PERGAMON

Mathematical and Computer Modelling 35 (2002) 361–374

MATHEMATICAL
AND
COMPUTER
MODELLING

www.elsevier.com/locate/mcm

Alternative Methods of Linear Regression

A. GILONI
Sy Syms School of Business, Yeshiva University
500 West 185th Street
New York, NY 10033, U.S.A.

M. PADBERG*
Operations Management Department, Stern School of Business
New York University, 40 West 4th Street
New York, NY 10012, U.S.A.

(Received January 2001; accepted May 2001)

Abstract—This paper is a survey on traditional linear regression techniques using the ℓ_1-, ℓ_2-, and ℓ_∞-norm. We derive the characterization of the respective regression estimates (including optimality and uniqueness criteria), as well as discuss some of their statistical properties. © 2002 Elsevier Science Ltd. All rights reserved.

Keywords—ℓ_1-, ℓ_2-, ℓ_∞-regression, Linear programming.

1. INTRODUCTION

Often one wishes to determine whether or not there exists a (causal) relationship between hypothesized predictor or *independent* variables and some response or *dependent* variable. Furthermore, at times, one needs to forecast the value of the response variable based on some assumed relationship. In many situations, it is not in our power to determine that such a relationship is valid just by assuming or asserting a mathematical model of the relationship. Instead, data must be collected from the population of available data of these variables and an *empirical* relationship between the dependent and independent variables in question must be established on the basis of the data.

The practitioner may wish to develop a concrete model from an assumed relationship, i.e., some arbitrary mathematical function $G(x_1, \dots, x_p) = y$, where $x_1, \dots, x_p$ are the independent variables and y is the dependent variable. In this paper, the function G will be assumed to be a linear function L. Although more general cases can be considered, often they can be dealt with by transforming the variables. In order to test the proposed model, the practitioner must obtain a sample of data as described. Recognizing the existence of imprecision, the model must be modified to include a random error term, thereby giving the *linear regression model* $y = L(x_1, \dots, x_p) + \varepsilon$.

In the case of forecasting, the objective is not to test the validity of a hypothesized relationship between the variables, but rather to "invent" a relationship which adequately describes the variation in the data. However, it is essential that the strength or statistical significance of the created model, as well as of the forecasts be described. This description is also useful for the

*Supported in part by a grant from the Office of Naval Research (N00014-96-0327).

PII: S0895-7177(01)00170-4

case of testing a proposed model, since one would want to know the level of assurance that a model is correct. In both cases, generally only linear relationships will be permitted, although at times transformations of the dependent and/or independent variables can be performed to include nonlinear ones.

Thus, one must determine some line or a hyperplane which is closest to the data under some distance criterion or function. Recognizing the importance of considering errors in a model of this type was already done by Galileo Galilei (1564–1642) in the 17th century and determining the line which best fits three or more points was studied as early as the 18th century by Roger Joseph Boscovich (1711–1787). The technique that resulted is formally known as linear regression and was developed by Adrien Marie Legendre (c. 1752–1833), Carl Friedrich Gauss (1777–1855), Joseph Fourier (1768–1830), and many, many other eminent mathematicians. See the voluminous papers by Harter [1–6] for an exhaustive treatment of the history of regression up to the mid-1970s.

There are different distance functions or metrics which can be utilized to perform linear regression. Therefore, the original problem is categorized under the class of mathematical problems. In order to solve these problems, a remarkably wide range of mathematical techniques are invoked. At times, classical analysis is sufficient, while other problems require the use of linear programming (LP) and even discrete optimization. Furthermore, several of the possible metrics are interrelated and approximation theory becomes a useful tool as well. In this chapter, several metrics will be discussed, as well as their respective statistical properties, quality of fit, and possible refinements.

2. TRADITIONAL LINEAR REGRESSION TECHNIQUES

To formalize the linear regression model, we assume that we have n measurements or observations on the dependent variable y and some number $p \geq 1$ of independent variables $x_1, \dots, x_p$ of each one for which we know n values as well. We denote

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} x_1^1 & \cdot & \cdot & \cdot & x_p^1 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ x_1^n & \cdot & \cdot & \cdot & x_p^n \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{x}^n \end{pmatrix} = (\mathbf{x}_1, \dots, \mathbf{x}_p), \tag{2.0.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of n observations and $\mathbf{X}$ is an $n \times p$ matrix of reals frequently referred to as the *design matrix*. Furthermore, $\mathbf{x}_1, \dots, \mathbf{x}_p$ are column vectors with n components and $\mathbf{x}^1, \dots, \mathbf{x}^n$ are row vectors with p components corresponding to the columns and rows of $\mathbf{X}$, respectively.

The statistical (or hypothesized) linear regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.0.2}$$

where $\boldsymbol{\beta}^\top = (\beta_1, \dots, \beta_p)$ is the vector of *parameters* of the linear model and $\boldsymbol{\varepsilon}^\top = (\varepsilon_1, \dots, \varepsilon_n)$ a vector of n random variables corresponding to the error terms in the asserted relationship. An upper index $\top$ denotes "transposition" of a vector or matrix throughout this work. In the statistical model, the dependent variable y, thus, is a random variable for which we obtain measurements or observations that contain some "noise" or measurement errors that are captured in the error terms $\boldsymbol{\varepsilon}$. On the other hand, for the numerical problem that we are facing, we write

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}, \tag{2.0.3}$$

where given some arbitrarily fixed parameter vector $\boldsymbol{\beta}$, the components r_i of the vector $\mathbf{r}^\top = (r_1, \dots, r_n)$ are the *residuals* that result, given the observations $\mathbf{y}$, a fixed design matrix $\mathbf{X}$, and the chosen vector $\boldsymbol{\beta} \in \mathbb{R}^p$. The residuals $\mathbf{r}$ are thus in terms of the statistical model, *realizations* of the random error terms $\boldsymbol{\varepsilon}$ given the particular observations $\mathbf{y}$ and parameter settings $\boldsymbol{\beta}$. Given $\mathbf{y}$ and $\mathbf{X}$, the general objective in linear regression is to find parameter settings $\boldsymbol{\beta} \in \mathbb{R}^p$ such that some appropriate measure of the dispersion of the resulting residuals $\mathbf{r} \in \mathbb{R}^n$ is as small as possible.

We note that it is entirely possible that, e.g., $x_1^j = 1$, for all $j \in \{1, \ldots, n\}$ in the design matrix $\mathbf{X}$. In this case, we refer to β_1 as the "intercept term" corresponding to the situation in the two parameter case, i.e., when $p = 2$. If $x_1^j = 1$, for all $j \in \{1, \ldots, n\}$ and $p = 1$, the problem of finding a "best" fitting scalar β_1 means that we want some good measure of "centrality" of the observations $\mathbf{y}$.

The notion of what is "best" can be made precise using different norms on $\mathbb{R}^n$ and we discuss next the most commonly used ones.

2.1. L$_2$ or Least Squares Regression

Least squares regression is, by far, the most well-known and utilized regression technique. The regression estimates $\boldsymbol{\beta}$ are found by minimizing the sum of squared residuals under the Euclidean (or ℓ_2-) norm $\|\mathbf{x}\|_2 = \sqrt{\sum x_j^2}$, i.e., we wish to find parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ such that

$$S = \mathbf{r}^\top \mathbf{r} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\boldsymbol{\beta}^\top (\mathbf{X}^\top \mathbf{y}) + \boldsymbol{\beta}^\top (\mathbf{X}^\top \mathbf{X})\, \boldsymbol{\beta} \tag{2.1.1}$$

is minimum. Note that the expression is not the value of the Euclidean norm of the residuals, but rather the square of the norm. This transformation is monotone, and thus, it does not affect optimality. The function to be minimized is positive semidefinite, and thus, the first-order conditions do the job, i.e., to minimize S, its gradient ∇S with respect to $\boldsymbol{\beta}$ must be calculated and set to zero or

$$\frac{1}{2}\nabla S = -\mathbf{X}^\top \mathbf{y} + (\mathbf{X}^\top \mathbf{X}) = 0\boldsymbol{\beta}. \tag{2.1.2}$$

The equations $(\mathbf{X}^\top \mathbf{X}) = \boldsymbol{\beta}\mathbf{X}^\top \mathbf{y}$ that must be solved for $\boldsymbol{\beta}$ are called the *normal equations* for ℓ_2-regression.

Assuming that the rank of $\mathbf{X}$ is p, i.e., that $r(\mathbf{X}) = p$, it follows that $(\mathbf{X}^\top \mathbf{X})^{-1}$ exists and the optimal $\boldsymbol{\beta} = \boldsymbol{\beta}^{\mathrm{LS}}$ is given by

$$\boldsymbol{\beta}^{\mathrm{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}, \tag{2.1.3}$$

i.e., the ℓ_2-norm yields a *unique* optimum. For matters of our analysis, we will make this rank assumption, the solution is not unique, but we could work with some "pseudo-inverse" of $\mathbf{X}^\top \mathbf{X}$, which we will not do. Clearly, the computational effort is centered about the inversion of $\mathbf{X}^\top \mathbf{X}$, but a numerical solver must be able to cope with the possible singularity of $\mathbf{X}^\top \mathbf{X}$.

