Sect 11.2 The Simple Linear Regression Model

Simple linear regression model

The response Y is related to the independent variable x through the equation

$$Y = \alpha + \beta x + \varepsilon,$$

In the above, α and β are unknown intercept and slope **parameters** respectively and ε is a random variable that is assumed to be normally distributed with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$. The quantity σ^2 is often called the error variance or residual variance.

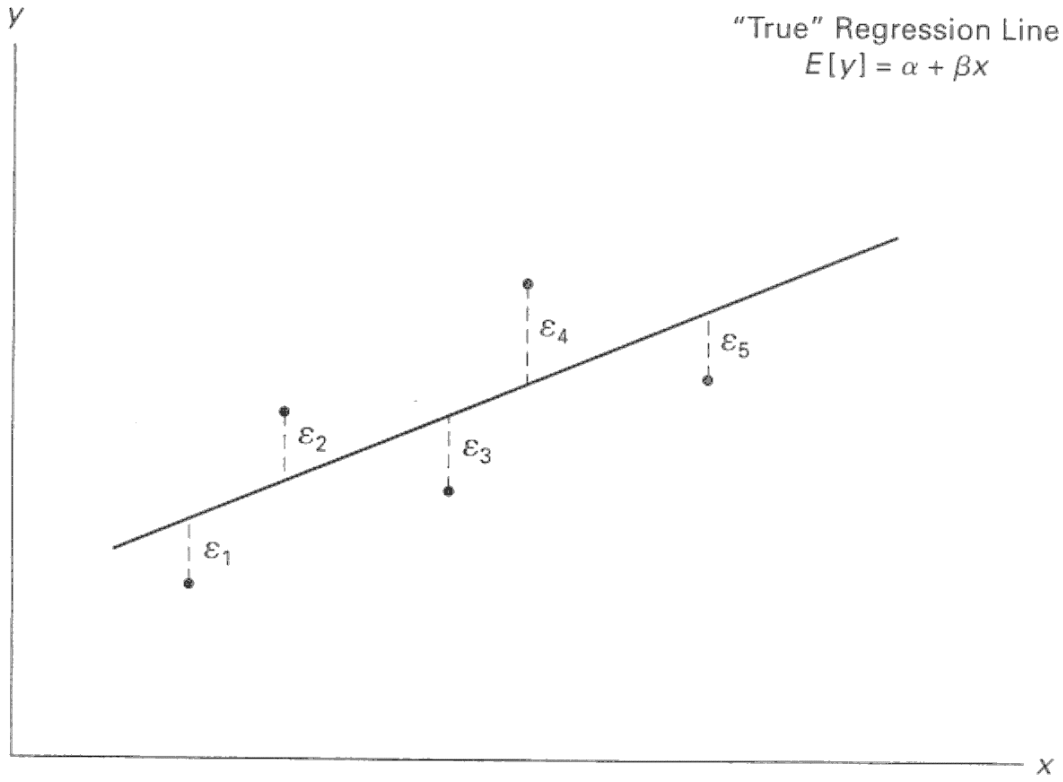

Figure 11.1 Hypothetical *(y, x)* data scattered around the true regression line for $n = 5$.

The Estimated/Fitted Regression Line

The estimated or fitted regression line is given by

$$\hat{y} = a + bx,$$

where $\hat{y}$ is the predicted or fitted value.

- the fitted line is an estimate of the true regression line. One expects that the fitted line should be closer to the true regression line when a large amount of data are available.
- a real life pollution study.

One of the more challenging problems confronting the water pollution control field is presented by the tanning industry. Tannery wastes are chemically complex. They are characterized by high values of biochemical oxygen demand, volatile solids, and other pollution measures. Consider the experimental data of Table 11.1, which was obtained from 33 samples of chemically treated waste in the study conducted at the Virginia Polytechnic Institute and State University. Readings on *x,* the percent reduction in total solids, and *y,* the percent reduction in chemical oxygen demand for the 33 samples, were recorded.

The data of Table 11.1 are plotted in Figure 11.2, showing a scatter diagram. From an inspection of this scatter diagram, it is seen that the points closely follow a straight line, indicating that the assumption of linearity between the two variables appears to be reasonable.

