JACKKNIFING IN UNBALANCED SITUATIONS

David V. Hinkley
Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

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

1. INTRODUCTION

During the two decades since Quenouille and Tukey introduced the jackknife technique for reducing bias and estimating standard error, 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 [6].

Miller [7] gives the first detailed 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, and simple illustrative examples are given in Section 2.3. The more important case of non-linear functions of linear model parameters is discussed, and an example given, in Section 3.

A second aspect of the jackknife is the use of pseudo-values in robust data analysis; a detailed account in the case of correlation estimation has been given by Hinkley [3]. In Section 4 we briefly discuss the potential 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.

The ideas are illustrated on a large data set in Section 5.

Throughout this paper the following model is assumed

$$Y = A\beta + e \tag{1.1}$$

where

$$Y^T = (Y_1, \cdots, Y_n), \beta^T = (\beta_1, \cdots, \beta_p), \quad e^T = (e_1, \cdots, e_n)$$

and

$$A = \begin{matrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{matrix} = \begin{matrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{matrix}, \tag{1.2}$$

such that A is of rank p when any single row is deleted. Unless otherwise stated, the e_j are taken to be i.i.d. with mean zero and constant variance σ^2.

If the model (1.1) includes a constant term, then we take this to be β_1 (not β_0) and hence define $x_{i1} \equiv 1$. It is important not to replace x_{ik} by $x_{ik} - \bar{x}_{\cdot k}$ $(k \geq 2)$ when $x_{i1} \equiv 1$, because all parameter values need to be independent of the design: one can check that incorrect jackknife results are obtained for a constant term when the x_{ik} are mean-adjusted.

Some standard statistical measures used in the sequel are the design matrix

$$D_0 = A^T A,$$

the least-squares estimate

$$\hat{\beta} = D_0^{-1} A^T Y$$

and the residual vector

$$R = Y - A\hat{\beta} = (I - AD_0^{-1}A^T)Y. \tag{1.3}$$

Another important set of quantities is

$$w_i = \mathbf{x}_i^T D_0^{-1} \mathbf{x}_i \quad (i = 1, \cdots, n) \tag{1.4}$$

Received July 1976; revised April 1977

which measure distances of single design points from the center of the design. In a replicated 2^k factorial all w_i are equal.

Many relevant calculations have been taken from Miller [7] and Cook [2], 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 β to be the least squares estimator

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

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

$$\begin{aligned}\hat\beta_{-i} &= \hat\beta - \frac{(A^TA)^{-1}\mathbf{x}_i(Y_i - \mathbf{x}_i^T)}{1 - \mathbf{x}_i^T(A^TA)^{-1}\mathbf{x}_i} \\ &= \hat\beta - \frac{D_0^{-1}\mathbf{x}_iR_i}{1 - w_i} \qquad (i = 1, \cdots, n),\end{aligned}$$

by definitions (1.3) and (1.4). Notice that $R_i/(1 - w_i)$ is the difference between Y_i and its least squares predictor from all other observations.

To describe the standard jackknife procedure we first define pseudovalues

$$P_i = n\hat\beta - (n - 1)\hat\beta_{-i} \quad (i = 1, \cdots, n), \qquad (2.2)$$

from which the jackknifed estimator is given by

$$\tilde\beta = n^{-1}\sum P_i. \qquad (2.3)$$

Using (2.1) we obtain

$$P_i = \hat\beta + (n - 1)\, D_0^{-1}\mathbf{x}_iR_i(1 - w_i)^{-1}$$

and

$$\tilde\beta = \hat\beta + (n - 1)n^{-1}D_0^{-1}\sum (1 - w_i)^{-1}\mathbf{x}_iR_i. \qquad (2.4)$$

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 $\hat\beta$,

$$\operatorname{var}(\tilde\beta) > \operatorname{var}(\hat\beta);$$

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 [7].

