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 Vermandele
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

List of tables **xi**

List of figures **xiii**

Preface **xv**

Notation and typography **xvii**

1 Robust linear regression **1**

1.1 The linear regression model 1

1.2 Different types of outliers 3

1.3 LS estimation 4

1.4 M estimation 6

1.4.1 L_1 or Least Absolute Deviation (LAD) estimation 6

1.4.2 The principle of M estimation 6

1.4.3 M estimation as a generalization of maximum likelihood (ML) estimation 9

1.4.4 Practical implementation of M estimates 11

Regression M estimate with preliminary scale estimation . . 11

1.4.5 Regression quantiles as regression M estimates 12

1.4.6 Monotone vs. redescending M estimators 12

1.4.7 GM estimation 13

1.5 Robust regression with a high breakdown point 14

1.5.1 LTS and LMS estimation 15

1.5.2 S estimation 16

1.5.3 MM estimation 18

Numerical computation of the S and MM estimate 19

1.5.4 MS estimation 21

1.6 Robust inference for M, S and MM estimators 22

1.6.1 Asymptotic distribution of M, S and MM estimators 23
1.6.2 Robust confidence intervals and tests with robust regression estimators 27
Inference for a single linear combination of the regression parameters 28
Inference for several linear combinations of the regression parameters 29
1.6.3 Robust R-squared 29
1.6.4 Extension of the Hausman test to check for the presence of outliers 33
Some preliminary results 34
Comparison of LS and S 35
Comparison of S and MM 36
1.7 Examples 37
Comparing estimators 37
Identifying outliers 42
1.8 Appendix 1: M-estimators of location and scale 46
1.8.1 M-estimator of location 46
1.8.2 M-estimator of scale 48
1.9 Appendix 2: Generalized Method of Moments (GMM) and asymptotic distributions of regression M, S and MM estimators 49
1.9.1 GMM-estimation principle 49
1.9.2 M-, S- and MM estimators as GMM estimators 50
1.9.3 Asymptotic variance matrix of an MM estimator 52
If the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are generated by a stationary and ergodic process, and are independent (Assumption A1) 52
In absence of heteroskedasticity (Assumption A2) 54
If the distribution of the error terms is symmetric around zero (Assumption A3) 54
1.9.4 Asymptotic variance matrix of an S-estimator 55
1.9.5 Asymptotic variance matrix of an M-estimator 55
References **57**

Author index **61**

Subject index **63**

Tables

Figures

1.1 Vertical outlier, good leverage point and bad leverage point 4

1.2 Huber loss function ρ_κ^{H} and score function ψ_κ^{H} 9

1.3 Tukey-Biweight loss function ρ_κ^{B} and score function ψ_κ^{B} 9

1.4 Caption needed 37

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

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 $\mathbf{x} = (x_1, \ldots, x_p)^t$, 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
$\mathbf{x}^t$, $\mathbf{X}^t$	transposition of a vector or a matrix
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
...	...

1 Robust linear regression

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

1.1 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

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon \tag{1.1}$$

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

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, n \tag{1.2}$$

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 (1.2) can be more compactly written as

$$y_i = \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.3}$$

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

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

equations (1.3) 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 (1.3) 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 (1.3) may be formulated as follows:

$$y_i = \mathbf{x}_i^t\boldsymbol{\beta} + \sigma\nu_i, \qquad i = 1, \ldots, n \tag{1.4}$$

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:

$$F_{y_i|\mathbf{x}_i}(y) = \Pr(y_i \leq y|\mathbf{x}_i) = \Pr\left(\nu_i \leq \frac{y - \mathbf{x}_i^t\boldsymbol{\beta}}{\sigma}\middle|\mathbf{x}_i\right) = F_{0,1}\left(\frac{y - \mathbf{x}_i^t\boldsymbol{\beta}}{\sigma}\right) \tag{1.5}$$

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

$$f_{y_i|\mathbf{x}_i}(y) = \frac{1}{\sigma} f_{0,1}\left(\frac{y - \mathbf{x}_i^t\boldsymbol{\beta}}{\sigma}\right). \tag{1.6}$$

Hence, $\mathbf{x}_i^t\boldsymbol{\beta}$ 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 1.6).

❑ Remark

The classic *location-scale model* may be seen as a particular case of regression model (1.4). 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

$$y_i = \beta_0 + \sigma\nu_i, \qquad i = 1, \ldots, n, \tag{1.7}$$

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.

1: 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.

2: 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.

❏

1.2 Different types of outliers

Model (1.1) 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 1.1 (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.

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

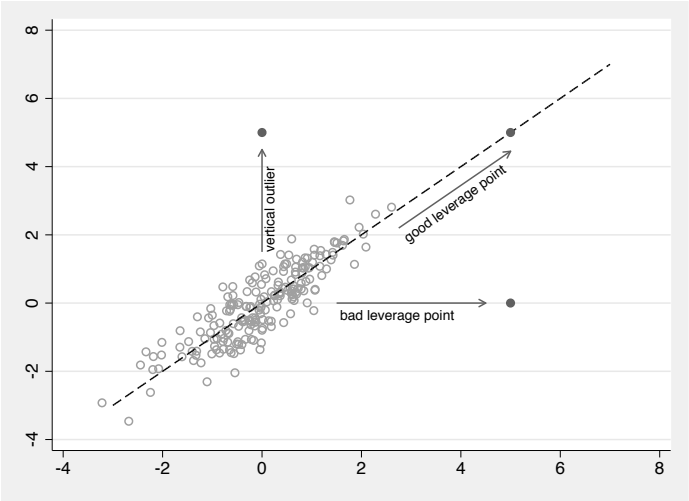

Figure 1.1. Vertical outlier, good leverage point and bad leverage point

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).

3: Please provide citation details!

The most popular estimation method in linear regression is certainly the *least squares* (LS) method introduced in 1805 by Legendre. 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.

1.3 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:

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

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:

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) \tag{1.8}$$

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

$$\sum_{i=1}^{n} r_i(\boldsymbol{\beta})\mathbf{x}_i = \mathbf{0} \tag{1.9}$$

which is equivalent to the linear equations system

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y}.$$

If $\mathbf{X}$ has full rank[1], then the solution of (1.9) is unique and is given by

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} = \widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}. \tag{1.10}$$

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 (1.9) that the residuals $r_i(\widehat{\boldsymbol{\beta}}_{\text{LS}})$, $i = 1, \ldots, n$, have zero average.

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

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\boldsymbol{\gamma}) = \widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) + \boldsymbol{\gamma} \qquad \text{for all } \boldsymbol{\gamma} \in \mathbb{R}^{p+1} \tag{1.11}$$

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \lambda\mathbf{y}) = \lambda\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}) \qquad \text{for all } \lambda \in \mathbb{R} \tag{1.12}$$

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

$$\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}\mathbf{A}, \mathbf{y}) = \mathbf{A}^{-1}\widehat{\boldsymbol{\beta}}_{\text{LS}}(\mathbf{X}, \mathbf{y}). \tag{1.13}$$

The properties (1.11), (1.12) and (1.13) 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.

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 $\mathbf{X}\mathbf{a} \neq \mathbf{0}$ for all $\mathbf{a} \neq \mathbf{0}$. This is equivalent to the nonsingularity of $\mathbf{X}'\mathbf{X}$.

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 (1.2). 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.

1.4 M estimation

1.4.1 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

$$\widehat{\boldsymbol{\beta}}_{\text{LAD}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} |r_i(\boldsymbol{\beta})|. \tag{1.14}$$

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 $\boldsymbol{\beta}$:

$$\sum_{i=1}^{n} \text{sign}(r_i(\widehat{\boldsymbol{\beta}}_{\text{LAD}}))\mathbf{x}_i = \mathbf{0} \tag{1.15}$$

If the model contains an intercept term, (1.15) 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 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 (1.14) and (1.15) 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).

1.4.2 The principle of M estimation

Huber (1964) hence generalized median regression to a wider class of estimators, called

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.

M estimators, by considering other functions than the absolute value in (1.14) in order to find a reasonable balance between robustness and Gaussian efficiency.

