1

Basic notions in classical data analysis

The goal of data analysis is to discover relations in a dataset. The basic ideas of probability distributions, and the mean and variance of a random variable are introduced first. The relations between two variables are then explored with correlation and regression analysis. Other basic notions introduced in this chapter include Bayes theorem, discriminant functions, classification and clustering.

1.1 Expectation and mean

Let x be a random variable which takes on discrete values. For example, x can be the outcome of a die cast, where the possible values are $x_i = i$, with $i = 1, \ldots, 6$. The *expectation* or expected value of x from a population is given by

$$\mathrm{E}[x] = \sum_i x_i P_i, \tag{1.1}$$

where P_i is the probability of x_i occurring. If the die is fair, $P_i = 1/6$ for all i, so $\mathrm{E}[x]$ is 3.5. We also write

$$\mathrm{E}[x] = \mu_x, \tag{1.2}$$

with μ_x denoting the *mean* of x for the population.

The expectation of a sum of random variables satisfies

$$\mathrm{E}[ax + by + c] = a\,\mathrm{E}[x] + b\,\mathrm{E}[y] + c, \tag{1.3}$$

where x and y are random variables, and a, b and c are constants.

For a random variable x which takes on continuous values over a domain Ω, the expection is given by an integral,

$$\mathrm{E}[x] = \int_\Omega x p(x)\,\mathrm{d}x, \tag{1.4}$$

where $p(x)$ is the *probability density* function. For any function $f(x)$, the expectation is

$$\begin{aligned} \mathrm{E}[f(x)] &= \int_{\Omega} f(x)\, p(x)\, \mathrm{d}x \quad \text{(continuous case)} \\ &= \sum_{i} f(x_i)\, P_i \quad \text{(discrete case)}. \end{aligned} \tag{1.5}$$

In practice, one can sample only N measurements of x $(x_1, \ldots, x_N)$ from the population. The *sample mean* $\overline{x}$ or $\langle x \rangle$ is calculated as

$$\overline{x} \equiv \langle x \rangle = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{1.6}$$

which is in general different from the population mean μ_x. As the sample size increases, the sample mean approaches the population mean.

1.2 Variance and covariance

Fluctuation about the mean value is commonly characterized by the variance of the population,

$$\mathrm{var}(x) \equiv \mathrm{E}[(x - \mu_x)^2] = \mathrm{E}[x^2 - 2x\mu_x + \mu_x^2] = \mathrm{E}[x^2] - \mu_x^2, \tag{1.7}$$

where (1.3) and (1.2) have been invoked. The standard deviation s is the positive square root of the population variance, i.e.

$$s^2 = \mathrm{var}(x). \tag{1.8}$$

The sample standard deviation σ is the positive square root of the sample variance, given by

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2. \tag{1.9}$$

As the sample size increases, the sample variance approaches the population variance. For large N, distinction is often not made between having $N - 1$ or N in the denominator of (1.9).

Often one would like to compare two very different variables, e.g. sea surface temperature and fish population. To avoid comparing apples with oranges, one usually standardizes the variables before making the comparison. The *standardized variable*

$$x_s = (x - \overline{x})/\sigma, \tag{1.10}$$

is obtained from the original variable by subtracting the sample mean and dividing by the sample standard deviation. The standardized variable is also called the *normalized* variable or the *standardized anomaly* (where *anomaly* means the deviation from the mean value).

For two random variables x and y, with mean μ_x and μ_y respectively, their *covariance* is given by

$$\text{cov}(x, y) = \text{E}[(x - \mu_x)(y - \mu_y)]. \tag{1.11}$$

The variance is simply a special case of the covariance, with

$$\text{var}(x) = \text{cov}(x, x). \tag{1.12}$$

The sample covariance is computed as

$$\text{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}). \tag{1.13}$$

1.3 Correlation

The (Pearson) correlation coefficient, widely used to represent the strength of the linear relationship between two variables x and y, is defined as

$$\hat{\rho}_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}, \tag{1.14}$$

where s_x and s_y are the population standard deviations for x and y, respectively.