Alternatively, the least squares regression estimates $\boldsymbol{\beta}^{\mathrm{LS}}$ can be found by linear programming, and thus, do not require an explicit inversion of $\mathbf{X}^\top \mathbf{X}$ nor consideration of singular matrices, as this is done automatically by any commercial LP solver. The gradient with respect to $\boldsymbol{\beta}$ is calculated componentwise

$$\frac{\partial S}{\partial \beta_k} = -2 \sum_{\ell=1}^{n} \left(y_\ell - \sum_{j=1}^{p} x_j^\ell \beta_j \right) x_k^\ell, \qquad \text{for } k = 1, \ldots, p. \tag{2.1.4}$$

The problem of finding a solution to the normal equations (2.1.2) then is the linear program

$$\begin{aligned} &\text{minimize} && z \\ &\text{subject to} && z + \sum_{j=1}^{p} \left(\sum_{i=1}^{n} x_j^i x_k^i \right) \beta_j = \sum_{i=1}^{n} y_i x_k^i, \qquad \text{for } k = 1, \ldots, p, \\ & && z \geq 0, \qquad \beta_1, \ldots, \beta_p \text{ free.} \end{aligned} \tag{2.1.5}$$

This linear program attains its optimal solution at $z = 0$ with the p normal equations enforced, thereby giving least squares estimates. Although there is no computational difficulty without full-column rank of the design matrix, in such a case, the linear program will not have a unique solution either. However, any commercial LP solver can be used to find least squares regression estimates, and in the case of nonuniqueness, alternative optima can be determined automatically.

Primarily due to the convenient closed form (2.1.3) of least squares regression estimators, the statistical linear regression model has been studied intensively for well over 200 years now and—under the assumption of a *normal* (or *Gaussian* or *exponential*) distribution of the error terms ε—an impressive statistical apparatus has been created to assess the goodness of fit, the quality of individual and/or subsets of the regression coefficients, as well as other statistical properties of the linear regression model. The very assumption of the *normality* of the distribution of the error terms has, however, been under attack from the very beginning as, for instance, in the words of Francis Ysidro Edgeworth (1845–1926) written in 1883 where he submits that "... the *ancient solitary reign* of the exponential (Gaussian) law of error should come to an end" (cited in [1, p. 167]). On the other hand, some of the properties of the statistical linear regression model alluded to above require seemingly weak assumptions. See, e.g., [7, p. 41].

Statistical properties of the least squares estimators β^{LS} of the "true" parameters β of the linear regression model, such as, e.g., their asymptotic "consistency", are easy to prove under the Gauss-Markov conditions (see [8]). The entire analysis of least squares regression has been extended to the case where the error terms ε follow a normal distribution with mean zero and covariance matrix $\mathbf{\Sigma}$. As in the Gauss-Markov case where the common error variance σ^2 may *a priori* be known or unknown, $\mathbf{\Sigma}$ may be known or unknown. In either case, for "small" sample sizes, the resulting distributions of the regression estimates β^{LS} when viewed as random variables can be calculated directly and tabulated, while for "large" sample sizes, asymptotic distributions can be found (under certain additional restrictions). Based on these distributional results, a multitude of tests for the significance of, as well as confidence intervals (or ellipsoids) for individual (or subsets of the) regression estimates—none of which we will summarize here and all of which are valid given the *assumed* normality of the distribution of the error terms ε—have been developed and are readily available to the practitioner of least squares linear regression.

If and when the error terms in the linear regression model do indeed follow a normal distribution, then the least squares regression estimates β^{LS} are "best estimators" under most acceptable criteria that the statistical profession has developed in the past two centuries or so. However, as we have pointed out above, the very assumption of the general applicability of the normal law of errors has been under attack from the very beginning of the development of linear regression and, in particular, the least squares analysis hinges critically on the existence of the second moment of the error distribution. Thus, if we must assume or believe that the error distribution follows, for instance, a Cauchy distribution or any "long-tailed" distribution having no finite second moment, then the elegant arguments made in favor of the least squares regression estimators become invalid, and thus, it may become mandatory to look for other criteria to find "best" estimators for the linear regression model (2.0.2).

2.2. $\mathbf{L_1}$ or Least Sum of Absolute Deviations Regression

Proposed apparently by the Jesuit Boscovich in the 18th century and studied, among many others, by Pierre Simon Laplace (1749–1827), Fourier, Gauss and Edgeworth in the 19th century, ℓ_1-regression came to a new life in the 1960s with the observation by Fama [9], Mandelbrot [10], Blattberg and Sargent [11], Sharpe [12], and others that stock-market prices, market-indices, and other economic time series cannot be explained nor predicted well enough in the traditional least squares setting. In most cases, remarkably better explanatory or predictive results were obtained through the use of ℓ_1-regression where the estimates of the parameters $\beta \in \mathbb{R}^p$ of the linear regression model are found by minimizing the sum of the *absolute* (rather than squared) residuals, i.e., here we wish to find $\beta \in \mathbb{R}^p$ such that

$$\|\mathbf{y} - \mathbf{X}\beta\|_1 = \sum_{i=1}^{n} |y_i - \mathbf{x}^i\beta| \tag{2.2.1}$$

is minimum and the ℓ_1-norm $\|\mathbf{x}\|_1 = \sum |x_j|$ instead of the ℓ_2-norm is employed in the process of minimizing the dispersion of the residuals. Clearly, this objective function is nondifferentiable, calculus does not help, and a closed form solution—unlike the least squares case—simply does not exist. Even worse is that it was not clear for a long time what solution method to use to find the minimum in (2.2.1) and that even under the assumption of a full rank design matrix, the uniqueness of an optimal solution cannot be guaranteed. With the advent of computers and numerical techniques such as linear programming, ℓ_1-regression has, however, become a viable alternative to least squares regression. As a result, both empirical and theoretical studies addressing the properties of ℓ_1-regression have again been augmented considerably in the past 30 years or so. L_1-regression is called in the literature by a multitude of names that all mean the same: LAD (least absolute deviation), LAE (least absolute error), LAV (least absolute value), LAR (least absolute residual), LSAD (least sum of absolute deviations), MAD (minimum absolute deviation), MSAE (minimum sum of absolute errors), and so forth. All but LSAD and MSAE are truly misnomers because they paraphrase the ℓ_∞-norm rather than as intended the ℓ_1-norm.

The key to the numerical solution of the ℓ_1-regression problem is the fact (known already to Fourier, Charles Jean Gustave Nicolas de la Vallée Poussin (1866–1962) and probably others) that the minimization problem (2.2.1) is a linear programming (LP) problem of the form

$$\begin{array}{ll} \text{minimize} & \mathbf{e}_n^T\mathbf{r}^+ + \mathbf{e}_n^T\mathbf{r}^- \\ \text{subject to} & \mathbf{X}\beta + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}, \\ & \beta \text{ free}, \qquad \mathbf{r}^+ \geq \mathbf{0}, \quad \mathbf{r}^- \geq \mathbf{0}. \end{array} \tag{2.2.2}$$

Other forms of writing the ℓ_1-regression problem as a linear program are possible and can be found in the literature. In the formulation (2.2.2), the residuals $\mathbf{r}$ of the general form (2.0.3) are simply replaced by a difference $\mathbf{r}^+ - \mathbf{r}^-$ of nonnegative variables, i.e., we require that $\mathbf{r}^+ \geq \mathbf{0}$ and $\mathbf{r}^- \geq \mathbf{0}$, whereas the parameters $\beta \in \mathbb{R}^p$ are "free" to assume positive, zero, or negative values. The objective function of (2.2.2) "hides" the nondifferentiability of the absolute value objective function in a clever way, but captures the objective function of ℓ_1-regression correctly, since we are minimizing. Moreover, from the mathematical properties of linear programming solution procedures, it follows readily that in any solution inspected by, e.g., the simplex algorithm, either $r_i^+ > 0$ or $r_i^- > 0$, but not both, thus giving $|r_i|$ in the objective function depending on whether $r_i > 0$ or $r_i < 0$ for any $i \in N$, where $N = \{1, \ldots, n\}$.

