11

Regression Models in Risk Management

Stan Uryasev
University of Florida, USA

This chapter discusses theory and application of generalized linear regression that minimizes a general error measure of regression residual subject to various constraints on regression coefficients and includes least-squares linear regression, median regression, quantile regression, mixed quantile regression, and robust regression as special cases. General error measures are nonnegative positively homogeneous convex functionals that generalize the notion of norm and, in general, are asymmetric with respect to ups and downs of a random variable, which allows one to treat gains and losses differently. Each nondegenerate error measure $\mathcal{E}$ yields the deviation measure $\mathcal{D}$ projected from $\mathcal{E}$ and the statistic $\mathcal{S}$ associated with $\mathcal{E}$. General deviation measures are also nonnegative positively homogeneous convex functionals, which, in contrast to error measures, are insensitive to a constant shift. They generalize the notion of standard deviation, but are not required to be symmetric. General deviation measures admit dual characterization in terms of risk envelopes, which is instrumental in devising efficient optimization formulations for minimization of deviation measures. The central theoretical result in generalized linear regression is the error decomposition theorem stating that minimization of an error measure of the regression residual can be decomposed into minimizing the projected deviation measure of the residual without the intercept and into setting the intercept to the statistic associated with the error measure. The value of this theorem is that minimization of deviation measures admits dual formulation in terms of risk envelopes and, as a result, yields efficient optimization formulations. Application of generalized linear regression includes examples of financial index tracking, sparse signal reconstruction, therapy treatment planning, collateralized debt obligation, mutual fund return-based style classification, and mortgage pipeline hedging. The examples also provide linear program formulations of the corresponding regressions.

Financial Signal Processing and Machine Learning, First Edition.
Edited by Ali N. Akansu, Sanjeev R. Kulkarni and Dmitry Malioutov.

The chapter is organized as follows. The introduction discusses a general setup of linear regression. Section 11.2 introduces general error and deviation measures, whereas Section 11.3 introduces risk envelopes and risk identifiers. Section 11.4 states the error decomposition theorem. Sections 11.5, 11.6, and 11.7 formulate least-squares linear regression, median regression, and quantile regression, respectively, and present application of these regressions in financial engineering and signal processing. Section 11.7 also formulates mixed quantile regression as a generalization of quantile regression. Section 11.8 introduces unbiased linear regression and risk acceptable linear regression with application to financial index tracking. Section 11.9 discusses robust regression with application to mortgage pipeline hedging.

11.1 Introduction

In statistics, regression analysis aims to find the best relationship between a *response* random variable Y (*regressant*) and n independent variables $x_1, \dots, x_n$ (*regressors*) in the form:

$$Y = f(x_1, \dots, x_n) + \epsilon,$$

based on m available simultaneous observations of $x_1, \dots, x_n$ and Y (regression data): $x_{1j}, \dots, x_{nj}, y_j, j = 1, \dots, m$, where ϵ is the *approximation error*.

There are two main classes of regression: *parametric* and *nonparametric*. If the function f is determined by a finite set of parameters, regression is called parametric; otherwise, it is called nonparametric. The class of parametric regressions is further divided into *linear* and *nonlinear*. In linear regression, f is a linear function with respect to unknown parameters, whereas $x_1, \dots, x_n$ can be involved nonlinearly (though, in some definitions, linear regression is assumed to be linear with respect to $x_1, \dots, x_n$; see, e.g., Hastie *et al.*, 2008). Typically, in linear regression, f has the form

$$f(x_1, \dots, x_n) = c_0 + c_1 x_1 + \dots + c_n x_n = c_0 + \sum_{k=1}^{n} c_k x_k, \tag{11.1}$$

where $c_k \in \mathbb{R}$, $k = 0, 1, \dots, n$, are unknown regression parameters with c_0 called *intercept* or *bias*. (Estimates of $c_0, c_1, \dots, c_n$ found from the regression data are denoted by $\widehat{c}_0, \widehat{c}_1, \dots, \widehat{c}_n$, respectively.) In nonlinear regression, f is a nonlinear function of specified unknown parameters, which are usually found iteratively.

One of the main approaches for finding estimates of regression parameters is to maximize the likelihood of the observations of $y_1, \dots, y_m$ under the assumption that the residuals $e_j = y_j - f(x_{1j}, \dots, x_{nj})$, $j = 1, \dots, m$, are realizations of *independent and identically distributed* (i.i.d.) random variables $\epsilon_1, \dots, \epsilon_m$ with zero mean; see van der Waerden (1957). For example, if $\epsilon_1, \dots, \epsilon_m$ are i.i.d. and have the normal distribution $N(0, \sigma^2)$, then the likelihood of observing $y_1, \dots, y_m$ is given by

$$\frac{1}{(\sqrt{2\pi}\sigma)^m} \prod_{j=1}^{m} \exp\left(-\frac{1}{2\sigma^2}\big(y_j - f(x_{1j}, \dots, x_{nj})\big)^2\right),$$

and its maximization simplifies to

$$\min \sum_{j=1}^{m} (y_j - f(x_{1j}, \dots, x_{nj}))^2,$$

which is called the *least-squares method*. With $f(x_1, \dots, x_n)$ in the form of (11.1), this minimization problem yields a system of linear equations for estimates $\widehat{c}_1, \dots, \widehat{c}_n$:

$$\sum_{k=1}^{n} \widehat{c}_k \sum_{j=1}^{n} (x_{ij} - \widetilde{x}_i)(x_{kj} - \widetilde{x}_k) = \sum_{j=1}^{m} (x_{ij} - \widetilde{x}_i)(y_j - \widetilde{y}), \quad i = 1, \dots, n, \tag{11.2}$$

with $\widehat{c}_0 = \widetilde{y} - \sum_{k=1}^{n} c_k \widetilde{x}_k$, where $\widetilde{x}_i = \frac{1}{n} \sum_{j=1}^{m} x_{ij}$ for $i = 1, \dots, n$ and $\widetilde{y} = \frac{1}{n} \sum_{j=1}^{m} y_j$.

Even if $\epsilon_1, \dots, \epsilon_m$ are only uncorrelated (not necessarily independent) with zero mean and same variance, the *Gauss–Markov theorem* states that the *best linear unbiased estimator* (BLUE) of the form (11.1) is determined by *least-squares linear regression*. If $\epsilon_1, \dots, \epsilon_m$ are correlated and/or not identically distributed random variables, then the least-squares regression may not be appropriate.

Statistical approximation theory takes a different perspective on regression: when the response random variable Y is not understood completely and is better to be treated as a function $f(X_1, \dots, X_n)$ of *random* variables $X_1, \dots, X_n$, the error $Y - f(X_1, \dots, X_n)$ is sought to minimize some loss function or *error measure* with respect to unknown regression parameters; see Rockafellar *et al.* (2008). In this approach, central to regression analysis is the choice of error measure that should conform to *risk preferences* of an analyst. For example, if the problem is to track a stock market index by a portfolio of selected financial instruments, whose returns are random variables $X_1, \dots, X_n$, the analyst may penalize only underperformance of the portfolio return $f(X_1, \dots, X_n)$ with respect to the index return Y, so that symmetric measures like $\|\cdot\|_2$ are not appropriate.

This chapter pursues the statistical approximation approach to regression. It focuses on a general theory of approximating an output random variable Y by a linear combination of input random variables $X_1, \dots, X_n$:

$$f(X_1, \dots, X_n) = c_0 + c_1 X_1 + \dots + c_n X_n = c_0 + \sum_{k=1}^{n} c_k X_k$$

with an arbitrary error measure $\mathcal{E}$ under additional constraints on regression coefficients.

11.2 Error and Deviation Measures

Let $(\Omega, \mathcal{M}, \mathbb{P})$ be a probability space of elementary events Ω with the sigma-algebra $\mathcal{M}$ over Ω and with a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{M})$. Random variables are assumed to be measurable real-valued functions from $\mathcal{L}^2(\Omega) = \mathcal{L}^2(\Omega, \mathcal{M}, \mathbb{P})$ unless otherwise specified,[1] and the relationships between random variables X and Y (e.g., $X \le Y$ and $X = Y$) are understood to hold in the almost-sure sense (i.e., $\mathbb{P}[X \le Y] = 1$ and $\mathbb{P}[X = Y] = 1$), respectively. Also, $\inf X$ and $\sup X$ mean *essential infimum* and *essential supremum* of X (i.e., $\operatorname{ess\,inf} X$ and $\operatorname{ess\,sup} X$), respectively. Two important integral characteristics of a random variable X are its *mean* and *variance*, defined by

$$\mu(X) = \int_{\Omega} X(\omega)\, d\mathbb{P}[\omega], \quad \sigma^2(X) = \int_{\Omega} (X(\omega) - \mu(X))^2\, d\mathbb{P}[\omega],$$

[1] $\mathcal{L}^2(\Omega)$ is the Lebesgue space of measurable square-integrable functions on Ω: $X \in \mathcal{L}^2(\Omega)$ is equivalent to $\int_{\Omega} |X(\omega)|^2 d\mathbb{P}[\omega] < \infty$.

respectively, and $\sigma(X)$ is called the *standard deviation* of X. If $X \in \mathcal{L}^2(\Omega)$, then $\mu(X)$ and $\sigma^2(X)$ are well defined (bounded), which explains the choice of $\mathcal{L}^2(\Omega)$.

