LINEAR REGRESSION

- Regression analysis is a statistical technique that involve exploring the relationship between two or more variables.

- We assume in this section that a random variable Y is a function of only one independent variable and their relationship is linear.

- By a linear relationship we mean that the mean of Y, $E[Y]$, is known to be a linear function of x, that is,

$$E[Y] = \alpha + \beta x \quad .$$

The two constants,

- intercept α and
- slope β

are unknown.

- We will estimate α and β from a sample of Y values with their associated values of x.

Remark:

- $E[Y]$ is a function of x. In any single experiment, x will assume a certain value x_i and the mean Y will take the value,

$$E[Y_i] = \alpha + \beta x_i \quad .$$

- If we define a random variable E by

$$E = Y - (\alpha + \beta x),$$

the random variable Y is a function of x. Indeed,

$$Y = \alpha + \beta x + E$$

where E has a mean, $E[E]=0$ and variance, $\sigma_E^2 = \sigma^2$. σ_E^2 is identical the variance of Y, namely,

$$\sigma_E^2 = \sigma_Y^2 = \sigma^2 .$$

The value of σ^2 is not known, in general, but it is assumed to be constant and not a function of x.

$$\mu_{Y|x} = E[\alpha + \beta x + \epsilon]$$

$$= \alpha + \beta x + \underbrace{E[\epsilon]}_{0}$$

$$= \alpha + \beta x$$

and

$$\sigma^2_{Y|x} = \sigma^2_{\alpha+\beta x+\epsilon}$$

$$= \sigma^2_{\alpha+\beta x} + \sigma^2_{\epsilon}$$

$$= 0 + \sigma^2 = \sigma^2 .$$

- The true regression model

$$E[Y|x] = \mu_{Y|x} = \alpha + \beta x$$

is a line of mean values, namely

the random variable Y is related to x by this straight-line relationship.

$$\boxed{E[Y] = \alpha + \beta x}$$

- The height of the regression line at any value of x just the expected value of Y for that x.
- The slope, β, can be interpreted as the change in the mean of Y for a unit change in x.
- On the other hand, the variability of Y at a particular value of x is determined by the error variance σ^2
  - This means that the distribution of Y-values at each x and that the variance of this distribution is the same at each x, namely, $\sigma^2_{Y|x} = \sigma^2$.

Example:

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

- Consider the data in Table 11.1. In this table y is the purity of oxygen produced in chemical distillation process and x is the percentage of hydrocarbons that are present. The following figure presents a scatter diagram of the data in the table.

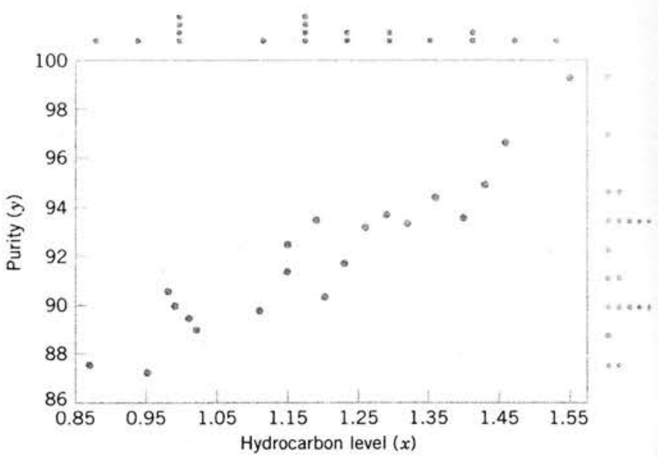


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

- Each (x_i, y_i) pair is represented as a point plotted in a two-dimensional coordinate system.

- Inspection of this scatter diagram indicates that, athough no simple curve will pass exactly through all points, there is a strong indication the points lie scattered randomly around a straight line.

- Therefore, it is reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

$$E[Y \mid x] = \mu_{Y|x} = \alpha + \beta x .$$

- Suppose that the true regression model relating oxygen purity to hydrocarbon level is

$$\mu_{Y|x} = 75 + 15x$$

and suppose that the variance is σ^2.

