

Statistical Facts

S.R. Carpenter, Zoology 535

In the following material, x_i and y_i are individual observations of random variables x and y , respectively. The number of samples is n . All summations are taken over n . Matrices are uppercase bold; vectors are lowercase bold.

Estimators

The estimator for the mean of the population (which has the same units as the observations x_i) is

$$\bar{x} = (1/n) \sum x_i$$

The sum of squared deviations from the mean is

$$SS_x = \sum (x_i - \bar{x})^2$$

The variance of the population (in the square of the units of the data) is estimated by

$$\text{var}(x) = SS_x / (n - 1)$$

The standard deviation is an index of the variability of the population in the same units as the mean;

$$s_x = \text{var}(x)^{0.5}$$

The standard error is an index of the precision of the estimate of the mean:

$$\text{standard error or s.e.} = [\text{var}(x)/n]^{0.5}$$

The sum of cross products of x and y is

$$SCP(x,y) = \sum (x_i - \bar{x})(y_i - \bar{y})$$

The covariance of x and y is

$$\text{cov}(x,y) = SCP(x,y)/(n-1)$$

The correlation coefficient of x and y (which is dimensionless) is

$$r_{x,y} = SCP(x,y)/(SS_x SS_y)^{0.5} = \text{cov}(x,y)/[\text{var}(x) \text{var}(y)]^{0.5}$$

Error Propagation

Suppose z is a random variable calculated from the random variables x and y ; $z = f(x,y)$. Then in general

$$\text{var}(z) \approx [(\text{df}/\text{dx})^2 \text{var}(x)] + [(\text{df}/\text{dy})^2 \text{var}(y)] + 2 [(\text{df}/\text{dx}) (\text{df}/\text{dy}) \text{cov}(x,y)]$$

All the derivatives are calculated at the mean values x^* and y^* . This equation is sometimes called "first-order error propagation" because it is obtained from the first-derivative term of a Taylor expansion of $f(x,y)$. Note that it is an approximation of $\text{var}(z)$ if second order and higher derivatives of f are not zero.

Below are listed some useful specific formulas for common situations. Note that if x and y can be assumed to be independent, their covariance is 0 and that term drops out of the equation.

Sum: $z = x + y$

$$\text{var}(z) = \text{var}(x) + \text{var}(y) + 2 \text{cov}(x,y)$$

Difference: $z = x - y$

$$\text{var}(z) = \text{var}(x) + \text{var}(y) - 2 \text{cov}(x,y)$$

Multiplication by a constant: $z = c x$

$$\text{var}(z) = c^2 \text{var}(x)$$

Product: $z = x y$

$$\text{var}(z) = y^{*2} \text{var}(x) + x^{*2} \text{var}(y) + 2 x^* y^* \text{cov}(x,y)$$

Quotient: $z = x/y$

$$\text{var}(z) = z^{*2} [(\text{var}(x)/x^{*2}) + (\text{var}(y)/y^{*2}) - 2 (\text{cov}(x,y)/x^* y^*)]$$

Product raised to powers: $z = x^m y^n$

$$\text{var}(z) = z^{*2} [m^2 (\text{var}(x)/x^{*2}) + n^2 (\text{var}(y)/y^{*2}) + 2 m n (\text{cov}(x,y)/x^* y^*)]$$

Linear Regression

Linear regression is the process of fitting equations linear in the parameters to data. The equation may involve nonlinear combinations of predictor (independent) variables. For example,

$$z = c x^2 y^{0.73} y^{2x} \sin(x^7) e^y$$

where x and y are predictors and c is a parameter to be estimated, is linear in c even though it is nonlinear in x and y . The fitting of equations nonlinear in parameters is a more complicated topic that will be covered elsewhere.

The equations below assume that sequential observations of the predictors are uncorrelated. It is also assumed that any errors in measuring the predictors are much smaller than the total errors in the response variable (i.e. the variance in observing the response variable, plus the variance due to factors not included in the regression model). When these assumptions are not met, the parameter estimates may be seriously biased and the estimates of the errors are unreliable. Remediation of these problems will be covered elsewhere.

In the simple case of a straight line model,

$$y = b_0 + b_1 x + \varepsilon$$

where y is the response (dependent) variable, x is the predictor (independent) variable, b_0 and b_1 are estimators of the parameters, and ε is the error of the regression, then the parameters are estimated as

$$b_1 = \text{SCP}(x,y)/\text{SS}_x = r_{x,y} (s_y/s_x)$$

$$b_0 = y^* - b_1 x^*$$

In the more general case of regression with any number of predictors or combinations of predictors, it is easiest to adopt matrix notation.

\mathbf{y} is a column vector of values of the response variable

\mathbf{X} is a matrix of predictors, where each column is values of a predictor or a combination of predictors. To include an intercept in the model, include a column of ones in \mathbf{X} .

\mathbf{b} is a column vector of parameters. Each row is the parameter for the corresponding column of \mathbf{X} .

ε is a column vector of errors with standard deviation s_ε .

In this notation, the model is written

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon}$$

and the estimated values of y are calculated as $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}$.

The parameter estimates are calculated as

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

The error standard deviation is estimated as

$$s_{\varepsilon}^2 = (\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y})/(n-p)$$

where p is the number of parameters estimated.

The variance of the parameters is

$$\text{var}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{X})^{-1} s_{\varepsilon}^2$$

The variance of a prediction $y_0 = \mathbf{x}_0'\hat{\mathbf{b}}$ is

$$\text{var}(y_0) = \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 s_{\varepsilon}^2$$

References

Draper, N. and H. Smith. 1981. Applied Regression Analysis. Second edition. Wiley, NY.

Meyer, S.L. 1975. Data Analysis for Scientists and Engineers. Wiley, NY.