TABLE 1.1 Measures of Solids and Chemical Oxygen Demand

Solids reduction, x (%)	Chemical oxygen demand, y (%)	Solids reduction, x (%)	Chemical oxygen demand, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

The fitted regression line and a hypothetical true regression line are shown on the scatter diagram of Figure 11.2.

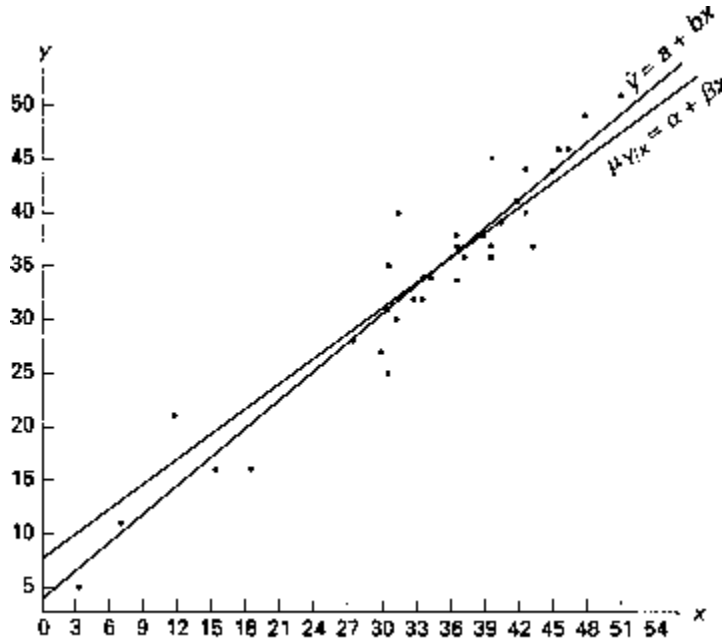

Figure 11.2 Scatter diagram with regression lines.

The Model Assumptions

- **The mean of *Y* given *x*, $\mu_{Y|x} = \alpha + \beta x$, fall on a straight line**
- **All the distribution of *Y* given *x* are normal distribution with the mean of *Y* given *x*, $\mu_{Y|x} = \alpha + \beta x$, ,as the center and variance $\sigma^2_{Y|x}$**
- **All the distribution of *Y* given *x* have the same error variance σ^2, that is, $\sigma^2_{Y|x} = \sigma^2$**