For a sample containing N pairs of (x, y) measurements or observations, the *sample correlation* is computed by

$$\rho \equiv \rho_{xy} = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{N} (x_i - \overline{x})^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{N} (y_i - \overline{y})^2\right]^{\frac{1}{2}}}, \tag{1.15}$$

which lies between -1 and $+1$. At the value $+1$, x and y will show a perfect straight-line relation with a positive slope; whereas at -1, the perfect straight line will have a negative slope. With increasing noise in the data, the sample correlation moves towards 0.

An important question is whether the obtained sample correlation can be considered significantly different from 0 – this is also called a test of the null (i.e. $\hat{\rho}_{xy} = 0$) hypothesis. A common approach involves transforming to the variable

$$t = \rho \sqrt{\frac{N-2}{1-\rho^2}}, \tag{1.16}$$

which in the null case is distributed as the Student's t distribution, with $\nu = N - 2$ degrees of freedom.

For example, with $N = 32$ data pairs, ρ was found to be 0.36. Is this correlation significant at the 5% level? In other words, if the true correlation is zero ($\hat{\rho}_{xy} = 0$), is there less than 5% chance that we could obtain $\rho \geq 0.36$ for our sample? To answer this, we need to find the value $t_{0.975}$ in the t-distribution, where $t > t_{0.975}$ occur less than 2.5% of the time and $t < -t_{0.975}$ occur less than 2.5% of the time (as the t-distribution is symmetrical), so altogether $|t| > t_{0.975}$ occur less than 5% of the time. From t-distribution tables, we find that with $\nu = 32 - 2 = 30$, $t_{0.975} = 2.04$.

From (1.16), we have

$$\rho^2 = \frac{t^2}{N - 2 + t^2}, \tag{1.17}$$

so substituting in $t_{0.975} = 2.04$ yields $\rho_{0.05} = 0.349$, i.e. less than 5% of the sample correlation values will indeed exceed $\rho_{0.05}$ in magnitude if $\hat{\rho}_{xy} = 0$. Hence our $\rho = 0.36 > \rho_{0.05}$ is significant at the 5% level based on a '2-tailed' t test. For moderately large N ($N \geq 10$), an alternative test involves using Fisher's z-transformation (Bickel and Doksum, 1977).

Often the observations are measurements at regular time intervals, i.e. time series, and there is *autocorrelation* in the time series – i.e. neighbouring data points in the time series are correlated. Autocorrelation is well illustrated by persistence in weather patterns, e.g. if it rains one day, it increases the probability of rain the following day. With autocorrelation, the effective sample size may be far smaller than the actual number of observations in the sample, and the value of N used in the significance tests will have to be adjusted to represent the effective sample size.

A statistical measure is said to be *robust* if the measure gives reasonable results even when the model assumptions (e.g. data obeying Gaussian distribution) are not satisfied. A statistical measure is said to be *resistant* if the measure gives reasonable results even when the dataset contains one or a few outliers (an *outlier* being an extreme data value arising from a measurement or recording error, or from an abnormal event).

Correlation assumes a linear relation between x and y; however, the sample correlation is not *robust* to deviations from linearity in the relation, as illustrated in Fig. 1.1a where $\rho \approx 0$ even though there is a strong (nonlinear) relationship between the two variables. Thus the correlation can be misleading when the underlying relation is nonlinear. Furthermore, the sample correlation is not *resistant* to

