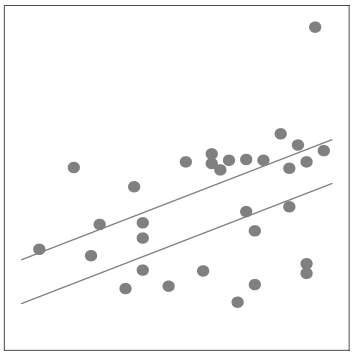

Chapter 14
Linear Least Squares Analysis

Linear least squares methods allow researchers to study how variables are related. For example, a researcher might be interested in determining the relationship between the weight of an individual and such variables as height, age, sex, and general body dimensions.

Sections 1 and 2 introduce methods used to analyze how one variable can be used to predict another (for example, how height can be used to predict weight). Section 3 introduces methods to analyze how several variables can be used to predict another (for example, how the combination of height, age, sex, and general body dimensions can be used to predict weight). Bootstrap applications are given in Section 4. Section 5 outlines the laboratory problems. References for regression diagnostic methods are [12], [28], [49].

14.1 Simple linear model

A *simple linear model* is a model of the form

$$Y = \alpha + \beta X + \epsilon,$$

where X and ϵ are independent random variables, and the distribution of ϵ has mean 0 and standard deviation σ. Y is called the *response* variable, and X is called the *predictor* variable. ϵ represents the measurement error.

The response variable Y can be written as a linear function of the predictor variable X plus an error term. The linear prediction function has slope β and intercept α.

The objective is to estimate the parameters in the conditional mean formula

$$E(Y|X = x) = \alpha + \beta x$$

using a list of paired observations. The observed pairs are assumed to be either the values of a random sample from the joint (X, Y) distribution or a collection of

independent responses made at predetermined levels of the predictor. Analysis is done conditional on the observed values of the predictor variable.

14.1.1 Least squares estimation

Assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means $E(Y_i) = \alpha + \beta x_i$, that the collection $\{\epsilon_i\}$ is a random sample from a distribution with mean 0 and standard deviation σ, and that all parameters (α, β, and σ) are unknown.

Least squares is a general estimation method introduced by A. Legendre in the early 1800's. In the simple linear case, the *least squares* (LS) estimators of α and β are obtained by minimizing the following sum of squared deviations of observed from expected responses:

$$S(\alpha, \beta) = \sum_{i=1}^{N} (Y_i - (\alpha + \beta x_i))^2 .$$

Multivariable calculus can be used to demonstrate that the LS estimators of slope and intercept can be written in the form

$$\widehat{\beta} = \sum_{i=1}^{N} \left[\frac{(x_i - \overline{x})}{S_{xx}} \right] Y_i \quad \text{and} \quad \widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{x} = \sum_{i=1}^{N} \left[\frac{1}{N} - \frac{(x_i - \overline{x})\overline{x}}{S_{xx}} \right] Y_i,$$

where $\overline{x}$ and $\overline{Y}$ are the mean values of predictor and response, respectively, and S_{xx} is the sum of squared deviations of observed predictors from their sample mean:

$$S_{xx} = \sum_{i=1}^{N} (x_i - \overline{x})^2.$$

Formulas for $\widehat{\alpha}$ and $\widehat{\beta}$ can be written in many different ways. The method used here emphasizes that each estimator is a linear combination of the response variables.

Example: Olympic winning times

To illustrate the computations, consider the following 20 data pairs, where x is the time in years since 1900 and y is the Olympic winning time in seconds for men in the final round of the 100-meter event [50, p. 248]:

x	0	4	8	12	20	24	28	32	36	48
y	10.8	11.0	10.8	10.8	10.8	10.6	10.8	10.3	10.3	10.3
x	52	56	60	64	68	72	76	80	84	88
y	10.4	10.5	10.2	10.0	9.95	10.14	10.06	10.25	9.99	9.92

The data set covers all Olympic events held between 1900 and 1988. (Olympic games were not held in 1916, 1940, and 1944.) For these data, $\overline{x} = 45.6$, $\overline{y} = 10.396$, and

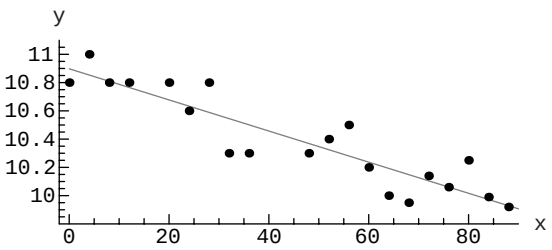

Figure 14.1. *Olympic winning time in seconds for men's 100-meter finals (vertical axis) versus year since* 1900 *(horizontal axis). The gray line is the linear least squares fit,* $y = 10.898 - 0.011x$.

the least squares estimates of slope and intercept are $\widehat{\beta} = -0.011$ and $\widehat{\alpha} = 10.898$, respectively. Figure 14.1 shows a scatter plot of the Olympic winning times data pairs superimposed on the least squares fitted line. The results suggest that the winning times have decreased at the rate of about 0.011 seconds per year during the 88 years of the study.

Properties of LS estimators

Theorem 4.4 can be used to demonstrate the following:

1. $E(\widehat{\beta}) = \beta$ and $Var(\widehat{\beta}) = \sigma^2/S_{xx}$.
2. $E(\widehat{\alpha}) = \alpha$ and $Var(\widehat{\alpha}) = \left(\sum_i x_i^2\right) \sigma^2 / (N\, S_{xx})$.

In addition, the following theorem, proven by Gauss and Markov, states that LS estimators are best (minimum variance) among all linear unbiased estimators of intercept and slope.

Theorem 14.1 (Gauss–Markov Theorem). *Under the assumptions of this section, the least squares (LS) estimators are the best linear unbiased estimators of* α *and* β.

For example, consider estimating β using a linear function of the response variables, say $W = c + \sum_i d_i Y_i$ for some constants c and $d_1, d_2, \ldots, d_N$. If W is an unbiased estimator of β, then

$$Var(W) = Var\left(c + \sum_i d_i Y_i\right) = \sum_i d_i^2\, Var(Y_i) = \left(\sum_i d_i^2\right) \sigma^2$$

is minimized when $d_i = (x_i - \overline{x})/S_{xx}$ and $c = 0$. That is, the variance is minimized when W is the LS estimator of β.

Although LS estimators are best among linear unbiased estimators, they may not be ML estimators. Thus, there may be other more efficient methods of estimation.