We denote by $\mathbf{P}$ the polyhedron associated with our linear program

$$\mathbf{P} = \left\{(\beta, \mathbf{r}^+, \mathbf{r}^-) \in \mathbb{R}^{p+2n} : \mathbf{X}\beta + \mathbf{r}^+ - \mathbf{r}^- = \mathbf{y}, \mathbf{r}^+ \geq \mathbf{0}, \mathbf{r}^- \geq \mathbf{0}\right\}, \tag{2.2.3}$$

and let $\mathbf{z} = (\beta, \mathbf{r}^+, \mathbf{r}^-)$ for short. As in least squares regression, we will make the blanket assumption that $r(\mathbf{X}) = p$. Consequently, the rank of the constraint matrix defining $\mathbf{P}$ equals $2n + p$ and $\mathbf{P}$ is a nonempty *pointed* polyhedron of dimension $n + p$, i.e., $\dim \mathbf{P} = n + p$. (For all undefined polyhedral terms, see, e.g., [13, Chapter 7].) From the least squares estimates β^{LS} defined in (2.1.3) and their residuals $\mathbf{r}^{\mathrm{LS}} = \mathbf{y} - \mathbf{X}\beta^{\mathrm{LS}}$, we find

$$\mathbf{z}^{\mathrm{LS}} = \left(\beta^{\mathrm{LS}}, \max\left\{\mathbf{0}, \mathbf{r}^{\mathrm{LS}}\right\}, -\min\left\{\mathbf{0}, \mathbf{r}^{\mathrm{LS}}\right\}\right) \in \mathbf{P}. \tag{2.2.4}$$

The face F_{LS} of *smallest* dimension of $\mathbf{P}$ containing $\mathbf{z}^{\mathrm{LS}}$ satisfies $\dim F_{\mathrm{LS}} = p - r(\mathbf{X}_Z)$, where $(\mathbf{X}_Z, \mathbf{y}^Z)$ is the largest submatrix of $(\mathbf{X}, \mathbf{y})$ such that $\mathbf{X}_Z\beta^{\mathrm{LS}} = \mathbf{y}^Z$, i.e., such that the corresponding least squares residuals are zero, and $r(\mathbf{X}_Z) = 0$ if $\mathbf{X}_Z = \emptyset$. This follows because the equation system satisfied by $\mathbf{z}^{\mathrm{LS}}$ has a rank of $2n + r(\mathbf{X}_Z)$. Consequently, $\mathbf{z}^{\mathrm{LS}}$ typically lies on some low-dimensional face of $\mathbf{P}$, but it is, in general, neither an extreme point of $\mathbf{P}$ nor an optimal solution to (2.2.2). To characterize optimality of $\boldsymbol{\beta}^{\mathrm{LS}}$ for ℓ_1-regression let

$$Z = \left\{i \in N : r_i^{\mathrm{LS}} = 0\right\}, \qquad U = \left\{i \in N : r_i^{\mathrm{LS}} > 0\right\}, \qquad L = \left\{i \in N : r_i^{\mathrm{LS}} < 0\right\}. \tag{2.2.5}$$

Furthermore, $\mathbf{X}_Z = (\mathbf{x}^i)_{i \in Z}$, $\mathbf{e}_Z = (1, \dots, 1)^{\mathrm{T}}$ with $|Z|$ components equal to one and $\mathbf{X}_U$, $\mathbf{e}_U$, $\mathbf{X}_L$, and $\mathbf{e}_L$ are defined likewise. Whenever we write $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1$, it is understood that the minimization is over all $\boldsymbol{\beta} \in \mathbb{R}^p$.

PROPOSITION 1.

(i) *The least squares estimate* $\boldsymbol{\beta}^{\mathrm{LS}}$ *is an optimal solution to* $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1$ *if and only if there exists* $\mathbf{v} \in \mathbb{R}^{|Z|}$ *such that*

$$\mathbf{v}\mathbf{X}_Z = -\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L, \qquad -\mathbf{e}_Z^{\mathrm{T}} \le \mathbf{v} \le \mathbf{e}_Z^{\mathrm{T}}. \tag{2.2.6}$$

If $Z = \emptyset$, *condition (2.2.6) simplifies to* $\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U = \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L$.

(ii) $\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 \ge \|\mathbf{r}^{\mathrm{LS}}\|_2^2 / \|\mathbf{r}^{\mathrm{LS}}\|_\infty$, *where* $\|\mathbf{x}\|_\infty = \max\{|x_j| : j \in N\}$ *is the* ℓ_∞*-norm.*

PROOF OF PROPOSITION 1(i). The dual linear program to (2.2.2) is given by

$$\max\left\{\mathbf{u}\mathbf{y} : \mathbf{u}\mathbf{X} = \mathbf{0}, -\mathbf{e}_n^{\mathrm{T}} \le \mathbf{u} \le \mathbf{e}_n^{\mathrm{T}}\right\} = \max\left\{\mathbf{u}\mathbf{r}^{\mathrm{LS}} : \mathbf{u}\mathbf{X} = \mathbf{0}, \ -\mathbf{e}_n^{\mathrm{T}} \le \mathbf{u} \le \mathbf{e}_n^{\mathrm{T}}\right\},$$

where the asserted equality follows because $\mathbf{u}\mathbf{y} = \mathbf{u}(\mathbf{X}\boldsymbol{\beta}^{\mathrm{LS}} + \mathbf{r}^{\mathrm{LS}}) = \mathbf{u}\mathbf{r}^{\mathrm{LS}}$ for all $\mathbf{u} \in \mathbb{R}^n$ satisfying $\mathbf{u}\mathbf{X} = \mathbf{0}$. Suppose now that condition (2.2.6) is satisfied. Define $\mathrm{u}_i = 1$ for $i \in U$, $\mathrm{u}_i = -1$ for $i \in L$, and $\mathbf{u}_Z = \mathbf{v}$. Then $\mathbf{u}$ is a feasible solution to the dual, $\mathbf{u}\mathbf{r}^{\mathrm{LS}} = \mathbf{e}_U^{\mathrm{T}}\mathbf{r}_U^{\mathrm{LS}} - \mathbf{e}_L^{\mathrm{T}}\mathbf{r}_L^{\mathrm{LS}} = \|\mathbf{r}^{\mathrm{LS}}\|_1$, and thus, $\boldsymbol{\beta}^{\mathrm{LS}}$ is an optimal solution to the ℓ_1-regression problem by the weak theorem of duality of linear programming. Suppose, on the other hand, that $\boldsymbol{\beta}^{\mathrm{LS}}$ is an optimal solution to the ℓ_1-regression problem, but that $\mathbf{v} \in \mathbb{R}^{|Z|}$ satisfying (2.2.6) does not exist. By Farkas' lemma, it follows (see, e.g., [13, Exercise 6.5]) that there exist $\boldsymbol{\xi} \in \mathbb{R}^p$, $\boldsymbol{\eta}^+$, $\boldsymbol{\eta}^- \in \mathbb{R}^{|Z|}$ such that

$$\begin{aligned} \mathbf{X}_Z\boldsymbol{\xi} + \boldsymbol{\eta}^+ - \boldsymbol{\eta}^- &= \mathbf{0}, & \left(-\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L\right)\boldsymbol{\xi} + \mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^+ + \mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^- &< 0, \\ \boldsymbol{\eta}^+ &\ge \mathbf{0}, & \boldsymbol{\eta}^- &\ge \mathbf{0}. \end{aligned}$$

If $\mathbf{X}_Z$ is empty, then $\mathbf{0} \ne -\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L$ and we choose any $\boldsymbol{\xi} \in \mathbb{R}^p$ such that $(-\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L)\boldsymbol{\xi} < 0$. Since $\mathbf{r}_U^{\mathrm{LS}} > \mathbf{0}$ and $\mathbf{r}_L^{\mathrm{LS}} < \mathbf{0}$, there exists $\lambda > 0$ such that $\mathbf{r}_U^+(\lambda) = \mathbf{r}_U^{\mathrm{LS}} - \lambda\mathbf{X}_U\boldsymbol{\xi} \ge \mathbf{0}$ and $\mathbf{r}_L^-(\lambda) = -\mathbf{r}_L^{\mathrm{LS}} + \lambda\mathbf{X}_L\boldsymbol{\xi} \ge \mathbf{0}$. Consequently, $\boldsymbol{\beta}(\lambda) = \boldsymbol{\beta}^{\mathrm{LS}} + \lambda\boldsymbol{\xi}$, $\mathbf{r}_Z^+(\lambda) = \lambda\boldsymbol{\eta}^+$, $\mathbf{r}_Z^-(\lambda) = \lambda\boldsymbol{\eta}^-$, $\mathbf{r}_U^-(\lambda) = \mathbf{0}$, and $\mathbf{r}_L^+(\lambda) = \mathbf{0}$, together with $\mathbf{r}_U^+(\lambda)$ and $\mathbf{r}_L^-(\lambda)$ define a feasible solution to the linear program (2.2.2). Calculating its objective function, we get

$$\begin{aligned} \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^+(\lambda) + \mathbf{e}_n^{\mathrm{T}}\mathbf{r}^-(\lambda) &= \lambda\left(\mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^+ + \mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^-\right) + \left\|\mathbf{r}_U^+(\lambda)\right\|_1 + \left\|\mathbf{r}_L^-(\lambda)\right\|_1 \\ &= \left\|\mathbf{r}^{\mathrm{LS}}\right\|_1 + \lambda\left(\mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^+ + \mathbf{e}_Z^{\mathrm{T}}\boldsymbol{\eta}^- - \mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U\boldsymbol{\xi} + \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L\boldsymbol{\xi}\right) < \left\|\mathbf{r}^{\mathrm{LS}}\right\|_1, \end{aligned}$$

and consequently, $\boldsymbol{\beta}^{\mathrm{LS}}$ is not optimal.

