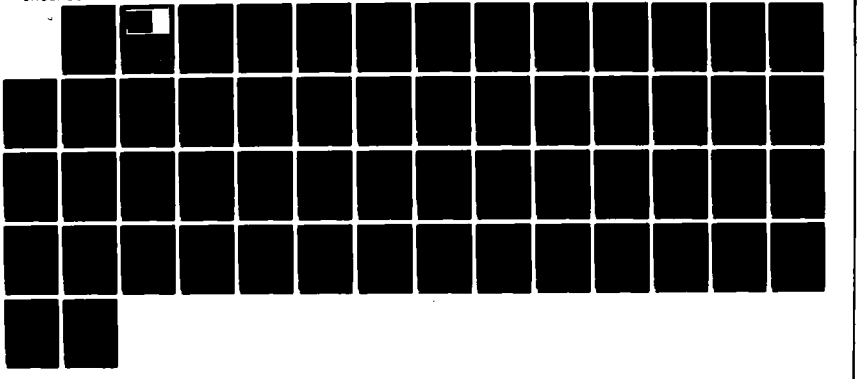
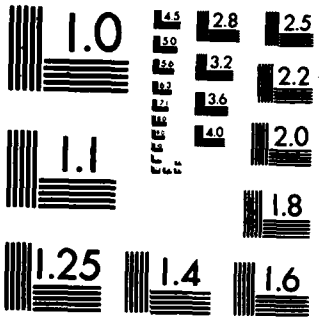
AD-A141 505 JACKKNIFE AND BOOTSTRAP INFERENCE IN REGRESSION AND A 1/1
CLASS OF REPRESENTA..(U) WISCONSIN UNIV-MADISON
MATHEMATICS RESEARCH CENTER C F WU APR 84 MSR-TSR-2675
UNCLASSIFIED DAAG29-80-C-0041 F/G 12/1 NL





MRC Technical Summary Report # 2675

JACKKNIFE AND BOOTSTRAP INFERENCE IN REGRESSION AND A CLASS OF REPRESENTATIONS FOR THE LSE

C. F. Jeff Wu

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53705

April 1984

(Received February 29, 1984)

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

JACKKNIFE AND BOOTSTRAP INFERENCE IN REGRESSION AND A CLASS OF REPRESENTATIONS FOR THE LSE

C. F. Jeff Wu

Technical Summary Report #2675

April 1984

ABSTRACT

A class of representations for the least squares estimator is presented and their applications sketched. Partly motivated by one such representation, the author ~~we~~ proposes a class of weighted jackknife estimators of variance of the least squares estimator by deleting any fixed number of observations at a time. These estimators are unbiased for homoscedastic errors and a special case, the delete-one jackknife variance estimator, is almost unbiased for heteroscedastic errors. The method is extended in various ways, including the use of the jackknife histogram, for variance and interval estimation with nonlinear parameters. Three bootstrap methods are considered. It is shown that none of them has the robustness property enjoyed by the (weighted) delete-one jackknife. Subset sampling with variable subset size is also considered. Several bias-reducing estimators are proposed. They are motivated by the observation that bias-reduction is mathematically equivalent to unbiased estimation of variance. Some simulation results on estimating the ratio of two normal parameters are reported.

AMS (MOS) Subject Classifications: 62J05, 62J02, 62G15

Key words: jackknife percentile method, subset sampling, variance estimator, bias reduction, Fieller's method, representation of the least squares estimator, robust regression.

Work Unit Number 4 - Statistics and Probability

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041.
Also, supported by the Alfred P. Sloan Foundation for Basic Research.

SIGNIFICANCE AND EXPLANATION

The Quenouille-Tukey jackknife is an old tool for bias reduction and non-parametric variance estimation. Recently Efron introduced the bootstrap method as a more versatile tool. It seems to have the potential to be useful in many kinds of problems involving estimation of error. These tools are not quite well developed for regression models. We propose a class of weighted jackknife methods that recompute the least squares estimates by deleting any fixed number of observations at a time. The key step is to weight each subset least squares estimate with the determinant of the Fisher information matrix of the subset. Some desirable properties of the procedures are proved. For nonlinear parameters, the methods are useful for bias reduction and variance estimation. Since we do not restrict to the classical delete-one jackknife, confidence intervals can be constructed from the histogram of some proper estimates from the resamples. The utility of the classical jackknife method is broadened with this new tool. On the other hand we show that the existing bootstrap methods may not work so well in the regression situation. Some simulation results are presented and further questions for research are raised.

The responsibility for the wording and views expressed in this descriptive summary lies with MRC, and not with the author of this report.

JACKKNIFE AND BOOTSTRAP INFERENCE IN REGRESSION AND A CLASS OF REPRESENTATIONS FOR THE LSE

C. F. Jeff Wu

1. Introduction

In the first part of this paper we show that the full-data least squares estimate $\hat{\beta}$ (LSE) can be represented as a weighted average of the LSE's $\hat{\beta}_s$ from all subsets s of a fixed size with the weight proportional to the determinant of the $X_s^T X_s$ matrix associated with the subset (Theorem 2), i.e.,

(1.1) $\hat{\beta}$ = sum of $w_s \hat{\beta}_s$ over all subsets of a fixed size, $w_s \propto |X_s^T X_s|$, $\sum_s w_s = 1$.

Instead of averaging over all subsets of a fixed size, we may consider drawing samples from the full data according to a resampling scheme, and computing the LSE and the determinant of the corresponding $X^T X$ matrix from each such sample. One main result (Theorem 1) states that the above representation still holds for any resampling method that is symmetric and nondegenerate with positive probability (Assumption B of Section 3), including the jackknife and the bootstrap. Several implications of the representation result are sketched in Section 3. A major one is in suggesting a new class of robust regression estimators. The details are in Section 4.

The representation (1.1) involves a linear function of $\hat{\beta}_s$. To estimate the variance-covariance matrix (henceforth abbreviated as variance) of $\hat{\beta}$, it seems natural to look at a quadratic function of $\hat{\beta}_s$. A quadratic extension of (1.1) is

(1.2) $$\xi \sum w_s (\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T, \quad w_s \text{ in (1.1)}$$

where the summation is over all subsets of size r. It turns out that the choice $\xi = (r-k+1)/(n-r)$, n = # of observations, k = # of regression parameters, makes (1.2) an unbiased estimator of the variance of $\hat{\beta}$ if the errors are uncorrelated with mean zero and constant variance (Theorem 3). This estimator is denoted as $v_{J,r}$ in (5.1).

Sponsored by the United States Army under Contract No. DAAG29-80-C-0041. Also, supported by the Alfred P. Sloan Foundation for Basic Research.

The second and major part of this paper deals with the jackknife and bootstrap resampling methods for variance and interval estimation and bias reduction. The method (1.2) can be viewed as a weighted jackknife by deleting every subset of size $n-r$ from the full-data. The purpose of the adjustment factor $\xi = (r-k+1)/(n-r)$ in (1.2) is to make the distance of $\sqrt{\xi}(\hat{\beta}_s - \hat{\beta})$ match the distance of $\hat{\beta}-\beta$. For example, if $r = n-1$ (delete-one-at-a-time), $\hat{\beta}_s$ is too close to $\hat{\beta}$. It is necessary to multiply the weighted sum of squares in (1.2) by a large factor $\xi = n-k$. Further attention is paid to the two extreme choices of the subset size r. If $r = k =$ # of regression parameters, it turns out that $v_{J,k}$ is identical to the usual variance estimator (by assuming equal variances). Theorem 4 provides the details, including the necessary modification of the definition (1.2) when some subsets are associated with singular X matrices. The other extreme is the delete-one jackknife, $r = n-1$. Our proposal is closely related to a delete-one jackknife proposed by Hinkley (1977). The main difference is that Hinkley's estimator uses weights proportional to the square of $|X_s^T X_s|$ and is therefore a biased estimator of the variance of $\hat{\beta}$. Both delete-one jackknife variance estimators are robust against error variance heterogeneity in that their biases converge to zero as $n \to \infty$ under (the same) weak regularity conditions. Hinkley's estimator does not fare well in the empirical study reported in Section 10.

In practice the resampling methods of inference are only used in situations where no closed form of the variance (or other measures of variability) of the point estimator is available. In Section 7 we consider extensions of the above method to parameters $\theta = g(\beta)$ which are nonlinear functions of the regression parameters β. An obvious extension is to replace $\hat{\beta}$, $\hat{\beta}_s$ in (1.2) by their counterparts $g(\hat{\beta})$, $g(\hat{\beta}_s)$. The scale factor ξ is applied after the nonlinear transformation g, (7.1). Another approach is to incorporate this scale adjustment internally before applying the transformation g, (7.2). To obtain confidence intervals for θ without computing variance estimates, we propose a jackknife percentile method through the construction of a weighted empirical distribution function of some estimates of θ based on the same subsets with the same weight w_s in (1.1). Here we find it more natural to estimate θ with the internal

adjustment method. The jackknife percentile method is similar in spirit to Efron's (1982) bootstrap percentile method. There is some theoretical advantage in using the percentile method since the possible skewness in the original point estimate $\hat{\theta}$ will be reflected in the histogram of the resampled estimates. Extensions to nonlinear regression models are briefly outlined. The jackknife methodology has long been associated with the delete-one jackknife. It has mainly been used for variance estimation and bias reduction. The method proposed here overcomes these limitations by allowing the deletion of more than one observation and the construction of the jackknife histogram. Further discussion is given in Section 7.

Other resampling methods are studied in Section 8. The subset sampling method is an extension of the jackknife by allowing different subset sizes. The variance estimator (1.2) is extended to this situation. Three bootstrap methods of variance estimation are considered. Two of them do not in general give unbiased variance estimators in the equal variance case, as is shown by a counterexample. The third one by bootstrapping the residuals is known to be identical to the usual variance estimator (8.19) in the case of linear parameters. The latter estimator is unbiased in the equal variance case but is biased for unequal variances.

The issue of bias reduction is studied in Section 9. It is shown that bias reduction is achievable if and only if the variance of $\hat{\beta}$ can be estimated unbiasedly (apart from a lower order term). Based on this connection, several estimators of the bias of $\hat{\theta}$ are proposed as natural counterparts of the variance estimators considered before. Conditions under which these estimators achieve bias reduction are given in Theorems 7, 8 and Corollaries 3, 4.

Several jackknife and bootstrap methods 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 the nonlinear parameter $\theta = -\beta_1/(2\beta_2)$, 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. Further questions are raised in Section 11.

2. Some matrix lemmas

For a matrix X of order $n \times k$, let X_s be its $r \times k$ submatrix consisting of the $i_1^{th},\ldots,i_r^{th}$ rows, $s = (i_1,\ldots,i_r)$, and $X^{(j)}$ be the $n \times (k-1)$ submatrix obtained from deleting the j^{th} column of X. For a square matrix A of order k, its adjoint is defined as the $k \times k$ matrix

$$\text{adj } A = [c_{ij}],\ 1 \leq i, j \leq k \tag{2.1}$$

with its (i,j) element $c_{ij} = (-1)^{i+j}M_{ji}$ and M_{ji} is the determinant of the $(k-1) \times (k-1)$ submatrix of A with the j^{th} row and i^{th} column deleted. Let $|A|$, A^{-1}, A^T be respectively the determinant, inverse and transpose of A. Recall $A^{-1} = \text{adj } A/|A|$, if A^{-1} exists.

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

(i) $$|X^TZ| = \sum_{s \in S_k} |X_s|\,|Z_s| \ ,$$

(ii) $$|X^TZ| = \binom{n-k}{r-k}^{-1} \sum_{s \in S_r} |X_s^TZ_s| \quad \text{for any } r \geq k \ ,$$

where

$$S_r = \text{all subsets } s \text{ of size } r \ . \tag{2.2}$$

(Note that X_s and Z_s are square matrices for $s \in S_k$).

Proof: Lemma 1(i) is in Noble (1969, p. 226). Lemma 1(ii) is obtained by applying Lemma 1(i) to each term $|X_s^TZ_s|$ and to $|X^TZ|$. □

Lemma 2. Let X be an $n \times k$ matrix, $n \geq k$. Then

$$\text{adj } X^TX = \binom{n-k+1}{r-k+1}^{-1} \sum_{s \in S_r} \text{adj } X_s^TX_s, \ r \geq k \ . \tag{2.3}$$

If $X_s^TX_s$ are nonsingular for all $s \in S_r$,

$$|X^TX|(X^TX)^{-1} = \binom{n-k+1}{r-k+1}^{-1} \sum_{s \in S_r} |X_s^TX_s|(X_s^TX_s)^{-1} \ . \tag{2.4}$$

Proof: From definition, the (i,j) elements of adj X^TX and adj $X_s^TX_s$ 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 from

deleting the i^{th} column of X_s. Therefore (2.3) is equivalent to

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

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 $k-1$. (2.4) follows from (2.3) since adj $A = |A|A^{-1}$. □

Lemma 3. Let Z be an $r \times k$ matrix, $r \geq k$. If $|Z^TZ| = 0$, then $|Z^TW| = 0$ for any $r \times k$ matrix W.

Proof: Since $|Z^TZ| = 0$, Z is not of full rank, which implies Z^TW is singular and $|Z^TW| = 0$. □

3. Representations of the least squares estimator

To motivate the general representation result, let us first consider the simple linear regression model

$$y_i = \alpha + \beta x_i + e_i, \ i = 1,\ldots,n \tag{3.1}$$

with $Ee_i = 0$, $Ee_i^2 = \sigma^2$ and $\text{cov}(e_i,e_j) = 0$ for $i \neq j$. The ordinary least squares estimator (LSE) $\hat{\beta}$ of β has several equivalent expressions,

$$\hat{\beta} = \sum_1^n (y_i - \bar{y})(x_i - \bar{x}) / \sum_1^n (x_i - \bar{x})^2$$

$$= \sum_{i<j} (y_i - y_j)(x_i - x_j) / \sum_{i<j} (x_i - x_j)^2 \tag{3.2}$$

$$= \sum_{i<j} u_{ij} \hat{\beta}_{ij} \ , \tag{3.3}$$

where

$$\hat{\beta}_{ij} = \frac{y_i - y_j}{x_i - x_j} \tag{3.4}$$

are the pairwise slopes for $x_i \neq x_j$ and

$$u_{ij} = \frac{(x_i - x_j)^2}{\sum_{i<j} (x_i - x_j)^2} \ .$$

To validate the step from (3.2) to (3.3), $u_{ij}\hat{\beta}_{ij}$ in (3.3) is defined to be zero for $x_i = x_j$. One can now interpret $\hat{\beta}$ as a weighted average of all the least squares estimates $\hat{\beta}_{ij}$ based on the (i,j) pairs of observations, with the weight proportional

to $(x_i - x_j)^2$, which happens to be the determinant of $X_{ij}^T X_{ij}$, where

$$X_{ij} = \begin{bmatrix} 1 & x_i \\ 1 & x_j \end{bmatrix}$$

is the design matrix corresponding to the (i,j) observations. It seems natural to guess the following extension for the general linear model: the LSE based on the full data set is equal to a weighted average of the LSE's based on all subsets of fixed size with the weight proportional to the determinant of the X^TX matrix corresponding to the subset. In fact a more general result will be shown to be true.

