Applied Robust Regression in Stata

Applied Robust Regression in Stata

Ben Jann
Institute of Sociology, University of Bern, Switzerland

Vincenzo Verardi
University of Namur and Université Libre de Bruxelles, Belgium

Catherine Vemandele
Université Libre de Bruxelles, Belgium

A Stata Press Publication
StataCorp LP
College Station, Texas

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: !!
ISBN-13: !!

Library of Congress Control Number: !!

The acknowledgments go here.

Contents

Tables

Figures

Preface

[The book introduces robust statistics in Stata from an applied perspective. We review existing commands and present a variety of new tools, give advice on how to choose among the different estimators and illustrate how they can be applied in practice. After a general introduction the book first discusses robust estimation of univariate location and scale and, along the way, briefly introduces the basic concepts of robust statistics. The book then moves on to simple and multiple robust regression and models for qualitative dependent variables, each time reviewing (briefly) the theory, presenting the algorithms, commands, and implementation details, and providing applied examples. Furthermore, we discuss multivariate identification of outliers and present robust versions of factor models] ...

Notation and typography

Math typesetting conventions and custom commands (to be removed later)

The following lists contain some notational conventions for equations and some custom commands. The conventions are guided by what is typically done in Stata Press books.

- omit punctuation in display-style equations
- use small caps for acronyms of estimators and other definitions, such as IF, ARE, ASV, BIC, MM estimator, S estimator, etc.
- use boldface lowercase letters for vectors ($\mathbf{x}$)
- use boldface uppercase letters for matrices ($\mathbf{X}$)
- use $E(X)$ for expectation
- use $\mathrm{Var}(X)$ and $\mathrm{Cov}(X, Y)$ for variance and covariance: `\mathrm{Var}` ...
- use $\mathcal{N}(a, b)$ for normal distribution: `\mathcal{N}` ...
- use $\Pr(x < y)$ for probability: `\Pr` ...
- use $'$ for both transposition $\mathbf{x}'$ (`\stvec{x}'`) and first derivative $F'(x)$ (`F'(x)`)
- use of parentheses
  - parentheses (): grouping/order of operations, functions, open intervals
  - square brackets []: matrices, closed intervals
  - nested [()] for grouping/order of operations: may use square brackets as second set of parentheses for visual distinction
  - braces {}: sets
- commands:

small caps acronyms	`\stsc{IQR}(X)`	$\textsc{iqr}(X)$
estimation hat	`\sthat{\theta}`	$\widehat{\theta}$
mean bar	`\stbar{x}`	$\overline{x}$
vectors	`\stvec{x}`	$\mathbf{x}$
matrices	`\stmat{X}`	$\mathbf{X}$
bold symbols	`\boldsymbol\theta`	$\boldsymbol{\theta}$
integral dif operator	`\int_a^b x \dif F(x)`	$\int_a^b x\, dF(x)$
indicator function	`\I(x<X)`	$\mathrm{I}(x < X)$
sign operator	`\sign(x)`	$\mathrm{sign}(x)$
median operator	`\med(x)`	$\mathrm{med}(x)$

Stata code, datasets, programs, and references to manuals

In this book we assume that you are somewhat familiar with Stata, that you know how to input data and to use previously created datasets, create new variables, run regressions, and the like. Generally, we use the `typewriter font` to refer to Stata commands, syntax, and variables. A "dot" prompt followed by a command indicates that you can type verbatim what is displayed after the dot (in context) to replicate the results in the book.

The data we use in this book are freely available for you to download, using a net-aware Stata, from the Stata Press website, http://www.stata-press.com. In fact, when we introduce new datasets, we merely load them into Stata the same way that you would. For example,

```
. use http://www.stata-press.com/data/!!!/football.dta, clear
```

In addition, the Stata packages presented in this book may be obtained by typing

```
. ssc install robstat
  (output omitted)
. ssc install robreg
  (output omitted)
. ssc install robmv
  (output omitted)
```

Also say what other packages need to be installed (if any), e.g. `moremata`, I think.

Throughout the book, we often refer to the Stata manuals using [R], [P], etc. For example, [R] **regress** refers to the *Stata Reference Manual* entry for `regress`, and [P] **matrix** refers to the entry for `matrix` in the *Stata Programming Manual*.

Mathematical and statistical symbols

We also assume that you have basic knowledge of mathematics and statistics, although we tried to keep the exposition as simple and non-technical as possible. Below is a list of some mathematical and statistical symbols that we will frequently use in the book.

$X, Y, Z, \ldots$	random variables
$x_i, y_i, z_i, \ldots$	realizations (observations) of random variables
n	number of observations
$x_{(i)}$	ith order statistic of $x_1, \ldots, x_n$ (ith observation in the list of observations sorted in ascending order)
$F(x)$	cumulative distribution function of a random variable; ...
$f(x)$	density ...

$F'(x)$	first derivative of function $F(x)$, that is $F'(x) = dF(x)/dx = f(x)$; we use $'$ for both the first derivative of a function and the transposition of a vector or matrix
$\mathcal{N}(\mu, \sigma)$	normal distribution with mean μ and standard deviation σ
$\mathcal{N}(0, 1)$	standard normal distribution
$\lvert x \rvert$	absolute value of x
$\lVert \mathbf{x} \rVert$	Euclidean norm of vector $\mathbf{x} = (x_1, \ldots, x_p)'$, that is, $\lVert \mathbf{x} \rVert = \sqrt{x_1^2 + \cdots + x_p^2}$
$\lceil x \rceil$	smallest integer greater or equal to x
$\lfloor x \rfloor$	largest integer smaller or equal to x
$\mathbf{x}'$, $\mathbf{X}'$	transposition of a vector or a matrix; we use $'$ for both the transposition of a vector or matrix and the first derivative of a function
i.i.d.	independent and identically distributed
$\lim_{x \to y} g(x)$	limiting value of function $g(x)$ as x approaches y
$\sup_x g(x)$	supremum (least upper bound) of function $g(x)$ with respect to argument x
$\text{sign}(x)$	the sign of x; to be precise, $\text{sign}(x) = -1$ if $x < 0$, $\text{sign}(x) = +1$ if $x > 0$, $\text{sign}(x) = 0$ if $x = 0$
$X \sim F$	random variable X is distributed as F
$X \approx F$	random variable X is approximately distributed as F
...	...

Part I

Introduction

1 Introduction

1.1 Motivation

Linear least-squares (LS) regression is, without doubt, the workhorse of data analysis in social sciences, economics and related fields. The reasons for the popularity of LS regression are obvious. The procedure convinces by its formal and practical simplicity. LS regression is easy to implement from a technical point of view and its results, the estimated regression coefficients, are easy to interpret. Furthermore, LS regression is easy to teach because its math is relatively simple and is didactically convenient because LS solutions for small datasets can easily be computed manually for purpose of exercise and understanding. From a statistical point of view, LS regression is favorable because it can be shown that under the assumption of homoscedastic (i.e., equal-variance) and normally distributed errors the LS estimator is the best (i.e., most efficient) unbiased estimator (BUE) for the coefficients of a linear regression model. That is, among all possible unbiased estimators, the LS estimator has the smallest sampling variance under these conditions.[1] Also under relaxed assumption, such as non-normal or heteroscedastic (i.e., non-equal-variance) errors, the LS estimator is consistent and has, in many cases, good efficiency properties.[2] For example, in case of homoscedastic non-normal errors, the LS estimator is the beast linear unbiased estimator (BLUE), that is, has the smallest sampling variance among all "linear" unbiased estimators.[3]

The outstanding usefulness of LS regression should not be challenged here. It is important, however, to realize that LS regression may not always be the best—or at least not the only—choice for analyzing a given dataset. The restrictiveness of the conditions under which the LS estimator is deemed best—homoscedasticity and normality of errors—implies that situations are possible in which alternative estimators can be valuable. For example, as mentioned above, if the errors are homoscedastic but non-normal, the LS estimator may be the best linear unbiased estimator, but this also means that there can be non-linear estimators that, depending on the nature of the deviation from

1. Noting the equivalence between the LS estimator and the arithmetic mean, the BUE property of the LS estimator is not much of a surprise given the fact that Carl Friedrich Gauß derived the normal distribution as a justification for the arithmetic mean. That is, the normal distribution is *defined* as the distribution under which the LS procedure leads to the best unbiased estimator for the expected value (for historical background see Huber 1972).
2. Although in the later case, ordinary LS estimates of the sampling variance are biased and need to be adjusted by applying heteroscedasticity-robust variance estimation; see White 1980)
3. The term "linear" does not refer to the fact that the coefficients of a linear regression model are to be estimated; it refers to a property of the estimation methodology. A linear estimators is ... explain the difference between linear and nonlinear estimators ...

normality, substantially outperform the LS estimator in terms of efficiency.[4] In particular, in case of distributions with heavy tails, that is, if extreme values are more frequent than in a normal distribution (an example being the t-distribution with few degrees of freedom), the efficiency of the LS estimator can quickly become poor. Furthermore, the LS estimator may yield misleading results if the data are "contaminated" by erroneous observations or, more generally, by a secondary data-generating process.

Efficiency under alternative error distributions

Assume, for now, that the data are not contaminated and, more or less, follow a uniform data-generating process that can be described by a linear regression model. Why, under such a condition, can a low efficiency of the LS estimator be a problem? Although, on average across multiple samples, the LS estimator is unbiased, more efficient estimators would be preferable because the precision of an estimator has a direct effect on the value of the results. For example, the power of a significance test and, therefore, the potential of the test to find an existing relation, decisively depends on the efficiency of the employed estimator.

In the context of error distributions with heavy tails the efficiency argument can also be motivated as follows. Although the LS estimator is unbiased on average, there is a good chance for a single sample—and in practice often only one sample is available—to contain extreme values that bias the regression results in one or the other direction. Robust regression methods that are less sensitive to such outliers will typically provide more valid results in such situations, being closer to the true value of the parameter to be estimated.

Figure 1.1 shows two examples of data sets that have been generated according to model

$$Y = \beta_1 + \beta_2 X + \epsilon$$

with $\beta_1 = \beta_2 = 0$ (that is, the "true" regression function is a horizontal line at $Y = 0$) and ϵ following a t distribution with two degrees of freedom, that is $\epsilon \sim t(2)$. Included as lines are the estimated regression fits using LS estimation, as well as two robust estimators (an M estimator and an MM estimator). As is evident, the LS solution is affected by the outliers and suggests a positive relation between X and Y in the two examples, whereas the two robust estimator are relatively stable. Robust methods, so to say, contain a safeguard against extreme data constellations that can occur at random due to sampling or a stochastic data-generating process. As a diagnostic by-product, robust methods inform about whether given data are characterized by an anomalous constellation or not, because only in the former case the results from LS and the robust methods will substantially differ.

4. The limitation to linear estimators is not much less restrictive than the limitation to normal errors.

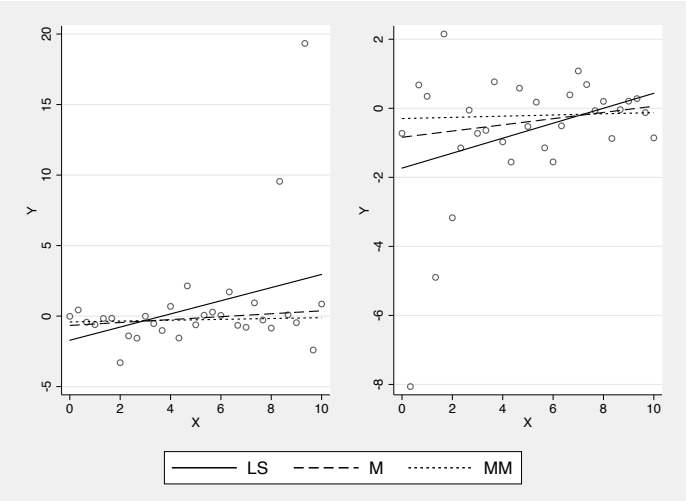

Figure 1.1. Example scatter plots with outliers and different regression fits

Bias due to data contamination

Now assume that the data are "contaminated", that is, that the majority of data points follows a well-defined model, but that there are also some observations that come from a different distribution. For example, while collecting the data, coding errors could have occurred for some of the observations. In a study by Jasso (1985) on the relation between marital duration and coital frequency there were four observations with a value of 88 for the monthly coital frequency. Although such values would not be impossible (as argued by Jasso 1985), the observations were highly suspicious as no other values of comparable magnitude existed in the data. As argued by Kahn and Udry (1986), the four observations probably were miscoded missing values, whose designated value was 99. The problem with such miscoded observations is that they can have strong effects on the results provided by a LS regression. That is, regression results and the substantive conclusions drawn from them may differ depending on whether the miscoded observations are kept in the data or not. It seems important to use methods for data analysis that are able to identify such problems because, in the words of Anscombe (1973, 18), "[w]e are usually happier about asserting a regression relation if the relation is still apparent after a few observations (any ones) have been deleted—that is, we are happier if the regression relation seems to permeate all the observations and does not derive largely from one or two."

Conceptually, contamination can be understood as a situation in which the observed data are the result of a mixture of two or more data-generating processes. In the case of coding errors there may be a main process of substantive interest (e.g., the relation between marital duration and coital frequency), as well as a secondary process (data miscoding by interviewers) that leads to observations that follow a different distribution

and have a different interpretation. LS regression will not be able to distinguish the two processes and its results will be valid for neither one of the processes. If, however, the data are dominated by one of the processes (that is, if one of the processes is responsible for the bulk of the data) and the two processes do lead to distinguishable data structures, statistical procedures to identify the main process are possible. This is where robust regression comes in. One of the goals of robust regression techniques is to provide estimates that are resistant against partial contamination of the data. Robust methods are supposed to correctly identify the primary relation in the data even if, for example, parts of the data are glaringly erroneous.

An illustrative example comes from astronomy. Figure 1.2 shows the Hertzsprung-Russell diagram of star cluster CYG OB1 (see Rousseeuw and Leroy 1987, 27). Displayed is the logarithm of the light intensity of the stars against the logarithm of their effective surface temperature (using a reversed axis). Furthermore, the graph shows as lines the results of three different regression estimators, the LS estimator (solid line), a low breakdown point M estimator (dashed line), and a high breakdown point MM estimator (dotted line). The results from the LS estimator and the low breakdown point M estimator are almost identical. They are strongly influenced by the group of four stars in the upper right corner of the diagram. In contrast, the high breakdown point MM estimator completely ignores the four outliers and adequately captures the trend in the main part of the data. Hence, at least one of the two employed robust estimators successfully identified the main process (due to the estimator's high breakdown point; see below).

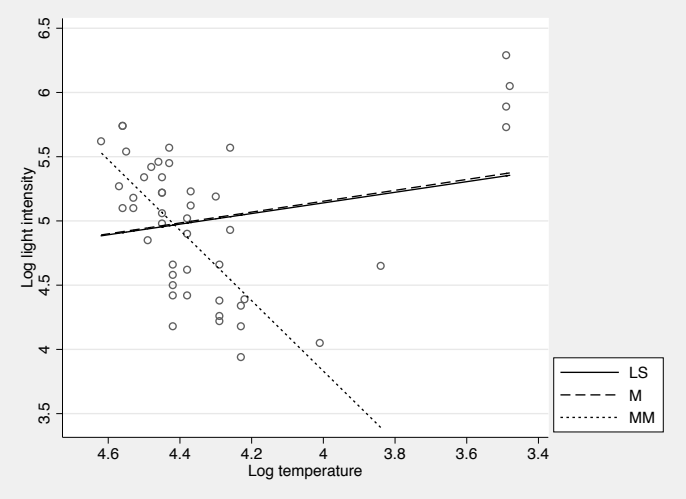

Figure 1.2. Hertzsprung-Russell diagram of the star cluster CYG OB1 including different regression fits (source: Rousseeuw and Leroy 1987, 27)

Again, from a diagnostic perspective, the interesting cases are the ones in which LS regression and robust estimators lead to differing results. Substantial differences

between robust regression and the LS estimator indicate that the data cannot be fully descried by a uniform model and that a part of the observations stands in stark contrast to the main trend in the data. With the help of the residuals from robust regression, the atypical observations can be identified and, for example, be subjected to a separate analysis. In this way, robust regression can contribute to a better understanding of the data and, potentially, give way to new insights and new hypotheses. In fact, according to Kruskal (1960, 1), the atypical observations may prove to be the most interesting part of the data: "An apparently wild (or otherwise anomalous) observation is a signal that says: 'Here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study.'" The four outliers in figure 1.2, by the way, are not errors. The explanation is that there are two different types of stars: main-sequence stars and giants. That is, conceptually, the observation stem from two different populations.

Goals and use of robust regression

To summarize, we can state that robust regression estimators (1) should achieve good efficiency also in case of non-normal errors and (2) should be resistant against contamination of the data by outliers. The maximum proportion of contamination a robust estimator is able to absorb is called the *breakdown point*.

Both aspects can be formalized with the help of the viewpoint coined by Huber (1964) that observed data follow a mixture distribution

$$F_\varepsilon = (1-\varepsilon)F_{\boldsymbol{\theta}} + \varepsilon G$$

where $F_{\boldsymbol{\theta}}$ is the distribution of interest according to the supposed model, G is an arbitrary alternative distribution, and $\varepsilon \in [0,1]$ determines the mixing proportion. For example, in line with the assumptions of classic linear regression, $F_{\boldsymbol{\theta}}$ could be a distribution according to the linear model

$$Y = \beta_1 + \beta_2 X + \cdots + \epsilon$$

where X has a given distribution and ϵ is an independent and identically normally distributed error term. The distribution of the observed data, however, is contaminated by observations from an unspecified alternative distribution G and does not fully follow this model. The goal of robust regression now is to deliver reasonable results for $F_{\boldsymbol{\theta}}$ even if the model is somewhat misspecified, that is, if $\varepsilon > 0$. In the words of Heritier et al. (2009, 7), robust methods are "a set of statistical tools for correct estimation and inference about F_θ when the data-generating process is F_ε, not only when $\varepsilon = 0$, as with classical methods, but also for relatively small ε and *any* G. As a by-product, data not fitting F_θ exactly can be easily identified, and the model can possibly be changed and refitted." In addition, to be of diagnostic value, robust estimators should be serious competitors of classic methods in case of $\varepsilon = 0$. In particular, robust estimators should achieve good "gaussian efficiency", that is, they should achieve a hight relative efficiency compared to LS estimation in the ideal case of normally distributed errors.[5]

5. Note that the estimation of "robust standard errors" is not the primary concern of robust regression.

Yet, robust regression should be seen as a complement and not so much as a substitute to LS regression. In our view, the main use of robust regression lies in its diagnostic potential. Classic regression techniques may lead to meaningful results in many situations, but a comparison to robust results is always advisable. Before drawing far-reaching conclusions based on classic methods one should evaluate whether the conclusions are "robust", that is, whether methods that rely on less restrictive assumptions and are less affected by outliers and atypical data constellations come to the same conclusions.

If classic procedures and robust regression lead to substantially diverging or even contradicting results, the robust results can provide an immediate contribution to a better understanding of the data. As a by-product of robust estimation, observations that do not fit the supposed model can easily be identified, offering clues about possible misspecification, the nature of outliers, and alternative data-generating processes. Compared to classic regression diagnostics for the identification of influential observations (see Belsley et al. 1980; Cook and Weisberg 1982; Chatterjee and Hadi 1988; Fox 1991) robust regression methods have the advantage that they can also identify "masked" multiple outliers that would go undetected by classic diagnostics. However, robust techniques are no panacea and cannot, for example, fully replace diagnostic methods that are concerned with the identification of structural misspecification (such as omitted variable bias, wrong functional form, or missing interaction terms).

[Should there also be some text giving a brief historical account of the development of robust statistics and robust regression?]

1.2 What is covered in this book?

...

1.3 Robust statistics in Stata

- Summary of existing tools
- Brief presentation of our new packages; basic usage and syntax

---

The term "robust standard errors" refers to estimators for the sampling variances of the coefficient estimates that are consistent also if the assumption of identically distributed errors is violated (i.e., if the errors are heteroscedastic; see White 1980). To prevent false conclusions with respect to confidence intervals and significance tests, it is always a good idea to consider "robust standard errors", be it with classic regression or with robust regression.

Part II

Robustness theory and basic robust statistics

2 Basic concepts in estimation

An estimation problem in statistics may have many potential solutions. To separate useful estimation strategies from approaches that are less feasible, criteria have to be defined by which different estimators can be evaluated and compared. In this chapter we first review a number of basic criteria typically used in classic statistics. We then discuss additional criteria that are important in the context of robust statistics. Our discussion in this chapter is conceptual in nature; it is supposed to establish a theoretical basis for the specific robust estimators that are discussed in the subsequent chapters from an applied perspective.

2.1 Classical properties of estimators

The goal of statistical estimation is to obtain a reasonable value for the unknown *parameter* of a statistical model, based data whose properties are assumed to be consistent with the suggested model. Let $\mathcal{X}^{(n)} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ that have a joint probability distribution $P_{\boldsymbol{\theta}}^{(n)}$ depending on the unknown parameter $\boldsymbol{\theta}$. Note that, depending on context, $\boldsymbol{\theta}$ may be scalar, or it may be a vector of multiple parameters, that is $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$. Likewise, $\mathbf{x}_i$ may be univariate, or it may contain multiple dimensions, that is $\mathbf{x}_i = (X_{i1}, \ldots, X_{ik})'$. [I think it is important to emphasize that X can be a collection of variables as the book is on regression and not on univariate statistics. Hence, I changed notation to vectors. But I'm unsure whether this is ok...]

1: Do we need the parentheses around n? (In general, to improve readability, we should use as little ink as possible in equations.) Later in the chapter, n is a subscript (without parentheses) not a superscript; what would be the best way to deal with n?

We assume that the joint distribution of the random variables $\mathbf{x}_i$, $i = 1, \ldots, n$, belongs to the (parametric) statistical model $\mathcal{P}^{(n)} = \{P_{\boldsymbol{\theta}}^{(n)} | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta}$ is the set of possible values of $\boldsymbol{\theta}$. The goal is to estimate $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ based on a realization of $\mathcal{X}^{(n)}$. In general, we will consider the case in which the statistical model $\mathcal{P}^{(n)}$ conforms to simple random sampling (SRS), that is, a situation in which the random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and identically distributed (i.i.d.). In this situation, each $\mathbf{x}_i$ independently follows a common distribution $P_{\boldsymbol{\theta}}$, which can be characterized by the distribution function $F_{\boldsymbol{\theta}}(\mathbf{x}) = \Pr(X_1 \leq x_1, \ldots, X_k \leq x_k)$ (or simply F when there is no risk of confusion about the parameter we have to estimate).

2: I do not really understand. Why do we need $\mathcal{P}$? What is the difference to $P_{\boldsymbol{\theta}}$?

3: Maybe be a bit more specific about $\boldsymbol{\Theta}$. I made Theta bold because if $\boldsymbol{\theta}$ is multidimensional then also $\boldsymbol{\Theta}$ has to be. Or do I misunderstand what Theta is?

An *estimator* of $\boldsymbol{\theta}$ can then be defined as follows: An estimator of the parameter $\boldsymbol{\theta}$ is any statistic $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ taking its value in $\boldsymbol{\Theta}$.

The value of $\widehat{\boldsymbol{\theta}}$ provided by a particular realization of the random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is called an *estimate* of $\boldsymbol{\theta}$. Note that, for simplicity, we will often use the notation $\mathbf{x}_i$,

$i = 1, \ldots, n$, to designate the ith random variable as well as a realization of it (i.e., a specific observation); the context will always clearly indicate if we have to consider $\mathbf{x}_i$ as a random variable or as a particular value.

The definition above contains no indication of the quality of an estimator; any statistic $\widehat{\boldsymbol{\theta}}$ that provides a value in $\boldsymbol{\Theta}$ is a valid estimator. To narrow down the set of estimators to estimators that can be considered useful we need quality criteria. Classic quality criteria are unbiasedness, efficiency, and consistency.

2.1.1 Unbiasedness

From a good estimator one may expect that, on average, it gives the "correct" answer. Let us denote by $E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)}))$ the expectation of statistic $\widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ when $\mathcal{X}^{(n)} \sim P_{\boldsymbol{\theta}}^{(n)}$. Think of $E_{\boldsymbol{\theta}}$ as the average value we would obtain for $\widehat{\boldsymbol{\theta}}$ from a large number of repeated realizations of $\mathcal{X}^{(n)}$, given that for each repetition $\mathcal{X}^{(n)}$ follows distribution $P_{\boldsymbol{\theta}}^{(n)}$ (as, for example, in repeated random sampling from the same population).

Unbiasedness can then be defined as follows: The estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is called unbiased if

$$E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and all } n$$

That is, no matter the sample size n, estimator $\widehat{\boldsymbol{\theta}}$ will, on average across a large number of repeated samples, provide the correct answer (given that our assumptions about the joint distribution of $\mathcal{X}^{(n)}$ are correct, such as, e.g., independent sampling of observations). The difference

$$B_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\theta}$$

is called the *bias* of estimator $\widehat{\boldsymbol{\theta}}$. The absence of bias indicates that the sampling distribution of $\widehat{\boldsymbol{\theta}}$ has a mean that coincides with the value of the parameter of interest.

Zero bias is often difficult to achieve in small samples. Therefore, another useful criterion is *asymptotic unbiasedness*: The estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is called asymptotically unbiased if

$$\lim_{n\to\infty} E_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \qquad \text{for all } \boldsymbol{\theta} \in \boldsymbol{\Theta}$$

That is, an estimator is asymptotically unbiased if the bias vanishes with increasing sample size. An important question in this context is, of course, how fast the bias vanishes (or how large the sample size has to be for the bias to be negligible).

2.1.2 Efficiency

For a specific estimation problem, several (asymptotically) unbiased estimators may exist. To choose the best among them we need further information about the performance of the different estimators. Furthermore, there may also be situations in which a biased estimator is to be preferred over an unbiased estimator. A key aspect in this regard

is the *efficiency* of an estimator. Efficiency has to do with how spread out about $\boldsymbol{\theta}$ the sampling distribution of the estimator is. The smaller the dispersion of estimator $\widehat{\boldsymbol{\theta}}$ around the true value $\boldsymbol{\theta}$ in repeated samples, the more "efficient" (or precise) is the estimator.

Mean squared error

First consider the case of a *scalar* parameter θ. The precision of estimator $\widehat{\theta}$ can be measured by its *mean squared error* (MSE):

$$\text{MSE}_\theta(\widehat{\theta}) = E_\theta\Big((\widehat{\theta} - \theta)^2\Big)$$

A small mean squared error for $\widehat{\theta}$ means that the sampling distribution of $\widehat{\theta}$ is well concentrated around the exact value of the parameter to estimate and hence that the estimator $\widehat{\theta}$ has a good precision.

It is easy to show that

$$\text{MSE}_\theta(\widehat{\theta}) = \text{Var}_\theta(\widehat{\theta}) + \Big(B_\theta(\widehat{\theta})\Big)^2$$

That is, the mean squared error of an estimator can be decomposed into its variance and its squared bias. Hence, if $\widehat{\theta}$ is unbiased, $\text{MSE}_\theta(\widehat{\theta})$ is simply equal to $\text{Var}_\theta(\widehat{\theta})$.

Relative efficiency

An estimator $\widehat{\theta}_A$ of θ is more precise—we will say *more efficient*—than another estimator $\widehat{\theta}_B$ if

$$\text{MSE}_\theta(\widehat{\theta}_A) \leq \text{MSE}_\theta(\widehat{\theta}_B) \qquad \text{for all } \theta \in \Theta$$

and

$$\text{MSE}_\theta(\widehat{\theta}_A) < \text{MSE}_\theta(\widehat{\theta}_B) \qquad \text{for at least one } \theta \in \Theta$$

4: Shouldn't it be "<" in at least one case?

In general, we consider the "large-sample" sampling distributions of asymptotically unbiased estimators. If, for large n, the estimators $\widehat{\theta}_A$ and $\widehat{\theta}_B$ are approximately $\mathcal{N}(\theta, \text{Var}(\widehat{\theta}_A))$ and $\mathcal{N}(\theta, \text{Var}(\widehat{\theta}_B))$, respectively, we define the *asymptotic relative efficiency* (ARE) of $\widehat{\theta}_B$ with respect to $\widehat{\theta}_A$ as the ratio

$$\text{ARE}_\theta(\widehat{\theta}_B, \widehat{\theta}_A) = \frac{\text{Var}(\widehat{\theta}_A)}{\text{Var}(\widehat{\theta}_B)}$$

(see Serfling 1980). If $\widehat{\theta}_B$ is (asymptotically) less efficient than $\widehat{\theta}_A$, then

$$\text{ARE}_\theta(\widehat{\theta}_B, \widehat{\theta}_A) < 1$$

Efficiency of the maximum likelihood estimator

Let us consider the case in which the random variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are univariate and i.i.d. with a common distribution function F_θ and a common density function f_θ that satisfies some differentiability conditions with respect to θ. Suppose also that the *Fisher information*

$$\mathcal{I}(F_\theta) = E_\theta\left(\left(\frac{\partial}{\partial\theta}\log f_\theta(\mathbf{x})\right)^2\right)$$

is strictly positive and finite. Then it follows that

(i) for large n, the maximum likelihood estimator $\widehat{\theta}_{\text{ML}}$ of θ is approximately distributed as $\mathcal{N}\big(\theta, (n\mathcal{I}(F_\theta))^{-1}\big)$

(ii) for a wide class of estimators $\widehat{\theta}$ that are approximately distributed as $\mathcal{N}(\theta, V)$, a *lower bound* to V is $(n\mathcal{I}(F_\theta))^{-1}$

(see Lehmann and Casella 1988). In this situation,

$$\text{ARE}_\theta(\widehat{\theta}, \widehat{\theta}_{\text{ML}}) = \frac{(n\mathcal{I}(F_\theta))^{-1}}{V} \leq 1 \tag{2.1}$$

making $\widehat{\theta}_{\text{ML}}$ the most (asymptotically) efficient among the given class of estimators $\widehat{\theta}$. Note, however, as will be discussed later, that (2.1) does not necessarily make $\widehat{\theta}_{\text{ML}}$ the estimator of choice, when certain other considerations are taken into account.

Notation in the multidimensional case

If $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is a vector of parameters we define the *mean squared error matrix* of the estimator $\widehat{\boldsymbol{\theta}}$ as follows:

$$\text{MSE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\Big((\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\Big)$$

If $\widehat{\boldsymbol{\theta}}$ is unbiased, the mean squared error matrix simply coincides with the covariance matrix of the estimator.

If, for large n, the p-variate estimators $\widehat{\boldsymbol{\theta}}_A$ and $\widehat{\boldsymbol{\theta}}_B$ are approximately normally distributed with mean $\boldsymbol{\theta}$ and nonsingular covariance matrices $\boldsymbol{\Sigma}_A$ and $\boldsymbol{\Sigma}_B$, respectively, it is usual to define the *asymptotic relative efficiency* (ARE) of $\widehat{\boldsymbol{\theta}}_B$ with respect to $\widehat{\boldsymbol{\theta}}_A$ as the ratio of the *generalized variances* (determinants of the covariance matrices), raised to the power $1/p$, that is

$$\text{ARE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_B, \widehat{\boldsymbol{\theta}}_A) = \left(\frac{\det(\boldsymbol{\Sigma}_A)}{\det(\boldsymbol{\Sigma}_B)}\right)^{1/p}$$

If $\text{ARE}_{\boldsymbol{\theta}}(\widehat{\boldsymbol{\theta}}_B, \widehat{\boldsymbol{\theta}}_A) < 1$, estimator $\widehat{\boldsymbol{\theta}}_B$ is (asymptotically) less efficient than estimator $\widehat{\boldsymbol{\theta}}_A$.

Here again the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_{\text{ML}}$ of $\boldsymbol{\theta}$ appears as the most (asymptotically) efficient estimator among a wide class of (asymptotically) unbiased estimators $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$. Moreover, for large n,

$$\widehat{\boldsymbol{\theta}}_{\text{ML}} \approx \mathcal{N}\big(\boldsymbol{\theta}, (n\mathbf{I}(F_{\boldsymbol{\theta}}))^{-1}\big)$$

where "$\approx$" stands for "is approximately distributed as" and $\mathbf{I}(F_{\boldsymbol{\theta}})$ is the $p \times p$ Fisher information matrix with its elements defined as

$$\mathcal{I}_{ij}(F_{\boldsymbol{\theta}}) = E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\theta_i}\log f_{\boldsymbol{\theta}}(\mathbf{x}) \cdot \frac{\partial}{\partial\theta_j}\log f_{\boldsymbol{\theta}}(\mathbf{x})\right)$$

2.1.3 Consistency

The consistency criterion comes in two flavors, as consistency in terms of convergence in probability and as consistency in terms of convergence in distribution. [Please explain here (or below) why (asymptotic) unbiasedness and efficiency is not enough an why the additional criterion of consistency is needed.]

Convergence in probability

An estimator $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(\mathcal{X}^{(n)})$ is *consistent* if, for $n \to \infty$, $\widehat{\boldsymbol{\theta}}$ *converges in probability* to $\boldsymbol{\theta}$. That is, for any $\epsilon > 0$ and for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$:

$$\lim_{n\to\infty} P_{\boldsymbol{\theta}}^{(n)}\Big(\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \le \epsilon\Big) = 1$$

This type of consistency means that, when the sample size grows, the probability that the estimator $\widehat{\boldsymbol{\theta}}$ takes a value near the exact value of the parameter and, consequently, provides a "good" estimate of the parameter $\boldsymbol{\theta}$, grows to 1. Consistency of $\widehat{\boldsymbol{\theta}}$ is required if we want an estimator to provide unambiguous statistical inference for $\boldsymbol{\theta}$. Note that consistency in terms convergence in probability is given if an estimator is asymptotically unbiased and if the variance of the estimator approaches zero as the sample size grows.

5: I don't really understand this sentence ("Consistency of . . . ")

For simplicity, consider the case of a scalar parameter θ. If $B_\theta(\widehat{\theta}) \overset{n\to\infty}{\longrightarrow} 0$ and $\text{Var}_\theta(\widehat{\theta}) \overset{n\to\infty}{\longrightarrow} 0$, then $\text{MSE}_\theta(\widehat{\theta}) \overset{n\to\infty}{\longrightarrow} 0$, so that $\widehat{\theta} = \widehat{\theta}(\mathcal{X}^{(n)})$ *converges in quadratic mean* to θ. That is, for $n \to \infty$, $\widehat{\theta} \overset{L_2}{\to} \theta$. It is well known that convergence in quadratic mean implies convergence in probability.

6: I don't understand. What is "convergence in quadratic mean"? Where does L_2 come from?

Convergence in distribution

The statistic $T^{(n)} = T(\mathcal{X}^{(n)})$ *converges in distribution* to the probability distribution $\mathcal{L}$ if, for $n \to \infty$, the distribution function $F_{T^{(n)}}$ of $T^{(n)}$ converges to the distribution function $F_{\mathcal{L}}$ of $\mathcal{L}$ in any point of continuity of $F_{\mathcal{L}}$. That is, convergence in probability is given if

$$F_{T^{(n)}}(x) \overset{n\to\infty}{\longrightarrow} F_{\mathcal{L}}(x) \quad \text{for any continuity point } x \text{ of distribution function } F_{\mathcal{L}}$$

As a shorthand, we will write

$$T^{(n)} \xrightarrow{d} \mathcal{L}$$

to denote convergence in probability for $n \to \infty$. In practice, convergence in probability means that, for large n, we may consider that $T^{(n)}$ is approximately distributed as $\mathcal{L}$, that is,

$$T^{(n)} \approx \mathcal{L}$$

[Please explain here why convergence in probability is a useful concept. What is it used for? Why do we need it in the context of robust statistics?]

[What about Fisher-consistency?]

2.1.4 Other aspects

Depending on context, a number of other criteria can be important. For example, from a practical perspective, *computational complexity* can be a relevant criterion to choose between different estimators. In general, estimators that require operations in the order of n^2 (that is, if the number of required computational operations grows quadratically with the sample size) lead to prohibitive computational costs in large samples. In many cases it is possible to design alternative estimators (or improved computational algorithms for a given estimator) that only require operations in the order of $\ln n$ and are thus much more efficient (with respect to computer time) in large samples. [Maybe expand a bit on this.]

Furthermore, consistent estimators may differ in their *rate of convergence*, that is, in how fast the mean squared error diminishes with growing sample size. [Please expand on this. How is the rate of convergence defined (and interpreted)? What are typical rates of convergence, e.g. of the mean etc.?] Naturally, an estimator with a faster rate of convergence, is usually to be preferred over an estimator with a slower rate of convergence. [Does this make sense?]

To enable approximately valid statistical inference in terms of confidence intervals and significance tests, a further important criterion is *asymptotic normality*. Asymptotic normality is given if it can be shown that the sampling distribution of an estimator approximates the normal distribution in large samples. For example, maximum-likelihood estimators can be shown to be asymptotically normal under general conditions. If, however, the asymptotic sampling distribution of an estimator is known to be non-normal, the viability of the estimator for statistical inference is limited.[Does this make sense? Please expand ...]

Finally, an estimator should be *equivariant* to transformations of data. That is, a transformation of the data should affect the estimator $\widehat{\boldsymbol{\theta}}$ in the same functional way as it affects the true parameter $\boldsymbol{\theta}$. For example, let θ_A be the expected value of variable X_A and θ_B be the expected value of variable X_B. If X_B can be expressed as a linear combination of X_A, that is, $X_B = a + b \cdot X_A$, then $\theta_B = a + b \cdot \theta_A$. In this case, also $\widehat{\theta_B} = a + b \cdot \widehat{\theta_A}$ should hold. In other words, whether you express your data in Dollars or

in Euros, whether you express your data in degrees Fahrenheit or degrees Celsius should only affect the scaling of your estimator, but should not affect your results otherwise.

[I just wrote some stuff in this section about points that seem relevant to me. I am not sure whether I always chose the right words. Also, maybe it would be good express some of it formally.]