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

$$\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{y_i - \mathbf{x}_i^t\boldsymbol{\beta}}{\widehat{\sigma}}\right) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right) \tag{1.16}$$

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 $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ scale equivariant; see (1.12). 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 (1.16) 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 1.2):

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

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 functions (see figure 1.3)

4: Please provide citation!

$$\rho_\kappa^{\mathrm{B}}(u) = \begin{cases} \frac{\kappa^2}{6}\left[1 - \left(1 - \left(\frac{u}{\kappa}\right)^2\right)^3\right] & \text{if } |u| \le \kappa \\ \frac{\kappa^2}{6} & \text{if } |u| > \kappa \end{cases} \tag{1.17}$$

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 $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ as a solution of the estimating equations system obtained by differentiating the function to minimize in (1.16) with respect to each component of $\boldsymbol{\beta}$, that is, as a solution of the equations system

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0} \tag{1.18}$$

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

$$\psi^{\text{H}}_{\kappa}(u) = \begin{cases} -2\kappa & \text{if } u < -\kappa \\ 2u & \text{if } -\kappa \le u \le \kappa \\ 2\kappa & \text{if } u > \kappa \end{cases}$$

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

$$\psi^{\text{B}}_{\kappa}(u) = \begin{cases} u\Big(1 - \big(\frac{u}{\kappa}\big)^2\Big)^2 & \text{if } |u| \le \kappa \\ 0 & \text{if } |u| > \kappa \end{cases}$$

(see figures 1.2 and 1.3). If the loss function ρ is convex on $\mathbb{R}$—this is the case for ρ^{H}_{κ}—the score function ψ is monotone (non decreasing) on $\mathbb{R}$ and $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ is called a *monotone* regression M estimator; if ρ is bounded—this is the case for ρ^{B}_{κ}—the score function ψ vanishes out of a certain interval of $\mathbb{R}$ and $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ is then called a *redescending* regression M estimator.

The main advantage of monotone score functions ψ is that all solutions of (1.18) are solutions of (1.16). In the case of redescending score functions ψ, the estimating equations (1.18) 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 $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ defined by (1.16), which makes the computation of the M estimate considerably more complex.

❑ Remark

Applying the M estimation procedure in the particular case of the location-scale model (1.7) 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 1.8 at the end of this chapter.

❑

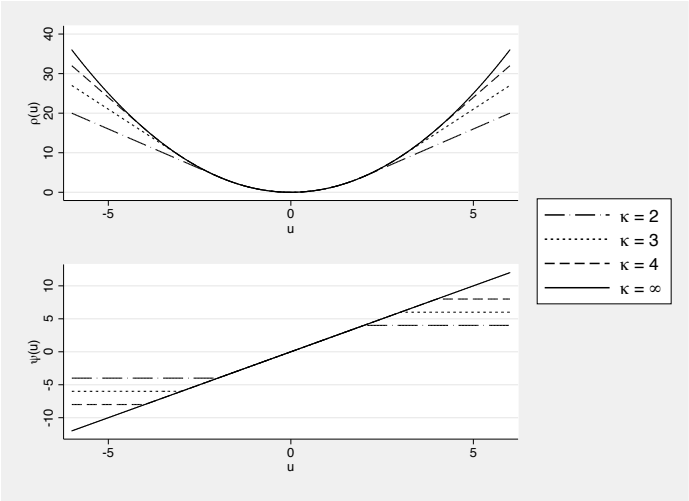

Figure 1.2. Huber loss function ρ_κ^{H} and score function ψ_κ^{H}

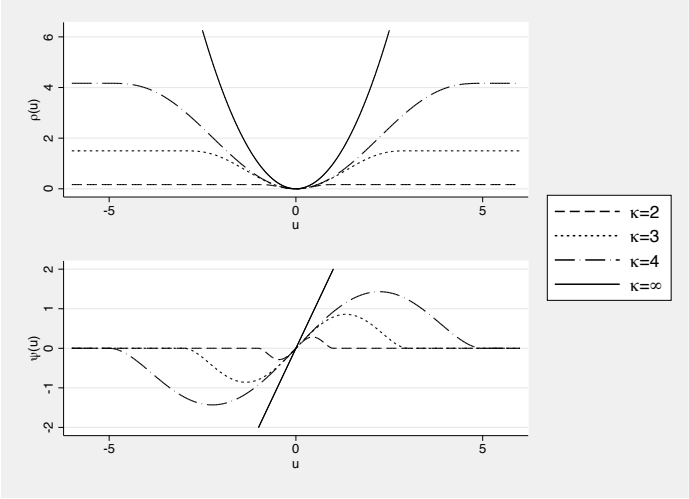

Figure 1.3. Tukey-Biweight loss function ρ_κ^{B} and score function ψ_κ^{B}

1.4.3 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 (1.4) with

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

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

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

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

$$\begin{aligned}
\left(\widehat{\boldsymbol{\beta}}_{\text{ML}}^t, \widehat{\sigma}_{\text{ML}}\right)^t &= \arg\max_{\boldsymbol{\beta},\sigma} \frac{1}{\sigma^n} \prod_{i=1}^{n} f_{0,1}\left(\frac{y_i - \mathbf{x}_i^t \boldsymbol{\beta}}{\sigma}\right) \\
&= \arg\max_{\boldsymbol{\beta},\sigma} \left[\sum_{i=1}^{n} \ln f_{0,1}\left(\frac{y_i - \mathbf{x}_i^t \boldsymbol{\beta}}{\sigma}\right) - n \ln \sigma\right] \\
&= \arg\min_{\boldsymbol{\beta},\sigma} \left[\sum_{i=1}^{n} \rho_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) + n \ln \sigma\right] \qquad (1.19)
\end{aligned}$$

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

$$\widehat{\boldsymbol{\beta}}_{\text{ML}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) \qquad (1.20)$$

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 (1.20), 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 (1.19) is

$$\begin{cases} \sum_{i=1}^{n} \psi_{\text{ML}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) \mathbf{x}_i = \mathbf{0} \\ \frac{1}{n} \sum_{i=1}^{n} \rho_{\text{ML;scale}}\left(\frac{r_i(\boldsymbol{\beta})}{\sigma}\right) = \delta \end{cases} \qquad (1.21)$$

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

1.4.4 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 $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ is solution of the estimating equations system (1.18) 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 (1.18) can be rewritten as

$$\sum_{i=1}^{n} w_i r_i(\boldsymbol{\beta}) \mathbf{x}_i = \mathbf{0} \tag{1.22}$$

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 (1.9) related to LS estimation, and if the w_i's were known, the equations (1.22) 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 (1.16) 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 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 $r_i(\widehat{\boldsymbol{\beta}}_0) = r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{LAD}})$ are equal to zero, determining the MAD of the n residuals could lead to underestimating σ when p is large.

3. In the case of a convex loss function ρ, the convergence of the algorithm to the global minimum of (1.16) is guaranteed whatever the starting point $\widehat{\boldsymbol{\beta}}_0$.

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 $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$.

1.4.5 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:

$$\widehat{\boldsymbol{\beta}}_\alpha = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\alpha(y_i - \mathbf{x}_i^t\boldsymbol{\beta}) \tag{1.23}$$

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

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

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 $\mathbf{x}_i^t\boldsymbol{\beta}_\alpha$. Then $\widehat{\boldsymbol{\beta}}_\alpha$ defined by (1.23) 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).

1.4.6 Monotone vs. redescending M estimators

As already mentioned, taking a loss function $\rho(u)$ in the minimization problem (1.16) 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 $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ with respect to leverage points? To answer to this question, let us recall that $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ is solution of the estimating equations system (1.18).

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 (1.18) 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 (1.18) to zero, the residual

$r_i(\widehat{\boldsymbol{\beta}}) = y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}$ 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 (1.18) 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.

1.4.7 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

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\widetilde{w}(d(\underline{\mathbf{x}}_i))\mathbf{x}_i = \mathbf{0} \tag{1.24}$$

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

$$\sum_{i=1}^{n} \eta\left(d(\underline{\mathbf{x}}_i), \frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}}\right)\mathbf{x}_i = \mathbf{0} \tag{1.25}$$

where for each s, $\eta(s,u)$ is a nondecreasing and bounded ψ-function of u. The estimating equations system (1.24) may be seen as a particular case of (1.25) 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

$$d(\underline{\mathbf{x}}_i) = \sqrt{\left(\underline{\mathbf{x}}_i - \widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}\right)^t \widehat{\boldsymbol{\Sigma}}_{\underline{\mathbf{x}}}^{-1}\left(\underline{\mathbf{x}}_i - \widehat{\boldsymbol{\mu}}_{\underline{\mathbf{x}}}\right)} \tag{1.26}$$

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 **??**). 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 (1.26) are affine equivariant, which presents the same computational difficulties as for redescending M estimates and reduce substantially the appeal of this estimator.

