# The Stata Journal



H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142; FAX 979-845-3144
jnewton@stata-journal.com

Nicholas J. Cox
Department of Geography
University of Durham
South Road
Durham City DH1 3LE UK
n.j.cox@stata-journal.com





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

The Stata Journal (2003)
**3**, Number 4, pp. 342–350

# Variance estimation for the instrumental variables approach to measurement error in generalized linear models

James W. Hardin
Arnold School of Public Health
University of South Carolina
Columbia, SC 29208

Raymond J. Carroll
Department of Statistics MS-3143
Texas A&M University
College Station, TX 77843-3143

**Abstract.** This paper derives and gives explicit formulas for a derived sandwich variance estimate. This variance estimate is appropriate for generalized linear additive measurement error models fitted using instrumental variables. We also generalize the known results for linear regression. As such, this article explains the theoretical justification for the sandwich estimate of variance utilized in the software for measurement error developed under the Small Business Innovation Research Grant (SBIR) by StataCorp. The results admit estimation of variance matrices for measurement error models where there is an instrument for the unknown covariate.

**Keywords:** st0048, sandwich estimate of variance, measurement error, White's estimator, robust variance, generalized linear models, instrumental variables

## 1 Introduction

This is the second of five papers describing software for fitting measurement error models. Software production by StataCorp was funded by a National Institutes of Health (NIH) Small Business Innovation Research Grant (SBIR). The goal of the work described in the grant is the production of software to analyze statistical models where one or more covariates are measured with error. The software development includes two major features. The first development feature is the development of Stata programs to support communication to dynamically linked user-written computer code. StataCorp was responsible for this development and support for user-written code in the C/C++ programming languages was added to Stata version 8. Stata refers to compiled user-written code as plugins and maintains documentation on their web site at *http://www.stata.com/support/plugins*.

In this paper, we investigate the derivation of the sandwich estimate of variance for a generalized linear additive measurement error model fitted using instrumental variables (IV). The general idea in this context has been proposed, for example, in Carroll, Ruppert, and Stefanski (1995), and explicitly described for two-stage models

The project described was supported by Grant Number R44 RR12435 from the National Institutes of Health, National Center for Research Resources. The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Research Resources.

 

in Hardin (2002); though this latter reference does not explicitly mention measurement error as motivation. We also include a restatement of the findings for the simple case of linear regression with instrumental variables and discuss the implications.

In section 2, we summarize the methods for fitting GLMs. In section 3, we state the instrumental variables technique and provide explicit formulas for the sandwich estimate of variance for GLMs with instrumental variables following the arguments of Hardin (2002) and Murphy and Topel (1985). In section 4, we summarize the simplification of our general formula for the case of linear regression. Section 5 presents an example application, and we present a summary in section 6.

We investigate the case of fitting a generalized linear model (GLM) where one or more of the covariates are measured with error. We have instruments available that are uncorrelated with the error term of the model but correlated with the covariates measured with error. Using the instrumental variables approach, we have an estimating equation that includes the instrumental variables regressions in the estimation of the GLM.

The usual variance estimate of the coefficient vector from the GLM does not take into account the estimation of the instrumental variables regressions. We must derive a variance estimate that takes into account these regressions as well as the GLM estimation. As we show, the sandwich estimate of variance for estimating equations is a valid estimator. This variance estimate defines estimating equations that include all of the parameters—the parameters from the instrumental variables regressions and the parameters from the GLM.

Hardin and Carroll (2003) present an overview of both generalized linear models and the notational conventions that we employ in the present discussion. A second source of information in the case of univariate measurement error is the `gllamm` command; see Rabe-Hesketh, Skrondal, and Pickles (2003).

## 2 Estimation

Hardin and Carroll (2003) reference sources for estimation algorithms for the classical GLM. In those references, it is assumed that the list of covariates is measured without error. Recall that we use $\mathbf{X}_z$ for covariates measured without error, $\mathbf{X}_w$ for covariates measured with error, $\mathbf{X}_s$ for the instruments of $\mathbf{X}_w$, and $\mathbf{X}_r$ for the augmented matrix of exogenous variables $[\mathbf{X}_z\ \mathbf{X}_s]$.

