Chapter 2

A Brief Introduction to Robust Statistics

This chapter provides a brief introduction to some of the key concepts and techniques in the area of outlier robust estimation and testing. The setup is as follows. Section 2.1 discusses the concept of outliers. Section 2.2 introduces some performance measures for assessing the robustness of estimation and testing procedures. Section 2.3 discusses several robust estimators, the most prominent of which is the generalized maximum likelihood type (GM) estimator. Section 2.4 touches upon the estimation of multivariate location and scatter. Finally, Section 2.5 contains some miscellaneous remarks on robust model selection, robust testing, residual analysis, and diagnostic plots.

Throughout this chapter, the i.i.d. setting is considered. Some concepts and techniques can be generalized to dependent and/or heterogeneous observations, and relevant references to these generalizations are found in the text.

2.1 Outliers

In this section the concept of an outlier is discussed in more detail. Subsection 2.1.1 recapitulates the definition of outliers provided by Davies and Gather (1993). Subsection 2.1.2 classifies outliers into several commonly used categories.

2.1.1 Definition of Outliers

As mentioned in Section 1.2, the term outlier is often used rather informally. Davies and Gather (1993) even state that the word *outlier* has never been given a precise definition. Barnett and Lewis (1979, p. 4) define an outlier as

> ...an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data, ...

which cannot be called a precise definition. According to Barnett and Lewis (1979), the decision whether an observation is an outlier, is left to the sub-

jective judgement of the researcher. Judge et al. (1988, p. 889) use the term outlier for a large value of the regression error. Krasker et al. (1983), and even Hampel et al. (1986) and Rousseeuw and Leroy (1987) do not provide a formal definition of outliers. They only present classifactions of outliers into different categories (see Subsection 2.1.2, below).

As Davies and Gather admit, there is not much novelty in their definition of outliers itself. Rather, the novelty is in the fact that the concept of an 'outlier' is defined at all. Still, it is useful to consider their definition in somewhat more detail, as it differs in some respects from other approaches that are followed in the literature.

Let F be the target distribution or null model, i.e., the distribution without outliers. For expositional purposes, F is taken to be the univariate normal distribution with (possibly unknown) mean μ and (possibly unknown) variance σ^2. For $0 < \alpha < 1$, Davies and Gather introduce the concept of an *outlier region* as

$$\text{out}(\alpha, \mu, \sigma^2) = \{x : \ |x - \mu| > z_{1-\alpha/2}\sigma\}, \tag{2.1}$$

with z_q the q-quantile of the standard normal distribution. An observation x_i is called an α outlier with respect to F if $x_i \in \text{out}(\alpha, \mu, \sigma^2)$. Consequently, it is possible that regular observations, i.e., observations that are drawn from the target distribution F, are classified as α outliers. Also note that x_i in this definition can be any real number and need not coincide with one of the observations.

An important feature of (2.1) is that it is based on the 'true' parameters μ and σ^2, instead of estimates of these parameters. As a result, the outlier region cannot be used directly in most practical circumstances. An advantage, however, of the use of true rather than estimated parameters, is that the definition of outliers does not depend on the method actually used for detecting the outliers. For example, one may estimate μ and σ by means of the arithmetic mean and the ordinary sample standard deviation and classify those observations x_i as outliers that satisfy

$$x_i \in \{x : \ |x - \hat{\mu}| > z_{1-\alpha/2}\hat{\sigma}\},$$

with $\hat{\mu}$ and $\hat{\sigma}$ the mentioned estimates of μ and σ, respectively. Alternatively, one could use different estimators for μ and σ, like the median and the scaled median absolute deviation (see (2.32) and the discussion below). This would give rise to different observations to be labeled outliers. The definition of Davies and Gather allows for the distinction between, on the one hand, 'real' outliers, and, on the other hand, observations that are labeled outliers on the basis of one of the above outlier identification methods. As a consequence, one can construct performance measures of outlier identification methods based on their ability to spot the observations in $\text{out}(\alpha, \mu, \sigma^2)$. Davies and Gather discuss several of these measures.

A different, more conventional approach for defining outliers is to postulate an outlier generating model (or mixture model). Here, the observations x_i are

drawn from the target distribution F with probability p. With probability $(1-p)$, the x_i are drawn from the contaminating distribution G. The distribution G can, in turn, be a mixture of several contaminating distributions $G_1, \ldots, G_k$. The observations are called regular if they are drawn from F, while they are called contaminants when drawn from G (see Davies and Gather (1993)). In this approach the identification of outliers coincides with the identification of the contaminants. This contrasts with the definition of Davies and Gather, in which observations from F may well be α outliers with respect to F, while observations from G need not be α outliers with respect to F. To illustrate this point, let F be the standard normal and G the standard Cauchy distribution. Then it is possible, although with low probability, that a drawing from F exceeds 5. For α equal to, e.g., 0.01, this drawing would be called an α outlier using the definition of Davies and Gather. This seems rather natural. Using the above alternative definition of outliers, however, the drawing would not be called an outlier, as it follows the target distribution F rather than the contaminating distribution G. Similarly, a drawing from G is inside the interval $[-1, 1]$ with probability $1/2$. Using the alternative definition, such an observation would be called an outlier. Using the definition of Davies and Gather, however, combined with the α value of 0.01, the observation is not an α outlier with respect to the standard normal. This illustrates the usefulness of the definition of Davies and Gather over some of the more conventional definitions employed in the literature. The conventional definition, however, has its own merits, as it provides a useful tool for developing concepts to assess the robustness of statistical procedures, see Sections 2.2 and 4.2.

2.1.2 Classification of Outliers

Following Krasker et al. (1983) and Hampel et al. (1986), outliers can be classified into two main categories, namely (i) gross errors and (ii) outliers due to model failure.

Gross errors come in the form of recording errors. These recording errors arise due to, e.g., plain keypunch errors, transcription errors, technical difficulties with the measurement equipment, a change in the unit of measurement, incompletely filled out questionaires, or misinterpreted questions. Observations that are gross errors can often safely be discarded. As mentioned in Chapter 1, however, detecting these observations may be very difficult, e.g., in a multivariate setting. Hampel et al. (1986, pp. 25–28) argue that the occurrence of gross errors in empirical data sets is the rule rather than the exception.[1] As a small number of gross errors can already cause great diffi-

[1] As Krasker et al. (1983) correctly point out, gross errors occur more regularly in some data sets than in others. For example, gross errors in the relatively short macroeconomic time series that are often used in empirical, macro-oriented econometric work, are extremely unlikely. In cross-sectional data, however, outliers are much more likely to occur, especially if the sample is large. The same holds for longitudinal data sets with a large cross-sectional dimension, compare Lucas et al. (1994) and van Dijk et al. (1994).

culties for the traditional OLS estimator, the use of outlier robust statistical procedures seems warranted.

The second main cause for the occurrence of outliers is the approximate nature of the statistical/econometric model itself (see Krasker et al. (1983, Section 2) and Hampel et al. (1986, Section 1.2)). As already mentioned in Chapter 1, the use of dummy variables in empirical econometric model building is common practice. Introducing a dummy variable that equals one for only a single data point, is equivalent to discarding that data point completely. Thus, it is common practice in empirical model building to state (implicitly) that the postulated model does not describe all observations. The next best thing one can do, and usually does, is then to build a model that describes the majority of the observations. This is exactly the aim of robust statistical procedures.

The approximate nature of the model can demonstrate itself in different forms, four of which are mentioned below. First, in time series analysis there can be special events that are too complex to be captured by simple statistical models. A typical example of this is the oil crisis in 1973. Although the oil crisis could be made endogenous in an econom(etr)ic time series model, it is generally treated as an exogenous event in most empirical studies. Also special government actions sometimes fall outside that part of reality one is trying to model (see, e.g., Juselius' (1995) example of the introduction of several new taxes). Second, in cross-section studies, one can easily think of situations where outliers show up, just because one is unable to capture the whole complexity of real life in a single model. For example, one can imagine an income study where persons are asked for their present salary and their expected salary within half a year. For most individuals, the discrepancy between these two numbers will be very small. It is possible, however, that one of the individuals is a Ph.D. student (low present salary), who possesses some external information: within four months he is going to graduate and accept a position at a banking institution (with a very high salary). As the presence of external information is difficult to capture in ordinary statistical models, this Ph.D. student will show up as an outlier in most of the models that are build for this income study. One can think of many similar examples, e.g., firms that are typical success-stories, firms that are unique in the sample (e.g., Fokker in a sample of Dutch companies), etc. Third, when using linear models, the possible nonlinearity of the relationship under study may give rise to outliers. For example, the relationship between consumption and income may be linear for small to moderate values of income, but nonlinear for high values of income. When fitting a linear model, the observations with high income are then likely to show up as outliers. Fourth, the omission of variables from the model may give rise to outliers. These variables may have been omitted because they were not available, or because the researcher was not aware of their relevance for the subject under study. As an example, one can think of a study on household expenditures. Some of the families under study may live in areas with extremely cold winters, thus having significantly higher heating and isolation expenditures than the remaining households in

the sample. If in a first instance a temperature variable is omitted from the model, then the families in the cold winter areas probably show up as outliers.

To close this section, I discuss a classification of regression outliers for the simple linear regression model $y_t = x_t\beta + \varepsilon_t$. The relevant outlier configurations are displayed in Figure 2.1.

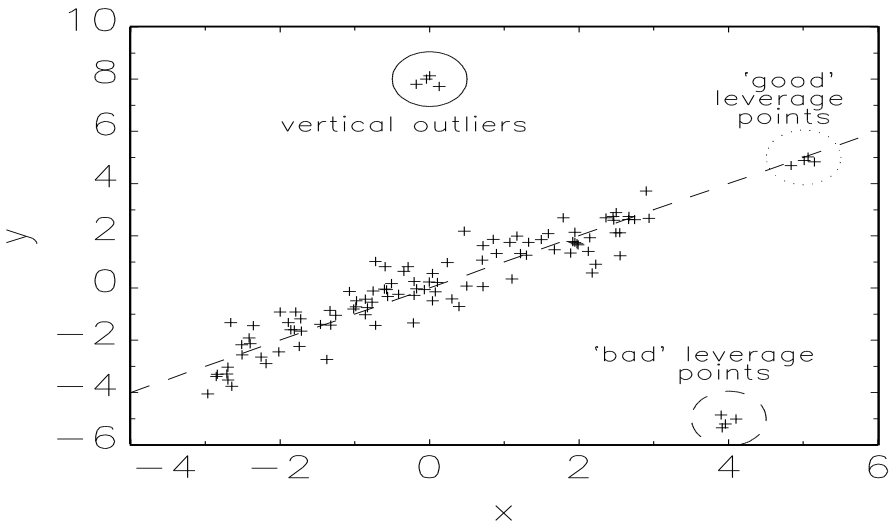

Figure 2.1.— A classification of regression outliers

The first set of outliers, the ones in the solid circle, are called vertical outliers. The x_t values of these observations fall inside the range of the bulk of the observations. However, the observations depart markedly from the linear relationship indicated by the majority of the data. For example, in the household expenditure example mentioned earlier, the households in the cold winter areas may show up as vertical outliers.