1.5 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 LAD regression protects against outlying y_i, it cannot cope with grossly aberrant values of $\mathbf{x}_i$: LAD 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.

1.5.1 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

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

where $1/2 \leq \alpha \leq 1$ and $r^2_{(1)}(\boldsymbol{\beta}) \leq \cdots \leq r^2_{(n)}(\boldsymbol{\beta})$ are the ordered squared residuals. 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 (1999); 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

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

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`.

6: Add paragraph on LQS, as this is also supported by robreg.

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?

1.5.2 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

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

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, 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 $E(\rho(u)) = E(u^2) = 1$.

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

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{s_\rho(e_1, \ldots, e_n)}\right) = \delta \tag{1.27}$$

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 $\delta = E(\rho(u))$ 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 (1.27)).

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, that is, satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{s_\rho(r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))}\right) = \delta \qquad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^{p+1}. \tag{1.28}$$

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 1.16) in that

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho})}{\widehat{\sigma}_\rho}\right) \leq \sum_{i=1}^{n} \rho\left(\frac{r_i(\widetilde{\boldsymbol{\beta}})}{\widehat{\sigma}_\rho}\right) \qquad \text{for all } \widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1} \tag{1.29}$$

where the residuals are standardized by the same scale M estimate $\widehat{\sigma}_\rho = s_\rho(r_1(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}), \ldots, r_n(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}))$ of σ on both sides of the inequality (1.29). 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 (1.28). This means that, if we denote $\widehat{\sigma}_\rho = s_\rho(r_1(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}), \ldots, r_n(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}))$ and $\widetilde{\sigma}_\rho = s_\rho(r_1(\widetilde{\boldsymbol{\beta}}), \ldots, r_n(\widetilde{\boldsymbol{\beta}}))$ for $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^{p+1}$, we have $\widehat{\sigma}_\rho \leq \widetilde{\sigma}_\rho$ and

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

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

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

which proves (1.29).

If ρ has a derivative ψ, it follows that $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ is also an M estimate in the sense of (1.18), 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 $\widehat{\sigma}_\rho = s_\rho(r_1(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}), \ldots, r_n(\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}))$, with $\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho}$ and $\widehat{\sigma}_\rho$ such that

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

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 (1.17), with $\kappa = 1.547$, is a common choice. This S estimator resists to a contamination of up to 50% of outliers and, hence, has 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%.

1.5.3 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%. Furthermore, 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 they 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 Gaussian efficiency) where the scale measure s_{ρ_0} is defined by

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

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

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

$$\sum_{i=1}^{n}\rho\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\rho_0}}\right) \tag{1.30}$$

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

such that

$$\sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}{\widehat{\sigma}_{\rho_0}}\right) \leq \sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0})}{\widehat{\sigma}_{\rho_0}}\right). \tag{1.31}$$

The key result is given in Yohai (1987). Recall that all local minima of (1.30) are solutions of the estimating equations (1.18) with $\psi(u) = \rho'(u)$ and $\widehat{\sigma} = \widehat{\sigma}_{\rho_0}$:

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\boldsymbol{\beta})}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i = \mathbf{0} \tag{1.32}$$

Yohai shows that if $\rho(u) \leq \rho_0(u)$ for all $u \in \mathbb{R}$ and if (1.31) is satisfied, then $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ is consistent. Moreover, it can be shown that the MM estimator $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ has the same breakdown point than the S estimator $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ of the first step, determined by the function $\rho_0(\cdot)$. If, furthermore, $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ is any solution of (1.32), then it has the same efficiency—this efficiency is determined by the choice of the function $\rho(\cdot)$—as the global minimum of (1.30). In conclusion, it is not necessary to find the absolute minimum of (1.30) to ensure consistency, a high breakdown point and a high efficiency.

It is common to use a Tukey-Biweight $\rho^{\text{B}}_{\kappa}(\cdot)$ 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 section 1.6.4.

Numerical computation of the S and MM estimate

The numerical computation of the estimate $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ at the second step of the procedure follows the approach described in section 1.4.4: starting with $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$, we use the iteratively reweighted least squares (IRWLS) algorithm to attain a solution of the equation (1.32). It may be shown (see Maronna et al. (2006) that (1.30) decreases at each iteration, which insures (1.31). Hence, once the initial S estimate is computed, $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min}\, s_\rho(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta})). \tag{1.33}$$

An approximate solution of (1.33) can be obtained by finding $\widehat{\boldsymbol{\beta}}$ equal to

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

where

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

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]

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

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

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

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 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

$$N \geq \frac{|\log\alpha|}{|\log(1-(1-\xi)^{p+1})|} \approx \frac{|\log\alpha|}{(1-\xi)^{p+1}} \tag{1.34}$$

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

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.

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

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

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 $\widehat{\sigma}_{\rho;M} = s_\rho(r_1(\widehat{\boldsymbol{\beta}}_M), \ldots, r_n(\widehat{\boldsymbol{\beta}}_M))$. Since ρ is a monotone function, the inequality $\widehat{\sigma}_{\rho;M} < \widehat{\sigma}_{\rho;M-1}$ implies that

$$n\delta = \sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_M)}{\widehat{\sigma}_{\rho;M}}\right) \geq \sum_{i=1}^{n} \rho\left(\frac{r_i(\widehat{\boldsymbol{\beta}}_M)}{\widehat{\sigma}_{\rho;M-1}}\right). \tag{1.35}$$

Consequently, if we observe that $\sum_{i=1}^{n} \rho(r_i(\widehat{\boldsymbol{\beta}}_M)/\widehat{\sigma}_{\rho;M-1}) > n\delta$, 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 (1.35).

Although the N given by (1.34) ensures that the approximation $\widehat{\boldsymbol{\beta}}$ of $\widehat{\boldsymbol{\beta}}_{\text{S};\rho}$ 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 using 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 1.6).

1.5.4 MS estimation

Explicit formulas for $\widehat{\boldsymbol{\beta}}_{\text{S};\rho}$ 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 (1.1). 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

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

continuous and dichotomous variables in (1.1) and rewrite the regression model equation as follows:

$$y = (\beta_0 + \beta_1 x_1 + \cdots + \beta_{p_1} x_{p_1}) + (\beta_1^* x_1^* + \cdots + \beta_{p_2}^* x_{p_2}^*) + \varepsilon \tag{1.36}$$

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 (1.36), then $\boldsymbol{\beta}^*(\beta_1^*, \ldots, \beta_{p_2}^*)^t$ 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 $\widehat{\boldsymbol{\beta}}_{\text{MS}}^{(k)}$ and $\widehat{\boldsymbol{\beta}}_{\text{MS}}^{*(k)}$ as follows. Let s_ρ be a measure of dispersion satisfying (1.28), $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip_1})^t$ and $\mathbf{x}_i^* = (x_{i1}^*, \ldots, x_{ip_2}^*)^t$:

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

where

$$\widehat{\sigma}^{(k-1)} = s_\rho\left(y_i - \mathbf{x}_i^t \widehat{\boldsymbol{\beta}}_{\text{MS}}^{(k-1)} - (\mathbf{x}_i^*)^t \widehat{\boldsymbol{\beta}}_{\text{MS}}^{*(k-1)}; i = 1, \ldots, n \right).$$

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).

1.6 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

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

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* and *ergodic* process H.[9] 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 coincide 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.

1.6.1 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 (Section 1.9) 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 (1.4). Here, the observations

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. Furthermore, 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.