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

$$\operatorname{var}(\hat\beta) = \sigma^2D_0^{-1}, \quad \operatorname{var}(R) = \sigma^2(I - AD_0^{-1}A^T),$$

so that from (2.4) we have immediately

$$\operatorname{var}(\tilde\beta) = \sigma^2\left\{D_0^{-1} + \left(\frac{n-1}{n}\right)^2 \cdot D_0^{-1}(D_2 - D_1D_0^{-1}D_1)D_0^{-1}\right\}, \qquad (2.5)$$

where

$$D_k = \sum (1 - w_j)^{-k}\mathbf{x}_j\mathbf{x}_j^T \quad (k = 0, 1, 2).$$

Supposing w_i to be of order n^{-1}, we may expand $(1 - w_i)^k$ in series and verify that $\operatorname{var}(\tilde\beta) - \operatorname{var}(\hat\beta)$ 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

$$V = \{n(n - 1)\}^{-1}\sum (P_i - \tilde\beta)(P_i - \tilde\beta)^T, \qquad (2.6)$$

which may be used to estimate both $\operatorname{var}(\hat\beta)$ and $\operatorname{var}(\tilde\beta)$; for a simple account of the rationale for this in the balanced case, see Hinkley [4]. It is not hard to show, under mild conditions including $nD_0^{-1} \to \sum > 0$, that $nV \to n \operatorname{var}(\hat\beta)$, i.e., that V is an accurate large-sample variance estimate; see Miller [7]. However V is not unbiased, and straightforward calculations show that

$$E(V) = \left(\frac{n-1}{n}\right)D_0^{-1} \cdot \{D_1 - n^{-1}(D_2 - D_1D_0^{-1}D_1)\}D_0^{-1}\sigma^2, \qquad (2.7)$$

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

$$D_k^* = \sum w_j^k\mathbf{x}_j\mathbf{x}_k^T \quad (k = 1, 2),$$

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

$$E(V) \cong \frac{n-1}{n}(D_0^{-1} + D_0^{-1}D_1^*D_0^{-1} + D_2D_2^*D_0^{-1})^2. \qquad (2.8)$$

To summarize these developments, we have found that for an exactly linear estimator *(i)* the jackknifed estimator $\tilde\beta$ 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 $\mathbf{x}_i$ are sampled from a multivariate normal population, the estimated likelihood of the value $\mathbf{x}_i$ is a decreasing function of w_i. In addition, tr {var $(\hat{\beta} - \hat{\beta}_{-i})$} is an increasing function of w_i. This suggests that in the contribution of the ith observation to the jackknifed estimator, $\hat{\beta} - \hat{\beta}_{-i}$ have a weight decreasing in w_i. A specific choice of weight is indicated by the fact that

$$n(1 - w_i)(\hat{\beta} - \hat{\beta}_{-i}) = nD_0^{-1}\mathbf{x}_i R_i = \hat{I}(\beta;\ \mathbf{x}_i^T,\ Y_i) \quad (2.9)$$

the estimated influence function of β at $(\mathbf{x}_i^T, Y_i)$; see Appendix, Lemma 1. (Roughly speaking, the influence function at $(\mathbf{x}, y)$ is proportional to the incremental change in β when a very small fraction of the measurement population is moved to $(\mathbf{x}, y)$.)

We therefore propose the weighted pseudo-value

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

the weighted jackknife estimator

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

and the variance estimate

$$V_w = \{n(n - p)\}^{-1} \sum (Q_i - \tilde{\beta}_w)(Q_i - \tilde{\beta}_w)^T$$
$$= n(n - p)^{-1} D_0^{-1}(\sum R_j^2 \mathbf{x}_j \mathbf{x}_j^T) D_0^{-1}, \quad (2.12)$$

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 definition (2.10) is essentially due to Quenouille [8].

