ON JACKKNIFING IN UNBALANCED SITUATIONS

BY

DAVID V. HINKLEY

TECHNICAL REPORT NO. 22
JUNE 28, 1976

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 RO1 GM21215-02

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON JACKKNIFING IN UNBALANCED SITUATIONS

BY

DAVID V. HINKLEY

TECHNICAL REPORT NO. 22
JUNE 28, 1976

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 5 RO1 GM21215-02

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
ON JACKKNIFING IN UNBALANCED SITUATIONS

By

David V. Hinkley

Summary

Both the standard jackknife and a weighted jackknife are investigated in the general linear model situation. Properties of bias reduction and standard error estimation are derived, and the weighted jackknife shown to be superior for unbalanced data. There is a preliminary discussion of robust regression fitting using jackknife pseudo-values.

Key words: Jackknife; linear model; regression; residual; robustness

Research supported by Public Health Service Grant 5 RO1 GM21215-02
1. INTRODUCTION

During the two decades since Quenouille and Tukey introduced the jackknife technique for reducing bias and estimating standard errors, an extensive literature has grown up dealing with large-sample properties and empirical validations in common applications; these include estimation of variances, correlations and ratios. With few exceptions, the jackknife has been applied to balanced models. An excellent review is given by Miller (1974a).

Miller (1974b) gives the first published account of jackknifing linear model estimates, and shows that the jackknife produces consistent results in large samples. The present paper examines the small-sample properties of the standard jackknife in the general linear model, and compares it to an alternative weighted jackknife procedure. The general linear model is a test case, the desired objective being a suitable version of the jackknife for use with unbalanced, or non-symmetric, statistics. Properties of the balanced and weighted jackknife procedures are derived for the linear model in Sections 2.1 and 2.2, with simple numerical examples in Section 2.3. Non-linear functions of linear model parameters are discussed and an example given in Section 3.

A second aspect of the jackknife is the use of pseudo-values in obtaining robust estimates; a detailed account in the case of correlation estimation has been given by Hinkley (1976b). In Section 4 we briefly discuss the potential role of the jackknife in obtaining robust regression estimates. The essential idea is to scale individual residuals according to the relative importance of corresponding design points.
Throughout this paper the following model is assumed

\[ Y = A\beta + e \quad (1.1) \]

where

\[ Y^T = (y_1, y_2, \ldots, y_n), \quad \beta^T = (\beta_1, \ldots, \beta_p), \quad e^T = (e_1, \ldots, e_n) \]

and

\[ A = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} x_{11}^T \\ \vdots \\ x_{n1}^T \end{pmatrix}; \quad (1.2) \]

unless otherwise stated, the \( e_j \) are taken to be i.i.d. with mean zero and constant variance \( \sigma^2 \). If the model includes a constant term, then \( x_{11} = 1 \); it is important not to replace \( x_{1k} \) by \( x_{1k} - \bar{x}_k \) \((k \geq 2)\) in \( A \), since this produces incorrect jackknife results for \( \beta_1 \). Some associated statistical measures are

\[ D_0 = A^TA, \quad \hat\beta = (A^TA)^{-1}A^TY, \quad R = (I - A(A^TA)^{-1}A^T)Y = Y - A\hat\beta, \]

and

\[ w_i = x_i^T(A^TA)^{-1}x_i \quad (i=1, \ldots, n). \]

Many relevant calculations have been taken from Miller (1974b) and Cook (1975), which articles suggested some of the ideas in this paper.
2. TWO JACKKNIFE PROCEDURES

2.1 The Balanced Jackknife

The basic components of the standard jackknife procedure are parameter estimates obtained by successively deleting single observations. For the linear regression model (1.1) we take the complete data estimator of \( \beta \) to be the least squares estimator

\[
\hat{\beta} = (A^TA)^{-1}A^TY;
\]

we shall not be concerned directly with estimating \( \sigma^2 \). The corresponding estimator obtained by deleting \((x_i^T, Y_i)\) is easily seen to be

\[
\hat{\beta}_{-i} = \hat{\beta} - \frac{(A^TA)^{-1}x_i(1-x_i^T\hat{\beta})}{1-x_i^T(A^TA)^{-1}x_i} = \hat{\beta} - \frac{D_0^{-1}W_iR_1}{1-w_i} \quad (i=1,\ldots,n), \tag{2.1}
\]

say.

To describe the standard jackknife procedure we first define pseudo-values

\[
P_i = n\hat{\beta}-(n-1)\hat{\beta}_{-i} \quad (i=1,\ldots,n), \tag{2.2}
\]

from which the jackknifed estimator is given by

\[
\hat{\beta} = n^{-1}\sum P_i. \tag{2.3}
\]

Using (2.1) we obtain

\[
P_i = \hat{\beta}+(n-1)D_0^{-1}x_iR_1(1-w_i)^{-1}
\]
and

\[ \tilde{\beta} = \hat{\beta} + (n-1) n^{-1} D_0^{-1} \sum (1-w_i)^{-1} X_i R_i. \quad (2.4) \]

Quite generally the jackknifed estimator removes bias of order \( n^{-1} \). Here, since \( \hat{\beta} \) is unbiased, this property is redundant. Clearly \( \tilde{\beta} \) is unbiased, since \( E(R_i) = 0 \), so the fact that \( \hat{\beta} \) and \( \tilde{\beta} \) are generally different, implies that, together with the Gauss-Markov property of \( \hat{\beta} \)

\[ \text{var}(\tilde{\beta}) > \text{var}(\hat{\beta}); \]

the exceptions to this occur in balanced linear models, where \( w_i \) is constant. A somewhat weaker property of \( \tilde{\beta} \) is general consistency, which holds if \( n^{-1} D_0 \) converges to a positive definite matrix, this implying \( \max w_i \to 0 \); see Miller (1974b).

