CHAPTER SIX

Fundamental Statistical Concepts

There are three kinds of lies: lies, damn lies, and statistics.
Benjamin Disraeli (1804–1881)

Statistics is vital to any scientific discipline that is confronted with the task of summarizing data and making inferences from them. This elementary chapter presents notations and results in probability and statistics that will be useful or extended later.

6.1 Basics

Several definitions are related to expressing the variability of a random variable. The **variance** σ_X^2 of a random variable X is defined as

$$\mathrm{Var}[\,X\,] \equiv E[\,(X - E[\,X\,])^2\,].$$

The **standard deviation** σ_X is the square root of the variance, $\sqrt{\mathrm{Var}[\,X\,]}$. The **skewness** of X with mean μ is $E[\,(X-\mu)^3/\sigma^3\,]$, and the **kurtosis** is $E[\,(X-\mu)^4/\sigma^4\,]$.

The **sample mean** of a random sample $X_1, X_2, \ldots, X_n$ is

$$\overline{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i.$$

A measure of a random sample's variability is its **sample variance**, defined as

$$\overline{\sigma^2} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{6.1}$$

The **sample standard deviation** $\overline{\sigma}$ is the square root of the sample variance:

$$\overline{\sigma} \equiv \sqrt{\overline{\sigma^2}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}\,. \tag{6.2}$$

An estimator for a parameter θ is said to be **unbiased** if its expected value equals θ. Sample variance (6.1) is an unbiased estimator of σ_X^2 when each random sample X_i has the same distribution as X. Although the sample standard deviation is a biased estimator of the standard deviation, it converges to the unbiased one.

The **covariance** between two random variables X and Y is defined by

$$\text{Cov}[\,X, Y\,] \equiv E\,[\,(X-\mu_X)(Y-\mu_Y)\,]\,,$$

where μ_X and μ_Y are the means of X and Y, respectively. If X and Y tend to move in the same direction, their covariance will be positive, whereas if they tend to move in opposite directions, their covariance will be negative. Call X and Y **uncorrelated** random variables if $\text{Cov}[\,X, Y\,] = 0$. A computational shortcut for covariance is

$$\text{Cov}[\,X, Y\,] = E[\,XY\,] - \mu_X\mu_Y. \tag{6.3}$$

The **correlation** (or **correlation coefficient**) between X and Y is

$$\rho_{X,Y} \equiv \frac{\text{Cov}[\,X, Y\,]}{\sigma_X\sigma_Y}, \tag{6.4}$$

provided that both have nonzero standard deviations. The variance of a weighted sum of random variables satisfies

$$\text{Var}\left[\,\sum_{i=1}^{n} a_i\,X_i\,\right] = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j\,\text{Cov}[\,X_i,\,X_j\,]. \tag{6.5}$$

The above becomes $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i^2\,\text{Var}[\,X_i\,]$ when X_i are uncorrelated.

Let $X\,|\,I$ denote X conditional on the **information set** I. The information set can be another random variable's value or the past values of X, for example. The **conditional expectation** $E[X\,|\,I\,]$ is the expected value of X conditional on I. Note that it is a random variable. The extremely useful **law of iterated conditional expectations** says that

$$E[\,X\,] = E[\,E[\,X\,|\,I\,]\,].$$

More generally, if I_2 contains at least as much information as I_1, then

$$E[\,X\,|\,I_1\,] = E[\,E[\,X\,|\,I_2\,]\,|\,I_1\,]. \tag{6.6}$$

A typical example is for I_1 to contain price information up to time t_1 and for I_2 to contain price information up to a later time t_2.

➤ **Exercise 6.1.1** Prove Eq. (6.3) by using the well-known identity $E[\,\sum_i a_i X_i\,] = \sum_i a_i\,E[\,X_i\,]$.

➤ **Exercise 6.1.2** Prove that if $E[\,X\,|\,Y = y\,] = E[\,X\,]$ for all realizations y, then X and Y are uncorrelated. (Hint: Use the law of iterated conditional expectations.)