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

The general expectation of V_w is easily seen to be

$$E(V_w) = n(n - p)^{-1}(\cdot (D_0^{-1} - D_0^{-1}D_1{}^*D_0^{-1})\sigma^2, \quad (2.13)$$

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

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

$$\hat{V} = (n - p)^{-1} \sum R_j^2 D_0^{-1}, \quad (2.14)$$

in that V_w (and V) are robust against nonhomogeneity of error variance. To see this, suppose that in (1.1) var (e) = diag $(\sigma_1^2, \cdots, \sigma_n^2) = \Lambda$. Then

$$\text{var}\ (\hat{\beta}) = D_0^{-1}A^T \Lambda\ A D_0^{-1}. \quad (2.15)$$

Since $E(R_j^2) \cong \sigma_j^2$ we have

$$E(V) \cong n^{-1}\ \text{tr}\ (\Lambda) D_0^{-1}$$

and

$$E(V_w) \cong D_0^{-1} \sum \sigma_j^2 \mathbf{x}_j \mathbf{x}_j^T D_0^{-1} = D_0^{-1} A^T \Lambda A D_0^{-1}; \quad (2.16)$$

justification of these results is given in the Appendix, Lemma 2.

Thus V_w approximates the true variance (2.15) of $\hat{\beta}$ when error variances are unequal, whereas the usual estimate does not. For the enlarger magnification example in Section 5, the estimates of variance for the least-squares slope estimate differ by a factor of 5 because of variance heterogeneity.

2.3 Illustrative examples

To illustrate the preceding results we consider two elementary linear examples. A non-linear example is given in Section 3.

Example 2.1 Two-point design

Perhaps the simplest instance of the linear model is that of simple linear regression with a two-point design. We suppose that

$$x_{i1} = 1\ (i = 1, \cdots, n);\quad x_{i2} = 0\ (i = 1, \cdots, n_0);$$
$$x_{i2} = 1\ (i = n_0 + 1, \cdots, n)$$

with $n = n_0 + n_1$. The model is thus $EY_i = \beta_1 + \beta_2 x_{i2}$, and the least-squares estimates are

$$\hat{\beta}_1 = \bar{Y}_0 \quad \text{and} \quad \hat{\beta}_2 = \bar{Y}_1 - \bar{Y}_0,$$

where

$$\bar{Y}_0 = n_0^{-1} \sum_{j=1}^{n_0} Y_j, \quad \bar{Y}_1 = n_1^{-1} \sum_{j=n_0+1}^{n} Y_j.$$

Further denote replicate sums of squares by

$$SS_0 = \sum_{j=1}^{n_0} (Y_j - \bar{Y}_0)^2,\ SS_1 = \sum_{j=n_0+1}^{n} (Y_j - \bar{Y}_1)^2.$$

It is straightforward to verify that

$$w_i = \begin{cases} n_0^{-1} & (i = 1, \cdots, n_0) \\ n_1^{-1} & (i = n_0 + 1, \cdots, n) \end{cases}$$

and that $\tilde{\beta} = \tilde{\beta}_w = \tilde{\beta}$.

Consider the estimates of var $(\hat{\beta}_2)$. From (2.6) and (2.12) we obtain

$$V_{22} = \frac{n - 1}{n} \left\{ \frac{SS_0}{(n_0 - 1)^2} + \frac{SS_1}{(n_1 - 1)^2} \right\}$$

$$\hat{V}_{22,w} = \frac{n}{n - 2} \left(\frac{SS_0}{n_0^2} + \frac{SS_1}{n_1^2} \right),$$

as compared to the usual mean-square estimate from (2.14)

$$\hat{V}_{22} = \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \left(\frac{SS_0 + SS_1}{n - 2} \right).$$

If the error variances at $x_2 = 0$ and $x_2 = 1$ are σ_0^2 and σ_1^2 respectively, then

$$\text{var}\,(\hat{\beta}_2) = \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\,.$$

Thus if $\sigma_0^2 \neq \sigma_1^2$, both V_{22} and $V_{22,w}$ reflect this.