2.2 Measures of robustness

Intuitively, the classical approach to statistics is about defining estimators that have desirable properties under clearly specified conditions. The goal of robust methods, however, is to develop estimators that perform well also in the "neighborhood" of such conditions. This leads to the proposition of so called "robust" estimators, that are, for instance, not affected too strongly or too quickly by the presence of outliers. Although outliers are only one of the main concerns of robust methods, we will make our first steps into robustness theory by presenting some basic concepts for measuring the degree to which estimators are affected by atypical observations.

▷ Example

Consider the following observations of the grades achieved by $n = 25$ students in fifth year of primary school (on a scale of 1 to 10):

7: Is 1-10 correct?

6.00	6.50	7.00	7.00	7.00
7.00	7.00	7.50	7.50	8.00
8.00	8.00	8.50	8.50	8.50
8.50	9.00	9.00	9.50	9.50
9.50	9.50	9.50	9.50	10.00

If we calculate two *measures of location*, the mean and the median, as well as two *measures of scale*, the standard deviation and the interquartile range, the results are as follows:

```
. drop _all
. matrix x = (6.00, 6.50, 7.00, 7.00, 7.00,    ///
>             7.00, 7.00, 7.50, 7.50, 8.00,    ///
>             8.00, 8.00, 8.50, 8.50, 8.50,    ///
>             8.50, 9.00, 9.00, 9.50, 9.50,    ///
>             9.50, 9.50, 9.50, 9.50, 10.00)'
. quietly svmat x
. tabstat x, statistics(mean median sd iqr)
    variable |      mean       p50        sd       iqr
-------------+--------------------------------------------
          x1 |      8.22       8.5  1.137248       2.5
-----------------------------------------------------------
```

Now, imagine the dot separating the decimals in the last observation is mistakenly removed, so that the last observation is coded as 1000. In this case the results are the following:

```
. replace x1 = 1000 in l
(1 real change made)
. tabstat x1, statistics(mean median sd iqr)
    variable |      mean       p50        sd       iqr
-------------+--------------------------------------------
          x1 |     47.82       8.5  198.3737       2.5
```

As is evident, the mean and the standard deviation strongly increased due to the introduction of the erroneous observation, whereas the median and the interquartile range remained unchanged. The example illustrates the fact that one single outlier may "break" the mean and the standard deviation, but does not affect the median or the interquartile range. Hence, these two later statistics can be considered as being more robust to erroneous data than the two first ones.

◁

How can the degree of robustness of different statistics be quantified? How can we compare the robustness of different estimators from various viewpoints? These are questions we will address in the rest of this section.

In robust estimation theory it is common to consider *parameters* as *functionals*. More precisely, the functional by which a parameter T is defined is a rule that maps every probability distribution F into a real number, that is $T = T(F)$.[1] Often, a natural *estimate* T_n of the parameter $T(F)$ based on sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$—where $x_1, \ldots, x_n$ are realizations of n independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ of distribution F—may be defined as the value of the functional at the empirical distribution F_n.[2] That is, $T_n = T(F_n)$. For example, if

$$T(F) = \int_{-\infty}^{\infty} x \, dF(x) = \mu$$

$$T(F) = \int_{-\infty}^{\infty} x \, dF(x) = \mu$$

is the expected value of the distribution F, then

$$T_n = T(F_n) = \int_{-\infty}^{\infty} x \, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} x_i = \mu_n$$

1. F is the cumulative distribution function of a random variable X. Evaluated at position x, the function returns the probability that the random variable will take on a value lower than or equal to x, that is, $F(x) = \Pr(X \leq x)$. In order to avoid unnecessary technical difficulties we will generally assume in this chapter that the distribution F is continuous with density f. The density is the first derivative of F; it is nonnegative and integrates to one.
2. The empirical distribution F_n is a discrete distribution with a probability mass of $1/n$ at each value of the sample $\mathbf{X}_n$.

is the arithmetic mean of a sample $\mathbf{X}_n$ from F. Likewise, if $T(F) = F^{-1}(0.5) = Q_{0.5}$ is the median of the distribution F, then $T_n = T(F_n) = F_n^{-1}(0.5) = Q_{0.5;n}$ is the empirical median of a sample $\mathbf{X}_n$ from F.

The robustness of a statistic (or estimator) T_n may be analyzed in a very intuitive way by studying how a contamination of the sample $\mathbf{X}_n$ affects T_n. This empirical approach leads to the notions of the *sensitivity curve* and the *finite-sample breakdown point* of T_n. But it is also of great interest to consider the limiting case where n tends to infinity. As the sample size n grows, the empirical distribution function F_n approaches the underlying population distribution function F, and the empirical measures of robustness of the statistic T_n move in a natural way to the concepts of the *influence function* and the *asymptotic breakdown point* of the functional T.

2.2.1 The sensitivity curve and the influence function

The sensitivity curve

The *sensitivity curve* (SC) is an empirical tool to quantify the robustness of a statistic in a given sample. Consider a data set $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ and the statistic $T_n = T_n(x_1, \ldots, x_n) = T(F_n)$. To study the impact of a potential outlier on this statistic, we may analyze the change in the value of the statistic once we add an extra data point x, where x is varied between $-\infty$ an $+\infty$. Hence, the *(standardized) sensitivity curve* of statistic T_n for the sample $\mathbf{X}_n$ is defined as

$$\text{SC}(x; T_n, \mathbf{X}_n) = \frac{T_{n+1}(x_1, \ldots, x_n, x) - T_n(x_1, \ldots, x_n)}{\frac{1}{n+1}}$$

That is, for each value of x we compare the statistic in the "contaminated" sample to its value in the original sample, and rescale the difference by dividing the difference by $1/(n+1)$, the proportion of contamination.

▷ Example

Consider a data set $\mathbf{X}_n$ of $n = 20$ (rounded) random numbers from a $\mathcal{N}(0, 1)$ (standard normal) distribution:

−0.49	0.14	1.54	0.63	−0.87	−0.86	1.65	−0.55	0.91	−0.03
−0.61	0.22	−1.61	0.15	0.36	1.96	1.04	0.24	−0.45	0.98

Figure 2.1 shows the standardized sensitivity curves for the mean and the median; Figure 2.2 displays the standardized sensitivity curves for the standard deviation and the interquartile range. As is evident, the mean and the standard deviation have unbounded sensitivity curves (the curves go off to minus or plus infinity as the outlier moves away from the center of the uncontaminated data), whereas the sensitivity curves of the median and the interquartile range are bounded. The "classic" location and scale measures

may be completely perturbed by the presence of one single outlying observation—this illustrates the non-robust character of these two statistics—, while the impact of the additional outlying observation on the quantile-based location and scale measures remains very limited.

◁

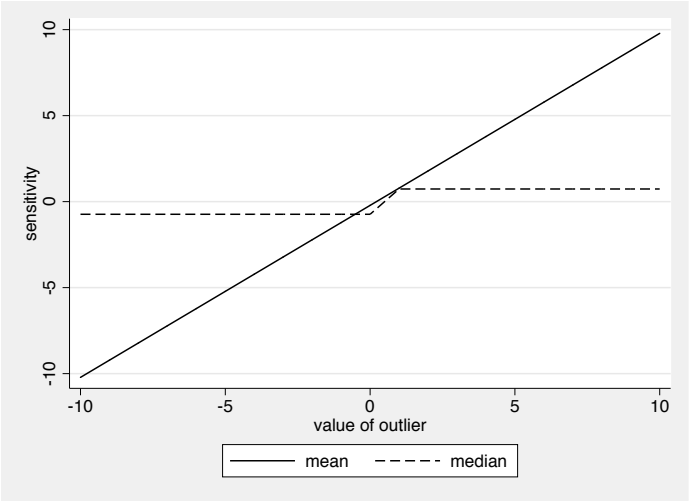

Figure 2.1. Standardized sensitivity curves of the mean and the median for a sample of $n = 20$ random $\mathcal{N}(0, 1)$ numbers

The influence function

An intuitive way to introduce the *influence function* (IF) of functional T at some distribution F is to think of the influence function as an asymptotic version of the sensitivity curve of statistic $T_n = T(F_n)$ when the sample size n grows, so that the empirical distribution function F_n tends to the underlying population distribution function F (cf. Hampel 1974). More precisely, the influence function is defined as

$$\begin{aligned} \text{IF}(x; T, F) &= \lim_{n\to\infty} \frac{T\Big(\Big(1 - \frac{1}{n+1}\Big)F + \frac{1}{n+1}\Delta_x\Big) - T(F)}{\frac{1}{n+1}} \\ &= \lim_{\varepsilon\to 0} \frac{T((1-\varepsilon)F + \varepsilon\Delta_x) - T(F)}{\varepsilon} \end{aligned}$$

where Δ_x is a probability distribution with all its mass at point x. That is, the influence function measures the effect on T of a perturbation of F obtained by adding a small probability mass at point x. $\text{IF}(x; T, F)$ can be found for most functionals T. In chapter 3 we will provide the influence functions of various measures of location, scale, skewness and tails heaviness.

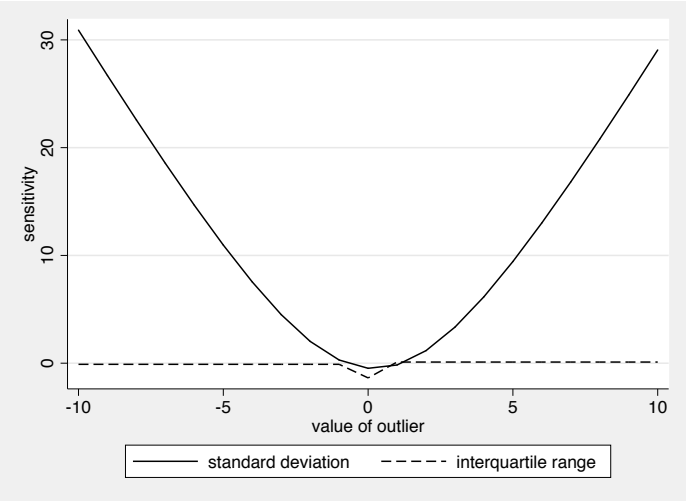

Figure 2.2. Standardized sensitivity curves of the standard deviation and the interquartile range for a sample of $n = 20$ random $\mathcal{N}(0, 1)$ numbers

The gross-error sensitivity

Since $\text{IF}(x; T, F)$ quantifies the influence on T of an infinitesimal contamination of the distribution F at point x, it is a *local* measure of robustness. It may be completed by a more global measure, the *gross-error sensitivity* of T at distribution F, defined as

$$\gamma^*(T, F) = \sup_x |\text{IF}(x; T, F)|$$

$\gamma^*(T, F)$ evaluates to the biggest influence an outlier can have on the functional T. With respect to robustness it is desirable to use an estimator that is associated with a functional T for which $\gamma^*(T, F)$ is finite (that is, for which the influence function is bounded).

The local-shift sensitivity

The local-shift sensitivity is another tool related to the influence function; it aims at measuring the effect of "wiggling" an observation, that is, of a small perturbation as opposed to gross error. This is useful to asses the effects of rounding, grouping, or other local inaccuracies.

Jumps in the influence function (IF) indicate that a small fluctuation of the value of x can cause an abrupt change in the estimate. Hence, from the perspective of robustness, we prefer a continuous IF with an appropriately bounded derivative (wherever the derivative exists). To appreciate this kind of characteristic of the *IF*, we may determine

the *local-shift sensitivity*:

$$\lambda^*(T,F) = \sup_{x \neq y} \frac{|\text{IF}(y;T,F) - \text{IF}(x;T,F)|}{|y-x|}$$

The asymptotic variance of an estimator

The influence function may also be used as a heuristic tool to determine the asymptotic variance of the estimators. Indeed, under some regularity conditions for the functional T, we have, under F,

$$\sqrt{n}(T(F_n) - T(F)) \longrightarrow^d \mathcal{N}(0, \text{ASV}(T,F))$$

where

$$\text{ASV}(T,F) = \int_{-\infty}^{\infty} \text{IF}(x;T,F)^2 \, dF(x) \tag{2.2}$$

(cf. Hampel et al. 1986, p. 85 and 226). Consequently, under F, the interval

$$\left[T(F_n) - z_{(1-\alpha/2)} \frac{\sqrt{\text{ASV}(T,F)}}{\sqrt{n}},\ T(F_n) + z_{(1-\alpha/2)} \frac{\sqrt{\text{ASV}(T,F)}}{\sqrt{n}} \right]$$

provides an asymptotic confidence interval for the parameter $T(F)$, at a confidence level of $(1-\alpha)\%$. If the distribution F, and hence the asymptotic variance $\text{ASV}(T,F)$, are not known, it is still possible to obtain a confidence interval for $T(F)$ by an appropriate resampling method.

8: Why is "under F" needed? Isn't this obvious? And: What does "$\longrightarrow^d$" mean? "approximately distributed as"? or "asymptotically distributed as"? or simply "distributed as"? Why not "$\sim$" or "$\overset{a}{\sim}$"? Furthermore: Why not include "$/n$" in the definition of ASV?

9: Hm, isn't it possible to simply use IF evaluated at the sample to get an estimate of ASV? (i.e. without resampling) That is, evaluate IF at each observed x using F_n instead of F and then compute the variance from these values. At least this is an approach I saw in other research.

2.2.2 The breakdown point

The sensitivity curve shows how an estimator reacts to the introduction of one single outlier. Some estimators cannot resist even against a single outlier. As we have seen, this is the case for the mean and the standard deviation. Other estimators, such as the median and the interquartile range, are robust against this type of contamination because their sensitivity curve (SC) is bounded. Possibly, however, the number of outliers in a sample is so large that even estimators with a bounded SC can no longer resist their effect. Hence, to evaluate different estimators, it is important to know what the amount of contamination is an estimator can tolerate. The *breakdown point* is a measure for such *resistance* of an estimator. It quantifies, roughly, the smallest amount of contamination in the sample that may cause the estimator to take on arbitrary values. Its definition is as follows.

The finite-sample breakdown point

The breakdown point $\epsilon_n^*(T_n; \mathbf{X}_n)$ of the statistic $T_n = T_n(x_1, \ldots, x_n) = T(F_n)$ at the sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$ refers to the smallest proportion of observations in $\mathbf{X}_n$ that

need to be replaced to cause the value of the statistic to be arbitrarily large or small, and hence, to make the statistic worthless or meaningless. Note that, typically, ϵ_n^* is independent of $x_1, \ldots, x_n$.

More formally, for a univariate location estimator T_n, which breaks down if its value becomes arbitrarily large, we may define the (finite-sample) breakdown point as follows (see Hampel and Stahel 1982; Donoho and Huber 1983). In a given sample $\mathbf{X}_n = \{x_1, \ldots, x_n\}$, let us replace m data points $x_{i_1}, \ldots, x_{i_m}$ by arbitrary values $y_1, \ldots, y_m$; let us call the new data set $\mathbf{Z}_n = \{z_1, \ldots, z_n\}$. Then the *(finite-sample gross-error) breakdown point* of the estimator is

10: I assume $\min\{x; y\}$ means "find the smallest x for which y is true". This should be explained somewhere.

$$\epsilon_n^*(T_n; \mathbf{X}_n) = \min\left\{\frac{m}{n}; \max_{i_i,\ldots,i_m} \sup_{y_1,\ldots,y_m} |T_n(z_1, \ldots, z_n)| = \infty\right\}$$

Following the same idea, we will say that a scale estimator breaks down if it takes on a value that is arbitrarily large (scale explosion) or close to zero (scale implosion). Furthermore, a skewness or kurtosis estimator, which is bounded by $[-1, 1]$, breaks down if the absolute value of the estimate attains the value of 1.

▷ Example

If the ith observation among $x_1, \ldots, x_n$ goes to infinity, the mean μ and the standard deviation σ go to infinity as well. This means that the finite-sample breakdown point of these two statistics is $1/n$. In contrast, the finite-sample breakdown point of the median $Q_{0.5}$ is $\frac{(n/2)}{n}$ if n is even and $\frac{(n+1)/2}{n}$ if n is odd. That is, half the data or a bit more must be replaced to make the median take on arbitrary values. The finite-sample breakdown point of the interquartile range IQR is equal to $\frac{\lfloor n/4 \rfloor + 1}{n}$, where $\lfloor n/4 \rfloor$ denotes the integer part of $n/4$. That is, a bit more than one fourth of the data needs to be replaced to make the IQR break down.

◁

The asymptotic breakdown point

The *asymptotic breakdown point* $\epsilon^*(T, F)$ of the functional T under the distribution F is defined as

$$\epsilon^*(T, F) = \lim_{n \to \infty} \epsilon_n^*(T_n; \mathbf{X}_n)$$

with the x_i's sampled from F (cf. Hampel 1971).

11: Maybe give more details: max ϵ in $(1-\epsilon)F + \epsilon G$ with any G before T breaks down.

2.2.3 Gaussian efficiency

In general, if F it is known, the ML-estimator for that F is most efficient (see above). Furthermore, for F equal Gaussian, the mean is the ML-estimator and hence most efficient. Robust estimators should not only be efficient with respect to Gaussian, but for a wide variety of distributions. However, Gaussian efficiency is also important if

robust estimators are viewed as competitors of standard estimators. Hence it is often good to know (relative) Gaussian efficiency and then complement this with relative efficiencies for other distributions.

[Give formal definitions etc. This is not only about Gaussian efficiency. For example, the mean has high variance under fat-tails distributions; robust estimators can have better efficiency in such cases...]

2.2.4 Aspects of Interpretation

[What about asymmetric distributions and interpretation? (e.g. Median vs. Mean) Need to address such aspects. The point is that robust estimators often estimate something that is conceptually different than the nonrobust counterpart (or something where robust and nonrobust counterparts coincide only in special situations, such as in a symmetric distribution or in a normal distribution).]

2.2.5 Summary

How do we choose a good (robust) estimator? We are clearly interested in estimators with

1. a *bounded* (low gross-error sensitivity) and *smooth* (low local-shift sensitivity) *influence function*
2. and a *high breakdown point*.

Moreover, we are generally considering estimators that are *(Fisher-)consistent*. To specify this property, let us consider a *parametric model*, that is, let us assume that the underlying population distribution F has a specified form but depends on one or more unknown parameters. Formally: $F \in \{F_\theta; \theta \in \Theta\}$. For instance, in the location model, the underlying distribution is assumed to be $F_\mu(x) = F_0(x - \mu)$, where F_0 is a generic distribution function. In the location-scale model, the population distribution is $F_\theta(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$ with $\theta = (\mu, \sigma)'$. In this parametric context, an estimator $T_n = T(F_n)$ of the parameter θ of a parametric family is said to be *Fisher-consistent* if this estimator is associated with a functional T such that, for $n \to \infty$,

12: I don't really understand. Is Fisher-consistency defined with respect to a specified F_0, or does it mean that the relation holds in general, for any arbitrary F_0?

$$T_n = T(F_n) \longrightarrow^{\mathrm{P}} T(F_\theta) = \theta \qquad \text{for all } \theta \in \Theta$$

where $\longrightarrow^{\mathrm{P}}$ denotes the convergence in probability. That is, irrespective of the value of θ, a *Fisher-consistent* estimator will always converge to θ.

Finally, we are also looking for estimators that are as *efficient* as possible at an assumed model. In general, compromises between robustness and efficiency must be made to achieve good overall performance, as is shown in the following section.

3 Basic robust statistics

Many measures of location, scale, skewness and kurtosis or heaviness of the tails have been proposed and studied in the statistical literature. The present chapter is devoted to the comparison of the (asymptotic) Gaussian efficiency and robustness performance of three different classes of estimators: (i) "classic" estimators, based on (centered) moments of the distribution F_n; (ii) estimators built from specific quantiles of the distribution; (iii) estimators defined on the basis of pairwise comparisons or combinations of the observations. In addition, we will discuss robust test of normality and robust boxplots.

3.1 Robust estimation of location

There is apparent consensus in applied statistics about the fact that the sample mean and the sample median are two complementary location estimators: the mean is very efficient in case of Gaussian (i.e. normally distributed) data but fragile to outliers (and problematic in case of highly asymmetric data) while the median is very robust (and meaningful in case of asymmetry) but rather inefficient. Both are extensively used in practice. Let us briefly recall their respective properties and introduce two other frequently used location estimators.

3.1.1 The mean and the α-trimmed mean

The *mean* corresponds to the functional $\mu = \mu(F) = \int_{-\infty}^{\infty} x\, dF(x)$; its empirical counterpart is $\mu_n = \mu(F_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$. It is well known that this location estimator is the most efficient estimator for Gaussian data; its asymptotic variance is $\text{ASV}(\mu, F) = \sigma^2$, where σ^2 denotes the variance of the distribution F (taking $F = \Phi$, the standard normal distribution, we have $\text{ASV}(\mu, \Phi) = 1$). Unfortunately, the mean lacks robustness. Indeed, one single outlying observation can move this estimator towards an arbitrarily large (absolute) value: its asymptotic breakdown point $\epsilon^*(\mu, F)$ is equal to 0. Likewise, its influence function is unbounded, leading to an infinite gross-error sensitivity: $\text{IF}(x; \mu, F) = x - \mu$ (see Wilcox 2005, p. 25). Figure 3.1a shows the influence function of μ under the standard normal distribution.

13: Comment on integrals: Why not $\int x f(x)\, dx$? Maybe show somewhere that $\int f(x)\, dx = \int dF(x)$

14: Why "under the standard normal distribution"? Isn't the influence function always the same irrespective of F?

A simple and classical way to "robustify" the sample mean consists in discarding a certain proportion α ($0 \leq \alpha < 0.5$) of the smallest and of the biggest observations in

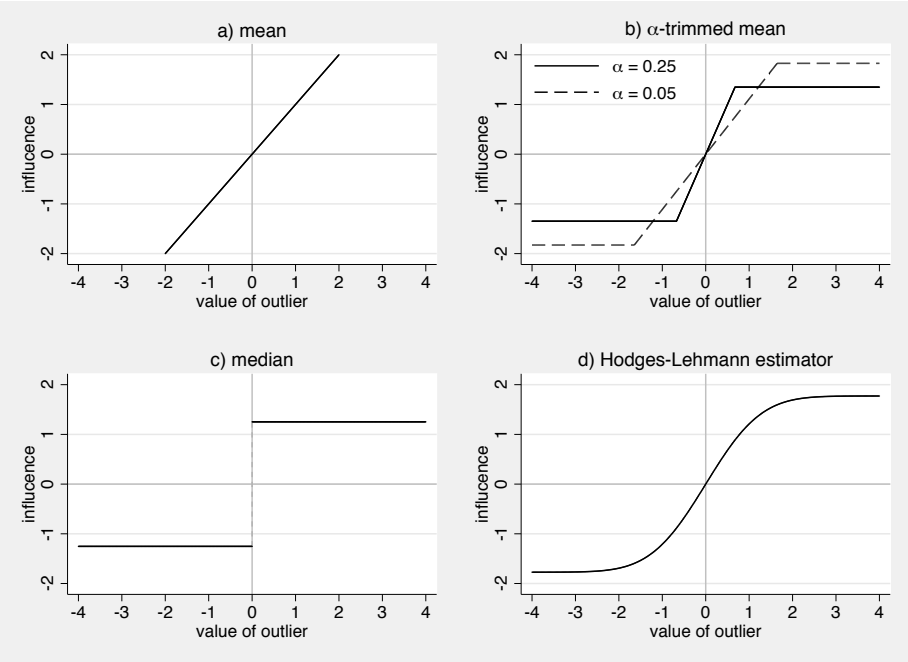

Figure 3.1. Influence functions of μ, $\mu^{0.25}$, $\mu^{0.05}$, $Q_{0.5}$, and HL under the standard normal distribution

the sample. This leads to the *α-trimmed mean* defined by

$$\mu_n^\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n - \lfloor \alpha n \rfloor} x_{(i)}$$

where $\lfloor x \rfloor$ denotes the integer part of x and $x_{(i)}$ is the ith order statistic (the observation at the ith position in the list of sorted observations in ascending order). Note that the sample mean μ_n is a special case of the α-trimmed mean μ_n^α corresponding to $\alpha = 0$. The functional associated with this location estimator is

$$\mu^\alpha(F) = \frac{1}{1 - 2\alpha} \int_{Q_\alpha}^{Q_{1-\alpha}} x \, dF(x)$$

where $Q_\alpha = F^{-1}(\alpha)$ and $Q_{1-\alpha} = F^{-1}(1 - \alpha)$ are the α and $(1 - \alpha)$ quantiles of distribution F.

The influence function of this functional has the advantage to be bounded. If F is

symmetric, then

$$\text{IF}(x;\mu^\alpha,F) = \begin{cases} \frac{1}{1-2\alpha}\big(F^{-1}(\alpha)-\mu\big) & \text{if } x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}(x-\mu) & \text{if } F^{-1}(\alpha) \le x \le F^{-1}(1-\alpha) \\ \frac{1}{1-2\alpha}\big(F^{-1}(1-\alpha)-\mu\big) & \text{if } x > F^{-1}(1-\alpha) \end{cases}$$

(see, e.g., Staudte and Sheather 1990). As an example, see Figure 3.1b for the influence functions of $\mu^{0.05}$ and $\mu^{0.25}$ under the standard Gaussian distribution Φ.

Moreover, the asymptotic breakdown point of μ^α is equal to $100\alpha\%$. Clearly, hence, the proportion α of trimming appears as a parameter allowing to choose the level of robustness of the trimmed mean. Of course, this gain in robustness goes hand in hand with a loss in efficiency. Yet, this loss is not as large as one may fear. Using the fact that

$$\text{ASV}(\mu^\alpha,F) = \int_{-\infty}^{\infty} \text{IF}(x;\mu^\alpha,F)^2\,dF(x)$$

we obtain, for example,

15: How did you compute these numbers? Are there closed form solutions?

$$\text{ASV}(\mu^{0.05},\Phi) \approx 1.0263 \qquad \text{ASV}(\mu^{0.10},\Phi) \approx 1.0604 \qquad \text{ASV}(\mu^{0.25},\Phi) \approx 1.1952$$

for the standard normal distribution. Hence, the asymptotic Gaussian relative efficiency of μ^α with respect to the mean, defined as $\text{ASV}(\mu,\Phi)/\text{ASV}(\mu^\alpha,\Phi) = 1/\text{ASV}(\mu^\alpha,\Phi)$, reaches 97% for $\alpha = 0.05$, 94% for $\alpha = 0.10$, and, despite the fact that half of the sample is discarded, 84% for $\alpha = 0.25$.

3.1.2 The median

The *median*—the quantile of order 0.5—corresponds to the functional $Q_{0.5}(F) = F^{-1}(0.5)$. Its empirical version $Q_{0.5;n}$ is simply the sample median $F_n^{-1}(0.5)$, typically computed as

$$Q_{0.5;n} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{x_{(n/2)}+x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

where $x_{(i)}$ again denotes the ith order statistic among $x_1,\ldots,x_n$.

The median performs better than the mean from the robustness point of view (see, e.g., Staudte and Sheather 1990, p. 56 and 59). First of all, its influence function is given by

$$\text{IF}(x;Q_{0.5},F) = \begin{cases} -\frac{1}{2f(F^{-1}(0.5))} & \text{if } x < F^{-1}(0.5) \\ 0 & \text{if } x = F^{-1}(0.5) \\ \frac{1}{2f(F^{-1}(0.5))} & \text{if } x > F^{-1}(0.5) \end{cases}$$

where f is the density function associated with F. In particular, as displayed in Figure 3.1c, $\text{IF}(x;Q_{0.5},\Phi) = \text{sign}(x)\sqrt{\pi/2}$ for standard Gaussian data, since $f\big(\Phi^{-1}(0.5)\big) = \sqrt{\pi/2}$. This influence function is bounded, leading to a bounded gross-error sensitivity,

but it has a discontinuity at $x = 0$. Furthermore, the asymptotic breakdown point of the median is equal to 50% and is thus higher than the asymptotic breakdown point of the α-trimmed mean, regardless of the value of α.

16: Say why the discontinuity is a problem.

Finally, the asymptotic variance of the median is given as

$$\text{ASV}(Q_{0.5}, F) = \frac{1}{4f(F^{-1}(0.5))^2}$$

Hence, relative Gaussian efficiency compared to the mean is equal to

$$\frac{\text{ASV}(\mu, \Phi)}{\text{ASV}(Q_{0.5}, \Phi)} = \frac{1}{\pi/2} = \frac{2}{\pi} \approx 64\%$$

3.1.3 The Hodges-Lehmann estimator

Hodges and Lehmann (1963) have introduced an alternative location estimator that has the advantage of a bounded, continuous and smooth influence function, but also a high asymptotic Gaussian relative efficiency with respect to the sample mean. The *Hodges-Lehmann estimator* at the sample $\mathbf{X}_n$ is defined by

17: Can't we type $\text{HL}_n = \underset{i<j}{\text{med}}\left(\frac{x_i+x_j}{2}\right)$? This would seem easier to understand to me. Also, why only $i < j$? Why not all possible combinations (even including $i = j$)?

$$\text{HL}_n = \text{med}\left\{\frac{x_i + x_j}{2}; i < j\right\}$$

It is the empirical version of the functional $\text{HL} = \text{HL}(F)$, which is defined as the median of the distribution of $(X+Y)/2$, where X and Y are i.i.d. random variables of distribution F.

18: I use $f(x)^2$ instead of $f^2(x)$. I know the latter is often used in stats literature, but to me the former much is clearer.

For symmetric F, the influence function of HL is given as

$$\text{IF}(x; \text{HL}, F) = \frac{2F(x-\mu) - 1}{2\int_{-\infty}^{\infty} f(y)^2\, dy}$$

Figure 3.1c presents the influence function for standard Gaussian data. It illustrates that outliers have a bounded influence. Also note that the sensitivity of the Hodges-Lehmann estimator depends on the smoothness of F. Hence, the local-shift sensitivity is small as long as the data have a smooth distribution function.

19: The original just said that the sensitivity depends on the smoothness of F without saying why this is relevant. I thus added some text. Please check, whether it is ok.

Because HL combines the robust properties of the median with the efficiency properties of averaging, it performs well for a variety of distributions. Based on (2.2) we obtain

$$\text{ASV}(\text{HL}, F) = \frac{1}{12}\left(\frac{1}{\int_{-\infty}^{\infty} f(y)^2\, dy}\right)^2$$

For example, for standard Gaussian date, $\text{ASV}(\text{HL}, \Phi) = \pi/3 = 1.0472$. Hence, the asymptotic efficiency of the Hodges-Lehmann estimator relative to the mean is equal

to $3/\pi \approx 95\%$ at the normal distribution. Moreover, the Hodges-Lehmann estimator reaches a relative efficiency with respect to the mean of at least 86% for any symmetric distribution (see, for example, Staudte and Sheather 1990, p. 120-121). Compared to the median, the Hodges-Lehmann estimator has a higher Gaussian efficiency (95% vs. 64%) but a lower asymptotic breakdown point (29% vs. 50%).

3.1.4 M estimate of location

[What about M estimators???]

3.1.5 Summary

Table 3.1 summarizes the different properties of the four location estimators. From the perspective of a good balance between high Gaussian efficiency and a high breakdown point, the Hodges-Lehmann estimator appears to perform particularly well.

Table 3.1. Characteristics of the four location estimators

Estimator	Class	Gaussian efficiency	Asymptotic breakdown point	Bounded influence function
mean μ_n	moment	100%	0%	no
α-trimmed mean μ_n^α	moment	$\alpha = 0.05$: 97% $\alpha = 0.10$: 94% $\alpha = 0.25$: 84%	$100\alpha\%$	yes
median $Q_{0.5;n}$	quantile	64%	50%	yes
Hodges-Lehmann HL_n	pairwise	95%	29%	yes

3.2 Robust estimation of scale

3.2.1 The standard deviation

The classic statistic to estimate the scale parameter of a distribution is the *standard deviation* σ_n, corresponding to the functional

$$\sigma = \sigma(F) = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 \, dF(x)}$$

At sample $\mathbf{X}_n$, σ_n is typically computed as

20: I changed this to the usual $1/(n-1)$ variant; if we use the $1/n$ version we need to explain why.

$$\sigma_n = \sigma(F_n) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_n)^2}$$

with μ_n as defined above. The standard deviation is the most efficient estimator of the scale parameter σ in case of Gaussian data (note that $\text{ASV}(\sigma, \Phi) = 0.5$). However, just like the mean, the standard deviation is very fragile to outliers. Its influence function, given as

21: Show why $\text{ASV}(\sigma, \Phi) = 0.5$

$$\text{IF}(x; \sigma, F) = \frac{1}{2\sigma}\left(x^2 - 2\mu x + \mu^2 - \sigma^2\right)$$

is unbounded and its asymptotic breakdown point is equal to 0% (e.g., Rousseeuw and Croux 1993, p. 1275). The influence function for standard Gaussian data, given as $\text{IF}(x; \sigma, \Phi) = \frac{1}{2}(x^2 - 1)$, is displayed in Figure 3.2a.

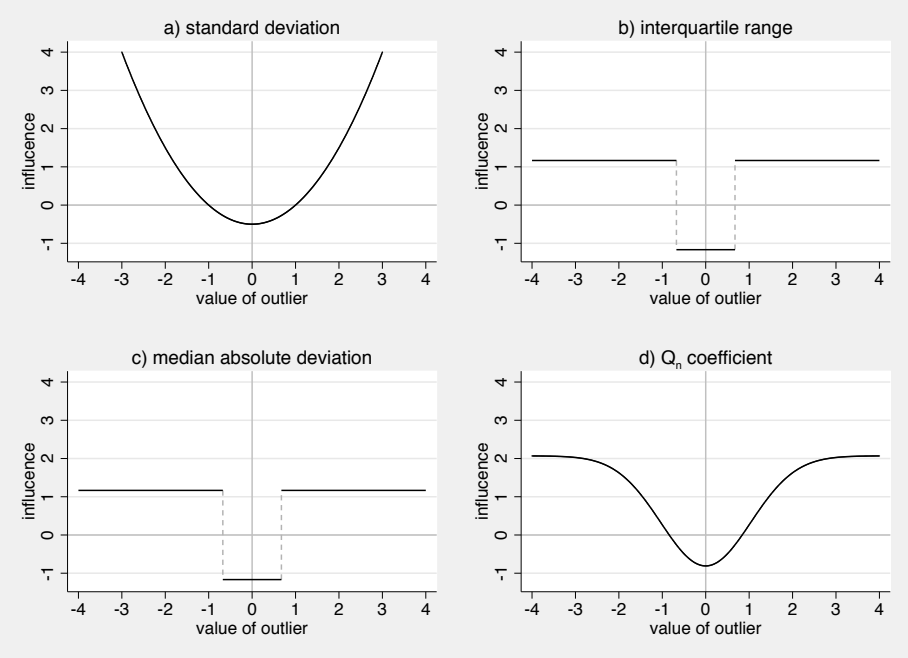

Figure 3.2. Influence functions of σ, IQR_c, MAD and Q under the standard normal distribution

3.2.2 The interquartile range

A common alternative scale measure, defined on the basis of quantiles, is the *interquartile range*

$$\text{IQR} = Q_{0.75} - Q_{0.25}$$

where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles of distribution F. Instead of IQR one frequently uses the *corrected interquartile range* IQR_c defined as

$$\text{IQR}_c = d \cdot \text{IQR}$$

where d is a constant chosen to make the estimator $\text{IQR}_{c;n}$ Fisher-consistent for the scale parameter of the underlying distribution. For example, for a Gaussian distribution, use $d = 1/\big(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\big) \approx 0.7413$ to make the corrected interquartile range a consistent estimator for the usual scale parameter σ.

The influence function of the interquartile range is bounded, but discontinuous (see, for example, Wilcox 2005, p. 35–36). It is given as

$$\text{IF}(x; \text{IQR}, F) = \begin{cases} \frac{1}{f(F^{-1}(0.25))} - C & \text{if } x < F^{-1}(0.25) \\ -C & \text{if } F^{-1}(0.25) \le x \le F^{-1}(0.75) \\ \frac{1}{f(F^{-1}(0.75))} - C & \text{if } x > F^{-1}(0.75) \end{cases}$$

where

$$C = \frac{1}{4}\left[\frac{1}{f(F^{-1}(0.25))} + \frac{1}{f(F^{-1}(0.75))}\right]$$

Note that $\text{IF}(x; \text{IQR}_c, F) = d \cdot \text{IF}(x; \text{IQR}, F)$. Figure 3.2b displays the influence function of IQR_c for standard Gaussian data. Like the two quartiles $Q_{0.25}$ and $Q_{0.75}$, IQR and IQR_c have an asymptotic breakdown point equal to 25%. This gain in robustness with respect to the standard deviation is accompanied by a high loss in Gaussian efficiency. Specifically, $\text{ASV}(\text{IQR}_c, \Phi) = 1.3605$ so that the asymptotic Gaussian efficiency of the corrected interquartile range with respect to the standard deviation is equal to $\text{ASV}(\sigma, \Phi)/\text{ASV}(\text{IQR}_c, \Phi) = 0.5/1.3605 \approx 37\%$.

22: Give details about $\text{ASV}(\text{IQR}_c, \Phi)$. How is it computed?

3.2.3 The median absolute deviation

Another robust alternative is the *median absolute deviation*, whose empirical version is defined as

$$\text{MAD}_n = d \cdot \underset{i}{\text{med}} \left| x_i - \underset{j}{\text{med}}\, x_j \right|$$

That is, the median absolute deviation is equal to the (rescaled) median of the absolute deviations from the median of the variable of interest. For Gaussian distributions, we need to set $d = 2/\big(\Phi^{-1}(0.75) - \Phi^{-1}(0.25)\big) \approx 1.4826$ to make MAD_n Fisher-consistent for the scale parameter σ.

The functional corresponding to MAD_n is

$$\text{MAD} = \text{MAD}(F) = d \cdot G_F^{-1}(0.5)$$

where G_F is the distribution function of $|X - Q_{0.5}(F)| = |X - F^{-1}(0.5)|$ with X as a random variable of distribution F. In other words, MAD is equal to d times the median of

the distribution associated with $|X - Q_{0.5}(F)|$, the magnitude of the difference between X and its median.