14.1.2 Permutation confidence interval for slope

Permutation methods can be used to construct confidence intervals for the slope parameter β in the simple linear model. Let

$$(x_i, y_i) \text{ for } i = 1, 2, \ldots, N$$

be the observed pairs and π be a permutation of the indices $1, 2, \ldots, N$ other than the identity. Then the quantity

$$b(\pi) = \frac{\sum_i (x_i - \overline{x})(y_{\pi(i)} - y_i)}{\sum_i (x_i - \overline{x})(x_{\pi(i)} - x_i)}$$

is an estimate of β, and the collection

$$\{b(\pi) \ : \ \pi \text{ is a permutation other then the identity}\}$$

is a list of $N! - 1$ estimates. The ordered estimates

$$b_{(1)} \ < \ b_{(2)} \ < \ b_{(3)} \ < \ \cdots \ < \ b_{(N!-1)}$$

are used in constructing confidence intervals.

Theorem 14.2 (Slope Confidence Intervals). *Under the assumptions of this section, the interval*

$$[b_{(k)}, b_{(N!-k)}]$$

is a $100(1 - 2k/N!)\%$ *confidence interval for* β.

The procedure given in Theorem 14.2 is an example of *inverting* a hypothesis test: A value β_o is in a $100(1 - \gamma)\%$ confidence interval if the two sided permutation test of

$$H_o: \text{ The correlation between } Y - \beta_o X \text{ and } X \text{ is zero}$$

is accepted at the γ significance level. For a proof, see [74, p. 120].

Since the number of permutations can be quite large, Monte Carlo analysis is used to estimate endpoints. For example, assume the Olympic times data (page 206) are the values of random variables satisfying the assumptions of this section. An approximate 95% confidence interval for the slope parameter (based on 5000 random permutations) is $[-0.014, -0.008]$.

14.2 Simple linear regression

In simple *linear regression*, the error distribution is assumed to be normal, and, as above, analyses are done conditional on the observed values of the predictor variable. Specifically, assume that

$$Y_i = \alpha + \beta x_i + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means $E(Y_i) = \alpha + \beta x_i$, that the collection $\{\epsilon_i\}$ is a random sample from a normal distribution with mean 0 and standard deviation σ, and that all parameters are unknown.

In this setting, LS estimators are ML estimators.

Theorem 14.3 (Parameter Estimation). *Given the assumptions and definitions above, the LS estimators of α and β given on page 206 are ML estimators, and the statistics*

$$\begin{aligned}\widehat{\epsilon_i} = Y_i - (\widehat{\alpha} + \widehat{\beta} x_i) &= Y_i - (\overline{Y} + \widehat{\beta}(x_i - \overline{x})) \\ &= Y_i - \sum_j \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j\end{aligned}$$

are ML estimators of the error terms for $i = 1, 2, \ldots, N$. Each estimator is a normal random variable, and each is unbiased. Further, the statistic

$$S^2 = \frac{1}{N-2} \sum_i (Y_i - (\widehat{\alpha} + \widehat{\beta} x_i))^2$$

is an unbiased estimator of the common variance σ^2.

14.2.1 Confidence interval procedures

This section develops confidence interval procedures for the slope and intercept parameters, and for the mean response at a fixed value of the predictor variable.

Hypothesis tests can also be developed. Most computer programs automatically include both types of analyses.

Confidence intervals for β

Since the LS estimator $\widehat{\beta}$ is a normal random variable with mean β and variance σ^2/S_{xx}, Theorem 6.2 can be used to demonstrate that

$$\widehat{\beta} \pm t_{N-2}(\gamma/2) \sqrt{\frac{S^2}{S_{xx}}}$$

is a $100(1-\gamma)\%$ confidence interval for β, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

Confidence intervals for α

Since the LS estimator $\widehat{\alpha}$ is a normal random variable with mean α and variance $\sigma^2 \left(\sum_i x_i^2\right) / (N\, S_{xx})$, Theorem 6.2 can be used to demonstrate that

$$\widehat{\alpha} \pm t_{N-2}(\gamma/2) \sqrt{\frac{S^2 \left(\sum_i x_i^2\right)}{N\, S_{xx}}}$$

is a $100(1-\gamma)\%$ confidence interval for α, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

For example, if the Olympic times data (page 206) are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for the slope parameter is $[-0.013, -0.009]$, and a 95% confidence interval for the intercept parameter is [10.765, 11.030].

Confidence intervals for mean response

The mean response $E(Y_o) = \alpha + \beta x_o$ at a new predictor-response pair, (x_o, Y_o), can be estimated using the statistic

$$\widehat{\alpha} + \widehat{\beta} x_o = \overline{Y} + \widehat{\beta}(x_o - \overline{x}) = \sum_{i=1}^{N} \left[\frac{1}{N} + \frac{(x_o - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_i.$$

This estimator is a normal random variable (by Theorem 4.6) with mean $\alpha + \beta x_o$ and

$$Var(\widehat{\alpha} + \widehat{\beta} x_o) = \sigma^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right).$$

Thus, Theorem 6.2 can be used to demonstrate that

$$(\widehat{\alpha} + \widehat{\beta} x_o) \ \pm \ t_{N-2}(\gamma/2) \sqrt{S^2 \left(\frac{1}{N} + \frac{(x_o - \overline{x})^2}{S_{xx}} \right)}$$

is a $100(1-\gamma)\%$ confidence interval for $\alpha + \beta x_o$, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-2}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-2)$ degrees of freedom.

Example: Percentage of dead or damaged spruce trees

For example, as part of a study on the relationship between environmental stresses and the decline of red spruce tree forests in the Appalachian Mountains, data were collected on the percentage of dead or damaged trees at various altitudes in forests in the northeast. The paired data were of interest because concentrations of airborne pollutants tend to be higher at higher altitudes [49, p. 102].

Figure 14.2 is based on information gathered in 53 areas. For these data, the least squares fitted line is $y = 8.24x - 33.66$, suggesting that the percentage of damaged or dead trees increases at the rate of 8.24 percentage points per 100 meters elevation.

An estimate of the mean response at 1000 meters ($x_o = 10$) is 48.76% damaged or dead. If these data are the values of independent random variables satisfying the assumptions of this section, then a 95% confidence interval for the mean response at 1000 meters is [48.44, 49.07].

