Iranian Journal of Science & Technology, Transaction A, Vol. 29, No. A2
Printed in The Islamic Republic of Iran, 2005

AN INFLUENCE APPROACH FOR SENSITIVITY ANALYSIS OF NON-RANDOM DROPOUT BASED ON THE COVARIANCE STRUCTURE*

M. GANJALI** AND M. REZAEI

Dept. of Statistics, Faculty of Mathematical Sciences,
Shahid Beheshti University, Tehran, I. R. of Iran
m-ganjali@sbu.ac.ir

Abstract – A generalized Heckman model is used for the joint modeling of longitudinal continuous responses and dropout in order to see the influence of a small perturbation of the elements of the covariance structure on displacement of the likelihood. The perturbation from random dropout in the direction of informative dropout is considered for Mastitis data.

Keywords – Longitudinal data, Informative dropout, Sensitivity analysis, Generalized Heckman model

1. INTRODUCTION

Recently joint modeling of response and non-response in cross-sectional and longitudinal data has been extensively used [1]. Examples of such models for cross sectional data with selection and longitudinal data with dropout are the selection model of Heckman [2] (hereafter, SMH) and the dropout model of Diggle and Kenward [3] (hereafter DK). DK [3] use a selection model where they decompose the joint distribution of response and dropout into a marginal distribution for longitudinal continuous responses and a conditional distribution of dropout given previous and current responses. In the DK [3] model, based on the observed likelihood function, dropout is completely random (CRD) if dropout neither depends on the previous response nor on the current response. Dropout is at random (RD) if, given the previous response, dropout is not dependent on the current response, and dropout is informative (ID), and so nonignorable (NID), if dropout depends on the current response.

However, the DK [3] model rests on strong assumptions (see the review paper of Little [4] and discussion of DK [3]). Thus it has been suggested that when using joint modeling, a sensitivity analysis should be performed [5-8]. Assessment of the influence of a small perturbation model component which links the two models of response and dropout is an important consideration. To this end, several tools have been discussed in the literature, such as the informal sensitivity analysis of Kenward [9] and a formal local influence based approach [7]. Molenberghs et.al [10] use DK's model and the approach of Cook [11] for measuring the influence of a small perturbation of the model components. This involves studying the curvature of the likelihood displacement resulting from the perturbation.

In this paper we start with the SMH [2] and study the influence of perturbation of the model components (which link the response model and dropout model) to likelihood displacement. As there

*Received by the editor November 5, 2003 and in final revised form April 30, 2005
**Corresponding author

is just one parameter in the link between dropout and response mechanisms, this approach will help us to provide a picture of the global influence.

Furthermore, we shall use the generalized Heckman model (GHM) presented by Crouchley and Ganjali [12] and the local influence of Cook [11] for measuring the influence of small perturbations of the model components for longitudinal data with dropout.

In the next section, the SMH and its generalization will be reviewed. In Section 3, likelihood displacement and global and local influence will be discussed and the approach will be explained for measuring the influence of a small perturbation of the covariance structure of the Heckman and generalized Heckman models. In Section 4, we use the Mastitis data and consider the perturbation from RD in the direction of ID. In Section 5 conclusions are given.

2. SELECTION MODEL AND ITS GENERALIZATION

a) Selection model

Heckman [2] proposed a joint model for a continuous response (y_i) and a sample selection mechanism. Sample selection is the complement to missing data and dropout. The Heckman [2] model is defined by the means of two equations,

$$R_i^* = \alpha^T \mathbf{W}_i + v_i \tag{1}$$

$$y_i^* = \beta^T \mathbf{X}_i + \varepsilon_i \tag{2}$$

where α and β are vectors of parameters, $\mathbf{W}_i$ and $\mathbf{X}_i$ are vectors of covariates ($\mathbf{W}_i$ usually contains covariates not present in $\mathbf{X}_i$), (v_i, ε_i) are i.i.d drawings from a bivariate normal distribution with zero means, variances $\sigma^2_{RR} = 1$, $\sigma^2_{Y_0Y_0}$ and covariance σ_{RY_0}. It is assumed that only the sign of R_i^* is observed and that y_i^* is observed only when $R_i^* > 0$. Define

