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 StataCorp LP.

The Stata Journal (2003)
3, Number 4, pp. 329–341

Measurement error, GLMs, and notational conventions

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 introduces additive measurement error in a generalized linear-model context. We discuss the types of measurement error along with their effects on fitted models. In addition, we present the notational conventions to be used in this and the accompanying papers.

Keywords: st0047, generalized linear models, transportability, measurement error

1 Introduction

This is the first 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*.

The software for measurement error analysis, the second development feature, was co-developed by StataCorp in conjunction with Raymond J. Carroll, James W. Hardin, Henrik Schmiediche, Tamara Stoner, and H. Joseph Newton. Professor Carroll was the design expert for the functionality of the software and provided formulas for known results, as well as deriving asymptotic standard error formula for estimators—work that had not previously appeared in the literature. Hardin, Schmiediche, and Stoner worked on the C/C++ programming languages and ado-code development, while Newton assisted in both design and certification of the resulting software.

This work was premiered in a workshop held at the 2003 Boston Stata Users' Group meeting in March of 2003.

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.

We investigate the nature and effects of additive measurement error on fitted generalized linear models (GLMs). Measurement error is defined and categorized along with discussion of the transportability of models. We include a restatement of the findings for the simple case of linear regression and discuss the implications. This restatement serves as a particularly illustrative vehicle for identifying the important issues. We follow closely the arguments and developments of Carroll, Ruppert, and Stefanski (1995) but deviate from the notational conventions in that reference.

We investigate analysis strategies of fitting a GLM where one or more of the covariates are measured with error. It is important to note that there are several avenues for identifying and estimating this measurement error and we must be able to estimate this source of variation to proceed with model fitting.

The main purpose of this initial paper is to provide a single source for identification of the jargon and notation used to describe various types of measurement error data. While measurement error itself is fairly easy to describe, the manner in which we can estimate the necessary quantities that allow proper analysis incorporates several different collections of measurements. Readers will find this initial article valuable for reminders of these details as well as for identifying the role various notational quantities play. This article is *software free* in the sense that we will not discuss any of the specific commands that were written in support of this grant. Later articles will provide detailed descriptions as well as syntax specifications for the developed programs.

2 Notational conventions

The usual notation in the measurement error literature involves naming individual matrices: $\mathbf{Z}$ for covariates measured without error, $\mathbf{W}$ for error-prone observed covariates, $\mathbf{S}$ for the instruments of $\mathbf{W}$, and $\mathbf{R}$ for the augmented matrix of exogenous variables $[\mathbf{Z}\,\mathbf{S}]$. To avoid confusion with the measurement error notation and the usual notation associated with GLMs (the $\mathbf{W}$ weight matrix in the IRLS algorithm), we denote the usual measurement error notational conventions as subscripts of $\mathbf{X}$:

Table 1: Notational conventions

n	the number of observations in the sample
p	the number of covariates in the analysis of interest $(p = p_z + p_x)$
$\mathbf{Y}$	the response variable in the analysis; $(n \times 1)$
$\mathbf{X}_z$	the covariates measured without error; $(n \times p_z)$
$\mathbf{X}_u$	the (unknown) covariates measured with error; $(n \times p_x)$
$\mathbf{X}_w$	the error-prone observed covariates for $\mathbf{X}_u$; $(n \times p_w), p_w = kp_x, k \geq 1$
$\mathbf{X}_s$	the instruments for $\mathbf{X}_w$; $(n \times p_s)$
$\mathbf{X}_\mathbf{T}$	the extra exogenous variables for the instruments $\mathbf{X}_s$; $(n \times p_t)$
$\mathbf{X}_r$	the augmented matrix of exogenous variables $[\mathbf{X}_z\ \mathbf{X}_\mathbf{T}]$; $\{n \times (p_z + p_t)\}$

Table 1 lists the notational conventions. Table 2 illustrates the allowable data organizations for the accompanying software.