The exact variance of \( \tilde{\beta} \) is easy to compute. Recall that \( \hat{\beta} \) and \( R^T = (R_1, \ldots, R_n)^T \) are uncorrelated with respective covariance matrices

\[ \text{var}(\hat{\beta}) = \sigma^2 D_0^{-1}, \quad \text{var}(R) = \sigma^2 (I - A D_0^{-1} A^T), \]

so that from (2.4) we have immediately

\[ \text{var}(\tilde{\beta}) = \sigma^2 \left( D_0^{-1} + \frac{n-1}{n} D_0^{-1} (D_2 D_0^{-1} D_1) D_0^{-1} \right), \quad (2.5) \]

where

\[ D_k = \sum (1-w_j)^{-k} X_j X_j^T \quad (k=0,1,2). \]

Supposing \( w_i \) to be of order \( n^{-1} \), we may expand \( (1-w_i)^k \) in series and verify that \( \text{var}(\tilde{\beta}) - \text{var}(\hat{\beta}) \) is of order \( n^{-2} \).
The second, and probably more important, feature of the jackknife procedure is the distribution-free estimate of variance for the parameter estimator. The standard definition is

\[ V = (n(n-1))^{-1} \Sigma (P_1 \hat{\beta}) (P_1 \hat{\beta})^T, \]  

(2.6)

which may be used to estimate both \( \text{var}(\hat{\beta}) \) and \( \text{var}(\hat{\beta}) \); for a simple account of the rationale for this in the balanced case, see Hinkley (1976a). It is not hard to show, under mild conditions including \( nD_1^{-1} \Rightarrow \Sigma > 0 \), that \( nV \Rightarrow n\text{var}(\hat{\beta}) \), i.e., that \( V \) is an accurate large-sample variance estimate; see Miller (1974b). However \( V \) is not unbiased, and straightforward calculations show that

\[ E(V) = \frac{(n-1)}{n} D_0^{-1} (D_2 - D_1 D_0^{-1} D_1) D_0^{-1} \sigma^2, \]  

(2.7)

as compared to \( \sigma^2 D_0^{-1} = \text{var}(\hat{\beta}) \). If we suppose the \( w_i \) are of order \( n^{-1} \) and define

\[ D_2^* = \Sigma w_j \xi_j \xi_j^T \quad (k=1,2), \]

then expansion of (2.7) gives the approximation to order \( n^{-2} \)

\[ E(V) \approx \frac{n^{-1}}{n} (D_0^{-1} + D_0^{-1} D_2^* D_0^{-1} + D_0^{-1} D_2^* D_0^{-1}) \sigma^2. \]  

(2.8)

To summarize these developments, we have found that for an exactly linear estimator (i) the jackknifed estimator \( \hat{\beta} \) is in general different from the original estimator and (ii) the jackknife variance estimate \( V \) is biased in general. These failures are due to the balanced form of the standard jackknife procedure, and occur only in the unbalanced model. Two numerical examples of these results are given in Section 2.3.
2.2 A Weighted Jackknife

The pseudo-values $P_i$ in (2.2) are defined symmetrically with respect to the observations, whereas the model is generally unbalanced. The lack of balance is reflected in the "distances" $w_i$. In the situation where the $x_i$ are sampled from a multivariate normal population, the estimated likelihood of the value $x_i$ is a decreasing function of $w_i$. In addition, $\text{tr}(\text{var}(\hat{\beta}_{-i}))$ is an increasing function of $w_i$. This suggests that in representing the contribution of the $i^{th}$ observation, $\hat{\beta}_{-i}$ has a weight decreasing in $w_i$. A specific choice of weight is indicated by the fact that

$$n(1-w_i)(\hat{\beta}_{-i}) = nD_0^{-1}x_iR_i = \hat{I}(\beta; x_i, Y_i), \quad (2.9)$$

the estimated influence function of $\beta$ at $(x_i, Y_i)$; see Appendix, Lemma 1.

We therefore propose the weighted pseudo-value

$$Q_i = \hat{\beta} + n(1-w_i)(\hat{\beta}_{-i}) = \hat{\beta} + nD_0^{-1}x_iR_i, \quad (2.10)$$

the weighted jackknife estimator

$$\tilde{\beta}_w = n^{-1} \sum Q_i = \hat{\beta} \quad (2.11)$$

and the variance estimate

$$V_w = (n(n-p))^{-1} \sum (Q_i - \tilde{\beta}_w)(Q_i - \tilde{\beta}_w)^T$$

$$= n(n-p)^{-1}D_0^{-1}(\Sigma X_j^T x_i^T D_0^{-1}, \quad (2.12)$$
where in each case the explicit form for the linear model is given. The denominator \( n-p \) used in \( V_w \) reflects the degrees of freedom in the residual vector, and makes \( V_w \) exactly unbiased in the balanced case, when \( w_i = pn^{-1} \).

The reproducing property (2.11) for linear estimates corresponds to that for \( \tilde{\beta} \) in the balanced case. This property may indicate superior performance of \( \tilde{V}_w \) in non-linear situations; see Section 2.4.

The general expectation of \( V_w \) is easily seen to be

\[
E(V_w) = n(n-p)^{-1}(D_0^{-1}D_0^{-1}D_1D_0^{-1})\sigma^2, \tag{2.13}
\]

which is biased in unbalanced cases. We compare this with \( E(V) \) in two examples in Section 2.3.

In one respect the jackknife variance estimate is superior to the usual estimate

\[
\hat{V} = (n-p)^{-1} \sum_j D_j^{-1},
\]

in that \( V_w \) (and \( V \)) are robust against non-homogeneity of error variance. To see this, suppose that in (1.1) \( \text{var}(e) = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) = \Lambda \). Then

\[
\text{var}(\hat{\beta}) = D_0^{-1}A^T \Lambda D_0^{-1}. \tag{2.14}
\]

Since \( E(R_j^2) \neq \sigma_j^2 \) we have

\[
E(\hat{V}) \neq n^{-1} \text{tr}(\Lambda) D_0^{-1}
\]

and
\[ E(V_w) = D_0^{-1} \sum_j \sum_i \sigma_j^2 x_j D_0^{-1} = D_0^{-1} A^{T} A D_0^{-1}; \]  

(2.15)

These results are discussed in the Appendix, Lemma 2.

