1

Robust Statistics and its Application in Linear Regression

This is an introductory chapter giving the mathematical background to the robust statistics that are used in the rest of the book. Robust linear regression methods are then generalized to nonlinear regression in the rest of the book.

The robust approach to linear regression is described in this chapter. It is the main motivation for extending statistical inference approaches used in linear regression to nonlinear regression. This is done by considering the gradient of a nonlinear model as the design matrix in a linear regression. Outlier detection methods used in linear regression are also extended to use in nonlinear regression.

In this chapter the consistency and asymptotic distributions of robust estimators and robust linear regression are presented. The validity of the results requires certain regularity conditions, which are presented here. Proofs of the theorems are very technical and since this book is about nonlinear regression, they have been omitted.

1.1 Robust Aspects of Data

Robust statistics were developed to interpret data for which classical assumptions, such as randomness, independence, distribution models, prior assumptions about parameters and other prior hypotheses do not apply. Robust statistics can be used in a wide range of problems.

The classical approach in statistics assumes that data are collected from a distribution function; that is, the observed values \((x_1, x_2, \ldots, x_n)\) follow the simultaneous distribution function \(F_n(x_1, \ldots, x_n)\). If the observations are identically independently distributed (i.i.d.) with distribution \(F\), we write \(x_i \overset{iid}{\sim} F(x_i), i = 1, \ldots, n\) (the tilde sign \(\sim\) designates a distribution). In real-life data, these explicit or other implicit assumptions might not be true. Outlier data effects are examples of situations that require robust statistics to be used for such null conditions.
1.2 Robust Statistics and the Mechanism for Producing Outliers

Robust statistics were developed to analyse data drawn from a wide range of distributions and particularly data that do not follow a normal distribution, for example when a normal distribution is mixed with another known statistical distribution:

\[ F = (1 - \epsilon)N(\cdot, \cdot) + \epsilon D \]  

(1.1)

where \( \epsilon \) is a small value representing the proportion of outliers, \( N(\cdot, \cdot) \) is the normal cumulative distribution function (CDF) with appropriate mean and variance, and \( D \) belongs to a suitable class of CDFs. A normal distribution \( (D) \) with a large variance can produce a wide distribution, such as:

\[ F = (1 - \epsilon)N(0, \sigma^2) + \epsilon N(0, \sigma_0^2) \]

for a large value of \( \sigma_0^2 \) (see Figure 1.1a). A mixture of two normal distributions with a large difference in their means can be generated by:

\[ F = (1 - \epsilon)N(0, \sigma^2) + \epsilon N(\mu_0, \sigma_0^2) \]

where the variance value \( \sigma_0^2 \) is much smaller than \( \sigma^2 \), and the mean \( \mu_0 \) is the mean of the shifted distribution (see Figure 1.1b). The models in this book will

![Figure 1.1](image1.png)

**Figure 1.1** Contaminated normal densities: (a) mixture of two normal distributions with different means; (b) mixture of two normal distributions with different variances.  

---

1 Some of the mathematical notation and definitions used in this chapter follow that in Maronna, et al. (2006).
be used to interpret data sets with outliers. Figure 1.1a shows the CDF of a mixture of two normal distributions with different means:

\[ F = 0.9 \times N(0, 1) + 0.1 \times N(9, 1) \]

and Figure 1.1b shows the CDF of a mixture of two normal distributions with different variances:

\[ F = 0.9 \times N(0, 1) + 0.1 \times N(0, 2) \]

1.3 Location and Scale Parameters

In this section we discuss the location and scale models for random sample data. In later chapters these concepts will be extended to nonlinear regression. The location model is a nonlinear regression model and the scale parameter describes the nonconstant variance case, which is common in nonlinear regression.

1.3.1 Location Parameter

Nonlinear regression, and linear regression in particular, can be represented by a location model, a scale model or simultaneously by a location model and a scale model (Maronna et al. 2006). Not only regression but also many other random models can be systematically studied using this probabilistic interpretation. We assume that an observation \( x_i, i = 1, \ldots, n \) depends on the unknown true value \( \mu \) and that a random process acts additively as

\[ x_i = \mu + \epsilon_i, \ i = 1, \ldots, n \]  

(1.2)

where the errors \( \epsilon_i \) are random variables. This is called the location model and was defined by Huber (1964). If the errors \( \epsilon_i \) are independent with common distribution \( F_0 \) then the \( x_i \) outcomes are independent, with common distribution function

\[ F(x) = F_0(x - \mu) \]

and density function \( f_0(x - \mu) = F'_0 \). An estimate \( \hat{\mu} \) is a function of the observations \( \hat{\mu} = \hat{\mu}(x_1, \ldots, x_n) \). We are looking for estimates that, with high probability, satisfy \( \hat{\mu} \approx \mu \). The maximum likelihood estimate (MLE) of \( \mu \) is a function of observations that maximize the likelihood function (joint density):

\[ L(x_1, \ldots, x_n; \mu) = \prod_{i=1}^{n} f_0(x_i - \mu) \]  

(1.3)

The estimate of the location can be obtained from:

\[ \hat{\mu}_{\text{MLE}} = \hat{\mu}(x_1, \ldots, x_n) = \arg \max_{\mu} L(x_1, \ldots, x_n; \mu) \]
Since $f_0$ is positive and the logarithm function is an increasing function, the MLE of a location can be calculated using a simple maximization logarithm statement:

$$
\hat{\mu}_{\text{MLE}} = \arg \max_{\mu} \ell(\mu) = \arg \max_{\mu} \sum_{i=1}^{n} \log(f_0(x_i - \mu))
$$

(1.4)

If the distribution $F_0$ is known then the MLE will have desirable mathematical and optimality properties, in the sense that among unbiased estimators it has the lowest variance and an approximately normal distribution. In the presence of outliers, since the distribution $F_0$ and, in particular, the mixture distribution (1.1) are unknown or only approximately known, statistically optimal properties might not be achieved. In this situation, some optimal estimates can still be found, however. Maronna et al. (2006, p. 22) state that to achieve optimality, the goal is to find estimates that are:

- nearly optimal when $F_0$ is normal
- nearly optimal when $F_0$ is approximately normal.

To this end, since MLEs have good properties such as sufficiency, known distribution and minimal bias within an unbiased estimator but are sensitive to the distribution assumptions, an MLE-type estimate of (1.4) can be defined. This is called an M-estimate. As well as the M-estimate for location, a more general definition can be developed. Let:

$$
\rho(t) = -\log f_0(t)
$$

(1.5)

The negative logarithm of (1.3) can then be written as $\sum_{i=1}^{n} \rho(x_i - \mu)$.

A more sophisticated form of M-estimate can be defined by generalizing to give an estimator for a multidimensional unknown parameter $\mu$ of an arbitrary modeling of a given random sample $(x_1, \ldots, x_n)$.

**Definition 1.1** If a random sample $(x_1, \ldots, x_n)$ is given, and $\mu \in \mathbb{R}^p$ is an unknown $p$-dimensional parameter of a statistical model describing the behavior of the data, any estimator of $\mu$ is a function of a random sample $\hat{\mu} = \mu(x_1, \ldots, x_n)$. The M-estimate of $\mu$ can be defined in two different ways: by a minimization problem of the form (estimating equation and functional form are represented together):

$$
\sum_{i=1}^{n} \rho(x_i; \hat{\mu}) = \int \rho(x, \hat{\mu})dF_n = \min
$$

(1.6)

or as the solution of the equation with the functional form

$$
\sum_{i=1}^{n} \psi(x_i; \hat{\mu}) = \int \psi(x, t)dF_n(x) = 0
$$

(1.7)
where the functional form means $E(\psi(x, t)) = 0$, $F_n$ is an empirical CDF, and $\rho$ (the robust loss function) and $\psi$ are arbitrary functions. If $\rho$ is partially differentiable, we can define the psi function as $\psi_{\rho, x_1}(x_i; \mu) = (\partial / \partial \mu)\rho(x_i; \mu)$, which is specifically proportional to the derivative ($\psi \propto \rho'$), and the results of Equations 1.6 and 1.7 are equal. In this section we are interested in the M-estimate of the location for which $\psi(x, t) = \psi(x - t)$.

