

6. Linear Regression Analysis

- *Simple linear regression*
- *Interpretation of the correlation coefficient*
- *Effects of measurement errors*

6. Linear Regression Analysis

a. Simple linear regression

The least-squares fit of a straight line to a set of measurements was discussed in Section 5b. The solution, for the case where the dependent variable y_i is measured with constant uncertainty σ , is

$$y = y_0 + b_x x \quad (6.1)$$

$$y_0 = \bar{y} - b_x \bar{x} \quad (6.2)$$

$$b_x = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \quad (6.3)$$

or, for the variables $x' = x - \bar{x}$ and $y' = y - \bar{y}$,

$$y'_0 = 0 \quad (6.4)$$

$$b_x = \frac{\overline{x'y'}}{\overline{x'^2}} . \quad (6.5)$$

While this solution can be found for any set of measurements, it is still necessary to consider if the solution is a useful representation of the measurements and if the assumed

dependency between y and x is valid. Indeed, if it is assumed that x is the dependent variable and y the dependent variable, the solution is different:

$$x'_0 = 0 \quad (6.6)$$

$$b_y = \frac{\overline{x'y'}}{y'^2}. \quad (6.7)$$

Linear regression analysis addresses the dual tasks of finding the best-fit relationships and testing for correlations in the data that indicate a linear relationship between the variables.

The key measure of correlation in regression analysis is the *correlation coefficient*, defined in terms of the variables $x' = x - \bar{x}$ and $y' = y - \bar{y}$ as

$$\begin{aligned} \rho &= \frac{V_{xy}}{\sqrt{V_x V_y}} = \frac{\overline{x'y'}}{\sqrt{\overline{x'^2} \overline{y'^2}}} \\ &= \frac{\overline{(x - \bar{x})(y - \bar{y})}}{\sqrt{\overline{(x - \bar{x})^2} \overline{(y - \bar{y})^2}}} = \sqrt{b_x b_y}. \end{aligned} \quad (6.8)$$

The correlation coefficient is thus not dependent on which variable is considered independent. However, the slope parameters b_x and b_y for the two cases are only equal in the case where the variables x and y are linearly related, i.e., $y_i = (\text{constant}) x_i$. If x and y are completely uncorrelated (so that $\overline{x'y'} = 0$), both b_x and b_y are zero and hence the best-fit lines are perpendicular to each other. Figure 6.1 shows an example of the relationships between slope parameters for a case with correlation coefficient 0.8. When the data are considered in vertical slices (as appropriate for the assumption that y is a function of x), each vertical slice appears centered on the regression line labeled $y(x)$, but when considered in horizontal slices each slice appears centered on the line labeled $x(y)$; this illustrates why the two regression fits must be different.

The following example helps illustrate the meaning of the correlation coefficient. Let u_1 and u_2 be two independent variables that obey normal distributions with standard deviations equal to unity and means of zero. Define variables y_1 and y_2 as

$$y_1 = a_1 + b_1 u_1 \quad (6.9)$$

$$y_2 = a_2 + b_2 u_2 + b_3 u_1 \quad (6.10)$$

where a_1 , a_2 , b_1 , and b_2 are non-zero constants. The variables y_1 and y_2 are then correlated, and the correlation coefficient is

$$\rho = \frac{\langle y'_1 y'_2 \rangle}{\sqrt{\langle y'_1 \rangle \langle y'_2 \rangle}} = \frac{b_3}{\sqrt{b_2^2 + b_3^2}}. \quad (6.11)$$

The variance in y_2 is $(b_2^2 + b_3^2)$, so the fraction of the variance contributed by the function u_1 is the square of the correlation coefficient. The square of the correlation coefficient is sometimes said to measure the fraction of the variance in one variable that can be “explained” or accounted for by correlation with another variable. The remainder of the variance results from other sources, perhaps from correlation with other variables.

