C. Radhakrishna Rao
Helge Toutenburg
Shalabh
Christian Heumann

Linear Models and Generalizations

Least Squares and Alternatives

With Contributions by Michael Schomaker

With 35 Illustrations
Third Extended Edition

Springer

Contents

$$\mathrm{E}(y) = \beta_0 + \beta_1 X$$

and

$$\mathrm{var}(y) = \sigma^2 \ .$$

Sometimes X can also be a random variable. Such an aspect is explained later in Section 2.15. In such a case, instead of simple mean and simple variance of y, we consider the conditional mean of y given $X = x$ as

$$\mathrm{E}(y|x) = \beta_0 + \beta_1 x$$

and the conditional variance of y given $X = x$ as

$$\mathrm{var}(y|x) = \sigma^2 \ .$$

The parameters β_0, β_1 and σ^2 are generally unknown and e is unobserved. The determination of the statistical model (2.1) depends on the determination (*i.e.*, estimation) of β_0, β_1 and σ^2.

Only T pairs of observations (x_t, y_t) $(t = 1, \ldots, T)$ on (X, y) are observed which are used to determine the unknown parameters.

Different methods of estimation can be used to determine the estimates of the parameters. Among them, the least squares and maximum likelihood principles are the most popular methods of estimation.

2.2 Least Squares Estimation

We observe a sample of T sets of observations (x_t, y_t) $(t = 1, \ldots, T)$ and in view of (2.1), we can write

$$y_t = \beta_0 + \beta_1 x_t + e_t \quad (t = 1, \ldots, T) \ . \tag{2.2}$$

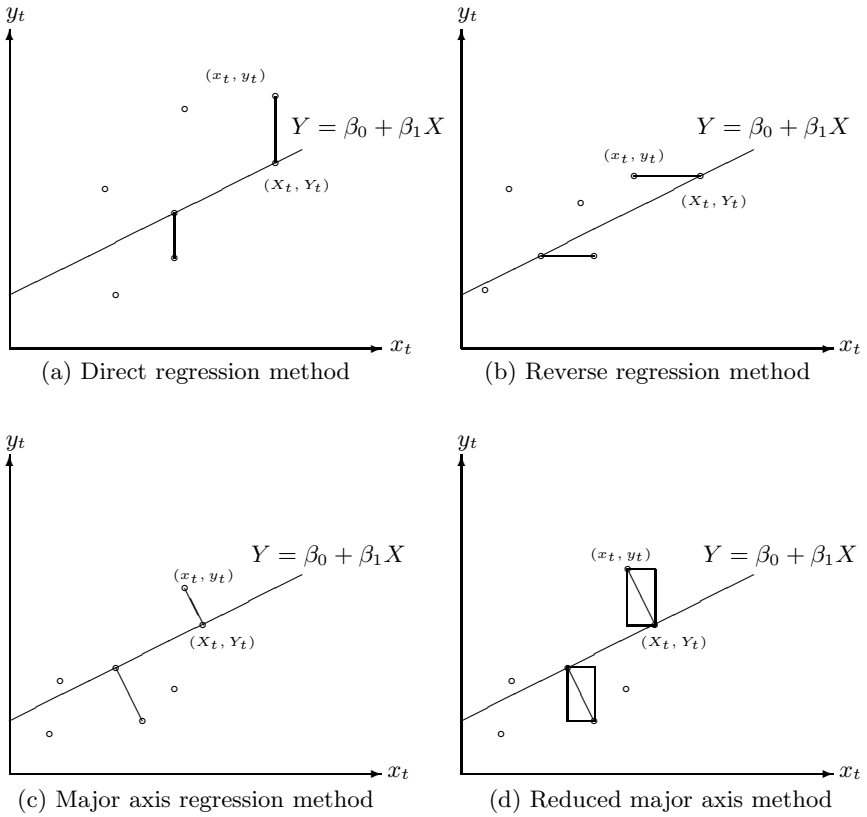
The principle of least squares aims at estimating β_0 and β_1 so that the sum of squares of difference between the observations and the line in the scatter diagram is minimum. Such an idea is viewed from different perspectives. When the vertical difference between the observations and the line in the scatter diagram (see Fig. 2.1(a)) is considered and its sum of squares is minimized to obtain the estimates of β_0 and β_1, the method is known as *direct regression*.

Another approach is to minimize the sum of squares of difference between the observations and the line in horizontal direction in the scatter diagram (see Fig. 2.1(b)) to obtain the estimates of β_0 and β_1. This is known as *reverse (or inverse) regression* method.

