Kernel Weighted Influence

N. Hens[1], M. Aerts[1], G. Molenberghs[1], H. Thijs[1], G. Verbeke[2]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium
[2] Biostatistical Centre, Katholieke Universiteit Leuven, Kapucijnenvoer 35, B-3000 Leuven, Belgium

Abstract: To asses the sensitivity for non-random dropout in a selection model framework, several methods were developed. None of them are without limitations. In this paper, a new method called kernel weighted influence is proposed. It uses several features of global and local influence approaches. Together with the use of nonparametric techniques, it provides a challenging new technique with a variety of options.

Keywords: Local Influence, Global Influence, Kernel Weights, Missing Data, Sensitivity Analysis.

1 Introduction

In a longitudinal study, each unit is measured on several occasions. It is not unusual for some sequences of measurements to terminate early for reasons outside the control of the investigator, any unit so affected is often called a dropout. Little and Rubin (1987) make important distinctions between different missing values processes. A dropout process is said to be completely random (MCAR) if the dropout is independent of both unobserved and observed data and random (MAR) if, conditional on the observed data, the dropout is independent of the unobserved measurements; otherwise the dropout process is termed non-random (MNAR) or non-ignorable.

To represent such a model, Diggle and Kenward (1994) proposed a selection model which consists of two parts: a measurement part and a missingness process part. Such a model, which tries to represent a non-random dropout mechanism, relies on strong and untestable assumptions. Not only the assumed distributional form can be misspecified but also the presence of influential observations can have a large impact. Two well known methods to investigate the influence of individual cases are global influence, based on case-deletion, and local influence, where the model is slightly perturbed to study the stability of the model, as is done by Lesaffre and Verbeke (1998). In Thijs et al (2000), Molenberghs et al (2001) and Verbeke et al (2001), the latter method was used to investigate the influence of non-random missingness as part of a sensitivity analysis in the selection modeling framework. In the next sections these methods will be introduced briefly and an extension will be proposed.

2 A Selection Model for Non-Random Dropout

Let us assume that for subject i, $i = 1, \cdots, N$, a sequence of responses Y_{ij} is measured at two occasions $j = 1, 2$. Let R_i be a missingness indicator and assume that y_{i1} is always observed. Then $r_i = 1$ if y_{i2} is missing and $r_i = 0$ if y_{i2} is observed. The measurement part of the model of Diggle and Kenward (1994), which is in fact a linear mixed model, is given by

$$Y_i = (Y_{i1}, Y_{i2}) \sim N(X_i\beta, V_i), \qquad i = 1, \ldots, N,$$

where β is a vector of fixed effects and $V_i = Z_i G Z_i^{'} + \Sigma_i$. The X_i and Z_i contain covariate values. The missingness process is described by

$$\text{logit}[Pr(R_i = 1|y_{i1}, y_{i2})] = \psi_0 + \psi_1 y_{i1} + \psi_2 y_{i2},$$

where $Pr(R_i = 1|y_{i1}, y_{i2})$ is the probability for the i^{th} subject to drop out. If ψ_2 differs from zero, the missingness process is non-random.

3 Influence Measures

3.1 Kernel Weights

Classical influence measures like the global influence and local influence approach are essentially based on cases. Global influence was based on the calculation of likelihood displacements when cases are left out one by one. In a local influence approach, one strategy is to examine the normal curvature in the direction of each observation. Our proposal is to extend these two approaches by looking in the neighborhood of the outcomes (y_{1i}, y_{2i}, r_i). Therefore we introduce the following weights. If $r_i = 0$,

$$w_i(y_{1j}, y_{2j}, r_j) = \begin{cases} \frac{K^2(0)}{\text{norm. denominator}_0} & r_j = 1 \\ \frac{K^2(0) - K(\frac{y_{1j}-y_{1i}}{h_1})K(\frac{y_{2j}-y_{2i}}{h_1})}{\text{norm. denominator}_0} & r_j = 0 \end{cases}$$

and if $r_i = 1$,

$$w_i(y_{1j}, y_{2j}, r_j) = \begin{cases} \frac{K^2(0) - K(\frac{y_{1j}-y_{1i}}{h_1})K(0)}{\text{norm. denominator}_1} & r_j = 1 \\ \frac{K^2(0)}{\text{norm. denominator}_1} & r_j = 0 \end{cases}$$

where K is a gaussian kernel function, h_1 and h_2 are two possibly different bandwidths and r_j is the missingness indicator for subject j. The motivation of the weights is as follows. If $r_i = 0$, (y_{1i}, y_{2i}) is a completer and all completers in the neighborhood get low weight. All other subjects get high weight, including the dropouts. If $r_i = 1$, y_{2i} is not observed and all dropouts in the neighborhood are given low weight, while all other subjects get high weight.