To summarize, the weighted jackknife procedure reproduces the exact properties of linear estimates somewhat more neatly, but both \( \hat{\beta} \) and \( \hat{\beta}_w \) are unbiased and both \( V \) and \( V_w \) are biased. In the next section we compare the two procedures in two examples. A brief comparison for non-linear statistics is given in Section 3.

2.3 Examples

To illustrate the preceding results we consider two simple linear examples; a non-linear example is given in Section 3.

Example 2.1. Two-sample comparison.

Perhaps the simplest instance of the linear model is the two-sample problem, where two independent samples \( (Z_{11}, \ldots, Z_{1n_1}) \) and \( (Z_{21}, \ldots, Z_{2n_2}) \) are expressed

\[ Y_j = Z_{1j} \quad (j=1, \ldots, n_1), \quad Y_{n_1+k} = Z_{2k} \quad (k=1, \ldots, n_2), \]

and \( n = n_1 + n_2 \). A convenient coding device is

\[ x_{i1} = 1 \quad (i=1, \ldots, n); \quad x_{i2} = 0 \quad (i=1, \ldots, n_1), \quad x_{i2} = 1 \quad (i=n_1+1, \ldots, n); \]

the corresponding parameters are

\[ \beta_1 = E(Z_1), \quad \beta_2 = E(Z_2) - E(Z_1), \]

whose least-squares estimates are \( \hat{\beta}_1 = \overline{Z}_1 \) and \( \hat{\beta}_2 = \overline{Z}_2 - \overline{Z}_1 \).
It is straightforward to verify that

\[ w_i = \begin{cases} 
  n_1^{-1} & (i=1,\ldots,n_1) \\
  n_2^{-1} & (i=n_1+1,\ldots,n) 
\end{cases} \]

and \( \tilde{\beta} = \tilde{\beta}_w = \beta \). The two jackknife estimates of \( \text{var}(\hat{\beta}) \) are

\[
V = \frac{n-1}{n} \begin{bmatrix}
\frac{1}{(n_1-1)^2} SS_1 & -\frac{1}{(n_1-1)^2} SS_1 \\
\frac{1}{(n_1-1)^2} SS_1 & \frac{1}{(n_2-1)^2} SS_2 + \frac{1}{(n_2-1)^2} SS_2
\end{bmatrix}
\]

and

\[
V_w = \frac{n}{n-2} \begin{bmatrix}
\frac{n_1}{n_1} SS_1 & -\frac{n_1}{n_1} SS_1 \\
\frac{n_1}{n_1} SS_1 & \frac{n_1}{n_1} SS_1 + \frac{n_2}{n_2} SS_2
\end{bmatrix}
\]

where \( SS_k = \frac{n_k}{\sum_{n=1}^{n_k} (Z_{k,n} - \bar{Z}_k)^2} \) (k=1,2). That \( V \) and \( V_w \) are robust against the situation of unequal variances is clear; correspondingly if variances are equal, the jackknife estimates of \( \text{var}(\hat{\beta}_1) \) are inefficient.

As one numerical example, suppose that \( n_1 = 5 \) and \( n_2 = 10 \) with \( \sigma^2 = 1 \). The exact variance of \( \hat{\beta}_2 \) is \( 0.30 \); the balanced and weighted
jackknife estimates of $\text{var}(\hat{\beta}_2)$ have means 0.29 and 0.33. If the
two samples have unequal variances, the weighted jackknife estimate of
$\text{var}(\hat{\beta}_2)$ has average error of at most 8%.

Of course the "correct" analysis of the model would pay attention
to the possibility of different variances, so that no new methodology
is obtained from the jackknife in this particular case.

**Example 2.2. Simple linear regression.**

We take the design and data from Miller's (1974b) second numerical
example. The model is simple linear regression, i.e. $p=2$ with $x_{11} = 1$.
Data and related computations are given in Table 2.1 below. The example
is of interest because the $x_2$ values are bunched at one end of the
range, so that the $w$ values vary greatly. Our immediate concern is the
estimate of regression slope and the behavior of the two jackknife
procedures. Estimates and estimated standard errors are given at the
foot of the table, from which the main impression is of poor standard
error given by the balanced jackknife.

[Table 2.1 here]

Turning to average performance for this particular design, we find,
using (2.5),

$$
\text{var}(\hat{\beta}) = \begin{pmatrix} 0.676 & -0.091 \\ -0.091 & 0.014 \end{pmatrix},
\text{var}(\hat{\beta}) = \text{var}(\hat{\beta}) + \begin{pmatrix} 0.078 & -0.010 \\ -0.010 & 0.001 \end{pmatrix}.
$$

Average properties of variance estimates, computed from (2.7) and (2.13),
are
\[
E(V) = \begin{pmatrix}
1.007 & -0.135 \\
0.021 & 0.012
\end{pmatrix}, \quad E(V_w) = \begin{pmatrix}
0.535 & -0.071 \\
0.012
\end{pmatrix}.
\]

to be compared with \( \text{var}(\hat{\beta}) \). Here the weighted jackknife appears much better, but \( V_w \) has an average error of about 20%.

The main effect of the weighted jackknife in this design is to de-emphasise observations at \( x_2 = 1 \) relative to the balanced jackknife.
3. NON-LINEAR STATISTICS

The principal motivation for the earlier discussion is the need for an appropriate jackknife procedure that will handle unbalanced statistics \( t((x_1,Y_1),..., (x_n,Y_n)) \). The linear estimator \( \hat{\beta} \) discussed in Section 2 is a test case, where ideally exact properties of the original estimator will be reproduced; this is true for \( \hat{\beta}_w \) but not for \( V_w \), although in the latter case we have unexpected robustness against error variance heterogeneity. In general, a principal question is whether or not a given jackknife procedure removes first-order bias. Here we examine the simplest practical non-linear case where the parameter of interest is a non-linear scalar function of the linear model parameter \( \beta \).