The median absolute deviation may appear more attractive than the interquartile range for certain purposes. It has the same asymptotic Gaussian efficiency as the corrected interquartile range, but performs better in terms of robustness (e.g., Rousseeuw and Croux 1993, p. 1273–1274): Its asymptotic breakdown point is as high as 50%. The influence function of MAD is given as $d \cdot \text{IF}(x; G_F^{-1}(0.5), F)$ with

$$\text{IF}(x; G_F^{-1}(0.5), F) = \frac{\text{sign}\big(|x - F^{-1}(0.5)| - G_F^{-1}(0.5)\big) - C \cdot \text{sign}\big(x - F^{-1}(0.5)\big)}{2\big[f\big(F^{-1}(0.5) + G_F^{-1}(0.5)\big) + f\big(F^{-1}(0.5) - G_F^{-1}(0.5)\big)\big]}$$

where

$$C = \frac{f\big(F^{-1}(0.5) + G_F^{-1}(0.5)\big) - f\big(F^{-1}(0.5) - G_F^{-1}(0.5)\big)}{f(F^{-1}(0.5))}$$

(see, e.g., Wilcox 2005). In particular, under the standard Gaussian distribution (with $F = \Phi$ and $f = \phi$), we have

23: Is it only for Gaussian data that the IF for IQR and MAD is the same, or is this true for any symmetric distribution?

$$\text{IF}(x; \text{MAD}, \Phi) = 1.4826 \cdot \frac{\text{sign}\big(|x| - \Phi^{-1}(0.75)\big)}{4\,\phi(\Phi^{-1}(0.75))}$$

24: In Rousseeuw/Croux the IF of MAD is closer to that one of Q; check that...

which is displayed in Figure 3.2c.

Despite its good robustness properties, MAD is primarily useful for *symmetric* distributions. In fact, the MAD corresponds to finding the symmetric interval around the median that contains 50% of the data (50% of the probability mass), which does not appear to be a very sensible approach for asymmetric distributions. The interquartile range does not have this restriction, as the quartiles need not be equally far away from the median.

3.2.4 The Q_n coefficient

Finally, a very interesting but relatively unknown scale estimator is the Q_n statistic introduced by Rousseeuw and Croux (1993):

$$Q_n = d \cdot \{|x_i - x_j|; i < j\}_{(k)}$$

where d is a constant factor allowing Q_n to be a Fisher-consistent estimator for the scale parameter of the underlying distribution F and $k = \binom{h}{2} \approx \binom{n}{2}/4$, with $h = \lfloor n/2 \rfloor + 1$ (h is roughly half the number of observations). In other words, if we omit the constant d, the statistic Q_n corresponds approximately to the 0.25 quantile of the $\binom{n}{2}$ distances $|x_i - x_j|$, $i < j$. Fisher-consistency for the scale parameter σ at Gaussian distributions can be achieved by setting $d = 1/\big(\sqrt{2}\Phi^{-1}(5/8)\big) \approx 2.2191$.

25: In Rousseeuw/Croux the value is 2.2219, which seems to be an error.

The functional counterpart of Q_n is

$$Q = Q(F) = d \cdot H_F^{-1}(0.25)$$

where H_F is the distribution function of $|X - Y|$ with X and Y being two independent random variables of distribution F. Note that $Q(F_n)$ is not exactly the same as Q_n, where we take an order statistic among $\binom{n}{2}$ elements instead of n^2 elements, but asymptotically this makes no difference.

The scale estimator Q_n has globally better properties than the previous scale estimators we have presented. Like the MAD, it has an asymptotic breakdown point equal to 50%. Yet, unlike the MAD, Q_n is not slanted towards symmetric distributions. Moreover, its influence function is not only bounded, but also smooth:

26: I use numerical integration to obtain the value of the denominator (used for the graph and for the computation of the efficiency). Is there also a closed-form solution?

$$\text{IF}(x; Q, F) = d \cdot \frac{0.25 - F(x + d^{-1}) + F(x - d^{-1})}{\int_{-\infty}^{\infty} f(y + d^{-1})\, dF(y)}$$

Figure 3.2d displays the influence function of Q_n for standard Gaussian data.

Finally, Q_n is asymptotically more efficient than the median absolute deviation under Gaussian distributions. In particular, numerical integration of $\int_{-\infty}^{\infty} \text{IF}(x; Q, \Phi)^2\, d\Phi(x)$ yields $\text{ASV}(Q, \Phi) \approx .6089$, corresponding to an asymptotic Gaussian relative efficiency with respect to the standard deviation of $\text{ASV}(\sigma, \Phi)/\text{ASV}(Q, \Phi) \approx 0.5/.6089 \approx 82\%$, which is surprisingly high.

27: I get 0.6089, not 0.6077.

28: The following output shows my computations. Results should be precise (e.g., they do not change if I increase the number of integration points to 100000). This is just for you. It will be removed later on.

```
. mata
------------------------------------------------- mata (type end to exit) -----
: d = 1/(sqrt(2)*invnormal(5/8))
: d
  2.219144466
: function myf(x) return(normalden(x+sqrt(2)*invnormal(5/8))*normalden(x))
: // note: normalden(37)=2.12e-298, normalden(38)=0
: dd = mm_integrate_sr(&myf(), -38, 38, 1000, 1)
: dd
  .2681315272
: function myf1(x, d, dd)
> {
>     return( (d * (0.25 - normal(x + 1/d) + normal(x - 1/d)) / dd)^2 *
>               normalden(x))
> }
: mm_integrate_sr(&myf1(), -38, 38, 1000, 1, d, dd)
  .6089006937
: end
--------------------------------------------------------------------------------
```

3.2.5 Summary

To conclude, Table 3.2 summarizes the different properties of the presented estimators of scale. As is evident, the Q_n coefficient has superior properties in terms of robustness and efficiency compared to the other robust estimators.

Table 3.2. Characteristics of the four scale estimators

Estimator	Class	Gaussian efficiency	Asymptotic breakdown point	Bounded influence function
standard deviation σ_n	moment	100%	0%	no
interquartile range IQR_n	quantile	37%	25%	yes
median absolute deviation MAD_n	quantile	37%	50%	yes
Q_n coefficient	pairwise	82%	50%	yes

3.3 Robust estimation of skewness

3.3.1 The Fisher coefficient

As far as skewness is concerned, the most classic estimator is the *Fisher coefficient*. Given sample $\mathbf{X}_n$ the Fisher coefficient is defined as

$$\gamma_{1;n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_n}{\sigma_n}\right)^3$$

with μ_n and σ_n as the sample mean and the standard deviation. This estimator is associated with the functional

$$\gamma_1 = \gamma_1(F) = \mu_3(F)/\sigma(F)^3 \qquad \text{where} \quad \mu_3(F) = \int_{-\infty}^{\infty}(x - \mu(F))^3\, dF(x)$$

which is equal to zero at symmetric F.

Since the Fisher coefficient relies on the mean and the standard deviation, it is not surprising that its resistance to outliers is poor. More precisely, its asymptotic breakdown point is equal to 0% and its influence function is unbounded (see, for example, Groeneveld 1991). For a *symmetric* distribution F, assuming $\mu(F) = 0$ and $\sigma(F) = 1$ without loss of generality, the influence function of the Fisher coefficient is given as

29: What exactly does this mean? Is the IF always like that irrespective of μ and σ or is it different, but does not change shape? (In this case: What would be the IF for $\mu \neq 0$ and $\sigma \neq 1$?)

$$\text{IF}(x; \gamma_1, F) = x^3 - 3x$$

See Figure 3.3a for a graphical display. The influence function for an *asymmetric* distribution is more complex. However, although no longer being an odd function of x, it has a quite similar form to that found for symmetric F. Also note, for comparisons with other skewness estimators, that $\text{ASV}(\gamma_1, \Phi) = 6$.

30: How is this found?

3.3.2 Yule and Kendall, and Hinkley skewness measures

Alternative estimators of skewness, such as $(\mu_n - \text{mode}_n)/\sigma_n$ and $(\mu_n - Q_{0.5;n})/\sigma_n$ as proposed by Karl Pearson, are just as fragile with respect to outliers as the standard

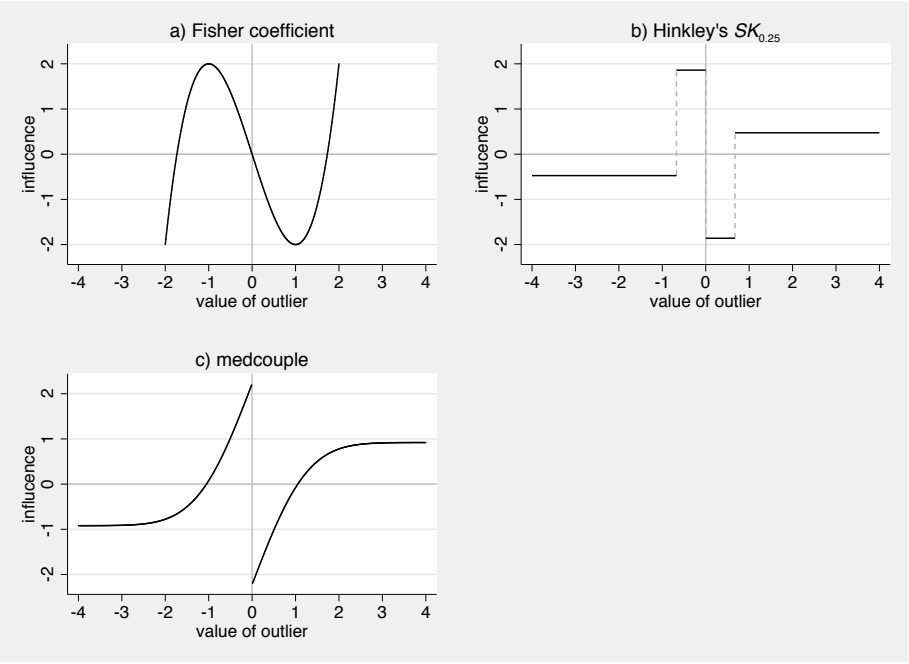

Figure 3.3. Influence functions of γ_1, $\mathrm{SK}_{0.25}$ and MC under the standard normal distribution

skewness estimator.

Fortunately, robust alternatives based on quantiles are available. For example, Yule and Kendall have proposed the skewness measure

31: Do you have a citation for this?

$$\mathrm{SK}_{\mathrm{YK}} = \frac{(Q_{0.75} - Q_{0.5}) - (Q_{0.5} - Q_{0.25})}{Q_{0.75} - Q_{0.25}} = \frac{Q_{0.25} + Q_{0.75} - 2Q_{0.5}}{Q_{0.75} - Q_{0.25}}$$

where $Q_{0.25}$, $Q_{0.5}$, and $Q_{0.75}$ are the three quartiles.

Hinkley (1975) generalized this formula to other quantiles:

$$\mathrm{SK}_p = \frac{(Q_{1-p} - Q_{0.5}) - (Q_{0.5} - Q_p)}{Q_{1-p} - Q_p} = \frac{Q_p + Q_{1-p} - 2Q_{0.5}}{Q_{1-p} - Q_p},$$

where Q_p and Q_{1-p} are the quantiles of order p and $1-p$ (with $0 < p < 0.5$). Hinkley's measure SK_p is equal to zero for symmetric distributions, is positive for right tailed (left skewed) and negative for left tailed (right skewed) distributions. It has a much smaller asymptotic variance under the standard normal distribution than γ_1. For instance, $\mathrm{ASV}(\mathrm{SK}_{0.25}, \Phi) = 1.8421$. The asymptotic breakdown point of SK_p is equal to $100p\%$ (in particular, the Yule and Kendall skewness estimator, corresponding to $p = 0.25$, has an asymptotic breakdown point equal to 25%).

32: How is the ASV derived? Plus: what is the expected value of $Q_{0.25}$ for Gaussian data?

For a *symmetric* distribution F with density f and, without loss of generality, $\mu(F) = F^{-1}(0.5) = 0$ and $\sigma(F) = 1$, the influence function of SK_p is

$$\text{IF}(x; \text{SK}_p, F) = \begin{cases} \frac{-1/f(0)}{F^{-1}(1-p)-F^{-1}(p)} & \text{if } 0 \leq x < F^{-1}(1-p) \\ \frac{1/f\left(F^{-1}(1-p)\right)-1/f(0)}{F^{-1}(1-p)-F^{-1}(p)} & \text{if } F^{-1}(1-p) \leq x \end{cases}$$

for $x \geq 0$ and $\text{IF}(x; \text{SK}_p, F) = -\text{IF}(-x; \text{SK}_p, F)$ for $x < 0$. The influence function for standard Gaussian data is displayed in Figure 3.3b. In case of an *asymmetric* distribution the influence function is much more complex and is no longer an odd function of x (see Groeneveld 1991, p. 101).

3.3.3 The medcouple

As usual when working with quantiles, the influence function of SK_p is not smooth. To tackle this problem, Brys et al. (2004a) propose to replace the quantiles Q_p and Q_{1-p} in SK_p by actual data points and introduce a new skewness measure called *medcouple*. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the n order statistics associated to the sample $\mathbf{X}_n$ and $Q_{0.5;n}$ be the sample median. The medcouple is then defined as

$$\text{MC}_n = \underset{x_{(i)} \leq Q_{0.5;n} \leq x_{(j)}}{\text{med}} h\left(x_{(i)}, x_{(j)}\right)$$

where, for all $x_{(i)} \neq x_{(j)}$, the kernel function h is given as

$$h\left(x_{(i)}, x_{(j)}\right) = \frac{\left(x_{(j)} - Q_{0.5;n}\right) - \left(Q_{0.5;n} - x_{(i)}\right)}{x_{(j)} - x_{(i)}}$$

For the special case $x_{(i)} = x_{(j)} = Q_{0.5;n}$, the kernel is defined as follows: let $m_1 < \cdots < m_k$ denote the indices of the order statistics that are tied to the median $Q_{0.5;n}$ (that is $x_{(m_l)} = Q_{0.5;n}$ for all $l = 1, \ldots, k$). Then,

33: What about other ties, i.e. $x_{(i)} = x_{(j)} \neq Q_{0.5;n}$? $h()$ is not defined for these cases because the denominator is 0. Should $h()$ be defined as 0 in this case?

$$h\left(x_{(m_i)}, x_{(m_j)}\right) = \begin{cases} -1 & \text{if } i + j < k + 1 \\ 0 & \text{if } i + j = k + 1 \\ 1 & \text{if } i + j > k + 1 \end{cases}$$

Due to the denominator it is clear that $h\left(x_{(i)}, x_{(j)}\right)$, and hence MC_n, will always lie between -1 and 1 (similar to SK_p).

The functional form of the medcouple is simply defined at any continuous distribution F as

$$\text{MC} = \text{MC}(F) = \underset{X \leq Q_{0.5} \leq Y}{\text{med}} h(X, Y)$$

where $Q_{0.5} = F^{-1}(0.5)$ is the median of F and X and Y are i.i.d. random variables of distribution F. The kernel h is the same as above with the finite-sample median $Q_{0.5;n}$ replaced by $Q_{0.5}$. This functional MC is equal to zero in case of a symmetric

distribution F. It is positive for right tailed (left skewed) and negative left tailed (right skewed) distributions.

The asymptotic breakdown point of the medcouple is equal to 25%, which is the same as for the quartile skewness $\mathrm{SK}_{0.25} = \mathrm{SK}_{\mathrm{YK}}$. The advantage of MC, however, lies in the fact that its influence function resembles a smoothed version of the influence function of SK_p $(0 < p < 0.5)$. In particular, for standard Gaussian F, the influence function is given as

34: Why SK_p? Why not $\mathrm{SK}_{0.25}$?

$$\mathrm{IF}(x; \mathrm{MC}, \Phi) = \pi\left(2\Phi(x) - 1 - \frac{\mathrm{sign}(x)}{\sqrt{2}}\right)$$

(see Figure 3.3c). This leads to an asymptotic variance for Gaussian data of

35: Is 1.25 an exact result? How do you compute that? Plus: what is the expected value of medcouple for Gaussian data?

$$\mathrm{ASV}(\mathrm{MC}, \Phi) = \int_{-\infty}^{\infty} \mathrm{IF}(x; \mathrm{MC}, \Phi)^2 \, d\Phi(x) = 1.25$$

3.3.4 Summary

Table 3.3 provides an overview of the properties of the discussed skewness estimators. As can be seen, both proposed robust estimators are more robust and, at the same time, more efficient than the standard skewness coefficient.

36: How to compute efficiency? Need to rescale...

Table 3.3. Characteristics of the three skewness estimators

Estimator	Class	Gaussian efficiency[a]	Asymptotic breakdown point	Bounded influence function
Fisher coefficient $\gamma_{1;n}$	moment	100%	0%	no
Hinkley's $\mathrm{SK}_{0.25;n}$	quantile	???	25%	yes
medcouple MC_n	pairwise	???	25%	yes

[a] relative to the efficiency of the Fisher coefficient

3.4 Robust estimation of the tails heaviness

3.4.1 The classical kurtosis coefficient

The classical *kurtosis coefficient* is defined by the functional

$$\gamma_2 = \gamma_2(F) = \mu_4(F)/\sigma(F)^4$$

with

$$\mu_4(F) = \int_{-\infty}^{\infty} (x - \mu(F))^4 \, dF(x)$$

Given a sample $\mathbf{X}_n$, $\gamma_{2;n}$ is computed as

$$\gamma_{2;n} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_n}{\sigma_n}\right)^4$$

with μ_n and σ_n as the sample mean and the standard deviation.

The kurtosis coefficient is often considered as a measure of the tail heaviness of a distribution relative to that of the normal distribution. In particular, γ_2 is equal to three in case of distributions with a tails heaviness similar to the normal distribution, is larger than three for leptokurtic distributions (i.e., distributions with heavier tails than the normal distribution) and is smaller than three for platokurtic distributions (i.e., distributions with lighter tails than the normal distribution). However, since the coefficient also measures the peakedness of a distribution, there is no agreement on what the kurtosis really estimates. Another disadvantage of the kurtosis is that its interpretation, and consequently its use, is restricted to symmetric distributions (due of its intrinsic comparison with the symmetric normal distribution). Moreover, as usual for estimators relying on the mean and the standard deviation, the kurtosis coefficient is very sensitive to outliers in the data. This is reflected in the asymptotic breakdown point being equal to zero. The influence function is unbounded and is given as

$$\text{IF}(x; \gamma_2, F) = (z^2 - \gamma_2)^2 - \gamma_2(\gamma_2 - 1) - 4\gamma_1 z,$$

where $z = (x - \mu(F))/\sigma(F)$, $\gamma_1 = \mu_3(F)/\sigma(F)^3$ and $\gamma_2 = \mu_4(F)/\sigma(F)^4$ (see Ruppert 1987). See Figure 3.4a for a graphical display of the influence function for standard Gaussian F, which is given as $\text{IF}(x; \gamma_2, \Phi) = (x^2 - 3)^2 - 6$. The form of the influence function indicates that contamination at the center has far less influence than that in the extreme tails. This suggests that γ_2 is primarily a measure of tail behavior, and only to a lesser extent of peakedness. The asymptotic variance of γ_2 for Gaussian data is given as $\text{ASV}(\gamma_2, \Phi) = 24$.

37: Closed form solution?

3.4.2 The quantile and medcouple tail weight measures

To overcome the problems of the kurtosis coefficient, Brys et al. (2006) have proposed two measures of *left* and *right* tail weight for univariate continuous distributions. As discussed below, these measures have the advantage that they can be applied to symmetric as well as asymmetric distributions that do not need to have finite moments. Moreover, their interpretation is unambiguous and they are robust against outlying values.

More precisely, Brys et al. (2006) defined *left* and *right* tail measures as measures of skewness that are applied to the half of the probability mass lying on the left side or on the right side of the median $Q_{0.5}$ of the distribution F, respectively. As candidate measures of skewness they use both, SK_p $(0 < p < 0.5)$ and MC (see above).

Recall that SK_p is a measure of skewness of the distribution F around $Q_{0.5}$, involving the quantiles Q_p and Q_{1-p} of orders p and $(1-p)$ of F. Applying $-\text{SK}_{p/2}$ to the left half of the distribution F (i.e., to $x < Q_{0.5}$) leads to the *Left Quantile Weight*

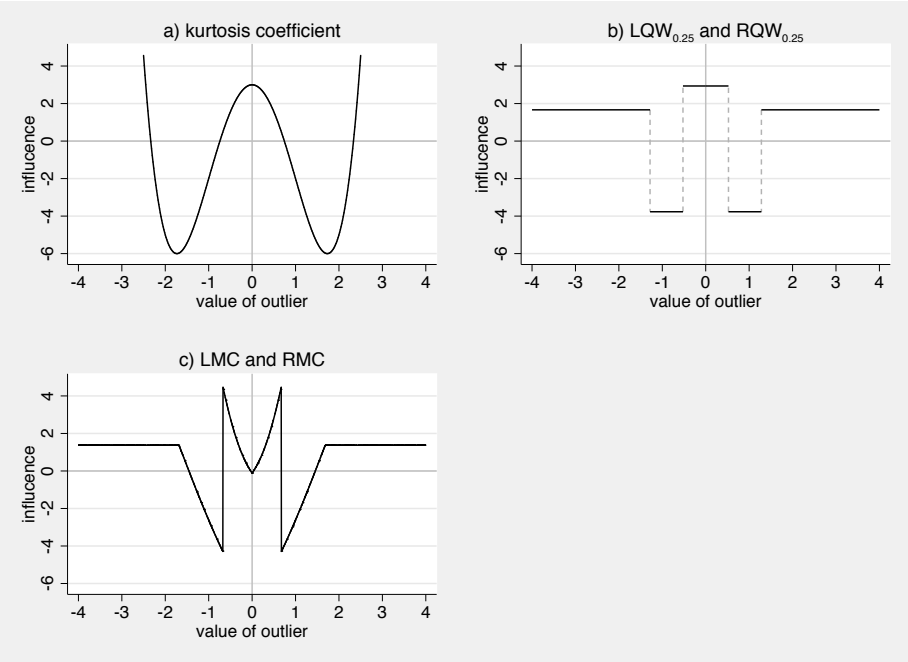

Figure 3.4. Influence functions of γ_2, $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$, LMC and RMC under the standard normal distribution

$\text{LQW}_p = \text{LQW}_p(F)$, which corresponds to (the opposite of) a measure of skewness of the left half of F around the first quartile $Q_{0.25}$ involving the quantiles $Q_{p/2}$ and $Q_{0.5-p/2}$:

$$\text{LQW}_p = -\frac{(Q_{0.5-p/2} - Q_{0.25}) - (Q_{0.25} - Q_{p/2})}{Q_{0.5-p/2} - Q_{p/2}} = -\frac{Q_{p/2} + Q_{0.5-p/2} - 2Q_{0.25}}{Q_{0.5-p/2} - Q_{p/2}}.$$

Similarly, applying $\text{SK}_{p/2}$ to the right half of the distribution F (i.e., to $x > Q_{0.5}$) provides the *Right Quantile Weight* $\text{RQW}_p = \text{RQW}_p(F)$ which corresponds to a measure of skewness of the right half of F around the third quartile $Q_{0.75}$ involving the quantiles $Q_{0.5+p/2}$ and $Q_{1-p/2}$:

$$\text{RQW}_p = \frac{(Q_{1-p/2} - Q_{0.75}) - (Q_{0.75} - Q_{0.5+p/2})}{Q_{1-p/2} - Q_{0.5+p/2}} = \frac{Q_{0.5+p/2} + Q_{1-p/2} - 2Q_{0.75}}{Q_{1-p/2} - Q_{0.5+p/2}}.$$

With $p = 1/4 = 0.25$, we obtain

$$\text{LQW}_{0.25} = -\frac{Q_{0.125} + Q_{0.375} - 2Q_{0.25}}{Q_{0.375} - Q_{0.125}}$$

$$\text{RQW}_{0.25} = \frac{Q_{0.625} + Q_{0.875} - 2Q_{0.75}}{Q_{0.875} - Q_{0.625}}$$

Note that the sample versions $\text{LQW}_{p;n}$ and $\text{RQW}_{p;n}$ are easily found by using the quantiles of F_n, the empirical distribution function of $\mathbf{X}_n$.

As with LQW and RQW, we can also apply the MC to each side of the distribution, leading to the *Left Medcouple* (LMC) and to the *Right Medcouple* (RMC), defined as

$$\text{LMC} = \text{LMC}(F) = -\text{MC}(x < Q_{0.5})$$

$$\text{RMC} = \text{RMC}(F) = \text{MC}(x > Q_{0.5})$$

By using MC_n, the finite-sample version of MC, we obtain the finite-sample versions LMC_n and RMC_n.

Since both the quantile and the medcouple tail weight measures only depend on quantiles, they are given for any distribution, even for distributions without finite moments. Note that LQW and RQW require to fix the parameter p in advance (depending on the degree of robustness one wants to attain), whereas LMC and RMC do not require any additional parameter to be set.

Some general properties of the tail weight measures are as follows. Let X be a random variable with continuous distribution F_X. Furthermore, let W stand for any of the defined tail weight measures; let LW stand for left tailed measures and RW for the right tailed measures. Then:

- Like the skewness measures SK_p and MC, W is location and scale invariant. That is, $\text{W}(F_{aX+b}) = \text{W}(F_X)$.
- $\text{LW}(F_{-X}) = \text{RW}(F_X)$.
- If F is symmetric, then $\text{LW}(F) = \text{RW}(F)$.
- $\text{W} \in [-1, 1]$.

The medcouple tail weight measures can resist up to 12.5% outliers in the data. In a similar way, it can be shown that the left and right quantile tail weight measures LQW_p and RQW_p have an asymptotic breakdown point equal to $(100p/2)\%$; in particular, $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$ have the same asymptotic breakdown point as LMC and RMC, and $\text{LQW}_{0.125}$ and $\text{RQW}_{0.125}$ have an asymptotic breakdown point of 6.25%.

Moreover, the influence functions of LMC and RMC are smooth versions of the influence functions of $\text{LQW}_{0.25}$ and $\text{RQW}_{0.25}$, as shown in Figure 3.4b and c for standard Gaussian F.[1] All these influence functions are bounded. More precisely, if F is a continuous distribution with density f and if we denote, as usual, the quantile of order p of F by $Q_p = F^{-1}(p)$, we have

$$\text{IF}(x; \text{LQW}_p, F) = 2\,\frac{\big(\text{IF}(x; Q_{0.25}, F) - \text{IF}(x; Q_{0.5-p/2}, F)\big)\big(q_{0.25} - Q_{p/2}\big) - \big(\text{IF}(x; Q_{p/2}, F) - \text{IF}(x; Q_{0.25}, F)\big)\big(Q_{0.5-p/2} - Q_{0.25}\big)}{\big(Q_{0.5-p/2} - Q_{p/2}\big)^2}$$

1. Figure 3.4 shows only the left part (defined on $\mathbb{R}^-$) of the influence functions of LMC and $\text{LQW}_{0.25}$, and the right part (defined on $\mathbb{R}^+$) of the influence functions of RMC and $\text{RQW}_{0.25}$.

and

$$\mathrm{IF}(x; \mathrm{RQW}_p, F) = 2\,\frac{\begin{array}{c}\big(\mathrm{IF}(x; Q_{0.5+p/2}, F) - \mathrm{IF}(x; Q_{0.75}, F)\big)\big(q_{1-p/2} - Q_{0.75}\big) \\ -\big(\mathrm{IF}(x; Q_{0.75}, F) - \mathrm{IF}(x; Q_{1-p/2}, F)\big)\big(Q_{0.75} - Q_{0.5+p/2}\big)\end{array}}{\big(Q_{1-p/2} - Q_{0.5+p/2}\big)^2}$$

with

$$\mathrm{IF}(x; Q_p, F) = \frac{p - \mathrm{I}[x < Q_p]}{f(Q_p)}$$

The expression of the influence functions of LMC and RMC is more complex and can be found in Brys et al. (2006, p. 740–741).

38: (maybe have a look and then check computation of graph; it seems IF has a discontinuity; need to use shortdash)

Finally, the asymptotic variances of the left and right tail weight measures under Gaussian distributions are much smaller than the variance of the classical kurtosis coefficient γ_2. In particular:

$$\mathrm{ASV}(\mathrm{LQW}_{0.25}, \Phi) = \mathrm{ASV}(\mathrm{RQW}_{0.25}, \Phi) = 3.71$$
$$\mathrm{ASV}(\mathrm{LQW}_{0.125}, \Phi) = \mathrm{ASV}(\mathrm{RQW}_{0.125}, \Phi) = 2.23$$

and

$$\mathrm{ASV}(\mathrm{LMC}, \Phi) = \mathrm{ASV}(\mathrm{RMC}, \Phi) = 2.62$$

39: How are these numbers computed? Furthermore, is it a fair comparison. That is, are L/RQW and L/RMC similar in size to the kurtosis for Gaussian data or do they have to be rescaled? Only if the measures have the same size the variances can be compared.

3.4.3 Summary

Table 3.4 summarizes the properties of the presented tails heaviness estimators. [Another sentence needed here!]

40: How to compute efficiency? (need to rescale)

Table 3.4. Characteristics of the three tails heaviness estimators

Estimator	Class	Gaussian efficiency[a]	Asymptotic breakdown point	Bounded influence function
kurtosis coefficient $\gamma_{2;n}$	moment	100%	0%	no
$\mathrm{LQW}_{0.25;n}$ and $\mathrm{RQW}_{0.25;n}$	quantile	???	12.5%	yes
LMC_n and RMC_n	pairwise	???	12.5%	yes

[a] relative to the efficiency of the kurtosis coefficient

3.5 Example

41: The examples will be replaced later by the `robstat` command. Possibly, it would be good to split the example into parts and include the parts at appropriate placed in the sections above.

As an illustrative example we will generate two datasets (of size $n = 1000$), one drawn from a standard normal distribution and one drawn from a chi-square distribution with one degree of freedom. All descriptive statistics presented above will be calculated for

both samples. To simplify interpretation we will present the excess kurtosis rather than the kurtosis (in other words, the reported tail heaviness statistics are equal to zero for the normal distribution). We then contaminate the datasets by replacing a random selection of 5% of the observations by value 5 for the normally distributed sample and by $F^{-1}_{\chi^2_1}(\Phi(5))$ for the chi-square distributed sample, $F^{-1}_{\chi^2_1}$ and Φ are the quantile function of the χ^2_1 distribution and the normal cumulative distribution function, respectively. In this way the degree of outlyingness is comparable between the two setups.

When we compare the classical, quantile-based and pairwise-based estimates obtained for the normally distributed dataset free of outliers (see the upper part of table 3.5), we do not see big differences between the three types of approaches. Indeed, the estimates all point towards a symmetrical distribution centered at zero with non-excessive tails and a dispersion of about one. If one looks at these statistics for the case of the chi-squared distributed data (see the lower part of table 3.5), the location estimate is, as expected, not the same since the mean is more attracted by the tail than the robust competitors. A similar phenomenon occurs for skewness and tails heaviness.

Table 3.5. The estimates of location, scale, skewness and tails heaviness in the original (uncontaminated) datasets

				Tails heaviness	
	Location	Scale	Skewness	Left	Right
Normally distributed sample					
Classical	0.015	0.977	−0.006	0.157	
Quantile-based	0.031	0.946	−0.088	0.025	0.0139
Pairwise-based	0.017	0.973	−0.024	0.038	0.075
Chi-square distributed sample					
Classical	0.993	1.406	2.322	6.323	
Quantile-based	0.414	0.908	2.491	−0.439	0.135
Pairwise-based	0.652	0.496	0.539	−0.491	0.190

When the dataset is contaminated by a small portion of outliers, the classical statistics change substantially, while the effect on their robust equivalent is only marginal (table 3.6). For example, for the normal case, classical statistics would point towards a right-tailed skewed distribution with relatively large dispersion and big-tail heaviness. The robust counterparts would still point towards the standard normal distribution. A similar phenomenon is observable for the chi-square.

```
set seed 1234
clear
set obs 1000
drawnorm z
gen x=invchi2(1,uniform())
```

Table 3.6. The estimates of location, scale, skewness and tails heaviness in the contaminated datasets

				Tails heaviness	
	Location	Scale	Skewness	Left	Right
Normally distributed sample					
Classical	0.263	1.447	1.530	3.465	
Quantile-based	0.086	1.012	0.016	0.023	0.055
Pairwise-based	0.109	1.072	0.041	0.030	0.145
Chi-square distributed sample					
Classical	1.011	1.416	2.311	6.276	
Quantile-based	0.437	0.923	2.305	−0.458	0.135
Pairwise-based	0.672	0.518	0.522	−0.492	0.189

```
**** UNCONTAMINATED NORMAL  ******
qui sum z, d
local meanN=r(mean)
local medianN=r(p50)
local sdN=r(sd)
local iqrN=(r(p75)-r(p25))*0.7413
local skewN=r(skewness)
local kurtN=r(kurtosis)-3
centile z, centile(25 50 75)
local qskewN=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)
qui hl z
local hlN=e(hl)
qui qn z
local qnN=e(qn)
qui medcouple z, lmc rmc
local mcN=e(mc)
local lmcN=e(lmc)
local rmcN=e(rmc)
qui centile z, centile(12.5 25 37.5 62.5 75 87.5)
local lqwN=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwN=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))
di in r "Location parameters Normal"
di "Mean: " `meanN'
di "Median: " `medianN'
di "Hodges-Lehman: " `hlN'
di in r"Scale parameters Normal"
di "Standard deviation: " `sdN'
di "Iqr: " `iqrN'
di "Qn: " `qnN'
di in r "Skewness parameters Normal"
di "Skewness: " `skewN'
di "Quantile skewness: " `qskewN'
```

```
di "Mecouple: " `mcN'

di in r "Tail heavyness parameters Normal"
di "Kurtosis: " `kurtN'
di "Quantile right heaviness: " `rqwN'
di "Quantile left heaviness: " `lqwN'
di "Right mecouple: " `rmcN'
di "Left mecouple: " `lmcN'

**** UNCONTAMINATED CHI2 *****

qui sum x, d
local meanCHI2=r(mean)
local medianCHI2=r(p50)

local sdCHI2=r(sd)
local iqrCHI2=(r(p75)-r(p25))*0.7413
local skewCHI2=r(skewness)
local kurtCHI2=r(kurtosis)-3

centile x, centile(25 50 75)
local qskewCHI2=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)

qui hl x
local hlCHI2=e(hl)

qui qn x
local qnCHI2=e(qn)

qui medcouple x, lmc rmc
local mcCHI2=e(mc)
local lmcCHI2=e(lmc)
local rmcCHI2=e(rmc)

qui centile x, centile(12.5 25 37.5 62.5 75 87.5)
local lqwCHI2=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwCHI2=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))

di in r "Location parameters CHI2"

di "Mean: " `meanCHI2'
di "Median: " `medianCHI2'
di "Hodges-Lehman: " `hlCHI2'

di in r"Scale parameters CHI2"
di "Standard deviation: " `sdCHI2'
di "Iqr: " `iqrCHI2'
di "Qn: " `qnCHI2'

di in r "Skewness parameters CHI2"
di "Skewness: " `skewCHI2'
di "Quantile skewness: " `qskewCHI2'
di "Mecouple: " `mcCHI2'

di in r "Tail heavyness parameters CHI2"
di "Kurtosis: " `kurtCHI2'
di "Quantile right heaviness: " `rqwCHI2'
di "Quantile left heaviness: " `lqwCHI2'
di "Right mecouple: " `rmcCHI2'
di "Left mecouple: " `lmcCHI2'

**** CONTAMINATED NORMAL  ******

replace z=5 in 1/50
qui sum z, d
local meanNc=r(mean)
local medianNc=r(p50)

local sdNc=r(sd)
local iqrNc=(r(p75)-r(p25))*0.7413
```

```
local skewNc=r(skewness)
local kurtNc=r(kurtosis)-3

centile z, centile(25 50 75)
local qskewNc=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)

qui hl z
local hlNc=e(hl)

qui qn z
local qnNc=e(qn)

qui medcouple z, lmc rmc
local mcNc=e(mc)
local lmcNc=e(lmc)
local rmcNc=e(rmc)

qui centile z, centile(12.5 25 37.5 62.5 75 87.5)
local lqwNc=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwNc=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))

di in r "Location parameters Normal"

di "Mean: " `meanNc'
di "Median: " `medianNc'
di "Hodges-Lehman: " `hlNc'

di in r"Scale parameters Normal"
di "Standard deviation: " `sdNc'
di "Iqr: " `iqrNc'
di "Qn: " `qnNc'

di in r "Skewness parameters Normal"
di "Skewness: " `skewNc'
di "Quantile skewness: " `qskewNc'
di "Mecouple: " `mcNc'

di in r "Tail heavyness parameters Normal"
di "Kurtosis: " `kurtNc'
di "Quantile right heaviness: " `rqwNc'
di "Quantile left heaviness: " `lqwNc'
di "Right mecouple: " `rmcNc'
di "Left mecouple: " `lmcNc'

**** CONTAMINATED CHI2 *****
replace x=invchi2(1,normal(5)) in 1/50
qui sum x, d
local meanCHI2c=r(mean)
local medianCHI2c=r(p50)

local sdCHI2c=r(sd)
local iqrCHI2c=(r(p75)-r(p25))*0.7413
local skewCHI2c=r(skewness)
local kurtCHI2c=r(kurtosis)-3

centile x, centile(25 50 75)
local qskewCHI2c=(r(c_1)+r(c_3)-2*r(c_2))/((r(c_2)-r(c_1))/1.349)

qui hl x
local hlCHI2c=e(hl)

qui qn x
local qnCHI2c=e(qn)

qui medcouple x, lmc rmc
local mcCHI2c=e(mc)
local lmcCHI2c=e(lmc)
local rmcCHI2c=e(rmc)

qui centile x, centile(12.5 25 37.5 62.5 75 87.5)
```

```
local lqwCHI2c=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqwCHI2c=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))

di in r "Location parameters CHI2c"

di "Mean: " `meanCHI2c'
di "Median: " `medianCHI2c'
di "Hodges-Lehman: " `hlCHI2c'

di in r"Scale parameters CHI2c"
di "Standard deviation: " `sdCHI2c'
di "Iqr: " `iqrCHI2c'
di "Qn: " `qnCHI2c'

di in r "Skewness parameters CHI2c"
di "Skewness: " `skewCHI2c'
di "Quantile skewness: " `qskewCHI2c'
di "Mecouple: " `mcCHI2c'

di in r "Tail heavyness parameters CHI2c"
di "Kurtosis: " `kurtCHI2c'
di "Quantile right heaviness: " `rqwCHI2c'
di "Quantile left heaviness: " `lqwCHI2c'
di "Right mecouple: " `rmcCHI2c'
di "Left mecouple: " `lmcCHI2c'
```

3.6 Robust tests of normality

The estimates of location, scale, skewness and tails heaviness can be used to characterize the underlying distribution. In particular, they can be used to test for normality. For example, Jarque and Bera (1980) have proposed a normality test relying on the classical skewness and kurtosis coefficients. More precisely, under the normality assumption ($\gamma_1 = 0$ and $\gamma_2 = 3$), we have

$$\sqrt{n}\begin{bmatrix} \gamma_{1;n} \\ \gamma_{2;n} - 3 \end{bmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix} \right)$$

which leads to the Jarque-Bera test statistic

$$T = n\left(\frac{\gamma_{1;n}^2}{6} + \frac{(\gamma_{2;n} - 3)^2}{24} \right) \approx \chi_2^2$$

The Jarque-Bera test is a very popular and interesting test for normality. It has been shown that, for a wide range of alternative distributions, it outperforms tests such as the Kolmogorov-Smirnov test, the Cramér-von Mises test and the Durbin test. Unfortunately, despite its good power properties and computational simplicity, the Jarque-Bera test is highly sensitive to outliers because it is constructed from the moment-based skewness and kurtosis measures.