6.1.1 Generalization to Higher Dimensions

It is straightforward to generalize the above concepts to higher dimensions. Let $\boldsymbol{X} \equiv [\,X_1, X_2, \ldots, X_n\,]^{\mathrm{T}}$ be a vector random variable (A^{T} means the transpose of A). Its **mean vector** and the $n \times n$ **covariance matrix** are defined, respectively, as

$$E[\,\boldsymbol{X}\,] \equiv [\,E[\,X_1\,], E[\,X_2\,], \ldots, E[\,X_n\,]\,]^{\mathrm{T}},$$
$$\text{Cov}[\,\boldsymbol{X}\,] \equiv [\,\text{Cov}[\,X_i,\,X_j\,]\,]_{1\le i,\,j\le n}.$$

In analogy with Eq. (6.3), $\mathrm{Cov}[\,\boldsymbol{X}\,] = E[\,\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\,] - E[\,\boldsymbol{X}\,]\,E[\,\boldsymbol{X}\,]^{\mathrm{T}}$. The **correlation matrix** is defined as $[\,\rho_{X_i,X_j}\,]_{1\le i,j\le n}$. Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$ be N observations on $\boldsymbol{X}$. The **sample mean vector** and the **sample covariance matrix** are defined, respectively, as

$$\overline{\boldsymbol{X}} \equiv \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{X}_i, \quad \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{X}_i - \overline{\boldsymbol{X}})(\boldsymbol{X}_i - \overline{\boldsymbol{X}})^{\mathrm{T}}.$$

The sample covariance matrix is an unbiased estimator of the covariance matrix.

➤ **Exercise 6.1.3** Prove that $E[\,A\boldsymbol{X}\,] = AE[\,\boldsymbol{X}\,]$ and that $\mathrm{Cov}[\,A\boldsymbol{X}\,] = A\,\mathrm{Cov}[\,\boldsymbol{X}\,]\,A^{\mathrm{T}}$.

6.1.2 The Normal Distribution

A random variable X is said to have **normal distribution** with mean μ and variance σ^2 if its probability density function is $e^{-(x-\mu)^2/(2\sigma^2)}/(\sigma\sqrt{2\pi})$. This fact is expressed by $X \sim N(\mu, \sigma^2)$ where $\sim$ means equality in distribution. The **standard normal distribution** has zero mean, unit variance, and the distribution function

$$N(z) \equiv \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-x^2/2}\,dx.$$

(The **distribution function** of a random variable X is defined as $F(z) \equiv \mathrm{Prob}[\,X \le z\,]$ for any real number z.) The normal distribution is due to de Moivre (1667–1754).

There are fast and accurate approximations to $N(z)$ [5, 423]. An example is

$$N(z) \approx 1 - \frac{1}{\sqrt{2\pi}}\,e^{-z^2/2}(a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5)$$

for $z \ge 0$, where $x \equiv 1/(1 + 0.2316419z)$ and

$$\begin{aligned} a_1 &= 0.319381530, \quad & a_3 &= 1.781477937, \quad & a_5 &= 1.330274429, \\ a_2 &= -0.356563782, \quad & a_4 &= -1.821255978. \end{aligned}$$

As for $z < 0$, use $N(z) = 1 - N(-z)$.

The **central moments** of the normal random variable X are

$$E[\,(X-\mu)^{2n}\,] = \frac{(2n)!}{2^n n!}\,\sigma^{2n}, \quad E[\,(X-\mu)^{2n+1}\,] = 0, \tag{6.7}$$

where $n = 0, 1, 2\ldots$. For example, the skewness and the kurtosis of the standard normal distribution are zero and three, respectively. The **moment generating function** of a random variable X is defined as $\theta_X(t) \equiv E[\,e^{tX}\,]$. The moment generating function of $X \sim N(\mu, \sigma^2)$ is known to be