The second set of outliers, the ones in the dotted circle, are called 'good' leverage points. They satisfy the linear relationship indicated by the bulk of the data, but have x_t values outside the usual range. Such observations tend to increase the efficiency of the OLS estimator, which is probably the reason why they are called *good* leverage points. The predicate good is controversial, however, because these observations also have a large influence on the OLS estimator. By perturbing them, the fitted regression line can alter substantially. Moreover, if the observations are shifted over a larger distance, they can easily turn into bad leverage points.

The third set of outliers, the ones in the dashed circle, are called 'bad' leverage points. Just as with the good leverage points, the adjective bad is placed between quotation marks. Bad leverage points have aberrant values for x_t and, moreover, do not fit the (linear) pattern set out by the bulk of the data. Bad leverage points are detrimental to the OLS estimator (see, e.g.,

Rousseeuw and Leroy (1987, Chapter 2)). They can be caused by gross errors, or by the presence of unusual individuals, as the Ph.D. student in the income study example described earlier in this section.

Ways to correct for these three types of regression errors during the estimation stage are discussed in Section 2.3.

2.2 Robustness Assessment

This section discusses several concepts for assessing the robustness of statistical procedures in an i.i.d. context. Most of these are asymptotic concepts, which requires a representation of estimators that can be used in an asymptotic context. Therefore, Subsection 2.2.1 introduces the representation of estimators as functionals that operate on the space of distribution functions. Subsection 2.2.2 introduces the first quantitative concept by which to judge statistical procedures, namely the influence function. This is a local robustness measure. Subsection 2.2.3 introduces a global robustness measure: the breakdown point. Subsection 2.2.4 treats the bias curve, which is a useful concept for worst-case analyses. This subsection also contains some comments on the recent approach of Horowitz and Manski (1995). For expositional purposes, the whole presentation in this section is for the simple univariate location/scale model. All concepts can, however, be generalized to more complex models, see Hampel et al. (1986). Moreover, as some of the notation may be nonstandard for econometricians, each subsection starts with a simple example. These examples serve to provide some intuition for the concepts that are developed in each subsection.

2.2.1 Estimators as Functionals

Example 2.1 Consider a sample $\{y_t\}_{t=1}^T$, with T denoting the sample size. Usually, one thinks of an estimator as a function $\mu_T(\cdot): \mathbb{R}^T \to \mathbb{R}$, depending on the arguments $y_1, \ldots, y_T$. For example, the mean is given by

$$\mu_T(y_1, \ldots, y_T) = T^{-1} \sum_{t=1}^{T} y_t, \tag{2.2}$$

and the median by

$$\mu_T(y_1, \ldots, y_T) = \arg\min_m T^{-1} \sum_{t=1}^{T} |y_t - m|. \tag{2.3}$$

As it is assumed throughout this chapter that we are working in an i.i.d. context, the ordering of the observations is irrelevant. Consequently, no information is lost by replacing the arguments $y_1, \ldots, y_T$ in (2.2) and (2.3) by the empirical distribution function $F_T(y)$, with

$$F_T(y) = T^{-1} \sum_{t=1}^{T} 1_{\{y \geq y_t\}}(y), \tag{2.4}$$

and 1_A the indicator function of the set A. For example, the mean can now be written as

$$\mu(F_T) = \int_{-\infty}^{+\infty} y dF_T(y) = T^{-1} \sum_{t=1}^{T} y_t. \tag{2.5}$$

Note that μ now depends on the empirical distribution function F_T instead of on the observations $y_1, \ldots, y_T$. It is now a small step towards viewing estimators as functionals that map the space of cumulative distribution functions (c.d.f.'s) to the real line. △

Let $\mathcal{F}$ denote the space of all distribution functions. Then an estimator is defined as a functional $\mu : \mathcal{F} \to \mathbb{R}$. For example, for the mean one obtains

$$\mu(F) = \int_{-\infty}^{+\infty} y dF(y), \tag{2.6}$$

while for the median one has

$$\mu(F) = \arg\min_m \int_{-\infty}^{+\infty} |y - m| dF(y). \tag{2.7}$$

Each estimator can also be evaluated at the sample, i.e., at F_T. In this way one obtains (2.5) for the mean, and

$$\mu(F_T) = \arg\min_m \int_{-\infty}^{+\infty} |y - m| dF_T(y) = \arg\min_m T^{-1} \sum_{t=1}^{T} |y_t - m|, \tag{2.8}$$

for the median. Note than an alternative way to define the median is given by $\mu(F) = F^{-1}(1/2)$, with $F^{-1}(\cdot)$ the inverse of the c.d.f. This alternative definition is used in Subsection 2.2.2.

The estimators used here are assumed to be Fisher consistent (see Rao (1973, p. 345)). This is important for the definition of the influence function in Subsection 2.2.2, below. Let F_θ denote a distribution that is indexed by a finite dimensional parameter vector θ. Moreover, let $g(\theta)$ be the parametric function one wants to estimate. Then the crucial condition for an estimator $\mu(\cdot)$ to be Fisher consistent is

$$\mu(F_\theta) \equiv g(\theta) \text{ identically in } \theta,$$

compare Hampel et al. (1986, p. 83). This requirement restricts the class of estimators to the estimators that produce the correct estimates when evaluated at the distribution F_θ. Rao (1973, p. 346) supplements this condition with the requirement that the functional μ has certain continuity properties and, moreover, that the functional is also used for obtaining parameter estimates in finite samples by means of the quantity $\mu(F_T)$.

The representation of an estimator as a functional automatically leads to the first robustness concept, namely that of qualitative robustness. For given F, one can derive the asymptotic distribution of $\mu(F)$. Call this distribution

G_F. For example, for a wide variety of estimators (μ) and data generating processes (F), one has that

$$\sqrt{T}(\mu(F_T) - \mu(F)) \xrightarrow{d} N(0, V(\mu, F)),$$

with $N(0, V)$ denoting a normal random variate with mean zero and variance V, $V(\mu, F)$ denoting the asymptotic variance of the estimator, and $\xrightarrow{d}$ denoting convergence in distribution (see Hampel et al. (1986, p. 85)). Using the distribution F of y_t and the asymptotic distribution of the estimator (G_F), one arrives at the following concept. The estimator μ is said to be *qualitatively robust* at F if for every $\varepsilon > 0$ there exists a $\delta > 0$, such that

$$d(F, \tilde{F}) < \delta \text{ implies } d(G_F, G_{\tilde{F}}) < \varepsilon, \tag{2.9}$$

with $d(\cdot, \cdot)$ a suitable metric and $\tilde{F}$ a distribution function. Hampel et al. (1986) suggest the Prohorov metric for $d(\cdot, \cdot)$, but other choices are also possible.

The intuition behind qualitative robustness is that if the distribution of the y_t's (F) is perturbed only slightly, then the corresponding change in the asymptotic distribution of the estimator (G_F) should also be small. Qualitative robustness is intimately connected to continuity of the functional $\mu(\cdot)$, see Hampel et al. (1986). Note that the mean is not qualitatively robust if $d(\cdot, \cdot)$ is the Prohorov metric, not even at the standard normal distribution. This is seen by letting $\tilde{F}$ in (2.9) be equal to $(1 - \eta)F + \eta\tilde{G}$, with F the standard normal, $\tilde{G}$ the standard Cauchy, and $0 < \eta < 1$. The mean of this distribution is undefined for every positive value of η. As a result, the mean does not satisfy (2.9) and is, therefore, not qualitatively robust.

Qualitative robustness is only a first measure to assess the robustness of statistical procedures. Other measures are discussed next.

2.2.2 Influence Function

Example 2.2 Consider the same setting as in Example 2.1, only assume that the observations are ordered such that $y_1 \leq y_2 \leq \ldots \leq y_T$. Moreover, assume for simplicity that the sample size, T, is odd. In this subsection, the influence function is studied. The influence function measures the effect of small[2] changes in the distribution (the sample) on the value of an estimator.

[2] The word 'small' is a subjective and rather vague notion. It has to be made precise in each particular problem that is studied. Assume that one replaces the observation y_T above by the observation $y_T + \zeta$. This change causes a deviation from the original i.i.d. assumption. The change is small in that only one out of the T original observations is replaced. In a quite different sense, however, the change is substantial, as ζ may diverge to infinity. So the number of changed observations (i.e., the fraction of contamination) is small, but the actual change in the contaminated observation may be excessively large. In the present subsection, the above change is labeled 'small,' but such labeling procedures have to be made explicit in each problem at hand. The user can then decide for him/herself whether the deviation is actually small in a sense that is relevant for his/her own research.