We begin with an $n \times p$ matrix of covariates measured without error given by the augmented matrix $\mathbf{X} = (\mathbf{X}_z\ \ \mathbf{X}_u)$, where $\mathbf{X}_u$ is unobserved, and $\mathbf{X}_w = \mathbf{X}_u$ plus measurement error. $\mathbf{X}_z$ is an $n \times p_z$ matrix of covariates measured without error (possibly including a constant), and $\mathbf{X}_w$ is an $n \times p_x$ $(p_z + p_x = p)$ matrix of covariates with classical measurement error that estimates $\mathbf{X}_u$. We wish to employ an $n \times p_s$ (where $p_s \geq p_x$) matrix of instruments $\mathbf{X}_s$ for $\mathbf{X}_w$.

Greene (2003) discusses instrumental variables and provides a clear presentation to supplement the following concise description. The method of instrumental variables assumes that some subset $\mathbf{X}_w$ of the independent variables is correlated with the error term in the model. In addition, we have a matrix $\mathbf{X}_s$ of independent variables that are correlated with $\mathbf{X}_w$. Given $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{X}_w$ are uncorrelated. Using these relationships, we can construct an approximately consistent estimator that may be succinctly described. One performs a regression for each of the independent variables (each column) of $\mathbf{X}_w$ on the instruments and the independent variables not correlated with the error term $(\mathbf{X}_z \;\; \mathbf{X}_s)$. Predicted values are obtained from each regression and substituted for the associated column of $\mathbf{X}_w$ in the analysis of the GLM of interest. This construction provides an approximately consistent estimator of the coefficients in the GLM (it is consistent in the linear case). Section 3 addresses forming a valid variance estimator for the coefficients in the GLM.

If we have access to the complete matrix of covariates measured without error (if we know $\mathbf{X}_u$ instead of using instruments $\mathbf{X}_s$), we denote the linear predictor $\eta = \sum_{j=1}^{p} [\mathbf{X}_z \;\; \mathbf{X}_u]_j \beta_j$, and the associated derivative as $\partial\eta/\partial\beta_j = [\mathbf{X}_z \;\; \mathbf{X}_u]_j$. The estimating equation for $\boldsymbol{\beta}$ is then $\sum_{i=1}^{n} (y_i - \mu_i)/\mathrm{V}(\mu_i)(\partial\mu/\partial\eta)_i [\mathbf{X}_z \;\; \mathbf{X}_u]_{ji}$.

However, since we do not observe $\mathbf{X}_u$, we use $\mathbf{X}_r = (\mathbf{X}_z \;\; \mathbf{X}_s)$ to denote the augmented matrix of exogenous variables, which combines the covariates measured without error and the instruments. We regress each of the $p_x$ components (each of the $j$ columns) of $\mathbf{X}_w$ on $\mathbf{X}_r$ to obtain an estimated $(p_z + p_s) \times 1$ coefficient vector $\boldsymbol{\gamma}_j$ for $j = 1, \ldots, p_x$.

The complete coefficient vector $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}} \;\; \boldsymbol{\gamma}_2^{\mathrm{T}} \;\; \cdots \;\; \boldsymbol{\gamma}_{p_x}^{\mathrm{T}})^{\mathrm{T}}$ for these IV regressions is described by the estimating equation

$$\Psi_2 = \begin{bmatrix} \mathbf{X}_r{}^{\mathrm{T}} (\mathbf{X}_{w\,1} - \mathbf{X}_r \gamma_1) \\ \mathbf{X}_r{}^{\mathrm{T}} (\mathbf{X}_{w\,2} - \mathbf{X}_r \gamma_2) \\ \vdots \\ \mathbf{X}_r{}^{\mathrm{T}} (\mathbf{X}_{w\,p_x} - \mathbf{X}_r \gamma_{p_x}) \end{bmatrix} \tag{1}$$