Throughout the paper we assume the following general linear model:

$$y_i = x_i^T\beta + e_i \quad , \quad i = 1,\ldots,n$$

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 $X = [x_1,\ldots,x_n]^T$, (3.4) can be rewritten as

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

We always assume X^TX is nonsingular. The ordinary least squares estimator (LSE) based on the full data (y,X) is

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

In Theorem 1 $\hat{\beta}$ is related to the LSE's based on values "resampled" from the full data (y,X). A brief discussion of resampling procedures is given next.

The full data, $z_1 = (y_1,x_1),\ldots,z_n = (y_n,x_n)$ are thought of as being observed and fixed. A resample of $(z_i)_1^n$ is a reweighted version of $(z_i)_1^n$ with weight $P_i^* \geq 0$. The vector $P^* = (P_1^*,\ldots,P_n^*)$ is called a <u>resampling vector</u>. For each P^*, the corresponding least squares estimate β^* is based on P_i^* "copies" of z_i, i.e.,

(3.7) $$\beta^* = (X^TD^*X)^{-1}X^TD^*y, \; D^* = \mathrm{diag}(P_1^*,\ldots,P_n^*)$$

is a weighted least squares estimate with weight proportional to P_i^*. Let "*" denote the joint distribution of $(P_i^*)_1^n$ under a resampling procedure. The expectation under repeated sampling according to the given resampling procedure 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 $(i_1,\ldots,i_k)$ of size k, k = # of parameters in (3.5).

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

(B) 1. The n random variable $\{P^*_i\}_1^n$ are exchangeable.

2. $\text{Prob}_*(\text{support size of } P^* > k) > 0$, where the support size of P^* is the total number of i's with $P^*_i > 0$.

It will be shown after Theorem 1 that several important resampling procedures satisfy the assumption (B).

Our first major result states that the full-data LSE $\hat{\beta}$ is a weighted average of the resampled-data LSE's β^* with weight proportional to $|X^TD^*X|$ for any resampling procedure satisfying (A).

Theorem 1. For any resampling method $*$ satisfying the assumption (A), the LSE $\hat{\beta}$ based on the full data can be represented as

$$\hat{\beta} = \frac{E_*|X^TD^*X|\beta^*}{E_*|X^TD^*X|} , \tag{3.8}$$

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

Proof: First consider the D^* with nonsingular X^TD^*X. Since β^* is the solution to the equation $(X^TD^*X)\beta^* = X^TD^*y$, from Cramer's rule (Noble, 1969, p.209), the j^{th} element of β^* is equal to the ratio of the determinant of the matrix obtained by replacing the j^{th} column of X^TD^*X by the vector X^TD^*y over the determinant of X^TD^*X. Notationally,

$$\beta^*_j = \frac{|X^TD^*X^{(j)}(y)|}{|X^TD^*X|} ,$$

where $X^{(j)}(y)$ is the $n \times k$ matrix obtained by replacing the j^{th} column of X by y. This establishes

$$|X^TD^*X^{(j)}(y)| = |X^TD^*X|\beta^*_j \quad \text{for nonsingular } X^TD^*X . \tag{3.9}$$

For singular $X^T D^* X$,

(3.10) $$|X^T D^* X^{(j)}(y)| = 0$$

follows from Lemma 3. From (3.9) and (3.10), the j^{th} element of the right hand side of (3.8) equals

(3.11) $$\frac{E_*|X^T D^* X^{(j)}(y)|}{E_*|X^T D^* X|} = \frac{E_* \sum' |X_s| \, |D_s^*| \, |X_s^{(j)}(y)|}{E_* \sum' |X_s|^2 |D_s^*|} ,$$

where $\sum'$ is summation over s in S_k, $D_s^* = \mathrm{diag}(P_{i_1}^*, \ldots, P_{i_k}^*)$ is the diagonal submatrix of D^* corresponding to the subset $s = (i_1, \ldots, i_k)$ and (3.11) follows from Lemma 1(i). Since $E_*|D_s^*| = E_*\left(\prod_{j=1}^{k} P_{i_j}^*\right) = a_k > 0$ independent of s from the assumption (A), (3.11) equals

$$\frac{\sum' |X_s| \, |X_s^{(j)}(y)|}{\sum' |X_s|^2} = \frac{|X^T X^{(j)}(y)|}{|X^T X|} ,$$

which is equal to the j^{th} element of $\hat{\beta}$. The proof is completed.

An important resampling procedure that satisfies (B) is the bootstrap (Efron, 1979). A simple random sample $z_1^*, \ldots, z_m^*$ of fixed size m, is drawn with replacement from the observed sample $z_1, \ldots, z_n$. Let $P_i^* = \#(z_j^* = z_i, \, j = 1, \ldots, m)$. Then $P^* = (P_1^*, \ldots, P_n^*)$ follows a multinomial distribution,

(3.12) $$P^* \sim \mathrm{Mult}_m(n, \tfrac{1}{n} \underset{\sim}{1}), \quad \underset{\sim}{1} = (1, \ldots, 1)$$

with m independent draws on n categories each having probability $1/n$. For $m > k$, (3.12) satisfies (B) and the representation result of Theorem 1 applies to the bootstrap method. Note that the bootstrap sample size m need not be the same as the original sample size n.

Another example is the jackknife method, which computes the LSE by deleting any subset of size d or equivalently by retaining any subset of size $r = n-d$. The jackknife resampling with fixed size r is defined by

(3.13) $$\mathrm{Prob}_*(P_i^* = 1 \text{ for } i \in s, = 0 \text{ for } i \notin s) = \binom{n}{r}^{-1} \text{ for all } s \in S_r .$$

Let $\hat{\beta}_s$ denote the LSE based on the subset of observations z_{i_j}, $i_j \in s$,

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

where $y_s = (y_{i_1}, \ldots, y_{i_r})^T$. As a special case of Theorem 1, we have

Theorem 2. For any $r > k$,

(3.15) $$\hat{\beta} = \frac{\sum_{s \in S_r} |X_s^T X_s| \hat{\beta}_s}{\sum_{s \in S_r} |X_s^T X_s|} = \frac{\sum_{s \in S_r} |X_s^T X_s| \hat{\beta}_s}{\binom{n-k}{r-k} |X^T X|} \quad ,$$

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

The second identity of (3.15) follows from Lemma 1(ii). Theorem 2 is the extension anticipated in the beginning of this section.

In general the bootstrap and the jackknife provide different representations of $\hat{\beta}$. But when the bootstrap sample size is k, $|X^T D^* X| = 0$ for any bootstrap sample with support size less than k. The remaining bootstrap samples with support size k are identical to the jackknife samples of size k. Therefore the bootstrap resampling and the jackknife resampling, both with resample size k, give the same representation of $\hat{\beta}$.

Theorem 2 was proved by Subrahmanyam (1972) for the special case $r = k$ and extended to general r by Hoerl and Kennard (1980). (They were kindly brought to my attention by D. B. Rubin and W. Y. Tsai.) The singularity problem of $X_s^T X_s$ was not rigorously handled in their theorems and proofs. The same problem will surface again in variance estimation (Section 5), where such a negligence will lead to incorrect results.

The subset LSE $\hat{\beta}_s$ is related to the full-data LSE $\hat{\beta}$ by (Bingham, 1977; Cook and Weisberg, 1982, p. 136)

(3.16) $$\hat{\beta}_s = \hat{\beta} - (X^T X)^{-1} X_{\bar{s}}^T r_{\bar{s},s} = \hat{\beta} - (X_s^T X_s)^{-1} X_{\bar{s}}^T r_{\bar{s}}$$

where $\bar{s}$ is the complement of s and

(3.17) $$r_{\bar{s}} = y_{\bar{s}} - X_{\bar{s}} \hat{\beta} \quad ,$$
$$r_{\bar{s},s} = y_{\bar{s}} - X_{\bar{s}} \hat{\beta}_s \quad ,$$

are the vectors of residuals in $\bar{s}$ from fitting the full-data LSE $\hat{\beta}$ and the subset LSE $\hat{\beta}_s$ respectively. Theorem 2 can be restated in terms of the residuals $r_{\bar{s}}$, $r_{\bar{s},s}$ associated with the discarded subsets $\bar{s}$.

Corollary 1. For any $r > k$, $\underset{s}{r}$, $\underset{s,s}{r}$ defined in (3.17),

$$\sum_{s \in S_r} |x_s^T x_s| x_s^T \underset{s,s}{r} = \sum_{s \in S_r} |x_s^T x_s| (x_s^T x_s)^{-1} x_s^T \underset{s}{r} = 0 \quad , \tag{3.18}$$

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

For $r = n-1$ (delete - 1 jackknife), (3.18) reduces to the familiar normal equation $\Sigma\, x_i r_i = 0$, $r_i = y_i - x_i^T \hat{\beta}$. Formula (3.18) may be useful in regression diagnostics when the diagnostic statistics involve deleting more than one observation.

Theorem 2 can be trivially extended to the weighted least squares estimators. Let $W = \mathrm{diag}(u_1, \ldots, u_n)$ 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

$$\hat{\beta}^* = (X^T W^{-1} X)^{-1} X^T W^{-1} y, \quad \hat{\beta}_s^* = (X_s^T W_s^{-1} X_s)^{-1} X_s^T W_s^{-1} y_s \quad . \tag{3.19}$$

By applying the transformation $W^{-1/2}$ to X and y, and $W_s^{-1/2}$ to X_s and y_s, Corollary 2 follows from Theorem 2.

Corollary 2. For $r > k$,

$$\hat{\beta}^* = \frac{\sum_{s \in S_r} |x_s^T W_s^{-1} x_s| \hat{\beta}_s^*}{\sum_{s \in S_r} |x_s^T W_s^{-1} x_s|} \quad , \tag{3.20}$$

where the terms with singular $x_s^T W_s^{-1} x_s$ are defined to be zero.

Formula (3.20) for $r = k$ is of particular interest, since $\hat{\beta}_s^*$ is identical to the unweighted LSE $\hat{\beta}_s = X_s^{-1} y_s$ if X_s^{-1} exists. Therefore the weighted LSE $\hat{\beta}^*$ is a convex combination of the unweighted LSE's $\hat{\beta}_s$ based on the subsets of size k. As a consequence, the collection of the weighted LSE's with any positive weight matrix is contained in the bounded convex hull spanned by the finite number of unweighted LSE's based on all subsets of size k. Rubin (1978) proved this result and noted its use in proving the convergence of certain iterative reweighted least squares algorithms as was later done in Dempster, Laird and Rubin (1980).

Koenker and Bassett (1978) introduced a concept of "regression quantiles" for the linear models as a natural generalization of the c[illegible]inary sample quantiles for the location model. As a special case, the least absolute deviation estimator is the regression median. It turns out that the set of regression quantiles is identical to the convex hull of $\hat{\beta}_s$ with s in S_k (Theorem 3.1, Koenker and Bassett, 1978). This connection suggests that the representation (3.15) may be relevant in robust regression. In the next section a class of robust regression estimators will be proposed by exploiting this representation.

4. An application: some new robust regression estimators

As in the previous section we shall use the simple linear regression model (3.1) to illustrate the main idea. The least squares estimator $\hat{\beta}$ of the slope parameter, being a weighted average of the pairwise slopes $\hat{\beta}_{ij}$ (3.4), is not robust in the sense that it can be heavily influenced by a few extreme values of y. Theil (1950) and Sen (1968) suggested the (unweighted) medians of $\hat{\beta}_{ij}$ as a more robust estimator. Jaeckel (1972) considered the weighted medians of $\hat{\beta}_{ij}$. Jaeckel (1972), Scholz (1978), and Sievers (1978) proved that an asymptotically optimal choice of weight is $c|x_j - x_i|$. Note that $|x_j - x_i|$ is different from the weight $(x_j - x_i)^2$ in the representation (3.3). From the optimality property of the least squares estimator, the weight $(x_j - x_i)^2$ is optimal among all linear unbiased estimators of β.

The representation of $\hat{\beta}$ in terms of $\hat{\beta}_{ij}$ suggests a host of robust modifications. An important class is the following weighted trimmed regression estimators. Let $I = \{(i,j) : 1 \le i < j \le n, x_i \ne x_j\}$ and $|I| = t$. Order the $\hat{\beta}_{ij}$, $(i,j) \in I$, into $\hat{\beta}^{(1)} \le \hat{\beta}^{(2)} \le \ldots \le \hat{\beta}^{(t)}$. Let w_{ij} be the weight associated with $\hat{\beta}_{ij}$ and $w_{(i)}$ the corresponding weight associated with $\hat{\beta}^{(i)}$. The (α_1,α_2) - trimmed regression estimator is defined as

$$\hat{\beta}_{tr} = \sum_{m_1+1}^{k-m_2} w_{(i)}\hat{\beta}^{(i)} \Big/ \sum_{m_1+1}^{k-m_2} w_{(i)} \ , \tag{4.1}$$

$$\sum_{1}^{m_1} w_{(i)} = \alpha_1, \ \sum_{k-m_2+1}^{k} w_{(i)} = \alpha_2 \ .$$

In $\hat{\beta}_{tr}$, the lower $100\ \alpha_1$% and the upper $100\ \alpha_2$% (according to the weighted empirical

distribution of $\hat{\beta}^{(i)}$ with weight $w_{(i)}$) of the pairwise slopes $\hat{\beta}_{ij}$ are trimmed. The trimmed regression estimator (4.1) covers both the least squares estimator and the weighted median estimator as α_1 and α_2 vary. From the above discussion, for α_1, α_2 close to zero $w_{ij} \propto (x_j - x_i)^2$ should be chosen; and for α_1, α_2 close to 0.5, $w_{ij} \propto |x_j - x_i|$ should be chosen. The optimal choice of w_{ij} for general α_1 and α_2 depends on the asymptotic distribution of $\hat{\beta}_{tr}$, which is beyond the scope of the paper.