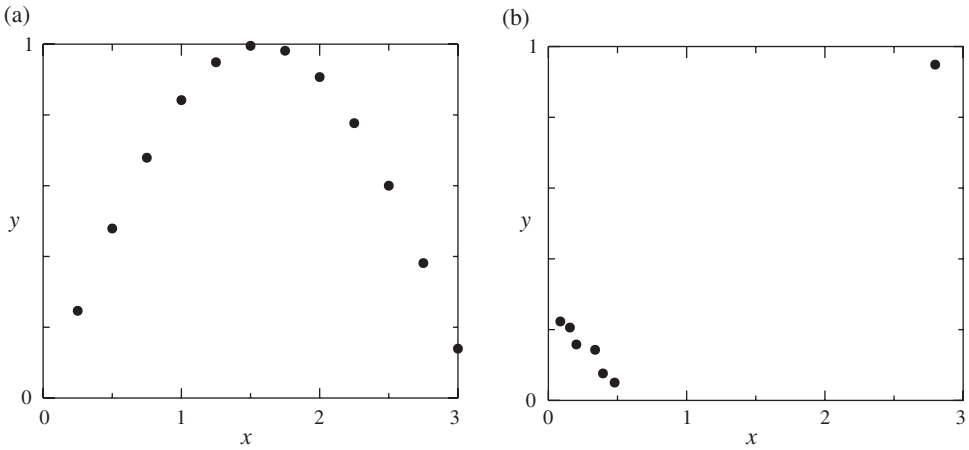

Fig. 1.1 (a) An example showing that correlation is not robust to deviations from linearity. Here the strong nonlinear relation between x and y is completely missed by the near-zero correlation coefficient. (b) An example showing that correlation is not resistant to outliers. Without the single outlier, the correlation coefficient changes from positive to negative.

outliers, where in Fig. 1.1b if the outlier datum is removed, ρ changes from being positive to negative.

1.3.1 Rank correlation

For the correlation to be more robust and resistant to outliers, the Spearman rank correlation is often used instead. If the data $\{x_1, \ldots, x_N\}$ are rearranged in order according to their size (starting with the smallest), and if x is the nth member, then $\text{rank}(x) \equiv r_x = n$. The correlation is then calculated for r_x and r_y instead, which can be shown to simplify to

$$\rho_{\text{rank}} = \rho_{r_x r_y} = 1 - \frac{6 \sum_{i=1}^{N} (r_{x_i} - r_{y_i})^2}{N(N^2 - 1)}. \tag{1.18}$$

For example, if six measurements of x yielded the values $1, 3, 0, 5, 3, 6$ then the corresponding r_x values are $2, 3.5, 1, 5, 3.5, 6$, (where the tied values were all assigned an averaged rank). If measurements of y yielded $2, 3, -1, 5, 4, -99$ (an outlier), then the corresponding r_y values are $3, 4, 2, 6, 5, 1$. The Spearman rank correlation is $+0.12$, whereas in contrast the Pearson correlation is -0.61, which shows the strong influence exerted by an outlier.

An alternative robust and resistant correlation is the biweight midcorrelation (see Section 11.2.1).

1.3.2 Autocorrelation

To determine the degree of autocorrelation in a time series, we use the autocorrelation coefficient, where a copy of the time series is shifted in time by a lag of l time intervals, and then correlated with the original time series. The lag-l autocorrelation coefficient is given by

$$\rho(l) = \frac{\sum_{i=1}^{N-l}[(x_i - \overline{x})(x_{i+l} - \overline{x})]}{\sum_{i=1}^{N}(x_i - \overline{x})^2}, \tag{1.19}$$

where $\overline{x}$ is the sample mean. There are other estimators of the autocorrelation function, besides the non-parametric estimator given here (von Storch and Zwiers, 1999, p. 252). The function $\rho(l)$, which has the value 1 at lag 0, begins to decrease as the lag increases. The lag where $\rho(l)$ first intersects the l-axis is l_0, the *first zero crossing*. A crude estimate for the *effective sample size* is $N_{\text{eff}} = N/l_0$. From symmetry, one defines $\rho(-l) = \rho(l)$. In practice, $\rho(l)$ cannot be estimated reliably when l approaches N, since the numerator of (1.19) would then involve summing over very few terms.

