Chapter 1

Introduction and historic overview

1.1 Classical regression

In a classical regression problem, one deals with a functional relation $y = g(x)$ between two variables, x and y. As an archetype example, let x represent time and $y = g(x)$ a certain quantity observed at time x (say, the outside temperature or the stock market index), then one would like to model the evolution of g.

One records a number of observations $(x_1, y_1), \dots, (x_n, y_n)$ and tries to approximate them by a relatively simple model function, such as linear $y = a + bx$ or quadratic $y = a + bx + cx^2$ or exponential $y = ae^{bx}$, etc., where $a, b, c, \dots$ are the respective coefficients (or parameters of the model).

Generally, let us denote the model function by $y = g(x; \Theta)$, where $\Theta = (a, b, \dots)$ is the vector of relevant parameters. The goal is to find a particular function $g(x; \hat{\Theta})$ in that class (i.e., choose a particular value $\hat{\Theta}$ of Θ) that approximates (fits) the observed data $(x_1, y_1), \dots, (x_n, y_n)$ best. It is not necessary to achieve the exact relations $y_i = g(x_i; \hat{\Theta})$ for all (or any) i, because y_i's are regarded as imprecise (or noisy) observations of the functional values.

A standard assumption in statistics is that y_i's are small random perturbations of the true values $\tilde{y}_i = g(x_i;\tilde{\Theta})$, i.e.

$$y_i = g(x_i;\tilde{\Theta}) + \varepsilon_i, \qquad i = 1,\ldots,n$$

where $\tilde{\Theta}$ stands for the true (but unknown) value of Θ, and (small) errors ε_i are independent normally distributed random variables with zero mean and, in the simplest case, common variance σ^2. Then the joint probability density function is

$$f(y_1,\ldots,y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - g(x_i;\Theta)\big)^2\Big],$$

so the log-likelihood function is

$$\log L(\Theta,\sigma^2) = -\ln(2\pi\sigma^2)^{n/2} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big[y_i - g(x_i;\Theta)\big]^2. \qquad (1.1)$$

Thus the maximum likelihood estimate $\hat{\Theta}$ of Θ is obtained by minimizing the sum of squares

$$\mathscr{F}(\Theta) = \sum_{i=1}^{n}\big[y_i - g(x_i;\Theta)\big]^2, \qquad (1.2)$$

which leads us to the classical least squares. This method for solving regression problems goes back to C.-F. Gauss [69] and A.-M. Legendre [121] in the early 1800s. It is now a part of every standard undergraduate statistics course.

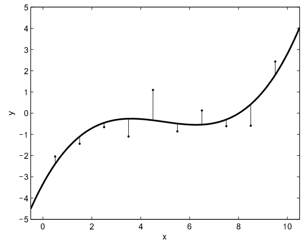

Figure 1.1 *Ordinary regression minimizes the sum of squares of vertical distances: a cubic polynomial fitted to 10 data points.*

We emphasize that the x and y variables play different roles: x is called a *control* variable (controlled by the experimenter), its values $x_1,\ldots,x_n$ are error-free, and y is called a *response* variable (observed as a response), its values $y_1,\ldots,y_n$ are imprecise (contaminated by noise). Geometrically, the regression procedure minimizes the sum of squares of *vertical* distances (measured along the y axis) from the data points (x_i,y_i) to the graph of the function $y = g(x;\Theta)$, see Fig. 1.1.

For example, if one deals with a linear relation $y = a + bx$, then the least squares estimates $\hat{a}$ and $\hat{b}$ minimize the function

$$\mathscr{F}(a,b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Solving equations $\partial\mathscr{F}/\partial a = 0$ and $\partial\mathscr{F}/\partial b = 0$ gives

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \qquad \text{and} \qquad \hat{b} = s_{xy}/s_{xx}, \tag{1.3}$$

where $\bar{x}$ and $\bar{y}$ are the "sample means"

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.4}$$

and

$$\begin{aligned} s_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ s_{yy} &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ s_{xy} &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \end{aligned}$$

are the components of the so called "scatter matrix"

$$\mathbf{S} = \begin{bmatrix} s_{xx} & s_{xy} \\ s_{xy} & s_{yy} \end{bmatrix}, \tag{1.5}$$

which characterizes the "spread" of the data set about its centroid $(\bar{x}, \bar{y})$.

Remark. To estimate a and b, one does not need to know the variance σ^2. It can be estimated separately by maximizing the log-likelihood function (1.1) with respect to σ^2, which gives

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2. \tag{1.6}$$

This estimate is slightly biased, as $\mathbb{E}(\hat{\sigma}^2) = \frac{n-2}{n}\,\sigma^2$. It is customary to replace n in the denominator with $n-2$, which gives an unbiased estimate of σ^2. Both versions of $\hat{\sigma}^2$ are strongly consistent, i.e., they converge to σ^2 with probability one.

The regression model has excellent statistical properties. The estimates $\hat{a}$ and $\hat{b}$ are strongly consistent, i.e., $\hat{a} \to a$ and $\hat{b} \to b$ as $n \to \infty$ (with probability one), and unbiased, i.e., $\mathbb{E}(\hat{a}) = a$ and $\mathbb{E}(\hat{b}) = b$. They have normal distributions with variances

$$\sigma_a^2 = \sigma^2 \left(\frac{\bar{x}^2}{s_{xx}} + \frac{1}{n} \right), \qquad \sigma_b^2 = \frac{\sigma^2}{s_{xx}}.$$

These variances are the smallest among the variances of unbiased estimators, i.e., they coincide with the Cramer-Rao lower bounds. Hence the estimates $\hat{a}$ and $\hat{b}$ are 100% efficient. All around, they are statistically optimal in every sense.

Remark. Suppose the errors ε_i are *heteroscedastic*, i.e., have different variances: $\varepsilon_i \sim N(0, \sigma_i^2)$. The maximum likelihood estimate of Θ is now obtained by the weighted least squares:

$$\mathscr{F}(\Theta) = \sum_{i=1}^{n} w_i \big[y_i - g(x_i; \Theta) \big]^2,$$

where the weights are set by $w_i = \sigma_i^{-2}$. In the linear case, $y = a + bx$, the estimates are still given by (1.3), but now the formulas for the sample mean and the scatter matrix should incorporate weights, e.g.,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} w_i x_i, \qquad s_{xx} = \sum_{i=1}^{n} w_i (x_i - \bar{x})^2, \quad \text{etc.} \tag{1.7}$$

Thus, heteroscedasticity only requires minor changes in the regression formulas.

1.2 Errors-in-variables (EIV) model

Recall that the classical regression problem was solved in the early 1800s. In the late nineteenth century statisticians encountered another problem, which looked very similar, but turned out to be substantially different and far more difficult. In fact the superficial similarity between the two caused a great deal of confusion and delayed the progress for several decades.

That new problem is reconstructing a functional relation $y = g(x)$ given observations $(x_1, y_1), \dots, (x_n, y_n)$ in which *both* variables are subject to errors. We start with an example and describe a formal statistical model later.