Rockafellar *et al.* (2002, 2006a, 2008) introduced *error measures* as functionals $\mathcal{E} : \mathcal{L}^2(\Omega) \to [0, \infty]$ satisfying

(E1) *Nonnegativity*: $\mathcal{E}(0) = 0$, but $\mathcal{E}(X) > 0$ for $X \neq 0$; also, $\mathcal{E}(c) < \infty$ for constants c.
(E2) *Positive homogeneity*: $\mathcal{E}(\lambda X) = \lambda \mathcal{E}(X)$ when $\lambda > 0$.
(E3) *Subadditivity*: $\mathcal{E}(X + Y) \leq \mathcal{E}(X) + \mathcal{E}(Y)$ for all X and Y.
(E4) *Lower semicontinuity*: set $\{X \in \mathcal{L}^2(\Omega) | \mathcal{E}(X) \leq c\}$ is closed for all $c < \infty$.

For example, $\mathcal{L}^p$ norms $\|X\|_p$ with $p \geq 1$ are error measures. However, error measures are not required to be symmetric; that is, in general, $\mathcal{E}(-X) \neq \mathcal{E}(X)$. An example of an *asymmetric* error measure is given by

$$\mathcal{E}_{a,b,p}(X) = \|a\, X_+ + b\, X_-\|_p, \quad a > 0, \quad b > 0, \quad 1 \leq p \leq \infty. \tag{11.3}$$

Another one is the asymmetric mean absolute error:

$$\mathcal{E}_\alpha(X) = \frac{1}{\alpha} E[\alpha\, X_+ + (1 - \alpha)\, X_-], \quad \alpha \in (0, 1), \tag{11.4}$$

where $X_\pm = \max\{0, \pm X\}$. For $\alpha = 1/2$, $\mathcal{E}_\alpha(X)$ simplifies to $\|X\|_1$. Observe that for $a = 1$ and $b = 1$, (11.3) simplifies to $\|X\|_p$, whereas for $p = 1$, $a = 1$, and $b = 1/\alpha - 1$, it reduces to (11.4).

An error measure $\mathcal{E}$ is *nondegenerate* if there exists $\delta > 0$ such that $\mathcal{E}(X) \geq \delta\, |E[X]|$ for all X. For example, (11.3) and (11.4) are both nondegenerate error measures with $\delta = \min\{a, b\}$ and $\delta = \min\{1, 1/\alpha - 1\}$, respectively; see Rockafellar *et al.* (2008).

Similar to error measures, Rockafellar *et al.* (2002, 2006a) introduced deviation measures as functionals $\mathcal{D} : \mathcal{L}^2(\Omega) \to [0, \infty]$ satisfying

(D1) *Nonnegativity*: $\mathcal{D}(X) = 0$ for constant X, but $\mathcal{D}(X) > 0$ otherwise.
(D2) *Positive homogeneity*: $\mathcal{D}(\lambda X) = \lambda \mathcal{D}(X)$ when $\lambda > 0$.
(D3) *Subadditivity*: $\mathcal{D}(X + Y) \leq \mathcal{D}(X) + \mathcal{D}(Y)$ for all X and Y.
(D4) *Lower semicontinuity*: set $\{X \in \mathcal{L}^2(\Omega) | \mathcal{D}(X) \leq c\}$ is closed for all $c < \infty$.

It follows from D1 and D3 that

$$\mathcal{D}(X - c) = \mathcal{D}(X) \qquad \text{for all constants } c,$$

which is known as *insensitivity to constant shift* (see Rockafellar *et al.*, 2006a). Axioms D1–D4 generalize well-known properties of the standard deviation; however, they do not imply symmetry, so that in general, $\mathcal{D}(-X) \neq \mathcal{D}(X)$.

Each error measure $\mathcal{E}$ yields a deviation measure through *penalties relative to expectation*

$$\mathcal{D}(X) = \mathcal{E}(X - E[X]), \tag{11.5}$$

and if $\mathcal{E}$ is nondegenerate, it furnishes another deviation through *error projection*

$$\mathcal{D}(X) = \inf_{c \in IR} \mathcal{E}(X - c), \tag{11.6}$$

which is called *the deviation of X projected from* $\mathcal{E}$; see Theorem 2.1 in Rockafellar *et al.* (2008). A solution to (11.6) is *the statistic of X associated with* $\mathcal{E}$

$$\mathcal{S}(X) = \arg\min_{c \in \mathbb{R}} \mathcal{E}(X - c), \tag{11.7}$$

which, in general, is an interval $[\mathcal{S}^-(X), \mathcal{S}^+(X)]$ of constants with $\mathcal{S}^-(X) = \min\{c | c \in \mathcal{S}(X)\}$ and $\mathcal{S}^+(X) = \max\{c | c \in \mathcal{S}(X)\}$, and has the following properties:

$$\mathcal{S}(X - c) = \mathcal{S}(X) - c \quad \text{for any constant } c,$$
$$\mathcal{S}(\lambda\, X) = \lambda\, \mathcal{S}(X) \quad \text{for any constant } \lambda > 0.$$

Well-known examples of the relationships (11.6) and (11.7) are given in the following table:

$\mathcal{E}(X)$	$\mathcal{D}(X)$	$\mathcal{S}(X)$
$\|X\|_2$	$\sigma(X)$	$E[X]$
$\|X\|_1$	$\|X - \text{med}\,(X)\|_1$	med (X)
$\frac{1}{\alpha}E[\alpha\, X_+ + (1-\alpha)\, X_-]$	$\text{CVaR}_\alpha^\Delta(X)$	$q_X(\alpha) = [q_X^-(\alpha), q_X^+(\alpha)]$

where med (X) is the median of X (possibly an interval),

$$q_X^-(\alpha) = \inf\{t \mid F_X(t) \geq \alpha\} \quad \text{and} \quad q_X^+(\alpha) = \sup\{t \mid F_X(t) \leq \alpha\}$$

are *lower* and *upper* α*-quantiles*, respectively, and $\text{CVaR}_\alpha^\Delta(X)$ is conditional value at risk (CVaR) deviation defined by

$$\text{CVaR}_\alpha^\Delta(X) = E[X] - \frac{1}{\alpha}\int_0^\alpha q_X^+(s)\, ds. \tag{11.8}$$

Observe that for $\mathcal{E}(X) = \|X\|_2$, deviations (11.5) and (11.6) coincide, whereas for $\mathcal{E}(X) = \|X\|_1$, they are different.

For a given deviation measure $\mathcal{D}$, a nondegenerate error measure can be obtained by *inverse projection*

$$\mathcal{E}(X) = \mathcal{D}(X) + |E[X]|,$$

which through (11.6) projects back to $\mathcal{D}$ with the associated statistic $\mathcal{S}(X) = E[X]$; see Rockafellar *et al.* (2008, Example 2.5).

If $\mathcal{E}_1, \ldots, \mathcal{E}_l$ are nondegenerate error measures that project to deviations $\mathcal{D}_1, \ldots, \mathcal{D}_l$, respectively, then, for any weights $\lambda_1 > 0, \ldots, \lambda_l > 0$ with $\sum_{k=1}^l \lambda_k = 1$,

$$\mathcal{E}(X) = \inf_{\substack{c_1, \ldots, c_l \\ \lambda_1 c_1 + \ldots + \lambda_l c_l = 0}} \sum_{k=1}^{l} \lambda_k\, \mathcal{E}_k(X - c_k)$$

is a nondegenerate error measure, which projects to the deviation measure

$$\mathcal{D}(X) = \sum_{k=1}^{l} \lambda_k \mathcal{D}_k(X)$$

with the associated statistic

$$\mathcal{S}(X) = \sum_{k=1}^{l} \lambda_k \mathcal{S}_k(X).$$

See Theorem 2.2 in Rockafellar *et al.* (2008). As an immediate consequence of this result, we restate Example 2.6 from Rockafellar *et al.* (2008).