The autocorrelation function can be integrated to yield a *decorrelation time scale* or *integral time scale*

$$\begin{aligned} T &= \int_{-\infty}^{\infty} \rho(l)\, \mathrm{d}l \quad \text{(continuous case)} \\ &= \left(1 + 2\sum_{l=1}^{L} \rho(l)\right) \Delta t \quad \text{(discrete case)}, \end{aligned} \tag{1.20}$$

where Δt is the time increment between adjacent data values, and the maximum lag L used in the summation is usually not more than $N/3$, as $\rho(l)$ cannot be estimated reliably when l becomes large. The effective sample size is then

$$N_{\text{eff}} = N\Delta t/T, \tag{1.21}$$

with $N\Delta t$ the data record length. When the decorrelation time scale is large, $N_{\text{eff}} \ll N$.

With two time series x and y, both with N samples, the effective sample size is often estimated by

$$N_{\text{eff}} = \frac{N}{\sum_{l=-L}^{L}\left[\rho_{xx}(l)\rho_{yy}(l) + \rho_{xy}(l)\rho_{yx}(l)\right]}, \tag{1.22}$$

(Emery and Thomson, 1997), though sometimes the $\rho_{xy}\rho_{yx}$ terms are ignored (Pyper and Peterman, 1998).

1.3.3 Correlation matrix

If there are M variables, e.g. M stations reporting the air pressure, then correlations between the variables lead to a correlation matrix

$$\mathbf{C} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2M} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{M1} & \rho_{M2} & \cdots & \rho_{MM} \end{bmatrix}, \tag{1.23}$$

where ρ_{ij} is the correlation between the ith and the jth variables. The diagonal elements of the matrix satisfy $\rho_{ii} = 1$, and the matrix is symmetric, i.e. $\rho_{ij} = \rho_{ji}$. The jth column of $\mathbf{C}$ gives the correlations between the variable j and all other variables.

1.4 Regression

Regression, introduced originally by Galton (1885), is used to find a linear relation between a dependent variable y and one or more independent variables $\mathbf{x}$.

1.4.1 Linear regression

For now, consider simple linear regression where there is only one independent variable x, and the dataset contains N pairs of (x, y) measurements. The relation is

$$y_i = \tilde{y}_i + e_i = a_0 + a_1 x_i + e_i, \qquad i = 1, \ldots, N, \tag{1.24}$$

where a_0 and a_1 are the regression parameters, $\tilde{y}_i$ is the y_i predicted or described by the linear regression relation, and e_i is the error or the residual unaccounted for by the regression (Fig. 1.2). As regression is commonly used as a prediction tool (i.e. given x, use the regression relation to predict y), x is referred to as the *predictor* or independent variable, and y, the *predictand*, response or dependent variable. Curiously, the term 'predictand', widely used within the atmospheric–oceanic community, is not well known outside.

The error

$$e_i = y_i - \tilde{y}_i = y_i - a_0 - a_1 x_i. \tag{1.25}$$

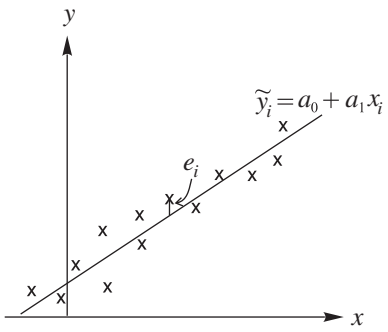

Fig. 1.2 Illustrating linear regression. A straight line $\tilde{y}_i = a_0 + a_1 x_i$ is fitted to the data, where the parameters a_0 and a_1 are determined from minimizing the sum of the square of the error e_i, which is the vertical distance between the ith data point and the line.

By finding the optimal values of the parameters a_0 and a_1, linear regression minimizes the sum of squared errors (SSE) ,

$$\text{SSE} = \sum_{i=1}^{N} {e_i}^2, \tag{1.26}$$

yielding the best straight line relation between y and x. Because the SSE is minimized, this method is also referred to as the *least squares* method.