Suppose (see Madansky [127]) we wish to determine ρ, the density of iron, by making use of the relation

$$\text{MASS} = \rho \times \text{VOLUME}. \tag{1.8}$$

We can pick n pieces of iron and measure their volumes $x_1, \dots, x_n$ and masses $y_1, \dots, y_n$. Given these data, we need to estimate the coefficient ρ in the functional relation $y = \rho x$. We cannot use the exact formula $y_i = \rho x_i$ for any i, because the measurements may be imprecise (our pieces of iron may be contaminated by other elements).

Similar problems commonly occur in economics (where, for instance, x may be the price of a certain good and y the demand, see Wald [187]) and in sociology. For a fascinating collection of other examples, including the studies of A-bomb survivors, see Chapter 1 in [28].

So how do we solve the iron density problem? For example, we can assume (or rather, pretend) that the volumes x_i's are measured precisely and apply the classical regression of y on x, i.e., determine $y = bx$ and set $\rho = b$. Alternatively, we can assume that our masses y_i's are error-free and do the regression of x on y, i.e., find $x = b'y$ and then set $\rho = 1/b'$.

This may sound like a good plan, but it gives us two different estimates, $\rho_1 = b$ and $\rho_2 = 1/b'$, which should make us at least suspicious. An objection was raised against this strategy as early as in 1901 by K. Pearson, see p. 559 in [144]: "we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable." See Fig. 1.2.

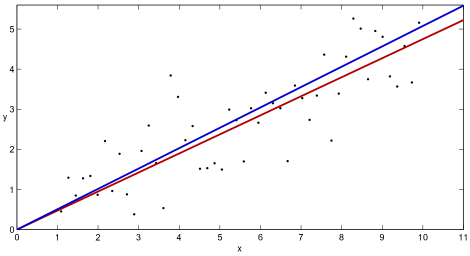

Figure 1.2 *50 data points (marked by dots) are fitted by two methods: the regression of y on x is the lower line and the regression of x on y is the upper line. Their slopes are* 0.494 *and* 0.508*, respectively.*

It was later determined that under natural statistical assumptions (to be described shortly) both estimates, ρ_1 and ρ_2, are inconsistent and may be heavily biased, see e.g., [8, 118, 142]; the consequences of this biasedness in econometrics are discussed in Chapter 10 of [128]. In fact, ρ_1 systematically underestimates the true density ρ, and ρ_2 systematically overestimates it.

Thus the new type of regression problem calls for nonclassical approaches. First we need to adopt an appropriate statistical model in which both x_i's and y_i's are subject to errors; it is called *errors-in-variables* (EIV) model[1]. It assumes that there are some 'true' values $\tilde{x}_i$ and $\tilde{y}_i$, that are linked by the (unknown) functional relation $\tilde{y}_i = g(\tilde{x}_i)$, and the experimenters observe their per-

[1] Another popular name is *measurement error* (ME) model, but we prefer EIV.

turbed values:

$$x_i = \tilde{x}_i + \delta_i, \qquad y_i = \tilde{y}_i + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{1.9}$$

Here $\delta_1, \ldots, \delta_n, \varepsilon_1, \ldots, \varepsilon_n$ are $2n$ independent random variables with zero mean.

In the simplest case, one can assume that δ_i's have a common variance σ_x^2 and ε_i's have a common variance σ_y^2. Furthermore, it is common to assume that δ_i and ε_i are normally distributed, i.e.

$$\delta_i \sim N(0, \sigma_x^2) \qquad \text{and} \qquad \varepsilon_i \sim N(0, \sigma_y^2). \tag{1.10}$$

We also need to make some assumptions about the true values $\tilde{x}_i$'s and $\tilde{y}_i$'s, as they are neither random observations nor the model parameters (yet). There are two basic ways of treating these 'intermediate' objects.

First, the true values $\tilde{x}_i$'s and $\tilde{y}_i$'s may be regarded as fixed (nonrandom), then they have to be treated as additional parameters. They are sometimes referred to as "incidental" or "latent" parameters, or even "nuisance" parameters (as their values are normally of little interest). This interpretation of $\tilde{x}_i$'s and $\tilde{y}_i$'s is known as the *functional model*.

Alternatively, one can regard $\tilde{x}_i$'s and $\tilde{y}_i$'s as realizations of some underlying random variables that have their own distribution. It is common to assume that $\tilde{x}_i$'s are sampled from a normal population $N(\mu, \sigma^2)$, and then $\tilde{y}_i$'s are computed by $\tilde{y}_i = g(\tilde{x}_i)$. In that case δ_i and ε_i's are usually assumed to be independent of $\tilde{x}_i$'s and $\tilde{y}_i$'s. The mean μ and variance σ^2 of the normal population of $\tilde{x}_i$'s can be then estimated along with the parameters of the unknown function $g(x)$. This treatment of the true values is known as the *structural model*.

This terminology is not quite intuitive, but it is currently adopted in the statistics literature. It goes back to Kendall's works [109, 110] in the 1950s and became popular after the first publication of Kendall and Stuart's book [111]. Fuller [66] suggests a simple way of remembering it: the model is Functional (F) if the true points are Fixed; and the model is Structural (S) if the true points are Stochastic.

Before we turn to the solution of the EIV regression problem (which is typified by the iron density example), we describe a special version of the EIV model, which constitutes the main subject of this book.

1.3 Geometric fit

In the late 1800s statisticians encountered a special case of the EIV regression that arose in the analysis of images (photographs, drawings, maps). For example, given an imperfect line on an image, one wants to straighten it up, i.e., find an ideal line approximating the visible line contour. To this end, one can mark several points on the contour and try to fit a perfect straight line to the marked points.

More generally, one may want to approximate a round object on an image by a perfect circle, or an oval by a perfect ellipse, or a box by a perfect rectangle, etc. We call this task *geometric fitting problem*. It consists of approximating a visible contour on an image by a simple geometric figure (line, curve, polygon, etc). We discuss approximation by lines in this section.

In a coordinate system, the given points on the visible contour can be recorded as $(x_1,y_1),\dots,(x_n,y_n)$, and one looks for the best fitting line in the form $y = a+bx$. Hence again the problem looks like a familiar regression. But a close look reveals that both x_i's and y_i's may be imprecise, hence we are in the framework of the EIV model.

Furthermore, there is a novel feature here: due to the geometric character of the problem, the errors in x and y directions should have the same magnitude, on average, hence we have a special case of the EIV model characterized by

$$\sigma_x^2 = \sigma_y^2. \tag{1.11}$$

In this case the "noise" vector $(\delta_i, \varepsilon_i)$ has a normal distribution with zero mean and a scalar covariance matrix, i.e., the random noise is *isotropic* in the *xy* plane. The isotropy means that the distribution of the noise vector is invariant under rotations. This property is natural in image processing applications, as the choice of coordinate axes on the image is often arbitrary, i.e., there should not be any differences between the x, or y, or any other directions.

Conversely, suppose that the random vector $(\delta_i, \varepsilon_i)$ has two basic properties (which naturally hold in image processing applications):

(a) it is isotropic, as described above,

(b) its components δ_i and ε_i are independent.

Then it necessarily has a normal distribution. This is a standard fact in probability theory, see e.g., [14] or Section III.4 of [60]. Thus the assumption about normal distribution (1.10) is not a luxury anymore, but a logical consequence of the more basic assumptions (a) and (b).

