Ch.1 Correlation & Regression

1.1 Mean and Variance [Book, Sect.1.1, 1.2]

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 $i = 1,2,3,4,5,6$.

The *expectation* or expected value of x from a population is given by

$$\mathrm{E}[x] = \sum_i x_i P_i, \tag{1}$$

where P_i is the probability of x_i occurring. If the die is fair, $P_i = 1/6$ for all i.
Q1: What is $\mathrm{E}[x]$ for a fair die?

We also write

$$\mathrm{E}[x] = \mu_x, \tag{2}$$

with μ_x denoting the *mean* of x for the population.
The expectation of a sum of random variables satisfies

$$\mathrm{E}[ax + by + c] = a\,\mathrm{E}[x] + b\,\mathrm{E}[y] + c, \tag{3}$$

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,

$$\mathrm{E}[x] = \int_\Omega x p(x)\,\mathrm{d}x, \tag{4}$$

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

$$\begin{aligned} \mathrm{E}[f(x)] &= \int_\Omega f(x)p(x)\,\mathrm{d}x \quad \text{(continuous case)} \\ &= \sum_i f(x_i)P_i \quad \text{(discrete case).} \end{aligned} \tag{5}$$

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

$$\overline{x} \equiv \langle x \rangle = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{6}$$

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

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

$$\begin{aligned}\mathrm{var}(x) &\equiv \mathrm{E}[(x-\mu_x)^2] = \mathrm{E}[x^2-2x\mu_x+\mu_x^2] = \mathrm{E}[x^2]+\mathrm{E}[-2x\mu_x]+\mathrm{E}[\mu_x^2] \\ &= \mathrm{E}[x^2] - 2\mu_x\mathrm{E}[x] + \mu_x^2 = \mathrm{E}[x^2] - \mu_x^2, \qquad (7)\end{aligned}$$

where (3) and (2) have been invoked.

The standard deviation s is the positive square root of the population variance, i.e.

$$s^2 = \mathrm{var}(x). \qquad (8)$$

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

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2. \qquad (9)$$

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 (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*

$$x_s = (x - \overline{x})/\sigma\,, \tag{10}$$

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

$$\mathrm{cov}(x, y) = \mathrm{E}[(x - \mu_x)(y - \mu_y)]. \tag{11}$$

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

$$\mathrm{var}(x) = \mathrm{cov}(x, x). \tag{12}$$

The sample covariance is computed as

$$\mathrm{cov}(x, y) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}). \tag{13}$$

1.2 Correlation: [Book, Sect.1.3]

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

$$\hat{\rho}_{xy} = \frac{\mathrm{cov}(x, y)}{s_x s_y} \,, \tag{14}$$

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

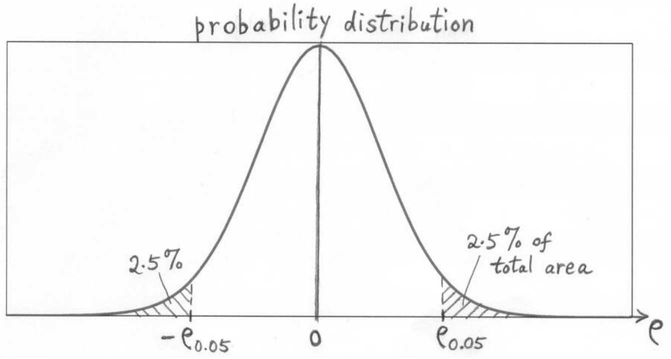
For a sample containing N pairs of $(x,\ y)$ measurements or observations, the *sample correlation* is computed by

$$\rho \equiv \rho_{xy} = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_{i=1}^{N}(x_i - \overline{x})^2\right]^{\frac{1}{2}} \left[\sum_{i=1}^{N}(y_i - \overline{y})^2\right]^{\frac{1}{2}}} \,, \tag{15}$$

which lies between -1 and $+1$.

Test of null hypothesis: Can the obtained sample correlation be considered significantly different from 0? This is also called a test of the null (i.e. $\hat{\rho}_{xy} = 0$) hypothesis.

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?



Use *t* test. [See Book, pp.3-4]. Significance test very dependent on sample size *N*.

Autocorrelation: Often the observations are measurements at regular time intervals, i.e. time series, and neighbouring data points in the time series are correlated. 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 fewer 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.