We may then form an $n \times p_x$ matrix $\widehat{\mathbf{X}}_u = [\mathbf{X}_r \widehat{\gamma}_1 \;\; \mathbf{X}_r \widehat{\gamma}_2 \;\; \cdots \;\; \mathbf{X}_r \widehat{\gamma}_{p_x}]$ of predicted values from the instrumental variables regressions to estimate $\mathbf{X}_u$. Combining the (predicted value) regressors with the independent variables measured without error, we may write the estimating equation of the GLM as

$$\Psi_1 = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathrm{V}(\mu_i)} \left( \frac{\partial\mu}{\partial\eta} \right)_i [\mathbf{X}_z \;\; \mathbf{X}_r \boldsymbol{\gamma}]_{ji} \tag{2}$$

where

$$[\mathbf{X}_z \;\; \mathbf{X}_r \boldsymbol{\gamma}]_{ji} = \begin{cases} (\mathbf{X}_z)_{ji} & \text{if } 1 \le j \le p_z \\ (\mathbf{X}_r \gamma_{j-p_z})_i & \text{if } p_z < j \le p_z + p_x \end{cases}$$

Operationally, we obtain a two-stage estimate $\widehat{\boldsymbol{\beta}}$ by first replacing each unknown covariate $\mathbf{X}_{w\,i}$ for $i = 1, \ldots, p_x$ with the fitted values of the regression of $\mathbf{X}_{w\,i}$ on $(\mathbf{X}_z \;\; \mathbf{X}_s)$. We call the resulting $n \times p_x$ matrix of fitted values $\widehat{\mathbf{X}}_u$. We then perform

the (second stage) usual GLM fit of $\mathbf{Y}$ on $(\mathbf{X}_z \;\; \widehat{\mathbf{X}}_u)$. This GLM fit provides an estimate of $\boldsymbol{\beta}$. Our goal is to construct a valid variance estimate of $\boldsymbol{\beta}$.

## 3 The sandwich variance estimate

The variance matrix estimate from the IRLS algorithm used to compute the (second stage) GLM fit assumes that $\widehat{\mathbf{X}}_u = \mathbf{X}_u$. This is clearly unacceptable. An alternative approach is accounted for in Murphy and Topel (1985); this variance estimator is included in the software developed as part of this small business innovation research project. We derive an appropriate sandwich estimate of variance that takes into account the estimation of $\mathbf{X}_u$. Excellent reviews of the sandwich variance estimator and its properties are given in Carroll and Kauermann (2001) and Carroll et al. (1998), while the classic references are Eicker (1963), Eicker (1967), Huber (1967), and White (1980). Application of the sandwich estimate of variance for panel data is discussed in Xie, Simpson, and Carroll (2000). Our application is a special case of Hardin (2002), which describes the general derivation of asymptotic and sandwich variance estimators for two-stage models. In fact, the instrumental variables approach to measurement error is a special case of two-stage estimation.

The two-stage derivation resulting in an estimate for $\boldsymbol{\beta}$ involves estimating the combined parameter vector given by $\Theta = (\boldsymbol{\beta}^{\mathrm{T}} \;\; \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. These results are from the estimating equations given in equations 1 and 2. While we are ultimately interested in $\boldsymbol{\beta}$, we must consider all of the parameters in forming the associated variance matrix.

Our goal is the construction of the sandwich estimate of variance given by $\mathbf{V}_{\mathrm{S}} = \mathbf{A}^{-1} \;\mathbf{B}\; \mathbf{A}^{-\mathrm{T}}$. We form the variance matrix, $\mathbf{A}$, for $\Theta$ by obtaining the necessary derivatives. The variance matrix $\mathbf{A}$ (information matrix) may be calculated numerically, but the analytic derivatives are not difficult and are given by

$$\mathbf{A} = \left[ \begin{array}{cc} -\dfrac{\partial \Psi_1}{\partial \boldsymbol{\beta}}_{(p_z+p_x)\times(p_z+p_x)} & -\dfrac{\partial \Psi_1}{\partial \boldsymbol{\gamma}}_{(p_z+p_x)\times\{p_x(p_z+p_s)\}} \\ -\dfrac{\partial \Psi_2}{\partial \boldsymbol{\beta}}_{\{p_x(p_s+p_z)\}\times(p_z+p_x)} & -\dfrac{\partial \Psi_2}{\partial \boldsymbol{\gamma}}_{\{p_x(p_z+p_s)\}\times\{p_x(p_z+p_s)\}} \end{array} \right]^{-1}$$

