β_1	0.813	0.882	0.888	0.942
β_2	0.924	0.963	0.912	0.960
β_3	0.916	0.959	0.911	0.958
β_4	0.930	0.971	0.921	0.971
	$n = 1000$		$n = 1000$	
β_0	0.910	0.960	0.911	0.960
β_1	0.807	0.883	0.898	0.950
β_2	0.921	0.963	0.906	0.954
β_3	0.915	0.960	0.904	0.954
β_4	0.910	0.960	0.906	0.956

Carroll and Kauermann (2002) point out that the sandwich variance estimator is more variable than its naive counterpart. They provide a useful investigation of the model properties that affect this variability and make several suggestions for altering the usual calculation of test statistics and/or degrees of freedom.

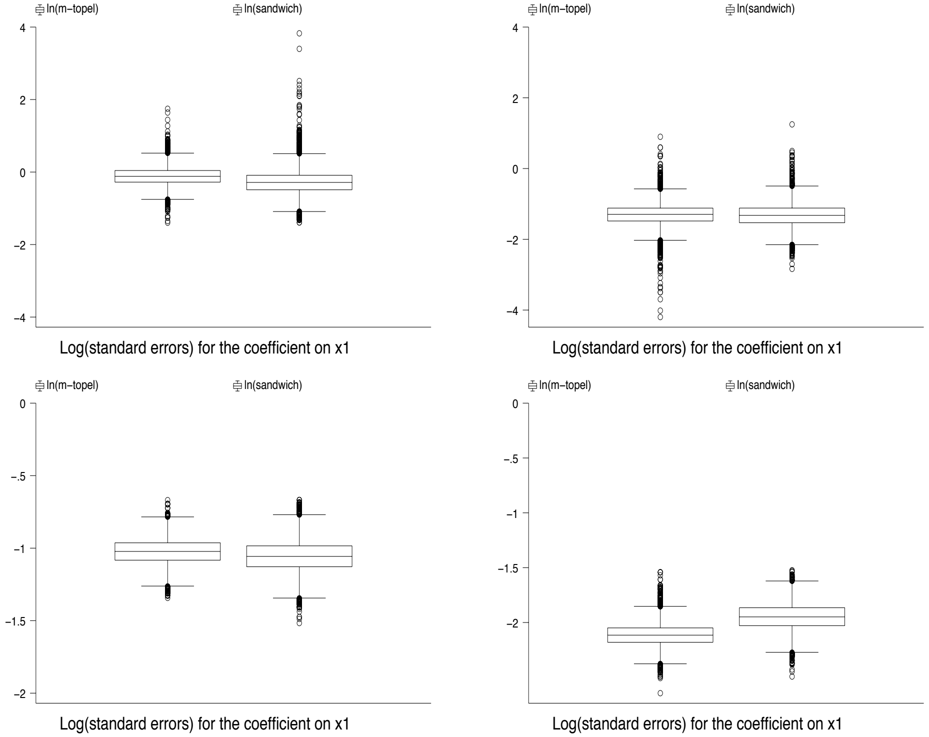

Figure 1: In each of the four graphics, the left boxplot is for the the log of the 10,000 Murphy–Topel standard errors, and the right boxplot is for the log of the 10,000 sandwich standard errors. The left column shows results for data generated under the correct models, and the right column shows results for data generated under the heteroskedastic models. The top row shows results for sample size equal to 20, and the bottom row shows results for sample size equal to 100. We note that the sandwich estimator is more variable than the Murphy–Topel estimator for small samples, and that the difference in variability appears to diminish as the sample sizes increase.

7 Summary

The sandwich estimate of variance for two-stage maximum likelihood models has a form similar to the familiar Murphy–Topel estimator. The use of the $\mathbf{C}^*$ matrix in the sandwich estimate of variance requires computing second derivatives of the second model's log likelihood. This is computationally more difficult than the corresponding matrix in the Murphy–Topel estimator. However, we gain three advantages with the sandwich estimator. First, we have an estimator with the same robustness properties of all sandwich estimates of variance. Second, we can easily calculate the full sandwich estimate of variance for the complete parameter vector using equation 4. Third, the full sandwich variance matrix admits Wald tests of hypotheses across the two models, which is not possible using the Murphy–Topel estimator.

8 Acknowledgments

The author wishes to thank an anonymous referee for bringing attention to the Stefanski and Boos reference. This referee also made other valuable suggestions to expand the discussion to include comparison of our results with those results for GEE.

9 References

Binder, D. A. 1983. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51: 279–292.

Carroll, R. J. and G. Kauermann. 2002. The sandwich variance estimator: Efficiency properties and coverage probability of confidence intervals. *Journal of the American Statistical Association* submitted.

Greene, W. 2000. *Econometric Analysis*. 4th ed. Upper Saddle River, NJ: Prentice–Hall.

Huber, P. J. 1967. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 221–233. Berkeley, CA: University of California Press.

Liang, K.-Y. and S. L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73: 13–22.

Murphy, K. M. and R. H. Topel. 1985. Estimation and inference in two-step econometric models. *Journal of Business and Economic Statistics* 3(4): 370–379.

Stefanski, L. A. and D. D. Boos. 2002. The calculus of M-estimation. *The American Statistician* 56(1): 29–38.

About the Author

James W. Hardin is a lecturer in the Department of Statistics and a Research Scientist in the Academy for Advanced Telecommunication and Learning Technologies at Texas A&M University.