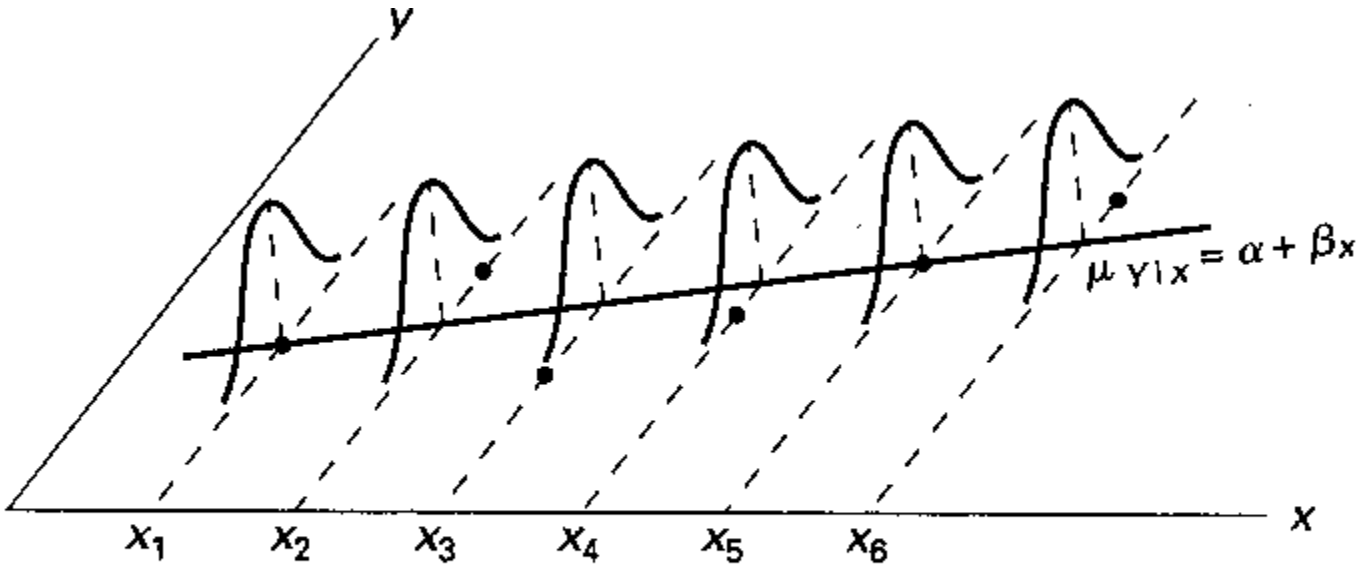

Figure 11.3 Individual observations around true regression line.

S 11.3 Least Squares and the Fitted Model

Residual: error in fit

Given a set of regression data $[(x_i, y_i); i = 1, 2, \ldots n]$, and a fitted model, $\hat{y}_i = a + bx_i$, the i^{th} residual e_i is given by

$$e_i = y_i - \hat{y}_i, \qquad i = 1, 2, \ldots, n.$$

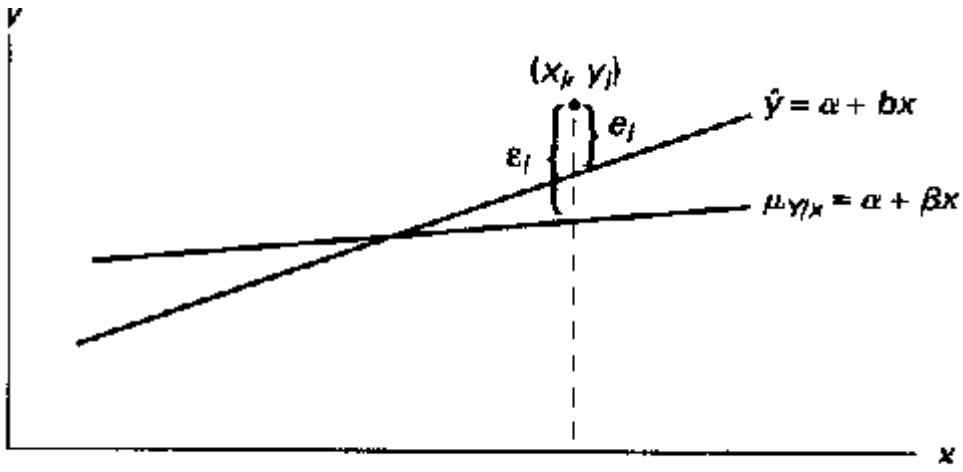

Figure 11.4 Comparing ε_i with the residual, e_i

- **The Method of Least Squares**
  - Target:

    find a and b, the estimates of α and β, so that the sum of the squares of the residuals is a minimum

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Differentiating SSE with respect to a and b, we have

$$\frac{\partial(SSE)}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i), \qquad \frac{\partial(SSE)}{\partial b} = -2\sum_{i=1}^{n} (y_i - a - bx_i)x_i.$$

Setting the partial derivatives equal to zero and rearranging the terms, we obtain the equations (called the **normal equations**)

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

which may be solved simultaneously to yield computing formulas for a and b.

Estimating the regression coefficients

Given the sample $\{(x_i, y_i); i = 1, 2, \ldots, n\}$, the least squares estimates a and b of the regression coefficients α and β are computed from the formulas

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x}.$$

Example 11.1 Estimate the regression line for the pollution data of Table 11.1.

Solution $\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41{,}355, \quad \sum_{i=1}^{33} x_i^2 = 41{,}086.$

Therefore,

$$b = \frac{(33)(41{,}355) - (1104)(1124)}{(33)(41{,}086) - (1104)^2} = 0.903643$$

and

$$a = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Thus the estimated regression line is given by

$$\hat{y} = 3.8296 + 0.9036x.$$

What is Good About "Least Squares"?

- designed to provide a fitted line that results in a "closeness" between the line and the plotted points.
- least squares procedure produces a line that **minimizes the sum of squares of vertical deviations** from the points to the line.