PROOF OF PROPOSITION 1(ii). If $\|\mathbf{r}^{\mathrm{LS}}\|_\infty = 0$, then the ratio is defined to be zero and the inequality holds. Otherwise, $\mathbf{u}^{\mathrm{T}} = \mathbf{r}^{\mathrm{LS}}/\|\mathbf{r}^{\mathrm{LS}}\|_\infty$ is a feasible solution to the dual of (2.2.2), and hence, by linear programming duality,

$$\min \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 = \max\left\{\mathbf{u}\mathbf{r}^{\mathrm{LS}} : \mathbf{u}\mathbf{X} = \mathbf{0}, -\mathbf{e}_n^{\mathrm{T}} \le \mathbf{u} \le \mathbf{e}_n^{\mathrm{T}}\right\} \ge \frac{\|\mathbf{r}^{\mathrm{LS}}\|_2^2}{\|\mathbf{r}^{\mathrm{LS}}\|_\infty}. \qquad \blacksquare$$

The proposition shows that even if the least squares solution $\mathbf{z}^{\mathrm{LS}}$ defines an extreme point of $\mathbf{P}$, then $\boldsymbol{\beta}^{\mathrm{LS}}$ is in general *not* an optimal solution to the ℓ_1-regression problem as condition (2.2.6) may be violated. Since by assumption, $r(\mathbf{X}) = p$, it follows by a standard argument from linear programming that $\mathbf{z}^{\mathrm{LS}}$ is an extreme point of $\mathbf{P}$ if and only if $r(\mathbf{X}_Z) = p$. Indeed, condition (2.2.6), together with such a rank condition, characterizes optimal extreme points of $\mathbf{P}$ completely, since the extreme points of a pointed polyhedron are precisely its faces of dimension zero.

PROPOSITION 2. *Let $\boldsymbol{\beta} \in \mathbb{R}^p$, $\mathbf{r}^{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, $\mathbf{z}^{\boldsymbol{\beta}}$, and Z, U, L be defined as in (2.2.4) and (2.2.5) with $\boldsymbol{\beta}^{\mathrm{LS}}$ replaced by $\boldsymbol{\beta}$. Then $\mathbf{z}^{\boldsymbol{\beta}}$ is an optimal extreme point of P if and only if $\mathrm{r}(\mathbf{X}_Z) = p$ and (2.2.6) is satisfied.*

Thus, for every optimal extreme point solution $\boldsymbol{\beta}^* \in \mathbb{R}^p$ of the ℓ_1-regression problem, there exists a nonsingular $p \times p$ submatrix $\mathbf{X}_B$ of $\mathbf{X}_Z$ such that

$$\boldsymbol{\beta}^* = \mathbf{X}_B^{-1}\mathbf{y}^B, \qquad \mathbf{r}^+ = \max\{\mathbf{0}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\}, \qquad \mathbf{r}^- = -\min\{\mathbf{0}, \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\}, \tag{2.2.7}$$

defines an extreme point $\mathbf{z}^* = (\boldsymbol{\beta}^*, \mathbf{r}^+, \mathbf{r}^-)$ of $\mathbf{P}$ satisfying condition (2.2.6), where $B \subseteq Z$, $|B| = p$ and $\mathbf{y}^B$ is the subvector of $\mathbf{y}$ corresponding to the rows of $\mathbf{X}_B$. Since the parameters $\boldsymbol{\beta} \in \mathbb{R}^p$ are *free* variables of the linear program (2.2.2), every (decent) commercial LP solver will automatically put all $\boldsymbol{\beta}$ variables into the LP *basis* (provided that $\mathrm{r}(\mathbf{X}) = p$ as assumed!) and produce an optimal solution of the form (2.2.7). An optimal solution to (2.2.2) need, of course, not be *unique*. The following proposition (and its proof) is adapted from Koenker and Bassett [14, Theorem 3.3] and gives a necessary and sufficient condition for the ℓ_1-regression problem to have a unique solution.

PROPOSITION 3. *$\boldsymbol{\beta}^* \in \mathbb{R}^p$ as defined in (2.2.7) uniquely solves the ℓ_1-regression problem if and only if*

$$-\mathbf{e}_p^{\mathrm{T}} - \sum_{i\in D} \left|\mathbf{x}^i\mathbf{X}_B^{-1}\right| < \left(\mathbf{e}_U^{\mathrm{T}}\mathbf{X}_U - \mathbf{e}_L^{\mathrm{T}}\mathbf{X}_L\right)\mathbf{X}_B^{-1} < \mathbf{e}_p^{\mathrm{T}} + \sum_{i\in D} \left|\mathbf{x}^i\mathbf{X}_B^{-1}\right|, \tag{2.2.8}$$

where $D = Z - B$ and $|\mathbf{x}| = (|x_1|, \ldots, |x_p|)^{\mathrm{T}}$, for any $\mathbf{x} \in \mathbb{R}^p$.

PROOF. $\boldsymbol{\beta}^*$ is unique if and only if $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*\|_1 < \|\mathbf{y} - \mathbf{X}(\boldsymbol{\beta}^* + \mathbf{w})\|_1$, for all $\mathbf{w} \in \mathbb{R}^p$ with $\mathbf{w} \neq \mathbf{0}$. Since by the triangle inequality, $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1$ is a convex function of $\boldsymbol{\beta}$, it is necessary and sufficient that the inequality holds for all "sufficiently small" nonzero $\mathbf{w} \in \mathbb{R}^p$. Because $\mathbf{X}_B$ is nonsingular, setting $\mathbf{v} = \mathbf{X}_B\mathbf{w}$, the inequality holds if and only if

$$\sum_{i\in R} |r_i^*| < \sum_{i\in B} |v_i| + \sum_{i\in R} \left|r_i^* - \mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v}\right|,$$

for all $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = \varepsilon$ and sufficiently small $\varepsilon > 0$, where $\mathbf{r}^* = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^*$ and $R = D \cup U \cup L$. Define

$$\operatorname{sign}(u; z) = \begin{cases} \operatorname{sign}(u), & \text{if } u \neq 0, \\ \operatorname{sign}(z), & \text{if } u = 0. \end{cases}$$

For small enough $\varepsilon > 0$, we get

$$\left|r_i^* - \mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v}\right| = \operatorname{sign}\left(r_i^*; -\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v}\right)\left(r_i^* - \mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v}\right)$$

and the last inequality simplifies to

$$\sum_{i\in R} \operatorname{sign}\left(r_i^*; -\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v}\right)\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{v} < \sum_{i\in B} |v_i|,$$

for all $\mathbf{v} \in \mathbb{R}^p$ with $\|\mathbf{v}\|_2 = \varepsilon$, which is equivalent to

$$\sum_{i\in D} \operatorname{sign}\left(-\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k\right)\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k + \sum_{i\in U\cup L} \operatorname{sign}(r_i^*)\,\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k < 1,$$

$$\sum_{i\in D} \operatorname{sign}\left(\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k\right)\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k + \sum_{i\in U\cup L} \operatorname{sign}(r_i^*)\,\mathbf{x}^i\mathbf{X}_B^{-1}\mathbf{u}_k > -1,$$

where $\mathbf{u}_k \in \mathbb{R}^p$ is the k^{th} unit vector and $k = 1, \ldots, p$. Thus, (2.2.8) follows. ▮

As in linear programming calculations, alternative optima are the rule rather than the exception. It becomes necessary to find a "compromise" between competing optimal extreme-point solutions to the ℓ_1-regression problem. So let $\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^q$ be *all* $q \geq 1$ optimal extreme point solutions to (2.2.2). Then we choose as the ℓ_1-regression estimate $\boldsymbol{\beta}^{L_1} \in \mathbb{R}^p$, the *center of gravity* of the extremal solutions,

$$\boldsymbol{\beta}^{L_1} = \frac{1}{q} \sum_{\ell=1}^{q} \boldsymbol{\beta}^{\ell}, \tag{2.2.9}$$

thereby getting a unique ℓ_1-regression estimator for any data. In practice, one will typically content oneself with some optimal extreme-point solution to the ℓ_1-regression problem, but rendering the ℓ_1-regression estimator unique like in (2.2.9) has some consequences for the associated statistical model.

Different from the least squares analysis and manifestly due to the lack of a convenient, mathematically tractable closed form solution for the ℓ_1-regression estimator, the statistical theory developed for ℓ_1-regression is less advanced than it is for ℓ_2-regression. Whereas for the ℓ_2-regression estimates $\boldsymbol{\beta}^{\mathrm{LS}}$, one has the formula $\boldsymbol{\beta}^{\mathrm{LS}} = \boldsymbol{\beta} + (\mathbf{X}^{\top}\mathbf{X})^{-1}\mathbf{X}^{\top}\boldsymbol{\varepsilon}$, which permits one to determine the sample distribution of $\boldsymbol{\beta}^{\mathrm{LS}}$ from the error distribution, the dependence of the ℓ_1-regression estimator on the errors is more complicated. Assuming, for simplicity, the uniqueness of $\boldsymbol{\beta}^{L_1}$, we get $\boldsymbol{\beta}^{L_1} = \boldsymbol{\beta} + \mathbf{X}_B^{-1}\boldsymbol{\varepsilon}_B$, and thus, different from $\boldsymbol{\beta}^{\mathrm{LS}}$, where *all* error terms $\boldsymbol{\varepsilon}$ enter at once. Here the dependence of $\boldsymbol{\beta}^{L_1}$ is "local" on only a subset $\boldsymbol{\varepsilon}_B$ of the error terms. But to obtain the sample distribution of the ℓ_1-regression estimates, we must vary over all possible values of the observations (and thus, all errors), and hence, the dependence of $\boldsymbol{\beta}^{L_1}$ on $\mathbf{X}_B$ (which varies as well!) may make a precise determination of the distribution of $\boldsymbol{\beta}^{L_1}$ computationally intractable; see also [15,16]. However, it has been proven that *asymptotically* the ℓ_1-regression estimator follows a normal distribution.

