

# Regression Analysis

Regression analysis “fits” or derives a model that describes the variation of a *response* (“dependent”) variable as a function of one or more *predictor* (or “independent”) variables. The general regression model is one of several that share the same basic conceptual model

$$\text{data} = \text{systematic component} + \text{irregular component}$$

where the systematic component is predictable or explainable by the predictor variables, and is represented by the regression model, while the irregular component is regarded as “noise” or prediction errors—variations in the response variable that can not be accounted for by the predictor variables.

## The regression equation

The specific *bivariate, linear*, regression model is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where  $Y_i$  is the  $i$ -th value of the dependent or response variable,  $\beta_0$  and  $\beta_1$  are the coefficients of the regression line (also known as the slope ( $\beta_1$ ) and intercept ( $\beta_0$ )),  $X_i$  is the independent or predictor variable, and  $\varepsilon_i$  is the prediction error, or “noise” or “residual.” (Note that usually in descriptions of regression analysis, upper-case  $X_i$ ’s and  $Y_i$ ’s stand for “raw” data values, while lower case  $x_i$ ’s and  $y_i$ ’s stand for deviations of  $X_i$  and  $Y_i$  about their respective means, i.e.

$$\begin{aligned}x_i &= X_i - \bar{X} \\y_i &= Y_i - \bar{Y}\end{aligned}$$

There are several alternative ways of writing the regression equation or model

- true model, no error:  $Y_i = \beta_0 + \beta_1 X_i$ ,
- true model, with error:  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ,
- true model, no subscripts:  $Y = \beta_0 + \beta_1 X + \varepsilon$ ,
- true model, with error:  $Y = a + bX + e$  (where  $a$  and  $b$  are the regression coefficients and  $e$  is the residuals,

- estimated model  $Y_i = b_0 + b_1 X_i + e_i$ , where  $b_0$  and  $b_1$  are estimates of  $\beta_0$  and  $\beta_1$ , and  $e_i$  is an estimate of  $\varepsilon_i$ ,
- estimated model  $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$ , where  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are estimates of  $\beta_0$  and  $\beta_1$ , and  $e_i$  is an estimate of  $\varepsilon_i$ .

### Other variables and quantities

There are a number of other quantities that are important in the analysis, including:

- the “fitted” or predicted values of the response variable  $Y_i$  (called “y-hat”)

$$\begin{aligned}\hat{Y}_i &= b_0 + b_1 X_i, \\ &= \bar{Y} + b_1(X_i - \bar{X})\end{aligned}$$

- the residuals or prediction errors

$$e_i = Y_i - \hat{Y}_i$$

- the sums of squared deviations and their cross products

$$\begin{aligned}S_X &= \sum_{i=1}^n (X_i - \bar{X})^2 \\ &= \sum_{i=1}^n x_i^2 \\ S_Y &= \sum_{i=1}^n (Y_i - \bar{Y})^2 \\ &= \sum_{i=1}^n y_i^2, \text{ and} \\ S_{XY} &= \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \sum_{i=1}^n x_i y_i\end{aligned}$$

- and the residual sum of squares

$$SSE = \sum_{i=1}^n e_i^2$$

### Fitting the regression equation (i.e. estimating parameters)

The regression equation is “fitted” by choosing the values of  $b_0$  and  $b_1$  in such a way that the sum of squares of the prediction errors,  $D$ , are minimized, i.e.

$$\begin{aligned} \text{Min } D &= \sum_{i=1}^n e_i^2 \\ &= \sum_{i=1}^n (Y_i - b_0 + b_1 X_i)^2 \end{aligned}$$

The specific values of  $b_0$  and  $b_1$  that minimize  $D$  could be found iteratively, or by trial and error, but it is known that the following “ordinary least-squares” (OLS) estimates of  $b_0$  and  $b_1$  do in fact minimize  $D$ :

$$\begin{aligned} b_1 &= \frac{S_{XY}}{S_X} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}, \text{ and} \\ b_0 &= \bar{Y} - b_1 \bar{X}. \end{aligned}$$

### Goodness-of-fit statistics

The “goodness of fit” of the regression equation, or a measure of the strength of the relationship between  $Y$  and  $X$  can be described in several ways. As in analysis of variance, the sum of squares of the dependent variable  $Y$  can be decomposed into two components

$$\begin{aligned} \text{TotalSS} &= \text{RegrSS} + \text{ErrorSS} \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, \end{aligned}$$

Where *TotalSS* is the “total sum of squares” (of deviations of individual dependent variable values of the mean), *RegrSS* is the “regression sum of squares” or that component of the total sum of squares “explained” by the regression equation, and *ErrorSS* is the “residual sum of squares,” or the sum of squares of the residual,

$$\sum_{i=1}^n e_i^2$$

An *F*-statistic that can be used to test the null hypothesis that the relationship between the predictor and response variables is *not* significant is

$$F = \frac{MS_{Regr}}{MS_{Error}}$$

$$= \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (k)}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n - k - 1)}$$

The denominator of this expression,  $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (n - k - 1)$ , is also known as the “mean-square error of the regression,” and is sometimes represented by  $s^2$ . The square root of the mean-square error,  $s$ , is called the “standard error of the regression,” and provides a measure of uncertainty in the estimates of  $Y$  produced by the regression equation. In general, the larger the  $F$ , the stronger the relationship.

Another measure of the strength of the relationship between the response and predictor variable is the “explained variance” (a proportion, but sometimes expressed as a percentage), also known as the “coefficient of determination,” or  $R^2$

$$R^2 = \frac{RegrSS}{TotalSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

$$= 1 - \frac{ErrorSS}{TotalSS} = \frac{\sum_{i=1}^n (\hat{Y}_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

### Significance of the regression coefficients

There are a number of other quantities that are useful in interpreting a regression equation. These include standard errors for the slope and intercept

$$se(b_0) = s \left( \frac{1}{n} + \frac{\bar{X}^2}{S_X} \right)^2, \text{ and}$$

$$se(b_1) = s / S_X$$

Using these standard errors,  $t$ -statistics that can be used to test hypotheses about the regression coefficients can be constructed:

$$t(b_0) = \frac{b_0 - b_0^*}{se(b_0)}, \text{ and}$$

$$t(b_1) = \frac{b_1 - b_1^*}{se(b_1)}$$

where  $b_0^*$  and  $b_1^*$  are hypothesized values of the regression coefficients, which are usually taken to be 0, so that large values of the  $t$ -statistics will signal that  $b_0$  and  $b_1$  values that are significant (i.e. not zero). The standard error or standard deviation of the predicted value of the response variable,  $\hat{Y}_*$ , given a particular value of the predicted variable,  $X_*$ , is

$$se(\hat{Y}_* | X_*) = s \left[ 1 + \frac{1}{n} + \frac{(X_* - \bar{X})^2}{S_X} \right]$$