$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ 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*.[10]

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

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

where s_{ρ_0} is a measure of dispersion satisfying

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

This leads to the scale M estimator

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

Let $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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): $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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 $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ and by $\boldsymbol{\beta}_0$ when it is estimated by $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$. Moreover, we will use the generic notations $u_0 = (y - \mathbf{x}^t\boldsymbol{\beta}_0)/\sigma$ and $u = (y - \mathbf{x}^t\boldsymbol{\beta})/\sigma$, 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

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

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

where

$$\mathbf{V}_{\text{MM}} = \mathbf{G}_{\text{MM}}^{-1}\boldsymbol{\Omega}_{\text{MM}}\big(\mathbf{G}_{\text{MM}}^{t}\big)^{-1} \tag{1.37}$$

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

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

$$\mathbf{G}_{\text{MM}} = -\frac{1}{\sigma} E \begin{pmatrix} \psi' \mathbf{x}\mathbf{x}^t & \mathbf{0} & \psi' u \mathbf{x} \\ \mathbf{0} & \rho_0'' \mathbf{x}\mathbf{x}^t & \rho_0'' u_0 \mathbf{x} \\ \mathbf{0} & \mathbf{0} & \rho_0' u_0 \end{pmatrix} \tag{1.38}$$

and

$$\mathbf{\Omega}_{\text{MM}} = E \begin{pmatrix} \psi^2 \mathbf{x}\mathbf{x}^t & \psi \rho_0' \mathbf{x}\mathbf{x}^t & \psi \rho_0 \mathbf{x} \\ \psi \rho_0' \mathbf{x}\mathbf{x}^t & (\rho_0')^2 \mathbf{x}\mathbf{x}^t & \rho_0 \rho_0' \mathbf{x} \\ \psi \rho_0 \mathbf{x}^t & \rho_0 \rho_0' \mathbf{x}^t & \rho_0^2 - \delta^2 \end{pmatrix}. \tag{1.39}$$

❑

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

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

$$\begin{aligned} \text{Avar}(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}) = \frac{1}{n} \big[& \mathbf{A} E(\psi^2 \mathbf{x}\mathbf{x}^t) \mathbf{A} - \mathbf{a} E(\psi \rho_0 \mathbf{x}^t) \mathbf{A} \\ & - \mathbf{A} E(\psi \rho_0 \mathbf{x}) \mathbf{a}^t + E(\rho_0^2 - \delta^2) \mathbf{a}\mathbf{a}^t \big] \end{aligned} \tag{1.40}$$

$$\begin{aligned} \text{Avar}(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}) = \frac{1}{n} \big[& \mathbf{A}_{\text{S}} E((\rho_0')^2 \mathbf{x}\mathbf{x}^t) \mathbf{A}_{\text{S}} - \mathbf{a}_{\text{S}} E(\rho_0 \rho_0' \mathbf{x}^t) \mathbf{A}_{\text{S}} \\ & - \mathbf{A}_{\text{S}} E(\rho_0 \rho_0' \mathbf{x}) \mathbf{a}_{\text{S}}^t + E(\rho_0^2 - \delta^2) \mathbf{a}_{\text{S}} \mathbf{a}_{\text{S}}^t \big] \end{aligned} \tag{1.41}$$

$$\begin{aligned} \text{Acov}(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}) = \frac{1}{n} \big[& \mathbf{A} E(\psi \rho_0' \mathbf{x}\mathbf{x}^t) \mathbf{A}_{\text{S}} - \mathbf{a} E(\rho_0 \rho_0' \mathbf{x}^t) \mathbf{A}_{\text{S}} \\ & - \mathbf{A} E(\psi \rho_0 \mathbf{x}) \mathbf{a}_{\text{S}}^t + E(\rho_0^2 - \delta^2) \mathbf{a} \mathbf{a}_{\text{S}}^t \big] \end{aligned} \tag{1.42}$$

with

$$\mathbf{A} = \sigma \big[E(\psi' \mathbf{x}\mathbf{x}^t) \big]^{-1} \tag{1.43}$$

$$\mathbf{a} = \mathbf{A} \frac{E(\psi' u \mathbf{x})}{E(\rho_0' u_0)} \tag{1.44}$$

$$\mathbf{A}_{\text{S}} = \sigma \big[E(\rho_0'' \mathbf{x}\mathbf{x}^t) \big]^{-1} \tag{1.45}$$

$$\mathbf{a}_{\text{S}} = \mathbf{A}_{\text{S}} \frac{E(\rho_0'' u_0 \mathbf{x})}{E(\rho_0' u_0)}. \tag{1.46}$$

❑ 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 $(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}^t, \widehat{\sigma}_{\rho_0})^t$ are presented in Appendix 2 (Section 1.9) 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 $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$.
2. Replace $E(\cdot)$ by $\frac{1}{n}\sum_{i=1}^{n}(\cdot)$.

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

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

with

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

It is interesting to note that the estimate $\widehat{\text{Avar}}(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})$ of the asymptotic variance $\text{Avar}(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})$ 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 $\psi((y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})/\widehat{\sigma}_{\rho_0})$ has a small value when ψ is a redescending function.[11] Hence, if there are bad leverage points in the sample, then their $\mathbf{x}_i$-value is large, but at the same time $\psi((y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})/\widehat{\sigma}_{\rho_0})$ will be zero. This explains intuitively why vertical outliers and bad leverage points have only a limited influence on the estimate $\widehat{\text{Avar}}(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})$.

❑ Remark

As previously explained, the LS estimator $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ 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 $\rho_0''(u_0)=2$) 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 $\text{Avar}(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0})$:

$$\begin{aligned}\text{Avar}(\widehat{\boldsymbol{\beta}}_{\text{LS}})=\frac{1}{n}\big[&\mathbf{A}_{\text{LS}}E(4u_0^2\mathbf{x}\mathbf{x}^t)\mathbf{A}_{\text{LS}}-\mathbf{a}_{\text{LS}}E(2u_0^3\mathbf{x}^t)\mathbf{A}_{\text{LS}}\\&-\mathbf{A}_{\text{LS}}E(2u_0^3\mathbf{x})\mathbf{a}_{\text{LS}}^t+E(u_0^4-1)\mathbf{a}_{\text{LS}}\mathbf{a}_{\text{LS}}^t\big]\end{aligned}\tag{1.47}$$

with

$$\mathbf{A}_{\text{LS}}=\frac{\sigma}{2}\big[E(\mathbf{x}\mathbf{x}^t)\big]^{-1}\qquad\text{and}\qquad\mathbf{a}_{\text{LS}}=\mathbf{A}_{\text{LS}}\frac{E(u_0\mathbf{x})}{E(u_0^2)}.\tag{1.48}$$

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

If, in addition, there is homoskedasticity[12] and if the distribution $F_{0,1}$ of the error terms is symmetric (around 0), we retrieve the well-known asymptotic variance matrix

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

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 not 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:

$$\begin{aligned}\mathrm{Acov}(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}) = \frac{1}{n}\big[&\mathbf{A}E(2u\rho_0'\mathbf{x}\mathbf{x}^t)\mathbf{A}_{\mathrm{S}} - \mathbf{a}E(\rho_0\rho_0'\mathbf{x}^t)\mathbf{A}_{\mathrm{S}} \\ &- \mathbf{A}E(2u\rho_0\mathbf{x})\mathbf{a}_{\mathrm{S}}^t + E(\rho_0^2 - \delta^2)\mathbf{a}\mathbf{a}_{\mathrm{S}}^t\big]\end{aligned} \tag{1.49}$$

with

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

while $\mathbf{A}_{\mathrm{S}}$ and $\mathbf{a}_{\mathrm{S}}$ remain unchanged with respect to (1.45) and (1.46).

❑

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 (Section 1.9). 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.

1.6.2 Robust confidence intervals and tests with robust regression estimators

As just explained, we may consider that, under the model (1.4) and Assumption A,[13] a robust M, S or MM estimator $\widehat{\boldsymbol{\beta}}$ is, for large n, approximately normally distributed with mean $\boldsymbol{\beta}$ and variance $\widehat{\mathrm{Avar}}(\widehat{\boldsymbol{\beta}})$, where $\widehat{\mathrm{Avar}}(\widehat{\boldsymbol{\beta}})$ corresponds to the empirical counterpart of the asymptotic matrix $\mathrm{Avar}(\widehat{\boldsymbol{\beta}})$ specified in the previous subsection. This result underlies the inference procedures developed for linear combinations of the regression parameters.

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

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 $\mathcal{N}(\gamma, \widehat{\sigma}_\gamma^2)$ where

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

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

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

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-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

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

or equivalently, since the approximate normal distribution of $\widehat{\gamma}$ implies that $((\widehat{\gamma} - \gamma)/\widehat{\sigma}_\gamma)^2 \approx \chi_1^2$, 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 chi-square distribution with one degree of freedom.