Assume $k = \binom{n}{2}$, i.e. $x_i < x_j$ for all $i < j$. The breakdown point (Huber, 1981) of the unweighted (α,α) - trimmed regression estimator (4.1) (with $w_{ij} = 1$) is computed as follows. Let m of the n y_i values be perturbed. The percentage of the pairwise slopes $\hat{\beta}_{ij}$ not affected by the perturbation is

$$1 - \frac{\binom{n-m}{2}}{\binom{n}{2}} = 1 - (1 - \frac{m}{n})(1 - \frac{m}{n-1}) \simeq 2f - f^2 \quad , \quad f = \frac{m}{n}$$

which equals α iff the percentage of the perturbed y values equals $1 - \sqrt{1-\alpha}$,

$$f^* = 1 - \sqrt{1-\alpha} \quad , \tag{4.2}$$

which is the desired breakdown point. For small α, $f^* \simeq \frac{\alpha}{2}$. The breakdown point f^* as a function of α is given in the following table.

α	0.5	0.4	0.3	0.2	0.1
f^*	0.293	0.225	0.163	0.105	0.051

Notice that the Theil-Sen median regression estimator has a breakdown point 0.293. For general weighted trimmed regression estimators, no simple formula like (4.2) is available since it depends on the particular weight system. Robust regression estimators with high breakdown point are considered in Siegel (1982) and Rousseeuw (1984).

One can also consider the (α_1,α_2) - Winsorized regression estimator by taking the weighted (α_1,α_2) - Winsorized (Huber, 1981) mean of $\hat{\beta}^{(i)}$. Other robust alternatives are straightforward.

For the general regression model (3.5) and the subset least squares estimates $\hat{\beta}_s$ (3.14), a weighted trimmed mean of $\hat{\beta}_s$, $s \in S_r$, can be obtained by (i) ordering the vector-valued $\hat{\beta}_s$ according to some criterion, e.g., the Mahalanobis distance, convex hull trimming or ellipsoidal trimming (Titterington, 1978), (ii) trimming the extreme values and

(iii) taking a weighted average of the remaining ones. The weight can be proportional to $|X_s^T X_s|$, or to $|X_s^T X_s|^\lambda$ where the power λ is chosen between 0 and 1. Another way is to apply the weighted univariate trimming to each component of $\hat{\beta}_s$ separately.

For the delete-one jackknife, $r = n-1$, our proposal results in the following estimator

$$\hat{\beta}_{tr,J(1)} = \sum_{i=m_1+1}^{n-m_2} (1 - w_{(i)})^\lambda \hat{\beta}^{(i)} \Big/ \sum_{i=m_1+1}^{n-m_2} (1 - w_{(i)})^\lambda \quad , \tag{4.3}$$

where $\hat{\beta}^{(i)}$ are the ordered values of $\hat{\beta}_{-i}$, $i = 1(1)n$, $\hat{\beta}_{-i}$ = LSE of β with the i^{th} observation deleted, $w_{(i)}$ is the $w_j = x_j^T (X^T X)^{-1} x_j$ associated with $\hat{\beta}^{(i)}$. (In later sections we shall denote $\hat{\beta}_{-i}$ by $\hat{\beta}_{(i)}$.) Hinkley (1977) proposed a similar estimator, the main difference being that his "ordering" is on the weighted pseudo-values $Q_i = \hat{\beta} + n(1-w_i)(\hat{\beta}-\hat{\beta}_{-i})$. Without a detailed study, it is hard to judge their relative merits. We merely point out that $|Q_i - \hat{\beta}| > |\hat{\beta}_{-i} - \hat{\beta}|$ if $w_i < 1-n^{-1}$. This follows from (5.15) and (6.8).

Other regression L-estimators have been considered by Bickel (1973), Koenker and Bassett (1978), Ruppert and Carroll (1980). The main difference between our proposal and theirs is that the former directly involves repeated estimators of β while the latter depend on the residuals. This may provide a clue to the possible advantages of our proposal.

Finally we consider a class of robust regression M-estimator obtained by minimizing

$$\sum_{s \in S_r} w_s \eta(\hat{\beta}_s - \hat{\beta}) \quad , \tag{4.4}$$

where w_s may be proportional to $|X_s^T X_s|^\lambda$ or some other weight function, $\eta(\hat{\beta}_s - \hat{\beta}) = \psi(\|\hat{\beta}_s - \hat{\beta}\|)$, $\|\hat{\beta}_s - \beta\|$ is a distance measure of $\hat{\beta}_s - \beta$, and ψ is a scalar function discussed in Huber (1981). For $w_s = |X_s^T X_s|$ and η = Euclidean distance, the regression M-estimate (4.4) reduces to the ordinary least squares estimate via Theorem 2. Hinkley (1977) proposed a similar estimator for the delete-one jackknife in terms of the weighted pseudo-values Q_i. Since (4.4) requires more computation than the usual M-estimator, it can not be recommended for practical use until further desirable properties are documented.

5. General weighted jackknife in regression

Miller (1974) extended the ordinary (unweighted) jackknife to the regression situation and proved some asymptotic properties. Since the subset LSE's are not exchangeable, unweighted jackknifes do not seem natural. As a consequence, the acclaimed "bias-reducing" property of the jackknife in the location case is lost here and the unweighted jackknife variance estimator is biased even for linear parameters. Recognizing this problem, Hinkley (1977) proposed a weighted jackknife and demonstrated its desirable properties. Both dealt with jackknifing by deleting <u>one</u> observation at a time from the full data. In this section we propose a class of weighted jackknife variance estimators of the LSE $\hat{\beta}$ by deleting any fixed number of observations at a time. Jack-knife estimation for nonlinear parameters will be considered in Section 7.

Since jackknifing by deleting d observations is equivalent to retaining $r = n-d$ observations, all the results of this section will be in terms of $\hat{\beta}_s$, the LSE based on the subset s. The proposed jackknife variance estimator by recomputing the LSE for each subset of size r is

$$v_{J,r} = \frac{r-k+1}{n-r} \frac{\sum_{s \in S_r} |X_s^T X_s| (\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T}{\sum_{s \in S_r} |X_s^T X_s|} \tag{5.1}$$

$$= \binom{n-k}{r-k+1}^{-1} |X^T X|^{-1} \sum_{s \in S_r} |X_s^T X_s| (\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T \quad , \tag{5.2}$$

where $X_s^T X_s$ are assumed nonsingular for all $s \in S_r$ and (5.2) follows from (5.1) via Lemma 1(ii). Under the ideal assumption that the errors e_i in (3.5) have constant variance σ^2, we prove that $v_{J,r}$ satisfies the minimal requirement that it is an unbiased estimator of the variance of $\hat{\beta}$ under the same assumption. Other properties will be taken up in Sections 6 and 7.

<u>Theorem 3</u>. If $\mathrm{Var}(e) = \sigma^2 I$ in (3.5),

$$E(v_{J,r}) = \sigma^2 (X^T X)^{-1} = \mathrm{Var}(\hat{\beta}) \quad . \tag{5.3}$$

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 the set s,

$$|X_s^T X_s|(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T = |X_s^T X_s|(X_s^T X_s)^{-1} X_{s^r}^T X_{s^r} (X_s^T X_s)^{-1} \tag{5.4}$$

and its expectation is

$$\begin{aligned} &|X_s^T X_s|(X_s^T X_s)^{-1} X_s^T (I_s - X_s (X^T X)^{-1} X_s^T) X_s (X_s^T X_s)^{-1} \\ &= |X_s^T X_s|(X_s^T X_s)^{-1} - |X_s^T X_s|(X^T X)^{-1} \quad , \end{aligned} \tag{5.5}$$

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

$$\sum_{s \in S_r} |X_s^T X_s|(X_s^T X_s)^{-1} = \binom{n-k+1}{r-k+1} |X^T X|(X^T X)^{-1} \quad , \tag{5.5}'$$

and from Lemma 1(ii),

$$\sum_{s \in S_r} |X_s^T X_s|(X^T X)^{-1} = \binom{n-k}{r-k} |X^T X|(X^T X)^{-1} \quad ,$$

which, together with (5.5), imply

$$E\Big(\sum_{s \in S_r} |X_s^T X_s|(\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T \Big) = \binom{n-k}{r-k+1} |X^T X|(X^T X)^{-1} \quad ,$$

and thus proving the result. □

The factor

$$\xi = \frac{r-k+1}{n-r}$$

in $v_{J,r}$ can now be given a statistical interpretation. The sampling error $\hat{\beta}-\beta$ has variance $\sigma^2(X^T X)^{-1}$. Given $\hat{\beta}$, the "resampling error" $\hat{\beta}_s - \hat{\beta}$ has $\xi^{-1} v_{J,r}$ as its (weighted) "resampling variance". Due to the unbiasedness result (5.3), the original sampling error $\hat{\beta}-\beta$ and the resampling error $\hat{\beta}_s - \hat{\beta}$ have different stochastic orders. The purpose of the scale factor $\sqrt{\xi}$ is to make the two errors of the same order in the following sense,

$$\mathrm{Var}(\sqrt{\xi}(\hat{\beta}_s - \hat{\beta})) = \mathrm{Var}(\hat{\beta}-\beta) + \text{lower order terms} \quad \text{for all } s \ . \tag{5.6}$$

In fact, we have, from (5.3),

$$\sum_{s \in S_r} w_s \, \mathrm{Var}(\sqrt{\xi}(\hat{\beta}_s - \hat{\beta})) = \mathrm{Var}(\hat{\beta}-\beta) \quad , \tag{5.6}'$$

where the weight w_s is proportional to $|X_s^T X_s|$. In particular, for $p = 1$, this reduces to

$$\mathrm{Var}(\sqrt{\xi}(\hat{\beta}_s - \hat{\beta})) = \mathrm{Var}(\hat{\beta}-\beta) \quad ,$$

since w_s is a constant. In general, if the weights w_s are uniformly bounded away from 0 and 1, (5.6) follows from (5.6)'.

The implementation of $v_{J,r}$, $r < n-1$, and its extensions in more complex situations (some described in Sections 7 and 9) may be too cumbersome since $\binom{n}{r}$ computations of $\hat{\beta}_s$ and $|X_s^T X_s|$ are required. As in the bootstrap method, it is suggested to use a Monte Carlo approximation:

(i) draw J subsets randomly without replacement from S_r; denote the collection of the selected subsets by S,

(ii) compute the variance estimate

$$v'_{J,r} = \frac{r-k+1}{n-r} \frac{\sum_{s \in S} |X_s^T X_s| (\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T}{\sum_{s \in S} |X_s^T X_s|} .$$

When the subset size r equals the number of parameters k, the jackknife variance estimator $v_{J,k}$ can be defined <u>without</u> the additional assumption that $X_s^T X_s$ is nonsingular for any $s \in S_k$. For $s \in S_k$, $\hat{\beta}_s = X_s^{-1} y_s$, $\hat{\beta}_s - \hat{\beta} = X_s^{-1} r_s$, r_s in (5.4), 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

(5.7) $$v_{J,k} = \frac{1}{(n-k)|X^T X|} \sum_{s \in S_k} |X_s|^2 (\hat{\beta}_s - \hat{\beta})(\hat{\beta}_s - \hat{\beta})^T$$

(5.8) $$= \frac{1}{(n-k)|X^T X|} \sum_{s \in S_k} (\text{adj } X_s) r_s \; r_s^T (\text{adj } X_s)^T ,$$

where (5.7), a special case of (5.2), requires $|X_s| \neq 0$ for all $s \in S_k$ while (5.8) is well-defined without any additional restriction. The proof of Theorem 3 works for the variance estimator (5.8). The only change involves replacing (5.5)' by a similar identity in terms of the adjoint matrices (see (2.3)). For $v_{J,k}$ we can establish the following coincidental result, which also implies the conclusion of Theorem 3.

Theorem 4. When the subset size is equal to the number of regression parameters, the jackknife variance estimator $v_{J,k}$ (5.8), is identical to the usual unbiased variance estimator

$$\text{var} = \hat{\sigma}^2(X^TX)^{-1},\ \hat{\sigma}^2 = \frac{\underset{\sim}{r}^T\underset{\sim}{r}}{n-k},\ \underset{\sim}{r} = y - X\hat{\beta} \quad . \tag{5.9}$$

Proof: Note that

$$\begin{aligned}(\text{adj } X_s)r_s &= [(-1)^{i+j}|X^{(i)}_{s(j)}|]_{i,j}(r_{sj})_j \\ &= (\sum_j (-1)^{i+j} r_{sj}|X^{(i)}_{s(j)}|)_i = (|X^{(i)}_s(r_s)|)_i \quad ,\end{aligned} \tag{5.10}$$

where $X^{(i)}_{s(j)}$ = matrix obtained from deleting the j^{th} row and the i^{th} column of X_s, r_{sj} = j^{th} element of r_s, $X^{(i)}_s(r_s)$ = matrix obtained by replacing the i^{th} column of X_s by r_s. The last equation of (5.10) follows from the usual result on expansion of determinant (Noble, 1969, p. 208). From (5.10), the (i,j) element of the matrix $\sum (\text{adj } X_s)r_s r_s^T(\text{adj } X_s)^T$ in (5.8) is equal to

$$\sum |X^{(i)}_s(r_s)|\ |X^{(j)}_s(r_s)| = |X^{(i)T}(\underset{\sim}{r})X^{(j)}(\underset{\sim}{r})| \quad , \tag{5.11}$$

where $X^{(i)}(\underset{\sim}{r})$ is the matrix obtained by replacing the i^{th} column of X by the residual vector $\underset{\sim}{r}$. Since $X^{(i)}_s(r_s)$ is the $k \times k$ submatrix of $X^{(i)}(\underset{\sim}{r})$ with rows corresponding to the subset s, (5.11) follows from Lemma 1(i). Noting that $\underset{\sim}{r}$ is orthogonal to the other columns of $X^{(i)}(\underset{\sim}{r})$ from the normal equation $X^T\underset{\sim}{r} = 0$, the (i,j) element of $X^{(i)T}(\underset{\sim}{r})X^{(j)}(\underset{\sim}{r})$ is $\underset{\sim}{r}^T\underset{\sim}{r}$, and the other elements in its i^{th} row and j^{th} column are zero. This gives

$$|X^{(i)T}(\underset{\sim}{r})X^{(j)}(\underset{\sim}{r})| = (-1)^{i+j}\underset{\sim}{r}^T\underset{\sim}{r}|X^{(i)T}X^{(j)}| \quad , \tag{5.12}$$

where $X^{(i)}$ is the submatrix of X with its i^{th} column deleted. From (5.8), (5.11) and (5.12), we have

$$\begin{aligned}v_{J,k} &= \frac{\underset{\sim}{r}^T\underset{\sim}{r}}{(n-k)|X^TX|}[(-1)^{i+j}|X^{(i)T}X^{(j)}|]_{i,j} \\ &= \frac{\underset{\sim}{r}^T\underset{\sim}{r}}{(n-k)}\frac{\text{adj } X^TX}{|X^TX|} = \hat{\sigma}^2(X^TX)^{-1} \quad .\end{aligned}$$

□

A bootstrap resampling method also leads to the estimator $\widehat{var}$, (5.9). Details are in Section 8.

