0.0.1 Key Definitions

Residual The difference between the predicted value (based on the regression equation) and the actual, observed value.

Outlier In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Leverage An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. Influence can be thought of as the product of leverage and outlierness.

Cook's distance A measure that combines the information of leverage and residual of the observation.

0.0.2 Leverage

In statistics, leverage is a term used in connection with regression analysis and, in particular, in analyses aimed at identifying those observations that are far away from corresponding average predictor values. Leverage points do not necessarily have a large effect on the outcome of fitting regression models.

Leverage points are those observations, if any, made at extreme or outlying values of the independent variables such that the lack of neighboring observations means that the fitted regression model will pass close to that particular observation.[1]

Modern computer packages for statistical analysis include, as part of their facilities for regression analysis, various quantitative measures for identifying influential observations: among these measures is partial leverage, a measure of how a variable contributes to the leverage of a datum.

0.0.3 Leverage in LME models

For the general mixed model, leverage can be defined through the projection matrix that results from a transformation of the model with the inverse of the Cholesky decomposition of , or through an oblique projector. The MIXED procedure follows the latter path in the computation of influence diagnostics. The leverage value reported for the th observation is the th diagonal entry of the matrix

which is the weight of the observation in contributing to its own predicted value, . While is idempotent, it is generally not symmetric and thus not a projection matrix in the narrow sense. The properties of these leverages are generalizations of the properties in models with diagonal variance-covariance matrices. For example, , and in a model with intercept and , the leverage values

are and . The lower bound for is achieved in an intercept-only model, and the upper bound is achieved in a saturated model. The trace of equals the rank of . If denotes the element in row , column of , then for a model containing only an intercept the diagonal elements of are

Because is a sum of elements in the th row of the inverse variance-covariance matrix, can be negative, even if the correlations among data points are nonnegative. In case of a saturated model with , .