Example 11.1 (mixed quantiles and mixed-CVaR deviation) *For any choice of probability thresholds* $\alpha_k \in (0,1)$ *and weights* $\lambda_1 > 0, \ldots, \lambda_l > 0$ *with* $\sum_{k=1}^{l} \lambda_k = 1$,

$$\mathcal{E}(X) = E[X] + \inf_{\substack{c_1,\ldots,c_l \\ \lambda_1 c_1 + \ldots + \lambda_l c_l = 0}} \sum_{k=1}^{l} \frac{\lambda_k}{\alpha_k} E[\max\{0, c_k - X\}] \tag{11.9}$$

is a nondegenerate error measure called mixed quantile error measure*, which projects to the mixed CVaR deviation*

$$\mathcal{D}(X) = \sum_{k=1}^{m} \lambda_k \, \mathrm{CVaR}^{\Delta}_{\alpha_k}(X), \quad \sum_{k=1}^{m} \lambda_k = 1, \lambda_k > 0, \quad k = 1, \ldots, m, \tag{11.10}$$

with the associated statistic

$$\mathcal{S}(X) = \sum_{k=1}^{l} \lambda_k \, q_X(\alpha_k), \quad q_X(\alpha_k) = [q_X^-(\alpha_k), q_X^+(\alpha_k)]. \tag{11.11}$$

11.3 Risk Envelopes and Risk Identifiers

Deviation measures have dual characterization in terms of *risk envelopes* $\mathcal{Q} \subset \mathcal{L}^2(\Omega)$ defined by the following properties:

(Q1) $\mathcal{Q}$ is nonempty, closed, and convex;
(Q2) for every nonconstant X, there is some $Q \in \mathcal{Q}$ such that $E[XQ] < E[X]$; and
(Q3) $E[Q] = 1$ for all $Q \in \mathcal{Q}$.

There is a one-to-one correspondence between deviation measures and risk envelopes (Rockafellar *et al.* 2006a, Theorem 1):

$$\mathcal{D}(X) = E[X] - \inf_{Q \in \mathcal{Q}} E[XQ],$$
$$\mathcal{Q} = \{Q \in \mathcal{L}^2(\Omega) \mid \mathcal{D}(X) \geq E[X] - E[XQ] \text{ for all } X\}. \tag{11.12}$$

The elements of $\mathcal{Q}$ at which $E[XQ]$ attains infimum for a given X are called *risk identifiers* for X:

$$\mathcal{Q}(X) = \arg\min_{Q \in \mathcal{Q}} E[XQ].$$

They are those elements of $\mathcal{Q}$ that track the downside of X as closely as possible.

The second relationship in (11.12) implies that the set of risk identifiers for X with respect to a deviation measure $\mathcal{D}$ is determined by

$$\mathcal{Q}_{\mathcal{D}}(X) = \{Q \in \mathcal{Q} \mid \mathcal{D}(X) = E[(E[X] - X)Q] \equiv \mathrm{Cov}(-X, Q)\}.$$

From the optimization perspective, $\mathcal{Q}_{\mathcal{D}}(X)$ is closely related to *subdifferential* $\partial\mathcal{D}(X)$ of a deviation measure $\mathcal{D}$ at X, which is the set of *subgradients* $Z \in \mathcal{L}^2(\Omega)$ such that

$$\mathcal{D}(Y) \geq \mathcal{D}(X) + E[(Y - X)Z] \quad \text{for all} \quad Y \in \mathcal{L}^2(\Omega).$$

Proposition 1 in Rockafellar *et al.* (2006b) shows that

$$\partial\mathcal{D}(X) = 1 - \mathcal{Q}_{\mathcal{D}}(X).$$

11.3.1 Examples of Deviation Measures $\mathcal{D}$, Corresponding Risk Envelopes $\mathcal{Q}$, and Sets of Risk Identifiers $\mathcal{Q}_{\mathcal{D}}(X)$

1. standard deviation $\mathcal{D}(X) = \sigma(X) \equiv \|X - E[X]\|_2$:

$$\mathcal{Q} = \{Q \mid E[Q] = 1, \sigma(Q) \leq 1\}, \quad \mathcal{Q}_\sigma(X) = \left\{1 - \frac{X - E[X]}{\sigma(X)}\right\}$$

2. standard lower semideviation $\mathcal{D}(X) = \sigma_-(X) \equiv \|[X - E[X]]_-\|_2$:

$$\mathcal{Q} = \{Q \mid E[Q] = 1, \|Q - \inf Q\|_2 \leq 1\}, \quad \mathcal{Q}_{\sigma_-}(X) = \left\{1 - \frac{E[Y] - Y}{\sigma_-(X)}\right\},$$

   where $Y = [E[X] - X]_+$

3. mean absolute deviation $\mathcal{D}(X) = \mathrm{MAD}(X) \equiv |X - E[X]|_1$:

$$\mathcal{Q} = \{Q \mid E[Q] = 1, \sup Q - \inf Q \leq 2\},$$
$$\mathcal{Q}_{\mathrm{MAD}}(X) = \{Q = 1 + E[Z] - Z \mid Z \in \mathrm{sign}[X - E[X]]\}$$

4. lower worst-case deviation $\mathcal{D}(X) = E[X] - \inf X$:

$$\mathcal{Q} = \{Q \mid E[Q] = 1, Q \geq 0\},$$
$$\mathcal{Q}_{\mathcal{D}}(X) = \{Q \mid E[Q] = 1, Q \geq 0, Q(\omega) = 0 \,\text{when}\, X(\omega) > \inf X\}$$

5. CVaR deviation $\mathcal{D}(X) = \mathrm{CVaR}^{\Delta}_{\alpha}(X)$:

$$\mathcal{Q} = \{Q \mid E[Q] = 1, 0 \leq Q \leq 1/\alpha\} \tag{11.13}$$

and $\mathcal{Q}_{\mathrm{CVaR}^{\Delta}_{\alpha}}(X)$ is the set of elements Q such that $E[Q] = 1$ and

$$Q(\omega)\begin{cases} = \alpha^{-1} & \text{on } \{\omega \in \Omega \mid X(\omega) < -\mathrm{VaR}_{\alpha}(X)\}, \\ \in [0, \alpha^{-1}] & \text{on } \{\omega \in \Omega \mid X(\omega) = -\mathrm{VaR}_{\alpha}(X)\}, \\ = 0 & \text{on } \{\omega \in \Omega \mid X(\omega) > -\mathrm{VaR}_{\alpha}(X)\}. \end{cases} \tag{11.14}$$

If $\mathcal{D}_1, \dots, \mathcal{D}_m$ are deviation measures, then

$$\mathcal{D}(X) = \sum_{k=1}^{m} \lambda_k \, \mathcal{D}_k(X), \quad \sum_{k=1}^{m} \lambda_k = 1, \quad \lambda_k > 0, \quad k = 1, \dots, m, \tag{11.15}$$

and

$$\mathcal{D}(X) = \max\{\mathcal{D}_1(X), \dots, \mathcal{D}_m(X)\}, \tag{11.16}$$

are deviation measures as well, for which the risk envelopes are given by Proposition 4 in Rockafellar *et al.* (2006a):

$$\mathcal{Q} = \begin{cases} \text{closure of } \sum_{k=1}^{m} \lambda_k \, \mathcal{Q}_k \text{ for (11.15)}, \\ \text{closure of the convex hull of } \cup_{k=1}^{m} \mathcal{Q}_k \text{ for (11.16)}, \end{cases}$$

where $\mathcal{Q}_1, \dots, \mathcal{Q}_m$ are the risk envelopes for the deviation measures $\mathcal{D}_1, \dots, \mathcal{D}_m$. This result and the formula (11.13) imply that the risk envelope for the mixed CVaR deviation (11.10) is determined by

$$\mathcal{Q} = \text{closure of} \sum_{k=1}^{m} \lambda_k \, Q_k, \quad \text{where} \quad E[Q_k] = 1, \quad 0 \le Q_k \le 1/\alpha_k, \quad k = 1, \dots, m. \tag{11.17}$$

Risk identifiers along with risk envelopes are instrumental in formulating optimality conditions and devising optimization procedures in applications involving deviation measures. For example, if X is discretely distributed with $\mathbb{P}[X = x_k] = p_k$, $k = 1, \dots, n$, then with the risk envelope representation (11.13), the CVaR deviation (11.8) is readily restated as a linear program

$$\mathrm{CVaR}^{\Delta}_{\alpha}(X) = E[X] - \min_{q_1, \dots, q_n} \left\{ \sum_{k=1}^{n} q_k p_k x_k \;\middle|\; q_k \in [0, \alpha^{-1}], \sum_{k=1}^{n} q_k p_k = 1 \right\},$$

whereas for the same X, mixed CVaR deviation (11.10) with (11.17) can be represented by

$$\sum_{i=1}^{m} \lambda_i \, \mathrm{CVaR}^{\Delta}_{\alpha_i}(X) = E[X] - \min_{q_{ik}} \left\{ \sum_{i,k=1}^{m,n} \lambda_i \, q_{ik} p_k x_k \;\middle|\; q_{ik} \in [0, \alpha_i^{-1}], \sum_{k=1}^{n} q_{ik} p_k = 1 \right\}.$$

11.4 Error Decomposition in Regression

An unconstrained generalized linear regression problem is formulated as follows: *approximate a random variable* $Y \in \mathcal{L}^2(\Omega)$ *by a linear combination* $c_0 + \sum_{k=1}^{n} c_k X_k$ *of given random variables* $X_k \in \mathcal{L}^2(\Omega)$, $k = 1, \dots, n$, *and minimize an error measure* $\mathcal{E}$ *of the error*

$Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k$ *with respect to* $c_0, c_1, \ldots, c_n$, *where* $\mathcal{E}$ *is assumed to be nondegenerate and finite everywhere on* $\mathcal{L}^2(\Omega)$, or, formally,

$$\min_{c_0,c_1,\ldots,c_n} \mathcal{E}(Z) \quad \text{with} \quad Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k. \tag{11.18}$$

Observe that because of possible asymmetry of $\mathcal{E}$, $\mathcal{E}(-Z) \neq \mathcal{E}(Z)$.