Theorem 4 was proved by Subrahmanyam (1972) for the variance estimator (5.7) (not the more general (5.8)) by assuming $|X_s| \neq 0$ for all s in S_k. It is important to distinguish (5.8) from (5.7) in case $|X_s| = 0$ for some s. For a subset 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 (5.7) to be zero as was done before in the representation theorem. This is obviously so since the more general expression $(\text{adj } X_s)r_s r_s^T(\text{adj } X_s)^T$ in (5.8) for singular or nonsingular X_s is nonnegative definite and is in general nonzero for singular X_s. Such an incorrect interpretation of (5.7) will lead to a variance estimator smaller than $\hat{\sigma}^2(X^TX)^{-1}$. A simple illustration follows.

Consider the simple linear regression model (3.1) with k = 2. The jackknife variance estimator $v_{J,2}$ for the slope parameter β has two forms

$$v_{J,2} = c \sum_{i<j}^{n} (x_i - x_j)^2 \left(\frac{y_i - y_j}{x_i - x_j} - \hat{\beta}\right)^2 \tag{5.13}$$

$$= c \sum_{i<j}^{n} (y_i - y_j - \hat{\beta}(x_i - x_j))^2 \quad , \tag{5.14}$$

where $c = (n(n-2)\Sigma_1^n(x_i - \bar{x})^2)^{-1}$, (5.13) comes from (5.7) and (5.14) from (5.8). In terms of the residuals $\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$, $(\hat{\alpha},\hat{\beta})$ = LSE of (α,β), (5.14) equals

$$c \sum_{i<j}^{n} (\hat{e}_i - \hat{e}_j)^2 = \left(\sum_{1}^{n} (x_i - \bar{x})^2\right)^{-1}(n-2)^{-1} \sum_{1}^{n} \hat{e}_i^2$$

Apart from the constant c, the contribution of the pair (i,j) with $x_i = x_j$ to the variance estimate is $(\hat{e}_i - \hat{e}_j)^2 = (y_i - y_j)^2$ which measures the variability within two repeated runs. Interpreting terms in (5.13) with $x_i = x_j$ as zero amounts to ignoring the internal variability of the responses with the same x value, thus leading to underestimation of the true error.

At the other end of the choice of r is the jackknife variance estimator $v_{J,n-1}$ obtained by taking every subset s of size n-1, or in a more familiar language, by deleting one observation at a time. For each s in S_{n-1}, let i denote the element not in s, and write $X_s = X_{(i)}$. We shall adopt the notation that the subscript "(i)" added

to a quantity means "with the i^{th} observation deleted," and in a similar spirit, use $v_{J(1)}$ for the "delete-one" jackknife variance estimator $v_{J,n-1}$. From $|X_{(i)}^T X_{(i)}| = (1-w_i)|X^T X|$, $w_i = x_i^T (X^T X)^{-1} x_i$, (5.2) and

$$\hat{\beta}_{(i)} = \hat{\beta} - (X^T X)^{-1} x_i r_i (1-w_i)^{-1} \quad , \tag{5.15}$$

where $\hat{\beta}_{(i)}$ is the LSE of β with the i^{th} observation deleted and $r_i = y_i - x_i^T \hat{\beta}$ is the i^{th} residual, $v_{J(1)}$ takes a simple form

$$v_{J(1)} = \sum_1^n (1-w_i)(\hat{\beta}_{(i)} - \hat{\beta})(\hat{\beta}_{(i)} - \hat{\beta})^T \tag{5.16}$$

$$= (X^T X)^{-1} \sum_1^n \frac{r_i^2}{1-w_i} x_i x_i^T (X^T X)^{-1} \quad . \tag{5.17}$$

It turns out that $v_{J(1)}$ enjoys a model-robustness property, which is the main theme of the next section.

6. Model-robustness of the weighted delete-one jackknife variance estimators

In Theorem 3 the general weighted jackknife variance estimators $v_{J,r}$ are shown to be unbiased for $\text{var}(\hat{\beta})$ if the errors are homoscedastic. It is natural to ask how $v_{J,r}$ will perform under violations of the homoscedasticity assumption. Since under the heteroscedasticity assumption $\text{var}(\underset{\sim}{e}) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, the variance of $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = (X^T X)^{-1} \sum_1^n \sigma_i^2 x_i x_i^T (X^T X)^{-1} \quad . \tag{6.1}$$

From comparing (5.17) and (6.1), it seems that $v_{J(1)}$ is robust for estimating $\text{Var}(\hat{\beta})$, (6.1), under the broader heteroscedasticity assumption. This remarkable aspect of both $v_{J(1)}$ and a related variance estimator will be treated in this section.

The asymptotic computations (as n becomes large) will be done under one or several of the following assumptions.

(C) 1. Let X_n denote the X matrix in (3.5) for n observations,

$\max_{1\le i\le n} x_i^T(X_n^TX_n)^{-1}x_i \le \frac{c}{n}$, c independent of n.

2. $\max_{1\le i\le\infty} \sigma_i^2 < \infty$.

3. The minimum and maximum eigenvalues of $\frac{1}{n}X_n^TX_n$ are uniformly bounded away from 0 and ∞.

4. The elements of X_n are uniformly bounded.

From comparing (5.17) and (6.1), 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 4. If

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

then

(6.3) $$Er_i^2 = (1-w_i)\sigma_i^2 , \quad w_i = x_i^T(X^TX)^{-1}x_i .$$

More generally, under the assumptions C1 and C2,

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

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

Proof: From $r_i = y_i - x_i^T\hat{\beta} = e_i - x_i^T(X^TX)^{-1}X^Te$,

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

where $w_{ij} = x_i^T(X^TX)^{-1}x_j$ and the second equality of (6.5) follows from $w_i = \sum_{j=1}^n w_{ij}^2$. It is now obvious that (6.3) follows from (6.2). Assuming C1 and C2,

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

which is of order n^{-1}. Therefore (6.4) follows from (6.5).

□

By comparing (5.17) and (6.1), the following result is obtained as a direct consequence of Lemma 4.

Theorem 5. Under (3.5)

(i) $Ev_{J(1)} = Var(\hat{\beta})$ under (6.2),

(ii) $Ev_{J(1)} = Var(\hat{\beta})(1 + O(n^{-1}))$ under (C1 - C2).

We are not able to prove a similar result for the more general $v_{J,r}$. We conjecture that $v_{J,r}$ is also robust in the above sense for r close to n. This is confirmed in 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 like $n^{-1}X_n^T X_n$ converging to a positive definite matrix is assumed (Miller, 1974). On the other hand, (6.2) is a more restrictive assumption. Let q be the number of different σ_i's in (6.2). Then the linear model (3.5) can be rewritten as

(6.6) $$y_{ik} = x_{ik}^T\beta + e_{ik},\ Ee_{ik} = 0,\ Var\ e_{ik} = \sigma_i^2,\ k = 1(1)n_i,\ i = 1(1)q\ ,$$

with uncorrelated errors. Let T_i be the subspace spanned by x_{ik}, $k = 1(1)n_i$. According to (6.2) T_i, $i = 1(1)q$, are <u>orthogonal</u> to each other with respect to the positive definite matrix $(X^TX)^{-1}$. Writing $\dim(T_i)$ = dimension of T_i in R^k, we have $\sum \dim(T_i) = k$ since T_i, $i = 1(1)q$, generates the column space of X, whose dimension is k. A special case of (6.6) is the k-sample problem with unequal variances. Let $x_{ik} = x_i$ for $k = 1(1)n_i$. Then x_i, $i = 1(1)q$, are orthogonal to each other w.r.t. $(X^TX)^{-1}$, which forces $q = k$. By writing $\theta_i = x_i^T\beta$, (6.6) becomes the k-sample problem

(6.7) $$y_{ik} = \theta_i + e_{ik},\ Ee_{ik} = 0,\ Var\ e_{ik} = \sigma_i^2,\ k = 1(1)n_i,\ i = 1(1)k\ .$$

The $k \times k$ matrix $Z^T = [x_1,\ldots,x_k]$ is nonsingular and $\theta = (\theta_1,\ldots,\theta_k)^T = Z\beta$ is a reparametrization of β. We shall come back to (6.7) later.

Closely related to our $v_{J(1)}$ is a weighted delete-one jackknife method first considered by Hinkley (1977). His approach is via the construction of pseudo-values in the hope that the nice properties of pseudo-values in the location model would carry over to the regression model. Specifically, define the i^{th} weighted pseudo-value

(6.8) $$Q_i = \hat{\beta} + n(1-w_i)(\hat{\beta}-\hat{\beta}_{(i)}) = \hat{\beta} + n(X^TX)^{-1}x_i r_i,\ w_i \text{ in (6.3)}\ .$$

Note that Q_i differs from the unweighted pseudo-value in that the weight $(1-w_i)$ is attached to $n(\hat{\beta}-\hat{\beta}_{(i)})$ and that $(1-w_i)$ is proportional to $|X_{(i)}^T X_{(i)}|$. Hinkley (1977) pointed out that

$$\hat{\beta} = n^{-1} \sum_1^n Q_i \quad . \tag{6.9}$$

The right hand expression of (6.9) is the usual jackknife point estimator in terms of the pseudo-values. (6.9) is also a special case of the general representation (3.15). He then defined the jackknife variance estimator in terms of Q_i

$$v_{H(1)} = \{n(n-k)\}^{-1} \sum_1^n (Q_i-\hat{\beta})(Q_i-\hat{\beta})^T$$

$$= \sum_1^n \frac{(1-w_i)^2}{1-n^{-1}k} (\hat{\beta}_{(i)} - \hat{\beta})(\hat{\beta}_{(i)} - \hat{\beta})^T \tag{6.10}$$

$$= (X^TX)^{-1} \sum_1^n \frac{r_i^2}{1-n^{-1}k} x_i x_i^T (X^TX)^{-1} \tag{6.11}$$

as a direct extension of a similar definition in the location case. From comparing $v_{H(1)}$ (6.11), and $v_{J(1)}$ (5.17) it seems that $v_{H(1)}$ is also robust in the sense of Theorem 5(ii). The comparison is however more favorable to the latter. 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 $\mathrm{Var}(e) = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, $Ev_{H(1)} \neq \mathrm{Var}(\hat{\beta}) = (X^TX)^{-1} \sum_1^n \sigma_i^2 x_i x_i^T (X^TX)^{-1}$ even under the restriction (6.2) of Theorem 5. As in Theorem 5(ii), $v_{H(1)}$ is approximately unbiased under (C1 - C2), i.e. $Ev_{H(1)} = \mathrm{Var}(\hat{\beta})(1 + O(n^{-1}))$. This is because

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

where the first equation follows from Lemma 4, (6.4), and the second equation follows from (C1). The results concerning $v_{H(1)}$ are summarized in Theorem 6.

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

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

(ii) Under $\mathrm{Var}(e) = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$ and (C1 - C2),

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

Unlike $v_{J(1)}$ the exact unbiasedness $Ev_{H(1)} = \mathrm{Var}(\hat{\beta})$ does not hold true even in special cases. In the simulation of Section 10, $v_{H(1)}$ is found to be more biased than other estimators for both equal and unequal variances. 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, Lemma 2 of Appendix) by following Miller's (1974) proof for the balanced jackknife. The strong consistency of $v_{J(1)}$ can be established in a similar manner.

Standard asymptotic justifications of 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 "robustness" of $v_{J(1)}$ and $v_{H(1)}$ (Theorems 5(ii) and 6(ii)) against the heterogeneity of errors, first recognized in Hinkley (1977), is a truly fresh and important property of the jackknife methodology. (Practically speaking this advantage should be put in the context of nonlinear estimation, Sections 7 and 9.)

To further appreciate this point, let us consider the robustness aspect of the usual variance estimator $\widehat{\mathrm{var}} = \hat{\sigma}^2(X^TX)^{-1}$. For $\mathrm{Var}(e) = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$, from (6.5)

$$E\hat{\sigma}^2 = \sum_1^n \frac{1-w_i}{n-k}\,\sigma_i^2 = \bar{\sigma}^2 \quad .$$

Therefore

(6.12) $$E\,\widehat{\mathrm{var}} = \bar{\sigma}^2(X^TX)^{-1}$$

is equal to

$$\mathrm{Var}(\hat{\beta})(1 + O(n^{-1})) = (X^TX)^{-1}\sum_1^n (\sigma_i^2 + O(n^{-1}))x_ix_i^T(X^TX)^{-1}$$

if $\max_i|\sigma_i^2 - \sigma^2| = O(n^{-1})$, or equivalently,

(6.13) $$\max_{1\le i\le n}\sigma_i^2 - \min_{1\le i\le n}\sigma_i^2 = O(n^{-1}) \quad ,$$

since σ^2 is a weighted average of σ_i^2. The condition (6.13) is sufficient for the

robustness of $\widehat{\text{var}}$ in the sense of Theorem 5(ii). However the result is quite uninteresting since (6.13) forces the variances to be nearly equal for large n. A detailed comparison of $v_{J(1)}$, $v_{H(1)}$, $\widehat{\text{var}}$ and other bootstrap variance estimators for the 2-sample problem will be given in Section 8.

To close this section, we shall make two other remarks.

1. Tukey's reformulation of Quenouille's jackknife in terms of the pseudo-values works well for the i.i.d. case. Its extension to the non-i.i.d. situations 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 some properly defined pseudo-values, both of which reduce to the usual jackknife point estimator in the unstratified case. It was found recently (Rao and Wu, 1983a) that neither estimator reduces bias as is typically claimed for the jackknife. On the other hand a truly bias-reducing jackknife estimator was not motivated by the pseudo-values.