Table 2: Allowable data organizations

Code	p_z	p_x	p_w	p_s	p_t	Restrictions
1	p_z	0	0	0	0	$p_z \geq 1$
2	p_z	p_x	p_x	0	0	$p_z \geq 0, p_x \geq 1$
3	p_z	p_x	kp_x	0	0	$p_z \geq 0, p_x \geq 1; k \geq 2$
4	p_z	p_x	0	p_x	p_t	$p_z \geq 0, p_x \geq 1; p_t \geq p_x$

Data organizations described by code 1 do not need special measurement error software. There are no covariates measured with error, and analyses proceed in a customary fashion. Data organizations described by code 2 require user specification of the measurement error variance since there are no data replicates from which to estimate the error variance. Data organizations described by code 3 will admit user specification of the measurement error variance, or we can estimate this variance from the replicate data. Note that if there is more than one covariate measured with error ($p_x > 1$), then each one of the covariates measured with error must have the same number and missingness of replicate measurements k ($p_w = kp_x$) per observation. Data organizations described by code 4 must have an instrument for each unknown covariate. In addition, there must be a list of extra exogenous variables with at least as many covariates as there are unknown covariates. Estimation may then use instrumental variables methods to generate the unknown covariates prior to estimating the model of interest. Standard errors may then be calculated using standard two-stage estimation methods; see Hardin (2002), for example.

Data organizations described by code 3 deserve further discussion. First, Stata users will note that this organization requires a wide format instead of a long format. Second, since the measurement error variance assumes that the order of the replicate error-prone measurements is important, the missingness pattern (per observation) must be the same for each replicate group.

Assume that the variables `w11` and `w12` are replicate measures for the unknown variable `xu1`. Likewise, assume that the variables `w21` and `w22` are replicate measures for the unknown variable `xu2`. For an observation i to be included in the analysis, the missingness must match for `w11` and `w21` as well as for `w12` and `w22`. Put another way, for a given observation to be included in the analysis, the missingness pattern of the ordered groups (`w11`,`w12`) and (`w21`,`w22`) must be the same; see table 3. Extensions to three or more replicates proceed under these same restrictions.

(*Continued on next page*)

Table 3: Missingness patterns for observation i

xu1		xu2		Included
w11	w12	w21	w22	in analysis?
observed	observed	observed	observed	yes
missing	observed	observed	observed	no
observed	missing	observed	observed	no
observed	observed	missing	observed	no
observed	observed	observed	missing	no
missing	missing	observed	observed	no
missing	observed	missing	observed	yes
missing	observed	observed	missing	no
observed	missing	missing	observed	no
observed	missing	observed	missing	yes
observed	observed	missing	missing	no
missing	missing	missing	observed	no
missing	missing	observed	missing	no
missing	observed	missing	missing	no
observed	missing	missing	missing	no
missing	missing	missing	missing	no

The method of regression calibration—Hardin, Schmiediche, and Carroll (2003a)—and the method of simulation extrapolation—Hardin, Schmiediche, and Carroll (2003b)—can be used to fit generalized linear models for data that are organized by codes 2 and 3. The `qvf` command, developed as part of the accompanying software, is a fast version of the [R] **glm** command that allows instrumental variables specifications so that it can be used for data that are organized by code 4; since it is a replacement for `glm`, the `qvf` command may also be used, in many instances, for fitting models to data organized by code 1.

3 Measurement error in simple linear regression

Many textbooks provide a cursory description of measurement error in the context of linear regression. Typically, this description focuses on simple linear regression and illustrates the effect of measurement error as a bias of the estimated slope toward zero. This type of bias is known as attenuation.

As Carroll, Ruppert, and Stefanski (1995) point out, this conclusion must be qualified for multiple regression, as the effect of measurement error depends on the relationship between the error-prone measurement, the unknown covariate we are trying to measure, and the remaining covariates measured without error. For nonlinear regression models, the measurement error effects will depend on whether there is one or more covariates, whether we have univariate or multivariate measurement error, and whether the error-prone proxy is biased.

We begin with an illustration of the effects of measurement error for the case of simple linear regression to facilitate the present discussion. The model is given by $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_u + \epsilon$, where the true covariate $\mathbf{X}_u$ has mean μ_x and variance σ_x^2, the error term ϵ is independent of the covariate, and the error term has mean zero and variance σ_ϵ^2. Under simple additive measurement error, we observe only $\mathbf{X}_w$ where $\mathbf{X}_w = \mathbf{X}_u + \mathbf{U}$. $\mathbf{U}$ has mean zero, variance σ_u^2, and is independent of the covariate $\mathbf{X}_u$.

Were we simply to regress $\mathbf{Y}$ on $\mathbf{X}_w$, we would not obtain a consistent estimate of β_1 but instead obtain an estimate of $\beta_{1*} = \lambda\beta_1$ where

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} < 1$$

is the attenuation factor. Stata users also call this factor the reliability ratio; see [R] **eivreg**.

