Figure 1.1 displays two aspects of the relationship between real personal saving (SAV) and real personal disposable income (INC) in the US

In Fig. 1.1a the value of each series is shown for the period 1959.1 to 1992.1

It is a typical example of a time series plot, in which time is displayed on the horizontal axis and the values of the series are displayed on the vertical axis

Income shows an upward trend throughout the period, and in the early years, saving does likewise

This pattern, however, is not replicated in the middle and later years

SEE FIGURE 1.1

Figure 1.2 illustrates various associations between the natural log of real personal expenditure on gasoline (GAS), the natural log of the real price of gasoline (PRICE), and the natural log of real disposable personal income

The real price series, shows the two dramatic price hikes of the early and late 1970s, which were subsequently eroded by reductions in the nominal price of oil and by US inflation

The income and expenditure series are both shown in per capita form, because US population increased about 44% over the period

Per capita real expenditure of gasoline increased steadily in the 1960s and early 1970s, as real income grew and real price declined

This steady price ended with the price shocks of the 1970s, and per capita gas consumption has never regained the peak levels of the early seventies

SEE FIGURE 1.2

An alternative display of the same information is in terms of a scatter plot, shown in Fig 1.1b

Here one series is plotted against the other

Both parts of Fig 1.1 indicate a positive association between the variables:

increases in one tend to be associated with increases in the other

It is clear that although the association is approximately linear in the early part of the period, it is not so in the second half

SEE FIGURE 1.1b

SAMPLE MEANS

The observations for two variables are denoted by (X_i, Y_i) with $i = 1, 2, \ldots, N$

Sample means are given by

$$\overline{X} = \frac{\sum_{i=1}^{N} X_i}{N}, \overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

Data in deviation form (deviations from the mean) are denoted by

$$x_i = X_i - \overline{X}, y_i = Y_i - \overline{Y}$$

SCATTER DIAGRAM

Figure 1 shows an illustrative point on a scatter diagram with the sample means as new axes, giving four quadrants

Positive relationship: points lying for the most part in QI and QIII.

Negative relationship:points lying for the most part in QII and QIV.

The sign of $\sum x_i y_i$ will indicate whether the scatter diagram slopes upward or downward

SEE FIGURE SCATTER DIAGRAM

SAMPLE COVARIANCE

It is better to express the sum in average terms, giving the sample covariance:

$$COV(X,Y) = \frac{\sum_{i=1}^{N} x_i y_i}{N} = \frac{\sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

The value of the covariance depends on the units in which the variables are measured

CORRELATION COEFFICIENT

To obtain a measure of association that is invariant with respect to units of measurement the deviations are expressed in standard deviation units:

$$\begin{aligned} s_x &= \sqrt{\frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \\ s_y &= \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{N}} \end{aligned}$$

The correlation coefficient is defined as

$$r = \sum_{i=1}^{N} \frac{x_i y_i}{s_x s_y N}$$

ALTERNATIVE FORM

Since

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}}, s_y = \sqrt{\frac{\sum_{i=1}^{N} y_i^2}{N}}$$

the correlation coefficient can also be expressed as

$$r = \sum_{i=1}^{N} \frac{x_i y_i}{s_x s_y N} = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2} \sqrt{\sum_{i=1}^{N} y_i^2}}$$

AN ALTERNATIVE EXPRESSION

Moreover using the fact that

$$\begin{aligned} N\sum_{i=1}^{N} x_i y_i &= N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i, \\ N\sum_{i=1}^{n} x_i^2 &= N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2, \\ N\sum_{i=1}^{n} y_i^2 &= N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2, \end{aligned}$$

we have

$$r = \frac{N\sum_{i=1}^{N} X_i Y_i - \sum_{i=1}^{N} X_i \sum_{i=1}^{N} Y_i}{\sqrt{N\sum_{i=1}^{N} X_i^2 - (\sum_{i=1}^{N} X_i)^2}\sqrt{N\sum_{i=1}^{N} Y_i^2 - (\sum_{i=1}^{N} Y_i)^2}}$$