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

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

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

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

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 $\mathcal{N}_q(\boldsymbol{\gamma}, \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}})$ with

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

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 chi-square 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

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

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

1.6.3 Robust R-squared

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

$$\begin{aligned} R^2 &= \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2}, \end{aligned} \tag{1.50}$$

where ESS, TSS and RSS are respectively the explained, total and residual sum of squares. Note that $y_i - \widehat{y}_i = r_i\left(\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}\right)$ are the LS residuals. Moreover, $\overline{y}$ is the LS estimate of $\mu = \mathrm{E}(y)$, that is the LS estimate of the intercept β_0 in the linear regression model (1.1) 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.,

$$R^2 = \left(\frac{\sum_{i=1}^n (y_i - \overline{y})\left(\widehat{y}_i - \overline{\widehat{y}}\right)}{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}\sqrt{\sum_{i=1}^n \left(\widehat{y}_i - \overline{\widehat{y}}\right)^2}} \right)^2, \tag{1.51}$$

with $\overline{\widehat{y}}$ the arithmetic mean of the predicted responses. Equation (1.51) 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

$$\phi^2 = \max_{\boldsymbol{\beta}} \mathrm{Corr}^2\big(y, \mathbf{x}^t\boldsymbol{\beta}\big), \tag{1.52}$$

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.,

$$R^2_{\mathrm{adj}} = 1 - \big(1 - R^2\big)\left(\frac{n-1}{n-(p+1)}\right). \tag{1.53}$$

It is rather obvious that the R^2 given by (1.50) can be driven by extreme observations, not only through the LS estimator $\widehat{\boldsymbol{\beta}}_{\mathrm{LS}}$ 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 (1.16), a robust coefficient of determination can be defined by

$$R^2_\rho = 1 - \frac{\displaystyle\sum_{i=1}^n \rho\left(\frac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}}{\widehat{\sigma}}\right)}{\displaystyle\sum_{i=1}^n \rho\left(\frac{y_i - \widehat{\mu}_{\mathrm{M};\rho}}{\widehat{\sigma}}\right)}, \tag{1.54}$$

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 (1.50) given by

$$R^2_{\mathrm{S}} = 1 - \frac{s\Big(y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}\ ; i = 1, \ldots, n\Big)}{s(y_i - \widehat{\mu}\ ; i = 1, \ldots, n)} \tag{1.55}$$

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

Although (1.54) and (1.55) are direct generalizations of (1.50) 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 (1.51) 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 1.4.4, 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

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

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 $\rho^B_\kappa(\cdot)$ given by (1.17), we have

$$w_i = \begin{cases} \left(1 - \left(\dfrac{r_i\Big(\widehat{\boldsymbol{\beta}}\Big)}{\kappa\widehat{\sigma}}\right)^2\right)^2 & \text{if } \left|\dfrac{r_i\Big(\widehat{\boldsymbol{\beta}}\Big)}{\widehat{\sigma}}\right| \leq \kappa \\ 0 & \text{if } \left|\dfrac{r_i\Big(\widehat{\boldsymbol{\beta}}\Big)}{\widehat{\sigma}}\right| > \kappa. \end{cases}$$

Then a robust version of (1.51) is given by

$$R_w^2 = \left(\frac{\sum_{i=1}^n w_i (y_i - \overline{y}_w)\left(\widehat{y}_i - \overline{\widehat{y}}_w\right)}{\sqrt{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2}\sqrt{\sum_{i=1}^n w_i \left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2}} \right)^2, \tag{1.56}$$

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 (1.50):

$$\widetilde{R}_w^2 = 1 - \frac{\sum_{i=1}^n w_i (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n w_i (y_i - \overline{y}_w)^2}. \tag{1.57}$$

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:

$$\widetilde{R}_{w,a}^2 = \frac{\sum_{i=1}^n w_i \left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2}{\sum_{i=1}^n w_i \left(\widehat{y}_i - \overline{\widehat{y}}_w\right)^2 + a \sum_{i=1}^n w_i (y_i - \widehat{y}_i)^2}, \tag{1.58}$$

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 (1.52) if we take[14]

$$a = \frac{\mathrm{E}\left[\frac{\psi(u)}{u}\right]}{\mathrm{E}[\psi(u)]}, \quad \text{with } u \sim \mathcal{N}(0,1).$$

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

$$\widetilde{R}_{w,a;\mathrm{adj}}^2 = 1 - \left(1 - \widetilde{R}_{w,a}^2\right)\left(\frac{n-1}{n-(p+1)}\right). \tag{1.59}$$

14. 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 $a = 1.2076$.

1.6.4 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 (1.1). 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 $\underline{\widehat{\boldsymbol{\beta}}} = \left(\widehat{\beta}_1, \ldots, \widehat{\beta}_p\right)^t$, such that $\boldsymbol{\beta} = \left(\beta_0, \underline{\boldsymbol{\beta}}^t\right)^t$ and $\widehat{\boldsymbol{\beta}} = \left(\beta_0, \underline{\widehat{\boldsymbol{\beta}}}^t\right)^t$.

Some preliminary results

The development of the tests proposed in Dehon et al. (2012) relies on the results presented in Subsection 1.6.1 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 1.6.1, 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 (1.4) 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,

$$\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho} \approx \mathcal{N}_{p+1}\Big(\boldsymbol{\beta}, \text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)\Big),$$
$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} \approx \mathcal{N}_{p+1}\Big(\boldsymbol{\beta}, \text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)\Big)$$

and

$$\widehat{\boldsymbol{\beta}}_{\text{LS}} \approx \mathcal{N}_{p+1}\Big(\boldsymbol{\beta}, \text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big)\Big),$$

where the matrices $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$, $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$ and $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big)$ are given by (1.40), (1.41) and (1.47), respectively. Moreover,

$$\begin{aligned}&\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\\ &\approx \mathcal{N}_{p+1}\Big(\mathbf{0}, \text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big) + \text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big) - 2\text{Acov}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)\Big),\end{aligned}$$

where $\text{Acov}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$ is given by (1.42). Since $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$, $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ and $\text{Acov}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$ may be consistently estimated by their empirical counterparts[15] $\widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$, $\widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ and $\widehat{\text{Acov}}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$, we have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho} \approx \mathcal{N}_{p+1}\Big(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}\Big)$$

where

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})} = \widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big) + \widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)\\ - 2\widehat{\text{Acov}}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big).\end{aligned} \tag{1.60}$$

If we only consider the slopes estimates, we simply have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\underline{\widehat{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\text{MM};\rho_0,\rho} \approx \mathcal{N}_p\Big(\underline{\mathbf{0}}, \underline{\widehat{\boldsymbol{\Sigma}}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}\Big) \tag{1.61}$$

15. As explained in Subsection 1.6.1, 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 $\underline{\widehat{\boldsymbol{\Sigma}}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}$ is the matrix $\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}$ without its first line and its first column.

Following a similar approach, we have, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{LS}} \approx \mathcal{N}_{p+1}\Big(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})}\Big)$$

with

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})} = \widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big) + \widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big) \\ - 2\widehat{\text{Acov}}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)\end{aligned} \tag{1.62}$$

where $\widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$, $\widehat{\text{Avar}}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big)$ and $\widehat{\text{Acov}}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$ are the empirical counterparts of the matrices $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$, $\text{Avar}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}\Big)$ and $\text{Acov}\Big(\widehat{\boldsymbol{\beta}}_{\text{LS}}, \widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big)$ given by (1.41), (1.47) and (1.49), respectively. As a consequence, under $\mathcal{H}_0$ and Assumption A, for large n:

$$\underline{\widehat{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\text{LS}} \approx \mathcal{N}_p\Big(\underline{\mathbf{0}}, \underline{\widehat{\boldsymbol{\Sigma}}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})}\Big) \tag{1.63}$$

where $\underline{\widehat{\boldsymbol{\Sigma}}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})}$ is the matrix $\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})}$ without its first line and its first column.

Comparison of LS and S

Let us consider the classical LS estimator $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ and the S-estimator $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ 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 (1.17): if the tuning constant κ is set at 1.547, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ has a breakdown point equal to 50% (but a rather low Gaussian efficiency of only 28%). Under the null hypothesis (and Assumption A), $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ and $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ are both consistant estimators of $\boldsymbol{\beta}$, but $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ has a higher Gaussian efficiency. Under the alternative hypothesis of a moderate contamination, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ still converges to $\boldsymbol{\beta}$ (see Omelka and Salibian-Barrera 2010) but it is not the case for $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ anymore (the outliers distort the LS estimate and introduce a bias, in such a way that $\widehat{\boldsymbol{\beta}}_{\text{LS}}$ possesses another limit in probability than $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$).

