

Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples

Robert M. Bell and Daniel F. McCaffrey¹

Abstract

Linearization (or Taylor series) methods are widely used to estimate standard errors for the coefficients of linear regression models fit to multi-stage samples. When the number of primary sampling units (PSUs) is large, linearization can produce accurate standard errors under quite general conditions. However, when the number of PSUs is small or a coefficient depends primarily on data from a small number of PSUs, linearization estimators can have large negative bias. In this paper, we characterize features of the design matrix that produce large bias in linearization standard errors for linear regression coefficients. We then propose a new method, bias reduced linearization (BRL), based on residuals adjusted to better approximate the covariance of the true errors. When the errors are i.i.d., the BRL estimator is unbiased for the variance. Furthermore, a simulation study shows that BRL can greatly reduce the bias even if the errors are not i.i.d. We also propose using a Satterthwaite approximation to determine the degrees of freedom of the reference distribution for tests and confidence intervals about linear combinations of coefficients based on the BRL estimator. We demonstrate that the jackknife estimator also tends to be biased in situations where linearization is biased. However, the jackknife's bias tends to be positive. Our bias reduced linearization estimator can be viewed as a compromise between the traditional linearization and jackknife estimators.

Key Words: Complex samples; Linearization; Jackknife; Satterthwaite approximation; Degrees of Freedom.

1. Introduction

Regression analysis of multi-stage samples has become very common in recent years (for example, Ellickson and McGuigan 2000; Shapiro, Morton, McCaffrey, Senterfitt, Fleishman, Perlman, Athey, Keesey, Goldman, Berry and Bozette 1999; Goldstein 1991; Landis, Lepkowski, Ekland and Stehouver 1982). Although hierarchical models (Bryk and Raudenbush 1992; Gelman, Carlin, Stern and Rubin 1995, Chapter 13) allow analysis of both fixed and random effects, many analysts prefer the simplicity of standard regression models when random effects are not of direct interest. Standard regression estimators produce unbiased parameter estimates that can be efficient, but the default standard error estimators do not account for the sample design, resulting in inconsistent standard errors (Kish 1965; Skinner 1989a). Various methods produce consistent standard error estimates applicable when the number of primary sampling units (PSUs) is sufficiently large. These include sample reuse methods such as the jackknife, bootstrap and balance repeated replication as well as linearization (or Taylor series) methods.

Linearization (Skinner 1989b) is a nonparametric method for estimating the standard errors of design-based statistics such as means and ratios as well as coefficients from linear and nonlinear regression models. By nonparametric, we mean that linearization does not rest on any assumptions about the within-PSU error structure, such as an assumption of constant intra-cluster correlation. When the number of PSUs can be considered large, linearization produces

consistent standard errors in the presence of multiple features of complex sample designs-stratification, multi-stage sampling, and sampling weights-as well as heteroskedastic errors (Fuller 1975). Because of these desirable properties and its increased availability in software such as SUDAAN, Stata, and SAS Version 8.0 (Shah, Barnwell, and Bieler 1997; StataCorp. 1999; SAS Institute, Inc. 1999), linearization has become a common method for estimating standard errors and confidence intervals and for conducting statistical tests on data from complex sample designs (for example, Ellickson and McGuigan 2000; Shapiro *et al.* 1999; Rust and Rao 1996). Linearization has also been proposed for estimating standard errors from Generalized Estimating Equations (GEE) fit to multi-stage data (Zeger and Liang 1986).

However, the linearization method has limitations. When the number of primary sampling units is small, standard error estimates can be severely biased low, they can have large coefficients of variation, and the standard degrees of freedom may be far too liberal (Kott 1994; Murray, Hannan, Wolfinger, Baker and Dwyer 1998). Consequently, standard linearization inference for coefficients based mainly on data from a small number of PSUs may produce confidence intervals that are too narrow and tests with Type I error rates that are substantially higher than their nominal values. Sample reuse methods like the jackknife have similar limitations.

In this paper, we characterize the design factors (*i.e.*, the distribution of explanatory variables within and between PSUs) that produce large bias in linearization and jackknife

1. Robert M. Bell, Statistics Research Department, AT&T Labs-Research, Room C211, 180 Park Ave., Florham Park, NJ 07932; Daniel F. McCaffrey, Statistics Group, RAND, 201 North Craig Street, Suite 202, Pittsburgh, PA 15213-1516.

standard errors for linear regression coefficients and demonstrate that the problem can persist even when the number of PSUs is quite large. We then propose an alternative to the standard linearization estimator that is unbiased for independent, identically distributed (i.i.d.) errors and tends to greatly reduce bias otherwise. We also present approximate degrees of freedom for use with tests and confidence intervals based on our variance estimator. Simulation results show improved small sample properties of our alternative estimator and test compared with those of more traditional methods. Finally, we present an example of our methods using data from a national experiment evaluating care for depression.

2. Bias of the Linearization Method

For simplicity, we restrict consideration in the body of this paper to unweighted linear regression for two-stage nonstratified samples. Extensions to weighted estimators and stratified samples are presented in McCaffrey, Bell and Botts (2001) and discussed further in section 8.

Let n equal the number of PSUs and m_i equal the number of final sampling units from the i^{th} PSU, for $i = 1, \dots, n$. The overall sample size is $M = \sum_i m_i$. We assume that $y_{ij} = \beta'x_{ij} + \varepsilon_{ij}$, where ε has mean 0 and covariance matrix \mathbf{V} , and where y_{ij} , x_{ij} , and ε_{ij} all refer to the j^{th} observation from the i^{th} PSU. We drop the standard OLS assumption of i.i.d. errors, assuming only that errors from distinct PSUs are uncorrelated. Specifically, we assume that \mathbf{V} is block diagonal, with $m_i \times m_i$ blocks \mathbf{V}_i for $i = 1, \dots, n$. In addition to the notation of this model, throughout the paper, we let \mathbf{I} denote an $M \times M$ identity matrix and \mathbf{I}_i equal an $m_i \times m_i$ identity matrix.

Let $\hat{\beta}$ denote the estimated coefficients of the linear regression model. To simplify presentation, we generally discuss a linear combination of the regression coefficients, $l'\hat{\beta}$, for an arbitrary column vector l . For the special case where one element of $l = 1$ and the rest are 0, $l'\hat{\beta}$ equals a single estimated coefficient. If errors are uncorrelated across PSUs, the variance of $l'\hat{\beta}$ is

$$\text{Var}(l'\hat{\beta}) = l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}'_i \mathbf{V}_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l, \quad (1)$$

where \mathbf{X} and \mathbf{X}_i are the design matrices for the entire sample and for PSU i , respectively.

The standard linearization estimator of the variance of $l'\hat{\beta}$ is given by:

$$v_L = l'(\mathbf{X}'\mathbf{X})^{-1} \left(c \sum_{i=1}^n \mathbf{X}'_i \mathbf{r}_i \mathbf{r}'_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \quad (2)$$

where \mathbf{r}_i is the vector of residuals for the i^{th} PSU. Comparison of (1) and (2) shows that linearization simply involves estimating \mathbf{V}_i by a constant c times the outer product of the residuals. The constant c is typically set equal to $n/(n-1)$, the value used by SUDAAN and the Stata svy procedures (Shah, Barnwell, and Bieler 1997; StataCorp. 1999). For GEE procedures, Zeger and Liang (1986) set $c = 1$.

Under fairly general conditions, nv_L converges in probability to the variance of the asymptotic distribution of $\sqrt{n}(l'\hat{\beta} - l'\beta)$ and the relative bias of v_L is $O(1/n)$ as the number of PSUs gets large (Fuller 1975; Kott 1994). To demonstrate convergence for the bias of v_L , Kott (1994) assumes that the number of observations from every PSU is bounded and that elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are bounded by B/n for a constant B . These assumptions effectively ensure that the influence of any PSU on the final estimate diminishes as the number of PSUs grows. Convergence of the bias of v_L holds for heteroskedastic data from stratified samples with unequal sampling weights and arbitrary correlation structure within PSUs. Unfortunately, consistency does not guarantee good properties for small to moderate numbers of PSUs.

Theorem 1. When $\mathbf{V} = \sigma^2\mathbf{I}$ and $c = n/(n-1)$, $E(v_L) \leq \text{Var}(l'\hat{\beta})$ with equality if and only if $l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i \mathbf{X}_i$ is constant across i .