For reasons that may be debatable, statisticians like "unbiased" estimators, i.e., they like to have a relation of the form $E(\boldsymbol{\beta}^{L_1}) = \boldsymbol{\beta}$, which essentially says that "on average" the ℓ_1-regression estimates $\boldsymbol{\beta}^{L_1}$ estimate the "true" underlying parameter $\boldsymbol{\beta}$ of the statistical linear regression model (2.0.2) correctly no matter what $\boldsymbol{\beta} \in \mathbb{R}^p$ may be. It is a fact that the restriction to optimal *extreme point* solutions $\boldsymbol{\beta}^*$ of the ℓ_1-regression problem may indeed produce "biased" estimators, see [17] for a pertaining small example with a *discrete* probability distribution of the error terms. However, defining the ℓ_1-regression estimator as in (2.2.9) gets one around this difficulty. Assuming that the error terms $\boldsymbol{\varepsilon}$ are *symmetrically distributed* random variables with mean zero, one shows the unbiasedness of $\boldsymbol{\beta}^{L_1}$ as follows. Let $\boldsymbol{\beta}^{L_1}(\boldsymbol{\varepsilon}) = (1/q)\sum_{\ell=1}^{q} \boldsymbol{\beta}^{\ell}(\boldsymbol{\varepsilon})$ with $q \geq 1$ as defined in (2.2.9). It follows that

$$\boldsymbol{\beta}^{L_1}(\boldsymbol{\varepsilon}) = \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\mathbf{y}^{B_\ell} = \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\left(\mathbf{X}_{B_\ell}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_{B_\ell}\right) = \boldsymbol{\beta} + \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\boldsymbol{\varepsilon}_{B_\ell}.$$

For $\bar{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} - \boldsymbol{\varepsilon}$, we get from the symmetry of the optimality condition

$$\boldsymbol{\beta}^{L_1}(-\boldsymbol{\varepsilon}) = \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\bar{\mathbf{y}}^{B_\ell} = \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\left(\mathbf{X}_{B_\ell}\boldsymbol{\beta} - \boldsymbol{\varepsilon}_{B_\ell}\right) = \boldsymbol{\beta} - \frac{1}{q}\sum_{\ell=1}^{q} \mathbf{X}_{B_\ell}^{-1}\boldsymbol{\varepsilon}_{B_\ell},$$

i.e., more precisely, that for $\bar{\mathbf{y}}$, we have precisely q optimal extreme point solutions $\bar{\boldsymbol{\beta}}^{\ell} = \boldsymbol{\beta}^{\ell}(-\boldsymbol{\varepsilon})$ defined by $B_1, \ldots, B_q$ as well. Consequently, $\boldsymbol{\beta} - \boldsymbol{\beta}^{L_1}(\boldsymbol{\varepsilon}) = -[\boldsymbol{\beta} - \boldsymbol{\beta}^{L_1}(-\boldsymbol{\varepsilon})]$, and thus, from the symmetry of the error distribution, we have $E(\boldsymbol{\beta}^{L_1}) = \boldsymbol{\beta}$ as argued in [17]. This is dependent upon whether the underlying error distribution is discrete or continuous. Moreover, in the case of a continuous distribution, one can rule out nonuniqueness of $\boldsymbol{\beta}^{L_1}$, and thus, biasedness as well, because they are events of probability measure zero in this case. Thus, the ℓ_1-regression estimator is unbiased ("for what it is worth" in the words of Sielken and Hartley [17, p. 641]).

The real question of the statistical analysis rests with whether or not *practical* tests of the significance of and confidence intervals for the ℓ_1-regression estimator can be found. As we have mentioned above, except perhaps for very small sample sizes, it seems next to impossible to determine the sample distribution of β^{L_1} exactly. Thus, the question arises whether or not these distributions can be found *approximately*. This is indeed the case, as was shown by way of a Monte Carlo simulation by Rosenberg and Carlson [15] and made analytically precise via an asymptotic distributional result by Bassett and Koenker [18], see also [19]. More precisely, Bassett and Koenker consider a sequence of linear models of the form $\mathbf{y}_n = \mathbf{X}_n\beta + \varepsilon_n$, where $\mathbf{y}_n \in \mathbb{R}^n$ and $\mathbf{X}_n$ is $n \times p$ with $\mathrm{r}(\mathbf{X}_n) = p < n$. The error terms are assumed to be i.i.d. random variables with a marginal distribution function F having a median of zero. The latter assumption can always be met by including an intercept term in the design matrix.

THEOREM 1. *(See [18].) Let $\{\beta_n^*\}$ denote a sequence of unique solutions to* $\min\|\mathbf{y}_n - \mathbf{X}_n\beta\|_1$ *and assume the following.*

(i) *F is continuous and has a continuous and positive density f at the median.*

(ii) *$\lim_{n\to+\infty}(1/n)\mathbf{X}_n^\top\mathbf{X}_n = \mathbf{Q}$ is a positive definite matrix.*

Then $\sqrt{n}(\beta_n^* - \beta)$ converges in distribution to a p-dimensional Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\omega^2\mathbf{Q}^{-1}$, where $\omega^2 = (2f(0))^{-2}$ is the asymptotic variance of the sample median of random samples from F.

In contrast to the corresponding asymptotic theory for ℓ_2-regression, see, e.g., [20, p. 398], the existence of the second moment of F is (as well as further technical conditions are) not required. Moreover, it follows from the theorem that the asymptotic confidence ellipsoids of the ℓ_1-regression estimators are strictly smaller than those for the least squares estimators for *all* distributions F, for which the sample median is a more efficient estimator of location than the sample mean (such as, e.g., the double exponential (Laplace) distribution or the Cauchy distribution). It follows, furthermore, that the ℓ_1-regression estimates are asymptotically consistent estimators of the parameters of the linear regression model. Finally, based on the theorem of Bassett and Koenker, asymptotic tests and confidence intervals for ℓ_1-regression estimators using the χ^2 distribution and the standardized normal distribution have been developed by Koenker and Bassett [21], Dielman and Pfaffenberger [22], and others. In other words, as in ℓ_2-regression analysis, an entire apparatus to judge the quality of fit of the estimates of the linear model (2.0.2) is available to the practitioner of large-scale ℓ_1-regression.

2.3. $\mathbf{L}_\infty$ or Chebychev Regression

According to Harter [1, p. 149], the idea of minimizing the *maximum residual error* in "solving" inconsistent (linear) systems of equations goes back to Leonhard Euler (1707–1783). Jean-Victor Poncelet (1788–1867) used this criterion in his work on approximating certain nonlinear functions over some finite interval by linear ones, which is what Pafnuty Lvovich Chebychev (1821–1894) generalized by considering polynomials in the approximation process. The criterion of minimizing the maximum residual remains a major criterion today in *approximation theory*, see, e.g., [23]. For the linear regression model (2.0.2) or (2.0.3), this criterion means that parameters $\beta \in \mathbb{R}^p$ are sought such that

$$\|\mathbf{y} - \mathbf{X}\beta\|_\infty = \max\left\{\left|y_i - \mathbf{x}^i\beta\right| : 1 \le i \le n\right\} \tag{2.3.1}$$

is minimum, i.e., the ℓ_∞-norm $\|\mathbf{x}\|_\infty = \max\{|x_i| : i \in N = \{1, \ldots, n\}\}$ on $\mathbb{R}^n$ is used instead of the ℓ_1- or ℓ_2-norm in the process of minimizing the dispersion of the residuals. As in the case of the ℓ_1-norm, this objective function is nondifferentiable and—except in very special cases, see, e.g., Proposition 4(iv) below—a general closed form solution to the problem simply is not known to exist. As a result, there is little treatment of ℓ_∞-regression in the statistical literature, even though the method of ℓ_∞-regression is recommended whenever the *sample midrange* is a more efficient estimator of the location or "centrality" parameter of the error distribution than either the sample mean or the sample median. This is the case, e.g., if the errors follow a uniform distribution, see, e.g., [5].