where

$$\begin{aligned} \frac{\partial \Psi_1}{\partial \beta_k} \;=\; & -\sum_{i=1}^{n} \left[ \frac{1}{\mathrm{V}(\mu_i)} \left( \frac{\partial \mu}{\partial \eta} \right)_i^2 \right. \\ & \left. - (\mu_i - y_i) \left\{ \frac{1}{\mathrm{V}(\mu_i)^2} \left( \frac{\partial \mu}{\partial \eta} \right)_i^2 \frac{\partial \mathrm{V}(\mu_i)}{\partial \mu} - \frac{1}{\mathrm{V}(\mu_i)} \left( \frac{\partial^2 \mu}{\partial \eta^2} \right)_i \right\} \right] \\ & [\mathbf{X}_{\mathrm{Z}} \;\; \mathbf{X}_{\mathrm{R}}\boldsymbol{\gamma}]_{ji} [\mathbf{X}_{\mathrm{Z}} \;\; \mathbf{X}_{\mathrm{R}}\boldsymbol{\gamma}]_{ki} \\ & j = 1, \ldots, p_z + p_x; \;\; k = 1, \ldots, p_z + p_x \\ & \text{yields a matrix of size } (p_z + p_x) \times (p_z + p_x) \end{aligned} \tag{3}$$

$$\begin{aligned}
\frac{\partial \Psi_1}{\partial \gamma_{\ell_k}} &= -\sum_{i=1}^{n}\left[\frac{1}{\mathrm{V}(\mu_i)}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\right. \\
&\quad \left.-(\mu_i - y_i)\left\{\frac{1}{\mathrm{V}(\mu_i)^2}\left(\frac{\partial \mu}{\partial \eta}\right)_i^2\frac{\partial \mathrm{V}(\mu_i)}{\partial \mu} - \frac{1}{\mathrm{V}(\mu_i)}\left(\frac{\partial^2 \mu}{\partial \eta^2}\right)_i\right\}\right] \\
&\quad [\mathbf{X}_{\mathrm{Z}}\ \ \mathbf{X}_{\mathrm{R}}\boldsymbol{\gamma}]_{ji}\mathbf{X}_{\mathrm{R}\,ki}\beta_{\ell+p_z} \\
&\quad j = 1,\ldots,p_z+p_x;\ k = 1,\ldots,p_z+p_s;\ \ell = 1,\ldots,p_x \\
&\quad \text{yields a matrix of size } (p_z+p_x)\times\{p_x(p_z+p_s)\} \qquad (4) \\
\frac{\partial \Psi_2}{\partial \beta_k} &= 0 \\
&\quad k = 1,\ldots,p_z+p_x \\
&\quad \text{yields a matrix of size } \{p_x(p_z+p_s)\}\times(p_z+p_x) \qquad (5) \\
\frac{\partial \Psi_2}{\partial \gamma_{\ell_k}} &= -\sum_{i=1}^{n}\mathbf{X}_{\mathrm{R}\,ji}\mathbf{X}_{\mathrm{R}\,ki} \\
&\quad j = 1,\ldots,p_z+p_s;\ k = 1,\ldots,p_z+p_s;\ \ell = 1,\ldots,p_x \\
&\quad \text{yields a block diagonal matrix of size } \{p_x(p_z+p_s)\}\times\{p_x(p_z+p_s)\} \\
&\quad \text{where each block matrix is of size } (p_z+p_s)\times(p_z+p_s) \qquad (6)
\end{aligned}$$

The elements of the variance matrix are formed from the definitions above. Mapping these equations is accomplished by defining the matrix $\mathbf{A}$ using

$$[\mathbf{X}_{\mathrm{Z}}\ \ \mathbf{X}_{\mathrm{R}}\boldsymbol{\gamma}]_{ji} = \begin{cases} \mathbf{X}_{\mathrm{Z}\,ji} & \text{if } 1 \le j \le p_z \\ \left(\mathbf{X}_{\mathrm{R}}\gamma_{(j-p_z)}\right)_i & \text{if } p_z < j \le p_z + p_x \end{cases}$$