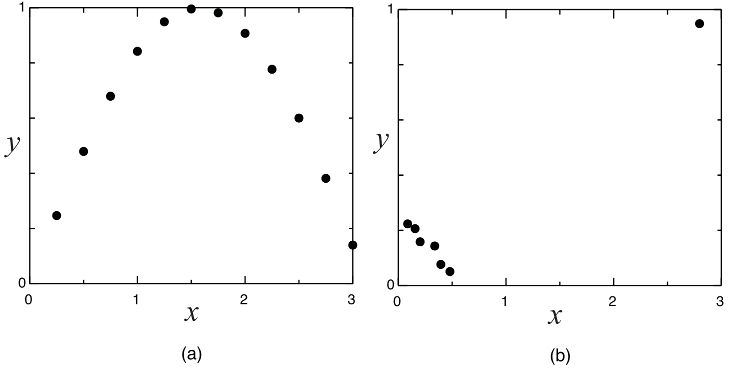


Figure : (a) Correlation is not robust to deviations from linearity, as the nonlinear relation between x and y is missed by $\rho \approx 0$. (b) Correlation is not resistant to outliers, without the single outlier, the correlation coefficient changes from positive to negative.

Spearman rank correlation

For the correlation to be more robust and resistant to outliers, the Spearman rank correlation is often used.
Arrange the data $\{x_1, \ldots, x_N\}$ in the 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.

E.g., 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.

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

$$\rho(l) = \frac{\displaystyle\sum_{i=1}^{N-l}[(x_i - \overline{x})(x_{i+l} - \overline{x})]}{\displaystyle\sum_{i=1}^{N}(x_i - \overline{x})^2} \quad , \tag{16}$$

where $\overline{x}$ is the sample mean.

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 $N_{\text{eff}} = N/l_0$.

Q2: Suppose for a time series with 500 monthly values the autocorrelation function has values 0.72, 0.50, 0.31, 0.18, 0.04, -0.21, -0.33, -0.45, -0.24, -0.11, 0.15, 0.29, 0.37, 0.22, 0.06, ... for lags of 1, 2, ..., 15. Estimate the first zero crossing l_0 and the effective sample size.

Correlation functions in Matlab

`www.mathworks.com/help/techdoc/ref/corrcoef.html`
rho = corrcoef(x,y)
[rho, p] = corrcoef(x,y),
where the p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero, based on the *t* test.

Alternatively, in the Matlab Statistics Toolbox:

`www.mathworks.com/help/toolbox/stats/corr.html`

rho = corr(x,y)

[rho, p] = corr(x,y)

This has Spearman rank correlation as an option.

1.3 Regression [Book, Sect.1.4]

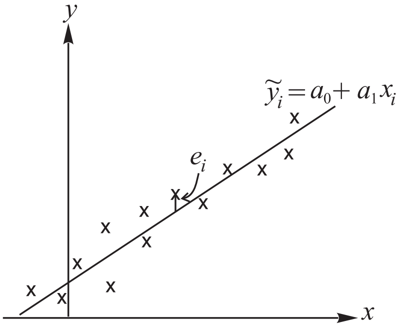
Regression is used to find a linear relation between a dependent variable y and one or more independent variables $\mathbf{x}$.

Linear regression

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

$$y_i = \tilde{y}_i + e_i = a_0 + a_1 x_i + e_i, \qquad i = 1, \ldots, N, \tag{17}$$

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



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.

The error

$$e_i = y_i - \tilde{y}_i = y_i - a_0 - a_1 x_i. \tag{18}$$

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

$$\mathrm{SSE} = \sum_{i=1}^{N} {e_i}^2 = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i)^2, \tag{19}$$

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.

$$\frac{\partial \mathrm{SSE}}{\partial a_0} = 0 \quad \Rightarrow \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i) = 0. \tag{20}$$

$$\frac{\partial \mathrm{SSE}}{\partial a_1} = 0 \quad \Rightarrow \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i) x_i = 0. \tag{21}$$

These two equations are called the *normal equations*, from which we will obtain the optimal values of a_0 and a_1.
From (20), we have

$$a_0 = \frac{1}{N}\sum y_i - \frac{a_1}{N}\sum x_i\,, \quad \text{i.e.}\;\; a_0 = \overline{y} - a_1\overline{x}. \tag{22}$$

Substituting (22) into (21) yields

$$a_1 = \frac{\sum x_i y_i - N\overline{x}\,\overline{y}}{\sum x_i^2 - N\overline{x}\,\overline{x}}. \tag{23}$$

Relating regression to correlation

As regression and correlation are two approaches to extract linear relations between two variables, the two methods are related.
Eq.(23) can be rewritten as

$$a_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2} \tag{24}$$

Using Eq. (15), we get

$$a_1 = \rho_{xy} \frac{\sigma_y}{\sigma_x}\,, \tag{25}$$

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 for the standard deviation of the error (σ_e):

$$\sigma_e^2 = \sigma_y^2 (1 - \rho_{xy}^2), \tag{26}$$

where $1 - \rho_{xy}^2$ is the fraction of the variance of y not accounted for by the regression.