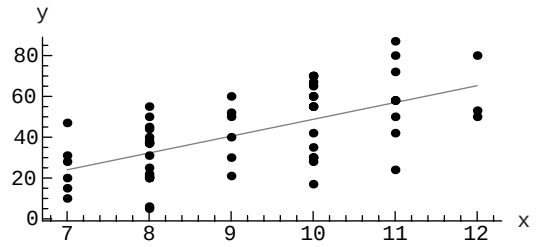

Figure 14.2. *Percentage dead or damaged red spruce trees (vertical axis) versus elevation in* 100 *meters (horizontal axis) at* 53 *locations in the northeast. The gray line is the linear least squares fit,* $y = 8.24x - 33.66$.

Comparison of procedures

The confidence interval procedure for β given in this section is valid when the error distribution is normal. When the error distribution is not normal, the permutation procedure given in Theorem 14.2 can be used.

The confidence interval procedures given in this section assume that the values of the predictor variable are known with certainty (the procedures are conditional on the observed values of the predictor) and assume that the error distributions are normal. Approximate bootstrap confidence interval procedures can also be developed under broader conditions; see Section 14.4.

14.2.2 Predicted responses and residuals

The i^{th} estimated mean (or *predicted response*) is the random variable

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} x_i = \sum_j \left[\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] Y_j \quad \text{for } i = 1, 2, \ldots, N,$$

and the i^{th} estimated error (or *residual*) is

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i \quad \text{for } i = 1, 2, \ldots, N.$$

Each random variable is a linear function of the response variables. Theorem 4.5 can be used to demonstrate that $Cov(\widehat{Y}_i, \widehat{\epsilon}_i) = 0$.

Although the error terms in the simple linear model have equal variances, the estimated errors do not. Specifically, the variance of the i^{th} residual is

$$Var(\widehat{\epsilon}_i) = \sigma^2 \left[\left(1 - \frac{1}{N} - \frac{(x_i - \overline{x})^2}{S_{xx}} \right)^2 + \sum_{j \neq i} \left(\frac{1}{N} + \frac{(x_j - \overline{x})(x_i - \overline{x})}{S_{xx}} \right)^2 \right] = \sigma^2 \, c_i.$$

The i^{th} estimated *standardized residual* is defined as follows:

$$r_i = \widehat{\epsilon}_i / \sqrt{S^2 \, c_i} \text{ for } i = 1, 2, \ldots, N,$$

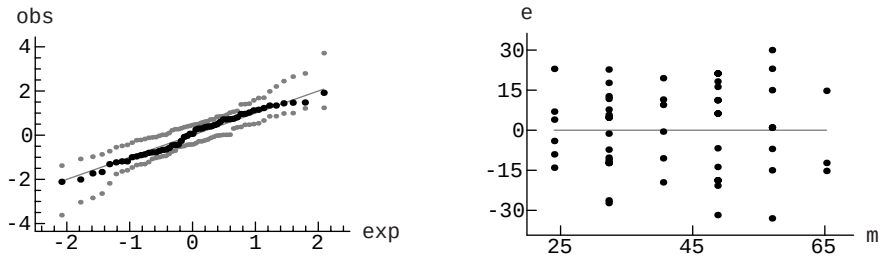

Figure 14.3. *Enhanced normal probability plot of standardized residuals (left plot) and scatter plot of residuals versus estimated means (right plot) for the spruce trees example.*

where S^2 is the estimate of the common variance given in Theorem 14.3 and c_i is the constant in brackets above.

Predicted responses, residuals, and estimated standardized residuals are used in diagnostic plots of model assumptions. For example, the left plot in Figure 14.3 is an enhanced normal probability of the estimated standardized residuals from the spruce trees example (page 210), and the right plot is a scatter plot of residuals (vertical axis) versus predicted responses (horizontal axis). The left plot suggests that the error distribution is approximately normally distributed; the right plot exhibits no relationship between the estimated errors and estimated means.

The scatter plot of residuals versus predicted responses should show no relationship between the variables. Of particular concern are the following:

1. If $\widehat{\epsilon_i} \approx h(\widehat{y_i})$ for some function h, then the assumption that the conditional mean is a linear function of the predictor may be wrong.
2. If $SD(\widehat{\epsilon_i}) \approx h(\widehat{y_i})$ for some function h, then the assumption of equal standard deviations may be wrong.

14.2.3 Goodness-of-fit

Suppose that the N predictor-response pairs can be written in the following form:

$$(x_i, Y_{i,j}) \text{ for } j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, I.$$

(There are a total of n_i observed responses at the i^{th} level of the predictor variable for $i = 1, 2, \ldots, I$, and $N = \sum_i n_i$.) Then it is possible to use an analysis of variance technique to test the goodness-of-fit of the simple linear model.

Assumptions

The responses are assumed to be the values of I independent random samples

$$Y_{i,1}, Y_{i,2}, \ldots, Y_{i,n_i} \text{ for } i = 1, 2, \ldots, I$$

from normal distributions with a common unknown standard deviation σ.

Let μ_i be the mean of responses in the i^{th} sample: $\mu_i = E(Y_{i,j})$ for all j. Of interest is to test the null hypothesis that $\mu_i = \alpha + \beta x_i$ for $i = 1, 2, \ldots, I$.

Sources of variation

The formal goodness-of-fit analysis is based on writing the sum of squared deviations of the response variables from the predicted responses using the linear model (known as the *error* sum of squares),

$$SS_e = \sum_{i,j} \left(Y_{i,j} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2,$$

as the sum of squared deviations of the response variables from the estimated group means (known as the *pure error* sum of squares),

$$SS_p = \sum_{i,j} \left(Y_{i,j} - \overline{Y}_{i\cdot}\right)^2,$$

plus the weighted sum of squared deviations of the group means from the predicted responses (known as the *lack-of-fit* sum of squares),

$$SS_\ell = \sum_i n_i \left(\overline{Y}_{i\cdot} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2.$$

Pure error and lack-of-fit mean squares

The *pure error* mean square, MS_p, is defined as follows:

$$MS_p = \frac{1}{N-I} SS_p = \frac{1}{N-I} \sum_{i,j} \left(Y_{i,j} - \overline{Y}_{i\cdot}\right)^2.$$

MS_p is equal to the pooled estimate of the common variance. Theorem 6.1 can be used to demonstrate that $(N-I)MS_p/\sigma^2$ is a chi-square random variable with $(N-I)$ degrees of freedom.

