The Annals of Statistics
1986, Vol. 14, No. 4, 1261–1295

INVITED PAPER

JACKKNIFE, BOOTSTRAP AND OTHER RESAMPLING METHODS IN REGRESSION ANALYSIS[1]

BY C. F. J. WU

University of Wisconsin-Madison

Motivated by a representation for the least squares estimator, we propose a class of weighted jackknife variance estimators for the least squares estimator by deleting any fixed number of observations at a time. They are unbiased for homoscedastic errors and a special case, the delete-one jackknife, is almost unbiased for heteroscedastic errors. The method is extended to cover nonlinear parameters, regression M-estimators, nonlinear regression and generalized linear models. Interval estimators can be constructed from the jackknife histogram. Three bootstrap methods are considered. Two are shown to give biased variance estimators and one does not have the bias-robustness property enjoyed by the weighted delete-one jackknife. A general method for resampling residuals is proposed. It gives variance estimators that are bias-robust. Several bias-reducing estimators are proposed. Some simulation results are reported.

Table of contents

1. Introduction and summary. Statistical inference based on data resampling has drawn a great deal of attention in recent years. Most of the theoretical work thus far has been for the independent and identically distributed (i.i.d.)

Received April 1984; revised March 1986.

[1]Supported by the Alfred P. Sloan Foundation for Basic Research. Also sponsored by the United States Army under contract No. DAAG29-80-C-0041.

AMS 1980 *subject classifications*. Primary 62J05, 62J02, 62G05.

Key words and phrases. Weighted jackknife, bootstrap, linear regression, variable jackknife, jackknife percentile, bias-robustness, bias reduction, Fieller's interval, representation of the least squares estimator, M-regression, nonlinear regression, generalized linear models, balanced residuals.

case. Resampling methods justifiable in the i.i.d. case may not work in more complex situations. The main objective of this paper is to study these methods in the context of regression models, and to propose new methods that take into account special features of regression data.

Four resampling methods for regression problems are reviewed in Section 2; three of them deal with resampling from the (y, x) pairs. Two main problems with this group of methods are that they neglect the unbalanced nature of regression data and the choice of the resample size is restrictive. The fourth method is to bootstrap the residuals. It depends on the exchangeability of the errors and is not robust against error variance heteroscedasticity.

We propose in Section 4 a class of weighted jackknife methods, which do not have the problems just mentioned. Its two salient features are the flexible choice of resample size and the weighting scheme. Advantages for the first will be discussed in the next paragraph. We now discuss the second feature. For linear regression models (2.1), the proposed variance estimators $v_{J,r}$ (4.1) and $\tilde{v}_{J,r}$ (4.3) are, apart from a scalar factor, weighted sums of squares of the difference between the subset estimate and the full-data estimate (over all the subsets of size r). The weight is proportional to the determinant of the X^TX matrix for the subset. This choice of weight is motivated by a representation result (Theorem 1 in Section 3), which says that the full-data least squares estimator (LSE) is a weighted average, using the same weight, of the subset LSE over the same subsets. (A more general representation for any symmetric resampling procedure is given in Theorem 2.) For linear parameters, $v_{J,r}$ and $\tilde{v}_{J,r}$ are unbiased for estimating the variance of the LSE for homoscedastic errors (Theorem 3). A special case, the delete-one jackknife (i.e., subset size = sample size $-$ 1), and another delete-one jackknife variance estimator (Hinkley (1977)) with a different choice of weights, are both almost unbiased for estimating the variance of the LSE for heteroscedastic errors (Theorems 5 and 6). The latter estimator is, however, biased for unbalance designs as shown in the simulation study. We also show in Theorem 4 that $v_{J,k}$, that is $v_{J,r}$ with the subset size r equal to the number of β parameters in the linear model (2.1), and a more general version $v'_{J,k}$ (4.12) are identical to the usual variance estimator $\hat{\sigma}^2(X^TX)^{-1}$ (2.9). Since $\hat{\sigma}^2(X^TX)^{-1}$ is not a consistent estimator for the variance of the LSE when the errors are heteroscedastic, the result suggests that, for the purpose of robustness, the subset size in $v_{J,r}$ should not be too small.

Why do we consider jackknifing with a general choice of subset size? To answer this, let us review the delete-one jackknife, which is often synonymous with the jackknife in the literature. The delete-one jackknife works fine for bias and variance estimation in the case of smooth estimators (see Miller's review (1974a)), but for nonsmooth estimators such as the sample median, it is known to give inconsistent variance estimators. Another disadvantage is that the normalized histogram of $\{\hat{\theta}_{(i)}\}_1^n$, where $\hat{\theta}_{(i)}$ is obtained from the original estimate $\hat{\theta}$ by deleting the ith observation, does not in general converge to standard normal. Therefore, one cannot construct valid confidence intervals without estimating variance. Wu (1985) proves that, for the one sample mean and nonnormal errors, asymptotic normality is obtained iff d, the number of observations deleted, and

$n - d$, both diverge to ∞. Beyond normality, interval estimators based on the histogram of the resample estimates, unlike the symmetric t-intervals, take into account the possible skewness of the original estimator $\hat{\theta}$ and often possess desirable second-order properties (Singh (1981); Beran (1982); Abramovitch and Singh (1985); Efron (1985); Wu (1985)). Jackknifing with a flexible choice of subset size will allow this prospect to be exploited (see (4.6)).

Other resampling methods are studied in Section 6. The variable jackknife is an extension of the jackknife by allowing different subset sizes. The variance estimator $v_{J,r}$ (4.1) is extended to this situation. Two bootstrap methods for variance estimation are considered. A simple example is given to show that they do not, in general, give unbiased variance estimators even in the equal variance case. Careless use of the unweighted bootstrap can lead to an inconsistent and upward-biased variance estimator (see (6.14) to (6.17)).

In Section 7 a general method for resampling residuals is proposed by retaining an important feature of the jackknife. Unlike the method of bootstrapping residuals (2.7)–(2.8), it gives variance estimators that are unbiased for heteroscedastic errors. A special case, called the balanced residuals method, resembles the balanced half-samples method (McCarthy (1969)) for stratified samples and is worth further investigation in view of its more efficient and systematic use of the resamples.

In Section 8 the weighted jackknife method of Section 4 is extended to regression M-estimators, nonlinear regression and generalized linear models. The only essential change for the last two models is in the choice of weights, i.e., to replace the determinant of the X^TX matrix of the subset by that of the (estimated) Fisher information matrix of the subset. Note that the Fisher information matrix for the linear model (2.1) with i.i.d. errors is proportional to X^TX.

The issue of bias reduction for nonlinear parameter $\theta = g(\beta)$ is studied in Section 9. It is shown that bias reduction is achievable if and only if there exists an unbiased estimator for the variance of the LSE $\hat{\beta}$ (apart from lower-order terms). Based on this connection, several estimators of the bias of $\hat{\theta} = g(\hat{\beta})$ are proposed as natural counterparts of the variance estimators considered before.

Several estimators are compared in a simulation study, assuming a quadratic regression model. Criteria for the simulation comparison include the bias of estimating the variance–covariance matrix of $\hat{\beta}$, the bias of estimating a nonlinear parameter θ, the coverage probability and length of the interval estimators of θ. For the last two criteria, Fieller's method and the t-interval with the linearization variance estimator are included for comparison. The simulation results are summarized at the end of Section 10. Concluding remarks and further questions are given in Section 11.

2. Review of past work. We define a linear model by $y_i = x_i^T\beta + e_i$, where x_i is a $k \times 1$ deterministic vector, β is the $k \times 1$ vector of parameters and e_i are uncorrelated errors with mean zero and variance σ_i^2. Writing $y = (y_1, \ldots, y_n)^T$, $e = (e_1, \ldots, e_n)^T$ and $X = [x_1, \ldots, x_n]^T$, it can be rewritten as

$$y = X\beta + e, \qquad \mathrm{Var}(e) = \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2). \tag{2.1}$$

We assume that X^TX is nonsingular. The ordinary LSE is

$$\hat{\beta} = (X^TX)^{-1}X^Ty. \tag{2.2}$$

The ordinary jackknife for linear models is given as follows. Let $\hat{\beta}_{(i)}$ be the estimate of β obtained for recomputing $\hat{\beta}$ in (2.2) with the ith pair (y_i, x_i) deleted from the sample. For estimating a nonlinear function of β, $\theta = g(\beta)$, define $\hat{\theta} = g(\hat{\beta})$, $\hat{\theta}_{(i)} = g(\hat{\beta}_{(i)})$ and the pseudovalues $p_i = n\hat{\theta} - (n-1)\hat{\theta}_{(i)}$. The ordinary jackknife point estimator of θ is given by $\tilde{\theta} = n^{-1}\sum_1^n p_i$ and the ordinary jackknife variance estimator for $\hat{\theta}$ is given by

$$v_J = \frac{1}{n(n-1)}\sum_1^n (p_i - \tilde{\theta})(p_i - \tilde{\theta})^T. \tag{2.3}$$

The asymptotic properties of $\tilde{\theta}$ and v_J were studied by Miller (1974b) under strong conditions that excluded cases with very unbalanced X. Hinkley (1977) pointed out three shortcomings of the method: For linear parameters $\theta = \beta$ and $\sigma_i^2 = \sigma^2$ in (2.1), (1) $\tilde{\beta}$ is unbiased but has in general bigger variance than the LSE $\hat{\beta}$; (2) v_J is, in general, biased for estimating $\mathrm{Var}(\hat{\beta})$ or $\mathrm{Var}(\tilde{\beta})$. (3) The order of the bias of $\tilde{\theta}$ is between n^{-1} and n^{-2}, depending on the balancedness of the X matrix.

This criticism is well supported by our empirical findings (Table 1, Section 10).

The problems just mentioned stem from the *balanced* nature of the ordinary jackknife, which neglects the *unbalanced* nature of the regression data. Hinkley (1977) proposed a weighted modification. Using the weighted pseudovalues

$$Q_i = \hat{\theta} + n(1 - w_i)(\hat{\theta} - \hat{\theta}_{(i)}), \qquad w_i = x_i^T(X^TX)^{-1}x_i,$$

he defined the jackknife point estimator

$$\tilde{\theta}_w = \frac{1}{n}\sum_1^n Q_i$$

and the jackknife variance estimator

$$v_{H(1)} = \frac{1}{n(n-k)}\sum_1^n (Q_i - \tilde{\theta}_w)(Q_i - \tilde{\theta}_w)^T. \tag{2.4}$$

The notation (1) denotes "delete-one." For $\theta = \beta$, $\tilde{\beta}_w$ is identical to $\hat{\beta}$, and $v_{H(1)}$ can be written as

$$\sum_1^n \frac{(1 - w_i)^2}{1 - n^{-1}k}(\hat{\beta}_{(i)} - \hat{\beta})(\hat{\beta}_{(i)} - \hat{\beta})^T \tag{2.5}$$

$$= (X^TX)^{-1}\sum_1^n \frac{r_i^2}{1 - n^{-1}k} x_i x_i^T (X^TX)^{-1}, \qquad r_i = y_i - x_i^T\hat{\beta}. \tag{2.6}$$

For equal variances $v_{H(1)}$ is biased (see Theorem 6), but is robust against error variance heterogeneity. He also showed heuristically that $\tilde{\theta}_w$ restores the

bias-reducing property for which the i.i.d. jackknife is acclaimed. A more general treatment of this property will be given in Section 9.

The delete-one jackknife method has, however, several disadvantages as described in Section 1. We propose instead a class of weighted modifications, allowing for the deletion of an *arbitrary* number of observations. Many of its members, including the delete-one method, share the desirable properties of Hinkley's weighted modification mentioned previously. However, unlike $v_{H(1)}$, all the variance estimators in the class are unbiased for $\theta = \beta$ and $\sigma_i^2 = \sigma^2$.

Instead of recomputing the point estimate by deleting observation(s) each time, Efron (1979) advocated the use of simple random sampling with replacement (i.i.d. sampling) for resampling data and gave it a catchy name "bootstrap." Two bootstrap methods were considered for the regression model (2.1).

One is based on drawing an i.i.d. sample $\{e_i^*\}_1^n$ from the normalized residuals $\{r_i/(1 - kn^{-1})^{1/2}\}_1^n$, where $r_i = y_i - x_i^T\hat{\beta}$ is the ith residual. Define the bootstrap observation $y_i^* = x_i^T\hat{\beta} + e_i^*$, by treating $\hat{\beta}$ as the "true" parameter and $\{r_i/(1 - kn^{-1})^{1/2}\}$ as the "population" of errors. The bootstrap LSE is

$$\beta^* = (X^TX)^{-1}X^Ty^*. \tag{2.7}$$

For a nonlinear estimator $\hat{\theta} = g(\hat{\beta})$, the bootstrap variance estimator is defined as

$$v_b = E_*(\theta^* - \hat{\theta})(\theta^* - \hat{\theta})^T, \qquad \theta^* = g(\beta^*), \qquad \beta^* \text{ in (2.7)}, \tag{2.8}$$

where the asterisk (*) denotes i.i.d. sampling (or bootstrap sampling) from the population of normalized residuals. Note that $\theta^* - \hat{\theta}$ is unweighted. For $\theta = \beta$, it is easy to see (Efron (1979)) that

$$E_*\beta^* = \hat{\beta},$$

$$v_b = \hat{v} = \hat{\sigma}^2(X^TX)^{-1}, \qquad \hat{\sigma}^2 = \frac{1}{n-k}\sum_1^n r_i^2, \tag{2.9}$$

that is, the bootstrap variance estimator is identical to the usual variance estimator $\hat{v}$. Therefore, for homoscedastic errors $\sigma_i^2 = \sigma^2$, v_b is unbiased. But for heteroscedastic errors (unequal σ_i^2 in (2.1)), v_b is in general *biased* and *inconsistent* since the true variance of the LSE $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = (X^TX)^{-1}\sum_1^n \sigma_i^2 x_i x_i (X^TX)^{-1}.$$

This difficulty with v_b is due to its very nature. The drawing of i.i.d. samples from $\{r_i/(1 - kn^{-1})^{1/2}\}$ depends on the assumption that the residuals r_i are fairly exchangeable. Any inherent heterogeneity among r_i is lost in the process of i.i.d. sampling and will not be reflected in v_b. On the other hand, the jackknife method, by not mixing the residuals, allows the heterogeneity of r_i and σ_i^2 to be reflected in its variance estimate. This robustness aspect will be studied in Section 5.