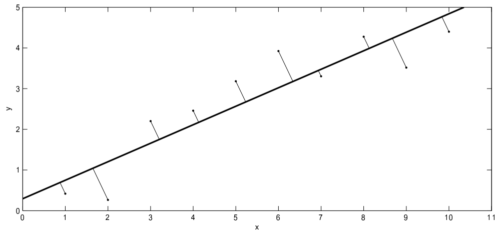

Figure 1.3 *Orthogonal regression minimizes the sum of squares of orthogonal distances.*

A practical solution to the special case $\sigma_x^2 = \sigma_y^2$ of the EIV model was proposed as early as in 1877 by Adcock [1] based on purely geometric (rather than statistical) considerations. He defines the fitting line $y = a + bx$ that is overall closest to the data points, i.e., the one which minimizes

$$\mathscr{F} = \sum_{i=1}^{n} d_i^2, \tag{1.12}$$

where d_i denotes the geometric (orthogonal) distance from the point (x_i, y_i) to the fitting line, see Fig. 1.3. By using elementary geometry, we obtain

$$\mathscr{F}(a,b) = \frac{1}{1+b^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2. \tag{1.13}$$

Solving the equation $\partial\mathscr{F}/\partial a = 0$ yields

$$a = \bar{y} - b\bar{x}, \tag{1.14}$$

where $\bar{x}$ and $\bar{y}$ are the sample means, cf. Section 1.1. By the way, recall that (1.14) also holds in the classical case, cf. (1.3). Now eliminating a from (1.13) gives us a function of one variable

$$\mathscr{F}(b) = \frac{s_{yy} - 2bs_{xy} + s_{xx}b^2}{1+b^2},$$

where s_{xx}, s_{xy}, s_{yy} are the components of the scatter matrix, cf. Section 1.1. Next, the equation $\partial\mathscr{F}/\partial b = 0$ reduces the problem to a quadratic equation,

$$s_{xy}b^2 - (s_{yy} - s_{xx})b - s_{xy} = 0. \tag{1.15}$$

It has two roots, but a careful examination reveals that the minimum of $\mathscr{F}$ corresponds to the following one:

$$b = \frac{s_{yy} - s_{xx} + \sqrt{(s_{yy} - s_{xx})^2 + 4s_{xy}^2}}{2s_{xy}}. \tag{1.16}$$

This formula applies whenever $s_{xy} \neq 0$. In the case $s_{xy} = 0$, we need to set $b = 0$ if $s_{xx} > s_{yy}$ and $b = \infty$ if $s_{xx} < s_{yy}$. We encourage the reader to derive the formula (1.16) and carefully examine the special case $s_{xy} = 0$.

The above solution may be elementary, by our modern standards, but it has a history showing its nontrivial character. It was first obtained in 1878 by Adcock [2], who incidentally made a simple calculational error. Adcock's error was corrected the next year by Kummell [117], but in turn, one of Kummell's formulas involved a more subtle error. Kummell's error was copied by some other authors in the 1940s and 1950s (see [89, 126]). Finally it was corrected in 1959 by Madansky [127]. Madansky's work [127] is perhaps the most cited in the early studies on the EIV regression.

We call the fitting method based on minimization of the sum of squares of orthogonal (geometric) distances from the data points to the fitted contour *orthogonal fit* or *geometric fit*. Despite the natural appeal of the orthogonal fitting line, the early publications [1, 2, 117] in 1877–79 passed unnoticed. Twenty years later the orthogonal fitting line was independently proposed by Pearson [144], and another 20 years later, by Gini [72].

Pearson and Gini made another important observation: the line which minimizes (1.12) is the major axis of the scattering ellipse associated with the data set. The scattering ellipse is defined by equation

$$\begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix}^T \mathbf{S} \begin{bmatrix} x-\bar{x} \\ y-\bar{y} \end{bmatrix} = 1,$$

its center is $(\bar{x},\bar{y})$ and its axes are spanned by the eigenvectors of the scatter matrix $\mathbf{S}$. This fact establishes a link between the orthogonal fit and the principal component analysis of linear algebra.

Pearson [144] also estimated the angle $\theta = \tan^{-1} b$ which the fitting line made with the x axis and found a simple formula for it:

$$\tan 2\theta = \frac{2s_{xy}}{s_{xx} - s_{yy}}. \tag{1.17}$$

We leave its verification to the reader as an exercise.

Adcock and Pearson were motivated by geometric considerations and did not use probabilities. Only in the 1930s their method was incorporated into the formal statistical analysis. Koopmans [113] (see also Lindley [126]) determined that the orthogonal fit provided the maximum likelihood estimate under the assumptions (1.9)–(1.11). Recall that the classical least squares fit (1.2) also maximizes the likelihood in the ordinary regression model (1.1). Thus there is a deep analogy between the two regression models.

The geometric nature of the orthogonal fit makes the resulting line independent of the choice of the coordinate system on the image. In other words, the geometric fit is invariant under orthogonal transformations (rotations and translations) of the coordinate frame.

The invariance under certain transformations is very important. We say that a fitting line is invariant under translations if changing the data coordinates by

$$T_{c,d}\colon\ (x,y) \mapsto (x+c, y+d) \tag{1.18}$$

will leave the line unchanged, i.e., its equation in the new coordinate system will be $y+d = a+b(x+c)$. Similarly we define invariance under rotations

$$R_\theta\colon\ (x,y) \mapsto (x\cos\theta + y\sin\theta, -x\sin\theta + y\cos\theta) \tag{1.19}$$

and under scaling of variables

$$S_{\alpha,\beta}\colon\ (x,y) \mapsto (\alpha x, \beta y). \tag{1.20}$$

An important special case of a scaling transformation is $\alpha = \beta$; it is called a similarity (or sometimes a dilation; in formal mathematics it is known as a homothety). We will denote it by

$$S_\alpha = S_{\alpha,\alpha}\colon (x,y) \mapsto (\alpha x, \alpha y). \tag{1.21}$$

It takes little effort to verify that the orthogonal fitting line is invariant under $T_{c,d}$ and R_θ, as well as S_α, but *not* invariant under general scaling transformations $S_{\alpha,\beta}$ with $\alpha \neq \beta$. We leave the verification of these facts to the reader.

The orthogonal fit has a clear appeal when applied to regular geometric patterns. Fig. 1.4 shows four data points placed at vertices of a rectangle. While classical regression lines are skewed upward or downward (the first and second panels of Fig. 1.4), the orthogonal regression line cuts right through the middle of the rectangle and lies on its axis of symmetry. Arguably, the orthogonal fitting line would “please the eye” more than any other line.

However, the orthogonal fit leads to an inconsistency if one applies it to a more general EIV model, where $\sigma_x^2 \neq \sigma_y^2$. This inconsistency stems from the noninvariance of the orthogonal fitting line under scaling transformations $S_{\alpha,\beta}$.