The *lack-of-fit* mean square, MS_ℓ, is defined as follows:

$$MS_\ell = \frac{1}{I-2} SS_\ell = \frac{1}{I-2} \sum_i n_i \left(\overline{Y}_{i\cdot} - (\widehat{\alpha} + \widehat{\beta} x_i)\right)^2.$$

Properties of expectation can be used to demonstrate that

$$E\left(MS_\ell\right) = \sigma^2 + \frac{1}{I-2} \sum_i n_i \left(\mu_i - (\alpha + \beta x_i)\right)^2.$$

If the null hypothesis that the means follow a simple linear model is true, then the expected value of MS_ℓ is σ^2; otherwise, values of MS_ℓ will tend to be larger than σ^2. The following theorem relates the pure error and lack-of-fit mean squares.

Theorem 14.4 (Distribution Theorem). *Under the general assumptions of this section and if the null hypothesis is true, then the ratio $F = MS_\ell/MS_p$ has an f ratio distribution with $(I-2)$ and $(N-I)$ degrees of freedom.*

Table 14.1. *Goodness-of-fit analysis of the spruce tree data.*

Source	df	SS	MS	F	p value
Lack-of-Fit	4	132.289	33.072	0.120	0.975
Pure Error	47	12964.3	275.835		
Error	51	13096.5			

Goodness-of-fit test: Observed significance level

Large values of $F = MS_\ell/MS_p$ support the alternative hypothesis that the simple linear model does not hold. For an observed ratio, f_{obs}, the p value is $P(F \geq f_{\text{obs}})$.

For example, assume the spruce trees data (page 210) satisfy the general assumptions of this section. Table 14.1 shows the results of the goodness-of-fit test. There were 6 observed predictor values. The observed ratio of the lack-of-fit mean square to the pure error mean square is 0.120. The observed significance level, based on the f ratio distribution with 4 and 47 degrees of freedom, is 0.975. The simple linear model fits the data quite well.

14.3 Multiple linear regression

A *linear model* is a model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \epsilon,$$

where each X_i is independent of ϵ, and the distribution of ϵ has mean 0 and standard deviation σ. Y is called the *response* variable, each X_i is a *predictor* variable, and ϵ represents the measurement error.

The response variable Y can be written as a linear function of the $(p-1)$ predictor variables plus an error term. The linear prediction function has p parameters.

In multiple *linear regression*, the error distribution is assumed to be normal, and analyses are done conditional on the observed values of the predictor variables. Observations are called *cases*.

14.3.1 Least squares estimation

Assume that

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are independent random variables with means

$$E(Y_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_{p-1} x_{p-1,i} \text{ for all } i,$$

that the collection of errors $\{\epsilon_i\}$ is a random sample from a normal distribution with mean 0 and standard deviation σ, and that all parameters (including σ) are unknown.

The *least squares* (LS) estimators of the coefficients in the linear prediction function are obtained by minimizing the following sum of squared deviations of observed from expected responses:

$$S\left(\beta_0, \beta_1, \ldots, \beta_{p-1}\right) = \sum_{k=1}^{N} \left(Y_k - \left(\beta_0 + \beta_1 x_{1,k} + \beta_2 x_{2,k} + \cdots + \beta_{p-1} x_{p-1,k}\right)\right)^2$$
$$= \sum_{k=1}^{N} \left(Y_k - \sum_{j=0}^{p-1} \beta_j x_{j,k} \right)^2, \quad \text{where } x_{0,k} = 1 \text{ for all } k.$$

The first step in the analysis is to compute the partial derivative with respect to β_i for each i. Partial derivatives have the following form:

$$\frac{\partial S}{\partial \beta_i} = -2 \left[\sum_{k=1}^{N} Y_k x_{i,k} - \sum_{j=0}^{p-1} \beta_j \left(\sum_{k=1}^{N} x_{j,k} x_{i,k} \right) \right].$$

The next step is to solve the p-by-p system of equations

$$\frac{\partial S}{\partial \beta_i} = 0, \text{ for } i = 0, 1, \ldots, p-1,$$

or, equivalently,

$$\sum_{j=0}^{p-1} \left(\sum_{k=1}^{N} x_{j,k} x_{i,k} \right) \beta_j = \sum_{k=1}^{N} Y_k x_{i,k}, \text{ for } i = 0, 1, \ldots, p-1.$$

In matrix notation, the system becomes

$$\left(\mathbf{X}^T \mathbf{X}\right) \underline{\beta} = \mathbf{X}^T \underline{Y},$$

where $\underline{\beta}$ is the p-by-1 vector of unknown parameters, $\underline{Y}$ is the N-by-1 vector of response variables, $\mathbf{X}$ is the N-by-p matrix whose (i, j) element is $x_{j,i}$, and $\mathbf{X}^T$ is the transpose of the $\mathbf{X}$ matrix. The $\mathbf{X}$ matrix is often called the *design matrix*. Finally, the p-by-1 vector of LS estimators is

$$\underline{\widehat{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \underline{Y}.$$

Estimates exist as long as $\left(\mathbf{X}^T \mathbf{X}\right)$ is invertible.

The rows of the design matrix correspond to the observations (or cases). The columns correspond to the predictors. The terms in the first column of the design matrix are identically equal to one.

For example, in the simple linear case, the matrix product

$$\left(\mathbf{X}^T \mathbf{X}\right) = \begin{bmatrix} N & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

has inverse

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \frac{1}{NS_{xx}}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & N \end{bmatrix}, \quad \text{where } S_{xx} = \sum_i (x_i - \overline{x})^2,$$

and the LS estimators are

$$\begin{bmatrix} \widehat{\alpha} \\ \widehat{\beta} \end{bmatrix} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\underline{Y} = \begin{bmatrix} \sum_i \left(\frac{1}{N} - \frac{(x_i-\overline{x})\overline{x}}{S_{xx}}\right) Y_i \\ \sum_i \left(\frac{(x_i-\overline{x})}{S_{xx}}\right) Y_i \end{bmatrix}.$$

The estimators here correspond exactly to those given on page 206.

Model in matrix form

The model can be written as

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\epsilon},$$

where $\underline{Y}$ and $\underline{\epsilon}$ are N-by-1 vectors of responses and errors, respectively, $\underline{\beta}$ is the p-by-1 coefficient vector, and $\mathbf{X}$ is the N-by-p design matrix.