$$\theta_X(t) = \exp\left[\,\mu t + \frac{\sigma^2 t^2}{2}\,\right]. \tag{6.8}$$

If $X_i \sim N(\mu_i, \sigma_i^2)$ are independent (or, equivalently for normal distributions, uncorrelated), then $\sum_i X_i$ has a normal distribution with mean $\sum_i \mu_i$ and variance $\sum_i \sigma_i^2$. In general, let $X_i \sim N(\mu_i, \sigma_i^2)$, which may not be independent. Then $\sum_{i=1}^n t_i X_i$ is normally distributed with mean $\sum_{i=1}^n t_i\mu_i$ and variance $\sum_{i=1}^n \sum_{j=1}^n t_i t_j\,\mathrm{Cov}[\,X_i, X_j\,]$ [343]. The joint distribution of $X_1, X_2, \ldots, X_n$ has this

joint moment generating function:

$$E\left[\exp\left[\sum_{i}^{n} t_i X_i\right]\right] = \exp\left[\sum_{i=1}^{n} t_i \mu_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} t_i t_j \operatorname{Cov}[X_i, X_j]\right].$$

These X_i are said to have a **multivariate normal distribution**. We use $\boldsymbol{X} \sim N(\boldsymbol{\mu}, C)$ to denote that $\boldsymbol{X} \equiv [X_1, X_2, \ldots, X_n]^{\mathrm{T}}$ has a multivariate normal distribution with mean $\boldsymbol{\mu} \equiv [\mu_1, \mu_2, \ldots, \mu_n]^{\mathrm{T}}$ and covariance matrix $C \equiv [\operatorname{Cov}[X_i, X_j]]_{1 \le i,j \le n}$. With $M \equiv C^{-1}$ and the (i, j)th entry of the matrix M being $M_{i,j}$, the $\boldsymbol{X}$'s probability density function is

$$\frac{1}{\sqrt{(2\pi)^n \det(C)}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - \mu_i) M_{ij} (X_j - \mu_j)\right],$$

where $\det(C)$ denotes the determinant of C [23].

In particular, if X_1 and X_2 have a **bivariate normal distribution** with correlation ρ, their joint probability density function is

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{\left(\frac{X_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{X_1-\mu_1}{\sigma_1}\right)\left(\frac{X_2-\mu_2}{\sigma_2}\right) + \left(\frac{X_2-\mu_2}{\sigma_2}\right)^2}{2(1-\rho^2)}\right].$$

The sum $\omega_1 X_1 + \omega_2 X_2$ is normally distributed with mean $\omega_1\mu_1 + \omega_2\mu_2$ and variance

$$\omega_1^2\sigma_1^2 + 2\omega_1\omega_2\rho\sigma_1\sigma_2 + \omega_2^2\sigma_2^2. \tag{6.9}$$

Fast and accurate approximations to the bivariate normal random variable's distribution function are available [470].

If $X_i \sim N(\mu_i, \sigma^2)$ are independent, then $Y \equiv \sum_{i=1}^{n} X_i^2/\sigma^2$ has a **noncentral chi-square distribution** with n **degrees of freedom** and **noncentrality parameter** $\theta \equiv (\sum_{i=1}^{n} \mu_i^2)/\sigma^2 > 0$, denoted by $Y \sim \chi(n, \theta)$. The mean and the variance are $n + \theta$ and $2n + 4\theta$, respectively [463]. When μ_i are zero, Y has the **central chi-square distribution**.

The **central limit theorem**, which is due to Laplace (1749–1827), is a cornerstone for probability and statistics. It says that, if X_i are independent with mean μ and variance σ^2, then

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \to N(0, 1).$$

Conditions for the theorem's applicability are rather mild [343].

➤ **Exercise 6.1.4** Prove that central moments (6.7) are equivalent to

$$E[(X-\mu)^n] = \begin{cases} 0, & \text{if } n \ge 1 \text{ is odd} \\ 1 \cdot 3 \cdot 5 \cdots (n-1)\,\sigma^n, & \text{if } n \ge 2 \text{ is even} \end{cases},$$

where $n = 1, 2, \ldots$.