2. We suppose that the purpose of jackknife variance estimation is to aid the point estimator $\hat{\beta}$ in making inference about β. The variance estimators are then required to be nonnegative and almost unbiased. However in situations like the determination of sample size, the variance itself is the parameter of primary interest and other risk criteria like the mean square error (MSE) will be more appropriate. In this context, a nonnegative biased estimator (J. N. K. Rao, 1973) and MINQUE (C. R. Rao, 1970) (which may take negative values) have been proposed. Horn, Horn and Duncan (1975) proposed $(1-w_i)^{-1}r_i^2$, which appears in $v_{J(1)}$, (5.17), 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). It is difficult to extend this comparison to estimation of the variance-covariance matrix.

7. *Jackknifing for nonlinear parameters*

So far we have confined our study of the jackknife to the linear parameters as an important test case. Their utility as a practical tool is more appreciated in the complex

situations where no exact results are available. In this section we first consider a simple nonlinear situation. The parameter of interest $\theta = g(\beta)$ is a nonlinear function of the linear parameter β in the model (3.5). The natural estimator of θ is $\hat{\theta} = g(\hat{\beta})$. In this and the next section we will consider variance and interval estimation for $\hat{\theta}$. Bias reduction of $\hat{\theta}$ will be considered in Section 9. Extensions to nonlinear regression models will be briefly outlined later.

A natural extension of the general weighted jackknife variance estimator $v_{J,r}$, (5.1), for the nonlinear estimator $\hat{\theta} = g(\hat{\beta})$ is

$$\hat{v}_{J,r}(\hat{\theta}) = \frac{r-k+1}{n-r} \frac{\sum_{s \in S_r} |x_s^T x_s| (\hat{\theta}_s - \hat{\theta})(\hat{\theta}_s - \hat{\theta})^T}{\sum_{s \in S_r} |x_s^T x_s|} , \tag{7.1}$$

where $\hat{\theta}_s = g(\hat{\beta}_s)$ and $\hat{\beta}_s$ is assumed to exist for any $s \in S_r$. Another extension of $v_{J,r}$ is

$$\tilde{v}_{J,r}(\hat{\theta}) = \frac{\sum_{s \in S_r} |x_s^T x_s| (\tilde{\theta}_s - \hat{\theta})(\tilde{\theta}_s - \hat{\theta})^T}{\sum_{s \in S_r} |x_s^T x_s|} , \tag{7.2}$$

$$\tilde{\theta}_s = g(\tilde{\beta}_s), \ \tilde{\beta}_s = \hat{\beta} + \sqrt{\frac{r-k+1}{n-r}} (\hat{\beta}_s - \hat{\beta}) . \tag{7.3}$$

Both can be implemented by Monte Carlo approximation as in the linear case. In (7.2) the scale factor $\sqrt{r-k+1}/\sqrt{n-r}$ is applied <u>internally</u> to $\hat{\beta}_s - \hat{\beta}$, while in (7.1) it is applied <u>externally</u> to $\hat{\theta}_s - \hat{\theta}$ after the transformation g. Under reasonable smoothness conditions on g, both $\hat{v}_{J,r}(\hat{\theta})$ and $\tilde{v}_{J,r}(\hat{\theta})$ will be close to the *linearization (or δ-method)* variance estimator

$$v_{lin} = g'(\hat{\beta})^T v_{J,r} g'(\hat{\beta}) , \tag{7.4}$$

where $g'(\hat{\beta})$ is the derivative vector of g evaluated at $\hat{\beta}$. For variance estimation there is perhaps little difference in choosing between $\hat{v}_{J,r}(\hat{\theta})$ and $\tilde{v}_{J,r}(\hat{\theta})$. The internal scaling (7.3) turns out to be instrumental in the following construction of the <u>jackknife distribution</u> based on repeated sampling of subsets of size r:

(i) draw subsets $s_1, \ldots, s_J$ randomly without replacement from S_r,

(ii) construct a weighted empirical distribution function $\hat{CDFJ}(t)$ based on $g(\tilde{\beta}_{s_i})$, $\tilde{\beta}_{s_i}$ defined in (7.3), $i = 1(1)J$, with weight proportional to $|x_{s_i}^T x_{s_i}|$.

Similar to Efron's (1982) bootstrap percentile method is the jackknife percentile method consisting of taking

$$[\hat{CDFJ}^{-1}(\alpha) \ , \ \hat{CDFJ}^{-1}(1-\alpha)] \tag{7.5}$$

as an appropriate $1 - 2\alpha$ central confidence interval for θ. Since $\hat{CDFJ}(t)$ is a discrete function, (7.5) 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 the smoothed percentile, bias-corrected percentile and bootstrap t as modifications of the bootstrap percentile method. The same idea can be applied to the jackknife percentile method in a straightforward manner. It is more natural to apply the internal scaling (7.3) since $\hat{\beta}$ is the center of the weighted distribution of $\hat{\beta}_s$ due to the representation result, Theorem 2, while $\hat{\theta}$ may be shifted from the center of the weighted distribution of $\hat{\theta}_s$ due to the nonlinear distortion g. For this reason we think (7.2) may be more natural than (7.1). The issue of internal or external scaling also arises in the context of bootstrap inference from stratified samples, where it is found that a standard bootstrap method involving a single external scale adjustment gives rise to incorrect variance estimate (Efron, 1982; Bickel and Freedman, 1984), since the corresponding internal scales vary from stratum to stratum. This observation has led Rao and Wu (1983b) to construct a valid bootstrap method by applying the internal scale factor within every stratum before applying the transformation g.

In a more complex situation like the nonlinear regression model

$$y_i = f_i(\beta) + e_i \ , \tag{7.6}$$

where f_i is a nonlinear smooth function of β and the error e_i satisfies the assumptions in (3.5), the jackknife variance estimator $v_{J,r}$, (5.1), has a natural extension, namely, to replace $x_s^T x_s$ by $\sum f_i'(\hat{\beta}) f_i'(\hat{\beta})^T$ with i summing over s and to interpret $\hat{\beta}$ and $\hat{\beta}_s$ as the nonlinear least squares estimates based on the full-data and the subset s, $f_i'(\beta)$ 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 computations was proposed by Fox, Hinkley and Larntz (1980). Confidence intervals can be constructed from the jackknife histogram based on $\tilde{\beta}_s$, (7.3), with the weight function discussed above. Their properties and a similar extension to the generalized linear models will be reported later.

The term "jackknife" is commonly identified in the literature with the delete-one jackknife. According to Efron's (1982) simulation results, the delete-one jackknife does not in general perform as well as the bootstrap. We think there are two reasons for this. First, the delete-one point estimate $\hat{\theta}_{(i)} = g(\hat{\beta}_{(i)})$ is too close to $\hat{\theta} = g(\hat{\beta})$ to reflect the true variability of $\hat{\theta} - \theta = g(\hat{\beta}) - g(\beta)$. The linearly adjusted $\sqrt{n-k}\ (\hat{\theta}_{(i)} - \hat{\theta})$, though correct to the first order, does not take account of the nonlinearity that the function g has undergone between $\hat{\beta}$ and β, since $\hat{\beta} - \beta$ is of larger order of magnitude than $\hat{\beta}_{(i)} - \hat{\beta}$. For nonsmooth θ like the median in the location case, the delete-one jackknife variance estimator is not even consistent. The second reason is that the delete-one jack-knife generates exactly n resampled estimates $\hat{\theta}_{(i)}$. Except for very large n, they do not provide enough values for constructing histograms. This is why the delete-one jackknife method is traditionally associated with variance estimation. The resulting symmetric confidence intervals of the form $[\hat{\theta} - t\hat{\sigma}, \hat{\theta} + t\hat{\sigma}]$ have a serious drawback, namely, they can not reflect the possible skewness inherent in the original estimate $\hat{\theta}$ around θ. On the other hand, the histogram-based confidence intervals do reflect the skewness in $\hat{\theta}$. For the bootstrap method, this was rigorously established in Singh (1980) for some estimators. This and the fact that $\beta^* - \hat{\beta}$ and $\hat{\beta} - \beta$ are of the same order of magnitude, where β^* is the bootstrap estimate of β, perhaps explain the general good performance of the bootstrap histo-gram methods over the delete-one jackknife method.

It should be clear by now why we propose the general jackknife method by deleting more than one observation. It generates more pseudo-replicates of $\hat{\theta}$ to allow for the construction of a histogram. For $n = 20$, the delete-two jackknife generates 190 values instead of the meager 20 values given by the delete-one jackknife. Regarding the question of the choice of r, we may choose r to make the scale factor $(r-k+1)/(n-r)$

near one, that is, $r \simeq (n+k-1)/2$. This choice guarantees that $\hat{\theta}_s - \hat{\theta}$ is of the same stochastic order as $\hat{\theta} - \theta$ and makes it unnecessary to perform the internal scale adjustment (7.3). For the location problem $k = 1$, choosing $r = \frac{n}{2}$ reduces to the half-sampling procedure (Efron, 1982, Chapter 8). Another advantage of choosing r near $(n+k-1)/2$ is in variance estimation when the parameter of interest is not a smooth function. The inconsistency of the delete-one jackknife variance estimator for nonsmooth estimates like the median can be avoided. In the context of complex sample surveys, the balanced half-sample method of McCarthy (1966) is found to provide more reliable confidence intervals than the linearization method and the jackknife method in the empirical study of Kish and Frankel (1973).

8. Bootstrap and subset sampling in regression

Can the previous results for the jackknife be extended to other resampling methods? For a given resampling method denoted by $*$, β^* and D^* defined in (3.7), we would like to find a variance estimator of the form

$$v = \lambda E_* w_* (\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T \quad , \tag{8.1}$$

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)

$$E(v|\mathrm{Var}(e) = \sigma^2 I) = \sigma^2 (X^T X)^{-1} \quad . \tag{8.2}$$

The left hand side of (8.2) is equal to

$$\lambda\sigma^2 E_* \{w_* (X^T D^* X)^{-1} X^T D^{*2} X (X^T D^* X)^{-1} - w_* (X^T X)^{-1}\}$$

$$= \lambda\sigma^2 \{E_* [w_* (X^T D^* X)^{-1} X^T D^{*2} X (X^T D^* X)^{-1}] - (X^T X)^{-1}\} \quad . \tag{8.3}$$

The first term inside the curly bracket of (8.3) seems intractable except for

$$D^{*2} = D^* = \mathrm{diag}(P_1^*, \ldots, P_n^*) \quad , \tag{8.4}$$

which is equivalent to $P_i^* = 0$ or 1 for all i. This prompts us to consider procedures satisfying Assumption (B) and condition (8.4), whose defining probabilities

$$\mathrm{Prob}_*(P_{i_1}^* = \cdots = P_{i_r}^* = 1, \text{ remaining } P_j^* = 0) = \frac{c_r}{\binom{n}{r}} \tag{8.5}$$

are independent of the subset $(i_1, \ldots, i_r)$, where c_r is the probability of resampling a

subset of size r, $c_k+\ldots+c_n=1$. When $c_r = 1$, (8.5) reduces to (3.13), the jackknife with fixed subset size r. Therefore we call any procedure satisfying (B) and (8.4) a subset sampling procedure. Some may prefer to call it a variable jackknife. Back to (8.3), its first term under (8.4) is

$$\frac{E_*|X^T D^* X|(X^T D^* X)^{-1}}{E_*|X^T D^* X|} = \frac{E_*\,\mathrm{adj}(X^T D^* X)}{E_*|X^T D^* X|} ,$$

whose (i,j) element is

$$(-1)^{i+j}\frac{E_*|X^{(j)^T} D^* X^{(i)}|}{E_*|X^T D^* X|}$$

$$= (-1)^{i+j}\frac{\sum_{s\in S_{k-1}}|X_s^{(j)}|E_*|D_s^*|\;|X_s^{(i)}|}{\sum_{s\in S_k}|X_s|^2 E_*|D_s^*|} \tag{8.6}$$

$$= \frac{\alpha_{k-1}}{\alpha_k}\frac{(-1)^{i+j}|X^{(j)^T}X^{(i)}|}{|X^TX|} = \frac{\alpha_{k-1}}{\alpha_k}\ (i,j)\ \text{element of}\ (X^TX)^{-1} , \tag{8.7}$$

where the expansion in (8.6) is justified by Lemma 1(i) and

$$\alpha_i = \mathrm{Prob}_*(P_1^* = P_2^* = \cdots = P_i^* = 1) = \sum_{r=k}^{n} c_r\binom{n}{r}^{-1}\binom{n-i}{r-i} . \tag{8.8}$$

From (8.1), (8.3) and (8.7), for a subset sampling procedure $*$, we have found that the variance estimator

$$\left(\frac{\alpha_{k-1}}{\alpha_k} - 1\right)^{-1}\frac{E_*|X^T D^* X|(\beta^*-\hat\beta)(\beta^*-\hat\beta)^T}{E_*|X^T D^* X|} \tag{8.9}$$

satisfies the unbiasedness requirement (8.2). For the special case of jackknifing with fixed subset size r, $\frac{\alpha_{k-1}}{\alpha_k} - 1 = \frac{n-r}{r-k+1}$ and (8.9) reduces to the jackknife variance estimator $v_{J,r}$, (5.1).

Note that the scale factor in (8.9)

$$\left(\frac{\alpha_{k-1}}{\alpha_k} - 1\right)^{-1} = \frac{\alpha_k/\alpha_{k-1}}{1-\alpha_k/\alpha_{k-1}} = \frac{\mathrm{Prob}_*(P_k^* = 1 | P_1^* = \cdots = P_{k-1}^* = 1)}{\mathrm{Prob}_*(P_k^* = 0 | P_1^* = \cdots = P_{k-1}^* = 1)} \tag{8.10}$$

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 useful.