Theorem 14.5 (Parameter Estimation). *Given the assumptions and definitions above, the vector of LS estimators of $\underline{\beta}$ given on page 215 is a vector of ML estimators, and the vector*

$$\widehat{\underline{\epsilon}} = \underline{Y} - \mathbf{X}\widehat{\beta} = (\mathbf{I} - \mathbf{H})\,\underline{Y},$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ and $\mathbf{I}$ is the N-by-N identity matrix, is a vector of ML estimators of the error terms. Each estimator is a normal random variable, and each is unbiased. Further, the statistic

$$S^2 = \frac{1}{N-p}\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2,$$

where $\widehat{Y}_i$ is the i^{th} estimated mean, is an unbiased estimator of σ^2.

The i^{th} estimated mean (or *predicted response*) is the random variable

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta_{p-1}} x_{i,p-1} \text{ for } i = 1, 2, \ldots, N.$$

Further, the matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is often called the *hat matrix* since it is the matrix that transforms the response vector to the predicted response vector

$$\widehat{\underline{Y}} = \mathbf{X}\,\widehat{\underline{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\underline{Y} = \mathbf{H}\,\underline{Y}$$

(the vector of Y_i's is transformed to the vector of Y_i *hat*s).

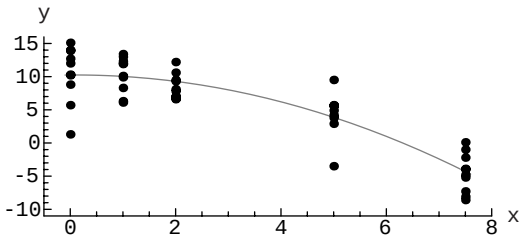

Figure 14.4. *Change in weight in grams (vertical axis) versus dosage level in* 100 *mg/kg/day (horizontal axis) for data from the toxicology study. The gray curve is the linear least squares fit,* $y = 10.2475 + 0.053421x - 0.2658x^2$.

Variability of LS estimators

If $\underline{V}$ is an m-by-1 vector of random variables and $\underline{W}$ is an n-by-1 vector of random variables, then $\Sigma(\underline{V}, \underline{W})$ is the m-by-n matrix whose (i, j) term is $Cov(V_i, W_j)$. The matrix $\Sigma(\underline{V}, \underline{W})$ is called a *covariance matrix.*

Theorem 14.6 (Covariance Matrices). *Under the assumptions of this section, the following hold:*

1. *The covariance matrix of the coefficient estimators is*

$$\sum\left(\underline{\widehat{\beta}}, \underline{\widehat{\beta}}\right) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

2. *The covariance matrix of the error estimators is*

$$\sum\left(\underline{\widehat{\epsilon}}, \underline{\widehat{\epsilon}}\right) = \sigma^2(\mathbf{I} - \mathbf{H}).$$

3. *The covariance matrix of error estimators and predicted responses is*

$$\sum\left(\underline{\widehat{\epsilon}}, \underline{\widehat{Y}}\right) = \mathbf{0},$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ *is the hat matrix,* $\mathbf{I}$ *is the N-by-N identity matrix, and* $\mathbf{0}$ *is the N-by-N matrix of zeros.*

The diagonal elements of $\sum\left(\underline{\widehat{\beta}}, \underline{\widehat{\beta}}\right)$ and $\sum\left(\underline{\widehat{\epsilon}}, \underline{\widehat{\epsilon}}\right)$ are the variances of the coefficient and error estimators, respectively. The last statement in the theorem says that error estimators and predicted responses are uncorrelated.

Example: Toxicology study

To illustrate some of the computations, consider the data pictured in Figure 14.4, collected as part of a study to assess the adverse effects of a proposed drug for the treatment of tuberculosis [40].

Ten female rats were given the drug for a period of 14 days at each of five dosage levels (in 100 milligrams per kilogram per day). The vertical axis in the plot

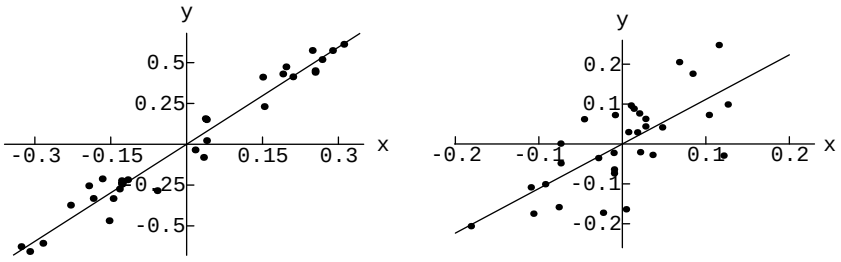

Figure 14.5. *Partial regression plots for the data from the timber yield study. The left plot pictures residuals of log-volume (vertical axis) versus log-diameter (horizontal axis) with the effect of log-height removed. The right plot pictures residuals of log-volume (vertical axis) versus log-height (horizontal axis) with the effect of log-diameter removed. The gray lines are $y = 1.983x$ in the left plot and $y = 1.117x$ in the right plot.*

shows the weight change in grams (WC), defined as the weight at the end of the period minus the weight at the beginning of the period; the horizontal axis shows the dose in 100 mg/kg/day. A linear model of the form

$$\text{WC} = \beta_0 + \beta_1 \text{ dose} + \beta_2 \text{ dose}^2 + \epsilon$$

was fit to the 50 (dose,WC) cases. (The model is linear in the unknown parameters and quadratic in the dosage level.) The LS prediction equation is shown in the plot.

Example: Timber yield study

As part of a study to find an estimate for the volume of a tree (and therefore its yield) given its diameter and height, data were collected on the volume (in cubic feet), diameter at 54 inches above the ground (in inches), and height (in feet) of 31 black cherry trees in the Allegheny National Forest [50, p. 159]. Since a multiplicative relationship is expected among these variables, a linear model of the form

$$\text{log-volume} = \beta_0 + \beta_1 \text{ log-diameter} + \beta_2 \text{ log-height} + \epsilon$$

was fit to the 31 (log-diameter,log-height,log-volume) cases, using the natural logarithm function to compute log values.

The LS prediction equation is

$$\text{log-volume} = -6.632 + 1.983 \text{ log-diameter} + 1.117 \text{ log-height}.$$

Figure 14.5 shows *partial regression* plots of the timber yield data.

(i) In the left plot, the log-volume and log-diameter variables are adjusted to remove the effects of log-height. Specifically, the residuals from the simple linear regression of log-volume on log-height (vertical axis) are plotted against the residuals from the simple linear regression of log-diameter on log-height (horizontal axis). The relationship between the adjusted variables can be described using the linear equation $y = 1.983x$.