Confidence intervals for θ can be obtained from the histogram of θ^* in (2.8). Repeat (2.7) and (2.8) B times. Define CDFB(t) to be the unweighted empirical distribution function based on the B bootstrap estimates $\{\theta^{*b}\}_{b=1}^{B}$. The *bootstrap percentile* method (Efron (1982)) consists of taking

$$[\mathrm{CDFB}^{-1}(\alpha), \mathrm{CDFB}^{-1}(1-\alpha)] \tag{2.10}$$

as an approximate $1-2\alpha$ central confidence interval for θ. The interval (2.10) is computed with a continuity correction.

The second bootstrap method is based on drawing i.i.d. sample $\{y_i^*, x_i^*\}_1^n$ from the pairs $\{(y_i, x_i)\}_1^n$. Compute the bootstrap LSE,

$$\beta^* = \left(\sum_1^n x_i^* x_i^{*T}\right)^{-1} \sum_1^n x_i^* y_i^*, \tag{2.11}$$

and the unweighted bootstrap variance estimator

$$\begin{aligned} v^* &= E_*(\theta^* - \hat{\theta})(\theta^* - \hat{\theta})^T, \qquad \hat{\theta} = g(\hat{\beta}), \\ \theta^* &= g(\beta^*), \qquad \beta^* \text{ in (2.11)}, \end{aligned} \tag{2.12}$$

where * denotes i.i.d. sampling from $\{(y_i, x_i)\}_1^n$. Recognizing the nonrobustness of the first bootstrap method, Efron and Gong (1983) seemed to favor the second method. However, like the unweighted jackknife v_J (2.3), the estimator (2.12) suffers from neglecting the unbalanced nature of the regression data $\{y_i, x_i\}$. Simple examples will be given in Section 6.2 to demonstrate its inconsistency and bias (even when $\sigma_i^2 = \sigma^2$). The simulation study of Section 10 does not paint a bright picture for v^*. A weighted modification to (2.12), inspired by the proposed weighted jackknife, is considered in (6.12). It gives generally biased variance estimators. However, the bias is very small for the particular example in the simulation study.

3. A class of representations for the least squares estimator. To motivate the general result, let us first consider the simple linear regression model $y_i = \alpha + \beta x_i + e_i$, $i = 1, \ldots, n$. The LSE $\hat{\beta}$ of the slope parameter has several equivalent expressions,

$$\begin{aligned} \hat{\beta} &= \sum_1^n (y_i - \bar{y})(x_i - \bar{x}) / \sum_1^n (x_i - \bar{x})^2 \\ &= \sum_{i<j}^n (y_i - y_j)(x_i - x_j) / \sum_{i<j}^n (x_i - x_j)^2 \\ &= \sum_{i<j} u_{ij} \hat{\beta}_{ij}, \end{aligned} \tag{3.1}$$

where $\hat{\beta}_{ij} = (y_i - y_j)/(x_i - x_j)$ are pairwise slopes for $x_i \neq x_j$, $u_{ij} = (x_i - x_j)^2 / \sum_{i<j}^n (x_i - x_j)^2$ and $u_{ij}\hat{\beta}_{ij}$ is defined to be zero for $x_i = x_j$. One can interpret $\hat{\beta}$ as a weighted average of the LSE β_{ij} based on the (i, j) pairs of observations, with weight proportional to $(x_i - x_j)^2$, which happens to be the determinant of the X^TX matrix of the (i, j) pair.

This representation has indeed a general version for any linear model (2.1). Let $s = (i_1, \ldots, i_r)$ be a subset of $(1, \ldots, n)$ and $\hat{\beta}_s$ be the LSE based on the (x_i, y_i) in s, i.e.,

$$\hat{\beta}_s = \left(X_s^T X_s\right)^{-1} X_s^T y_s,$$

where $y_s = (y_{i_1}, \ldots, y_{i_r})^T$ and $X_s = [x_{i_1}, \ldots, x_{i_r}]^T$. The following representation (3.2) says that the full-data LSE $\hat{\beta}$ is a weighted average of the subset LSE $\hat{\beta}_s$ over all the subsets of a fixed size with weight proportional to the determinant of $X_s^T X_s$. Throughout the paper, Σ_r denotes *the summation over all the subsets s of size r.*

THEOREM 1. *For any $r \geq k$,*

$$\hat{\beta} = \frac{\Sigma_r |X_s^T X_s| \hat{\beta}_s}{\Sigma_r |X_s^T X_s|} = \frac{\Sigma_r |X_s^T X_s| \hat{\beta}_s}{\binom{n-k}{r-k} |X^T X|}, \tag{3.2}$$

where $|X_s^T X_s| \hat{\beta}_s$ is defined to be zero for singular $X_s^T X_s$.

For a bibliographical note on this result see Farebrother (1985). Note that the second identity of (3.2) follows from Lemma 1(ii) to be stated later. The representation (3.2) can be extended to very general resampling plans. Before stating and proving this result in Theorem 2, we outline some further aspects of Theorem 1.

Theorem 1 can be extended to cover the weighted least squares estimator. Let $W = \operatorname{diag}(u_1, \ldots, u_n)$ be the diagonal matrix with elements $u_i > 0$ and W_s be its square submatrix corresponding to the set s. Let the full-data weighted LSE and subset weighted LSE be denoted by $\mathring{\beta} = (X^T W X)^{-1} X^T W y$ and $\mathring{\beta}_s = (X_s^T W_s X_s)^{-1} X_s^T W_s y_s$. By applying the transformation $W^{1/2}$ to X and y, and $W_s^{1/2}$ to X_s and y_s, Corollary 1 follows from Theorem 1.

COROLLARY 1. *For $r \geq k$,*

$$\mathring{\beta} = \frac{\Sigma_r |X_s^T W_s X_s| \mathring{\beta}_s}{\Sigma_r |X_s^T W_s X_s|}, \tag{3.3}$$

where the terms with singular $X_s^T W_s X_s$ are defined to be zero.

Formula (3.3) for $r = k$ is of particular interest, since $\mathring{\beta}_s$ is identical to the unweighted LSE $\hat{\beta}_s = X_s^{-1} y_s$ when X_s^{-1} exists. In this case the weighted LSE $\mathring{\beta}$ is a convex combination of the unweighted LSE $\hat{\beta}_s$. As a consequence, the collection of the *weighted* LSE's is contained in the bounded convex hull spanned by the *finite* number of *unweighted* LSE's based on all the subsets of size k. Rubin (1978) proved this result and noted its use in proving the convergence of certain iterative reweighted least squares algorithms (Dempster, Laird and Rubin (1980)).

The representation (3.2) also suggests a host of robust modifications to the ordinary LSE. The basic idea is to trim the more extreme $\hat{\beta}_s$ and take a weighted average of the remaining ones or to use weighted M-estimation based on the $\hat{\beta}_s$. Details can be found in Wu (1984).

Before stating Theorem 2, we need to introduce the concepts of resampling. A resample of $z_1 = (y_1, x_1), \ldots, z_n = (y_n, x_n)$ is a reweighted version of $(z_i)_1^n$ with weight $P_i^* > 0$. The vector $P^* = (P_1^*, \ldots, P_n^*)$ is called a *resampling vector*. For each P^*, the corresponding LSE β^* is based on P_i^* "copies" of z_i, i.e.,

$$\beta^* = (X^T D^* X)^{-1} X^T D^* y, \qquad D^* = \operatorname{diag}(P_1^*, \ldots, P_n^*) \tag{3.4}$$

is a weighted LSE with weight proportional to P_i^*. The full-data LSE $\hat{\beta}$ corresponds to $P_i^* = 1/n$. Let * denote the joint distribution of $(P_i^*)_1^n$ under a resampling procedure. The expectation under * is denoted by E_*.

ASSUMPTIONS ON THE RESAMPLING PROCEDURE *:

(A) $E_*(\prod_{j=1}^k P_{i_j}^*) = a_k > 0$, independent of the subset $(i_1, \ldots, i_k)$, k = number of parameters in (2.1).

It is easy to see that (A) is implied by (B).

(B) 1. The n random variables $\{P_i^*\}_1^n$ are exchangeable.
2. $\text{Prob}_*(\text{support size of } P^* \geq k) > 0$, where the support size of P^* is the total number of i's with $P_i^* > 0$.

Condition (B1) says that the resampling plan is symmetric; (B2) is a minimal condition to ensure that at least some resamples have support of size $\geq k$.

The next result states that the full-data LSE $\geq$ is a weighted average of the resampled-data LSE β^* (3.4) with weight proportional to $|X^T D^* X|$ for any resampling procedure satisfying (A). Another interpretation (suggested by B. Efron) is that, in the resampling world, $|X^T D^* X|\beta^*/E_*|X^T D^* X|$ is an "unbiased" estimator of $\hat{\beta}$.

THEOREM 2. *For any resampling method * satisfying the assumption* (A),

$$\hat{\beta} = \frac{E_*|X^T D^* X|\beta^*}{E_*|X^T D^* X|}, \tag{3.5}$$

where $|X^T D^* X|\beta^*$ is defined to be zero if $X^T D^* X$ is singular.

Theorem 1 is a special case of Theorem 2 since the resampling procedure in Theorem 1, which selects every subset of size $r \geq k$ with equal probability, satisfies (B). Another important resampling procedure that satisfies (B) is the bootstrap (Efron (1979)). A bootstrap sample is obtained from drawing an i.i.d. sample $z_1^*, \ldots, z_n^*$ from $(z_i)_1^n$. The resampling vector $P^* = (P_1^*, \ldots, P_n^*)$ has a multinomial distribution with n draws on n categories each having probability $1/n$.

Before proving Theorem 2, we state two matrix lemmas. For an $n \times k$ matrix X let X_s be its $r \times k$ submatrix consisting of the $(i_1, \ldots, i_r)$th rows, $s = (i_1, \ldots, i_r)$, and let $X^{(j)}$ be the $n \times (k-1)$ submatrix obtained from deleting the jth column of X. For a square matrix A, let adj A be its adjoint.

LEMMA 1. *Let X and Z be $n \times k$ matrices, $n \geq k$. Then*

(i) $$|X^T Z| = \Sigma_k |X_s||Z_s|,$$

(ii) $$|X^T Z| = \binom{n-k}{r-k}^{-1} \Sigma_r |X_s^T Z_s| \quad \text{for any } r \geq k,$$

where Σ_r denotes the summation over all the subsets of size r.

Lemma 1(i) is the Binet–Cauchy expansion (Noble (1969), page 226). Lemma 1(ii) is obtained by applying Lemma 1(i) to each term $|X_s^T Z_s|$ and to $|X^T Z|$.

LEMMA 2. *Let X be an $n \times k$ matrix, $n \geq k$. Then for $r \geq k$,*

$$\text{adj } X^T X = \binom{n-k+1}{r-k+1}^{-1} \Sigma_r \text{adj } X_s^T X_s. \tag{3.6}$$

If $X_s^T X_s$ are nonsingular for all s of size r, $r \geq k$,

$$|X^T X|(X^T X)^{-1} = \binom{n-k+1}{r-k+1}^{-1} \Sigma_r |X_s^T X_s|(X_s^T X_s)^{-1}. \tag{3.7}$$

PROOF. From definition, the (i, j)th elements of adj $X^T X$ and adj $X_s^T X_s$ are $(-1)^{i+j}|X^{(j)T}X^{(i)}|$ and $(-1)^{i+j}|X_s^{(j)T}X_s^{(i)}|$, respectively, where $X_s^{(i)}$ is obtained by deleting the ith column of X_s. Therefore, (3.6) is equivalent to

$$|X^{(j)T}X^{(i)}| = \binom{n-k+1}{r-k+1}^{-1} \Sigma_r |X_s^{(j)T}X_s^{(i)}|,$$

which follows from Lemma 1(ii), noting that $X^{(j)T}X^{(i)}$ and $X_s^{(j)T}X_s^{(i)}$ are both of order $k-1$. (3.7) follows from (3.6) since adj $A = |A|A^{-1}$. □

PROOF OF THEOREM 2. First consider the D^* with nonsingular $X^T D^* X$. Since β^* is the solution to the equation $(X^T D^* X)\beta^* = X^T D^* y$, from Cramer's rule (Noble (1969), page 209), the jth element of β^* is equal to the ratio of the determinant of the matrix obtained by replacing the jth column of $X^T D^* X$ by the vector $X^T D^* y$ over the determinant of $X^T D^* X$. Notationally,

$$\beta_j^* = |X^T D^* X^{(j)}(y)|/|X^T D^* X|,$$

where $X^{(j)}(y)$ is the matrix obtained by replacing the jth column of X by y. This establishes

$$|X^T D^* X^{(j)}(y)| = |X^T D^* X|\beta_j^* \quad \text{for nonsingular } X^T D^* X. \tag{3.8}$$

For singular $X^T D^* X$, $X^T D^*$ is not of full rank and , therefore,

$$|X^T D^* X^{(j)}(y)| = 0. \tag{3.9}$$

From (3.8) and (3.9), the jth element of the right side of (3.5) equals

$$\frac{E_*|X^TD^*X^{(j)}(y)|}{E_*|X^TD^*X|} = \frac{E_*\Sigma_k|X_s||D_s^*||X_s^{(j)}(y)|}{E_*\Sigma_k|X_s|^2|D_s^*|}, \tag{3.10}$$

where $D_s^* = \text{diag}(P_{i_1}^*, \ldots, P_{i_k}^*)$ is the diagonal submatrix of D^* corresponding to s and (3.10) follows from Lemma 1(i). Since $E_*|D_s^*| = E_*(\Pi_{j=1}^k P_{i_j}^*) = a_k > 0$ independent of s from the assumption (A), (3.10) equals

$$\frac{\Sigma_k|X_s|\,|X_s^{(j)}(y)|}{\Sigma_k|X_s|^2} = \frac{|X^TX^{(j)}(y)|}{|X^TX|},$$

which is the jth element of $\hat{\beta}$. This completes the proof. □

4. General weighted jackknife in regression. We propose in this section a general weighted jackknife method that does not have some of the undesirable properties possessed by the three existing methods discussed in Section 2.