Among those resampling procedures that do not satisfy (8.4), i.e. $\text{Prob}_*(P_i^* > 2$ for some $i) > 0$, we single out the bootstrap, (3.12), for further study. Unfortunately it will be shown next via a simple counterexample that no variance estimator (8.1) for the bootstrap can satisfy (8.2) in general. Consider the following regression model:

$$\begin{aligned} y_{i1} &= \beta_1 + e_{i1},\ i = 1,\ldots,n_1 \\ y_{i2} &= \beta_2 + e_{i2},\ i = 1,\ldots,n_2 \end{aligned} \qquad n = n_1 + n_2 \tag{8.11}$$

with uncorrelated errors, $Ee_i = 0$ and $\text{Var}\, e_{i1} = \sigma_1^2$, $\text{Var}\, e_{i2} = \sigma_2^2$. Let $P^* = (P_1^*,\ldots,P_n^*)$ be a resampling vector from the bootstrap method, (3.12), with the bootstrap sample size $m = n$. Rewrite $P^* = (P_{11}^*,\ldots,P_{n_1 1}^*, P_{12}^*,\ldots,P_{n_2 2}^*)$ to correspond to the two samples of (8.11) and define $n_1^* = \Sigma_1^{n_1} P_{i1}^*$, $n_2^* = n - n_1^* = \Sigma_1^{n_2} P_{i2}^*$. Then

$$n_j^* \overset{*}{\sim} B(n, \frac{n_j}{n}),\ j = 1,2 \tag{8.12}$$

is a binomial distribution with parameters n and n_j/n, and the conditional distribution

$$((P_{ij}^*)_{i=1}^{n_j} | n_j^*) \overset{*}{\sim} \text{Mult}_{n_j^*}(n_j, \frac{1}{n_j}\underset{\sim}{1}),\ j = 1,2, \tag{8.13}$$

is a multinomial distribution with n_j^* independent draws on n_j categories each having probability $1/n_j$. Then $\beta_j^* = n_j^{*-1} \sum_{i=1}^{n_j} P_{ij}^* y_{ij}$, $j = 1,2$, $|X^T D^* X| = n_1^* n_2^*$ and

$$|X^T D^* X|(\beta_1^* - \hat{\beta}_1)^2 = \frac{n_2^*}{n_1^*} \{\sum_1^{n_1} P_{i1}^*(y_{i1} - \bar{y}_1)\}^2 \tag{8.14}$$

where $\bar{y}_1 = n_1^{-1} \Sigma_1^{n_1} y_{i1}$. From (8.13),

$$\begin{aligned} E_*(|X^T D^* X|(\beta_1^* - \hat{\beta}_1)^2 | n_1^*) &= \frac{n_2^*}{n_1^*} \text{Var}_*(\sum_1^{n_1} P_{i1}^*(y_{i1} - \bar{y}_1) | n_1^*) \\ &= \frac{n_2^*}{n_1^*} \{\sum_1^{n_1} n_1^* \frac{1}{n_1}(1 - \frac{1}{n_1})(y_{i1} - \bar{y}_1)^2 - \sum_{i \neq j}^{n_1} n_1^* \frac{1}{n_1^2}(y_{i1} - \bar{y}_1)(y_{j1} - \bar{y}_1)\} \\ &= \frac{n_2^*}{n_1^*} \frac{n_1^*}{n_1} \sum_1^{n_1} (y_{i1} - \bar{y}_1)^2 = \frac{n_2^*}{n_1} SS_1 , \end{aligned} \tag{8.15}$$

where

$$SS_j = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \quad .$$

From (8.14) - (8.15),

$$E_*(|X^T D^* X|(\beta_1^* - \hat{\beta}_1)^2) = E_*(\frac{n_2^*}{n_1} SS_1) = \frac{n_2}{n_1} SS_1 \quad ,$$

Similarly,

$$E_*(|X^T D^* X|(\beta_2^* - \hat{\beta}_2)^2) = \frac{n_1}{n_2} SS_2 \quad .$$

The cross term

$$E_*(|X^T D^* X|(\beta_1^* - \hat{\beta}_1)(\beta_2^* - \hat{\beta}_2)) = 0$$

since its conditional expectation given n_1^*, n_2^* is easily shown to be zero.

Therefore we have

$$E_*|X^T D^* X|(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T = \text{diag}(\frac{n_2}{n_1} SS_1, \frac{n_1}{n_2} SS_2) \quad .$$

Its expectation under (8.11) is

$$\text{diag}(\frac{n_2(n_1-1)}{n_1} \sigma_1^2, \frac{n_1(n_2-1)}{n_2} \sigma_2^2) \quad ,$$

which is not proportional to the variance of $(\hat{\beta}_1, \hat{\beta}_2)$,

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

unless $n_1 = n_2$. Therefore, no matter how λ is chosen in (8.1), (8.2) cannot be satisfied. In fact, its bias does not go to zero as $n \to \infty$ unless $n_1/n_2 \to 1$.

From similar computations, it can be shown that the unweighted bootstrap variance estimator

$$E_*(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T = \text{diag}(E_*(\frac{1}{n_1^*}) \frac{SS_1}{n_1}, E_*(\frac{1}{n_2^*}) \frac{SS_2}{n_2}) \quad , \tag{8.16}$$

which is well-defined if n_1^* and $n_2^* > 1$. For small or moderate n_i, $E_*(n_i^{*-1} | n_i^* > 1)$ is not close to $(n_i - 1)^{-1}$ and the unweighted estimator (8.16) is biased. If n_1 and n_2 are both large, $E_*(n_i^{*-1} | n_i^* > 1) \simeq (n_i - 1)^{-1}$ and (8.16) is almost unbiased. It is, however, unclear whether this can be extended to general linear models even when the error variances are equal.

It is not surprising that the unweighted bootstrap does not provide an unbiased variance estimator, since, as in the case of the unweighted jackknife, the LSE's β^* based

on the bootstrap resamples are not exchangeable. What is more disappointing is the failure of the weighted bootstrap. One would expect it to perform well since the same weight function was used in jackknife resampling with satisfactory results.

If (8.11) is recognized as a two-sample problem rather than a regression problem, unbiased variance estimators can be obtained by bootstrapping and rescaling within each sample. But the main point we have tried to make here is that a result like Theorem 3 cannot be extended to the bootstrap method for the general linear model (3.5). It is however unclear what will happen if the weight w_* in (8.1) is chosen differently from $|X^T D^* X|$.

On the other hand, the jackknife works quite well here. Routine computation gives

$$v_{J(1)} = \text{diag}\left(\frac{SS_1}{n_1(n_1-1)}, \frac{SS_2}{n_2(n_2-1)}\right) ,$$

and

$$Ev_{J(1)} = \text{diag}\left(\frac{\sigma_1^2}{n_1}, \frac{\sigma_2^2}{n_2}\right) .$$

The latter also follows from Theorem 5(i) since the model (8.11) satisfies (6.2). It can be shown that the delete-two jackknife

$$v_{J,n-2} = v_{J(1)}$$

and that

$$v_{H(1)} = \frac{n}{n-2} \text{diag}\left(\frac{SS_1}{n_1^2}, \frac{SS_2}{n_2^2}\right) ,$$

which is biased but becomes approximately unbiased as n_1 and n_2 become large. The usual variance estimator

$$\widehat{\text{var}} = \frac{SS_1 + SS_2}{n-2} \text{diag}\left(\frac{1}{n_1}, \frac{1}{n_2}\right)$$

is unbiased for $\sigma_1 = \sigma_2$, approximately unbiased for σ_1 near σ_2, and biased otherwise.

To obtain valid bootstrap variance estimators, we can draw a simple random sample $\{e_i^*\}_1^n$ with replacement from the "population" $\{r_i/\sqrt{1-k/n}\}_1^n$, $r_i = y_i - x_i^T\hat{\beta}$ is the i^{th} residual. Define the bootstrap data $y_i^* = x_i^T\hat{\beta} + e_i^*$, $i = 1(1)n$, by treating $\hat{\beta}$ as the true parameter with the above "population" of errors, and the bootstrap LSE is

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

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

(8.18) $$v_b = E_*(\theta^*-\hat{\theta})(\theta^*-\hat{\theta})^T \quad , \quad \theta^* = g(\beta^*), \ \beta^* \text{ in (8.17)} \quad .$$

Note that $\theta^*-\hat{\theta}$ is unweighted. When $\theta = \beta$, it is easy to see (Efron, 1979) that

(8.19) $$v_b = \widehat{var} = \hat{\sigma}^2(X^TX)^{-1}, \ \hat{\sigma}^2 = \frac{1}{n-k}\sum_1^n r_i^2 \quad .$$

Therefore, for homoscedastic errors $Var(e) = \sigma^2 I$, v_b is a valid variance estimator. For constructing confidence intervals for θ, note that each y^* vector is associated with the same X matrix. The unweighted percentile method of Efron (1982) is as follows. Repeat the above procedure for B times. Define $\widehat{CDFB}(t)$ to be the unweighted empirical distribution function based on the B bootstrap estimates θ^{*b}, $b = 1(1)B$. The <u>bootstrap percentile</u> method consists of taking

(8.20) $$[\widehat{CDFB}^{-1}(\alpha), \ \widehat{CDFB}^{-1}(1-\alpha)]$$

as an approximate $1 - 2\alpha$ central confidence interval for θ. The interval (8.20) is computed with a continuity correction.

But for heteroscedastic errors $Var(e) = diag(\sigma_1^2,\ldots,\sigma_n^2)$, $v_b = \widehat{var}$ does poorly as demonstrated in Section 6 for the linear parameters. This should be quite clear from the nature of the procedure. The assumption underlying the drawing of i.i.d. samples from $\{r_i/\sqrt{1-k/n}\}$ is that the residuals r_i are viewed as exchangeable. The first bootstrap residual e_1^* may come from the tenth residual r_{10}, and so forth. The heterogeneity among r_i is lost in this mixing process. On the other hand the delete-one jackknife, by retaining the identity of the residuals, reflects the possible heterogeneity of r_i and of the error variance σ_i^2.

Recognizing the model-dependent nature of the bootstrap residual method, Efron and Gong (1983, p. 43) seemed to favor the unweighted bootstrap method since it "takes less advantage of the special structure of the regression problem." However, their next statement that "the (unweighted bootstrap) method gives a trustworthy estimate of $\hat{\beta}$'s variability <u>even if the regression model is not correct</u>" cannot be substantiated as one can easily infer from our counterexample.

Since the basic principle of the bootstrap is to simulate samples that resemble the unknown population, we must point out that the "population" $\{r_i/\sqrt{1-k/n}\}$ does not resemble the true population of errors $\{e_i\}$ in that r_i are mildly correlated with nonconstant variances $(1-w_i)/(1-k/n)$ if $\mathrm{Var}(e) = \sigma^2 I$. One possibility is to replace the n values $r_i/\sqrt{1-k/n}$ by $n-k$ uncorrelated residuals $\hat{e}_i$ with variance σ^2, e.g. the BLUS residuals (Theil, 1971). If the errors e_i are assumed to be normal, $\hat{e}_i$ are also normal. One may first apply a random orthogonal transformation T to $\{\hat{e}_i\}_1^{n-k}$ to obtain $\{T\hat{e}_i\}_1^{n-k}$, and then draw i.i.d. sample $\{e_i^*\}_{i=1}^n$ from $\{T\hat{e}_i\}_1^{n-k}$. A major problem is that the i.i.d. property of $\{\hat{e}_i\}$ depends critically on the homoscedasticity and normality assumption.

9. Bias reduction

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

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

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

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

where tr is the trace of a matrix. Since the reduction of bias of $\hat{\theta}$ amounts to estimating $B(\hat{\theta})$ unbiasedly 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 $g''(\beta)$. First we consider the jackknife resampling. Let $\tilde{\theta}_s = g(\tilde{\beta}_s)$ be defined in (7.3). Taylor expansion gives

$$\tilde{\theta}_s = \hat{\theta} + g'(\hat{\beta})^T(\tilde{\beta}_s-\hat{\beta}) + \frac{1}{2}(\tilde{\beta}_s-\hat{\beta})^T g''(\hat{\beta})(\tilde{\beta}_s-\hat{\beta}) + \eta_s \quad , \tag{9.3}$$

where η_s is the remainder term. If the weights w_s are uniformly bounded away from 0 and 1, the discussion around (5.6) and (5.6)' implies that $\tilde{\beta}_s - \hat{\beta}$ and $\hat{\beta} - \beta$ have the same stochastic order. If we further assume

$$\hat{\beta} - \beta = O_p(n^{-0.5}) \tag{9.4}$$

and the continuous third differentiability of g around β, we have

$$\eta_s = O_p(n^{-1.5}) \quad . \tag{9.5}$$

For the jackknife with subset size r, we propose to consider the following estimator of $B(\hat{\theta})$,

$$\tilde{B}_{J,r} = \sum_{s \in S_r} w_s(\tilde{\theta}_s - \hat{\theta}), \quad w_s = \frac{|X_s^T X_s|}{\sum' |X_s^T X_s|} \quad . \tag{9.6}$$

where $\sum'$ is summation over s in S_r. From (9.3) and Theorem 2,

$$\begin{aligned} \tilde{B}_{J,r} &= \frac{1}{2} tr(g''(\hat{\beta}) \sum' w_s(\tilde{\beta}_s - \hat{\beta})(\tilde{\beta}_s - \hat{\beta})^T) + \sum' w_s \eta_s \\ &= \frac{1}{2} tr(g''(\hat{\beta}) v_{J,r}) + \sum' w_s \eta_s \quad . \end{aligned} \tag{9.7}$$

Since $g''(\hat{\beta}) = g''(\beta) + O_p(n^{-0.5})$ and $E(v_{J,r}) = Var(\hat{\beta})$ under (9.4) and $Var(\underset{\sim}{e}) = \sigma^2 I$, the first term of (9.7) estimates $B(\hat{\theta})$ *unbiasedly up to order* n^{-2}; the second term of (9.7) is of order $O_p(n^{-1.5})$ under the same assumption that led to $\eta_s = O_p(n^{-1.5})$ since w_s are assumed to be bounded away from 0 and 1. This leads us to the following theorem.

Theorem 7. Assume (C3), the continuous third differentiability of g around β and that the weights w_s are uniformly bounded away from 0 and 1. For homoscedastic errors $Var(\underset{\sim}{e}) = \sigma^2 I$,

$$E(\tilde{B}_{J,r}) = B(\hat{\theta}) + O(n^{-2}) \quad .$$

We assume (C3) in Theorem 7 since it implies (9.4) under $Var(\underset{\sim}{e}) = \sigma^2 I$.

It is clear from the arguments leading to Theorem 7 that, for any general $Var(\underset{\sim}{e}) = V$, as long as (9.4) and

$$Ev_{J,r} = Var(\hat{\beta}) + O(n^{-2}) \quad \text{under} \quad Var(\underset{\sim}{e}) = V \tag{9.8}$$

are satisfied, the conclusion of Theorem 7 holds true. One such candidate is the delete-one jackknife variance estimator $v_{J(1)}$. For $V = diag(\sigma_1^2, \ldots, \sigma_n^2)$, according to Theorem 5(ii),

$$Ev_{J(1)} = Var(\hat{\beta})(1 + O(n^{-1})) \quad \text{under} \quad (C1 - C2)$$

$$= Var(\hat{\beta}) + O(n^{-2}) \quad \text{under} \quad (C1 - C3) \ ,$$

where the second equality follows from $Var(\hat{\beta}) = O(n^{-1})$ under (C3). Since (C1) is implied by (C3 - C4), we have the following corollary.

