TABLE 1.1
Determinants of Marital Coital Frequency

| | *Model 1* | *Model 2* | *Model 3* | *Model 4* |
|---|---|---|---|---|
| *Period* | −.72*** | −.67*** | –3.06** | –0.08 |
| *Log Wife's Age* | 27.61** | 13.56 | 29.49 | –1.62 |
| *Log Husband's Age* | –6.43 | 7.87 | 57.89 | –5.23 |
| *Log Marital Duration* | –1.50*** | –1.56*** | –1.51* | 1.29 |
| *Wife Pregnant* | –3.71 | –3.74*** | –2.88*** | –3.95* |
| *Child Under 6* | –0.56** | –0.68*** | –2.91*** | –0.55** |
| *Wife Employed* | 0.37 | 0.23 | 0.86 | 0.02 |
| *Husband Employed* | –1.28** | –1.10** | –4.11*** | –0.38 |
| $R^2$ | 0.0475 | 0.0612 | 0.2172 | 0.0411 |
| $n$ | 2062 | 2055 | 243 | 1812 |

SOURCE: Adapted from Kahn and Udry (1986: table 1).
NOTES: Model 1: Jasso's original analysis; Model 2: four "miscodes" and four other outliers dropped; Model 3: Marital duration $\leq$ 2 years (excluding miscodes and outliers); Model 4: Marital duration $>$ 2 years (excluding miscodes and outliers).
$^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$.

This example illustrates three important points. First, it shows the value of using diagnostic tools to uncover potentially problematic observations. Second, it shows how outliers can influence regression estimates even in large data sets. Third, the debate reflects the fact that there is no universally accepted method for handling unusual observations. The decision on what action should be taken when influential observations are detected should be based on substantive knowledge. In other words, the researcher must make a judgment call. With respect to this particular example, I leave it to those with better substantive knowledge of the topic to decide the best way to handle the outliers. It is sufficient for the purpose of this book to show that, despite a sample of more than 2,000 observations, as few as eight outliers drastically altered the results. If one decides that the observations should not be ignored, they can be handled by removing them, as did Kahn and Udry, or by robust regression.

## 2. IMPORTANT BACKGROUND

We now turn to various concepts important to assessing the robustness of an estimator. In this regard, bias, consistency, efficiency, breakdown point, and the influence function will be defined. All of these will be used throughout the book.

## Bias and Consistency

Assume a sample, *Z,* with *n* observations. Let $T_n(Z_1, \ldots, Z_n)$ with probability distribution P represent an estimator for the parameter $\theta$. In other words, applying *T* to *Z* gives the estimate of the population parameter:

$$T(Z) = \hat{\theta} \qquad [2.1]$$

The estimator is unbiased if

$$E[T(Z)] = E(\hat{\theta}) = \theta. \qquad [2.2]$$

In other words, the *average of an unbiased statistic equals the population parameter.* It follows, then, that the bias of an estimator $T(Z) = \hat{\theta}$ is given by

$$\text{bias } E[T(Z) - \theta]. \qquad [2.3]$$

Unbiasedness is certainly important, but consistency is also of concern when determining the "best" estimator to use. *An estimator $\hat{\theta}$ is consistent if it converges to θ as the sample size grows to infinity.* We can also consider consistency in terms of the mean squared error (MSE) of an estimate. In this respect, $\hat{\theta}$ is consistent if

$$\lim_{n\to\infty} \text{MSE}\left(\hat{\theta}\right) = 0. \qquad [2.4]$$

## Breakdown Point

The breakdown point (BDP)[1] is a *global measure of the resistance* of an estimator. More specifically, it is the smallest fraction or percentage of discrepant data (i.e., outliers or data grouped at the extreme end of the tail of the distribution) that the estimator can tolerate without producing an arbitrary result (Hampel 1974; Huber 2004). Assume all possible "corrupted" samples $Z'$ that replace *m* observations in the data set with arbitrary values (i.e., observations that do not fit the general trend in the data). The maximum effect[2] that could arise from these substitutions is

$$\text{effect } (m; T, Z) = \overset{\text{sup}}{Z'} \left\| T(Z') - T(Z) \right\|, \qquad [2.5]$$

where the supremum is over all possible $Z'$. If the effect (*m; T, Z*) is infinite, the *m* outliers have an arbitrarily large impact on *T*. In other words, the estimator "breaks down" and fails to adequately represent the pattern in the bulk of the data. More generally, the breakdown point for an estimator $T$ for a finite sample $Z$ is defined as

$$BDP(T,Z) = \min\left\{\frac{m}{n} : \text{effect } (m;T,Z) \text{ is infinite}\right\}. \qquad [2.6]$$

The highest possible breakdown point for an estimator is 50%, which indicates that as many as half the observations could be discounted. A breakdown point higher than 0.5 is undesirable because it would mean that the estimate could pertain to less than half of the data.

The goal of a robust estimator is to sufficiently capture the pattern in the bulk of the data. In other words, a breakdown point greater than zero is a desirable attribute. In fact, Hampel et al. (1986) argue that data sets typically contain as much as 10% of observations that deviate from the general pattern characterized by the bulk of the data, suggesting that a robust estimator should have a breakdown point of at least 10%. As we shall see later, however, some of the first proposed robust regression estimators have a breakdown point of 0 or very close to it.

## Influence Function

Originally proposed by Hampel (1974; see also Hoaglin, Mosteller, and Tukey 1983:350–358; Jurečková and Picek 2006:27–32), the *influence function* of an estimator measures the impact of a single observation $y_i$ that contaminates the theoretically assumed distribution $F$ of an estimator $T$. In other words, whereas the breakdown point measures global robustness, the influence function (*IF*) measures *local resistance or, more specifically, infinitesimal perturbations* on the estimator. Also referred to as the influence curve (or sensitivity curve when viewed with respect to a single sample), the influence function for an estimator $T$ is defined by

$$IF(Y,F,T) = \lim_{\lambda \to 0} \frac{T\left[(1-\lambda)F + \lambda\delta_y\right] - T(F)}{\lambda}, \qquad [2.7]$$

where $\delta_Y$ is the point of contamination at $y$ (i.e., at $y$ and 0 otherwise) with probability mass $\lambda$. In other words, $\lambda$ gives the proportion of contamination at $y$. Simply put, the *IF* indicates the change in an estimate caused by adding arbitrary outliers at the point $y$, standardized by the proportion of contamination.