For example, let us again consider the task of determining the iron density by using (1.8) and measuring volumes x_i's and masses y_i's of some iron pieces, cf. the previous section. If we employ the orthogonal fit to the measurements $(x_1,y_1),\dots,(x_n,y_n)$, then the fitting line $y = bx$, and the resulting estimate of the iron density $\rho = b$, would depend on the choice of units in which the measurements x_i's and y_i's are recorded. That is, if we rescale the variables by $(x,y) \mapsto (\alpha x, \beta y)$, the equation of the orthogonal fitting line in the new coordinate system would be $\beta y = b'(\alpha x)$, where $b' \neq b$. In other words, a different density would be obtained if we change pounds to kilograms or tons, and similarly liters to bushels or cubic meters.

This objection was raised in 1937 by Roos [156] and further discussed in the statistics literature in the 1940s [89, 187]. Thus the orthogonal fit has its limitations, it is essentially restricted to the special case $\sigma_x^2 = \sigma_y^2$ of the EIV model. Some modern books, see e.g., [28], strongly warn against the use of orthogonal fitting line in EIV applications with $\sigma_x^2 \neq \sigma_y^2$, and more generally, against the use of other techniques that are based on any unreliable assumptions about σ_x^2 and σ_y^2.

We briefly overview basic features of the general EIV model in the next section (though not attempting anything close to a comprehensive coverage).

1.4 Solving a general EIV problem

Let us turn back to the EIV model (1.9)–(1.10) without assuming (1.11), i.e., leaving σ_x^2 and σ_y^2 unconstrained.

Kummell [117] was perhaps the first who examined, in 1879, the task of

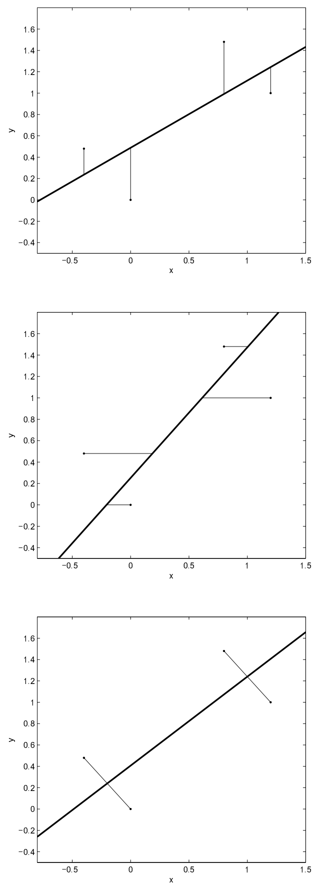

Figure 1.4 *The regression of y on x minimizes the sum of squares of vertical distances (top); the regression of x on y does the same with horizontal distances (middle); the orthogonal regression minimizes the sum of squares of orthogonal distances (bottom).*

determining the underlying functional relation $y = g(x)$ in the EIV context, and realized that this could not be done in any reasonable sense (!), unless one makes an extra assumption on the relation between σ_x^2 and σ_y^2. Even in the simplest, linear case $y = a + bx$, there is no sensible way to estimate the parameters a and b without extra assumptions. The problem is just *unsolvable*, however simple it may appear!

Many other researchers arrived at the same conclusion in the early twentieth century. The realization of this stunning fact produced a long turmoil in the community lasting until about the 1950s and marked by confusion and controversy. A. Madansky, for example, devotes a few pages of his 1959 paper [127] describing the shock of an average physicist who would learn about the unsolvability of the "simple" regression problem, and how statisticians could explain it to him.

Later the insolvability of this problem was proved in mathematical terms. First, it was established in 1956 by Anderson and Rubin [9] (see also [73]) that even in the linear case $y = a + bx$ the likelihood function was unbounded (its supremum was infinite), thus maximum likelihood estimates could not be determined. Interestingly, the likelihood function has critical points, which have been occasionally mistaken for maxima; only in 1969 the issue was resolved: M. Solari [168] proved that all critical points were just saddle points.

Second (and more importantly), it was shown in 1977 by Nussbaum [139] (see also page 7 in [40]) that no statistical procedure could produce strongly consistent estimates $\hat{a}$ and $\hat{b}$ (which would converge to the true values of a and b as $n \to \infty$). See also the discussion of identifiability in the book [40] by Cheng and Van Ness.

To make the EIV regression model solvable, Kummel [117] assumed that

$$\text{the ratio} \qquad \kappa = \sigma_x/\sigma_y \qquad \text{is known.} \tag{1.22}$$

He justified his assumption by arguing that experimenters "usually know this ratio from experience." Later this assumption was commonly adopted in the statistics literature. Recently Fuller [66] called the EIV model satisfying the assumptions (1.9), (1.10), and (1.22) the "classical EIV model."

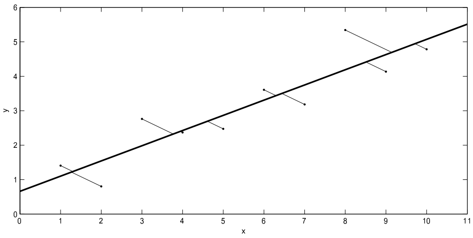

Figure 1.5 *The EIV fit minimizes the sum of squares of "skewed" distances from the data points to the line. Here $\kappa = 2$.*

Now the EIV regression problem has a well defined solution. In 1879 Kummell [117] gave formulas for the best fitting line that involved κ. His line $y = a + bx$ minimizes

$$\mathscr{F} = \frac{1}{1 + \kappa^2 b^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{1.23}$$

and its slope is estimated by

$$b = \frac{\kappa^2 s_{yy} - s_{xx} + \sqrt{(\kappa^2 s_{yy} - s_{xx})^2 + 4\kappa^2 s_{xy}^2}}{2\kappa^2 s_{xy}}, \tag{1.24}$$

compare this to (1.16). The intercept is again $a = \bar{y} - b\bar{x}$, as in (1.14).

This line minimizes the sum of squares of the distances to the data points (x_i, y_i) measured along the vector $(\kappa b, -1)$, see Fig. 1.5. Kummell arrived at his formula rather intuitively, but later it was determined that he actually found the maximum likelihood solution, cf. [113, 126].

In the special case $\kappa = 1$, i.e., $\sigma_x^2 = \sigma_y^2$, the vector $(\kappa b, -1) = (b, -1)$ is normal to the line $y = a + bx$, thus we arrive at the familiar orthogonal fit. Hence, the EIV linear regression (1.24) includes the orthogonal fit as a particular case.

The slope b given by (1.24) is monotonically increasing with κ (this follows from the standard fact $s_{xy}^2 \leq s_{xx}s_{yy}$ by some algebraic manipulations, which we leave to the reader as an exercise). In the limit $\kappa \to 0$, the EIV regression line converges to the classical regression of y on x with the slope $b = s_{xy}/s_{xx}$, cf. (1.3). Similarly, in the limit $\kappa \to \infty$, the EIV regression line converges to the classical regression of x on y with the slope $b = s_{yy}/s_{xy}$. Thus the classical regressions (of y on x and of x on y) are the extreme cases of the EIV regression.

The EIV line minimizing (1.23) can be made invariant under rescaling of coordinates $x \mapsto \alpha x$ and $y \mapsto \beta y$, as the scaling factors α and β can be incorporated into the ratio κ by the obvious rule $\kappa \mapsto \kappa\alpha/\beta$. This fact was pointed out in 1947 by Lindley [126], who concluded that the estimate (1.24) thus conformed to the basic requirement of the EIV model: it does not depend on the units in which the measurements $x_1, \dots, x_n$ and $y_1, \dots, y_n$ are recorded.