(ii) In the right plot, the log-volume and log-height variables are adjusted to remove the effects of log-diameter. The relationship between the adjusted variables can be described using the linear equation $y = 1.117x$.

The slopes of the lines in the partial regression plots correspond to the LS estimates in the prediction equation above. The plots suggest that a linear relationship between the response variable and each of the predictors is reasonable.

14.3.2 Analysis of variance

The linear regression model can be reparametrized as follows:

$$Y_i = \mu + \sum_{j=1}^{p-1} \beta_j (x_{j,i} - \overline{x}_{j\cdot}) + \epsilon_i \text{ for } i = 1, 2, \ldots, N,$$

where μ is the overall mean

$$\mu = \frac{1}{N} \sum_{i=1}^{N} E(Y_i) = \beta_0 + \sum_{j=1}^{p-1} \beta_j \overline{x}_{j\cdot}$$

and $\overline{x}_{j\cdot}$ is the mean of the j^{th} predictor for all j. The difference $E(Y_i) - \mu$ is called the i^{th} *deviation* (or the i^{th} *regression effect*). The sum of the regression effects is zero.

This section develops an analysis of variance f test for the null hypothesis that the regression effects are identically zero (equivalently, a test of the null hypothesis that $\beta_i = 0$ for $i = 1, 2, \ldots, p-1$).

If the null hypothesis is accepted, then the $(p-1)$ predictor variables have no predictive ability; otherwise, they have some predictive ability.

Sources of variation; coefficient of determination

In the first step of the analysis, the sum of squared deviations of the response variables from the mean response (the *total* sum of squares),

$$SS_t = \sum_{i=1}^{N} (Y_i - \overline{Y})^2,$$

is written as the sum of squared deviations of the response variables from the predicted responses (the *error* sum of squares),

$$SS_e = \sum_{i=1}^{N} (Y_i - \widehat{Y}_i)^2,$$

plus the sum of squared deviations of the predicted responses from the mean response (the *model* sum of squares),

$$SS_m = \sum_{i=1}^{N} (\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{N} \left(\sum_{j=1}^{p-1} \widehat{\beta}_j (x_{j,i} - \overline{x}_{j\cdot}) \right)^2 .$$

The ratio of the model to the total sums of squares, $R^2 = SS_m/SS_t$, is called the *coefficient of determination*. R^2 is the proportion of the total variation in the response variable that is explained by the model.

In the simple linear case, R^2 is the same as the square of the sample correlation.

Analysis of variance f test

The *error* mean square is the ratio $MS_e = SS_e/(N-p)$, and the *model* mean square is the ratio $MS_m = SS_m/(p-1)$. The following theorem relates these random variables.

Theorem 14.7 (Distribution Theorem). *Under the general assumptions of this section and if the null hypothesis is true, then the ratio $F = MS_m/MS_e$ has an f ratio distribution with $(p-1)$ and $(N-p)$ degrees of freedom.*

Large values of $F = MS_m/MS_e$ support the hypothesis that the proposed predictor variables have some predictive ability. For an observed ratio, f_{obs}, the p value is $P(F \geq f_{\text{obs}})$.

For the toxicology study example (page 217), $f_{\text{obs}} = 82.3$ and the p value, based on the f ratio distribution with 2 and 47 degrees of freedom, is virtually zero. The coefficient of determination is 0.778; the estimated linear model explains about 77.8% of the variation in weight change.

For the timber yield example (page 218), $f_{\text{obs}} = 613.2$ and the p value, based on the f ratio distribution with 2 and 28 degrees of freedom, is virtually zero. The coefficient of determination is 0.978; the estimated linear model explains about 97.8% of the variation in log-volume.

It is possible for the f test to reject the null hypothesis and the value of R^2 to be close to zero. In this case, the potential predictors have some predictive ability, but additional (or different) predictor variables are needed to adequately model the response.

14.3.3 Confidence interval procedures

This section develops confidence interval procedures for the β parameters and for the mean response at a fixed combination of the predictor variables.

Hypothesis tests can also be developed. Most computer programs automatically include both types of analyses.

Confidence intervals for β_i

Let v_i be the element in the (i,i) position of $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$, and let S^2 be the estimate of the common variance given in Theorem 14.5. Since the LS estimator $\widehat{\beta}_i$ is a normal random variable with mean β_i and variance $\sigma^2 v_i$, Theorem 6.2 can be used to demonstrate that

$$\widehat{\beta}_i \pm t_{N-p}(\gamma/2)\sqrt{S^2 v_i}$$

is a $100(1-\gamma)\%$ confidence interval for β_i, where $t_{N-p}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-p)$ degrees of freedom.

For example, if the data in the toxicology study (page 217) are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for β_2 is $[-0.43153, -0.10008]$. Since 0 is not in the confidence interval, the result suggests that the model with the dose^2 term is significantly better than a simple linear model relating dose to weight change.

Confidence intervals for mean response

The mean response $E(Y_o) = \sum_{i=0}^{p-1} \beta_i x_{i,0}$ at a new predictors-response case can be estimated using the statistic

$$\sum_{i=0}^{p-1} \widehat{\beta}_i x_{i,0} = \underline{x}_0^T \underline{\widehat{\beta}} = \underline{x}_0^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \underline{Y},$$

where $\underline{x}_0^T = (1, x_{1,0}, x_{2,0}, \ldots, x_{p-1,0})$. This estimator is a normal random variable with mean $E(Y_o)$ and variance

$$\sigma^2 \left(\mathbf{x_0}^T \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{x_0}\right) = \sigma^2\, v_o.$$

Thus, Theorem 6.2 can be used to demonstrate that

$$\sum_{i=0}^{p-1} \widehat{\beta}_i x_{i,0} \;\pm\; t_{N-p}(\gamma/2)\sqrt{S^2\, v_0}$$

is a $100(1-\gamma)\%$ confidence interval for $E(Y_o)$, where S^2 is the estimate of the common variance given in Theorem 14.3 and $t_{N-p}(\gamma/2)$ is the $100(1-\gamma/2)\%$ point on the Student t distribution with $(N-p)$ degrees of freedom.

For example, an estimate of the mean log-volume of a tree with diameter 11.5 inches and height 80 inches is 3.106 log-cubic inches. If these data are the values of random variables satisfying the assumptions of this section, then a 95% confidence interval for the mean response at this combination of the predictors is [3.05944, 3.1525].