Q3: If $\rho_{xy} = 0.7$, what % of the variance of y is not accounted for by the regression?

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) :

$$\text{SST} = \text{SSR}+\text{SSE}, \tag{27}$$

where

$$\text{SST} = \sum_{i=1}^{N} (y_i - \overline{y})^2, \tag{28}$$

$$\text{SSR} = \sum_{i=1}^{N} (\tilde{y}_i - \overline{y})^2, \tag{29}$$

$$\text{SSE} = \sum_{i=1}^{N} (y_i - \tilde{y}_i)^2. \tag{30}$$

How well the regression fitted the data can be characterized by

$$R^2 = \frac{\mathsf{SSR}}{\mathsf{SST}} = 1 - \frac{\mathsf{SSE}}{\mathsf{SST}}\,, \tag{31}$$

where R^2 approaches 1 when the fit is very good. 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.

Multiple linear regression (MLR)

Multiple predictors $x_l, (l = 1, \cdots, k)$ for the response variable y:

$$y_i = a_0 + \sum_{l=1}^{k} x_{il} a_l + e_i, \qquad i = 1, \cdots, N. \tag{32}$$

In vector form,

$$\mathbf{y} = \mathbf{X}\mathbf{a} + \mathbf{e}, \tag{33}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1N} & \cdots & x_{kN} \end{bmatrix}, \tag{34}$$

$$\mathbf{a} = \begin{bmatrix} a_0 \\ \vdots \\ a_k \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}. \tag{35}$$

$$\mathrm{SSE} = \mathbf{e}^{\mathrm{T}}\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{a})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{a}), \tag{36}$$

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

$$\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{a}) = \mathbf{0}. \tag{37}$$

$$\Rightarrow \quad \mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{a} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \quad \Rightarrow \quad \mathbf{a} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{y}. \tag{38}$$

Stepwise regression

Sometimes there are many possible predictors. When all possible predictors are used in building an MLR model, one often '*overfits*' the data (esp. when the sample size is relatively small), i.e. too many parameters are used in the model so that one is simply fitting to the noise in the data.

Stepwise regression automatically eliminates insignificant predictors:
(1) Start with no predictors.
(2) Forward selection: try out the predictors one by one, include them if they are 'statistically significant'.
(3) Backward elimination: test the candidate predictors one by one for statistical significance, deleting any which are not significant.
(4) Iterate steps (2) and (3) until no more changes.

Perfect Prog and MOS

Physical (or dynamical) prediction models have surpassed statistical models in many fields.
E.g. in numerical weather forecasting, dynamical models can be integrated forward in time to give weather forecasts. Nevertheless regression is commonly used to improve the raw forecasts from dynamical models.
Why? Variables in the dynamical model usually have poor resolution and are sometimes too idealized. E.g., the lowest temperature level in the model may be some distance above the ground. Some local variable (e.g. ozone concentration) may not even be variables carried in the dynamical model.

The **Perfect Prog** (abbreviation for perfect prognosis) scheme computes an MLR from the historical data archive:

$$y(t) = \mathbf{x}(t)^{\mathrm{T}}\mathbf{a} + e(t), \tag{39}$$

During actual forecasting, $\mathbf{x}(t)$ is provided by the forecasts from the dynamical model, and $y(t)$ is predicted by the MLR.

Problem: while the regression model was developed or trained using historical data for $\mathbf{x}$, the actual forecasts used the dynamical model forecasts for $\mathbf{x}$. Hence, the systematic error between the dynamical model forecasts and real data have not been taken into account—i.e. perfect prognosis is assumed.

A better approach is the **model output statistics** (MOS) scheme: the dynamical model forecasts have been archived, so the MLR was developed using $y(t)$ from the data archive and $\mathbf{x}(t)$ from the dynamical model forecast archive.

Since **x** was from the dynamical model forecasts during both model training and actual forecasting, the model bias in Perfect Prog has been eliminated.

While MOS is more accurate than Perfect Prog, it is considerable more difficult to implement since a slight modification of the dynamical model would require the regeneration of the dynamical model forecast archive and the recalculation of the regression relations.

Regression functions in Matlab

`www.mathworks.com/help/toolbox/stats/regress.html`
a = regress(y, X)

`www.mathworks.com/help/toolbox/stats/regstats.html` regstats(y, X, model)

Stepwise regression:
`www.mathworks.com/help/toolbox/stats/stepwisefit.html`
a = stepwisefit(X, y)

Interactive stepwise regression:
`www.mathworks.com/help/toolbox/stats/stepwise.html`
stepwise(X, y)

References

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*. Wiley, New York, 2nd edition.