Actually, if one rescales the coordinates by $x \mapsto x$ and $y \mapsto \kappa y$, then in the new variables we have $\kappa = 1$. Thus the EIV regression line can be transformed to the orthogonal fitting line. Therefore, the EIV linear regression model (with the known ratio $\kappa = \sigma_x/\sigma_y$) can be converted to the orthogonal regression model by a simple rescaling of the coordinates, and vice versa.

But we emphasize that the general EIV regression and the orthogonal fit must conform to different requirements:

- The EIV regression must be invariant under scaling of the variables x and y (resulting from the change of units in which these variables are measured).
- The orthogonal fit (due to its geometric nature) must be invariant under rotations and translations of the coordinate frame on the xy plane, as well as under similarities resulting from a change of unit of length.

This difference has important consequences for nonlinear regression discussed in Section 1.9.

1.5 Nonlinear nature of the "linear" EIV

It may be enlightening to interpret the orthogonal regression problem geometrically in the space $\mathbb{R}^{2n}$ with coordinates $x_1, y_1, \ldots, x_n, y_n$. We follow Malinvaud (Chapter 10 of [128]). Our observations $(x_1, y_1), \ldots, (x_n, y_n)$ are represented by one point (we denote it by $\mathscr{X}$) in this multidimensional space. To understand the construction of the orthogonal fitting line, consider the subset $\mathbb{P} \subset \mathbb{R}^{2n}$ defined by

$$(x_1, y_1, \ldots, x_n, y_n) \in \mathbb{P} \iff \exists a, b \colon y_i = a + bx_i \;\; \forall i,$$

i.e., $\mathbb{P}$ consists of all $(x_1, y_1, \ldots, x_n, y_n) \in \mathbb{R}^{2n}$ such that the n planar points $(x_1, y_1), \ldots, (x_n, y_n)$ are collinear. Note that the true values $\tilde{x}_1, \tilde{y}_1, \ldots, \tilde{x}_n, \tilde{y}_n$ are represented by one point (we denote it by $\tilde{\mathscr{X}}$) in $\mathbb{P}$, i.e., $\tilde{\mathscr{X}} \in \mathbb{P}$.

The orthogonal fitting line minimizes the sum

$$\sum_{i=1}^{n} (x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2,$$

which is the square of the distance (in the Euclidean metric) between the points $\mathscr{X}$ and $\tilde{\mathscr{X}}$. Thus, the orthogonal fitting procedure corresponds to choosing a point $\tilde{\mathscr{X}} \in \mathbb{P}$ closest to the point $\mathscr{X} \in \mathbb{R}^{2n}$ representing the data. In other words, we simply project the given point $\mathscr{X}$ onto $\mathbb{P}$ orthogonally. Or is it that simple?

It takes a little effort to verify that $\mathbb{P}$ is a *nonlinear* submanifold ('surface') in $\mathbb{R}^{2n}$. Indeed, it is specified by $n-2$ independent relations

$$\frac{y_{i+1} - y_i}{x_{i+1} - x_i} = \frac{y_{i+2} - y_i}{x_{i+2} - x_i}, \qquad i = 1, \ldots, n-2, \tag{1.25}$$

each of which means the collinearity of the three planar points (x_i, y_i), (x_{i+1}, y_{i+1}), and (x_{i+2}, y_{i+2}). The relations (1.25) are obviously quadratic, hence $\mathbb{P}$ is an $(n+2)$-dimensional *quadratic* surface (variety) in $\mathbb{R}^{2n}$.

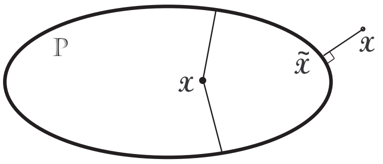

Figure 1.6: *Projection of the point $\mathscr{X}$ onto the quadratic manifold $\mathbb{P}$.*

Projecting a point $\mathscr{X}$ onto a quadratic surface is not a trivial, and definitely not a linear, problem. This geometric interpretation should dispel our illusion (if we still have any) that we deal with a linear problem, it unmasks its truly nonlinear character.

Imagine, for example, the task of projecting a point $\mathscr{X} \in \mathbb{R}^2$ onto a quadric, say an ellipse $a^2x^2 + b^2y^2 = 1$. This is not a simple problem, its exact solution involves finding roots of a 4th degree polynomial [162]. In a sense, we are lucky that the projection of our data point $\mathscr{X} \in \mathbb{R}^{2n}$ onto $\mathbb{P}$ reduces to just a quadratic equation (1.15).

Besides, the projection may not be unique (for example when $\mathscr{X}$ lies on the major axis of the ellipse near the center, see Fig. 1.6). We will actually see that the orthogonal fitting line may not be unique either, cf. Section 2.3.

To further emphasize the nonlinear nature of the EIV regression, suppose for a moment that the errors are heteroscedastic, i.e.

$$\delta_i \sim N(0, \sigma_{x,i}^2) \qquad \text{and} \qquad \varepsilon_i \sim N(0, \sigma_{y,i}^2),$$

where the ratio of variances is known, but it differs from point to point, i.e., we assume that

$$\kappa_i = \sigma_{x,i}/\sigma_{y,i}$$

is known for every $i = 1, \dots, n$. Recall that in the classical regression the heteroscedasticity of errors does not affect the linear nature of the problem. Now, in the EIV model, the best fitting line should minimize

$$\mathscr{F}(a,b) = \sum_{i=1}^{n} \frac{(y_i - a - bx_i)^2}{1 + \kappa_i^2 b^2}.$$

Despite its resemblance to (1.23), the minimization of this $\mathscr{F}$ cannot be reduced to a quadratic (or any finite degree) polynomial equation. Here “finite degree” means a degree independent of the sample size n. This is a hard-core nonlinear problem that has no closed form solution; its numerical solution requires iterative algorithms.

In other words, the hidden nonlinear nature of the “linear” EIV fit may come in different ways at different stages. The more general assumptions on errors one makes the more serious difficulties one faces.

Yet another explanation why the linear EIV regression has an essentially nonlinear character was given by Boggs et al., see [23].

Overall, the linear EIV model, though superficially resembling the classical linear regression, turns out to be dissimilar in many crucial ways. The sharp contrast between these two models is now recognized by many authors. As the textbook [29] puts it, “Regression with errors in variables (EIV) ... is so fundamentally different from the simple linear regression ... that it is probably best thought of as a completely different topic.”

1.6 Statistical properties of the orthogonal fit

Our book is devoted to the orthogonal fitting problem, and from now on we adopt the statistical model assumptions (1.9), (1.10), and (1.11). Under these assumptions the orthogonal fit maximizes the likelihood function, i.e., provides the Maximum Likelihood Estimate (a formal proof of this fact will be given in Section 6.3).

In this section we touch upon the basic statistical properties of the linear orthogonal fit, i.e., the behavior of the estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters of the fitting line $y = \alpha + \beta x$ (we use α and β here, instead of the previous a and b, to be consistent with the notation in the papers we will refer to).

