M-ESTIMATES FOR THE HETEROSEDASTIC
LINEAR MODEL

David Ruppert*
and
Raymond J. Carroll**
University of North Carolina at Chapel Hill

We treat the linear model \( Y_i = C_i^T \beta + Z_i \) where \( C_i \) is a known vector, \( \beta \) is an unknown parameter, and \( \text{Var } Z_i \) is a function of \( |C_i^T \beta| \) which is known, except for a parameter \( \theta \). For simultaneous M-estimates, \( \hat{\beta} \) and \( \hat{\theta} \), we show that \( \sqrt{n} (\hat{\theta} - \theta) = O_p (\sqrt{n} \hat{\sigma}) \), and find the limit distribution of \( N^\frac{1}{2} (\hat{\beta} - \beta) \). For the special case of least squares estimation, this limit distribution is the same as the limit distribution of the weighted least squares using the weights, \( w_i = (\text{Var } Z_i)^{-1} \), and in general the distribution is that of a "weighted M-estimate" using these weights. Moreover, the covariance matrix of the limit distribution can be consistently estimated, so large sample confidence ellipsoids and tests of hypotheses concerning \( \beta \) are feasible.

AMS 1970 subject classifications
Primary 62J05; Secondary 62G35.

Key words and phrases. Weighted M-estimates, weighted least squares, estimated weights, linear regression in scale, robustness, asymptotic representation, asymptotically normal, confidence ellipsoids.

*Supported by National Science Foundation Grant NSF MCS78-01240

**Supported by Air Force Office of Scientific Research under Grant AFOSR-75-2796
1. **Introduction.** Consider the linear model

\[ Y_i = \tau_i + \sigma_i e_i \]

and

\[ \tau_i = C_i^T \beta, \text{ for } i = 1, \ldots, N, \]

where \( e_1, \ldots, e_N \) are i.i.d. with distribution \( F, C_i \) (\( = C_i \)) are known elements in \( \mathbb{R}^p \), \( \beta \) is a vector parameter in \( \mathbb{R}^p \), and \( \sigma_1, \ldots, \sigma_N \) are constants, which express the possible heteroscedasticity of the model. If \( F \) were a normal distribution function and the \( \sigma_i \) were known then the minimum variance unbiased estimator of \( \beta \) would be the weighted least squares estimator \( \hat{\beta}_W \), which is the solution to

\[
(1.1) \quad \sum_{i=1}^{N} \rho((Y_i - C_i^T \hat{\beta}_W) / \sigma_i) = \text{minimum},
\]

where \( \rho(x) = x^2 \). If \( F \) has heavy tails compared with the Gaussian distribution or gross errors are possibly present in the data, then one may wish to replace \( \rho(x) = x^2 \) with a function such that \( \psi = \rho' \) is bounded. Such estimators have been called M-estimators by Huber (1964, 1973), because they are generalizations of maximum likelihood estimators. If the \( \sigma_i \) are not known it may still be possible to estimate them. For example, Fuller and Rao (1978) consider the case where the \( Y_i \) occur in groups for which \( \sigma_i \) is constant, and Box and Hill (1974) assume that

\[
(1.2) \quad \sigma_i(\theta) = \sigma_i = \sigma |\tau_i|^{\theta}
\]

for an unknown parameter \( \theta \) and estimate both \( \theta \) and \( \beta \) by Bayesian methods.

Both Fuller and Rao and Box and Hill treat only Gaussian errors.

Anscombe (1961) has proposed tests of heteroscedasticity, and Bickel (1978) has developed robust versions of these, but neither has considered modifying the estimate of \( \beta \) if the null hypothesis of homoscedasticity is rejected.

In many empirical studies, one finds that the dispersion of the residuals
increases with the magnitude of the fitted values, so that it is reasonable to assume that \( \sigma_i = \sigma(|\tau_i|) \) for some nondecreasing function \( \sigma \). While the power family (1.2) may be sufficiently rich to model this phenomenon (although it has obvious difficulties in practice when \( \tau_i = 0 \)), we prefer to study the general class,

\[
\sigma_i = \exp(h(\tau_i)^T \theta),
\]

where \( \theta \) is a parameter in \( \mathbb{R}^q \) and \( h \) is a known function from \( \mathbb{R}^1 \) to \( \mathbb{R}^q \).

Notice that choosing \( h_1(\tau_i) = 1 \) (\( h_1 \) is the first coordinate of \( h \)), \( \exp(\theta_1) \) becomes a scale parameter; as a result, our estimate of \( \theta \) will be scale equivariant regardless of the choice of \( \rho \). Two choices of \( h \) which have no practical difficulties when \( \tau_i = 0 \) are \( h(\tau) = (1, \log(1 + |\tau|)) \) and \( h(\tau) = (1, |\tau|) \).