LIMITS OF r

The correlation coefficient must lie in the range from -1 to $+1$

From the Cauchy-Schwarz inequality we have:

$$(\sum_{i=1}^{N} x_i y_i) \leq (\sum_{i=1}^{N} x_i^2)(\sum_{i=1}^{N} y_i^2)$$

Since

$$r = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}$$

It follows that $r^2 \leq 1$

NUMERICAL EXAMPLE

$$r = \frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}} = \frac{70}{\sqrt{40}\sqrt{124}} = 0.9879$$

If no theory exists or can be devised that connects the two variables, the correlation may be classed as a nonsense correlation

Yule took annual data from 1866 to 1911 for the death rate in England and Wales and for the proportion of all marriages solemnized in the Church of England and found the correlation coefficient to be 0.95

However, no British politician proposed closing down the Church of England to confer immortality on the electorate

ECONOMIC THEORY

Lets say that we have two variables: expenditure (Y) and income denoted by (X)

There are N observations: $i = 1, \ldots, N$

Actual values: $Y_1, \ldots, Y_N$; $X_1, \ldots, X_N$

Economic theory suggests a linear relationship between the two variables:

$$\widetilde{Y}_i = a + bX_i$$

$\widetilde{Y}_i$ are the values suggested by the theory

The differences between the actual values and the suggested ones are the errors

$$e_i = Y_i - \widetilde{Y}_i$$

LEAST SQUARES ESTIMATORS

Thus, we have the actual values: $Y_1, \ldots, Y_N$

and the values suggested by the theory: $\widetilde{Y}_1, \ldots, \widetilde{Y}_N$

Their difference are the errors $e_1, \ldots, e_N$

We want to get an estimate of the two coefficients: a and b

We calculate the sum of the squared errors:

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \widetilde{Y}_i)^2 = \sum_{i=1}^{N} (Y_i - a - bX_i)^2$$

This sum is a function of the two parameters a and b

We choose the values of a and b that minimises this sum.

These are the least squares estimates: α and β

MINIMIZATION PROBLEM

Take the first derivatives with respect to a and b and set them equal to zero:

$$\begin{aligned} \frac{\partial[\sum_{i=1}^{N} e_i^2]}{\partial a} &= \frac{\partial[\sum_{i=1}^{N}(Y_i - a - bX_i)^2]}{\partial a} \\ &= -2\sum_{i=1}^{N}(Y_i - a - bX_i) = 0, \quad (1) \end{aligned}$$

$$\begin{aligned} \frac{\partial[\sum_{i=1}^{N} e_i^2]}{\partial b} &= \frac{\partial[\sum_{i=1}^{N}(Y_i - a - bX_i)^2]}{\partial b} \\ &= -2\sum_{i=1}^{N} X_i(Y_i - a - bX_i) = 0 \end{aligned}$$

$$(2)$$

These two equations imply that

$$\begin{aligned} \sum_{i=1}^{N} Y_i &= Na + b\sum_{i=1}^{N} X_i, \\ \sum_{i=1}^{N} X_i Y_i &= a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_i^2 \end{aligned} \tag{3}$$

Dividing the first equation by N gives

$$\overline{Y} = \alpha + \beta\overline{X} \Rightarrow \alpha = \overline{Y} - \beta\overline{X} \tag{4}$$

It can be shown that if we substitute the above expression into equation (3) and solve for β we get

$$\beta = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} \tag{5}$$

ALTERNATIVE EXPRESSIONS

Recall that

$$COV(X,Y) = \frac{\sum_{i=1}^{N} x_i y_i}{N}, VAR(X) = \frac{\sum_{i=1}^{N} x_i^2}{N}$$