The M-estimate was first introduced for the location parameter by Huber (1964). Later, Huber (1972) developed the general form of the M-estimate, and the mathematical properties of the estimator (1973; 1981).

**Definition 1.2** The M-estimate of location $\mu$ is defined as the answer to the minimization problem:

$$\hat{\mu} = \text{arg min}_\mu \sum_{i=1}^{n} \rho(x_i - \mu)$$ (1.8)

or the answer to the equation:

$$\sum_{i=1}^{n} \psi(x_i - \hat{\mu}_n) = 0$$ (1.9)

If the function $\rho$ is differentiable, with derivative $\psi(t) = d\rho/dt$, the M-estimate of the location (1.8) can be computed from the implicit equation (1.9).

If $F_0$ is a normal distribution, the $\rho$ function, ignoring constants, is a quadratic function $\rho(x) = x^2$ and the parameter estimate is equivalent to the least squares estimate, given by:

$$\hat{\mu} = \text{arg min}_\mu \sum_{i=1}^{n} (x_i - \mu)^2$$

which has the average solution $\hat{\mu} = \bar{x} = (1/n) \sum_{i=1}^{n} x_i$.

If $F_0$ is a double exponential distribution with density $f_0(x) = (1/2)e^{-|x|}$, the rho function, apart from constants, is the absolute value function $\rho(x) = |x|$, and the parameter estimate is equivalent to the least median estimate given by:

$$\hat{\mu} = \text{arg min}_\mu \sum_{i=1}^{n} |x_i - \mu|$$ (1.10)

which has median solution $\hat{\mu} = \text{median}(x_i)$ (see Exercise 1). Apart from the mean and median, the distribution of the M-estimate is not known, but the convergence properties and distribution can be derived. The M-estimate is defined under two different formulations: the $\psi$ approach from the estimating equation $\sum_{i=1}^{n} \psi(x_i, \mu) = 0$ or by minimization of $\sum_{i=1}^{n} \rho(x_i ; \mu)$, where $\rho$ is a primitive function of $\psi$ with respect to $\mu$. The consistency and asymptotic assumptions of the M-estimate depend on a variety of assumptions. The $\psi$ approach does
not have a unique root or an exact root, and a rule is required for selecting a root when multiple roots exist.

**Theorem 1.3**  Let \( \lambda_F(t) = \int \psi(t, x) dF(x) = 0 \). Assume that:

**A 1.4**

(i) \( \lambda_{F_n}(t) = 0 \) has unique root \( \mu_0 \)  
(ii) \( \psi \) is continuous and either bounded or monotone.

Then the equation \( \lambda_{F_n}(t) = 0 \) has a sequence of roots \( \hat{\mu}_n \) that converge in probability \( \hat{\mu}_n \to \mu_0 \).

In most cases, the equation \( \lambda_{F_n}(t) = 0 \) does not have an explicit answer and has to be estimated using numerical iteration methods. Starting from \( \sqrt{n} \) consistent estimates \( \hat{\mu}_n \), one step of the Newton–Raphson estimate is \( \delta_n = \hat{\mu}_n - \frac{\psi_{F_n}(\hat{\mu}_n)}{\psi'_{F_n}(\hat{\mu}_n)} \), where \( \psi'_{F_n}(t) = \frac{\partial}{\partial t} \sum \psi(x_i - t) \). The consistency and normality of \( \delta_n \) are automatic, but there is no root of \( \lambda_{F_n} = 0 \) and furthermore the iteration does not change the first-order asymptotic properties.

**Theorem 1.5**  Suppose the following assumptions are satisfied:

**A 1.6**

(i) \( \lambda_{F(t)} = 0 \) has unique root \( t_0 \)  
(ii) \( \psi(x, t) \) is monotone in \( t \)  
(iii) \( \lambda'_{F}(t_0) \) exists and \( \neq 0 \)  
(iv) \( \int_{t_0} \psi^2(x, t) dF(x) < \infty \) in some neighborhood of \( t_0 \) and is continuous at \( t_0 \).

Then any sequence of the root of \( \lambda_{F_n}(t) = 0 \) satisfies

\[
\sqrt{n}(\hat{\mu} - t_0) \xrightarrow{D} \mathcal{N} \left( 0, \frac{\int \psi^2(x, t_0) dF(x)}{\left( \int \psi'(x, t) dF(x) \right)^2}_{t=t_0} \right)
\]

(1.11)

For a proof, see DasGupta (2008) and Serfling (2002).

Thus the location estimate \( \mu \) from (1.9) or (1.8), under the conditions of Theorem 1.3, will converge in probability to the exact solution of \( \sum \psi(x_i - \mu_0) = 0 \) as \( n \to \infty \). Under the conditions of Theorem 1.5 it will have a normal distribution:

\[
\mu \sim \mathcal{N} \left( \mu_0, \frac{\nu}{n} \right) \text{ with } \nu = \frac{E_F(\psi(x - \mu_0)^2)}{(E_F \psi'(x - \mu_0))^2}
\]

(1.12)
In Equation (1.12), the parameter $\mu_0$ is unknown and one cannot calculate the variance of the estimate. Instead, for inference purposes, we replace the expectations in the equation with the average, and parameter $\mu_0$ by its estimate $\hat{\mu}_n$:

$$\hat{\nu} = \frac{\sum \psi(x_i - \hat{\mu}_n)^2}{\sum [\psi'(x_i - \hat{\mu}_n)]^2}$$  \hspace{1cm} (1.13)

In Appendix A.3 several robust loss functions are presented.

### 1.3.2 Scale Parameters

In this section we discuss the scale model and parameters. The scale parameter estimate value is not only important in applications, but also plays a crucial role in computational iterations and the heteroscedasticity of variance cases. Consider the observations $x_i$ that satisfy the multiplicative model known as the scale model:

$$x_i = \sigma \varepsilon_i$$ \hspace{1cm} (1.14)

The values $\varepsilon_i$ are i.i.d., with density $f_0$, and $\sigma > 0$ is an unknown parameter called the scale parameter. The distribution of $x_i$ follows the scale family:

$$\frac{1}{\sigma} f_0 \left( \frac{x}{\sigma} \right).$$

Examples are the exponential family $f_0(x) = \exp(-x)I(x > 0)$ and the normal scale family $N(0, \sigma^2)$.

Thus the MLE of $\sigma$ is:

$$\hat{\sigma} = \arg \max_{\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0(x_i/\sigma)$$

Let $e_i = x_i/\sigma$. Taking logs and differentiating with respect to $\sigma$ yields

$$\frac{1}{n} \sum_{i=1}^{n} \frac{f'_0(e_i)}{f_0(e_i)} \frac{x_i}{\sigma} = 1$$

Let $\rho(t) = t\psi(t)$. If $\psi(t) = -f'_0(t)/f_0$, the estimating equation for $\sigma$ can be written as

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{x_i}{\sigma} \right) = 1$$ \hspace{1cm} (1.15)

**Definition 1.7** Assume in the multiplicative model that $(x_1, \ldots, x_n)$ are $n$ i.i.d. random samples of random variable $X$ and $\varepsilon_1, \ldots, \varepsilon_n$ are $n$ random samples of error random variable $\varepsilon$. $X$ follows the multiplicative model $X = \sigma \varepsilon$. 
For an appropriate $\rho$ function and constant $\delta$ the M-estimate of scale $\sigma$ is defined as (Huber 1964)

$$\int \rho \left( \frac{x}{\sigma(F)} \right) dF = \text{E} \left( \frac{x}{\sigma(F)} \right) = \delta$$

(1.16)

with sequence of estimation

$$\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{x_i}{\sigma} \right) = \delta$$

(1.17)

Under regularity conditions, this converges to the functional form of (1.16).

