Print

Verifying Assumptions Using Residual Plots

To verify the assumptions of linear regression, you can use the residual values from the regression analysis as your best estimates of the error terms. Remember that the residuals are the difference between each observed value of Y and its predicted value. Residuals are defined as follows: $r_i = Y_i - \widehat{Y}_i$, where $\widehat{Y}_i$ is the predicted value of the i-th value of the response variable.

To check for violations of equal variances, you use the residuals versus predicted values plot. You can also use this plot to check for violations of linearity and independence.

To further examine any violations of equal variances, you use the residuals versus the values of the independent variables. If you plot the residuals versus X-1, versus X-2, and so on, you can determine which predictor contributes to the violation of the assumption.

Let's go over some examples of residual plots to determine whether any of the linear model assumptions are being violated. Suppose you have plots for four models that are fit to four sets of data. Each plot has the residual values on the Y-axis and the predicted values on the X-axis. What you want is a random scatter of the residual values above and below the reference line at 0. This indicates that the model assumptions are valid. If there are patterns or trends in the residual values, the assumptions might not be valid and the models might have problems. When you check the plots, check for obvious violations by simply analyzing the shape of the scatter.

In the first plot, the residuals are randomly scattered around the reference line at 0, no patterns appear in the residuals, and the model form appears to be adequate. These features indicate that the assumptions of linearity, equal variances, and independence are all reasonable.

In the second plot, the residual values have a quadratic or curved shape, and therefore, the linearity assumption is violated. The model doesn't fit the data. To account for the curvature in the data, one possible solution is to add a quadratic term into the model as an additional predictor variable.

In the third plot, the residuals have a funnel shape. The variance of the residuals is not constant. From left to right, the variance increases, and therefore, the equal variance assumption is being violated. The response variable in the model might need some sort of transformation. The natural log and square root transformations are common remedies to the departure from equal variances in this scenario.

The residuals in the fourth plot have a cyclical shape. The observations are not independent, so the independence assumption is being violated. Cyclical patterns such as this can appear when the predictor variable is a measure of time. The residuals are autocorrelated, meaning correlated over time. More specifically, the residuals are more correlated with observations nearer in time than with observations farther away in time.

In addition to verifying assumptions, its also important to check for outliers. These observations are often data errors or reflect unusual circumstances, and they can heavily affect your regression results. You need to investigate outliers to see whether they result from data entry error or some other problem that you can correct.

Close