Suppose again that observations are available on model (1.1), and that \( \theta = f(\beta) \), where \( f \) has continuous first and second derivatives \( \nabla f \) and \( \nabla^2 f \) respectively. The estimator for \( \beta \) is \( \hat{\beta} = (A^T A)^{-1} A^T Y \), and \( \hat{\theta} = f(\hat{\beta}) \). If we assume \( n^{-1} D_0 \to \Sigma \), then \( \hat{\beta} - \beta \) is of order \( n^{-1} \) and by Taylor expansion

\[
\hat{\theta} - \theta = (\hat{\beta} - \beta) \nabla f(\hat{\beta}) + \frac{1}{2} (\hat{\beta} - \beta) \nabla^2 f(\hat{\beta})(\hat{\beta} - \beta) + o_p(n^{-2}). 
\]  

For the weighted jackknife procedure pseudo-values are defined as

\[
Q_i = \hat{\theta} + n(1-w_i)(\hat{\theta} - \hat{\theta}_i).
\]

Substitution of (2.1) in (3.1) and evaluation of \( Q_i \) gives the jackknifed estimator

\[
\hat{\theta}_w = n^{-1} \sum Q_i = f(\hat{\beta}) - \frac{1}{2} \sum (1-w_j) \nabla^2 f(\hat{\beta}) D_0^{-1} \nabla f(\hat{\beta}) D_0^{-1} X_j R_j^2.
\]
It follows by Taylor expansion of \( f(\hat{\beta}) \) and taking expectations that, to order \( n^{-1} \),

\[
E(\tilde{\epsilon}_w) = \theta + \frac{1}{2} \sigma^2 \sum \nabla^2 f(\beta) \nabla f(\beta) \Sigma_{\delta}^{-1} S_{\beta} D_0 \Sigma_{\beta}^{-1} S_{\beta} \ .
\]  

(3.2)

This is exactly \( \theta \) since the last two terms both equal \( \frac{1}{2} \sigma^2 \text{tr}(\nabla^2 f(\beta) D_0^{-1}) \). Thus the leading bias term, assumed order \( n^{-1} \), is removed.

Calculation of \( E(\tilde{\epsilon}_w) \) to the next order of magnitude is generally complicated, and since we only have interest in the order, the case \( p=1 \) is satisfactory. For \( p=1 \) we readily calculate

\[
E(\tilde{\epsilon}_w) - \theta = - \frac{1}{6} \sigma^2 f'''(\beta) x_j^2 / (\text{v}_j^2)^3 .
\]

(3.3)

Note that this term is of order \( n^{-2} \), and vanishes if errors are symmetrically distributed or if \( f \) is quadratic.

The corresponding development for the balanced jackknife is very similar, the essential difference being that for \( \tilde{\theta} \) (3.2) holds with final term replaced by

\[
- \frac{1}{2} \left( \frac{n-1}{n} \right) \Sigma_{1-w_j}^{-1} S_{\delta} D_0 \Sigma_{\beta}^{-1} S_{\beta} \ .
\]

Since \( \max w_j \Rightarrow 0 \) we may expand \( (1-w_j)^{-1} \) and conclude that

\[
E(\tilde{\theta}) = \theta - \frac{1}{2} \sigma^2 \text{tr}(D_0^{-1} \Sigma_{\beta}^{-1} S_{\beta} \Sigma_{\beta}^{-1} S_{\beta}) + o(n^{-2}) \ .
\]

(3.4)

That is, the \( n^{-1} \) bias term is removed, but the remaining bias is of higher order than that for \( \tilde{\epsilon}_w \) unless all \( w_j \) are of order \( n^{-1} \), which is not automatic.
The last remark is not of purely mathematical interest. From a practical viewpoint it suggests that when the \( \chi \) vectors have severe non-uniform dispersion in the observed design, so that the \( w_1 \) are of different orders of magnitude, then \( \widetilde{\theta}_w \) is superior to \( \widetilde{\theta} \), i.e. weighting is profitable.

A simple illustration is offered by Example 2.2 with \( \theta = \beta_1/\beta_2 \). The required computations are given in Table 3.1.

The delta method for evaluating bias gives approximately 0.1, which makes \( \widetilde{\theta}_w \) seem reasonable, as does the comparison of estimated standard errors. The main difference between the two jackknifes is due to the first observation, which has a heavy weight and modest residual.
4. ROBUST REGRESSION USING PSEUDO-VALUES

In a recent article Hinkley (1976b) has shown that the jackknife pseudo-values may be used to define robust measures of correlation, essentially by treating pseudo-values as observations on a location model. For the linear model, Cook (1975) has discussed the use of \( \hat{\beta}_i \) in exhibiting important large residuals, the implicit purpose being to isolate data points that might be de-emphasised or omitted in a re-fit of the model. These ideas clearly suggest possible methods of robust regression based on pseudo-values \( Q_i \) and \( P_i \). Only a brief discussion will be given here.

By analogy with the classical location problem, where robust estimators (Huber, 1972) are designed to reduce large values of the sample influence function, we may propose regression estimates to replace \( n^{-1} \sum Q_i \) that reduce large values of \( Q_i \). Of course \( Q_i \) is a \( p \)-dimensional vector, so that we can work either on individual coordinates or on all coordinates simultaneously. In any event we must suppose that the error terms \( e_j \) in the model (1.1) have symmetric distributions, so that the \( Q_i \) have symmetric distributions about \( \beta \).

If we estimate separately on each coordinate of \( \beta \), denote the components of \( Q_i \) by \( Q_{ji} \) with corresponding order statistics \( Q_j(i) \) \( (j=1, \ldots, p; i=1, \ldots, n) \). Then two standard forms of robust estimate are

\[
\beta_j^* = \frac{1}{\sum_{i=1}^{n} h_i Q_j(i)}, \quad h_i = h_{n-i+1}, \quad \sum h_i = 1, \tag{4.1}
\]

and
solution to \( \sum_{i=1}^{n} \psi(Q_j(1)-\hat{P}_j^*) = 0 \), \( \psi(-u) = -\psi(u) \); \( (4.2) \)

see Huber (1972). Either form of estimate is unbiased under our model assumptions; further theoretical study would be somewhat easier for (4.2), as in Hinkley (1976b). There is some conceptual difficulty in this approach, in that each observation is treated in \( p \) different ways, rather than being classified once as "good" or "bad", but this is often useful flexibility.

