it easy to program new statistical methods. The graphics of the language allow easy production of advanced, publication-quality graphics. Since a wide variety of experts use the program, R includes a comprehensive library of statistical functions, including many cutting-edge statistical methods. In addition to this, many third-party specialized methods are publicly available. And most important, R is free and open source.

A common concern of beginning users of R is the steep learning curve involved in using it. Such concern stems from the fact that R is a command-driven environment. Consequently, the statistical analysis is performed in a series of steps, in which commands are typed out and the results from each step are stored in objects that can be used by further inquiries. This is contrary to other programs, such as SPSS and SAS, which require users to determine all characteristics of the analysis up front and provide extensive output, thus relying on the users to identify what is relevant to their initial question.

Another source of complaints relates to the difficulty of writing new functions. The more complex the function, the more difficult it becomes to identify errors in syntax or logic. R will prompt the user with an error message, but no indication is given of the nature of the problem or its location within the new code. Consequently, despite the advantage afforded by being able to add new functions to R, many users may find it frustrating to write new routines. In addition, complex analyses and simulations in R tend to be very demanding on the computer memory and processor; thus, the more complex the analysis, the longer the time necessary to complete the task, sometimes days.

Large data sets or complex tasks place heavy demands on computer RAM, resulting in slow output.

Brandon K. Vaughn and Aline Orr

See also SAS; SPSS; Statistica; Systat

Web Sites

Comprehensive R Archive Network (CRAN): http://CRAN.R-project.org

The R Project for Statistical Computing: http://www.r-project.org

R^2

R-squared (R^2) is a statistic that explains the amount of variance accounted for in the relationship between two (or more) variables. Sometime R^2 is called the coefficient of determination, and it is given as the square of a correlation coefficient.

Given paired variables (X_i, Y_i), a linear model that explains the relationship between the variables is given by

$$Y = \beta_0 + \beta_1 X + e,$$

where e is a mean zero error. The parameters of the linear model can be estimated using the least squares method and denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The parameters are estimated by minimizing the sum of squared residuals between variable Y_i and the model $\beta_0 + \beta_1 X_i$, that is, $(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} (Y_i - \beta_0 + \beta_1 X_i)^2$.

It can be shown that the least squares estimations are

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$

where the sample cross-covariance S_{xy} is defined as

$$S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \overline{XY} - \bar{X}\bar{Y}.$$

Statistical packages such as SAS, SPLUS, and R provide a routine for obtaining the least squares estimation. The estimated model is denoted as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X.$$

With the above notations, the sum of squared errors (*SSE*), or the sum of squared residuals, is given by

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

SSE measures the amount of variability in Y that is not explained by the model. Then how does one measure the amount of variability in Y that is explained by the model? To answer this question,

one needs to know the total variability present in the data. The total sum of squares (*SST*) is the measure of total variation in the *Y* variable and is defined as

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

where $\bar{Y}$ is the sample mean of *Y* variables, that is,

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Since *SSE* is the minimum of the sum of squared residuals of any linear model, *SSE* is always smaller than *SST*. Then the amount of variability explained by the model is *SST* − *SSE*, which is denoted as the regression sum of squares (*SSR*), that is,

$$SST = SST - SSE.$$

The ratio $SSR/SST = (SST - SSE)/SST$ measures the proportion of variability explained by the model. The coefficient of determination (R^2) is defined as the ratio

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SSE}.$$

The coefficient of determination is given as the ratio of variations explained by the model to the total variations present in *Y*. Note that the coefficient of determination ranges between 0 and 1. R^2 value is interpreted as the proportion of variation in *Y* that is explained by the model. $R^2 = 1$ indicates that the model exactly explains the variability in *Y*, and hence the model must pass through every measurement (X_i, Y_i). On the other hand, $R^2 = 0$ indicates that the model does not explain any variability in *Y*. R^2 value larger than .5 is usually considered a significant relationship.