As a particular numerical example, let $n = 10$ and $\sigma_1^2 = 2\sigma_0^2$. Then as n_0 varies we obtain the values in Table 2.1 for the ratios

$$p = \frac{E(V_{22})}{\text{var}\,(\hat{\beta}_2)}\,, p_w = \frac{E(V_{22,w})}{\text{var}\,(\hat{\beta}_2)} \text{ and } \hat{p} = \frac{E(\hat{V}_{22})}{\text{var}\,(\hat{\beta}_2)}$$

TABLE 2.1—*Comparison of variance estimates in Example 2.1*

n_0	2	3	4	5	6	7	8
p	1.54	1.21	1.13	1.13	1.17	1.30	1.54
p_w	0.78	0.94	1.00	1.00	0.96	0.88	0.61
$\hat{p}$	1.56	1.35	1.16	1.00	0.70	0.73	0.56

We emphasize that the example is purely illustrative; in practice the possibility of different variances would of course by explored.

Example 2.2 Simple linear regression

We take the design and data from Miller's [7] second numerical example. The model is simple linear regression, i.e., $p = 2$ with $x_{i1} \equiv 1$. Data and related computations are given in Table 2.2 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.

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

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

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

Average properties of variance estimates, computed from (2.7) and (2.13), are

$$E(V) = \begin{pmatrix} 1.007 & -0.135 \\ & 0.021 \end{pmatrix},$$

$$E(V_w) = \begin{pmatrix} 0.535 & -0.071 \\ & 0.012 \end{pmatrix}$$

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

The main effect of the weighted jackknife in this design is to deemphasize 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 $t((\mathbf{x}_1^T, Y_1), \cdots, (\mathbf{x}_n^T, Y_n))$. 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 nonlinear case where the parameter of interest is a non-

TABLE 2.2—*Jackknife analysis of artificial simple linear regression data* ($\beta_1 = \beta_2 = \sigma = 1$).

data	x_2	1	3	5	6	6	7	8	8.5	9	10
	Y	0.91	4.24	6.59	8.22	7.53	7.89	10.13	9.25	8.92	11.35
residual	R	-0.99	0.24	0.50	1.08	0.39	-0.29	0.90	-0.50	-1.36	0.03
weight	w	0.51	0.26	0.13	0.10	0.10	0.11	0.14	0.17	0.20	0.29
balanced	P_1	-9.78	2.06	2.00	2.28	1.38	0.73	0.39	1.37	3.00	0.78
pseudo-values	P_2	2.44	0.91	0.95	0.99	1.03	1.02	1.27	0.88	0.47	1.06
weighted	Q_i	-4.94	1.84	1.97	2.28	1.37	0.73	0.41	1.33	2.76	0.79
pseudo-values	Q_2	1.81	0.93	0.95	0.99	1.03	1.02	1.26	0.89	0.53	1.06
discrepancy	c	5.01	0.16	0.32	1.20	0.16	0.09	1.12	0.42	3.69	0.00

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

Estimated s.e.'s: $\sqrt{\hat{V}_{22}} = 0.100$, $\sqrt{V_{w,22}} = 0.102$; $\sqrt{V_{22}} = 0.161$

linear scalar function of the linear model parameter β.

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

$$\hat{\theta} - \hat{\theta}_{-1} = (\hat{\beta} - \hat{\beta}_{-i})^T \nabla f(\hat{\beta}) - \tfrac{1}{2}(\hat{\beta} - \hat{\beta}_{-i})^T \nabla^2 f(\hat{\beta})\,(\hat{\beta} - \hat{\beta}_{-i}) + o_p(n^{-2})\,. \quad (3.1)$$

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

$$\hat{\theta}_w = n^{-1} \sum Q_i = f(\hat{\beta}) - \tfrac{1}{2} \sum (1 - w_j)^{-1} \mathbf{x}_j^T D_0^{-1} \nabla^2 f(\hat{\beta}) D_0^{-1} \mathbf{x}_j R_j^2\,.$$

It follows by Taylor expansion of $f(\hat{\beta})$ and taking expectations that, to order n^{-1},

$$E(\tilde{\theta}_w) = \theta + \tfrac{1}{2} E\{(\hat{\beta} - \beta)^T \nabla^2 f(\beta)\,(\hat{\beta} - \beta)\} - \tfrac{1}{2}\sigma^2 \sum \mathbf{x}_j^T D_0^{-1} \nabla^2 f(\beta) D_0^{-1} \mathbf{x}_j\,. \quad (3.2)$$