Robust alternatives to the Jarque-Bera test have been proposed and studied in Brys et al. (2004b). The authors start from the fact that the Jarque-Bera test can be seen as a special case of the following general testing procedure. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_p)'$ be a vector of estimators of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ (a vector of characteristic parameters of the

underlying distribution) such that, under the null hypothesis of normality,

$$\sqrt{n}\Big(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\Big)' \xrightarrow{d} \mathcal{N}(\mathbf{0},\boldsymbol{\Omega})$$

Then, the general test consists in rejecting, at level α, the null hypothesis of normality if

$$T = n\Big(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\Big)'\boldsymbol{\Omega}^{-1}\Big(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\Big) > \chi^2_{p;1-\alpha}$$

where $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$-quantile of the chi-square distribution with p degrees of freedom. Brys et al. (2004b) then propose to use, in this general testing procedure, the robust skewness estimator MC_n or the tails heaviness estimators LMC_n and RMC_n.

Three tests have been studied. The first one is only based on the skewness estimator MC_n (the medcouple). In this case, $k=1$, $\widehat{\theta}=\text{MC}_n$ and $\Omega=1.25$. The second one is based on the left and right tail heaviness estimators LMC_n (left medcouple) and RMC_n (right medcouple). In this case, $k=2$, $\widehat{\boldsymbol{\theta}}=(\text{LMC}_n,\text{RMC}_n)'$, $\boldsymbol{\theta}=(0.199,0.199)'$ and

$$\boldsymbol{\Omega}=\begin{bmatrix} 2.62 & -0.0123 \\ -0.0123 & 2.62 \end{bmatrix}$$

The third test combines MC_n, LMC_n and RMC_n. In this case, $k=3$, $\widehat{\boldsymbol{\theta}}=(\text{MC}_n,\text{LMC}_n,\text{RMC}_n)'$, $\boldsymbol{\theta}=(0,0.199,0.199)'$ and

$$\boldsymbol{\Omega}=\begin{bmatrix} 1.25 & 0.323 & -0.323 \\ 0.323 & 2.62 & -0.0123 \\ -0.323 & -0.0123 & 2.62 \end{bmatrix}$$

This last test seems to have the best overall performance.

Example

42: The example will be revised to so that the relevant Stata code is visible.

We will analyze the body weight of 64 different animal species. The dataset we use is available online.[2] These data have been made available by Rice University, University of Houston Clear Lake and Tufts University.

To start the analysis, we first calculate the classic estimators of location, scale, skewness and kurtosis using Stata's `summary` command with the `detail` option. The results are presented in the first line of Table 3.7. In the second line of the table we present the results obtained using the commands for robust estimators (that is, `hl`, `qn`, `medcouple` and `medcouple` with the `lmc` and `rmc` options).

If we would only look at the classic estimators, we would conclude that the average animal weight is very high, but with a huge dispersion. The asymmetry is large and positive and tails are very heavy. When we look at the equivalent robust statistics, we see that the median weight is much lower than the mean weight. The robust dispersion

2. See http://onlinestatbook.com/stat_sim/transformations/body_weight.html

Table 3.7. Classic estimates of location, scale, skewness and tails heaviness as well as estimates based on pairwise combinations

	Location	Dispersion	Skewness	Tails
Classic	$\mu_n = 3{,}111{,}355$	$\sigma_n = 1.3 \times 10^7$	$\gamma_{1;n} = 5.461$	$\gamma_{2;n} = 32.77$
Robust	$Q_{0.5;n} = 3{,}500$ $\text{HL}_n = 94{,}307$	$Q_n = 6{,}667.5$	$\text{MC}_n = 0.985$	$\text{LMC}_n = -0.090$ $\text{RMC}_n = 0.915$

is also much smaller than that suggested by the standard deviation and right skewness is extreme. As far as the heaviness of the tails is concerned, the right tail is extremely heavy while the left one is similar to the left tail of the normal (even slightly lighter). When looking at the difference between classical and robust estimators, it is evident that outliers are present in the dataset.

A first way to tackle this problem is to transform the data to reduce the excessive importance of very big animals (such as dinosaurs). Given that weights are strictly positive, we consider a logarithmic transformation and redo the above descriptive statistics analysis (see Table 3.8).

Table 3.8. Classic estimates of location, scale, skewness and tails heaviness as well as estimates based on pairwise combinations based on transformed data

	Location	Dispersion	Skewness	Tails
Classic	$\mu_n = 9.313$	$\sigma_n = 4.135$	$\gamma_{1;n} = 0.304$	$\gamma_{2;n} = 2.192$
Robust	$Q_{0.5;n} = 8.161$ $\text{HL}_n = 9.289$	$Q_n = 4.281$	$\text{MC}_n = 0.386$	$\text{LMC}_n = 0.515$ $\text{RMC}_n = 0.241$

When we do this transformation, we see that the differences between classic and robust estimators become much smaller. Indeed the mean is only slightly larger than the median, the dispersion estimate is very similar for both methods as well as the skewness estimate that only points towards evidence of very moderate positive skewness. As far as the heaviness of the tails is concerned, the classic estimator is close to 3 which is the value of the kurtosis of the normal distribution and therefore points towards standard tails. Nevertheless when we look at the robust estimate for the latter, there is evidence of a heavy left tail. This last point is very important.

The classic and the robust tests for the normality of the log-transformed body weight variable lead to different findings. The standard Jarque-Bera statistic is 2.726, which is much smaller than the critical value of $\chi^2_{2;0.95} = 5.99$. That is, the standard Jarque-Bera test does not reject the null hypothesis of normality. On the other hand, the robust test statistic involving MC_n, LMC_n and RMC_n is equal to 9.266, which is larger than the critical value of $\chi^2_{3;0.95} = 7.815$. That is, the null hypothesis of normality is rejected

by the robust test. Even though the logarithmic transformation substantially reduces the effect of atypical observations, outliers still bias the classic test. In particular, we believe that the heaviness of the left tail is not satisfactorily identified by the classic kurtosis coefficient, and this affects the result of the normality test.

◁

3.7 Robust boxplots

As stated by Bruffaerts et al. (2014), among others, the boxplot is without any doubt the most commonly used tool to represent the distribution of the data and identify atypical observations in a univariate dataset. An observation is considered as atypical (or extreme) when it is above the upper whisker or below the lower whisker. An important issue with the standard boxplot is that, as soon as asymmetry or tail heaviness appears, the percentage of values identified as atypical becomes excessive. To cope with this, Hubert and Vandervieren (2008) proposed an *adjusted* boxplot for skewed data. Their idea is to modify the whiskers according to the degree of asymmetry in the data, which can be robustly measured by the medcouple. Alternatively, Bruffaerts et al. (2014) propose to apply a simple rank-preserving transformation on the original data so that the transformed observations can be adjusted by a so-called *Tukey g-and-h distribution*. Using the quantiles of this distribution, it is then relatively easy to recover whiskers of the boxplot related to the original data. Given the result of simulations, the latter seems to be more efficient and we therefore concentrate on that too here.

3.7.1 The classic boxplot and the adjusted boxplot

In a univariate setup, an observation is often considered as atypical as soon as its value does not belong to the interval $[Q_{0.25} - 1.5\,\text{IQR}; Q_{0.75} + 1.5\,\text{IQR}]$, where $Q_{0.25}$ and $Q_{0.75}$ are the first and third quartiles, and IQR is the interquartile range. For Gaussian data, approximately 0.7% of the observations will lie outside this interval. Unfortunately, as soon as asymmetry or tail heaviness appears, the percentage of values detected as atypical becomes excessively high. To deal with the above drawbacks of the standard boxplot, Hubert and Vandervieren (2008) suggested to use an alternative boxplot, called the adjusted boxplot, where the interval for the boxplot is

$$\left[Q_{0.25} - 1.5e^{-4\,\text{MC}}\text{IQR},\, Q_{0.75} + 1.5e^{3\,\text{MC}}\text{IQR}\right] \quad \text{if MC} \geq 0$$

and

$$\left[Q_{0.25} - 1.5e^{-3\,\text{MC}}\text{IQR},\, Q_{0.75} + 1.5e^{4\,\text{MC}}\text{IQR}\right] \quad \text{if MC} < 0$$

where MC is the medcouple.

Although this rule works well for most commonly used distributions, it presents some limitations and drawbacks (see Bruffaerts et al. 2014), the most restrictive probably being that it does not deal with excessive tail heaviness. Bruffaerts et al. (2014) deal with most of these limitations. Since this method relies on the Tukey g-and-h distribution,

we briefly describe this distribution before explaining the methodology of Bruffaerts et al. (2014).

3.7.2 The Tukey g-and-h distribution

The Tukey g-and-h family of distributions covers a large variety of distributions which can substantially differ from normality in both skewness and heaviness of the tails. If Z is a random variable with standard normal distribution, and g and h are two constants ($g \neq 0$, $h \in \mathbb{R}$), then the random variable Y given by

$$Y = \frac{1}{g}(\exp(gZ) - 1)\exp(hZ^2/2)$$

is distributed as a Tukey g-and-h distribution, that is $Y \sim T(g, h)$.[3] The constants g and h control the skewness and the tail weight (or elongation) of the distribution, respectively. They can be estimated from the empirical quantiles[4] $Q_{1-p}(\{y_j\})$ and $Q_p(\{y_j\})$ of order $(1-p)$ and p ($0.5 < p < 1$) of n independent realizations $\{y_1, \ldots, y_n\}$ of Y (see Jiménez and Arunachalam 2011). Following Jiménez and Arunachalam (2011), g and h can be estimated as follows:

43: Why such complicated notation?

$$\widehat{g} = \frac{1}{z_p}\ln\left(-\frac{Q_p(\{y_j\})}{Q_{1-p}(\{y_j\})}\right) \quad \text{and} \quad \widehat{h} = \frac{2\ln\left(-\widehat{g}\,\frac{Q_p(\{y_j\})Q_{1-p}(\{y_j\})}{Q_p(\{y_j\})+Q_{1-p}(\{y_j\})}\right)}{z_p^2} \tag{3.1}$$

where z_p is the quantile of order p of the standard normal distribution. Note that the chosen order p for the quantiles determines the robustness of the method with respect to outliers. For example, if we set $p = 0.9$, the breakdown point of the estimators of g and h is set to $1-p = 10\%$, as the method provides meaningful results if there are up to 10% of outliers. For estimation purposes, we suggest using $p = 0.9$, except if one believes that the contamination rate is larger than 10%. Working with a lower value for p would increase robustness with respect to outliers, but at the cost of lowering the efficiency. Furthermore, it would not make sense to work with a $p \leq 0.75$, as a contamination of more than 25% would make the first and/or third quartile of the boxplot break down.

3.7.3 A generalized boxplot

The method proposed by Bruffaerts et al. (2014) overcomes most of the limitations of the adjusted boxplot. Based on an initial dataset $\{x_1, \ldots, x_n\}$, their procedure is as follows.

3. Note that Y is a strictly increasing transformation of Z, driven by the values of g and h. Hence, for every order $p \in (0, 1)$, $y_p = \frac{1}{g}(\exp(gz_p) - 1)\exp(hz_p^2/2)$, where y_p and z_p are the quantiles of order p of the distributions of Y and Z respectively. This implies in particular that the median $y_{0.5}$ of Y is equal to zero.
4. The expression $Q_p(\{y_j\})$ denotes the empirical quantile of order p related to the series $\{y_1, \ldots, y_n\}$. The notation $\min(\{x_j\})$ and $\max(\{x_j\})$ is later used to define the minimum and maximum values of the series $\{x_1, \ldots, x_n\}$.

1. Reduce the data by a scale factor s_0:

$$x_i^* = \frac{x_i}{s_0}, \quad i = 1, \ldots, n$$

with $s_0 = \text{IQR}(\{x_j\})$, where $\text{IQR}(\{x_j\})$ is the interquartile range of the series $\{x_1, \ldots, x_n\}$.

2. Shift the dataset to obtain only strictly positive values: compute

$$r_i = x_i^* - \min(\{x_j^*\}) + \zeta, \quad i = 1, \ldots, n$$

where $\zeta > 0$ is a small quantity. They propose to use $\zeta = 0.1$

3. Standardize the values obtained in step 2 in order to obtain new values belonging to the open interval $(0, 1)$: compute

$$\widetilde{r}_i = \frac{r_i}{\min(\{r_j\}) + \max(\{r_j\})}, \quad i = 1, \ldots, n$$

4. Apply the inverse normal (probit) transformation

$$w_i = \Phi^{-1}(\widetilde{r}_i), \quad i = 1, \ldots, n$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal.

5. Center and reduce the values w_i: compute

$$w_i^* = \frac{w_i - Q_{0.5}(\{w_j\})}{\text{IQR}(\{w_j\})/1.3426}, \quad i = 1, \ldots, n$$

where $Q_{0.5}(\{w_j\})$ and $\text{IQR}(\{w_j\})$ are the median and the interquartile range of the series $\{w_1, \ldots, w_n\}$. The constant 1.3426 ensures, in the Gaussian case, the consistency of the scale estimator $\text{IQR}(\{x_j\})$ with the scale parameter σ (the standard deviation).

6. Adjust the distribution of the values w_i^*, $i = 1, \ldots, n$, by the Tukey $T(\widehat{g}^*, \widehat{h}^*)$ distribution, where $\widehat{g}^*$ and $\widehat{h}^*$ are the estimates of the skewness and tail weight parameters g and h obtained by applying equation (3.1) to the empirical quantiles $Q_{0.1}(\{w_j^*\})$ and $Q_{0.9}(\{w_j^*\})$ of orders 0.1 and 0.9 of the series $\{w_1^*, \ldots, w_n^*\}$.

44: I don't understand. What is exactly done in this step?

7. Determine the quantiles $\xi^*_{\alpha/2}$ and $\xi^*_{1-\alpha/2}$ of orders $\alpha/2$ and $1 - \alpha/2$, $\alpha \in (0, 1)$, of the $T_{\widehat{g}^*, \widehat{h}^*}$ distribution specified in the previous step, where α corresponds to the desired detection rate of atypical values in the absence of contamination with outliers. Let

$$\mathcal{I} = \left\{ i = 1, \ldots, n \middle| w_i^* \notin \left[\xi^*_{\alpha/2}, \xi^*_{1-\alpha/2} \right] \right\}$$

be the set of indices of the values w_i^* that are detected as atypical in the series $\{w_1^*, \ldots, w_n^*\}$. The values x_i for which $i \in \mathcal{I}$ are considered as atypical observations in the initial dataset.

8. Based on the above steps, it is possible to come up with a *generalized* boxplot which is associated to the original dataset. From the detection bounds $L_-^* = \xi_{\alpha/2}^*$ and $L_+^* = \xi_{1-\alpha/2}^*$ computed in step 7, one can build the respective detection bounds B_-^* and B_+^* for the related original dataset (B_-^* and B_+^* are the extremities of the lower and upper whiskers of the generalized boxplot):

$$B_\pm^* = \left(\Phi\left(Q_{0.5}(\{w_j\}) + \frac{\text{IQR}(\{w_j\})}{1.3426} L_\pm^*\right) \times \{\min(\{r_j\}) + \max(\{r_j\})\} + \min(\{x_j^*\}) - \zeta\right) s_0$$

Example

We illustrate the use of the robust boxplot by an example from Bruffaerts et al. 2014. The data contain daily earnings (in British pounds) of 50 top soccer players.[5] The left panel in Figure 3.5 displays the kernel density estimate of the earnings variable.[6] The right panel displays the three flavors of boxplots. Daily earnings appear to be slightly asymmetrically distributed with a relatively heavy tail. A medcouple measure of 0.12 indicates that the distribution is slightly asymmetric, but not too much. This explains why the upper whisker of the generalized boxplot goes beyond the upper whisker of the two other boxplots.

◁

5. Source: http://www.paywizard.co.uk/main/pay/vip-celebrity-salary/football-players-s alary
6. We use an Epanechnikov kernel with Silverman's rule-of-thumb bandwidth.

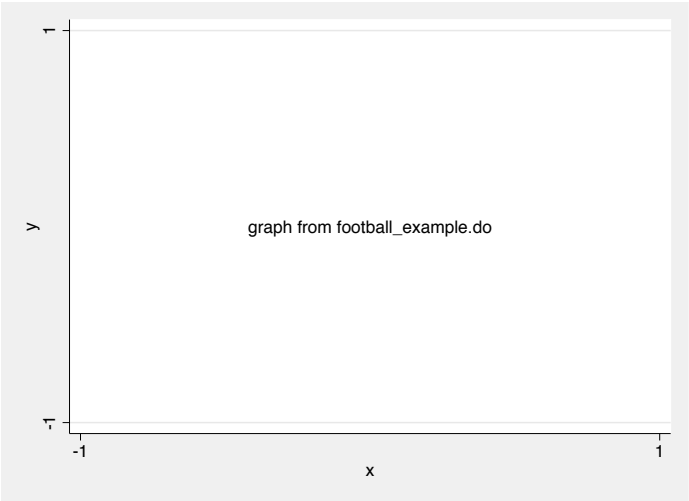

Figure 3.5. Classic, adjusted, and generalized boxplot [will be inserted after revising the boxplot program]

Part III

Robust regression

4 Robust linear regression

This chapter is devoted to the estimation of the parameters of linear regression models. Let us first precise some notations.

4.1 The linear regression model

In a linear regression model, we try to explain a variable y—the *dependent* variable—as a linear function of some *explanatory* variables (or *predictors*) $x_1, \ldots, x_p$: we assume that

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{4.1}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown regression coefficients—β_0 is called the *intercept* and $\beta_1, \ldots, \beta_p$ are the *slopes*—that have to be estimated and ε is a random error term (the error of the statistical model, due to omitted factors, errors of measurement, random effects, etc.).

To estimate the regression coefficients, we need a random sample of realizations of $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$, where n is the sample size. We have

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{4.2}$$

where ε_i's are generally assumed to be i.i.d. random variables. That is,

$$\varepsilon_i \overset{\text{i.i.d.}}{\sim} F_{0,\sigma}$$

where distribution $F_{0,\sigma}$ has a location (centrality) parameter equal to zero and a scale parameter equal to σ.

Denoting by $\mathbf{x}_i$ and $\boldsymbol{\beta}$ the $(p+1)$ dimensional column vectors with coordinates $(1, x_{i1}, \ldots, x_{ip})$ and $(\beta_0, \beta_1, \ldots, \beta_p)$, respectively, equation (4.2) can be more compactly written as

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{4.3}$$

Furthermore, letting $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$ and

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ 1 & x_{21} & \ldots & x_{2p} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

equations (4.3) takes the matrix-notation form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Let us note here that, *conditionally to the predictors*, the linear regression model (4.3) may be considered as a location-scale model. Indeed, under the assumption of homoscedasticity—an identical scale parameter σ for each error term ε_i—regression model (4.3) may be formulated as follows:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \sigma\nu_i, \qquad i = 1, \ldots, n \tag{4.4}$$

where the ν_i's are i.i.d. with distribution function $F_{0,1}$ (a distribution with a location parameter equal to zero and a scale parameter equal to one). In this case, the conditional distribution of y_i given $\mathbf{x}_i$ is of the form:

$$F_{y_i|\mathbf{x}_i}(y) = \Pr(y_i \leq y|\mathbf{x}_i) = \Pr\left(\nu_i \leq \frac{y - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\middle|\mathbf{x}_i\right) = F_{0,1}\left(\frac{y - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \tag{4.5}$$

Furthermore, if $f_{0,1}$ denotes the density function of the error terms ν_i, that is,

$$f_{0,1}(u) = \frac{dF_{0,1}(u)}{du} = F_{0,1}'(u)$$

then

$$f_{y_i|\mathbf{x}_i}(y) = \frac{1}{\sigma} f_{0,1}\left(\frac{y - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \tag{4.6}$$

Hence, $\mathbf{x}_i'\boldsymbol{\beta}$ corresponds to the unknown location parameter of the distribution of y_i and σ is the scale parameter of the distribution of y_i. For simplicity, we will consider that the distribution $F_{0,1}$ is continuous and symmetric around zero (and exception is Section 4.6).

❑ Remark

The classic *location-scale model* may be seen as a particular case of regression model (4.4). It suffices to set $\beta_1 = \cdots = \beta_p = 0$ and to assume that the ν_i's are i.i.d. with distribution $F_{0,1}$ (such that $E(\nu_i) = 0$). In this case, the observations

$$y_i = \beta_0 + \sigma\nu_i, \qquad i = 1, \ldots, n, \tag{4.7}$$

are i.i.d. with a common distribution F characterized by mean $\mu = \beta_0$ and scale parameter σ.

❑

❑ Remark

Most textbook presentations of the linear regression model assume the explanatory variables to be fixed (and measured without error). That is, the explanatory variables

are not assumed to be random variables. In the context of a *designed experiment*, this assumption is reasonable since the values of the experimental factors are determined *a priori* by the researchers. In other contexts such as, for example, when using social-science survey data, the assumption makes no sense. [Say here what the consequence is: The fact that the X's are random doesn't really change anything. (unlike measurement error which attenuates the estimates)]

Nonetheless, since we focus on the problem of outlying values, we will ignore the issue in this chapter and consider x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$, as predetermined. That is, results will always be conditional on the particular *values* taken by the explanatory variables.

I'm not sure whether this remark makes sense. First, LS results are valid also if the X's are random. Second, if we talk about X outliers it makes not much sense to assume X fixed.

❑

4.2 Different types of outliers

Model (4.1) assume that *all* units of the population and, *de facto*, all units of the sample are consistent with the supposed linear model. If a unit has a behavior that does not respect the underlying theoretical model, we define it as an *outlying* unit with respect to the model.

Of course, in the case of *simple* linear regression model ($p = 1$), a visual inspection of the scatterplot is generally sufficient to detect the outliers. But, when the number of explanatory variables is greater than two, it becomes impossible to visualize all the data set and the use of robust methods to estimate the regression parameters is then essential. More precisely, we aim at developing procedures that provide a good fit to the bulk of the data without being perturbed by a small proportion of outliers, and that do not require deciding previously which observations are outliers. Moreover, the comparison between the estimations provided by the classical least squares estimator and those obtained using a robust estimation procedure will allow to bring to the fore the outlyingness of some data.

In cross-sectional regression analysis, three types of outliers may influence the estimations. Rousseeuw and Leroy (1987) define them as *vertical outliers*, *good leverage points* and *bad leverage points*. To illustrate this terminology, consider a simple linear regression as shown in figure 4.1 (the generalization to higher dimensions is straightforward). *Vertical outliers* are those observations that have outlying values for the corresponding error term (that is, in the y-dimension) but are not outlying in the space of explanatory variables (in the x-dimension). *Good leverage points* are observations that are outlying in the space of explanatory variables but that are located close to the regression hyperplane. Finally, *bad leverage points* are observations that are both outlying in the space of explanatory variables and located far from the true regression hyperplane.

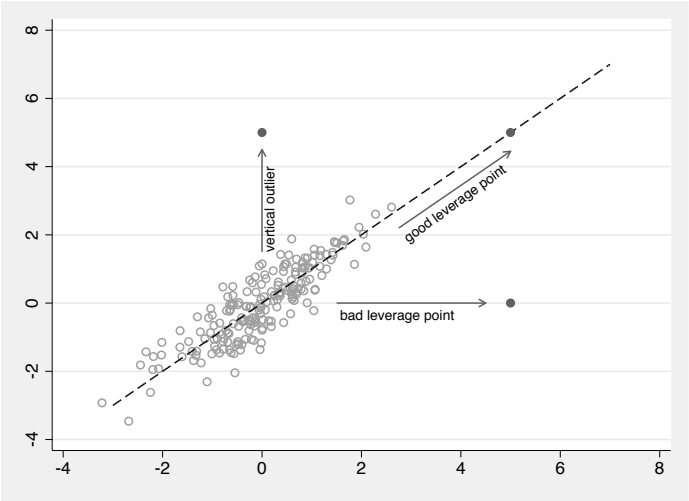

Figure 4.1. Vertical outlier, good leverage point and bad leverage point

All these types of outliers risk to affect the estimation of the regression hyperplane but their effect changes according to the estimator we will consider and the type of outlyingness. For the classical least squares estimation method, for instance, the bad leverage points are considered as the most dangerous outliers because their presence can change the sign of the slope of the regression line (in simple regression); the good leverage points have little influence on the estimation of the regression coefficients but they have an impact on the variances and covariances of the regression coefficients' estimators and, consequently, risk to influence the inferential procedures (tests and confidence intervals).

The most popular estimation method in linear regression is certainly the *least squares* (LS) method introduced in 1805 by Legendre [Citation?]. One of its principal advantage is the simplicity of the computation of the LS estimates. Its popularity has also be reinforced by the fact that, under the normality of the error terms, LS estimates of the regression coefficients coincide with the maximum likelihood estimates. We will first briefly review the logic behind least squares (LS) estimation and recall why the LS estimator is particularly affected by the presence of atypical individuals. We will thereafter introduce some alternative estimation methods that have been proposed to try to cope with outliers.

4.3 LS estimation

Let us denote by $\widehat{y}_i(\boldsymbol{\beta})$ the value fitted by the regression model for the ith statistical unit of the sample when taking $\boldsymbol{\beta}$ as value for the vector of regression coefficients:

$$\widehat{y}_i(\boldsymbol{\beta}) = \mathbf{x}_i'\boldsymbol{\beta}, \qquad i = 1, \ldots, n$$

The difference between the observed value y_i and the fitted value $\widehat{y}_i(\boldsymbol{\beta})$ is the residual $r_i(\boldsymbol{\beta})$:

$$r_i(\boldsymbol{\beta}) = y_i - \widehat{y}_i(\boldsymbol{\beta}), \qquad i = 1, \ldots, n$$

Although $\boldsymbol{\beta}$ can be estimated in several ways, the underlying idea is often to take an estimate $\widehat{\boldsymbol{\beta}}$ in such a way that the fitted values $\widehat{y}_i(\widehat{\boldsymbol{\beta}})$ for the dependent variable are as close as possible to the observed values y_i $(i = 1, \ldots, n)$, i.e., in such a way that we minimize globally the magnitude of the residuals $r_i(\widehat{\boldsymbol{\beta}})$. This idea leads to try to find the estimate $\widehat{\boldsymbol{\beta}}$ that minimizes a specific aggregate prediction error.

In the case of the well-known ordinary least squares (LS), this aggregate prediction error is defined as the sum of squared residuals:

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) \tag{4.8}$$

where "arg min" stands for "the value minimizing". In other terms, $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is solution of the so called *normal equations* system — we will also call it the *estimating equations* system — obtained by differentiating the function $\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta})$ to minimize with respect to each component of $\boldsymbol{\beta}$, that is, $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is the solution of

$$\sum_{i=1}^{n} r_i(\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0} \tag{4.9}$$

which is equivalent to the linear equations system

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}.$$

If $\mathbf{X}$ has full rank[1], then the solution of (4.9) is unique and is given by

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{4.10}$$

This estimate can be computed in Stata using the `regress` command (see [R] **regress**).

Note here that, if the model contains a constant term β_0, that is, if the first component of the vectors $\mathbf{x}_i$, $i = 1, \ldots, n$, is equal to one, it follows from (4.9) that the residuals $r_i(\widehat{\boldsymbol{\beta}}_{\text{LS}})$, $i = 1, \ldots, n$, have zero average.

1. The matrix of predictors $\mathbf{X}$ is said to have *full rank* if its columns are linearly independent (absence of multicollinearity), that is, if $\mathbf{Xa} \neq \mathbf{0}$ for all $\mathbf{a} \neq \mathbf{0}$. This is equivalent to the nonsingularity of $\mathbf{X}'\mathbf{X}$.

It is easy to verify that the LS estimator satisfies (see Maronna et al. 2006, 92)

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\boldsymbol{\gamma}) = \widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) + \boldsymbol{\gamma} \qquad \text{for all } \boldsymbol{\gamma} \in \mathbb{R}^{p+1} \tag{4.11}$$

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \lambda\mathbf{y}) = \lambda\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) \qquad \text{for all } \lambda \in \mathbb{R} \tag{4.12}$$

and, for any nonsingular $(p+1) \times (p+1)$ matrix $\mathbf{A}$,

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}\mathbf{A}, \mathbf{y}) = \mathbf{A}^{-1}\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) \tag{4.13}$$

The properties (4.11), (4.12) and (4.13) are called *regression*, *scale* and *affine equivariance* of $\widehat{\boldsymbol{\beta}}_{\text{LS}}$, respectively. In the sequence, it will be desirable that every other estimator of $\boldsymbol{\beta}$ also satisfies these natural properties.

It is also well known that the LS estimator of $\boldsymbol{\beta}$ coincides with the maximum likelihood estimator in case of normally distributed error terms in (4.2). Hence, $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is the most efficient estimator of $\boldsymbol{\beta}$ in the Gaussian regression model.

However, an important drawback of LS is that, by considering squared residuals, it tends to award an excessive importance to observations with large residuals and, consequently, distort parameters estimation when outliers exist.

4.4 M estimation

4.4.1 L_1 or Least Absolute Deviation (LAD) estimation

Edgeworth (1887) realized that due to the squaring of the residuals, LS becomes extremely vulnerable to the presence of outliers. To cope with this, he proposed a method consisting in minimizing the sum of the absolute values of the residuals rather than the sum of their squares. More precisely, his method defines the L_1 or *least absolute deviation* (LAD) estimate as

$$\widehat{\boldsymbol{\beta}}_{\text{LAD}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| \tag{4.14}$$

This estimate is solution of the estimating equations system obtained by differentiating the sum of the absolute values of the residuals with respect to each component of $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} \text{sign}(r_i(\widehat{\boldsymbol{\beta}}_{\text{LAD}}))\mathbf{x}_i = \mathbf{0} \tag{4.15}$$

If the model contains an intercept term, (4.15) implies that the residuals $r_i(\widehat{\boldsymbol{\beta}}_{\text{LAD}})$, $i = 1, \ldots, n$, have a median equal to zero; this motivates the fact that the LAD regression estimator is also sometimes called the *median regression estimator*.

Unlike for $\widehat{\boldsymbol{\beta}}_{\text{LS}}$, there is no explicit expression for $\widehat{\boldsymbol{\beta}}_{\text{LAD}}$.[2] However, there exist very

2. Note also that the LAD estimate of $\boldsymbol{\beta}$ may not be unique and has the property that at least $(p+1)$ residuals are equal to zero.

fast algorithms to compute it and $\widehat{\boldsymbol{\beta}}_{\text{LAD}}$ is available in Stata via the `qreg` command as a standard function (see [R] **qreg**).

Finally, it can easily be seen from (4.14) and (4.15) that this estimator does protect against vertical outliers (but not against bad leverage points). However, this gain in robustness with respect to the LS estimator comes with an important loss of efficiency: the asymptotic relative efficiency of $\widehat{\boldsymbol{\beta}}_{\text{LAD}}$ with respect to $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ is equal to $2/\pi = 63.7\%$ at a Gaussian error distribution (see Huber 1981).

4.4.2 The principle of M estimation

Huber (1964) hence generalized median regression to a wider class of estimators, called M estimators, by considering other functions than the absolute value in (4.14) in order to find a reasonable balance between robustness and Gaussian efficiency.

An M estimate $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ of $\boldsymbol{\beta}$ is defined by

$$\widehat{\boldsymbol{\beta}}_{\text{M};\rho} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\widehat{\sigma}}\right) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) \tag{4.16}$$

where $\rho(u)$ is a loss function that is positive, even such that $\rho(0) = 0$, and non decreasing for positive values u, and $\widehat{\sigma}$ is an auxiliary estimate of the scale parameter σ required to standardize the residuals and to make $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ scale equivariant; see (4.12). In most situations, $\widehat{\sigma}$ is computed in advance, but it can also be computed simultaneously through a scale M estimating equation. This problem will be discussed in more details later.

❑ Remark

The LS estimate and the LAD estimate correspond respectively to $\rho(u) = u^2$ and $\rho(u) = |u|$. In these two cases, $\widehat{\sigma}$ becomes a constant factor outside the summation sign in (4.16) and

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta}))$$

Thus neither the LS nor the LAD estimate require an auxiliary scale estimate.

❑

Of course, if we want a M estimator more robust against vertical outliers than the LS estimator, we have to take a loss function ρ that is less rapidly increasing than the square function in order to give less weight to big (in absolute value) residuals in the minimization problem. In order to combine robustness and efficiency under a Gaussian error distribution, Huber (1964) has suggested to use for ρ a function of the form (see figure 4.2):

$$\rho_{\kappa}^{\text{H}}(u) = \begin{cases} u^2 & \text{if } |u| \leq \kappa \\ 2\kappa|u| - \kappa^2 & \text{if } |u| > \kappa \end{cases}$$

where κ is a constant determining the trade-off between robustness and efficiency. These functions of Huber are convex on the whole real line and may be seen as intermediate functions between the quadratic function (leading to the non robust but efficient LS estimate) and the absolute value function (associated with the robust but poorly efficient LAD estimate).

Another class of loss functions ρ widely used in the literature is the class of the Tukey-Biweight [Citation?] functions (see figure 4.3)

$$\rho_\kappa^{\mathrm{B}}(u) = \begin{cases} \frac{\kappa^2}{6}\left[1 - \left(1 - \left(\frac{u}{\kappa}\right)^2\right)^3\right] & \text{if } |u| \leq \kappa \\ \frac{\kappa^2}{6} & \text{if } |u| > \kappa \end{cases} \tag{4.17}$$

These functions are bounded. Once again, the constant κ allows the trade-off between robustness and Gaussian efficiency. We will show the advantage and disadvantage to use a bounded function ρ hereafter.

We may also characterize $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ as a solution of the estimating equations system obtained by differentiating the function to minimize in (4.16) with respect to each component of $\boldsymbol{\beta}$, that is, as a solution of the equations system

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0} \tag{4.18}$$

where $\psi(u) = d\rho(u)/du = \rho'(u)$. For instance, taking $\rho(u) = \rho_\kappa^{\mathrm{H}}(u)$, we have

$$\psi_\kappa^{\mathrm{H}}(u) = \begin{cases} -2\kappa & \text{if } u < -\kappa \\ 2u & \text{if } -\kappa \leq u \leq \kappa \\ 2\kappa & \text{if } u > \kappa \end{cases}$$

for $\rho(u) = \rho_\kappa^{\mathrm{B}}(u)$, we obtain

$$\psi_\kappa^{\mathrm{B}}(u) = \begin{cases} u\left(1 - \left(\frac{u}{\kappa}\right)^2\right)^2 & \text{if } |u| \leq \kappa \\ 0 & \text{if } |u| > \kappa \end{cases}$$

(see figures 4.2 and 4.3). If the loss function ρ is convex on $\mathbb{R}$—this is the case for ρ_κ^{H}—the score function ψ is monotone (non decreasing) on $\mathbb{R}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ is called a *monotone* regression M estimator; if ρ is bounded—this is the case for ρ_κ^{B}—the score function ψ vanishes out of a certain interval of $\mathbb{R}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ is then called a *redescending* regression M estimator.

The main advantage of monotone score functions ψ is that all solutions of (4.18) are solutions of (4.16). In the case of redescending score functions ψ, the estimating equations (4.18) may have multiple solutions corresponding to multiple local minima of $\sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})/\widehat{\sigma})$, and generally only one of them (the "good" solution) corresponds to the global minimizer $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ defined by (4.16), which makes the computation of the M estimate considerably more complex.

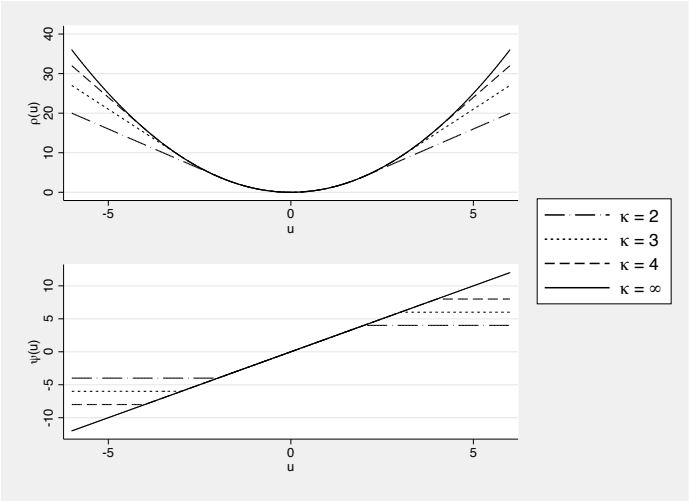

Figure 4.2. Huber loss function ρ_κ^{H} and score function ψ_κ^{H}

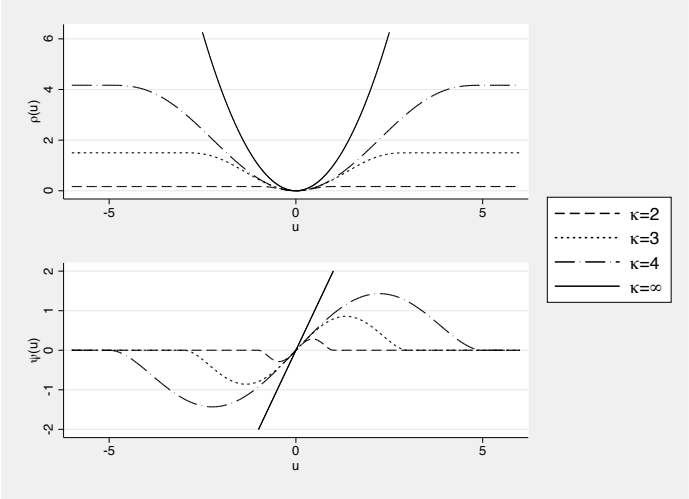

Figure 4.3. Tukey-Biweight loss function ρ_κ^{B} and score function ψ_κ^{B}

❑ Remark

Applying the M estimation procedure in the particular case of the location-scale model (4.7) leads to the M *estimators of location and scale*—we just mentioned the existence of these estimators in the previous chapter. The interested reader will find some results relative to these specific estimators in appendix 4.8 at the end of this chapter.