Proof. Without loss of generality, we assume that $\sigma^2 = 1$ so that $\mathbf{V} = \mathbf{I}$. The residual vector \mathbf{r} can be written as $(\mathbf{I} - \mathbf{H})\varepsilon$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the hat or projection matrix for \mathbf{X} . Thus, we have that $\mathbf{r}_i = (\mathbf{I} - \mathbf{H})_i \varepsilon$, where $(\mathbf{I} - \mathbf{H})_i$ contains the m_i rows of $(\mathbf{I} - \mathbf{H})$ for the i^{th} PSU. Consequently,

$$\begin{aligned} E(v_L) &= \left(\frac{n}{n-1} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \\ &\left(\sum_{i=1}^n \mathbf{X}'_i (\mathbf{I} - \mathbf{H})_i E(\varepsilon\varepsilon') (\mathbf{I} - \mathbf{H})'_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \\ &= \left(\frac{n}{n-1} \right) l'(\mathbf{X}'\mathbf{X})^{-1} \\ &\sum_{i=1}^n \left(\mathbf{X}'_i \mathbf{X}_i - \mathbf{X}'_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \quad (3) \end{aligned}$$

because $E(\varepsilon\varepsilon') = \mathbf{I}$ and $(\mathbf{I} - \mathbf{H})_i (\mathbf{I} - \mathbf{H})'_i = (\mathbf{I}_i - \mathbf{H}_{ii})$ for $\mathbf{H}_{ii} = \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_i$. Let $\mathbf{D}_i = \mathbf{X}'_i \mathbf{X}_i - (1/n) (\mathbf{X}'\mathbf{X})$. Note that $\sum_i \mathbf{D}_i = \sum_i \mathbf{X}'_i \mathbf{X}_i - \mathbf{X}'\mathbf{X} = 0$. Thus,

$$\begin{aligned}
 E(v_L) &= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
 &= \sum_{i=1}^n \left(\mathbf{X}'_i \mathbf{X}_i - (1/n) \mathbf{X}' \mathbf{X} + \mathbf{D}_i \right) (\mathbf{X}' \mathbf{X})^{-1} [(1/n) \mathbf{X}' \mathbf{X} + \mathbf{D}_i] \\
 &\quad (\mathbf{X}' \mathbf{X})^{-1} l \\
 &= \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
 &= \left(\mathbf{X}' \mathbf{X} - (1/n) \mathbf{X}' \mathbf{X} - \sum_{i=1}^n \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \\
 &= l' (\mathbf{X}' \mathbf{X})^{-1} l - \left(\frac{n}{n-1} \right) l' (\mathbf{X}' \mathbf{X})^{-1} \\
 &\quad \left(\sum_{i=1}^n \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{D}_i \right) (\mathbf{X}' \mathbf{X})^{-1} l \\
 &= \text{Var}(l' \hat{\beta}) - \left(\frac{n}{n-1} \right) \left(\sum_{i=1}^n a'_i (\mathbf{X}' \mathbf{X})^{-1} a_i \right) \quad (4)
 \end{aligned}$$

for $a_i = \mathbf{D}_i (\mathbf{X}' \mathbf{X})^{-1} l = [\mathbf{X}'_i \mathbf{X}_i - (1/n) (\mathbf{X}' \mathbf{X})] (\mathbf{X}' \mathbf{X})^{-1} l$. Because $(\mathbf{X}' \mathbf{X})^{-1}$ is positive definite, $E(v_L) \leq \text{Var}(l' \hat{\beta})$ with equality if and only if $a_i \equiv 0$, or equivalently, $\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l$ is constant across the i .

Replication methods do not necessarily avoid the problem of bias for regression variance estimators. A jackknife estimator for multi-stage samples can be derived from the set of pseudo values $\{\tilde{\beta}_{[i]}\}$, estimates of β from data that exclude the i^{th} PSU:

$$v_{JK} = [(n-1)/n] \sum_i l' (\tilde{\beta}_{[i]} - \hat{\beta}) (\tilde{\beta}_{[i]} - \hat{\beta})' l \quad (5)$$

(Cochran 1977; Rust and Rao 1996). If $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i , then

$$\begin{aligned}
 v_{JK} &= [(n-1)/n] l' (\mathbf{X}' \mathbf{X})^{-1} \sum_i \mathbf{X}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \\
 &\quad \mathbf{r}_i \mathbf{r}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l, \quad (6)
 \end{aligned}$$

which follows from the updating formula $(\mathbf{X}' \mathbf{X} - \mathbf{X}'_i \mathbf{X}_i)^{-1} = (\mathbf{X}' \mathbf{X})^{-1} + (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i (\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1}$ (Cook and Weisberg 1982; Bell and McCaffrey 2002, page 34). Some authors (Efron and Tibshirani 1993) suggest an alternative jackknife estimator with $\hat{\beta}$ replaced by the mean of the $\tilde{\beta}_{[i]}$'s in (5). These two methods provide very similar estimates in our simulations, so we discuss only the version based on (5) in what follows.

Theorem 2. When $\mathbf{V} = \sigma^2 \mathbf{I}$ and $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ exists for all i , then $E(v_{JK}) \geq \text{Var}(l' \hat{\beta})$ with equality if and only if $l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{X}_i$ is constant across i (proof in appendix).

The following example shows that the conditions for linearization and the jackknife estimators to be unbiased are very restrictive even for simple linear regression.

Example 1. Consider simple linear regression. We have that

$$\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l = \frac{m_i}{Ms^2} \begin{bmatrix} 1 & \bar{x}_i \\ \bar{x}_i & s_i^2 + \bar{x}_i^2 \end{bmatrix} \begin{bmatrix} s^2 + \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} l$$

where s^2 and $\{s_i^2\}$ are ML estimates for the overall and within-PSU variances of x , with divisors M and $\{m_i\}$, respectively. So we have

$$\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}' \mathbf{X})^{-1} l = \frac{m_i}{Ms^2} \begin{bmatrix} s^2 + \bar{x}^2 - \bar{x}_i \bar{x} & \bar{x}_i - \bar{x} \\ (s^2 + \bar{x}^2) \bar{x}_i - (s_i^2 + \bar{x}_i^2) \bar{x} & s_i^2 + \bar{x}_i^2 - \bar{x}_i \bar{x} \end{bmatrix} l.$$

To have v_L and v_{JK} unbiased for the slope, *i.e.*, for $l' = (0, 1)$, we must have that $m_i(\bar{x}_i - \bar{x})$ and $m_i(s_i^2 + \bar{x}_i^2 - \bar{x}_i \bar{x})$ are both constant across i . The former implies that $\bar{x}_i \equiv \bar{x}$, and together they imply that $m_i s_i^2 = \sum_j (x_{ij} - \bar{x})^2$ is constant. Note that m_i need not be constant. These two conditions are not sufficient to guarantee unbiasedness for $l' = (1, 0)$, however. Additional algebra shows that the bias in the linearization estimator for the variance of the slope equals

$$-\frac{n}{(n-1)M^3 s^4} \left\{ \sum_{i=1}^n [m_i (\bar{x}_i - \bar{x})]^2 + \sum_{i=1}^n \left[\sum_{j=1}^{m_i} (x_{ij} - \bar{x})^2 - \bar{m} s^2 \right]^2 \right\}.$$

Consequently, the bias includes a part that is proportional to the weighted variance of the PSU means of x and another that is proportional to the variance of the within-PSU sums of squares.

The example shows that when the errors are i.i.d., v_L is unbiased only under very restrictive conditions. When $\mathbf{V} \neq \mathbf{I}$, Theorems 1 and 2 do not hold, and the bias in v_L can even be positive (see Example 2 of Bell and McCaffrey 2002).

In general, v_L tends to have negative bias. The estimator is the sum over PSUs of squares of linear combinations of residuals, $c^{1/2} l' (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}'_i \mathbf{r}_i$. These sums of squares tend to be too small for two reasons: residuals are generally smaller than true errors due to overfitting, and residuals tend to have lower intra-cluster correlation than the errors. The factor $c = n/(n-1)$ corrects completely for these problems only in very restricted circumstances like the conditions in Theorem 1.

The bias of the linearization estimator (or the jackknife) increases with the between-PSU variance of the explanatory variables. Consequently, explanatory variables that are (nearly) constant within PSUs tend to exhibit the largest bias. When there are several such explanatory variables, there can be substantial underestimation of intra-cluster

correlations, leading to large bias in estimated variances for all the corresponding coefficients. Even greater bias potential appears to occur when certain PSUs account for most of the variability in the covariates and have disproportionate impact on the determination of $l' \hat{\beta}$.

3. The Bias Reduced Linearization Method

Phillip Kott has proposed two methods for reducing the bias in linearization. Kott (1994) suggested correcting the bias in v_L by using the residuals and the design matrix to estimate the negative of the bias of v_L by \hat{R} ($\hat{R} > 0$, typically) and setting $v_{K94} = v_L / (1 - \hat{R}/v_L)$. Kott suggested the estimator v_{K94} rather than the more obvious $(v_L + \hat{R})$ as *ad hoc* compensation for the relative bias in \hat{R} as an estimator of the true negative bias, R .