$$y_i = y_i^* \text{ if } R_i^* > 0$$
$$y_i = 0 \text{ if } R_i^* \le 0,$$

for $i = 1,...,n$. And also define

$$R_i = 1 \text{ if } R_i^* > 0$$
$$R_i = 0 \text{ if } R_i^* \le 0$$

so that (y_i, R_i) constitute the observations for subject i. A joint analysis of (y_i, R_i) is required when $\sigma_{RY} \neq 0$. Heckman [2] provides a two step estimator for this model.

b) Generalized selection model

The Heckman [2] model has been generalized by Crouchley and Ganjali [12] to the situation of repeated responses with dropout. This is

$$R_{it}^* = \alpha_t^T \mathbf{W}_{it} + v_{it} \tag{3}$$

$$y_{it}^* = \beta_t^T \mathbf{X}_{it} + \varepsilon_{it} \tag{4}$$

where t=1,...,T. Now

$$\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT})$$
$$\mathbf{R}_i = (R_{i2}, ..., R_{iT}).$$

where $y_{i1} = y_{i1}^*$ for i=1,...,n,

$$y_{it} = y_{it}^* \text{ if } R_{it}^* > 0$$
$$y_{it} = 0 \text{ if } R_{it}^* \le 0$$

for i=1,...,n and t=2,...,T and also

$$R_{it} = 1 \text{ if } R_{it}^* > 0$$
$$R_{it} = 0 \text{ if } R_{it}^* \le 0.$$

It is assumed that all the subjects at the start of the study are observed, i.e. $R_{i1} = 1, \ \forall i$. The observations for subject i take the form

$$(\mathbf{y}_i, \mathbf{R}_i) = (\left[y_{i1}^*, ..., y_{it-1}^*, 0, ..., 0 \right], [1, ..., 1, 0, ..., 0]),$$

if dropout occurs at time t and

$$(\mathbf{y}_i, \mathbf{R}_i) = (\left[y_{i1}^*, ..., y_{iT}^* \right], [1, ..., 1])$$

if responses of a subject are completely observed. Dropout is a monotone missing data pattern, i.e. $\Pr(R_{it} = 0 \mid R_{it-1} = 0) = 1$ and $\Pr(R_{it} = 1 \mid R_{it-1} = 0) = 0$, i.e. a subject cannot re-appear once they have dropped out.

For the vector of errors ε_i, let $\text{Var}(\varepsilon_i) = \Sigma_{YY}$. Let Σ_{YY} be unstructured so that $\text{Var}(\varepsilon_{it}) = \sigma^2_{YY_t}$ and $\text{cov}(\varepsilon_{is}, \varepsilon_{it}) = \sigma_{YYs,t}$. It is also assumed that the subjects are independent of each other so that $\text{cov}(\varepsilon_{is}, \varepsilon_{i't}) = 0$ for i ≠ i' for all s and t. Write $\text{Var}(v_i) = \Sigma_{RR}$, where $\text{diag}(\Sigma_{RR}) = 1$. The off diagonal elements of Σ_{RR} are unstructured so that $\Sigma_{RR_{s,t}} = \text{cov}(v_{is}, v_{it}) = \sigma_{RR_{s,t}}$. Subscript RR is used to indicate the non-response sub matrix of the joint response and non-response variance-covariance structure for the stochastic errors. Let $\Sigma_{YR} = \left[\text{cov}(\varepsilon_{is}, v_{it})\right]$. In this GHM both Σ_{YR} and the off diagonal elements of Σ_{RR} are unstructured.

c) The dropout mechanism

When dropout occurs y_{it}^* is not observed. We need a slightly different notation if we want to study the relationship between the unobserved response y_{it}^* and the dropout mechanism. Rubin [13] and Little and Rubin [1] note that for CRD the dropout process must be independent of both the observed responses $\mathbf{y}_{io}^* = (y_{i1}^*, ..., y_{it-1}^*)$ and y_{it}^*, while for RD the dropout process, conditional on $\mathbf{y}_{io}^*$, must be independent of y_{it}^*.