For illustration, consider the following do-file:

```
clear
set seed 12345
if "‘1’" != "" {
        set seed ‘1’
}

set obs 10

gen double xu  = invnorm(uniform())
summ xu
replace xu = xu/sqrt(r(Var))
label var xu " "
* sigma^2_x = 1.00

gen double err = invnorm(uniform())
summ err
replace err = err/sqrt(r(Var))*.5
* sigma^2_e = 0.25

gen double u = invnorm(uniform())
summ u
replace u = u/sqrt(r(Var))
* sigma^2_u = 1.00

* generate error-prone version of "unknown" xu
gen double xw = xu + u
label var xw " "

* beta0 = 0
* beta1 = 1
gen y  = 0 + 1*xu + err
label var y " "

regress y xu
mat btrue = e(b)
gen double ytrue = btrue[1,2] + btrue[1,1]*xw
label var ytrue " "
```

```
regress y xw
predict double yhat
label var yhat " "

twoway (scatter y xu, m(+)) (scatter y xw, m(Oh)) (line ytrue xw, clp(dot)) /*
*/ (line yhat xw), /*
        */ legend( label(1 "True values") label(2 "Error-prone values") /*
        */ label(3 "True least-squares line") /*
        */ label(4 "Error-prone least-squares line") )
```

Running this do-file produces the illustration of the issues at hand. The graph resulting from the execution of this do-file is presented in figure 1.

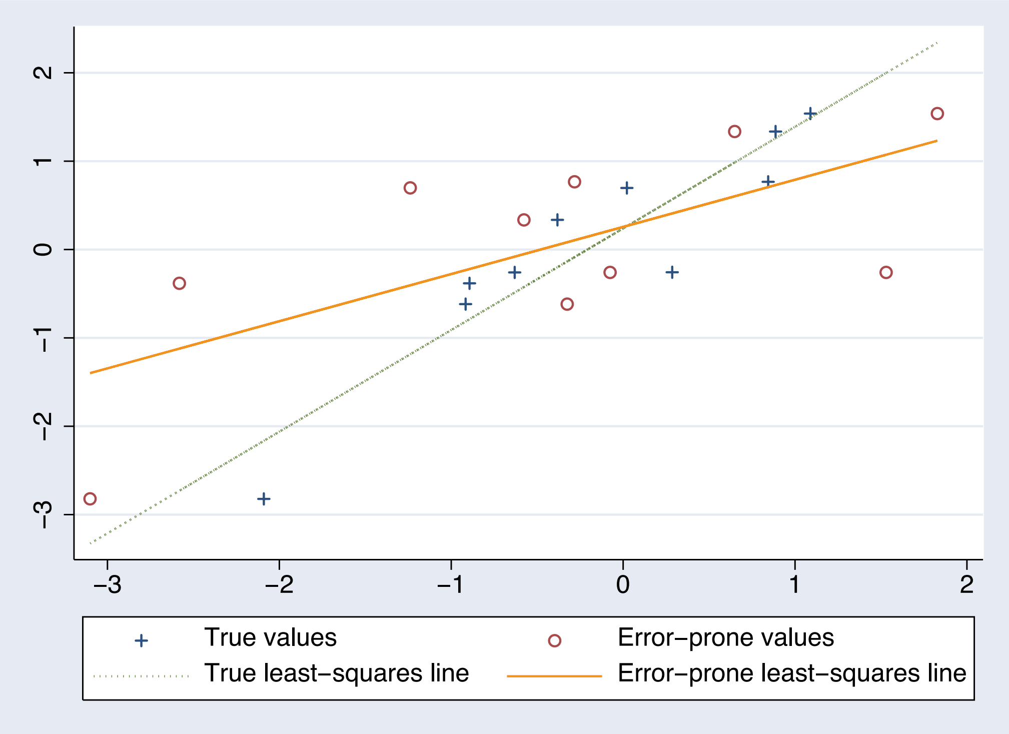

Figure 1: Illustration of the additive measurement error model. For these data, we have $\sigma_x^2 = \sigma_u^2 = 1$ and $\sigma_\epsilon^2 = .25$. The least-squares slope using the true covariate would be $\widehat{\beta}_1 = 1.1502$, and the least-squares slope using the error-prone measurements is $\widehat{\beta}_{1*} = 0.5337 \approx 1.1502\{\sigma_x^2/(\sigma_x^2 + \sigma_u^2)\} = 0.5751$. For each true value represented by the plus sign, there is an observed circle at the same height, but it is measured with horizontal error (error in the covariate). The measurement error tends to widen the scatterplot, thus pulling the fitted regression slope toward a horizontal line.