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

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

$$E(\tilde{\theta}_w) - \theta = -\tfrac{1}{6} E(e^3) f'''(\beta) \sum x_j^3 / (\Sigma\, x_j^2)^3\,. \quad (3.3)$$

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

$$-\frac{1}{2}\left(\frac{n-1}{n}\right)\sigma^2 \sum (1 - w_j)^{-1} \mathbf{x}_j^T D_0^{-1} \nabla^2 f(\beta) D_0^{-1} \mathbf{x}_j.$$

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

$$E(\tilde{\theta}) - \theta = -\tfrac{1}{2}\sigma^2 \text{ tr } \{D_0^{-1} \nabla^2 f(\beta) D_0^{-1} \sum w_j \mathbf{x}_j \mathbf{x}_j^T\} + O(n^{-2}) \quad (3.4)$$

That is, the n^{-1} bias term is removed, but the remaining bias is of higher order than that for $\tilde{\theta}_w$ *unless* all w_j are of order n^{-1}, which is not automatic.

The last remark is not of purely theoretical interest. From a practical viewpoint it suggests that when x vectors have severe non-uniform dispersion in the observed design, so that the w_i are of different orders of magnitude, then $\tilde{\theta}_w$ is superior to $\tilde{\theta}$.

Example 3.1 Ratio parameter in simple linear regression

We consider the problem in Example 2.2 with $\theta = \beta_1/\beta_2$ as the parameter of interest. Required computations are given in Table 3.1. Theoretical approximations for bias and standard error of $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_2$ can be obtained by the delta method: with $\hat{\beta}_i = \beta_i + \delta_i$ ($i = 1, 2$) write

$$\hat{\theta} = (\beta_1 + \delta_1)\,(\beta_2 + \delta_2)^{-1} \approx (\beta_1 + \delta_1) \Big/ \left\{\left(1 - \frac{\delta_2}{\beta_2} + \frac{1}{2}\,\frac{\delta_2^2}{\beta_2^2}\right)\right\} \beta_2,$$

so that

$$E(\hat{\theta}) - \theta \approx \tfrac{1}{2}\beta_1 \text{ var } (\hat{\beta}_2)/\beta_2^3 - \text{cov } (\hat{\beta}_1, \hat{\beta}_2)/\beta_2^2 \quad (3.5)$$

and

$$\text{var } (\hat{\theta}) \approx \frac{1}{\beta_2^2} \text{ var } (\hat{\beta}_1) + \frac{\beta_1^2}{\beta_2^4} \text{ var } (\hat{\beta}_2) - \frac{2 \text{ cov } (\hat{\beta}_1, \hat{\beta}_2)}{\beta_2^3}\,. \quad (3.6)$$

The approximate bias (3.5) is 0.1 in the example, which makes $\tilde{\theta}_w$ seem reasonable, as does the comparison of standard errors: (3.6) gives estimated standard error of 0.93, quite close to $\sqrt{V_w}$.

For this same design we have obtained simulation

TABLE 3.1—*Jackknife analysis of β_1/β_2 from regression data in Table 2.1.*

x_2	1	3	5	6 ·	6	7	8	8.5	9	10
w	0.5087	0.2603	0.1260	0.1017	0.1017	0.1060	0.1389	0.1660	0.2003	0.2903
P	-12.368	2.055	1.975	2.215	1.327	0.719	0.184	1.426	3.168	0.729
Q	- 6.381	1.833	1.941	2.212	1.326	0.720	0.212	1.381	2.906	0.747

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

balanced jackknife: $\tilde{\theta} = 0.143$, est. s.e. $= \sqrt{V} = 1.417$

weighted jackknife: $\tilde{\theta}_w = 0.690$, est. s.e. $= \sqrt{V_w} = 0.825$

results for $\hat\theta$, $\tilde\theta$ and $\tilde\theta_w$ using 10,000 samples in which the model was $\beta_1 = 1, \beta_2 = 2, \sigma^2 = 1$. The errors were pseudo-random normal. Table 3.2 contains a summary of the simulations, from which the superiority of $\tilde\theta_w$ and V_w is clear.