Hence, it follows from equation (5)

$$\beta = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{COV(X,Y)}{VAR(X)} \tag{6}$$

It can also be shown that

$$\beta = r\frac{s_y}{s_x} \tag{7}$$

NUMERICAL EXAMPLE

$$\begin{aligned}\sum_{i=1}^{N} Y_i &= Na + b\sum_{i=1}^{N} X_i, \\ \sum_{i=1}^{N} X_i Y_i &= a\sum_{i=1}^{N} X_i + b\sum_{i=1}^{N} X_i^2\end{aligned}$$

$$\begin{aligned}40 &= 5a + 20b, \\ 230 &= 20a + 120b\end{aligned}$$

We have 2 equations and 2 unknowns. Or

$$\begin{aligned}\beta &= \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{70}{40} = 1.75, \\ \alpha &= \overline{Y} - \beta\overline{X} = 8 - 1.75(4) = 1\end{aligned}$$

3 IMPORTANT PROPERTIES

The Least square estimates minimizes the sum of the squared errors: $\sum_{i=1}^{N} e_i^2$

It passes through the mean point (equation 4): $\overline{Y} = \alpha + \beta\overline{X}$

From equation (1) we have

$$\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i) = 0 \Rightarrow \sum_{i=1}^{N} \widehat{e}_i = 0$$

From equation (2) we have

$$\sum_{i=1}^{N} X_i(Y_i - \alpha - \beta X_i) = 0 \Rightarrow \sum_{i=1}^{N} X_i \widehat{e}_i = 0$$

The least squares residuals have no covariance in the sample with the values of the independent variable

The theoretical covariance is given by

$$COV(X, \widehat{e}) = E(X\widehat{e}) - E(X)E(\widehat{e}),$$

The sample covariance is given by

$$\frac{\sum_{i=1}^{N} X_i \widehat{e}_i}{N} - \frac{\sum_{i=1}^{N} X_i}{N} \frac{\sum_{i=1}^{N} \widehat{e}_i}{N} = 0$$

In view of equations (1) and (2) this is 0.

MODEL

Our bivariate regression is given by

$$Y_i = a + bX_i + e_i \tag{8}$$

The errors e_i are stochastic

Assumptions about the errors:

i) They are identically and independently distributed

ii) They have expected value 0, $E(e_i) = 0$, and variance σ^2, $VAR(e_i) = \sigma^2$

That is $e_i \sim iid(0, \sigma^2)$

ESTIMATORS AND ESTIMATES

There are thus three parameters to be estimated in the model, namely, a,b and σ^2

Once the two parameters a and b has been estimated and a line has been fitted, the residuals from this line may be used to form an estimate of σ^2

An estimator is a formula, method, or recipe for estimating an unknown population parameter

An estimate is the numerical value obtained when sample data are substituted in the formula

Thus the least square formula $\beta = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$ is an estimator

For each particular sample (values of X_i, Y_i, $i = 1, \ldots, N$) we have one estimate (a numerical value of β)

However, before the sampling the fact that e_i is stochastic implies that Y_i is stochastic

Thus the estimator β is stochastic as well

It is a linear combination of the y (or Y) variable, $\beta = \sum_{i=1}^{N} w_i y_i$ where $w_i = \frac{x_i}{\sum_{i=1}^{N} x_i^2}$, and hence a linear combination of the stochastic e variable

TWO IMPORTANT QUESTIONS

There are two important questions regarding the least squares estimators of a and b: α and β respectively

1. What are the properties of these estimators

2. How may these estimators be used to make inferences about a and b

The answers to both these questions depend on the sampling distribution of the least squares estimators.