Our proposal is mathematically motivated by the representation (3.2) for the LSE $\hat{\beta}$, which can be rewritten as

$$\Sigma_r|X_s^TX_s|(\hat{\beta}_s - \hat{\beta}) = 0.$$

As its second-order analog, we propose

$$v_{J,r}(\hat{\theta}) = \frac{r-k+1}{n-r}\Sigma_r w_s(\hat{\theta}_s - \hat{\theta})(\hat{\theta}_s - \hat{\theta})^T, \qquad w_s \propto |X_s^TX_s|,\ \Sigma_r w_s = 1, \tag{4.1}$$

$$= \binom{n-k}{r-k+1}^{-1}|X^TX|^{-1}\Sigma_r|X_s^TX_s|(\hat{\theta}_s - \hat{\theta})(\hat{\theta}_s - \hat{\theta})^T \tag{4.2}$$

as a jackknife estimator of the variance of $\hat{\theta} = g(\hat{\beta})$, where g is a smooth function of β, $\hat{\theta}_s = g(\hat{\beta}_s)$ and Σ_r is the summation over all the subsets of size r. Throughout the paper, w_s denotes the weight defined in (4.1). Formula (4.2) follows from (4.1) by Lemma 1(ii). In (4.1) and (4.2), $X_s^TX_s$ are assumed nonsingular for all s. In (4.1) the scale factor $\sqrt{r-k+1}/\sqrt{n-r}$ is applied *externally* to $\hat{\theta}_s - \hat{\theta}$ after the transformation g. Another variance estimator is obtained by applying the same factor *internally* to $\hat{\beta}_s - \hat{\beta}$,

$$\tilde{v}_{J,r}(\hat{\theta}) = \Sigma_r w_s(\tilde{\theta}_s - \hat{\theta})(\tilde{\theta}_s - \hat{\theta})^T, \tag{4.3}$$

$$\tilde{\theta}_s = g(\tilde{\beta}_s), \qquad \tilde{\beta}_s = \hat{\beta} + \left(\frac{r-k+1}{n-r}\right)^{1/2}(\hat{\beta}_s - \hat{\beta}). \tag{4.4}$$

Under reasonable smoothness conditions on g, both $v_{J,r}(\hat{\theta})$ and $\tilde{v}_{J,r}(\hat{\theta})$ will be close to the linearized jackknife variance estimator $g'(\hat{\beta})^Tv_{J,r}(\hat{\beta})g'(\hat{\beta})$, where $g'(\hat{\beta})$ is the derivative of g at $\hat{\beta}$.

The linearization (or δ-method) variance estimator is given by

$$v_{\text{lin}} = g'(\hat{\beta})^T\hat{v}g'(\hat{\beta}), \qquad \hat{v} \text{ given in (2.9)}. \tag{4.5}$$

It is based on the assumption of homoscedastic errors. A simulation comparison of these estimators will be given in Section 10.

If r is chosen to be $(n + k - 1)/2$, the scale factor $(r - k + 1)/(n - r)$ becomes one and $v_{J,r} = \tilde{v}_{J,r}$. For $k = 1$, the subset is a half-sample.

Three features of $v_{J,r}$ and $\tilde{v}_{J,r}$ deserve further discussion.

(i) *The weight is proportional to the determinant of the Fisher information matrix of the subset.* This is because the information matrix of the model $y_s = X_s^T\beta + e_s$ for s is $cX_s^TX_s$, where c is the Fisher information of the error term e_i. It is a *scalar* weight no matter what the dimension of $\theta = g(\beta)$. This interpretation of w_s in (4.1) allows $v_{J,r}$ or $\tilde{v}_{J,r}$ to be extended to general nonlinear situations. See Section 8.

(ii) *Flexible choice of subset size and construction of histogram for interval estimation.* The choice of subset size is general. As pointed out in Section 1, there are theoretical difficulties in using the delete-one jackknife for interval estimation and for variance estimation for nonsmooth estimators such as the sample median. For the one-sample problem the jackknife with r around $0.72n$ was shown to possess a desirable second-order asymptotic property (Wu (1985)).

One advantage of the bootstrap over the delete-one jackknife is that the former can construct interval estimates (see (2.10)) based on the bootstrap histogram, which reflects the skewness of the original point estimator. This advantage is shared by the general jackknife. A weighted jackknife distribution can be constructed as follows.

(1) Draw subsets $s_1, \ldots, s_J$ of size r randomly without replacement.
(2) Construct a weighted empirical distribution function CDFJ(t) based on $g(\tilde{\beta}_{s_i})$, $\tilde{\beta}_{s_i}$ defined in (4.4), $i = 1(1)J$, with weight proportional to $|X_{s_i}^TX_{s_i}|$.

Similar to the bootstrap percentile method (2.10) is the *jackknife percentile* method consisting of taking

$$[\mathrm{CDFJ}^{-1}(\alpha), \mathrm{CDFJ}^{-1}(1 - \alpha)] \tag{4.6}$$

as an appropriate $1 - 2\alpha$ central confidence interval for θ. Since CDFJ(t) is discrete, (4.6) is computed with a continuity correction. For multiparameters θ, a confidence region can be similarly constructed once the shape of the region is determined. Efron ((1982), Chapter 10) considered other modifications to the bootstrap percentile method. Similar modifications to the jackknife percentile method can be obtained in a straightforward manner. Here the external scaling method (4.1) is not appealing since the histogram of $\{g(\hat{\beta}_{s_i})\}$ has a wrong scale and stretching or shrinking by the factor $\sqrt{r - k + 1}\,/\,\sqrt{n - r}$ seems arbitrary. On the other hand, the internal scaling method (4.3) does not perform well in the simulation study (Section 10). One may use $r = (n + k - 1)/2$ to avoid use of the scale factor.

(iii) *The scale factor* $\xi = (r - k + 1)/(n - r)$. For general r, the resampling error $\hat{\beta}_s - \hat{\beta}$ has a different stochastic order from that of the sampling error $\hat{\beta} - \beta$. This is corrected by multiplying $\hat{\beta}_s - \hat{\beta}$ by $\sqrt{\xi}$. In fact, from the

unbiasedness of $v_{J,r}(\hat{\beta})$ (to be proved in Theorem 3),

$$\Sigma_r w_s \mathrm{Var}\big(\sqrt{\xi}\,(\hat{\beta}_s - \hat{\beta})\big) = \mathrm{Var}(\hat{\beta} - \beta), \tag{4.7}$$

where the weight w_s is proportional to $|X_s^T X_s|$. In particular, for $k = 1$, since w_s is constant, this reduces to $\mathrm{Var}(\sqrt{\xi}\,(\hat{\beta}_s - \hat{\beta})) = \mathrm{Var}(\hat{\beta} - \beta)$. In general if the weights w_s in (4.7) are uniformly bounded away from 0 and 1, (4.7) implies

$$\mathrm{Var}\big(\sqrt{\xi}\,(\hat{\beta}_s - \hat{\beta})\big) = \mathrm{Var}(\hat{\beta} - \beta)(1 + O(1)) \quad \text{for all } s. \tag{4.8}$$

The implementation of the proposed jackknife method may be quite cumbersome since $\binom{n}{r}$ computations are required. As in the bootstrap method, Monte Carlo approximation by randomly selecting J distinct subsets, $J \ll \binom{n}{r}$, can be used. Construction of $v_{J,r}$ or $\tilde{v}_{J,r}$ from these J subsamples is obvious.

We now turn to the theoretical aspect of the proposed jackknife method. Its use for bias reduction will be studied in Section 9. For the linear parameter $\theta = \beta$, $v_{J,r}(\hat{\beta})$, which is identical to $\tilde{v}_{J,r}(\hat{\beta})$, has three properties. For simplicity we use $v_{J,r}$ for $v_{J,r}(\hat{\beta})$.

(i) It is unbiased for $\mathrm{Var}(\hat{\beta})$ if $\mathrm{Var}(e) = \sigma^2 I$ (Theorem 3).
(ii) A suitably defined version of $v_{J,k}$ is identical to the usual variance estimator $\hat{\sigma}^2(X^TX)^{-1}$ (2.9) (Theorem 4).
(iii) $v_{J,n-1}$ is robust against error variance heteroscedasticity (Section 5).

THEOREM 3. *If* $\mathrm{Var}(e) = \sigma^2 I$ *in* (2.1),

$$E(v_{J,r}) = \sigma^2(X^TX)^{-1} = \mathrm{Var}(\hat{\beta}). \tag{4.9}$$

PROOF. From

$$\hat{\beta}_s - \hat{\beta} = (X_s^T X_s)^{-1} X_s^T (y_s - X_s\hat{\beta}) = (X_s^T X_s)^{-1} X_s^T r_s,$$

where $r_s = y_s - X_s\hat{\beta}$ is the residual vector for s,

$$|X_s^T X_s|(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T = |X_s^T X_s|(X_s^T X_s)^{-1} X_s^T r_s r_s^T X_s (X_s^T X_s)^{-1}$$

and its expectation is

$$\begin{aligned} &|X_s^T X_s|(X_s^T X_s)^{-1} X_s^T \big(I_s - X_s(X^TX)^{-1}X_s^T\big) X_s (X_s^T X_s)^{-1} \\ &\quad = |X_s^T X_s|(X_s^T X_s)^{-1} - |X_s^T X_s|(X^TX)^{-1}, \end{aligned} \tag{4.10}$$

where I_s is the identity matrix for s. From Lemma 2,

$$\Sigma_r |X_s^T X_s|(X_s^T X_s)^{-1} = \binom{n-k+1}{r-k+1} |X^TX|(X^TX)^{-1}, \tag{4.11}$$

and from Lemma 1(ii),

$$\Sigma_r |X_s^T X_s|(X^TX)^{-1} = \binom{n-k}{r-k} |X^TX|(X^TX)^{-1},$$

which, together with (4.10), imply

$$E\left(\Sigma_r |X_s^T X_s|(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \beta)^T\right) = \binom{n-k}{r-k+1} |X^T X|(X^T X)^{-1},$$

and thus the result. □

When the subset size r equals the number of parameters k, the jackknife variance estimator $v_{J,k}$ can be redefined *without* the additional assumption that $X_s^T X_s$ is nonsingular for any s. For a subset s of size k,

$$\hat{\beta}_s = X_s^{-1} y_s, \qquad \hat{\beta}_s - \hat{\beta} = X_s^{-1} r_s, \qquad r_s = y_s - X_s \hat{\beta}$$

and

$$|X_s^T X_s|(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T = |X_s|^2 X_s^{-1} r_s r_s^T (X_s^T)^{-1} = (\text{adj } X_s) r_s r_s^T (\text{adj } X_s)^T.$$

Note that adj X_s, the adjoint of X_s, is always defined, whereas $|X_s| X_s^{-1}$ is defined only for nonsingular X_s. This suggests defining a more general variance estimator for $r = k$,

$$v'_{J,k} = \frac{1}{(n-k)|X^T X|} \Sigma_k (\text{adj } X_s) r_s r_s^T (\text{adj } X_s)^T. \tag{4.12}$$

Note that $v_{J,k}$ in (4.1) requires the nonsingularity of $X_s^T X_s$ for every s while $v'_{J,k}$ is well-defined without additional restriction. For $v'_{J,k}$ we can establish the following coincidental result, which also implies the conclusion of Theorem 3 for $v_{J,k}$.

THEOREM 4. *The estimator $v'_{J,k}$ is identical to the estimator $\hat{v} = \hat{\sigma}^2 (X^T X)^{-1}$.*

PROOF. Note that

$$(\text{adj } X_s) r_s = \left[(-1)^{i+j} |X_{s(j)}^{(i)}|\right]_{i,j} (r_{sj})_j \tag{4.13}$$

$$= \left(\sum_j (-1)^{i+j} r_{sj} |X_{s(j)}^{(i)}|\right)_i = \left(|X_s^{(i)}(r_s)|\right)_i, \tag{4.14}$$

where $X_{s(j)}^{(i)}$ is the matrix obtained by deleting the jth row and the ith column of X_s, $r_{sj} = j$th element of r_s and $X_s^{(i)}(r_s)$ is the matrix obtained by replacing the ith column of X_s by r_s. The last equation of (4.14) follows from a standard expansion of determinant (Noble (1969), page 208). From (4.14), the (i, j)th element of the matrix $\Sigma_k (\text{adj } X_s) r_s r_s^T (\text{adj } X_s)^T$ in (4.12) is equal to

$$\Sigma_k |X_s^{(i)}(r_s)||X_s^{(j)}(r_s)| = |X^{(i)}(\mathbf{r})^T X^{(j)}(\mathbf{r})|, \tag{4.15}$$

where $X^{(i)}(\mathbf{r})$ is the matrix obtained by replacing the ith column of X by the residual vector $\mathbf{r} = y - X\hat{\beta}$. Since $X_s^{(i)}(r_s)$ is the $k \times k$ submatrix of $X^{(i)}(\mathbf{r})$ with rows corresponding to s, (4.15) follows from Lemma 1(i). Noting that $\mathbf{r}$ is

orthogonal to the other columns of $X^{(i)}(\mathbf{r})$ from the normal equation $X^T\mathbf{r} = 0$, the (i, j)th element of $X^{(i)}(\mathbf{r})^T X^{(j)}(\mathbf{r})$ is $\mathbf{r}^T\mathbf{r}$, and the other elements in its ith row and jth column are zero. This gives

$$|X^{(i)}(\mathbf{r})^T X^{(j)}(\mathbf{r})| = (-1)^{i+j}\mathbf{r}^T\mathbf{r}|X^{(i)T}X^{(j)}|, \tag{4.16}$$

where $X^{(i)}$ is the submatrix of X with its ith column deleted. From (4.12), (4.15) and (4.16) we have

$$\begin{aligned} v'_{J,k} &= \frac{\mathbf{r}^T\mathbf{r}}{(n-k)|X^TX|}\left[(-1)^{i+j}|X^{(i)T}X^{(j)}|\right]_{i,j} \\ &= \frac{\mathbf{r}^T\mathbf{r}}{(n-k)}\frac{\operatorname{adj} X^TX}{|X^TX|} = \hat{\sigma}^2(X^TX)^{-1}. \end{aligned}$$ □

Theorem 4 was proved by Subrahmanyam (1972) for $v_{J,k}$ (not the more general $v'_{J,k}$) by assuming $|X_s| \neq 0$ for all s. For s with $|X_s| = 0$, it is incorrect to interpret $|X_s|^2(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T$ in $v_{J,k}$ to be zero as was done before for the representation theorem. This is obvious since the more general expression $(\operatorname{adj} X_s)r_s r_s^T(\operatorname{adj} X_s)^T$ in (4.12) is nonnegative definite and is generally nonzero for singular X_s. Such an incorrect interpretation of $v_{J,k}$ will lead to a variance estimator smaller than $\hat{\sigma}^2(X^TX)^{-1}$. A simple example was given in Wu (1984).