Well-known particular cases of the linear regression (11.18) include:

1. Least-squares linear regression with $\mathcal{E}(Z) = \|Z\|_2$;
2. Median regression with $\mathcal{E}(Z) = \|Z\|_1$; and
3. Quantile regression with the asymmetric mean absolute error

$$\mathcal{E}(Z) = \frac{1}{\alpha} E[\alpha\, Z_+ + (1-\alpha)\, Z_-],$$

where $Z_\pm = \max\{0, \pm Z\}$ and $\alpha \in (0, 1)$.

The choice of error measure to be used in a given regression problem is determined by the particular application and risk preferences of a decision maker.

Theorem 3.2 in Rockafellar *et al.* (2008) shows that the generalized linear regression (11.18) can be decomposed into minimizing the projected deviation measure of $Y - \sum_{k=1}^{n} c_k X_k$ with respect to $c_1, \ldots, c_n$ and into setting the intercept c_0 to the associated statistic of optimal $Y - \sum_{k=1}^{n} c_k X_k$. In other words, (11.18) is reduced to

$$\min_{c_1,\ldots,c_n} \mathcal{D}(\widetilde{Z}) \quad \text{and} \quad c_0 \in \mathcal{S}(\widetilde{Z}) \quad \text{with} \quad \widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k, \tag{11.19}$$

where $\mathcal{D}(\widetilde{Z}) = \inf_{c \in \mathbb{R}} \mathcal{E}(\widetilde{Z} - c)$ is the deviation projected from $\mathcal{E}$, and $\mathcal{S}(\widetilde{Z}) = \arg\min_{c \in \mathbb{R}} \mathcal{E}(\widetilde{Z} - c)$ is the statistic associated with $\mathcal{E}$; see Rockafellar *et al.* (2008). This result is known as *error decomposition*.

Furthermore, Theorem 4.1 in Rockafellar *et al.* (2008) states that $c_1, \ldots, c_n$ is a solution to (11.19) if and only if

$$\text{there exists } Q \in \mathcal{Q}_{\mathcal{D}}(\tilde{Z}) \text{ such that } E[(1-Q)X_j] = 0 \text{ for } j = 1, \ldots, n, \tag{11.20}$$

where $\mathcal{Q}_{\mathcal{D}}(\widetilde{Z})$ is the risk identifier for $\widetilde{Z}$ with respect to deviation measure $\mathcal{D}$; see Rockafellar *et al.* (2008).

In many applications (e.g., factor models, index tracking, and replication problems), the coefficients $c_0, c_1, \ldots, c_n$ are often required to satisfy additional constraints. Let $\mathcal{C}$ be a feasible set of $n+1$ dimensional vector $c = (c_0, c_1, \ldots, c_n)$. For example, the requirement of $c_0, c_1, \ldots, c_n$ to be nonnegative translates into having $\mathcal{C} = \{c \in \mathbb{R}^{n+1} \mid c \geq 0\}$. In this case, the generalized linear regression takes the form

$$\min_{c_0,c_1,\ldots,c_n} \mathcal{E}\left(Y - c_0 - \sum_{k=1}^{n} c_k X_k\right) \quad \text{subject to} \quad (c_0, c_1, \ldots, c_n) \in \mathcal{C}. \tag{11.21}$$

Sections 11.5, 11.6, and 11.7 discuss the problem (11.21) with different error measures $\mathcal{E}$ and feasible sets $\mathcal{C}$ frequently arising in various statistical decision applications.

11.5 Least-Squares Linear Regression

A least-squares linear regression is one of the basic and most widely used statistical tools that finds its applications in virtually all areas of science dealing with data analysis and statistics (e.g., physics, biology, medicine, finance, and economics).

Unconstrained least-squares linear regression is a particular case of (11.18) with $\mathcal{E}(\cdot) = \|\cdot\|_2$ and is given by

$$\min_{c_0,c_1,\dots,c_n} \left\| Y - c_0 - \sum_{k=1}^{n} c_k X_k \right\|_2^2 . \tag{11.22}$$

The first-order necessary optimality conditions for the optimization problem (11.22) yield a system of linear equations for $c_0, c_1, \dots, c_n$:

$$\begin{cases} \sum\limits_{k=1}^{n} c_k \operatorname{Cov}(X_k, X_j) = \operatorname{Cov}(Y, X_j), j = 1, \dots, n, \\ c_0 = E[Y] - \sum\limits_{k=1}^{n} c_k \, E[X_k]. \end{cases} \tag{11.23}$$

If the covariance matrix Λ of $X_1, \dots, X_n$ is nonsingular, then the system can be solved either numerically or in a closed form through the inverse Λ^{-1} :

$$(c_1, \dots, c_n)^\top = \Lambda^{-1}(\operatorname{Cov}(Y, X_1), \dots, \operatorname{Cov}(Y, X_n))^\top.$$

This is the main advantage of the least-squares linear regression.

The system (11.23) shows that the least-squares linear regression is solved in two steps: finding $c_1, \dots, c_n$ and then determining c_0. In fact, for $\mathcal{D} = \sigma$, the error decomposition formulation (11.19) takes the form

$$\min_{c_1,\dots,c_n} \sigma(\widetilde{Z}) \quad \text{and} \quad c_0 = E[\widetilde{Z}], \quad \text{where} \quad \widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k,$$

which states that the least-squares linear regression is equivalent to minimizing variance of $Y - \sum_{k=1}^{n} c_k X_k$ with respect to $c_1, \dots, c_n$ and then setting intercept c_0 to the mean of $Y - \sum_{k=1}^{n} c_k X_k$. This fact is often taken for granted and may create the impression that the linear regression with another error measure $\mathcal{E}$ also leads to c_0 being $E[Y - \sum_{k=1}^{n} c_k X_k]$. However, this is possible only if the *deviation projected from* $\mathcal{E}$ coincides with the *deviation from the penalties relative to expectation*; see Rockafellar *et al.* (2008).

With the risk identifier corresponding to the standard deviation, that is,

$$\mathcal{Q}_\sigma(X) = \left\{ 1 - \frac{X - E[X]}{\sigma(X)} \right\},$$

the optimality conditions (11.20) can be recast in the form

$$E\left[\left(Y - \sum_{k=1}^{n} c_k X_k \right) (X_j - E[X_j]) \right] = 0, \quad j = 1, \dots, n,$$

which with $c_0 = E\left[Y - \sum_{k=1}^{n} c_k X_k\right]$ are equivalent to the system (11.23).

In contrast to (11.2), the system (11.23) yields "true" $c_0, c_1, \dots, c_n$ (not estimates) provided that the expected values $E[Y]$ and $E[X_k]$ and the covariances $\mathrm{Cov}(X_k, X_j)$ and $\mathrm{Cov}(Y, X_j)$ are known. However, in real-life problems, this is almost never the case: we are only given simultaneous observations of $X_1, \dots, X_n$ and Y: $x_{1j}, \dots, x_{nj}, y_j, j = 1, \dots, m$, so that the expected values and covariances should be estimated through the given data.

In applications, least-squares linear regression is often solved subject to additional constraints on regression coefficients and, in general, can be formulated by

$$\min_{c_0, c_1, \dots, c_n} \left\| Y - c_0 - \sum_{k=1}^{n} c_k X_k \right\|_2 \quad \text{subject to} \quad (c_0, c_1, \dots, c_n) \in \mathcal{C}, \tag{11.24}$$

where $\mathcal{C}$ is some feasible set of $(c_0, c_1, \dots, c_n)$. This problem admits a closed-form solution only in a few simple cases, for example when $\mathcal{C}$ is determined by a set of linear equalities. In a general case, (11.24) is solved numerically.

Example 11.2 (index tracking with mean square error) *Let Y be the daily rate of return of a stock market index (e.g., S&P 500 and Nasdaq), and let $X_1, \dots, X_n$ be the daily rates of return of chosen financial instruments. Suppose a unit capital is to be allocated among these instruments with capital weights $c_1, \dots, c_n$ to replicate the index's rate of return by a linear combination of $X_1, \dots, X_n$ without shorting of the instruments. The imposed requirements on $c_1, \dots, c_n$ correspond to the feasible set*

$$\mathcal{C} = \left\{ (c_1, \dots, c_n) \in \mathbb{R}^n \;\middle|\; \sum_{k=1}^{n} c_k = 1,\ c_k \geq 0, k = 1, \dots, n \right\}. \tag{11.25}$$

In this case, optimal allocation positions $c_1, \dots, c_n$ can be found through the least-squares linear regression (11.24) with $c_0 = 0$ and $\mathcal{C}$ given by (11.25), which is a quadratic optimization problem.