If we estimate the vector \( \beta \) simultaneously, we use a scalar measure of the sample influence vector \( Q_1 \cdot \hat{\beta} \), one possibility being

\[
c_i = n^{-1}(Q_1 \cdot \hat{\beta})^T D_0 (Q_1 \cdot \hat{\beta}) = nw_i R_i^2 (i=1,\ldots,n); \]

see Cook (1975) for a similar measure. Note that \( c_i \) is a measure of the information content in the \( i \)th observation under the assumption of equal error variance. Then, denoting the ordered values by \( c(i) \) \((i=1,\ldots,n)\), a simple form of robust estimator is the trimmed mean

\[
\beta^* = \frac{1}{n-m} \sum_{c(i) \leq c(n-m)} Q_1 , \quad (4.3)
\]

where \( m \) is to be chosen. Theoretical study of such estimators is likely to be challenging.

The standard form of robust regression estimate (Huber, 1973; Andrews, 1974) corresponding to (4.2) is

solution to \( \sum_{i=1}^{n} \psi(Y_j - x_j^T \beta^*) = 0 \), \( (4.4) \)

which clearly treats all residuals with equal weight. The robust
jackknife estimates (4.1)-(4.3) seem potentially superior in this regard, since "harmless" large residuals are ignored. Also, (4.1) and (4.3) in particular are probably easier to calculate, in part because no scaling of residuals is needed.

Estimate (4.4) is computed iteratively, possibly starting with the median estimator of Andrews (1974) rather than the least squares estimate. The robust jackknife estimates (4.1)-(4.3) might be replaced by analogous iterative procedures, replacing \( \hat{\beta} \) by an initial estimate \( \beta^+ \) in the definition of \( Q_1 \) and then working iteratively through the defining equation; fast convergence might be anticipated from the fact that when \( \hat{\beta} \) is replaced by \( \beta^+ \) in defining \( Q_1 \), it remains true that \( n^{-1} \sum Q_1 = \hat{\beta} \).

The above account is clearly no more than an outline of robust jackknife regression. To take a simple numerical example we return to the simple linear regression data of Example 2.1. The required data is given in the last three rows of Table 2.1, from which we are alerted to the first and ninth observations by their large values of \( c_1^* \). Note that the corresponding residuals do not appear unusual, but these data points are very influential in the regression fit. If we use the estimate (4.3) with \( m=2 \), resulting estimates are

\[
\begin{align*}
b_1^* &= 1.341 \\
b_2^* &= 1.017.
\end{align*}
\]

The same points stand out in the estimation of the ratio \( \theta = \beta_1/\beta_2 \), as we see in the final row of Table 3.1; the trimmed mean estimate of \( \theta \) is \( \theta^* = 1.29 \).
An interesting real example with severe imbalance is provided by the data in Table 4.1. The data comes from measurements of enlarger magnification \( m \) and object-to-image distance \( d \) for a lens of unknown focal length \( \varphi \). In fact the distance \( d \) is measured from a point distant \( b \) (unknown) from the image, so that we observe \( y = d - b \). Simple physics provides the relationship

\[
y = \alpha + \varphi(m + m^{-1}) = \beta_1 + \beta_2x,
\]

and our interest is primarily in \( \beta_2 = \varphi \). For the present purpose we assume \( x \) to have no error. The unweighted least squares analysis is summarized as follows

\[
\begin{align*}
\hat{\beta}_1 &= -5.300, & \hat{\beta}_2 &= 18.878, & \Sigma x_1^2 &= 2.81 \\
\bar{x} &= 2.385 & \Sigma(x_j - \bar{x})^2 &= 20.49, & \text{est. s.e.}(\hat{\beta}_2) &= 0.05
\end{align*}
\]

Individual residuals are given in the table, from which it is clear that for \( x > 2.5 \) the residuals are systematically large. If these are reflective of larger variability, rather than lack of model fit, then of course the estimated standard error for \( \hat{\beta}_2 \) will be too small. This can be seen from the jackknife estimate \( V_{\hat{\beta}_2} \), which gives estimated s.e.\( (\hat{\beta}_2) = 0.12 \).

Since the weights \( w \) are also large for \( x > 2.5 \), the weighted residuals \( c = nw^2 \) are very large in that range. The largest ten values are asterisked. Weighted pseudo-value components defined by \( q_1 = \frac{Q_{-1}}{\hat{\beta}_2} \) are given in the final column of Table 4.1. Discarding points with ten largest values of \( c \) and averaging the remaining pseudo-values gives \( \beta_2^* = 18.90 \).
The alternative coordinate approach essentially treats the second pseudo-value components as observations on a location model with mean $\beta_2$. Trimming the five largest and five smallest components (double asterisked in the table), we obtain the trimmed $q$ mean value 0.06, corresponding to $\beta_{2}^{**} = 18.94$. The standard error computed from the winsorized sample variance of the pseudo-values is 0.02.
5. DISCUSSION

This paper really does little more than scratch the surface of two problems: the suitable definition of a jackknife procedure for unbalanced data, and the application to robust estimation.

The balanced jackknife certainly looks inferior to the proposal in Section 2.2, but the standard error estimates provided by that proposal are not truly satisfactory despite their robustness property.

The method of robust regression sketched out in Section 4 is intuitively promising, in that residuals are weighted by their real effect on the estimation. Detailed theoretical study would be of interest, but numerical comparison with recent unpublished methods of Tukey, Hoaglin and others is of more importance.

An alternative to the weighted jackknife procedure of Section 2.2 is to omit data in small groups, where groups are chosen so as to equalise information content. Initial results for this are complicated, and in any event grouping is likely to lose information; see Hinkley (1976a).