To motivate our method, suppose \( F \) is standard normal so that the log-likelihood is

\[
L(Y; \theta, \beta) = -\frac{1}{2} \sum_{i=1}^{N} \left\{ \log 2\pi + \log \sigma_i^2 + \left( (Y_i - \tau_i)/\sigma_i \right)^2 \right\}.
\]

If \( \theta \) were known, (1.1) would yield the MLE for \( \beta \), while if \( \beta \) were known, the MLE of \( \theta \) would solve

\[
\sum_{i=1}^{N} \left\{ \left( (Y_i - \tau_i)/\sigma_i(\theta) \right)^2 - 1 \right\} h(\tau_i) = 0.
\]

A reasonable computational alternative to solving (1.1) and (1.4) simultaneously might consist of (i) obtaining a preliminary estimate of \( \beta \) (such as the least squares estimate) and hence estimates for \( \tau_1, \ldots, \tau_n \) (ii) solve (1.4) using these estimates, thus obtaining estimates of \( \sigma_1, \ldots, \sigma_n \), which (iii) are used to solve (1.1).

We thus suggest the following procedure. First, a preliminary estimate \( \hat{\beta}_0 \) of \( \beta \) is calculated and assumed to satisfy

\[
(\hat{\beta}_0 - \beta) = O_p(N^{-\frac{k}{2}}).
\]
Examples of estimates satisfying (1.5) are given in Section 5. At the second stage, define $t_i = C_i^T \hat{\beta}$ and obtain robustified estimates of $\theta$ by solving the following analogue of (1.4):

$$\sum_{i=1}^{N} \left\{ \psi^2((y_i - t_i) \exp(-h(t_i)^T \hat{\theta})) - \xi \right\} h(t_i) = 0,$$

(1.6)

where $\psi$ is monotone nondecreasing and $\xi = E\psi^2(Z_i)$, the expectation being taken under the standard normal distribution. Clearly, $\psi(x) = x$ leads to (1.4).

At the third stage, we now solve a robust version of (1.1):

$$\sum_{i=1}^{N} \psi((y_i - C_i^T \hat{\beta})/\hat{\sigma}_i) C_i = 0,$$

(1.7)

where

$$\hat{\sigma}_i = \exp(-h(t_i)^T \hat{\theta}).$$

(1.8)

Our main result (Theorem 4.1) lists conditions under which the limit distribution of $\hat{\beta}$ defined by (1.5)-(1.8) is the same as that of the estimate which could be found by solving (1.7) when $\sigma_1, \ldots, \sigma_N$ are known. In Section 2 we introduce notation and assumptions. Section 3 demonstrates that

$$\hat{\theta} - \theta = o_p(N^{-\frac{1}{2}}).$$

(1.9)

In Section 4 we state and prove the main result.

Remarks.

A If one assumes homoscedasticity ($\sigma_i = \sigma$), then $h(t_i) = 1$. The estimate $\hat{\beta}$ becomes an ordinary robust regression estimate with preliminary estimate of scale given by (1.5)-(1.6). See Maronna and Yohai (1979) for further details.

B We do not know if iterating (1.5)-(1.8) will lead to convergence. Further, we do not know whether simultaneous solution of (1.6)-(1.7) is possible.
Throughout the paper, we will use the following convention. For a real function $f$ on a space $X$, we say that $x_0$ solves $f(x) = 0$ if

$$|f(x_0)| < 2 \inf_X |f(x)|.$$  \hspace{1cm} (1.10)

With this convention, all our estimators will exist but need not be unique. However, our asymptotic results hold for every appropriate sequence of estimators.

The Box and Cox (1964) transformations form an alternative method for dealing with heteroscedasticity, as well as other deviations from the normal linear model. Defining

$$y^{(\lambda)} = \frac{y^{\lambda-1}}{\lambda} \hspace{1cm} (\lambda \neq 0)$$

$$= \log y \hspace{1cm} (\lambda = 0),$$

ey postulate that for some $\lambda$, $y^{(\lambda)}$ satisfies a homoscedastic normal linear model. Their methodology is based upon an entirely different model from ours, but a practitioner might consider using both on a given data set. Carroll (1978) and Bickel and Doksum (1978) independently studied both Box-Cox and robustified Box-Cox methods and concluded that variances of the estimated coefficients in the linear model are often much larger when $\lambda$ is estimated than when it is known. Therefore, confidence intervals and tests of hypotheses which are constructed as if $\lambda$ were known and not estimated are invalid.

In contrast, we show in Section 4 that, if the errors, $e_1$, are symmetrically distributed or $\psi(x) = x$, then for our method, the variances of estimated coefficients when $\theta$ is estimated are similar to those when $\theta$ is known, and confidence intervals and tests can be validly constructed as if $\theta$ were known.
2. Assumptions and notations. Let \( \psi \) be a nondecreasing function from \( \mathbb{R}^1 \) to \( \mathbb{R}^1 \) which satisfies

B1. \( E\psi(e_1) = 0, \)

B2. \( 0 < E\psi^2(e_1) = \xi < \infty, \)

for some \( \alpha > 0 \)

B3. \( E\psi((e_1+r)(1+s)) = A(\psi) r + O(|r|^{1+\alpha} + |s|^{1+\alpha}) \)

as \( r \to 0 \) and \( s \to 0 \) with \( A(\psi) > 0, \)

B4. \( E\psi^2((e_1+r)(1+s)) - \xi = A(\psi^2) s + O(|r|^{1+\alpha} + |s|^{1+\alpha}) \)

as \( r \to 0 \) and \( s \to 0 \) with \( A(\psi^2) > 0, \)