We included a check for a first argument in the preceding do-file. Interested readers can run the do-file with a single argument specifying a random number seed to see slightly different results.

It is a common misconception that the presence of measurement error always attenuates (biases the estimated coefficient toward the null) the estimated slope. In fact, this conclusion depends on the assumption of additive measurement error. Carroll, Ruppert, and Stefanski (1995) point out circumstances and error models where this conclusion is incorrect.

4 Estimating the measurement error variance

In assessing the error process, we might use external data sources. In such cases, the *transportability* of the results is in question. In other words, if the external data sources are not produced under a similar measurement error process, then using that external data source for estimation of measurement error quantities in the current dataset is not appropriate. The use of external sources for estimating the error process carries with it the assumption that the error process in the data of interest is the same as for the external source.

An estimate of the measurement error variance is needed to adjust for the bias in estimated coefficients in a measurement error analysis. There are several techniques for obtaining this estimated variance. In the following subsections, we describe each of these techniques. From one perspective, the models we seek to fit are special cases of missing-data problems. We seek to include a covariate in our model for which we do not have observations, or for which we do not have complete observations. Using other sources of information, we must build up an unbiased representation of this unknown covariate to use in model fitting. Standard errors must then be adjusted to account for this extra source of variability.

There are, thus, two steps to addressing measurement error. The first step is the construction of a proxy for the missing covariate. In general, we may have no observations or some subset of observations for the covariate of interest. We must have some other source of information to substitute for the missing observations for this covariate. This information could come from unbiased replicate error-prone measurements. This is useful as it provides an observation-wise point estimate of the missing covariate as well as a source for estimating the error variance. Alternatively, we might have an unbiased instrument for the missing covariate along with additional covariates to use in an instrumental variables regression approach. A third approach utilizes an error-prone measurement error proxy with an externally determined estimate of the measurement error.

4.1 User-specified variance

The easiest adjustment is provided by a user-specified estimate of the measurement error variance. This specification may be the result of some external study though one must be concerned with the transportability of the estimate. The analysis may concern a subset of a much larger collection of replicate measures from which a sound estimate of the measurement error variance may be obtained. Finally, the analysis may include covariates that have a long history of estimation in other related studies from which the measurement error variance is largely believed to be "known".

For any of these cases, the analyst may choose to specify an estimate of measurement error variance despite the presence of information in the data that would otherwise allow calculation of an estimate. In effect, this specification of the error variance is exactly what is behind the Stata `eivreg` command except that the command refers to the reliability (a function of the error variance).

4.2 Replicate measures

Replicate measures of the unknown covariate provide a very good means for estimating the measurement error covariance. Carroll, Ruppert, and Stefanski (1995) point out that certain types of covariates can be subject to drift such that later measurements from proxy variables tend to exhibit a shift in the mean across the observations. When $\mathbf{X}_u$ has multiple columns, it is assumed that for each observation, there are the same number of replicate proxy measures obtained. As such, the order of these measures is important in regard to drift. The measurement error covariance is calculated by summing a function of

$$\Delta_{ij}\left(\mathbf{X}_{w\,ij}-\overline{\mathbf{X}}_{w\,i\cdot}\right)\left(\mathbf{X}_{w\,ij}-\overline{\mathbf{X}}_{w\,i\cdot}\right)^{\mathrm{T}}$$

where i is the observation and j is the replicate number; Δ_{ij} is an indicator that replicate j is nonmissing (for each column of $\mathbf{X}_w$) for observation i. It may be difficult to see initially, but this formula is the reason for the results listed in table 3.

4.3 Instrumental variables

We may have an unbiased measure for the unknown covariate along with a list of exogenous variables. Using the covariates measured without error $\mathbf{X}_z$ along with the other exogenous variables $\mathbf{X}_{\mathbf{T}}$, we can address the bias in the estimated coefficients through standard instrumental variables techniques. First, the instrument $\mathbf{X}_s$ is regressed on the augmented matrix of covariates $\mathbf{X}_r = [\mathbf{X}_z\ \mathbf{X}_{\mathbf{T}}]$. Fitted values from this regression are then used in place of the instruments in a standard analysis.