4. ROBUST REGRESSION USING PSEUDO-VALUES

In a recent article Hinkley [3] 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 [2] has discussed the use of $\hat\beta - \hat\beta_{-i}$ in exhibiting important large residuals, the implicit purpose being to isolate data points that might be de-emphasized or omitted in a re-fit of the model. These ideas clearly suggest possible methods of robust regression based on the pseudo-values Q_i. Only a brief discussion will be given here.

By analogy with the classical location problem, where robust estimators (Huber, [5]) 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 the influence of extreme values of Q_i. Of course Q_i is a vector here, so that we have the choice of working on coordinates separately or simultaneously. In the latter case we might work in terms of the squared standardized residual

$$c_i = n^{-1}(Q_i - \hat\beta)^T D_0 (Q_i - \hat\beta) = n_i w_i R_i^2 \quad (i = 1, \cdots, n)\,, \qquad (4.1)$$

extreme values indicating points to be trimmed from the analysis; see Cook [2].

On the whole it seems best to define robust estimates separately for each component. If we denote components of Q_i by Q_{ji} with corresponding order statistics $Q_{j(i)}$ $(j = 1, \cdots, p;\ i = 1, \cdots, n)$, then two standard forms of robust estimates are

$$\beta_j^* = \sum_{i=1}^{n} h_i Q_{j(i)}, \quad h_i = h_{n-i+1}, \quad \sum h_i = 1 \qquad (4.2)$$

and

$$\text{solution to } \sum_{i=1}^{n} \psi(Q_{j(i)} - \beta_j^*) = 0,\ \psi(-u) = -\psi(u); \qquad (4.3)$$

see Huber [5]. Either form of estimate is unbiased if the Q_i are symmetrically distributed, which is implied by symmetrically distributed errors e_j. Theoretical study of such estimators is complicated by the correlation between the Q_i, but a similar situation has been handled by Hinkley [3].

The estimators (4.2) and (4.3) differ from those proposed by Huber [5], Andrews [1], and others in that "harmless' large residuals are ignored: Q_i will not be extreme if w_i is small. One possible modification of (4.2) and (4.3) is to substitute β_j^* for $\hat\beta_j$ in the definition of Q_i, making the estimation interative. This procedure will be reported on elsewhere.

Example 4.1 Trimmed mean pseudo-value

A simple numerical illustration of the above ideas may be obtained from the data of Example 2.1. Relevant quantities are given in the last three rows of Table 2.2. Note that the first and ninth observations have large values of c_i (see (4.1)): these observations are very influential in the regression fit. The 10% trimmed mean estimates ((4.2) with $h_1 = h_{10} = 0, h_2 = \cdots = h_9 = \frac{1}{8}$) of β_1 and β_2 are $\beta_1^* = 1.34$ and $\beta_2^* = 1.02$. Corresponding calculations for θ are given in Table 3.1, from which we obtain the 10% trimmed mean estimate $\theta^* = 1.29$.

5. PHOTOGRAPHIC DATA EXAMPLE

An interesting real example with severe imbalance is given by the data in Table 5.1, kindly provided by Lincoln Moses. The data comes from measurements of enlarger magnification m and object-to-image distance d for a lens of unknown focal length ϕ. 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

$$y \doteq \alpha + \phi(m + n^{-1}) = \beta_1 + \beta_2 x\,,$$

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

$$\hat\beta_1 = -5.272,\ \hat\beta_2 = 18.870,\ \sum r_i^2 = 3.0729$$

$$\bar x = 2.385,\ \sum (x_j - \bar x)^2 = 20.49,\ \text{est. s.e. } (\hat\beta_2) = 0.05\,.$$

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 the usual estimated standard error for $\hat\beta_2$ will be too

TABLE 3.2—*Simulation results for jackknife analysis of $\theta = \beta_1/\beta_2$ when $\beta_1 = 1, \beta_2 = 2, \sigma^2 = 1$; design as in Table 3.1; 10000 cases*

Statistic	$\hat\theta$	$\tilde\theta$	$\tilde\theta_W$	V	V_W
mean	0.526	0.486	0.501	0.301	0.142
variance	0.198	0.220	0.195	0.110	0.016

small. This can be seen from the jackknife estimate V_w, which gives estimated s.e. $(\hat{\beta}_2) = 0.12$.

