Sociology 405/805
Revised February 24, 2004

Summary of Formulae for Bivariate Regression and Correlation

Let X be an independent variable and Y a dependent variable, with n observations for each of the values of these two variables. Preferably, both X and Y are measured at the interval or ratio level, although it is also common to estimate correlation coefficients and regression lines when one or both variables are measured at only the ordinal level.

The first stage in obtaining the estimates of correlation and regression statistics is to compute ΣX, ΣY, ΣX^2, ΣY^2 and ΣXY. Each summation is across all n values of X and Y. Then use these summations to calculate the following expressions:

$$S_{XX} = \Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

$$S_{XY} = \Sigma XY - \frac{(\Sigma X)(\Sigma Y)}{n}$$

$$S_{YY} = \Sigma Y^2 - \frac{(\Sigma Y)^2}{n}$$

Correlation coefficient

Using the above expressions, the correlation coefficient is

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}.$$

Regression line

The slope b and the intercept a of the regression line are

$$b = \frac{S_{XY}}{S_{XX}}$$

$$a = \bar{Y} - b\bar{X}$$

where $\bar{Y} = \Sigma Y/n$ and $\bar{X} = \Sigma X/n$.
The estimate of the regression line expressing the relationship between the dependent variable Y and the independent variable X is

$$\hat{Y} = a + bX.$$

Standard errors

For this regression line, the standard error of estimate is

$$s_e = \sqrt{\frac{\Sigma Y^2 - a\Sigma Y - b\Sigma XY}{n-2}}$$

and the standard deviation of b is

$$s_b = \frac{s_e}{\sqrt{S_{XX}}}.$$

The standard deviation of the mean predicted value $\hat{Y}$ is

$$s_{\hat{Y}} = s_e\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}}$$

and the standard deviation for an individual predicted value, $\hat{Y}_i$, is

$$s_{\hat{Y}_i} = s_e\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{S_{XX}}}.$$

Components of the Variation in the Dependent Variable Y

The total variation in the dependent variable is

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

This total variation can be broken into two components, the explained variation, or regression sum of squares, and the unexplained, or residual sum of squares.

The regression sum of squares is

$$SS_r = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = b^2 S_{XX}.$$

The unexplained variation, or the residual or error sum of squares is

$$SS_e = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY.$$

These two components of the total variation can be used to determine R^2, the goodness of fit of the regression equation.

$$R^2 = \frac{SS_r}{SS_t}$$

Tests of Statistical Significance

There are various ways of testing for the statistical significance of the regression line. For each test, the null hypothesis is that there is no relationship between X and Y. The alternative hypothesis can be constructed as either a one or two directional statement. These can be stated in general as:

H_0: No relationship between X and Y

H_1: Some relationship between X and Y

Alternatively, the research hypothesis can be stated as a one directional relationship, either a positive or a negative relationship between X and Y.

For an hypothesis test about the goodness of fit R^2, the hypotheses are:

$H_0 : R^2 = 0$

$H_1 : R^2 \neq 0$

The test for R^2 is an F test with 1 and $(n-2)$ degrees of freedom and can be written as:

$$F = \frac{SS_r}{SS_e/(n-2)} = \frac{R^2}{1-R^2}(n-2).$$

The tests of significance for the correlation coefficient r, and for the slope of the line b, are usually constructed as one directional tests. The null hypothesis is that there is no relationship between X and Y, and the research

hypothesis is either a positive relationship between X and Y, or a negative relationship between the two variables. If the test is to determine whether there is a positive relationship between the two variables, the hypotheses for the test of significance on the Pearson correlation coefficient r would be as follows. Let ρ be the true correlation between X and Y.

$H_0 : \rho = 0$

$H_1 : \rho > 0$

The following t-test with $(n-2)$ degrees of freedom tests these hypotheses:

$$t = r\sqrt{\frac{n-2}{1-r^2}}.$$

To test for a positive slope for the regression line, b, the hypotheses are:

$H_0 : \beta = 0$

$H_1 : \beta > 0$

where β is the slope of the true regression line when Y is regressed on X. This test is usually written as a t-test with $n-2$ degrees of freedom, where

$$t = \frac{b-\beta}{s_b}.$$

If the null hypothesis is that $\beta = 0$, then this test is simply

$$t = \frac{b}{s_b}.$$

Note that for a bivariate relationship, involving only two variables, each of the above three tests is really the same test, so not more than one of these tests need be reported. That is,

$$t = \frac{b}{s_b} = r\sqrt{\frac{n-2}{1-r^2}}$$

and

$$F = \frac{R^2}{1-R^2}(n-2) = t^2.$$

Analysis of variance

The decomposition of the variation of Y is presented as an analysis of variance table in Table 1. Recalling that

$$SS_r = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = b^2 S_{XX}$$

and

$$SS_e = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \Sigma Y^2 - a\Sigma Y - b\Sigma XY,$$

the F test is

$$F = \frac{SS_r}{SS_e/(n-2)} = \frac{R^2}{1-R^2}(n-2).$$

Table 1: Analysis of Variance Table

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
Regression	$SS_r = \sum(\hat{Y}_i - \bar{Y})^2$	1	SS_r	$R^2(n-2)/(1-R^2)$
Residual	$SS_e = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$SS_e/(n-2)$	
Total	$SS_t = \sum(Y_i - \bar{Y})^2$	$n-1$		

Last edited February 24, 2004.