### 1.3.3 Location and Dispersion Models

In an alternative approach, the location–dispersion model with two unknown parameters is defined as:

$$x_i = \mu + \sigma \varepsilon_i, \ i = 1, \ldots, n$$

(1.18)

where $\varepsilon$ has density $f_0$ and hence $x_i$ has density

$$f(x) = \frac{1}{\sigma} f_0 \left( \frac{x - \mu}{\sigma} \right)$$

In this case, $\sigma$ is a scale parameter of $\sigma \varepsilon_i$, but a dispersion parameter of $x_i$. In practice, the parameter estimate for $\mu$ depends on $\sigma$, which might be known or unknown. The MLE for estimating $(\mu, \sigma)$ simultaneously is:

$$(\hat{\mu}, \hat{\sigma}) = \arg \max_{\mu,\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_0 \left( \frac{x_i - \mu}{\sigma} \right)$$

(1.19)

which, after taking logs and changing the sign, can be written as an optimization problem:

$$(\hat{\mu}, \hat{\sigma}) = \arg \min_{\mu,\sigma} \left\{ \frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{x_i - \mu}{\sigma} \right) + \log(\sigma) \right\}$$

(1.20)

with $\rho_0 = -\log f_0$. Equations 1.19 and 1.20 are extensively used in this book to develop the underlying theory and also in reaching parameter estimates in problems of nonlinear and robust nonlinear regression. By calculating the derivative of (1.20) with respect to location and dispersion, the MLE of simultaneous $(\mu, \sigma)$ can be defined, as in 1.9 and 1.17, by the simultaneous equations:

$$\begin{cases}
\text{estimating } \mu : & \sum_{i=1}^{n} \psi \left( \frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) = 0 \\
\text{estimating } \sigma : & \frac{1}{n} \sum_{i=1}^{n} \rho_{\text{scale}} \left( \frac{x_i - \hat{\mu}}{\hat{\sigma}} \right) = \delta
\end{cases}$$

(1.21a) (1.21b)
where \( \psi(t) = -\rho'_0(t), \rho_{\text{scale}}(t) = t\psi(t) \) and \( \delta = 1 \). The functional form can be written as:

\[
\begin{align*}
\text{estimating } \mu : & \quad E\psi \left( \frac{x-\mu}{\sigma} \right) = 0 \quad (1.22a) \\
\text{estimating } \sigma : & \quad E\rho_{\text{scale}} \left( \frac{x-\mu}{\sigma} \right) = \delta \quad (1.22b)
\end{align*}
\]

It can be proved that if \( F \) is symmetric then \( \hat{\mu} \sim N(\mu, \nu/n) \), where:

\[
\nu = \sigma^2 \frac{E[\psi(x-\mu)/\sigma]^2}{[E\psi(x-\mu)/\sigma]^2}
\]

For computational purposes, the expectation can be estimated by the average, and unknown parameters can be replaced by their estimated values.

### 1.3.4 Numerical Computation of M-estimates

The calculation of M-estimates, as discussed in the last three sections, requires numerical solutions to optimization or root-finding problems. In this section we will derive these using the iterative reweighting method. The system of simultaneous location and dispersion introduced in Section 1.3.3 is discussed in this section. The special cases of univariate location (Section 1.3.1) and scale estimates (Section 1.3.2) can easily be simplified from the implemented algorithms by considering another parameter as the known value. Note that in practice both of the parameters are unknown. The role of numerical methods is vital in nonlinear regression parameter estimation and software design because nonlinear models have to be approximated by linear expansion and the performing of multiple numerical iterations.

The computation of location parameters from the simultaneous equations (1.21) depends on the value of the dispersion parameter. This can be computed before the location, or a simultaneous algorithm can be used to compute both. However, it is important to note that in numerical procedures, the initial estimate of the scale parameter will break the robustness of the location estimates. This is critical in nonlinear regression, which will be discussed in Chapter 3.

In general, consider an estimating iterative procedure, with values of location and dispersion in the \( k \)th iteration \( \hat{\mu}_k, \hat{\sigma}_k \), and starting from initial values where \( k = 1 \). Let \( \hat{t}_i = ((x_i - \hat{\mu})/\hat{\sigma}) \). Depending on the known value of the dispersion we can solve three problems.

**Problem 1: Location estimation with previously estimated dispersion:**
Assuming that the dispersion estimate computed previously is \( \hat{\sigma} \), the location estimate is the answer to the estimating equation (1.21a) for a fixed known value \( \hat{\sigma} \). Note that we can write Equation 1.21a as:

\[
\frac{1}{\hat{\sigma}} \sum_{i=1}^{n} \psi_{0}(\hat{t}_i) \frac{(x_i-\hat{\mu})}{\hat{t}_i} = 0
\]

(1.23)
Algorithm 1.1 Location with previously computed variance.

**Step 1:** Compute $\hat{\sigma} = \text{MADN}(x)$ and $\mu_0 = \text{Median}(x)$.

**Step 2:** For $k = 0, 1, \ldots$ compute the weights $w_{k,j}$ and then $\mu_{k+1}$ in (1.24).

**Step 3:** Stop iteration when $|\mu_{k+1} - \hat{\mu}_k| < \varepsilon \hat{\sigma}$.

Starting from an initial value for the location, the new iterative value can be computed as:

$$\hat{\mu}_{k+1} = \frac{\sum_{i=1}^{n} w_{k,i} x_i}{\sum_{i=1}^{n} w_{k,i}} \quad (1.24)$$

where $w_{k,i} = W_1((x_i - \hat{\mu}_k) / \hat{\sigma})$ is the weight at iteration $k$ for point $x_i$ and

$$W_1(t) = \begin{cases} \psi(t)/t & t \neq 0 \\ \psi'(0) & t = 0 \end{cases} \quad (1.25)$$

If $W_1(t)$ is bounded and nonincreasing for $t > 0$, then the sequence converges to the solution of (1.21a), hence the algorithm for estimating the location to achieve a tolerance $\varepsilon$ as the precision of the parameter can be defined:

**Problem 2: Scale estimate:** Suppose the location is a fixed known value, then the estimating equation of scale (1.21b) is equivalent to (1.17). It can be rewritten in the weighted form:

$$\frac{1}{n} \sum_{i=1}^{n} \rho_{\text{scale}}(t_i) \left( \frac{x_i - \hat{\mu}}{\hat{\sigma}} \right)^2 = \delta$$

An iterative solution is then:

$$\hat{\sigma}_{k+1}^2 = \frac{1}{n \hat{\sigma}} \sum_{i=1}^{n} w_{k,i}(x_i - \mu)^2 \quad (1.26)$$

where $w_{k,i} = W_2((x_i - \hat{\mu}_k) / \hat{\sigma})$ is the weight at iteration $k$ for point $x_i$ and

$$W_2(t) = \begin{cases} \rho(t)/t^2 & t \neq 0 \\ \rho'(0) & t = 0 \end{cases} \quad (1.27)$$

This formula can be used to derive the iterative reweighting for estimating location and dispersion simultaneously. Meanwhile, for estimating the univariate scale estimate, without loss of generality, we can replace the zero value for location. If $W_2(t)$ is a bounded, even, continuous and nonincreasing function for $t > 0$, the sequence $\sigma_k$ converges to the solution of (1.17), so the algorithm is as 1.2:

**Problem 3: Simultaneous location and dispersion estimate:** The simultaneous location and dispersion estimate in (1.21) can be derived from two old algorithms.
Algorithm 1.2 Location with previously computed variance.

Step 1: For \( k = 0, 1, \ldots \) compute the weights \( w_{2k,j} \) and then \( \sigma_{k+1} \) in (1.26).

Step 2: Stop iteration when \( |\hat{\sigma}_{k+1}/\hat{\sigma}_k| < \epsilon \).

Algorithm 1.3 Simultaneous location and dispersion estimate.

Step 1: Compute the starting values \( \hat{\mu}_0, \hat{\sigma}_0 \),

\[ t_{k,i} = \frac{x_i - \hat{\mu}_k}{\hat{\sigma}_k}, \quad i = 1, \ldots, n \]

and weights \( W_1(t_{k,i}) \) and \( W_2(t_{k,i}) \) from (1.25) and (1.27), respectively.