A bounded influence function is a desirable attribute of a robust estimator because it means that the influence of a particular observation can get only so high. An unbounded influence function allows the influence of ‘‘contaminated’’ observations to continue to grow, regardless of how unusual they are. In other words, there is no limit on the effect of discrepancy. As we shall see later, the influence function for OLS regression is unbounded and proportional to the size of the residual, meaning that a highly discrepant residual can completely destroy the OLS estimator. Many early robust regression methods also have unbounded influence functions, resulting in a resistance that is sometimes no better than that of OLS. Most robust estimators commonly employed today, however, have both a high breakdown point and bounded influence function.

## Relative Efficiency

Another important concept for understanding robust estimation is efficiency. If the goal is to make inferences about a larger population from sample data, we desire an unbiased estimator that is as efficient as possible. In the strictest sense, the efficiency of an estimator is determined by the ratio of its minimum possible variance to its actual variance. Only when the ratio is equal to one—that is, when it has the lowest possible variance—is an estimate considered efficient.[3] An estimator is asymptotically efficient if it reaches efficiency with large samples. More generally, an estimator is considered to be efficient if its sampling variance is relatively small, resulting in small standard errors. It follows that some estimators are more efficient than others, and thus the concept of *relative efficiency* is useful for assessing competing estimators.

For most kinds of estimation, there is one estimator that has maximum efficiency under some particular assumptions. We can use this estimator as a benchmark to which we compare the efficiency of other estimators. Assume that we have two estimators $T_1$, and $T_2$, for the population parameter $\theta$. If $T_1$ has maximum efficiency and $T_2$ is less efficient, $T_1$ will also have a smaller mean squared error. The relative efficiency of $T_2$ is determined by the ratio of its mean squared error to the mean squared error for $T_1$:

$$\text{Efficiency } (T_1, T_2) = \frac{E\left[(T_2 - \theta)^2\right]}{E\left[(T_1 - \theta)^2\right]} \qquad [2.8]$$

If the assumptions of linearity, constant error variance, and uncorrelated errors are met, OLS estimates are the most efficient of unbiased linear estimators. As a result, relative efficiency of robust estimators is assessed in comparison to the OLS estimators under these conditions. Although no robust regression estimator is more efficient than OLS under these conditions, several estimators are nearly as efficient, and at the same time have the desirable property of high resistance to outliers. The relative efficiency of robust regression estimators should be considered cautiously, however, because it is asymptotic efficiency that is typically assessed (Ryan 1997:354). In other words, relative efficiency is meaningful only with sufficiently large sample sizes. Little is known about the small sample properties of most robust regression estimators, resulting in the common practice of using bootstrapping to find standard errors in these situations.

## Measures of Location

Although there are various types of regression, all predict conditional values of a dependent variable from some predictor(s) by taking into account some measure of location and scale of the response variable. OLS, for example, estimates the conditional mean of a dependent variable *y* from one or more independent variable *x*s. Because OLS is based on the mean, which is not resistant to outliers, its estimates can also be affected by outliers. Similarly, estimates from generalized linear models (GLMs) are not completely resistant to outliers because they estimate the conditional mean of a linear predictor. Robust regression methods rely on more robust measures of location and/or scale. It is helpful, then, to discuss various measures of location and scale before exploring the regression techniques that use them.

A measure of *location* is a quantity that characterizes a position in a distribution. Typically, measures of center are of most concern, although other measures of location (quantiles, for example) can also be considered. Assume a random variable *Y* with distribution *F*. An estimate $\theta(Y)$ is a measure of location of *F* if, for any constants *a* and *b*, four conditions[4] are met (Wilcox 2005:20–21):

a. $\theta(Y+a)=\theta(Y)+a$
b. $\theta(-Y)=-\theta(Y)$
c. $Y\geq\theta$ implies $\theta(Y)\geq 0$
d. $\theta(bY)=b\theta(Y)$

Condition (a), which requires that when a constant is added to all values of *Y*, the measure of location will increase by the same amount, is referred to as *location equivariance*. Taken together, Conditions (a), (b), and (c) require that the value of the measure is within the range of *Y*. Condition (d) means that the measure has *scale equivariance*. In other words, if all values of *Y* are multiplied by a particular value (i.e., if the scale is altered), the measure of location will be altered by the same factor.

## The Mean

The most common measure of location is the mean. Consider independent observations $y_i$ and a simple model estimating the center $\mu$ of a population distribution

$$y_i = \mu + e_i, \tag{2.9}$$

where the $e_i$ represent the residuals. If the underlying distribution is normal, the sample mean is the maximally efficient estimator of $\mu$, producing the fitted model

$$y_i = \bar{y} + e_i. \tag{2.10}$$

Despite its widespread use, including in OLS regression, the mean is not a robust measure of location. If the distribution has heavy tails or outliers, the mean is less efficient than many other measures of center and, more important, can often be misleading. Even the addition of a single badly miscoded observation can alter its estimate.

Consider the following five observations for the variable $y$:

$$\begin{array}{lll} y_1 = 3 & y_2 = 3 & y_3 = 4 \\ y_4 = 5 & y_5 = 5 & \end{array}$$

Applying the well-known formula for the sample mean, $\bar{y} = \frac{1}{n}\sum\limits_{i=1}^{n} y_i$, produces $\bar{y} = 4$. We now replace just one of the observations, $y_3$, with a "bad" observation (assume that it is a miscode), giving the following values of $y$:

$$\begin{array}{lll} y_1 = 3 & y_2 = 3 & y_3 = 44 \\ y_4 = 5 & y_5 = 5 & \end{array}$$

For these new data, $\bar{y} = (3 + 3 + 44 + 5 + 5)/5 = 12$. The mean has been dramatically pulled toward the outlier, taking a value three times larger than when the outlier is excluded. In fact, the "contaminated" mean is much larger than any of the observed values *except* the "bad" observation.

Because even a single observation can cause the mean to break down, its breakdown point is $BDP = \frac{1}{n}$, and thus effectively 0 when $n$ is large. Just as problematic, the influence of each observation on the mean is proportional to the size of $y$. The mean is found by minimizing the least squares objective function:

$$\sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 0 \tag{2.11}$$

Taking the derivative with respect to $y$ produces the influence function

$$IF_{\bar{y}}(y) = 2y. \tag{2.12}$$

This, of course, is not an attractive attribute for data that are not "well behaved" (i.e., that have outliers or a heavy tail).