Alternatively, the sum of squares of perpendicular distance between the observations and the line in the scatter diagram (see Fig. 2.1(c)) is minimized to obtain the estimates of β_0 and β_1. This is known as *orthogonal regression* or *major axis regression* method.

The *least absolute deviation regression* method considers the sum of the absolute deviation of the observations from the line in the vertical direction in the scatter diagram (see Fig. 2.1(a)) to obtain the estimates of β_0 and β_1.

The *reduced major axis regression* method proposes to minimize the sum of the areas of rectangles defined between the observed data points and the nearest point on the line in the scatter diagram to obtain the estimates of the regression coefficients (see Fig. 2.1(d)).



(a) Direct regression method (b) Reverse regression method

(c) Major axis regression method (d) Reduced major axis method

FIGURE 2.1. Scatter diagrams of different methods of regression

One may note that the principle of least squares does not require any assumption about the form of probability distribution of e_t in deriving the least squares estimates. For the purpose of deriving the statistical inferences only, we assume that e_t's are observed as random variable ϵ_t with $\mathrm{E}(\epsilon_t) = 0$, $\mathrm{var}(\epsilon_t) = \sigma^2$ and $\mathrm{cov}(\epsilon_t, \epsilon_{t^*}) = 0$ for all $t \neq t^*$ $(t, t^* = 1, \ldots, T)$. This

assumption is needed to find the mean and variance of the least squares estimates. The assumption that ϵ_t's are normally distributed is utilized while constructing the tests of hypotheses and confidence intervals of the parameters.

Based on these approaches, different estimates of β_0 and β_1 are obtained which have different statistical properties. Among them the direct regression approach is more popular. Generally, the direct regression estimates are referred as the least squares estimates. We will consider here the direct regression approach in more detail. Other approaches are also discussed.

2.3 Direct Regression Method

This method is also known as the *ordinary least squares estimation*. The regression models (2.1) and (2.2) can be viewed as the regression models for population and sample, respectively. The direct regression approach minimizes the sum of squares

$$S(\beta_0, \beta_1) = \sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_t)^2 \tag{2.3}$$

with respect to β_0 and β_1.

The partial derivatives of (2.3) with respect to β_0 and β_1 are

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_t) \tag{2.4}$$

and

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_t) x_t, \tag{2.5}$$

respectively. The solution of β_0 and β_1 is obtained by setting (2.4) and (2.5) equal to zero. Thus obtained solutions are called the direct regression estimators, or usually called as the Ordinary Least Squares (OLS) estimators of β_0 and β_1.

This gives the ordinary least squares estimates b_0 of β_0 and b_1 of β_1 as

$$b_0 = \bar{y} - b_1 \bar{x} \tag{2.6}$$

$$b_1 = \frac{SXY}{SXX} \tag{2.7}$$

where

$$SXY = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y}), \qquad SXX = \frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2, \qquad \bar{x} = \frac{1}{T} \sum_{t=1}^{T} x_t$$

and $\bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$.

Further, using (2.4) and (2.5), we have

$$\begin{aligned}
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0^2} &= -2\sum_{t=1}^{T}(-1) = 2T\,,\\
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_1^2} &= 2\sum_{t=1}^{T}x_t^2\,,\\
\frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0\partial\beta_1} &= 2\sum_{t=1}^{T}x_t = 2T\bar{x}\,.
\end{aligned}$$

Thus we get the Hessian matrix which is the matrix of second order partial derivatives as

$$\begin{aligned}
H &= \begin{pmatrix} \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0^2} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0\partial\beta_1} \\ \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_0\partial\beta_1} & \frac{\partial^2 S(\beta_0,\beta_1)}{\partial\beta_1^2} \end{pmatrix}\\
&= 2\begin{pmatrix} T & T\bar{x} \\ T\bar{x} & \sum_{t=1}^{T}x_t^2 \end{pmatrix}\\
&= 2\begin{pmatrix} \mathbf{1}' \\ x' \end{pmatrix}(\mathbf{1}, x) \qquad (2.8)
\end{aligned}$$

where $\mathbf{1} = (1,\ldots,1)'$ is a T-vector of elements unity and and $x = (x_1,\ldots,x_T)'$ is a T-vector of observations on X. The matrix (2.8) is positive definite if its determinant and the element in the first row and column of H are positive. The determinant of H is

$$\begin{aligned}
|H| &= 2\left(T\sum_{t=1}^{T}x_t^2 - T^2\bar{x}^2\right)\\
&= 2T\sum_{t=1}^{n}(x_t-\bar{x})^2\\
&\geq 0. \qquad (2.9)
\end{aligned}$$