6.1.3 Generation of Univariate and Bivariate Normal Distributions

Let X be uniformly distributed over $(0, 1]$ so that $\text{Prob}[X \le x] = x$ for $0 < x \le 1$. Repeatedly draw two samples x_1 and x_2 from X until $\omega \equiv (2x_1 - 1)^2 + (2x_2 - 1)^2 < 1$. Then $c(2x_1 - 1)$ and $c(2x_2 - 1)$ are independent standard normal variables, where

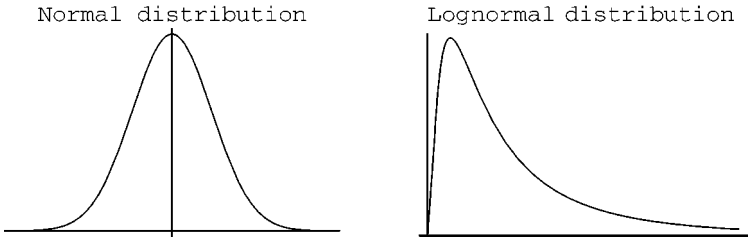

Figure 6.1: Normal and lognormal distributions: the standard normal distribution X and the lognormal distribution e^X.

$c \equiv \sqrt{-2(\ln \omega)/\omega}$. This is called the **polar rejection method** [727]. Pairs of normally distributed variables with correlation ρ can be generated as follows. Let X_1 and X_2 be independent standard normal variables. Then

$$U \equiv a X_1,$$
$$V \equiv \rho U + \sqrt{1-\rho^2}\, a X_2$$

are the desired random variables because $\mathrm{Var}[\,U\,] = \mathrm{Var}[\,V\,] = a^2$ and $\mathrm{Cov}[\,U, V\,] = \rho a^2$.

6.1.4 The Lognormal Distribution

A random variable Y is said to have a **lognormal distribution** if $\ln Y$ has a normal distribution (see Fig. 6.1). This distribution is due to Bachelier [147].

If X is normally distributed with mean μ and variance σ^2, then the density function of the lognormally distributed random variable $Y \equiv e^X$ is

$$f(y) \equiv \begin{cases} \frac{1}{\sigma y \sqrt{2\pi}}\, e^{-(\ln y - \mu)^2/(2\sigma^2)}, & \text{if } y > 0 \\ 0, & \text{if } y \le 0 \end{cases}. \tag{6.10}$$

The mean and the variance of Y are

$$\mu_Y = e^{\mu + \sigma^2/2}, \quad \sigma_Y^2 = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right), \tag{6.11}$$

respectively. Furthermore,

$$\mathrm{Prob}[\,Y \le y\,] = \mathrm{Prob}[\,X \le \ln y\,] = N\left(\frac{\ln y - \mu}{\sigma}\right). \tag{6.12}$$

The nth **moment** about the origin, defined as $\int_{-\infty}^{\infty} x^n f(x)\, dx$ for a random variable x with density function $f(x)$, is $e^{n\mu + n^2\sigma^2/2}$ for Y. A version of the central limit theorem states that the product of n independent positive random variables approaches a lognormal distribution as n goes to infinity.

➤ **Exercise 6.1.5** Let Y be lognormally distributed with mean μ and variance σ^2. Show that $\ln Y$ has mean $\ln[\,\mu/\sqrt{1+(\sigma/\mu)^2}\,]$ and variance $\ln[\,1+(\sigma/\mu)^2\,]$.

➤ **Exercise 6.1.6** Let X be a lognormal random variable such that $\ln X$ has mean μ and variance σ^2. Prove the identity $\int_a^{\infty} x f(x)\, dx = e^{\mu+\sigma^2/2}\, N(\frac{\mu - \ln a}{\sigma} + \sigma)$.