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

$$H = \Big(\underline{\widehat{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\text{LS}}\Big)^t \underline{\widehat{\boldsymbol{\Sigma}}}^{-1}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}-\widehat{\boldsymbol{\beta}}_{\text{LS}})} \Big(\underline{\widehat{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \underline{\widehat{\boldsymbol{\beta}}}_{\text{LS}}\Big), \tag{1.64}$$

with $\widehat{\boldsymbol{\Sigma}}_{\Big(\underline{\widehat{\boldsymbol{\beta}}}_{\text{S};\rho_0}-\underline{\widehat{\boldsymbol{\beta}}}_{\text{LS}}\Big)}$ computed from (1.62). It follows from (1.63) 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 $\widehat{\underline{\boldsymbol{\beta}}}_{\text{LS}}$ and $\widehat{\underline{\boldsymbol{\beta}}}_{\text{S};\rho_0}$ 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 $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ by an MM estimator $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$, 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 $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ [16]. For instance, if we take for ρ the Tukey-Biweight function (1.17) with the tuning constant κ equal to 4.685, the Gaussian efficiency of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ attains 95%, and for $\kappa = 6.256$, the Gaussian efficiency of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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 $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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 $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ (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

$$H = \Big(\widehat{\underline{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\text{MM};\rho_0,\rho}\Big)^t \widehat{\underline{\boldsymbol{\Sigma}}}^{-1}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})} \Big(\widehat{\underline{\boldsymbol{\beta}}}_{\text{S};\rho_0} - \widehat{\underline{\boldsymbol{\beta}}}_{\text{MM};\rho_0,\rho}\Big), \tag{1.65}$$

with $\widehat{\boldsymbol{\Sigma}}_{(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0} - \widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho})}$ given by (1.60). Under the null hypothesis $\mathcal{H}_0$, $\widehat{\underline{\boldsymbol{\beta}}}_{\text{S};\rho_0}$ and $\widehat{\underline{\boldsymbol{\beta}}}_{\text{MM};\rho_0,\rho}$ are both consistant estimators of $\underline{\boldsymbol{\beta}}$ and $H \approx \chi^2_p$. Under the alternative hypothesis $\mathcal{H}_1$, i.e., under a moderate contamination of the sample, the bias of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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 $\widehat{\underline{\boldsymbol{\beta}}}_{\text{MM};\rho_0,\rho}$ and hence distorts the MM-estimation with respect to the S-estimation — if

$$H > \chi^2_{p;1-\alpha},$$

16. Recall here that $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ possesses the same breakdown point as $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$.

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 $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$) and to retain ultimately the MM estimator that, while not being significantly different from $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ 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.

1.7 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 1.4, 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).

7: Graph-Ex-1.do was missing; I recreated the graph from the description.

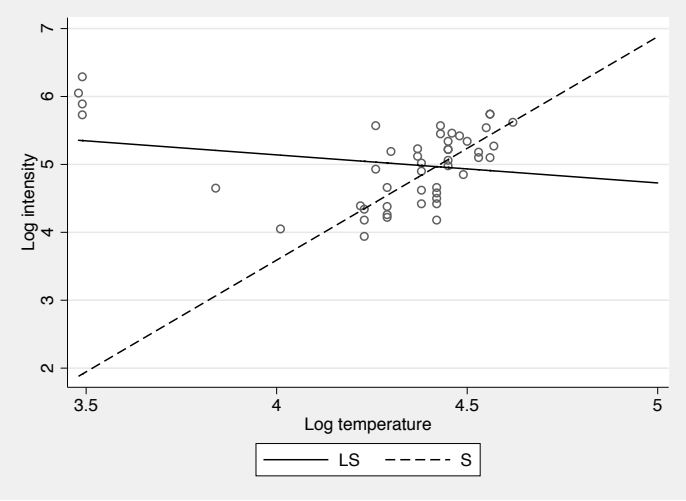

Figure 1.4. Caption needed

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
------------------------------------------------------------------------------
```

The results of the LS estimation (solid line in figure 1.4) 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
------------------------------------------------------------------------------
              |               Robust
log_intensity |      Coef.   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.

8: 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
------------------------------------------------------------------------------
             |               Robust
log_intensity|      Coef.   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 1.6.4, 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
```

log_intensity	Coef.	Robust 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 1.6.4) 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 1.6.4 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 prefered. 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 (1.26)). 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 provided 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 1.5.3 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.

◁

1.8 Appendix 1: M-estimators of location and scale

The application of the M-estimation approach in the particular case of the location-scale model (1.7) leads to the M-estimators of location and scale.

1.8.1 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:

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \mu}{\widehat{\sigma}}\right) = 0, \tag{1.66}$$

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 (1.66) 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[17] (see Huber and Ronchetti 2009, 54). The asymptotic breakdown point is nul if ψ is unbounded. If ψ is equal to the function $\psi^{\mathrm{B}}_{\kappa}$ and hence is redescending, the breakdown point of $\widehat{\mu}_{\mathrm{M};\rho}$ 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)[18].

17. The breakdown point of $\widehat{\boldsymbol{\beta}}_{\mathrm{M};\rho}$ is actually equal to the breakdown point of the preliminary estimator of the scale parameter σ.
18. Note however that it is possible to prove that, using the MAD as initial scale estimator, the breakdown point of $\widehat{\mu}_{\mathrm{M};\rho^{\mathrm{B}}_{\kappa}}$ is strictly greater than 0.49 in the Gaussian case.

1.8.2 M-estimator of scale

A M-estimate $\widehat{\sigma}_{\mathrm{M};\rho}$ of the scale parameter σ is defined as the solution of the equation

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i-\widehat{\mu}}{\sigma}\right)=\delta \tag{1.67}$$

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 $\widehat{\sigma}_{\mathrm{M};\rho}$ for σ, we have to take $\delta=\mathrm{E}_{F_{0,1}}[\rho(\nu)]$. An usual choice for the loss function ρ is the Tukey-Biweight function ρ_κ^B defined by (1.17).

The M-estimators of scale are translation invariant and scale equivariant. The influence function of the functional S associated to the scale M-estimator $\widehat{\sigma}_{\mathrm{M};\rho}$ under the distribution $F_{0,1}$ of the error term ν of the location-scale model is given by

$$\mathrm{IF}(u;S,F_{0,1})=\frac{\rho(u)-\delta}{\mathrm{E}_{F_{0,1}}[\rho'(\nu)\nu]}.$$

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 (1.66) and (1.67) (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 $\widehat{\mu}_{\mathrm{M};\rho}$ an asymptotic breakdown point smaller than 50% — hence, smaller than the breakdown point attainable by using the MAD as preliminary 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.

❑

1.9 Appendix 2: Generalized Method of Moments (GMM) and asymptotic distributions of regression M, S and MM estimators

1.9.1 GMM-estimation principle

For simplicity, let us consider immediately the context of the regression model (1.1). 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*[19] and *ergodic*[20] 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*.[21]

Suppose that our objective is to estimate the functional $\boldsymbol{\theta} = \boldsymbol{\theta}(H)$ that is implicitly defined by the equation

$$\mathrm{E}_H[\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})] = \mathbf{0}, \tag{1.68}$$

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 (1.68) 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 (1.68), that is,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{m}\Big(y_i, \mathbf{x}_i, \widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}\Big) = \mathbf{0}. \tag{1.69}$$

Under regularity conditions detailed in Hansen (1982), the GMM estimator $\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}}$ defined by (1.69) has a limiting normal distribution:

$$\sqrt{n}\Big(\widehat{\boldsymbol{\theta}}_{\mathrm{GMM}} - \boldsymbol{\theta}\Big) \to^d \mathcal{N}(\mathbf{0}, \mathbf{V}), \tag{1.70}$$

where, in the exactly-identified case,

$$\mathbf{V} = \mathbf{G}^{-1}\boldsymbol{\Omega}\big(\mathbf{G}^t\big)^{-1}, \tag{1.71}$$

with[22]

$$\mathbf{G} = \mathrm{E}\left[\frac{\partial \mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t}\right] \quad \text{and} \quad \boldsymbol{\Omega} = \mathrm{E}\big[\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})\mathbf{m}^t(y, \mathbf{x}, \boldsymbol{\theta})\big]. \tag{1.72}$$

19. 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.
20. 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.
21. The interested reader can find very general results, valid in presence of autocorrelation, in Croux et al. (2003).
22. Here and later, we simply write $\mathrm{E}[\cdot]$ for $\mathrm{E}_H[\cdot]$.