5. Bias-robustness of weighted delete-one jackknife variance estimators. If the homoscedasticity assumption in Theorem 3 is violated, will $v_{J,r}$ be approximately unbiased? In this section we will show that the two weighted delete-one jackknife estimators $v_{J,n-1}$ (4.1) and $v_{H(1)}$ (2.4) have this desirable property.We adopt the conventional notation that the subscript (i) denotes "with the ith observation deleted," and in a similar spirit, use $v_{J(1)}$ for $v_{J,n-1}$. It is easy to show that

$$v_{J(1)} = \sum_1^n (1 - w_i)\big(\hat{\beta}_{(i)} - \hat{\beta}\big)\big(\hat{\beta}_{(i)} - \hat{\beta}\big)^T \tag{5.1}$$

$$= (X^TX)^{-1}\sum_1^n \frac{r_i^2}{1 - w_i}x_i x_i^T(X^TX)^{-1}, \tag{5.2}$$

where $\hat{\beta}_{(i)}$ is the LSE of β with the ith observation deleted, $r_i = y_i - x_i^T\hat{\beta}$ is the ith residual and $w_i = x_i^T(X^TX)^{-1}x_i$.

Under $\operatorname{Var}(e) = \operatorname{diag}(\sigma_i^2)$, the variance of $\hat{\beta}$ is

$$\operatorname{Var}(\hat{\beta}) = (X^TX)^{-1}\sum_1^n \sigma_i^2 x_i^T(X^TX)^{-1}. \tag{5.3}$$

Comparison of (5.2) and (5.3) suggests that $v_{J(1)}$ is robust for estimating $\operatorname{Var}(\hat{\beta})$ under the broader heteroscedasticity assumption.

The asymptotic computations will be done under several of the following assumptions.

(C) 1. Let X_n denote the X matrix in (2.1) for n observations; $\max_{1 \le i \le n} x_i^T (X_n^T X_n)^{-1} x_i \le c/n$, c independent of n.
2. $\max_{1 \le i < \infty} \sigma_i^2 < \infty$.
3. The minimum and maximum eigenvalues of $n^{-1} X_n^T X_n$ are uniformly bounded away from 0 and ∞.
4. The elements of X_n are uniformly bounded.

The unbiasedness of $V_{J(1)}$ for estimating $\mathrm{Var}(\hat{\beta})$ hinges on the relation $Er_i^2 = (1 - w_i)\sigma_i^2$. Conditions for its validity or approximate validity are given in the next lemma.

LEMMA 3. *If*

$$w_{ij} = x_i^T (X^T X)^{-1} x_j = 0 \quad \text{for any } i, j \text{ with } \sigma_i \ne \sigma_j, \tag{5.4}$$

then

$$Er_i^2 = (1 - w_i)\sigma_i^2, \qquad w_i = x_i^T (X^T X)^{-1} x_i. \tag{5.5}$$

More generally, under the assumptions (C1) *and* (C2),

$$Er_i^2 = (1 - w_i)\sigma_i^2 + O(n^{-1}), \tag{5.6}$$

where the big O-notation $O(n^{-1})$ denotes terms of order n^{-1}.

PROOF. From $r_i = y_i - x_i^T \hat{\beta} = e_i - x_i^T (X^T X)^{-1} X^T e$,

$$Er_i^2 = \sigma_i^2 - 2w_i\sigma_i^2 + \sum_{j=1}^{n} w_{ij}^2 \sigma_j^2 = (1 - w_i)\sigma_i^2 + \sum_{j=1}^{n} w_{ij}^2 \left(\sigma_j^2 - \sigma_i^2\right), \tag{5.7}$$

where $w_{ij} = x_i^T (X^T X)^{-1} x_j$ and the second equality of (5.7) follows from $w_i = \sum_{j=1}^{n} w_{ij}^2$. It is now obvious that (5.5) follows from (5.4). Assuming (C1) and (C2),

$$\left| \sum_j w_{ij}^2 \left(\sigma_j^2 - \sigma_i^2\right) \right| \le 2\left(\max_i \sigma_i^2\right) \sum_j w_{ij}^2 = 2\left(\max_i \sigma_i^2\right) w_i,$$

which is of order n^{-1}. Therefore, (5.6) follows from (5.7). □

By comparing (5.2) and (5.3), the following result is obtained as a direct consequence of Lemma 3.

THEOREM 5. *Under* (2.1),

(i) $Ev_{J(1)} = \mathrm{Var}(\hat{\beta})$ *under* (5.4);
(ii) $Ev_{J(1)} = \mathrm{Var}(\hat{\beta})(1 + O(n^{-1}))$ *under* (C1) *and* (C2).

A similar result for $v_{J,r}$ for a broad range of r values was proved in Shao and Wu (1985). It is also supported by the simulation study of Section 10.

The assumption (C2) is weak; (C1) is also reasonable since it is easy to show that it is implied by (C3) and (C4). (C3) says that $X_n^T X_n$ grows to infinity at the rate n. Usually a stronger condition such as $n^{-1}X_n^T X_n$ converging to a positive definite matrix is assumed (Miller (1974b)). On the other hand, (5.4) is more restrictive. Let q be the number of different σ_i's in (5.4). Then the linear model (2.1) can be rewritten as

$$(5.8) \qquad y_{ij} = x_{ij}^T\beta + e_{ij}, \qquad \mathrm{Var}(e_{ij}) = \sigma_i^2, \qquad j = 1(1)n_i,\ i = 1(1)q.$$

Let T_i be the subspace spanned by x_{ij}, $j = 1(1)n_i$. According to (5.4) T_i, $i = 1(1)q$, are orthogonal to each other with respect to the positive definite matrix $(X^TX)^{-1}$. A special case of (5.8) is the k-sample problem with unequal variances

$$(5.9) \qquad y_{ij} = \theta_i + e_{ij}, \qquad \mathrm{Var}(e_{ij}) = \sigma_i^2, \qquad j = 1(1)n_i,\ i = 1(1)k.$$

Closely related to $v_{J(1)}$ is $v_{H(1)}$. From comparing $v_{H(1)}$ (2.4) and $v_{J(1)}$ (5.2), it seems that $v_{H(1)}$ is also robust in the sense of Theorem 5(ii). The comparison is, however, more favorable for $v_{J(1)}$. Under the ideal assumption $\mathrm{Var}(e) = \sigma^2 I$, $Er_i^2 = (1 - w_i)\sigma^2 \neq (1 - n^{-1}k)\sigma^2$. Therefore, $Ev_{H(1)} \neq \sigma^2(X^TX)^{-1}$, although under (C1) the difference is of lower order, i.e., $Ev_{H(1)} = \sigma^2(X^TX)^{-1}(1 + O(n^{-1}))$ since $Er_i^2 - (1 - n^{-1}k)\sigma^2 = (n^{-1}k - w_i)\sigma^2 = O(n^{-1})$ under (C1). Under the broader assumption

$$\mathrm{Var}(e) = \mathrm{diag}(\sigma_i^2), \qquad Ev_{H(1)} \neq \mathrm{Var}(\hat{\beta})$$

even under the condition (5.4) in Theorem 5. As in Theorem 5(ii), $v_{H(1)}$ is approximately unbiased under (C1) and (C2), i.e., $Ev_{H(1)} = \mathrm{Var}(\hat{\beta})(1 + O(n^{-1}))$. This is because

$$E\frac{r_i^2}{1 - n^{-1}k} = \frac{1 - w_i}{1 - n^{-1}k}\sigma_i^2 + O(n^{-1}) = \sigma_i^2 + O(n^{-1}),$$

where the first equation follows from Lemma 3, (5.6), and the second equation follows from (C1). The results concerning $v_{H(1)}$ are summarized as follows.

THEOREM 6. (i) *Under* $\mathrm{Var}(e) = \sigma^2 I$ *and* (C1), $Ev_{H(1)} \neq \mathrm{Var}(\hat{\beta})$ *but*

$$Ev_{H(1)} = \mathrm{Var}(\hat{\beta})\big(1 + O(n^{-1})\big).$$

(ii) *Under* $\mathrm{Var}(e) = \mathrm{diag}(\sigma_i^2)$, (C1) *and* (C2),

$$Ev_{H(1)} = \mathrm{Var}(\hat{\beta})\big(1 + O(n^{-1})\big).$$

Unlike $v_{J(1)}$, the exact unbiasedness $Ev_{H(1)} = \mathrm{Var}(\hat{\beta})$ does not hold true even in special cases. Theorem 6 is a more rigorous version of what is essentially in Hinkley (1977). The strong consistency of $v_{H(1)}$ was established in Hinkley (1977) by following Miller's (1974b) proof for the balanced jackknife. The strong consistency of $v_{J(1)}$ can be established in a similar manner.

Standard asymptotic justifications for the jackknife variance estimators are in terms of its consistency and the normality of the associated t-statistics. They confirm that the jackknife method works asymptotically as well as the classical δ-method. Then why should the jackknife be chosen over the δ-method except possibly for computational or other practical reasons? The bias-robustness of $v_{J(1)}$ and $v_{H(1)}$ (Theorems 5(ii) and 6(ii)) against the heteroscedasticity of errors, first recognized in Hinkley (1977), is a fresh and important property of the jackknife methodology.

Before closing this section, we make two other remarks.

(i) *Is the concept of pseudovalues appropriate for non-i.i.d. situations*? Tukey's (1958) reformulation of Quenouille's (1956) jackknife in terms of the pseudovalues works well for the i.i.d. case. Its extension to non-i.i.d. situations lack firm theoretical foundation. In fact it may lead to less desirable results as is evidenced by the slight inferiority of $v_{H(1)}$ to $v_{J(1)}$. A more striking example is offered in the context of inference from stratified samples. Two jackknife point estimators have been proposed in terms of pseudovalues, both of which reduce to the usual jackknife point estimator in the unstratified case. It was found (Rao and Wu (1985b)) that neither estimator reduces bias as is typically claimed for the jackknife.

(ii) *Relation to quadratic unbiased variance estimators.* If the purpose of jackknife variance estimation is to aid the point estimator $\hat{\beta}$ in making inference about β, the variance estimators are required to be nonnegative and almost unbiased. In situations such as the determination of sample size, the variance itself is the parameter of primary interest and other criteria such as the mean squared error (MSE) will be more appropriate. In this context, a nonnegative biased estimator (Rao (1973)) and MINQUE (Rao (1970)) were proposed. Horn, Horn and Duncan (1975) proposed $(1 - w_i)^{-1}r_i^2$, which appears in $v_{J(1)}$ (5.2), as an estimator of σ_i^2 and called it AUE (almost unbiased estimator). The MSE of $(1 - w_i)^{-1}r_i^2$ was shown to be smaller than that of MINQUE in a wide range of situations (Horn and Horn (1975)). However, it is difficult to extend this comparison to estimation of the variance–covariance matrix.

6. Variable jackknife and bootstrap. Can the results in Sections 4 and 5 be extended to other resampling methods? For a given resampling method denoted by *, β^* and D^* defined in (3.4), we would like to find a variance estimator of the form

$$v = \lambda E_* w_* (\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T, \tag{6.1}$$

where the weight w_* is proportional to $|X^T D^* X|$ and $E_* w_* = 1$, such that it satisfies the minimal requirement (as in Theorem 3)

$$E\big(v|\mathrm{Var}(e) = \sigma^2 I\big) = \sigma^2 (X^T X)^{-1}. \tag{6.2}$$

The left side of (6.2) is equal to

$$\lambda\sigma^2 E_*\{w_*(X^TD^*X)^{-1}X^TD^{*2}X(X^TD^*X)^{-1} - w_*(X^TX)^{-1}\}$$
$$= \lambda\sigma^2\{E_*[w_*(X^TD^*X)^{-1}X^TD^{*2}X(X^TD^*X)^{-1}] - (X^TX)^{-1}\}. \tag{6.3}$$

The first term inside the curly bracket of (6.3) seems intractable except for

$$D^{*2} = D^* = \operatorname{diag}(P_1^*, \ldots, P_n^*), \tag{6.4}$$

which is equivalent to $P_i^* = 0$ or 1 for all i. This prompts us to consider procedures satisfying assumption (B) and (6.4), whose defining probabilities

$$\operatorname{Prob}_*(P_{i_1}^* = \cdots = P_{i_r}^* = 1, \text{remaining } P_j^* = 0) = c_r \Big/ \binom{n}{r} \tag{6.5}$$

are independent of the subset $(i_1, \ldots, i_r)$, where c_r is the probability of selecting a subset of size r, $c_k + \cdots + c_n = 1$. When $c_r = 1$, (6.5) reduces to the jackknife with subset size r.

6.1. *Variable jackknife.* Since the constant c_r in (6.5) varies with r, we shall call any procedure satisfying (B) and (6.4) a *variable jackknife*. Continuing from (6.3), its first term under (6.4), apart from $\lambda\sigma^2$, is

$$E_*|X^TD^*X|(X^TD^*X)^{-1}/E_*|X^TD^*X| = E_*\operatorname{adj}(X^TD^*X)/E_*|X^TD^*X|,$$

whose (i, j)th element is

$$(-1)^{i+j}E_*|X^{(j)T}D^*X^{(i)}|/E_*|X^TD^*X|$$
$$= (-1)^{i+j}\Sigma_{k-1}|X_s^{(j)}|E_*|D_s^*||X_s^{(i)}|/\Sigma_k|X_s|^2E_*|D_s^*| \tag{6.6}$$
$$= \frac{\alpha_{k-1}}{\alpha_k}\frac{(-1)^{i+j}|X^{(j)T}X^{(i)}|}{|X^TX|} = \frac{\alpha_{k-1}}{\alpha_k}, \tag{6.7}$$

where the expansion in (6.6) is justified by Lemma 1(i) and

$$\alpha_i = \operatorname{Prob}_*(P_1^* = P_2^* = \cdots = P_i^* = 1) = \sum_{r=k}^{n} c_r\binom{n}{r}^{-1}\binom{n-i}{r-i}. \tag{6.8}$$

From (6.1), (6.3) and (6.7), the variance estimator

$$\left(\frac{\alpha_{k-1}}{\alpha_k} - 1\right)^{-1}\frac{E_*|X^TD^*X|(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T}{E_*|X^TD^*X|} \tag{6.9}$$

satisfies the unbiasedness requirement (6.2). For the special case of jackknifing with subset size r, $\alpha_{k-1}/\alpha_k - 1 = (n-r)/(r-k+1)$ and (6.9) reduces to $v_{J,r}$.