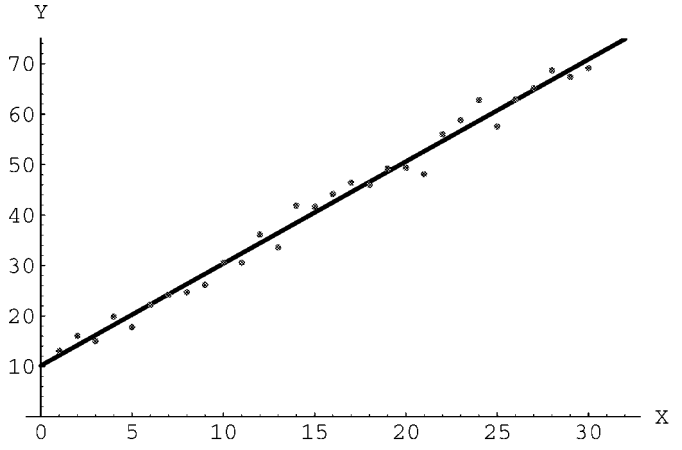

Figure 6.2: Linear regression. The linear function $Y = 10.1402 + 2.0238X$ is fit to the data under the least-squares criterion.

6.2 Regression

Suppose we are presented with the data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. The data can be plotted on a rectangular coordinate system, resulting in the so-called **scatter diagram**, such as the dots in Fig. 6.2. If the scatter diagram suggests a linear relation between the variables, we can fit a simple straight-line model $y = \beta_0 + \beta_1 x$ to the data. The problem of finding such a fit is called **linear regression**.[1] To estimate the model parameters β_0 and β_1 with the **least-squares principle**, we find $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} [\, y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x) \,]^2. \tag{6.13}$$

This line is called the linear regression of y on x [632]. It is well known that

$$\widehat{\beta}_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (x_i - \overline{x})^2} = \frac{n \sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{n \sum_i x_i^2 - \left(\sum_i x_i\right)^2}, \tag{6.14}$$

$$\widehat{\beta}_0 = \frac{\sum_i y_i - \widehat{\beta}_1 \sum_i x_i}{n} = \overline{y} - \widehat{\beta}_1 \overline{x}. \tag{6.15}$$

The resulting line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is called the **estimated regression line** or the **least-squares line**. The ith **fitted value** is $\widehat{y}_i \equiv \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Note that $(\overline{x}, \overline{y})$ is on the estimated regression line by virtue of Eq. (6.15).

A few statistics are commonly used. The **error sum of squares** (**SSE**) is the sum of the squared deviation about the estimated regression line:

$$\text{SSE} \equiv \sum_i (y_i - \widehat{y}_i)^2 = \sum_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Because the SSE measures how much variation in y is *not* explained by the linear regression model, it is also called the **residual sum of squares** or the **unexplained variation**. The **total sum of squares** (**SST**) is defined as $\text{SST} \equiv \sum_i (y_i - \overline{y})^2$, which measures the total amount of variation in observed y values. This value is also

known as the **total variation**. By the least-squares criterion, SSE $\leq$ SST. The ratio SSE/SST is the proportion of the total variation that is left unexplained by the linear regression model. The **coefficient of determination** is defined as

$$R^2 \equiv 1 - \frac{\text{SSE}}{\text{SST}}; \tag{6.16}$$

it is the proportion of the total variation that can be explained by the linear regression model. A high R^2 is typically a sign of success of the linear regression model in explaining the y variation. Finally, the **regression sum of squares** (**SSR**) is defined as $\text{SSR} \equiv \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$. It is well known that

$$\text{SSR} = \text{SST} - \text{SSE} = \widehat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{6.17}$$

Thus $R^2 = \text{SSR/SST}$. Because the SSR is large when the estimated regression line fits the data closely (as SSE is small), it is interpreted as the amount of total variation that is explained by the linear regression model. For this reason it is sometimes called the **explained variation**.

The more general linear regression, also known as **multiple regression**, fits

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

to the data. Equation (6.17) holds for multiple regression as well [422, 523]. Nonlinear regression uses nonlinear regression functions. In **polynomial regression**, for example, the problem is to fit

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

to the data. See Fig. 6.3 for the $k = 2$ case.

➤ **Exercise 6.2.1** Prove that $\text{SSE} = \sum_i y_i^2 - \widehat{\beta}_0 \sum_i y_i - \widehat{\beta}_1 \sum_i x_i y_i$.