Corollary 3. Under the conditions of Theorem 7, *(C2) and (C4), for heteroscedastic errors* $Var(\underset{\sim}{e}) = diag(\sigma_1^2,\ldots,\sigma_n^2)$,

$$E\,\tilde{B}_{J(1)} = B(\hat{\theta}) + O(n^{-2}) \ , \tag{9.9}$$

where, $w_i = x_i^T(X^TX)^{-1}x_i$, $\tilde{\beta}_{(i)} = \hat{\beta} + \sqrt{n-k}\ (\hat{\beta}_{(i)}-\hat{\beta})$ and

$$\tilde{B}_{J(1)} = \tilde{B}_{J,n-1} = \sum_1^n \frac{(1-w_i)}{n-k}\{g(\tilde{\beta}_{(i)}) - g(\hat{\beta})\} \ . \tag{9.10}$$

Since $v_{H(1)}$ also satisfies (9.8) *for* $V = diag(\sigma_1^2,\ldots,\sigma_n^2)$, one would expect a result similar to Corollary 3. Hinkley (1977) considered the following estimator of $B(\hat{\theta})$,

$$\hat{B}_{J(1)} = \sum_1^n (1-w_i)\{g(\hat{\beta}_{(i)}) - g(\hat{\beta})\} \tag{9.11}$$

and demonstrated its unbiasedness (without spelling out proper regularity conditions) for the homoscedastic case. A stronger result will be proved next. Consider the expansion

$$g(\hat{\beta}_{(i)}) = g(\hat{\beta}) + g'(\hat{\beta})^T(\hat{\beta}_{(i)}-\hat{\beta}) + \frac{1}{2}(\hat{\beta}_{(i)}-\hat{\beta})^T g''(\hat{\beta})(\hat{\beta}_{(i)}-\hat{\beta}) + \eta_{(i)} \ , \tag{9.12}$$

where the remainder term $\eta_{(i)} = o_p(n^{-3})$ since g''' is bounded in a neighborhood of β and $\hat{\beta}_{(i)} - \hat{\beta} = o_p(n^{-1})$. (9.12) gives

$$\hat{B}_{J(1)} = \frac{1}{2}\,tr\{g''(\hat{\beta})v_{J(1)}\} + \sum_1^n (1-w_i)\eta_{(i)} \ , \tag{9.13}$$

which reveals the surprising connection of the bias estimator $\hat{B}_{J(1)}$ with $v_{J(1)}$, instead of with its twin $v_{H(1)}$. Both $\hat{B}_{J(1)}$ and $v_{H(1)}$ were motivated by pseudo-values (Hinkley, 1977). (The connection seems to suggest that $v_{J(1)}$ is a more natural estimator than $v_{H(1)}$.) In Lemma 5, to be stated later, we will prove $\Sigma_1^n(1-w_i)\eta_{(i)} = o_p(n^{-1.5})$, which in conjunction with (9.13) and Theorem 6(ii) yields the following result.

Corollary 4. Assume the continuous third differentiability of g near β and (C2 - C4). For heteroscedastic errors $\mathrm{Var}(\underset{\sim}{e}) = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$,

$$E\hat{B}_{J(1)} = B(\hat{\theta}) + O(n^{-2}) \quad .$$

The proof parallels that of Corollary 3, the key steps being associated with (9.13), Theorem 6(ii) and Lemma 5. The main difference between Corollaries 3 and 4 is that some of the regularity conditions required in Corollary 3 are automatically satisfied in Corollary 4. The reason is that $\hat{\beta}_{(i)}$ is much closer to $\hat{\beta}$ than $\tilde{\beta}_{(i)}$ to $\hat{\beta}$. This brings home the problem of choosing 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 $B(\hat{\theta})$, we prefer $\tilde{B}_{J(1)}$ since it uses $\tilde{\beta}_{(i)}$ 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 $\hat{\beta} - \beta$. See the relevant discussion in Section 7. This difference will probably not be detectable quantitatively unless g is markedly nonlinear. On the other hand, for very smooth g, heuristic (in contrast to rigorous) computations show that the error term $\Sigma(1-w_i)\eta_{(i)}$ in $\hat{B}_{J(1)}$, (9.13), is of stochastic order n^{-2} while the error term $\Sigma\, w_s\eta_s$ in $\tilde{B}_{J,r}$ (including $\tilde{B}_{J(1)}$), (9.7), is of stochastic order $n^{-1.5}$, suggesting that $\hat{B}_{J(1)}$ is a better approximation to $B(\hat{\theta})$.

Lemma 5. Under the conditions of Corollary 4,

$$\Sigma(1-w_i)\eta_{(i)} = o_p(n^{-1.5}) \quad ,$$

where $\eta_{(i)}$ is defined in (9.12).

Proof: Denote $\hat{\beta}_{(i)} - \hat{\beta} = (d_{ij})_{j=1}^{k}$. Since g''' is bounded near β,

$$|\Sigma(1-w_i)\eta_{(i)}| \le M \sum_{j,\ell,m=1}^{k} \sum_{i=1}^{n} (1-w_i)|d_{ij}d_{i\ell}d_{im}| \tag{9.14}$$

where $M < \infty$ independent of n. Under (C2 - C4), it is proved in Lemma 6 that $\max_i|\hat{\beta}_{(i)} - \hat{\beta}| = o_p(n^{-0.5})$. Continuing (9.14), we have

$$|\Sigma(1-w_i)\eta_{(i)}| \le o_p(n^{-0.5}) \sum_{\ell,m=1}^{k} \sum_{i=1}^{n} (1-w_i)|d_{i\ell}d_{im}| \quad . \tag{9.15}$$

For $\ell = m$,

$$\sum_{1}^{n} (1-w_i)|d_{i\ell}^2| = (\ell,\ell) \text{ element of } v_{J(1)} = o_p(n^{-1}) \quad , \tag{9.16}$$

for $\ell \neq m$,

$$\sum_{1}^{n} (1-w_i)|d_{i\ell}d_{im}| = o_p(n^{-1}) \tag{9.17}$$

follows from (9.16) and the Cauchy-Schwarz inequality. Combining (9.15) - (9.17), we have the desired result. □

Note that $o_p(n^{-1.5})$ is only an upper bound of the order of $\Sigma(1-w_i)\eta_{(i)}$ since $\eta_{(i)} = o_p(n^{-3})$ and the sum of n $o_p(n^{-3})$ terms is likely to be of $o_p(n^{-2})$.

Lemma 6. Under (C2 - C4), and $\mathrm{Var}(\underset{\sim}{e}) = \mathrm{diag}(\sigma_1^2,\ldots,\sigma_n^2)$

$$\max_i(\hat{\beta}_{(i)} - \hat{\beta}) = o_p(n^{-0.5}) \quad .$$

For $\mathrm{Var}(\underset{\sim}{e}) = \sigma^2 I$, this follows from the proof of Lemma 3.3 of Miller (1974). For unequal σ_i^2, the only change involves using (6.4) of our Lemma 4. Note that (C2 - C4) implies (C1 - C3) as shown after Theorem 5. (C1 - C2) guarantees (6.4) and (C3) guarantees $(X^TX)^{-1} = O(n^{-1})$.

For a general jackknife with subset size r, $\hat{B}_{J(1)}$ can be extended to

$$\hat{B}_{J,r} = \frac{r-k+1}{n-r} \sum_{s \in S_r} w_s(g(\hat{\beta}_s) - g(\hat{\beta})), \quad w_s \text{ in (9.6)} \quad . \tag{9.18}$$

The difference between $\hat{B}_{J,r}$ and $\tilde{B}_{J,r}$ is analogous to that between $\hat{v}_{J,r}(\hat{\theta})$, (7.1), and $\tilde{v}_{J,r}(\hat{\theta})$, (7.2). The former applies the scale adjustment (r-k+1)/(n-r) externally and the latter internally. As $\tilde{B}_{J,r}$ does in Theorem 7, $\hat{B}_{J,r}$ also estimates the leading term of $B(\hat{\theta})$ unbiasedly.

Let us now turn our attention to the bootstrap sampling. Since the unbiased estimation for $B(\hat{\theta})$ hinges on the unbiased estimation for $\mathrm{Var}(\hat{\beta})$, from the study of Section 8, we need only consider the last bootstrap method, (8.16) - (8.18), considered there. Let $\beta^* = (X^TX)^{-1}X^Ty^*$ be the bootstrap LSE defined in (8.16), where * denotes the bootstrap (or i.i.d) sampling from the rescaled residuals. From the unbiasedness of the LSE,

$$E_*\beta^* = \hat{\beta} \quad , \tag{9.19}$$

and from (8.19),

(9.20) $$E_*(\beta^* - \hat{\beta})(\beta^* - \hat{\beta})^T = \hat{\sigma}^2(X^TX)^{-1}$$

is unbiased for $\mathrm{Var}(\hat{\beta})$ under $\mathrm{Var}(\underset{\sim}{e}) = \sigma^2 I$. By repeating the steps (9.3) - (9.7) and using (9.19) and (9.20), the proposed bootstrap estimator of bias

(9.21) $$\hat{B}_{boot} = E_*\theta^* - \hat{\theta},\ \theta^* = g(\beta^*)$$

is equal to

(9.22) $$\frac{1}{2}\hat{\sigma}^2 \mathrm{tr}(g''(\hat{\beta})(X^TX)^{-1}) + E_*\eta_* \ ,$$

where η_* is the remainder term of the expansion

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

From (9.21) - (9.22), we have

Theorem 8. Assume (C3), the continuous third differentiability of g near β and

(9.23) $$E_*\eta_* = O_p(n^{-1.5}) \ .$$

For homoscedastic errors $\mathrm{Var}(\underset{\sim}{e}) = \sigma^2 I$,

$$E\hat{B}_{boot} = B(\hat{\theta}) + O(n^{-2}) \ .$$

This unbiasedness result cannot be extended to the heteroscedastic case because of (9.20). The condition (9.23) is a reasonable one since $\eta_* = O_p(n^{-1.5})$ follows from $\beta^* - \hat{\beta} = O_p(n^{-0.5})$, which is a consequence of (C3) and the conditional central limit theorem of β^* (Freedman, 1981, Theorem 2.2).

10. Simulation results

In this section we examine the Monte-Carlo behavior of (i) the bias of several estimators of the variance-covariance matrix of the least squares estimator, $\mathrm{Var}(\hat{\beta})$, (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 of the same nonlinear parameter.

Under consideration is the following quadratic regression model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \ , \quad i = 1(1)12$$

$$x_i = 1,\ 1.5,\ 2,\ 2.5,\ 3,\ 3.5,\ 4,\ 5,\ 6,\ 7,\ 8,\ 10 \ .$$

Two variance patterns are considered:

Unequal variances: $e_i = \sqrt{\frac{x_i}{2}}\, N(0,1)$

Equal variances : $e_i = N(0,1)$.

The e_i's are independent. For unequal variances the variance-covariance matrix of the ordinary least squares estimator $\hat{\beta}$ is

$$\mathrm{Var}(\hat{\beta}) = \begin{bmatrix} 1.50 & -0.79 & 0.08 \\ & 0.48 & -0.05 \\ & & 0.01 \end{bmatrix}$$

while the expectation (see (6.12)) of the usual variance estimator $\widehat{\mathrm{var}}$, is

$$\bar{\sigma}^2 (X^T X)^{-1} = \begin{bmatrix} 2.09 & -0.87 & 0.07 \\ & 0.42 & -0.04 \\ & & 0.00 \end{bmatrix}$$

Because of the heterogeneity of errors, the two matrices are quite different. For a variance estimator v, its bias is defined as

$$B(v) = E(v) - \mathrm{Var}(\hat{\beta}) \quad .$$

Four variance estimators are considered: (1) the usual variance estimator $\widehat{\mathrm{var}}$ (5.9), which is identical to the bootstrap variance estimator v_b (8.18), (2) the delete-one jackknife variance estimator $v_{J(1)}$ (5.16), (3) Hinkley's delete-one jackknife variance estimator $v_{H(1)}$ (6.10), (4) the retain-eight jackknife variance estimator $v_{J,8}$ (5.7). The following results 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. The same set of normal random numbers is used throughout the study.

$$B(\widehat{\mathrm{var}}) = \begin{bmatrix} 0.58 & -0.07 & -0.00 \\ & -0.05 & 0.01 \\ & & -0.00 \end{bmatrix}$$

$$B(v_{J(1)}) = \begin{bmatrix} 0.02 & 0.03 & -0.00 \\ & -0.04 & 0.01 \\ & & -0.00 \end{bmatrix}$$

$$B(v_{H(1)}) = \begin{bmatrix} -0.23 & 0.19 & -0.03 \\ & -0.14 & 0.02 \\ & & -0.00 \end{bmatrix}$$

$$B(v_{J,8}) = \begin{bmatrix} 0.09 & 0.01 & -0.01 \\ & -0.04 & 0.01 \\ & & -0.00 \end{bmatrix}$$

In the unequal variance case, $\widehat{var}$ is known to be biased (6.12) - (6.13), $v_{J(1)}$ to be almost unbiased (Theorem 5). Both are confirmed by the simulations. The robustness of $v_{J,8}$ conjectured in Section 6 is also confirmed. The only surprise is the poor performance of $v_{H(1)}$. The claimed robustness (as n becomes large) of $v_{H(1)}$ in Theorem 6 does not hold up here. Its bias is quite nontrivial. This prompted us to examine the bias behavior of $v_{H(1)}$ in the <u>equal variance</u> case since $v_{H(1)}$ is the only one that is not exactly unbiased (Theorem 6(i)). In this case, $\text{Var}(\hat{\beta}) = \bar{\sigma}^2(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$. The bias of $v_{H(1)}$ is not negligible,

$$B(v_{H(1)}) = \begin{bmatrix} -0.13 & 0.07 & -0.01 \\ & -0.04 & 0.00 \\ & & -0.00 \end{bmatrix}$$

and the biases of $\widehat{var}$, $v_{J(1)}$ and $v_{J,8}$ are all very small (none of the entries exceeds 0.0102 in magnitude). Another thing to note is that all the diagonal elements of $v_{H(1)}$ are downward-biased in the simulation. The poor performance of $v_{H(1)}$ in both cases should cause its users some concern at least in the small sample situations.