Another application of constrained least-squares linear regression is sparse signal reconstruction, whose objective is to find a decision vector that has few nonzero components and satisfies certain linear constraints. The SPARCO toolbox offers a wide range of test problems for benchmarking of algorithms for sparse signal reconstruction; see http://www.cs.ubc.ca/labs/scl/sparco/. Typically, SPARCO toolbox problems are formulated in one of three closely related forms: *L1Relaxed*, *L1Relaxed D*, and *L2 D* (or *LASSO*). Both "L1Relaxed" and "L1Relaxed D" formulations minimize the $\mathcal{L}_1$-error of the regression residual subject to box constraints on decision variables and subject to a constraint on the $\mathcal{L}_1$-norm of the decision vector.[2] The difference in these two formulations is that "L1Relaxed D" splits each decision variable c_i into two nonnegative variables $c_i^+ = \max\{c_i, 0\}$ and $c_i^- = \max\{-c_i, 0\}$ ($c_i = c_i^+ - c_i^-$ and $|c_i| = c_i^+ + c_i^-$) and, as a result, has all decision variables nonnegative. Since "L1Relaxed D" doubles the number of the decision variables, in some problems, it may be less efficient than "L1Relaxed." The "L2 D" formulation minimizes the weighted sum of the squared $\mathcal{L}_2$-norm of the regression residual and the $\mathcal{L}_1$-norm of the vector of regression coefficients subject to box constraints on the coefficients. As the "L1Relaxed D" formulation, this one also splits each regression coefficient into two nonnegative parts.

[2] $\mathcal{L}_1$-norm of a vector is the sum of absolute values of vector components.

Example 11.3 (sparse reconstruction problem from the SPARCO toolbox) *Let $L(c, X)$ be an error function that linearly depends on a decision vector $c = (c_1, \dots, c_n)$ and on a given random vector $X = (X_1, \dots, X_n)$. The "L2 D" formulation of the sparse reconstruction problem from the SPARCO toolbox is a regression that minimizes a linear combination of $\|L(c, X)\|_2^2$ and the regularization part $\sum_{i=1}^n |c_i|$ subject to box constraints $l_i \le c_i \le u_i$, $i = 1, \dots, n$, where l_i and u_i are given bounds with $u_i \ge 0$ and $l_i \le 0$. Let $c^{\pm} = (c_1^{\pm}, \dots, c_n^{\pm})$ with $c_i^{\pm} = \max\{\pm c_i, 0\}$, then $c_i = c_i^+ - c_i^-$ and $|c_i| = c_i^+ + c_i^-$, and the "L2 D" formulation takes the form*

$$\min_{c^+, c^-} \ \|L(c^+ - c^-, X)\|_2^2 + \lambda \sum_{i=1}^n (c_i^+ + c_i^-)$$

$$\text{subject to} \ \ 0 \le c_i^+ \le u_i, \quad 0 \le c_i^- \le -l_i, i = 1, \dots, n, \tag{11.26}$$

where λ is a given parameter.

Constrained least-squares linear regression is also used in an intensity-modulated radiation therapy (IMRT) treatment-planning problem formulated in Men *et al.* (2008). To penalize underdosing and overdosing with respect to a given threshold, the problem uses *quadratic one-sided penalties* or, equivalently, *second-order lower and upper partial moments*.

Example 11.4 (therapy treatment planning problem) *Let $[L(c, X)]_+ \equiv \max\{0, L(c, X)\}$ be a loss function, where $L(x, \theta)$ linearly depends on a decision vector $c = (c_1, \dots, c_n)$ and on a given random vector $X = (X_1, \dots, X_n)$. The regression problem, arising in intensity-modulated radiation therapy treatment, minimizes $\|[L(c, X)]_+\|_2^2$ subject to box constraints $l_i \le c_i \le u_i$ with given bounds l_i and u_i:*

$$\min_{c_1, \dots, c_n} \ \|[L(c, X)]_+\|_2^2 \quad \text{subject to} \quad l_i \le c_i \le u_i, \quad i = 1, \dots, n. \tag{11.27}$$

11.6 Median Regression

In the least-squares linear regression, large values of the error $Z = Y - c_0 - \sum_{k=1}^n c_k X_k$ are penalized heavier than small values, which makes the regression coefficients quite sensitive to outliers. In applications that require equal treatment of small and large errors, the median regression can be used instead.

Unconstrained median regression is a particular case of (11.18) with $\mathcal{E}(\cdot) = \|\cdot\|_1$:

$$\min_{c_0, c_1, \dots, c_n} \left\| Y - c_0 - \sum_{k=1}^n c_k X_k \right\|_1, \tag{11.28}$$

for which the error decomposition formulation (11.19) takes the form

$$\min_{c_1, \dots, c_n} E|\widetilde{Z} - \operatorname{med} \widetilde{Z}| \quad \text{and} \quad c_0 \in \operatorname{med} \widetilde{Z} \quad \text{with} \quad \widetilde{Z} = Y - \sum_{k=1}^n c_k X_k, \tag{11.29}$$

where $\operatorname{med} \widetilde{Z}$ is the median of $\widetilde{Z}$, which, in general, is any number in the closed interval $[q_{\widetilde{Z}}^-(1/2), q_{\widetilde{Z}}^+(1/2)]$. Observe that the median regression does not reduce to minimization of the mean-absolute deviation (MAD) and that c_0 is not the mean of $Y - \sum_{k=1}^n c_k X_k$.

Let $c_1, \ldots, c_n$ be an optimal solution to the problem (11.29), and let the random variable $\widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k$ have no probability "atom" at $q_{\widetilde{Z}}^{+}(1/2)$; then, the interval med $\widetilde{Z}$ is a singleton, and the optimality conditions (11.20) reduce to

$$E[X_j - E[X_j] | \widetilde{Z} \leq \text{med } \widetilde{Z}] = 0, j = 1, \ldots, n.$$

These conditions are, however, rarely used in practice.

In applications, $X_1, \ldots, X_n$ and Y are often assumed to be discretely distributed with joint probability distribution $\mathbb{P}[X_1 = x_{1j}, \ldots, X_n = x_{nj}, Y = y_j] = p_j > 0,\ j = 1, \ldots, m,$ with $\sum_{j=1}^{m} p_j = 1$. In this case,

$$\|Z\|_1 = \sum_{j=1}^{m} p_j \left| y_j - c_0 - \sum_{k=1}^{n} c_k x_{kj} \right|,$$

and the median regression (11.28) reduces to the linear program

$$\begin{aligned} \min_{\substack{c_0, c_1, \ldots, c_n, \\ \zeta_1, \ldots, \zeta_m}} & \sum_{j=1}^{m} p_j \zeta_j \\ \text{subject to} \quad & \zeta_j \geq y_j - c_0 - \sum_{k=1}^{n} c_k x_{kj}, \quad j = 1, \ldots, m, \\ & \zeta_j \geq c_0 + \sum_{k=1}^{n} c_k x_{kj} - y_j, \quad j = 1, \ldots, m, \end{aligned} \tag{11.30}$$

where $\zeta_1, \ldots, \zeta_m$ are auxiliary variables.

The median regression with constraints on regression coefficients is formulated by

$$\min_{c_0, c_1, \ldots, c_n} \left\| Y - c_0 - \sum_{k=1}^{n} c_k X_k \right\|_1 \quad \text{subject to} \quad (c_0, c_1, \ldots, c_n) \in \mathcal{C}, \tag{11.31}$$

where $\mathcal{C}$ is a given feasible set of $(c_0, c_1, \ldots, c_n)$. For an arbitrary joint probability distribution of $X_1, \ldots, X_n$ and Y, the necessary optimality conditions for (11.31) are given in Rockafellar *et al.* (2006a).

If $X_1, \ldots, X_n$ and Y are discretely distributed, and $\mathcal{C}$ is determined by a set of linear constraints, the Equation (11.31) reduces to a linear program.

Example 11.5 (index tracking with mean absolute error) *The setting is identical to that in Example 11.2. But this time, the optimal allocation positions $c_1, \ldots, c_n$ are found through the median regression (11.31) with $\mathcal{C}$ given by (11.25). If $X_1, \ldots, X_n$ and Y are assumed to be discretely distributed with joint probability distribution $\mathbb{P}[X_1 = x_{1j}, \ldots, X_n = x_{nj}, Y = y_j] = p_j > 0, j = 1, \ldots, m$, where $\sum_{j=1}^{m} p_j = 1$, then this regression problem can be formulated as the linear program*

$$\begin{aligned} \min_{\substack{c_1, \ldots, c_n \\ \zeta_1, \ldots, \zeta_m}} & \sum_{j=1}^{m} p_j \zeta_j \end{aligned}$$

$$\begin{aligned} \textit{subject to} \quad & \zeta_j \geq y_j - \sum_{k=1}^{n} c_k x_{kj}, \quad j = 1, \ldots, m, \\ & \zeta_j \geq \sum_{k=1}^{n} c_k x_{kj} - y_j, \quad j = 1, \ldots, m, \\ & \sum_{k=1}^{n} c_k = 1, \ c_k \geq 0, \quad k = 1, \ldots, n, \end{aligned}$$

where $\zeta_1, \ldots, \zeta_m$ are auxiliary variables.