The correlation coefficient is notoriously dangerous to interpret, especially in the sense of statistical inference. To gain some sense of the variability to be expected in

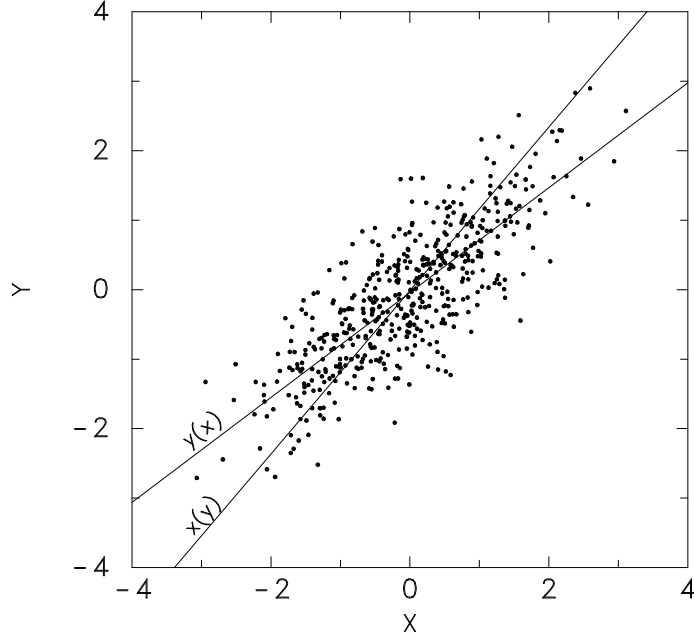


Fig. 6.1: The linear regression lines for the assumptions that y is the dependent variable ($y(x)$) and that x is the dependent variable ($x(y)$). The data have a correlation coefficient of 0.80 and were generated from random Gaussian distributions with standard deviations of 1.0 and means of zero.

measurements of this parameter, consider the model of a general bivariate Gaussian distribution:

$$P(x_1, x_2) = \frac{\exp \left\{ \frac{-1}{(1-\rho^2)} \left[\frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{(x_2-\mu_2)^2}{2\sigma_2^2} - \rho \frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} \right] \right\}}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}. \quad (6.12)$$

It can be verified by direct integration that $\langle x_1 \rangle = \mu_1$ and $\langle x_2 \rangle = \mu_2$, that $\sigma_{x_1} = \sigma_1$ and $\sigma_{x_2} = \sigma_2$, and that

$$\rho = \frac{\langle (x_1 - \mu_1)(x_2 - \mu_2) \rangle}{\sigma_1\sigma_2}, \quad (6.13)$$

so this distribution has appropriate properties to serve as a model for regression results. Using this model, we can ask what the probability will be for observing specific values r of the correlation coefficient. (We will distinguish r , the result of calculations with finite samples from the population, from the population correlation coefficient ρ .)

Two properties derivable from the bivariate Gaussian distribution illustrate the meaning of the correlation coefficient. To obtain them, consider the conditional probability of x_2 given x_1 :

$$P(x_2|x_1) = AP(x_1, x_2) \quad (6.14)$$

where A is defined to normalize the probability distribution when integrated over x_2 :

$$\int P(x_2|x_1)dx_2 = 1. \quad (6.15)$$

If $x'_1 = x_1 - \mu_1$ and $x'_2 = x_2 - \mu_2$ are the deviations from the means μ_1 and μ_2 , then the conditional probability distribution obtained by integrating (6.15) to determine A is

$$P(x'_2|x'_1) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\left(\frac{x'_2 - \rho x'_1 \frac{\sigma_2}{\sigma_1}}{\sqrt{2}\sigma_2\sqrt{1-\rho^2}}\right)^2\right\}. \quad (6.16)$$

This probability distribution has the following two properties:

- If one variable is specified as $x_1 = x'_1 + \mu_1$, then the conditional probability distribution for x_2 , given this value of x_1 , has Gaussian form with mean $\mu_2 + \rho x'_1 \sigma_2 / \sigma_1$. In particular, the conditional expectation value for x_2 is

$$\langle x_2 | x_1 \rangle = \mu_2 + \rho x_1 \frac{\sigma_2}{\sigma_1}. \quad (6.17)$$

- The variance in x_2 for a fixed value of x_1 is

$$\langle (x_2 - \langle x_2 \rangle)^2 \rangle = \sigma_2^2 (1 - \rho^2). \quad (6.18)$$

Thus, if σ_2 represents the total variance in x_2 , ρ^2 represents the fraction of this variance that is removed once x_1 is fixed.

These properties of the bivariate Gaussian distribution make it possible to generate simulated experiments to study the expected distribution in r by Monte Carlo techniques. For a given population correlation coefficient ρ and sample size N , one can generate many random samples from the appropriate bivariate distribution and compute the sample correlation coefficient r for each sample. Some results from such calculations are shown in Figures 6.2 and 6.3. These figures illustrate that observed correlation coefficient will often differ significantly from the true population correlation coefficient, especially for small sample sizes, so it is important not to attributed unwarranted significance to correlation coefficients obtained from small samples.