The case when $\sum_{t=1}^{T}(x_t-\bar{x})^2 = 0$ is not interesting because then all the observations are identical, *i.e.*, $x_t = c$ (some constant). In such a case there is no relationship between x and y in the context of regression analysis. Since $\sum_{t=1}^{T}(x_t-\bar{x})^2 > 0$, therefore $|H| > 0$. So H is positive definite for any (β_0,β_1); therefore $S(\beta_0,\beta_1)$ has a global minimum at (b_0,b_1).

The fitted line or the fitted linear regression model is

$$y = b_0 + b_1 X \qquad (2.10)$$

and the predicted values are

$$\hat{y}_t = b_0 + b_1 x_t \quad (t = 1,\ldots,T)\,. \qquad (2.11)$$

3.20 Confidence Interval for the Ratio of Two Linear Parametric Functions

Let $\theta_1 = P_1'\beta$ and $\theta_2 = P_2'\beta$ be two linear parametric functions and we wish to find a confidence interval of $\lambda = \frac{\theta_1}{\theta_2}$.

The least squares estimators of θ_1 and θ_2 are

$$\hat{\theta}_1 = P_1'\hat{\beta} \quad \text{and} \quad \hat{\theta}_2 = P_2'\hat{\beta}$$

with the variance-covariance matrix

$$\sigma^2 \begin{pmatrix} P_1'HP_1 & P_1'HP_2 \\ P_2'HP_1 & P_2'HP_2 \end{pmatrix} = \sigma^2 \begin{pmatrix} a & b \\ b' & c \end{pmatrix}, \text{ say.}$$

Then

$$\mathrm{E}(\hat{\theta}_1 - \lambda\hat{\theta}_2) = 0\,, \quad \mathrm{var}(\hat{\theta}_1 - \lambda\hat{\theta}_2) = \sigma^2(a - 2\lambda b + \lambda^2 c)\,.$$

Hence

$$F = \frac{(\hat{\theta}_1 - \lambda\hat{\theta}_2)^2}{s^2(a - 2\lambda b + \lambda^2 c)} \sim F_{1,T-K}$$

and

$$P\left\{(\hat{\theta}_1 - \lambda\hat{\theta}_2)^2 - F_{1-\alpha}s^2(a - 2\lambda b + \lambda^2 c) \le 0\right\} = 1 - \alpha\,. \tag{3.377}$$

The inequality within the brackets in (3.377) provides a $(1-\alpha)$ confidence region for λ. Because the expression in (3.377) is quadratic in λ, the confidence region is the interval between the roots of the quadratic equation or outside the interval, depending on the nature of the coefficients of the quadratic equation.

3.21 Nonparametric Regression

The nonparametric regression model describes the dependence of study variable on explanatory variables without specifying the function that relates them. The general nonparametric regression model is expressed as

$$y = \psi(X) + \epsilon \tag{3.378}$$

where y is a study variable, X is a vector of explanatory variables, ϵ is disturbance term and $\psi(x)$ is the unspecified real valued function of X at some fixed value x given by

$$\psi(x) = E(y|X = x)\,. \tag{3.379}$$

The first derivative of $\psi(x)$ indicates the response or regression coefficient of y with respect to x and the second derivative of $\psi(x)$ indicates the curvature of $\psi(x)$.

We assume that both y and $X_1, \ldots, X_K$ are stochastic and related as

$$y = \psi(X_1, \ldots, X_K) + \epsilon \tag{3.380}$$

with $\mathrm{E}(\epsilon|X = x) = 0$ and $\mathrm{V}(\epsilon|X = x) = \sigma^2 I$. We observe T identically and independently distributed observations $(y_t, x_{t1}, \ldots, x_{tK})$, $t = 1, \ldots, T$ from an absolutely continuous $(K + 1)$-variate distribution with density $f(y, X_1, \ldots, X_K) = f(y, X)$. If $\mathrm{E}(|y|) < \infty$, then the conditional mean of y given $X = x$ exists as in (3.379).