In his 1996 paper, Kott suggests calculating the ratio of $\text{Var}(l' \hat{\beta})$ to $E(v_L)$ under the assumption that $\mathbf{V} = \mathbf{I}$ and adjusting v_L by the ratio. If $\mathbf{V} = \mathbf{I}$ then the resulting estimator v_{K96} will be unbiased.

In the context of generalized estimating equations, Mancl and DeRouen (2001) take a different approach to correcting the bias in the linearization estimator. They suggest adjusting the residuals from each PSU to reduce the bias in $\mathbf{r}_i \mathbf{r}_i'$ as an estimator of \mathbf{V}_i . For the unweighted linear model given in section 2, they approximate $E(\mathbf{r}_i \mathbf{r}_i')$ by $(\mathbf{I}_i - \mathbf{H}_{ii}) \mathbf{V}_i (\mathbf{I}_i - \mathbf{H}_{ii})$ and suggest replacing \mathbf{r}_i by $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{r}_i$ in equation (2). Thus, for unweighted linear models the Mancl and DeRouen estimator equals $n/(n-1) v_{JK}$ and the properties on this estimator follow from the properties of the jackknife estimator.

We present an alternative approach that we first proposed in 1997 (McCaffrey and Bell 1997). The method is also based on replacing \mathbf{r}_i in equation (2) with adjusted residuals of the form $\mathbf{r}_i^* = \mathbf{A}_i \mathbf{r}_i$ intended to act more like the true errors $\boldsymbol{\varepsilon}_i$. Like Kott (1996), we derive an estimator that eliminates the bias of v_L when \mathbf{V} equals \mathbf{U} , a specified block-diagonal covariance matrix, and reduces the bias for other \mathbf{V} . Like Mancl and DeRouen (2001) we adjust the residuals from each PSU. However, using \mathbf{U} we derive an alternative approximation to the $E(\mathbf{r}_i \mathbf{r}_i')$ and our resulting estimator is not proportional to the jackknife but rather can be seen as a compromise between the linearization and jackknife estimators. Our approach is also a generalization of the method of MacKinnon and White (1985), who adjust individual residuals to produce a heteroskedastically-consistent variance estimator (in the sense of White 1980) that is unbiased when the errors are independent and homoskedastic.

Theorem 3. For a specified block-diagonal covariance matrix \mathbf{U} , consider the class of estimators $v_{L^*} = l'(\mathbf{X}'\mathbf{X})^{-1} (\sum_{i=1}^n \mathbf{X}'_i \mathbf{A}_i \mathbf{r}_i \mathbf{r}'_i \mathbf{A}'_i \mathbf{X}_i) (\mathbf{X}'\mathbf{X})^{-1} l$, where \mathbf{A}_i satisfies $\mathbf{A}_i [(\mathbf{I} - \mathbf{H})_i \mathbf{U} (\mathbf{I} - \mathbf{H})'_i] \mathbf{A}'_i = \mathbf{U}_i$ for $i = 1, \dots, n$. If $\mathbf{V} = k\mathbf{U}$ for some scalar k , then $E(v_{L^*}) = \text{Var}(l' \hat{\beta})$.

Proof. The expected value of v_{L^*} is given by

$$\begin{aligned} E(v_{L^*}) &= l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}'_i \mathbf{A}_i (\mathbf{I} - \mathbf{H})_i (k\mathbf{U}) (\mathbf{I} - \mathbf{H})'_i \mathbf{A}'_i \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l \\ &= l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}'_i (k\mathbf{U}_i) \mathbf{X}_i \right) (\mathbf{X}'\mathbf{X})^{-1} l = \text{Var}(l' \hat{\beta}). \end{aligned}$$

Without external evidence to the contrary, an analyst is likely to use a working covariance matrix of the form $\mathbf{U} = \sigma^2 \mathbf{I}$, which simplifies the condition on \mathbf{A}_i to $\mathbf{A}_i (\mathbf{I}_i - \mathbf{H}_{ii}) \mathbf{A}'_i = \mathbf{I}_i$ or

$$\mathbf{A}'_i \mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}. \tag{7}$$

We set $\mathbf{U} = \mathbf{I}$ in what follows.

A solution to equation (7) exists for PSU i whenever $(\mathbf{I}_i - \mathbf{H}_{ii})$ is full rank, which is true if all the eigenvalues of \mathbf{H}_{ii} are strictly less than 1 (the eigenvalues of \mathbf{H}_{ii} are always between 0 and 1). An eigenvalue of \mathbf{H}_{ii} may equal 1 – e.g., when the model includes a dichotomous explanatory variable that is one if and only if an observation falls in the i^{th} PSU.

For $m_i > 1$, \mathbf{A}_i is not unique. If \mathbf{A}_i satisfies $\mathbf{A}'_i \mathbf{A}_i = (\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, then so does $\mathbf{O} \mathbf{A}_i$, for any $m_i \times m_i$ orthogonal matrix \mathbf{O} . If $\mathbf{V} = \sigma^2 \mathbf{I}$, the choice of \mathbf{A}_i is unimportant because any solution to (7) will produce an unbiased variance estimator. However, the resulting estimators are biased when $\mathbf{V} \neq \sigma^2 \mathbf{I}$, and the bias can vary greatly with the choice of \mathbf{A}_i . Heuristically, it makes sense to choose the solution \mathbf{A}_i “closest” to the identity matrix, so as to “mix” the residuals as little as possible. Two promising candidates are the Cholesky decomposition of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, which has all 0’s below the diagonal, and the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$. Let \mathbf{P} be an orthogonal matrix whose columns are the eigenvectors of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ and $\boldsymbol{\Lambda}$ be a diagonal matrix containing the corresponding eigenvalues of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, so that $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} = \mathbf{P} \boldsymbol{\Lambda} \mathbf{P}'$. Then for $\boldsymbol{\Lambda}^{1/2}$ equal to the elementwise square root of $\boldsymbol{\Lambda}$, $\mathbf{P} \boldsymbol{\Lambda}^{1/2} \mathbf{P}'$ is symmetric and solves (7). In contrast, multiplying either of these two solutions by a random orthogonal matrix could greatly distort the residuals.

Among the class of adjusted residuals of the form $\mathbf{A}_i \mathbf{r}_i$ where \mathbf{A}_i satisfies (7), those based on the symmetric square root of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, $\mathbf{r}_i^* = \mathbf{P} \boldsymbol{\Lambda}^{1/2} \mathbf{P}' \mathbf{r}_i$, are “best” in the sense of Theil (1971) – i.e., they minimize the expected sum of the squared differences between the estimated and true i.i.d. errors (see pages 36–37 of Bell and McCaffrey 2002 for details). When there is intra-cluster correlation, simulation results in section 6 suggest that the bias of v_{L^*} based on the

symmetric square root is greatly reduced compared with that of the traditional linearization estimator, v_L . For these reasons, we consider only the symmetric root in the remainder of the paper and refer to the estimator using this root as the biased reduced linearization estimator, v_{BRL} .

As Kott (1994) proved for v_L , if the number of units in every PSU is bounded and the elements of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ are bounded by B/n for some constant B (i.e., $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = O(1/n)$), then the bias in v_{BRL} is $O(n^{-2})$ and the relative bias is $O(1/n)$ (Bell and McCaffrey 2002, page 15).

4. Variance of the Estimators and Testing

We note that v_L , v_{BRL} , and v_{JK} can all be written in the form

$$v^* = c l'(\mathbf{X}'\mathbf{X})^{-1} \sum_i \mathbf{X}'_i \mathbf{A}_i \mathbf{r}_i \mathbf{r}'_i \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l,$$

where: $c = n/(n - 1)$, 1 , or $(n - 1)/n$, respectively, and $\mathbf{A}_i = \mathbf{I}_i$, $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1/2}$, or $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$, respectively. This formulation of the estimators shows that v_{BRL} can be viewed as a compromise between v_L and v_{JK} , chosen to offset their opposing biases.

Theorem 4. Let the error terms be distributed as multivariate normal with mean $\mathbf{0}$ and nonsingular covariance matrix \mathbf{V} . Then for any variance estimator of the form

$$v^* = c l'(\mathbf{X}'\mathbf{X})^{-1} \sum_i \mathbf{X}'_i \mathbf{A}_i \mathbf{r}_i \mathbf{r}'_i \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l,$$

v^* equals the weighted sum of independent χ^2_1 random variables where the weights are the eigenvalues of the $n \times n$ matrix $\mathbf{G} = \{\mathbf{g}'_i \mathbf{V} \mathbf{g}_i\}$, for $\mathbf{g}_i = c^{1/2}(\mathbf{I} - \mathbf{H})'_i \mathbf{A}_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ (proof in appendix).