Some of the characteristics of the correlation coefficient illustrated by examples in this chapter are:

- The correlation coefficient does not characterize the grouping of the data about the best-fit line, but rather the fraction of the variability that can be attributed to linear dependence. Data that are tightly grouped about a line will nevertheless have zero correlation coefficient if that line has zero slope. The same degree of scatter about a line with unity slope can give a high correlation coefficient.
- The two possible linear relationships are equal only for a completely correlated data set. The two regression relationships differ more as the correlation coefficient becomes smaller, and the two lines are orthogonal when the correlation coefficient is zero. The regression relationships also differ significantly from the relationship that would minimize the perpendicular distance of the best-fit line to the data; e.g., in Fig. 6.1, the data were generated using a relationship that would give unity slope.
- Figures 6.2 and 6.3 show that, for small samples, large values of the correlation coefficient can arise purely from statistical fluctuations. For example, 10 measurements will show correlation coefficients having absolute magnitudes greater than 0.5 in more than 10% of the samples, even if there is no true correlation. Correlation coefficients

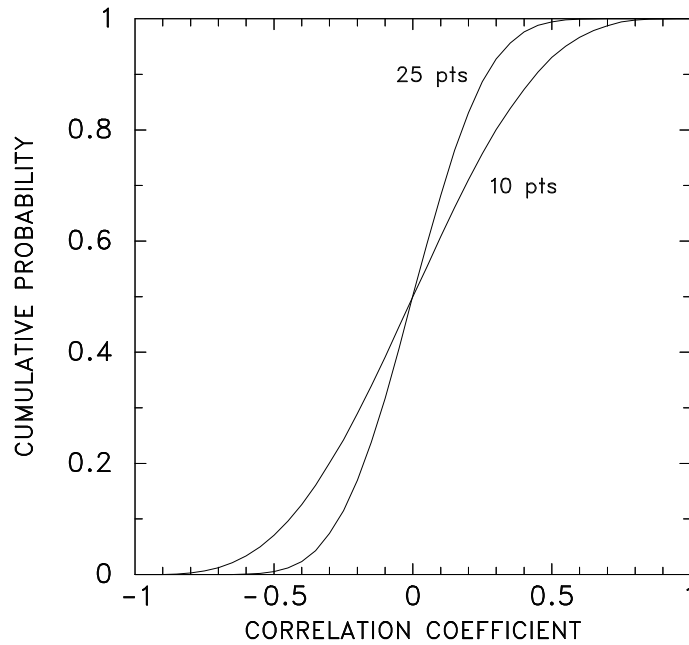


Fig. 6.2: Simulation results generated for a correlation coefficient of $r=0$, using repeated random sequences. The plot shows the fraction of the results for which the calculated correlation coefficient was smaller than the plotted value, for sequences of 10 points and 25 points.

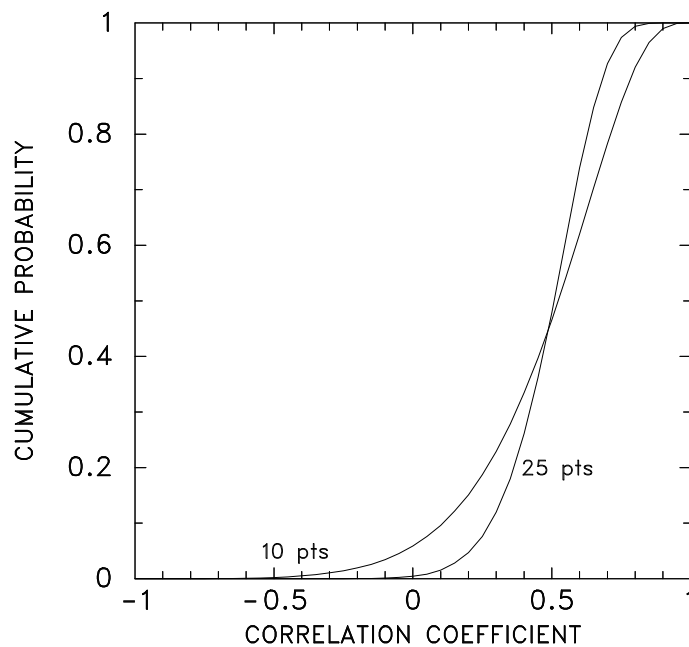


Fig. 6.3: Simulation results similar to those in Fig. 6.2, for a population correlation coefficient of 0.5 and for sequences each containing 10 points or 25 points.

calculated using small samples must be interpreted carefully to avoid falsely attributing too much significance to them.

