ESTIMATING THE STANDARD ERROR OF ROBUST REGRESSION ESTIMATES

Simon J. Sheather
University of Melbourne

and

Thomas P. Hettmansperger*
Penn State University
ESTIMATING THE STANDARD ERROR
OF ROBUST REGRESSION ESTIMATES

Simon J. Sheather
University of Melbourne
and
Thomas P. Hettmansperger*
Penn State University

*The work of this author was partially supported by ONR Contract N00014-80-C0741.
Abstract

In this paper we provide a review of the available methods for estimating the standard error of $M$- and $\ell_1$-estimates in regression. In the case of $M$-estimates, we show how to use MINITAB to compute these estimates along with estimates of their standard errors.

Key words: $M$-estimate, $\ell_1$-estimate, robust regression, standard error, bootstrap.
1. Introduction

Over the last two decades there has been much interest in the statistical literature in alternative methods to least squares for fitting equations to data. During this time a large number of estimates of regression coefficients have been proposed that are not unduly affected by a small percentage of the data (so-called robust estimates). Although the robustness properties of these estimates have been studied in great detail, little attention has been paid to the problem of estimating the asymptotic covariance matrices of these estimates. Such estimates are necessary if inferences are to be made about the unknown regression parameters.

In this paper we provide a brief description of two popular robust regression estimates, namely M- and T-estimates. We review the available methods for estimating the asymptotic covariance matrices of each of these estimates. In the case of M-estimates, we show how to use MINITAB to compute the robust estimates along with an estimate of their asymptotic covariance matrix. Finally, the different robust estimates and their estimated covariance matrices are compared via an example.


Consider the linear regression model specified by

\[ \mathbf{y} = \mathbf{X}\beta + \xi \]

where \( \mathbf{y}^T = (y_1, y_2, \ldots, y_n) \), \( \mathbf{X} \) is an \( n \times (p+1) \) full rank matrix of known constants, \( \beta^T = (\beta_0, \beta_1, \ldots, \beta_p) \) a vector of unknown
parameters, and \( \xi^T = (\xi_1, \xi_2, \ldots, \xi_n) \) a vector of i.i.d. random errors from a distribution with median 0 and density \( f \).

2.1. M-estimates.

Throughout this section we shall assume that the distribution of the random errors is symmetric.

An M-estimate is defined as the solution \( \hat{\xi}_M \) of the vector of estimating equations

\[
\sum_{i=1}^{n} \eta(x_i, r_i) x_i = 0
\]

(1)

where \( r_i = y_i - x_i^T \hat{\beta} \) and \( x_i^T \) is the \( i \)th row of \( X \). We will consider only \( \eta(x, r) \) functions of the form

\[
\eta(x, r) = \nu(x) \frac{r}{\sigma \nu(x)}
\]

(2)

where \( \sigma \) is a scale factor that may be estimated from the data, \( \nu(x) \) is a nonnegative weight function, and

\[
\nu_c(t) = \begin{cases} 
  t & \text{if } |t| \leq c \\
  \text{csign}(t) & \text{if } |t| > c
\end{cases}
\]

which is known as Huber's \( \psi \) function. This form of \( \eta(x, r) \), for the special case with \( \nu(x_i) = 1 - h_i \), \( h_i \) the leverage of \( x_i \), was proposed in Handschin, Schweppe, Kohlas and Fiechter (1975) and is referred to as Schweppe's form. It is discussed by Hampel (1978, Section 6) where he says that this is the most intuitive way to bound influence in both the residual and design space. Huber (1981, Section 7.9) recommends this choice for \( \eta \), and again, in Huber (1983, Section 6), he reaffirms this recommendation. If we
take $v(x) = 1$, we get Huber's M-estimate which has unbounded influence in the design points. If, in addition, we specify a large value for $c$, the resulting estimate is essentially least squares. For a more complete description of M-estimates see Hampel et al (1986, Chapter 6) and Hettmansperger (1987).