Our discussion will shed more light on a stark dissimilarity between the orthogonal fit and the classical regression, whose nice features we mentioned in Section 1.1. Even a quick look reveals a totally different (and somewhat shocking) picture.

To begin with, the distribution of the estimates $\hat{\alpha}$ and $\hat{\beta}$ is not normal and does not belong to any standard family of probability distributions. Only in 1976, explicit formulas for their density functions were found by Anderson and others [7, 10]; see Section 2.4. Those expressions are overly complicated, involve double-infinite series, and Anderson [7] promptly conceded that they are not very useful for practical purposes. Instead, he and Kunitomo [118] derived various approximations to the distribution functions of $\hat{\alpha}$ and $\hat{\beta}$, which turned out to be practically accurate.

Second, and worse, the estimates $\hat{\alpha}$ and $\hat{\beta}$ do not have finite moments, i.e.,

$$\mathbb{E}(|\hat{\alpha}|) = \infty \qquad \text{and} \qquad \mathbb{E}(|\hat{\beta}|) = \infty.$$

Thus they also have infinite mean squared errors:

$$\mathbb{E}\big([\hat{\alpha} - \alpha]^2\big) = \infty \qquad \text{and} \qquad \mathbb{E}\big([\hat{\beta} - \beta]^2\big) = \infty.$$

These stunning facts were also revealed in 1976 by Anderson [7]. Intuitively, one can see why this happens from (1.16), where the denominator can take value zero, and its probability density does not vanish at zero. We encourage the reader to closely examine this observation.

Until Anderson's discovery, researchers "approximated" the moments of the estimates $\hat{\alpha}$ and $\hat{\beta}$ as follows. They employed Taylor expansion, dropped higher order terms, and obtained some "approximate" formulas for the moments of $\hat{\alpha}$ and $\hat{\beta}$ (including their means and variances). Anderson demonstrated that all those formulas were fundamentally flawed, as the actual moments did not exist. Anderson said that those formulas should be regarded as "moments of some approximations," rather than "approximate moments."

Once Anderson made his discovery, it immediately lead to fundamental methodological questions: how can one trust a statistical estimate that has an infinite mean squared error (not to mention infinite bias)? Should these facts imply that the estimate is totally unreliable? Why did not anybody notice these bad features in practice? Can an estimate with infinite moments be practically better than others which have finite moments? These questions lead to further studies, see next.

In the late 1970s, Anderson [7, 8], Patefield [142], and Kunitomo [118] compared the slope $\hat{\beta}$ of the orthogonal fitting line, given by (1.16), with the slope $\hat{\beta}$ of the classical regression line, given by (1.3) (of course, both estimates were used in the framework of the same model (1.9), (1.10), and (1.11)). They denote the former by $\hat{\beta}_{\mathrm{M}}$ (Maximum likelihood) and the latter by $\hat{\beta}_{\mathrm{L}}$ (Least squares). Their results can be summarized in two seemingly conflicting verdicts:

(a) The mean squared error of $\hat{\beta}_{\mathrm{M}}$ is infinite, and that of $\hat{\beta}_{\mathrm{L}}$ is finite (whenever $n \geq 4$), thus $\hat{\beta}_{\mathrm{L}}$ appears (infinitely!) more accurate;

(b) The estimate $\hat{\beta}_{\mathrm{M}}$ is consistent and asymptotically unbiased, while $\hat{\beta}_{\mathrm{L}}$ is inconsistent and asymptotically biased (it is consistent and unbiased only in the special case $\beta = 0$), thus $\hat{\beta}_{\mathrm{M}}$ appears more appropriate.

Going further, Anderson shows that

$$\mathsf{Prob}\bigl\{|\hat{\beta}_{\mathrm{M}} - \beta| > t\bigr\} < \mathsf{Prob}\bigl\{|\hat{\beta}_{\mathrm{L}} - \beta| > t\bigr\}$$

for all $t > 0$ that are not too large, i.e., for all $t > 0$ of practical interest. In other words, the accuracy of $\hat{\beta}_{\mathrm{M}}$ *dominates* that of $\hat{\beta}_{\mathrm{L}}$ everywhere, except for very large deviations (large t). It is the heavy tails of $\hat{\beta}_{\mathrm{M}}$ that make its mean squared error infinite, otherwise it tends to be closer to β than its rival $\hat{\beta}_{\mathrm{L}}$.

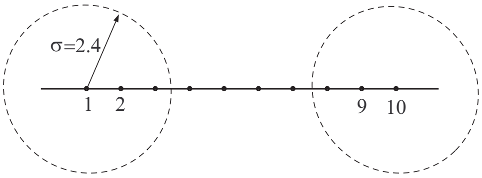

Figure 1.7: *The true points location and the noise level in our experiment.*

Furthermore, when one observes values of $\hat{\beta}_{\mathrm{M}}$ in practice, or in simulated experiments, nothing indicates that $\hat{\beta}_{\mathrm{M}}$ has infinite moments; its values group around a certain center and have a seemingly normal distribution. Large deviations occur so rarely that they usually pass unregistered. However, those large deviations are, ultimately, responsible for the lack of moments. In order to make them visible, i.e., have them appear at a noticeable rate in computer experiments, one needs to increase the noise level $\sigma = \sigma_x = \sigma_y$ way above what it normally is in image processing applications, see next.

For example, we generated 10^6 random samples of $n = 10$ points on the line $y = x$ whose true positions were equally spaced on a stretch of length 10, with $\sigma = 2.4$ (note how high the noise is: its standard deviation is a quarter of the length of the interval where the data are observed; see Fig. 1.7). Fig. 1.8 plots the average estimate $\hat{\beta}_{\mathrm{M}}$ over k samples, as k runs from 1 to 10^6. It behaves very much like the sample mean of the Cauchy random variable (whose moments do not exist either). Thus one can see, indeed, that the estimate $\hat{\beta}_{\mathrm{M}}$ has infinite moments. But if one decreases the noise level to $\sigma = 2$ or less, then the erratic behavior disappears, and the solid line in Fig. 1.8 turns just flat, as it is for the finite moment estimate $\hat{\beta}_{\mathrm{L}}$.

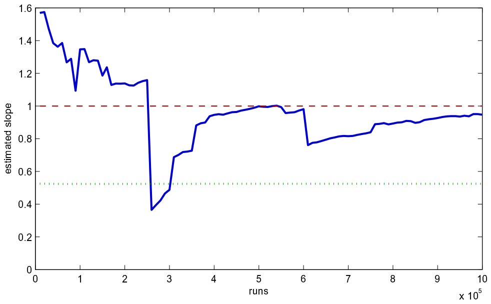

Figure 1.8 *The average estimate $\hat{\beta}_{\mathrm{M}}$ over k randomly generated samples (solid line), as k runs from 1 to 10^6. The true slope $\beta = 1$ is marked by the dashed line. The average estimate $\hat{\beta}_{\mathrm{L}}$ is the dotted line, it remains stable at level 0.52, systematically underestimating β.*

Now which estimate, $\hat{\beta}_{\mathrm{M}}$ or $\hat{\beta}_{\mathrm{L}}$, should we prefer? This may be quite a dilemma for a practitioner who is used to trusting the mean squared error as an absolute and ultimate criterion. Anderson argues that in this situation one has to make an exception and choose $\hat{\beta}_{\mathrm{M}}$ over $\hat{\beta}_{\mathrm{L}}$, despite its infinite mean squared error.