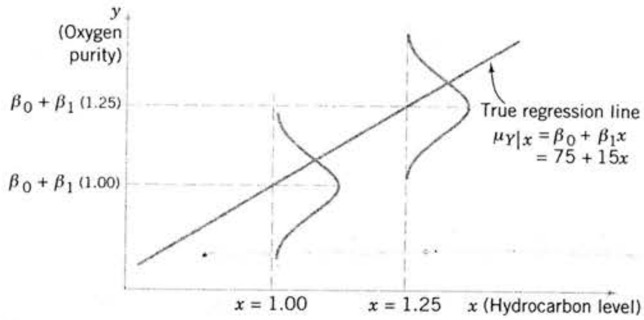


Figure 11-2 The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

- This figure illustrates this situation. Indeed here we have used a normal distribution to describe random variation in E.
  - Since E is normally distributed, $E \sim N(0, \sigma^2)$, Y is a normally distributed random variable.
  - The variance σ^2 determines the variability in the observations Y on oxygen purity.
    - When σ^2 is small, the observed values of Y will fall close to the line.
    - When σ^2 is large, the observed values of Y may deviate considerably from the line.
    - Becouse σ^2 is constant, the variability in Y at any value of x is the same.

- The regression model describes the relationship between oxygen purity Y and hydrocarbon level x.
  - Therefore, for any value of hydrocarbon level, oxygen purity has a normal distribution with mean $75 + 15x$ and variance 2.
    - For example, if $x = 1.25$, Y has mean value

$$\mu_{Y|x} = 75 + 15(1.25) = 93.75$$

and variance 2.

Remark:

- I most real-world problems the values of the intercept and slope (α, β) and the error variance σ^2 will not be known.
- They must be estimated from sample data.
- Gauss proposed the method of least squares in order to estimate the parameters α and β.

Least-Squares Method of Estimation

The estimation of regression parameters α and β can be made by the method of least squares. Their estimation $\hat{\alpha}$ and $\hat{\beta}$, be chosen so that the sum of the squared differences between observed sample values y_i and the estimated expected value of Y, $\hat{\alpha}+\hat{\beta}x_i$, is minimized.

Let us write

$$e_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

The least-square estimates $\hat{\alpha}$ and $\hat{\beta}$ are found by minimizing

$$Q = \sum_{i=1}^{n} e_i^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta}x_i)\right)^2,$$

where $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are n pairs of observations and e_i, $i=1,2,\ldots,n$ are called residuals.

The following figure gives a grafical representation of the least-squares method.

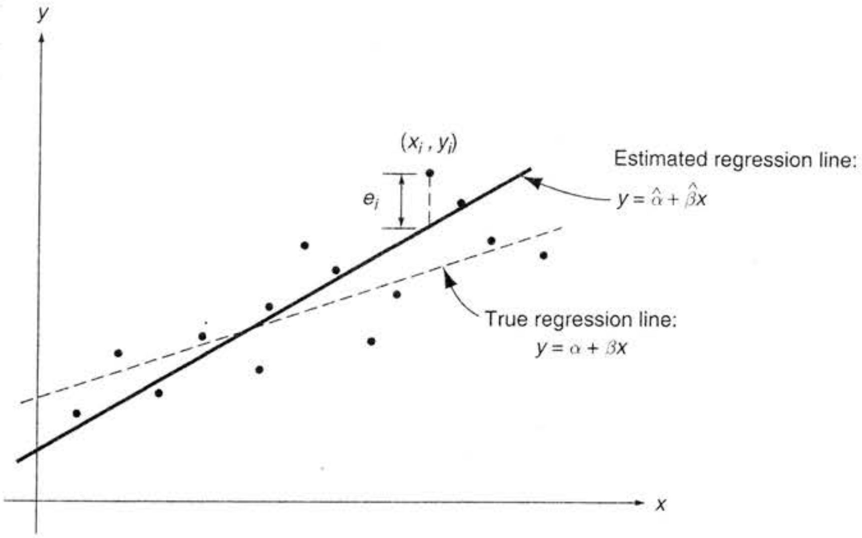


Figure 11.1 The least squares method of estimation

- We observe that the residuals are the vertical distances between the observed values of Y, y_i and the least-square estimate $\hat{\alpha} + \hat{\beta}x$ of the true regression line $\alpha + \beta x$.

Theorem:

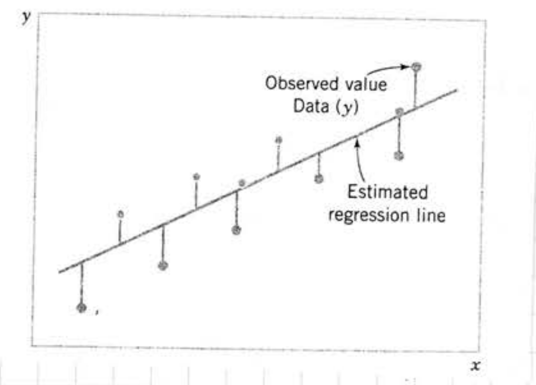
The least-squares estimates of α and β in the simple linear regression model are,

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\bar{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



and

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad .$$

We can also write,

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

and

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

$$= \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} .$$

ibis

Example:

We can fit a simple linear regression model to the oxygen purity data in Table 11.1.