B5. there exists \( C_0 \) such that for all \( \delta < 1 \)

\[
\limsup_{k \to 0} E[\sup_{k \to 0} |\phi((e_1+r)(1+s)) - \phi((e_1+r')(1+s'))| : \]

\[
|r|, |r'|, |s|, |s'| \leq k \text{ and } |r-r'|, |s-s'| \leq k\delta \}

\[
\leq C_0 \delta
\]

for both \( \phi = \psi \) and \( \phi = \psi^2, \) and

B6. \( \lim_{r,s \to 0} E(\phi((e_1+r)(1+s)) - \phi(e_1))^2 = 0 \)

for both \( \phi = \psi \) and \( \phi = \psi^2. \)

The function \( h \) from \( \mathbb{R}^1 \) to \( \mathbb{R}^q, C_i \) in \( \mathbb{R}^p \) and \( \tau_i = C_i^T B, \) satisfy

B7. \( \lim_{N \to 0} (\sup_{N \to 0} (|C_i| + h(\tau_i))N^{-\frac{1}{2}}) = 0, \)

B8. \( \sup_{N} (N^{-1} \sum_{i=1}^{N} |C_i|^2 + h(\tau_i)^2) < \infty, \)

B9. Letting \( \lambda_n \) be the minimum eigenvalue of

\[
H_N = N^{-1} \sum_{i=1}^{N} h(\tau_i) h(\tau_i)^T
\]

then

\[
\liminf_{N \to \infty} \lambda_n = \lambda_\infty > 0
\]

B10. \( h \) is Lipschitz continuous on an interval (possibly infinite), I, such that \( \tau_i \) is in I for all i and N,
B11. \[ \inf_N \inf_{\theta \in \Theta} \mathbf{h}(\tau_i)^T \theta > -\infty \]
and \[ \mathbf{d}_i = C_i \exp(-\mathbf{h}(\tau_i)^T \theta), \]
\[ N^{-1} \sum_{i=1}^{N} d_i d_i^T = S_N + S \text{ (positive definite)}. \]

Remark. The monotonicity of \( \psi \) is needed only to prove \( \sqrt{N} \)-consistency of \( \hat{\beta} \) and \( \hat{\theta} \).

The conditions B1-B6 are notationally complex but widely applicable. They clearly hold for \( \rho(x) = x^2/2 \) (\( \psi(x) = x \)) if \( Ee_1 = 0, Ee_1^4 < \infty \), where \( A(\psi) = 1, A(\psi^2) = 2Ee_1^2 \). If \( F \) is symmetric and \( \psi \) is odd, B1-B6 can be verified if (i) \( \psi \) is constant outside an interval (ii) \( \psi \) is Lipschitz continuous and twice boundedly differentiable except possibly at a finite number of points, \( \alpha_1, \ldots, \alpha_k \) (iii) \( F \) is Lipschitz continuous in neighborhoods of \( \alpha_1, \ldots, \alpha_k \) and (iv) \( E\psi^2(e_1), E\psi'(e_1) \) and \( Ee_1 \psi(e_1) \psi'(e_1) \) are all positive. Then \( A(\psi) = E\psi'(e_1) \) and \( A(\psi^2) = 2Ee_1 \psi(e_1) \psi'(e_1) \).

For the power model (1.2), \( \mathbf{h}(x) = (1, \log |x|), \theta = (\log \sigma, \theta) \) and B10, B11 hold if

\[ \inf_N \inf_{1 \leq i \leq N} |\tau_i| > 0. \]

Since the power model (1.2) makes little sense when \( \tau_i = 0 \), (2.1) is reasonable.

To avoid this difficulty, in Section 1 we suggested \( \mathbf{h}(x) = (1, \log(1+|x|)) \), for which B10, B11 hold. The other example \( \mathbf{h}(x) = (1, |x|) \) will satisfy B10, B11 as long as the \( \{\tau_i\} \) are uniformly bounded.
3. Estimation of $\varphi$.

**Theorem 3.1.** Suppose B1-B8 hold, $(\hat{\beta}_0-\beta) = O_p(N^{-\frac{k}{2}})$, $t_i = C_i^T \hat{\theta}$ and $\hat{\theta}$ is any solution (see 1.10) to

\[
\sum_{i=1}^{N} \{ \psi^2((Y_i - t_i) \exp(-\frac{1}{2} (t_i^T \hat{\theta})) - \xi) \} h(t_i) = 0.
\]

Then

\[
(3.1) \quad \hat{\theta} - \theta = O_p(N^{-\frac{k}{2}}).
\]

**Proof.** For $\Delta_1$ in $\mathbb{R}^p$, $\Delta_2$ in $\mathbb{R}^q$ and

\[
\Delta = (\Delta_1, \Delta_2), \text{ define } h_i(\Delta) = h(t_i + C_i^T \Delta, N^{-\frac{k}{2}}),
\]

\[
(3.3) \quad \alpha_i^{(1)}(\Delta) = \exp\left\{ -h_i(\Delta) \Delta_2 N^{-\frac{k}{2}} + (h_i(\Delta) - h_i(\hat{\theta}))^T \hat{\theta} \right\} - 1
\]

\[
(3.4) \quad \alpha_i^{(2)}(\Delta) = N^{-\frac{k}{2}} d_i^T \Delta_1 \quad \text{(see B12)}.
\]