The regression coefficient related to x_j $(j = 1, \ldots, K)$ is

$$\beta_j(x) = \beta(x) = \frac{\partial \psi(x)}{\partial x_j} = \lim_{h \to 0} \frac{\psi(x + h) - \psi(x - h)}{2h} \tag{3.381}$$

where $\psi(x-h) = \psi(x_1, \ldots, x_j - h, \ldots, x_k)$. When $\psi(x)$ is linear, then $\beta_j(x)$ is the j^{th} regression coefficient and is fixed for all x. When $\psi(x)$ is non–linear, then $\beta_j(x)$ depends on x and $\beta_j(x)$ is a varying regression coefficient. The fixed regression coefficient can be defined as $\beta(\bar{x})$, *i.e.*, $\beta(x)$ evaluated at $x = \bar{x} = (\bar{x}_1, \ldots, \bar{x}_K)$. Similarly the second order partial derivative is

$$\beta^{(2)}(x) = \frac{\partial^2}{\partial x_j^2} \psi(x) = \lim_{h \to 0} \frac{\psi(x + 2h) - 2\psi(x) - \psi(x - 2h)}{(2h)^2}, \tag{3.382}$$

and, in general, the p^{th} order partial derivative $(p = 1, 2, \ldots)$ is

$$\begin{aligned} \beta^{(p)}(x) &= \frac{\partial^p}{\partial x_j^p} \psi(x) \\ &= \lim_{h \to 0} \Big[\Big(\frac{1}{2h} \Big)^p \sum_{m=0}^{p} (-1)^m \binom{p}{m} \psi\Big(x + (p - 2m)h \Big) \Big] \end{aligned} \tag{3.383}$$

and the cross partial derivative is

$$\begin{aligned} &\frac{\partial^{p_1 + \ldots + p_r}}{\partial x_{j1}^{p_1}, \ldots, \partial x_{jr}^{p_r}} = \\ &\lim_{h \to 0} \Big[\Big(\frac{1}{2h} \Big)^{p_1 + \ldots + p_r} \sum_{m_1 = 0}^{p_1} \cdots \sum_{m_r = 0}^{p_r} (-1)^{m_1 + \ldots + m_r} \\ &\times \binom{p_1}{m_1} \ldots \binom{p_r}{m_r} \psi\Big(x + (p_1 - 2m_1)h, \ldots, x + (p_r - 2m_r)h \Big) \Big] \end{aligned} \tag{3.384}$$

respectively, where each of $j_1, \ldots, j_r = 1, \ldots, K$ $(j_1 \neq \ldots \neq j_r)$, $x + (p_1 - 2m_1)h = (x_1, \ldots, x_{j1} + (p_1 - 2m_1)h, \ldots, x_K)$, $x + (p_r - 2m_r)h = (x_1, \ldots, x_{jr} + (p_r - 2m_r)h, \ldots, x_K)$.

Now we consider the nonparametric estimation of partial derivatives of $\psi(x)$ without specifying its form. We first consider a nonparametric estimator of $\psi(x)$ and then take its partial derivatives.

3.21.1 Estimation of the Regression Function

Most of the estimation methods of nonparametric regression assume that the regression function is smooth in some sense.

Since $\psi(x)$ depends on unknown densities, so we use the data of $(K+1)$ variables as $z_t = (x_t, y_t)$, $t = 1, \ldots T$ and note that

$$g_T(z) = \frac{1}{Th^{K+1}} \sum_{t=1}^{T} \mathcal{K}\Big(\frac{z_t - z}{h}\Big) , \tag{3.385}$$

$$g_{1T}(x) = \int g_T(z)dy = \frac{1}{Th^K} \sum_{t=1}^{T} \mathcal{K}_1\Big(\frac{x_t - x}{h}\Big) , \tag{3.386}$$

where h is the window width (also called as band width or smoothing parameter) which is a positive function of T that goes to zero as $T \to \infty$, $\mathcal{K}$ is a kernel or a weight function such that $\int \mathcal{K}(z)dz = 1$ and $\mathcal{K}_1(x) = \int \mathcal{K}(z)dy$. The kernel $\mathcal{K}$ determines the shape of the curve and h determines their width. See, Prakasa-Rao (1983), Silverman (1986), Ullah and Vinod (1988) and Pagan and Ullah (1999) for the details on kernel density estimation. Substituting (3.385) and (3.386) in (3.379), we have

$$\psi_T(x) = \psi_T = \int y \frac{g_T(z)}{g_{1T}(x)} dy = \sum_{t=1}^{T} y_t w_t(x) \tag{3.387}$$

where

$$w_t(x) = \frac{\mathcal{K}_1\Big(\frac{x_t - x}{h}\Big)}{\sum_{t=1}^{T} \mathcal{K}_1\Big(\frac{x_t - x}{h}\Big)} . \tag{3.388}$$

The estimator ψ_T of ψ in (3.387) is known as Nadaraya–Watson type estimator due to Nadaraya (1964) and Watson (1964) and is a kernel nonparametric regression estimate. Note that (3.387) is a weighted average of the observed values y_t where the weight of t^{th} observation depends on the distance x_t to x through the kernel $\mathcal{K}$. The fitted nonparametric regression model that is obtained by without making any assumption about the functional form of $\psi(x)$ is

$$y = \psi_T(x) + \hat{\epsilon} \tag{3.389}$$

where $\hat{\epsilon}$ is the nonparametric residual.