We can write v_L as a quadratic form $\mathbf{y}'\mathbf{G}^*\mathbf{y}$, where the M -by- M matrix $\mathbf{G}^* = \sum_{i=1}^n \mathbf{g}_i \mathbf{g}'_i$, so that v_L is a weighted sum of independent chi-square random variables with weights equal to the eigenvalues of $\mathbf{G}^*\mathbf{V}$. The proof consists of showing that the nonzero eigenvalues of $\mathbf{G}^*\mathbf{V}$ equal the nonzero eigenvalues of \mathbf{G} .

The mean and variance of v^* are simple functions of the eigenvalues of \mathbf{G} , namely $E(v^*) = \sum_{i=1}^n \lambda_i E(u_i^2) = \sum_{i=1}^n \lambda_i$ and $\text{Var}(v^*) = \sum_{i=1}^n \lambda_i^2 \text{Var}(u_i^2) = \sum_{i=1}^n 2\lambda_i^2$. If $\mathbf{V} = \sigma^2 \mathbf{I}$ and $\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ for $i = 1, \dots, n$ are constant, conditions for v_L and v_{JK} to be unbiased, then Theorem 4 implies that av_L , av_{JK} , and av_{BRL} are all distributed χ^2_{n-1} for $a = (n-1)/\text{Var}(l'\hat{\beta})$ (Bell and McCaffrey 2002, pages 41–42). However, in general, the $\mathbf{X}'_i \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} l$ will not be constant and the squared coefficient of variation will exceed $2/(n - 1)$, the corresponding statistic for a χ^2_{n-1} random variable.

This excess variability is of particular concern when considering reference distributions for testing the null hypothesis that $l'\beta = 0$, with test statistics of the form

$t = l'\hat{\beta}/\sqrt{v^*}$. For v_L , Shah, Holt and Folsom (1977) suggested comparing t to a reference t -distribution with $n - 1$ degrees of freedom, which is now the default in Stata (Stata Corp. 1999), SUDAAN (Shah, Barnwell and Bieler 1997) and SAS (SAS Institute 1999). The choice of $n - 1$ degrees of freedom is motivated by the fact that v_L can be written as the sum of squares of n random variables $c^{1/2} l'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'_i \mathbf{r}_i$. However, because the variance of $(n - 1)v_L/E(v_L)$ tends to be greater than $2(n - 1)$, tests that use a t -distribution with $n - 1$ degrees of freedom would tend to have Type I error rates that exceed the nominal value, even if v_L were unbiased.

Satterthwaite (1946) suggested approximating the distribution of a linear combination of χ^2_1 variables by χ^2_f (up to a constant) where the first two moments of the linear combination match those of χ^2_f . We would approximate v_L , v_{BRL} or v_{JK} by a χ^2_f where $f = 2/cv^2 = (\sum_{i=1}^n \lambda_i)^2 / \sum_{i=1}^n \lambda_i^2$ and the λ_i are the eigenvalues of the corresponding matrix \mathbf{G} . Tests based on reference t -distributions with f degrees of freedom would be expected to provide better Type I error rates than tests based on $n - 1$ degrees of freedom. Rust and Rao (1996) also suggest using a Satterthwaite approximation to estimate the degrees of freedom for the jackknife estimator. They present results for the estimator of a mean, while Theorem 4 extends this approach to testing linear combinations of regression coefficients. Kott (1994, 1996) suggests using the Satterthwaite approximation to estimate the degrees of freedom for tests based on his alternatives to linearization.

The coefficient of variation for any of the nonparametric variance estimators can be very large for certain designs. High variability occurs under the same conditions that v_L and v_{JK} are most biased – when residuals from only a few PSUs effectively determine the final variance estimate. This variability of the estimators is an inherent cost of using nonparametric techniques.

Because the Satterthwaite degrees of freedom f requires specifying the unknown matrix \mathbf{V} , we have investigated two methods for setting \mathbf{V} . The first treats \mathbf{V} as block-diagonal and estimates each block with the outer-product of the residuals for the PSU. Because preliminary simulation results showed that degrees of freedom based on this empirical estimate of \mathbf{V} produced tests that were extremely conservative, we do not present any simulation results for this method. Kott (1994) also found that estimating \mathbf{V} for use in the formula for estimated degrees of freedom proved unsatisfactory. Instead, we used a second method that sets \mathbf{V} identically equal to the identity matrix – i.e., it assumes independent, homoskedastic errors for purposes of determining degrees of freedom.

The distribution of v_{BRL} (and the other variance estimators) tends to be less skewed and have less mass in the lower tail than the distribution of a χ^2_f where f equals the Satterthwaite degrees of freedom. Hence, reference t -distributions based on the Satterthwaite approximation tend to overestimate tail probabilities. For example, when data from a couple of PSUs nearly determine the value of a

coefficient, the Satterthwaite degrees of freedom can be less than two, incorrectly implying a chi-square density that is infinite at zero. Consequently, the probability of very large t -statistics may not be as large as the Satterthwaite approximation would imply, especially when the Satterthwaite degrees of freedom are less than 4 or 5.

5. Simulation Methods

We use a Monte Carlo simulation to study the properties of alternative variance estimators and tests for a balanced two-stage cluster sample with $n = 20$ PSUs and a constant $m = 10$ observations in each PSU. All simulation replications use a common design matrix \mathbf{X} with four explanatory variables chosen to represent a range of difficulty for nonparametric variance estimators. The first two explanatory variables, x_1 and x_2 , are dichotomous (0 or 1) and constant within PSU. The variable x_1 is 1 in half the clusters: 1, 3, ..., 19, while x_2 is 1 in just three clusters: 9, 10, and 11. Both x_3 and x_4 were generated from standard normal distributions. They differ in that x_3 was generated from a multivariate normal with intra-cluster correlation of 0.5 within PSU, while x_4 was generated from independent normal distributions. Observed intra-cluster correlations are 1.00, 1.00, 0.62 and -0.04 , respectively. Observed correlations among the explanatory variables are all very small with the exception of $\text{Corr}(x_1, x_2) = 0.14$, $\text{Corr}(x_1, x_3) = 0.25$ and $\text{Corr}(x_1, x_4) = -0.11$. The estimated regression

coefficients are linear combinations of the dependent variable with multipliers given by the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which are shown in Figure 1. For the first three coefficients, and to a lesser extent β_3 , observations from the same PSU tend to have similar multipliers. Of more importance, $\hat{\beta}_0$, $\hat{\beta}_2$, and $\hat{\beta}_3$ are determined primarily by results in a small number of PSUs with relatively large multipliers (in absolute value). For example, Figure 1 shows that the multipliers for β_3 are large for the second PSU, which has a mean that is over two standard deviations from the average PSU mean. In general, variance in the PSU means gives some PSUs greater weight for estimating β_3 .

The dependent variable was generated from the equation $y_{ij} = \beta' x_{ij} + \epsilon_{ij}$, where $\beta = 0$ and the ϵ_i 's are standard multivariate normal random variables with intra-cluster correlation ρ . We use three alternative values of $\rho = 0, 1/9$, and $1/3$, corresponding to design effects for the sample mean of $\text{DEFF} = 1, 2$, and 4 , respectively ($\text{DEFF} = 1 + (m - 1)\rho$). Monte Carlo results are based on 100,000 replications of \mathbf{y} for our fixed \mathbf{X} .

We evaluated the ordinary least squares (OLS) variance estimator, $s^2 l' (\mathbf{X}'\mathbf{X})^{-1} l$, and five nonparametric variance estimators: the standard linearization estimator given in equation (2) with $c = n / (n - 1)$; the jackknife estimator given in (5); bias reduced linearization; and Kott's two adjustments to linearization. BRL and the Kott adjustments are all based on working intra-cluster correlations of $\rho = 0$.