Since the weights w are also large for $x > 2.5$, the weighted residuals $c = nwr^2$ are very large in that range. The largest ten values are asterisked. Weighted pseudo-value components for β_2 defined by $q_i = Q_{2i} - \hat{\beta}_2$ are given in the final column of Table 5.1.

The robust estimation approach of Section 4 essentially treats the second pseudo-value components q_{2i} as observations on a location model with mean β_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 $\beta_2{}^* = 18.935$. The standard error computed from the winsorized sample variance of the pseudo-values is 0.02.

In this particular example the estimates of β_2 do not differ much, but use of the jackknife and its pseudovalues is definitely of value in assessing precision. The high variability of Q_{2i} for $x > 2.5$ leads one to believe that the trimmed mean estimate $\beta_2{}^*$ is more precise than $\hat{\beta}_2$, as the estimated standard errors suggest.

An alternative analysis is two-stage weighted least squares in which the first ten observations ($x > 2.5$) are assumed to have error variance $\sigma_1{}^2$ different from the error variance $\sigma_2{}^2$ for the last forty-six observations.

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 will be of interest, but numerical comparison with other recent methods 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 equalize information content. Initial results for this are complicated, and in any event grouping is likely to lose information; see Hinkley [4].

6. ACKNOWLEDGEMENTS

The author is grateful to R. G. Miller, Jr. and Lincoln E. Moses for providing numerical examples, to Elizabeth Hinkley for programming some of the calculations, and to the referees for helping to clarify the presentation of the material.

TABLE 5.1—*Regression analysis of photographic enlarger data*

x	y	r	w	c	$Q_2-\hat{\beta}$
5.200	92.05	-.803	0.405	14.62*	-6.180**
4.250	75.00	.073	.188	.06	0.374
3.851	67.51	.113	.123	.09	0.452**
3.633	63.50	.216	.094	.25*	0.738**
3.320	57.69	.313	.061	.33*	0.799**
3.072	53.50	.803	.041	1.48*	1.508**
2.900	49.38	-.072	.031	.01	-0.101**
2.874	49.26	.299	.030	.15*	0.400
2.672	45.64	.491	.022	.30*	0.385
2.634	45.05	.618	.021	.45*	0.421**
2.485	41.78	.159	.018	.03	0.044
2.392	39.81	-.056	.018	.00	-0.001
2.368	39.38	-.033	.018	.00	0.001
2.285	37.50	-.347	.018	.12*	0.094
2.249	37.13	-.037	.019	.00	0.014
2.225	36.56	-.154	.019	.03	0.067
2.188	35.94	-.076	.020	.01	0.041
2.166	35.50	-.101	.020	.01	0.060
2.129	34.84	-.063	.021	.00	0.044
2.118	34.55	-.145	.021	.03	0.106
2.109	34.49	-.035	.022	.00	0.027
2.091	34.19	.004	.022	.00	-0.003
2.083	34.06	.025	.022	.00	-0.021
2.050	33.25	-.162	.023	.03	0.148
2.040	33.13	-.093	.024	.01	0.088
2.028	32.78	-.217	.024	.06	0.211
2.026	32.93	-.029	.024	.00	0.029
2.022	32.91	.026	.024	.00	-0.026
2.011	32.66	-.016	.025	.00	0.016
2.003	32.56	.035	.025	.00	-0.036
2.001	32.51	.023	.025	.00	-0.024
2.000	32.61	.142	.025	.03	-0.149**
2.002	32.44	-.066	.025	.00	0.069
2.002	32.13	-.376	.025	.20*	0.393
2.009	32.57	-.068	.025	.00	0.070
2.011	32.63	-.046	.025	.00	0.047
2.020	32.79	-.056	.024	.00	0.056
2.033	33.06	-.031	.024	.00	0.030
2.050	33.88	.468	.023	.29*	-0.428**
2.054	33.50	.013	.023	.00	-0.011
2.091	34.13	-.056	.022	.00	0.045
2.100	34.38	.245	.022	.00	-0.019
2.107	34.44	-.048	.022	.00	0.036
2.129	34.84	-.063	.021	.00	0.044
2.140	35.00	-.110	.021	.01	0.074
2.150	35.25	-.049	.021	.00	0.031
2.167	35.63	.010	.020	.00	-0.006
2.195	36.11	-.038	.020	.00	0.020
2.213	36.63	.142	.019	.02	-0.067**
2.256	37.00	-.299	.019	.09	0.105
2.288	38.00	.097	.018	.01	-0.026
2.318	38.26	-.209	.018	.04	0.038
2.362	39.25	-.050	.018	.00	0.003
2.391	39.88	.033	.018	.00	0.001
2.478	41.38	-.108	.018	.01	-0.028
2.500	41.89	-.014	.019	.00	-0.004