This estimator (3.387) is also the weighted least squares estimator of $\psi(x)$ because $\psi_T(x)$ is the value of $\psi(x)$ for which the weighted squared error

$$\sum_{t=1}^{T} \mathcal{K}\Big(\frac{x_t - x}{h}\Big)(y_t - \psi(x))^2$$

is minimum. The method of moments also yields the same estimator of $\psi(x)$ as in (3.387).

When the window width h is not the same for all data points, then some alternative estimators of $\psi(x)$ are suggested. The recursive regression estimator of $\psi(x)$ in such a case is

$$\hat{\psi}_T(x) = \frac{\sum_{t=1}^{T} \frac{y_t}{h_t^K} \mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)}{\sum_{t=1}^{T} \frac{1}{h_t^K} \mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)} \tag{3.390}$$

where h_t denotes a sequence of positive numbers, assumed to satisfy $\sum h_t^K \to \infty$ as $T \to \infty$. An alternative estimator is

$$\tilde{\psi}_T(x) = \frac{\sum_{t=1}^{T} y_t \, \mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)}{\sum_{t=1}^{T} \mathcal{K}\Big(\frac{x_t - x}{h_t}\Big)} \,. \tag{3.391}$$

Both (3.390) and (3.391) are recursive as

$$\hat{\psi}_T(x) = \hat{\psi}_{T-1}(x) + \frac{y_T - \hat{\psi}_{T-1}(x)}{1 + (T-1)\frac{\hat{f}_{T-1}(x)}{h_T^K} \mathcal{K}\Big(\frac{x_T - x}{h_T}\Big)} \tag{3.392}$$

$$\tilde{\psi}_T(x) = \tilde{\psi}_{T-1}(x) + \vartheta_T^{-1} \left[y_T - \tilde{\psi}_{T-1}(x) \mathcal{K}\left(\frac{x_T - x}{h_T}\right)\right] \tag{3.393}$$

where

$$\vartheta_T = \vartheta_{T-1} + \mathcal{K}\Big(\frac{x_T - x}{h_T}\Big), \; \vartheta_0 = 0 \,.$$

Both (3.392) and (3.393) can be updated as additional data points are available.

When ϵ's are such that $\mathrm{V}(\epsilon) = \Sigma \; (\neq \sigma^2 I)$, a $T \times T$ positive definite matrix, then the generalized least squares estimator of $\psi(x)$ is obtained by minimizing $\epsilon' \mathcal{K}^{1/2} \Sigma^{-1} \mathcal{K}^{1/2} \epsilon$ with respect to $\psi(x)$ as

$$\psi_T^* = (\mathbf{1}' \mathcal{K}^{\frac{1}{2}} \Sigma^{-1} \mathcal{K}^{\frac{1}{2}} \mathbf{1})^{-1} \mathbf{1}' \mathcal{K}^{\frac{1}{2}} \Sigma^{-1} y \tag{3.394}$$

where $\mathbf{1} = (1, \ldots, 1)'$, $\mathcal{K} = \mathrm{diag}(\mathcal{K}_1, \ldots, \mathcal{K}_T)$ is a diagonal matrix with $\mathcal{K}_T = \mathcal{K}\Big(\frac{x_t - x}{h}\Big)$.

An operational version of a consistent estimator of $\beta(x)$ in (3.381) is

$$b_T(x) = \frac{\psi_T(x+h) - \psi_T(x-h)}{2h} \tag{3.395}$$

where $\psi_T(x)$ is given by (3.387). Similarly, the estimators of the p^{th} order partial derivative and cross partial derivatives can be obtained by replacing $\psi(\cdot)$ with $\psi_T(\cdot)$ in (3.383) and (3.384), respectively.