❑

4.4.3 M estimation as a generalization of maximum likelihood (ML) estimation

The M estimation as defined above may be seen, as already explained, as a generalization of the LS or LAD estimation, but also as a generalization of the maximum-likelihood (ML) estimation (see, for instance, Maronna et al. 2006). Indeed, assuming model (4.4) with fixed $\mathbf{x}_i$ and with ν_i, $i = 1, \ldots, n$, i.i.d. of density $f_{0,1}$, the likelihood of the sample $\{y_1, \ldots, y_n\}$ is given by

45: Couldn't we also just use unspecified f instead of $f_{0,1}$?

$$\frac{1}{\sigma^n} \prod_{i=1}^{n} f_{0,1}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)$$

Hence, maximum likelihood estimation of the parameters $\boldsymbol{\beta}$ and σ consists in looking for

$$\begin{aligned}\left(\widehat{\boldsymbol{\beta}}_{\text{ML}}', \widehat{\sigma}_{\text{ML}}\right)' &= \arg\max_{\boldsymbol{\beta},\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_{0,1}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) \\ &= \arg\max_{\boldsymbol{\beta},\sigma} \left[\sum_{i=1}^{n} \ln f_{0,1}\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right) - n \ln \sigma\right] \\ &= \arg\min_{\boldsymbol{\beta},\sigma} \left[\sum_{i=1}^{n} \rho_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) + n \ln \sigma\right] \end{aligned} \tag{4.19}$$

where $\rho_{\text{ML}}(u) = -\ln f_{0,1}(u)$. If σ is known, the minimization problem simply becomes

$$\widehat{\boldsymbol{\beta}}_{\text{ML}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) \tag{4.20}$$

and $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ is solution of the estimating equations system

$$\sum_{i=1}^{n} \psi_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right)\mathbf{x}_i = \mathbf{0}$$

where $\psi_{\text{ML}}(u) = \rho'_{\text{ML}}(u) = -(1/f_{0,1}(u))f'_{0,1}(u)$. If $f_{0,1}$ is the standard normal density function, $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ coincides with $\widehat{\boldsymbol{\beta}}_{\text{LS}}$. If $f_{0,1}$ is the density function of the Laplace distribution, that is, if $f_{0,1}(u) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|u|)$, $u \in \mathbb{R}$, then $\widehat{\boldsymbol{\beta}}_{\text{ML}}$ is equal to $\widehat{\boldsymbol{\beta}}_{\text{LAD}}$.

If σ is not known but is estimated beforehand and fixed in (4.20), the estimating equations system becomes

$$\sum_{i=1}^{n} \psi_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0}$$

Note that, if $\boldsymbol{\beta}$ and σ are estimated simultaneously, the estimating equations system related to (4.19) is

$$\begin{cases} \sum_{i=1}^{n} \psi_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right)\mathbf{x}_i = \mathbf{0} \\ \frac{1}{n}\sum_{i=1}^{n} \rho_{\text{ML;scale}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta \end{cases} \tag{4.21}$$

where $\rho_{\text{ML;scale}}(u) = u\psi_{\text{ML}}(u)$ and $\delta = 1$.

4.4.4 Practical implementation of M estimates

Let us first assume, for simplicity, that the scale parameter σ is known. In that case, the regression M-estimate $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ is solution of the estimating equations system (4.18) where $\widehat{\sigma}$ is replaced by σ. Defining the weight function w by

$$w(u) = \begin{cases} \frac{\psi(u)}{u} & \text{if } u \neq 0 \\ \psi'(0) & \text{if } u = 0 \end{cases}$$

the estimating system (4.18) can be rewritten as

$$\sum_{i=1}^{n} w_i r_i(\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0} \tag{4.22}$$

where $w_i = w(r_i(\boldsymbol{\beta})/\sigma)$. Hence, the equations to solve in the M estimation procedure appear as *weighted* versions of the normal equations (4.9) related to LS estimation, and if the w_i's were known, the equations (4.22) could be solved by applying LS to $\sqrt{w_i}y_i$ and $\sqrt{w_i}\mathbf{x}_i$. But the weights w_i are functions of $\boldsymbol{\beta}$ and depend upon the data, and hence are not known. So we have to use an iterative procedure. Using an initial estimate $\widehat{\boldsymbol{\beta}}_0$ for $\boldsymbol{\beta}$ (for instance, the LAD estimate of $\boldsymbol{\beta}$), the weights can be computed and serve as the start of an *iteratively reweighted least squares algorithm* (IRWLS). Note however that the latter is guaranteed to converge to the global minimum of (4.16) only if the loss function ρ is convex on the whole real line $\mathbb{R}$ (which is the case for the ρ_c^{H} functions introduced by Huber).[3]

If σ is not known, it can be estimated (in a robust way) beforehand using the residuals $r_i(\widehat{\boldsymbol{\beta}}_0)$, $i = 1, \ldots, n$, and then fixed in the iterative procedure described above. It is of course also possible to estimate simultaneously $\boldsymbol{\beta}$ and σ in this procedure, by updating $\widehat{\sigma}$ at each iteration (see Maronna et al. 2006 for more details).

Regression M estimate with preliminary scale estimation

In practice, we may take the LAD estimate as initial estimate $\widehat{\boldsymbol{\beta}}_0$ for $\boldsymbol{\beta}$ (recall that the LAD estimate does not require estimating a scale). Then we may estimate σ using

3. In the case of a convex loss function ρ, the convergence of the algorithm to the global minimum of (4.16) is guaranteed whatever the starting point $\widehat{\boldsymbol{\beta}}_0$.

normalized MAD of the residuals $r_i(\widehat{\boldsymbol{\beta}}_0)$. More precisely, we may take

$$\widehat{\sigma} = 1.4826 \cdot \operatorname*{med}_i \Big(\Big| r_i(\widehat{\boldsymbol{\beta}}_0) \Big| ; r_i(\widehat{\boldsymbol{\beta}}_0) \neq 0 \Big)$$

The reason for using only *non null* residuals is that, since at least $(p+1)$ residuals $r_i(\widehat{\boldsymbol{\beta}}_0) = r_i(\widehat{\boldsymbol{\beta}}_{\text{LAD}})$ are equal to zero, determining the MAD of the n residuals could lead to underestimating σ when p is large.

Since $\widehat{\boldsymbol{\beta}}_{\text{LAD}}$ is regression, scale and affine equivariant, it is easy to show that

$$\begin{aligned} \widehat{\sigma}(\mathbf{X}, \mathbf{y} + \mathbf{X}\boldsymbol{\gamma}) &= \widehat{\sigma}(\mathbf{X}, \mathbf{y}) \qquad \text{for all } \boldsymbol{\gamma} \in \mathbb{R}^{p+1} \\ \widehat{\sigma}(\mathbf{X}\mathbf{A}, \mathbf{y}) &= \widehat{\sigma}(\mathbf{X}, \mathbf{y}) \qquad \text{for any nonsingular } \mathbf{A} \in \mathbb{R}^{(p+1)\times(p+1)} \end{aligned}$$

and

$$\widehat{\sigma}(\mathbf{X}, \lambda\mathbf{y}) = |\lambda| \widehat{\sigma}(\mathbf{X}, \mathbf{y}) \qquad \text{for all } \lambda \in \mathbb{R}$$

Hence, $\widehat{\sigma}$ is regression and affine invariant, as well as scale equivariant, which ensures the regression, scale end affine equivariance of the M estimator $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$.

4.4.5 Regression quantiles as regression M estimates

Let

$$\rho_\alpha(u) = \begin{cases} \alpha u & \text{if } u \geq 0 \\ -(1-\alpha)u & \text{if } u < 0 \end{cases}$$

for $\alpha \in (0,1)$. Koenker and Bassett (1978) defined the *regression α-quantile* $\widehat{\boldsymbol{\beta}}_\alpha$ as follows:

$$\widehat{\boldsymbol{\beta}}_\alpha = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\alpha(y_i - \mathbf{x}_i'\boldsymbol{\beta}) \tag{4.23}$$

The case $\alpha = 0.5$ corresponds to the LAD estimate. Assume the model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta}_\alpha + \epsilon_i, \qquad i = 1, \ldots, n$$

where the $\mathbf{x}_i$'s are fixed and the α-quantile of ϵ_i is zero; this is equivalent to assuming that the α-quantile of y_i is, conditionally to $\mathbf{x}_i$, equal to $\mathbf{x}_i'\boldsymbol{\beta}_\alpha$. Then $\widehat{\boldsymbol{\beta}}_\alpha$ defined by (4.23) is an estimate of $\boldsymbol{\beta}_\alpha$. It may be seen as a generalization of the LAD estimate as well as a specific case of M estimate.

Regression quantiles are especially useful with heteroskedastic data. There is a very large literature on regression quantiles; see, for instance, Koenker (2005).

4.4.6 Monotone vs. redescending M estimators

As already mentioned, taking a loss function $\rho(u)$ in the minimization problem (4.16) that is less rapidly increasing than the square function provides a certain robustness

of the regression M estimate with respect to the vertical points. But what about the robustness of $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ with respect to leverage points? To answer to this question, let us recall that $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ is solution of the estimating equations system (4.18).

It is easy to see that *monotone* M estimates break down in presence of a single bad leverage point. Indeed, if $\psi(u)$ is a monotone function, an $\mathbf{x}$-outlier will dominate the solution of (4.18) in the following sense: if for some i, $\mathbf{x}_i$ is "much larger than the rest", then in order to make the sum in the left part of (4.18) to zero, the residual $r_i(\widehat{\boldsymbol{\beta}}) = y_i - \mathbf{x}_i'\widehat{\boldsymbol{\beta}}$ must be near zero, that is, the regression hyperplane has to fit the point $(\mathbf{x}_i, y_i)$ as well as possible, and hence $\widehat{\boldsymbol{\beta}}$ is essentially determined by this leverage point $(\mathbf{x}_i, y_i)$.

This does not happen with the *redescending* M estimate since the use of a function ψ that vanishes for "outlying" residuals allows to find a solution $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ of (4.18) which is not affected by the presence of a bad leverage point $(\mathbf{x}_i, y_i)$ in the data set. Hence, from the robustness point of view, the redescending regression M estimators are more interesting than the monotone M estimators. Unfortunately, as already explained, the practical implementation of M estimators is less easy for the redescending than for the monotone ones.

4.4.7 GM estimation

Other approaches have been considered to limit the influence of leverage points on the estimation of the regression coefficients. For instance, defining $\underline{\mathbf{x}}_i = (x_{i1}, \ldots, x_{ip})'$ such that $\mathbf{x}_i = (1, \underline{\mathbf{x}}_i')'$, a simple way to robustify a monotone M estimate is to downweight the influential $\underline{\mathbf{x}}_i$'s to prevent them from dominating the estimating equations. Hence we may define an estimate as solution of

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) \widetilde{w}(d(\underline{\mathbf{x}}_i))\mathbf{x}_i = \mathbf{0} \tag{4.24}$$

where $\widetilde{w}$ is a weight function and $d(\underline{\mathbf{x}}_i)$ is some measure of the "largeness" of $\underline{\mathbf{x}}_i$. Here ψ is monotone and $\widehat{\sigma}$ is simultaneously estimated by an M estimating equation of the form

$$\frac{1}{n}\sum_{i=1}^{n} \rho_{\mathrm{scale}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta$$

In order to bound the effect of influential points, $\widetilde{w}$ must be such that $\widetilde{w}(t)t$ is bounded.

More generally, we may let the weights depend on the residuals as well as on the predictor variables, and use a *generalized* M estimate (GM estimate) $\widehat{\boldsymbol{\beta}}_{\mathrm{GM}}$ defined as solution of

$$\sum_{i=1}^{n} \eta\left(d(\underline{\mathbf{x}}_i), \frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0} \tag{4.25}$$

where for each s, $\eta(s, u)$ is a nondecreasing and bounded ψ-function of u. The estimating equations system (4.24) may be seen as a particular case of (4.25) when choosing

$\eta(s,u) = \widetilde{w}(s)\psi(u)$. This particular choice corresponds to the class of *Mallows estimates* (see Mallows 1975) which has been extensively studied in the literature.

The most usual way to measure the "largeness" of $\underline{\mathbf{x}}_i$, $i = 1, \ldots, n$, is to take the *leverage* of $\underline{\mathbf{x}}_i$, that is, to consider

$$d(\underline{\mathbf{x}}_i) = \sqrt{(\underline{\mathbf{x}}_i - \widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})' \widehat{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}}}^{-1} (\underline{\mathbf{x}}_i - \widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}})} \tag{4.26}$$

where $\widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}$ and $\widehat{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}}}$ are a robust location vector and robust dispersion matrix of the $\underline{\mathbf{x}}_i$'s, respectively (see chapter 8). If $\widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}$ and $\widehat{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}}}$ are the sample mean and covariance matrix, $d(\cdot)$ is known as the Mahalanobis distance.

As stated in Rousseeuw and Leroy (1987), the GM estimators were constructed in the hope of bounding the influence of a single outlying observation. Relying on this, optimal choices of ψ and $\widetilde{w}$ were made (see, among others, Ronchetti and Rousseeuw 1985 for a survey). However, Maronna et al. (1979) have proven that the breakdown point of all GM estimators is non-zero but decreases as a function of p (i.e., the breakdown point is less or equal to $1/(p+1)$) pointing out that a GM estimator is interesting to be used only when the number of explanatory variables is very small. Furthermore, Maronna et al. (2006) show that, to obtain affine equivariance of $\widehat{\boldsymbol{\beta}}_{\text{GM}}$, it is necessary that $\widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}$ and $\widehat{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}}}$ used in (4.26) are affine equivariant, which presents the same computational difficulties as for redescending M estimates and reduce substantially the appeal of this estimator.

4.5 Robust regression with a high breakdown point

As explained previously, LS regression is now being criticized more and more for its dramatic lack of robustness. Indeed, one single outlier can have an arbitrarily large effect on the estimate: the breakdown point ε^* of $\widehat{\beta}_{\text{LS}}$ is clearly equal to zero. Although L_1-regression protects against outlying y_i, it cannot cope with grossly aberrant values of $\mathbf{x}_i$: L_1-regression yields the same value $\varepsilon^* = 0$ as LS. M-estimation provides a certain robustness with respect to vertical points, but not with respect to bad leverage points when the loss function ρ is unbounded: the breakdown point ε^* associated with a monotone M-estimator is then still equal to zero.

Because of this vulnerability to bad leverage points, generalized M-estimators (GM-estimators) were introduced, with the basic purpose of bounding the influence of outlying $\mathbf{x}_i$. It turns out, however, that the GM-estimators now in use have a breakdown point of at most $1/(p+1)$, where $(p+1)$ is the dimension of $\mathbf{x}_i$. Various other estimators have been proposed by Theil (1950), Brown and Mood (1951), Sen (1968), Jaeckel (1972), and Andrews (1974), but none of them achieves $\varepsilon^* = 30\%$ in the case of simple regression ($p = 1$).

All of this raises the question whether robust regression with a high breakdown point is at all possible. The affirmative answer was given by Siegel (1982), who proposed an estimator (the *repeated median*) with a 50% breakdown point. Note that 50% is the best that can be expected: for larger amounts of contamination, it becomes impossible

to distinguish between the "good" and the "bad" parts of the sample. Siegel's estimator can be calculated explicitly but is not equivariant for linear transformations of the $\mathbf{x}_i$ (it is not affine equivariant). This explains why we do not study this estimator in more details and prefer to present other estimators introduced by Rousseeuw and Yohai, and all based on a robust scale measure.

4.5.1 LTS- and LMS-estimation

Robustness can be achieved by tackling the estimation of the regression parameters vector $\boldsymbol{\beta}$ from a different perspective. We know that LS estimation is based on the minimization of the variance of the residuals. However, since the variance is highly sensitive to outliers, LS-estimate will be sensitive to them as well. An interesting idea would then consist in minimizing a measure of the residual dispersion $s(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$ that is less sensitive to extreme residuals.

Relying on this idea, Rousseeuw (1983) introduced the *Least Trimmed Sum of Squares* (LTS) estimator which is based on the minimization of a trimmed variance of the residuals:

$$\widehat{\boldsymbol{\beta}}_{\text{LTS}} = \arg\min_{\boldsymbol{\beta}} s_{\text{LTS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$$

with

$$s_{\text{LTS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})) = \left(\frac{1}{\lceil \alpha n \rceil} \sum_{i=1}^{\lceil \alpha n \rceil} r^2_{(i)}(\boldsymbol{\beta}) \right)^{1/2},$$

where $1/2 \leq \alpha \leq 1$, $r^2_{(1)}(\boldsymbol{\beta}) \leq r^2_{(2)}(\boldsymbol{\beta}) \leq \cdots \leq r^2_{(n)}(\boldsymbol{\beta})$ are the ordered squared residuals, and $\lceil \cdot \rceil$ is the ceil function. The constant α determines the trade-off between the robustness and the efficiency of the estimator. Indeed, if α tends to one, the LTS-estimator tends to the LS estimator. In contrast, if $\alpha = 1/2$, the LTS-estimator will resist up to 50% of outlying data and, consequently, will have a breakdown point equal to 50%. Unfortunately, even if $\widehat{\boldsymbol{\beta}}_{\text{LTS}}$ converges to $\boldsymbol{\beta}$ at a rate of $1/\sqrt{n}$, its efficiency is low (under Gaussian conditions, the asymptotic relative efficiency of $\widehat{\boldsymbol{\beta}}_{\text{LTS}}$ with respect to $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ reaches only 7% when 50% of the data are trimmed).

Despite its relatively low efficiency, the LTS-estimator is quite popular because it can be quickly computed using the *Fast-lts algorithm* developed by Rousseeuw and Van Driessen (2006); this estimator is available in Stata through the command `robreg lts`.

Following the same idea, Rousseeuw (1984) introduced the *Least Median Squares* (LMS) estimator based on the minimization of the median of the squared residuals[4]:

$$\widehat{\boldsymbol{\beta}}_{\text{LMS}} = \arg\min_{\boldsymbol{\beta}} s_{\text{LMS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$$

with

$$s_{\text{LMS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})) = \left(\text{med}_i\, r_i^2(\boldsymbol{\beta})\right)^{1/2}.$$

4. The variance of the residuals corresponds to the arithmetic mean of the squared residuals; why not replace the mean by the more robust median?

LMS satisfies $\varepsilon^* = 50\%$ but has unfortunately a very low efficiency because of its $1/\sqrt[3]{n}$ convergence rate. The LMS-estimator is available in Stata through the command `robreg lms`.

4.5.2 S-estimation

Following always the same principle, Rousseeuw and Yohai (1984) have introduced a more general class of estimators: the regression *S-estimators*.

In order to well understand the basic intuition behind the S-estimation, let us consider once again the LS estimation. For LS estimation, we actually are looking for the value of the regression coefficients vector $\boldsymbol{\beta}$ that minimizes the variance (or standard deviation) of the residuals $r_i(\boldsymbol{\beta})$ $(i = 1, \ldots, n)$. More formally, we have

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}} s_{\text{LS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$$

with

$$s_{\text{LS}}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta})) = \left(\frac{1}{n} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) \right)^{1/2}.$$

The dispersion measure s_{LS} may be characterized as follows: given the realizations $e_1, \ldots, e_n$ of n i.i.d. random variables whose distribution is characterized by a mean equal to zero and a scale parameter σ, the dispersion measure $s_{\text{LS}}(e_1, \ldots, e_n)$ of the sample is an estimate of σ satisfying the equality

$$\frac{1}{n} \sum_{i=1}^{n} \left(\frac{e_i}{s_{\text{LS}}(e_1, \ldots, e_n)} \right)^2 = 1,$$

or still, taking $\rho(u) = u^2$,

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s_{\text{LS}}(e_1, \ldots, e_n)} \right) = 1.$$

Moreover, if $u \sim \mathcal{N}(0, 1)$, then $\mathrm{E}[\rho(u)] = \mathrm{E}\big[u^2\big] = 1$.

The S-estimation procedure proposed by Rousseeuw and Yohai (1984) relies on the same philosophy as the one underlying the LS estimation, but introduces robustness by using specific robust residual dispersion measures which correspond to M-estimators of the scale parameter σ. More formally, given the realizations $e_1, \ldots, e_n$ of n i.i.d. random variables with scale parameter σ (and a location parameter equal to zero), the *M-estimate* $\widehat{\sigma}_\rho$ of σ is the measure of dispersion $s_\rho(e_1, \ldots, e_n)$ defined as the solution of the equation

$$\frac{1}{n} \sum_{i=1}^{n} \rho\left(\frac{e_i}{s_\rho(e_1, \ldots, e_n)} \right) = \delta \tag{4.27}$$

where

- the function $\rho(\cdot)$ is positive, even, such that $\rho(0) = 0$, non decreasing for positive values and bounded;
- the constant δ is defined such that $\widehat{\sigma}_\rho = s_\rho(e_1, \ldots, e_n)$ is a consistent estimate of σ for the Gaussian regression model (generally δ is defined by $\delta = \mathrm{E}[\rho(u)]$ for $u \sim \mathcal{N}(0,1)$; the consistency parameter δ would therefore be nothing else than the population counterpart of the lefthand side of equation (4.27)).

Then Rousseeuw and Yohai (1984) defined an S-estimate of $\boldsymbol{\beta}$ by

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho} = \arg\min_{\boldsymbol{\beta}} s_\rho(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$$

where s_ρ is a measure of dispersion defining a scale M-estimator, i.e. satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_\rho(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))}\right) = \delta \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}. \tag{4.28}$$

One important fact is that an S-estimate of $\boldsymbol{\beta}$ is also an M-estimate. More precisely, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ is an M-estimate (in the sense of (4.16)) in that

$$\sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right)}{\widehat{\sigma}_\rho}\right) \leq \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widetilde{\boldsymbol{\beta}}\right)}{\widehat{\sigma}_\rho}\right) \qquad \text{for all } \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}, \tag{4.29}$$

where the residuals are standardized by the same scale M-estimate $\widehat{\sigma}_\rho = s_\rho\left(r_1\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right), \ldots, r_n\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right)\right)$ of σ on both sides of the inequation (4.29). Indeed, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ minimizes the residual dispersion measure $s_\rho(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$ which satisfies (4.28). This means that, if we denote $\widehat{\sigma}_\rho = s_\rho\left(r_1\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right), \ldots, r_n\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right)\right)$ and $\widetilde{\sigma}_\rho = s_\rho\left(r_1\left(\widetilde{\boldsymbol{\beta}}\right), \ldots, r_n\left(\widetilde{\boldsymbol{\beta}}\right)\right)$ for $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$, we have $\widehat{\sigma}_\rho \leq \widetilde{\sigma}_\rho$ and

$$\sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right)}{\widehat{\sigma}_\rho}\right) = n\delta = \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widetilde{\boldsymbol{\beta}}\right)}{\widetilde{\sigma}_\rho}\right).$$

Then, since ρ is monotone and $\widehat{\sigma}_\rho \leq \widetilde{\sigma}_\rho$, we necessarily have

$$\sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\right)}{\widehat{\sigma}_\rho}\right) = \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widetilde{\boldsymbol{\beta}}\right)}{\widetilde{\sigma}_\rho}\right) \leq \sum_{i=1}^{n} \rho\left(\frac{r_i\left(\widetilde{\boldsymbol{\beta}}\right)}{\widehat{\sigma}_\rho}\right),$$

which proves (4.29).

If ρ has a derivative ψ, it follows that $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ is also an M-estimate in the sense of (4.18), but with the condition that the scale parameter σ is estimated simultaneously with $\boldsymbol{\beta}$.

More formally, $\boldsymbol{\beta}$ is estimated by $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ and σ by $\widehat{\sigma}_\rho = s_\rho\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\Big), \ldots, r_n\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\Big)\Big)$, with $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ and $\widehat{\sigma}_\rho$ such that

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \psi\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\Big)}{\widehat{\sigma}_\rho}\right)\mathbf{x}_i = \mathbf{0} \\ \displaystyle\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}\Big)}{\widehat{\sigma}_\rho}\right) = \delta. \end{cases}$$

Note that, taking $\rho(u) = u^2$ and $\delta = 1$, we retrieve the standard LS minimization problem.

The choice of $\rho(\cdot)$ is crucial to have good robustness properties[5] and a high Gaussian efficiency. The Tukey-Biweight function defined in (4.17), with $\kappa = 1.547$, is a common choice. This S-estimator resists to a contamination of up-to 50% of outliers and, hence, have a breakdown point of 50%. Unfortunately, this S-estimator has a Gaussian efficiency of only 28.7%. If $\kappa = 5.182$, the Gaussian efficiency raises to 96.6% but the breakdown point drops to 10%. Actually an S-estimator cannot simultaneously have a high breakdown point and a high efficiency. In particular, Hössjer (1992) has shown that the maximum Gaussian asymptotic efficiency of an S-estimator with a breakdown point of 50% is 33%.

4.5.3 MM-estimation

We have just seen that S-estimation does not allow to reach jointly a high breakdown point and a high Gaussian efficiency. How should we then estimate the parameters of the regression model if we aim to combine high efficiency under normal errors with a high breakdown point? Several proposals have been made: the *MM-estimators* of Yohai (1987), the *τ-estimators* of Yohai and Zamar (1988), the *constrained M (CM) estimators* of Mendes and Tyler (1996). All these estimators can have a Gaussian asymptotic efficiency as close to 1 as desired, and simultaneously a breakdown point of 50%. Gervini and Yohai (2002) proposed one estimator that has a breakdown point of 50% and an efficiency equal to 1.

Let us here focus our attention on the regression MM-estimators since there are based on the M- and S-estimation procedures studied in the previous sections. An MM-estimator is defined in two successive steps:

1. Take an S-estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ with high breakdown point (but possibly low normal efficiency) where the scale measure s_{ρ_0} is defined by

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i(\boldsymbol{\beta})}{s_{\rho_0}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))}\right) = \delta \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}$$

5. Note that the function ρ defining the S-estimator needs to be *bounded* to get a positive breakdown point for the regression estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$.

(s_{ρ_0} is associated with the function $\rho_0(\cdot)$ and the constant δ). Let $\widehat{\sigma}_{\rho_0} = s_{\rho_0}\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big), \ldots, r_n\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big)\Big)$.

2. Take any other function $\rho(\cdot) \leq \rho_0(\cdot)$ and find the MM-estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ as a local minimum of
$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\rho_0}}\right) \tag{4.30}$$
such that
$$\sum_{i=1}^{n} \rho\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)}{\widehat{\sigma}_{\rho_0}}\right) \leq \sum_{i=1}^{n} \rho\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big)}{\widehat{\sigma}_{\rho_0}}\right). \tag{4.31}$$

❑ **Remark**

The key result is given in Yohai (1987). Recall that all local minima of (4.30) is solution of the estimating equations (4.18) with $\psi(u) = \rho'(u)$ and $\widehat{\sigma} = \widehat{\sigma}_{\rho_0}$:
$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i = \mathbf{0}. \tag{4.32}$$
Yohai shows that if $\rho(u) \leq \rho_0(u)$ for all $u \in \mathbb{R}$ and if (4.31) is satisfied, then $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ is consistent. Moreover, it can be shown that the MM-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ has the same breakdown point than the S-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ of the first step, determined by the function $\rho_0(\cdot)$. If furthermore $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ is any solution of (4.32), then it has the same efficiency — this efficiency is determined by the choice of the function $\rho(\cdot)$ — as the global minimum of (4.30). In conclusion, it is not necessary to find the absolute minimum of (4.30) to ensure consistency, a high breakdown point and a high efficiency.

❑

It is common to use a Tukey-Biweight $\rho_\kappa^B(\cdot)$ function for both the preliminary S-estimator and the final MM-estimator. The tuning constant κ can be set to 1.547 for the preliminary S-estimator to guarantee a 50% breakdown point, and it can be set to 4.685 for the second step MM-estimator to guarantee a 95% asymptotic Gaussian efficiency of this final estimator. Note however that though not breaking-down, an MM-estimator with a very high efficiency may have a high bias *under moderate contamination*: the larger the efficiency, the larger the bias. It is therefore important to choose the efficiency so as to maintain reasonable bias control. Results in Section 5.9 of Maronna et al. (2006) show that an efficiency of 0.95 yields too high a bias, and hence it is safer to choose an efficiency of 0.85 which gives a smaller bias while retaining a sufficiently high efficiency. We will raise once again this problem of bias in Subsection 4.6.4.

Numerical computation of the S- and MM-estimate

The numerical computation of the estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ at the second step of the procedure follows the approach described in Section 4.4.4: starting with $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$, we use the iter-

atively reweighted least squares algorithm to attain a solution of the equation (4.32). It may be shown (see Maronna et al. (2006) that (4.30) decreases at each iteration, which insures (4.31). Hence, once the initial S-estimate is computed, $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ comes at almost no additional computational cost.

We programmed an S- and an MM-estimator in Stata (with Tukey-Biweight loss function) using the fast algorithm of Salibian-Barrera and Yohai (2006) for computing the S-estimator. Explicit formulas for the estimators are not available and it is necessary to call on numerical optimization to compute them. We present just below a sketch of the fast algorithm for regression S-estimates we implemented in Stata.

Consider an estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ defined as

$$\arg \min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} s_\rho(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta})). \tag{4.33}$$

An approximate solution of (4.33) can be obtained by finding $\widehat{\boldsymbol{\beta}}$ equal to

$$\arg \min_{\boldsymbol{\beta}\in\mathcal{D}_N} s_\rho(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta}))$$

where

$$\mathcal{D}_N = \left\{\widehat{\boldsymbol{\beta}}_1,\ldots,\widehat{\boldsymbol{\beta}}_N\right\}$$

is a finite set of well selected candidates for $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$. One way to select these candidates is by subsampling elementary sets among the sample $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)$ (see Rousseeuw 1984). More formally, take a first random subsample of $(p+1)$ observations[6]

$$(\mathbf{x}_{i_1},y_{i_1}),\ldots,\left(\mathbf{x}_{i_{(p+1)}},y_{i_{(p+1)}}\right);$$

then the candidate $\widehat{\boldsymbol{\beta}}_1$ is obtained by fitting a hyperplane containing these $(p+1)$ points:

$$\mathbf{x}_{i_j}^t\widehat{\boldsymbol{\beta}}_1 = y_{i_j}, \qquad j=1,\ldots,p+1.$$

Taking N subsamples we obtain the N candidates. Note that if a subsample is collinear, it is replaced by another.

How large should N be? We have to guarantee that $\mathcal{D}_N$ includes at least one "good" candidate with high probability, say $(1-\alpha)$ (with, for example, $\alpha=0.01$). A necessary condition to have a "good" candidate is that it comes from a clean subsample, i.e., a subsample without outliers.

The probability of getting a clean subsample depends on the fraction of outliers in the sample and on p. When the fraction of outliers in the sample increases, the probability of getting a clean subsample decreases. Suppose the sample contains a proportion ξ of outliers. Then the probability of an outlier-free subsample is $\gamma=(1-\xi)^{p+1}$, and the

6. Recall that $(p+1)$ is the number of regression parameters to estimate, that is the dimension of the regression coefficients vector $\boldsymbol{\beta}$ to estimate.

probability of at least one clean subsample among the N selected subsamples is equal to $1-(1-\gamma)^N$. If we want this probability to be larger than $(1-\alpha)$, we must have

$$\log\alpha \geq N\log(1-\gamma) \approx -N\gamma$$

and hence

$$N \geq \frac{|\log\alpha|}{\left|\log\left(1-(1-\xi)^{p+1}\right)\right|} \approx \frac{|\log\alpha|}{(1-\xi)^{p+1}} \tag{4.34}$$

for p not too small (see Salibian-Barrera and Zamar 2004). Therefore N must grow exponentially with p.

The following observation allows to save much computing time. Suppose we have examined $(M-1)$ subsamples and

$$\widehat{\sigma}_{\rho;M-1} = s_\rho\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{M-1}\Big),\ldots,r_n\Big(\widehat{\boldsymbol{\beta}}_{M-1}\Big)\Big)$$

is the current minimum of the residual dispersion measure s_ρ. Now we draw the M-th subsample which yields the candidate $\widehat{\boldsymbol{\beta}}_M$. Let us consider $\widehat{\sigma}_{\rho;M} = s_\rho\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_M\Big),\ldots,r_n\Big(\widehat{\boldsymbol{\beta}}_M\Big)\Big)$. Since ρ is a monotone function, the inequality $\widehat{\sigma}_{\rho;M} < \widehat{\sigma}_{\rho;M-1}$ implies that

$$n\delta = \sum_{i=1}^{n}\rho\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_M\Big)}{\widehat{\sigma}_{\rho;M}}\right) \geq \sum_{i=1}^{n}\rho\left(\frac{r_i\Big(\widehat{\boldsymbol{\beta}}_M\Big)}{\widehat{\sigma}_{\rho;M-1}}\right). \tag{4.35}$$

Consequently, if we observe that $\sum_{i=1}^{n}\rho\left(\frac{r_i\left(\widehat{\boldsymbol{\beta}}_M\right)}{\widehat{\sigma}_{\rho;M-1}}\right) > n\delta$, this necessarily means that $\widehat{\sigma}_{\rho;M} \geq \widehat{\sigma}_{\rho;M-1}$ and we may spare the effort of computing the scale estimate $\widehat{\sigma}_{\rho;M}$ and discard $\widehat{\boldsymbol{\beta}}_M$. Therefore $\widehat{\sigma}_\rho$ has to be computed only for those subsamples that verify the inequality (4.35).

Although the N given by (4.34) ensures that the approximation $\widehat{\boldsymbol{\beta}}$ of $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ has the desired breakdown point, it does not imply that it is a good approximation to the exact S-estimate. To solve this problem, Salibian-Barrera and Yohai (2006) have proposed a procedure based on a "local improvement" step of the resampling initial candidates. This allows for a substantial reduction of the number of candidates required to obtain a good approximation to the optimal solution.

This algorithm can be called in Stata either directly using the `robreg s` function[7] or indirectly ussing the `robreg mm` function developed to compute MM-estimate, and invoking the `initial` option. Once the S-estimate is obtained, the MM-estimate directly follows by applying the iteratively reweighted least squares algorithm up to convergence. As far as inference is concerned, standard errors robust to heteroskedasticity (and asymmetric errors) are computed according to the formulas available in the literature (see Section 4.6).

7. The default values that are used in Stata for the implementation of the fast-S algorithm are $\xi = 0.2$ and $\alpha = 0.01$.

4.5.4 MS-estimation

Explicit formulas for $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ are generally not available and, as explained in the previous section, empirical implementation of S-estimation requires numerical optimization based on a subsampling algorithm. But this method presents an Achille's heel: it becomes inapplicable in practice when several *dummy* explanatory variables are involved in the regression model (4.1). Indeed, when several of the explanatory variables are binary, there is a high probability that random selection of subsamples yields collinear subsamples.

To cope with this, Maronna and Yohai (2000) have introduced the MS-estimator. The intuition behind this estimator is simple. For the sake of clarity, let us separate continuous and dichotomous variables in (4.1) and rewrite the regression model equation as follows:

$$y = (\beta_0 + \beta_1 x_1 + \ldots + \beta_{p_1} x_{p_1}) + \left(\beta_1^* x_1^* + \ldots + \beta_{p_2}^* x_{p_2}^*\right) + \varepsilon \tag{4.36}$$

where $x_1, \ldots, x_{p_1}$ are p_1 continuous explanatory variables and $x_1^*, \ldots, x_{p_2}^*$ are p_2 dichotomous explanatory variables ($p = p_1 + p_2$). If $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p_1})^t$ was known in equation (4.36), then $\boldsymbol{\beta}^* = \left(\beta_1^*, \ldots, \beta_{p_2}^*\right)^t$ would be robustly estimated using a monotone M-estimator (since $x_1^*, \ldots, x_{p_2}^*$ are all dummy variables, the data set can only contain, at worst, vertical outliers). On the other hand, if $\boldsymbol{\beta}^*$ was known, then $\boldsymbol{\beta}$ should be estimated using an S-estimator[8] and the subsampling algorithm should not generate collinear subsamples since all explanatory variables are continuous. The idea is then to alternate these two estimators till convergence.

Technically speaking, an MS-regression estimate is obtained iteratively; at the k-th step, we define $\widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(k)}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{*(k)}$ as follows. Let s_ρ be a measure of dispersion satisfying (4.28), $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip_1})^t$ and $\mathbf{x}_i^* = \left(x_{i1}^*, \ldots, x_{ip_2}^*\right)^t$:

$$\begin{cases} \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(k)} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p_1+1}} s_\rho\left(\left[y_i - (\mathbf{x}_i^*)^t \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{*(k-1)}\right] - \mathbf{x}_i^t \boldsymbol{\beta}\ ; i = 1, \ldots, n\right) \\ \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{*(k)} = \arg\min_{\boldsymbol{\beta}^* \in \mathbb{R}^{p_2}} \sum_{i=1}^n \rho\left(\dfrac{\left[y_i - (\mathbf{x}_i)^t \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(k-1)}\right] - (\mathbf{x}_i^*)^t \boldsymbol{\beta}^*}{\widehat{\sigma}^{(k-1)}}\right), \end{cases}$$

where $\widehat{\sigma}^{(k-1)} = s_\rho\left(y_i - (\mathbf{x}_i)^t \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{(k-1)} - (\mathbf{x}_i^*)^t \widehat{\boldsymbol{\beta}}_{\mathrm{MS}}^{*(k-1)}\ ; i = 1, \ldots, n\right)$. Note that `robreg s` and `robreg mm` automatically recognize the presence of dummy variables among the explanatory variables and, if appropriate, automatically apply the MS-procedure.