The following form of $\eta(x, r)$ can be used in a weighted least squares algorithm to compute $\hat{\beta}_M$

$$
\eta(x, r) = w(x, r)r/\sigma
$$

(3)

where

$$
w(x, r) = \min\{1, c|v(x)|/|r|\}.
$$

In the Appendix, for a particular choice of $v(x)$, we give the MINITAB commands to compute a 1-step version of $\hat{\beta}_M$ using weighted least squares to solve (1) with weights given by (3).

Maronna and Yohai (1981) show, under mild regularity conditions, that $n^{1/2}(\hat{\beta}_M - \beta)$ is asymptotically normal with mean $0$ and covariance matrix $U = \Sigma^{-1}Q\Sigma^{-1}$ where

$$
M = \frac{3}{\sigma^2} E[\eta(x, r)] = E[\frac{1}{\sigma} \psi'(\frac{r}{v(x)})v(x)]
$$

$$
Q = E[\eta^2(x, r)x x^T] = E[w^2(x, r) \frac{r^2}{\sigma^2} x x^T]
$$

$$
\psi'(t) = \frac{dv(t)}{dt} = \begin{cases} 1 & \text{if } |t| \leq c \\ 0 & \text{if } |t| > c. \end{cases}
$$

Obvious estimates of $M$ and $Q$ are given by

$$
M = \frac{1}{n\sigma} \sum_{i=1}^{n} \psi'(\frac{r_i}{v(x_i)})x_i x_i^T
$$

and
\[ \hat{Q} = \frac{1}{n \sigma^2} \sum_{i=1}^{n} w^2(x_i, r_i) r_i 2 x_i x_i^T, \]
respectively, where \( r_i = x_i - \hat{x}_i \hat{\beta}_M \). Thus the asymptotic
covariance matrix of \( \hat{\beta}_M \), \( V_M = n^{-1} U \) can be estimated by
\[ \hat{V}_M = \left( \frac{n}{n(n-p-1)} \right) n^{-1} \hat{V}^{-1} Q_M^{-1}. \]
The bias correction factor \( n/(n-p-1) \) was recommended by Huber
(1981, page 173) and is there to recapture the classical formula
in the least squares case (\( v(x) = 1 \) and \( c = \infty \)). If we let \( d_i = \frac{r_i}{\sigma v(x_i)} \) and \( w_i = w(x_i, r_i) \) then \( V \) is given by
\[ \hat{V}_M = \frac{n}{n-p-1} \left( \sum_{i=1}^{n} \psi'_c (d_i) x_i x_i^T \right) - 1 \left\{ \sum_{i=1}^{n} w_i^2 r_i 2 x_i x_i^T \right\}. \]
To implement \( \hat{\beta}_M \) the user has to decide on \( v(x) \), \( \sigma \) and \( c \).
Welsch (1980) proposed the following choices for \( v(x) \) and \( \sigma \)
\[ v(x) = \frac{(1-h_i)}{\sqrt{h_i}}, \]
\[ \sigma = S(i), \]
where \( h_i \) is the leverage of \( x_i \), defined to be the ith diagonal
element from the least squares projection matrix and \( S(i) \) is
the root mean square error from the least squares fit with the
ith case deleted. These choices can be motivated as follows.
First notice that in this case
\[ d_i = DFITS_i = t_i (h_i/(1-h_i))^{1/2} \]
where \( t_i \) is the ith studentized or t-residual and is given by
\[ t_i = r_i / \{s(i)(1-h_i)^{1/2}\}. \]
deleted. It is therefore an important diagnostic quantity for determining cases with large influence on the least squares fit. Referring back to (3), we see that as long as \( |d_i| \leq c \) the \( i \)th case is not downweighted in the robust fit; otherwise, it is downweighted in proportion to the excess of \( |d_i| \) over \( c \). For the choice of \( c \), we take
\[
c = 2\left(\frac{(p+1)/n}{1}\right)^{1/2}
\]
as recommended by Belsley, Kuh and Welsch (1980, page 28) for diagnostic purposes in conjunction with least squares.