in which we apply the notation

$$\begin{aligned}
\mathbf{X}_{\mathrm{Z}\,ji} &= i\text{th observation of the } j\text{th column of } \mathbf{X}_{\mathrm{Z}} \\
\mathbf{X}_{\mathrm{R}\,ji} &= i\text{th observation of the } j\text{th column of } \mathbf{X}_{\mathrm{R}} \\
\boldsymbol{\gamma}_{j-p_z} &= \text{IV coefficient vector from regressing } \mathbf{X}_{\mathrm{W}(j-p_z)} \text{ on } \mathbf{X}_{\mathrm{R}} \\
\beta_{\ell+p_z} &= (\ell+p_z)\text{th coefficient of } \boldsymbol{\beta} \\
\left(\mathbf{X}_{\mathrm{R}}\boldsymbol{\gamma}_{(j-p_z)}\right)_i &= i\text{th observation of (the predicted values from) } \mathbf{X}_{\mathrm{R}}\gamma_{j-p_z} \\
\gamma_{\ell_k} &= k\text{th coefficient of the } \ell\text{th IV coefficient vector}
\end{aligned}$$

Equation 3 defines the $(j,k)$ elements of $\mathbf{A}$ for $j,k = 1,\ldots,p_z+p_x$. Equation 4 defines the $(j,k)$ elements of $\mathbf{A}$ for $j = 1,\ldots,p_z+p_x$ and $k = p_z+p_x+\ell$, where $\ell = 1,\ldots,p_x(p_z+p_s)$. This notation addresses the cross derivatives for all of the $(p_z+p_s)\times 1$ coefficient vectors $\boldsymbol{\gamma}_m$ for $m = 1,\ldots,p_x$. Equation 5 calculates the $(j,k)$ elements of $\mathbf{A}$ for $j = p_z+p_x+\ell$, where $\ell = 1,\ldots,p_x(p_z+p_s)$ and $k = 1,\ldots,p_z+p_x$. Equation 6 defines the $(j,k)$ elements of $\mathbf{A}$ for $j,k = p_z+p_x+\ell$, where $\ell = 1,\ldots,p_x(p_z+p_s)$. These are the covariances of all of the $(p_z+p_s)\times 1$ coefficient vectors $\boldsymbol{\gamma}_m$ for $m = 1,\ldots,p_x$.

Noting that $\Psi_i = [\Psi_{1i}\ \Psi_{2i}]$, the middle of the sandwich estimate of variance is then given by $\mathbf{B} = \sum_{i=1}^{n} \Psi_i \Psi_i^{\mathrm{T}}$. A suitable estimate may be formed using

$$\widehat{\Psi}_{1i} = \left\{ \frac{y_i - \mu_i}{\mathrm{V}(\mu_i)} \left( \frac{\partial \mu}{\partial \eta} \right)_i [\mathbf{X}_{\mathrm{Z}}\ \ \mathbf{X}_{\mathrm{R}} \widehat{\boldsymbol{\gamma}}]_{ji} \right\}_{(p_z+p_x)\times 1}^{j=1,\ldots,(p_z+p_x)}$$

$$\widehat{\Psi}_{2i} = \begin{pmatrix} \left[ (\mathbf{X}_{\mathrm{W}\,1} - \mathbf{X}_{\mathrm{R}} \widehat{\gamma}_{1j})\, \mathbf{X}_{\mathrm{R}\,ji} \right]_{(p_z+p_s)\times 1}^{j=1,\ldots,(p_z+p_s)} \\ \left[ (\mathbf{X}_{\mathrm{W}\,2} - \mathbf{X}_{\mathrm{R}} \widehat{\gamma}_{2j})\, \mathbf{X}_{\mathrm{R}\,ji} \right]_{(p_z+p_s)\times 1}^{j=1,\ldots,(p_z+p_s)} \\ \vdots \\ \left[ (\mathbf{X}_{\mathrm{W}\,p_x} - \mathbf{X}_{\mathrm{R}} \widehat{\gamma}_{p_x j})\, \mathbf{X}_{\mathrm{R}\,ji} \right]_{(p_z+p_s)\times 1}^{j=1,\ldots,(p_z+p_s)} \end{pmatrix}_{\{p_x(p_z+p_s)\}\times 1}$$

The sandwich estimate of variance for $\boldsymbol{\beta}$ is then the upper $(p_z + p_x) \times (p_z + p_x)$ matrix of $\mathbf{V}_{\mathrm{S}}$ obtained from the derived estimates of $\mathbf{A}$ and $\mathbf{B}$.