Case Study and Data

Consider the following paired measurements from Moore et al. (1989), based on occupational mortality records from 1970 to 1972 in England and Wales. The figures represent smoking rates and deaths from lung cancer for a number of occupational groups.

77	84
137	116
117	123
94	128
116	155
102	101
111	118
93	113
88	104
102	88
91	104
104	129
107	86
112	96
113	144
110	139
125	113
133	146
115	128
105	115
87	79
91	85
100	120
76	60
66	51

For a set of occupational groups, the first variable is the smoking index (average 100), and the second variable is the lung cancer mortality index (average 100). Suppose we are interested in determining how much the lung cancer mortality index (*Y* variable) is influenced by the smoking index (*X* variable). Figure 1 shows the scatterplot of the smoking index versus the lung cancer mortality index. The straight line is the estimated linear model, and it is given by

$$Y = -2.8853 + 1.0875X.$$

SSE can be easily computed using the formula

$$SSE = \sum_{i=1}^{n} Y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} Y_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i Y_i, \qquad (1)$$

and *SST* can be computed using the formula

$$SST = \sum_{i=1}^{n} Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} Y_i \right)^2. \qquad (2)$$

In this example the coefficient of determination is .5121, indicating that the smoking index can explain the lung cancer mortality index.

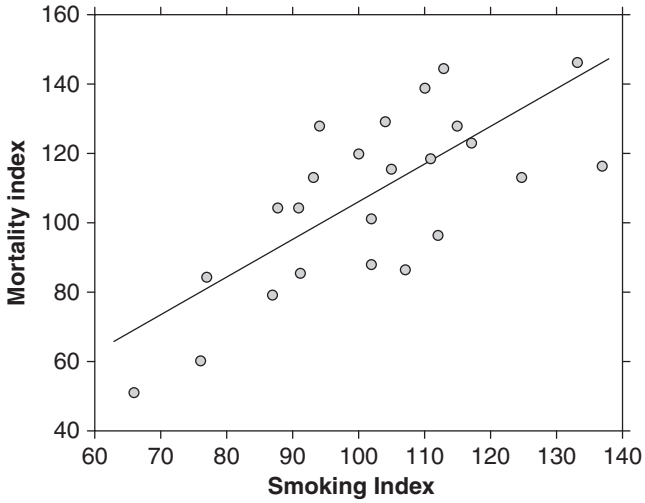

Figure 1 Scatterplot of Smoking Index Versus Lung Cancer Mortality Index

Note: The straight line is the linear model fit obtained by the least squares estimation.

Relation to Correlation Coefficient

With the previous Equations 1 and 2, R^2 can also be written as a function of the sample cross-covariance:

$$SSE = nS_{yy} - n\frac{{S_{xy}}^2}{S_{xx}} \quad \text{and} \quad SST = nS_{yy}.$$

Then the coefficient of determination can be written as

$$\mathrm{R}^2 = \frac{{S_{xy}}^2}{S_{xx}S_{yy}} = \frac{\left(\overline{XY} - \bar{X}\bar{Y}\right)^2}{(\overline{X^2} - \bar{X}^2)(\overline{Y^2} - \bar{Y}^2)},$$

which is the square of the Pearson product-moment correlation coefficient.

$$\mathrm{R} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})\sum_{i=1}^{n}(Y_i - \bar{Y})}}.$$

In the above example, the correlation coefficient is .7162, so the correlation square is .5121, the R^2 value.

R^2 for General Cases

The definition of the coefficient of determination can be further expanded in the case of multiple regression. Consider the following multiple regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e,$$

where Y is the response variable and $X_1, X_2, \cdots, X_p$ are p regressors, and e is a mean zero error. The unknown parameters $\beta_1, \beta_2, \cdots, \beta_p$ are estimated by the least squares method. The sum of squared residuals is given by

$$\begin{aligned} SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n}\left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)\right]^2, \end{aligned}$$

while the total sum of squares is given by

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Then the coefficient of multiple determination is given by

$$R^2 = \frac{SST - SSE}{SST},$$

which is the square of the multiple correlation coefficient R. As the number of regressors increases, the R^2 value also increases, so R^2 cannot be a useful measure for the goodness of model fit. Therefore, R^2 is adjusted for the number of explanatory variables in the model. The adjusted R^2 is defined as

$$R^2{}_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1} = \frac{(n-1)R^2 - p}{n-p+1}.$$

It can be shown that $R^2{}_{\text{adj}} \leq R^2$. The coefficient of determination can be further generalized in more general cases using the likelihood method.