A given sample yields a specific numerical estimate

Another sample from the same population will yield another numerical estimate

A sampling distribution describes the results that will be obtained from the estimator(s) over the potentially infinite set of samples that may be drawn from the population

MAIN ASSUMPTIONS

Recall that we assume: $e_i \sim iid(0, \sigma^2)$

The derivation of inference procedures requires an assumption about the distribution of the $e's$

The standard assumption is that of normality:

$e_i \sim N(0, \sigma^2)$

Recall that the least square estimator β is a linear combination of e and thus it is also stochastic

It can be shown that

$$\beta \sim N(b, \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2})$$

That is $E(\beta) = b$; In other words β is an unbiased estimator

In addition $VAR(\beta) = \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2}$. It can be shown that this is the smallest variance amongst all other unbiased linear estimators

The least squares estimator that has the minimum variance in the class of linear unbiased estimators is called best linear unbiased estimator: BLUE

THE α ESTIMATOR

It can be shown that

$$\alpha \sim N(a, \sigma^2(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} x_i^2}))$$

That is $E(\alpha) = a$; In other words α is an unbiased estimator

In addition $VAR(\alpha) = \sigma^2(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} x_i^2})$

HYPOTHESIS TESTING

The fact that $\beta \sim N(b, \frac{\sigma^2}{\sum_{i=1}^{N} x_i^2})$ implies that

$$\frac{\beta - b}{se(\beta)} \sim N(0, 1)$$

where $se(\beta) = \sqrt{VAR(\beta)} = \frac{\sigma}{\sqrt{\sum_{i=1}^{N} x_i^2}}$

We can use this information to do hypothesis testing

For example, we can test the null hypothesis: $H_0 : b = b_0$ against the alternative $H_0 : b \neq b_0$

$$\text{IF } \left| \frac{\beta - b_0}{se(\beta)} \right| > 1.96 \text{ REJECT } H_0$$

ESTIMATE OF σ^2

Because we do not know the true variance of the error (σ^2) we have to use an estimate

The estimate of σ^2, denoted by s^2 is the sample variance of the residuals:

$$s^2 = \frac{\sum_{i=1}^{N} \widehat{e}_i^2}{N-2}$$

Because we use an estimate of the true variance it follows that

$$\begin{aligned} \beta &\sim t_{N-2}(b, \frac{s^2}{\sum_{i=1}^{N} x_i^2}), \\ \alpha &\sim t_{N-2}(a, s^2(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} x_i^2})) \end{aligned}$$

For example, we can test the null hypothesis: $H_0 : b = b_0$ against the alternative $H_0 : b \neq b_0$

$$\text{IF } \left| \frac{\beta - b_0}{\widehat{se(\beta)}} \right| > \text{5\% Crit Val } t_{N-2} \text{ REJECT } H_0$$

NUMERICAL EXAMPLE

$$
\begin{aligned}
s^2 &= \frac{\sum_{i=1}^{N} \widehat{e}_i^2}{N-2} = \frac{1.5}{3} = 0.5, \\
\widehat{VAR}(\beta) &= \frac{s^2}{\sum_{i=1}^{N} x_i^2} = \frac{0.5}{40} = 0.0125, \\
\widehat{VAR}(\alpha) &= s^2\left(\frac{1}{N} + \frac{\overline{X}^2}{\sum_{i=1}^{N} x_i^2}\right) = 0.5\left(\frac{1}{5} + \frac{16}{40}\right) = 0.3
\end{aligned}
$$

The estimated standard errors of the regression coefficients are

$$\widehat{se(\beta)} = 0.1118,\ \widehat{se(a)} = 0.5477$$

Testing the null hypotheses: $H_0 : b = 0$ and $H_0 : a = 0$

The 5% critical value for t distribution with $N - 2 = 3$ degrees of freedom is 3.182. Thus

$$\left|\frac{\beta}{\widehat{se(\beta)}}\right| = \frac{1.75}{0.1118} = 15.653 > 3.182, \text{ REJECT } H_0,$$

$$\left|\frac{\alpha}{\widehat{se(a)}}\right| = \frac{1}{0.5477} = 1.826 < 3.182, \text{ ACCEPT } H_0$$