For example, consider the mean, which is given by $\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$, and the median, which is given by $\hat{\mu} = y_{(T+1)/2}$. One can now ask the question how the mean and the median are changed if one of the observations, y_T for simplicity, is changed to $y_T + \zeta$, with ζ some real number. Let $\hat{\mu}^\zeta$ denote the estimator for the revised sample. Note that $\hat{\mu}^0$ produces the expressions for the mean and median for the original sample. One obtains $\hat{\mu}^\zeta = \hat{\mu}^0 + \zeta/T$ for the mean, and

$$\hat{\mu}^\zeta = \begin{cases} \hat{\mu}^0 & \text{for } \zeta > y_{(T+1)/2} - y_T, \\ y_T + \zeta & \text{for } y_{(T-1)/2} - y_T \leq \zeta \leq y_{(T+1)/2} - y_T, \\ y_{(T-1)/2} & \text{otherwise,} \end{cases} \tag{2.10}$$

for the median. If one considers the difference between $\hat{\mu}^\zeta$ and $\hat{\mu}^0$, standardized by the fraction of changed data points $(1/T)$, one obtains $T(\hat{\mu}^\zeta - \hat{\mu}^0) = \zeta$ for the mean, and

$$T(\hat{\mu}^\zeta - \hat{\mu}^0) = \begin{cases} 0 & \text{for } \zeta > y_{(T+1)/2} - y_T, \\ T(y_T + \zeta - y_{(T+1)/2}) & \text{for } y_{(T-1)/2} - y_T \leq \\ & \quad \zeta \leq y_{(T+1)/2} - y_T, \\ T(y_{(T-1)/2} - y_{(T+1)/2}) & \text{otherwise,} \end{cases} \tag{2.11}$$

for the median. This standardized difference is closely related to the finite sample versions of the influence function as presented in Hampel et al. (1986, Section 2.1e). Note that for the mean, the standardized difference is an unbounded function in ζ. Consequently, by choosing a suitable value for ζ, the mean can attain any value by changing only one observation. Stated differently, one outlier can have an arbitrarily large impact on the mean. In contrast, the standardized difference for the median is a bounded function of ζ. Therefore, one outlier can only have a limited impact on the median as an estimate of the location of the sample. In this way, one can see that the median is more robust than the mean. The influence function tries to provide the same information as above by looking at the (standardized) effect of (infinitesimally) small perturbations in a distribution on the value of an estimator, where an estimator is now again seen as a mapping from the space of c.d.f.'s to the real line (see Subsection 2.2.1). △

Following Hampel et al. (1986, p. 84), the influence function is defined as follows.[3] Let Δ_y denote the c.d.f. with a point mass at y, i.e., $\Delta_y(\tilde{y}) = 1_{\{\tilde{y} \geq y\}}(\tilde{y})$. Then the influence function of the estimator μ at F is given by

$$IF(y, \mu, F) = \lim_{\eta \downarrow 0} \frac{\mu((1-\eta)F + \eta\Delta_y) - \mu(F)}{\eta} \tag{2.12}$$

in those y where this limit exists. Note that $\mu(F)$ denotes the value of the estimator at the original distribution F. Similarly, $\mu((1-\eta)F + \eta\Delta_y)$ denotes

[3]Note again that this definition is only valid for the i.i.d. setting. With dependent observations, different definitions of the IF are available, see Künsch (1984), Martin and Yohai (1986), and Chapter 4.

the value of the estimator at the slightly perturbed distribution $(1-\eta)F+\eta\Delta_y$. So the numerator in (2.12) captures the difference between the value of the estimator for the original distribution F and a slightly perturbed version of this original distribution. The denominator takes care of the standardization by the fraction of contamination. By looking at the definition of the IF, one can see that it is a type of directional derivative of μ, evaluated at F in the direction $\Delta_y - F$.

The IF measures the effect of an (infinitesimally) small fraction of contamination on the value of the estimator and is, therefore, called a quantitative robustness concept. In fact, one can say that the IF measures the asymptotic (standardized) bias of the estimator caused by contaminating the target distribution F. Using simple approximations, the IF can be used to compute the numerical effect of small fractions of contamination, e.g., one outlier, on the estimator. Taking a finite value of η and rewriting[4] (2.12), one obtains[5]

$$\mu((1-\eta)F + \eta\Delta_y) \approx \mu(F) + \eta IF(y, \mu, F). \tag{2.13}$$

Thus, the IF helps to quantify the change in the estimator caused by the presence of contaminated data. Moreover, if F is equal to the empirical c.d.f. F_T and if $\eta = (1+T)^{-1}$, then (2.13) presents the change in the estimator caused by adding a single observation y to a sample of size T. (2.13) also illustrates that estimators with a bounded IF are desirable from a robustness point of view. If the IF is bounded, then the effect of a small fraction of contamination (e.g., a small number of outliers) on the estimator is also bounded.

As the boundedness of its IF is an important feature of an estimator, it is natural to look at the supremum of the IF. This leads us to the the notion of *gross error sensitivity*. The gross error sensitivity γ^* is defined as the supremum of the IF with respect to y. As mentioned below (2.13), finite values of γ^* are desirable, because they induce that the estimator changes only slightly if a small fraction of contamination is added.

It is easy to show that γ^* is infinite if μ is the arithmetic mean. Using (2.6) and (2.12), one obtains

$$IF(y, \mu, F) = \lim_{\eta \downarrow 0} \frac{\int_{-\infty}^{\infty} \eta \tilde{y} d(\Delta_y - F)(\tilde{y})}{\eta} = y - \mu(F). \tag{2.14}$$

This function is unbounded and monotonically increasing in y, leading to the conclusion that γ^* is infinite. For the median, in contrast, one obtains for values of y that are greater than the median of F,

$$\begin{aligned} IF(y, \mu, F) &= \lim_{\eta \downarrow 0} \frac{F^{-1}((1-\eta)^{-1}/2) - F^{-1}(1/2)}{\eta} \\ &= \lim_{\eta \downarrow 0} \frac{F^{-1}((1/2) + \eta/(2(1-\eta))) - F^{-1}(1/2)}{\eta/(2(1-\eta))} \cdot \frac{\eta/(2(1-\eta))}{\eta} \\ &= \frac{1}{2f(F^{-1}(1/2))}, \end{aligned} \tag{2.15}$$

[4] Davies (1993) shows some examples where these approximations cannot be used.

[5] See Chapter 4 for a similar derivation in the time series context.

with F^{-1} the inverse c.d.f. and f the p.d.f. corresponding to F. Similarly, for values of y smaller than the median of F, one obtains $IF(y, \mu, F) = -1/(2f(F^{-1}(1/2)))$. Thus, if $f(F^{-1}(1/2)) \neq 0$, the IF of the median is clearly bounded and its gross error sensitivity is finite.

2.2.3 Breakdown Point

Example 2.3 Whereas Example 2.2 considered the effect of a change in one of the observations on the value of an estimator, the present example considers the maximum number of outliers an estimator can cope with before producing nonsensical values. The maximum fraction of outliers an estimator can cope with is closely related to the breakdown point, which is the topic of the present subsection.

First, consider the arithmetic mean. From Example 2.2 it follows that if one adds ζ to one of the original observations, then the mean changes from $\hat{\mu}^0$ to $\hat{\mu}^0 + \zeta/T$. By letting ζ diverge to infinity, one sees that the mean also diverges to infinity and, thus, produces a nonsensical value if there is only one (extreme) outlier. One can conclude that the maximum fraction of outliers the mean can cope with is zero, as a fraction of $1/T$ already suffices to drive the estimator over all bounds.

Second, consider the median. It follows from Example 2.2 that the effect of one extreme outlier on the median is bounded, at least if the sample size is greater than two.[6] First consider the case in which the sample size T is odd. Now in order to let the median diverge to infinity, one must let the $((T+1)/2)$th sample order statistic diverge to infinity. The minimum number of observations one has to change in order to achieve this objective, is $(T+1)/2$. If one adds ζ to $(T+1)/2$ of the original observations, the median becomes equal to $y_{t_0} + \zeta$ for some t_0 satisfying $1 \leq t_0 \leq T$. Letting ζ diverge to infinity, the median diverges. So for odd sample sizes, the maximum fraction of outliers the median can cope with is $((T+1)/2-1)/T$. Similarly, for even sample sizes one can argue that the maximum tolerable fraction of outliers is $(T/2-1)/T$.

The maximum fraction of outliers an estimator can cope with is formally studied in the present subsection using the notion of the breakdown point. $\triangle$

By its definition in (2.12), the influence function (IF) discussed in the previous subsection is a local concept. It measures the effect on the estimator of an *infinitesimally small* fraction of contamination. In practice, however, one is usually interested in the effect of a small, but positive fraction of contamination (see the approximation in (2.13)). Therefore, the IF has to be supplemented with a global robustness measure. Such a global measure can be used to indicate the size of the neighborhood in which approximations of the form (2.13) are allowed. A suitable global robustness measure is the breakdown point of an estimator.

[6] Following the ordinary convention, the median is set to $y_{(T+1)/2}$ for odd values of T, and to $(y_{T/2} + y_{T/2+1})/2$ for even values of T.

In the context of the simple i.i.d. location/scale model, the breakdown point ε^* of an estimator μ is defined as

$$\begin{aligned}\varepsilon^* \quad = \quad & \sup_{0\leq\eta\leq 1} \{\eta : \exists\ K_\eta \subset \mathbb{R}, \text{ with } K_\eta \text{ a compact set, such that} \\ & d(F,\tilde{F}) < \eta \text{ implies } \tilde{P}_F(\{\mu(F_T) \in K_\eta\}) \rightarrow 1 \text{ for } T \rightarrow \infty\}, \quad (2.16)\end{aligned}$$

with $d(\cdot,\cdot)$ a distance function as in (2.9) (compare Hampel et al. (1986)), $\tilde{F}$ a c.d.f., and $\tilde{P}_F$ the probability measure associated with $\tilde{F}$. A special case of the breakdown point that is often encountered in the literature, is the gross error breakdown point, defined by letting $\tilde{F}$ in (2.16) be equal to $(1-\eta)F + \eta G$, with G some distribution function. At first sight, the definition in (2.16) may be difficult to grasp. Therefore, some comments are in order. For a given value of η, one considers distributions $\tilde{F}$ that are in a well-specified neighborhood of the original distribution F. For each of these distributions it must hold that the probability that the location estimator is in some compact set, tends to one as the sample size diverges to infinity. The supremum value of η that satisfies these requirements, is the breakdown point. Informally, the breakdown point gives the maximum perturbation to the original distribution F that still guarantees a finite value of the estimator μ.

The breakdown point and the IF provide different pieces of information concerning the robustness of an estimator. As Hampel et al. (1986, pp. 175–176) show, the maximum bias of an estimator over gross error neighborhoods (i.e., $\tilde{F} = (1-\eta)F + \eta G$) is approximately equal to $\eta\gamma^*$, with γ^* the gross error sensitivity defined in Subsection 2.2.2. Thus, the maximum bias can be approximated using a concept derived from the IF, namely γ^*. As will be explained in the next subsection, this approximation to the maximum bias is very bad if η is chosen in the neighborhood of the breakdown point ε^*. It is even invalid if $\eta \geq \varepsilon^*$. As a rule of thumb, Hampel et al. suggest to use the approximation only for fractions of contamination $\eta < \varepsilon^*/2$. This nicely illlustrates that the IF and the breakdown point provide complementary pieces of information. The IF can be used to approximate the bias, while the neighborhood in which this approximation is useful, is determined by the breakdown point.