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

$$\theta = -\frac{\beta_1}{2\beta_2} ,$$

which maximizes the quadratic function $\beta_0 + \beta_1 x + \beta_2 x^2$ over x. Six point estimators are considered: $\hat{\theta}$; $\hat{\theta}_{J(1)} = \hat{\theta} - \hat{B}_{J(1)}$, (9.13); $\tilde{\theta}_{J(1)} = \hat{\theta} - \tilde{B}_{J(1)}$, (9.10); $\hat{\theta}_{J,8} = \hat{\theta} - \hat{B}_{J,8}$, (9.18); $\tilde{\theta}_{J,8} = \hat{\theta} - \tilde{B}_{J,8}$, (9.6); $\hat{\theta}_{boot} = \hat{\theta} - \hat{B}_{boot}$, (9.21). In drawing the bootstrap samples, the uniform random integers are generated according to the IMSL subroutine GGUD. The number of bootstrap samples B is 480, which is comparable to 495, the total number of jackknife subsets of size 8.

Their average biases are given in Table 1 for $\beta_0 = 0$, $\beta_1 = 4$ and several values of β_2. Bias reduction is more difficult to achieve when β_2 gets closer to 0 since θ becomes a more <u>curved</u> 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 $\hat{\theta}_{boot}$ 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 $\tilde{\theta}_{J,8}$ based on internal adjustment of distance do not perform as well. This is consistent with the asymptotic comparison given before Lemma 5. What puzzles us is the unpredictable behavior of the bootstrap estimator $\hat{\theta}_{boot}$ for $\beta_2 = -0.25$. According to Theorem 8, $\hat{\theta}_{boot}$ reduces bias for equal variances but not for unequal variances. What we see in Table 1 is the contrary. It appears that the curvature effect is the dominant factor here.

Table 1. Biases of six estimators of $\hat{\theta}$
(based on 3000 simulation samples)

$\beta_0 = 0$, $\beta_1 = 4$

estimator	Unequal variances β_2				Equal variances β_2	
	-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
$\hat{\theta}_{J(1)}$	-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
$\hat{\theta}_{boot}$	0.16	0.02	0.01	-0.00	-0.12	-0.00

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

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

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

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

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

and t_α is the upper α percentage point of a t-distribution with $n - 3$ (here 9) degrees of freedom, $\hat{\sigma}^2$ is the usual variance estimator (5.9) (by assuming equal variances). Fieller's interval estimate is unbounded in the case of (I) or (II) of (10.1). The method is <u>not</u> exact if the variances are unequal.

Altogether nine methods are compared in our simulation. A description is given below.

symbol	interval estimate	
Fieller	Fieller's interval,	(10.1)
VCJ(1)	Delete-1 jackknife	$\hat{\theta} \pm t_\alpha \sqrt{\tilde{v}_{J,n-1}(\hat{\theta})}$, (7.2)
VHJ(1)	Delete-1 jackknife	$\hat{\theta} \pm t_\alpha \sqrt{\hat{v}_{J,n-1}(\hat{\theta})}$, (7.1)
VCJ8	Retain-8 jackknife	$\hat{\theta} \pm t_\alpha \sqrt{\tilde{v}_{J,8}(\hat{\theta})}$, (7.2)
VHJ8	Retain-8 jackknife	$\hat{\theta} \pm t_\alpha \sqrt{\hat{v}_{J,8}(\hat{\theta})}$, (7.1)
VBOOT	Bootstrap variance	$\hat{\theta} \pm t_\alpha \sqrt{v_b}$, (8.18)
VLIN	Linear approximation	$\hat{\theta} \pm t_\alpha \sqrt{v_{lin}}$, (7.4)
PBOOT	Bootstrap percentile	$[\widehat{CDFB}^{-1}(\alpha), \widehat{CDFB}^{-1}(1-\alpha)]$, (8.20)
PJ8	Jackknife percentile (retain-8)	$[\widehat{CDFJ}^{-1}(\alpha), \widehat{CDFJ}^{-1}(1-\alpha)]$, (7.5)

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

The average coverage probabilities (based on 3000 samples) for these nine methods are given in Table 2 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 which category the corresponding Fieller's

intervals belong to. 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 3. For the rest, the median length over 3000 samples is given inside the parenthesis in Table 2. We do not report the average lengths since they are too much 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 3.73. The three methods perform unstably in highly nonlinear situations.

Table 2. Average coverage probabilities and median lengths for nine interval estimation methods (3000 simulation samples) Nominal level = 0.95, $\beta_0 = 0$, $\beta_1 = 4$

method	Unequal variances β_2				Equal variances β_2	
	-0.25	-0.35	-0.5	-1.0	-0.25	-1.0
Fieller	.858	.866	.968 (.98)	.952 (.92)	.947 (2.48)	.950 (.64)
VCJ(1)	.887	.848	.961 (.91)	.950 (.89)	.904 (2.03)	.935 (.62)
VHJ(1)	.866	.845	.950 (.87)	.947 (.87)	.899 (1.94)	.935 (.62)
VCJ8	.946	.920	.968 (.97)	.953 (.90)	.947 (3.19)	.939 (.63)
VHJ8	.931	.908	.965 (.93)	.953 (.90)	.941 (2.69)	.939 (.63)
VBOOT	.866	.902	.973 (.97)	.955 (.91)	.956 (2.42)	.946 (.64)
VLIN	.865	.891	.969 (.93)	.952 (.90)	.949 (2.18)	.948 (.64)
PBOOT	.829	.814	.940 (.84)	.921 (.79)	.912 (2.05)	.916 (.56)
PJ8	.809	.755	.909 (.78)	.912 (.78)	.831 (1.90)	.900 (.55)

(length of interval estimate inside the parenthesis)

Table 3. Median lengths of nine interval estimates of category (II) and category (III) $\beta_0 = 0$, $\beta_1 = 4$, unequal variances

method	$\beta_2 = -0.25$		$\beta_2 = -0.35$	
	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

*The figure inside the parenthesis is the number of simulation samples belonging to the category

The results 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 converage probabilities smaller. The phenomenon is especially pronounced for unequal variances 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 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 heterogeneity. As the theory indicates, the general 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 heterogeneity 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) which are used as evidence of the superiority of the bootstrap approximation over the classical t approximation. The undercoverage of the bootstrap percentile and the jackknife percentile methods, with the latter being the more serious one, is partly due to the fact that their associated intervals are shorter. But noting from Tables 2 and 3, the linearization variance method (VLIN) has comparable interval length and yet higher coverage probabilities. We think the problem is a more intrinsic one. We speculate that this shortcoming has something to do with the skewness and light-tailedness of the bootstrap and jackknife histograms.

4. Fieller's method *is exact in the equal variance case* even when the parameter is considerably nonlinear, but is quite vulnerable to error variance heterogeneity.

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. In fact 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 3).

6. Internal (curl) or external (hat) adjustment in jackknife variance estimation? In general the curl jackknife gives wider intervals than the hat jackknife. On the other hand the coverage probabilities of the two methods are very comparable. Further research is needed to sort out the relative merits of the two adjustment methods in more general situations.

11. Concluding remarks and further questions

The main ideas and results of this paper can be summarized as follows:

1. The general representation of the full-data least squares estimate as a weighted average of the resample-data least squares estimates for general resampling plans. We expect to see further applications of this representation.
2. The proper weight for each subset least squares estimate is proportional to the determinant of the X^TX matrix of the subset. Since the latter matrix is proportional to the Fisher information matrix of the subset, it immediately suggests an extension of our general jackknife procedure to nonlinear regression models and generalized linear models (McCullagh and Nelder, 1983). For each subset, the corresponding nonlinear least squares estimate or maximum likelihood estimate is computed and the Fisher information matrix of the subset is evaluated at the estimated parameter value. The formulae developed in the paper can be applied in a straightforward manner.
3. The delete-one jackknife variance estimator is robust against error heterogeneity. None of the bootstrap methods under consideration is robust. Bootstrapping the residuals is too model-dependent to be a robust tool.
4. The scope of the jackknife method is broadened with the introduction of the (weighted) jackknife histogram and the interval estimation method based on its percentiles. It is made possible by the more flexible choice of subset size and the weighting factor discussed in 2. Although the two percentile methods do not perform well in the simulation, an effective modification of the jackknife percentile method will probably have to incorporate the above two elements.
5. The problem of bias reduction is intimately related to unbiased estimation of variance. This is especially interesting when the latter is not easy to achieve, e.g. in the heteroscedastic situation.

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

I. We conjecture that Theorem 5 is still true for $v_{J,r}$ with small $d = n-r$, that is, the delete-d jackknife variance estimator is robust against error heterogeneity. Does $v_{J,r}$ enjoy other desirable properties? For example, is $v_{J,n-2}$ (delete-two jackknife) robust against certain forms of error correlation?

II. Does there exist a bootstrap variance estimator that is robust against error heterogeneity? For the bootstrap method to be model-free or model-robust as is sometimes claimed (Efron and Gong, 1983), this is a very basic requirement.

III. The methods based on the bootstrap-histogram and the jackknife-histogram perform disappointingly in the simulation. Refinements of these methods are called for. One obvious defect of the resample histograms is that they have shorter tails than their population counterparts. The handling of skewness may also be improper. The poor performance of the percentile methods raises our doubt about the relevance of the present asymptotic results on the bootstrap. Mathematical results that can explain the empirical behavior are urgently needed.

IV. Is it possible to find a theoretical guide on the choice of subset size for the jackknife method? One interesting possibility may start with the concept of "distance matching" given in Section 7.

V. The scale factor $(r-k+1)/(n-r)$ in the retain-r jackknife method is used for "distance matching". It can be applied either before or after the nonlinear transformation (see (7.1) and (7.2)). It would be interesting to sort out the relative merits of these two scaling methods.

REFERENCES

Beran, R. (1982). Estimated sampling distributions: the bootstrap and competitors. Ann. Statist. 10, 212-225.

Bickel, P. J. (1973). On some analogues to linear combinations of order statistics in the linear model. Ann. Statist. 1, 597-616.

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling. Ann. Statist. 12, to appear.

Bingham, C. (1977). Some identities useful in the analysis of residuals from linear regression. University of Minnesota, School of Statistics, T.R. No. 300.

Cook, R. D. and Weisberg, S. (1982). Residuals and Influence in Regression. Chapman and Hall, London.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1980). Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. Multivariate Analysis 5, 35-57.

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. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. The American Statistician 37, 36-48.

Fox, T., Hinkley, D. and Larntz, K. (1980). Jackknifing in nonlinear regression. Technometrics 22, 29-33.

Freedman, D. A. (1981). Bootstrapping regression models. Ann. Statist. 9, 1218-1228.

Hinkley, D. V. (1977). Jackknifing in unbalanced situations. Technometrics 19, 285-292.

Hoerl, A. E. and Kennard, R. W. (1980). A note on least squares estimates. Communication in Statistics B 9, 315-317.

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.

Jaeckel, L. A. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. Ann. Math. Statist. 43, 1449-1458.

Kish, L. and Frankel, M. (1974). Inference from complex samples (with discussion). J. Roy. Statist. Soc. B, 36, 1-37.

Koenker, R. and Bassett, G. Jr. (1978). Regression quantiles. Econometrika 46, 33-50.

McCarthy, P. J. (1969). Pseudo-replication: half-samples. Review of the Int. Statist. Institute 37, 239-264.

McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.

Miller, R. G. (1974). An unbalanced jackknife. Ann. Statist. 2, 880-891.

Noble, B. (1969). Applied Linear Algebra. Prentice Hall, New York.

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. (1983a). Inference from stratified samples: second order analysis of three methods for nonlinear statistics. T.R. No. 7, Lab. for Research in Statistics and Probability, Carleton University, Ottawa.

Rao, J. N. K. and Wu, C. F. J. (1983b). Bootstrap inference with stratified samples. T.R. No. 19, Lab. for Research in Statistics and Probability, Carleton University, Ottawa.

Rousseeuw, P. J. (1984). *Least median of squares regression.* Preprint.

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

Ruppert, D. and Carroll, R. J. (1980). Trimmed least squares estimation in the linear model. J. Amer. Statist. Assoc. 75, 828-838.

Scholz, F. (1978). Weighted median regression estimates. Ann. Statist. 6, 603-609.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. J. Amer. Statist. Assoc. 63, 1379-1389.

Siegel, A. F. (1982). Robust regression using repeated medians. Biometrika 69, 242-244.

Sievers, G. L. (1978). Weighted rank statistics for simple linear regression. J. Amer. Statist. Assoc. 73, 628-631.

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. Sankhya B, 34, 355-356.

Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis. Proc. Kon. Nederl. Akad. Wetensch. A 53, 386-392, 521-525, 1397-1412.

Theil, H. (1971). Econometrics. Wiley, New York.

Titterington, D. M. (1978). Estimation of correlation coefficients by ellipsoidal trimming. Appl. Statist. 27, 227-234.

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

CFJW/jvs

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER # 2675	2. GOVT ACCESSION NO. AD-A141505	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Jackknife and Bootstrap Inference in Regression and a Class of Representations for the LSE		5. TYPE OF REPORT & PERIOD COVERED Summary Report - no specific reporting period 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) C. F. Jeff Wu		8. CONTRACT OR GRANT NUMBER(s) DAAG29-80-C-0041
9. PERFORMING ORGANIZATION NAME AND ADDRESS Mathematics Research Center, University of Wisconsin 610 Walnut Street Madison, Wisconsin 53706		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS Work Unit Number 4 - Statistics & Probability
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office P.O. Box 12211 Research Triangle Park, North Carolina 27709		12. REPORT DATE April 1984 13. NUMBER OF PAGES 51
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

jackknife percentile method, subset sampling, variance estimator, bias reduction, Fieller's method, representation of the least squares estimator, robust regression.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A class of representations for the least squares estimator is presented and their applications sketched. Partly motivated by one such representation, we propose a class of weighted jackknife estimators of variance of the least squares estimator by deleting any fixed number of observations at a time. These estimators are unbiased for homoscedastic errors and a special case, the delete-one jackknife variance estimator, is almost unbiased for heteroscedastic errors. The method is extended in various ways, including the use of the jackknife histogram, for variance and interval estimation with nonlinear parameters.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED (continued)

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ABSTRACT (continued)

Three bootstrap methods are considered. It is shown that none of them has the robustness property enjoyed by the (weighted) delete-one jackknife. Subset sampling with variable subset size is also considered. Several bias-reducing estimators are proposed. They are motivated by the observation that bias-reduction is mathematically equivalent to unbiased estimation of variance. Some simulation results on estimating the ratio of two normal parameters are reported.