Step 2: Given \( \mu_k, \sigma_k \) renew the iteration by

\[ \hat{\mu}_{k+1} = \frac{\sum_{i=1}^{n} W_1(t_{k,i})x_i}{\sum_{i=1}^{n} W_1(t_{k,i})}, \quad \hat{\sigma}_{k+1} = \frac{\hat{\sigma}_2}{n\hat{\sigma}} \sum_{i=1}^{n} W_2(t_{k,i})t_{k,i}^2 \]

1.4 Redescending M-estimates

Less informative distributions tend to have an exponential tail; that is, they might be too narrow and not wide enough to solve the outlier problem in practice. Better performance can be obtained with a long tailed distribution, and this may even increase the maximum risk slightly beyond its minimax value (Huber 1981). Consider the M-estimate subject to the side condition:

\[ \psi(t) = 0 \text{ for } |t| > c \quad (1.28) \]

for arbitrary \( c \). Figure 1.2a shows the \( \rho \) and \( \psi \) functions for the Huber, Hampel, Tukey bi-square and Andrew functions (see Table A.11 for definitions). The Huber rho function is unbounded while the Hampel, bi-square and Andrew functions are bounded and, when redescending, satisfy (1.28). Redescending \( \psi \) functions are beneficial when there are extreme outliers, but the improvement is relatively minor and their sensitivity increases to an incorrect scale value and the minimum of objective function \( \sum \rho(x_i - \mu_n) \) can be a trapped, local minimum. Removing impossible outliers by careful data screening based on physical knowledge might be more effective and reduce the risk.

1.5 Breakdown Point

Intuitively, the breakdown point of an estimate is defined as the maximum proportion of arbitrarily large outliers (so, mathematically, with infinite values) that can be tolerated. In other words, it can be expressed as the smallest fraction of outliers that would cause the estimator to have an improperly large value. It
is known that the breakdown point concept is more useful for small-sample situations, so a finite sample breakdown point should be defined.

**Definition 1.8**  Finite sample breakdown point.

Let \( X = (x_1, \ldots, x_n) \) be a fixed sample of size \( n \). Huber (1981, Ch. 11.2) defined three ways of corrupting a sample:

(i) \( \varepsilon \) contamination: The sample is adjoined with \( m \) additional values \( Y = (y_1, \ldots, y_m) \). Thus the fraction of “bad” values in the corrupted sample \( Z = X \cup Y \) is \( \varepsilon = m/(m + n) \).

(ii) \( \varepsilon \) replacement: A subset of size \( m \) is replaced by arbitrary values \( (y_1, \ldots, y_m) \), hence the fraction of “bad” values in the corrupted sample \( Z \) is \( \varepsilon = m/n \).

(iii) \( \varepsilon \) modification: Let \( \pi \) be an arbitrary distance function defined in the space of empirical measures. Let \( F_n \) be an empirical measure of sample \( X \) and \( G_m \) an empirical measure of sample \( Z \) with size \( m \), such that \( \pi(F_n, G_m) \leq \varepsilon \).

Let \( T_n \) be an estimator on the same Euclidean space for estimating parameter \( \theta \), \( T(X) \) its value for sample \( X \), and \( T(Z) \) the value calculated for the \( \varepsilon \) modified/replaced/contaminated sample \( Z \). The maximum bias caused by \( \varepsilon \) corruption is:

\[
 b(\varepsilon; X, T) = \sup_{Z} \| T(Z) - T(X) \| \tag{1.29}
\]
where the supremum is taken over the set of all \( \varepsilon \)-corrupted samples \( Z \). The breakdown point is then defined as:

\[
\varepsilon^*(X, T) = \inf \{ \varepsilon | b(\varepsilon; X, T) = \infty \}
\]

which is a function of the original sample \( X \) and statistics \( T \). Unless specified otherwise, we will work here with \( \varepsilon \) contamination.

For example, for estimating the location, say the mean is affected by a single outlier and has breakdown point zero. The median has breakdown point \( \frac{1}{2} \), and this is the highest possible breakdown point. A breakdown point of more than 50% is unreasonable because then it would appear that the good data have changed into bad data. Huber (1984) proved that the location M-estimator under regularity conditions has a breakdown point of 50%, or negligibly lower. If \( \psi \) is monotone, bounded and \( \psi(-\infty) = -\psi(\infty) \), the breakdown point for the M-estimate (1.21a) is \( \varepsilon^* = \frac{1}{2} \). The same is true if the scale parameter is unknown but estimated by a high breakdown point estimator “MAD”, that is,

\[
\hat{\mu}_M = \arg \min_{\mu} \sum_{i=1}^{n} \rho \left( \frac{x_i - \mu}{\text{MAD}} \right)
\]

Huber (1984) proved the following theorem for redescending M-estimates with bounded \( \rho \).

**Theorem 1.9** Let \( \min \rho \) occur at 0, \( \rho(0) = -1 \), \( \rho \) be a monotonic function increasing toward both sides, and \( \lim \rho(0) \). If we put

\[
\sum_{x_i \in X} \rho(x_i - T(X)) = -A
\]

then the \( \varepsilon \)-contamination breakdown point of \( T \) is

\[
\varepsilon^* = \frac{m^*}{n + m^*}
\]

where \( m^* \) is an integer satisfying \([A] \leq m^* \leq [A]\). If \( c < \infty \) such that \( \rho(x) = 0 \), for \( |x| \geq c \) we have \( m^* = [A] \).

Note that the breakdown point calculated using the above theorem depends on the \( \psi \) function and the data itself. For a redescending M-estimate and unbounded \( \rho \), the following argument is satisfied.

**Theorem 1.10** Assume \( \rho \) to be symmetric, \( \rho(0) = 0 \), and \( \rho \) increasing on both sides:

\[
\lim_{|t| \to \infty} \rho(t) = \infty
\]
Robust Statistics

but suppose
\[ \lim_{|t| \to \infty} \frac{\rho(t)}{|t|} = 0 \]

Moreover, assume \( \psi = \rho' \) to be continuous, and there exists \( t_0 \) such that \( \psi \) is weakly increasing in \( 0 < t < t_0 \) and weakly decreasing in \( t_0 < t < \infty \). The breakdown point of the \( \epsilon \) contamination for the M-estimate is then \( \epsilon^* = \frac{1}{2} \).

**Remark 1.11** If \( \psi \) is nonmonotone and redescends to 0, as defined in Section 1.4, and if \( \rho \) is unbounded, then the breakdown point of the redescending M-estimate is high, with \( \epsilon^* = \frac{1}{2} \).

If \( \rho \) is bounded, the breakdown point is less than 0.5 but negligible.

### 1.6 Linear Regression

Multivariate linear regression is an intuitive approach to nonlinear inference, using a linear approximation of a nonlinear model (Chapter 2). Robust methods for location can be extended to linear regression and subsequently to nonlinear regression, and outlier detection methods for linear regression can similarly be extended to identification of outliers in nonlinear regressions. In the rest of this chapter, the theories of linear regression, robust linear regression, and outlier detection will be discussed.

Assume \( n \) couples of a random sample \( z_i = (x_i, y_i) \), \( i = 1, \ldots, n \) are observed, where \( x_i \) is an independent (predictor) variable and \( y_i \) is a dependent (response) variable. The linear regression model is defined as:

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ i = 1, \ldots, n \]  

where \((\beta_0, \beta_1)\) are unknown parameters, \( \epsilon_i \) are errors, which in classical statistics are assumed to be independent with mean \( E(\epsilon_i) = 0 \), and to have constant variance \( Var(\epsilon_i) = \sigma^2 \). If there is a normal distribution, the MLE of parameter estimates can be employed. A more general linear model can be defined for when \( n \) observations contain \( p \) independent variables \((x_{i1}, \ldots, x_{ip}, y_i)\) and \( p \) unknown parameters \( \beta_1, \ldots, \beta_p \):

\[ y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \ i = 1, \ldots, n \]  

\[ = \beta' x_i + \epsilon_i \]  

The multivariate regression model can be written in a simple matrix form. Write the unknown parameters in a column vector \( \beta = [\beta_1, \ldots, \beta_p]' \). The \( i \)th predictor observation with coordinates \( x_i = [x_{i1}, \ldots, x_{ip}]' \) is inserted in the columns of a design matrix \( X_{n \times p} = [x_1, \ldots, x_n]' \), the response values
in a column vector $Y = [y_1, \ldots, y_n]$, and the residuals in an error vector $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]$. Finally the multivariate linear regression model can be written in matrix form as:

$$Y = X\beta + \varepsilon$$  \hspace{1cm} (1.34)

$$= \mu(\beta) + \varepsilon$$  \hspace{1cm} (1.35)