Two other properties of the bivariate Gaussian distribution are useful in some interpretations of regression relationships: (i) a rotation can always be selected that transforms the variables into new variables that are uncorrelated; and (ii) the contours of constant probability in a plot of y_0 vs b are ellipses in the general case.

EXAMPLE 6.1: *It is a common error to show a regression fit as evidence that there is a difference in the measurements of two instruments. Suppose two wind vanes behave identically, and each has the same measurement error σ . They may produce a set of measurements like those shown in Fig. 6.1, for which both x and y have the same mean and standard deviation. It is an error to conclude that, because the regression line for $y(x)$ has a slope of about 0.8, the response of the instrument measuring y is only 80% as large as the response of the instrument measuring x .*

The distribution of r about ρ is asymmetrical, so it is useful to transform r to another variable that will have an error distribution that is approximately Gaussian. A common example is the Fisher z transformation, based on the variable

$$z_f = 0.5 \ln \left(\frac{1+r}{1-r} \right) . \quad (6.19)$$

This variable is approximately Gaussian-distributed with standard deviation

$$\sigma_z = \frac{1}{\sqrt{N-3}} . \quad (6.20)$$

The inverse transformation is

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} . \quad (6.21)$$

EXAMPLE 6.2: *Consider the case where a sample of 25 measurements gives a correlation coefficient of 0.5, as shown in Fig. 6.3. Find the one-standard-deviation uncertainty range for the correlation coefficient.*

From (6.19) and (6.20) $z_f=0.549$ and $\sigma_z = 0.213$, so the one-standard-deviation limits for z_f are 0.336 and 0.762. The corresponding values of r from (6.21) are 0.324 and 0.643, as shown on the plot.

The estimated uncertainty that would result from error propagation from the above formulas is

$$\delta r \approx \frac{1}{\sqrt{N-3}}(1+r)(1-r) \approx 0.16 . \quad (6.22)$$

This is approximately the same as the average of the two standard deviations found using the transformation equations. However, because of the skewed nature of the distribution functions, this estimate should only serve as a preliminary guide to the uncertainty.

The uncertainties in the slope and intercept for the regression can be estimated using the results from section 5b. The elements of the covariance matrix, for the case where x is the independent variable, are

$$V_{y'_0 y'_0} = \frac{\sigma^2}{N} \quad (6.23)$$

$$V_{b_x b_x} = \frac{\sigma^2}{N \overline{x'^2}} \quad (6.24)$$

$$V_{y'_0 b_x} = 0, \quad (6.25)$$

so in the primed coordinates the results for the slope and intercept are not correlated. This is not generally true in the original coordinates.

The uncertainty in the parameter b , the slope of the regression line, can also be determined using the analysis previously applied to linear least-squares fitting. In that case, the error matrix (cf. 5.22) was

$$H^{-1} = \frac{\sigma^2}{N(\overline{x^2} - \bar{x}^2)} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix} \quad (6.26)$$

or, with $\sigma_x^2 = (\overline{x^2} - \bar{x}^2)$,

$$\begin{pmatrix} V_{y_0 y_0} & V_{y_0 b} \\ V_{b y_0} & V_{bb} \end{pmatrix} = \frac{\sigma^2}{N \sigma_x^2} \begin{pmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \quad (6.27)$$

In particular,

$$V_{bb} = \frac{\sigma^2}{N \sigma_x^2} \quad (6.28)$$

where the standard deviation (for a true linear relationship) can be estimated from $\sigma \approx s$ where

$$s^2 = \sum_i (y_i - y_0 - b x_i)^2 / (N - 2). \quad (6.29)$$

b. Effects of measurement errors

In the preceding section, we assumed that the correlation between variables was the result of a physical relationship, and ignored the possible effects of measurement uncertainties. However, measurement errors will tend to obscure the true correlation, especially if there are correlations among the measurement errors. If the measurement uncertainty is large compared to the true range of variation in a variable, it may be difficult to determine the true correlation coefficient.