Constrained median regression is also used to design a portfolio of *credit default swaps* (CDS) and credit indices to hedge against changes in a *collateralized debt obligation* (CDO) book. A CDS provides insurance against the risk of default (credit event) of a particular company. A buyer of the CDS has the right to sell bonds issued by the company for their face value when the company is in default. The buyer makes periodic payments to the seller until the end of the life of the CDS or until a default occurs. The total amount paid per year, as a percentage of the notional principal, is known as the *CDS spread*, which is tracked by credit indices. A CDO is a credit derivative based on defaults of a pool of assets. Its common structure involves tranching or slicing the credit risk of the reference pool into different risk levels of increasing seniority. The losses first affect the *equity* (first loss) tranche, then the *mezzanine* tranche, and finally the *senior* and *super senior* tranches. *The hedging problem is to minimize risk of portfolio losses subject to budget and cardinality constraints on hedge positions.* The risk is measured by mean absolute deviation (MAD) and by $\mathcal{L}_1$-norm (mean absolute penalty).

Example 11.6 (median regression and CDO) *Let $L(c, X)$ be a loss function in hedging against changes in a collateralized debt obligation (CDO) book, where $L(c, X)$ linearly depends on a decision vector $c = (c_1, \ldots, c_n)$ (positions in financial instruments) and on a given random vector $X = (X_1, \ldots, X_n)$. A regression problem then minimizes the mean absolute error of $L(c, X)$ subject to the budget constraint $\sum_{i=1}^{n} a_i |c_i| \leq C$ with given C and $a_i > 0$, $i = 1, \ldots, n$, and subject to a constraint on cardinality of the decision variables not to exceed a positive integer S:*

$$\begin{aligned} & \min_{c_1, \ldots, c_n} \ \|L(c, X)\|_1 \\ \textit{subject to} \quad & \sum_{i=1}^{n} a_i |c_i| \leq C, \\ & \sum_{i=1}^{n} I_{\{a_i |c_i| \geq w\}} \leq S, \\ & |c_i| \leq k_i, \quad i = 1, \ldots, n, \end{aligned} \tag{11.32}$$

where w is a given threshold; $I_{\{\cdot\}}$ is the indicator function equal to 1 if the condition in the curly brackets is true, and equal to 0 otherwise; and $|c_i| \leq k_i$, $i = 1, \ldots, n$, are bounds on decision variables (positions).

The next three examples formulate regression problems arising in sparse signal reconstruction. In all of them, $L(c, X)$ is an error function that linearly depends on a decision vector $c = (c_1, \dots, c_n)$ and on a given random vector $X = (X_1, \dots, X_n)$, and $l_i \le c_i \le u_i$, $i = 1, \dots, n$, are box constraints with given bounds l_i and u_i ($l_i \le u_i$).

Example 11.7 (sparse signal reconstruction I: "L1Relaxed" formulation) *This regression problem minimizes the mean absolute error of $L(c, X)$ subject to a constraint on cardinality of c with given integer bound S and subject to box constraints on c:*

$$\begin{aligned} &\min_{c_1,\dots,c_n} \ \|L(c,X)\|_1 \\ &\textit{subject to} \ \ \sum_{i=1}^{n} \left(I_{\{a_i c_i \ge w\}} + I_{\{b_i c_i \le -w\}} \right) \le S, \\ &\qquad l_i \le c_i \le u_i, \quad i = 1, \dots, n. \end{aligned} \tag{11.33}$$

Example 11.8 (sparse signal reconstruction II) *This regression problem minimizes the mean absolute error of $L(c, X)$ subject to a constraint on the $\mathcal{L}_1$-norm of c, that is, $\sum_{i=1}^{n} |c_i| \le U$ with given bound U, and subject to box constraints on c:*

$$\begin{aligned} &\min_{c_1,\dots,c_n} \ \|L(c,X)\|_1 \\ &\textit{subject to} \ \ \sum_{i=1}^{n} |c_i| \le U, \\ &\qquad l_i \le c_i \le u_i, \quad i = 1, \dots, n. \end{aligned} \tag{11.34}$$

Example 11.9 (sparse signal reconstruction III) *This estimation problems minimizes the cardinality of c subject to constraints on the mean absolute error of $L(c, X)$ and on the $\mathcal{L}_1$-norm of c with given bounds ϵ and U, respectively, and subject to box constraints on c:*

$$\begin{aligned} &\min_{c_1,\dots,c_n} \ \sum_{i=1}^{n} \left(I_{\{a_i c_i \ge w\}} + I_{\{b_i c_i \le -w\}} \right) \\ &\textit{subject to} \ \ \|L(c,X)\|_1 \le \epsilon, \\ &\qquad \sum_{i=1}^{n} |c_i| \le U, \\ &\qquad l_i \le c_i \le u_i, \quad i = 1, \dots, n. \end{aligned} \tag{11.35}$$

Example 11.10 presents a reformulation of the regression problem (11.34).

Example 11.10 (sparse signal reconstruction from SPARCO toolbox) *Suppose the random vector X is discretely distributed and takes on values $X^{(1)}, \dots, X^{(m)}$ with corresponding positive probabilities $p_1, \dots, p_m$ summing into 1, so that $\|L(c, X)\|_1 = \sum_{j=1}^{m} p_j |L(c, X^{(j)})|$. Let*

$c_i^{\pm} = \max\{\pm c_i, 0\}$, $i = 1, \dots, n$, *then* $c_i = c_i^+ - c_i^-$ *and* $|c_i| = c_i^+ + c_i^-$. *Given that* $L(c, X^{(j)})$ *is linear with respect to* c *and that* $u_i \geq 0$ *and* $l_i \leq 0$, $i = 1, \dots, n$, *the problem (11.34) can be restated as the linear program*

$$\begin{aligned} \min_{\substack{c^+, c^- \\ \zeta_1, \dots, \zeta_m}} \quad & \sum_{j=1}^{m} p_j \, \zeta_j \\ \text{subject to} \quad & \sum_{i=1}^{n} (c_i^+ + c_i^-) \leq U, \\ & \zeta_j \geq L(c^+ - c^-, X^{(j)}), \quad i = 1, \dots, n, \\ & \zeta_j \geq -L(c^+ - c^-, X^{(j)}), \quad i = 1, \dots, n, \\ & 0 \leq c_i^+ \leq u_i, \quad 0 \leq c_i^- \leq -l_i, \quad i = 1, \dots, n, \end{aligned} \tag{11.36}$$

where $\zeta_1, \dots, \zeta_m$ *are auxiliary variables.*

11.7 Quantile Regression and Mixed Quantile Regression

Both the least-squares linear regression and median regression treat ups and downs of the regression error equally, which might not be desirable in some applications. For example, in the index tracking problem from Example 11.2, a decision maker (financial analyst) may use a quantile regression that minimizes the asymmetric mean absolute error $\mathcal{E}_\alpha(Z) = \alpha^{-1} E[\alpha \, Z_+ + (1-\alpha) \, Z_-]$ of $Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k$ for some $\alpha \in (0, 1)$.

Unconstrained quantile regression is a particular case of the generalized linear regression (11.18) with the asymmetric mean absolute error measure:

$$\min_{c_0, c_1, \dots, c_n} E[\alpha \, Z_+ + (1-\alpha) Z_-] \quad \text{with} \quad Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k, \tag{11.37}$$

where $Z_{\pm} = \max\{\pm Z, 0\}$; and the multiplier α^{-1} in the objective function is omitted. Observe that for $\alpha = 1/2$, (11.37) is equivalent to the median regression (11.28).

In this case, the error decomposition formulation (11.19) takes the form

$$\min_{c_1, \dots, c_n} \mathrm{CVaR}_\alpha^{\Delta}(\widetilde{Z}) \quad \text{and} \quad c_0 \in [q_{\widetilde{Z}}^-(\alpha), q_{\widetilde{Z}}^+(\alpha)] \quad \text{with} \quad \widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k. \tag{11.38}$$

In other words, the quantile regression (11.37) reduces to minimizing *CVaR deviation* of $Y - \sum_{k=1}^{n} c_k X_k$ with respect to $c_1, \dots, c_n$ and to setting c_0 to any value from the α-quantile interval of $Y - \sum_{k=1}^{n} c_k X_k$.