Constant terms can be readily added to this formulation, and ignoring them in our inference here does not reduce the generality. Let $z_i = (y_i, x_i), i = 1, \ldots, n$ be i.i.d. observations of the linear regression. Let $G_0(x)$ be the distribution of predictor $x_i$ and $F_0$ be the distribution of the error. Then the distribution of $z_i$ is given by

$$H_0(z) = G_0(x)F_0(y - \beta_0^tx_i)$$  \hspace{1cm} (1.36)

where $\beta_0$ is the true value of parameter $\beta$. Similarly, the true value of $\sigma$ is denoted by $\sigma_0$. Fitted values, defined as $\hat{y}_i = x_i\hat{\beta}$, and residuals, defined as $r_i = y_i - \hat{y}_i$, can be represented in matrix form too:

$$\hat{Y} = X\hat{\beta}, \quad r = Y - \hat{Y}$$  \hspace{1cm} (1.37)

The least squares method is popular with scientists because of its ease of use. The sum of square errors (SSE) is defined as:

$$SSE(\beta) = \sum_{i=1}^{n} (r_i^2)$$

$$= (Y - \hat{Y})'(Y - \hat{Y})$$

so the least squares estimate of the parameters is the answer to the optimizing sum of squares, which requires the following normal equations to be solved:

$$(X'X)\hat{\beta} = X'Y$$  \hspace{1cm} (1.38)

Therefore:

$$\hat{\beta}_{LS} = \arg \min_{\beta} SSE(\beta)$$  \hspace{1cm} (1.39)

$$= (X'X)^{-1}X'Y$$  \hspace{1cm} (1.40)

If the errors $\varepsilon_i$ are i.i.d. with mean zero and constant variance $\sigma^2$, the covariance matrix of the error vector is $\text{Cov}(\varepsilon) = \sigma^2I$, where $I$ is an $n \times n$ identity matrix, and we have the following identities:

$$E(Y) = X\beta$$

which means the linear regression problem (1.34) is a more general form of the location model (1.2). Therefore a robust M-estimate can be defined. Table 1.1 shows a summary of formulas in linear regression. These can be used to develop intuitive formulas for nonlinear regression by replacing the design matrix in linear regression with the gradient of the nonlinear function.
### Table 1.1 Linear regression formulas.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Formula</th>
<th>Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Location model</strong></td>
<td>( \mu(\beta) = X\beta )</td>
<td><strong>Unbiasedness and limit</strong></td>
<td>( E(\hat{\beta}_{LS}) = \beta ) ( Var(\hat{\beta}) = \sigma^2(X'X)^{-1} ) ( \hat{\beta} \sim N_p(0, \sigma^2(X'X)^{-1}) )</td>
</tr>
<tr>
<td><strong>Parameter estimate</strong></td>
<td>( \hat{\beta}_{LS} = (X'X)^{-1}X'Y )</td>
<td><strong>Variance</strong></td>
<td>( \hat{\sigma}^2 = \frac{SSE(\hat{\beta})}{n-p} )</td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>( \hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y )</td>
<td><strong>SS-estimate</strong></td>
<td>( SS(\hat{\beta}) = Y'(I - H)Y )</td>
</tr>
<tr>
<td><strong>SSE</strong></td>
<td>( SSE(\hat{\beta}) =</td>
<td></td>
<td>Y - X\hat{\beta}</td>
</tr>
<tr>
<td><strong>Hat matrix</strong></td>
<td>( H = X(X'X)^{-1}X' )</td>
<td><strong>100(1 - \alpha)%</strong></td>
<td>( \hat{Y}_0 \in (X'<em>0\hat{\beta} \pm s\sqrt{X'(X'X)^{-1}X_0}t</em>{(n-p,n/2)}) )</td>
</tr>
<tr>
<td><strong>Predictor covariance</strong></td>
<td><strong>Var(\hat{Y}) = \sigma^2H</strong></td>
<td><strong>Prediction interval</strong></td>
<td><strong>statistics</strong></td>
</tr>
<tr>
<td><strong>100(1 - \alpha)%</strong></td>
<td><strong>{ \beta : (\beta - \hat{\beta})'X'(\beta - \hat{\beta}) \leq pF_{p,n-p}^x }</strong></td>
<td><strong>(\hat{\beta} - \beta)'X'(\hat{\beta} - \beta) / \sigma^2 \sim \chi^2_p )</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Parameter CI</strong></td>
<td><strong>R^2 = SSR/SSTO</strong></td>
<td><strong>Error SS</strong></td>
<td>( (\hat{\beta} - \beta)'X'(\hat{\beta} - \beta) / ps^2 \sim F_{p,n-p} )</td>
</tr>
<tr>
<td><strong>R square</strong></td>
<td><strong>SSR = (\hat{Y} - \bar{Y}'(\hat{Y} - \bar{Y})</strong></td>
<td><strong>MSE = SSE/(n - p - 1)</strong></td>
<td><strong>F ratio</strong></td>
</tr>
<tr>
<td><strong>Regression SS</strong></td>
<td><strong>MSR = SSR/(p)</strong></td>
<td><strong>MSTO = SSE/(n - 1)</strong></td>
<td>( F = MSR/MSE \sim F(p, n - p - 1) )</td>
</tr>
<tr>
<td><strong>Total SS</strong></td>
<td><strong>SSTO = (Y - \hat{Y}'(Y - \bar{Y})</strong></td>
<td></td>
<td><strong>SS, sum of squares.</strong></td>
</tr>
</tbody>
</table>
1.7 The Robust Approach in Linear Regression

There are several robust methods for estimating the location and scale parameters in a linear regression. This section discuss the MM-estimate, which was defined by Yohai (1987). The logical path of the section is to end up with a simultaneous estimate of the location and scale parameters – the MM-estimate – which is a high breakdown point (HBP) estimate with asymptotic normality and strong consistency.

Simple linear regression (1.31) and general multiple linear regression (1.34), can both be considered as types of location and scale model (1.18) with multiple parameters. Consequently a robust M-estimate can be defined.

Analogous to Section 1.3.3, if the scale parameter \( \sigma \) is known, the location M-estimate is defined in the following way.

Let \( \rho \) be a real function satisfying the following assumptions.

A 1.12

(i) \( \rho(0) = 0 \)
(ii) \( \rho(-t) = \rho(t) \)
(iii) \( 0 \leq u < v \) implies \( \rho(u) \leq \rho(v) \)
(iv) \( \rho \) is continuous
(v) let \( a = \sup \rho(t) \) then \( 0 < a < \infty \)
(vi) let \( \rho(t) < a \) and \( 0 \leq u < v \) then \( \rho(u) < \rho(v) \).

Given a sample of size \( n \) \( (y_i, x_i) \), \( i = 1, \ldots, n \) and residuals defined as \( r_i(\beta) = y_i - \beta'x_i \), the M-estimate of location parameter \( \beta \), for known variance \( \sigma^2 \), is the solution of the optimization problem:

\[
\hat{\mu} = \arg \min_{\mu} \sum_{i=1}^{n} \rho \left( \frac{r_i(\beta)}{\sigma} \right)
\]

When the variance is unknown and the location has been previously calculated or is known, the M-estimate of scale \( \hat{\sigma} \) is defined as the answer to the equation:

\[
\frac{1}{n} \sum \rho \left( \frac{r_i(\beta)}{\hat{\sigma}} \right) = b
\]

where \( b \) is a constant, which may be defined by

\[
E_\phi(\rho(t)) = b
\]

We have to estimate the location and scale parameters simultaneously. In order to achieve efficiency and an HBP estimate, the MM-estimate algorithm for both location and scale parameters is defined as shown in Algorithm 1.4.
Algorithm 1.4 Robust M-estimate for linear regression

Stage 1: Take an estimate $\hat{\beta}_{0,n}$ of $\beta$ with high breakdown point, possibly 0.5.

Stage 2: Compute the residuals $r_i(\hat{\beta}_{0,n}) = y_i - x_i\hat{\beta}_{0,n}$.

Stage 3: Compute the M-estimate of scale $\hat{\sigma}_n = \sigma(r_i(\hat{\beta}_{0,n}))$ using a function $\rho_0$ obtained from

$$
\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left( \frac{r_i(\hat{\beta}_{0,n})}{\hat{\sigma}_n} \right) = b,
$$

where $b = E\phi(\rho_0(t))$. Define the constant $a$ such that:

1. $b/a = 0.5$

2. $a = \max \rho_0(t)$. As Huber proves, this implies that this scale estimate is (50%) HBP.