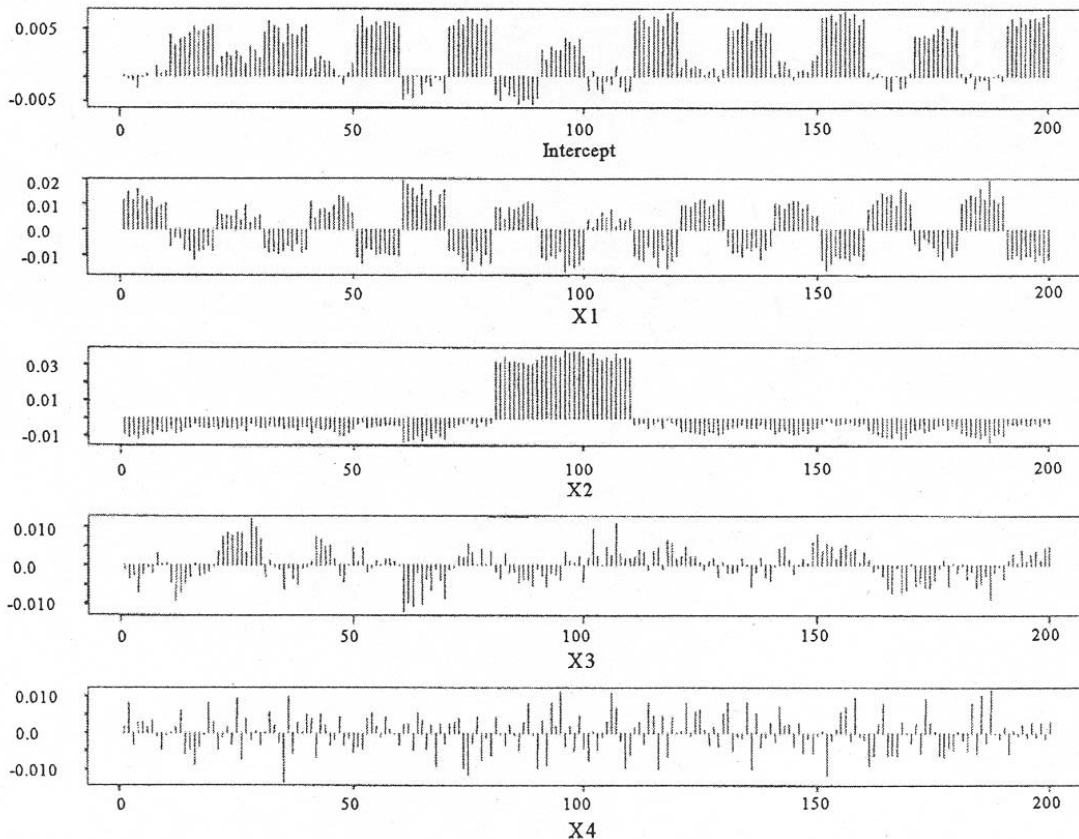


Figure 1. Values of the rows of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ for the design matrix used in simulations.

We estimated Type I error rates for eight alternative test procedures based on 100,000 replications from the null hypothesis where each $\beta_k = 0$, for $k = 0$ to 4. Each procedure compares a “ t -statistic” against a reference t -distribution. For the t 's based on linearization, the jackknife, and BRL, we use critical values from t -distributions with both $(n - 1) = 19$ degrees of freedom and the corresponding Satterthwaite approximation. For Kott's methods, we use his proposed degrees of freedom. All computations were implemented in SAS.

6. Simulation Results

Table 1 shows the bias of several variance estimators for the five regression coefficients (including the intercept) for $\rho = 0, 1/9$, and $1/3$. Except for Kott (1994), all values are exact based on the \mathbf{X} matrix described above. Because Kott (1994) cannot be written as a linear functional, its bias is estimated from the Monte Carlo simulations, and the standard error of the bias is shown in parentheses.

Table 1
Bias of Variance Estimators
(as a Percentage of the True Variance)

Estimator	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\rho = 0$					
OLS	0.0	0.0	0.0	0.0	0.0
Linearization	-9.6	-13.2	-32.5	-13.3	-1.8
Jackknife	11.7	17.2	51.2	17.6	2.1
Kott (1994)	4.0	2.5	-1.0	2.2	4.7
(standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	0.0	0.0	0.0	0.0	0.0
BRL	0.0	0.0	0.0	0.0	0.0
$\rho = 1/9$					
OLS	-50.2	-49.7	-50.7	-37.7	4.1
Linearization	-10.3	-14.2	-33.2	-17.1	-2.5
Jackknife	11.0	16.4	50.1	19.8	3.2
Kott (1994)	3.9	2.7	-0.8	1.5	4.6
(standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-0.8	-1.2	-1.0	-4.4	-0.7
BRL	-0.7	-1.0	-0.8	-1.2	0.1
$\rho = 1/3$					
OLS	-75.8	-75.5	-76.2	-65.3	13.8
Linearization	-10.7	-14.8	-33.5	-19.9	-4.1
Jackknife	10.7	15.9	49.5	21.4	5.9
Kott (1994)	3.6	2.4	-0.6	1.4	4.4
(standard error)	(0.2)	(0.1)	(0.3)	(0.2)	(0.1)
Kott (1996)	-1.2	-1.9	-1.5	-7.7	-2.3
BRL	-1.0	-1.5	-1.3	-2.1	0.4

Note: All values are exact except for Kott (1994), which is based on 100,000 simulation replications.

The OLS variances are unbiased for $\rho = 0$, but they are badly biased for $\rho = 1/9$ and $1/3$. As discussed in Wu, Holt, and Holmes (1988), the OLS variances are too small by

roughly a factor of $1/[1 + \rho(m - 1)ICC_x]$, where ICC_x denotes the intra-cluster correlation for an x variable. Hence, for PSU-level variables (including the intercept), the OLS variances are too small by roughly a factor of $1/DEFF$. Similarly, the bias is smaller, but still substantial for x_3 , the individual-level variable with large intra-cluster correlation. The positive bias for the OLS variance of $\hat{\beta}_4$ results from the slight negative intra-cluster correlation for x_4 .

Linearization and the jackknife each suffer from large biases, relatively independent of ρ , but the biases point in opposite directions. For each estimator, the magnitude of the bias varies greatly among the coefficients. The largest biases (in absolute value) occur for $\hat{\beta}_2$, which depends mainly on the data from three PSUs. The next greatest biases occur for $\hat{\beta}_3$, followed closely by $\hat{\beta}_1$ and $\hat{\beta}_0$.

Except for $\hat{\beta}_4$, Kott (1994) has much smaller magnitude bias than linearization. However, the method tends to over-compensate, often resulting in notable positive bias. An exception is $\hat{\beta}_2$, for which Kott's estimator remains biased low.

By design, Kott (1996) and BRL eliminate the bias for $\rho = 0$. Consequently, choice among these alternatives should rest mainly on how well they hold down bias for $\mathbf{V} \neq \mathbf{I}$. Both methods reduce the magnitude of bias dramatically relative to linearization for $\rho = 1/9$ and $1/3$. Although differences between the two methods are often small, BRL does uniformly better, with its worst bias being -2.1 percent. While Kott (1996) is practically indistinguishable from BRL for the PSU-level variables, it performs substantially worse for $\hat{\beta}_3$ and $\hat{\beta}_4$.

The linearization, jackknife, BRL and Kott estimators are highly correlated with similar coefficients of variation. For any given regression coefficient, the correlation among the variance estimators always exceeded 0.969, with most exceeding 0.99 (not shown). The smallest correlations tended to be between the jackknife and other estimators. The coefficients of variation (also not shown) were largest for Kott (1994) and tended to be smallest for linearization and Kott (1996) (except for the intercept). For the intercept, the jackknife had the smallest coefficient of variation. The relative variance of the BRL estimator was similar to that of the alternative nonparametric methods. Its coefficient of variation was between 1 and 6 percent larger than that of the linearization estimator but about 5 to 10 percent smaller than that of Kott (1994). Thus, the five nonparametric variance estimators tend to differ from each other mainly by constant factors, and Table 1 summarizes the main difference among these variance estimators.

Table 2 shows the Satterthwaite degrees of freedom for each of the five coefficients for the linearization, jackknife, BRL and Kott variance estimators. For all estimators the degrees of freedom were calculated assuming $\mathbf{V} = \mathbf{I}$ and consequently depend only on the design matrix and not on the values of \mathbf{y} . The approximations are similar for linearization and BRL although the linearization degrees of freedom tend to be slightly larger reflecting the fact that for

this design matrix the relative variances of the BRL estimators are marginally larger than those for linearization. Kott's approximation derives the coefficient of variation for a linearization-type estimator based on the true errors rather than the residuals. As a result, Kott's approximate degrees of freedom, which are larger than those for linearization or BRL, tend to overstate the precision of his estimator (see Kott 1994, section 6). Across all four estimators, the approximations are smallest for $\hat{\beta}_2$.

Table 2
Degrees-of-Freedom for Selected Estimators

Method	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
Satterthwaite (LIN)	9.02	14.45	3.30	11.56	16.65
Satterthwaite (Jackknife)	9.52	13.30	2.62	9.06	16.23
Satterthwaite (BRL)	9.24	14.08	2.90	10.26	16.45
Kott's method	10.33	16.41	4.32	11.36	17.44

Table 3 shows that Type I error rates for the standard linearization method with $(n - 1)$ degrees of freedom consistently exceed 5 percent for all three values of ρ . Type I errors are most common for $\hat{\beta}_2$, where they reach as high as 16 percent, but they also occur much too frequently for $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_3$, ranging from 7.0 to 8.8 percent. The magnitude of this problem correlates closely with the size of the bias of the linearization estimator (see Table 1). Type I error rates are much lower, 5.7 to 6.4 percent, for tests based on the Satterthwaite degrees of freedom. Thus using the alternative degrees of freedom improved the Type I error rates by about 30 to 88 percent.