Let $c_1, \dots, c_n$ be an optimal solution to the problem (11.38), and let the random variable $\widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k$ have no probability "atom" at $q_{\widetilde{Z}}^+(\alpha)$, then the interval $[q_{\widetilde{Z}}^-(\alpha), q_{\widetilde{Z}}^+(\alpha)]$ is a singleton, and the optimality conditions (11.20) simplify to

$$E\left[X_j - E[X_j] \,\middle|\, \widetilde{Z} \leq q_{\widetilde{Z}}^+(\alpha) \right] = 0, \quad j = 1, \dots, n.$$

However, in this form, they are rarely used in practice.

If $X_1, \dots, X_n$ and Y are discretely distributed with joint probability distribution $\mathbb{P}[X_1 = x_{1j}, \dots, X_n = x_{nj}, Y = y_j] = p_j > 0$, $j = 1, \dots, m$, where $\sum_{j=1}^m p_j = 1$, then with the formula (5) in Rockafellar *et al.* (2006a), the quantile regression (11.37) can be restated as the linear program

$$\min_{\substack{c_0, c_1, \dots, c_n \\ \zeta_1, \dots, \zeta_m}} \quad \sum_{j=1}^m p_j \left(y_j - c_0 + \alpha^{-1} \zeta_j - \sum_{k=1}^n c_k x_{kj} \right)$$

$$\text{subject to} \quad \zeta_j \geq c_0 + \sum_{k=1}^n c_k x_{kj} - y_j, \quad \zeta_j \geq 0, \quad j = 1, \dots, m, \tag{11.39}$$

where $\zeta_1, \dots, \zeta_m$ are auxiliary variables.

The *return-based style classification* for a mutual fund is a regression of the fund return on several indices as explanatory variables, where regression coefficients represent the fund's style with respect to each of the indices. In contrast to the least-squares regression, the quantile regression can assess the impact of explanatory variables on various parts of the regressand distribution, for example on the 95th and 99th percentiles. Moreover, for a portfolio with exposure to derivatives, the mean and quantiles of the portfolio return distribution may have quite different regression coefficients for the same explanatory variables. For example, in most cases, the strategy of investing into naked deep out-of-the-money options behaves like a bond paying some interest; however, in some rare cases, this strategy may lose significant amounts of money. With the quantile regression, a fund manager can analyze the impact of a particular factor on any part of the return distribution. Example 11.11 presents an unconstrained quantile regression problem arising in the return-based style classification of a mutual fund.

Example 11.11 (quantile regression in style classification) *Let $L(c, X)$ be a loss function that linearly depends on a decision vector $c = (c_1, \dots, c_n)$ and on a random vector $X = (X_1, \dots, X_n)$ representing uncertain rates of return of n indices as explanatory variables. The quantile regression (11.37) with $L(c, X)$ in place of Z takes the form*

$$\min_{c_1, \dots, c_n} E[\alpha [L(c, X)]_+ + (1 - \alpha)[L(c, X)]_-]. \tag{11.40}$$

A constrained quantile regression is formulated similarly to (11.37):

$$\min_{c_0, c_1, \dots, c_n} \quad E[\alpha \, Z_+ + (1 - \alpha) \, Z_-] \quad \text{with} \quad Z = Y - c_0 - \sum_{k=1}^n c_k X_k$$

$$\text{subject to} \quad (c_0, c_1, \dots, c_n) \in \mathcal{C}, \tag{11.41}$$

where $\mathcal{C}$ is a given feasible set for regression coefficients $c_0, c_1, \dots, c_n$.

Example 11.12 (index tracking with asymmetric mean absolute error) *The setting is identical to that in Example 11.2. But this time, the allocation positions $c_1, \dots, c_n$ are found from the constrained quantile regression (11.41) with $\mathcal{C}$ given by (11.25). If $X_1, \dots, X_n$ and*

Y are assumed to be discretely distributed with joint probability distribution $\mathbb{P}[X_1 = x_{1j}, \ldots, X_n = x_{nj}, Y = y_j] = p_j > 0$, $j = 1, \ldots, m$, *where* $\sum_{j=1}^{m} p_j = 1$, *then this regression problem can be formulated as the linear program*

$$\begin{aligned} \min_{\substack{c_1,\ldots,c_n \\ \zeta_1,\ldots,\zeta_m}} \quad & \sum_{j=1}^{m} p_j \left(y_j - \sum_{k=1}^{n} c_k x_{kj} + \alpha^{-1} \zeta_j \right) \\ \textit{subject to} \quad & \zeta_j \geq \sum_{k=1}^{n} c_k x_{kj} - y_j, \quad \zeta_j \geq 0, \quad j = 1, \ldots, m, \\ & \sum_{k=1}^{n} c_k = 1, \; c_k \geq 0, \quad k = 1, \ldots, n, \end{aligned}$$

where $\zeta_1, \ldots, \zeta_m$ *are auxiliary variables.*

The linear regression with the *mixed quantile error measure* (11.9) is called *mixed quantile regression*. It generalizes quantile regression and, through error decomoposition, takes the form

$$\min_{\substack{c_1,\ldots,c_n, \\ C_1,\ldots,C_l}} E\left[Y - \sum_{j=1}^{n} c_j X_j \right] + \sum_{k=1}^{l} \lambda_k \left(\frac{1}{\alpha_k} E\left[\max\left\{ 0, C_k - \sum_{j=1}^{n} c_j X_j \right\} \right] - C_k \right) \tag{11.42}$$

with the intercept c_0 determined by

$$c_0 = \sum_{k=1}^{l} \lambda_k \, C_k,$$

where $C_1, \ldots, C_l$ are a solution to (11.42); see Example 3.1 in Rockafellar *et al.* (2008).

The optimality conditions (11.20) for (11.42) are complicated. However, as the quantile regression, (11.42) can be reduced to a linear program.

11.8 Special Types of Linear Regression

This section discusses special types of unconstrained and constrained linear regressions encountered in statistical decision problems.

Often, it is required to find an unbiased linear approximation of an output random variable Y by a linear combination of input random variables $X_1, \ldots, X_n$, in which case the approximation error has zero expected value: $E[Y - c_0 - \sum_{k=1}^{n} c_k X_k] = 0$. A classical example of an *unbiased linear regression* is minimizing variance or, equivalently, standard deviation with the intercept c_0 set to $c_0 = E[Y - \sum_{k=1}^{n} c_k X_k]$. If, in this example, the standard deviation is replaced by a general deviation measure $\mathcal{D}$, we obtain a generalized unbiased linear regression:

$$\min_{c_1,\ldots,c_n} \mathcal{D}(\widetilde{Z}) \quad \text{and} \quad c_0 = E[\widetilde{Z}], \quad \text{where} \quad \widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k. \tag{11.43}$$

In fact, (11.43) is equivalent to minimizing the error measure $\mathcal{E}(Z) = \mathcal{D}(Z) + |E[Z]|$ of $Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k$. Observe that in view of the error decomposition theorem (Rockafellar *et al.*, 2008, Theorem 3.2), the generalized linear regression (11.18) with a nondegenerate error measure $\mathcal{E}$ and the unbiased linear regression (11.43) with the deviation measure $\mathcal{D}$ projected from $\mathcal{E}$ yield the same $c_1, \ldots, c_n$ but, in general, different intercepts c_0.

Rockafellar *et al.* (2008) introduced *risk acceptable linear regression* in which a deviation measure $\mathcal{D}$ of the approximation error $Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k$ is minimized subject to a constraint on the averse measure of risk $\mathcal{R}$ related to $\mathcal{D}$ by $\mathcal{R}(X) = \mathcal{D}(X) - E[X]$:

$$\min_{c_1,\ldots,c_n} \quad \mathcal{D}(Z) \quad \text{subject to} \quad \mathcal{R}(Z) = 0 \quad \text{with} \quad Z = Y - c_0 - \sum_{k=1}^{n} c_k X_k, \tag{11.44}$$

which is equivalent to

$$\min_{c_1,\ldots,c_n} \mathcal{D}(\widetilde{Z}) \quad \text{and} \quad c_0 = E(\widetilde{Z}) - \mathcal{D}(\widetilde{Z}), \quad \text{where} \quad \widetilde{Z} = Y - \sum_{k=1}^{n} c_k X_k. \tag{11.45}$$

The unbiased linear regression (11.43) and risk acceptable linear regression (11.45) show that the intercept c_0 could be set based on different requirements.

In general, the risk acceptable regression may minimize either an error measure $\mathcal{E}$ or a deviation measure $\mathcal{D}$ of the error Z subject to a constraint on a risk measure $\mathcal{R}$ of Z not necessarily related to $\mathcal{E}$ or $\mathcal{D}$. Example 11.13 illustrates a risk acceptable regression arising in a portfolio replication problem with a constraint on CVaR.