In most cases the measurement errors are not correlated with fluctuations in the values being measured. In this case, the observed covariance matrix is just the sum of the true covariance matrix and the covariance matrix describing the measurement errors:

$$\mathbf{H}_{observed}^{-1} = \mathbf{H}_{natural}^{-1} + \mathbf{H}_{measurement}^{-1}. \quad (6.30)$$

To show this, let x^* and y^* be observed values and let x and y be the true values, so that the respective measurement errors in x and y are

$$u = x^* - x \quad (6.31)$$

$$v = y^* - y. \quad (6.32)$$

Then the observed covariance has the expectation value

$$V_{x^*y^*} = \langle (x^* - \bar{x}^*)(y^* - \bar{y}^*) \rangle \quad (6.33)$$

$$= \langle (x + u - \bar{x} - \bar{u})(y + v - \bar{y} - \bar{v}) \rangle \quad (6.34)$$

$$= \langle (x - \bar{x})(y - \bar{y}) \rangle + \langle (u - \bar{u})(v - \bar{v}) \rangle = V_{xy} + V_{uv} \quad (6.35)$$

because other terms in (6.34) have expectation values of zero if the errors are uncorrelated with the values. The other elements of the covariance matrix are similarly related to the individual contributions.

Because x^* and y^* are the measured quantities, the estimator of the correlation coefficient that is obtained from them is

$$r_{x^*y^*} = \frac{V_{x^*y^*}}{\sqrt{V_{x^*x^*}V_{y^*y^*}}} = \frac{V_{xy} + V_{uv}}{\sqrt{(V_{xx} + V_{uu})(V_{yy} + V_{vv})}} \quad (6.36)$$

$$\neq \left(\frac{V_{xy}}{\sqrt{V_{xx}V_{yy}}} = \rho_{xy} \right). \quad (6.37)$$

Similarly,

$$b_{y'} = \frac{V_{x'y'}}{V_{x'x'}} = \frac{V_{xy} + V_{uv}}{V_{xx} + V_{uu}} \quad (6.38)$$

$$\neq \left(\frac{V_{xy}}{V_{xx}} = b_y \right). \quad (6.39)$$

Thus measurement errors can introduce biases in the estimated slope and correlation coefficient from a regression analysis.

EXERCISE 6.1: A set of 25 corresponding measurements of $\{x\}$ and $\{y\}$ give a correlation coefficient of 0.7. The estimated measurement uncertainty is 50% of the measured standard deviation for both x and y . What is the best estimate of the true correlation coefficient between x and y , and what are the one-standard-deviation error limits in this estimate?

c. Linear regression with several independent variables

Suppose that y is a linear function of a set of variables $\{x\}$:

$$y = a + \sum_j b_j x_j. \quad (6.40)$$

Define new variables z_j that have zero mean and unity variance:

$$z_j = (x_j - \bar{x}_j) / \sigma_j \quad (6.41)$$

where σ_j^2 is the variance in x_j . Then

$$y' = y - \bar{y} = \sum_j b'_j z_j \quad (6.42)$$

gives the fluctuation from the mean value in terms of the normalized variables $\{z\}$.

The least-squares fit conditions reduce to

$$\mathbf{b}' = \mathbf{H}^{-1} \mathbf{M} \quad (6.43)$$

where

$$H_{jk} = \sum_i \frac{z_{ji} z_{ki}}{\sigma_i^2} \quad (6.44)$$

$$M_k = \sum_i \frac{y_i z_{ki}}{\sigma_i^2} \quad (6.45)$$

and the summation index i includes each measurement, so for example z_{ki} is the i th measurement of normalized variable z_k . In the case where σ_i is a constant,

$$H_{jk} = \frac{N}{\sigma^2} \overline{z_j z_k} = \frac{N}{\sigma^2} r_{jk} \quad (6.46)$$

where r_{jk} is the correlation coefficient between the (normalized) variables z_j and z_k . Thus, the information matrix is related to the matrix of correlation coefficients via

$$\mathbf{H} = \frac{N}{\sigma^2} \mathbf{r}. \quad (6.47)$$

The information matrix thus provides information that helps determine the uncertainty in the regression coefficients and determine if an added parameter is significant in improving the fit. Further discussion of these applications will be postponed until tests of hypotheses are first considered in more general terms.

SOURCES AND FURTHER READING

Bevington, P. R., 1969: Data Reduction and Error Analysis for the Physical Sciences. McGraw-Hill, New York, 336 pp.

Brownlee, K. A., 1965: Statistical Theory and Methodology in Science and Engineering. John Wiley and Sons, New York, 590 pp.

Mosteller and Tukey, 1977: Data Analysis and Regression.

page intentionally left blank