The preceding description is implemented in the `qvf` command provided as part of the newly developed software.

## 4 The linear-regression case

In the linear-regression case (identity link and Gaussian variance), the derivation of the sandwich estimate of variance greatly simplifies. This simplification is shown in detail in White (1982). Here, we present the results of the simplification and summarize the implications.

The naive (model-based) covariance matrix utilizes the fact that $\mathrm{V}(\mathbf{Y} - \mathbf{X}_{\mathrm{W}}\boldsymbol{\beta}) = \mathrm{V}\{Y_i - X_{\mathrm{W}\,i}\boldsymbol{\beta}\}\ \mathbf{I} = \sigma^2\mathbf{I}$ where $\sigma^2$ is the mean square of $Y_i - X_{\mathrm{W}\,i}\widehat{\boldsymbol{\beta}}$. Thus, $\mathrm{V}(\widehat{\boldsymbol{\beta}}) \approx \sigma^2(\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\mathbf{X}_{\mathrm{P}})^{-1}$, where $\mathbf{X}_{\mathrm{P}} = \mathbf{X}_{\mathrm{S}} \left(\mathbf{X}_{\mathrm{S}}^{\mathrm{T}}\mathbf{X}_{\mathrm{S}}\right)^{-1} \mathbf{X}_{\mathrm{S}}^{\mathrm{T}}\mathbf{X}_{\mathrm{W}}$ is a matrix of the predicted values from using the instruments $\mathbf{X}_{\mathrm{S}}$. Therefore, the correct asymptotic variance can be obtained simply by performing a standard linear regression of $\mathbf{Y}$ on $\mathbf{X}_{\mathrm{P}}$.

The sandwich estimate of variance is then clearly given by

$$\left(\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\mathbf{X}_{\mathrm{P}}\right)^{-1} \mathbf{X}_{\mathrm{P}}^{\mathrm{T}} \left(\mathbf{Y} - \mathbf{X}_{\mathrm{W}}\boldsymbol{\beta}\right) \left(\mathbf{Y} - \mathbf{X}_{\mathrm{W}}\boldsymbol{\beta}\right)^{\mathrm{T}} \mathbf{X}_{\mathrm{P}} \left(\mathbf{X}_{\mathrm{P}}^{\mathrm{T}}\mathbf{X}_{\mathrm{P}}\right)^{-1}$$

such that the usual sandwich estimate from the linear regression of $\mathbf{Y}$ on $\mathbf{X}_{\mathrm{P}}$ is correct.

Thus, for the standard linear regression case, we may obtain both a model-based and a sandwich estimate of variance by considering only the second stage regression. These simplifications are not true in the general case of a GLM and may be discerned from the two-stage regression formula given in StataCorp (2003).

# 5 Example application

Carroll, Ruppert, and Stefanski (1995) (hereafter, CRS) present several examples using data from the Framingham Heart Study (see chapters 4 and 5). This dataset consists of three measurements taken two years apart on 1,615 men aged 31–65. The outcome variable is `chd`, indicating the presence of coronary heart disease within an eight-year period following the third set of collected measurements. The predictors include `age`, the patient's age in years; `smoke`, an indicator of whether the patient smokes; `sbp`, the systolic blood pressure; and `chol`, a categorical (three categories) of cholesterol.

In the examples, CRS uses the transformed predictor `lbsp` given by log(`sbp`−50). For illustration, we use the indicator variable `smoke` as an instrument for the transformed systolic blood pressure `lbsp` where the systolic blood pressure is the mean of two measurements for the patient by different technicians. Our use of the `smoke` variable as an instrument is for illustrative purposes, as this is not a good instrument. The purpose in choosing `smoke` as the instrument is to magnify the comparative results of the naive (GLM variance estimate ignoring the estimation from the IV regressions) and sandwich estimate of variance. Table 1 lists the coefficient estimates and standard errors.