Note that the scale factor in (6.9),

$$\left(\frac{\alpha_{k-1}}{\alpha_k} - 1\right)^{-1} = \frac{\alpha_k/\alpha_{k-1}}{1 - \alpha_k/\alpha_{k-1}}$$
$$= \frac{\operatorname{Prob}_*(P_k^* = 1 | P_1^* = \cdots = P_{k-1}^* = 1)}{\operatorname{Prob}_*(P_k^* = 0 | P_1^* = \cdots = P_{k-1}^* = 1)}, \tag{6.10}$$

is a conditional *odds ratio* given that the first $k-1$ units have been selected. For jackknifing with subset size r, this alternative interpretation of the scale factor $(r-k+1)/(n-r)$ may be of interest.

6.2. *Bootstrap.* Among the resampling procedures that do not satisfy (6.4), i.e., $\text{Prob}_*(P_i^* \geq 2 \text{ for some } i) > 0$, we single out the bootstrap for further study. The resampling vector $P^* = (P_1^*, \ldots, P_n^*)$ has the multinomial distribution $\text{Mult}_n(n, (1/n)\mathbf{1})$.

The unweighted and weighted bootstrap variance estimators are, respectively,

$$v^* = E_*(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T, \qquad \beta^* \text{ defined in (2.11)}, \tag{6.11}$$

and

$$v_w^* = E_*|X^TD^*X|(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T / E_*|X^TD^*X|. \tag{6.12}$$

That v^* and v_w^* are generally *biased* will be illustrated with the following regression model:

$$\begin{aligned} y_{i1} &= \beta_1 + e_{i1}, \quad i = 1, \ldots, n_1, \\ y_{i2} &= \beta_2 + e_{i2}, \quad i = 1, \ldots, n_2, \end{aligned} \qquad n = n_1 + n_2, \tag{6.13}$$

with uncorrelated errors, $Ee_i = 0$ and $\text{Var}(e_{i1}) = \sigma_1^2$, $\text{Var}(e_{i2}) = \sigma_2^2$. Rewrite the resampling vector P^* as $(P_{11}^*, \ldots, P_{n_1 1}^*, P_{12}^*, \ldots, P_{n_2 2}^*)$ to correspond to the two samples of (6.13) and define $n_1^* = \sum_1^{n_1} P_{i1}^*$, $n_2^* = n - n_1^*$. Then $\beta_j^* = n_j^{*-1} \sum_{i=1}^{n_j} P_{ij}^* y_{ij}$, $j = 1, 2$ and $|X^TD^*X| = n_1^* n_2^*$. For this example the computation * excludes the resamples with $n_1^* = 0$ or $n_2^* = 0$ to ensure that β_j^* are well-defined. After some tedious algebras (Wu (1984)), it can be shown that

$$v^* = \text{diag}\left(E_*\left(\frac{1}{n_1^*}\right)\frac{SS_1}{n_1}, E_*\left(\frac{1}{n_2^*}\right)\frac{SS_2}{n_2} \right)$$

and

$$v_w^* = \frac{1}{E_*(n_1^* n_2^*)} \text{diag}\left(E_*(n_2^*)\frac{SS_1}{n_1}, E_*(n_1^*)\frac{SS_2}{n_2} \right),$$

where $SS_j = \sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2$. Recall that, in this case,

$$\text{Var}(\hat{\beta}_1, \hat{\beta}_2) = \text{diag}\left(\frac{\sigma_1^2}{n_1}, \frac{\sigma_2^2}{n_2} \right).$$

Since for small or moderate n_i, E_* is not close to $(n_i - 1)^{-1}$, v^* is biased. Similarly v_w^* is biased. For large n_i, by using the approximation $E_*(n_i^*) \simeq n_i$ and $E_*(n_1^* n_2^*) = n_1 n_2$, v_w^* can be approximated by

$$\text{diag}\left(\frac{SS_1}{n_1^2}, \frac{SS_2}{n_2^2} \right),$$

which, for estimating $\text{Var}(\hat{\beta}_1, \hat{\beta}_2)$, has the relative bias $\text{diag}(-1/n_1, -1/n_2)$.

This relative bias is slightly worse than that of $v_{H(1)}$, which can be shown to be

$$\frac{n}{n-2}\operatorname{diag}\left(\frac{SS_1}{n_1^2}, \frac{SS_2}{n_2^2}\right)$$

with the relative bias

$$\operatorname{diag}\left(-\frac{1}{n_1}+\frac{2}{n-2}\left(1-\frac{1}{n_1}\right), -\frac{1}{n_2}+\frac{2}{n-2}\left(1-\frac{1}{n_2}\right)\right).$$

Of course, as n_1 and n_2 become large, the biases of the three estimators diminish.

It is not surprising that the unweighted bootstrap does not give an unbiased variance estimator, since, as in the case of the unweighted jackknife, the LSE β^* based on the bootstrap resamples are not exchangeable. It is mildly disappointing that the weighted bootstrap does not correct the bias. One may expect it to perform well since the same weight was used in the jackknife with satisfactory results. If (6.13) is recognized as a two-sample problem rather than a regression problem, unbiased variance estimators can be obtained by bootstrapping and rescaling each sample. The main point made here is that a result like Theorem 3 cannot be extended to the bootstrap method for the general linear model (2.1).

Careless use of the unweighted bootstrap can lead to inconsistent variance estimators as the following example shows. In standard statistical packages the regression model (2.1) (with a nonzero intercept) is automatically reexpressed as

$$y_i = \mu + (x_i - \bar{x})^T\beta + e_i, \tag{6.14}$$

where x_i and β are of dimension $k-1$. For $\sigma_i^2 = \sigma^2$, the standard variance estimator for the LSE $\hat{\mu} = \bar{y}$ is $(1/n)\hat{\sigma}^2$, where

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-k}\sum_1^n\left(y_i - \bar{y} - (x_i-\bar{x})^T\hat{\beta}\right)^2 \\ &= \frac{1}{n-k}\left[\sum_1^n(y_i-\bar{y})^2 - \sum_1^n(x_i-\bar{x})^T\hat{\beta}\hat{\beta}^T(x_i-\bar{x})\right].\end{aligned} \tag{6.15}$$

Let $\{(y_i^*, x_i^*)\}_1^n$ be a bootstrap sample from $\{(y_i, x_i)\}_1^n$. The LSE μ^* based on $\{(y_i^*, x_i^*)\}_1^n$ and the model (6.14) is $\bar{y}^* = (1/n)\sum_1^n y_i^*$. The unweighted bootstrap variance estimator

$$\begin{aligned}v^*(\hat{\mu}) &= E_*(\mu^*-\hat{\mu})^2 = \frac{1}{n^2}\sum_1^n(y_i-\bar{y})^2 \\ &= \frac{1}{n}\left[\frac{n-k}{n}\hat{\sigma}^2 + \frac{1}{n}\sum_1^n(x_i-\bar{x})^T\hat{\beta}\hat{\beta}^T(x_i-\bar{x})\right],\end{aligned} \tag{6.16}$$

which is bigger than $(1/n)\hat{\sigma}^2$. For $n \gg k$, the relative increase

$$\frac{v^*(\hat{\mu}) - n^{-1}\hat{\sigma}^2}{n^{-1}\hat{\sigma}^2} \simeq \frac{R^2}{1-R^2},$$

is an increasing function of R^2, which is the multiple correlation coefficient of

(6.14), i.e., the percent variation explained by the model (6.14). Therefore, $v^*(\hat{\mu})$ is asymptotically inconsistent for estimating $\mathrm{Var}(\hat{\mu})$ and the percent relative bias $R^2/(1 - R^2)$ is monotonically increasing in R^2.

The inconsistency can be explained as follows. Rewrite the model (6.14) as

$$y_i = \alpha + x_i^T\beta + e_i. \tag{6.17}$$

Then $\mu = \alpha + \bar{x}^T\beta$ is a linear combination of α and β. If μ is estimated by another bootstrap estimator

$$\tilde{\mu} = \alpha^* + \bar{x}^T\beta^*,$$

where (α^*, β^*) is the LSE based on $\{(y_i^*, x_i^*)\}_1^n$ and the model (6.17), and

$$E_*\begin{pmatrix} \alpha^* - \hat{\alpha} \\ \beta^* - \hat{\beta} \end{pmatrix}\begin{pmatrix} \alpha^* - \hat{\alpha} \\ \beta^* - \hat{\beta} \end{pmatrix}^T \to \mathrm{Var}\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$$

is satisfied, then $E_*(\tilde{\mu} - \hat{\mu})^2 \to \mathrm{Var}(\hat{\mu})$. On the other hand, the bootstrap estimator μ^* can be rewritten as

$$\mu^* = \alpha^* + \bar{x}^{*T}\beta^*, \qquad \bar{x}^* = \frac{1}{n}\sum_1^n x_i^*.$$

In $\mu^* - \hat{\mu} = \tilde{\mu} - \hat{\mu} + (\bar{x}^* - \bar{x})^T\beta^*$, the second term $(\bar{x}^* - \bar{x})^T\beta^*$ has the "bootstrap" variance approximately equal to $n^{-2}\Sigma_1^n(x_i - \bar{x})^T\hat{\beta}\hat{\beta}^T(x_i - \bar{x})$, which is identical to the second term of $v^*(\hat{\mu})$ (6.16). The upward bias and inconsistency of $v^*(\hat{\mu})$ is caused by the "bootstrap" variation of $\bar{x}^*$ in $\alpha^* + \bar{x}^{*T}\beta^*$.

Finally we show heuristically that the unweighted bootstrap variance estimator v^* (6.11) is closely related to $v_{H(1)}$. Note that, for β^* in (2.11),

$$\beta^* - \hat{\beta} = \left(\sum_1^n P_i^* x_i x_i^T\right)^{-1}\sum_1^n P_i^* x_i r_i, \qquad r_i = y_i - x_i^T\hat{\beta},$$

where $(P_i^*)_1^n$ has the multinomial distribution $\mathrm{Mult}_n(n, (1/n)\mathbf{1})$. If the "bootstrap" variation in $\Sigma_1^n P_i^* x_i x_i^T$ is ignored, using $E_*(P_i^*) = 1$ and $\Sigma_1^n x_i r_i = 0$, $v^* = E_*(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T$ can be approximated by

$$\begin{aligned} &\left(\sum_1^n x_i x_i^T\right)^{-1}\mathrm{Var}_*\left(\sum_1^n P_i^* x_i r_i\right)\left(\sum_1^n x_i x_i^T\right)^{-1} \\ &= (X^TX)^{-1}\sum_1^n r_i^2 x_i x_i^T (X^TX)^{-1}, \end{aligned} \tag{6.18}$$

which is close to $V_H(1)$ and suggests a simple modification to v^*, namely,

$$v^{*\prime} = \frac{n}{n - k}v^*. \tag{6.19}$$

The connection with $v_{H(1)}$ also suggests that, at least for X satisfying some balance conditions, $v^{*\prime}$ may be robust against error variance heteroscedasticity. The correction factor $n/(n - k)$ can be applied to v_w^*.

In view of the much poorer performance of v^* versus $v_{H(1)}$ in the simulation study (Table 1), this connection should not be carried too far and the "bootstrap"

variation in $\Sigma_1^n P_i^* x_i x_i^T$ should not be ignored. In that case, v^* will be bigger than the expression (6.18) because of convexity.

7. A general method for resampling residuals. As shown in (2.8) and (2.9), the method of bootstrapping residuals does not adapt well to the possibility of error variance heteroscedasticity. A different method for resampling residuals is proposed here. It resembles the jackknife in that the ith residual goes with the ith fitted value. Define

$$y_i^* = x_i^T\hat{\beta} + \frac{r_i}{\sqrt{1-w_i}}t_i^*, \qquad i = 1, \ldots, n, \tag{7.1}$$

where $r_i = y_i - x_i^T\hat{\beta}$ is the ith residual, $w_i = x_i^T(X^TX)^{-1}x_i$ and $t^* = (t_i^*)_1^n$ is obtained according to a resampling method * with the requirement

$$E_* t^* = 0, \qquad \mathrm{Var}_*(t^*) = I. \tag{7.2}$$

The LSE based on y_i^* is

$$\hat{\beta}^* = (X^TX)^{-1}X^Ty^*, \qquad y^* = (y_1^*, \ldots, y_n^*) \quad \text{in (7.1)}. \tag{7.3}$$

Inference about β or $\theta = g(\beta)$ can be made from the variability among $\hat{\beta}^*$ in (7.3). For example, $\mathrm{Var}(\hat{\theta})$ can be estimated by

$$v_* = E_*(\hat{\theta}^* - \hat{\theta})(\hat{\theta}^* - \hat{\theta})^T, \qquad \hat{\theta}^* = g(\hat{\beta}^*), \ \hat{\theta} = g(\hat{\beta}). \tag{7.4}$$

For the linear parameters $\theta = \beta$, it is easy to show, from (7.2), that

$$E_*(\hat{\beta}^*) = \hat{\beta}$$

and

$$v_* = (X^TX)^{-1}\sum_1^n \frac{r_i^2}{1-w_i}x_i x_i^T(X^TX)^{-1},$$

which is identical to $v_{J(1)}$ and therefore enjoys all the desirable properties of $v_{J(1)}$ discussed in Section 5, including the bias-robustness against error variance heteroscedasticity. For nonlinear $\theta = g(\beta)$, the estimator v_* depends on the choice of the resampling plan * and is in general different from $v_{J(1)}$. Other, especially higher order, properties of the proposed method need further study.

Two examples of the resampling method * satisfying (7.2) are given as follows.