1.9.2 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 $\widehat{\boldsymbol{\beta}}_{\text{M};\rho}$ and the M-scale estimator $\widehat{\sigma}_{\rho_0}$ are such that

$$\left\{\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\psi\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{M};\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{M};\rho}}{\widehat{\sigma}_{\rho_0}}\right)-\delta=0 \end{array}\right. \tag{1.73}$$

where $\psi(u)=\rho'(u)$, δ is a selected constant and, using similar notations as in the previous sections, $\widehat{\sigma}_{\rho_0}=s_{\rho_0}\Big(r_1\Big(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}\Big),\ldots,r_n\Big(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}\Big)\Big)$. This shows that the M-estimator $\Big(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}^t,\widehat{\sigma}_{\rho_0}\Big)^t$ is an exactly-identified Gmm estimator for $\boldsymbol{\theta}=\big(\boldsymbol{\beta}^t,\sigma\big)^t$, with

$$\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})=\left(\begin{array}{c}\psi\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)\mathbf{x}\\ \rho_0\left(\dfrac{y-\mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)-\delta\end{array}\right). \tag{1.74}$$

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 $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ as follows:

$$\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}=\arg\min_{\boldsymbol{\beta}} s_{\rho_0}(r_1(\boldsymbol{\beta}),\ldots,r_n(\boldsymbol{\beta})) \tag{1.75}$$

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

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

As previously explained, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$ satisfy the first order conditions

$$\left\{\begin{array}{l} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0'\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i=\mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\rho_0\left(\dfrac{y_i-\mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)-\delta=0. \end{array}\right. \tag{1.76}$$

Note that the equations (1.76) are of the same form as (1.73). 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 (1.76) — the set of equations involving ρ_0' — may have multiple solutions. Therefore one usually uses (1.75) to compute the S-estimate but (1.76) to determine its asymptotic distribution. Actually, (1.76) 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

$$\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta}) = \begin{pmatrix} \rho_0'\left(\dfrac{y - \mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right)\mathbf{x} \\ \rho_0\left(\dfrac{y - \mathbf{x}^t\boldsymbol{\beta}}{\sigma}\right) - \delta \end{pmatrix}.$$

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

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \psi\left(\dfrac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{MM};\rho_0,\rho}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i = \mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \rho_0'\left(\dfrac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right)\mathbf{x}_i = \mathbf{0} \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \rho_0\left(\dfrac{y_i - \mathbf{x}_i^t\widehat{\boldsymbol{\beta}}_{\mathrm{S};\rho_0}}{\widehat{\sigma}_{\rho_0}}\right) - \delta = 0. \end{cases} \tag{1.77}$$

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 (1.77) 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,

$$\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta}) = \begin{pmatrix} \psi\mathbf{x} \\ \rho_0'\mathbf{x} \\ \rho_0 - \delta \end{pmatrix}, \tag{1.78}$$

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.

1.9.3 Asymptotic variance matrix of an mm estimator

If the observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, are generated by a stationary and ergodic process, and are independent (Assumption A1)

The first-order equivalence of $(\widehat{\boldsymbol{\beta}}^t_{\text{MM};\rho_0,\rho}, \widehat{\boldsymbol{\beta}}^t_{\text{S};\rho_0}, \widehat{\sigma}_{\rho_0})^t$ with a GMM estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \boldsymbol{\beta}_0^t, \sigma)^t$ allows us to conclude that, if the observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are generated by a *stationary* and *ergodic* process, and are *independent* (Assumption A1)[23],

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

where

$$\mathbf{V}_{\text{MM}} = \mathbf{G}_{\text{MM}}^{-1}\boldsymbol{\Omega}_{\text{MM}}\left(\mathbf{G}_{\text{MM}}^t\right)^{-1},$$

with the matrices $\mathbf{G}_{\text{MM}}$ and $\boldsymbol{\Omega}_{\text{MM}}$ obtained by applying relations (1.72) to the moment function (1.78):

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

and

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

In particular, this result establishes the consistency of the MM-regression estimator $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$. Moreover, using the upper left $(p+1)\times(p+1)$ submatrice of $\mathbf{V}_{\text{MM}}$, we obtain

23. This Assumption A1 coincides with Assumption A in Section 1.6.1. 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.

that the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ is equal to

$$\begin{aligned}\text{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big) = \frac{1}{n}[&\mathbf{A}\text{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A} - \mathbf{a}\text{E}\big(\psi\rho_0\mathbf{x}^t\big)\mathbf{A} \\ &- \mathbf{A}\text{E}(\psi\rho_0\mathbf{x})\mathbf{a}^t + \text{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}\mathbf{a}^t],\end{aligned}$$

where

$$\mathbf{A} = \sigma\big[\text{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \quad \text{and} \quad \mathbf{a} = \mathbf{A}\frac{\text{E}(\psi' u\mathbf{x})}{\text{E}(\rho_0' u_0)}.$$

This expression of $\text{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ is then estimated by its empirical counterpart $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$, 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 $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$, $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ and $\widehat{\sigma}_{\rho_0}$.
2. Replace $\text{E}(\cdot)$ by $\frac{1}{n}\sum_{i=1}^n(\cdot)$.

For example, the first term of $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ is given by

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

with

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

Using standard asymptotic arguments, it can be shown that $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ is a consistent estimate of $\text{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$. From $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$, standard errors for the regression coefficients are obtained in the usual way: for $j = 0, 1, \ldots, p$,

$$\widehat{\text{se}}\Big(\Big[\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big]_j\Big) = \sqrt{\Big[\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)\Big]_{jj}}.$$

Moreover, the estimate $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ of the asymptotic variance $\text{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ 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[24]. 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 $\widehat{\text{Avar}}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$.

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

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

$$\begin{aligned}\text{Avar}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big) = \frac{1}{n}[&\text{E}\big(\psi^2\big)\mathbf{A}_2\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\mathbf{A}_2 - \text{E}(\psi\rho_0)\mathbf{a}_2\text{E}\big(\mathbf{x}^t\big)\mathbf{A}_2 \\ &- \text{E}(\psi\rho_0)\mathbf{A}_2\text{E}(\mathbf{x})\mathbf{a}_2^t + \text{E}\big(\rho_0^2 - \delta^2\big)\mathbf{a}_2\mathbf{a}_2^t],\end{aligned}$$

where

$$\mathbf{A}_2 = \sigma\frac{[\text{E}(\mathbf{x}\mathbf{x}^t)]^{-1}}{\text{E}(\psi')} \quad \text{and} \quad \mathbf{a}_2 = \mathbf{A}_2\frac{\text{E}(\psi' u)\text{E}(\mathbf{x})}{\text{E}(\rho_0' u_0)}.$$

Taking the empirical counterpart yields $\widehat{\text{Avar}}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$. 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 $\widehat{\text{Avar}}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ 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 $\mathbf{a} = \mathbf{0}$.

Under Assumptions A1 and A3, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ becomes

$$\begin{aligned}\text{Avar}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big) &= \frac{1}{n}\mathbf{A}\text{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\mathbf{A} \\ &= \frac{\sigma^2}{n}\big[\text{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1}\text{E}\big(\psi^2\mathbf{x}\mathbf{x}^t\big)\big[\text{E}\big(\psi'\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.\end{aligned}$$

The empirical counterpart of the latter expression, $\widehat{\text{Avar}}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$, is an estimate of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ 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 $\widehat{\text{Avar}}_{13}$ compared to $\widehat{\text{Avar}}_1$: the authors of Croux et al. (2003) then recommend to use $\widehat{\text{Avar}}_1$ in any case.

When all of Assumptions A1, A2 and A3 hold, then $\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}$ has asymptotic variance

$$\text{Avar}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big) = \frac{\sigma^2}{n}\frac{\text{E}\big(\psi^2\big)}{\big[\text{E}(\psi')\big]^2}\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.$$

This corresponds to the expression for the variance of the MM-regression estimator that was derived in Yohai (1987). The empirical counterpart $\widehat{\text{Avar}}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$ is an estimate of this asymptotic variance that, as $\widehat{\text{Avar}}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\text{MM};\rho_0,\rho}\Big)$, lacks robustness with respect to leverage points.