Ullah and Vinod (1988) analytically derived the estimator of $\beta(x) = \partial\psi(x)/\partial x$ as

$$\hat{\beta}_T(x) = \frac{\partial}{\partial x}\psi_T(x) = \sum_{t=1}^{T} y_t(\omega_{1t} - \omega_{2t})$$

where

$$\omega_{1t} = \frac{\mathcal{K}'\left(\frac{x_t - x}{h}\right)}{\sum_{t=1}^{T}\mathcal{K}\left(\frac{x_t - x}{h}\right)}$$

and $\omega_{2t} = \omega_t(x)\sum\omega_{1t}$; $\omega_t(x)$ is as in (3.388) and

$$\mathcal{K}'\left(\frac{x_t - x}{h}\right) = \frac{\partial}{\partial x_j}\mathcal{K}\left(\frac{x_t - x}{h}\right).$$

Alternatively, $\hat{\beta}_T(x)$ and its generalization for p^{th} order derivatives of $\psi(x)$, $\hat{\beta}^{(p)}(x)$ can be obtained as a solution of

$$\sum_{m=o}^{p}\binom{p}{m}\beta_T^{(m)}(x)f_T^{(p-1)}(x) = g_T^{(p)}(x), \quad (p = 1, 2, \ldots)$$

where $g^{(p)}(x)$ is the p^{th} order partial derivative of $g(x) = \int yf(y, x)dy$ with respect to x_j.

The restricted least squares estimator of $\psi(x)$ under the exact linear restrictions $R\beta(x) = r$ is

$$\hat{\beta}_T(x) = b_T(x) - R'(RR')^{-1}[Rb_T(x) - r] \tag{3.396}$$

where $b_T(x)$ is given by (3.395).

The regression function can also be estimated by using various nonparametric procedures like nearest neighbor kernel estimation, local polynomial regression, and smoothing splines.

The method of nearest neighborhood kernel estimation is based on defining a symmetric unimodal weight function $W(x)$ which is centered on the focal observation and goes to zero at the boundaries of the neighborhood around the focal value. Let x_{fo} be a focal x-value at which $\psi(x)$ is to be estimated. Now find ν nearest x-neighbors of x_{fo} where ν/x_{fo} is the span of the kernel smoother. The larger the span, smoother is the estimated regression function. Using the weights defined by $W(x)$, calculate the weighted average of y and obtain the fitted value

$$\hat{y}_{fo} = \hat{f}(x_{fo}) = \frac{\sum_{t=1}^{T} y_t W(x_t)}{\sum_{t=1}^{T} W(x_t)}. \tag{3.397}$$

Repetition of this procedure at a range of x-values spanning the data and connecting the fitted values produces an estimate of the regression function.

In local polynomial regression, the fitted values are produced by locally weighted regression rather than by locally weighted averaging. Another method of nonparametric regression is smoothing splines which are the solution to the penalized regression problem. Additive regression models are an alternate to nonparametric regression with several explanatory variables.

The readers are referred to Prakasa-Rao (1983), Silverman (1986), Ullah (1989a), Ullah (1989b), Härdle (1990) and Pagan and Ullah (1999) for the asymptotic properties of the estimators, related testing of hypothesis and other aspects on nonparametric regression.

3.22 Classification and Regression Trees (CART)

Nonparametric regression with multiple explanatory variables suffers from the problem of *curse of dimensionality*. This means that if the number of explanatory variables is high, then it may be difficult to catch the relevant features of the problem in hand, e.g. the influence of interactions of explanatory variables on the study variable may be difficult to study. We have only a finite sample available, but there may be big volumes in the space of explanatory variables where there may be no observation or only a few observations are obtained (*sparseness problem*). Therefore a reliable statistical estimation is not possible in these volumes. Parametric models, such as simple linear models, or additive models as proposed by Hastie and Tibshirani (1990) try to catch at least the *main effects* of the explanatory variables and discard any global or local interactions of the explanatory variables on the study variable. Furthermore, the results from the approaches like Projection Pursuit Regression (see Section 3.24) or Neural Networks (see Section 3.25) may be hard to interpret. In such situations, CART is more useful. CART tries to catch the relevant interactions of the explanatory variables in their influence on the study variable and present the results in a simple way.