There is a less consistent pattern for the Type I error probabilities for the jackknife. The jackknife with $(n - 1)$ degrees of freedom tends to be conservative for $\hat{\beta}_1$ and $\hat{\beta}_3$, in accord with the positive bias in the jackknife variance. In contrast, the probability of Type I error is much too large for $\hat{\beta}_2$, and a bit too large in two of three cases for the intercept $\hat{\beta}_0$. The apparent explanation is that the choice of $(n - 1)$ as the degrees of freedom for the reference t -distribution sometimes counteracts the bias in the jackknife variance. This conclusion is supported by the very low Type I error rates for the jackknife with Satterthwaite degrees of freedom; smaller degrees of freedom combined with large positive biases result in very conservative tests.

BRL with $(n - 1)$ degrees of freedom improves substantially on linearization with the same degrees of freedom. Because BRL is unbiased when $\rho = 0$, comparing the fifth row of the table against the first demonstrates the reduction in Type I errors that results from removing the bias of linearization. Excluding $\hat{\beta}_4$, BRL reduces Type I error rates by about 45 to 88 percent. However, BRL with $(n - 1)$ degrees of freedom remains consistently liberal, especially for $\hat{\beta}_2$. Comparison of rows 2 and 5 of each section shows the relative impact of bias reduction and the Satterthwaite

adjustment. For $\hat{\beta}_0$ and $\hat{\beta}_2$, degrees of freedom are more important, while bias matters more for $\hat{\beta}_1$ and $\hat{\beta}_3$. Performance for BRL with the Satterthwaite approximation is very good, except for $\hat{\beta}_2$, where the Type I error falls to about 3 percent.

Table 3
Type I Error Rates for Tests of the Null Hypothesis that $\beta = 0$

Estimator	Df	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\rho = 0$						
Linearization	$n - 1$	7.54	7.00	15.99	7.35	5.38
Linearization	Satt	5.75	6.45	6.33	6.28	5.18
Jackknife	$n - 1$	5.01	3.92	7.58	4.52	5.02
Jackknife	Satt	3.80	3.43	1.41	3.26	4.77
Kott (1994)	Kott	4.87	5.03	7.13	5.21	4.67
Kott (1996)	Kott	5.11	5.08	4.85	4.76	5.07
BRL	$n - 1$	6.28	5.37	11.25	5.90	5.21
BRL	Satt	4.73	4.86	3.12	4.72	5.00
$\rho = 1/9$						
Linearization	$n - 1$	7.81	7.14	16.19	8.18	5.34
Linearization	Satt	6.03	6.60	6.43	7.05	5.14
Jackknife	$n - 1$	5.31	4.06	7.63	4.49	4.77
Jackknife	Satt	4.11	3.61	1.48	3.24	4.51
Kott (1994)	Kott	5.07	5.03	7.00	5.51	4.56
Kott (1996)	Kott	5.42	5.28	5.14	5.32	5.01
BRL	$n - 1$	6.52	5.50	11.27	6.23	5.08
BRL	Satt	5.04	5.00	3.19	4.93	4.84
$\rho = 1/3$						
Linearization	$n - 1$	8.10	7.28	16.39	8.79	5.66
Linearization	Satt	6.30	6.78	6.62	7.53	5.44
Jackknife	$n - 1$	5.45	4.11	7.76	4.56	4.67
Jackknife	Satt	4.13	3.61	1.51	3.35	4.46
Kott (1994)	Kott	5.14	5.06	7.02	5.80	4.84
Kott (1996)	Kott	5.59	5.44	5.14	5.88	5.31
BRL	$n - 1$	6.76	5.63	11.55	6.45	5.19
BRL	Satt	5.18	5.14	3.30	5.26	4.98

Note: Entries with a true value of 5.00 percent have standard errors of 0.07 percent.

Tests based on Kott's 1994 estimator with his proposed degrees of freedom perform very well for the coefficients where the variance estimator is biased upward. It appears that the upward bias in the variance estimator is offset by the upward bias in the approximate degrees of freedom. Kott's variance estimator is slightly negatively biased for $\hat{\beta}_2$ and therefore the upward bias in the degrees of freedom compounds the bias in the estimator resulting in a Type I error rate of about 7 percent for all three values of ρ .

Tests based on Kott's 1996 estimator also perform well. For almost all the coefficients and all values of ρ the Type I error rate is close to 5 percent. The exception is the test for $\hat{\beta}_3$ when $\rho = 1/3$, which has an error rate of 5.88 percent as a result of the moderate bias in the variance estimator.

7. Example from the Partners in Care Experiment

We illustrate the methods in this paper using data from Partners in Care, a longitudinal experiment assessing the effect of “quality improvement” programs on care for depression in managed care organizations (MCOs) (Wells *et al.* 2000). The experiment followed 1,356 patients who screened positive for depression in 1996–1997 in 43 clinics of seven MCOs. Clinics were assigned at random to one of three experimental cells: usual care, a quality improvement program supplemented by resources for medication follow-up, or a quality improvement program supplemented by resources for access to psychotherapists. Clinics were assigned at random after forming 27 clinic sets – three for each of nine blocks (six MCOs constituted single blocks, and one MCO was divided into three blocks based on ethnic mix of the clinics). Within blocks of more than three clinics,

clinic sets were combined to match as closely as possible on anticipated sample size and patient characteristics. See Wells *et al.* (2000) for additional details.

We present results from an OLS regression on the mental health summary score from the SF-12 (Ware, Kosinski, and Keller 1995) for 1,048 patients at 6-month follow-up. Scores were standardized to have mean 50 and standard deviation 10 in a general population, with higher scores indicating better health. As in Wells *et al.* (2000), the explanatory variable of primary interest is an intervention indicator that estimates the combined effect of medication or therapy versus care as usual. The first two columns of Table 4 show OLS coefficients and standard errors for the intervention effect and all the covariates used by, but not reported in, Wells *et al.* (2000). Our regression differs from theirs because we do not weight for nonresponse or impute for missing values of the outcome variable, but the results for the intervention effect agree reasonably closely.

Table 4
Comparison of OLS, Linearization, and BRL Inference for Partner-in-Care Example

Explanatory Variable	$\hat{\beta}_j$	SE_{OLS}	$\frac{SE_{LIN}}{SE_{OLS}}$	$\frac{SE_{BRL}}{SE_{OLS}}$	DF_{BRL}	OLS	P-value	
							LIN	BRL
PSU-Level								
Intercept	28.795	3.409	1.03	1.06	23.7	0.000	0.000	0.000
Intervention	1.724	0.746	0.73	0.84	15.4	0.021	0.003	0.015
Block 1	1.386	1.867	0.63	0.80	2.7	0.458	0.244	0.426
Block 2	-0.031	1.576	0.88	1.07	3.6	0.984	0.982	0.986
Block 3	-1.042	1.230	0.53	0.61	3.9	0.397	0.117	0.241
Block 4	0.038	1.231	0.62	0.73	4.5	0.976	0.961	0.968
Block 5	-3.707	1.503	0.66	0.78	4.7	0.014	0.001	0.027
Block 6	-0.025	1.562	1.15	1.32	4.9	0.987	0.989	0.991
Block 7	-2.784	1.644	0.84	0.97	7.0	0.090	0.051	0.126
Block 8	0.822	1.233	0.93	1.03	12.0	0.505	0.476	0.527
Demographic								
Black	0.972	1.448	0.74	0.79	7.6	0.502	0.369	0.419
Hispanic	0.202	1.004	0.73	0.75	24.3	0.841	0.785	0.791
Other nonwhite	-1.033	1.409	0.77	0.80	21.6	0.463	0.349	0.369
Female	-0.502	0.803	1.09	1.12	23.1	0.532	0.571	0.581
Log of net worth + \$1,000	0.015	0.215	0.87	0.89	23.6	0.943	0.936	0.937
Less than high school	-1.690	1.217	1.00	1.04	25.3	0.165	0.173	0.192
Some college	-1.140	0.879	0.77	0.78	26.0	0.195	0.097	0.108
College graduate	-0.703	1.047	0.78	0.79	21.1	0.502	0.393	0.404
Age	0.059	0.032	0.91	0.93	26.5	0.064	0.047	0.056
Married	0.541	0.748	1.05	1.07	28.5	0.470	0.496	0.504
Baseline Health								
1 chronic condition (of 19)	-0.973	1.039	0.92	0.94	23.7	0.349	0.313	0.327
2 chronic conditions	0.198	1.116	0.87	0.90	23.0	0.859	0.840	0.846
3+ chronic conditions	-0.201	1.132	0.90	0.91	24.0	0.859	0.844	0.847
Depression and dysthymia	-5.305	1.335	0.93	0.95	25.8	0.000	0.000	0.000
Depression or dysthymia	-3.882	0.982	1.12	1.15	23.7	0.000	0.001	0.002
Prior depression only	-2.396	1.109	1.02	1.05	21.2	0.031	0.040	0.052
Mental component of SF-12	0.287	0.036	1.11	1.14	26.6	0.000	0.000	0.000
Physical comp of SF-12	0.079	0.036	0.88	0.89	24.6	0.029	0.017	0.022
Anxiety disorder	-2.438	0.749	1.20	1.23	26.3	0.001	0.010	0.014

Because patients from the same clinics could have similar outcomes, OLS standard errors could easily be too low – especially for PSU-level variables like Intervention. Columns 3 and 4 of Table 4 show the ratios of linearization and BRL standard errors to the OLS standard errors. We use clinic as the PSU because there is very little reason to expect correlations of errors across clinics after controlling for block.