(i) *Method of balanced residuals.* The errors t_i^* are chosen from a Hadamard matrix $[\delta_i^{(k)}]$, $\delta_i^{(k)} = \pm 1$, $1 \le i \le n$, $1 \le k \le R$, satisfying

$$\sum_{k=1}^R \delta_i^{(k)} = 0 \quad \text{for all } i,$$
$$\frac{1}{R}\sum_{k=1}^R \delta_i^{(k)}\delta_j^{(k)} = 0 \quad \text{for } i \ne j. \tag{7.5}$$

An extensive review on the existence and construction of Hadamard matrices can

be found in Hedayat and Wallis (1978). Typically, $n + 1 \le R \le n + 4$. The kth resampled $y^{(k)} = (y_i^{(k)})_{i=1}^n$ is defined as

$$y_i^{(k)} = x_i^T\hat{\beta} + \frac{r_i}{\sqrt{1 - w_i}}\delta_i^{(k)}, \qquad i = 1, \ldots, n, \tag{7.6}$$

and the corresponding LSE as

$$\hat{\beta}^{(k)} = (X^TX)^{-1}X^Ty^{(k)}. \tag{7.7}$$

The variance estimator v_* is of the form

$$\frac{1}{R}\sum_1^R(\hat{\theta}^{(k)} - \hat{\theta})(\hat{\theta}^{(k)} - \hat{\theta})^T, \qquad \hat{\theta}^{(k)} = g(\hat{\beta}^{(k)}). \tag{7.8}$$

Note that this method is similar in spirit to McCarthy's (1969) "balanced half-samples" method for stratified random samples. Here the unbiasedness of v_* is for an estimator with R recomputations, R roughly equal to n, whereas the bootstrap requires (at least conceptually) infinite recomputations for the unbiasedness result or other small sample results to hold.

The method (7.6)–(7.8) seems to assume the symmetry of the underlying errors e_i because, for each residual r_i, half of the R resamples have r_i and the other half have $-r_i$ in (7.6).

(ii) *A jackknife-bootstrap hybrid.* The errors $\{t_i^*\}_1^n$ are a bootstrap (i.e., i.i.d) sample from a finite population $\{a_j\}_{j=1}^M$ with

$$\sum_1^M a_j = 0, \qquad \frac{1}{M}\sum_1^M a_j^2 = 1.$$

The choice of $\{a_j\}$ will influence the higher-order performance of the method. One possibility is to choose

$$a_j = (r_j - \bar{r})\Big/\left[\frac{1}{n}\sum_1^n(r_j - \bar{r})^2\right]^{1/2}, \qquad j = 1, \ldots, n.$$

8. Jackknifing in nonlinear situations. We outline extensions of the general weighted jackknife method in three nonlinear situations. A rigorous treatment of their asymptotic properties is not attempted here since it would require very careful handling of the lower-order terms.

(i) *Regression M-estimator.* An M-estimator $\tilde{\beta}$ is obtained by minimizing

$$\sum_1^n \rho(y_i - x_i^T\beta) \tag{8.1}$$

over β, where ρ is usually assumed to be symmetric. Choices of ρ can be found in Huber (1981). Let $\tilde{\beta}_s$ be the M-estimator obtained from minimizing (8.1) with (y_i, x_i) in the subset s. The jackknife variance estimators $v_{J,r}$ (4.1) and $\tilde{v}_{J,r}$ (4.3) can be extended in a straightforward manner by replacing the LSE's $\hat{\beta}$ and $\hat{\beta}_s$ by

the M-estimators $\tilde{\beta}$ and $\tilde{\beta}_s$. The weight $|X_s^T X_s|$ remains the same. This can be justified in the case of i.i.d. errors e_i, since the approximation

$$\tilde{\beta} - \beta \simeq [E\rho''(e_i)]^{-1}\left(\sum_1^n x_i x_i^T\right)^{-1}\left(\sum_1^n x_i \rho'(e_i)\right) \tag{8.2}$$

resembles the expansion $\hat{\beta} - \beta = (\sum_1^n x_i x_i^T)^{-1}\sum_1^n x_i e_i$ for the LSE. Conditions under which (8.2) is justified can be found in Huber (1981). For independent but not identically distributed errors, if $E(\rho'(e_i))^2/[E\rho''(e_i)]^2$ is a constant, the weight $|X_s^T X_s|$ is still justifiable; otherwise not much is known.

(ii) *Nonlinear regression.* In a nonlinear regression model

$$y_i = f_i(\beta) + e_i, \tag{8.3}$$

where f_i is a nonlinear smooth function of β and e_i satisfies (2.1), the jackknife variance estimators $v_{J,r}$ (4.1) and $\tilde{v}_{J,r}$ (4.3) have a natural extension, namely to replace x_i by $f_i'(\hat{\beta})$ and to interpret $\hat{\beta}$ and $\hat{\beta}_s$ as their nonlinear least squares counterparts. Here $f_i'(\beta)$ is the vector of derivatives of f_i with respect to β. We may consider alternative weight functions to avoid the computation of f_i' or by evaluating it at other estimates $\hat{\beta}_s$. Another approach that requires less computation was proposed by Fox, Hinkley and Larntz (1980) for the delete-one jackknife.

(iii) *Generalized linear models.* We consider generalized linear models (McCullagh and Nelder (1983)) with uncorrelated errors. Let

$$y = (y_1, \ldots, y_n)^T, \qquad Ey = \mu = (\mu_1, \ldots, \mu_n)^T$$

and

$$\operatorname{Var}(y) = \sigma^2 V(\mu) = \sigma^2 \operatorname{diag}(v_i(\mu)).$$

The mean μ_i is related to the regressor x_i via a link function η, i.e., $\mu_i = \eta(x_i^T\beta)$. The full likelihood may not be available. Inference is instead based on the log quasilikelihood (Wedderburn (1974); McCullagh (1983)) $L(\mu; y)$ defined by

$$\frac{\partial L(\mu; y)}{\partial \mu} = V(\mu)^{-1}(y - \mu).$$

A generalized least squares estimator (GLS) $\hat{\beta}$ is defined as a solution to

$$D^T V^{-1}(y - \mu(\beta)) = 0,$$

where

$$\mu(\beta) = (\mu_i)_1^n = \left(\eta\left(x_i^T\beta\right)\right)_1^n \quad \text{and} \quad D = \frac{d\mu}{d\beta} = \operatorname{diag}\left(\eta'\left(x_i^T\beta\right)\right)X.$$

For the estimation of $\operatorname{Var}(\hat{\theta})$, $\hat{\theta} = g(\hat{\beta})$, the jackknife variance estimator $v_{J,r}$

(4.1) can be extended as

$$\frac{r-k+1}{n-r}\Sigma_r w_s(\hat{\theta}_s-\hat{\theta})(\hat{\theta}_s-\hat{\theta})^T, \qquad \hat{\theta}_s=g(\hat{\beta}_s),$$

$$w_s \propto |\hat{D}_s^T\hat{V}_s^{-1}\hat{D}_s|, \qquad \Sigma_r w_s=1,$$

and $\hat{\beta}_s$ is the GLS based on s. Extension of $\tilde{v}_{J,r}$ (4.3) is obvious. The weight $|\hat{D}_s^T\hat{V}_s^{-1}\hat{D}_s|$ is justified by noting that the Fisher information matrix for β is proportional to $D^TV^{-1}D$ and the approximation (McCullagh (1983))

$$\hat{\beta}-\beta \simeq (D^TV^{-1}D)^{-1}D^TV^{-1}(y-\mu). \tag{8.4}$$

The right side of (8.4) resembles the expansion $\hat{\beta}-\beta=(X^TX)^{-1}X^T(y-X\beta)$ in ordinary regression with $V^{-1/2}D=\operatorname{diag}(\eta'(x_i^T\beta)/v_i^{1/2}(\mu))X$ playing the role of X.

9. Bias reduction. The nonlinear estimator $\hat{\theta}=g(\hat{\beta})$ of $\theta=g(\beta)$ has bias of order n^{-1}. In this section we will show that bias reduction is closely connected with the existence of an almost unbiased estimator of variance. Under (C3) and the continuous third differentiability of g in a neighborhood of β, Taylor expansion gives

$$\hat{\theta}=\theta+g'(\beta)^T(\hat{\beta}-\beta)+\tfrac{1}{2}(\hat{\beta}-\beta)^Tg''(\beta)(\hat{\beta}-\beta)+O_p(n^{-1.5}), \tag{9.1}$$

where $O_p(n^{-1.5})$ denotes terms of stochastic order $n^{-1.5}$. From (9.1), the bias of $\hat{\theta}$

$$B(\hat{\theta})=E\hat{\theta}-\theta=\tfrac{1}{2}\operatorname{tr}(g''(\beta)\operatorname{Var}(\hat{\beta}))+O(n^{-2}), \tag{9.2}$$

where tr is the trace of a matrix. By elaborating on standard proofs (Fuller (1976), Theorem 5.4.3; Lehmann (1983), Theorem 2.5.1), it is easy to justify (9.2) under (C1), (C3), that e_i are independent with $\sup_i Ee_i^4<\infty$, and that g has third Lipschitz-continuous derivatives. Since the reduction of bias of $\hat{\theta}$ amounts to finding an unbiased estimator for $B(\hat{\theta})$ up to order n^{-2}, we will focus on the latter problem for the rest of the section.

Data resampling makes it possible to estimate $B(\hat{\theta})$ without computing the Hessian matrix $g''(\beta)$. Several estimators are proposed in this section. Heuristic justifications are given. A rigorous treatment of the bias-reducing property for these estimators would require a careful handling of the remainder terms and is not attempted here.

Consider first the jackknife resampling. Let $\tilde{\theta}_s=g(\tilde{\beta}_s)$, $\tilde{\beta}_s$ given in (4.4). The proposed estimator of $B(\hat{\theta})$ is

$$\tilde{B}_{J,r}=\Sigma_r w_s(\tilde{\theta}_s-\hat{\theta}), \quad w_s \text{ in (4.1)}. \tag{9.3}$$

For the linear parameters $\theta=\beta$, $\tilde{B}_{J,r}=0$ from the representation for the LSE (Theorem 1). Note that in this case $\hat{\theta}=\hat{\beta}$ has zero bias. From the expansion

$$\tilde{\theta}_s=\hat{\theta}+g'(\hat{\beta})^T(\tilde{\beta}_s-\hat{\beta})+\tfrac{1}{2}(\tilde{\beta}_s-\hat{\beta})^Tg''(\hat{\beta})(\tilde{\beta}_s-\hat{\beta})+\eta_s,$$

where η_s is the remainder term, and Theorem 1,

$$\tilde{B}_{J,r}=\tfrac{1}{2}\operatorname{tr}(g''(\hat{\beta})v_{J,r})+\Sigma_r w_s\eta_s. \tag{9.4}$$

Since $E(v_{J,r}) = \mathrm{Var}(\hat{\beta})$ under $\mathrm{Var}(e) = \sigma^2 I$, one would expect that $\tilde{B}_{J,r}$ captures the leading term of $B(\hat{\theta})$ in (9.2), i.e., $E(\tilde{B}_{J,r}) = B(\hat{\theta})(1 + O(n^{-1}))$ under reasonable assumptions on r, g, X and e_i. A similar result should hold in the case of heteroscedastic errors for those r with

$$E(v_{J,r}) = \mathrm{Var}(\hat{\beta})(1 + O(n^{-1})) \tag{9.5}$$

under $\mathrm{Var}(e) = \mathrm{diag}(\sigma_i^2)$. The relation (9.5) is satisfied for $r = n - 1$ (Theorem 5(ii)) and for other values of r (Shao and Wu (1985)).

Since $v_{H(1)}$ also satisfies (9.5) (Theorem 6(ii)), one would expect to find a bias-reducing estimator related to $v_{H(1)}$. Hinkley (1977) indeed considered the estimator,

$$\hat{B}_{J(1)} = \sum_1^n (1 - w_i)\{g(\hat{\beta}_{(i)}) - g(\hat{\beta})\}, \tag{9.6}$$

and demonstrated its almost unbiasedness for estimating $B(\hat{\theta})$ in the homoscedastic case. Using the expansion

$$g(\hat{\beta}_{(i)}) = g(\hat{\beta}) + g'(\hat{\beta})^T(\hat{\beta}_{(i)} - \hat{\beta}) + \tfrac{1}{2}(\hat{\beta}_{(i)} - \hat{\beta})^T g''(\hat{\beta})(\hat{\beta}_{(i)} - \hat{\beta}) + \eta_{(i)},$$

where $\eta_{(i)}$ is the remainder term, and Theorem 1, it is easy to show that

$$\hat{B}_{J(1)} = \tfrac{1}{2}\mathrm{tr}\{g''(\hat{\beta})v_{J(1)}\} + \sum_1^n (1 - w_i)\eta_{(i)}, \tag{9.7}$$

which reveals the surprising connection of the bias estimator $\hat{B}_{J(1)}$ with $v_{J(1)}$, rather than with its twin $v_{H(1)}$. From the unbiasedness of $v_{J(1)}$ (Theorem 5), one would expect that

$$E(\hat{B}_{J(1)}) = B(\hat{\theta})(1 + O(n^{-1}))$$

holds under reasonable assumptions.

Consistent with our usage of (1) to denote delete-one, we use $\tilde{B}_{J(1)}$ for $\tilde{B}_{J,n-1}$, i.e.,

$$\tilde{B}_{J(1)} = \tilde{B}_{J,n-1}. \tag{9.8}$$

A natural question is how to choose between $\hat{B}_{J(1)}$ and $\tilde{B}_{J(1)}$. In terms of imitating the behavior of $g(\hat{\beta}) - g(\beta)$, whose expectation is the bias $B(\hat{\theta})$, $\tilde{B}_{J(1)}$ is preferred since it uses $\hat{\beta}_{(i)}$ and $\hat{\beta}$, whose distance matches that of $\hat{\beta} - \beta$ whereas $\hat{\beta}_{(i)} - \hat{\beta}$ in $\hat{B}_{J(1)}$ is much smaller than $\hat{\beta} - \beta$. On the other hand, for smooth g, heuristic computations show that the error term $\Sigma(1 - w_i)\eta_{(i)}$ in $\hat{B}_{J(1)}$ is of order n^{-2}, while the error term $\Sigma w_s \eta_s$ in $\tilde{B}_{J,r}$ (including $\tilde{B}_{J(1)}$) is of order $n^{-1.5}$, suggesting that $\hat{B}_{J(1)}$ is a better approximation to $B(\hat{\theta})$.