14.3.4 Regression diagnostics

Recall that the hat matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ is the matrix that transforms the vector of observed responses $\underline{Y}$ to the vector of predicted responses $\underline{\widehat{Y}}$. Each predicted response is a linear combination of the observed responses:

$$\widehat{Y}_i = \sum_{j=1}^{N} h_{i,j} Y_j \;\text{ for } i = 1, 2, \ldots, N,$$

where $h_{i,j}$ is the (i, j) element of $\mathbf{H}$. In particular, the diagonal element $h_{i,i}$ is the coefficient of Y_i in the formula for $\widehat{Y}_i$.

Leverage

The *leverage* of the i^{th} response is the value $h_i = h_{i,i}$. Leverages satisfy the following properties:

1. For each i, $0 \leq h_i \leq 1$.
2. $\sum_{i=1}^{N} h_i = p$, where p is the number of parameters.

Ideally, the leverages should be about p/N each (the average value).

Residuals and standardized residuals

Theorem 14.6 implies that the variance of the i^{th} estimated error (or *residual*) is $Var(\widehat{\epsilon}_i) = \sigma^2(1 - h_i)$, where h_i is the leverage. The i^{th} estimated *standardized residual* is defined as follows:

$$r_i = \widehat{\epsilon}_i / \sqrt{S^2(1 - h_i)} \text{ for } i = 1, 2, \ldots, N,$$

where S^2 is the estimate of the common variance given in Theorem 14.5.

Residuals and standardized residuals are used in diagnostic plots of model assumptions. See Section 14.2.2 for examples in the simple linear case.

Standardized influences

The *influence* of the i^{th} observation is the change in prediction if the i^{th} observation is deleted from the data set.

Specifically, the influence is the difference $\widehat{Y}_i - \widehat{Y}_i(i)$, where $\widehat{Y}_i$ is the predicted response using all N cases to compute parameter estimates, and $\widehat{Y}_i(i)$ is the prediction at a "new" predictor vector $\underline{x}_i$, where parameter estimates have been computed using the list of $N - 1$ cases with the i^{th} case removed.

For the model estimated using $N - 1$ cases only, linear algebra methods can be used to demonstrate that the predicted response is

$$\widehat{Y}_i(i) = \widehat{Y}_i - \widehat{\epsilon}_i \frac{h_i}{1 - h_i}$$

and the estimated common variance is

$$S^2(i) = \frac{1}{N - p - 1} \left((N - p)S^2 - \frac{\widehat{\epsilon}_i^2}{1 - h_i} \right).$$

The i^{th} standardized influence is the ratio of the influence to the standard deviation of the predicted response,

$$\frac{\widehat{Y}_i - \widehat{Y}_i(i)}{SD(\widehat{Y}_i)} = \frac{\widehat{\epsilon}_i h_i/(1 - h_i)}{\sqrt{\sigma^2\, h_i}},$$

and the i^{th} estimated *standardized influence* is the value obtained by substituting $S^2(i)$ for σ^2:

$$\delta_i = \frac{\widehat{Y}_i - \widehat{Y}_i(i)}{\widehat{SD}(\widehat{Y}_i)} = \frac{\widehat{\epsilon}_i h_i/(1 - h_i)}{\sqrt{S(i)^2\, h_i}} = \frac{\widehat{\epsilon}_i \sqrt{h_i}}{S(i)(1 - h_i)}.$$

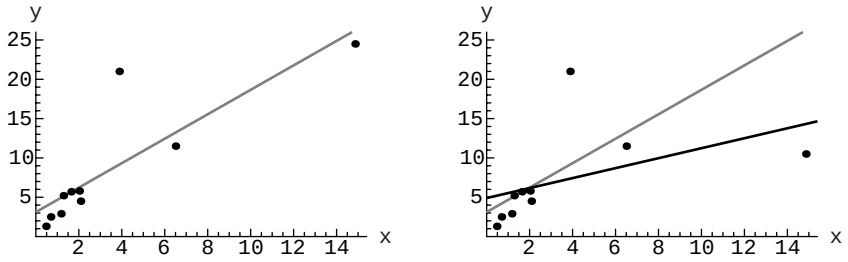

Figure 14.6. *Scatter plots of example pairs (left plot) and altered example pairs (right plot). The gray line in both plots has equation* $y = 3.11 + 1.55x$. *The black line in the right plot has equation* $y = 4.90 + 0.63x$.

Ideally, predicted responses should change very little if one case is removed from the list of N cases, and each δ_i should be close to zero. A general rule of thumb is that if $|\delta_i|$ is much greater than $2\sqrt{p/N}$, then the i^{th} case is highly influential.

Illustration

To illustrate the computations in the simple linear case, consider the following list of 10 (x, y) pairs:

x	0.47	0.69	1.17	1.28	1.64	2.02	2.08	3.88	6.50	14.86
y	1.30	2.50	2.90	5.20	5.70	5.80	4.50	21.00	11.50	24.50

The left plot in Figure 14.6 shows a scatter plot of the data pairs superimposed on the least squares fitted line, $y = 3.11 + 1.55x$. The following table gives the residuals, leverages, and standardized influences for each case:

i	1	2	3	4	5	6	7	8	9	10
$\widehat{\epsilon_i}$	−2.54	−1.68	−2.03	0.10	0.04	−0.45	−1.85	11.86	−1.72	−1.72
h_i	0.15	0.14	0.13	0.13	0.12	0.11	0.11	0.10	0.15	0.85
δ_i	−0.25	−0.16	−0.18	0.01	0.00	−0.04	−0.14	4.15	−0.17	−2.33

Based on the rule of thumb above, cases 8 and 10 are highly influential. Case 8 has a very large residual, and case 10 has a very large leverage value.

The right plot in Figure 14.6 illustrates the concept of leverage. If the observed response in case 10 is changed from 24.5 to 10.5, then the predicted response changes from 26.2 to 14.32. The entire line has moved to accommodate the change in case 10.

Different definitions of δ_i appear in the literature, although most books use the definition above. The rule of thumb is from [12], where the notation DFFITS_i is used for δ_i.

14.4 Bootstrap methods

Bootstrap resampling methods can be applied to analyzing the relationship between one or more predictors and a response. This section introduces two examples.