The key to the solution of the ℓ_∞-regression problem is the fact—known probably to Euler and other mathematicians of the 19th century—that the minimization problem (2.3.1) is a linear program of the form

$$\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & \mathbf{X}\beta + \gamma \mathbf{e}_n \geq \mathbf{y}, \\ & -\mathbf{X}\beta + \gamma \mathbf{e}_n \geq -\mathbf{y}, \\ & \beta \text{ free}, \ \gamma \text{ free}. \end{array} \tag{2.3.2}$$

Other formulations of the ℓ_∞-regression problem as a linear program are possible and can be found in the literature. In our analysis of the ℓ_∞-regression problem (2.3.2), we will again make the blanket assumption that $r(\mathbf{X}) = p$ and an intercept term may or may not be present in the design matrix. Moreover, we will make the assumption that $\beta \in \mathbb{R}^p$ such that $\mathbf{X}\beta = \mathbf{y}$ does not exist, i.e., we rule out the possibility of a "perfect fit", for convenience in the following analysis.

We denote by $\mathbf{Q}$ the polyhedron associated with our linear program

$$\mathbf{Q} = \left\{(\beta, \gamma) \in \mathbb{R}^{p+1} : \mathbf{X}\beta + \gamma \mathbf{e}_n \geq \mathbf{y}, -\mathbf{X}\beta + \gamma \mathbf{e}_n \geq -\mathbf{y}\right\}. \tag{2.3.3}$$

From our blanket assumption, it follows immediately that the rank of the constraint matrix of (2.3.2) equals $p+1$. Since $\mathbf{Q}$ is evidently nonempty, it is a pointed polyhedron in $\mathbb{R}^{p+1}$, i.e., $\mathbf{Q}$ has extreme points, and moreover, $\dim \mathbf{Q} = p+1$. For any $\beta \in \mathbb{R}^p$, we define

$$\gamma^\beta = \|\mathbf{y} - \mathbf{X}\beta\|_\infty, \qquad A = \left\{i \in N : \left|r_i^\beta\right| = \gamma^\beta\right\}, \tag{2.3.4}$$

where $\mathbf{r}^\beta = \mathbf{y} - \mathbf{X}\beta$ are the corresponding residuals, we define $\mathbf{f} \in \mathbb{R}^n$ by

$$f_i = 1, \quad \text{if } r_i^\beta \geq 0, \qquad f_i = -1, \quad \text{if } r_i^\beta < 0, \tag{2.3.5}$$

and let $\mathbf{X}_A = (\mathbf{x}^i)_{i \in A}$ and $\mathbf{f}^A = (f_i)_{i \in A}$.

PROPOSITION 4.

(i) $(\beta, \gamma^\beta) \in \mathbb{R}^{p+1}$ *is an extreme point of* Q *if and only if* $r(\mathbf{X}_A \mathbf{f}^A) = p+1$.

(ii) $(\beta, \gamma^\beta) \in \mathbb{R}^{p+1}$ *is an optimal solution to (2.3.2) if and only if there exists* $\mathbf{w} \in \mathbb{R}^{|A|}$ *such that*

$$\mathbf{w}\mathbf{X}_A = \mathbf{0}, \qquad \mathbf{w}\mathbf{f}^A = 1. \tag{2.3.6}$$

(iii) $\min \|\mathbf{y} - \mathbf{X}\beta\|_\infty \geq \|\mathbf{r}^{\mathrm{LS}}\|_2^2 / \|\mathbf{r}^{\mathrm{LS}}\|_1$, *where* $\mathbf{r}^{\mathrm{LS}} = \mathbf{y} - \mathbf{X}\beta^{\mathrm{LS}}$ *are the least squares residuals and* β^{LS} *are the least squares estimates defined in (2.1.3).*

(iv) *If* $n = p+1$, *then an optimal solution to (2.3.2) is* (β^c, γ^c), *where*

$$\beta^c = \beta^{\mathrm{LS}} - \gamma^c \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{f}^{\mathrm{LS}}, \qquad \gamma^c = \frac{\left\|\mathbf{r}^{\mathrm{LS}}\right\|_2^2}{\left\|\mathbf{r}^{\mathrm{LS}}\right\|_1}, \tag{2.3.7}$$

β^{LS} *is the least squares solution (2.1.3) and* $\mathbf{f}^{\mathrm{LS}}$ *is defined as in (2.3.5) for* $\beta = \beta^{\mathrm{LS}}$.

PROOF OF PROPOSITION 4(i). If $r(\mathbf{X}_A \mathbf{f}^A) = p+1$, then (β, γ^β) lies on a face of dimension zero of $\mathbf{Q}$, and thus, it is an extreme point of $\mathbf{Q}$. Suppose now that (β, γ^β) is an extreme point of $\mathbf{Q}$, but that $r(\mathbf{X}_A \mathbf{f}^A) < p+1$. Consequently, there exists $(\boldsymbol{\lambda}, \lambda_0) \in \mathbb{R}^{p+1}$, $(\boldsymbol{\lambda}, \lambda_0) \neq \mathbf{0}$, such that $\mathbf{X}_A \boldsymbol{\lambda} + \lambda_0 \mathbf{f}^A = \mathbf{0}$. But then $(\beta + \varepsilon\boldsymbol{\lambda}, \gamma^\beta + \varepsilon\lambda_0) \in \mathbf{Q}$ and $(\beta - \varepsilon\boldsymbol{\lambda}, \gamma^\beta - \varepsilon\lambda_0) \in \mathbf{Q}$ for some $\varepsilon > 0$ as is readily checked. Thus,

$$(\beta, \gamma^\beta) = \frac{1}{2}\left(\beta + \varepsilon\boldsymbol{\lambda}, \gamma^\beta + \varepsilon\lambda_0\right) + \frac{1}{2}\left(\beta - \varepsilon\boldsymbol{\lambda}, \gamma^\beta - \varepsilon\lambda_0\right),$$

and hence, (β, γ^β) is not an extreme point of $\mathbf{Q}$ since $(\boldsymbol{\lambda}, \lambda_0) \neq \mathbf{0}$.

Proof of Proposition 4(ii). The dual linear program to (2.3.2) is given by

$$\max\{\mathbf{u}\mathbf{y} - \mathbf{v}\mathbf{y} : \mathbf{u}\mathbf{X} - \mathbf{v}\mathbf{X} = \mathbf{0},\ \mathbf{u}\mathbf{e}_n + \mathbf{v}\mathbf{e}_n = 1,\ \mathbf{u} \geq \mathbf{0},\ \mathbf{v} \geq \mathbf{0}\}. \tag{2.3.8}$$

Suppose (2.3.6) is satisfied. Then setting

$$\mathbf{u}_A = \max\{\mathbf{0}, \mathbf{w}\}, \quad \mathbf{u}_{N-A} = \mathbf{0}, \qquad \mathbf{v}_A = \{\mathbf{0}, -\mathbf{w}\}, \quad \mathbf{v}_{N-A} = \mathbf{0},$$

we have a feasible solution to (2.3.8). Moreover, $(\mathbf{u} - \mathbf{v})\mathbf{y} = (\mathbf{u} - \mathbf{v})\mathbf{r}^{\boldsymbol{\beta}}$ for all $(\mathbf{u}, \mathbf{v})$ satisfying $\mathbf{u}\mathbf{X} - \mathbf{v}\mathbf{X} = \mathbf{0}$. From the definition of A, $(\mathbf{u}_A - \mathbf{v}_A)\mathbf{r}_A^{\boldsymbol{\beta}} = \gamma^{\boldsymbol{\beta}}$, and thus, the assertion follows from the weak duality theorem of linear programming. Suppose now that $(\boldsymbol{\beta}, \gamma^{\boldsymbol{\beta}})$ is optimal, but that (2.3.6) has no solution. It follows from Farkas' lemma that there exist ξ_0 and $\boldsymbol{\xi} \in \mathbb{R}^p$ such that $\mathbf{X}_A\boldsymbol{\xi} + \xi_0\mathbf{f}^A = \mathbf{0}$ and $\xi_0 < 0$. Consequently, $(\boldsymbol{\beta} + \varepsilon\boldsymbol{\xi}, \gamma^{\boldsymbol{\beta}} + \varepsilon\xi_0) \in \mathbf{Q}$ for some $\varepsilon > 0$ as is readily verified. But $\gamma^{\boldsymbol{\beta}} + \varepsilon\xi_0 < \gamma^{\boldsymbol{\beta}}$, and thus, $(\boldsymbol{\beta}, \gamma^{\boldsymbol{\beta}})$ is not optimal.

Proof of Proposition 4(iii). As we ruled out a perfect fit, $\mathbf{r}^{\mathrm{LS}} \neq \mathbf{0}$, and thus, $\mathbf{u} = \alpha\max\{\mathbf{0}, \mathbf{r}^{\mathrm{LS}}\}$, $\mathbf{v} = \alpha\max\{\mathbf{0}, -\mathbf{r}^{\mathrm{LS}}\}$, where $\alpha^{-1} = \|\mathbf{r}^{\mathrm{LS}}\|_1$ defines a feasible solution to the dual (2.3.8) of (2.3.2) with an objective function value of$\alpha\|\mathbf{r}^{\mathrm{LS}}\|_2^2$. So the assertion follows from the duality theorem of linear programming.