ACKNOWLEDGEMENTS

This work was partially supported by Public Health Service Grant 1 RO1 GM21215-01 at Stanford University. Regression data for the examples were kindly provided by R.G. Miller, Jr. and L. Moses.
APPENDIX. DETAILS OF MATHEMATICAL RESULTS

Some results quoted in the text are given fuller explanation here.

Lemma 1 (Regression influence function)

Let the design point \( \mathbf{x} \) and the response variable \( Y \) have a joint distribution function \( G \) such that

\[
E_G \{ \left[ \begin{array}{c} \mathbf{x} \\ Y \end{array} \right] | (\mathbf{x}, Y) \} = \left( \begin{array}{cc} \Sigma(G) & \chi(G) \\ \tau(G) \end{array} \right)
\]

and define \( \beta(G) = \Sigma^{-1}(G) \chi(G) \). Then the influence function of \( \beta \) at \( (\mathbf{x}, y) \) is

\[
I_G(\beta; \mathbf{x}, y) = \Sigma^{-1} \chi(y - \mathbf{x}^T \beta).
\]

Proof follows by direct calculation, using the definition of influence function (first von Mises derivative) for arbitrary \( S(G) \):

\[
I(S; z) = \lim_{\epsilon \to 0} \frac{S((1-\epsilon)G + \epsilon U_z) - S(G)}{\epsilon}
\]

where \( U_z \) has mass 1 at the point \( z \). Note that \( \mathbf{x} \) may have probability or design measure.

The sample influence function (2.9) is obtained by substituting estimates \( \hat{\beta} \) and \( \hat{\Sigma} = n^{-1} \mathbf{A}^T \mathbf{A} \).

A corollary result is that the influence function of \( \theta = f(\beta) \) is

\[
I_G(\theta; \mathbf{x}, y) = (\nabla f(\beta))^T I_G(\beta; \mathbf{x}, y),
\]

22
which implies that in Section 3

\[ Q_1 = \hat{\theta} + \hat{I}(\theta; X_1, Y_1) + o(n^{-1}); \]

the remainder term involves second von Mises derivatives and may be used to obtain (3.3). See Hinkley (1976b).

**Lemma 2 (Consistency of variance estimate)**

For the model (1.1) with \( \text{var}(\varepsilon) = \Lambda = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2) \), let \( R = Y - A\hat{\beta} \) and \( L = \text{diag}(RR^T) \). Then if

\[ n^{-1} A^T A \rightarrow \Sigma \text{ p.d., } n^{-1} A^T A \wedge A \rightarrow \Gamma \text{ p.d. } \quad (n \rightarrow \infty) \]

and if \( E(\varepsilon_i^4) \) is uniformly bounded,

(i) \( n \text{ var}(\hat{\beta}) \rightarrow \Sigma^{-1} \Gamma \Sigma^{-1} \)

(ii) \( nV_w = \frac{n}{n-p} (A^T A)^{-1} A^T L A (A^T A)^{-1} \rightarrow \Sigma^{-1} \Gamma \Sigma^{-1} \).

**Proof.** Part (i) follows by assumption, since \( \text{var}(\hat{\beta}) = (A^T A)^{-1} A^T \Lambda A (A^T A)^{-1} \).

Part (ii) follows by using a minor variation of the proof of Lemma 3.14 in Miller (1974b) to establish \( n^{-1} A^T L A \rightarrow \Gamma \).

A corollary is that if \( \Lambda = \sigma^2 I \), then \( nV_w \rightarrow \sigma^2 \Sigma^{-1} \).

23
Table 2.1 Jackknife analysis of artificial simple linear regression data ($\beta_1 = \beta_2 = \sigma = 1$).

<table>
<thead>
<tr>
<th></th>
<th>x_2</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>8.5</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td></td>
<td>0.91</td>
<td>4.24</td>
<td>6.59</td>
<td>8.22</td>
<td>7.53</td>
<td>7.89</td>
<td>10.13</td>
<td>9.25</td>
<td>8.92</td>
<td>11.35</td>
</tr>
<tr>
<td>residual R</td>
<td>-0.99</td>
<td>0.24</td>
<td>0.80</td>
<td>1.38</td>
<td>0.39</td>
<td>-0.29</td>
<td>0.90</td>
<td>-0.50</td>
<td>-1.56</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>weight w</td>
<td>0.51</td>
<td>0.26</td>
<td>0.13</td>
<td>0.10</td>
<td>0.10</td>
<td>0.11</td>
<td>0.14</td>
<td>0.17</td>
<td>0.20</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>balanced</td>
<td>$P_1$</td>
<td>-9.78</td>
<td>2.06</td>
<td>2.00</td>
<td>2.28</td>
<td>1.38</td>
<td>0.73</td>
<td>0.39</td>
<td>1.37</td>
<td>3.00</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>$P_2$</td>
<td>2.44</td>
<td>0.91</td>
<td>0.95</td>
<td>0.99</td>
<td>1.03</td>
<td>1.02</td>
<td>1.27</td>
<td>0.88</td>
<td>0.47</td>
<td>1.06</td>
</tr>
<tr>
<td>weighted</td>
<td>$Q_1$</td>
<td>-4.94</td>
<td>1.84</td>
<td>1.97</td>
<td>2.28</td>
<td>1.37</td>
<td>0.73</td>
<td>0.41</td>
<td>1.33</td>
<td>2.76</td>
<td>0.79</td>
</tr>
<tr>
<td></td>
<td>$Q_2$</td>
<td>1.51</td>
<td>0.93</td>
<td>0.95</td>
<td>0.99</td>
<td>1.03</td>
<td>1.02</td>
<td>1.26</td>
<td>0.89</td>
<td>0.55</td>
<td>1.06</td>
</tr>
<tr>
<td>discrepancy c</td>
<td>5.01</td>
<td>0.16</td>
<td>0.32</td>
<td>1.20</td>
<td>0.16</td>
<td>0.09</td>
<td>1.12</td>
<td>0.42</td>
<td>3.69</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

Slope estimates: least squares $\hat{\beta}_2 = 1.047$; balanced jackknife $\bar{\beta}_2 = 1.101$

Estimated s.e.'s: $\hat{\sigma}_{22} = 0.100$, $\hat{\sigma}_{w,22} = 0.102$, $\hat{\sigma}_{22} = 0.161$

Table 3.1 Jackknife analysis of $\beta_1 / \beta_2$ from regression data in Table 2.1.