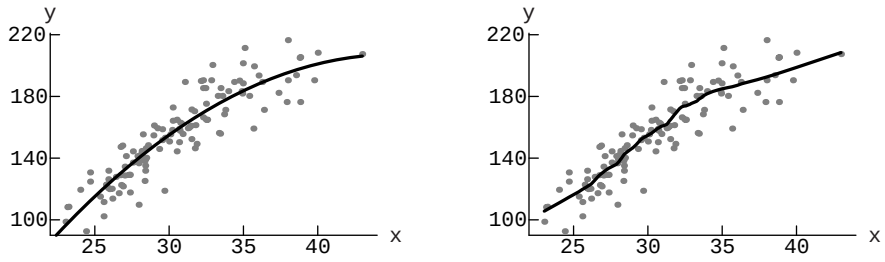

Figure 14.7. *Scatter plots of weight in pounds (vertical axis) versus waist circumference in inches (horizontal axis) for* 120 *physically active young adults. In the left plot, the curve* $y = -252.569 + 20.322x - 0.225x^2$ *is superimposed. In the right plot, the 25% lowess smooth is superimposed.*

Example: Unconditional analysis of linear models

If the observed cases are the values of a random sample from a joint distribution, then nonparametric bootstrap methods can be used to construct confidence intervals for parameters of interest without additional assumptions. (In particular, it is not necessary to condition on the observed values of the predictor variables.) Resampling is done from the list of N observed cases.

For example, the left plot in Figure 14.7 is a scatter plot of waist-weight measurements for 120 physically active young adults (derived from [53]) with a least squares fitted quadratic polynomial superimposed. An estimate of the mean weight for an individual with a 33-inch waist is 174.6 pounds. If the observed (x, y) pairs are the values of a random sample from a joint distribution satisfying a linear model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon,$$

then an approximate 95% confidence interval (based on 5000 resamples) for the mean weight when the waist size is 33 inches is [169.735, 176.944].

Example: Locally weighted regression

Locally weighted regression was introduced by W. Cleveland in the 1970's. Analysis is done conditional on the observed predictor values. In the single predictor case,

$$Y_i = g(x_i) + \epsilon_i \text{ for } i = 1, 2, \ldots, N$$

are assumed to be independent random variables, the function g is assumed to be a differentiable function of *unknown* form, and the collection $\{\epsilon_i\}$ is assumed to be a random sample from a distribution with mean 0 and standard deviation σ.

The goal is to estimate the conditional mean function, $y = g(x)$. Since $g(x)$ is differentiable, and a differentiable function is approximately linear on a small x-interval, the curve can be estimated as follows:

(i) For a given value of the predictor, say x_o, estimate the tangent line to $y = g(x)$ at $x = x_o$, and use the value predicted by the tangent line to estimate $g(x_o)$.

(ii) Repeat this process for all observed predictor values.

For a given x_o, the tangent line is estimated using a method known as weighted linear least squares. Specifically, the intercept and slope of the tangent line are obtained by minimizing the weighted sum of squared deviations

$$S(\alpha, \beta) = \sum_{i=1}^{N} w_i \left(Y_i - (\alpha + \beta x_i)\right)^2,$$

where the weights (w_i) are chosen so that pairs with x-coordinate near x_o have weight approximately 1; pairs with x-coordinate far from x_o have weight 0; and the weights decrease smoothly from 1 to 0 in a "window" centered at x_o.

The user chooses the proportion p of data pairs that will be in the "window" centered at x_o. When the process is repeated for each observed value of the predictor, the resulting estimated curve is called the $100p\%$ *lowess smooth*.

The right plot in Figure 14.7 shows the scatter plot of waist-weight measurements for the 120 physically active young adults with a 25% lowess smooth superimposed. The smoothed curve picks up the general pattern of the relationship between waist and weight measurements.

Lowess smooths allow researchers to approximate the relationship between predictor and response without specifying the function g. A bootstrap analysis can then be done, for example, to construct confidence intervals for the mean response at a fixed value of the predictor.

For the waist-weight pairs, a 25% smooth when $x = 33$ inches produced an estimated mean weight of 175.8 pounds. A bootstrap analysis (with 5000 random resamples) produced an approximate 95% confidence interval for mean weight when the waist size is 33 inches of [168.394, 182.650].

The lowess smooth algorithm implemented above uses tricube weights for smoothing and omits Cleveland's robustness step. For details about the algorithm, see [25, p. 121].

14.5 Laboratory problems

Laboratory problems for this chapter introduce *Mathematica* commands for linear regression analysis, permutation analysis of slope in the simple linear case, locally weighted regression, and diagnostic plots. Problems are designed to reinforce the ideas of this chapter.

14.5.1 Laboratory: Linear least squares analysis

In the main laboratory notebook (Problems 1 to 5), you will use simulation and graphics to study the components of linear least squares analyses; solve a problem on correlated and uncorrelated factors in polynomial regression; and apply linear least squares methods to three data sets from a study of sleep in mammals [3], [30].

14.5.2 Additional problem notebooks

Problems 6, 7, and 8 are applications of simple linear least squares (and other) methods. Problem 6 uses several data sets from an ecology study [32], [77]. Problem 7

uses data from an arsenic study [103]. Problem 8 uses data from a study on ozone exposure in children [113].

Problems 9, 10, and 11 are applications of multiple linear regression (and other) methods. In each case, the *adjusted* coefficient of determination is used to help choose an appropriate prediction model. Problem 9 uses data from a hydrocarbon-emissions study [90]. Problem 10 uses data from a study of factors affecting plasma beta-carotene levels in women [104]. Problem 11 uses data from a study designed to find an empirical formula for predicting body fat in men using easily measured quantities only [59].

Problem 12 applies the goodness-of-fit analysis in simple linear regression to several data sets from a physical anthropology study [50].

Problems 13 and 14 introduce the use of "dummy" variables in linear regression problems. In Problem 13, the methods are applied to a study of the relationship between age and height in two groups of children [5]. In Problem 14, the methods are applied to a study of the pricing of diamonds [26]. Problem 13 also introduces a permutation method for the same problem.

Note that the use of dummy variables in Problem 13 is an example of a *covariance analysis* and the use of dummy variables in Problem 14 is an example of the analysis of an *unbalanced* two-way layout.

Problems 15 and 16 are applications of locally weighted regression and bootstrap methods. Problem 15 uses data from a study of ozone levels in the greater Los Angeles area [28]. Problem 16 uses data from a cholesterol-reduction study [36].