A variance estimate of the resulting coefficients in the standard analysis may be calculated using standard instrumental variables techniques. These variance estimators include the Murphy–Topel, sandwich, and bootstrap estimators. All are included in the accompanying software.

5 Generalized linear models

This section provides a short review of generalized linear models without reference to measurement error. The section is meant as a short introduction to the full complement of models that are supported in the accompanying software wherein we describe the analyses of interest apart from the considerations we must make with measurement error. All of the techniques that we describe for addressing the bias induced by measurement error are addressed for the class of generalized linear models.

Nelder and Wedderburn (1972) introduced theory and an algorithm appropriate for obtaining maximum likelihood estimates where the response follows a distribution in the exponential family. This idea is extended in Wedderburn (1974), where it is shown that one need only assume moment properties (without relying on an underlying distribution), thus implying a quasi-likelihood model.

The treatment of generalized linear models (GLMs) has received attention in many articles and books such as McCullagh and Nelder (1989), Lindsey (1997), and Hardin and Hilbe (2001). These sources provide the detailed derivations of appropriate estimation algorithms.

The derivation of the iteratively reweighted least squares (IRLS) algorithm appropriate for fitting GLMs usually begins with the likelihood for an exponential family. The useful conclusion of the derivation is that a new estimate of the coefficient vector may be obtained via weighted ordinary least squares. Estimates are obtained by iterating the process to convergence.

The usual method for deriving classical likelihood-based regression models is to choose a distribution that matches the outcome variable. Once a distribution is chosen, a linear combination of covariates with unknown coefficients is substituted for the expected value of the distribution. In some cases, we also reparameterize this linear combination to enforce any range restrictions on the location parameter imposed by the distribution.

This mechanical description is an easy method for deriving specific regression models (linear regression, Poisson regression, logistic regression, etc.). However, it turns out that we can start by specifying an entire family of distributions for the outcome rather than one specific distribution. In so doing, we not only derive a general model, but we also derive a very efficient estimation algorithm.

Generalized linear models are the result of assuming the exponential family of distributions for the outcome variable. Continuing in the usual mechanical fashion of deriving a model, there are two helpful assumptions we can make. First, we can generalize the parameterization of the linear combination of covariates and associated coefficients as a *link function*, and second, we can derive a simple algorithm for estimation by substituting the Fisher scoring matrix for the Hessian. The resulting estimation method is known as iteratively reweighted least squares. Since all of the members of the exponential family have variance that is proportional to the mean, we can fully describe any member through a link and variance function. A final generalization from Wedderburn

(1974) allows us to choose these two functions without constraining them to originate from the same family member.

This section does not contain enough material to fully appreciate the study of generalized linear models. See the classic reference McCullagh and Nelder (1989) or Hardin and Hilbe (2001), which is specific to using Stata. The advantage of discussing measurement error models as extensions of generalized linear models is that the discussion then admits a wide range of models without introducing more theoretical justifications.

The exponential family of distributions includes a location parameter θ, a scale parameter $a(\phi)$, and a normalizing term $c(y,\phi)$. The associated probability density function is then given by

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \tag{1}$$

where $\mathrm{E}(y) = b'(\theta) = \mu$ and $\mathrm{V}(y) = b''(\theta)a(\phi)$. The normalizing term is independent of the parameter vector θ and ensures that the density integrates to one. Range restrictions on the parameter vector are addressed after the estimating equation has been constructed. The variance function is in terms of the expected value of the distribution and is also a function of the (possibly unknown) scale parameter $a(\phi)$. The density for a single observation is

$$f(y_i;\theta,\phi) = \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}$$

and the joint density for n independent outcomes is the product of the individual outcome densities:

$$f(y_1,\ldots,y_n;\theta,\phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}$$

The likelihood is simply a restatement of the joint density, where the outcomes are taken as given, and we model the parameters as unknown:

$$L(\theta,\phi|y_1,\ldots,y_n) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i,\phi)\right\}$$

In many model-building derivations, covariates are introduced into consideration in terms of the expected value of the outcome. In this case, we will wait to introduce the covariates into the estimating equation. We introduce a subscript for θ in anticipation of introducing the covariates.

The log likelihood for the exponential family is

$$\mathcal{L}(\theta,\phi|y_1,\ldots,y_n) = \sum_{i=1}^{n} \left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right\}$$

The goal is to obtain a maximum likelihood estimator for θ. Since our focus is only on θ, we derive a LIML estimating equation where we treat the dispersion parameter $a(\phi)$ as ancillary.