One strategy to combat the influence of outliers on the mean is to use a two-step procedure, where the outliers are first identified and removed before calculating the mean. Rather than calculating the mean for the distribution excluding the outliers, Hampel (1974) shows that using a robust measure of location is usually a better way to proceed. Many measures of location are less vulnerable than the mean to outliers. In other words, many estimators are more robust.

## $\alpha$-Trimmed Mean

A relatively robust measure of center is the *trimmed mean,* which reduces the impact of outliers or heavy tails by removing the observations at the tails of the distribution. Let $y_1, \ldots y_n$ represent observations on a variable from a random sample. We start by ordering the values of $y$ from lowest to highest, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, and determining the desired amount of trimming, $0 = \alpha < 0.5$. the mean is then calculated for all observations *except* the $g$ smallest and largest observations $g = [\alpha n]$, where $[\alpha n]$ is rounded to the nearest integer. The formula for the trimmed mean can be written as[5]

$$y_t = \frac{y_{(g+1)} + \cdots + y_{(n-g)}}{n - 2g}. \qquad [2.13]$$

The breakdown point of the trimmed mean is determined by the amount of trimming, and thus is $BDP = \alpha$. A simple rule of thumb is to remove 10% of the observations from each tail of the distribution (i.e., set $\alpha = 0.2$). Leger and Romano (1990) further suggest calculating the mean for $\alpha = 0, 0.1$, and 0.2 and choosing the value that gives the lowest standard error for the final calculation. The amount of trimming also determines the influence function. Unlike for the mean, the influence for the trimmed mean is bounded, although there are marked increases at $y_\alpha$ and $y_{1-\alpha}$.[6] Its influence function can be written as

$$IF_{\bar{y}_t}(y) = \begin{cases} \frac{y_\alpha - \hat{\mu}_t}{1-2\alpha} & \text{for } y < y_\alpha \\ \frac{y - \hat{\mu}_t}{1-2\alpha} & \text{for } y_\alpha \leq y \leq y_{1-\alpha} \\ \frac{y_{1-\alpha} - \hat{\mu}_t}{1-2\alpha} & \text{for } y > y_{1-\alpha} \end{cases} \qquad [2.14]$$

where $\hat{\mu}_t$ is the trimmed mean (see Wilcox 2005:29). The relative efficiency of the trimmed mean depends on the distribution. If the distribution is normal and too much trimming is done, precision will be reduced because it results in greater spread relative to the smaller *n,* thus increasing the estimate of the spread of its sampling distribution. On the other hand, if the distribution has heavy tails and extreme outliers, trimming can result in improved efficiency because the variance of *y*—and hence the estimated variance of the sampling distribution of its mean—is decreased. Judgments on the amount of trimming should be made only after careful examination of the distribution.

## The Median

The *median M* is simply the value of *y* that occupies the middle position when the data are ordered from smallest to largest. To find the median, we start by ordering the observations from lowest to highest value, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$. The median is given by

$$M = y_{((n+2)/2)} \quad \text{if } n \text{ is an odd number}$$

and

$$M = .5y_{(n/2)} + .5y_{(n/2+1)} \quad \text{if } n \text{ is an even number.}$$

Equivalently, the median minimizes the absolute values objective function

$$\sum_{i=1}^{n} |y_i - \hat{\mu}| = 0. \tag{2.15}$$

Taking the derivative of Equation 2.15 gives the shape of the influence function

$$IF_{\mathrm{M}}(y) = \begin{cases} 1 & \text{for } y > 0 \\ 0 & \text{for } y = 0 \\ -1 & \text{for } y < 0. \end{cases} \tag{2.16}$$

As the bounded influence function indicates, the median is highly resistant to outliers. Its robustness is also reflected in its breakdown point of $BDP = 0.5$. The disadvantage of the median is that it has relatively low efficiency compared to the mean when the distribution is normal. In these situations, the sampling variance for the mean is $\sigma^2/n$, whereas the sampling variance for the median is $\pi/2 = 1.57$ times larger at $\pi\sigma^2/2n$ (Kenney and Keeping 1962:211).

# Measures of Scale

Let *Y* represent a random variable. A measure of scale is any nonnegative functional $\tau(Y)$ that satisfies the following conditions (Wilcox 2005:34)[7]:

a. The measure is *scale equivariant,* meaning that $\tau(aY) = a\tau(Y)$, where $a$ is a constant that is greater than 0.
b. The measure is *location invariant,* meaning that $\tau(Y + b) = \tau(Y)$, where $b$ is a constant.
c. The measure is *sign invariant*, $\tau(Y) = \tau(-Y)$.

There are too many measures of scale to include them all, so we concentrate on those that are most relevant to robust regression. We explore mostly how outliers affect the magnitude of the scale estimate, paying little attention to efficiency. For more discussion on the latter, see Wilcox (2005).

## Standard Deviation

The most commonly employed measure of scale is the standard deviation $s$, which is defined by

$$s_y = \sqrt{\frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}. \qquad [2.17]$$

If the distribution of $y$ is normal, this is the most appropriate measure of scale because of its superior efficiency. On the other hand, the standard deviation is not robust to heavy-tailed distributions or distributions with outliers. Because it is based on the mean—which has an unbounded influence function and breakdown point of 0—the standard deviation inherits these qualities. As a result, robust regression techniques typically use other measures of scale.

## Mean Deviation From the Mean

The mean deviation from the mean (MD), sometimes known more simply as the *mean deviation*, is given by

$$MD = \frac{\sum\limits_{i=1}^{n}|y_i - \bar{y}|}{n}. \qquad [2.18]$$

The MD is relatively efficient compared to the standard deviation when the distribution of $y$ has heavy tails, but it also has the undesirable property of a breakdown point of 0 and an unbounded influence function. Although important for some early robust regression techniques, the MD should generally be seen as obsolete given that there are now much more robust measures of scale.

## Mean Deviation From the Median

The mean deviation from the median, MDM, is a slight improvement over the MD in terms of robustness. Rather than find the absolute difference of $y$ from the mean, the MDM finds the absolute differences from the median $M$, resulting in

$$MDM = \frac{\sum_{i=1}^{n} |y_i - M|}{n}. \qquad [2.19]$$

Although it also uses the median, MDM still relies on mean deviations, and thus has a breakdown point of $BDP = 0$ and an unbounded influence function (see Wilcox 2005:35 for more details). In other words, the mean deviation from the median is not immune to extreme outliers and heavy tails, and thus it is not ideal for use in robust regression.