- We need following quantities:

$n = 20$

$$\sum_{i=1}^{20} x_i = 23.92 \Rightarrow \bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i$$

$$= \frac{1}{20}(23.92)$$

$$= 1.1960$$

$$\sum_{i=1}^{20} y_i = 1843.21 \Rightarrow \bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i$$

$$= \frac{1}{20}(1843.21)$$

$$= 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170\,044.531$$

$$\sum_{i=1}^{20} x_i^2 = 29.2892$$

$$\sum_{i=1}^{20} x_i y_i = 2\,214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20}$$

$$= 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20}$$

$$= 2\,214.6566 - \frac{(23.92)(1\,843.21)}{20} = 10.17744$$

- The substitution of these values into equations gives

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.947$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\,\bar{x}$$
$$= 92.1605 - (14.947)(1.1960) = 74.283$$

The fitted simple regression model is

$$\hat{y} = 74.283 + 14.947x .$$

The estimated regression line together with observed data is shown in Figure 11.4

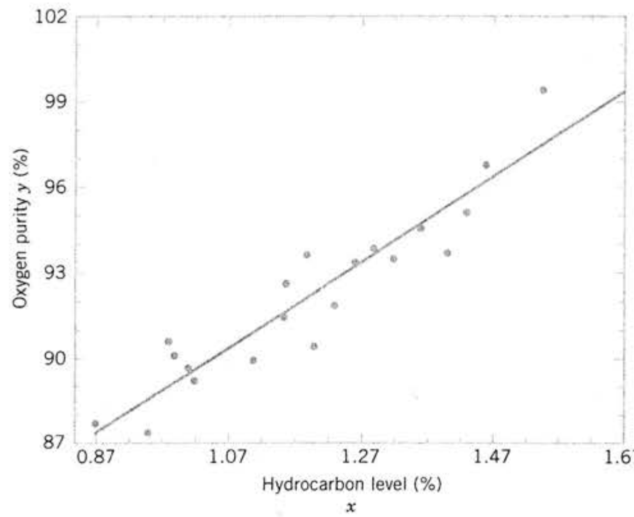


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

Remark:

- Using this model, we can predict oxygen purity of $\hat{y} = 89.23\%$ when the hydrocarbon level is $x = 1.00\%$.
- The purity 89.23 % may be interpreted as an estimate of the true population mean purity when $x = 1.00\%$, or as an estimate of a new observation when $x = 1.00\%$. These estimates are, of course, subject to error.

Estimating the variance σ^2:

- The variance of the error term E can be estimated from the residuals

$$e_i = y_i - \hat{y}_i .$$

- The unbies estimate of the variance is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - (\hat{\alpha} + \hat{\beta} x_i) \right)^2 .$$

- For the oxygen purity data in the above example, we get $\hat{\sigma} = 1.18$.

Properties of The Least Squares Estimators

- We have assumed that the error term E in the model, $Y = \alpha + \beta x + E$ is a random variable with zero mean and variance σ^2.
- Since the values of x are fixed, Y is a random variable with mean $\mu_{Y|x} = \alpha + \beta x$ and variance σ^2.

- The regression coefficients $\hat{\alpha}$ and $\hat{\beta}$ depend on the observed y's. Hence, the least-squares estimates $\hat{\alpha}$ and $\hat{\beta}$ may be viewed as random variables.

The Mean of $\hat{\beta}$, $E[\hat{\beta}]$:

- $\hat{\beta}$ is a linear combination of the observation Y_i, therefore it can be shown that

$$E[\hat{\beta}] = \beta .$$

This means that $\hat{\beta}$ is an unbised estimator of the true slope β.

The variance of $\hat{\beta}$, $\sigma_{\hat{\beta}}^2$:

- We have assumed that $\sigma_{E_i}^2 = \sigma^2$, it follows that $\sigma_{Y_i}^2 = \sigma^2$. The we show that

$$\sigma_{\hat{\beta}}^2 = \frac{\sigma^2}{S_{xx}} .$$

- We can show also,

$$E[\hat{\alpha}] = \alpha$$

and

$$\sigma_{\hat{\alpha}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$$

- $\hat{\alpha}$ is an unbiased estimator of α.

- The covariance of the random variables $\hat{\alpha}$ and $\hat{\beta}$ is not zero. It can be shown that

$$Cov(\hat{\alpha}, \hat{\beta}) = -\frac{\sigma^2 \bar{x}}{S_{xx}} .$$

Remark:

The estimate of σ^2 can be used in these above equations. We call the square roots of the resulting variance estimator as the estimated standard errors of $\hat{\alpha}$ and $\hat{\beta}$:

$$se(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

and $$se(\hat{\alpha}) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} .$$