1 Case Deletion

Case-deleted analysis is a popular method for evaluating the inuence of a subset of cases on inference.

Cite: CPJ develops case-deletion diagnostics for detecting influential observations in mixed linear models. Diagnostics for both fixed effects and variance components are proposed. Computational formulas are given that make the procedures feasible. The methods are illustrated using examples.

1.1 Case Deletion Diagnostic Statistics

For ordinary generalized linear models, regression diagnostic statistics developed by Williams (1987) are commonly used in statistical platforms such as SS. These diagnostics measure the influence of an individual observation on model fit, and generalize the one-step diagnostics developed by Pregibon (1981) for the logistic regression model for binary data. Preisser and Qaqish (1996) further generalize regression diagnostics to apply to models for correlated data fit by generalized estimating equations (GEEs), where the influence of entire clusters of correlated observations is measured.

1.2 Matrix Notation for Case deletion notation

For notational simplicity, $\boldsymbol{A}(i)$ denotes an $n \times m$ matrix $\boldsymbol{A}$ with the i-th row removed, a_i denotes the i-th row of $\boldsymbol{A}$, and a_{ij} denotes the $(i,j)-$th element of $\boldsymbol{A}$.

1.3 Partitioning Matrices

Without loss of generality, matrices can be partitioned as if the $i-$th omitted observation is the first row; i.e. $i = 1$.

1.4 CPJ's Three Propositions

1.4.1 Proposition 1

$$\boldsymbol{V}^{-1} = \begin{bmatrix} \nu^{ii} & \lambda_i' \\ \lambda_i & \Lambda_{[i]} \end{bmatrix}$$

$$\boldsymbol{V}_{[i]}^{-1} = \boldsymbol{\Lambda}_{[i]} - \frac{\lambda_i \lambda_i'}{\lambda_i}$$

1.5 Proposition 2

(i) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{X}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{X}$

(ii) $= (\boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y})^{-1}$

(iii) $\boldsymbol{X}_{[i]}^T \boldsymbol{V}_{[i]}^{-1} \boldsymbol{Y}_{[i]} = \boldsymbol{X}' \boldsymbol{V}^{-1} \boldsymbol{Y}$

1.6 Proposition 3

This proposition is similar to the formula for the one-step Newtown Raphson estimate of the logistic regression coefficients given by Pregibon (1981) and discussed in Cook Weisberg.

2 Measures of Influence

The impact of an observation on a regression fitting can be determined by the difference between the estimated regression coefficient of a model with all observations and the estimated coefficient when the particular observation is deleted. The measure DFBETA is the studentized value of this difference.

Influence arises at two stages of the LME model. Firstly when V is estimated by $\hat{V}$, and subsequent estimations of the fixed and random regression coefficients β and u, given $\hat{V}$.

2.1 Cook's Distance

Cook's distance or Cook's D is a commonly used estimate of the influence of a data point when performing least squares regression analysis.

In a practical ordinary least squares analysis, Cook's distance can be used in several ways: to indicate data points that are particularly worth checking for validity; to indicate regions of the design space where it would be good to be able to obtain more data points.

It is named after the American statistician R. Dennis Cook, who introduced the concept in 1977.

Cook's distance measures the effect of deleting a given observation. Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Points with a large Cook's distance are considered to merit closer examination in the analysis.

2.1.1 Computation

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j \ - \hat{Y}_{j(i)})^2}{p \ \mathrm{MSE}}.$$

The following are the algebraically equivalent expressions (in case of simple linear regression):

$$D_i = \frac{e_i^2}{p \ \mathrm{MSE}} \left[\frac{h_{ii}}{(1-h_{ii})^2} \right],$$

$$D_i = \frac{(\hat{\beta} - \hat{\beta}^{(-i)})^T (X^T X)(\hat{\beta} - \hat{\beta}^{(-i)})}{(1+p)s^2}.$$

In the above equations:

- $\hat{Y}_j$ is the prediction from the full regression model for observation j;
- $\hat{Y}_{j(i)}$ is the prediction for observation j from a refitted regression model in which observation i has been omitted;
- h_{ii} is the i-th diagonal element of the hat matrix

$$\mathbf{X} \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T;$$

- e_i is the crude residual (i.e., the difference between the observed value and the value fitted by the proposed model);
- MSE is the mean square error of the regression model;
- p is the number of fitted parameters in the model

2.1.2 DFBETA

The DFBETA statistic for measuring the influence of the th observation is defined as the one-step approximation to the difference in the MLE of the regression parameter vector and the MLE of the regression parameter vector without the th observation. This one-step approximation assumes a Fisher scoring step, and is given by

2.1.3 DFFITS

DFFITS is a statistical measured designed to a show how influential an observation is in a statistical model. It is closely related to the studentized residual.

$$DFFITS = \frac{\widehat{y_i} - \widehat{y_{i(k)}}}{s_{(k)}\sqrt{h_{ii}}}$$

2.1.4 PRESS

The prediction residual sum of squares (PRESS) is an value associated with this calculation. When fitting linear models, PRESS can be used as a criterion for model selection, with smaller values indicating better model fits.

$$PRESS = \sum (y - y^{(k)})^2 \tag{1}$$

- $e_{-Q} = y_Q - x_Q \hat{\beta}^{-Q}$
- $PRESS_{(U)} = y_i - x\hat{\beta}_{(U)}$

2.1.5 DFBETA

$$\begin{aligned} DFBETA_a &= \hat{\beta} - \hat{\beta}_{(a)} && (2) \\ &= B(Y - Y_{\bar{a}} && (3) \end{aligned}$$