Using the method of Wu, Holt and Holmes (1988), we estimate the intra-clinic correlation of the errors as -0.0026 , easily consistent with a true value of 0. Nonetheless, there is no reason to expect any of the correct standard errors to fall much below those obtained from OLS. Column 3 of Table 4 shows that the linearization standard errors frequently fall far below those obtained from OLS – especially for the PSU-level explanatory variables at the top of the table. Similarly, linearization with a reference t_{n-1} often produces much smaller P -values than does OLS.

BRL improves over linearization. BRL standard errors are always larger and sometimes substantially larger than the linearization standard errors. For example, the BRL estimates for PSU-level explanatory variables are on average 15 percent larger than the linearization estimates. On the other hand, BRL standard errors for PSU-level variables are still often smaller than the OLS estimates. Thus, even though BRL estimators should be nearly unbiased, the variability in the estimators results in estimates for some coefficients that are small. The variability is also reflected in degrees of freedom that are very small for the block indicators and, while larger for patient level variables, are still considerably less than 42, the number of clusters minus one. The degrees of freedom are especially small, 7.6, for the indicator variable Black (equal to one if the patient was African American and zero otherwise). Plots analogous to Figure 1 show that Black was concentrated in three clusters. The Black indicator equals zero for all the patients in 24 of 43 clusters, and 48 of the 78 African Americans in the sample were found in just three clusters. As discussed in sections 2 and 4, the concentration of Black into a small number of clusters results in high variance for both estimators and large bias in the linearization estimator, both of which can be seen in Table 4.

8. Discussion

Although linearization is a valuable tool that provides consistent standard errors and valid inference as the number of PSUs grows large in multi-stage samples, users should recognize problems with the method. Estimated variances of linear regression coefficients (including domain means) tend to be biased low – especially for coefficients (or linear combinations of coefficients) that depend largely on data from a small number of PSUs. Depending on the design, large biases can persist even when the total number of PSUs is quite large. The standard jackknife for multi-stage samples tends to have at least as large bias in the opposite direction. Similarly, using a reference t distribution with

degrees of freedom equal to one less than the number of PSUs may greatly understate the uncertainty in the estimated variance. Because the two problems (bias and overstated degrees of freedom) tend to occur in tandem for linearization, confidence intervals and statistical tests based on that method may be far too liberal.

Bias reduced linearization (BRL) produces unbiased variance estimates in the event that errors are homoskedastic and uncorrelated, and it tends to greatly reduce bias for other covariance structures investigated in our simulations. In our simulations, BRL consistently exhibited smaller biases than linearization by 90 percent or more and tended to improve substantially on Kott's 1994 adjusted linearization method. Results for BRL were comparable to those for Kott's 1996 method.

When BRL was used with the estimated Satterthwaite degrees of freedom, statistical inference improved greatly in comparison with the standard use of linearization. Bias reduction and Satterthwaite degrees of freedom seemed to contribute about equally to the improved performance. Although Satterthwaite's approximation may overcompensate, leading to conservative inference in certain situations, the problem does not seem noteworthy until the Satterthwaite degrees of freedom drop below 5 (based, in part, on simulations not reported in this paper). In such cases, analysts might choose to estimate critical values using simulations based on Theorem 4.

It is important to note some limitations of our simulation results. First, we only report results for four distinct explanatory variables plus an intercept. We choose those variables to span a wide variety of situations. Although some might describe x_2 as extreme or pathological, it is not outside the range of situations that we have seen in our own consulting work. Variables like x_2 can result from group-randomized trials (see section 7) or observational data where only a few PSUs exhibit a particular trait or from use of a series of dummy variables to represent levels of a categorical variable. Second, we present results only for $n = 20$ PSUs. To the extent that \mathbf{X} remains similar as n increases (e.g., by replication), Equation (4) implies that the bias declines in proportion to $1/(n - 1)$. Also, the results observed for $n = 20$ could occur for much larger n if the bulk of the variation in \mathbf{X} is contributed by a few PSUs, and the determination of $l' \hat{\beta}$ depends similarly on a small number of PSUs. Finally, to reduce the number of factors affecting the results, we simplified the design in several ways: constant PSU sizes, no weights or strata, and little multicollinearity. We suspect that relaxing any of those constraints would actually tend to make standard linearization and the jackknife perform worse. We do not believe that the choice of $m = 10$ for the PSU size had much impact either way on our findings.

Although we believe that our proposed methods will prove valuable to analysts of multi-stage samples, these

methods will not completely solve the inference problem for unweighted linear regression. Both authors have frequently observed the disturbing situation where standard linearization methods produced shorter confidence intervals than methods that ignore the design. Certainly, the bias of v_L and improper use of $n - 1$ degrees of freedom contribute to the frequency of this phenomenon, but our methods would not eliminate its occurrence (see section 7). Linearization, like sample reuse methods, necessarily produces estimators with high variance for some or possibly all coefficients in certain designs. When confronted with situations like the coefficients for our x_2 , where the Satterthwaite degrees of freedom fall near 3 or lower, analysts should seriously consider whether they can afford the large variability, and corresponding loss of power, that comes with nonparametric variance estimators. Parametric alternatives like hierarchical linear models or inference based on estimating a common intra-class correlation across all the PSUs (Wu, Holt and Holmes 1988) should produce more stable results.

Although this paper has focused on unweighted linear regression for samples without stratification, we have no reason to expect that the bias and degrees-of-freedom problems of linearization would be lessened by stratification or for either weighted least squares or generalized linear models (GLMs). As shown in McCaffrey, Bell and Botts (2001) the BRL method extends immediately to weighted linear regression by using $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ in the main condition of Theorem 3. Because solutions to GLMs, such as logistic regression, are equivalent to the final steps of iteratively reweighted least squares (McCullagh and Nelder 1989), the obvious choice for these models is to use BRL based on the final weights and to set $\mathbf{U} = \mathbf{W}^{-1}$. Nevertheless, Theorem 3 does not extend to GLMs because the weights are estimated from the data, and we have not investigated the properties of BRL in this context.

Korn and Graubard (1995) suggest $v_L^{1/2}$ as a standard error estimator for stratified samples in situations where the stratification is non-informative. The same reasoning applies to $v_{BRL}^{1/2}$. Fuller (1975) proposed an alternative design consistent standard error estimator for stratified samples. Bell and McCaffrey (2002, pages 32–33) show that by adjusting the vector of residuals for each stratum, BRL can reduce or remove the model bias that can exist in Fuller's estimator.

Acknowledgements

We thank the referees and associate editor for valuable comments on an earlier draft. This work is supported in part by NSF Grant 0001763.

Appendix

Proofs of Theorems 2 and 4

Proof of Theorem 2. Following the first steps of the proof of Theorem 1, equation (6) implies that

$$E(v_{JK}) = \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \mathbf{X}_i'(\mathbf{I}_i - \mathbf{H}_{ii})^{-1} \mathbf{X}_i\right) (\mathbf{X}'\mathbf{X})^{-1} l.$$

The existence of $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ implies that the eigenvalues of \mathbf{H}_{ii} are strictly less than 1, so that $(\mathbf{I}_i - \mathbf{H}_{ii})^{-1}$ can be written as $\sum_{j=0}^{\infty} \mathbf{H}_{ii}^j$. Consequently, letting $\mathbf{D} = (1/n)(\mathbf{X}'\mathbf{X})$ and $\mathbf{D}_i = (\mathbf{X}_i' \mathbf{X}_i) - \mathbf{D}$, we have

$$\begin{aligned} E(v_{JK}) &= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{k=1}^{\infty} [(\mathbf{D} + \mathbf{D}_i)(\mathbf{X}'\mathbf{X})^{-1}]^k\right) l \\ &= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{k=1}^{\infty} \sum_{r=0}^k \binom{k}{r} \frac{1}{n^{k-r}} [\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}]^r l\right) \\ &= \left(\frac{n-1}{n}\right) l'(\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \left(\sum_{r=0}^{\infty} \sum_{\substack{s=0 \\ r+s>0}}^{\infty} \binom{r+s}{r} \frac{1}{n^s} [\mathbf{D}_i(\mathbf{X}'\mathbf{X})^{-1}]^r l\right). \end{aligned}$$