Unfortunately, as stated above, the price to pay for robustness is efficiency. However this MS-estimator can be particularly helpful in the fixed effects panel data models, as suggested by Bramati and Croux (2007).

8. Since $x_1, \ldots, x_{p_1}$ are continuous explanatory variables, we cannot assume that there are no leverage points.

4.6 Robust inference for M-, S- and MM-estimators

Consistency and asymptotic normality of M-estimators under the assumption of i.i.d. error terms have been studied by Yohai and Maronna (1979) and for MM-estimators by Yohai (1987). Under fairly general conditions, allowing also for heteroskedasticity, asymptotic normality for S- and MM-estimators has been shown by Salibian-Barrera and Zamar (2004) in the location case. Some of these results are summarized in Maronna et al. (2006) with a distinction made between the case of *fixed* predictors and the case of *random* predictors.

Croux et al. (2003) have established the asymptotic normality of M-, S- and MM-estimators in the regression case under quite general conditions: they only assume that the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are generated by a *stationary*[9] and *ergodic*[10] process H. Under this assumption, the observations do not need to be independent, we may have heteroskedasticity (the processes $\mathbf{x}_i$ and ε_i are not necessarily independent) and the distribution of the error terms is not necessarily symmetric. In this context, the authors of Croux et al. (2003) have showed that the M-, S- and MM-estimators of the regression parameters $\boldsymbol{\beta}$ and of the scale parameter σ are first-order equivalent with exactly-identified GMM (Generalized Method of Moments) estimators and have then deduced the asymptotic variance matrix of the M-, S- and MM-estimators of $\boldsymbol{\beta}$ from results established for GMM (see Hansen 1982). The interest of the results of Croux et al. (2003) is multiple. They propose explicit formulas for the asymptotic variance matrices of the robust regression estimators, so recourse to bootstrap techniques is not necessary. Moreover, these variances are valid in the presence of autocorrelation and heteroskedasticity; as we will show it, if we impose the independence between the observations, the absence of heteroskedasticity or the symmetry of the distribution of the error terms, the expressions of the variances become much simpler and coïncide with the results previously proved by other authors. The robustness with respect to outliers of the estimates of the variance matrices is also taken into account. Finally, the results of Croux et al. (2003) may be used to develop robust confidence intervals and robust tests for the regression parameters; they are also on the basis of the extension of the Hausman test presented at the end of this section, which allows to check for the presence of outliers — by comparing the regression coefficients estimated by least squares and by a robust S-procedure — and to fix the maximal efficiency that may have an MM-estimator without suffering of significant bias in the presence of contamination of the data set by (moderately) bad leverage points — by comparing an S-estimate of $\boldsymbol{\beta}$ with several MM-estimates of different efficiencies.

9. A *stationary* process is a stochastic process whose joint probability distribution does not change when shifted in time or space. Consequently, parameters such as the mean and the variance, if they exist, also do not change over time or position . Hence, the mean and the variance of the process do not follow trends.
10. A stochastic process is said to be *ergodic* if its statistical properties (such as its mean and variance) can be estimated consistently from a single, sufficiently long sample (realization) of the process.

4.6.1 Asymptotic distribution of M-, S- and MM-estimators

Let us here present some of the fundamental results established by Croux et al. (2003) for the asymptotic distribution of M-, S- and MM-estimators. The interested reader will find some details about the main steps of the approach used to demonstrate these results in Appendix 2 at the end of this chapter.

Let y be the scalar dependent variable and $\mathbf{x} = (1, x_1, \ldots, x_p)^t$ be the $(p+1)$-vector of covariates. Consider once again the regression model (4.4). Here, the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are assumed to be generated by a *stationary* and *ergodic* process. To avoid too much technicalities, we also assume that the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are *independent*[11].

Let us denote by $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ the S-estimator of $\boldsymbol{\beta}$ associated with the loss function ρ_0:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} = \arg\min_{\boldsymbol{\beta}} s_{\rho_0}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$$

where s_{ρ_0} is a measure of dispersion satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{r_i(\boldsymbol{\beta})}{s_{\rho_0}(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))}\right) = \delta \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}.$$

This leads to the scale M-estimator

$$\widehat{\sigma}_{\rho_0} = s_{\rho_0}\Big(r_1(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}), \ldots, r_n(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0})\Big).$$

Let $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ be the MM-estimator of $\boldsymbol{\beta}$ associated with the loss function ρ_0 for the first step of the estimation procedure (S-estimation) and with the loss function ρ for the second step (M-estimation): $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ is a (local) minimum of

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\rho_0}}\right).$$

To avoid any ambiguity in the formulation of the results, we will denote the vector of regression parameters by $\boldsymbol{\beta}$ when it is estimated by $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and by $\boldsymbol{\beta}_0$ when it is estimated by $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$. Moreover, we will use the generic notations $u_0 = \frac{y - \mathbf{x}^t\boldsymbol{\beta}_0}{\sigma}$ and $u = \frac{y - \mathbf{x}^t\boldsymbol{\beta}}{\sigma}$, and we will simply replace $\psi(u) = \rho'(u)$ by ψ, and $\rho_0(u_0)$ by ρ_0.

Using these notations, we may formulate the results shown by Croux et al. (2003) as follows.

❑ **Proposition**

11. The interested reader can find very general results, valid in presence of *autocorrelation*, in Croux et al. (2003).

If the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are generated by a stationary and ergodic process, and are independent (Assumption A), then

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} \\ \widehat{\sigma}_{\rho_0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \\ \sigma \end{pmatrix}\right) \to^d \mathcal{N}(\mathbf{0}, \mathbf{V}_{\mathrm{MM}})$$

where

$$\mathbf{V}_{\mathrm{MM}} = \mathbf{G}_{\mathrm{MM}}^{-1} \boldsymbol{\Omega}_{\mathrm{MM}} \big(\mathbf{G}_{\mathrm{MM}}^t\big)^{-1}, \tag{4.37}$$

with the matrices $\mathbf{G}_{\mathrm{MM}}$ and $\boldsymbol{\Omega}_{\mathrm{MM}}$ given by:

$$\mathbf{G}_{\mathrm{MM}} = -\frac{1}{\sigma}\mathrm{E}\begin{pmatrix} \psi'\mathbf{x}\mathbf{x}^t & \mathbf{0} & \psi' u\mathbf{x} \\ \mathbf{0} & \rho_0''\mathbf{x}\mathbf{x}^t & \rho_0'' u_0 \mathbf{x} \\ \mathbf{0} & \mathbf{0} & \rho_0' u_0 \end{pmatrix} \tag{4.38}$$

and

$$\boldsymbol{\Omega}_{\mathrm{MM}} = \mathrm{E}\begin{pmatrix} \psi^2\mathbf{x}\mathbf{x}^t & \psi\rho_0'\mathbf{x}\mathbf{x}^t & \psi\rho_0\mathbf{x} \\ \psi\rho_0'\mathbf{x}\mathbf{x}^t & (\rho_0')^2\mathbf{x}\mathbf{x}^t & \rho_0\rho_0'\mathbf{x} \\ \psi\rho_0\mathbf{x}^t & \rho_0\rho_0'\mathbf{x}^t & \rho_0^2 - \delta^2 \end{pmatrix}. \tag{4.39}$$

❑

In particular, this result establishes the consistency of the regression MM-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and S-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$, and of the scale M-estimator $\widehat{\sigma}_{\rho_0}$.

Moreover, il allows to derive explicit formulas for the asymptotic variances of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ — denoted hereafter by $\mathrm{Avar}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\big)$ and $\mathrm{Avar}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\big)$, respectively —, and for the asymptotic covariance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ — denoted by $\mathrm{Acov}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\big)$:

$$\begin{aligned} \mathrm{Avar}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\big) = \frac{1}{n}[&\mathbf{A}\mathrm{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A} - \mathbf{a}\mathrm{E}\big(\psi\rho_0\mathbf{x}^t\big)\mathbf{A} \\ &-\mathbf{A}\mathrm{E}(\psi\rho_0\mathbf{x})\mathbf{a}^t + \mathrm{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}\mathbf{a}^t], \end{aligned} \tag{4.40}$$

$$\begin{aligned} \mathrm{Avar}\big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\big) = \frac{1}{n}[&\mathbf{A}_{\mathrm{S}}\mathrm{E}\Big((\rho_0')^2\mathbf{x}\mathbf{x}^t\Big)\mathbf{A}_{\mathrm{S}} - \mathbf{a}_{\mathrm{S}}\mathrm{E}\big(\rho_0\rho_0'\mathbf{x}^t\big)\mathbf{A}_{\mathrm{S}} \\ &-\mathbf{A}_{\mathrm{S}}\mathrm{E}(\rho_0\rho_0'\mathbf{x})\mathbf{a}_{\mathrm{S}}^t + \mathrm{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}_{\mathrm{S}}\mathbf{a}_{\mathrm{S}}^t], \end{aligned} \tag{4.41}$$

$$\begin{aligned} \mathrm{Acov}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big) = \frac{1}{n}[&\mathbf{A}\mathrm{E}\big(\psi\rho_0'\mathbf{x}\mathbf{x}^t\big)\mathbf{A}_{\mathrm{S}} - \mathbf{a}\mathrm{E}\big(\rho_0\rho_0'\mathbf{x}^t\big)\mathbf{A}_{\mathrm{S}} \\ &-\mathbf{A}\mathrm{E}(\psi\rho_0\mathbf{x})\mathbf{a}_{\mathrm{S}}^t + \mathrm{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}\mathbf{a}_{\mathrm{S}}^t], \end{aligned} \tag{4.42}$$

with

$$\mathbf{A} = \sigma\big[\mathrm{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1}, \tag{4.43}$$

$$\mathbf{a} = \mathbf{A}\frac{\mathrm{E}(\psi' u\mathbf{x})}{\mathrm{E}(\rho_0' u_0)}, \tag{4.44}$$

$$\mathbf{A}_{\mathrm{S}} = \sigma\big[\mathrm{E}\big(\rho_0''\mathbf{x}\mathbf{x}^t\big)\big]^{-1}, \tag{4.45}$$

$$\mathbf{a}_{\mathrm{S}} = \mathbf{A}_{\mathrm{S}}\frac{\mathrm{E}(\rho_0'' u_0\mathbf{x})}{\mathrm{E}(\rho_0' u_0)}. \tag{4.46}$$

❑ **Remark**

Note that Croux et al. (2003) have also considered the case where we estimate the parameters $\boldsymbol{\beta}$ and σ simultaneously by an M-estimation procedure. Some results about the asymptotic distribution of $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}^t, \widehat{\sigma}_{\rho_0}\right)^t$ are presented in Appendix 2 at the end of this chapter.

❑

The authors have also shown that the asymptotic variances and covariances can be estimated consistently by taking their empirical counterpart. More precisely, the estimates are obtained by applying the following two rules:

1. Replace, in u and u_0, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ and σ by the estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$.
2. Replace $\mathrm{E}(\cdot)$ by $\frac{1}{n}\sum_{i=1}^n(\cdot)$.

For example, the first term of $\widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)$ is given by

$$\frac{1}{n}\left[\widehat{\mathbf{A}}\left(\frac{1}{n}\sum_{i=1}^n\left[\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\right]^2\mathbf{x}_i\mathbf{x}_i^t\right)\widehat{\mathbf{A}}\right]$$

with

$$\widehat{\mathbf{A}}=\widehat{\sigma}_{\rho_0}\left[\frac{1}{n}\sum_{i=1}^n\psi'\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i\mathbf{x}_i^t\right]^{-1}.$$

It is interesting to note that the estimate $\widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)$ of the asymptotic variance $\mathrm{Avar}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)$ is robust with respect to bad leverage points and vertical outliers. Indeed, if there are observations yielding large residuals with respect to the robust MM-fit, then $\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)$ has a small value when ψ is a redescending function[12]. Hence, if there are bad leverage points in the sample, then their $\mathbf{x}_i$-value is large, but at the same time $\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)$ will be zero. This explains intuitively why vertical outliers and bad leverage points have only a limited influence on the estimate $\widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)$.

❑ **Remark**

As previously explained, the LS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ may be seen as a particular S-estimator of $\boldsymbol{\beta}$ associated with the loss function $\rho_0(u_0)=u_0^2$ (such that $\rho_0'(u_0)=2u_0$ and $\rho_0''(u_0)=$

12. Recall that, if ψ is redescending, it has the property to be equal to zero for large arguments.

2) and with the constant $\delta = 1$. The expression of the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ may then be simply derived from the one obtained for $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$:

$$\begin{aligned}\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{LS}}\right) = \frac{1}{n}[&\mathbf{A}_{\text{LS}}\text{E}\big(4u_0^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A}_{\text{LS}} - \mathbf{a}_{\text{LS}}\text{E}\big(2u_0^3\mathbf{x}^t\big)\mathbf{A}_{\text{LS}} \\ &-\mathbf{A}_{\text{LS}}\text{E}\big(2u_0^3\mathbf{x}\big)\mathbf{a}_{\text{LS}}^t + \text{E}\big(u_0^4-1\big)\mathbf{a}_{\text{LS}}\mathbf{a}_{\text{LS}}^t\big],\end{aligned} \tag{4.47}$$

with

$$\mathbf{A}_{\text{LS}} = \frac{\sigma}{2}\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \qquad \text{and} \qquad \mathbf{a}_{\text{LS}} = \mathbf{A}_{\text{LS}}\frac{\text{E}(u_0\mathbf{x})}{\text{E}(u_0^2)}. \tag{4.48}$$

If, in addition, there is homoskedasticity[13] and if the distribution $F_{0,1}$ of the error terms is symmetric (around 0), we retrieve the well-known asymptotic variance matrix

$$\frac{\sigma^2}{n}\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1},$$

that we may estimate by

$$\frac{\widehat{\sigma}^2_{\rho_0}}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^t\right]^{-1}.$$

Note that this latter estimator is absolutely no robust with respect to leverage points.

Finally, since the LS estimation can be considered as the special case of the MM-estimation associated with $\rho(u) = u^2$, it can be shown that:

$$\begin{aligned}\text{Acov}\left(\widehat{\boldsymbol{\beta}}_{\text{LS}},\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right) = \frac{1}{n}[&\mathbf{A}\text{E}\big(2u\rho_0'\mathbf{x}\mathbf{x}^t\big)\mathbf{A}_{\text{S}} - \mathbf{a}\text{E}\big(\rho_0\rho_0'\mathbf{x}^t\big)\mathbf{A}_{\text{S}} \\ &-\mathbf{A}\text{E}(2u\rho_0\mathbf{x})\mathbf{a}_{\text{S}}^t + \text{E}\big(\rho_0^2-\delta^2\big)\mathbf{a}\mathbf{a}_{\text{S}}^t\big],\end{aligned} \tag{4.49}$$

with

$$\mathbf{A} = \mathbf{A}_{\text{LS}} = \frac{\sigma}{2}\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \qquad \text{and} \qquad \mathbf{a} = \mathbf{A}_{\text{LS}}\frac{\text{E}(2u\mathbf{x})}{\text{E}(\rho_0'u_0)},$$

while $\mathbf{A}_{\text{S}}$ and $\mathbf{a}_{\text{S}}$ remain unchanged with respect to (4.45) and (4.46).

❑

Of course, in absence of heteroskedasticity or if the distribution $F_{0,1}$ of the error terms is symmetric (around 0), the expressions of the asymptotic variances and covariances simplify quite considerably, as shown in Appendix 2. Unfortunately, their estimates — their empirical counterparts — are not robust anymore with respect to (good and bad) leverage points. Hence, Croux et al. (2003) do advise against the use of these simplified variances and covariances, even when the assumptions of absence of heteroskedasticity and symmetry hold.

13. There is *homoskedasticity* when the processes $\mathbf{x}_i$ and (u_i, u_{0i}) are independent.

4.6.2 Robust confidence intervals and tests with robust regression estimators

As just explained, we may consider that, under the model (4.4) and Assumption A[14], a robust M-, S- or MM-estimator $\widehat{\boldsymbol{\beta}}$ is, for large n, approximately normally distributed with mean $\boldsymbol{\beta}$ and variance $\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)$, where $\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)$ corresponds to the empirical counterpart of the asymptotic matrix $\text{Avar}\left(\widehat{\boldsymbol{\beta}}\right)$ specified in the previous subsection. This result underlies the inference procedures developed for linear combinations of the regression parameters.

Inference for a single linear combination of the regression parameters

Let γ be a linear combination of the regression coefficients:

$$\gamma = \mathbf{b}^t \boldsymbol{\beta}$$

with $\mathbf{b}$ a constant (non random) vector. Then the natural estimate of γ is $\widehat{\gamma} = \mathbf{b}^t \widehat{\boldsymbol{\beta}}$, which is, under Assumption A and for large n, approximately $\mathcal{N}\left(\gamma, \widehat{\sigma}_\gamma^2\right)$ where

$$\widehat{\sigma}_\gamma^2 = \mathbf{b}^t \widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right) \mathbf{b}.$$

Hence an approximate two-sided confidence interval for γ with confidence level $(1-\alpha)$ is given by

$$\left[\widehat{\gamma} \pm z_{1-\frac{\alpha}{2}} \widehat{\sigma}_\gamma\right],$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard normal distribution.

Similarly, the test of level α for the null hypothesis $\mathcal{H}_0 : \gamma = \gamma_0$ against the two-sided alternative $\mathcal{H}_1 : \gamma \neq \gamma_0$ has the rejection region

$$|\widehat{\gamma} - \gamma_0| > z_{1-\frac{\alpha}{2}} \widehat{\sigma}_\gamma,$$

or equivalently, since the approximate normal distribution of $\widehat{\gamma}$ implies that $\left(\frac{\widehat{\gamma}-\gamma}{\widehat{\sigma}_\gamma}\right)^2 \approx \chi_1^2$, rejects $\mathcal{H}_0$ when

$$T > \chi^2_{1;1-\alpha}$$

where

$$T = \left(\frac{\widehat{\gamma} - \gamma_0}{\widehat{\sigma}_\gamma}\right)^2$$

and $\chi^2_{1;1-\alpha}$ is the $(1-\alpha)$-quantile of the chisquare distribution with one degree of freedom.

14. Recall that Assumption A specifies that the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are generated by a stationary and ergodic process, and are mutually independent.

In particular, if $\mathbf{b} = (0, \ldots, 0, 1, 0, \ldots, 0)^t$ (all the components of $\mathbf{b}$ are equal to zero except the jth component, equal to 1), we have $\gamma = \beta_j$ and $\widehat{\sigma}^2_\gamma = \left[\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)\right]_{jj}$. Then, the two-sided confidence interval for β_j with confidence level $(1-\alpha)$ is given by

$$\left[\widehat{\beta}_j \pm z_{1-\frac{\alpha}{2}} \sqrt{\left[\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)\right]_{jj}}\right]$$

and the test of level α for the null hypothesis $\mathcal{H}_0 : \beta_j = 0$ against the alternative $\mathcal{H}_1 : \beta_j \neq 0$ has the rejection region

$$\frac{\widehat{\beta}_j^2}{\left[\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)\right]_{jj}} > \chi^2_{1;1-\alpha}.$$

Inference for several linear combinations of the regression parameters

Let us now consider several linear combinations of the β_j's represented by the vector $\boldsymbol{\gamma} = \mathbf{B}\boldsymbol{\beta}$ where $\mathbf{B}$ is a $q \times (p+1)$ matrix of rank q. Then $\widehat{\boldsymbol{\gamma}} = \mathbf{B}\widehat{\boldsymbol{\beta}}$ is, under Assumption A and for large n, approximately $\mathcal{N}_q\left(\boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}\right)$ with

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}} = \mathbf{B}\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}\right)\mathbf{B}^t.$$

This implies that

$$(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^t \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}^{-1} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) \approx \chi^2_q$$

where χ^2_q is the chisquare distribution with q degrees of freedom. Hence, to test the linear hypothesis $\mathcal{H}_0 : \boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ for a given $\boldsymbol{\gamma}_0$, with level α, we may use the test that rejects $\mathcal{H}_0$ if

$$T > \chi^2_{q;1-\alpha}$$

where

$$T = (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)^t \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}^{-1} (\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$$

and $\chi^2_{q;1-\alpha}$ is the $(1-\alpha)$- quantile of the χ^2_q distribution. The most common application of this test is when $\mathcal{H}_0$ is the hypothesis that some of the coefficients β_j are equal to zero. If, for example, the null hypothesis is

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0,$$

then $\boldsymbol{\gamma} = \mathbf{B}\boldsymbol{\beta}$ with $\mathbf{B} = \left(\mathbf{I}_{q\times q}, \mathbf{0}_{q\times(p+1-q)}\right)$ (where $\mathbf{I}_{q\times q}$ is the $(q \times q)$ identity matrix) and $\mathcal{H}_0$ takes the form $\mathcal{H}_0 : \mathbf{B}\boldsymbol{\beta} = \mathbf{0}$.

4.6.3 Robust R^2

The coefficient of determination or R^2 is a very simple tool — probably the most used by practitioners — to assess the quality of fit in a multiple linear regression. It provides

an indication of the suitability of the chosen explanatory variables in predicting the response. In the classical setting, R^2 is usually presented as the quantity that estimates the percentage of variance of the response variable explained by its (linear) relationship with the explanatory variables. It is defined as the ratio

$$\begin{aligned} R^2 &= \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}, \end{aligned} \tag{4.50}$$

where ESS, TSS and RSS are respectively the explained, total and residual sum of squares. Note that $y_i - \widehat{y}_i = r_i\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big)$ are the LS residuals. Moreover, $\overline{y}$ is the LS estimate of $\mu = \text{E}(y)$, that is the LS estimate of the intercept β_0 in the linear regression model (4.1) in which $\beta_1 = \ldots = \beta_p = 0$.

When there is an intercept term in the linear model, this coefficient of determination R^2 is actually equal to the square of the correlation coefficient between the observed y_i's and the predicted $\widehat{y}_i$'s (see, e.g., Greene 1997), i.e.,

$$R^2 = \left(\frac{\sum_{i=1}^n (y_i - \overline{y})\Big(\widehat{y}_i - \overline{\widehat{y}}\Big)}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}\sqrt{\sum_{i=1}^n \Big(\widehat{y}_i - \overline{\widehat{y}}\Big)^2}} \right)^2, \tag{4.51}$$

with $\overline{\widehat{y}}$ the arithmetic mean of the predicted responses. Equation (4.51) has a nice interpretation in that R^2 measures the goodness of fit of the regression model by its ability to predict the response variable, ability measured by the correlation. Note that R^2 is a consistent estimator of the population parameter

$$\phi^2 = \max_{\boldsymbol{\beta}} \text{Corr}^2\big(y, \mathbf{x}^t\boldsymbol{\beta}\big), \tag{4.52}$$

that is, of the squared correlation between y and the best linear combination of the $\mathbf{x}$ (cf. Anderson 1984). In finite samples, R^2 is biased upward and is generally adjusted, e.g.,

$$R^2_{\text{adj}} = 1 - \big(1 - R^2\big)\left(\frac{n-1}{n-(p+1)}\right). \tag{4.53}$$

It is rather obvious that the R^2 given by (4.50) can be driven by extreme observations, not only through the LS estimator $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ used to compute the predicted responses $\widehat{y}_i$, but also through the average response $\overline{y}$ and the possible large residuals $y_i - \widehat{y}_i$ or deviations $y_i - \overline{y}$. Several robust R^2 have then been proposed in the literature (see Renaud and Victoria-Feser 2010). A *robust* R^2 should give an indication of the fit for the *majority* of the data, possibly leaving aside a few outlying observations. In other words, the (robust) goodness-of-fit criterion is used to choose a good model for the majority of the data rather than an "average" model for all the data. Let us focus our attention here on the two robust coefficients of determination available in Stata: R^2_ρ and R^2_w.

If instead of the LS estimate we use an M-estimate (associated with the loss function ρ) with general scale, defined as in (4.16), a robust coefficient of determination can be defined by

$$R^2_\rho = 1 - \frac{\displaystyle\sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^t \widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}}{\widehat{\sigma}}\right)}{\displaystyle\sum_{i=1}^{n} \rho\left(\frac{y_i - \widehat{\mu}_{\mathrm{M};\rho}}{\widehat{\sigma}}\right)}, \tag{4.54}$$

where $\widehat{\mu}_{\mathrm{M};\rho}$ is the M-estimate of the location parameter $\mu = \mathrm{E}(y)$, solution of

$$\arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mu}{\widehat{\sigma}}\right),$$

and $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ and $\widehat{\sigma}$ are robust estimates of $\boldsymbol{\beta}$ and σ for the full model (see Maronna et al. 2006).

Note that, independently, Croux and Dehon (2003) have proposed a class of robust R^2 which generalizes (4.50) given by

$$R^2_{\mathrm{S}} = 1 - \frac{s\left(y_i - \mathbf{x}_i^t \widehat{\boldsymbol{\beta}}\ ; i = 1, \ldots, n\right)}{s(y_i - \widehat{\mu}\ ; i = 1, \ldots, n)} \tag{4.55}$$

where $s(\cdot)$ is a robust dispersion measure (see Croux and Dehon 2003).

Although (4.54) and (4.55) are direct generalizations of (4.50) to the robust framework, they suffer from an important drawback: in practice, they are often biased. One possible reason why this phenomenon happens is that the computation of R^2_ρ or R^2_{S} requires and uses the estimation of two models: the full regression model and a location model. The associate residuals $y_i - \mathbf{x}_i^t \widehat{\boldsymbol{\beta}}$ and $y_i - \widehat{\mu}$ are not influenced by model deviation (as presence of outliers, for instance) in the same way, so that bounding these quantities directly and separately is not necessarily appropriate in the regression model framework.

To remedy this problem, Renaud and Victoria-Freser (2010) have proposed to "robustify" the expression (4.51) of the coefficient of determination. Suppose $\boldsymbol{\beta}$ has been estimated by an M-, S- or MM-estimate $\widehat{\boldsymbol{\beta}}$ using a loss function $\rho(\cdot)$, and let $\widehat{\sigma}$ be the final robust estimate of the scale parameter σ. Let, as usual, $\psi(u) = \rho'(u)$ for $u \in \mathbb{R}$. Define, as in section 4.4.4, the weight function W by

$$W(u) = \begin{cases} \frac{\psi(u)}{u} & \text{if } \ u \neq 0 \\ \psi'(0) & \text{if } u = 0, \end{cases}$$

and the weights

$$w_i = W\left(\frac{r_i\left(\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}}\right), \quad i = 1, \ldots, n.$$

Note that these weights w_i coïncide with those used in the last iteration of the *iteratively reweighted least squares algorithm* used to implement the M-estimation procedure. In particular, if $\rho(\cdot)$ is the Tukey-Biweight function $\rho_\kappa^B(\cdot)$ given by (4.17), we have

$$w_i = \begin{cases} \left(1 - \left(\dfrac{r_i\left(\widehat{\boldsymbol{\beta}}\right)}{\kappa\widehat{\sigma}}\right)^2\right)^2 & \text{if } \left|\dfrac{r_i\left(\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}}\right| \leq \kappa \\ 0 & \text{if } \left|\dfrac{r_i\left(\widehat{\boldsymbol{\beta}}\right)}{\widehat{\sigma}}\right| > \kappa. \end{cases}$$

Then a robust version of (4.51) is given by

$$R_w^2 = \left(\frac{\sum_{i=1}^n w_i(y_i - \overline{y}_w)\left(\widehat{y}_i - \overline{\widehat{y}}_w\right)}{\sqrt{\sum_{i=1}^n w_i(y_i - \overline{y}_w)^2}\sqrt{\sum_{i=1}^n w_i\left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2}}\right)^2, \tag{4.56}$$

where $\widehat{y}_i = y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}$ $(i = 1, \ldots, n)$, $\overline{y}_w = (1/\sum w_i)\sum w_i y_i$ and $\overline{\widehat{y}}_w = (1/\sum w_i)\sum w_i\widehat{y}_i$.

With the same weights and predictions, another robust coefficient of determination can be defined from (4.50):

$$\widetilde{R}_w^2 = 1 - \frac{\sum_{i=1}^n w_i(y_i - \widehat{y}_i)^2}{\sum_{i=1}^n w_i(y_i - \overline{y}_w)^2}. \tag{4.57}$$

It is shown in Renaud and Victoria-Feser (2010) that

$$R_w^2 = \widetilde{R}_w^2.$$

Renaud and Victoria-Freser (2010) have also proposed the following more general formulation for a robust coefficient of determination in order to take into account consistency considerations:

$$\widetilde{R}_{w,a}^2 = \frac{\sum_{i=1}^n w_i\left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2}{\sum_{i=1}^n w_i\left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2 + a\sum_{i=1}^n w_i(y_i - \widehat{y}_i)^2}, \tag{4.58}$$

where a is a constant factor. It has been shown that R_w^2 and $\widetilde{R}_w^2$ are both equal to $\widetilde{R}_{w,a}^2$ with $a = 1$. Moreover, with no assumption on the distribution of the explanatory variables, but under the assumption of normality of the errors and for a consistent estimator $\widehat{\sigma}$ of the residual scale, $\widetilde{R}_{w,a}^2$ is a consistent estimator of the population coefficient of determination (4.52) if we take[15]

$$a = \frac{\mathrm{E}\left[\frac{\psi(u)}{u}\right]}{\mathrm{E}[\psi(u)]}, \quad \text{with } u \sim \mathcal{N}(0,1).$$

15. For example, choosing, in particular, $\psi(u) = \psi_\kappa^B(u) = \left(\rho_\kappa^B\right)'(u)$ where ρ_κ^B is the Tukey-Biweight loss function with $\kappa = 4.685$, leads to $a = 1.2076$.

As shown by a simulation study in Renaud and Victoria-Feser (2010), for small samples and a relatively large number of covariates, using the same rationale than for the classical R^2, the robust coefficient might benefit of being adjusted, hence leading to the adjusted coefficient

$$\widetilde{R}^2_{w,a;\text{adj}} = 1 - \left(1 - \widetilde{R}^2_{w,a}\right)\left(\frac{n-1}{n-(p+1)}\right). \tag{4.59}$$

4.6.4 Extension of the Hausman test to check for the presence of outliers

In practice, it is usual to ask oneself if it is necessary to use a robust regression estimator or if it is preferable to use a classical estimator that is more efficient under the model and more easy to compute. When the data are not contaminated by outliers, classical and robust estimations of the regression coefficients are quite similar, while a moderate contamination of the sample may imply a possible clear difference between classical and robust estimations. Hence, no significant difference between the classical and robust estimations of $\boldsymbol{\beta}$ may lead us to conclude that the data do not contain outliers or that the influence of the outliers is rather limited: in such a case, we will prefer to retain the classical estimator given its higher efficiency (its higher statistical precision). On the contrary, a significant difference between the classical and robust estimations of $\boldsymbol{\beta}$ indicates that the data are contaminated by outliers in such a way that it biases the classical estimator: a robust estimator should then be preferred.

But which tool may we use to compare adequately two regression estimators and to judge if their values are significantly different or not?

To solve this question, Dehon *et al.* (2009, 2012) have proposed a statistical test, based on the methodology developed by Hausman (see Hausman 1978). Their testing procedure allows to compare a robust S-estimate and the classical LS estimate (in order to detect the presence of outliers). But it also allows to compare an S-estimator with an MM-estimator with a given efficiency level; repeating this test by considering different efficiency levels for the MM-estimator may be seen as a procedure allowing, in the presence of moderate contamination of the sample, to find in an appropriate way the maximum efficiency level that may have this MM-estimator without suffering from too large bias.

In all the cases, the problem of test may be formalized as follows. Consider the regression model (4.1). The null hypothesis $\mathcal{H}_0$ is that this model is valid for the entire population. Thus, at the sample level, under the null, no outliers are present. The alternative hypothesis $\mathcal{H}_1$ is that the model is misspecified for a minority of the population, implying a potentially moderate contamination of the sample. Note that we will also systematically consider that, under the null hypothesis $\mathcal{H}_0$, Assumption A1 is satisfied.

Before to describe the test statistics and the decision rules, let us precise a last point: since Gervini and Yohai (2002) showed that, *in the presence of outliers*, only

the p slopes $\beta_1, \ldots, \beta_p$ of the regression model can be satisfactorily estimated when the error distribution is assymmetric, the test will be based on the comparison of the slopes estimations and the estimations of the intercept β_0 will be disregarded. Hence, in the sequel of this section, we will use the following notations to take this characteristic into account: $\underline{\boldsymbol{\beta}} = (\beta_1, \ldots, \beta_p)^t$ and $\widehat{\underline{\boldsymbol{\beta}}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^t$, such that $\boldsymbol{\beta} = \left(\beta_0, \underline{\boldsymbol{\beta}}^t\right)^t$ and $\widehat{\boldsymbol{\beta}} = \left(\beta_0, \widehat{\underline{\boldsymbol{\beta}}}^t\right)^t$.

Some preliminary results

The development of the tests proposed in Dehon et al. (2012) relies on the results presented in Subsection 4.6.1 providing the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_{\text{LS}}$, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ under $\mathcal{H}_0$ and Assumption A. In Subsection 4.6.1, to avoid any ambiguity, the regression parameters vector was denoted by $\boldsymbol{\beta}_0$ if it was estimated by the S-estimator $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$, and by $\boldsymbol{\beta}$ if it was estimated by the MM-estimator $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$. From now on, we will exclusively denote the regression parameters vector in model (4.4) by $\boldsymbol{\beta}$ as soon as there is no risk of confusion anymore.

We have seen that, under $\mathcal{H}_0$ and Assumption A, for large n,

$$\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho} \approx \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)\right),$$
$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} \approx \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)\right)$$

and

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} \approx \mathcal{N}_{p+1}\left(\boldsymbol{\beta}, \text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{LS}}\right)\right),$$

where the matrices $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$, $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$ and $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{LS}}\right)$ are given by (4.40), (4.41) and (4.47), respectively. Moreover,

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\\ &\approx \mathcal{N}_{p+1}\left(\mathbf{0}, \text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right) + \text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right) - 2\text{Acov}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)\right),\end{aligned}$$

where $\text{Acov}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$ is given by (4.42). Since $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$, $\text{Avar}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$ and $\text{Acov}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$ may be consistently estimated by their empirical counterparts[16] $\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$, $\widehat{\text{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$ and $\widehat{\text{Acov}}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\right)$, we have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho} \approx \mathcal{N}_{p+1}\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)}\right)$$

16. As explained in Subsection 4.6.1, these empirical counterparts are simply obtained by replacing, in u and u_0, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ and σ by the estimates $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$, and the mathematical esperance $\text{E}(\cdot)$ by $\frac{1}{n}\sum_{i=1}^{n}(\cdot)$.

where

$$\widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)} = \widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right) + \widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right) - 2\widehat{\mathrm{Acov}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right). \tag{4.60}$$

If we only consider the slopes estimates, we simply have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho} \approx \mathcal{N}_p\left(\underline{\mathbf{0}}, \underline{\widehat{\boldsymbol{\Sigma}}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)}\right) \tag{4.61}$$

where $\underline{\widehat{\boldsymbol{\Sigma}}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)}$ is the matrix $\widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\right)}$ without its first line and its first column.

Following a similar approach, we have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\mathrm{LS}} \approx \mathcal{N}_{p+1}\left(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)}\right)$$

with

$$\widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)} = \widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right) + \widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right) - 2\widehat{\mathrm{Acov}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right) \tag{4.62}$$

where $\widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right)$, $\widehat{\mathrm{Avar}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)$ and $\widehat{\mathrm{Acov}}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right)$ are the empirical counterparts of the matrices $\mathrm{Avar}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right)$, $\mathrm{Avar}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)$ and $\mathrm{Acov}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\right)$ given by (4.41), (4.47) and (4.49), respectively. As a consequence, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{LS}} \approx \mathcal{N}_p\left(\underline{\mathbf{0}}, \underline{\widehat{\boldsymbol{\Sigma}}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)}\right) \tag{4.63}$$

where $\underline{\widehat{\boldsymbol{\Sigma}}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)}$ is the matrix $\widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)}$ without its first line and its first column.

Comparison of LS and S

Let us consider the classical LS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and the S-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ associated with the loss function $\rho_0(\cdot)$. As already mentioned, the choice of ρ_0 is crucial to guarantee robustness. The function ρ_0 usually used in the present context is the Tukey-Biweight function (4.17): if the tuning constant κ is set at 1.547, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ has a breakdown point equal to 50% (but a rather low Gaussian efficiency of only 28%). Under the null hypothesis (and Assumption A), $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ are both consistant estimators of $\boldsymbol{\beta}$, but $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ has a higher Gaussian efficiency. Under the alternative hypothesis of a moderate contamination, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ still converges to $\boldsymbol{\beta}$ (see Omelka and Salibian-Barrera 2010) but it is not the case for $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ anymore (the outliers distort the LS estimate and introduce a bias, in such a way that $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ possesses another limit in probability than $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$).

The test statistics proposed by Dehon *et al.* (2012) to check if the LS and S-estimates of the regression coefficients are statistically different is defined as

$$H = \left(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{LS}}\right)^t \widehat{\underline{\boldsymbol{\Sigma}}}^{-1}_{(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{LS}})} \left(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{LS}}\right), \tag{4.64}$$

with $\widehat{\boldsymbol{\Sigma}}_{\left(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{LS}}\right)}$ computed from (4.62). It follows from (4.63) that H is, under the null hyptothesis $\mathcal{H}_0$ (and Assumption A), asymptotically distributed as a χ^2_p (a chisquare distribution with p degrees of freedom). Consequently, we may consider that the classical estimate and the S-estimate of the regression slopes are significantly different, and hence decide to reject the null hypothesis $\mathcal{H}_0$, if

$$H > \chi^2_{p;1-\alpha},$$

where α is the given significance level and $\chi^2_{p;1-\alpha}$ is the $(1-\alpha)$- quantile of the χ^2_p distribution.