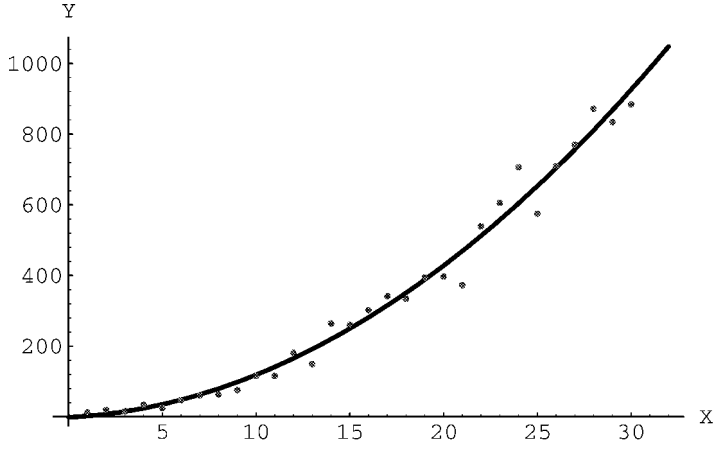

Figure 6.3: Nonlinear regression. The quadratic function $Y = -1.28204 + 2.52945X + 0.945518X^2$ is fit to the data under the least-squares criterion.

6.3 Correlation

Given n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on (X, Y), their **sample correlation coefficient** or **Pearson's** r is definedas

$$r \equiv \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2}\sqrt{\sum_i (y_i - \overline{y})^2}}. \tag{6.18}$$

The sample correlation coefficient is a point estimator for $\rho_{X,Y}$ and is traditionally used to summarize the strength of correlation. It can be shown that $-1 \le r \le 1$. In particular, $r = 1$ when the data lie on a straight line with positive slope and $r = -1$ when the data lie on a straight line with negative slope. In some sense r measures the *linear* relation between the variables.

In regression, one variable is considered dependent and the others independent; the purpose is to predict. Correlation analysis, in contrast, studies how strongly two or more variables are related, and the variables are treated symmetrically; it does not matter which of the two variables is called x and which y.

We used the symbol r deliberately: Squaring r gives exactly the coefficient of determination R^2. Indeed, from Eqs. (6.14) and (6.17),

$$r^2 = \widehat{\beta}_1^2 \frac{\sum_i (x_i - \overline{x})^2}{\sum_i (y_i - \overline{y})^2} = \widehat{\beta}_1 \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sum_i (y_i - \overline{y})^2} = \frac{\text{SSR}}{\text{SST}} = R^2. \tag{6.19}$$

Interestingly, Eq. (6.16) implies that $\text{SSE} = \text{SST} \times (1 - r^2)$.

EXAMPLE 6.3.1 Figure 6.4 plots the stock prices of Intel, Silicon Graphics, Inc. (SGI), VLSI Technology, and Wal-Mart from August 30, 1993, to August 30, 1995. The sample correlation coefficient between VLSI Technology and Intel is extremely high at 0.950376. The sample correlation coefficient between Intel and SGI is also high at 0.883291. Technology stocks seem to move together. In contrast, the sample correlation coefficient between Intel and Wal-Mart is low at 0.14917. From these numbers and Eq. (6.19), we can deduce, for example, that 90.3215% of the total variations between Intel's and VLSI Technology's stock prices can be explained by a linear regression model.

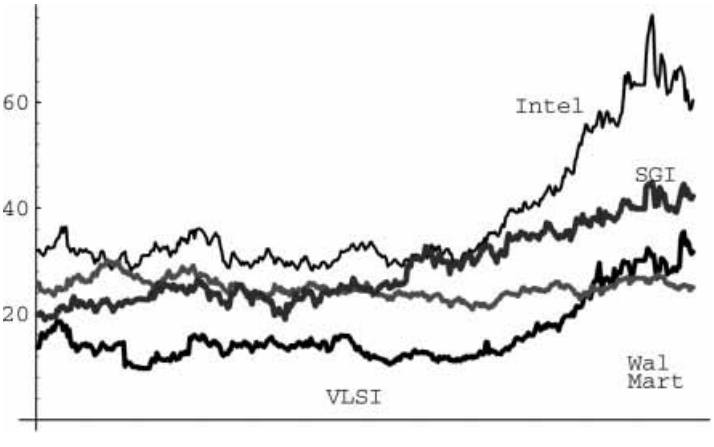

Figure 6.4: Correlation among stock prices. See Example 6.3.1.