* ten largest values of c
** five smallest and five largest pseudo-values

7. 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

$$E_G\left\{\begin{pmatrix}\mathbf{x}\\ Y\end{pmatrix}(\mathbf{x}^T, Y)\right\} = \begin{pmatrix}\Sigma(G) & \gamma(G)\\ & \tau(G)\end{pmatrix}$$

and define $\beta(G) = \Sigma^{-1}(G)\,\gamma(G)$. The the influence function of β at $(\mathbf{x}^T, y)$ is

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

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 $\hat{\Sigma} = n^{-1}A^TA$.

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

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

which implies that in Section 3

$$Q_i = \hat{\theta} + \hat{I}(\theta; \mathbf{x}_i, Y_i) + O(n^{-1})\,;$$

the remainder term involves second von Mises derivatives and may be used to obtain (3.3). See Hinkley [3].

Lemma 2 (Consistency of variance estimate)

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

$$n^{-1}\, A^TA \to \sum \text{p.d.}, n^{-1}\, A^T\Lambda\, A \to \Gamma \text{ p.d.} \quad (n \to \infty)$$

and if $E(e_i^4)$ is uniformly bounded,

(i) n var $(\hat{\beta}) \to \Sigma^{-1}\Gamma\Sigma^{-1}$

(ii) $nV_w = \dfrac{n^2}{n - p}\,(A^TA)^{-1}A^TLA(A^TA)^{-1} \to \Sigma^{-1}\Gamma\Sigma^{-}$

Proof. Part (i) follows by assumption, since var $(\hat{\beta}) = (A^TA)^{-1}A^T\Lambda A(A^TA)^{-1}$. Part (ii) follows by using a minor variation of the proof of Lemma 3.4 in Miller [7] to establish $n^{-1}A^TLA \to \Gamma$.

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

REFERENCES

[1] ANDREWS, D. F. (1974). A robust method for multiple linear regression. *Technometrics, 16,* 523–31.

[2] COOK, R. D. (1977). Detection of influential observations in linear regression. *Technometrics, 19,* 15–18.

[3] HINKLEY, D. V. (1976). Robust jackknife correlation. Stanford Univ., Biostat. Tech. Rep., 19.

[4] HINKLEY, D. V. (1977). Jackknife confidence limits using Student-t approximations. *Biometrika, 64,* 21–28.

[5] HUBER, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist., 43,* 1041–67.

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

[7] MILLER, R. G. (1974b). An unbalanced jackknife. *Ann. Statist., 2,* 880–91.

[8] QUENOUILLE, M. H. (1956). Notes on bias in estimation. *Biometrika, 43,* 353–60.