Example 11.13 (risk acceptable regression) *Let $L(c, X)$ be a portfolio replication error (loss function) that linearly depends on a decision vector $c = (c_1, \ldots, c_n)$ and on a random vector $X = (X_1, \ldots, X_n)$ representing uncertain rates of return of n instruments in a portfolio replicating the S&P 100 index. The risk acceptable regression minimizes the mean absolute error of $L(c, X)$ subject to the budget constraint $\sum_{i=1}^{n} a_i\, c_i \le U$ with known $a_1, \ldots, a_n$ and U and subject to a CVaR constraint on the underperformance of the portfolio compared to the index:*

$$\begin{aligned} &\min_{c_1,\ldots,c_n} \ \|L(c, X)\|_1 \\ &\textit{subject to} \ \ \sum_{i=1}^{n} a_i\, c_i \le U, \\ &\qquad CVaR_\alpha(L(c, X)) \le w, \\ &\qquad c_i \ge 0, \quad i = 1, \ldots, n, \end{aligned} \tag{11.46}$$

where α and w are given.

11.9 Robust Regression

Robust regression aims to reduce influence of sample outliers on regression parameters, especially when regression error has heavy tails.

In statistics, robustness of an estimator is a well-established notion and is assessed by the so-called *estimator's breakdown point*, which is the proportion of additional arbitrarily large

observations (outliers) needed to make the estimator unbounded. For example, the sample mean requires just a single such observation, while the sample median would still be finite until the proportion of such observations reaches 50%. Consequently, the mean's breakdown point is 0%, whereas the median's breakdown point is 50%.

As in the previous regression setting, suppose Y is approximated by a linear combination of input random variables $X_1, \dots, X_n$ with the regression error defined by $Z = Y - c_0 - \sum_{i=1}^{n} c_i X_i$, where $c_0, c_1, \dots, c_n$ are unknown regression coefficients. A robust regression minimizes an error measure of Z that has a nonzero breakdown point. Thus, in this setting, the regression's breakdown point is that of the error measure.

Often, a robust regression relies on order statistics of Z and on "trimmed" error measures. Two popular robust regressions are the *least median of squares (LMS) regression*, which minimizes the median of Z^2 and has a 50% breakdown point:

$$\min_{c_0,c_1,\dots,c_n} \operatorname{med}(Z^2) \quad \text{with} \quad Z = Y - c_0 - \sum_{i=1}^{n} c_i X_i. \tag{11.47}$$

and the *least-trimmed-squares (LTS) regression*, which minimizes the average α-quantile of Z^2 and has a $(1-\alpha)\cdot 100\%$ breakdown point:

$$\min_{c_0,c_1,\dots,c_n} \bar{q}_{Z^2}(\alpha) \quad \text{with} \quad Z = Y - c_0 - \sum_{i=1}^{n} c_i X_i. \tag{11.48}$$

Rousseeuw and Driessen (2006) referred to (11.48) as a challenging optimization problem.

Typically, in the LTS regression, α is set to be slightly larger than $1/2$. For $\alpha = 1$, $\bar{q}_{Z^2}(\alpha) = \|Z\|_2^2$, and (11.48) reduces to the standard least-squares regression. The LTS regression is reported to have advantages over the LMS regression or the one that minimizes the α-quantile of Z^2; see Rousseeuw and Driessen (2006), Rousseeuw and Leroy (1987), and Venables and Ripley (2002).

Let h be such that $h(t) > 0$ for $t \neq 0$ and $h(0) = 0$, but not necessarily symmetric (i.e., $h(-t) \neq h(t)$ in general). Then, the LMS and LTS regressions have the following generalization:

1. *Minimizing the upper α-quantile of $h(Z)$*:

$$\min_{c_0,c_1,\dots,c_n} q^{+}_{h(Z)}(\alpha) \quad \text{with} \quad Z = Y - c_0 - \sum_{i=1}^{n} c_i X_i, \tag{11.49}$$

2. *Minimizing the average α-quantile of $h(Z)$*:

$$\min_{c_0,c_1,\dots,c_n} \bar{q}_{h(Z)}(\alpha) \quad \text{with} \quad Z = Y - c_0 - \sum_{i=1}^{n} c_i X_i. \tag{11.50}$$

For example, in both (11.49) and (11.50), we may use $h(Z) = |Z|^p$, $p \geq 1$. In particular, for $h(Z) = Z^2$, (11.49) with $\alpha = 1/2$ corresponds to the LMS regression (11.47), whereas (11.50) reduces to the LTS regression (11.48).

When $h(-t) = h(t)$, (11.49) and (11.50) do not discriminate positive and negative errors. This, however, is unlikely to be appropriate for errors with significantly skewed distributions.

For example, instead of med (Z^2) and $\bar{q}_{Z^2}(\alpha)$, we can use *two-tailed α-value-at-risk (VaR) deviation* of the error Z defined by

$$\begin{aligned}\text{TwoTailVaR}_{\alpha}^{\Delta}(Z) &= \text{VaR}_{1-\alpha}(Z) + \text{VaR}_{1-\alpha}(-Z)\\ &\equiv q_Z^-(\alpha) - q_Z^+(1-\alpha), \quad \alpha \in (1/2, 1].\end{aligned} \tag{11.51}$$

The definition (11.51) shows that the two-tailed α-VaR deviation is, in fact, the range between the upper and lower $(1-\alpha)$-tails of the error Z, which is equivalent to the support of the random variable Z with truncated $(1-\alpha)\cdot 100\%$ of the "outperformances" and $(1-\alpha)\cdot 100\%$ of "underperformances." Consequently, the two-tailed α-VaR deviation has the breakdown point of $(1-\alpha)\cdot 100\%$. Typically, α is chosen to be 0.75 and 0.9.

Robust regression is used in mortgage pipeline hedging. Usually, mortgage lenders sell mortgages in the secondary market. Alternatively, they can exchange mortgages for mortgage-backed securities (MBSs) and then sell MBSs in the secondary market. The mortgage-underwriting process is known as the "pipeline." Mortgage lenders commit to a mortgage interest rate while the loan is in process, typically for a period of 30–60 days. If the rate rises before the loan goes to closing, the value of the loan declines and the lender sells the loan at a lower price. The risk that mortgages in process will fall in value prior to their sale is known as *mortgage pipeline risk*. Lenders often hedge this exposure either by selling forward their expected closing volume or by shorting either US Treasury notes or futures contracts. *Fallout* refers to the percentage of loan commitments that do not go to closing. It affects the mortgage pipeline risk. As interest rates fall, the fallout rises because borrowers locked in a mortgage rate are more likely to find better rates with another lender. Conversely, as rates rise, the percentage of loans that close increases. So, the fallout alters the size of the pipeline position to be hedged against and, as a result, affects the required size of the hedging instrument: at lower rates, fewer rate loans will close and a smaller position in the hedging instrument is needed. To hedge against the fallout risk, lenders often use options on US Treasury note futures.

Suppose a hedging portfolio is formed out of n hedging instruments with random returns $X_1, \ldots, X_n$. A pipeline risk hedging problem is to *minimize a deviation measure $\mathcal{D}$ of the underperformance of the hedging portfolio with respect to a random hedging target Y*, where short sales are allowed and transaction costs are ignored. Example 11.14 formulates a robust regression with the two-tailed α-VaR deviation used in a mortgage pipeline hedging problem.

Example 11.14 (robust regression with two-tailed α-VaR deviation) *Let a target random variable Y be approximated by a linear combination of n random variables $X_1, \ldots, X_n$, then the robust regression minimizes the two-tailed α-VaR deviation of the error $Y - c_0 - \sum_{i=1}^n c_i X_i$:*

$$\min_{c_0, c_1, \ldots, c_n} \text{TwoTailVaR}_{\alpha}^{\Delta}\left(Y - c_0 - \sum_{i=1}^{n} c_i X_i\right). \tag{11.52}$$

It has a $(1-\alpha)\cdot 100\%$-breakdown point.

References, Further Reading, and Bibliography

Hastie, T., Tibshirani, R. and Friedman, J. (2008). *The elements of statistical learning: data mining, inference, and prediction*, 2nd ed. New York: Springer.

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

Men, C., Romeijn, E., Taskin, C. and Dempsey, J. (2008). An exact approach to direct aperture optimization in IMRT treatment planning. *Physics in Medicine and Biology*, 52, 7333–7352.

Rockafellar, R.T., Uryasev, S. and Zabarankin, M. (2002). Deviation measures in risk analysis and optimization. Technical Report 2002-7, ISE Department, University of Florida, Gainesville, FL.

Rockafellar, R.T., Uryasev, S. and Zabarankin, M. (2006a). Generalized deviations in risk analysis. *Finance and Stochastics*, 10, 51–74.

Rockafellar, R.T., Uryasev, S. and Zabarankin, M. (2006b). Optimality conditions in portfolio analysis with general deviation measures. *Mathematical Programming, Series B*, 108, 515–540.

Rockafellar, R.T., Uryasev, S. and Zabarankin, M. (2008). Risk tuning with generalized linear regression. *Mathematics of Operations Research*, 33, 712–729.

Rousseeuw, P.J. and Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12, 29–45.

Rousseeuw, P. and Leroy, A. (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons.

van der Waerden, B. (1957). *Mathematische Statistik*. Berlin: Springer-Verlag.

Venables, W. and Ripley, B. (2002). *Modern applied statistics with S-PLUS*, 4th ed. New York: Springer.