Comparison of S and MM

Suppose now that the previous test has rejected the null hypothesis $\mathcal{H}_0$: the significant difference between $\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{LS}}$ and $\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0}$ indicates the presence of influential outliers in the sample and a robust regression estimator should then be preferred. In this case, it might be a good strategy to replace the S-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ by an MM-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$, since a good choice of the loss function $\rho(\cdot)$ allows this MM-estimator to reach a much higher efficiency than the initial S-estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$[17]. For instance, if we take for ρ the Tukey-Biweight function (4.17) with the tuning constant κ equal to 4.685, the Gaussian efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ attains 95%, and for $\kappa = 6.256$, the Gaussian efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ is equal to 99%. However, as already mentioned when we have studied the MM-estimation procedure, it is not advised to consider too highly efficient MM-estimators: indeed, a moderate contamination of the sample induces a bias for $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and, for a fixed sample, this bias grows when the efficiency of the estimator raises (see Maronna et al. 2006 and Omelka and Salibian-Barrera 2010). As a consequence, it is of the utmost importance to find the highest efficiency we may fix for the MM-estimator without paying the price of an excessive bias.

The statistical comparison of $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ (with a fixed value of the tuning constant κ for the loss function ρ, hence a fixed Gaussian efficiency for the MM-estimator) can be made using the statistics

$$H = \left(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho}\right)^t \widehat{\underline{\boldsymbol{\Sigma}}}^{-1}_{(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho})} \left(\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho}\right), \tag{4.65}$$

with $\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho})}$ given by (4.60). Under the null hypothesis $\mathcal{H}_0$, $\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0}$ and $\widehat{\underline{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho}$ are both consistant estimators of $\underline{\boldsymbol{\beta}}$ and $H \approx \chi^2_p$. Under the alternative hypothesis $\mathcal{H}_1$,

17. Recall here that $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ possesses the same breakdown point as $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$.

i.e., under a moderate contamination of the sample, the bias of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ risks to be large (the magnitude of the bias depends of the fixed efficiency of the MM-estimator) and a potentially significant difference may appear between the S-estimate and the MM-estimate of the regression slopes. As a consequence, we will decide to reject $\mathcal{H}_0$ — that is, in practice, we will conclude that the contamination of the sample by outliers significantly biases $\underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{MM};\rho_0,\rho}$ and hence distorts the MM-estimation with respect to the S-estimation — if

$$H > \chi^2_{p;1-\alpha},$$

where α is the chosen significance level. Dehon *et al.* (2012) propose to repeat this test by considering successively different values for the constant κ in function ρ (that is, different levels for the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$) and to retain ultimately the MM-estimator that, while not being significantly different from $\underline{\widehat{\boldsymbol{\beta}}}_{\mathrm{S};\rho_0}$ and hence not rejecting the null, has the highest efficiency. This way of proceeding allows to find heuristically the highest efficiency that may have the MM-estimator without suffering from an excessive bias in presence of moderate contamination of the sample by outliers.

4.7 Examples

Comparing estimators

In the first example, we will use a dataset made available by Rousseeuw and Leroy (1987). The dataset contains 47 stars in the direction of Cygnus. The explanatory variable is the logarithm of the effective temperature at the surface of the star (T_e), and the dependent variable is the logarithm of its light intensity (L/L_0). In the scatterplot of figure 4.4, it is evident that some stars (represented by hollow circles) have a very different behavior than the bulk of the data. To illustrate graphically the influence that these stars have on the estimation of the regression line, we superpose to the scatterplot two lines estimated by (1) ordinary least squares (solid line) and (2) a robust regression estimation method (more precisely, S-estimation; dashed line).

46: Graph-Ex-1.do was missing; I recreated the graph from the description.

To obtain the LS estimates for the intercept and slope of the regression of log intensity on log temperature in Stata, without making any difference between stars, we can type:

```
. use Star.dta, clear
. regress log_intensity log_temperature
```

Source	SS	df	MS	Number of obs	=	47
				F(1, 45)	=	2.08
Model	.664593334	1	.664593334	Prob > F	=	0.1557
Residual	14.3463934	45	.318808743	R-squared	=	0.0443
				Adj R-squared	=	0.0230
Total	15.0109868	46	.326325799	Root MSE	=	.56463

log_intensity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
log_tempera~e	-.4133041	.2862575	-1.44	0.156	-.9898562	.163248
_cons	6.793468	1.236516	5.49	0.000	4.302998	9.283939

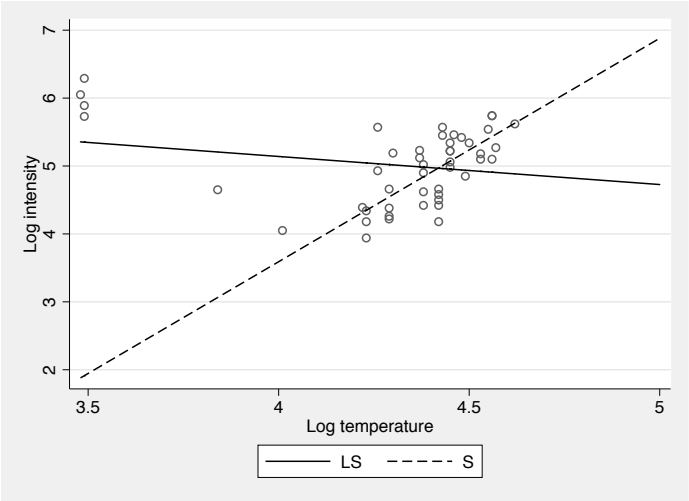

Figure 4.4. Caption needed

The results of the LS estimation (solid line in figure 4.4) indicate that, against intuition, if the temperature of a star increases, its light intensity decreases in average (although the effect is not significantly different from zero). If instead of using a classical estimator we use a robust estimator, the result changes drastically. For illustrative purposes we will estimate the above model using LTS, LMS, M, GM, S and MM (with an efficiency fixed at 95%) estimators. However, since it has now been widely accepted that S and MM estimators are preferable to LTS and LMS because of their higher efficiency and to M and GM estimator because of their higher robustness with respect to outliers (see above), in the subsequent examples we will focus exclusively on S and MM estimates.

LTS

A robust LTS estimator can be easily fit using the `robreg` package running the following command:

```
. robreg lts log_intensity log_temperature
enumerating 500 samples (percent completed)
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
LTS regression                                  Number of obs     =         47
                                                  Breakdown point =         .5
                                                  Subsamples      =        500
                                                  Scale estimate  =  .40961741
```

log_intensity	Coef.
log_tempera~e	4.727267
_cons	-15.81634

The results from the LTS estimator are very different from those obtained for the LS estimation. Indeed what we observe here is that if the log of the temperature of a star increases, its luminosity will increase as well. In term of size of effect, the LTS estimator suggests that an increase of 100% of the temperature is associated to an increase of the luminosity of approximately 473%.

LMS

If instead of the LTS estimator we wish to use the LMS estimator, we can type:

```
. robreg lms log_intensity log_temperature
enumerating 500 samples (percent completed)
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
LMS regression                              Number of obs     =         47
                                              Subsamples      =        500
                                              Scale estimate  =  .36429398
```

log_intensity	Coef.
log_tempera~e	3.636368
_cons	-11.20184

Even if the size of effect seems to be slightly smaller than with LTS, the sign of the relation is the same pointing towards a positive association between log-temperature and lightness of stars.

M-estimator

If we use the M estimator (with a Huber loss function), we do not expect the estimation to resist to outliers. Indeed, in the theoretical section it has been shown how this estimator resists to vertical outliers but not to bad leverage points (i.e. points outlying in the space of the explanatory variables). As expected, the M estimation provides results very similar to those of LS and we can conclude that the estimator breaks down.

The command to run the M estimator is:

```
. robreg m log_intensity log_temperature
fitting initial LAV estimate ... done
iterating RWLS estimate ........... done
```

```
M-Regression (95% efficiency)                   Number of obs     =         47
                                                  Huber k         =  1.3449986
                                                  Scale estimate  =  .63061122
                                                  Robust R2 (w)   =  .04761698
                                                  Robust R2 (rho) =  .03486282
------------------------------------------------------------------------------
log_intensity |      Coef.   Robust Std. Err.      z    P>|z|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
log_tempera~e |  -.4235066   .3992983    -1.06   0.289    -1.206117    .3591036
        _cons |   6.84754   1.758148     3.89   0.000     3.401634    10.29345
------------------------------------------------------------------------------
```

GM-estimator

The Generalized M estimate is slightly more complicated to compute than the M estimate. We first need to estimate the outlyingness of each individual in the x-dimension, and then downweight leverage points while estimating the model using an M estimator. In this example, given that there is a single explanatory variable, the outlyingness in the horizontal dimension can be measured by centering the data around a robustly estimated location parameter (e.g. the Hodges-Lehman estimate or the median) and reducing it using a robustly estimated measure of dispersion (e.g. the Croux and Rousseeuw Q_n estimate). In the case of multiple explanatory variables, the outlyingness in the space of the explanatory variables will have to be measured using robust multivariate estimates of location and scatter described in chapter XXX. As far as the down-weighting scheme for outliers is concerned, several alternatives have been proposed in the literature. In this example we award a weight equal to zero to any star associated to a leverage larger than 2.5 and equal to one otherwise. Given that there is one single explanatory variable, the GM estimator should behave satisfactory.

47: To be updated

The commands used for GM estimation are:

```
hl log_temperature
local mu=e(hl)
qn log_temperature
local s=e(qn)
gen leverage=(log_temperature-`mu')/`s'
robreg m log_intensity log_temperature if abs(leverage)<=2.5
```

S-estimator

If we estimate the model using an S estimator, we do not expect to have large differences with respect to LTS, LMS and GM in terms of point estimates. However, its higher efficiency makes it theoretically more appealing. The command to run the S estimator is:

```
. robreg s log_intensity log_temperature
enumerating 50 candidates (percent completed)
```

```
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
refining 2 best candidates ... done
S-Regression (28.7% efficiency)                 Number of obs     =         47
                                                  Subsamples      =         50
                                                  Breakdown point =         .5
                                                  Bisquare k      =   1.547645
                                                  Scale estimate  =  .47145696
```

log_intensity	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
log_tempera~e	3.290339	1.64075	2.01	0.045	.0745278	6.506151
_cons	-9.570732	7.373867	-1.30	0.194	-24.02325	4.881783

The results indicate that if the temperature of a star doubles, its light intensity increases by approximately 329%. As stated in the theoretical section, the gaussian efficiency of the S estimator with a 50% breakdown point (and a Tukey biweight loss function) is only 28%. In order to increase the efficiency while keeping the breakdown point at 50%, we can use MM estimators.

MM-estimator

It is well-known that even if an MM estimator has a breakdown point of 50%, it can be associated to a relatively large bias if its efficiency is set too high. As explained in Subsection 4.6.4, a general procedure is therefore to compare the MM estimate with a given level of efficiency to the S-estimate, and see if there is a significant difference. If the difference is small, this means that the bias should not be too big.

We compute here an MM estimator with an efficiency set at 95%:

```
. robreg mm log_intensity log_temperature, efficiency(95)
Step 1: fitting S-estimate
enumerating 50 candidates (percent completed)
0 ——— 20 ——— 40 ——— 60 ——— 80 ——— 100
..................................................
refining 2 best candidates ... done
Step 2: fitting redescending M-estimate
iterating RWLS estimate .............. done
MM-Regression (95% efficiency)                  Number of obs     =         47
                                                  Subsamples      =         50
                                                  Breakdown point =         .5
                                                  M-estimate: k   =   4.685045
                                                  S-estimate: k   =   1.547645
                                                  Scale estimate  =  .47145688
                                                  Robust R2 (w)   =  .41883093
                                                  Robust R2 (rho) =  .02050865
```

	Robust

log_intensity	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
log_tempera~e	2.253165	.7690643	2.93	0.003	.7458263	3.760503
_cons	-4.969402	3.410051	-1.46	0.145	-11.65298	1.714175

We see that the MM estimation leads to results comparable to the S estimation in terms of point estimates but is associated to a much higher efficiency. As explained above, a formal test could have been used but we leave this for another example. The MM estimated model suggests that an increase of 100% of the temperature of a star is associated with an increase of its luminosity by approximately 225%. In terms of the quality of the fit, if we rely on the robust $R^2(w)$ described previously, we see that the model is pretty good in predicting the luminosity of stars for the vast majority of the observations. Indeed close to 42% of the variations in terms of light intensity for the vast majority of the observations can be explained by the differences in temperatures.

Identifying outliers

In this second example where the objective is to unmask outliers, we use a dataset made available by Jeffrey D. Sachs and Andrew M. Warner in their article "Natural Resource Abundance and Economic Growth" (1997). In this paper, the authors show that economies with a high ratio of natural resource exports to GDP in 1970 (the base year) tended to grow slowly during the subsequent 20-year period 1970-1990. In the article the authors aknowledge the existence of outliers and try to deal with them working with differences in fits. More precisely, they look at how the predicted value for each observation varies when this specific observation is removed from the sample when fitting the model and compare the results with the model estimated using all of the observations. They expect to see big differences in fits for outlying individuals. However, if there are clusters of outliers, atypical individuals will mask one the other and will most probably not be detected with this technique. The outliers they identify are Chad, Gabon, Guyana, and Malaysia.

We propose here to use another procedure to identify the outliers. This procedure is simply based on the examination of the standardized residuals related to a regression S-estimator.

To identify the outliers, we fist estimate the regression model by running the command :

```
robreg s gea7090 lgdpea70 sxp sopen linv7089 rl dtt7090
```

The results of this S-estimation are presented in Figure ?.

[Insert here S graph from Ch3-Ex-2.do]

We ask for the predicted values of the dependent variable by the command: `predict yhat`. The robust residuals can then be easily determined using the command: `gen res=yhat-gea7090`. They are standardized by dividing them by the scale parameter

estimated from the regression S-estimate:`replace res=res/e(scale)`. We can now plot the standardized residuals and identify those that are larger or smaller than two given cut-off points corresponding to two specific quantiles of the normal distribution. We use here the percentiles 2.5 and 97.5 which are respectively equal to -1.96 and 1.96.

We see in Figure ?that, among the four countries identified as outliers by Sachs and Warner, only Malaysia is still emerging as outlier when using the S-estimation procedure. On the other hand, other countries such as Hong Kong, Ecuador or Iran seem to be atypical countries in the S-regression but were not detected by the original authors.

[Insert here S graph from Ch3-Ex-2.do]

▷ Example

Testing for the presence of outliers and setting the efficiency for MM-estimation. For this example, we again use the dataset relating the logarithm of the effective temperature at the surface of the star (explanatory variable T_e) and the logarithm of its light intensity (dependent variable L/L_0). The first question one might raise is: is there a significant difference between the classical estimate and the robust one? To answer this question we simply compute an S-estimate using the `robreg s`command and use `hausman`as an option. This implies that the testing procedure comparing the S-estimate with the LS estimate (see Subsection 4.6.4) is implemented.

[Insert here S regression from Ch3-Ex-3.do]

The results of the Hausman test (see Figure ?) clearly indicate that the difference between the S-estimate and the LS-estimate is significant (p-value$<$0.05) and thus that outliers distort the LS estimation. We should therefore use a robust estimator.

As stated previously S-estimators are very robust against outlier contamination but are relatively inefficient. MM-estimators on the other hand are more efficient than S-estimators but might be associated with a large bias if efficiency is set too high. To choose the level of efficiency to use in practice we have to apply the testing procedure described in Subsection 4.6.4 that compares the MM-estimates related to some given levels of efficiency with respect to an S-estimate. We can then finally set the efficiency of the MM-estimator at the highest efficiency level that does not lead to a rejection of the equality between the MM-estimate and the S-estimate. Doing this in practice is very simple as the testing procedure is implemented in the `robreg mm`command. For example, if the `robreg mm` command is run with the efficiency set at 75% , the `hausman`option compares the MM-estimate with 75% efficiency to the S-estimate obtained at the first step of the MM-estimation procedure. Similarly, if the efficiency is set at 85%, the `hausman`option compares the MM-estimate with 85% efficiency to the S-estimate, and so on. In this example we control if we can set the efficiency at 75%, 85%, 95%, and 99%. We obtain the following results:

- **MM-estimation with 75% efficiency:**

  `robreg mm log_intensity log_temperature, hausman efficiency(75)`

  [Insert here MM(75) regression from Ch3-Ex-3.do]

  The results of the Hausman test (see Figure ?) indicate that there is no significant difference between the MM-estimate and the S-estimate. It would therefore be preferable to work with the MM-estimator with an efficiency equal to 75% as it provides results comparable to the S-estimator in terms of bias but has a much higher efficiency.

- **MM-estimation with 85% efficiency:**

  `robreg mm log_intensity log_temperature, hausman efficiency(85)`

  [Insert here MM(85) regression from Ch3-Ex-3.do]

  Here again the Hausman test statistics takes a low value which tells us that there is no significant difference between the MM-estimate and the S-estimate.

- **MM-estimation with 95% efficiency:**

  `robreg mm log_intensity log_temperature, hausman efficiency(95)`

  [Insert here MM(95) regression from Ch3-Ex-3.do]

  If we set the efficiency of the MM-estimator to 95%, we still do not observe any significant difference between the MM-estimate and the S-estimate.

- **MM-estimation with 99% efficieny:**

  `robreg mm log_intensity log_temperature, hausman efficiency(99)`

  [Insert here MM(99) regression from Ch3-Ex-3.do]

  In the case of an efficiency of the MM-estimator equal to 99%, the Hausman test rejects the null hypothesis of equality between the MM-estimate and the S-estimate, which means that for this very high level of efficiency, the MM-estimator suffers from a too large bias.

  To summarize, it is clear that a classical estimator cannot be used in the example because the LS estimates are clearly distorted. A robust estimator should be preferred. We may use an MM-estimator with an efficiency equal to 95% instead of the less efficient S-estimator since despite the bias from which the MM-estimator potentially suffers, the MM-estimates of the regression parameters appear no significantly different from the S-estimates. It is not recommended to consider a higher level of efficiency for the MM-estimator (99%, for instance), since the statistical test indicates that the bias becomes too big in that case.

◁

▷ **Example**

Recognizing the type of outliers For this example, we will use the very famous auto dataset available from Stata. This dataset contains the price of a set of cars as well as a series of characteristics. To see if outliers are present in the dataset, we regress the price on all the available characteristics and compute the robust standardized residuals. Obviously this will not allow to recognize the types of outliers. To do so, we will use the graphical tool of Rousseuw and Van Zomeren (1990). The idea here is to use a scatter plot considering on the vertical dimension the standardized residuals and on the horizontal dimension the leverage of the observations measured using the robust Mahalanobis distance (as described in (4.26)). For gaussian data it is well known that the standardized residuals are normally distributed while the robust distances are distributed as a χ^2_p where p is the number of continuous explanatory variables. It is then natural to compare the standardized residuals and the leverages to some specific quantiles of the $\mathcal{N}(0,1)$ or χ^2_p distributions in order to detect if an individual has to be considered as an outlier and, if it is the case, to which type of outlier it corresponds. We decide to choose here the 2.5th and 97.5th percentiles of the $\mathcal{N}(0,1)$ distribution, and the 95th percentile of the χ^2_p distribution. Those individuals leading to small robust standardized residuals in absolute value and small leverages are considered as standard individuals; those giving large standardized residuals in absolute value and large leverages are defined as bad leverage points; those that coincide with large standardized residuals in absolute value but small leverages are considered as vertical outliers and, finally, those that give small standardized residuals in absolute value but large leverages are good leverage points.

[Insert here MM(95) regression from Ch3-Ex-4.do]
[Insert here MM(95) graphn from Ch3-Ex-4.do]

Figure ?clearly shows, for example, that the Cadillac Seville is a bad leverage point. This auto is indeed associated with a very large positive robust standardized residual and has a big leverage effect which means that its characteristics in the space of the explanatory variables are very different from the bulk of the data. On the other hand, the Cadillac Eldorado, the Lincoln Versaille and some other cars have a small leverage effect — their characteristics do not appear as different from the vast majority of the observations — but are highly overpriced given their large positive residuals; these cars are identified as vertical outliers. Finally some other cars such as the Plymouth Arrow or the Volkswagen (VW Diesel) *inter alia* are not outliers in terms of prices but have characteristics very different from the others. They are thus good leverage points. Note that even if these good leverage points do not have major effect on the estimation of the slope parameter and the constant, they might affect inference and shrink standard errors. It is hence important for researchers to identify them.

◁

▷ **Example**

Dealing with dummies. For this example, we use the "fertil1.dta" data set pro-

vided provided by Wooldridge (2001) which is a pooled cross section on more than a thousand U.S. women for the even years between 1972 and 1984. These data are used to study the relationship between women's education and fertility. We estimate a model relating the number of children ever born to a woman (kids) to the years of education, age, age squared, regional dummies, race dummies, the type of environment in which the women have been reared and year dummies, using an MS-estimator. Given the large number of dummy variables, it is very likely that the subsampling algorithm described in Subsection 4.5.3 leads to perfectly collinear subsamples. Using an MS-estimator should tackle the problem.

[Insert here MS regression from Ch3-Ex-5.do]

The results presented in Figure fig:fertil1_MS_results clearly point towards a robust and statistically significant negative relationship between education and fertility. Indeed, each additionnal year of schooling is associated to an average reduction of fertility (i.e. number of children) equal to 0.19. To identify the outliers and recognize their type, we again call on the graphical tool proposed by Rousseuw and Van Zomeren (1990). The only difference with the previous example is that dummy explanatory variables cannot create any leverage effect and should therefore be treated differently from the other explanatory variables. To estimate robust distances, we rely on the Stahel and Donoho multivariate estimator of location and scatter (this estimator will be described in details in Chapter ??). The latter is a projection based estimator that allows the pratialling out of dummy variables to calculate leverage effects. As before we can choose a quantile above which individuals can be seen as potentially outlying. We use here the 0.5th and 99.5th percentiles of the $\mathcal{N}(0,1)$ distribution as cut-off points for the robust standardized residuals, and the 99th percentile of the chi-square distribution with p_1 degrees of freedom, where p_1 is the number of continuous explanatory variables, as cut-off point for the robust distances. In Figure , we highlight the women for which the robust standardized residuals and (or) the robust distances exceed the cut-off points.

[Insert here graphn from Ch3-Ex-5.do]

It is evident that individuals such as 565 have more children that one would expect given their characteristics (which are not quite different from the bulk of the data). On the other hand individuals such as 706, 767 or 1063 have characteristics that are very different from the vast majority of the individuals; however their number of children is in accordance with her characteristics. Finally individual such as 519, or 490 or 967 have characteristics that are very different from the others. The first one has a number of children that is much smaller than one would expect according to the estimated model while the two others have more children than expected.

◁

4.8 Appendix 1: M-estimators of location and scale

The application of the M-estimation approach in the particular case of the location-scale model (4.7) leads to the M-estimators of location and scale.

4.8.1 M-estimator of location

An M-estimate $\widehat{\mu}_{\mathrm{M};\rho}$ of μ is defined by

$$\widehat{\mu}_{\mathrm{M};\rho} = \arg\min_{\mu} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mu}{\widehat{\sigma}}\right)$$

where $\rho(\cdot)$ is a loss function that is positive, even (such that $\rho(0) = 0$) and not decreasing for positive values u, and $\widehat{\sigma}$ is a preliminary robust estimate of σ if this scale parameter is unknown (the MAD, for example). We may also characterize $\widehat{\mu}_{\mathrm{M};\rho}$ as a solution of the following estimating equation:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mu}{\widehat{\sigma}}\right) = 0, \tag{4.66}$$

where $\psi(u) = \rho'(u)$.

Taking $\rho(u) = u^2$, we obtain $\psi(u) = 2u$ and hence

$$\sum_{i=1}^{n} (y_i - \widehat{\mu}_{\mathrm{M};\rho}) = 0,$$

implying that $\widehat{\mu}_{\mathrm{M};\rho} = \frac{1}{n}\sum_{i=1}^{n} y_i = \widehat{\mu}_{\mathrm{LS}}$. Taking $\rho(u) = |u|$, we have $\psi(u) = \mathrm{sgn}(u)$ and $\sum_{i=1}^{n} \mathrm{sgn}(y_i - \widehat{\mu}_{\mathrm{M};\rho}) = 0$; this leads to $\widehat{\mu}_{\mathrm{M};\rho} = \mathrm{med}\{y_i\} = \widehat{\mu}_{\mathrm{L}_1}$.

In general, if ψ is not redescending, the equation (4.66) may be solved using the Newton-Raphson algorithm with a robust estimate of μ — the empirical median $\mathrm{med}\{y_i\}$, for instance — as initial value for μ.

The influence function of the functional T associated to the location M-estimator $\widehat{\mu}_{\mathrm{M};\rho}$ under the distribution $F_{0,1}$ of the error term ν in the location-scale model — recall here that $F_{0,1}$ is assumed to be symmetric around zero — takes the form:

$$\mathrm{IF}(u; T, F_{0,1}) = \frac{\psi(u)}{\mathrm{E}_{F_{0,1}}[\psi'(\nu)]}.$$

Consequently, the choice of the function ρ, and hence of the function ψ, completely conditions the form of the influence function.

Moreover, it has been proven that an univariate location M-estimator has an asymptotic breakdown point equal to 50% whenever the function ψ is *non decreasing*, bounded

and symmetric, and the preliminary estimator of the scale parameter σ is the MAD[18] (see Huber and Ronchetti 2009, 54). The asymptotic breakdown point is nul if ψ is unbounded. If ψ is equal to the function ψ_κ^{B} and hence is redescending, the breakdown point of $\widehat{\mu}_{\text{M};\rho}$ is strictly smaller than 50% and depends upon the breakdown point of the preliminary scale estimator, upon the constant κ, but also upon the configuration of the sample (see Maronna et al. 2006, 78)[19].

4.8.2 M-estimator of scale

A M-estimate $\widehat{\sigma}_{\text{M};\rho}$ of the scale parameter σ is defined as the solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\widehat{\mu}}{\sigma}\right)=\delta \tag{4.67}$$

where $\rho(\cdot)$ is a loss function that is positive, even, not decreasing for positive values and bounded, and $\widehat{\mu}$ is a preliminary robust estimate of μ if this location parameter is unknown (the median, for instance). To ensure the consistency of $\widehat{\sigma}_{\text{M};\rho}$ for σ, we have to take $\delta = \text{E}_{F_{0,1}}[\rho(\nu)]$. An usual choice for the loss function ρ is the Tukey-Biweight function ρ_κ^B defined by (4.17).

The M-estimators of scale are translation invariant and scale equivariant. The influence function of the functional S associated to the scale M-estimator $\widehat{\sigma}_{\text{M};\rho}$ under the distribution $F_{0,1}$ of the error term ν of the location-scale model is given by

$$\text{IF}(u;S,F_{0,1})=\frac{\rho(u)-\delta}{\text{E}_{F_{0,1}}[\rho'(\nu)\nu]}.$$

Hence, the choice of a bounded function ρ implies that the influence function is also bounded. The asymptotic breakdown point of the scale M-estimator is:

$$\varepsilon^*(S,F_{0,1})=\min\left(\frac{\delta}{\rho(\infty)},1-\frac{\delta}{\rho(\infty)}\right),$$

which is strictly positive but not always equal to 50%, even if ρ is bounded.

❑ Remark

We may try to jointly estimate μ and σ by solving simultaneously two equations of the type (4.66) and (4.67) (see, for example, Huber and Ronchetti 2009, chapter 6). This complexifies the computations. Moreover, as explained in Maronna et al. (2006), it generally provides for $\widehat{\mu}_{\text{M};\rho}$ an asymptotic breakdown point smaller than 50% — hence, smaller than the breakdown point attainable by using the MAD as preliminary

18. The breakdown point of $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ is actually equal to the breakdown point of the preliminary estimator of the scale parameter σ.
19. Note however that it is possible to prove that, using the MAD as initial scale estimator, the breakdown point of $\widehat{\mu}_{\text{M};\rho_\kappa^{\text{B}}}$ is strictly greater than 0.49 in the Gaussian case.

estimator of σ. Consequently, the joint estimation of μ and σ is not recommended, especially when the scale parameter σ is considered as a nuisance parameter in the location-scale model.

❑

4.9 Appendix 2: Generalized Method of Moments (GMM) and asymptotic distributions of regression M-, S- and MM-estimators

4.9.1 GMM-estimation principle

For simplicity, let us consider immediately the context of the regression model (4.1). Let y be the scalar dependent variable and $\mathbf{x} = (1, x_1, \ldots, x_p)^t$ be the $(p+1)$-vector of covariates. We assume here that the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are generated by a *stationary*[20] and *ergodic*[21] process H. We also assume, to avoid too much technicalities, that there is *no autocorrelation*, that is, that the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are *independent*.[22]

Suppose that our objective is to estimate the functional $\boldsymbol{\theta} = \boldsymbol{\theta}(H)$ that is implicitly defined by the equation

$$\mathrm{E}_H[\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})] = \mathbf{0}, \tag{4.68}$$

where $\mathbf{m}$ is a known k-valued function, and $\mathrm{E}_H[\cdot]$ denotes the mathematical expectation with respect to H. If k equals the dimension of the parameter $\boldsymbol{\theta}$ to estimate, i.e., if the number of moments conditions specified by (4.68) coïncides with the dimension of $\boldsymbol{\theta}$, then the GMM estimation problem is said to be *exactly-identified*. Note that it is the case in the setting studied hereafter. The GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ of $\boldsymbol{\theta}$ is then simply obtained by solving the sample analogue of (4.68), that is,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{m}\Big(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}\Big) = \mathbf{0}. \tag{4.69}$$

Under regularity conditions detailed in Hansen (1982), the GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ defined by (4.69) has a limiting normal distribution:

$$\sqrt{n}\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}} - \boldsymbol{\theta}\Big) \to^d \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{4.70}$$

20. A *stationary* process is a stochastic process whose joint probability distribution does not change when shifted in time or space. Consequently, parameters such as the mean and the variance, if they exist, also do not change over time or position . Hence, the mean and the variance of the process do not follow trends.

21. A stochastic process is said to be *ergodic* if its statistical properties (such as its mean and variance) can be estimated consistently from a single, sufficiently long sample (realization) of the process.

22. The interested reader can find very general results, valid in presence of autocorrelation, in Croux et al. (2003).

where, in the exactly-identified case,

$$\mathbf{V} = \mathbf{G}^{-1}\boldsymbol{\Omega}\big(\mathbf{G}^t\big)^{-1}, \tag{4.71}$$

with[23]

$$\mathbf{G} = \mathrm{E}\Big[\frac{\partial \mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}\Big] \quad \text{and} \quad \boldsymbol{\Omega} = \mathrm{E}\big[\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})\mathbf{m}^t(y,\mathbf{x},\boldsymbol{\theta})\big]. \tag{4.72}$$

4.9.2 M-, S- and MM-estimators as GMM-estimators

Let us first consider the case where we estimate the parameters $\boldsymbol{\beta}$ and σ simultaneously by an M-estimation procedure. Let us denote by $\rho(\cdot)$ and $\rho_0(\cdot)$ the loss functions used for the M-estimation of $\boldsymbol{\beta}$ and σ, respectively. Then the M-regression estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ and the M-scale estimator $\widehat{\sigma}_{\rho_0}$ are such that

$$\left\{\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\psi\left(\frac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i = \mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0\left(\frac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}}{\widehat{\sigma}_{\rho_0}}\right) - \delta = 0 \end{array}\right. \tag{4.73}$$

where $\psi(u) = \rho'(u)$, δ is a selected constant and, using similar notations as in the previous sections, $\widehat{\sigma}_{\rho_0} = s_{\rho_0}\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}\Big),\ldots,r_n\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}\Big)\Big)$. This shows that the M-estimator $\Big(\widehat{\boldsymbol{\beta}}^t_{\mathrm{M};\rho},\widehat{\sigma}_{\rho_0}\Big)^t$ is an exactly-identified GMM-estimator for $\boldsymbol{\theta} = \big(\boldsymbol{\beta}^t,\sigma\big)^t$, with

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta}) = \left(\begin{array}{c} \psi\Big(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\Big)\mathbf{x} \\ \rho_0\Big(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\Big) - \delta \end{array}\right). \tag{4.74}$$

S-estimators of regression and scale depend only on a chosen loss function ρ_0 and on a constant δ. We have defined the S-regression estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ as follows:

$$\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} = \arg\min_{\boldsymbol{\beta}} s_{\rho_0}(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta})) \tag{4.75}$$

where s_{ρ_0} is a measure of dispersion satisfying

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\boldsymbol{\beta})}{s_{\rho_0}(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta}))}\right) - \delta = 0 \quad \text{for all } \boldsymbol{\beta}\in\mathbb{R}^{p+1}.$$

The scale estimator is then simply given by

$$\widehat{\sigma}_{\rho_0} = s_{\rho_0}\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big),\ldots,r_n\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big)\Big).$$

23. Here and later, we simply write $\mathrm{E}[\cdot]$ for $\mathrm{E}_H[\cdot]$.

As previously explained, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$ satisfy the first order conditions

$$\left\{\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0'\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)-\delta=0. \end{array}\right. \tag{4.76}$$

Note that the equations (4.76) are of the same form as (4.73). Hence an S-estimator is first-order equivalent with an M-estimator where $\rho(\cdot)=\rho_0(\cdot)$, and has the same asymptotic distribution (see Rousseeuw and Yohai 1984). Note however that the function ρ_0 defining the S-estimator needs to be bounded to get a positive breakdown point for the regression estimator. But if ρ_0 is bounded, ρ_0' is redescending and the first set of equations in (4.76) — the set of equations involving ρ_0' — may have multiple solutions. Therefore one usually uses (4.75) to compute the S-estimate but (4.76) to determine its asymptotic distribution. Actually, (4.76) implies that $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}^t,\widehat{\sigma}_{\rho_0}\right)^t$ is first-order equivalent with the GMM-estimator for $\boldsymbol{\theta}=\left(\boldsymbol{\beta}^t,\sigma\right)^t$, with

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})=\left(\begin{array}{c} \rho_0'\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)\mathbf{x} \\ \rho_0\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)-\delta \end{array}\right).$$

Let us now focus on MM-estimators of regression. First one needs to compute S-estimators $\left(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}^t,\widehat{\sigma}_{\rho_0}\right)^t$ for a given function ρ_0 and a constant δ. Secondly, for a given function $\psi=\rho'$, the MM-estimator of regression solves

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0}.$$

Note that ρ needs to be different from ρ_0, otherwise the MM-estimator would be equivalent with an S-estimator and share the low efficiency of the latter. In this MM-estimation procedure, $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$ are such that

$$\left\{\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\psi\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0'\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)-\delta=0. \end{array}\right. \tag{4.77}$$

Defining $\boldsymbol{\theta}=(\boldsymbol{\beta}^t,\boldsymbol{\beta}_0^t,\sigma)^t$, where the first parameter $\boldsymbol{\beta}$ will be estimated by $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ and the latter two by $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$, equations (4.77) show that $(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}^t,\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}^t,\widehat{\sigma}_{\rho_0})^t$

is first-order equivalent with the GMM-estimator for $\boldsymbol{\theta}$, with

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta}) = \begin{pmatrix} \psi\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)\mathbf{x} \\ \rho_0'\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}_0}{\sigma}\right)\mathbf{x} \\ \rho_0\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}_0}{\sigma}\right)-\delta \end{pmatrix}.$$

Using the generic notations $u_0 = \frac{y-\mathbf{x}^t\boldsymbol{\beta}_0}{\sigma}$ and $u = \frac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}$, the moment function $\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})$ takes the simpler form

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta}) = \begin{pmatrix} \psi(u)\mathbf{x} \\ \rho_0'(u_0)\mathbf{x} \\ \rho_0(u_0)-\delta \end{pmatrix},$$

or still more shortly,

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta}) = \begin{pmatrix} \psi\mathbf{x} \\ \rho_0'\mathbf{x} \\ \rho_0-\delta \end{pmatrix}, \tag{4.78}$$

if we simply replace $\psi(u)$ by ψ, $\rho_0(u_0)$ by ρ_0, and $\rho_0'(u_0)$ by ρ_0'. This compact notation for the moment function will be more practice to use in the sequel.