## Interquartile Range

The $q$-quantile range $QR_q$ is a set of bounded influence measures of scale that can have a very high breakdown point. Any particular $q$-quantile range is given by

$$QR_q = y_{1-q} - y_q, \text{ where } 0 < q < .5.$$

Setting $q = .25$ (i.e., the difference between the .25 and .75 quantiles) produces the interquartile range (IQR), which, with a breakdown point of $BDP = 0.25$, is the most robust and thus most commonly used of the quantile ranges (Wilcox 2005:35–36). The influence function for the IQR is given by the influence function at the third quartile minus the influence function at the first quartile (i.e., $IF_{.75} - IF_{.25}$):

$$IF_{IQR}(y) = \begin{cases} \frac{1}{f(y_{.25})} - C & \text{if } y < y_{.25} \text{ or } y > y_{.75} \\ -C & \text{if } y_{.25} \leq y \leq y_{.75} \end{cases} \qquad [2.20]$$

where

$$C = q\left\{\frac{1}{f(y_{.25})} + \frac{1}{f(y_{.75})}\right\} \qquad [2.21]$$

The high breakdown point and bounded influence function of the IQR are desirable properties, leading to its use in some early robust regression techniques. It still plays a role in quantile regression, which will be introduced later. There are more robust measures of scale, however, so despite its simplicity, the IQR is seldom incorporated in more recent developments in robust regression.

### Median Absolute Deviation

The median absolute deviation (MAD) is defined by

$$\text{MAD} = \text{median}|y_i - M|.$$

Based entirely on variation around the median, the MAD is far more resistant to outliers than the standard deviation and measures of absolute deviation associated with the mean.[8] The MAD achieves the highest possible breakdown point of $BDP = 0.5$ and has a bounded influence defined by

$$IF_{MAD}(y) = \frac{\text{sign}(|y-M| - MAD) - \frac{f(M+MAD) - f(M-MAD)}{f(M)} \text{sign}(y-M)}{2[f(M+MAD) + f(M-MAD)]} \qquad [2.22]$$

where $f(y)$ is the probability density function for $y$ (see Wilcox, 2005:35 for more details). An attractive attribute of the MAD is that it can be adjusted to ensure consistency for large sample sizes under the assumptions that $y \sim N(\mu, \sigma^2)$ by multiplying by 1.4826 (approximately $1/\Phi^{-1}(3/4)$, where $\Phi$ is the normal probability density function). All of these attributes make the MAD an attractive measure of scale for robust regression, at least as an initial estimate.

## *M*-Estimation

*M*-estimation includes a large class of estimators that generalize the idea of maximum likelihood to robust measures of scale and location (Huber 2004). *M*-estimation is also the foundation for many robust regression estimates, including those classified as *M*-estimates, *GM*-estimates, *S*-estimates, and *MM*-estimates. All of these will be discussed in Chapter 4. When formulated properly, *M*-estimates are very robust, especially with respect to estimating location. They are also relatively efficient compared to other robust measures for large samples ($n \geq 40$), becoming more efficient as $n$ gets larger (Hogg 1974; see also Wu 1985).

Assume that $y_1, \ldots, y_n$ is independently and identically distributed according to $F(y;\theta)$. Let $T_n(y_1, \ldots, y_n)$ be an estimate of an unknown parameter $\theta$ that characterizes the distribution $F(y;\theta)$. The likelihood of the estimator is given by

$$L(\theta; y_i, \ldots, y_n) = \prod_{i=1}^{n} f(y;\theta), \qquad [2.23]$$

where $f(y;\theta)$ is the probability density function corresponding to $F(y;\theta)$. The maximum likelihood estimator is the value of $\theta$ that maximizes the likelihood function or, equivalently, minimizes the objective function $\rho(y;\theta)$:

$$-\log l = \sum_{i=1}^{n} \rho(y;\theta) \qquad [2.24]$$

Restricting the objective function $\rho(y;\theta)$ to any function that is differentiable with an absolutely continuous derivative $\Psi(.)$ results in the maximum likelihood estimator $T_n$,

$$\sum_{i=1}^{n} \Psi(y;\theta) = 0, \qquad [2.25]$$

where

$$\begin{aligned}\Psi(y;\theta) &= -(\partial/\partial\theta)\rho(y;\theta)\\ &= (\partial/\partial\theta)\log f(y;\theta).\end{aligned} \qquad [2.26]$$

In order for the maximum likelihood estimator—or *M*-estimator—to be uniquely determined, $\rho(y;\theta)$ must be strictly convex, and thus the score function $\Psi(y;\theta)$ must be strictly increasing. Using $\rho(y;\theta) = -\log f(y;\theta)$ gives the ordinary maximum likelihood estimate (see Huber 2004: chap. 3).

*M*-estimates take on many different forms, the properties of which are determined by the choice of $\rho(.)$ or, equivalently, $\Psi(.)$. If $\Psi(.)$ is unbounded, the breakdown point of the estimator is $BDP = \lim_{n\to\infty} BDP = 0$. Conversely, if $\Psi(.)$ is odd and bounded, and thus $\rho(.)$ is symmetric around 0, the breakdown point of the estimator is $BDP = 0.5$. The score function $\Psi(.)$ has the same shape as the influence function proposed by Hampel (1974). More specifically, $IF(y;F,T) = \Psi(y)/\gamma(F)$, where $\gamma(F) = \int f(y)\, d\Psi(y)$. The proportionality constant $[\gamma(F)]^{-1}$depends on both $\Psi$ and the probability density function $f(y)$. In other words, the *IF* is the negative of the score function (see Jurečková and Sen 1996; Hoaglin et al. 1983:356).

## *M*-Estimation of Location

Consider the population mean $\mu$ as the expectation of the random variable *Y*. Let $\rho(y-\hat{\mu})$ be an objective function that measures distance from an estimate of location $\hat{\mu}$. The *M*-estimate is found by minimizing the objective function

$$\sum_{i=1}^{n} \rho(y;\theta) = \sum_{i=1}^{n} \rho\left(\frac{y_i-\hat{\mu}}{cS}\right), \qquad [2.27]$$

where $S$ is a measure of scale of the distribution and $c$ is a tuning constant that adjusts the degree of resistance of the estimator by defining the center and tails of the distribution. Although $M$-estimates are location equivariant, *they are not scale equivariant* and thus the tuning constant is required. The smaller the value of $c$, the greater the resistance the estimate has to outliers.