We have from basic principles that $\mathrm{E}(\partial\mathcal{L}/\partial\theta) = 0$. The LIML estimating equation is then $\Psi(\Theta) = \partial\mathcal{L}/\partial\theta = \mathbf{0}$ where $\partial\mathcal{L}/\partial\theta = \sum_i (y_i - b'(\theta))/a(\phi)$. Utilizing the GLM result that in canonical form $b'(\theta) = \mu$, we write

$$\frac{\partial\mathcal{L}}{\partial\theta} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)}$$

substituting μ for the expected value.

Since our goal is to introduce covariates that model the outcome, we include a subscript on μ allowing the mean to reflect a dependence on a linear combination of the covariates and their associated coefficients. Later, we consider the effects on fitted models when one or more of the covariates are measured with error. We use the chain rule to obtain a more useful form of the estimating equation; this is the usual presentation in discussions of the GLM:

$$\begin{aligned}
\frac{\partial\mathcal{L}}{\partial\beta_j} &= \left(\frac{\partial\mathcal{L}}{\partial\theta}\right)\left(\frac{\partial\theta}{\partial\mu}\right)\left(\frac{\partial\mu}{\partial\eta}\right)\left(\frac{\partial\eta}{\partial\beta_j}\right) \\
&= \sum_{i=1}^{n}\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)\left(\frac{1}{\mathrm{V}(\mu_i)}\right)\left(\frac{\partial\mu}{\partial\eta}\right)_i (x_{ji}) \\
&= \sum_{i=1}^{n}\frac{y_i - \mu_i}{a(\phi)\mathrm{V}(\mu_i)}\left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji}
\end{aligned}$$

This presentation utilizes $\mathcal{L}$ for the quasilikelihood, $g()$ for the link function relating the expected outcome to the linear predictor (the sum of the products of the covariates with their associated coefficient), and $\mathrm{V}()$ for the variance function in terms of the expected outcome. The derivation of the estimating equation along with the derivation and justification of the iteratively reweighted least squares algorithm is given in detail in Hardin and Hilbe (2001).

In terms of measurement error models, we must consider the case for which one or more of the $\mathbf{X}$ variables are measured with error. We differentiate the covariates using the notation presented earlier, $\mathbf{X}_z$ for covariates measured without error and $\mathbf{X}_w$ for covariates measured with error, and note that the first subscript j of the covariate in the estimating equation is in $\{1, \ldots, p_z + p_x\}$ referring to each covariate in the model. The benefit of utilizing the generalized linear model framework is that we simultaneously produce valid results for the entire collection of models represented by the exponential family.

Additional information can be found in the documentation for the `glm` command as well as in a forthcoming article on the `qvf` command produced under the associated grant. What is important about the derivation in terms of the exponential family of distributions is that we have an estimation equation for which we can discuss measurement error. Discussion of measurement error in this framework allows us to derive results for the entire family. Thus, support for the generalized linear models framework ultimately allows us to simultaneously address needs in linear regression, Poisson, logistic, probit, negative binomial, and many other models.

The software developed in support of this work is written in terms of generalized linear models. Each new command that we introduce is written in terms of a specific algorithm for addressing measurement error and applies these algorithms in terms of GLMs so that users may apply the specific technique to a wide collection of models.

6 Summary

Analyses with measurement error data must address the bias induced through the extra variance component. We have provided a notational convention that allows us to describe a collection of various data organizations. Depending on the amount of information contained in the data, analyses may be adjusted for bias and standard errors calculated. Various methods for this adjustment are described in Carroll, Ruppert, and Stefanski (1995).

7 References

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

Hardin, J. and J. Hilbe. 2001. *Generalized Linear Models and Extensions*. College Station, TX: Stata Press.

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

Hardin, J. W., H. Schmiediche, and R. J. Carroll. 2003a. The regression-calibration method for fitting generalized linear models with additive measurement error. *Stata Journal* 3(4): 360–371.

—. 2003b. The simulation extrapolation method for fitting generalized linear models with additive measurement error. *Stata Journal* 3(4): 372–384.

Lindsey, J. K. 1997. *Applying generalized linear models*. Berlin: Springer-Verlag.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2d ed. London: Chapman & Hall.

Nelder, J. A. and R. W. M. Wedderburn. 1972. Generalized linear models. *Journal of the Royal Statistical Society: Series A* 135(3): 370–384.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* 61(3): 439–447.

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).