Then let $\phi_i(x,y,z) = \psi^2((x+z)(1+y)) - \xi$ and define the process

\[
W_N(\Delta) = -N^{-\frac{k}{2}} \sum_{i=1}^{n} \phi(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) h_i(\Delta).
\]

Note that (3.1) can be rewritten as

\[
W_N(N^{\frac{k}{2}}(\hat{\beta}_0-\beta), N^{\frac{k}{2}}(\hat{\theta}-\theta)) = 0.
\]

By B1 and Chebyshev's inequality,

\[
(3.5) \quad W_N(0) = O_p(1),
\]

so that by our convention (1.10),

\[
W_N(N^{\frac{k}{2}}(\hat{\beta}_0-\beta), N^{\frac{k}{2}}(\hat{\theta}-\theta)) = O_p(1).
\]

We can therefore prove (3.2) by showing that for each $M_1 > 0$, $\varepsilon > 0$, $Q > 0$ there exists $M_2 > 0$ such that
(3.6) \[ p \left( \inf \left\{ \left\| W_N(\Delta) \right\| \geq Q \right\} \right) \geq 1 - \varepsilon \]

We will prove (3.6) by modifying the proof of Jurečková's (1977) Lemma 5.2.

We first apply Theorem A.1 of the appendix, with \( x_i = h_1(0) \) and \( \alpha_i(\Delta) = h_1(\Delta) - h_1(0) \). Then B1-B8, B10, B11 imply that the assumptions of Theorem A.1 are met with \( g_1 = 1 \), \( A(\phi, i) = A(\psi^2) \), so that for all \( M > 0 \),

\[ \sup_{\left\| \Delta \right\| \leq M} \left\| W_N(\Delta) - W_N(0) - A(\psi^2) N^{-\frac{1}{2}} \sum_{i=1}^{n} h(\tau_i) \alpha_i(\Delta) \right\| = o_p(1). \]

By a Taylor series expansion,

\[ \alpha_i(\Delta) = -N^{-\frac{1}{2}} h(\tau_i) \Delta_2 + (h_1(0) - h_1(\Delta)) \Delta \theta + o_p(N^{-1}). \]

Thus, by B9, setting

\[ G_N(\Delta) = N^{-\frac{1}{2}} \sum_{i=1}^{n} h(\tau_i) (h(0) - h(\Delta)) \Delta \theta, \]

(3.7) \[ \sup_{\left\| \Delta \right\| \leq M} \left\| W_N(\Delta) - W_N(0) - A(\psi^2) H \Delta_2 + G_N(\Delta) \right\| = o_p(1). \]

Now fix \( \varepsilon > 0 \), \( M_1 > 0 \), \( Q > 0 \). Use (3.5) to choose \( \gamma \) such that

\[ p(\left\| W_N(0) \right\| \geq \gamma/2) < \varepsilon/2. \]

Define

\[ D = \sup_N \sup_{\left\| \Delta_1 \right\| \leq M_1} \left\| G_N(\Delta) \right\|. \]

Then \( D < \infty \) (\( G_N \) depends only on \( \Delta_1 \)). Define \( M_2 \) by

\[ [A(\psi^2) \lambda_\gamma M_2/2 - \gamma - D] = Q. \]

Using B9 and (3.7), find \( N_0 \) such that \( \lambda_N \geq \lambda_\gamma / 2 \) and

\[ p\left( \sup_{\left\| \Delta_2 \right\| \leq M_2} \sup_{\left\| \Delta_1 \right\| \leq M_1} \left\| W_N(\Delta) - W_N(0) - A(\psi^2) H \Delta_2 - G_N(\Delta) \right\| < \frac{\gamma}{2} \right) \geq 1 - \varepsilon/2 \]

\( (N \geq N_0) \).
If $||\Delta_2|| = M_2$, $||\Delta_1|| \leq M_1$, and $N \geq N_0$, then with probability at least 1-ε,

$$
\Delta_2 W_N(\Delta) \\
\geq -M_2 ||W_N(\Delta)|| + \Delta_2^T H \Delta_2 A(\psi^2) - M_2 D - M_2 \gamma/2 \\
\geq [A(\psi^2) \lambda_\infty M_2/2 - \gamma - D]M_2 = QM_2.
$$

Since $\psi$ is nondecreasing, $\Delta_2 W_N(\Delta_1, \Delta_2 s)$ is a nondecreasing function of $s$. Thus, $||\Delta_2|| \geq M_2$ implies

$$
\Delta_2 W_N(\Delta) \geq \Delta_2 W_N(\Delta_1, M_2 \Delta_2 ||\Delta_2||^{-1}) \\
\geq (||\Delta_2||/M_2)(M_2 \Delta_2 ||\Delta_2||^{-1})W_N(\Delta_1, M_2 \Delta_2 ||\Delta_2||^{-1}) \\
\geq ||\Delta_2|| Q
$$

Thus,

$$
p \left( \inf_{||\Delta_1|| \leq M_1} \inf_{||\Delta_2|| \geq M_2} \frac{\Delta_2 W_N(\Delta)}{||\Delta_2||} \geq Q \right) \geq 1-\varepsilon
$$

which with the Cauchy-Schwarz inequality proves (3.2). □

4. Estimation of $\hat{\beta}$. The limiting distribution of $\hat{\beta}$ is a simple consequence of the following representation.