In the early 1980s, as if responding to Anderson's appeal, several statisticians (most notably, Gleser [73, 74, 75], Malinvaud [128], and Patefield [143]) independently established strong asymptotic properties of the orthogonal fitting line (and more generally, the classical EIV fitting line):

(a) the estimates $\hat{\alpha}$ and $\hat{\beta}$ are strongly consistent[2] and asymptotically normal;

[2] However, the maximum likelihood estimates of σ_x^2 and σ_y^2 are not consistent, in fact

$$\hat{\sigma}_x^2 \to \tfrac{1}{2}\,\sigma_x^2 \qquad \text{and} \qquad \hat{\sigma}_y^2 \to \tfrac{1}{2}\,\sigma_y^2$$

as $n \to \infty$, in the functional model. This odd feature was noticed and explained in 1947 by Lindley

(b) in a certain sense, these estimates are efficient.

They also constructed confidence regions for α and β. More details can be found in [66], [40], [128], and our Chapter 2.

These results assert very firmly that the maximum likelihood estimate $\hat{\beta}_{\mathrm{M}}$ is the best possible. Certain formal statements to this extent were published by Gleser [74], see also [39, 40, 128] and our Chapter 2.

After all these magnificent achievements, the studies of the linear EIV regression seem to have subsided in the late 1990s; perhaps the topic exhausted itself. The statistical community turned its attention to nonlinear EIV models.

For further reading on the linear EIV regression, see excellent surveys in [8, 73, 126, 127, 132, 187], and books [66], [40], [128] (Chapter 10), and [111] (Chapter 29). We give a summary of the orthogonal line fitting in Chapter 2.

1.7 Relation to total least squares (TLS)

The EIV linear regression is often associated with the so-called *total least squares* (TLS) techniques in computational linear algebra. The latter solve an overdetermined linear system

$$\mathbf{A}\mathbf{x} \approx \mathbf{b}, \qquad \mathbf{x} \in \mathbb{R}^m, \;\; \mathbf{b} \in \mathbb{R}^n, \;\; n > m, \tag{1.26}$$

where not only the vector **b**, but also the matrix **A** (or at least some of its columns) are assumed to be contaminated by errors. If only **b** is corrupted by noise, the solution of (1.26) is given by the ordinary least squares

$$\mathbf{x} = \operatorname{argmin} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2,$$

where $\|\cdot\|$ denotes the 2-norm. Equivalently, it can be paraphrased by

$$\mathbf{x} = \operatorname{argmin} \|\Delta\mathbf{b}\|^2 \qquad \text{subject to} \qquad \mathbf{A}\mathbf{x} = \mathbf{b} + \Delta\mathbf{b}. \tag{1.27}$$

If **A** has full rank, the (unique) explicit solution is

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

If A is rank deficient, the solution is not unique anymore, and one usually picks the minimum-norm solution

$$\mathbf{x} = \mathbf{A}^{-}\mathbf{b},$$

where $\mathbf{A}^{-}$ denotes the Moore-Penrose pseudoinverse.

If both **b** and **A** are corrupted by noise, the solution of (1.26) is more complicated, and it is the subject of the TLS techniques. In the simplest case, where

[126]; the factor 1/2 here is related to the degrees of freedom: we deal with $2n$ random observations and $n+2$ parameters of the model, thus the correct number of degrees of freedom is $n-2$, rather than $2n-2$.

all errors in **A** and **b** are independent and have the same order of magnitude, the solution is given by

$$\mathbf{x} = \operatorname{argmin} \big\| [\Delta\mathbf{A}\ \Delta\mathbf{b}] \big\|_F^2 \qquad \text{subject to} \qquad (\mathbf{A}+\Delta\mathbf{A})\mathbf{x} = \mathbf{b}+\Delta\mathbf{b}, \qquad (1.28)$$

where $[\Delta\mathbf{A}\ \Delta\mathbf{b}]$ denotes the "augmented" $n \times (m+1)$ matrix and $\|\cdot\|_F$ stands for the Frobenius norm (the "length" of the $[n(m+1)]$-dimensional vector). Note the similarities between (1.27) and (1.28).

To compute **x** from (1.28), one uses the singular values (and vectors) of the augmented matrix $[\mathbf{A}\ \mathbf{b}]$. In the basic case, see Chapter 2 of [185], it is given by

$$\mathbf{x} = (\mathbf{A}^T\mathbf{A} - \sigma_{m+1}^2 \mathbf{I})^{-1}\mathbf{A}^T\mathbf{b},$$

where σ_{m+1} is the smallest singular value of $[\mathbf{A}\ \mathbf{b}]$, and **I** denotes the identity matrix. This is the TLS in the "nutshell;" we refer to [77, 185, 160, 161] for an extensive treatment.

To see how the EIV and TLS models are related, consider an EIV problem of fitting a line $y = a + bx$ to data points (x_i, y_i), $i = 1, \dots, n$. This problem is equivalent to (1.26) with

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} a \\ b \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

We see that the vector **b** and the second column of **A** are corrupted by noise, thus we arrive at a particular TLS problem. If the errors in x_i's and y_i's are independent and have the same variance, then we can solve it by (1.28), and this solution is equivalent to the orthogonal least squares fit.

The link between the EIV regression models and the TLS techniques of computational linear algebra is very helpful. Many efficient tools of the TLS (especially, the SVD) can be employed to solve linear (or nonlinear but linearized) EIV problems, see [185, 160, 161].

1.8 Nonlinear models: General overview

> *...the errors in variables are bad enough in linear models. They are likely to be disastrous to any attempts to estimate additional nonlinearity or curvature parameters...*
>
> Z. Griliches and V. Ringstad; see [79]

Fitting a straight line to observed points may appear as a "linear" regression problem, but it has a truly nonlinear character (Section 1.5). Its solution is given by an irrational formula (1.16), and it may not be unique (Section 2.2). The probability distributions of the resulting estimates do not belong to any standard family and are described by overly complicated expressions (Section 2.4). The estimates do not have moments, i.e., their bias is indeterminate and their mean square errors are infinite. One might just wonder if things could get any worse.

Sadly, things do get worse when one has to fit *nonlinear* functions to data with errors in variables. We only overview some new troubles here. First of all, the nonlinear fitting problem may not even have a solution. More precisely, if one fits a curve of a certain type (say, a circle) by minimizing the orthogonal distances to the data points, then such a curve may not exist; we will see examples in Section 3.3. The nonexistence is a phenomenon specific to nonlinear problems only. Next, even if the best fitting curve exists, it may not be unique, there may be multiple solutions, all of which are "equally good;" see examples in Section 3.5. This leads to confusion in theoretical analysis.

Furthermore, even when the best fit exists and is unique, nothing is known about the distribution of the resulting parameter estimates; there are no explicit formulas for their densities or moments. In fact, theoretical moments quite often fail to exist. This happens even in the linear case, see Section 1.6. For the problem of fitting circles, see Section 6.4. The nonexistence of moments appears to be a common feature of the EIV regression and orthogonal fitting problems. To resolve these difficulties, statisticians have developed a nontraditional error analysis based on approximating distributions. We devote almost the entire Chapter 6 to those new statistical theories.

