CHAPTER 1

Evaluating the Fit of a Linear Regression Model

The fit of a linear regression model can be evaluated using several statistical metrics. Three common ones include the residuals, the coefficient of determination (R-squared or R^2), and the root mean squared error (RMSE).

1.1 Residuals

Residuals are the differences between the observed and predicted values. For an observation i, the residual e_i is calculated as

$$e_i = y_i - \hat{y}_i$$

where y_i is the observed value and $\hat{y}_i$ is the predicted value. By plotting these residuals against the predicted values, we can visually inspect the model's fit. Ideally, the residuals

should be randomly scattered around zero, and there should be no clear pattern in the residual plot.

1.2   R-Squared (R^2)

R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that can be predicted from the independent variables. It provides a measure of how well the model's predictions fit the data. R^2 is calculated as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where SS_{res} is the sum of squares of residuals and SS_{tot} is the total sum of squares. An R^2 value of 1 indicates a perfect fit, while an R^2 of 0 indicates that the model does not explain any of the variability of the response data around its mean.

1.3   Root Mean Squared Error (RMSE)

RMSE is a frequently used measure of the differences between the values predicted by a model and the values actually observed. It's the square root of the average of squared differences between prediction and actual observation. RMSE is calculated as

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

A lower RMSE indicates a better fit to the data.

It's important to note that these metrics should not be used in isolation to evaluate the model's fit. They should be used in combination along with the understanding of the underlying problem and domain knowledge.

Answers to Exercises

INDEX