In the Appendix, we present the MINITAB commands to compute \( \hat{\xi}_M \) and \( \hat{V}_M \) based on the above choices of \( v(x) \), \( \sigma \) and \( c \). Computing \( \hat{V}_M \) using MINITAB can be made easier by reexpressing \( \hat{V} \). First recall that \( \sum_{i=1}^{n} a_i x_i^T = X'AX \) where \( A \) is a diagonal matrix with diagonal elements \( a_1, \ldots, a_n \). Then letting \( D_1 = \text{diagonal} \{ \psi_C'(d_1), \ldots, \psi_C'(d_n) \} \) and \( D_2 = \text{diagonal} \{ w_1^2 r_1^2, \ldots, w_n^2 r_n^2 \} \) we can reexpress \( \hat{V}_M \) as
\[
\hat{V}_M = \frac{n}{n-p-1} (X'D_1X)^{-1}(X'D_2X)(X'D_1X)^{-1}.
\]

2.2. \( \ell_1 \)-estimates.

The \( \ell_1 \)-estimate \( \hat{\xi}_{\ell_1} \) is defined as the value of \( \xi \) that minimizes
\[
\sum_{i=1}^{n} |\hat{y}_i - x_i^T \xi|.
\]
For a review of the historical development of \( \ell_1 \)-estimates the reader is referred to Bloomfield and Steiger (1983).

Bassett and Koenker (1978) show, under mild regularity
conditions, that $n^{1/2}(\hat{\tau}_1 - \tau)$ is asymptotically normal with mean 0 and covariance matrix $U = \tau^2(X'X)^{-1}$ where $\tau = 1/(2f(0))$.

The problem of estimating $\tau$ has been, extensively and almost exclusively, studied in the one-sample location model setting, which occurs when $p = 0$ and $X$ consists of a column of ones. We now review the results from this setting. Let $Y(1) \leq Y(2) \leq \ldots \leq Y(n)$ denote the order statistics and $\hat{\theta}$ the sample median of $Y_1, Y_2, \ldots, Y_n$.

For the case that $n$ is odd (i.e. $n=2m+1$), Maritz and Jarrett (1978) and Efron (1979) independently proposed the following estimator of $\tau^2$

$$\hat{\tau}^2 = \frac{n}{2} \sum_{i=1}^{n} w_i (Y_{(i)})^2 - \left( \frac{n}{2} \sum_{i=1}^{n} w_i Y_{(i)} \right)^2$$

where

$$w_i = \frac{n}{(m!)^2} \int_{(i-1)/n}^{i/n} u^m (1-u)^m du.$$ 

The following related estimate was proposed by Sheather (1986)

$$\hat{\tau}^2 = \frac{n}{2} \sum_{i=1}^{n} w_i^* (Y_{(i)})^2 - \left( \frac{n}{2} \sum_{i=1}^{n} w_i^* Y_{(i)} \right)^2 \quad (4)$$

where

$$w_i^* = J(i-1/2)/n \sum_{k=1}^{n} J(k-1/2)/n$$

and

$$J(u) = \frac{n}{(m!)^2} u^m (1-u)^m.$$ 

Under the conditions that $f$ is positive and continuous in a neighborhood of 0 and $E[\log(1+|\xi|)] < \infty$, Babu (1986) has shown that $\hat{\tau}^2_{MJE} \to \tau^2$ almost surely as $n \to \infty$.

In a large Monte Carlo study, Sheather and McKean (1987)
compared various estimates of \( \tau \) in terms of their ability to studentize \( \hat{\theta} \). They found both \( \hat{\tau}_{\text{MJE}}^2 \) and \( \hat{\tau}_s^2 \) performed well with little difference between them.