Theorem 4.1. Suppose B1 to B12 hold and $A(\psi) > 0$. Then if $\hat{\beta}$ is any solution (see (1.10)) to

$$
\sum_{i=1}^{N} \psi((Y_i - C_i^T \hat{\beta}) \exp(-h(t_i)^T \hat{\beta})C_i \exp(-h(t_i)^T \hat{\beta})) = 0
$$

with $t_i = C_i^T \hat{\beta}$, $(\hat{\beta}_0 - \hat{\beta}) = O_p(N^{-\frac{1}{2}})$, and $\hat{\theta}$ satisfying (3.2), then

$$
N^{\frac{1}{2}} (\hat{\beta} - \beta) = N^{-\frac{1}{2}} \sum_{i=1}^{N} S^{-1} d_i \psi(e_i)(A(\psi))^{-1} + o_p(1).
$$

This result, strengthened so that the remainder is $o(1)$ almost surely, has been established in the homoscedastic case by Carroll and Ruppert (1979).
Proof. For $\Delta_1$ and $\Delta_3$ in $\mathbb{R}^p$, $\Delta_2$ in $\mathbb{R}^q$ and $\Delta = (\Delta_1, \Delta_2, \Delta_3)$, define

$$\alpha_1^{(1)}(\Delta) = N^{-\frac{k}{2}} d_1^T \Delta_1,$$

$$h_1(\Delta) = h(\tau_1 + C_1^T \Delta_3 N^{-\frac{k}{2}}),$$

$$\alpha_1^{(2)}(\Delta) = \exp(-h_1(\Delta)^T \Delta_2 N^{-\frac{k}{2}} + (h_1(\Delta) - h_1(\Delta))^T \theta) - 1,$$

and

$$\alpha_1^{(3)}(\Delta) = d_1 \alpha_1^{(2)}(\Delta).$$

Define the process

$$U_N(\Delta) = N^{-\frac{k}{2}} \sum_{i=1}^{N} \psi((e_i + \alpha_1^{(1)}(\Delta))(1 + \alpha_1^{(2)}(\Delta))(d_1 + \alpha_1^{(3)}(\Delta))).$$

Note that (4.1) can be rewritten as

$$U_N(\Delta) = N^{-\frac{k}{2}} \sum_{i=1}^{N} \psi((e_i + \alpha_1^{(1)}(\Delta))(1 + \alpha_1^{(2)}(\Delta))(d_1 + \alpha_1^{(3)}(\Delta))).$$

Letting $g_i = 1, \phi_i(e_i, r, s) = \psi((e_i + r)(1 + s))$, and $A(\psi, i) = A(\psi)$, the conditions of Theorem A.1 are implied by B1, B3, B5 to B8, B11, and B12, so for all $M > 0$

$$\sup_{|\Delta| \leq M} ||U_N(\Delta) - U_N(\Delta) - A(\psi) S \Delta_1|| = o_p(1).$$

Now by Chebyshev's theorem and B1,

$$U_N(\Delta) = O_p(1).$$

Therefore, if we set

$$\Delta^* = -(A(\psi) S)^{-1} U_N(\Delta),$$

then

$$\Delta^* = O_p(1)$$

and therefore, by (4.4)

$$U(\Delta^*) = o_p(1).$$
Consequently, by (4.3) and our convention (1.10),

\[(4.5) \quad U_N(N^{\hat{k}}(\hat{\beta} - \beta), N^{\hat{k}}(\hat{\theta} - \theta), N^{\hat{k}}(\hat{\beta}_0 - \beta)) = o_p(1).\]

By Theorem 3.1, \((\hat{\theta} - \theta) = O_p(N^{-\frac{1}{2}})\), so we need only establish that

\[(4.6) \quad (\hat{\beta} - \beta) = O_p(N^{-\frac{1}{2}})\]

to conclude from (4.3) and (4.4) that (4.2) holds. But by (4.5), (4.6) holds if for each \(\eta > 0, \varepsilon > 0\), and \(M_1\) there exists \(M_2\) satisfying

\[(4.7) \quad P(\inf \|A_1\| \geq M_2, \inf \|A_2\| \leq M_1, \inf \|A_3\| \leq M_1, |U_N(\Delta)| > \eta) > 1 - \varepsilon.\]

Now (4.7) follows from (4.4) in a manner quite similar to Jureckova's (1977) proof of her Lemma 5.2. \(\square\)

**Corollary 4.2.** If \(E\Psi^2(e_1) < \infty\) and the assumptions of Theorem 4.1 are met, then

\[N^{\hat{k}}(\hat{\beta} - \beta)^D \rightarrow N(0, S^{-1}E\Psi^2(e_1) A(\psi)^{-2}).\]

**Proof.** Use (4.2) and Loève's (1963, p 316) Normal Convergence Criterion to show that \(N^{\hat{k}}(\hat{\beta} - \beta)^T X^D \rightarrow N(0, X^TS^{-1}X E\Psi^2(e_1) A(\psi)^{-2})\) for each \(X\) in \(\mathbb{R}^p\).