Stage 4: Let $\rho_1$ be another function satisfying regularity conditions

1. $\rho_1(t) \leq \rho_0(t)$

and

1. $\sup \rho_1(t) = \sup \rho_0(t) = a$

Let $\psi_1 = \rho'_1$. Then the MM-estimate $\hat{\beta}_{1,n}$ is defined as the solution of

$$
\sum_{i=1}^{n} \psi_1 \left( \frac{r_i(\beta)}{\hat{\sigma}_n} \right) x_i = 0
$$

which verifies

$$
S_n(\hat{\beta}_{1,n}) < S_n(\hat{\beta}_{0,n})
$$

where

$$
S_n(\beta) = \sum_{i=1}^{n} \rho_1 \left( \frac{r_i(\beta)}{\hat{\sigma}_n} \right)
$$

and $\rho_1(0/0)$ is defined as 0.

The MM-estimate was defined by Yohai (1987), who proved that the estimate is HBP, asymptotic normal, and strongly consistent. This method is used in other parts of the book to develop estimators for nonlinear regression.

Definition 1.16 Analogous to $\varepsilon$ contamination (Definition 1.8) we can define the contamination and breakdown point for linear regression. Let $Z_n = \{(y_i, x_i), i = 1, \ldots, n\}$ be a set of $n$ observations. The corrupted sample, with $m$ additional values $W_m$ where $Y_n \cup W_m$, and with sample of size $n + m$, contains observations of both samples. Let $T_n$ be the estimate corresponding to a sample of size $n$. The bias can be written as

$$
b(m, T, Z_n) = \sup_{Z} \| T_{m+n}(Z_n \cup W_m) - T_n(Z_n) \| $$
and the finite sample breakdown point (1.30) is

$$\varepsilon^*(Z_n, T) = \inf \{ m(m + n) | b(m, T, Z_n) = \infty \}$$

Yohai (1987) proved the following theorems (we omit the proofs, which may be found in Yohai’s papers (1987; 1985):

**Theorem 1.17**  For $\varepsilon$ contamination of size $m$, define $c_n$ as

$$c_n = \max_{\theta \in \mathbb{R}^p} \# \{ i : 1 \leq i \leq n \text{ and } \beta'X_i = 0 \} / n$$  \hfill (1.44)

Suppose assumptions A1.12–A1.15 are satisfied. Then, for given $\varepsilon < (1 - 2c_n) / (2 - 2c_n)$ and $k_0$, there exists a $k_1$ such that $m/n + m \leq \varepsilon$, and $S_{m+n} \leq k_0$ implies:

$$\inf_{||\theta|| \leq k_1} \sum_{i=1}^{m+n} \rho_1(r_i(\theta)/S_{m+n}) > \sum_{i=1}^{m+n} \rho_1(r_i(\hat{\theta}_{0,m+n})/S_{m+n})$$

for all samples $Z_n \cup W_m$.

Note that Theorem 1.17 implies that the absolute minimum of $S_n(\beta)$ exists and obviously satisfies (1.42) and (1.43). However, any other value of $\beta$ that satisfies (1.42) and (1.43) is a local minimum and is an MM-estimate with HBP and high efficiency.

**Theorem 1.18**  Suppose $\rho_0$ and $\rho_1$ satisfy assumptions A1.12–1.15 and $c_n < 0.5$. Then, if $\hat{\beta}_0 = \{ \hat{\beta}_{0,n} \}_{n \geq p}$ is any sequence of estimates that satisfies (1.43):

$$\varepsilon^*(\hat{\beta}_1, Z_n) \geq \min(\varepsilon^*(\hat{\beta}_0, Z_n), (1 - 2c_n) / (2 - 2c_n))$$  \hfill (1.45)

Note that $c_n \to 0$ as $n \to \infty$, so $(1 - 2c_n) / (2 - 2c_n) \to 0$. Therefore, the above theorem implies that if $\varepsilon^*(\hat{\beta}_0, Z_n)$ is asymptotically 0.5, $\varepsilon^*(\hat{\beta}_1, Z_n)$ is asymptotically 0.5.

**Theorem 1.19**  Let i.i.d. observations $z_i = (y_i, x_i)$ be given, with distribution $G_0$. Assume that $\rho_0$ satisfies A 1.12 and $\{ \hat{\beta}_{0,n} \}_{n \geq p}$ is a sequence of estimates that is strongly consistent for $\theta_0$. Then the M-estimate of scale $(\hat{\sigma}_n)$ obtained from Stage 2 of Algorithm 1.4 (see (1.41)) is strongly consistent for $\sigma_0$.

The consistency of $\hat{\beta}_{1,n}$ requires the following two assumptions to hold:

**A 1.20**  The function $g(a) = E_{F_0}[\rho_1((u - a)/\sigma_0)]$, where $\sigma_0$ is defined by $E_{F_0}[\rho_0(u/\sigma_0)] = b$, has a unique minimum at $a = 0$.

**A 1.21**  $P_{G_0}(\beta'x = 0) < 0.5$ for all $\beta \in \mathbb{R}^p$. 
If \( \rho_1 \) satisfies A 1.12, then the sufficient condition for 1.20 is given as below:

**A 1.22** The error distribution \( F_0 \) has density \( f_0 \) with the following properties:

(i) \( f_0 \) is even
(ii) \( f_0(u) \) is monotone nonincreasing in \( |u| \)
(iii) \( f_0(u) \) is strictly decreasing in \( |u| \) in the neighborhood of 0.

**Theorem 1.23** Let i.i.d. observations \( z_i = (y_i, x_i) \) be given with distribution \( G_0 \). Assume \( \rho_0 \) and \( \rho_1 \) satisfy assumptions A 1.12–A 1.21. Assume also that sequence \( \{\hat{\beta}_{0,n}\}_{n \geq p} \) is strongly consistent for \( \theta_0 \). Then any other sequence \( \{\hat{\beta}_{1,n}\}_{n \geq p} \) for which \( S_n(\hat{\beta}_{1,n}) < S_n(\hat{\beta}_{0,n}) \) (Equation 1.43) is strongly consistent too.

Asymptotic normality of the M-estimate requires four of the \( x_i \)s, but the MM-estimate requires a second moment. We need some additional assumptions:

**A 1.24** \( \rho_1 \) is an even, twice continuously differentiable function and there exists \( m \) such that \( |u| \geq m \) implies \( \rho_1(u) = a \).

**A 1.25** \( G_0 \) has second moments and \( V = E_{G_0}(x_i x_i') \) is nonsingular.

**Theorem 1.26** Let \( z_i, i = 1, \ldots, n \) be i.i.d. with distribution \( H_0 \). Assume \( \rho_1 \) satisfies 1.24 and \( G_0 \) satisfies 1.25. Let \( \hat{\sigma}_n \) be an estimate of error scale that converges strongly to \( \sigma_0 \). Let \( \hat{\beta}_n \) be a sequence of estimates that satisfies (1.42) and which is strongly consistent to the true value \( \beta_0 \). Then

\[
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{r_i(\hat{\beta}_{0,n})}{k_0 \hat{\sigma}_n} \right) = b
\]

where

\[
A(\psi, F) = E_F(\psi^2(u/\sigma_0))
\]

and

\[
B(\psi, F) = E_F(\psi'(u/\sigma_0))
\]

One way to choose \( \rho_0 \) and \( \rho_1 \) satisfying A 1.12, A 1.14 and A 1.15 is as follows. Let \( \rho \) satisfy A.1.14, and let \( 0 < k_0 < k_1 \). Let \( \rho_0(t) = \rho(t/k_0) \) and \( \rho_1(t) = \rho(t/k_1) \). Then Assumption 1.12iii implies \( \rho(t/k_1) < \rho(t/k_0) \).

The scale estimate \( \hat{\sigma}_n \) is the answer to the equation

\[
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{r_i(\hat{\beta}_{0,n})}{k_0 \hat{\sigma}_n} \right) = b
\]
The value of $k_0$ should be chosen such that $b/a = 0.5$, $b = E_{F_0}(\rho_1(t))$, and $a = \max \rho_0(t)$. Thus, from Theorem 1.18, the MM-estimate $\hat{\beta}_{1,n}$ is HBP.