The breakdown point as defined in (2.16) is an asymptotic concept. Moreover, the use of the metric $d(\cdot,\cdot)$ in the definition can lead to differences of opinion (see the debate on the appropriate metric in the time series literature, e.g., Papantoni-Kazakos (1984) and Boente et al. (1987)). Donoho and Huber (1983) provided a finite sample version of the breakdown point. As this definition only requires a metric in a Euclidean space, no controversy arises on the appropriate specification of $d(\cdot,\cdot)$. The finite sample version of the breakdown point is equal to one minus the maximum fraction of uncontaminated observations that is needed to keep the estimator bounded. For example, the finite sample breakdown point of the mean is zero, as one outlier suffices to drive the estimator over all bounds (see Example 2.3). This immediately illustrates that the breakdown point is concerned with extreme outlier configurations, which

might be deemed unrealistic in situations of practical interest.

The definition of the breakdown point as stated above is strictly applicable to the i.i.d. setting and the linear model. In a nonlinear setting, a different definition is needed. Consider for example the simple location model, where the location parameter is now defined as $\tan(\tilde{\mu})$, with $\tilde{\mu} \in (-\pi/2, \pi/2)$. Then by construction, no reasonable estimator of $\tilde{\mu}$ can diverge to infinity for any number of outliers. As a solution, an alternative definition of the breakdown point can be used, namely, the minimum fraction of contamination that is needed to drive the estimator to the edge of the parameter space. In the linearly parameterized location model, these edges are plus and minus infinity, while in the nonlinearly parameterized location model presented above, the edges are $\pi/2$ and $-\pi/2$. Using this definition, the least-squares estimator again has a breakdown point of zero. Note that this alternative definition is also suitable for scale estimators, which are usually not allowed to implode to zero.

Stromberg and Ruppert (1992) argue that the definition of the breakdown point in a nonlinear context should not be based on the edge of the parameter space. Their alternative definition considers the minimum fraction of outliers that is needed to drive the fitted regression function to the edge of its range. For the location model, these edges are plus and minus infinity. Consequently, the least squares estimator again has a breakdown point of zero: by placing one outlier at infinity, the estimate of $\tilde{\mu}$ (see the nonlinear parameterization above) tends to $\pi/2$, while the fitted regression function tends to infinity.

Neither of these alternative definitions proposed in the nonlinear context overcomes the problem one encounters when dropping the i.i.d. assumption. In a simple autoregressive model of order one, $y_t = \phi y_{t-1} + \varepsilon_t$, with $\phi \in (-1, 1)$, the OLS estimator is still nonrobust and has breakdown point zero. One outlier is sufficient to completely corrupt the OLS estimates (see Chapters 4 and 5). The OLS estimator of ϕ cannot diverge to infinity due to the requirement $-1 < \phi < 1$. This excludes the usefulness of the definition in (2.16) in this context. Moreover, the regression function does not diverge to the edge of its range, such that the definition of Stromberg and Ruppert (1992) does not apply either. Instead, for one extremely large outlier, the OLS estimator of ϕ settles near the center of the parameter space, at zero (see Example 4.1 in Chapter 4). Extensions of the breakdown point and of the notion of qualitative robustness to the non-i.i.d. setting can be found in, e.g., Papantoni-Kazakos (1984) and Boente et al. (1987).

To conclude this subsection, it is worth mentioning that the breakdown point can also be defined for other statistical procedures than estimators, e.g., test statistics (see He et al. (1990)).

2.2.4 Bias Curve

This subsection studies the maximum (absolute) change in an estimator brought about by arbitrarily changing a fraction of the original observations. The finite

sample bias curve plots the maximum change against the fraction of altered observations. The bias curve does the same thing, only in an asymptotic context. In order to introduce the bias curve, the finite sample bias curve is discussed first in Example 2.4 for the simple examples of the mean and the median.

Example 2.4 Again, consider the i.i.d. sample $y_1, \ldots, y_T$ from Example 2.1. Following Example 2.2, one obtains that the maximum possible effect on the mean of changing only one original observation, is infinite. Therefore, the finite sample bias curve of the mean assigns zero to the point $\eta_T = 0$, and infinity to each of the points $\eta_T = 1/T, 2/T, \ldots, 1$, where η_T denotes the fraction of altered observations.

For the median, attention is restricted to odd values of the sample size T. Assume that the sample is ordered as in Example 2.2. Again, for $\eta_T = 0$ one obtains that the maximum possible effect is zero. Using (2.10), one obtains that the maximum absolute change in the median is

$$\max\{|y_{(T-1)/2} - y_{(T+1)/2}|, |y_{(T+3)/2} - y_{(T+1)/2}|\}$$

if only one observation is altered. The first absolute difference arises if the largest order statistic is moved towards minus infinity, while the second absolute difference follows by moving the smallest order statistic towards plus infinity. Similarly, for $\eta_T = 2/T$ one obtains a maximum change of

$$\max\{|y_{(T-3)/2} - y_{(T+1)/2}|, |y_{(T+5)/2} - y_{(T+1)/2}|\},$$

where the first difference follows by moving the two largest order statistics towards minus infinity, and the second difference follows by moving the two smallest order statistics towards plus infinity. So the finite sample bias curve for the median assigns the value

$$\max\{|y_{(T+1)/2-n} - y_{(T+1)/2}|, |y_{(T+1)/2+n} - y_{(T+1)/2}|\}$$

to $\eta_T = n/T$, $n = 0, \ldots, (T+1)/2 - 1$. For the values of η_T greater than or equal to the breakdown point (see Example 2.3), the maximum possible absolute change in the median is infinite. △

The bias curve discussed in the present subsection encompasses the information provided by the breakdown point. The breakdown point only provides the maximum fraction of contamination an estimator can tolerate. The bias curve, in contrast, gives the maximum bias of the estimator as a function of the fraction of contamination. Following Hampel et al. (1986, p. 175), the bias curve is given by

$$\sup_G |\mu((1-\eta)F + \eta G) - \mu(F)|, \tag{2.17}$$

with denoting G an arbitrary distribution. Figure 2.2 gives the bias curve for the median, evaluated at the standard normal distribution.

The bias curve in Figure 2.2 has a vertical asymptote at the breakdown point of the median, 0.5. This is evident, as the maximum bias is infinite for

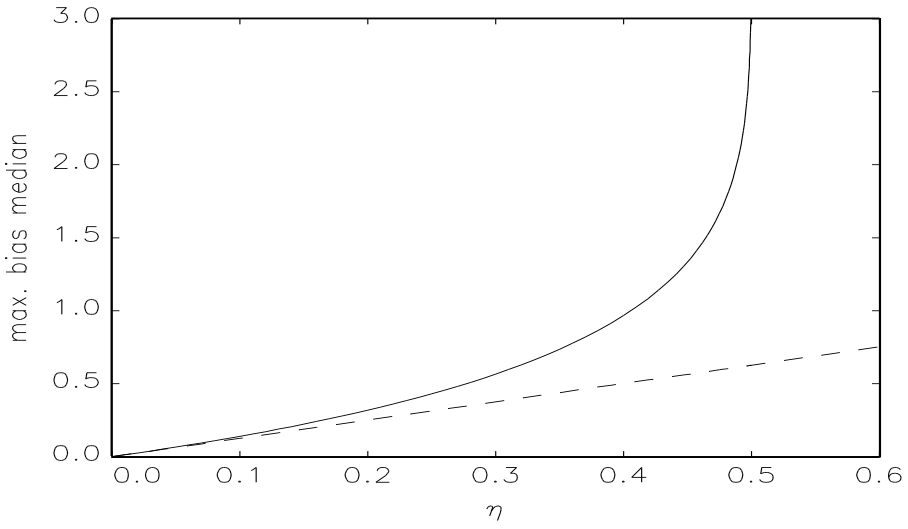

Figure 2.2.— Bias curve of the median (solid) at the standard normal distribution, and a linear approximation to the bias curve (dashed)

values of η greater than the breakdown point ε^*. The dashed curve gives the approximation to the bias curve mentioned in the previous subsection, $\eta\gamma^*$ (see page 26). It is easy to show that for many estimators the line $\eta\gamma^*$ is tangent to the bias curve in the point $\eta = 0$. Therefore, $\eta\gamma^*$ provides a reasonable approximation to the maximum bias if η is not too big, say smaller than $\varepsilon^*/2$.

The main advantage of the bias curve over the breakdown point, is that the bias curve provides more information. It does not only provide the maximum fraction of contamination that can be tolerated, but also the maximum bias for all values of η below the breakdown point. Therefore, if one is willing to assume an upper bound for the fraction of contamination, one can derive an upper bound for the possible bias from the bias curve.

It is illustrative to relate the information provided by the bias curve to the ideas put forward in the recent paper of Horowitz and Manski (1995). The main point of Horowitz and Manski is that in the presence of (possibly) contaminated data, one should not report point estimates, but rather interval estimates. The boundaries of this interval should reflect the ignorance about the precise form of the contamination. For example, if one is willing to assume that the fraction of contamination (η) encountered in a specific study does not exceed 0.25, then one can derive that the median of F must lie between the 0.375 and the 0.625 quantiles of $\tilde{F} = (1-\eta)F + \eta G$. Horowitz and Manski suggest to report this interval instead of a point estimate, such as the median of $\tilde{F}$.

If one is interested in interval estimates, one could alternatively report the median of $\tilde{F}$ plus or minus the value of the bias curve in the point $\eta = 0.25$.

This reflects the fact that one is aware of the possible bias in the obtained point estimate. The disadvantage of reporting this alternative interval as opposed to the interval put forward by Horowitz and Manski, is that one has to specify the target distribution F. From this perspective, the approach of Horowitz and Manski seems more promising for reporting bounds on parameter estimates. The paper of Horowitz and Manski (1995), however, has two major drawbacks. First, it only discusses several very simple cases, such that it remains to be seen how easily the approach can be generalized to the more complex models that are often encountered in econometrics. Second, Horowitz and Manski do not pay any attention to the need for robust estimation. They only focus on reporting intervals rather than point estimates. Consider the following example. Assume that the maximum fraction of contamination (η) is positive and that one wants to estimate the mean. Then the bounds on the mean derived by Horowitz and Manski are minus and plus infinity. Reporting such bounds is nonsensical. In general, one should always check whether the maximum fraction of contamination one is willing to assume, is below the breakdown point of the estimator one uses. Otherwise, it makes no sense to report the bounds of Horowitz and Manski. So, the approach of Horowitz and Manski provides complementary information to the three robustness concepts discussed earlier.

2.3 Robust Estimators

This section discusses several robust estimators. Subsection 2.3.1 treats the class of generalized maximum likelihood (GM) type estimators. Subsection 2.3.2 briefly touches upon some high breakdown estimators (compare Chapter 5).