Taking the derivative of Equation 2.27 gives the shape of the influence function. The $M$-estimator is then the value of $\hat{\mu}$ that solves

$$\sum_{i=1}^{n} \Psi\left(\frac{y-\hat{\mu}}{cS}\right) = 0. \qquad [2.28]$$

The measure of scale and the measure of location are estimated simultaneously, and thus an iterative estimation procedure is required (see Huber 2004 for extensive details). More details of estimation will be given with respect to $M$-estimates for regression in Chapter 5. For now, we continue with a general explanation extending from the mean.

$M$-estimation of the mean relies on the least squares objective function

$$\rho(y;\theta) = \frac{1}{2}(y-\hat{\mu})^2. \qquad [2.29]$$

The derivative of Equation 2.28 shows that influence is proportional to the value of $y$

$$\Psi(y;\theta) = (y-\hat{\mu}). \qquad [2.30]$$

To compute a more robust $M$-estimate than the mean, we simply replace the least squares objective function with another function that gives less weight to extreme values. The Huber weight function and biweight functions are two common choices.

## Huber Estimates

At the center of the distribution, the Huber weight function behaves like the mean and the least squares objective function associated with it (i.e., observations are given equal weight), but at the extremes it behaves like the median, and the least absolute values objective function associated with it, giving decreasing weight to observations as they get farther out on the tails:

$$\rho_{\mathrm{H}}(y;\theta) = \begin{cases} \frac{1}{2}y^2 & \text{if } y \le c \\ c|y| - \frac{1}{2}c^2 & \text{if } y > c \end{cases} \qquad [2.31]$$

Because the goal is to produce an estimate that is resistant to outliers, the MAD is typically used to calculate the measure of scale, $S$. Defining $S = MAD/0.6745$ results in $S$ estimating $\sigma$ when the population is normally distributed.

Following Huber (1964), it is convention (and standard in statistical software) to set $c = 1.345$, which gives substantial resistance to outliers (1.345/ 0.6745 $\cong$ 2MADs) and produces a relative efficiency of approximately 95%.

Taking the derivative of Equation 2.31 gives the shape of the influence function

$$\Psi_{\mathrm{H}}(y;\theta) = \begin{cases} c & \text{if } y > c \\ y & \text{if } y \leq |c| \\ -c & \text{if } y < -c. \end{cases} \qquad [2.32]$$

Finally, the derivative of $\Psi(.)$ gives the weights that are given to individual observations:

$$w_{\mathrm{H}_i}(y) = \begin{cases} 1 & \text{if } y \leq c \\ c/|y| & \text{if } y > c \end{cases} \qquad [2.33]$$

## Biweight Estimates

The major difference between the bisquare weight, also referred to as Tukey's bisquare, and the Huber weight occurs at the extreme ends of the tails of the distribution, where the biweight objective function is somewhat more resistant to outliers

$$\rho_{\mathrm{BW}}(y;\theta) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - \left(\frac{y}{c}\right)^2\right]^3\right\} & \text{if } |y| \leq c \\ \frac{c^2}{6} & \text{if } |y| > c. \end{cases} \qquad [2.34]$$

A tuning constant of $c = 4.685$ results in $4.685 \times S \cong$ 7MADs, which produces 95% efficiency when sampling from a normal population (Huber 1964). Taking the derivative of Equation 2.34, we see that the influence function tends rapidly toward zero

$$\Psi_{\mathrm{BW}}(y;\theta) = \begin{cases} y\left[1 - \left(\frac{y}{c}\right)^2\right]^2 & \text{if } |y| \leq c \\ 0 & \text{if } |y| > c. \end{cases} \qquad [2.35]$$

Taking the derivative of Equation 2.35 gives the weight function

$$w_{\mathrm{BW}_i}(y) = \begin{cases} \left[1 - \left(\frac{y}{c}\right)^2\right]^2 & \text{if } |y| \leq c \\ 0 & \text{if } |y| > c. \end{cases} \qquad [2.36]$$

Figure 2.1 displays the Huber and biweight functions with their default tuning constants, applied to a uniform distribution ranging from $-10$ to

10. We see that the two *M*-estimators behave much more similarly to each other than they do to the mean, which gives all observations equal weight. The Huber and biweight functions work in a similar manner for most of the distribution, except in the very center and at the extreme tails. For the biweight function, all observations with an absolute value greater than five, $|y_i| > 5$, are given a weight of zero, and only observations directly in the middle receive a weight of one. On the other hand, the Huber weight gives none of the observations a weight of zero, and a significantly larger proportion of observations a weight of one.

Although the Huber weight function and the biweight function are the most commonly used in *M*-estimation, there are many other options, some of which are shown in Table 2.1. For more details about these estimators, especially regarding recommendations for the tuning constants, see Andrews et al. (1972) and Ramsay (1977).

## *M*-Estimators of Scale

It is relatively straightforward to extend *M*-estimation to estimation of scale (Wilcox 2005:92–98). Again, the general idea is to find a function that gives less weight to extreme observations. The general class of *M*-estimators of scale are defined by the asymptotic variance of the *M*-estimate of location

$$\begin{aligned} \zeta^2 &= \frac{K^2\tau^2 E\left[\Psi^2(Z_i)\right]}{\{E[\Psi'(Z_i)]\}^2} \\ Z_i &= \frac{y_i - \mu_m}{cS}, \end{aligned} \qquad [2.37]$$

where $\mu_m$ is the *M*-estimate of location, $c$ is a positive tuning constant, $S$ is the initial measure of scale typically set to the MAD, and $\Psi$ is the score function. As with *M*-estimation of location, the Huber weight function and the biweight function are typical choices. Because it is used more often and has been shown to be more efficient, we concentrate on the latter, which results in the *biweight midvariance* (see Lax 1985).