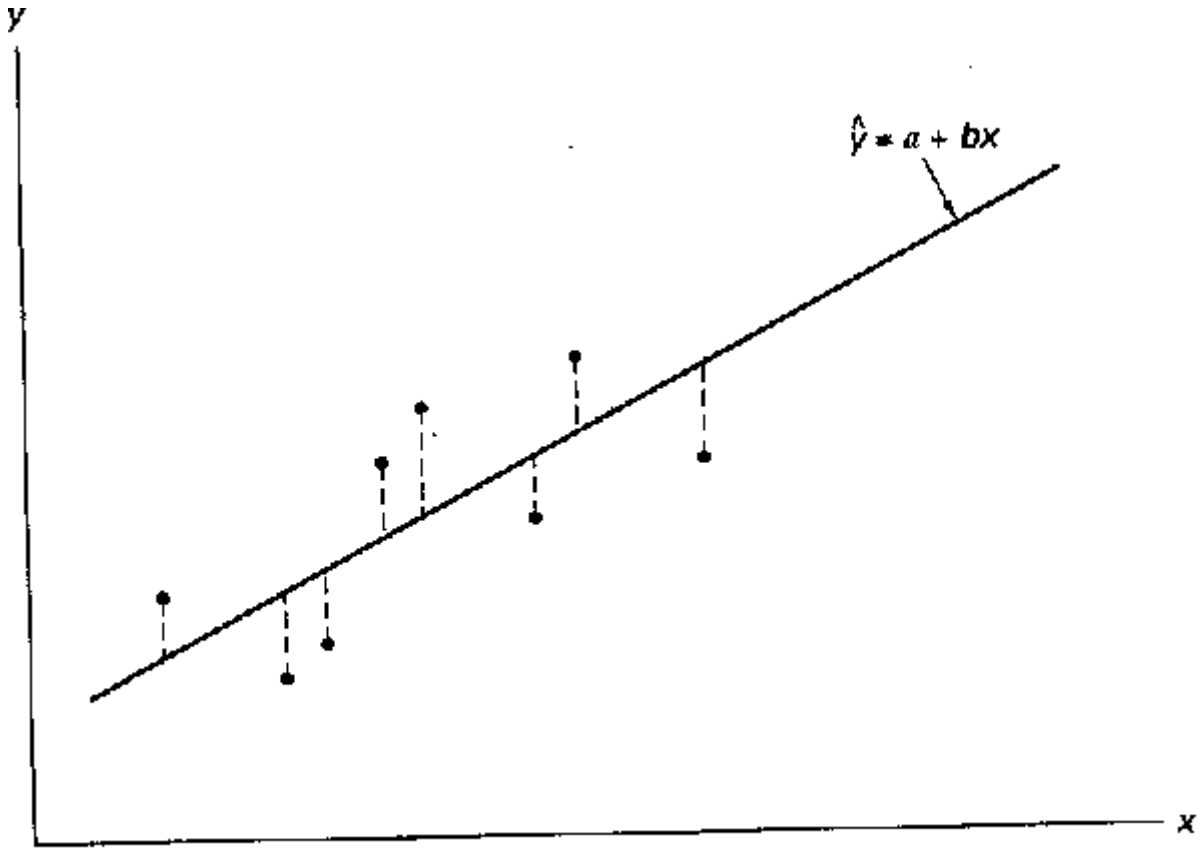

Figure 11.5 Residuals as vertical deviations.

Section 11.4 Properties of the Least Squares Estimators

- **Mean and Variance of Estimators**

Slope parameter

$$\mu_B = E(B) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})\,E(Y_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\alpha + \beta x_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta,$$

and then, using Corollary 3 of Theorem 4.10,

$$\sigma_B^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2\sigma_{Y_i}^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Intercept parameter

$$\mu_A = \alpha \qquad \text{and variance} \qquad \sigma_A^2 = \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2.$$

Both are unbiased estimators.

Partition of total variability and estimation of variance σ^2

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Theorem 11.1 An unbiased estimate of σ^2 is

$$s^2 = \frac{SSE}{n-2} = \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - bS_{xy}}{n-2}.$$

The estimator of σ^2 as Mean square error

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)$$

is called a **mean squared error,** depicting a type of mean (division by $n - 2$) of the squared residuals.