A natural extension of $\hat{B}_{J(1)}$ to the general jackknife with subset size r is

$$\hat{B}_{J,r} = \frac{r - k + 1}{n - r}\Sigma_r w_s(g(\hat{\beta}_s) - g(\hat{\beta})), \quad w_s \text{ in (4.1)}. \tag{9.9}$$

The difference between $\hat{B}_{J,r}$ and $\tilde{B}_{J,r}$ is analogous to that between $\hat{v}_{J,r}(\hat{\theta})$ (4.1) and $\tilde{v}_{J,r}(\hat{\theta})$ (4.4). The former applies the scale adjustment $(r - k + 1)/(n - r)$

externally and the latter internally. Like $\tilde{B}_{J,r}$, $\hat{B}_{J,r}$ also captures the leading term of $B(\hat{\theta})$ under reasonable assumptions.

Let us now consider the bootstrap. Since the unbiased estimation for $B(\hat{\theta})$ hinges on the unbiased estimation for $\mathrm{Var}(\hat{\beta})$, from the study of Section 6, we need only consider the bootstrap residual method. Let $\beta^* = (X^TX)^{-1}X^Ty^*$ be the bootstrap LSE defined in (2.7). The proposed bootstrap estimator of bias is

$$\hat{B}_b = E_*\theta^* - \hat{\theta}, \qquad \theta^* = g(\beta^*). \tag{9.10}$$

From the expansion

$$\theta^* = \hat{\theta} + g'(\hat{\beta})^T(\beta^* - \hat{\beta}) + \tfrac{1}{2}(\beta^* - \hat{\beta})^T g''(\hat{\beta})(\beta^* - \hat{\beta}) + \eta_*,$$

it is easy to show that

$$\hat{B}_b = \tfrac{1}{2}\mathrm{tr}(g''(\hat{\beta})v_b)E_*(\eta_*). \tag{9.11}$$

Since $E(v_b) = \mathrm{Var}(\hat{\beta})$ for $\mathrm{Var}(e) = \sigma^2 I$, one would expect that

$$E\hat{B}_b = B(\hat{\theta})(1 + O(n^{-1})) \tag{9.12}$$

holds under $\mathrm{Var}(e) = \sigma^2 I$ and other conditions which ensure $E(E_*\eta_*') = O(n^{-2})$. The result (9.12) cannot be extended to the heteroscedastic case because v_b is biased.

10. Simulation results. In this section we examine the Monte Carlo behavior of (i) the relative bias of several variance estimators, (ii) the bias of several estimators of a nonlinear parameter $\theta = g(\beta)$, and (iii) the coverage probability and length of the associated interval estimators for the same nonlinear parameter.

Under consideration is the following quadratic regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i, \qquad i = 1(1)12,$$
$$x_i = 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 6, 7, 8, 10.$$

Two variance patterens are considered: *unequal variances* $e_i = (\frac{1}{2}x_i)^{1/2}N(0,1)$ and *equal variances* $e_i = N(0,1)$. The e_i's are independent.

Seven variance estimators are considered:

(1) the usual variance estimator $\hat{v}$ (2.9), which is identical to the bootstrap variance estimator v_b (2.8);
(2) the unweighted jackknife variance estimator v_J (2.3);
(3) the delete-one jackknife variance estimator $v_{J(1)}$ (5.1);
(4) Hinkley's delete-one jackknife variance estimator $v_{H(1)}$ (2.6);
(5) the retain-eight jackknife variance estimator $v_{J,8}$ (4.1);
(6) the unweighted bootstrap variance estimator v^* (6.11);
(7) the weighted bootstrap variance estimator v_w^* (6.12).

The results in Table 1 are based on 3000 simulations on a VAX 11/780 at the University of Wisconsin-Madison. The normal random numbers are generated according to the IMSL subroutine GGNML. In drawing the bootstrap samples, the uniform random integers are generated according to the IMSL subroutine

TABLE 1

Relative biases of seven variance estimators. The relative bias of an estimator v is defined as $(E(V) - \mathrm{Var}(\hat{\beta}))/|\mathrm{Var}(\hat{\beta})|$; (i, j) denotes the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$.

	(0,0)	(0,1)	(0,2)	(1,1)	(1,2)	(2,2)
			Equal variances			
$\hat{v}$	−0.01	0.01	−0.01	−0.01	0.01	0.00
v_J	0.61	−0.78	1.03	0.93	−1.18	1.53
$v_{J(1)}$	−0.01	0.01	−0.00	−0.00	−0.00	0.00
$v_{H(1)}$	−0.13	0.16	−0.21	−0.17	0.22	−0.29
$v_{J,8}$	−0.01	0.01	−0.00	−0.00	−0.00	0.00
v^*	0.63	−0.85	1.22	1.04	−1.49	2.18
v_w^*	−0.07	0.07	−0.08	−0.06	0.07	−0.06
			Unequal variances			
$\hat{v}$	0.39	−0.09	−0.04	−0.11	0.20	−0.29
v_J	0.97	−1.07	1.29	1.10	−1.29	1.45
$v_{J(1)}$	0.02	0.04	−0.09	−0.08	0.12	−0.16
$v_{H(1)}$	−0.16	0.24	−0.35	−0.29	0.39	−0.47
$v_{J,8}$	0.06	0.02	−0.08	−0.08	0.13	−0.18
v^*	1.02	−0.98	1.17	0.91	−1.13	1.39
v_w^*	0.03	0.07	−0.14	−0.13	0.19	−0.26

GGUD. The number of bootstrap samples B is 480, which is comparable to 495, the total number of jackknife subsets of size 8. The same set of random numbers is used throughout the study. We thank Mike Hamada for computational assistance and John Tukey for comments on Table 1.

The results of Table 1 can be summarized as follows. The worst two are the unweighted estimators v_J and v^* with their relative biases ranging from 60 to 210%. The biases of v_J and of v^* are very close. This demonstrates a serious weakness of the unweighted procedures in unbalanced situations. The next poor performer is $v_{H(1)}$ with its relative biases ranging from 13 to 47%. (The signs of $\mathrm{Bias}(v_{H(1)})$ are the *opposite* of the signs of $\mathrm{Bias}(v_J)$ and $\mathrm{Bias}(v^*)$.) This is somewhat disappointing since its claimed robustness against error variance heteroscedasticity (Theorem 6(ii)) does not hold up here. On the other hand, as predicted by Theorem 5 and the result of Shao and Wu (1985), $v_{J(1)}$ and $v_{J,8}$ are nearly unbiased. This shows that the two weighting schemes can lead to significantly different results in unbalanced situations. The weighted bootstrap estimator v_w^* does almost as well as $v_{J(1)}$ and $v_{J,8}$. A rigorous justification is called for. As expected, $\hat{v}$ $(= v_b)$ does well for equal variances, but is severely biased for estimating $\mathrm{Var}(\hat{\beta}_0)$ for unequal variances. With one exception, $\hat{v}$ $(= v_b)$ is less biased than $v_{H(1)}$, which cannot be explained by the present theory.

We next consider bias reduction and interval estimation for the nonlinear parameter

$$\theta = -\beta_1/(2\beta_2),$$

which maximizes the quadratic function $\beta_0 + \beta_1 x + \beta_2 x^2$ over x. Seven point estimators are considered: $\hat{\theta}$; $\tilde{\theta}$ defined before (2.3); $\hat{\theta}_{J(1)} = \hat{\theta} - \hat{B}_{J(1)}$ (9.6);

$\tilde{\theta}_{J(1)} = \hat{\theta} - \tilde{B}_{J(1)}$ (9.8); $\hat{\theta}_{J,8} = \hat{\theta} - \hat{B}_{J,8}$ (9.9); $\tilde{\theta}_{J,8} = \hat{\theta} - \tilde{B}_{J,8}$ (9.3); $\hat{\theta}_b = \hat{\theta} - \hat{B}_b$ (9.10). Their average biases are given in Table 2. Bias reduction is more difficult to achieve when β_2 gets closer to 0 (since θ becomes more *curved* as a function of β_2) and when the variances are unequal. In the most nonlinear situation $\beta_2 = -0.25$ and unequal variances, only $\hat{\theta}_{J(1)}$ and $\hat{\theta}_b$ achieve mild reduction of bias and other estimators in fact have bigger biases. In all the other situations, the two jackknife estimators, $\hat{\theta}_{J(1)}$ and $\hat{\theta}_{J,8}$, achieve substantial reduction of bias. On the other hand, the other two jackknife estimators, $\tilde{\theta}_{J(1)}$ and $\tilde{\theta}_{J,8}$, do not perform as well. This is consistent with the asymptotic comparison given after (9.8). The worst performer is the unweighted jackknife estimator $\tilde{\theta}$. For $\beta_2 = -0.25$ and -0.35, it has bigger biases than the original estimator $\hat{\theta}$. The behavior of the bootstrap estimator $\hat{\theta}_b$ for $\beta_2 = -0.25$ is unpredictable. According to the discussion at the end of Section 9, $\hat{\theta}_b$ reduces bias for equal variances but not for unequal variances. What we see in Table 2 is the contrary.

We now consider interval estimation for θ. For equal variances, the classical Fieller interval is exact. In the context of maximizing the quadratic function, the exact $(1 - 2\alpha)$ Fieller interval is (Williams (1959), page 111)

$$\begin{array}{lll} \text{(I)} & (-\infty, \infty) & \text{if } (1-g_{12})^2 < (1-g_{11})(1-g_{22}), \\ \text{(II)} & (-\infty, \theta_L) \cup (\theta_U, \infty) & \text{if } (1-g_{12})^2 \geq (1-g_{11})(1-g_{22}),\ g_{22} > 1, \\ \text{(III)} & [\theta_L, \theta_U] & \text{otherwise,} \end{array} \tag{10.1}$$

where θ_L and θ_U are the smaller and larger values, respectively, of

$$\hat{\theta}\Big\{1 - g_{12} \pm \left[(1-g_{12})^2 - (1-g_{11})(1-g_{22})\right]^{1/2}\Big\}\Big/(1-g_{22}),$$

$$g_{ij} = \frac{t_\alpha^2 \hat{\sigma}^2 c^{ij}}{\hat{\beta}_i \hat{\beta}_j}, \qquad (X^T X)^{-1} = [c^{ij}]_{0 \leq i,\, j \leq 2}, \tag{10.2}$$

and t_α is the upper α percentage point of a t-distribution with $n - 3$ (here 9)

TABLE 2

Biases of six estimators of $\hat{\theta}$ (based on 3000 simulation samples). $\beta_0 = 0$; $\beta_1 = 4$.

	Unequal variances β_2				Equal variances β_2	
Estimator	**−0.25**	**−0.35**	**−0.5**	**−1.0**	**−0.25**	**−1.0**
$\hat{\theta}$	0.41	0.05	−0.02	−0.01	0.08	−0.01
$\tilde{\theta}$	−1.91	−0.16	−0.00	0.01	−0.38	0.01
$\hat{\theta}_{J(1)}$	−0.22	−0.01	0.00	−0.00	−0.05	−0.00
$\tilde{\theta}_{J(1)}$	0.63	0.06	0.02	0.00	0.02	−0.00
$\hat{\theta}_{J,8}$	1.48	0.00	0.00	−0.00	0.01	−0.00
$\tilde{\theta}_{J,8}$	2.39	0.05	−0.01	−0.00	−0.08	−0.00
$\hat{\theta}_b$	0.16	0.02	0.01	−0.00	−0.12	−0.00

degrees of freedom; $\hat{\sigma}^2$ is given in (2.9) (assuming equal variances). (If the data analyst is concerned with the possibility of unequal variances, larger degrees of freedom for t_α may be used.) Fieller's interval estimate is unbounded in the case of (I) or (II) of (10.1). The method is *not* exact for unequal variances.

Altogether nine methods are compared in our simulation. A description is given below. (The tenth method TBOOT in Table 3 is described in the rejoinder.)

Symbol		Interval estimate
Fieller	Fieller's interval,	(10.1)
VCJ(1)	Delete-1 jackknife	$\hat{\theta} \pm t_\alpha\sqrt{\tilde{v}_{J,n-1}(\hat{\theta})}$ (4.3)
VHJ(1)	Delete-1 jackknife	$\hat{\theta} \pm t_\alpha\sqrt{v_{J,n-1}(\hat{\theta})}$ (4.1)
VCJ8	Retain-8 jackknife	$\hat{\theta} \pm t_\alpha\sqrt{\tilde{v}_{J,8}(\hat{\theta})}$ (4.3)
VHJ8	Retain-8 jackknife	$\hat{\theta} \pm t_\alpha\sqrt{v_{J,8}(\hat{\theta})}$ (4.1)
VBOOT	Bootstrap variance	$\hat{\theta} \pm t_\alpha\sqrt{v_b}$ (2.8)
VLIN	Linear approximation	$\hat{\theta} \pm t_\alpha\sqrt{v_{\text{lin}}}$ (4.5)
PBOOT	Bootstrap percentile	$[\text{CDFB}^{-1}(\alpha), \text{CDFB}^{-1}(1-\alpha)]$ (2.10)
PJ8	Jackknife percentile (retain-8)	$[\text{CDFJ}^{-1}(\alpha), \text{CDFJ}^{-1}(1-\alpha)]$ (4.6)

(V: variance, C: curl, H: hat, P: percentile)

The Monte Carlo coverage probabilities are given in Table 3 for five sets of parameters. Since Fieller's interval in the case of (I) and (II) of (10.1) has infinite length, we break the 3000 simulation samples into categories (I), (II) and (III) according to the corresponding Fieller's intervals. In our simulation samples (I) never happens; (II) happens only when $\beta_2 = -0.25$ and -0.35. In these two cases, the median length of each interval estimate is computed separately for category (II) and category (III) and is given in Table 4. For the rest, the median length over 3000 samples is given in the parentheses in Table 3.

We do not report the average lengths since they are greatly influenced by a few extreme values. Take $\beta_2 = -0.25$ and unequal variances as an example. The average lengths for VCJ8, VHJ8 and VBOOT in category (III) are 176.85, 365.76 and 39.54, respectively, while the medians are 10.65, 6.64 and 3.37. The performance of the three methods is unstable in highly nonlinear situations.

The results of Tables 2 to 4 can be summarized as follows:

1. *Effect of parameter nonlinearity.* When the parameter θ becomes more nonlinear (β_2 closer to 0), all the intervals become wider and the associated coverage probabilities smaller. The phenomenon is especially noticeable for unequal varinces and $\beta_2 = -0.25, -0.35$, where we observe the Fieller paradox (i.e., Fieller's intervals take the form (10.1)(II)). In these two cases, only the two retain-8 jackknife methods provide intervals with good coverage probabilities. But the price is dear. Both the mean and median lengths of

TABLE 3

Average coverage probabilities and median lengths for nine interval estimation methods (3000 simulation samples). Nominal level = 0.95; $\beta_0 = 0$; $\beta_1 = 4$. The length of interval estimate is given in parentheses.

