Multiple Linear Regression(MLR) (Chapter 9)

Definition

A multiple regression model uses a linear surface, such as a plane or hyperplane, to approximate the relationship between a continuous response (target) variable, and a set of predictor variables. While the predictor variables are typically continuous, categorical predictor variables may be included as well, through the use of indicator (dummy) variables.Compared to simple linear regression, multiple regression models provide improved precision for estimation and prediction, analogous to the improved precision of regression estimates over univariate estimates.

The MLR model is given as:-

$y= B_0+ B_1 x_1+ B_2 x_2+.....+ B_p x_p+e$ **when the number of predictors is p.**

The Assumptions About The Error Term
1. Zero-Mean Assumption. The error term is a random variable, with mean or expected value equal to zero. In other words.
2. Constant Variance Assumption. The variance of the error term is constant
3. Independence Assumption. The values of the erroe term are independent.
4. Normality Assumption. The error term is a normally distributed random variable.

Usefulness of the model

However, the F-test considers the linear relationship between the target variable y and the set of predictors taken as a whole. The hypotheses for the F-test are given by the null hypothesis asserts that there is no linear relationship between the target variable y, and the set of predictors. Thus, the null hypothesis states that the coefficient for each predictor exactly equals zero,e F-statistic consists of a ratio of two means squares, the mean square regression (MSR) and the mean square error (MSE). A mean square represents a sum of squares divided by the degrees of freedom associated with that sum of squares statistic. As the sums of squares are always nonnegative, then so are the mean squares. Therefore, for the F-test, we shall reject the null hypothesis when the value of the test statistic F is significantly large.

Recall that adding a variable to the model will increase the value of the coefficient of determination , regardless of the usefulness of the variable. This is not a particularly attractive feature of this measure, because it may lead us to prefer models with marginally larger values for , simply because they have more variables, and not because the extra variables are useful. Therefore, in the interests of parsimony, we should find some way to penalize the measure for models that include predictors that are not useful. Fortunately, such a penalized form for does exist, and is known as the adjusted Coeffiecient of Determination. The formula for adjusted is as follows:

$R^2_{adj}=1-(1- R^2)*\{(n-1)/(n-p-1\}$ where n denotes the no.of examples and p denotes the no.of regressors.

Use case

For example, returning to the cereals data set, suppose we are interested in trying to estimate the value of the target variable, nutritional rating, but this time using two variables, sugars and fiber, rather than sugars alone.

Thus, the regression equation for this example is

Y=52.174-2.2436*sugar+2.8665*fiber

That is, the estimated nutritional rating equals 52.174 minus 2.2436 times the grams of sugar plus 2.8665 times the grams of fiber. Note that the coefficient for sugars is negative, indicating a negative relationship between sugars and rating, while the coefficient for fiber is positive, indicating a positive relationship. These results concur with the characteristics of the graphs in Figures 9.1 and 9.2. The straight lines shown in Figure 9.2 represent the value of the slope coefficients for each variable, −2.2436 for sugars and 2.8665 for fiber.

Interpretations

The interpretations of the slope coefficients and are slightly different than for the simple linear regression case. For example, to interpret the coefficient of sugar , we say that “the estimated decrease in nutritional rating for a unit increase in sugar content is 2.2436 points, when fiber content is held constant.” Similarly, we interpret the coefficient of fiber as follows: “the estimated increase in nutritional rating for a unit increase in fiber content is 2.8408 points, when sugar content is held constant.” In general, for a multiple regression with m predictor variables, we would interpret coefficient as follows: “the estimated change in the response variable for a unit increase in variable is , when all other predictor variables are held constant.” Recall that errors in prediction are measured by the residual(Y_i-y_i) .In simple linear regression, this residual represented the vertical distance between the actual data point and the regression line. In multiple regression, the residual is represented by the vertical distance between the data point and the regression plane or hyperplane.

Whenever a new predictor variable is added to the model, the value coefficient of determination always goes up. If the new variable is useful, the value will increase significantly; if the new variable is not useful, the value may barely increase at all.