Proof of Proposition 4(iv). Define $\mathbf{F} = \operatorname{diag}(f_1^{\mathrm{LS}}, \dots, f_n^{\mathrm{LS}})$ and note that by the definition of $\mathbf{f}^{\mathrm{LS}}$, we have $\mathbf{F}\mathbf{F} = \mathbf{I}_n$. The constraint set of (2.3.2) is equivalent to $\mathbf{F}\mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{e}_n \geq \mathbf{F}\mathbf{y}$, $-\mathbf{F}\mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{e}_n \geq -\mathbf{F}\mathbf{y}$, where γ and $\boldsymbol{\beta} \in \mathbb{R}^p$ are free variables. Since $n = p + 1$, the matrix $(\mathbf{F}\mathbf{X}\mathbf{e}_n)$ is of size $(p + 1) \times (p + 1)$ and we claim that it is nonsingular. For if it is not, then there exists $\boldsymbol{\lambda} \in \mathbb{R}^{p+1}$ such that $\boldsymbol{\lambda}\mathbf{F}\mathbf{X} = \mathbf{0}$ and $\boldsymbol{\lambda}\mathbf{e}_n = 0$. Letting $\boldsymbol{\mu} = \boldsymbol{\lambda}\mathbf{F}$, we have $\boldsymbol{\mu}\mathbf{X} = \mathbf{0}$ and $\boldsymbol{\mu}\mathbf{f}^{\mathrm{LS}} = 0$. But $\mathrm{r}(\mathbf{X}) = p$ and $(\mathbf{r}^{\mathrm{LS}})^{\mathrm{T}}\mathbf{X} = \mathbf{0}$. Thus, $\boldsymbol{\mu} = \delta(\mathbf{r}^{\mathrm{LS}})^{\mathrm{T}}$, where δ is a scalar because $n = p + 1$. Consequently, from $\boldsymbol{\mu}\mathbf{f}^{\mathrm{LS}} = 0$, we get $\delta\|\mathbf{r}^{\mathrm{LS}}\|_1 = 0$. Since we have ruled out a perfect fit $\|\mathbf{r}^{\mathrm{LS}}\|_1 \neq 0$, consequently, $\delta = 0$, and thus, the claim follows. Hence, $\mathbf{F}\mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{e}_n = \mathbf{F}\mathbf{y}$ has a unique solution. Multiplying by $\mathbf{X}^{\mathrm{T}}\mathbf{F}$, we find $\boldsymbol{\beta} + \gamma(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{f}^{\mathrm{LS}} = \boldsymbol{\beta}^{\mathrm{LS}}$, while multiplication by $(\mathbf{r}^{\mathrm{LS}})^{\mathrm{T}}\mathbf{F}$ yields $\gamma\|\mathbf{r}^{\mathrm{LS}}\|_1 = \|\mathbf{r}^{\mathrm{LS}}\|_2^2$. Consequently, (2.3.7) follows. Moreover, $-\mathbf{F}\mathbf{X}\boldsymbol{\beta}^c + \gamma^c\mathbf{e}_n \geq -\mathbf{F}\mathbf{y}$, because $\gamma^c\mathbf{e}_n \geq -(\mathbf{F}\mathbf{y} - \mathbf{F}\mathbf{X}\boldsymbol{\beta}^c) = -|\mathbf{r}^{\boldsymbol{\beta}}|$, and thus, $(\boldsymbol{\beta}^c, \gamma^c)$ is an extreme point of $\mathbf{Q}$. The optimality of $(\boldsymbol{\beta}^c, \gamma^c)$ follows from Part (ii) because $(\mathbf{X}_A\mathbf{f}^A) = (\mathbf{X}\mathbf{F}\mathbf{e})$, and thus, $\mathbf{w} = (1/\|\mathbf{r}^{\mathrm{LS}}\|_1)(\mathbf{r}^{\mathrm{LS}})^{\mathrm{T}}$ solves (2.3.6). ■

Part (i) of Proposition 4 is from [24], Parts (iii) and (iv) can be found in [25, p. 41]. Formula (2.3.7) is of interest only if ℓ_2-norm calculations are carried out, while Parts (i) and (ii) imply that there exists a $(p + 1) \times (p + 1)$ nonsingular submatrix of the constraint system of the linear program (2.3.2) which defines an optimal extreme point solution to the ℓ_∞-regression problem. Naively, one can thus solve the linear program (2.3.2) by enumerating all $(p+1) \times (p+1)$ submatrices of the constraint set of (2.3.2), checking their nonsingularity and feasibility of the corresponding solution, and picking anyone for which the resulting value of γ is minimal. Evidently, solving (2.3.2) by any commercial LP solver is far more efficient.

For every optimal extreme point solution $\boldsymbol{\beta}^* \in \mathbb{R}^p$ of the ℓ_∞-regression problem, there exists, hence, a nonsingular $(p + 1) \times (p + 1)$ submatrix $(\mathbf{X}_B\mathbf{f}^B)$ of $(\mathbf{X}_A\mathbf{f}^A)$ such that

$$\begin{pmatrix} \boldsymbol{\beta}^* \\ \gamma^* \end{pmatrix} = \left(\mathbf{X}_B\mathbf{f}^B\right)^{-1}\mathbf{y}^B \tag{2.3.9}$$

is an extreme point of $\mathbf{Q}$ satisfying condition (2.3.6), where $B \subseteq A$, $|B| = p + 1$, and $\mathbf{y}^B$ is the subvector of $\mathbf{y}$ corresponding to the rows of $\mathbf{X}_B$. Optimal solutions to the linear program (2.3.2) need, of course, not be unique. The following proposition gives a necessary and sufficient condition for the ℓ_∞-regression problem to have a unique solution.

Proposition 5. *Let $\boldsymbol{\beta}^* \in \mathbb{R}^p$ as defined in (2.3.9) be an optimal solution to (2.3.2). Then $\boldsymbol{\beta}^*$ is unique if and only if*

$$\mathbf{F}_A^*\mathbf{X}_A\boldsymbol{\xi} \geq \mathbf{0}, \qquad \xi_k < 0, \tag{2.3.10}$$

is inconsistent for $k = 1, \dots, p$, *where* $\mathbf{F}^*_A = \operatorname{diag}(f^*_i)_{i \in A}$ *and* A, $\mathbf{f}^*$ *are defined as in (2.3.4), (2.3.5) with* $\boldsymbol{\beta}$ *replaced with* $\boldsymbol{\beta}^*$.

PROOF. $\boldsymbol{\beta}^*$ uniquely solves (2.3.2) if and only if $\min\{\mathbf{c}\boldsymbol{\beta} : (\boldsymbol{\beta}, \gamma^*) \in \mathbf{Q}\} = \mathbf{c}\boldsymbol{\beta}^*$ for all $\mathbf{c} \in \mathbb{R}^p$, i.e., if and only if the optimal face of $\mathbf{Q}$ is exactly the extreme point $(\boldsymbol{\beta}^*, \gamma^*)$ of $\mathbf{Q}$. Consequently, by the strong duality theorem of linear programming,

$$\max\{\mathbf{u}(\mathbf{y} - \gamma^*\mathbf{e}_n) + \mathbf{v}(-\mathbf{y} - \gamma^*\mathbf{e}_n) : \mathbf{u}\mathbf{X} - \mathbf{v}\mathbf{X} = \mathbf{c},\ \mathbf{u} \geq \mathbf{0},\ \mathbf{v} \geq \mathbf{0}\} = \mathbf{c}\boldsymbol{\beta}^*,$$

for all $\mathbf{c} \in \mathbb{R}^p$. Writing $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathbf{r}^*$, $(\boldsymbol{\beta}^*, \gamma^*)$ is thus unique if and only if

$$\max\{\mathbf{u}(\mathbf{r}^* - \gamma^*\mathbf{e}_n) + \mathbf{v}(-\mathbf{r}^* - \gamma^*\mathbf{e}_n) : \mathbf{u}\mathbf{X} - \mathbf{v}\mathbf{X} = \mathbf{c},\ \mathbf{u} \geq \mathbf{0},\ \mathbf{v} \geq \mathbf{0}\} = 0,$$

for all $\mathbf{c} \in \mathbb{R}^p$. Since $\mathbf{r}^*_{N-A} - \gamma^*\mathbf{e}_{N-A} < \mathbf{0}$ and $-\mathbf{r}^*_{N-A} - \gamma^*\mathbf{e}_{N-A} < \mathbf{0}$, it follows that $\mathbf{u}_{N-A} = \mathbf{v}_{N-A} = \mathbf{0}$ in every optimal solution to this linear program. Moreover, for all $i \in A$, if $\mathrm{r}^*_i < 0$, then $u_i = 0$ and if $\mathrm{r}^*_i > 0$, then $\mathrm{v}_i = 0$ in every optimal solution. Thus, $(\boldsymbol{\beta}^*, \gamma^*)$ is unique if and only if

$$\mathbf{u}_A \mathbf{F}^*_A \mathbf{X}_A = \mathbf{c}, \qquad \mathbf{u}_A \geq \mathbf{0},$$

has a solution for every $\mathbf{c} \in \mathbb{R}^p$, or equivalently that

$$\mathbf{u}_A \mathbf{F}^*_A \mathbf{X}_A = \mathbf{u}_k^T, \qquad \mathbf{u}_A \geq \mathbf{0},$$

has a solution for every $k \in \{1, \dots, p\}$, where $\mathbf{u}_k \in \mathbb{R}^p$ is the k^{th} unit vector. Applying Farkas' lemma shows that condition (2.3.10) is necessary and sufficient for the uniqueness of $(\boldsymbol{\beta}^*, \gamma^*)$. ∎