➤ **Exercise 6.3.1** Find the estimated regression line for $\{(1, 1.0), (2, 1.5), (3, 1.7), (4, 2.0)\}$. Check that the coefficient of determination indeed equals the sample correlation coefficient.

6.4 Parameter Estimation

After a family of stochastic models has been chosen to capture the reality, the values of their parameters must be found to completely specify the distribution. Inferring those parameters constitutes the major goal of financial econometrics [147]. Three estimation techniques are mentioned below.

6.4.1 The Least-Squares Method

This method is due to Legendre (1752–1833) in 1806 and Gauss (1777–1855) in 1809 [582].[2] It works by minimizing the sum of squares of the deviations, in other words, the SSE. For example, the least-squares estimate of X, given the measurements x_i on it, is the number $\widehat{X}$ that minimizes

$$f(\widehat{X}) \equiv \sum_{i=1}^{n} (x_i - \widehat{X})^2. \tag{6.20}$$

This method was also used in the derivation of the estimated regression line in Section 6.2 by the minimization of (6.13). Recall that no stochastic models were assumed there.

Suppose that the following **linear regression model** is postulated between x and y:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where ϵ is a random variable with zero mean and finite variance. In other words, added to each observation of y is some uncorrelated noise ϵ. Then the estimated parameters of the estimated regression line, which are now random variables, have the smallest variances among all unbiased *linear* estimators. This is the **Gauss–Markov theorem**, which is due to Gauss in 1821 and Markov (1856–1922) in 1912 [75, 632]. It is interesting to observe that the least-squares estimate of β_1 – the $\widehat{\beta}_1$ in Eq. (6.14) – can be interpreted as the sample covariance between x and y divided by the sample variance of x (see also Exercise 6.4.1).

EXAMPLE 6.4.1 Two nice properties of the bivariate normal distribution are

$$E[X_2 \mid X_1] = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(X_1 - \mu_1), \quad \text{Var}[X_2 \mid X_1] = (1 - \rho^2)\sigma_2^2.$$

Hence the regressions are *linear* functions, and linear regression is justified. In fact, the fitted (predicted) value for X_2, given $X_1 = x$ for any two random variables X_1 and X_2, is exactly $E[X_2 \mid X_1 = x]$ under the least-squares principle (see Exercise 6.4.3) [846].

➤ **Exercise 6.4.1** Let X_1 and X_2 be random variables. The random variable

$$Y \equiv (X_2 - E[X_2]) - \{\alpha + \beta(X_1 - E[X_1])\}$$

is the prediction error of the linear prediction $\alpha + \beta(X_1 - E[X_1])$ of X_2 based on X_1. Show that (1) $\text{Var}[Y] = E[Y^2]$ is minimized at $\alpha \equiv 0$ and $\beta \equiv (\text{Cov}[X_1, X_2])/(\text{Var}[X_1])$, which is called **beta**, and (2) X_1 and Y are uncorrelated if the optimal linear prediction is used.

➤ **Exercise 6.4.2** Verify that the f in Eq. (6.20) is minimized at $\widehat{X} = (1/n)\sum_{i=1}^{n} x_i$.

➤ **Exercise 6.4.3** (1) Prove that a minimizes the **mean-square error** $E[(X-a)^2]$ when $a = E[X]$. (2) Show that the best predictor a of X_k based on $X_1, X_2, \ldots, X_{k-1}$ in the mean-square-error sense, that is, with minimum $E[(X_k - a)^2 \mid X_1, X_2, \ldots, X_{k-1}]$, is the **conditional least-squares estimator** $E[X_k \mid X_1, X_2, \ldots, X_{k-1}]$.