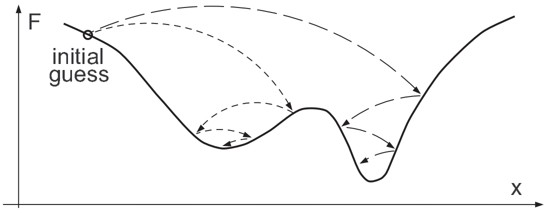

Figure 1.9 *Two algorithms minimizing a function $F(x)$. One makes shorter steps and converges to a local minimum. The other makes longer steps and converges to the global minimum.*

Down to more practical issues, the estimates of parameters in nonlinear EIV regression cannot be found in closed form, by explicit formulas like (1.16). They can only be computed by numerical algorithms, i.e., approximately. Numerical schemes, at best, converge to the desired estimate iteratively. However, in practice, the iterations may very well diverge, and even if they do converge, one never knows if they arrive at the desired estimate (the procedure may just terminate at a local minimum of the objective function; see Fig. 1.9).

Quite often, different numerical algorithms return different estimates. Fig. 1.9 shows an example where an iterative procedure is trapped by a local minimum. Another example is shown in Fig. 1.10: there is no local minima, but the second (slow) algorithm takes a large number of steps to reach the area near the minimum. In computer programs the number of iterations is always limited (usually, the limit is set to 50 or 100), thus the returned estimate may be still far from the actual minimum.

The estimates returned by different algorithms may even have seemingly different statistical characteristics (bias, variance, etc.). Thus the choice of the algorithm becomes a critical factor in practical applications, as well as in many theoretical studies. There is simply little point of studying an abstract "solution" that is not accessible in practice, while practical solutions heavily depend on the particular algorithm used to compute them.

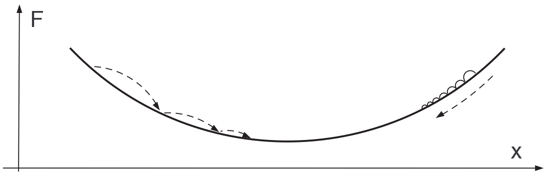

Figure 1.10 *Two algorithms minimizing a function $F(x)$ with a unique minimum. One approaches it fast (from the left) and arrives in a vicinity of the minimum in 5–10 steps. The other moves very slowly (from the right); it may take 100 or 1000 iterations to get near the minimum.*

Thus the analysis of numerical schemes becomes an integral part of the research. Sizable portions of published articles and books are now devoted to computer algorithms, their underlying ideas, performance, limitations, numerical stability, etc. This is all unavoidable, due to the nature of the subject.

1.9 Nonlinear models: EIV versus orthogonal fit

So far we have discussed two large topics—the orthogonal (geometric) fit and the EIV regression (with the known ratio of variances)—in parallel. In the linear context, these models can be transformed to one another by a simple scaling the variables x and y (Section 1.4), and both models have very similar properties.

In the nonlinear context, a strong link between these two models is lost. They can no longer be transformed to one another. The crucial disparity derives from the different requirements stated at the end of Section 1.4: the EIV regression must be invariant under scaling of the variables x and y, and the orthogonal (geometric) fit — under rotations and translations on the xy plane.

These requirements affect the very classes of nonlinear models used in each case.

For example, one may fit polynomials

$$y = a_0 + a_1 x + \cdots + a_k x^k \tag{1.29}$$

to observed points, which is common in the EIV context [79, 140]. Scaling of variables $S_{\alpha,\beta}$, cf. Section 1.3, transforms one polynomial to another, so the class of polynomial remains conveniently invariant.

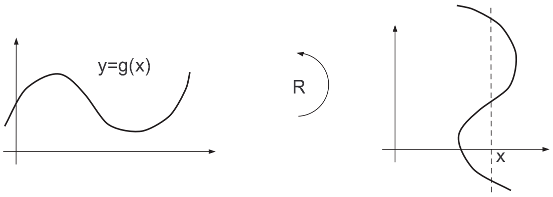

Figure 1.11 *The graph of an explicit nonlinear function $y = g(x)$ (left). After rotation, the same curve (right) does not represent any explicit function.*

However, a rotation R_θ of the coordinate plane transforms a polynomial to a different function; it becomes an implicit polynomial curve, which may not even allow an explicit representation $y = g(x)$. Thus explicit polynomials are *not* suited for orthogonal (geometric) fitting. The same applies to any other class of nonlinear explicit functions $y = g(x)$: the graph of a nonlinear function can always be rotated so that the resulting curve does not represent any explicit functional relation; see Fig. 1.11.

A natural class of models that remain invariant under rotations and translations consists of implicit polynomials of a certain degree $k \geq 1$. Polynomials of degree $k = 1$ are given by equation

$$Ax + By + C = 0, \tag{1.30}$$

which represents all straight lines on the plane (including vertical and horizontal lines). Polynomials of degree $k = 2$ are given by equation

$$Ax^2 + By^2 + Cxy + Dx + Ey + F = 0, \tag{1.31}$$

which represents all conic sections: ellipses, hyperbolas, and parabolas, in addition to straight lines. This class is large enough to cover a vast majority of the existing applications in computer vision and pattern recognition.

Implicit polynomials of higher degree $k \geq 3$ are occasionally used to describe more complex objects, see examples in [150, 176] where polynomials of degree $k = 3, 4$, and even $k = 6$, are mentioned. But the use of polynomials of degree $k \geq 3$ remains extremely rare, as most practitioners prefer to divide complex shapes into small segments that can be well approximated by lines and arcs of conics, or even arcs of circles. What one gets in the end is a sequence of circular arcs stitched together ('circular splines'); see [12, 145, 158, 164, 165]. Some authors plainly assert that "most of the objects in the world are made up of circular arc segments and straight lines;" see [146, 195].

Thus fitting circles and conics to observed data is practically the most important task in image processing applications, besides fitting lines.

We note that often one deals with objects in images that have rectangular or other polygonal shape, see e.g., [186, 189]. In that case a polygon of the right shape can be fit to data. Polygon consists of segments of straight lines, so that general line fitting algorithms can be used, but there are also vertices and corners that may require a special treatment. Such problems are not discussed in this book.

To summarize, we see that in the nonlinear context, the two large topics, (i) the EIV regression used in general statistics and (ii) the geometric fit used in image processing, go separate ways and become very different. There is another significant distinction here: these topics adopt different asymptotic models. In the general EIV regression, it is common to study properties of estimators as the sample size grows, i.e., as $n \to \infty$ (at the same time the noise level σ remains constant). In the image processing applications, the sample size is usually very limited, but the noise is quite small, hence a more appropriate asymptotic model is $\sigma \to 0$ while n is fixed. This issue will be discussed at length in Section 2.5.

We reiterate that our main subject is geometric curve fitting in image processing, i.e., the topic (ii) above. For a comprehensive presentation of the topic (i), i.e., the general nonlinear EIV regression, see a recent book [27] and its second edition [28], updated and expanded.