4.9.3 Asymptotic variance matrix of an MM-estimator

If the observations $(\mathbf{x}_i, y_i)$, $i = 1,\dots,n$, are generated by a stationary and ergodic process, and are independent (Assumption A1)

The first-order equivalence of $(\widehat{\boldsymbol{\beta}}^t_{\mathrm{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}^t_{\mathrm{S};\rho_0}, \widehat{\sigma}_{\rho_0})^t$ with a GMM-estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\beta}_0^t, \sigma)^t$ allows us to conclude that, if the observations $(\mathbf{x}_1, y_1),\dots,(\mathbf{x}_n, y_n)$ are generated by a *stationary* and *ergodic* process, and are *independent* (Assumption A1)[24],

$$\sqrt{n}\left(\begin{pmatrix} \widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0} \\ \widehat{\sigma}_{\rho_0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\beta}_0 \\ \sigma \end{pmatrix}\right) \to^d \mathcal{N}(\mathbf{0}, \mathbf{V}_{\mathrm{MM}})$$

where

$$\mathbf{V}_{\mathrm{MM}} = \mathbf{G}_{\mathrm{MM}}^{-1}\boldsymbol{\Omega}_{\mathrm{MM}}\left(\mathbf{G}_{\mathrm{MM}}^t\right)^{-1},$$

with the matrices $\mathbf{G}_{\mathrm{MM}}$ and $\boldsymbol{\Omega}_{\mathrm{MM}}$ obtained by applying relations (4.72) to the moment function (4.78):

$$\mathbf{G}_{\mathrm{MM}} = -\frac{1}{\sigma}\mathrm{E}\begin{pmatrix} \psi'\mathbf{x}\mathbf{x}^t & \mathbf{0} & \psi' u\mathbf{x} \\ \mathbf{0} & \rho_0''\mathbf{x}\mathbf{x}^t & \rho_0'' u_0\mathbf{x} \\ \mathbf{0} & \mathbf{0} & \rho_0' u_0 \end{pmatrix}$$

24. This Assumption A1 coincides with Assumption A in Section 4.6.1. We add here an number to the letter "A" in order to clearly distinguish the various assumptions we will consider in the sequel of this appendix.

and

$$\mathbf{\Omega}_{\mathrm{MM}} = \mathrm{E}\begin{pmatrix} \psi^2\mathbf{x}\mathbf{x}^t & \psi\rho_0'\mathbf{x}\mathbf{x}^t & \psi\rho_0\mathbf{x} \\ \psi\rho_0'\mathbf{x}\mathbf{x}^t & (\rho_0')^2\mathbf{x}\mathbf{x}^t & \rho_0\rho_0'\mathbf{x} \\ \psi\rho_0\mathbf{x}^t & \rho_0\rho_0'\mathbf{x}^t & \rho_0^2 - \delta^2 \end{pmatrix}.$$

In particular, this result establishes the consistency of the MM-regression estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$. Moreover, using the upper left $(p+1)\times(p+1)$ submatrice of $\mathbf{V}_{\mathrm{MM}}$, we obtain that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ is equal to

$$\begin{aligned}\mathrm{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big) = \frac{1}{n}[&\mathbf{A}\mathrm{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A} - \mathbf{a}\mathrm{E}\big(\psi\rho_0\mathbf{x}^t\big)\mathbf{A} \\ &- \mathbf{A}\mathrm{E}(\psi\rho_0\mathbf{x})\mathbf{a}^t + \mathrm{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}\mathbf{a}^t],\end{aligned}$$

where

$$\mathbf{A} = \sigma\big[\mathrm{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \quad \text{and} \quad \mathbf{a} = \mathbf{A}\frac{\mathrm{E}(\psi' u\mathbf{x})}{\mathrm{E}(\rho_0' u_0)}.$$

This expression of $\mathrm{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$ is then estimated by its empirical counterpart $\widehat{\mathrm{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$, by applying the following two rules:

1. Replace, in u and u_0, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\beta}_0$ and σ by the estimates $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$.
2. Replace $\mathrm{E}(\cdot)$ by $\frac{1}{n}\sum_{i=1}^n(\cdot)$.

For example, the first term of $\widehat{\mathrm{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$ is given by

$$\frac{1}{n}\left[\widehat{\mathbf{A}}\left(\frac{1}{n}\sum_{i=1}^n\left[\psi\left(\frac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\right]^2\mathbf{x}_i\mathbf{x}_i^t\right)\widehat{\mathbf{A}}\right]$$

with

$$\widehat{\mathbf{A}} = \widehat{\sigma}_{\rho_0}\left[\frac{1}{n}\sum_{i=1}^n\psi'\left(\frac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i\mathbf{x}_i^t\right]^{-1}.$$

Using standard asymptotic arguments, it can be shown that $\widehat{\mathrm{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$ is a consistent estimate of $\mathrm{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$. From $\widehat{\mathrm{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$, standard errors for the regression coefficients are obtained in the usual way: for $j = 0, 1, \ldots, p$,

$$\widehat{\mathrm{se}}\left(\Big[\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big]_j\right) = \sqrt{\Big[\widehat{\mathrm{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)\Big]_{jj}}.$$

Moreover, the estimate $\widehat{\text{Avar}}_1\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$ of the asymptotic variance $\text{Avar}_1\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$ is robust with respect to bad leverage points and vertical outliers. Indeed, if there are observations yielding large residuals with respect to the robust MM-fit, then $\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)$ has a small value when ψ is a redescending function[25]. Hence, if there are bad leverage points in the sample, then their $\mathbf{x}_i$-value is large, but at the same time $\psi\left(\frac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)$ will be zero. This explains intuitively why vertical outliers and bad leverage points have only a limited influence on the estimate $\widehat{\text{Avar}}_1\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$.

In absence of heteroskedasticity (Assumption A2)

A simplification of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ occurs when, in addition to Assumption A1, we also assume that there is *no heteroskedasticity*, i.e., we assume that the processes $\mathbf{x}_i$ and (u_i, u_{0i}) are independent (Assumption A2). In that case, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ becomes

$$\begin{aligned}\text{Avar}_{12}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right) = \frac{1}{n}[&\text{E}\left(\psi^2\right)\mathbf{A}_2\text{E}\left(\mathbf{x}\mathbf{x}^t\right)\mathbf{A}_2 - \text{E}(\psi\rho_0)\mathbf{a}_2\text{E}\left(\mathbf{x}^t\right)\mathbf{A}_2 \\ &- \text{E}(\psi\rho_0)\mathbf{A}_2\text{E}(\mathbf{x})\mathbf{a}_2^t + \text{E}\left(\rho_0^2 - \delta^2\right)\mathbf{a}_2\mathbf{a}_2^t],\end{aligned}$$

where

$$\mathbf{A}_2 = \sigma\frac{\left[\text{E}(\mathbf{x}\mathbf{x}^t)\right]^{-1}}{\text{E}(\psi')} \quad \text{and} \quad \mathbf{a}_2 = \mathbf{A}_2\frac{\text{E}(\psi' u)\text{E}(\mathbf{x})}{\text{E}(\rho_0' u_0)}.$$

Taking the empirical counterpart yields $\widehat{\text{Avar}}_{12}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$. However, Croux *et al.* (2003) do advise against the use of this variance matrix estimator in practice, even when assumptions A1 and A2 holds. The reason is that this estimator will not be robust with respect to (good and bad) leverage points. Indeed, $\widehat{\mathbf{A}}_2$, for example, is proportional to the inverse of an empirical second moment matrix of the observations $\mathbf{x}_i$. Leverage points are outlying in the covariates' space, and will then have a strong influence on $\widehat{\mathbf{A}}_2$. This can even lead $\widehat{\text{Avar}}_{12}\left(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\right)$ to break down, where breakdown of a variance matrix estimator means that the latter has a determinant close to zero or enormously large.

If the distribution of the error terms is symmetric around zero (Assumption A3)

A condition often imposed in the literature is that the distribution of $u_i = \frac{y_i-\mathbf{x}_i^t\boldsymbol{\beta}}{\sigma}$, given $\mathbf{x}_i$, is symmetric (Assumption A3). If this condition is met, the regression parameter estimator and the estimator of residual scale are asymptotically independent, and the different expressions simplify considerably, due to the fact that $\mathbf{a} = \mathbf{0}$.

25. Recall that, if ψ is redescending, it has the property to be equal to zero for large arguments.

Under Assumptions A1 and A3, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ becomes

$$\begin{aligned}\mathrm{Avar}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big) &= \frac{1}{n}\mathbf{A}\mathrm{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A} \\ &= \frac{\sigma^2}{n}\big[\mathrm{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1}\mathrm{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\big[\mathrm{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.\end{aligned}$$

The empirical counterpart of the latter expression, $\widehat{\mathrm{Avar}}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$, is an estimate of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ that is robust against vertical outliers and bad leverage points. But it relies on symmetry of the errors distribution, a quite strong assumption. A simulation study in Croux et al. (2003) shows that, even when symmetry is present, there is no gain in using $\widehat{\mathrm{Avar}}_{13}$ compared to $\widehat{\mathrm{Avar}}_1$: the authors of Croux et al. (2003) then recommend to use $\widehat{\mathrm{Avar}}_1$ in any case.

When all of Assumptions A1, A2 and A3 hold, then $\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}$ has asymptotic variance

$$\mathrm{Avar}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big) = \frac{\sigma^2}{n}\frac{\mathrm{E}\big(\psi^2\big)}{\big[\mathrm{E}(\psi')\big]^2}\big[\mathrm{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.$$

This corresponds to the expression for the variance of the MM-regression estimator that was derived in Yohai (1987). The empirical counterpart $\widehat{\mathrm{Avar}}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$ is an estimate of this asymptotic variance that, as $\widehat{\mathrm{Avar}}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}\Big)$, lacks robustness with respect to leverage points.

4.9.4 Asymptotic variance matrix of an S-estimator

If Assumption A1 holds, the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ is simply derived from the central $(p+1)\times(p+1)$ submatrice of $\mathbf{V}_{\mathrm{MM}}$ (cf. (4.37), (4.38) and (4.39)):

$$\begin{aligned}\mathrm{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big) = \frac{1}{n}[&\mathbf{A}_{\mathrm{S}}\mathrm{E}\Big((\rho_0')^2\mathbf{x}\mathbf{x}^t\Big)\mathbf{A}_{\mathrm{S}} - \mathbf{a}_{\mathrm{S}}\mathrm{E}\big(\rho_0\rho_0'\mathbf{x}^t\big)\mathbf{A}_{\mathrm{S}} \\ &- \mathbf{A}_{\mathrm{S}}\mathrm{E}(\rho_0\rho_0'\mathbf{x})\mathbf{a}_{\mathrm{S}}^t + \mathrm{E}\big(\rho_0^2-\delta^2\big)\mathbf{a}_{\mathrm{S}}\mathbf{a}_{\mathrm{S}}^t],\end{aligned}$$

where

$$\mathbf{A}_{\mathrm{S}} = \sigma\big[\mathrm{E}\big(\rho_0''\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \quad \text{and} \quad \mathbf{a}_{\mathrm{S}} = \mathbf{A}_{\mathrm{S}}\frac{\mathrm{E}(\rho_0''u_0\mathbf{x})}{\mathrm{E}(\rho_0'u_0)}.$$

If, in addition, A2 holds, then the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}$ takes the form

$$\begin{aligned}\mathrm{Avar}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}\Big) = &\frac{\sigma^2}{n}\frac{\mathrm{E}\Big((\rho_0')^2\Big)}{\big[\mathrm{E}(\rho_0'')\big]^2}\big[\mathrm{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1} + \frac{\sigma^2}{n}\frac{\mathrm{E}(\rho_0''u_0)}{\big[\mathrm{E}(\rho_0'')\big]^2\mathrm{E}(\rho_0'u_0)} \\ &\times\left\{\frac{\mathrm{E}(\rho_0''u_0)\mathrm{E}\big(\rho_0^2-\delta^2\big)}{\mathrm{E}(\rho_0'u_0)} - 2\mathrm{E}(\rho_0\rho_0')\right\} \\ &\times\big[\mathrm{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}\mathrm{E}(\mathbf{x})\mathrm{E}\big(\mathbf{x}^t\big)\big[\mathrm{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.\end{aligned}$$

Under A3, the expressions are the same as those for the MM-estimator, with ψ replaced by ρ_0'.

4.9.5 Asymptotic variance matrix of an M-estimator

Here the expressions are less explicit. Under Assumption A1, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{M};,\rho}$ is derived from the upper left $(p+1)\times(p+1)$ block of $\mathbf{G}_{\mathrm{M}}^{-1}\boldsymbol{\Omega}_{\mathrm{M}}(\mathbf{G}_{\mathrm{M}}^{t})^{-1}$ where

$$\mathbf{G}_{\mathrm{M}} = \mathrm{E}\left[\frac{\partial \mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}\right] \quad \text{and} \quad \boldsymbol{\Omega}_{\mathrm{M}} = \mathrm{E}\big[\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})\mathbf{m}^t(y,\mathbf{x},\boldsymbol{\theta})\big],$$

with $\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})$ given by (4.74). Defining $u = \frac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}$, and denoting shortly $\rho'(u) = \psi(u)$ and $\rho_0(u)$ by ψ and ρ_0, respectively, we have

$$\mathbf{G}_{\mathrm{M}} = -\frac{1}{\sigma}\mathrm{E}\begin{pmatrix} \psi'\mathbf{x}\mathbf{x}^t & \psi' u\mathbf{x} \\ \rho_0'\mathbf{x}^t & \rho_0' u \end{pmatrix}$$

and

$$\boldsymbol{\Omega}_{\mathrm{M}} = \mathrm{E}\begin{pmatrix} \psi^2\mathbf{x}\mathbf{x}^t & \psi\rho_0\mathbf{x} \\ \psi\rho_0\mathbf{x}^t & \rho_0^2 - \delta^2 \end{pmatrix}.$$

If in addition Assumption A2 holds, then

$$\mathbf{G}_{\mathrm{M}} = -\frac{1}{\sigma}\begin{pmatrix} \mathrm{E}(\psi')\mathrm{E}(\mathbf{x}\mathbf{x}^t) & \mathrm{E}(\psi' u)\mathrm{E}(\mathbf{x}) \\ \mathrm{E}(\rho_0')\mathrm{E}(\mathbf{x}^t) & \mathrm{E}(\rho_0' u) \end{pmatrix}$$

and

$$\boldsymbol{\Omega}_{\mathrm{M}} = \begin{pmatrix} \mathrm{E}\big(\psi^2\big)\mathrm{E}(\mathbf{x}\mathbf{x}^t) & \mathrm{E}(\psi\rho_0)\mathrm{E}(\mathbf{x}) \\ \mathrm{E}(\psi\rho_0)\mathrm{E}(\mathbf{x}^t) & \mathrm{E}\big(\rho_0^2\big) - \delta^2 \end{pmatrix}.$$

Under Assumption A3, the expressions of $\mathrm{Avar}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}\Big)$ and $\mathrm{Avar}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}\Big)$ are exactly similar to those for the MM-estimator.

5 Robust estimators for panel data

- Some basic FE and RE panel data robust models

6 Robust instrumental variables estimation

- Robust IV/2SLS estimators

7 Robust estimators for categorical and limited dependent variables

- Robust logit
- Possibly: Robust models for limited dependent variables
- Possibly: more on robust GLM

Part IV

Robust multivariate statistics

8 Robust estimation of location and scatter

- Minimum covariance determinant (MCD) and minimum volume ellipsoid (MVE) estimators of location and scatter
- Possibly: Other multivariate location and scatter estimators such as the Stahel-Donoho estimate
- S-estimators and MM-estimators of location and scatter

9 Robust principal component analysis

Part V

Appendix

A Syntax and options

A.1 Robust statistics (robstat)

48: This section will be replaced. The plan is to collect the different tools in new command `robstat`

A.1.1 Classical estimators

Classical estimators are readily available in Stata via the `summarize` command using the `detail` option.

A.1.2 Quantile-based estimators

Quantile based estimators are not readily available in Stata but can easily be calculated using the `centile` command. For example, for a statistical series x, we compute:

the *median* $Q_{0.5;n}$ by

```
centile x, centile(50)
local med=r(c_1)
```

the *corrected interquartile range* IQR_c by

```
centile x, centile(25 75)
local iqr=(r(c_2)-r(c_1))*0.7413
```

the Yule and Kendall skewness coefficient $\mathrm{SK}_{0.25;n}$ by

```
centile x, centile(25 50 75)
local sk=(r(c_1)+r(c_3)-2*r(c_2))/(r(c_2)-r(c_1))
```

the quantile tail weight measures $\mathrm{LQW}_{0.25;n}$ and $\mathrm{RQW}_{0.25;n}$ by

```
centile body, centile(12.5 25 37.5 62.5 75 87.5)
local lqw=-(r(c_1)+r(c_3)-2*r(c_2))/(r(c_3)-r(c_1))
local rqw=(r(c_4)+r(c_6)-2*r(c_5))/(r(c_6)-r(c_4))
```

A.1.3 Pairwise-based estimators

As explained in the previous sections, the location, scale, skewness and tails heaviness estimators based on pairwise combinations or comparisons of the observations are particularly interesting because they perform better than the classical or quantile-based estimators, namely in terms of robustness. But they require at first sight a heavy computation time of an order of n^2. Fortunately, as explained in **?**, an efficient algorithm proposed by Johnson and Mizoguchi (see **?**) allows to substantially reduce the computation time to an order of $n \log n$ and, hence, allows to use these robust estimators even in very large datasets. This algorithm has been programmed in Stata (cf. **?**) and is involved in the Stata computation of the pairwise-based estimators.

For these pairwise-based estimators, the Stata commands are `hl` (for the Hodges-Lehman location estimator HL_n), `qn` (for the Rousseeuw and Croux Q_n estimator), `medcouple` (for Brys' medcouple MC_n, left medcouple LMC_n and right medcouple RMC_n). The corresponding syntaxes are:

Location

Title

hl – Hodges and Lehman (1963) robust measure of location

Syntax

hl varname [if] [in]

Dispersion

Title

qn – Rousseeuw and Croux (1993) robust measure of dispersion

Syntax

qn varname [if] [in]

Skewness and heavyness of the tails

Title

medcouple – medcouple measure of asymmetry and heaviness of the tails

Syntax

medcouple varname [if] [in] [, lmc rmc nomc]

---

options description

Main

`lmc` specifies to calculate the medcouple only for those informations smaller than the median. This is an indicator of the heaviness of the left tail.

`rmc` Specifies to calculate the medcouple only for those informations larger than the median. This is an indicator of the heaviness of the right tail.

`nomc` Specifies not to calculate the global medcouple. This is for instance useful when one is only interested in the heaviness of the tails.

A.1.4 Normality tests

The Stata command "`robjb` *varname* [if] [in]" has been created to apply these robust tests of normality. This command implements the test considering both skewness and heaviness of tails by default. Two mutually exclusive options are available: `skewness` and `kurtosis`. If the former is used, a test based exclusively on the skewness is performed while if the latter is called, a test based exclusively on the heaviness of the tails is performed.

A.1.5 Boxplot

DESCRIBE HERE THE COMMAND, FOR THE MOMENT WE HAVE

Title

box_out - Boxplot for skewed and/or heavy-tailed distributions

Syntax

box_out varname [if] [in] [, out(varname) bdp(#) perc(#) nograph]

Description

box_out Creates the boxplot for skewed and/or heavy-tailed distributions

---

options description

Main

`out(varname)` Identifies the new variable to be created to identify individuals outside the fence defined by the whiskers

`bdp(integer)` Sets the desired Break-down point (in %). It is 10% by default

`perc(real)` Sets the desired percentage of points outside the whiskers in case of uncontaminated data.

`nograph` Suppresses the graph

box_out saves the following in e():

`e(g)` Estimated skewness parameter of the underlying Tukey g and h distribution

`e(h)` Estimated elongation parameter of the underlying Tukey g and h distribution

`e(upperW)` Value of the upper whisker

`e(lowerW)` Value of the lower whisker

A.2 Robust linear regression (robreg)

`robreg` provides a number of robust estimators for linear regression models. The syntax of `robreg` is:

`robreg` *estimator depvar* [*indepvars*] [*if*] [*in*] [, *options*]

where *estimator* is one of the following:

`mm` to fit the efficient high breakdown MM estimator proposed by Yohai (1987). On the first stage, a high breakdown S estimator is applied to estimate the residual scale and derive starting values for the coefficients vector. On the second stage, an efficient bisquare M estimator is applied to obtain the final coefficient estimates.

`gm` to be implemented

`m` to fit regression `M` estimators (Huber 1973) using iteratively reweighted least squares (IRWLS).

`s` to fit the high breakdown `S` estimator introduced by Rousseeuw and Yohai (1984) using the fast algorithm by Salibian-Barrera and Yohai (2006) with a nonsingular subsampling refinement as proposed by Koller (2012).

`lms`, `lqs`, or `lts` to fit the least median of squares (LMS), the least quantile of squares (LQS; a generalization of LMS), or the least trimmed squares (LTS) estimator, respectively (Rousseeuw and Leroy 1987). Estimation is carried out using simple resampling without local improvement (e.g. Rousseeuw and Leroy 1987, 197). Computation of standard errors is not supported for `lms`, `lqs`, and `lts`.

`robreg` without arguments replays the previous results; see [U] **20.3 Replaying prior results**. `robreg` saves its results in `e()`, so that post estimation commands can be applied; see [U] **20 Estimation and postestimation commands**. Type `ereturn list` to list the `e()`-returns after estimation; see [P] **ereturn**.

A.2.1 Options for robreg mm

Main

efficiency(#) sets the gaussian efficiency of the MM estimator (i.e., the asymptotic relative efficiency compared to the LS or ML estimator in case of i.i.d. normal errors). The efficiency is determined by appropriate choice of the tuning constant for the bisquare M estimator in the second stage of the MM algorithm. # must be between 0.1 and 99.9. The default for the MM estimator is `efficiency(85)`, as suggested by Maronna et al. (2006, 144).

`bp(#)` sets the breakdown point of the MM estimator. The breakdown point is determined by appropriate choice of the tuning constant for the S estimator in the first stage of the MM algorithm. # must be between 1 and 50. The default is `bp(50)`.

hausman performs a generalized Hausman test of the MM estimate against the S estimate. A significant Hausman test indicates that the MM estimate significantly deviates from the S estimate.

Biweight M estimate

`k(#)` specifies the tuning constant for the bisquare M estimator in the second stage of the MM algorithm. `k()` not allowed if `efficiency()` is specified.

tolerance(#) specifies the tolerance for the weights of the IRWLS algorithm used to fit the bisquare M estimator. When the maximum absolute change in the weights from one iteration to the next is less than or equal to `tolerance()`, the convergence criterion is satisfied. The default is `tolerance(1e-6)`.

iterate(#) specifies the maximum number of iterations for the IRWLS algorithm used to fit the bisquare M estimator. If convergence is not reached within `iterate()` iterations, the algorithm stops and returns error. The default is `iterate(16000)` or as set by `set maxiter` (see [R] **maximize**).

`relax` causes the IRWLS algorithm to return the current results instead of returning error if convergence is not reached.

generate(*newvar*) stores the final weights of the IRWLS algorithm in variable *newvar*.

replace permits `robreg` to overwrite existing variables.

Initial S estimate

nsamp(#) specifies the number of trial samples for the search algorithm of the S estimator in the first stage of the MM algorithm. The default value is determined according to formula $\lceil \ln(\alpha)/\ln(1-(1-\varepsilon)^p) \rceil$ within a range of 50 to 10000, where p is the number of coefficients in the model and $\alpha = 0.01$ and $\varepsilon = 0.2$ (see Salibian-Barrera and Yohai 2006 for a justification of the formula). The default values for α and ε

can be changed via `sopts()` (see below).

`sopts(`*options*`)` specifies additional options to be passed through to the S estimator. See the section on options for `robreg s` below.

`save(`*name*`)` saves the results of the S estimator under *name* using `estimates store` (see [R] **estimates**).

Standard errors

`vce(norobust)` causes standard errors to be computed using traditional formulas assuming constant error variance. The default is to compute robust standard errors as suggested by Croux et al. (2003) (using formula AVAR_1; the traditional formula is equivalent to AVAR_{2s}).

`norobust` is a synonym for `vce(norobust)`

Reporting

`level(#)` specifies the level for confidence intervals. The default is `level(95)` or as set by `set level` (see [R] **level**).

`first` causes the first stage S estimate to be displayed.

`nodots` suppresses the progress dots of the S estimator search algorithm.

`log` displays the iteration log of the second stage IRWLS algorithm.

display_options are various display options; see [R] **estimation options**.

A.2.2 Options for robreg gm

To be completed.

A.2.3 Options for robreg m

Main

`huber` causes the Huber objective function to be used (monotone M estimator). This is the default.

`biweight` causes the biweight or bisquare objective function to be used (redescending M estimator). `bisquare` is a synonym for `biweight`. The solution of a redescending M estimator may depend on the starting values.

`efficiency(#)` sets the gaussian efficiency (i.e. the asymptotic relative efficiency compared to the LS or ML estimator in case of i.i.d. normal errors) by appropriate choice of the tuning constant. # must be between 63.7 and 99.9 for `huber` and between

0.1 and 99.9 for `biweight`. The default is `efficiency(95)`.

`k(#)` specifies the tuning constant. `k()` not allowed if `efficiency()` is specified.

IRWLS algorithm

`tolerance(#)` specifies the tolerance for the weights of the IRWLS algorithm. When the maximum absolute change in the weights from one iteration to the next is less than or equal to `tolerance()`, the convergence criterion is satisfied. The default is `tolerance(1e-6)`.

`iterate(#)` specifies the maximum number of iterations for the IRWLS algorithm. If convergence is not reached within `iterate()` iterations, the algorithm stops and returns error. The default is `iterate(16000)` or as set by `set maxiter` (see [R] **maximize**).

`relax` causes the IRWLS algorithm to return the current results instead of returning error if convergence is not reached. For example, to fit a one-step M estimate specify `relax` together with `iterate(1)`.

`generate(`*newvar*`)` stores the final weights of the IRWLS algorithm in variable *newvar*.

`replace` permits `robreg` to overwrite existing variables.

Initial estimate

`init(`*arg*`)` determines the choice of the initial estimate that provides the starting values for the IRWLS algorithm. *arg* may be `lav` for the LAV estimator (a.k.a. median regression; fitted using `qreg`, see [R] **qreg**), `ols` for the least squares estimator (fitted using `regress`, see [R] **regress**), *name* for an estimation set stored under *name*, or `.` for the currently active estimation results. The default is `init(lav)`.

`save(`*name*`)` saves initial `lav` or `ols` estimate under *name* using `estimates store` (see [R] **estimates**).

Scale estimate

`scale(#)` provides a preliminary value for the residual scale that will be held constant. The default is to use the normalized median of the $(n-p)$ largest absolute residuals from the initial fit, where n is the sample size and p is the number of coefficients, as an estimate of the residual scale (MADN).

`updatescale` causes the MADN scale estimate to be updated in each iteration of the IRWLS algorithm. `updatescale` has no effect if `scale()` is specified.

`center` causes the MADN scale estimate to be computed based on median centered residuals. `center` has no effect if `scale()` is specified.

Standard errors

`vce(norobust)` causes standard errors to be computed using traditional formulas assuming constant error variance. The default is to compute robust standard errors as suggested by Croux et al. (2003) (using formula AVAR_1; the traditional formula is equivalent to AVAR_{2s}).

`vce(pv)` causes traditional standard errors to be computed using the pseudo-values approach (Street et al. 1988). `vce(pv)` is equivalent to `vce(norobust)` but includes some small sample correction.

`norobust` is a synonym for `vce(norobust)`

`nose` skips the computation of standard errors.

Reporting

`level(#)` specifies the level for confidence intervals. The default is `level(95)` or as set by `set level` (see [R] **level**).

`first` causes the initial estimate to be displayed.

`log` displays the iteration log of the second stage IRWLS algorithm.

display_options are various display options; see [R] **estimation options**.

A.2.4 Options for robreg s

Main

`bp(#)` sets the breakdown point by appropriate choice of the tuning constant (this also determines the gaussian efficiency). # must be between 1 and 50. The default is `bp(50)`.

`k(#)` specifies the tuning constant. `k()` not allowed if `bp()` is specified.

`hausman` performs a generalized Hausman test of the least squares estimate against the S estimate. A significant Hausman test indicates that the least squares estimate significantly deviates from the S estimate.

Resampling algorithm

`categorical(`*varlist*`)` specifies the variables to be treated as categorical. The default is to detect and treat all dummy variables as categorical. If `categorical()` is specified, the detection of dummy variables is deactivated and only the variables identified by `categorical()` are treated as categorical. Variables from `categorical()` that are not found in *indepvars* will be automatically added to the end of *indepvars*. `categorical()` may contain factor variables; see [U] **11.4.3 Factor variables**. I

changed the default behavior, if I remember right. Needs to be updated.

`nocategorical` treats all variables as continuous.

`nonsingular` use nonsingular subsampling needs to be completed; nonsingular should be default

`nsamp(#)` specifies the number of trial samples for the search algorithm. The default value is determined according to formula $\lceil \ln(\alpha)/\ln(1-(1-\varepsilon)^p) \rceil$ within a range of 50 to 10000, where p is the number of coefficients in the model and α and ε are set by `alpha()` and `epsilon()` (see Salibian-Barrera and Yohai 2006 for a justification of the formula).

`alpha(#)` specifies the maximum admissible risk of drawing a set of samples of which none is free of outliers. This is a parameter in the formula for the computation of the required number samples (see above). The default is `alpha(0.01)` (i.e. 1 percent). `alpha()` has no effect if `nsamp()` is specified.

`epsilon(#)` specifies the assumed maximum fraction of contaminated data. This is a parameter in the formula for the computation of the required number samples (see above). The default is `epsilon(0.2)` (i.e. 20 percent). `epsilon()` has no effect if `nsamp()` is specified.

`nkeep(#)` specifies the number of best candidates to be kept for final refinement. The default is `nkeep(2)`.

`rsteps(#)` specifies the number of local improvement steps applied to the candidates. The default is `rsteps(1)`.

`stolerance(#)` specifies the tolerance for the scale estimate of the candidates. When the absolute relative change in the scale from one iteration to the next is less than or equal to `stolerance()`, the convergence criterion is satisfied. The default is `stolerance(1e-6)`.

`siterate(#)` specifies the maximum number of iterations for the scale estimate of the candidates. If convergence is not reached within `siterate()` iterations, the algorithm stops and returns error. The default is `siterate(16000)` or as set by `set maxiter` (see [R] **maximize**).

`tolerance(#)` specifies the tolerance for the coefficients in the refinement IRWLS algorithm. When the maximum relative change in the coefficient vector from one iteration to the next is less than or equal to `tolerance()`, the convergence criterion is satisfied. The default is `tolerance(1e-6)`.

`iterate(#)` specifies the maximum number of iterations for the refinement IRWLS algorithm. If convergence is not reached within `iterate()` iterations, the algorithm stops and returns error. The default is `iterate(16000)` or as set by `set maxiter` (see [R] **maximize**).

`ssteps(#)` specifies the number of approximation steps for the scale estimate within each RWLS iteration. The default is `ssteps(1)`.

`srsteps(#)` n of iterations for p-subset scale; default: until convergence revise!

`cefficiency(#)` efficiency of M for catvars; default `cefficiency(95)` revise!

`ck(#)` k of M for catvars; not allowed with `cefficiency()` revise!

`csteps(#)` approx. steps for catvar M within p-subset; default `cstep(0)` revise!

`noxresid` do not residualize continuous variables revise!

`fsteps(#)` approx. iteration in final backfit rounds; default: until convergence revise!

`nbackfit(#)` n of final backfit rounds; default: 20 revise!

`nobreak` do not exit final backfitting if scale increases revise!

`generate(`*newvar*`)` stores the final IRWLS weights from the best solution in variable *newvar*.

`replace` permits `robreg` to overwrite existing variables.

Standard errors

`vce(norobust)` causes standard errors to be computed using traditional formulas assuming constant error variance. The default is to compute robust standard errors as suggested by Croux et al. (2003) (using formula AVAR_1; the traditional formula is equivalent to AVAR_{2s}).

`norobust` is a synonym for `vce(norobust)`

`nose` skips the computation of standard errors.

Reporting

`level(#)` specifies the level for confidence intervals. The default is `level(95)` or as set by `set level` (see [R] **level**).

`nodots` suppresses the progress dots of the search algorithm.

display_options are various display options; see [R] **estimation options**.

A.2.5 Options for robreg lms, robreg lqs, and robreg lts

Main

`bp(#)` sets the breakdown point, where # may be in $(0, 0.5]$. `bp()` determines the h parameter for the LQS and LTS estimators as $h = \lfloor (1-\#) \cdot n \rfloor + \lfloor \# \cdot (p+1) \rfloor$ where n is the sample size and p is the number of coefficients. The default is `bp(0.5)`. `bp()` is not allowed with `robreg lms`.

Resampling algorithm

nsamp(#) specifies the number of trial samples for the search algorithm. The default value is determined according to formula $\lceil \ln(\alpha)/\ln(1-(1-\varepsilon)^p) \rceil$ within a range of 50 to 10000, where p is the number of coefficients in the model and α and ε are set by `alpha()` and `epsilon()`.

`alpha(#)` specifies the maximum admissible risk of drawing a set of samples of which none is free of outliers. This is a parameter in the formula for the computation of the required number samples (see above). The default is `alpha(0.01)` (i.e. 1 percent). `alpha()` has no effect if `nsamp()` is specified.

epsilon(#) specifies the assumed maximum fraction of contaminated data. This is a parameter in the formula for the computation of the required number samples (see above). The default is `epsilon(0.2)` (i.e. 20 percent). `epsilon()` has no effect if `nsamp()` is specified.

generate(*newvar*) stores a variable *newvar* that marks the minimizing trial sample.

replace permits `robreg` to overwrite existing variables.

Reporting

nodots suppresses the progress dots of the search algorithm.

display_options are various display options; see [R] **estimation options**.

A.3 Robust logistics regression (roblogit)

A.4 Robust multivariate statistics (robmv)

References

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley & Sons.

Anscombe, F. J. 1973. Graphs in Statistical Analysis. *The American Statistician* 27(1): 17–21.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley series in probability and mathematical statistics, New York: Wiley.

Bruffaerts et al. 2014. Details missing!!! .

Brys, Hubert, and Struyf. 2004a. Details missing!!! .

———. 2004b. Details missing!!! .

———. 2006. Details missing!!! .

Chatterjee, S., and A. S. Hadi. 1988. *Sensitivity Analysis in Linear Regression*. New York: John Wiley & Sons.

Cook, R. D., and S. Weisberg. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.

Croux, C., and C. Dehon. 2003. Estimators of the multiple correlation coefficient: Local robustness and confidence intervals. *Statistical Paper* 44: 315–334.

Croux, C., G. Dhaene, and D. Hoorelbeke. 2003. Robust Standard Errors for Robust Estimators. Discussions Paper Series (DPS) 03.16, Center for Economic Studies, KULeuven. Available from http://www.econ.kuleuven.be/eng/ew/discussionpapers/Dps03/Dps0316.pdf.

Dehon, C., M. Gassner, and V. Verardi. 2012. Extending the Hausman test to check for the presence of outliers. *Advances in Econometrics* 29(Essays in Honor of Jerry Hausman): 435–453.

Donoho, and P. J. Huber. 1983. Details missing!!! .

Edgeworth, F. Y. 1887. On Observations Relating to Several Quantities. *Hermathena* 6: 279–285.

Fox, J. 1991. *Regression Diagnostics*. Quantitative applications in the social sciences, Newbury Park, CA: Sage.

Greene, W. 1997. *Econometric Analysis*. 3rd ed. Prentice Hall.

Groeneveld. 1991. Details missing!!! .

Hampel, F. R. 1971. Details missing!!! .

———. 1974. Details missing!!! .

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics. The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hampel, F. R., and W. A. Stahel. 1982. Details missing!!! .

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica* 46(6): 1251–1271.

Heritier, S., E. Cantoni, S. Copt, and M.-P. Victoria-Feser. 2009. *Robust Statistics in Biostatistics*. West Sussey: Wiley.

Hinkley. 1975. Details missing!!! .

Hodges, and Lehmann. 1963. Details missing!!! .

Huber, and Ronchetti. 2009. Details missing!!! .

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics* 35(1): 73–101.

———. 1972. The 1972 Wald Lecture. Robust Statistics: A Review. *The Annals of Mathematical Statistics* 43(4): 1041–1067.

———. 1973. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics* 1(5): 799–821.

———. 1981. *Robust Statistics*. New York: John Wiley & Sons.

Hubert, M., and E. Vandervieren. 2008. An Adjusted Boxplot for Skewed Distributions. *Computational Statistics and Data Analysis* 52: 5186–5201.

Jarque, and Bera. 1980. Details missing!!! .

Jasso, G. 1985. Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. *American Sociological Review* 50(2): 224–241.

Jiménez, J. A., and V. Arunachalam. 2011. Using Tukey's g and h family of distributions to calculate value-at-risk and conditional value-at-risk. *Journal of Risk* 13(4): 95–116.

Kahn, J. R., and J. R. Udry. 1986. Mariatl Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. *American Sociological Review* 51(5): 734–737.

Koenker, R. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R., and G. Bassett. 1978. Regression Quantiles. *Econometrica* 46(1): 33–50.

Koller, M. 2012. Nonsingular subsampling for S-estimators with categorical predictors. arXiv:1208.5595v1 [stat.CO].

Kruskal, W. H. 1960. Some Remarks on Wild Observations. *Technometrics* 2(1): 1–3.

Lehmann, E. L., and G. Casella. 1988. *Theory of Point Estimation*. 2nd ed. Springer.

Mallows, C. L. 1975. On some topics in robustness. Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ.

Maronna, R. A., O. H. Bustos, and V. J. Yohai. 1979. Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In *Smoothing techniques for curve estimation*, ed. T. Gasser and J. M. Rossenblat, 91–116. Lecture Notes in Mathematics 757, New York: Springer.

Maronna, R. A., D. R. Martin, and V. J. Yohai. 2006. *Robust Statistics. Theory and Methods*. Chichester: John Wiley & Sons.

Omelka, M., and M. Salibian-Barrera. 2010. Uniform asymptotics for S- and MM-regression estimators. *Annals of the Institute of Statistical Mathematics* 62(5): 897–927.

Renaud, O., and M.-P. Victoria-Feser. 2010. A robust coefficient of determination for regression. *Journal of Statistical Planning and Inference* 140(7): 1852–1862.

Ronchetti, E., and P. J. Rousseeuw. 1985. Change-of-Variance Sensitivities in Regression Analysis. *Probability Theory and Related Fields* 68: 503–519.

Rousseeuw, P., and V. Yohai. 1984. Robust Regression by Means of S-Estimators. In *Robust and Nonlinear Time Series Analysis. Lecture Notes in Statistics Vol. 26*, ed. J. Franke, W. Hardle, and D. Martin, 256–272. Berlin: Springer.

Rousseeuw, P. J. 1984. Least Median of Squares Regression. *Journal of the American Statistical Association* 79(388): 871–880.

Rousseeuw, P. J., and C. Croux. 1993. Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association* 88(424): 1273–1283.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Ruppert. 1987. Details missing!!! .

Salibian-Barrera, M., and V. J. Yohai. 2006. A Fast Algorithm for S-Regression Estimates. *Journal of Computational and Graphical Statistics* 15(2): 414–427.

Salibian-Barrera, M., and R. H. Zamar. 2004. Uniform asymptotics for robust location estimates when the scale is unknown. *The Annals of Statistics* 32(4): 1434–1447.

Serfling, R. 1980. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley & Sons.

Staudte, R. G., and S. J. Sheather. 1990. *Robust Estimation and Testing*. New York: John Wiley & Sons.

Street, J. O., R. J. Carroll, and D. Ruppert. 1988. A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares. *The American Statistician* 42(2): 152–154.

White, H. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* 48(4): 817–838.

Wilcox, R. R. 2005. *Introduction to Robust Estimation and Hypothesis Testing*. 2nd ed. New York: Elsevier Academic Press.

Yohai, V. J. 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* 15(2): 642–656.

Author index

Subject index