If we let f(.) denote a multivariate normal distribution then

$$\begin{aligned} f(\mathbf{y}_i^*, \mathbf{R}_i^*) &= f(y_{it}^*, \mathbf{R}_i^* \mid \mathbf{y}_{io}^*) f(\mathbf{y}_{io}^*) \\ &= f(y_{it}^* \mid \mathbf{R}_i^*, \mathbf{y}_{io}^*) f(\mathbf{R}_i^* \mid \mathbf{y}_{io}^*) f(\mathbf{y}_{io}^*) \end{aligned} \tag{5}$$

We have CRD if $f(y_{it}^* \mid \mathbf{R}_i^*, \mathbf{y}_{io}^*) = f(y_{it}^*)$ and $f(\mathbf{R}_i^* \mid \mathbf{y}_{io}^*) = f(\mathbf{R}_i^*)$. We have RD if $f(y_{it}^* \mid \mathbf{R}_i^*, \mathbf{y}_{io}^*) = f(y_{it}^* \mid \mathbf{y}_{io}^*)$. With either CRD or RD the joint probability of $(\mathbf{y}_i^*, \mathbf{R}_i^*)$ factors so that we can use $f(\mathbf{y}_{io})$ on its own for unbiased inference about β. If $f(y_{it}^* \mid \mathbf{R}_i^*, \mathbf{y}_{io}^*)$ does not simplify for CRD or RD we have NID.

Crouchley and Ganjali [12] denote the variance-covariance matrix Σ_{GH} for the elements of $(\mathbf{y}^*_{io}, y^*_{it}, \mathbf{R}^*_i)$, i.e.

$$\Sigma_{GH} = \begin{bmatrix} \Sigma_{Y_0Y_0} & \Sigma_{Y_0Y_t} & \Sigma_{Y_0R} \\ \Sigma_{Y_tY_0} & \sigma^2{}_{YY_t} & \Sigma_{Y_tR} \\ \Sigma_{RY_0} & \Sigma_{RY_t} & \Sigma_{RR} \end{bmatrix}$$

and they found that if both $\Sigma_{Y_tR} = \mathbf{0}$ (missing at random) and $\Sigma_{Y_0R} = \mathbf{0}$ (observed at random) we have CRD, i.e. $\Sigma_{Y_tR|Y_o} = \mathbf{0}$ and

$$f(y^*_{it}, \mathbf{R}^*_i \mid \mathbf{y}^*_{io}) = f(y^*_{it} \mid \mathbf{y}^*_{io}).f(\mathbf{R}^*_i)$$

Therefore we can estimate a model under CRD by imposing constraints $\Sigma_{Y_tR} = \mathbf{0}$ and $\Sigma_{Y_oR} = \mathbf{0}$. It can be seen that if

$$\Sigma_{Y_tR} - \Sigma_{Y_tY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR} = \mathbf{0} \tag{6}$$

We have RD and

$$f(y^*_{it}, \mathbf{R}^*_i \mid \mathbf{y}^*_{io}) = f(y^*_{it} \mid \mathbf{y}^*_{o}).f(\mathbf{R}^*_i \mid \mathbf{y}^*_{o}).$$

So we can estimate a model under RD by imposing the constraint $\Sigma_{Y_tR} = \Sigma_{Y_tY_o}\Sigma^{-1}_{Y_oY_o}\Sigma_{Y_oR}$. Note that neither CRD nor RD impose any constraints on the off diagonal elements of Σ_{RR}. ID or NID occurs when neither of the CRD or RD conditions applies. Consider as an example the case of two period longitudinal data where the response at the first time is observed for all individuals. In this case

$$\mathbf{y}_i = (y_{i1}, y_{i2})$$
$$R_i = R_{i2}$$

and let, for simplicity, Σ_{GH} be

$$\Sigma_{GH} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{bmatrix},$$

where $\text{cov}(y_{i1}, y_{i2}) = \sigma_{12}$ and $\text{cov}(y_{ij}, R_{i2}) = \sigma_{j3}$ for j=1,2. Consequently, for $\sigma_{22} > 0$, $\rho_{23} - \rho_{12}\rho_{13} = 0$ gives the conditions for ignorable dropout where $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$ and for j=1,2, $\rho_{j3} = \dfrac{\sigma_{j3}}{\sqrt{\sigma_{jj}}}$. This can occur in the following ways:

1. $\rho_{23} = \rho_{13} = 0$, which is completely random dropout (CRD).
2. $\rho_{23} = 0$ and $\rho_{12} = 0$, where the latter implies there is no correlation between the two responses.
3. $\rho_{23} = \rho_{12}\rho_{13}$.

In the next section we shall examine the likelihood displacement as a measure of sensitivity for perturbations of the CRD and RD model.

3. LIKELIHOOD DISPLACEMENT AND LOCAL INFLUENCE

We are interested in the influence that selection exerts on the parameters of interest in the GHM. If $\Sigma_{Y_tR} - \Sigma_{Y_tY_o}\Sigma_{Y_oY_o}^{-1}\Sigma_{Y_oR} = \mathbf{0}$, we have an RD process and in this case, measurement model parameters can not be influenced by selection. Modification of $\mathbf{H} = \Sigma_{Y_tR} - \Sigma_{Y_tY_o}\Sigma_{Y_oY_o}^{-1}\Sigma_{Y_oR}$ may lead to large differences in the model parameters. Denote the log-likelihood function corresponding to GHM by

$$l(\gamma \mid \mathbf{H}) = \sum_{i=1}^{n} l_i(\gamma \mid \mathbf{H}) \tag{7}$$

in which $l_i(\gamma \mid \mathbf{H})$ is the contribution of the i-th individual to the log-likelihood and $\gamma = (\alpha^T, \beta^T)$ is the parameter vector for the measurement and dropout mechanisms. Let $l(\gamma) = l(\gamma \mid \mathbf{H} = \mathbf{0})$. Here $l(\gamma)$ is the log-likelihood function which corresponds to a RD model. Suppose $\mathbf{H}$ can be perturbed around $\mathbf{0}$. Let $\hat{\gamma}$ be MLE for γ obtained by maximizing $l(\gamma \mid \mathbf{H} = \mathbf{0})$ and let $\hat{\gamma}_H$ denote the MLE for γ under $l(\gamma \mid H)$. Now one can compare $\hat{\gamma}_H$ and $\hat{\gamma}$ as local influence. If $\hat{\gamma}_H$ and $\hat{\gamma}$ are similar, parameter estimates are robust to the perturbation of RD in the direction of ID. Strongly different estimates show that the estimation procedure is highly sensitive to such modification. We can quantify the differences using Cook's likelihood displacement defined as

$$LD(\mathbf{H}) = 2\left[l(\hat{\gamma}) - l(\hat{\gamma}_H)\right]. \tag{8}$$

$LD(\mathbf{H})$ will be large if $l(\gamma \mid \mathbf{H} = \mathbf{0})$ is strongly curved at $\hat{\gamma}$ which means that γ is estimated with high precision, and small otherwise. A graph of $LD(\mathbf{H})$ versus $\mathbf{H}$ can be used to assess the influence of perturbations. As it is shown in the following paragraph, for cross sectional and longitudinal data with two periods, LD(H) can be plotted against H (which is a scalar in these cases) and it gives a global sensitivity analysis. For longitudinal data with more periods, the local influence of Cook [11] can be used in the same way as Molenberghs et al. [10] and Jansen et al. [14] to find more influential elements of $\mathbf{H}$.

For example, if we have cross-sectional data, the selection model of Heckman gives the following likelihood displacement

$$LD(\rho) = 2\left[l(\hat{\gamma}) - l(\hat{\gamma}_\rho)\right]. \tag{9}$$

where ρ (with values in [-1,1]) is the correlation between v_i and ε_i in system (1 & 2). In this case LD(ρ) can be plotted against ρ. With two-period longitudinal data with dropout, the likelihood displacement for $H = \rho_{23} - \rho_{12}\rho_{13}$ can be obtained by equation (8) and in this case LD(H) can be plotted against H.