The other estimator of \( \tau \) that performed well in the Monte Carlo study of Sheather and McKean (1987) was first proposed by Siddiqui (1960). This estimator of \( \tau \) is given by

\[
\hat{\tau}_d = n\{Y([n/2]+d) - Y([n/2]-d+1)\}/(4d)
\]

where \( d = o(n) \). Bloch and Gastwirth (1968) found that the value of \( d \) that minimized the first order term in the mean square error is \( O(n^{4/5}) \). In another Monte Carlo study, McKean and Schrader (1984) found that the tests resulting from studentizing \( \hat{\theta} \) by \( \hat{\tau}_d/n^{1/2} \) with \( d = O(n^{4/5}) \) were very liberal. Following a proposal by Lehmann (1963), McKean and Schrader (1984) found that \( d = O(n^{1/2}) \) was an improvement over \( d = O(n^{4/5}) \). Recently, Hall and Sheather (1987) have developed an edgeworth expansion for the studentized version of \( \hat{\theta} \). They found that the value of \( d \) that minimizes the first order correction term in this expansion is \( O(n^{2/3}) \). Unfortunately, the constant involved depends on the underlying density in a complicated manner, making it difficult to estimate in practice. A number of other estimators of \( \tau \) exist. For a review of such estimators in the one-sample setting see Sheather (1987).

We now return to the more general problem of estimating \( \tau \) based on the residuals \( r_i = y_i - \hat{x}_i \). Since \( p + 1 \) of these residuals will be exactly zero, McKean and Schrader (1987), following a suggestion by Hill and Holland (1977), suggest that these zero...
residuals be eliminated when estimating \( \tau \). Let \( n^* = n-p-1 \) and \( r^* \leq r^{(1)} \leq \ldots \leq r^{(n-p-1)} \) denote the ordered remaining residuals. Then we recommend the estimate of Sheather (1986) given by (4) with \( Y_{(i)} \) replaced by \( r_{(i)}^* \) and \( n \) replaced by \( n^* \). The resulting estimate of the asymptotic covariance matrix of \( \hat{\beta}_{\xi_1} \), \( V_{\xi_1} = n^{-1} C \) is given by

\[
V_{\xi_1} = \left( \frac{n}{n-p-1} \right) n^{-1} \text{Cov}(X'X)^{-1}
\]

where again the factor \( n/(n-p-1) \) acts as a bias correction.

3. Example.

The data in Table 1 are taken from Simkin (1978) and are annual rates of growth of average prices in the main cities of Free China from 1940 to 1946. In this example, interest is clearly in the rate of change of the growth in prices which corresponds to \( \beta_1 \) in the model below.

<table>
<thead>
<tr>
<th>Year (x)</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>Growth of prices (y)</td>
<td>1.62</td>
<td>1.63</td>
<td>1.90</td>
<td>2.64</td>
<td>2.05</td>
<td>2.13</td>
<td>1.94</td>
</tr>
</tbody>
</table>

We considered the simple linear regression model

\[ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

and calculated a one-step version of the \( M \)-estimate of \( \hat{\beta}_1(\hat{\beta}_{1M}) \) proposed by Welsch (1980) along with the \( \xi_1 \)-estimate of \( \hat{\beta}_1(\hat{\beta}_{1\xi_1}) \).