The *biweight midvariance* is both efficient and highly resistant to outliers, achieving a breakdown point of approximately 0.5 (Hoaglin et al. 1983). It is defined by

$$\hat{\zeta}^2_{\text{bimid}} = \frac{\sum_{i:y_i^2 \le 1} \left(y_i - M_y\right)^2 \left(1 - Z_i^2\right)^4}{\left[\sum_{i:y_i^2 \le 1} \left(1 - Z_i^2\right)\left(1 - 5Z_i^2\right)\right]^2}, \qquad [2.38]$$

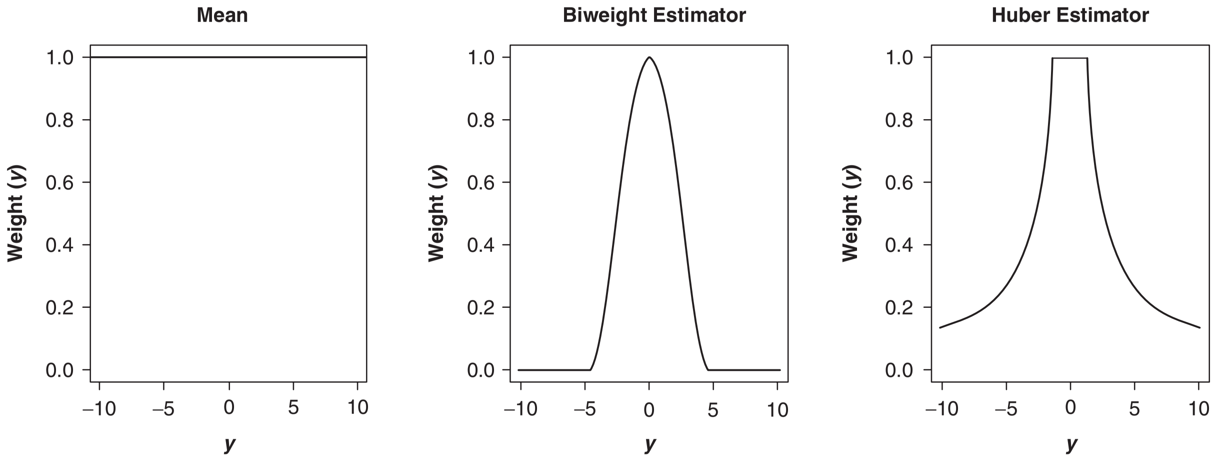

**Figure 2.1** Commonly Used *M*-Estimator Weight Functions Compared to the Mean

TABLE 2.1
A Selection of Some Possible Functions for *M*-Estimation

| | *Objective Function*, $\rho(\mu)$ | *Influence Function*, $\psi(\mu)$ | *Weight Function*, $w(\mu)$ |
|---|---|---|---|
| *Least Squares* | $\rho_{LS}(\mu) = \frac{1}{2}\mu^2$ | $\Phi_{LS}(\mu) = \mu$ | $w_{LS}(\mu) = 1$ |
| *Huber* | $\rho_H(\mu) = \begin{cases} \frac{1}{2}\mu^2 & \text{if } \mu \le c \\ c\lvert\mu\rvert - \frac{1}{2}m^2 & \text{if } \mu > c \end{cases}$ | $\Phi_H(\mu) = \begin{cases} c & \text{if } \mu > c \\ \mu & \text{if } \mu \le c \\ -c & \text{if } \mu < -c \end{cases}$ | $w_H(\mu) = \begin{cases} 1 & \text{if } \mu \le c \\ c/\lvert\mu\rvert & \text{if } \mu > c \end{cases}$ |
| *Biweight* | $\rho_{BW}(\mu) = \begin{cases} \frac{c^2}{6}\left\{1 - \left[1 - \left(\frac{\mu}{c}\right)^2\right]^3\right\} & \text{if } \lvert\mu\rvert \le c \\ \frac{c^2}{6} & \text{if } \lvert\mu\rvert > c \end{cases}$ | $\Phi_{BW}(\mu) = \begin{cases} \mu\left[1 - \left(\frac{\mu}{c}\right)^2\right]^2 & \text{if } \lvert\mu\rvert \le c \\ 0 & \text{if } \lvert\mu\rvert > c \end{cases}$ | $w_{BW}(\mu) = \begin{cases} \left[1 - \left(\frac{\mu}{c}\right)^2\right]^2 & \text{if } \lvert\mu\rvert \le c \\ 0 & \text{if } \lvert\mu\rvert > c \end{cases}$ |
| *Andrew* | $\rho_A(\mu) = \begin{cases} c\{1 - \cos(\mu/c)\} & \text{if } \lvert\mu\rvert \le c\pi \\ 2c & \text{if } \lvert\mu\rvert > c\pi \end{cases}$ | $\Phi_A(\mu) = \begin{cases} \sin(\mu/c) & \text{if } \lvert\mu\rvert \le c \\ 0 & \text{if } \lvert\mu\rvert > c \end{cases}$ | $w_A(\mu) = \begin{cases} \frac{\sin(\mu/c)}{(\mu/c)} & \text{if } \lvert\mu\rvert \le c \\ 0 & \text{if } \lvert\mu\rvert > c \end{cases}$ |
| *Ramsay* | $\rho_R(\mu) = \frac{1 - e^{-c\lvert\mu\rvert}(1+c\lvert\mu\rvert)}{c^2}$ | $\Phi_R(\mu) = \mu e^{-c\lvert\mu\rvert}$(maximum at $c^{-1}$) | $w_R(\mu) = e^{-c\lvert\mu\rvert}$ |

SOURCE: Adapted from Draper and Smith (1998: table 25.1).

where $M_y$ is the median of $y$ and

$$Z_i = \frac{y_i - \mu_m}{cS}. \qquad [2.39]$$

It is important to note that the summation in the equation is restricted to $y_i^2 \leq 1$. The tuning constant, $c$, is typically set to 9 and the scale to MAD, resulting in maximum efficiency.

## Comparing Various Estimates

### Example 2.1: Simulated Data

Table 2.2 compares the resistance of some of the estimators discussed thus far, applying them to simulated data. The first column applies the estimators to 20 random observations that were generated from the standard normal distribution $y_i \sim N(0,1)$, having a range from $-2.2$ to 1.7. In other words, these data are well-behaved, containing no outliers. The second column applies the estimators to the same data but with the addition of an extreme outlier taking a value of 60, assumed to be a miscoded observation. The breakdown point of the estimators is shown in the third column.