2.3.1 GM Estimators

This subsection discusses the class of GM estimators for outlier robust estimation. This class of estimators contains the class of maximum likelihood type estimators (M estimators). Therefore, the class of M estimators is not dealt with, separately. In the appropriate places, it is mentioned how the results for GM estimators reduce to those for M estimators. The discussion is organized as follows. First, the class of GM estimators is introduced. Second, the IF and the breakdown point of GM estimators are discussed. Next, the effects of the use of GM estimators for testing are briefly treated. Finally, some attention is paid to computational aspects.

A. Definition

Consider the linear regression model $y_t = x_t^\top \beta + \varepsilon_t$, $t = 1, \ldots, T$, where x_t is a p-dimensional vector. The class of GM estimators is defined implicitly by

the first order condition

$$\sum_{t=1}^{T} x_t \zeta(x_t, (y_t - x_t^\top \beta)/\sigma) = 0, \tag{2.18}$$

see Maronna and Yohai (1981). The parameter σ denotes the scale of ε_t. The function $\zeta(\cdot,\cdot)$ depends on both the set of regressors (x_t) and the standardized residual. The precise conditions that must be satisfied by $\zeta(\cdot,\cdot)$ in order for the GM estimator to have nice asymptotic properties (like consistency and asymptotic normality), can be found in Hampel et al. (1986, p. 315). The most important requirements are that for all $x \in \mathbb{R}^p$ $\zeta(x,\cdot)$ has to be continuous and continuously differentiable except in a finite number of points, that $\zeta(x,\cdot)$ has no vertical asymptotes, and that $\zeta(x,\cdot)$ is odd.[7] Moreover, $E((\zeta(x_t, \varepsilon_t/\sigma))^2 x_t x_t^\top)$ and $E(\zeta'(x_t, \varepsilon_t/\sigma) x_t x_t^\top)$ must exist and be nonsingular, where $\zeta'(x_t, r) = \partial\zeta(x_t, r)/\partial r$, and r denotes the *standardized* residual ε_t/σ. Note that (2.18) is nonlinear in β, in general. The estimation is, therefore, usually carried out using numerical techniques.

The OLS estimator is obtained as a special case of (2.18) by setting $\zeta(x,r) = r$. Also M estimators are a special case of (2.18), namely $\zeta(x,r) = \psi(r)$ for some function ψ satisfying the above regularity conditions (see, e.g., Huber (1964)).

Instead of defining GM estimators as the solution to a first order condition of the type (2.18), one can also define them as the minimand of the objective function

$$\sum_{t=1}^{T} \tau(x_t, (y_t - x_t^\top \beta)/\sigma), \tag{2.19}$$

with $\partial\tau(x,r)/\partial r = \zeta(x,r)$. The focus in this chapter, however, is on the definition as implied by (2.18). Note that the OLS estimator is defined by setting $\tau(x,r) = r^2/2$, while the class of M estimators is obtained by setting $\tau(x,r) = \rho(r)$, with $d\rho(r)/dr = \psi(r)$.

The easiest way to explain the intuition behind GM estimators is by considering the class of Mallows' GM estimators, given by $\zeta(x,r) = w_x(x)\psi(r)$, with $\psi(r)$ as introduced above, and $w_x(x)$ a weight function that assigns weights to the vectors of regressors, $w_x : \mathbb{R}^p \to [0,1]$ (see (2.25) and below for more details on weight functions for the regressors). Using this specification of $\zeta(\cdot,\cdot)$, (2.18) can be rewritten as

$$\sum_{t=1}^{T} w_x(x_t) x_t \cdot w_r((y_t - x_t^\top \beta)/\sigma)(y_t - x_t^\top \beta) = 0, \tag{2.20}$$

with $w_r(r) = \psi(r)/r$ for $r \neq 0$, and $w_r(0) = 1$. The functions $\psi(\cdot)$ and $w_x(\cdot)$ can now be chosen such that the weight of the tth observation decreases if either $(y_t - x_t^\top \beta)/\sigma$ becomes extremely large (vertical outliers and bad leverage

[7] A function $f(x)$ is called odd if $f(-x) = -f(x)$. It is called even if $f(x) = f(-x)$.

points), or x_t becomes large (leverage points). In this way, outliers and influential observations automatically receive less weight. For the OLS estimator, $w_x(x) \equiv 1$ and $w_r(r) \equiv 1$, such that all observations receive the same weight.

A disadvantage of Mallows' proposal for GM estimators it that it assigns less weight to both good and bad leverage points. As was mentioned in Section 2.1, good leverage points often increase the efficiency of the employed estimator. As an alternative to Mallows' proposal for GM estimators, one can consider the proposal of Schweppe. The Schweppe form of the GM estimator only downweights vertical outliers and bad leverage points, but not good leverage points. This generally increases the efficiency of the Schweppe estimator over the Mallows version. The Schweppe specification of $\zeta(\cdot,\cdot)$ is given by

$$\zeta(x,r) = w_x(x) \cdot \psi(r/w_x(x)), \tag{2.21}$$

see Hampel et al. (1986). Using (2.21), (2.18) can be written as

$$\sum_{t=1}^{T} x_t \cdot w_r((y_t - x_t^\top\beta)/(\sigma w_x(x_t)))(y_t - x_t^\top\beta) = 0, \tag{2.22}$$

with $w_r(r) = \psi(r)/r$ for $r \neq 0$, and $w_r(0) = 1$. Assume that $w_x(\cdot)$ and $\psi(\cdot)$ are chosen such that outliers receive less weight. For a leverage point (y,x), $w_x(x)$ will then be small. The weight for the tth observation in the estimation process is given by the value of $w_r(\cdot)$ in (2.22). Note that this weight may be close to one if the standardized residual is close to zero, irrespective of whether the observation is a leverage point or not. The requirement that the standardized residual is close to zero becomes stricter if $w_x(x_t)$ is small, i.e., if x_t is a leverage point.[8] Some other specifications of $\zeta(\cdot,\cdot)$ can be found in Hampel et al. (1986, Section 6.3).

If the weights on the regressors ($w_x(\cdot)$) are dropped, the class of GM estimators reduces to the class of M estimators. This class is also well studied in econometrics, where it is more common to use the term pseudo or quasi maximum likelihood estimators (White (1982), Gouriéroux et al. (1984)). As a result, many of the properties of (G)M estimators can be found in both the statistical and econometric literature (see also the discussion on testing below).

So far, nothing has been said about the specifications of $\psi(\cdot)$ and $w_x(\cdot)$. The OLS specification for $\psi(\cdot)$, $\psi(r) = r$, is the most familiar one. As shown in Section 2.2, however, this estimator is not robust. The most important reason for this is that the function $\psi(r) = r$ is unbounded (see also the derivation of

[8] It is worth noting that the Schweppe version of the GM estimator also has some practical disadvantages. First, the bias in the Schweppe estimator may be larger than that of the Mallows estimator, see, e.g., Hampel et al. (1986, p. 323, Figure 2). Second, the Schweppe estimator more easily displays convergence problems than the Mallows variant, especially if strongly redescending specifications of ψ are used. Even if no convergence problems arise, moderately bad leverage points tend to have a larger influence on the Schweppe version of the GM estimator than on the Mallows version. Therefore, if the Schweppe version of the GM estimator is used, more attention has to be devoted to starting values and iteration schemes than when the Mallows version is used.

the IF of GM estimators below). Several forms of bounded ψ functions are suggested in the literature, e.g., the Huber, the bisquare, and the Student t specification.

The Huber ψ function is given by $\psi(r) = \text{median}(-c, c, r)$, where $c > 0$ is a tuning constant. The lower c, the more robust the resulting estimator. As a special case of the Huber estimator, one can obtain the OLS estimator ($c \to \infty$) and the least absolute deviations (LAD) estimator ($c \downarrow 0$). The constant c not only determines the robustness of the corresponding estimator, but also its efficiency. For Gaussian ε_t, for example, the efficiency of the estimator is an increasing function of c. This illustrates that there is a tradeoff between efficiency and robustness. There are two common approaches for choosing c. First, one can specify a maximum amount of efficiency loss at a specific target model, e.g., the normal, and fix c accordingly. Second, one can fix the maximum influence of single observations on the estimator, i.e., impose a bound on the IF. This can lead to a different value of c.

The remaining two specifications of $\psi(\cdot)$ mentioned above, are the bisquare function,

$$\psi(r) = \begin{cases} r(1-(r/c)^2)^2 & \text{for } |r| \leq c, \\ 0 & \text{for } |r| > c, \end{cases} \tag{2.23}$$

and the Student t function,

$$\psi(r) = (1 + c^{-1})r/(1 + r^2/c). \tag{2.24}$$

Common values for c are 1.345 for the Huber function and 4.685 for the bisquare function. These values produce estimators that have an efficiency of 95% in case $x_t \equiv 1$ and ε_t is normally distributed. For the Student t, one can either estimate the degrees of freedom parameter c along with the other parameters (see also Chapter 3), or fix it at approximately 6 for an efficiency of 95% at the normal.

As a specification for the weight function $w_x(\cdot)$ for the regressors, one usually encounters the specification

$$w_x(x) = \psi(d(x_t)^\alpha)/d(x_t)^\alpha, \tag{2.25}$$

with $\alpha > 0$ and $d(x_t)$ the Mahalanobis distance of x_t,

$$d(x_t) = \sqrt{(x_t - m)^\top V^{-1}(x_t - m)}, \tag{2.26}$$

with m and V a location and scatter/covariance measure, respectively. More on the use of the Mahalanobis distance and on the estimation of m and V can be found in Subsection 2.4. Simpson et al. (1992) propose $\alpha = 2$ in (2.25) with $\psi(\cdot)$ equal to the Huber function. The tuning constant c of the Huber function is set to some high quantile, e.g., 0.975, of the χ^2 distribution with p degrees of freedom. Note that if the set of regressors contains a constant, this constant has to be excluded from both (2.25) and (2.26).