Consider a general regression setup in which the study variable y is either real-valued or categorical and $X_1, \ldots, X_K$ are the explanatory variables. In the usual nonparametric regression setup, we assume

$$y = \psi(X_1, \ldots, X_K) + \epsilon \ , \tag{3.398}$$

with $\mathrm{E}(\epsilon|X) = 0$. If the function ψ is unknown and not parameterized by a finite dimensional parameter, Breiman, H., Olshen and Stone (1984) suggested a recursive partitioning algorithm of the covariate space which results in a tree structure. If y is real-valued, as in (3.398), then the resulting

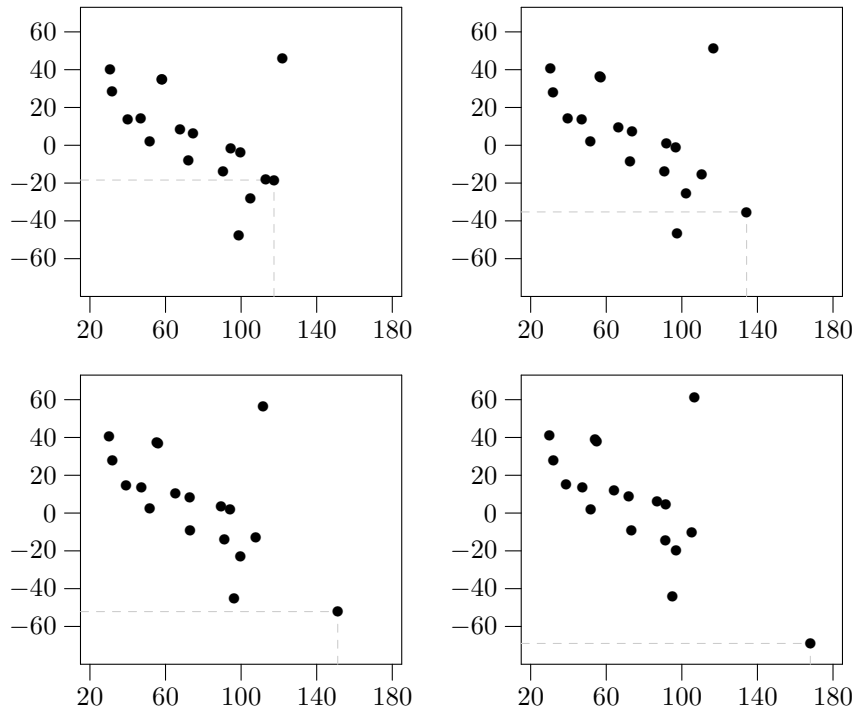


FIGURE 7.8. Four frames $\lambda = 0$ (a), $\lambda = \frac{1}{3}$ (b), $\lambda = \frac{2}{3}$ (c) and $\lambda = 1$ (d) (left to right, top down) of an animated plot of $\hat{\epsilon}(\lambda)$ versus $\hat{y}(\lambda)$ for data in Example 7.4 when removing the 17th observation (marked by dotted lines).

7.8 Model Selection Criteria

Consider a set of regression models with a study variable y and K explanatory variables $X_1, \ldots, X_K$ that can be fitted to a given set of data. The question then arises how to select a good model that has good agreement with the observed data. This also pertains to how to select the number of explanatory variables in the model which yield a 'good' model. A simple approach for model selection is to choose a model with smallest residual sum of squares (RSS). Another criterion is based on the coefficient of determination R^2 and adjusted R^2 which suggest to select the model with the higher coefficient of determination. Some other popular criteria of model selection are Akaikes Information Criterion (AIC, Akaike (1973)), the Bayesian Information Criterion (BIC, Schwarz (1978)) and Mallows C_p (Mallows (1973)) which we introduce in the following subsections.

7.8.1 Akaikes Information Criterion

One approach to select a good model from a set of candidate models is to use Akaikes Information Criterion (AIC) due to Akaike (1973). Its concept is based on the relationship between information theory and likelihood theory.

Kullback and Leibler (1951) presented an approach to quantify the information and defined the Kullback-Leibler distance function as

$$\begin{aligned} I(f,g) &= \int f(x)\log\Big(\frac{f(x)}{g(x)}\Big)dx \\ &= \int f(x)\cdot\log(f(x))dx - \int f(x)\cdot\log(g(x))dx \end{aligned} \tag{7.98}$$

which can be considered as a measure of distance between the two functions f and g. These functions can be two different models from the selection point of view in the context of linear regression analysis. Let f denote the underlying but unknown true model and $g(x|\theta)$ denote any approximation to it which depends on the unknown parameters to be estimated. Then the Kullback-Leibler distance function is

$$\begin{aligned} I(f,g(x|\theta)) &= \int f(x)\cdot\log(f(x))dx - \int f(x)\cdot\log(g(x|\theta))dx \\ &= \mathrm{E}_f[\log(f(x))] - \mathrm{E}_f[\log(g(x|\theta))]\,. \end{aligned} \tag{7.99}$$