	Unequal variances β_2				Equal variances β_2	
Method	**−0.25**	**−0.35**	**−0.5**	**−1.0**	**−0.25**	**−1.0**
Fieller	0.858	0.866	0.968	0.952	0.947	0.950
			(0.98)	(0.92)	(2.48)	(0.64)
VCJ(1)	0.887	0.848	0.961	0.950	0.904	0.935
			(0.91)	(0.89)	(2.03)	(0.62)
VHJ(1)	0.866	0.845	0.950	0.947	0.899	0.935
			(0.87)	(0.87)	(1.94)	(0.62)
VCJ8	0.946	0.920	0.968	0.953	0.947	0.939
			(0.97)	(0.90)	(3.19)	(0.63)
VHJ8	0.931	0.908	0.965	0.953	0.941	0.939
			(0.93)	(0.90)	(2.69)	(0.63)
VBOOT	0.886	0.902	0.973	0.955	0.956	0.946
			(0.97)	(0.91)	(2.42)	(0.64)
VLIN	0.865	0.891	0.969	0.952	0.949	0.948
			(0.93)	(0.90)	(2.18)	(0.64)
PBOOT	0.829	0.814	0.940	0.921	0.912	0.916
			(0.84)	(0.79)	(2.05)	(0.56)
PJ8	0.809	0.755	0.909	0.912	0.831	0.900
			(0.78)	(0.78)	(1.90)	(0.55)
TBOOT	0.755	0.816	0.960	0.948	0.918	0.949
			(0.92)	(0.90)	(2.14)	(0.64)

TABLE 4

Median lengths of nine interval estimates of category (II) *and category* (III). *$\beta_0 = 0$; $\beta_1 = 4$; unequal variances. The figure in the parentheses is the number of simulation samples belonging to the category.*

	$\beta_2 = -0.25$		$\beta_2 = -0.35$	
Method	**II (199)**	**III (2801)**	**II (7)**	**III (2993)**
Fieller	∞	3.81	∞	1.10
VCJ(1)	29.08	3.87	8.92	1.04
VHJ(1)	15.17	3.13	5.63	0.98
VCJ8	223.67	10.65	38.08	1.59
VHJ8	166.81	6.64	49.80	1.37
VBOOT	313.17	3.73	86.63	1.07
VLIN	14.75	2.91	5.82	1.02
PBOOT	55.05	3.07	17.78	0.93
PJ8	28.54	3.34	8.22	0.92
TBOOT	17.16	2.89	5.56	0.98

their intervals are quite big even in category (III) where Fieller's interval is reasonably tight but, of course, with poor coverage probability. In the other cases, the first seven methods all do reasonably well.

2. *Effect of error variance heteroscedasticity.* As the theory indicates, the performance is less desirable in the unequal variance case. Fieller's interval is far from being exact for $\beta_2 = -0.25, -0.35$ and unequal variances. For equal variances Fieller's method is almost exact and the next six methods (t-intervals with various variance estimates) perform reasonably well even in the most nonlinear case $\beta_2 = -0.25$. The two retain-8 jackknife methods are least affected by the heteroscedasticity of variances.
3. *Undercoverage of the percentile methods.* This is very disappointing in view of the second-order asymptotic results on the bootstrap (Singh (1981); Beran (1982)) that are used as evidence of the superiority of the bootstrap approximation over the classical t-approximation.
4. *Fieller's method* is exact in the equal variance case even when the parameter is considerably nonlinear, but is vulnerable to error variance heteroscedasticity.
5. *The linearization method is a winner.* This is most surprising since we cannot find a theoretical justification. The intervals are consistently among the shortest, and the coverage probabilities are quite comparable to the others (except for $\beta_2 = -0.25, -0.35$ and unequal variances where VCJ8 and VHJ8 are the best). The linearization method is compared favorably with Fieller's method. The former has consistently shorter intervals than the latter and the coverage probabilities are very close. For $\beta_2 = -0.25, -0.35$ and unequal variances, VLIN has much shorter intervals and much higher coverage probabilities. Note that Fieller's intervals are *unbounded* in 199 ($\beta_2 = -0.25$) and 7($\beta_2 = -0.35$) out of 3000 samples (Table 4).
6. *Internal (curl) or external (hat) adjustment in jackknife variance estimation?* In general the curl jackknife gives wider intervals than the hat jackknife, but the coverage probabilities of the two methods are comparable. One possible explanation is that, if the β_2 component of $\hat{\beta}_s - \hat{\beta}$ in the definition of $\tilde{\beta}_s$ (4.4) is negative, it may push $\tilde{\beta}_2$ closer to zero, resulting in a large value of $\tilde{\theta}_s$.

11. Concluding remarks and further questions. Although the jackknife method has been around for a long time, the recent surge of interest in resampling inference is primarily due to Efron's (1979, 1982) contributions and his pursuasive arguments in favor of the bootstrap. In fact, there is no clear-cut choice between the two methods for bias reduction or variance estimation in the i.i.d. case. The main difference is in histogram-based interval estimation, where the delete-one jackknife is definitely inferior.

Our main contribution to the jackknife methodology is twofold: (i) emphasis on a more flexible choice of the subset size and (ii) proposal of a general weighting scheme for independent but nonexchangeable situations such as regression. Because of (i), interval estimators based on the jackknife histogram can be constructed. Standard results on the large sample behavior of the bootstrap

histogram probably hold for the jackknife with a proper choice of subset size (see Wu (1985) for such results). Because of (ii), our proposed variance estimators are unbiased for homoscedastic errors and robust against heteroscedasticity. The weighting scheme is applicable to general independent situations. The scope of the jackknife methodology is substantially broadened.

Our analysis of the bootstrap gives a different picture from that of Efron (1982), which deals primarily with the i.i.d. problem. The bootstrap method seems to depend on the existence of an exchangeable component of a model. If such a component does not exist (e.g., heteroscedastic linear models, generalized linear models), it may result in serious problems such as bias and inconsistency as our examples show. It encounters similar difficulties in handling complex survey samples, where neither the independence nor the exchangeability assumption is satisfied (Rao and Wu (1985a)). We therefore *advise against its indiscriminate use in complex situations*. Important features of a problem should be taken into account in the choice of resampling methods. In fairness to the bootstrap, it is intuitively appealing, easy to program and works well when an exchangeable component exists and is bootstrapped. The last point is supported by the studies of Freedman and Peters (1984) and Carroll, Ruppért and Wu (1986). Bootstrap will undoubtedly remain a major tool in the resampling arsenal.

Several questions have been raised in the course of our study. We hope they will generate further interests and research in this area.

1. How can the subset size r in the weighted jackknife be determined? Choice of r appears to depend on computational consideration and purpose of analysis, e.g., interval estimation and bias-robustness of variance estimators.
2. The method for resampling residuals in Section 7 gives variance estimators that have better theoretical properties than the bootstrap. What resampling plans in (7.1) should be chosen? The choice depends on computational consideration and further theoretical aspects.
3. Is the weighted bootstrap variance estimator v_w^* (6.12), in general, nearly unbiased as the encouraging simulation result may suggest?
4. The methods based on the bootstrap-histogram and the jackknife-histogram perform disappointingly in the simulation. Refinements of these methods are called for (e.g., Efron (1985); Loh (1987)). One obvious defect of the resample histogram is that they have shorter tails than their population counterparts. The handling of skewness may also be improper. Theoretical results that can explain small-sample behavior are needed.
5. A careful study of resampling methods for more complex problems such as those in Section 8 is needed. Are there any shortcuts for computing the weights (or approximate weights) and/or higher-order asymptotic properties? How can models with correlated errors be handled?
6. The factor $(r - k + 1)/(n - r)$ in the weighted jackknife is used for scale adjustment. It can be applied either before or after the nonlinear transformation (see (4.1) and (4.3)). What are the relative merits of the two scaling methods?

REFERENCES

ABRAMOVITCH, L. and SINGH, K. (1985). Edgeworth corrected pivotal statistics and the bootstrap. *Ann. Statist.* **13** 116–132.

BERAN, R. (1982). Estimated sampling distributions: The bootstrap and competitors. *Ann. Statist.* **10** 212–225.

CARROLL, R. J., RUPPERT, D. and WU, C. F. J. (1986). Generalized least squares: Variance expansions, the bootstrap and the number of cycles. Unpublished.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. In *Multivariate Analysis* (P. R. Krishnaiah, ed.) **5** 35–57. North-Holland, Amsterdam.

EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia.

EFRON, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika* **72** 45–58.

EFRON, B. and GONG, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *Amer. Statist.* **37** 36–48.

FAREBROTHER, R. W. (1985). Relations among subset estimators: A bibliographical note. *Technometrics* **27** 85–86.

FOX, T., HINKLEY, D. and LARNTZ, K. (1980). Jackknifing in nonlinear regression. *Technometrics* **22** 29–33.

FREEDMAN, D. A. and PETERS, S. C. (1984). Bootstrapping a regression equation: Some empirical results. *J. Amer. Statist. Assoc.* **79** 97–106.

FULLER, W. A. (1976). *Introduction to Statistical Time Series.* Wiley, New York.

HEDAYAT, A. and WALLIS, W. D. (1978). Hadamard matrices and their applications. *Ann. Statist.* **6** 1184–1238.

HINKLEY, D. V. (1977). Jackknifing in unbalanced situations. *Technometrics* **19** 285–292.

HORN, S. D. and HORN, R. A. (1975). Comparison of estimators of heteroscedastic variances in linear models. *J. Amer. Statist. Assoc.* **70** 872–879.

HORN, S. D., HORN, R. A. and DUNCAN, D. B. (1975). Estimating heteroscedastic variances in linear models. *J. Amer. Statist. Assoc.* **70** 380–385.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

KISH, L. and FRANKEL, M. (1974). Inference from complex samples (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 1–37.

LEHMANN, E. L. (1983). *Theory of Point Estimation.* Wiley, New York.

LOH, W. Y. (1987). Calibrating confidence coefficients. *J. Amer. Statist. Assoc.* To appear.

MCCARTHY, P. J. (1969). Pseudo-replication: Half-samples. *Rev. Internat. Statist. Inst.* **37** 239–264.

MCCULLAGH, P. (1983). Quasi-likelihood functions. *Ann. Statist.* **11** 59–67.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MILLER, R. G. (1974a). The jackknife–a review. *Biometrika* **61** 1–15.

MILLER, R. G. (1974b). An unbalanced jackknife. *Ann. Statist.* **2** 880–891.

NOBLE, B. (1969). *Applied Linear Algebra.* Prentice-Hall, New York.

QUENOUILLE, M. (1956). Notes on bias in estimation. *Biometrika* **43** 353–360.

RAO, C. R. (1970). Estimation of heteroscedastic variances in linear models. *J. Amer. Statist. Assoc.* **65** 161–172.

RAO, J. N. K. (1973). On the estimation of heteroscedastic variances. *Biometrics* **29** 11–24.

RAO, J. N. K. and WU, C. F. J. (1985a). Bootstrap inference for sample surveys. In *Proc. 1984 ASA Meeting, Section on Survey Research Methods* 106–112.

RAO, J. N. K. and WU, C. F. J. (1985b). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *J. Amer. Statist. Assoc.* **80** 620–630.

RUBIN, D. B. (1978). A representation for the regression coefficients in weighted least squares. ETS Research Bulletin RB-78-1, Princeton, N.J.

SHAO, J. and WU, C. F. J. (1985). Robustness of jackknife variance estimators in linear models. Technical Report No. 778, Univ. of Wisconsin-Madison.

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195.

SUBRAHMANYAM, M. (1972). A property of simple least squares estimates. *Sankhyā Ser. B* **34** 355–356.

TUKEY, J. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Statist.* **29** 614.

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

WILLIAMS, E. J. (1959). *Regression Analysis*. Wiley, New York.

WU, C. F. J. (1984). Jackknife and bootstrap inference in regression and a class of representations for the LSE. Technical Report No. 2675, Mathematics Research Center, Univ. of Wisconsin-Madison.

WU, C. F. J. (1985). Statistical methods based on data resampling. Special invited paper presented at IMS meeting in Stony Brook.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN 53706

DISCUSSION

RUDOLF BERAN

University of California, Berkeley

My comments center on three topics: the resampling algorithm of Section 7 as a bootstrap algorithm; criteria for assessing performance of a confidence set; and robustifying jackknife or bootstrap estimates for variance and bias. It will be apparent that I do not accept several of Wu's conclusions, particularly those concerning the bootstrap. The implied criticism does not diminish the paper's merit in advancing jackknife theory for the heteroscedastic linear model.

1. The bootstrap idea is a statistical realization of the simulation concept: one fits a plausible probability model to the data and acts thereafter as though the fitted model were true. Suppose that the errors $\{e_i\}$ in the linear model (2.1) are independent and that the c.d.f. of e_i is $F(\cdot/\sigma_i)$, where F has mean zero and variance one. Consistent estimates of the $\{\sigma_i\}$ and of F are not available, in general. Nevertheless, let $\hat{\sigma}_{n,i}$ be an esimate of σ_i, such as $\hat{\sigma}_{n,i} = |r_i|(1 - w_i)^{-1/2}$ or $\hat{\sigma}_{n,i} = |r_i|(1 - n^{-1}k)^{-1/2}$, and let $\hat{F}_n$ be any c.d.f. with mean zero and variance one. The fitted model here is the heteroscedastic linear model parametrized by the estimates $\hat{\beta}_n, \{\hat{\sigma}_{n,i}\}$ and $\hat{F}_n$. The appropriate bootstrap algorithm, which I will call the heteroscedastic bootstrap, draws samples from this fitted model.

Section 7 of the paper describes just this resampling procedure, without recognizing it as a bootstrap algorithm suitable for the heteroscedastic linear model. The two bootstrap algorithms that are discussed critically in Section 2 are not even intended for the heteroscedastic linear model. The first is designed for the homoscedastic linear model; the second for linear predictors based on multivariate i.i.d. samples (Freedman (1981)).

Let $B_n(\beta, \{\sigma_i\}, F)$ and $V_n(\beta, \{\sigma_i\}, F)$ be the bias and variance of $g(\hat{\beta}_n)$ under the heteroscedastic model described in the preceding paragraphs. The ap-