B. Influence function and breakdown point

The IF of GM estimators is fairly easy to derive. It is assumed throughout that the necessary regularity conditions for the steps below (e.g., interchanging differentiation and integration) are satisfied. The functional version of the GM estimator is given by $\beta(F)$, where $\beta(F)$ solves

$$E(x\zeta(x,(y-x^\top\beta(F))/\sigma))=0, \tag{2.27}$$

with F the joint distribution of ε and x, and with the expectation taken both with respect to ε and x. Assume for simplicitly that σ is known and equal to one. Define $\tilde{F}=(1-\eta)F+\eta\Delta_{(\varepsilon_0,x_0)}$, as in Subsection 2.2.2. Replacing F in (2.27) by $\tilde{F}$, taking derivatives with respect to η, and evaluating in $\eta=0$, produces

$$\frac{d}{d\eta}\int x\zeta(x,y-x^\top\beta(\tilde{F}))d(\eta(\Delta_{(\varepsilon_0,x_0)}-F)+F)(y)\Big|_{\eta=0}=0\Longleftrightarrow$$

$$\int x\zeta(x,\varepsilon)d(\Delta_{(\varepsilon_0,x_0)}-F)(y)+\int x\frac{\partial\zeta(x,\varepsilon)}{\partial\varepsilon}(-x^\top)\,\frac{d\beta(\tilde{F})}{d\eta}\Big|_{\eta=0}dF(y)=0\Longleftrightarrow$$

$$x_0\zeta(x_0,\varepsilon_0)-E(\zeta'(x,\varepsilon)xx^\top)\,\frac{d\beta(\tilde{F})}{d\eta}\Big|_{\eta=0}=0, \tag{2.28}$$

where $\zeta'(x,\varepsilon)=\partial\zeta(x,\varepsilon)/\partial\varepsilon$. Note that $d\beta(\tilde{F})/d\eta|_{\eta=0}$ equals the IF of $\beta(\cdot)$. Therefore,

$$IF((\varepsilon_0,x_0),\beta,F)=\left[E(\zeta'(x,\varepsilon)xx^\top)\right]^{-1}x_0\zeta(x_0,\varepsilon_0). \tag{2.29}$$

According to the assumptions mentioned in the subsubsection on the definition of GM estimators, the matrix in brackets is nonsingular and finite. Therefore, the boundedness of the IF completely depends on the behavior of $x\zeta(x,\varepsilon)$. By choosing $\zeta(\cdot,\cdot)$ such that $x\zeta(x,\varepsilon)$ is bounded in both x and ε, a GM estimator with a bounded IF is constructed.

A discussion of the breakdown point of GM estimators is much more delicate, as the breakdown point depends on both the initial estimator that is used to solve (2.18) numerically, and on the estimators that are used for m and V in the Mahalanobis distance (2.26). Following Maronna et al. (1979), the breakdown point of GM estimators decreases rapidly with p if $\zeta(x,\cdot)$ is a monotone function for all $x\in\mathbb{R}^p$. Simpson et al. (1992) and Coakley and Hettmansperger (1993), however, show that certain one-step versions of GM estimators that use high breakdown starting values, have a high breakdown point. Such one-step GM estimators have the same asymptotic efficiency as their fully iterated counterparts.

C. Testing

GM estimators can be used to constuct Wald, Lagrange Multiplier (LM), and Likelihood Ratio (LR) type test statistics. Hampel et al. (1986, Chapter

7) show that these test statistics have a bounded IF if the underlying GM estimator has a bounded IF. Define the function $\tau(x,r)$ such that $\partial\tau(x,r)/\partial r = \zeta(x,r)$. Then the easiest way to think about tests based on GM estimators, is to regard the function $\tau(\cdot,\cdot)$ as the log likelihood. This also establishes a link between the statistical literature on robust testing (see, e.g., Hampel et al. (1986) and the references cited therein) and the econometric literature on likelihood based testing under misspecification of the likelihood (see White (1982) and Chapter 7).

Both from White (1982) and from Hampel et al. (1986), it follows that the Wald and LM tests based on GM estimators have standard χ^2 limiting distributions under conventional assumptions. The LR type test, however, has certain nuisance parameters in the limiting distribution. These nuisance parameters arise due to the discrepancy between the correct maximum likelihood estimator and the employed GM estimator. If the GM estimator coincides with the true maximum likelihood estimator, then there are no nuisance parameters and the LR type test also has a standard χ^2 limiting distribution. Inference based on the LR principle, thus, usually involves more complications in a context with GM estimators than inference based on the Wald or LM testing principle.

D. Computational aspects

It was already mentioned that GM estimators are mostly computed by means of numerical techniques. Most of these techniques employ iteration schemes (see, e.g., Marazzi (1991)). Therefore, an initial estimate is required to start up the iterations. I first discuss some possibilities for constructing a starting value. Then, I devote some attention to possible iteration schemes. Finally, a few words are spent on the estimation of the scale parameter σ in (2.18).

A starting value should, preferably, be easy to calculate. From this perspective, the OLS estimator seems a first good candidate. It is, however, not a good idea to use the OLS estimator as a starting value for the GM estimator, because this estimator is not robust.[9] As a consequence, one might start the iterations in a region of the parameter space that is too far from the true parameter values. Especially if the objective function defining the GM estimator has several local optima, there is a high risk that the GM estimator ends up in an incorrect (local) optimum and produces nonrobust estimates. As an alternative to the OLS estimator, one can use the least absolute deviations (LAD) estimator. This estimator is also reasonably easy to compute and does not require the additional estimation of a scale parameter σ. The main disadvantage of the LAD estimator is that it does not provide protection against leverage points. In order to remedy this problem, a weighted version of the LAD estimator can be used, with weights $w_x(x_t)$. Yet another alternative is to use one of the high breakdown point (HBP) estimators of Subsection 2.3.2.

[9] It is not always possible, however, to come up with a robust initial estimator to start up the iterations. If no robust estimator can be found, one just has to be satisfied with nonrobust parameter estimates as starting values, see, e.g., Lucas et al. (1994).

The advantage of these HBP estimators is that it is much more likely that the GM estimator produces parameter estimates that describe the bulk of the data. The main disadvantage of using HBP estimators as starting values is that they are often time consuming to compute.

Once the starting values have been obtained, one can start an iteration scheme for solving (2.18). It is, of course, possible to use general techniques for solving sets of nonlinear equations. The special structure of (2.18), however, also allows a different iteration scheme. Let $\hat{\beta}^{(n)}$ denote the trial estimate of β in the nth iteration, and let $\hat{\beta}^{(0)}$ denote the initial estimate. Rewrite (2.18) as

$$\sum_{t=1}^{T} x_t \tilde{w}_r(\beta, \sigma)(y_t - x_t^\top \beta) = 0, \tag{2.30}$$

with

$$\tilde{w}_r(\beta, \sigma) = \zeta(x_t, (y_t - x_t^\top \beta)/\sigma)/(y_t - x_t^\top \beta).$$

Assume for the moment that the scale parameter σ is known. For notational simplicity, σ is now deleted from the weigth function $\tilde{w}_r$, which is now denoted as $\tilde{w}_r(\beta)$. Estimation of σ is discussed below. By replacing $\tilde{w}_r(\beta)$ and $x_t^\top \beta$ in (2.30) by $\tilde{w}_r(\hat{\beta}^{(n)})$ and $x_t^\top \hat{\beta}^{(n+1)}$, respectively, one obtains that $\hat{\beta}^{(n+1)}$ is the weighted least-squares estimate

$$\hat{\beta}^{(n+1)} = \left(\sum_{t=1}^{T} \tilde{w}_r(\hat{\beta}^{(n)}) x_t x_t^\top \right)^{-1} \sum_{t=1}^{T} \tilde{w}_r(\hat{\beta}^{(n)}) x_t y_t, \tag{2.31}$$

for $n \geq 0$. So starting from the initial estimate $\hat{\beta}^{(0)}$, one obtains the next trial estimate by means of a weigthed least-squares regression. This provides a quick iteration scheme. Convergence, however, is not always guaranteed, and problems can be expected if $\zeta(x, \cdot)$ is discontinuous or if $\zeta'(x, \cdot)$ alternates sign.

A final computational aspect concerns the estimation of the scale parameter σ. If σ is omitted from (2.18), the GM estimator is not scale invariant, i.e., the estimates would change if both y_t and x_t were multiplied by a constant $c > 0$. To estimate σ, one cannot safely use the ordinary standard deviation, as this estimator is not robust. An often used alternative is the median absolute deviation, defined as

$$MAD(\{\varepsilon_t\}_{t=1}^{T}) = \text{median}|\varepsilon_t - \text{median}(\varepsilon_t)|. \tag{2.32}$$

The MAD is usually multiplied by 1.4826 to make it a consistent estimator of the standard deviation for Gaussian ε_t. (2.32) reveals that the MAD is a scale equivariant estimator of σ, i.e., if both y_t and x_t are multiplied by a positive constant c, then the MAD also has to be multiplied by c. The use of a scale equivariant estimator for σ in (2.18) renders the GM estimator for β scale invariant. For more scale equivariant estimators for σ, see Hampel et al. (1986).

Given an estimate $\hat{\beta}^{(n)}$, one can construct a scale estimate $\hat{\sigma}^{(n)}$, e.g., by using (2.32). This scale estimate can be put into (2.18) or (2.30), after which

$\hat{\beta}^{(n+1)}$ and $\hat{\sigma}^{(n+1)}$ can be computed. In certain cases, it is better to iterate over β only, while keeping σ fixed at $\hat{\sigma}^{(0)}$, see Andrews et al. (1972), Yohai (1987), and Maronna and Yohai (1991).

2.3.2 High Breakdown Estimators

As mentioned in the previous subsection, the breakdown point of GM estimators is, in general, a decreasing function of the dimension of the problem at hand. This means that for high dimensional problems, e.g., many regressors, the breakdown point is unacceptably low. In order to solve this defect of GM estimators, other robust estimators were introduced with a breakdown point independent of the dimension of the problem. Most of these estimators obtain a breakdown point near 0.5. Quite a few of these high breakdown point (HBP) estimators are available in the literature (see the references in Chapter 5). This subsection discusses some of them in more detail. The estimators that are discussed are the least median of squares (LMS) estimator, the class of S estimators, and two HBP-GM estimators. The class of MM estimators is treated in Chapter 5.

A. LMS

The OLS estimator minimizes the sum of squared residuals. As explained in Subsection 2.3.1, GM estimators replace the squaring operator by a function that increases less rapidly for large residuals. The breakdown point of GM estimators, however, can still be quite low if the number of regressors is large. Therefore, Rousseeuw (1984) follows a different line for robustifying the OLS estimator. By dividing the objective function of the OLS estimator, i.e., the sum of squared residuals, by the sample size, the OLS estimator is seen to minimize the *mean* of squared residuals. Rousseeuw then replaces the mean by a robust estimator of location, namely the median. The resulting estimator, least median of squares (LMS), thus minimizes the median of the squared residuals.