The term for $r = 0$ equals $l'(\mathbf{X}'\mathbf{X})^{-1} l = \text{Var}(l' \hat{\beta})$. The term for $r = 1$ equals 0. By the binomial theorem,

$$\sum_{s=0}^{\infty} \binom{r+s}{r} \frac{1}{n^s} = \left(\frac{n}{n-1}\right)^{r+1},$$

so that the remaining terms can be paired, for $r = 2, 4, 6, \dots$, to give

$$\left(\frac{n}{n-1}\right)^r l' \sum_{i=1}^n \left\{ \begin{aligned} & [(\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i]^{r/2} \\ & \left[(\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1}\right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1} \right] \\ & [\mathbf{D}_i (\mathbf{X}'\mathbf{X})^{-1}]^{r/2} \end{aligned} \right\} l.$$

The middle factor in the summation can be written as,

$$\left(\frac{n-2}{n-1}\right) (\mathbf{X}'\mathbf{X})^{-1} + \left(\frac{n}{n-1}\right) (\mathbf{X}'\mathbf{X})^{-1} (\mathbf{X}_i' \mathbf{X}_i) (\mathbf{X}'\mathbf{X})^{-1},$$

which is positive definite, so that the whole expression must be positive. Consequently, we have shown that $E(v_{JK}) \geq \text{Var}(l' \hat{\beta})$ with equality if and only if $l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{D}_i = 0$, which is true if and only if $l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i\mathbf{X}_i$ is constant across i .

Proof of Theorem 4.

$$\begin{aligned} v^* &= c \sum_{i=1}^n l'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_i\mathbf{A}_i(\mathbf{I}-\mathbf{H})_i\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{I}-\mathbf{H})'_i \\ &\quad \mathbf{A}_i\mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}l \\ &= \boldsymbol{\varepsilon}' \sum_{i=1}^n \mathbf{g}_i \mathbf{g}'_i \boldsymbol{\varepsilon}. \end{aligned}$$

Let \mathbf{P} equal the matrix of eigenvectors and $\mathbf{\Lambda}$ denote the diagonal matrix with elements $\lambda_1, \dots, \lambda_M$ equal to the eigenvalues of $\mathbf{V}^{1/2} \sum_{i=1}^n \mathbf{g}_i \mathbf{g}'_i \mathbf{V}^{1/2} = \mathbf{B}'\mathbf{B}$ where $\mathbf{B}' = \mathbf{V}^{1/2} [\mathbf{g}_1 \mathbf{g}_2 \dots \mathbf{g}_n]$. Let $\mathbf{u} = \mathbf{P}'\mathbf{V}^{-1/2}\mathbf{y}$ where $\mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2} = \mathbf{I}$ defines $\mathbf{V}^{-1/2}$, then the elements of \mathbf{u} are independent normal variables with variance 1 and

$$v^* = \mathbf{u}'\mathbf{\Lambda}\mathbf{u} = \sum_{i=1}^M \lambda_i u_i^2.$$

Let λ_i be any nonzero eigenvalue of $\mathbf{B}'\mathbf{B}$, then there exists a nonzero vector \mathbf{z} such that $\mathbf{B}'\mathbf{B}\mathbf{z} = \lambda_i\mathbf{z}$ and $\mathbf{B}\mathbf{B}'\mathbf{z} = \lambda_i\mathbf{z}$. Because $\mathbf{B}\mathbf{z} \neq \mathbf{0}$, λ_i is an eigenvalue of $\mathbf{B}\mathbf{B}'$. Similarly, any nonzero eigenvalue of $\mathbf{B}\mathbf{B}'$ is also an eigenvalue of $\mathbf{B}'\mathbf{B}$. Therefore, the nonzero eigenvalues of $\mathbf{B}'\mathbf{B}$ equal the nonzero eigenvalues of $\mathbf{B}\mathbf{B}' = \{\mathbf{g}'_i \mathbf{V}\mathbf{g}_j\}$.

References

- Bell, R.M., and McCaffrey, D.F. (2002). *Bias Reduction in Linearization Standard Errors for Linear Regression with Multi-Stage Samples*. AT&T Labs-Research, Florham Park, NJ, TD-4S9H9T, www.research.att.com/~rbell.
- Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newberry Park, CA: Sage.
- Cochran, W.G. (1977). *Sampling Techniques*. Third Edition, New York: John Wiley & Sons, Inc.
- Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.
- Efron, B., and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Ellickson, P.L., and McGuigan, K.A. (2000). Early predictors of adolescent violence. *American Journal of Public Health*, 90, 566-572.
- Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhyā*, C, 37, 117-32.
- Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis*. London: Chapman and Hall.
- Goldstein, H. (1991). Multilevel Modeling of Survey Data. *The Statistician*, 40, 235-244.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.
- Korn, E.L., and Graubard, B.I. (1995). Analysis of large health surveys: Accounting for the sampling design. *Journal of the Royal Statistical Society, Series A, General*, 158, 263-295.
- Kott, P.S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology*, 20, 159-64.
- Kott, P.S. (1996). Linear regression in the face of specification error: model-based exploration of randomization-based techniques. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 39-47.
- Landis, J.R., Lepkowski, J.M., Ekland, S.A. and Stehouwer, S.A. (1982). A statistical methodology for analyzing data from a complex survey: the first national health and nutrition examination survey. *Vital and Health Statistics*. Washington, D.C: US Government Printing Office. Series 2, 92.
- Mackinnon, J.G., and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305-325.
- Mancl, L.A., and Derouen, T.A. (2001). A covariance estimator for gee with improved small-sample properties. *Biometrics*, 57, 126-134.
- McCaffrey, D.F., and Bell, R.M. (1997). Bias reduction in standard error estimates for regression analyses from multi-stage designs with few primary sampling units. Paper presented at the Joint Statistical Meetings, Anaheim CA.
- McCaffrey, D.F., Bell, R.M. and Botts, C.H. (2001). Generalizations of bias reduced linearization. *Proceeding of the Survey Research Methods Section, American Statistical Association*.
- McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. Second Edition, London: Chapman and Hall.
- Murray, D. M., Hannan, P. J., Wolfinger, R. D., Baker, W.L. and Dwyer, J.H. (1998). Analysis of data from group-randomized trials with repeat observations on the same groups. *Statistics in Medicine*, 17, 1581-1600.
- Rust, K.F., and Rao, J.N.K. (1996). Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.
- Sas Institute Inc. (1999). *SAS/STAT* User's Guide, Version 8*. Cary, NC: Author.
- Satterthwaite, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.
- Searle, S.R., Casella, G. and McCulloch, C.E. (1992). *Variance Components*. New York: John Wiley & Sons, Inc.
- Shah, B.V., Bamwell, B.G. and Bieler, G.S. (1997). *SUDAAN User's Manual, Release 7.5*. Research Triangle Park, NC: Research Triangle Institute.
- Shah, B.V., Holt, M. M. and Folsom, R.E. (1977). Inference About Regression Models from Survey Data. *Bulletin of the International Statistical Institute*, 41, 43-57.

- Shapiro, M.F., Morton, S.C., McCaffrey, D.F., Senterfitt, J.W., Fleishman, J.A., Perlman, J.F., Athey, L.A., Keesey, J.W., Goldman, D.P., Berry, S.H. and Bozzette, S.A. (1999). Variations in the care of hiv-infected adults in the United States; results from the hiv cost and services utilization study. *Journal of the American Medical Association*, 281, 2305-2315.
- Skinner, C.J. (1989a). Introduction to Part A. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons, Inc. 23-57.
- Skinner, C.J. (1989b). Domain means, regression and multivariate analyses. *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt and T.M.F. Smith). New York: John Wiley & Sons, Inc. 59-88.
- Statacorp. (1999). *Stata Statistical Software: Release 6.0*. College Station, TX: Author.
- Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley & Sons, Inc.
- Ware, J.E., Jr., Kosinski, M. and Keller, S.D. (1995). *SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales*. Boston, Mass: The Health Institute, New England Medical Center.
- Wells, K.B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., Miranda, J., Carney, M. and Rubenstein, L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association*. 283, 212-220.
- Wells, K.B., Sherbourne, C., Schoenbaum, M., Duan, N., Meredith, L., Unutzer, J., Miranda, J., Carney, M. and Rubenstein, L.V. (2000). Impact of disseminating quality improvement programs for depression in managed primary care: a randomized controlled trial. *Journal of the American Medical Association*, 283, 212-220.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817-838.
- Wu, C.J.F., Holt, D. and Holmes, D.J. (1988). The effect of two stage sampling on the F statistic. *Journal of the American Statistical Association*, 83, 150-9.
- Zeger, S.L., and Liang, K.Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42, 121-130.