3.2 Kernel Weighted Global Influence

Let us introduce a weighted loglikelihood

$$l(\gamma; w_i) = \sum_{j=1}^{N} w_{ij} l_j(\gamma),$$

with $\gamma = (\theta, \psi)$, grouping the parameters of the measurement and dropout model and w_{ij} the j^{th} component of the vector w_i. The global influence measure CD_i compares the loglikelihood $l(\gamma) = \sum_{j=1}^{N} l_j(\gamma)$ with the weighted loglikelihood $l(\gamma; w_i)$ with $w_i = (1, \ldots, 1, 0, 1, \ldots, 1)$ where the 0 is located at the i^{th} entry. Thus CD_i is given by

$$CD_i = 2[l(\hat{\gamma}) - l(\hat{\gamma}_{(-i)}; w_i)]. \tag{1}$$

To explore a neighbourhood of the outcome (y_{1j}, y_{2j}, r_j), one can look at a vector w_i where the j^{th} component obtains weight $w_{ij} = w_i(y_{1j}, y_{2j}, r_j)$ as introduced in Section 3.1. This extension of the well known global influence approach is able to allocate groups of influential cases with similar outcomes, thus avoiding the problem of masking.

3.3 Kernel Weighted Local Influence

The principle is to investigate how the results of an analysis are changed under infinitesimal perturbations of the model. Analytically this method looks locally on the MNAR parameter in the dropout model. Let us denote this MNAR parameter by ω. The MAR assumption corresponds to the case where ω equals the null vector, denoted by ω_0. The likelihood displacement then considered is given by

$$LD(\omega) = 2[l(\hat{\gamma}|\omega_0) - l(\hat{\gamma}_\omega|\omega_0)]$$

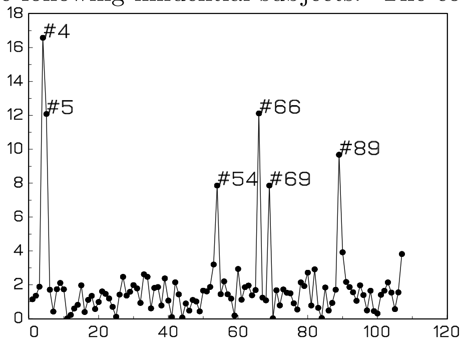
with

$$l(\gamma|\omega) = \sum_{i=1}^{N} l_i(\gamma|\omega_i).$$

This takes into account the variability of $\widehat{\gamma}$. Indeed, $LD(\omega)$ will be large if $\ell(\gamma|\omega_0)$ is strongly curved at $\widehat{\gamma}$, which means that γ is estimated with high precision, and small otherwise. Cook (1986) proposed to look at local influence, i.e., at the normal curvatures C_h in the direction of some N dimensional vector h of unit length. One evident choice is the vector h_i containing 1 in the i^{th} position and 0 elsewhere, corresponding to the perturbation of the i^{th} weight only. This reflects the influence of allowing the i^{th} subject to drop out non-randomly, while the others can only drop out at random. The local influence approach can be extended by looking at h_i where the j^{th} component equals $1 - w_i(y_{1j}, y_{2j}, r_j)$ with w_i as defined in Section 3.1. This method is more general and can provide new insights in the method of local influence.

4 The Mastitis Data

In the mastitis dataset the reduction in milk yield from cows who suffer from an infectious disease of the udder is registered. At two different time-points 107 cows were followed. There were 20 dropouts. A kernel weighted global influence analysis with bandwidths $h_1 = h_2 = 0.2$ on this data leads to the following influential subjects. The cows corresponding to numbers 4, 5, 54, 66, 69 and 89 seem to have a large influence. Subjects 54 and 69 were not found with the classical global influence, due to masking.



References

Cook, R.D. (1986) Assessment of local influence. *Journal of the Royal Statistical Society, Series B*, **48**, 133–169.

Diggle, P.J. and Kenward, M.G. (1994) Informative dropout in longitudinal data analysis (with discussion). *Applied Statistics*, **43**, 49–93.

Lesaffre, E. and Verbeke, G. (1998) Local influence in linear mixed models. *Biometrics*, **54**, 570-582.

Little, R.J.A. & Rubin, D.B. (1987) Statistical Analysis with Missing Data. *New York: Wiley.*

Molenberghs, G., Verbeke, G., Thijs, T., Lesaffre, E. and Kenward, M.G. (2001) Influence analysis to assess sensitivity of the dropout process. *Computational Statistics and Data Analysis*, **37**, 93–113.

Thijs, H., Molenberghs, G. and Verbeke, G. (2000) The Milk Protein Trial: Influence analysis of the Dropout Process. *Biometrical Journal*, **42**, 1–30.

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M.G. (2001) Sensitivity Analysis for Non-Random Dropout: A Local Influence Approach. *Biometrics*, **57**, 7–14.