These appear in Table 2 along with estimates of their standard error.
The standard error estimates denoted by $\hat{se}(\hat{\beta}_{1M})$ and $\hat{se}(\hat{\beta}_{1\ell_1})$ were obtained by taking the square root of the second diagonal element of $\hat{V}_M$ and $\hat{V}_{\ell_1}$, respectively. As a check on the accuracy of these estimates, we also calculated estimates of the standard errors of $\hat{\beta}_{1M}$ and $\hat{\beta}_{1\ell_1}$ using the bootstrap. A description of the bootstrap algorithm as it is applied to residuals in the regression setting can be found in Efron and Tibshirani (1986). In the case of $\ell_1$-estimates, the bootstrap algorithm was applied to the five residuals that remained after the two that were identically zero were eliminated. For the $M$- and $\ell_1$-estimates, 601 and 1000 repetitions of the bootstrap algorithm were performed, respectively. The standard error estimates $\hat{se}_B(\hat{\beta}_{1M})$ and $\hat{se}(\hat{\beta}_{1\ell_1})$ were each calculated as 0.75 times the interquartile range of the bootstrap estimates of $\beta_1$. This function of the interquartile range was used in preference to the standard deviation because both histograms of bootstrap estimates, although normal in shape, had many more outliers than one would expect from a normal distribution. Note for both the $M$- and $\ell_1$-estimate of $\beta_1$ the close agreement between the estimates of standard error obtained from $\hat{\beta}$ and the bootstrap. For the purposes of comparison we also report in Table 2 the least squares estimate of $\beta_1(\hat{\beta}_{1LS})$ and its estimated standard error. The efficiency gain by using the $M$- or $\ell_1$-estimate of $\beta_1$ instead of the least squares estimate is quite striking.
Table 2

\[
\hat{\beta}_{1M} = 0.075 \quad \hat{\beta}_{1L} = 0.102 \quad \hat{\beta}_{1LS} = 0.075
\]
\[
\hat{se}(\hat{\beta}_{1M}) = 0.033 \quad \hat{se}(\hat{\beta}_{1L}) = 0.049 \quad \hat{se}(\hat{\beta}_{1LS}) = 0.063
\]
\[
\hat{se}_B(\hat{\beta}_{1M}) = 0.033 \quad \hat{se}_B(\hat{\beta}_{1L}) = 0.045
\]

Appendix

Least Squares:

NAME 'Y', 'X1', ..., 'XP', 'SRI', 'YHI', 'TR', 'DF', 'HI'
REGR 'Y' on p in 'X1', ..., 'XP', put std resids in 'SRI', fit in 'YHI';
TRESIDS in 'TR';
DFITS in 'DF';
HI in 'HI'.
PRINT 'Y' 'X1' ... 'XP' 'YHI' 'TR' 'DF' 'HI'
PLOT 'TR' vs 'YHI'

Robust:

NAME 'WEL' 'W' 'RESIDS' 'SR2' 'YH2'
LET K1 = 2*SQRT((p+1)/n))
LET 'WEL' = K1/ABSO('DF')
RMIN 1 'WEL' into 'W'
REGR 'Y' on p in 'X1' ... 'XP' 'SR2' 'YH2';
WEIGHTS in 'W';
RESIDS in 'RESIDS'.
PRINT 'Y' 'YH2' 'W'
AVERAGE 'W'

V:

NAME 'DIFF' 'IND' 'WT'
LET 'DIFF' = K1 - ABSO('DF')
LET 'IND' = .5*(SIGN('DIFF')+1)
REGR 'Y' on p in 'X1' ... 'XP';
WEIGHTS in 'IND';
XPXINV in M1.
\[
\text{LET 'WT'} = ('W'**2)*('RESIDS'**2)
\]
\[
\text{LET 'WT'} = ('W'**2)*('RESIDS'**2)
\]

\[
\text{LET 'WT'} = ('W'**2)*('RESIDS'**2)
\]

\[
\text{LET 'WT'} = ('W'**2)*('RESIDS'**2)
\]

\[
\text{LET 'WT'} = ('W'**2)*('RESIDS'**2)
\]
REGR 'Y' on p in 'XL' ... 'XP';
WEIGHTS in 'WT';
XPXINV M2.

INVERSE M2 into M3
MULT M1 by M3 into M4
MULT M4 by M1 into M5
LET K2 = n/(n-p-1)
MULT K2 by M5 into M6

References


In this paper we provide a review of the available methods for estimating the standard error of M- and \( \ell_1 \)-estimates in regression. In the case of M-estimates, we show how to use MINITAB to compute these estimates along with estimates of their standard errors.
END
5-87
DTIC