The LMS estimator has a high breakdown point of approximately 0.5. The efficiency properties of the estimator, however, are very poor. It converges at the rate $T^{1/3}$ instead of the usual $T^{1/2}$. Moreover, all present algorithms for computing the estimator are very time consuming. Techniques for approximating the LMS estimator are provided in Rousseeuw and Leroy (1987). In a model with only one regressor and an intercept, the LMS estimator can be computed exactly (Edelsbrunner and Souvaine (1990)). Finally, Davies (1993) shows that the IF of the LMS estimator is unbounded. The estimator can be very sensitive to small changes in observations that lie near the center of the design space.

A direct generalization of both the OLS and LMS estimator is the class of S estimators. The mean of squared residuals produces the traditional estimate of the variance for a sample with known mean zero. Similarly, the median of the squared residuals produces the square of the (unscaled) MAD for a

sample with known median zero. Both the OLS and the LMS estimator, thus, minimize an estimator for the scale of ε_t. Other estimators for the scale of ε_t are readily available and can also be used as objective functions for estimators of β. This produces the class of S estimators (Rousseeuw and Yohai (1984)). The main advantage of S estimators over the LMS estimator is that the the objective function of the S estimator can be chosen such that the S estimator becomes $T^{1/2}$ consistent, while retaining the high breakdown point of 0.5. This high breakdown point, however, comes at a considerable efficiency loss at the Gaussian distribution.

B. HBP-GM estimators

As the high breakdown point of the LMS estimator and of S estimators is, in general, counterbalanced by a low efficiency, researchers have sought for more efficient estimators that retain the high breakdown point. This resulted in the introduction of MM estimators (Yohai (1987)), HBP-GM estimators of the Mallows form (Simpson et al. (1992)), and HBP-GM estimators of the Schweppe form (Coakley and Hettmansperger (1993)). The main idea behind these estimators is to use the inefficient high breakdown estimates as starting values for an efficient (G)M estimation procedure. By performing only one Newton step of the iteration process to compute the (G)M estimator, an efficient estimator can be constructed that retains the high breakdown point of the initial estimator.

The MM estimator is discussed in more detail in Chapter 5. The other two estimators can be constructed such that they have a bounded IF, are $T^{1/2}$ consistent, have a breakdown point of approximately 0.5, and have a high efficiency at the central model, e.g., the normal. These estimators, therefore, seem very promising. Their main disadvantage is the required computational cost for computing the estimators. First, one has to compute HBP initial estimates. As mentioned earlier, this is often time consuming. Second, the weights for the regressors x_t also have to be based on HBP estimators for multivariate scatter and location. Computing these estimators is again computer intensive, thus further increasing the computational costs of the HBP-GM estimators (see also Section 2.4).

2.4 Multivariate Location and Scatter

One of the elements of the GM estimator in Section 2.3 is the weight function for the set of regressors, $w_x(\cdot)$. This function depends on the Mahalanobis distance given in (2.26). In order to compute this distance, estimators for multivariate location (m) and scatter (V) are needed. It is, of course, possible to use the sample mean and the sample covariance matrix for m and V, respectively. These estimators are, however, not robust, thus making the weight function $w_x(\cdot)$ and the complete GM procedure nonrobust. Therefore, robust estimators of location and scatter are needed. This section discusses some of

these estimators. Subsections 2.4.1 and 2.4.2 discuss low and high breakdown robust estimators for location and scatter, respectively. Some miscellaneous remarks on location and scatter estimation are gathered in Subsection 2.4.3.

2.4.1 Low Breakdown Estimators

Maronna (1976) generalizes M estimators for the univariate location/scale model to the multivariate context. Let $\{x_t\}_{t=1}^T$ denote a set of p-dimensional i.i.d. random vectors. Maronna (1976) suggests to estimate m and V in (2.26) by solving

$$T^{-1}\sum_{t=1}^{T} u_1(d_t)(x_t - m) = 0, \tag{2.33}$$

$$T^{-1}\sum_{t=1}^{T} u_2(d_t)(x_t - m)(x_t - m)^\top = V, \tag{2.34}$$

with repect to m and V, where $d_t = d(x_t)$ is the Mahalanobis distance given in (2.26), and where $u_1(\cdot)$ and $u_2(\cdot)$ are weight functions. The precise assumptions that must be satisfied by $u_1(\cdot)$ and $u_2(\cdot)$ can be found in Maronna (1976) and in Hampel et al. (1986). One can again employ an iterated weighted least squares algorithm to compute m and V (compare Subsection 2.3.1).

The intuition behind (2.33) and (2.34) is similar to that behind (2.30). Vectors x_t that lie far from the bulk of the data automatically receive a smaller weight for appropriate choices of $u_1(\cdot)$ and $u_2(\cdot)$. The nonrobust mean and standard covariance matrix are obtained by assigning weight one to all observations, i.e., $u_1(x) \equiv 1$ and $u_2(x) \equiv 1$. By letting the weight functions become smaller for large d_t, robust estimators of location and scatter can be obtained, i.e., estimators with a bounded IF (see Maronna (1976)). An obvious way to construct suitable weight functions is by setting $u_i(d_t) = \psi(d_t)/d_t$ for $i = 1, 2$, with $\psi(\cdot)$ a bounded function. For example, if $\psi(\cdot)$ is the Huber function with tuning constant c, observations that are near the bulk of the data are fully taken into account. Observations that are too distant, i.e., $d_t > c$, receive an ever decreasing weight.

The tuning constant c can be chosen as follows. Let $F(x)$ denote the target model for the x_t vectors. Then c can be chosen such that the estimators for m and V assign unit weight to 95% of the observations under the target model. If, for example, F is the multivariate normal distribution, $u_i(d_t) = \psi(d_t)/d_t$, and $\psi(\cdot)$ is the Huber function, then c can be set to the 0.95 quantile of the χ^2 distribution with p degrees of freedom. Other percentages than 95% can also be used.

A final note concerns the consistency of V as implicitly defined in (2.33) and (2.34). Consistency of estimators for m and V is typically only proved for spherically symmetric target distributions. If the only condition on $u_1(\cdot)$ and $u_2(\cdot)$ is that these functions are nonincreasing for $d_t > 0$, the estimator for V may be robust, but inconsistent. Consider the simple example with

$u_1(\cdot) = u_2(\cdot)$ and $u_1(d_t) = 1_{\{d_t^2 \leq c\}}(d_t)$, with $1_A(\cdot)$ the indicator function of the set A. If the target model is the multivariate normal distribution, then the estimator for m is a consistent estimator of the mean of the multivariate normal. The estimator for V, however, does not consistently estimate the covariance matrix of the target distribution, as it discards observations in the tails for $c < \infty$. This causes the estimate of V to be 'too low,' in general. A solution to this problem is to redefine $u_2(\cdot)$ as $u_2(x) = \tau \cdot u_1(x)$, with τ such that

$$\int \tau u_2(\sqrt{x^\top x}) x x^\top dF(x) = I,$$

where F is the multivariate *standard* normal distribution. In the example above, one needs $\tau^{-1} = p^{-1} \int_0^c z d\chi_p^2$, with χ_p^2 the χ^2 distribution function with p degrees of freedom. The M estimator for V then consistently estimates the variance-covariance matrix of the normal distribution. Note that $\tau \to 1$ for $c \to \infty$: the ordinary covariance matrix estimator is consistent for the Gaussian distribution.

2.4.2 High Breakdown Estimators

Maronna (1976) showed that the breakdown point of ordinary M estimators of multivariate location and scatter is bounded from above by $(1 + p)^{-1}$. Therefore, if the number of regressors is moderately large, the breakdown point of the estimators is quite low. This has obvious consequences for the breakdown point of GM estimators that are based on M estimates of location and scatter.

In order to improve the breakdown behavior of M estimators for location and scatter, various alternative estimators were developed. Most of these attain a breakdown point of approximately 0.5 in large samples. Rousseeuw (1985) introduced the minimum volume ellipsoid (MVE) estimator. The MVE estimator resembles the LMS estimator. It looks for the ellipsoid with the smallest volume covering at least half of the observations. The center of this ellipsoid is taken as an estimator of location, while the metric matrix defining the ellipsoid is used to construct an estimate of the scatter matrix V. Just as with the LMS estimator, the MVE estimator only converges at the rate $T^{1/3}$, see Davies (1992). Quicker convergence is achieved by the class of S estimators for multivariate location and scatter, Davies (1987).[10] Under usual regularity conditions, these estimators converge at a rate $T^{1/2}$ to a Gaussian limiting distribution, while retaining the high breakdown point of 0.5. Lopuhaä (1989) shows that S estimators satisfy the same type of first order conditions as M estimators, namely (2.33) and (2.34). Therefore, an asymptotic analysis that only rests upon simple expansions of the first order conditions (2.33) and (2.34), produces results that apply to both S and M estimators (see also the remarks

[10] See Chapter 5 for some more details on S estimators. Chapter 5 uses S estimators in the regression setting, but the principles are similar to the ones used in the context of estimation of multivariate location and scatter.

in Chapters 5 and 6). Other high breakdown estimators for multivariate location and scatter include the minimum covariance matrix determinant (MCD) estimator, MM estimators and τ estimators (Lopuhaä (1990)), and projection based estimators (see, e.g., Hampel et al. (1986, Section 5.5)).

The computation times of high breakdown estimators of multivariate location and scatter are often high. This holds especially for large values of p. Therefore, one usually relies on heuristic algorithms for approximating the estimators. Some of these algorithms can be found in, e.g., Rousseeuw and Leroy (1987), Rousseeuw and van Zomeren (1990) and Woodruff and Rocke (1994).

2.4.3 Miscellaneous Remarks

All of the estimators described in Subsections 2.4.1 and 2.4.2 are affine equivariant. Let $\tilde{x}_t = Ax_t + \mu$ for some nonsingular matrix A. Moreover, let $\hat{m}$ and $\hat{V}$ denote the estimates of m and V based on the x_t vectors, and let $\tilde{m}$ and $\tilde{V}$ denote the estimates for the $\tilde{x}_t$ vectors. Then the estimators for m and V are called affine equivariant if $\tilde{m} = A\hat{m} + \mu$ and $\tilde{V} = A\hat{V}A^\top$. Affine equivariance is a natural condition for estimators of multivariate location and scatter. Affine equivariant estimators transform in a natural way under linear transformations of the data. If one drops the requirement of affine equivariance, other estimators can be constructed. Some of these have a high breakdown point and are easy to compute, for example, the coordinate-wise median of the x_t vectors. Such estimators can be used as starting values for the more computer intensive estimators. This illustrates that the large computational burden of the high breakdown procedures of Subsection 2.4.2 is partly caused by the requirement that the estimators be affine equivariant.