<table>
<thead>
<tr>
<th></th>
<th>x_2</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>6</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>8.5</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>w</td>
<td></td>
<td>0.5087</td>
<td>0.2603</td>
<td>0.1260</td>
<td>0.1017</td>
<td>0.1017</td>
<td>0.1060</td>
<td>0.1589</td>
<td>0.1660</td>
<td>0.2005</td>
<td>0.2903</td>
</tr>
<tr>
<td>p</td>
<td>-12.368</td>
<td>2.055</td>
<td>1.975</td>
<td>2.215</td>
<td>1.327</td>
<td>0.719</td>
<td>0.184</td>
<td>1.426</td>
<td>3.168</td>
<td>0.729</td>
<td></td>
</tr>
<tr>
<td>q</td>
<td>-6.381</td>
<td>1.833</td>
<td>1.941</td>
<td>2.212</td>
<td>1.326</td>
<td>0.720</td>
<td>0.212</td>
<td>1.381</td>
<td>2.906</td>
<td>0.747</td>
<td></td>
</tr>
</tbody>
</table>

$\hat{\beta}_1 = 0.8546$, $\hat{\beta}_2 = 1.047$, $\hat{\beta} = 0.816$, est. s.e.($\hat{\beta}$) = 0.93 (using delta method)

balanced jackknife: $\bar{\beta} = 0.145$, est. s.e. = $\hat{\sigma} = 1.417$

weighted jackknife: $\hat{\beta}_w = 0.690$, est. s.e. = $\hat{\sigma}_w = 0.825$
Table 4.1: Jackknife regression analysis of photographic enlarger data.