1.9.4 Asymptotic variance matrix of an S-estimator

If Assumption A1 holds, the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ is simply derived from the central $(p+1)\times(p+1)$ submatrice of $\mathbf{V}_{\text{MM}}$ (cf. (1.37), (1.38) and (1.39)):

$$\begin{aligned}\text{Avar}_1\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big) = \frac{1}{n}[&\mathbf{A}_{\text{S}}\text{E}\Big(\big(\rho_0'\big)^2\mathbf{x}\mathbf{x}^t\Big)\mathbf{A}_{\text{S}} - \mathbf{a}_{\text{S}}\text{E}\big(\rho_0\rho_0'\mathbf{x}^t\big)\mathbf{A}_{\text{S}}\\ &- \mathbf{A}_{\text{S}}\text{E}(\rho_0\rho_0'\mathbf{x})\mathbf{a}_{\text{S}}^t + \text{E}\big(\rho_0^2-\delta^2\big)\mathbf{a}_{\text{S}}\mathbf{a}_{\text{S}}^t],\end{aligned}$$

where

$$\mathbf{A}_{\text{S}} = \sigma\big[\text{E}\big(\rho_0''\mathbf{x}\mathbf{x}^t\big)\big]^{-1} \quad \text{and} \quad \mathbf{a}_{\text{S}} = \mathbf{A}_{\text{S}}\frac{\text{E}(\rho_0''u_0\mathbf{x})}{\text{E}(\rho_0'u_0)}.$$

If, in addition, A2 holds, then the asymptotic variance matrix of $\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}$ takes the form

$$\begin{aligned}\text{Avar}_{12}\Big(\widehat{\boldsymbol{\beta}}_{\text{S};\rho_0}\Big) = &\frac{\sigma^2}{n}\frac{\text{E}\Big(\big(\rho_0'\big)^2\Big)}{\big[\text{E}(\rho_0'')\big]^2}\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1} + \frac{\sigma^2}{n}\frac{\text{E}(\rho_0''u_0)}{\big[\text{E}(\rho_0'')\big]^2\text{E}(\rho_0'u_0)}\\ &\times\left\{\frac{\text{E}(\rho_0''u_0)\text{E}\big(\rho_0^2-\delta^2\big)}{\text{E}(\rho_0'u_0)} - 2\text{E}(\rho_0\rho_0')\right\}\\ &\times\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}\text{E}(\mathbf{x})\text{E}\big(\mathbf{x}^t\big)\big[\text{E}\big(\mathbf{x}\mathbf{x}^t\big)\big]^{-1}.\end{aligned}$$

Under A3, the expressions are the same as those for the MM estimator, with ψ replaced by ρ_0'.

1.9.5 Asymptotic variance matrix of an M-estimator

Here the expressions are less explicit. Under Assumption A1, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{M};,\rho}$ is derived from the upper left $(p+1)\times(p+1)$ block of $\mathbf{G}_{\text{M}}^{-1}\boldsymbol{\Omega}_{\text{M}}(\mathbf{G}_{\text{M}}^t)^{-1}$ where

$$\mathbf{G}_{\text{M}} = \text{E}\left[\frac{\partial\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^t}\right] \quad \text{and} \quad \boldsymbol{\Omega}_{\text{M}} = \text{E}\big[\mathbf{m}(y,\mathbf{x},\boldsymbol{\theta})\mathbf{m}^t(y,\mathbf{x},\boldsymbol{\theta})\big],$$

with $\mathbf{m}(y, \mathbf{x}, \boldsymbol{\theta})$ given by (1.74). 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

$$\mathbf{G}_{\text{M}} = -\frac{1}{\sigma}\mathrm{E}\begin{pmatrix} \psi'\mathbf{x}\mathbf{x}^t & \psi' u\mathbf{x} \\ \rho_0'\mathbf{x}^t & \rho_0' u \end{pmatrix}$$

and

$$\boldsymbol{\Omega}_{\text{M}} = \mathrm{E}\begin{pmatrix} \psi^2\mathbf{x}\mathbf{x}^t & \psi\rho_0\mathbf{x} \\ \psi\rho_0\mathbf{x}^t & \rho_0^2 - \delta^2 \end{pmatrix}.$$

If in addition Assumption A2 holds, then

$$\mathbf{G}_{\text{M}} = -\frac{1}{\sigma}\begin{pmatrix} \mathrm{E}(\psi')\mathrm{E}(\mathbf{x}\mathbf{x}^t) & \mathrm{E}(\psi' u)\mathrm{E}(\mathbf{x}) \\ \mathrm{E}(\rho_0')\mathrm{E}(\mathbf{x}^t) & \mathrm{E}(\rho_0' u) \end{pmatrix}$$

and

$$\boldsymbol{\Omega}_{\text{M}} = \begin{pmatrix} \mathrm{E}\big(\psi^2\big)\mathrm{E}(\mathbf{x}\mathbf{x}^t) & \mathrm{E}(\psi\rho_0)\mathrm{E}(\mathbf{x}) \\ \mathrm{E}(\psi\rho_0)\mathrm{E}(\mathbf{x}^t) & \mathrm{E}\big(\rho_0^2\big) - \delta^2 \end{pmatrix}.$$

Under Assumption A3, the expressions of $\mathrm{Avar}_{13}\Big(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}\Big)$ and $\mathrm{Avar}_{123}\Big(\widehat{\boldsymbol{\beta}}_{\text{M};\rho}\Big)$ are exactly similar to those for the MM estimator.

References

Anderson, T. W. 1984. *An Introduction to Multivariate Statistical Analysis*. 2nd ed. John Wiley & Sons.

Andrews, D. F. 1974. A Robust Method for Multiple Linear Regression. *Technometrics* 16: 523–531.

Bramati, M. C., and C. Croux. 2007. Robust estimators for the fixed effects panel data model. *The Econometrics Journal* 10(3): 521–540.

Brown, G. W., and A. M. Mood. 1951. On Median Tests for Linear Hypotheses. In *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*, 159–166.

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.

Edgeworth, F. Y. 1887. On Observations Relating to Several Quantities. *Hermathena* 6: 279–285.

Gervini, D., and V. J. Yohai. 2002. A class of robust and fully efficient regression estimators. *The Annals of Statistics* 30: 583–616.

Greene, W. 1997. *Econometric Analysis*. 3rd ed. Prentice Hall.

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.

Hössjer, O. 1992. On the optimality of S-estimators. *Statistics and Probability Letters* 14(5): 413–419.

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.

———. 1981. *Robust Statistics*. New York: John Wiley & Sons.

Jaeckel, L. A. 1972. Estimating Regression Coefficients by Minimizing the Dispersion of Residuals. *Annals of Mathematical Statistics* 5: 1449–1458.

Koenker, R. 2005. *Quantile Regression*. Cambridge: Cambridge University Press.

Koenker, R., and G. Bassett. 1978. Regression Quantiles. *Econometrica* 46(1): 33–50.

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.

Maronna, R. A., and V. J. Yohai. 2000. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference* 89(1-2): 197–214.

Mendes, B. V. M., and D. E. Tyler. 1996. Constrained M-estimation for regression. In *Robust Statistics, Data Analysis and Computer Intensive Methods (Schloss Thurnau, 1994)*, ed. EDITORS?, 299–320. Lecture Notes in Statistics 109, New York: Springer.

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. 1983. Regression Techniques with High Breakdown Point. *The Institute of Mathematical Statistics Bulletin* 12: 155–???

———. 1984. Least Median of Squares Regression. *Journal of the American Statistical Association* 79(388): 871–880.

Rousseeuw, P. J., and A. M. Leroy. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley & Sons.

Rousseeuw, P. J., and K. Van Driessen. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics* 41(3): 212–223.

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.

Sen, P. K. 1968. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association* 63: 1379–1389.

Siegel, A. F. 1982. Robust Regression Using Repeated Medians. *Biometrika* 69: 242–244.

Theil, H. 1950. A Rank-Invariant Method of Linear and Polynomial Regression Analysis (Parts 1-3). *Nederlandsche Akademie van Wetenschappen Proceedings (Ser. A)* 53: 386–392; 521–525; 1397–1412.

Yohai, V. J. 1987. High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics* 15(2): 642–656.

Yohai, V. J., and R. A. Maronna. 1979. Asymptotic behavior of M-estimates for the linear model. *The Annals of Statistics* 7: 258–268.

Yohai, V. J., and R. H. Zamar. 1988. High breakdown estimates of regression by means of the minimization of an efficient scale. *Journal of the American Statistical Association* 83: 406–413.

Author index

Subject index