Another point concerns the estimation of location and scatter for x_t vectors that are not spherically distributed. The Mahalanobis distance implicitly builds upon the assumption that the x_t vectors follow an elliptical distribution. Based on this assumption, one can construct a 'tolerance ellipsoid.' Points outside this ellipsoid are viewed as outliers. For non-spherical x_t, however, the Mahalanobis distance seems inappropriate, and one must think of alternative ways for constructing tolerance regions. In multivariate settings, this problem is far from trivial and satisfactory solutions have yet to be proposed. One possibility is to base the tolerance region on contours of a robust nonparametric density estimate of the x_t vectors. Constructing such a density estimate, however, is difficult and computationally demanding, especially for large p. The fact that the Mahalanobis distance is not really appropriate for non-spherical x_t should always be kept in mind when applying GM estimators and robust estimators of location and scatter to empirical data sets.

A final remark relates to the specification of the functions $u_1(\cdot)$ and $u_2(\cdot)$. The specification one usually encounters in the literature is $u_i(x) = \psi(d_t^\alpha)/d_t^\alpha$, with α equal to one or two, and $\psi(\cdot)$ equal to the Huber function. The tuning constant of the Huber function is usually some high quantile of the χ^2 distribu-

tion with p degrees of freedom. The advantage of this specification of the weight functions is that one obtains interpretable weights. Observations near the center receive weight one, while more distant data points get a smaller weight. As an alternative to the Huber function, one can use $\psi(d_t) = d_t \cdot 1_{d_t<c}(d_t)$. This function also yields interpretable weights, with distant observations now receiving a weight equal to zero.

Different specifications of the function ψ are, of course, also possible. Examples are the bisquare function and the Student t function (see Subsection 2.3.1). These functions are scarcely used, however. Two possible reasons for this are the lack of interpretation of the resulting weights and the difficulty of choosing an appropriate value for the tuning constant. To illustrate these points, consider the example of the Student t function: $\psi(d_t) = (1+c^{-1})d_t/(1+d_t^2/c)$ and $u_i(d_t) = \psi(d_t)/d_t$. For $d_t > 1$ the weight of the tth observation is smaller than one, while the reverse holds for $0 \leq d_t < 1$. This does not depend on the value of c. The problem is now how to relate the value of the weights to statements about the 'outlyingness' of the corresponding observation. Clearly, weights below one are no longer a signal that the observation is an outlier. The weight must be *much* smaller than one. How much exactly remains open to debate. Also note that perfectly regular observations can easily receive a weight below unity and, thus, not be fully taken into account. Related to this point is the difficulty of choosing the appropriate tuning constant. For the Huber function mentioned above, a suitable tuning constant is some high quantile of the χ^2 distribution. For such a value, centrally located observations get a unit weight. For other functions, such as the Student t function, the tuning constant must be chosen such that the weight of non-outlying observations is as high as possible, while that of outliers is (much) lower. In view of the discussion above, this can be a problem for some specifications of $\psi(\cdot)$.

2.5 Model Selection and Evaluation

This section gathers some remarks on model selection and model evaluation in a context with outlier robust estimators. Subsection 2.5.1 treats outlier robust model selection, while Subsection 2.5.2 deals with residual analysis and outlier robust diagnostic testing.

2.5.1 Robust Model Selection

In the process of building a model, one is frequently confronted with the question whether one model fits the data better than another model. Various techniques have been developed to answer this question. Most of these are equally applicable in a context with robust and a context with nonrobust estimators. Some others have to be slightly revised in order to account for the use of other estimation principles than least-squares.

As mentioned in Subsection 2.3.1, one can construct robust Wald, LM, and LR type test statistics based on outlier robust estimators. These test statistics

can be used to perform significance tests on (groups of) variables. Moreover, they can be used to test restrictions on the parameters in the model. In sum, these robust test statistics provide one of the means for choosing between different models. Under usual regularity conditions the Wald and LM type tests have limiting χ^2 distributions, while the limiting distribution of the LR test is a weighted sum of independent χ^2 variables.

Instead of using statistical tests to discriminate between models, one can look at other model selection criteria. Two well-known statistics that are used in this context are the Akaike information criterion (AIC) and the Schwarz criterion (see, e.g., Judge et al. (1988)). Both of these criteria are based on the value of the likelihood at the parameter estimates. As explained in Chapter 7, the objective function defining a (G)M estimator can be interpreted as a (pseudo) likelihood. By using this pseudo likelihood in the above model selection criteria, robust versions of the AIC and SC can be constructed. This is the idea underlying the papers of Ronchetti (1985) and Machado (1993) (see also Hampel et al. (1986, Section 7.3d)). Strongly related to the AIC is Mallows' C_p statistic, a robust version of which is proposed in Ronchetti and Staudte (1994).

As explained in Chapter 1, there is a philosophical problem when selecting a model in a context with outliers. Robust procedures try to find the best model for the bulk of the data. This might result in the wrong model[11] being chosen, as was illustrated by the example in Section 1.2. The empirical support for the correct model, however, might consist of only a few extreme observations. This requires that one has a firm belief in such a model on a priori grounds, for if these few extreme observations turn out to be incorrect, the whole model can collapse. In any event, it is useful to know whether the selected model hinges on only a few observations or whether it is supported by the majority of the data. Robust model selection procedures provide useful tools to answer such questions (see also the examples in Ronchetti and Staudte (1994)).

2.5.2 Residual Analysis and Model Diagnostics

Related to the topic of model selection is the topic of model evaluation. The fitted model is then checked for possible misspecification. Directions of misspecification include neglected heteroskedasticity, serial correlation, omitted variables, nonlinearity, simultaneity, and so on. Tests for these types of misspecification are amply available in the literature, see, e.g., Godfrey (1988) and Pagan and Vella (1989) and the references cited therein. Most of these tests check the properties of the regression residuals. Therefore, this subsection concentrates on some of the properties of residuals from robust regressions.

The traditional OLS estimator minimizes the sum of squared errors. Every error, thus, obtains the same weight. Robust estimation procedures, in

[11]Recently, Davies (1995) has objected against the terminology of 'true model' and the objective of looking for the true model. In his terminology, the above so called 'wrong' model might still give an adequate description of the data set at hand.

contrast, weight large errors less heavily. As a result, the residuals for outlying observations may be much larger when an outlier robust estimator is used than when the OLS estimator is employed. Putting residuals from an outlier robust modeling exercise into a traditional diagnostic test can, therefore, cause problems if there are outliers in the data. Moreover, the OLS estimator is linear and has nice geometric properties, while robust estimators are, in general, nonlinear and fail the geometric interpretation of the OLS estimator. This can also cause problems when using robust residuals in traditional model evaluation procedures, see, e.g., Cook et al. (1992) and McKean et al. (1993). In sum, one must take great care when using robust residuals in standard model evaluation tests.

The fact that the robust residuals for outlying observations are generally larger in absolute value than residuals based on the OLS estimator, can also be turned into an advantage. By inspecting these residuals, one gets a much clearer signal than with the OLS residuals about which observations do not fit into the model. Moreover, in a linear regression setting, one can also plot the estimated Mahalanobis distances of the observations. Plotting the robust residuals versus the robust distances provides valuable information about the types of outliers one is dealing with (see Rousseeuw and van Zomeren (1990) and Fung (1993)).

The difference between robust and OLS-based residuals requires a modification of traditional diagnostic tests. Sometimes, these modifications are quite natural and straightforward. For example, the Breusch-Pagan test for heteroskedasticity regresses the log of the squared residuals on a set of explanatory variables (see Judge et al. (1988, p. 372)). Usually, the OLS estimator is employed to perform the estimation. In a context with outlier robust estimators, however, it quite natural to replace the OLS estimator by a robust estimation technique. The performance of diagnostic tests based on robust auxiliary regressions can be evaluated both by means of asymptotic analyses and by means of Monte-Carlo simulations. This is left as a topic for future research.

Another straightforward modification applies when the diagnostic test is an LM test based on a Gaussian likelihood. In that case, the LM test can be based on a different likelihood or on the objective function of a GM estimator. This produces a pseudo LM test. For example, when testing for first order autocorrelation in the residuals $\hat{\varepsilon}_t$, the LM test based on the Gaussian likelihood is given by

$$T \frac{\left(\sum_{t=1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2 \hat{\varepsilon}_{t-1}^2}. \tag{2.35}$$

Using the GM objective function discussed in Subsection 2.3.1, a robust variant of this test is

$$T \frac{\left(\sum_{t=1}^{T} \zeta(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t/\hat{\sigma}) \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T} \zeta(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t/\hat{\sigma})^2 \hat{\varepsilon}_{t-1}^2} = T \frac{\left(\sum_{t=1}^{T} \hat{w}_t \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=1}^{T} \hat{w}_t^2 \hat{\varepsilon}_t^2 \hat{\varepsilon}_{t-1}^2}, \tag{2.36}$$

with $\zeta(\cdot,\cdot)$ as in (2.18), $\hat{\sigma}$ a scale estimate, and $\hat{w}_t = \hat{\sigma}\zeta(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_t/\hat{\sigma})/\hat{\varepsilon}_t$. Large residuals are automatically downweighted in (2.36), whereas they are fully weighted in (2.35). It is again left for future research to systematically evaluate the properties of robust diagnostic LM tests under various circumstances.

To conclude, outlier robust model evaluation still remains a largely open area. Some traditional diagnostic procedures can directly be applied in a context with robust estimators. The properties of these diagnostic procedures, however, have not yet been extensively studied in a context where the data contain outliers. Other diagnostic tests require some modification before they can be applied to robust residuals. Sometimes, the required modification is straightforward, but again the properties of these modified diagnostic tests have not yet been well documented in the literature. Moreover, great care has to be taken when proposing simple modifications of standard diagnostic tests, especially in the context of high breakdown estimation (see Cook et al. (1992) and McKean et al. (1993)). The lack of attention paid to robust diagnostic tests is partly due to the lack of serious empirical exercises performed with outlier robust procedures. Therefore, in order to stimulate the development of outlier robust model evaluation procedures, one can also try to stimulate applied researchers to use outlier robust techniques for their empirical work.[12]

[12]Note that this argument can also be turned the other way around: in order to stimulate the use of outlier robust procedures in applied work, the development of outlier robust model evaluation procedures is a first requirement. Otherwise, applied researchers will not use the robust methods, because they have no techniques for evaluating their finally selected model.