6.4.2 The Maximum Likelihood Estimator

Suppose that the sample has the probability density function $p(z \mid \theta)$. If Z is observed, $p(Z \mid \theta)$ is called the **likelihood** of θ.[3] The **maximum likelihood** (**ML**) **method** estimates θ by the number $\widehat{\theta}$ that maximizes the likelihood. Formally the **likelihood function** as the joint probability of the event $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ is

$$L(\boldsymbol{\theta}) \equiv \text{Prob}[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \boldsymbol{\theta}],$$

where $\boldsymbol{\theta} \equiv (\theta_1, \theta_2, \ldots, \theta_k)$ is the vector of parameters to be estimated. The likelihood function product equals $\prod_{i=1}^{n} p_{X_i}(x_i \mid \boldsymbol{\theta})$, where $p_{X_i}(x_i \mid \boldsymbol{\theta})$ is the probability density function of $X_i = x_i$ when the samples are drawn independently. The ML method estimates $\boldsymbol{\theta}$ with $\widehat{\boldsymbol{\theta}} \equiv (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_m)$ such that $L(\widehat{\boldsymbol{\theta}}) \geq L(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. It may be biased, however.

An estimator is **consistent** if it converges in probability to the true parameter as the sample size increases. The ML method, among consistent estimators, enjoys such optimality properties as minimum asymptotic variance and asymptotic normality under certain regularity conditions. These properties carry over to samples from a stochastic process [413, 422]. Unlike the least-squares method, which uses only the first two moments of the observations, the ML method utilizes the complete distribution of the observations.

Under certain regularity conditions, the ML estimate of $\boldsymbol{\theta}$ is the solution to the simultaneous equations $\partial L(\boldsymbol{\theta})/\partial\theta_i = 0$. Often it is the logarithm of $L(\boldsymbol{\theta})$, called the **log-likelihood function**, that is more convenient to work with. Numerical techniques are needed when a closed-form solution for $\boldsymbol{\theta}$ is not available.

EXAMPLE 6.4.2 Based on n independent observations $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma^2)$, the log-likelihood function is

$$\ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

After setting $\partial \ln L/\partial\mu$ and $\partial \ln L/\partial\sigma^2$ to zero, we obtain $\widehat{\mu} = (1/n)\sum_i x_i$, the sample mean, and $\widehat{\sigma^2} = (1/n)\sum_i (y_i - \widehat{\mu})$. The ML estimator of variance is biased because it differs from Eq. (6.1). It is consistent, however.

6.4.3 The Method of Moments

The **method of moments** estimates the parameters of a distribution by equating the population moments with their sample moments. Let $X_1, X_2, \ldots, X_n$ be random samples from a distribution characterized by k parameters $\theta_1, \theta_2, \ldots, \theta_k$. The method of moments estimates these parameters by solving k of the following equations:

$$M_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad M_3 = \frac{1}{n}\sum_{i=1}^{n} X_i^3, \ldots,$$

where the moments $M_i \equiv E[\, X^i \,]$ are functions of the parameters.

The name method of moments comes from the notion that parameters should be estimated by using moments. Also called the **analog method**, the method of moments requires no knowledge of the likelihood function. Although only certain moments of the observations instead of the full probability density function are used, this method is convenient and usually leads to simple calculations as well as to consistent estimators. Furthermore, it is the only approach of wide applicability in some situations.

Additional Reading

This chapter draws on [12, 23, 195, 273, 343, 463, 802, 816, 846] for probability theory, statistics, and statistical inferences. A very accurate approximation to the normal distribution appears in [678]. Regression analysis is covered by many books [317, 422, 632, 799]. See [273, 522, 584, 846] for more information about the lognormal distribution. The method of moments was introduced by Pearson (1857–1936) in 1894 [415].

NOTES

1. The idea of regression is due to Galton (1822–1911) [65].
2. Gauss claimed to have made the discovery in 1795 [75, 339].
3. The idea of likelihood is due to Ronald Fisher (1890–1962) [671].