The model $g_a(x|\theta)$ is said to be better than $g_b(x|\theta)$ in the sense of Kullback-Leibler distance if $I(f,g_a(x|\theta)) < I(f,g_b(x|\theta))$. In such a case the distance between model $g_a(x|\theta)$ and f is smaller than the distance between $g_b(x|\theta)$ and f. Therefore the aim is to search a model $g(x|\theta)$ which minimizes the Kullback–Leibler distance $I(f,g(x|\theta))$. Using (7.99), we note that

$$\begin{aligned} & I(f,g_a(x|\theta)) < I(f,g_b(x|\theta)) \\ \Leftrightarrow\ & \mathrm{E}_f[\log(f(x))] - \mathrm{E}_f[\log(g_a(x|\theta))] < \mathrm{E}_f[\log(f(x))] - \mathrm{E}_f[\log(g_b(x|\theta))] \\ \Leftrightarrow\ & -\mathrm{E}_f[\log(g_a(x|\theta)) < -\mathrm{E}_f[\log(g_b(x|\theta))]\,. \end{aligned} \tag{7.100}$$

Thus (7.100) indicates that the term $\mathrm{E}_f[\log(f(x))]$ can be treated as a constant for the comparison between two models and therefore $\mathrm{E}_f[\log(g(x|\theta))]$ becomes the function of interest. Let $\hat{\theta}(y)$ denote an estimator of θ based on observations y. Since θ is usually unknown and has to be estimated from the data, so one can consider on minimizing the *expected* Kullback–Leibler distance $\mathrm{E}_y\left[I(f,g(\cdot|\hat{\theta}(y)))\right]$ instead of (7.99). One good approach in such cases is to use the maximum likelihood estimate of θ.

So, using (7.100),

$$\begin{aligned} \mathrm{E}_y\left[I(f, g(\cdot|\hat{\theta}(y)))\right] &= \int f(x)\log(f(x))dx \\ &\quad - \mathrm{E}_y\left[\int f(x)\log[g(x|\hat{\theta}(y))]dx\right] \\ &= C - \mathrm{E}_y\,\mathrm{E}_x\left[\log[g(x|\hat{\theta}(y))]\right] \end{aligned} \tag{7.101}$$

where C is a constant term. Akaike found a relationship between the second term of (7.101) and the log-likelihood function. He showed, that

$$\mathrm{E}_y\,\mathrm{E}_x\left[\log[g(x|\hat{\theta}(y))]\right] \approx L(\hat{\theta}|y) - K \tag{7.102}$$

where $L(\hat{\theta}|y)$ denotes the maximized log-likelihood for the model g and K is the the number of estimable parameters in the approximating model. The readers are referred to Burnham and Anderson (2002) for the derivation of (7.102). The AIC can further be expressed as

$$AIC = -2L(\hat{\theta}|y) + 2K\,. \tag{7.103}$$

Burnham and Anderson (2002) state that Akaike multiplied the bias–corrected log-likelihood by -2 for the reasons like e.g., that it is well known that -2 times the logarithm of the ratio of two maximized likelihood values is asymptotically chi-squared under certain conditions and assumptions. Two points frequently arise and we note this here. Firstly, the model associated with the minimum AIC remains unchanged if the bias corrected likelihood (*i.e.*, $\log L - K$) is multiplied by -0.17, -34, or -51.3, or any other negative number. Thus the minimization is not changed by the multiplication of both the terms by a negative constant; Akaike merely chooses -2. Secondly, some investigators have not realized the formal link between Kullback-Leibler information and AIC and believed that the number 2 in the second term in AIC was somehow arbitrary and other numbers should also be considered. This error has led to considerable confusion in the technical literature; clearly, K is the asymptotic bias correction and is not arbitrary. Akaike chooses to work with $-2\log L$ rather than $\log L$; thus the term $+2K$ is theoretically correct for large sample size. As long as both the terms (the log–likelihood and the bias correction) are multiplied by the same negative constant, the minimization of AIC in (7.103) gives the same result and nothing is arbitrary.

The model with smallest AIC, which in turn is closest to the unknown true model, is said to be the best model under AIC criterion. In case of a linear regression model, AIC can be written as:

$$AIC_{LM} = T \cdot \ln(RSS) + 2K - T \cdot \ln(T)\,. \tag{7.104}$$

The second terms in (7.103) and (7.104) are often termed as penalty term.