Moo K. Chung

Further Readings

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika, 10,* 507–521.

Moore, D. S., & McCabe, G. P. (1989). *Introduction to the practice of statistics.* New York: W. H. Freeman.

Nagelkerke, N. J. D. (1992). Maximum likelihood estimation of functional relationships. *Lecture Notes in Statistics, 69,* 110.

Radial Plot

The radial plot is a graphical method for displaying and comparing observations that have differing precisions. Standardized observations are plotted against the precisions, where precision is defined as the reciprocal of the standard error. The original observations are given by slopes of lines through the origin. A scale of slopes is sometimes drawn explicitly.

Suppose, for example, that data are available on the degree classes obtained by students graduating from a university and that we wish to compare, for different major subjects, the proportions of students who achieved upper second-class honors or higher. Typically, different numbers of students graduate in different subjects. A radial plot will display the data as proportions so that they may be compared easily. Similarly, a radial plot can be used to compare other summary statistics (such as means, regression coefficients, odds ratios) observed for different sized groups, or event rates observed for differing time periods.

Sometimes, particularly in the natural and physical sciences, measurements intrinsically have differing precisions because of natural variation in the source material and experimental procedure. For example, archaeological and geochronological dating methods usually produce an age estimate and its standard error for each of several crystal grains or rock samples, and the standard errors differ substantially. In this case, the age estimates may be displayed and compared using a radial plot in order to examine whether they agree or how they differ. A third type of application is in *meta-analysis,* such as in medicine, to compare estimated treatment effects from different studies. Here the precisions of the estimates can vary greatly because of the differing study sizes and designs. In this context the graph is often called a *Galbraith plot.* In general, a radial plot is applicable when one wants to compare a number of estimates of some parameter of interest, for which the estimates have different standard errors.

A basic question is, Do the estimates agree (within statistical variation) with a common value? If so, what value? A radial plot provides a visual assessment of the answer. Also, like many graphs, it allows other features of the data to be seen, such as whether the estimates differ systematically in some way, perhaps due to an underlying factor or mixture of populations, or whether there are anomalous values that need explanation. It is inherently not straightforward to compare individual estimates, either numerically or graphically, when their precisions vary. In particular, simply plotting estimates with error bars does not allow such questions to be assessed.

The term *radial plot* is also used for a display of *directional data,* such as wind directions and velocities or quantities observed at different times of day, via radial lines of different lengths emanating from a central point. This type of display is not discussed in this entry.

Mathematical Properties

Let $z_1, z_2, \ldots, z_n$ denote n observations or estimates having standard errors $\sigma_1, \sigma_2, \ldots, \sigma_n$, which are either known or well estimated. Then we plot the points (x_i, y_i) given by $x_i = 1/\sigma_i$ and $y_i = (z_i - {}_0)/\sigma_i$, where z_0 is a convenient reference value. Each y_i has unit standard deviation, so each point has the same standard error with respect to the y scale, but estimates with higher precision plot farther from the origin on the x scale. The (centered) observation $(z_i - z_0)$ is equal to y_i/x_i, which is the slope of the line joining (0, 0) and (x_i, y_i), so that values of z can be shown on a scale of slopes. Figure 1 illustrates these principles.

Furthermore, if each z_i is an unbiased estimate of the *same* quantity μ, say, then the points will scatter with unit standard deviation about a line from (0, 0) with slope $\mu - z_0$. In particular, points scattering with unit standard deviation about the *horizontal* radius agree with the reference value z_0. This provides a simple visual assessment of how