The value of $k_1$ can be chosen to determine the asymptotic efficiency. The MM-estimate of $\beta$ in Stage 3 of the MM-estimate algorithm is computed as

$$
\hat{\beta}_{1,n} = \arg \min_{\beta} \sum_{i=1}^{n} \rho \left( \frac{r_i(\beta)}{k_1\hat{\sigma}_n} \right)
$$

or equivalently by the equation

$$
\sum_{i=1}^{n} \psi \left( \frac{r_i(\beta)}{k_1\hat{\sigma}_n} \right) x_i = 0
$$

From the asymptotic normal distribution of the MM-estimate $\hat{\beta}_n$ (Equation 1.46), if the errors have an $N(0, 1)$ distribution, $\sigma_0 = 1$ and the variance depends only on $k_1$ and not $k_0$ explicitly. To achieve a certain amount of efficiency, for example 95%, the Fisher information for linear regression is $I(\beta) = (1/\sigma_0^2)X'X$ and the efficiency of the MM-estimate is equal to

$$
eff(\hat{\beta}_{1,n}) = \frac{1/I(\beta)}{\Var(\hat{\beta}_{1,n})} = \frac{\sigma_0^2(X'X)^{-1}}{\sigma_0^2[A(\psi_1, F_0)/B^2(\psi_1, F_0)]V^{-1}} = \frac{B^2(\psi_1, F_0)}{A(\psi_1, F_0)}$$

### 1.8 S-estimator

S-estimate stands for scale estimate, which is the equivalent of the M-estimator for the scale parameter $\sigma$. It is used in defining the $\tau$-estimate. The S-estimate was defined by Rousseeuw and Yohai (1984) for linear regression, and they proved that it had HBP and asymptotic normal properties. The S-estimate is difficult to compute, although good algorithms have been proposed. The importance of the S-estimate is in helping to compute tuning constants in methods for obtaining HBP estimators.

Define the dispersion function $S(r_1, \ldots, r_n)$ to be a function of any sample $(r_1, \ldots, r_n)$ of real numbers. It is the solution of the Huber equation of the form:

$$
\sum_{i=1}^{n} \rho \left( \frac{r_i(\beta)}{S} \right) = nk
$$

where $k = E_{\phi}(\rho)$ is the mean of the real valued function $\rho$ and $\phi$ is the standard normal density function. The real valued function $\rho$ was defined in Chapter 1,
and for the S-estimate we assume it satisfies the following assumptions:

**A 1.27** \( \rho \) is symmetric and continuously differentiable and \( \rho(0) = 0 \).

**A 1.28** There exists \( c > 0 \) such that \( \rho \) is strictly increasing on \([0, c]\) and constant on \([c, \infty)\).

We want to estimate the function model parameter \( \beta \) and scale \( \sigma \). For each vector \( \beta \), we obtain residuals \( r_i(\beta) = y_i - x_i \beta \). The dispersion function \( S(r_1(\beta), \ldots, r_n(\beta)) \) is then calculated from Equation 3.7. After that, we define the S-estimate of \( \beta \) (denoted by \( \hat{\beta}_S \)) as:

\[
\hat{\beta}_{n,S} = \arg \min_\beta S(r_1(\beta), \ldots, r_n(\beta))
\]  

(1.48)

and the final scale estimator is

\[
\hat{\sigma}_{n,S} = S(r_1(\hat{\beta}_S), \ldots, r_n(\hat{\beta}_S))
\]  

(1.49)

S-estimators are affine equivariant, but their calculation is difficult.

Rousseeuw and Yohai (1984) studied the breakdown property of the S-estimate and proved that under certain conditions it is consistent and asymptotically normal (see the following theorems, but the proofs are omitted).

**Theorem 1.29** Breakdown point of the S-estimate. Assume:

**A 1.30** \( \frac{E_\phi(\rho)}{\rho(c)} = \frac{1}{2} \), where \( c \) is the consistency factor.

Under A 1.27–A 1.30, the solution (1.48) exists and the breakdown of the S-estimate is equal to

\[
\varepsilon^* = \left( \frac{n}{2} - p + 2 \right) / n
\]

which tends to 0.5 as \( n \to \infty \). If condition 1.30 is replaced by \( \frac{E_\phi(\rho)}{\rho(c)} = \lambda \), where \( 0 < \lambda < 1/2 \), then the corresponding S-estimate has a breakdown point tending to \( \lambda \).

For the proofs, see Rousseeuw and Yohai (1984).

**Theorem 1.31** Consistency Let \( \rho \) be a function satisfying 1.27 and 1.28, with derivative \( \rho' = \psi \). Assume that

(i) \( \psi(t)/t \) is nonincreasing for \( t > 0 \)
(ii) \( E(\|X\|) < \infty \).
Let $\hat{\beta}_{n,S}$ and $\hat{\sigma}_{n,S}$ be solutions of (1.48) and (1.49), respectively. Then:
\begin{align*}
\hat{\beta}_{n,S} &\to \beta_0, \text{ a.s.} \quad (1.50) \\
\hat{\sigma}_{n,S} &\to \sigma_0, \text{ a.s.} \quad (1.51)
\end{align*}

where $\beta_0$ and $\sigma_0$ are the actual values of the parameters. For the proofs, see Maronna and Yohai (1981, Thm 2.2 and 3.1). Note that the S-estimate satisfies the same first-order necessary conditions as M-estimates.

**Theorem 1.32**  Asymptotic normality.

Let $\beta_0 = 0$ and $\sigma_0 = 0$ for simplicity. If the conditions of Theorem 1.31 hold and also:

(iii) $\psi$ is differentiable in all but a finite number of points, $|\psi'|$ is bounded and $\psi' d\phi > 0$

(iv) $E[XX^T]$ is nonsingular and $E[||X||^3] < \infty$

then
\begin{align*}
n^{1/2}(\hat{\beta}_{n,S} - \beta_0) &\to \mathcal{N}\left(0, E[XX^T]^{-1} \int \frac{\psi^2 d\phi}{(\int \psi' d\phi)^2}\right) \quad (1.52) \\
n^{1/2}(\hat{\sigma}_{n,S} - \sigma_0) &\to \mathcal{N}\left(0, E[XX^T]^{-1} \int \frac{(\rho(y) - k)^2 d\phi(y)}{(\int y\psi(y) d\phi(y))^2}\right) \quad (1.53)
\end{align*}

The asymptotic distributions (1.52) and (1.53) are useful for practical problem inferences so they help us to compute the covariance matrix of parameter estimates. In particular, in Section 3.9 we extend them to nonlinear regression.

### 1.9 Least Absolute and Quantile Estimates

About a century ago, Edgeworth (1887) observed that methods of estimation based on minimizing sums of absolute residuals could be far superior to least squares methods under non-Gaussian error conditions. However, computation remained a major problem until simple linear programming algorithms were developed by researchers such as Wagner (1959) and Charnes et al. (1955), and algorithms for the $l_1$ norm were provided by Barrodale and Roberts (1974), Bartels and Conn (1980), and others. These algorithms are readily extended to linear quantile regression, as introduced by Koenker and Bassett (1978) and Koenker and D’Orey (1987), and can also be applied to nonlinear regression, as was done by Koenker and Park (1996).

The least absolute (LA) value estimates, known as $l_1$ norm estimates, minimize the $L_1$ criterion, defined as
\begin{equation}
\hat{\theta}_{LA} = \arg \min_{\theta} \sum_{i=1}^{n} |r_i(\theta)| \quad (1.54)
\end{equation}
where the residuals are defined as
\[ r_i(\beta) = (y_i - X_i \beta), i = 1, \ldots, n \] (1.55)
The LA estimator of \( \beta \) may be formulated as a linear program. The primal for the \( l_1 \) linear program may be written as:
\[
\max \{ 1_n^\prime \varepsilon^+ + 1_n^\prime \varepsilon^- \mid (\beta, \varepsilon^+, \varepsilon^-) \in \mathbb{R}^p \times \mathbb{R}^{2n}, X \beta + \varepsilon^+ - \varepsilon^- = Y \} \tag{1.56}
\]
where \( 1_n \) is an \( n \)-vector of 1s, \( \varepsilon_i^+ = \max(\varepsilon_i, 0) \), and \( \varepsilon_i^- = -\min(\varepsilon_i, 0) \). The simplex or interior point method can be used to solve the linear programming problem (1.56). Having distinguished the positive and negative parts of the residual vector, we are simply minimizing a linear function subject to linear constraints. The dual problem may be written as
\[
\max \{ Y^\prime d \mid d \in \Omega = d \in [-1, 1]^n, X^\prime d = 0 \}
\]
The dual variables, \( d \), may be viewed as Lagrange multipliers on the constraints; that is, the marginal costs of relaxing the constraints. If \( \varepsilon_i \) is nonzero, then \( d_i = \text{sgn}(\varepsilon_i) \); otherwise, when \( \varepsilon_i = 0 \), \( d_i \in (-1, 1) \). By complementary slackness there will be, barring degeneracy in the primal problem, exactly \( p \) of the \( \varepsilon_i \)'s equal to zero at an optimum. Consequently, \( p \) of the \( d_i \)'s are not equal to \( \pm 1 \). For more detail see Koenker and D'Orey (1987) and Koenker and Bassett (1978) for linear, and Kennedy and Gentle (1980) for nonlinear regression formulation and computation.