Since B7, B11 and B12 show that

\[\lim_{N \rightarrow \infty} \left( \sum_{i=1}^{N} ||d_i||^{-2} - \sum_{i=1}^{N} ||d_i||^{-2} \right) = 0,

it is easy to see that the criterion is met. \(\square\)

Recall that \(A(\psi) = E\Psi'(e_1)\), for commonly used \(\Psi\).

Define

\[\hat{\sigma}_i = \exp(h(C_i^T \hat{\beta}_0)^T \hat{\theta})\]

(one could use \(\hat{\beta}_0\) instead of \(\hat{\beta}_0\)). Then when \((\hat{\beta} - \theta) = O_p(N^{-\frac{1}{2}})\) and \((\hat{\beta}_0 - \beta) = O_p(N^{-\frac{1}{2}})\), B7 and B10 imply

\[\sup_{i \leq N} |\hat{\sigma}_i - \sigma_i| = o_p(1).\]
Suppose for both $\phi = \psi$ and $\phi = \psi'$

$$\lim_{\epsilon \to 0} \sup_{x, y} |\phi((1 + x)(e_1 + y)) - \phi(e_1)| : |x| \leq \epsilon \text{ and } |y| \leq \epsilon = 0.$$  

Then it is easy to prove that

$$|N^{-1} \sum_{i=1}^{N} \phi(\hat{\sigma}_i^{-1} (Y_i - C_i \hat{\beta})) - \phi(e_1)| = o_p(1),$$

and by the strong law of large numbers,

$$N^{-1} \sum_{i=1}^{N} \phi(e_1) \to E\phi(e_1).$$

Moreover,

$$\left| \left| N^{-1} \sum_{i=1}^{N} \sigma_i^{-2} C_i^T - N^{-1} \sum_{i=1}^{N} d_i d_i^T \right| \right|$$

$$\leq \sup_{i \in S} |\sigma_i - \sigma_i| (N^{-1} \sum_{i=1}^{N} ||C_i||^2) = o_p(1)$$

(with $||A||$ equal to, say, the Euclidean norm of the matrix $A$),

so

$$N^{-1} \sum_{i=1}^{N} \sigma_i^{-2} C_i^T \to S.$$  

Therefore, $E\psi^2(e_1) A(\psi)^{-2} S$ can be consistently estimated, and large sample confidence ellipsoids and tests of hypotheses for $\beta$ can be constructed.

5. $\sqrt{N}$-consistency of preliminary estimators. Since we require that our preliminary estimate satisfy (1.5), we now give conditions which insure that

an $M$-estimate $\hat{\beta}_0$ solving

$$\sum_{i=1}^{N} \psi(Y_i - C_i \hat{\beta}_0) C_i = 0$$

satisfy (1.5). This $\hat{\beta}_0$ is not scale equivariant, but this does not affect the limit distribution of $\hat{\beta}$.

B13. $\psi$ is odd.
B14. $\psi$ is Lipschitz continuous with Lipschitz constant $L_\psi$

B15. the Radon-Nikodym derivative $\psi'$ satisfies

$E(\psi(x(e_1 + y)) - \psi(xe_1) - \psi'(xe_1)xy) \leq K|xy|^{1+\alpha}$

for some $K$ and $\alpha > 0$ and all $x$ and $y$.

B16. $N^{-\frac{1}{2}}(\sigma_i + 1)||d_i|| \to 0$

B17. $\sup(N^{-1} \sum_{i=1}^{N} (\sigma_i^2 + 1)||d_i||^2) < \infty$

B18. the minimum eigenvalue, $\lambda_N$, of

$N^{-1} \sum_{i=1}^{N} \sigma_i E\psi'(\sigma_i e_i) d_i d_i^T$

satisfies $\liminf_{N \to \infty} \lambda_N > 0$.

**Theorem 5.1.** Under B13 to B18, (1.9) holds.

**Proof.** We will apply Theorem A.1 with $\phi_i(e_1, r, s) = \psi(\sigma_i(e_1 - r))(\text{so } \phi_i \text{ does}

not depend on } s), g_i = \sigma_i, k_i = N^{-\frac{1}{2}}||c_i||, A(\psi, i) = \sigma_i E\psi'(\sigma_i e_i), \alpha_i^{(3)} = 0,$

and $\alpha_i^{(2)}$ left undefined. By B13 and B14, (A.3) holds. By B15, for all $r$ and $r'$