The first panel of the table shows the results for various measures of location. Consistent with its $BDP = 0$, the mean is badly distorted by the outlier as it is pulled toward it (changing from 0 to 2.85). On the other hand, the trimmed mean—which, following convention, has trimmed 20% of the data from the tails and thus removed the outlier—has performed very well, taking on almost identical values for the good data and the contaminated data ($-0.09$ versus $-0.04$). The median and $M$-estimate (using bisquare weights), which both have BDP $= 0.5$, are also virtually unaffected by the outlier.

Turning now to the measures of scale, we see that those involving a mean in their calculation—that is, the standard deviation, the mean deviation from the mean, and the mean deviation from the median—are all badly distorted by the outlier. Of course, this is not surprising given that they all have $BDP = 0$. The standard deviation is most affected, taking on a value more than 13 times as large as it does in the absence of the outlier. On the other hand, the outlier has very little influence on the interquartile range ($BDP = 0.25$) and the median absolute deviation ($BDP = 0.5$), the two measures based on the median. Similar to the $M$-estimate of location, the outlier does not hinder the performance of the biweight midvariance, which has $BDP = 0.5$.

TABLE 2.2
Measures of Location and for Simulated Data
With and Without an Extreme Outlier

| | *Breakdown* | *All Observations,* | *Outlier Removed,* |
|---|---|---|---|
| *Estimator* | *Point* | $\hat{\theta}_1$ | $\hat{\theta}_2$ |
| *Measures of Location* | | | |
| Mean | 0 | 0 | 2.85 |
| $\alpha$-trimmed mean | $\alpha$ (proportion of trimming) | −0.09 | −0.04 |
| Median | .5 | −0.02 | 0.005 |
| *M*-estimation | .5 | −0.12 | −0.03 |
| *Measures of Scale* | | | |
| Standard deviation | 0 | 1 | 13.13 |
| Mean deviation from mean | 0 | 0.71 | 5.44 |
| Mean deviation from median | 0 | 0.61 | 2.89 |
| Interquartile range | .25 | 1.07 | 1.21 |
| Median absolute deviation | .5 | 0.61 | 0.66 |
| Biweight midvariance | .5 | 0.89 | 1.06 |

## EXAMPLE 2.2: Public Opinion Toward Pay Inequality in Cross-National Perspective

We now turn to an example using real social science data. The data in Table 2.3 are from Weakliem, Andersen, and Heath's (2005) cross-national study of the relationship between income inequality and public opinion on pay inequality. The data set contains information measured during the 1990s on 48 countries. The variables are as follows:

- *Secpay.* The average score on an item from the World Values Survey (Inglehart et al. 2000) that asked respondents their opinions about pay inequality (secpay). The wording of the question is as follows: "Imagine two secretaries, of the same age, doing practically the same job. One finds out that the other earns considerably more than she does. The better paid secretary, however, is quicker, more efficient and more reliable at her job. In your opinion, is it fair or not fair that one secretary is paid more than the other?" Respondents were given two response choices: "Fair" (coded 0), or "Not Fair" (coded 1). As a result, *high average scores reflect public opinion that favors equality* (i.e., a majority of respondents in the country answered that it was not fair for the two secretaries to have different salaries). The averaged score across countries ranges from 0.054 to 0.622 and has a mean of 0.2.

TABLE 2.3
Public Opinion and Economic and Political Variables for 48 Countries

| *Country* | *Public Opinion (Secpay)* | *Gini Coefficient* | *Per Capita GDP* | *Democracy* |
|---|---|---|---|---|
| Armenia | .061 | 44.4 | 2072 | 0 |
| Australia | .179 | 31.7 | 22451 | 1 |
| Austria | .112 | 23.1 | 23166 | 1 |
| Azerbaijan | .070 | 36.0 | 2175 | 0 |
| Bangladesh | .057 | 28.3 | 1361 | 0 |
| Belarus | .075 | 28.8 | 6319 | 0 |
| Belgium | .302 | 27.2 | 23223 | 1 |
| Brazil | .232 | 60.1 | 6625 | 0 |
| Britain | .211 | 34.6 | 20336 | 1 |
| Bulgaria | .164 | 30.8 | 4809 | 0 |
| Canada | .176 | 28.6 | 23582 | 1 |
| Chile | .361 | 56.5 | 8787 | 0 |
| China | .131 | 41.5 | 3105 | 0 |
| Croatia | .092 | 29.0 | 6749 | 0 |
| Czech Republic | .557 | 26.6 | 12362 | 0 |
| Denmark | .248 | 21.7 | 24217 | 1 |
| Dominican Republic | .089 | 50.5 | 4598 | 1 |
| Estonia | .054 | 35.4 | 7682 | 0 |
| Finland | .354 | 22.6 | 20847 | 1 |
| France | .231 | 32.7 | 21175 | 1 |
| Georgia | .086 | 37.1 | 3353 | 0 |
| Hungary | .115 | 28.9 | 10232 | 0 |
| India | .226 | 29.7 | 2077 | 1 |
| Ireland | .289 | 35.9 | 21482 | 1 |
| Italy | .226 | 34.6 | 20585 | 1 |
| Japan | .284 | 24.9 | 23257 | 1 |
| Latvia | .070 | 28.5 | 5728 | 0 |
| Lithuania | .096 | 33.6 | 6436 | 0 |
| Mexico | .211 | 53.7 | 7704 | 0 |
| Moldova | .127 | 34.4 | 1947 | 0 |
| Netherlands | .328 | 31.5 | 22176 | 1 |
| Norway | .441 | 24.2 | 26342 | 1 |
| Peru | .175 | 46.2 | 4282 | 1 |
| Portugal | .265 | 35.6 | 14701 | 1 |
| Romania | .133 | 28.2 | 5648 | 0 |
| Russia | .076 | 48.0 | 6460 | 0 |
| Slovakia | .622 | 19.5 | 9699 | 0 |
| Slovenia | .108 | 29.2 | 14293 | 0 |
| Spain | .286 | 32.5 | 16212 | 1 |
| Sweden | .401 | 25.0 | 20659 | 1 |
| Switzerland | .149 | 36.1 | 25512 | 1 |
| Taiwan | .075 | 27.7 | 12090 | 0 |

*(Continued)*

TABLE 2.3 (Continued)

| *Country* | *Public Opinion (Secpay)* | *Gini Coefficient* | *Per Capita GDP* | *Democracy* |
|---|---|---|---|---|
| Turkey | .207 | 41.5 | 6422 | 0 |
| Ukraine | .085 | 47.3 | 3194 | 0 |
| Uruguay | .273 | 42.3 | 8623 | 0 |
| USA | .148 | 36.9 | 29605 | 1 |
| Venezuela | .208 | 46.8 | 5808 | 1 |
| West Germany[a] | .149 | 30.0 | 22169 | 1 |

[a] The survey was administered to respondents in West Germany only, and the data set uses the term "West Germany."