The quantile estimate can be applied in linear regression, a process known as linear quantile regression. Suppose that the \( \alpha \)th conditional quantile function is \( Q_{Y|X}(\alpha) = X \beta_\alpha \). Given the distribution function of \( Y \), \( \beta_\alpha \) can be obtained:
\[
\beta_\alpha = \arg \min_\beta E[\rho(Y - X \beta)] \quad \rho_\alpha(u) = u(\alpha - I(u < 0))
\]
For an indicator function \( I \) and loss function \( \rho \) we have
\[
\min_u E[\rho(Y - u)] = \min(\alpha - 1) \int_{-\infty}^u (y - u) dF_Y(y) + \alpha \int_u^\infty (y - u) dF_Y(y)
\]
which is the \( \alpha \)th quantile of the distribution of \( Y \). Replacing the sample analog, symmetric \( l_1 \) criterion, with an asymmetric linear criterion, gives an estimate of \( \beta \nabla_{\alpha}(\beta) = \sum_{i=1}^n \rho_\alpha[r_i(\beta)] \quad \rho_\alpha(u) = y(\alpha - I(u < 0))
\]
In this way, the quantile regression of Koenker and Bassett (1978) will be obtained. The dual problem is now
\[
\max \{ Y^\prime d \mid d \in \Omega = d \in \alpha, X^\prime d = 0 \}
\]
1.10 Outlier Detection in Linear Regression

In this section, methods for outlier detection in linear regression are discussed. Some of these methods are used in Chapter 6 to identify outliers in nonlinear regressions.

There are several approaches to detecting outliers. There are statistics that measure the influences of different kinds of data points, or that measure the distance of the data points from the fitted regression line. Another approach is to isolate outlying data using robust methods. Since the robust fitted regression line is not swamped by outlier data, the computed residuals of the outliers can be large, revealing the large distance to the regression line.

Another popular method is to study what happens when a single observation or a group of observations is deleted.

Consider the general linear regression (1.34). Using the hat matrix, defined by $H = X(X'X)^{-1}X'$, the prediction expression (see Table 1.1) can be written as:

$$\hat{Y} = HY$$

$$\hat{y}_i = h_{ii}y_i + \sum_{j \neq i=1}^n h_{ij}y_j$$  \hspace{1cm} (1.57)

In this equation, if $h_{ii}$ is large relative to the remaining terms, the fitted value $\hat{y}_i$ is dominated by response $\hat{y}_j$, so $h_{ij}$ is interpreted as the amount of influence or leverage of $y_j$ on $\hat{y}_i$. Following this reasoning, Hoaglin and Welsch (1978) suggested direct use of $h_{ij}$ as a diagnostic to identify high leverage points.

A common way of developing an influence detection method is to refit a model by deleting a special case or a set of cases. The amount of change of certain statistics – the parameter estimates, predicted likelihoods, residuals, and so on – can be observed for the measure when recalculated with the $i$th data point removed. The notation ($-i$) is used for each removed observation. Subsequently, the new estimates can be utilized in the computation of the influence measures.

Some statistical measures for identifying outliers are briefly discussed in the following sections.

1.10.1 Studentized and Deletion Studentized Residuals

Studentized residuals (hereafter referred to as $t_i$) are used for identifying outliers. They are standardized by dividing the residuals by their standard error. Using the prediction form given by the hat matrix and the diagonal form of the covariance matrix of residuals (see Table 1.1), the standardized residuals can be written as:

$$t_i = \frac{r_i}{\sigma \sqrt{1 - h_{ii}}}.$$  \hspace{1cm} (1.58)
Note that the variance $\sigma^2$ is unknown, so we replace it by its estimate $\hat{\sigma}$. The studentized residuals are then defined as

$$t_i = \frac{r_i}{\hat{\sigma} \sqrt{1 - h_{ii}}} \quad (1.59)$$

where $h_{ii}$ is the diagonal of the hat matrix $H$. The deleted studentized residuals ($d_i$) are defined as

$$d_i = \frac{r_i}{\hat{\sigma}_{(-i)} \sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}_{(-i)}$ is the estimated standard deviation in the absence of the $i$th observation. The residuals, denoted by $r_i = y_i - f(x_i; \hat{\theta})$, are obtained from the ordinary least squares-, M- or MM-estimates. The $i$th observation is considered an outlier if $|t_i|$ or $|d_i| > 2.5$ or $3$ (Anscombe and Tukey 1963; Srikantan 1961)

### 1.10.2 Hadi Potential

Hadi (1992) proposed the Hadi potential, given by $p_{ii}$, to detect high leverage points or large residuals:

$$p_{ii} = \frac{h_{ii}}{1 - h_{ii}}$$

He proposed the cut-off point for $p_{ii}$ to be $\text{Median}(p_{ii}) + c \times \text{MADN}(p_{ii})$, where MADN represents the normalized mean absolute deviance, defined by:

$$\text{MADN}(p_{ii}) = \frac{\text{Median}\{p_{ii} - \text{Median}(p_{ii})\}}{0.6745} \quad (1.60)$$

c is an appropriately chosen constant, such as 2 or 3.

### 1.10.3 Elliptic Norm (Cook Distance)

The Cook distance (denoted by $CD$), which was defined by Cook and Weisberg (1982), is used to assess influential observations. An observation is influential if the value of $CD$ is greater than 1. They defined $CD$ as:

$$CD_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T (X^TX)(\hat{\beta} - \hat{\beta}_{(-i)})}{p\hat{\sigma}^2}$$

where $\hat{\beta}_{(-i)}$ is the parameter estimate when the $i$th observation is removed. It can be shown that the $CD_i$ can be written as:

$$CD_i = \frac{t_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

since this form avoids deleting observations and is numerically more efficient, especially when we extend it to nonlinear regression. The cut-off point is equal to 1; that is, the expectation of a 50% confidence ellipsoid of parameter estimates.
1.10.4 Difference in Fits

Difference in fits (DFFITS) is another diagnostic parameter used when measuring influence and was defined by Belsley et al. (1980). For the $i$th observation, DFFITS is defined as

$$DFFITS_i = \left( \sqrt{\frac{h_{ii}}{1 - h_{ii}}} \right) |d_i|$$

Belsley et al. considered an observation to be an outlier when DFFITS exceeds a cut-off point of $2\sqrt{p/n}$.

1.10.5 Atkinson’s Distance

The Atkinson distance (denoted by $C_i$ for observation $i$) was developed by Atkinson (1981), who also studied its properties (Atkinson 1982; 1986). It is used to detect influential observations, defined as:

$$C_i = \left( \sqrt{\frac{n - p}{p} \frac{h_{ii}}{1 - h_{ii}}} \right) |d_i|$$

The cut-off point is suggested to be equal to 2.

1.10.6 DFBETAS

DFBETAS is a measure of how much an observation has affected an estimate of a regression coefficient (there is one DFBETA for each regression coefficient, including the intercept). For linear regression with design matrix $X$, the DFBETAS for the $i$th data point and $j$th parameter can be computed as:

$$DFBETAS_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{\sigma}^2(\hat{\beta}) (X^T X)^{-1}}}$$

(1.61)

For small/medium datasets, an absolute value of 1 or greater is ’suspicious’. For large datasets, absolute values larger than $2/\sqrt{n}$ are considered highly influential.