$|\phi_i(e_1, r) - \phi_i(e_1, r')| \leq K \sigma|r - r'|$

so that (A.4) and (A.5) hold. Also, (A.5) is implied by B15, and (A.1), (A.2),

(A.7), (A.8), and (A.9) are easily checked. Therefore, for all $M > 0$,

$$\sup_{||\Delta|| \leq M} ||N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi(\sigma_i (e_1 - d_i^T \Delta)) - N^{-\frac{1}{2}} \sum_{i=1}^{N} \psi(\sigma_i e_i)$$

$$+ N^{-1} \sum_{i=1}^{N} \sigma_i E\psi'(\sigma_i e_i) d_i d_i^T \Delta|| = O_p(1).$$

Now (5.1) follows exactly as Jurečková's (1977) Lemma 5.2, since

$Y_i - C_i^T \hat{\beta}_0 = \sigma_i (e_i - C_i \hat{\beta}_0 - \hat{\beta}).$
APPENDIX

The following general theorem will be used when studying \( \hat{\beta}_0, \hat{\theta} \) and \( \hat{\beta} \).

Theorem A.1. Let \( g_i \) and \( k_i \) (=g_{iN} \) and \( k_{iN} \) be sequences of positive constants such that

\[
(A.1) \quad \lim_{N \to \infty} \left( \sup_{i \leq N} \left( k_i + k_i g_i \right) \right) = 0
\]

and

\[
(A.2) \quad \sup_{N} \left( \sum_{i=1}^{N} k_i^2 + k_i^2 g_i^2 \right) = c_1 < \infty.
\]

Let \( \phi_i \) be a function from \( \mathbb{R}^3 \) to \( \mathbb{R}^1 \) satisfying

\[
(A.3) \quad \mathbb{E}\phi_i(e_1,0,0) = 0 \quad \text{for all } i,
\]

\[
(A.4) \quad \lim_{k \to 0} \sup \mathbb{E}\{ \sup \left( |\phi_i(e_1,r,s) - \phi_i(e_1,r',s')| \right) : \}
\]

\[
|r|, |r'|, |s|, |s'| \leq k \quad \text{and} \quad |r-r'|, |s-s'| \leq k\delta \}
\]

\[
\leq C_0 \delta g_i
\]

for some \( C_0 \) and all \( \delta < 1 \) and all \( i \),

\[
(A.5) \quad \sup_{i \leq N} g_i^{-1} \mathbb{E}(\phi(e_1,r,s) - \phi(e_1,0,0) - A(\phi,i)r) = O(|r|^{1+\alpha} + |s|^{1+\alpha})
\]

for positive constants \( A(\phi,i) \) and \( \alpha > 0 \), and

\[
(A.6) \quad \lim_{r,s \to 0} \sup_{i \leq N} \mathbb{E}(\phi_i(e_1,r,s) - \phi_i(e_1,0,0))^2 = 0.
\]

Let \( \alpha^{(1)}_i, \alpha^{(2)}_i, \) and \( \alpha^{(3)}_i \) be functions from \( \mathbb{R}^m \) to \( \mathbb{R}^1, \mathbb{R}^1 \) and \( \mathbb{R}^n \), respectively, such that

\[
(A.7) \quad \alpha^{(\ell)}_i(0) = 0, \quad \text{and}
\]

\[
(A.8) \quad ||\alpha^{(\ell)}_i(x) - \alpha^{(\ell)}_i(y)|| \leq k_i ||x-y||
\]

for all \( x \) and \( y \) in \( \mathbb{R}^m, i=1, \ldots, N, \) and \( \ell=1,2, \) or 3. Let \( x_i \) (=x_{iN} \) be elements of \( \mathbb{R}^n \) satisfying

\[
(A.9) \quad N^{-\frac{1}{2}} ||x_i|| \leq k_i.
\]
For \( \Delta \in \mathbb{R}^m \), define the process

\[
U_N(\Delta) = N^{-\frac{k}{2}} \sum_{i=1}^{N} \phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta))(x_i + \alpha_i^{(3)}(\Delta)).
\]

Then, for all \( M > 0 \)

\[
(A.10) \quad \sup_{|\Delta| \leq M} |U_N(\Delta) - U_N(0) - N^{-\frac{k}{2}} \sum_{i=1}^{N} A(\phi, i)\alpha_i^{(1)}(\Delta)x_i| = o_p(1).
\]

**Proof of Theorem A.1.** Fix \( M > 0 \). We will show that

\[
(A.11) \quad E(U_N(\Delta) - U_N(0)) = N^{-\frac{k}{2}} \sum_{i=1}^{N} A(\phi, i)\alpha_i^{(1)}(\Delta)x_i + o_p(1)
\]

and

\[
(A.12) \quad U_N(\Delta) - U_N(0) - E(U_N(\Delta) - U_N(0)) = o_p(1)
\]

for each fixed \( \Delta \), and that there exists \( K \) depending upon \( M \) but not \( \delta \) such that for all \( \delta > 0 \) and all \( N \)

\[
(A.13) \quad E \sup \{ |U_N(\Delta) - U_N(\Delta^*)| : |\Delta| \leq M, |\Delta^*| \leq M, |\Delta - \Delta^*| \leq \delta \} \leq K\delta.
\]

Since for any \( \delta \), we can cover a ball of radius \( M \) in \( \mathbb{R}^m \) with a finite number of balls of radius \( \delta \), (A.11), (A.12), and (A.13) prove that for each \( \delta > 0 \),

\[
\lim_{N \to \infty} P \left( \sup_{|\Delta| \leq M} |U_N(\Delta) - U_N(0) - N^{-\frac{k}{2}} \sum_{i=1}^{N} A(\phi, i)\alpha_i^{(1)}(\Delta)x_i| \leq K\delta \right) = 1,
\]

which proves the theorem.