- *Gini.* The Gini coefficient, which theoretically ranges from 0 (perfect income equality where income is divided equally among all citizens) and 1 (perfect inequality, where one individual has all of the income). In other words, high values indicate high levels of income inequality.
- *Per Capita GDP/1000.* The per capita gross domestic product of the country in U.S. dollars.
- *Democracy.* A dummy variable coded 1 for "Old Democracies" (i.e., the country had experienced democratic rule for at least 10 years at the time of the data collection), and 0 for "New Democracies."

For more detailed information on the sources used to construct the measures, see Weakliem et al. (2005).

Of interest is the distribution of public opinion toward pay inequality (often referred to simply as "public opinion" from here onward) for those countries that were democratic for less than 10 years at the time of the study ($n = 26$). Given that the public opinion variable will be used as a dependent variable in regression analyses later, it is important to explore its distribution in a preliminary attempt to identify any features—such as a skew or outliers—that might be problematic. We start by examining Figure 2.2, which displays a kernel density estimate (i.e., a smoothed histogram) of the distribution of the public opinion variable. With the exception of a small bump at the extreme positive end of the distribution, the rest of the distribution is fairly symmetric. Further exploratory analysis indicates that two countries—the Czech Republic and Slovakia—have unusually high values. As can be seen in Table 2.4, these countries have values of 0.557 and 0.622, whereas no other country has a value reaching 0.4. Given that the two countries were joined until very recently, it seems likely that the uniqueness of these countries is due to a common cultural and historical heritage.

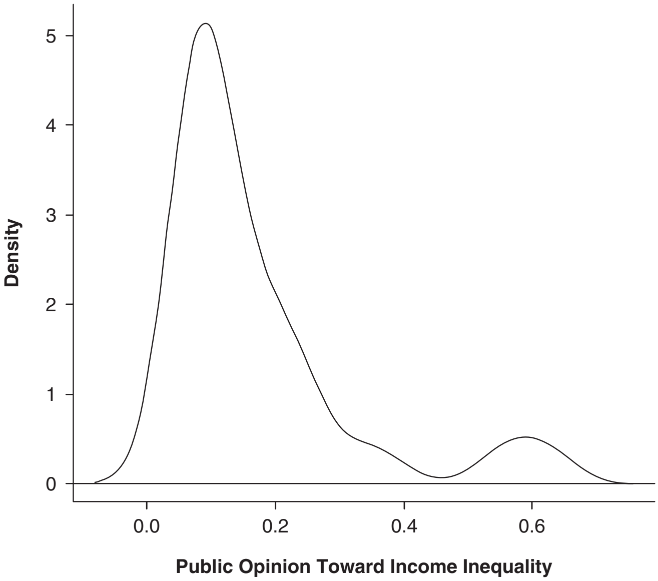

**Figure 2.2** Distribution of Public Opinion Toward Pay Inequality for 26 New Democracies

We now turn to Table 2.4, which explores how various estimators of location and scale behave when the Czech Republic and Slovakia are included and excluded. Starting with the mean, we see that it decreases substantially (from 0.167 to 0.131) when the outliers are excluded. Similarly, the measures of scale based on the mean decrease substantially when the outliers are removed (e.g., the standard deviation is 1.86 times as large when the outliers are included than when they are excluded). On the other hand, the high resistance of the median and the *M*-estimate is evident in that they are virtually unchanged when the outliers are removed. Similarly, the differences in estimates between the two data sets are much smaller for the median absolute deviation and the *M*-estimate (biweight midvariance), two measures of location with high breakdown points.

In concluding this chapter, a cautionary note about examining the univariate distributions of the variables used in a regression analysis is appropriate. OLS regression estimates the *conditional mean* of *y* given the *x*s. As a result, an outlier for *y* is not necessarily a regression outlier. Conversely, it is not necessary that an influential observation in terms of the regression estimates is an outlier in terms of *y*. Still, this does not mean that we should

TABLE 2.4
Measures of Location and Scale for Public Opinion Variable, New Democracies

| *Estimator* | *All Observations, $\hat{\theta}_1$* | *Czech Republic and Slovakia Removed, $\hat{\theta}_2$* |
|---|---|---|
| *Measures of Location* | | |
| Mean | 0.167 | 0.131 |
| $\alpha$-trimmed mean | 0.123 | 0.114 |
| Median | 0.112 | 0.102 |
| *M*-estimation | 0.127 | 0.112 |
| *Measures of Scale* | | |
| Standard deviation | 0.145 | 0.078 |
| Mean deviation from mean | 0.102 | 0.060 |
| Mean deviation from median | 0.081 | 0.056 |
| Interquartile range | 0.129 | 0.097 |
| Median absolute deviation | 0.042 | 0.032 |
| Biweight midvariance | 0.005 | 0.004 |

ignore the univariate distributions. Failing to explore the univariate distributions could prevent the researcher from detecting important features of the data. But it is best to refrain from any remedies for the unusual observations until the relationships between the variables have been explored. With this in mind, we now turn to the OLS estimation of linear regression, exploring in detail how unusual observations can affect its estimates and how they can be detected. We return to measures of scale and location later in the context of robust regression methods.

## Notes

[1] Typically, $\varepsilon_n^*$ is used instead of BDP to denote the breakdown point. I intentionally avoid the use of $\varepsilon_n^*$ in order to prevent confusion with the errors for a regression model, which is unrelated to the breakdown point.

[2] Although discussions of the breakdown point often use the term *bias,* the term *effect* is used here to avoid confusion with the usual statistical meaning of bias discussed earlier. If influential outliers do not reflect miscoding, an estimator can still be unbiased—that is, the average of the estimator from repeated random sampling will equal the population parameter—regardless of the effect the outliers have on the estimate. Still, this does not mean that the estimate will be a useful summary of the data.

[3] If the estimator satisfies this condition, it is considered to have reached the Cramer-Rao lower bound (see Cramer 1946 for more details).