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>r</th>
<th>w</th>
<th>c</th>
<th>q</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.200</td>
<td>92.05</td>
<td>-0.817</td>
<td>0.105</td>
<td>15.12*</td>
<td>-6.285**</td>
</tr>
<tr>
<td>4.250</td>
<td>75.00</td>
<td>0.067</td>
<td>0.186</td>
<td>0.05</td>
<td>0.341</td>
</tr>
<tr>
<td>3.851</td>
<td>67.51</td>
<td>0.110</td>
<td>0.123</td>
<td>0.08</td>
<td>0.411**</td>
</tr>
<tr>
<td>3.633</td>
<td>63.50</td>
<td>0.215</td>
<td>0.094</td>
<td>0.24*</td>
<td>0.733**</td>
</tr>
<tr>
<td>3.520</td>
<td>57.69</td>
<td>0.314</td>
<td>0.061</td>
<td>0.34*</td>
<td>0.802**</td>
</tr>
<tr>
<td>3.072</td>
<td>55.50</td>
<td>0.806</td>
<td>0.041</td>
<td>1.49*</td>
<td>1.513**</td>
</tr>
<tr>
<td>2.900</td>
<td>49.28</td>
<td>-0.067</td>
<td>0.031</td>
<td>0.01</td>
<td>-0.094**</td>
</tr>
<tr>
<td>2.874</td>
<td>49.26</td>
<td>0.304</td>
<td>0.030</td>
<td>0.16*</td>
<td>0.406</td>
</tr>
<tr>
<td>2.672</td>
<td>45.64</td>
<td>0.197</td>
<td>0.022</td>
<td>0.50*</td>
<td>0.390</td>
</tr>
<tr>
<td>2.634</td>
<td>45.05</td>
<td>0.625</td>
<td>0.021</td>
<td>0.46*</td>
<td>0.425**</td>
</tr>
<tr>
<td>2.485</td>
<td>41.78</td>
<td>0.168</td>
<td>0.018</td>
<td>0.03</td>
<td>0.046</td>
</tr>
<tr>
<td>2.392</td>
<td>39.81</td>
<td>-0.047</td>
<td>0.018</td>
<td>0.00</td>
<td>-0.001</td>
</tr>
<tr>
<td>2.368</td>
<td>39.38</td>
<td>-0.024</td>
<td>0.018</td>
<td>0.00</td>
<td>0.001</td>
</tr>
<tr>
<td>2.285</td>
<td>37.50</td>
<td>-0.337</td>
<td>0.018</td>
<td>0.11*</td>
<td>0.092</td>
</tr>
<tr>
<td>2.249</td>
<td>37.12</td>
<td>-0.027</td>
<td>0.019</td>
<td>0.00</td>
<td>0.010</td>
</tr>
<tr>
<td>2.225</td>
<td>36.56</td>
<td>-0.144</td>
<td>0.019</td>
<td>0.02</td>
<td>0.063</td>
</tr>
<tr>
<td>2.188</td>
<td>35.94</td>
<td>-0.066</td>
<td>0.020</td>
<td>0.00</td>
<td>0.036</td>
</tr>
<tr>
<td>2.156</td>
<td>35.50</td>
<td>-0.090</td>
<td>0.020</td>
<td>0.01</td>
<td>0.054</td>
</tr>
<tr>
<td>2.129</td>
<td>34.84</td>
<td>0.052</td>
<td>0.021</td>
<td>0.02</td>
<td>0.036</td>
</tr>
<tr>
<td>2.118</td>
<td>34.55</td>
<td>-0.134</td>
<td>0.021</td>
<td>0.02</td>
<td>0.098</td>
</tr>
<tr>
<td>2.109</td>
<td>34.49</td>
<td>-0.024</td>
<td>0.022</td>
<td>0.00</td>
<td>0.018</td>
</tr>
<tr>
<td>2.091</td>
<td>34.19</td>
<td>0.016</td>
<td>0.022</td>
<td>0.00</td>
<td>-0.013</td>
</tr>
<tr>
<td>2.083</td>
<td>34.06</td>
<td>0.037</td>
<td>0.022</td>
<td>0.00</td>
<td>0.051</td>
</tr>
<tr>
<td>2.050</td>
<td>33.25</td>
<td>-0.150</td>
<td>0.023</td>
<td>0.03</td>
<td>0.137</td>
</tr>
<tr>
<td>2.040</td>
<td>33.15</td>
<td>-0.081</td>
<td>0.024</td>
<td>0.00</td>
<td>0.076</td>
</tr>
<tr>
<td>2.028</td>
<td>32.78</td>
<td>-0.205</td>
<td>0.024</td>
<td>0.04</td>
<td>0.200</td>
</tr>
<tr>
<td>2.025</td>
<td>32.59</td>
<td>-0.017</td>
<td>0.024</td>
<td>0.00</td>
<td>0.017</td>
</tr>
<tr>
<td>2.022</td>
<td>32.91</td>
<td>0.038</td>
<td>0.024</td>
<td>0.00</td>
<td>0.038</td>
</tr>
<tr>
<td>2.011</td>
<td>32.66</td>
<td>-0.004</td>
<td>0.025</td>
<td>0.00</td>
<td>0.004</td>
</tr>
<tr>
<td>2.003</td>
<td>32.56</td>
<td>0.047</td>
<td>0.025</td>
<td>0.00</td>
<td>-0.049**</td>
</tr>
<tr>
<td>2.001</td>
<td>32.51</td>
<td>0.035</td>
<td>0.025</td>
<td>0.00</td>
<td>-0.057</td>
</tr>
<tr>
<td>2.000</td>
<td>32.61</td>
<td>0.154</td>
<td>0.025</td>
<td>0.03</td>
<td>-0.162**</td>
</tr>
<tr>
<td>2.002</td>
<td>32.44</td>
<td>-0.054</td>
<td>0.025</td>
<td>0.00</td>
<td>0.057</td>
</tr>
<tr>
<td>2.002</td>
<td>32.13</td>
<td>-0.364</td>
<td>0.025</td>
<td>0.19*</td>
<td>0.361</td>
</tr>
<tr>
<td>2.009</td>
<td>32.57</td>
<td>-0.056</td>
<td>0.025</td>
<td>0.00</td>
<td>0.058</td>
</tr>
<tr>
<td>2.011</td>
<td>32.63</td>
<td>-0.034</td>
<td>0.025</td>
<td>0.00</td>
<td>0.035</td>
</tr>
<tr>
<td>2.020</td>
<td>32.79</td>
<td>-0.044</td>
<td>0.024</td>
<td>0.00</td>
<td>0.044</td>
</tr>
<tr>
<td>2.013</td>
<td>32.06</td>
<td>-0.019</td>
<td>0.024</td>
<td>0.00</td>
<td>0.018</td>
</tr>
<tr>
<td>2.050</td>
<td>33.88</td>
<td>-0.020</td>
<td>0.023</td>
<td>0.00</td>
<td>0.018</td>
</tr>
<tr>
<td>2.054</td>
<td>33.50</td>
<td>0.024</td>
<td>0.023</td>
<td>0.00</td>
<td>-0.022</td>
</tr>
<tr>
<td>2.091</td>
<td>34.13</td>
<td>-0.044</td>
<td>0.022</td>
<td>0.00</td>
<td>0.035</td>
</tr>
<tr>
<td>2.100</td>
<td>34.38</td>
<td>0.036</td>
<td>0.022</td>
<td>0.00</td>
<td>-0.028</td>
</tr>
<tr>
<td>2.107</td>
<td>34.44</td>
<td>-0.036</td>
<td>0.022</td>
<td>0.00</td>
<td>0.027</td>
</tr>
<tr>
<td>2.129</td>
<td>34.84</td>
<td>-0.052</td>
<td>0.021</td>
<td>0.00</td>
<td>0.036</td>
</tr>
<tr>
<td>2.140</td>
<td>35.00</td>
<td>-0.099</td>
<td>0.021</td>
<td>0.01</td>
<td>0.066</td>
</tr>
<tr>
<td>2.150</td>
<td>35.25</td>
<td>-0.058</td>
<td>0.021</td>
<td>0.00</td>
<td>0.024</td>
</tr>
<tr>
<td>2.167</td>
<td>35.65</td>
<td>0.021</td>
<td>0.020</td>
<td>0.00</td>
<td>-0.013</td>
</tr>
<tr>
<td>2.195</td>
<td>36.11</td>
<td>-0.208</td>
<td>0.020</td>
<td>0.00</td>
<td>0.015</td>
</tr>
<tr>
<td>2.213</td>
<td>36.62</td>
<td>0.152</td>
<td>0.019</td>
<td>0.02</td>
<td>-0.071**</td>
</tr>
<tr>
<td>2.256</td>
<td>37.00</td>
<td>-0.289</td>
<td>0.019</td>
<td>0.09*</td>
<td>0.102</td>
</tr>
<tr>
<td>2.288</td>
<td>38.00</td>
<td>0.107</td>
<td>0.018</td>
<td>0.01</td>
<td>-0.031</td>
</tr>
<tr>
<td>2.318</td>
<td>38.26</td>
<td>-0.020</td>
<td>0.018</td>
<td>0.00</td>
<td>0.004</td>
</tr>
<tr>
<td>2.352</td>
<td>39.25</td>
<td>0.040</td>
<td>0.018</td>
<td>0.00</td>
<td>0.003</td>
</tr>
<tr>
<td>2.391</td>
<td>39.88</td>
<td>0.042</td>
<td>0.018</td>
<td>0.00</td>
<td>-0.001</td>
</tr>
<tr>
<td>2.478</td>
<td>41.36</td>
<td>-0.100</td>
<td>0.018</td>
<td>0.01</td>
<td>-0.025</td>
</tr>
<tr>
<td>2.500</td>
<td>41.89</td>
<td>-0.006</td>
<td>0.019</td>
<td>0.00</td>
<td>-0.002</td>
</tr>
</tbody>
</table>

* ten largest values of c  ** five largest and five smallest values of q
REFERENCES