Our condition for uniqueness requires that p systems of inequalities must be checked for inconsistency which is, of course, a laborious computation. The literature on ℓ_∞-regression does not offer—to the best of our knowledge—any condition. Uniqueness of an optimal solution to (2.3.2) is, of course, not to be expected. So as in the case of ℓ_1-regression, we let $\boldsymbol{\beta}^1, \dots, \boldsymbol{\beta}^q$ denote all $q \geq 1$ optimal extreme point solutions to the linear program (2.3.2) and define the ℓ_∞-regression estimate to be

$$\boldsymbol{\beta}^{L_\infty} = \frac{1}{q} \sum_{\ell=1}^{q} \boldsymbol{\beta}^\ell, \tag{2.3.11}$$

thereby getting a unique ℓ_∞-regression estimator for any data. As in the case of ℓ_1-regression, in the practice of ℓ_∞-regression, one will usually content oneself with finding a single optimal extreme point solution to the linear program (2.3.2).

We have been unable to locate any substantial statistical analysis of ℓ_∞-regression in the literature. The only pertaining result (see [17]) concerns the biasedness or unbiasedness of the ℓ_∞-regression estimator. As in the case of the ℓ_1-regression estimator, one establishes the unbiasedness of (2.3.11) along the lines of the arguments employed in Section 2.2. By analogy to the cases of the ℓ_1-norm and ℓ_2-norm, one might think that an asymptotic distributional result for the ℓ_∞-regression estimator similar to the theorem of Bassett and Koenker (see Section 2.2) can be proven. In other words, such a result would be that if the marginal error distribution F is centered such that the midrange of the errors is zero, then the asymptotic distribution of the ℓ_∞-regression estimator is normal as in the Bassett-Koenker theorem with the quantity ω^2 replaced by the *asymptotic variance of the sample midrange* of random samples from F. This is motivated by the well-known fact that the sample midrange is an optimal estimate of centrality under the ℓ_∞-norm, whereas the median is optimal for the ℓ_1-norm and the arithmetic mean is optimal for the ℓ_2-norm. However, the sample midrange does not have nice statistical properties like those of the sample median and the sample mean. More specifically, in the case of univariate location, the ℓ_∞-location estimator (sample midrange) is not even $\sqrt{n}$ consistent, nor is its limiting distribution the normal distribution. Rather, the sample midrange of a standard normal distribution converges at a rate proportional to $1/\log n$ (cf. [26]). Furthermore, the limiting cumulative distribution function (cdf) of the sample midrange from a standard normal distribution is the logistic distribution (cf. [27])

$$F(x) = \frac{1}{1+e^{-x}}. \tag{2.3.12}$$

Therefore, there is little reason to believe that in the case of regression, the ℓ_∞-estimator is $\sqrt{n}$ consistent and/or asymptotically normal.

2.4. Summary

Looking at virtually every textbook in statistics today, one is left with the impression that linear regression consists only of *least squares* regression, i.e., the treatment of the general linear regression model (2.0.2) or (2.0.3) by way of the ℓ_2-norm. Hardly any text that we have consulted deals with ℓ_1-regression where, however, in the past 20 years or so, a solid body of statistical knowledge has been assembled—in response to historical attempts and to needs that are expressed and documented in the econometric literature for well over 30 years now. L_∞-regression—though historically an exciting topic—appears to have been completely neglected by the statistical (textbook) literature. This is all the more astonishing because the computational problems of ℓ_1- and ℓ_∞-regression of yesteryear have long been overcome by the advent of linear programming and easily available commercial software to solve such problems very efficiently indeed. In our survey, we have purposely restricted ourself to a discussion of methods and models for which efficient computing software, given today's machinery, is readily available. In particular, we have left out a summary of linear regression models using the more general ℓ_p-norms with $p \notin \{1,2,\infty\}$ for which the computational requirements are considerably more burdensome than in the linear programming case (as they generally require methods from *convex programming* where machine computations are far more limited today). Even with this self-imposed restriction on ℓ_1-, ℓ_2- and ℓ_∞-regression, we have managed to clarify certain issues like the *uniqueness* of ℓ_1- and ℓ_∞-regression estimates in our review, for which we give *deterministic* necessary and sufficient conditions in Sections 2.2 and 2.3, respectively, which appear to be new.

REFERENCES

1. H.L. Harter, The method of least squares and some alternatives—Part I, *Int. Stat. Rev.* **42**, 147–174 (1974).
2. H.L. Harter, The method of least squares and some alternatives—Part II, *Int. Stat. Rev.* **42**, 235–264 (1974).
3. H.L. Harter, The method of least squares and some alternatives—Part III, *Int. Stat. Rev.* **43**, 1–44 (1975).
4. H.L. Harter, The method of least squares and some alternatives—Part IV, *Int. Stat. Rev.* **43**, 125–190 (1975).
5. H.L. Harter, The method of least squares and some alternatives—Part V, *Int. Stat. Rev.* **43**, 269–272 (1975).
6. H.L. Harter, The method of least squares and some alternatives—Addendum to Part IV, *Int. Stat. Rev.* **43**, 273–278 (1975).
7. A. Sen and M. Srivastava, *Regression Analysis: Theory, Methods, and Applications*, Springer-Verlag, New York, (1990).
8. H. White, *Asymptotic Theory for Econometricians*, Academic Press, Orlando, FL, (1984).
9. E.F. Fama, Portfolio analysis in a stable Paretian market, *Management Science* **11**, 404–419 (1965).
10. B. Mandelbrot, The variation of some other speculative prices, *Journal of Business* **XL**, 393–413 (1967).
11. R.C. Blattberg and T. Sargent, Regression with non-Gaussian stable disturbances: Some sampling results, *Econometrica* **39**, 501–510 (1971).
12. W.F. Sharpe, Mean-absolute-deviation characteristic lines for securities and portfolios, *Management Science* **18**, B1–B13 (1971).
13. M. Padberg, *Linear Optimization and Extensions*, Springer-Verlag, Berlin, (1995).
14. R. Koenker and G. Bassett, Regression quantiles, *Econometrica* **46**, 33–50 (1978).
15. B. Rosenberg and D. Carlson, A simple approximation of the sampling distibution of least absolute residuals regression estimates, *Communications in Statistics—Simula. Computa.* **B6** (4), 421–437 (1977).
16. L.D. Taylor, Estimation by minimizing the sum of absolute errors, In *Frontiers in Econometrics*, (Edited by P. Zarembka), pp. 169–190, Academic Press, New York, (1974).

17. R.L. Sielken and H.O. Hartley, Two linear programming algorithms for unbiased estimation of linear models, *Journal of the American Statistical Association* **68** (1973).
18. G. Bassett and R. Koenker, Asymptotic theory of least absolute error regression, *Journal of the American Statistical Association* **73**, 618–622 (1978).
19. D. Pollard, Asymptotics for least absolute deviation regression estimators, *Econometric Theory* **7**, 186–199 (1991).
20. H. Theil, *Principles of Econometrics*, Wiley, New York, (1971).
21. R. Koenker and G. Bassett, Tests of linear hypotheses and ℓ_1 estimation, *Econometrica* **50**, 1577–1583 (1982).
22. T. Dielman and R. Pfaffenberger, LAV (least absolute value) estimation in linear regression: A review, *TIMS/Studies in the Management Sciences* **19**, 31–52 (1982).
23. A.M. Ostrowski, *Solutions of Equations in Euclidean and Banach Spaces*, Academic Press, New York, (1973).
24. P. Kirchberger, Über Tschebyschefsche Annäherungsmethoden, *Mathematische Annalen* **57**, 509–540 (1903).
25. E.W. Cheney, *Introduction to Approximation Theory*, McGraw-Hill, New York, (1966).
26. J.D. Broffitt, An example of the large sample behavior of the midrange, *The American Statistician* **28**, 69–70 (1974).
27. H.A. David, *Order Statistics*, John Wiley & Sons, (1981).
28. H.L. Harter, Nonuniqueness of least absolute values regression, *Communications in Statistics—Theory and Methods* **A6** (9), 829–838 (1977).
29. S.C. Narula and J.F. Wellington, The minimum sum of absolute errors regression: A state of the art survey, *International Statistical Review* **50**, 317–326 (1982).
30. M. Padberg and J. Wigington, Efficient computation of MSAE regression estimates using the dual simplex bounded variables technique, Graduate School of Industrial Administration, Carnegie-Mellon University (1969).
31. J.C. Wigington, MSAE estimation: An alternative approach to regression analysis for economic forecasting applications, *Applied Economics* **4**, 11–21 (1972).