To prove (A.11), note that by (A.3),

\[
N^{-\frac{k}{2}} \sum_{i=1}^{N} E(U_N(\Delta) - U_N(0)) =
\]

\[
N^{-\frac{k}{2}} \sum_{i=1}^{N} E\left( \phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) - \phi_i(e_i, 0, 0) \right)(x_i + \alpha_i^{(3)}(\Delta)),
\]

by (A.1) and (A.8).

\[
(A.14) \quad |x_i + \alpha_i^{(3)}(\Delta)| \leq 2|x_i| \quad \text{for all large } N, \quad \text{and by } (A.5),
\]

\[
E(\phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) - \phi_i(e_i, 0, 0)) =
\]

\[
A(\phi, i)\alpha_i^{(1)}(\Delta) + O(g_i k_i |\Delta|)^{1+\alpha}
\]
uniformly in $i$. Therefore
\[ E((U_N(\Delta) - U_N(0)) = N^{-\frac{1}{2}} \sum_{i=1}^{N} A(\phi_i, \alpha_i^{(1)}(\Delta) x_i \]
\[ + O(N^{-\frac{1}{2}} \sum_{i=1}^{N} (g_{i,k_i})^{1+\alpha} ||x_i||), \]
and by (A.2) and (A.9),
\[ N^{-\frac{1}{2}} \sum_{i=1}^{N} (g_{i,k_i})^{1+\alpha} ||x_i|| \leq g_{i}^{\alpha} \sum_{i=1}^{N} g_{i}^{2} = \]
\[ \leq (g_{i,k_i})^{2} (\sum_{i=1}^{N} g_{i}^{2})^{\frac{1}{2}} (\sum_{i=1}^{N} k_{i}^{2})^{\frac{1}{2}} = o(1). \]
Thus (A.1) holds.

Next, by (A.14) we have that for $N$ large
\[ \text{Var}(U_N(\Delta) - U_N(0)) \leq \]
\[ (2N^{-\frac{1}{2}} \sum_{i=1}^{N} g_{i}^{2} ||x_i||^2) \sup_{\Delta \in N} g_{i}^{-2} E(\phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) - \phi_i(e_i, 0, 0))^2. \]
It follows from (A.4), (A.5), (A.7), and (A.8) that
\[ \sup_{\Delta \in N} g_{i}^{-2} E(\phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) - \phi_i(e_i, 0, 0))^2 = o(1). \]
Therefore, (A.12) is proved by applying (A.2). Finally, by (A.14) the RHS of (A.13) is less than or equal to
\[ 2N^{-\frac{1}{2}} \sum_{i=1}^{N} ||x_i|| E \sup \{ |\phi_i(e_i, \alpha_i^{(1)}(\Delta), \alpha_i^{(2)}(\Delta)) - \phi_i(e_i, \alpha_i(\Delta^*), \alpha_i(\Delta^{*})| : ||\Delta|| \leq M, ||\Delta^{*}|| \leq M, ||\Delta - \Delta^{*}|| \leq \delta \} \]
which by (A.2), (A.4) to (A.9), and the Cauchy-Schwarz inequality does not exceed
\[ \sup_{\Delta \in N} 2(\sum_{i=1}^{N} (g_{i,k_i})^{2})^{\frac{1}{2}} C_0^{\delta} = 2C_0^{\frac{1}{2}} \delta. \]
Therefore (A.13) is verified.
REFERENCES


Carroll, Raymond J. (1979), "Robust transformations to achieve approximate normality", to appear in J Roy Statist Soc, Series B.

Carroll, Raymond J. and Ruppert, David (1979), "Almost sure properties of robust regression estimates", unpublished manuscript.


**M-Estimates for the Heteroscedastic Linear Model**

**Authors:**
Raymond J. Carroll and David Ruppert

**Performing Organization:**
Department of Statistics
University of North Carolina at Chapel Hill

**Sponsor:**
Air Force Office of Scientific Research
Bolling AFB, DC

**Report Date:**
July 1979

**Security Class:**
Unclassified

**Abstract:**
We treat the linear model $Y_i = C_i \beta + Z_i$ where $C_i$ is a known vector, $\beta$ is an unknown parameter, and $\text{Var}(Z_i)$ is a function of $C_i \beta$, which is known, except for a parameter $\theta$. For simultaneous M-estimates, $\hat{\beta}$ and $\hat{\theta}$, we show that $D(\hat{\theta}, \theta) = O_p(N^{-1/2})$, and find the limit distribution of $N^{1/2}(\hat{\beta} - \beta)$. Moreover, the covariance matrix of the limit distribution can be consistently estimated, so large sample confidence ellipsoids and tests of hypotheses concerning $\beta$ are feasible.