Differentiation of (1.26) by a_0 yields

$$\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i) = 0. \tag{1.27}$$

Differentiation of (1.26) by a_1 gives

$$\sum_{i=1}^{N} (y_i - a_0 - a_1 x_i) x_i = 0. \tag{1.28}$$

These two equations are called the *normal equations*, from which we will obtain the optimal values of a_0 and a_1.

From (1.27), we have

$$a_0 = \frac{1}{N} \sum y_i - \frac{a_1}{N} \sum x_i, \quad \text{i.e. } a_0 = \overline{y} - a_1 \overline{x}. \tag{1.29}$$

Substituting (1.29) into (1.28) yields

$$a_1 = \frac{\sum x_i y_i - N \overline{x}\, \overline{y}}{\sum x_i^2 - N \overline{x}\, \overline{x}}. \tag{1.30}$$

Equations (1.29) and (1.30) provide the optimal values of a_0 and a_1 for minimizing the SSE, thereby yielding the best straight line fit to the data in the x-y plane. The parameter a_1 gives the slope of the regression line, while a_0 gives the y-intercept.

1.4.2 Relating regression to correlation

Since regression and correlation are two approaches to extract linear relations between two variables, one would expect the two to be related. Equation (1.30) can be rewritten as

$$a_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}. \tag{1.31}$$

Comparing with the expression for the sample correlation, (1.15), we see that

$$a_1 = \rho_{xy}\frac{\sigma_y}{\sigma_x}, \tag{1.32}$$

i.e. the slope of the regression line is the correlation coefficient times the ratio of the standard deviation of y to that of x.

It can also be shown that

$$\sigma_e^2 = \sigma_y^2(1 - \rho_{xy}^2), \tag{1.33}$$

where $1-\rho_{xy}^2$ is the fraction of the variance of y not accounted for by the regression. For example, if $\rho_{xy} = 0.5$, then $1 - \rho_{xy}^2 = 0.75$, i.e. 75% of the variance of y is not accounted for by the regression.

1.4.3 Partitioning the variance

It can be shown that the variance, i.e. the total sum of squares (SST), can be partitioned into two: the first part is that accounted for by the regression relation, i.e. the sum of squares due to regression (SSR), and the remainder is the sum of squared errors (SSE):

$$\text{SST} = \text{SSR} + \text{SSE}, \tag{1.34}$$

where

$$\text{SST} = \sum_{i=1}^{N}(y_i - \overline{y})^2, \tag{1.35}$$

$$\text{SSR} = \sum_{i=1}^{N}(\tilde{y}_i - \overline{y})^2, \tag{1.36}$$

$$\text{SSE} = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2. \tag{1.37}$$

How well the regression fitted the data can be characterized by

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}, \tag{1.38}$$

where R^2 approaches 1 when the fit is very good. Note that R is called the *multiple correlation coefficient*, as it can be shown that it is the correlation between $\tilde{y}$ and y (Draper and Smith, 1981, p. 46), and this holds even when there are multiple predictors in the regression, a situation to be considered in the next subsection.

1.4.4 Multiple linear regression

Often one encounters situations where there are multiple predictors x_l, $(l = 1, \ldots, k)$ for the response variable y. This type of multiple linear regression (MLR) has the form

$$y_i = a_0 + \sum_{l=1}^{k} x_{il} a_l + e_i, \qquad i = 1, \ldots, N. \tag{1.39}$$

In vector form,

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}, \tag{1.40}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & \cdots & x_{kN} \end{bmatrix}, \tag{1.41}$$

$$\mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}. \tag{1.42}$$

The SSE is then

$$\text{SSE} = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{a})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{a}), \tag{1.43}$$

where the superscript T denotes the transpose. To minimize SSE with respect to $\mathbf{a}$, we differentiate the SSE by $\mathbf{a}$ and set the derivatives to zero, yielding the *normal equations*,

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{a}) = \mathbf{0}. \tag{1.44}$$