Table 1: Instrumental variables logistic regression results. Naive standard errors are the result of ignoring the estimation from the instruments. Sandwich standard errors are the result from the sandwich estimate of variance presented in the previous section.

| Variable | Coeff | Naive Std. Error | Sandwich Std. Error |
|---|---|---|---|
| lbsp | −20.6183 | 9.3636 | 14.3143 |
| age | 0.1914 | 0.0590 | 0.0884 |
| chol | 0.0171 | 0.0046 | 0.0069 |
| constant | 74.7127 | 37.3035 | 57.1535 |

This example demonstrates the difference between the two approaches. The naive standard errors are calculated from the variance matrix resulting from the GLM fit using the fitted values for the `lbsp` variable. This naive variance estimate does not take into account the estimation of `lbsp` and is invalid.

The associated software developed for measurement error analysis includes support for various models and variance estimates. These estimates include the sandwich variance estimate described here, as well as bootstrap, asymptotic (model-based), and Murphy–Topel (for instrumental variables approach to measurement error).

# 6 Summary

This paper presents a sandwich estimate of variance for generalized linear models with instrumental variables. The presentation includes detailed formulas for the calculation of the variance estimate. These formulas admit the calculation of a valid variance

estimate for the coefficient vector for any regression model within the generalized linear models framework. These results may be extended to quasi-likelihood cases as well.

As Binder (1992) points out, the bread of the sandwich estimate of variance is not, in general, symmetric. We have asymmetry for the case of GLMs with instrumental variables due to the augmented matrices of cross derivatives. We have also shown that in the special case of linear regression, our derivation simplifies such that we may use the results of the second stage regression without modification.

## 7 References

Binder, D. A. 1992. Fitting Cox's proportional hazards models from survey data. *Biometrika* 79(1): 139–147.

Carroll, R. J. and G. Kauermann. 2001. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 96: 1387–1396.

Carroll, R. J., D. Ruppert, and L. A. Stefanski. 1995. *Measurement Error in Nonlinear Models*. London: Chapman & Hall.

Carroll, R. J., S. Wang, D. G. Simpson, A. J. Stromberg, and D. Ruppert. 1998. The sandwich (robust covariance matrix) estimator. Unpublished Technical Report. *http://stat.tamu.edu/ftp/pub/rjcarroll/sandwich.pdf*.

Eicker, F. 1963. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics* 34(2): 447–456.

—. 1967. Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, 59–82. Berkeley, CA: University of California Press.

Greene, W. H. 2003. *Econometric Analysis*. 5th ed. Upper Saddle River, NJ: Prentice–Hall.

Hardin, J. W. 2002. The robust variance estimator for two-stage models. *Stata Journal* 2(3): 253–265.

Hardin, J. W. and R. J. Carroll. 2003. Measurement error, GLMs, and notational conventions. *Stata Journal* 3(4): 328–340.

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.

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.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles. 2003. Maximum likelihood estimation of generalized linear models with covariate measurement error. *Stata Journal* 3(4): 385–410.

StataCorp. 2003. *The Stata Reference Manual*. Version 8 ed. College Station, TX: Stata Press.

White, H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48(4): 817–838.

—. 1982. Instrumental variables regression with independent observations. *Econometrica* 50(2): 483–499.

Xie, M., D. G. Simpson, and R. J. Carroll. 2000. Random effects in censored ordinal regression: Latent structure and Bayesian approach. *Biometrics* 56: 376–383.

**About the Authors**

James W. Hardin (jhardin@gwm.sc.edu), is an Associate Research Professor, Department of Epidemiology and Biostatistics, and a Research Scientist, Center for Health Services and Policy Research, Arnold School of Public Health, Carolina Plaza Suite 1120, University of South Carolina, Columbia, SC 29208, USA.

Raymond J. Carroll (carroll@stat.tamu.edu) is a Distinguished Professor, Department of Statistics, MS 3143, Texas A&M University, College Station, TX 77843-3143, USA.

Research by StataCorp was supported by the National Institutes of Health (NIH) Small Business Innovation Research Grant (SBIR) (2R44RR12435-02).