4. MASTITIS DATA: MODEL AND RESULTS

Mastitis is the occurrence of the infectious diseases of the udder and can reduce the milk yield of infected animals. We shall use data of the total milk yield for 107 cows from a single herd, in two consecutive years, to investigate the relationship between yield and mastitis. Of 107 animals, 27 were infected in their second year which will be treated as missing.

Table 1 shows the number of the observations and the sample mean of the milk yield of non-infected animals in the j th year for $j = 1,2,...,5$ and for both responses (Y_1 and Y_2).

Table 1. Descriptive summary of the data

Selected year	1	2	3	4	5
No. of cows for year 1	9	27	25	23	23
Sample mean of Y_1	5.875	5.568	6.007	5.915	5.541
No. of cows for year 2	6	19	19	15	21
Sample mean of Y_2	6.064	6.563	6.319	6.579	6.460

As Table 1 suggests, there may be no significant effect of the selected year on the mean of the responses, but there may be the effect of time on the responses (neglecting the selected year, sample mean of Y_1 =5.765 and sample mean of Y_2 =6.444). However, as dropout may be informative, no final conclusion can be reached before a joint and sensitivity analyses are done. For these data the GHM is in the form

$$y^*_{i1} = \beta_0 + \varepsilon_{i1}, \tag{10}$$

$$y^*_{i2} = \beta_0 + \eta + \varepsilon_{i2}, \tag{11}$$

$$R^*_{i2} = \alpha_0 + v_{i3} \tag{12}$$

where η gives the effect of time on the mean of the response. We omit the effect of explanatory variable, selected year, as a previous analysis [12] showed no significant effect of this variable on response. We used NAG [15] routine E04UCF to obtain the likelihood displacements for these data.

Results from the GHM, System (10-12) for ID, RD, and CRD models are presented in Table 2.

Table 2: Results for mastitis data (^{I} : Informative dropout model, ^{II} : Random dropout model, ^{III} : Complete random dropout model and ^{IV} : Informative dropout model for data without outliers)

	ID model ^{I}		RD model ^{II}		CRD model ^{III}		IDWO model ^{IV}	
Par	Est	Se	Est	Se	Est	Se	Est	Se
β_0	5.765	0.090	5.765	0.090	5.765	0.090	5.798	0.086
η	0.315	0.138	0.719	0.107	0.719	0.107	0.617	0.434
ρ_{12}	0.470	0.087	0.581	0.071	0.581	0.071	0.727	0.054
ρ_{13}	-0.157	0.125	-0.149	0.013	-	-	-0.127	0.131
ρ_{23}	0.676	0.117	-	-	-	-	-0.127	0.934
σ_{11}	0.931	0.064	0.931	0.064	0.931	0.064	0.872	0.060
σ_{22}	1.274	0.113	1.138	0.088	1.138	0.087	1.044	0.100
α_0	0.634	0.130	0.667	0.131	0.667	0.132	0.645	0.133
-logL	308.771		311.389		312.013		275.998	

We get an increase in deviance of 6.484 for 2 d.f. (p=0.039) for a test of CRD ($\rho_{13} = \rho_{23} = 0$) in System (10-12) and an increase in deviance of 5.236 for 1 d.f. (p=0.022) for a test of RD ($\rho_{23} = \rho_{12}\rho_{13}$) in System (10-12). Table 2 shows that for the ID model, dropout is informative because of the stochastic dependency (ρ_{23} =0.676) between the dropout process and the response in the second period. The value of ρ_{23} implies that a large value of the response in the second period (which may be missing) will increase the probability of being present in that same second period. All

the models give a significant change in mean response in the second period, but the CRD and RD model overestimate it.

Using Pearson residuals Crouchley and Ganjali [12] found 3 outliers in responses (cows 4, 5, 66). Deleting these observations shows no sign of ID (see results of IDWO in Table 1). Figure 1 shows the LD against different values of H for full data and data without outliers.

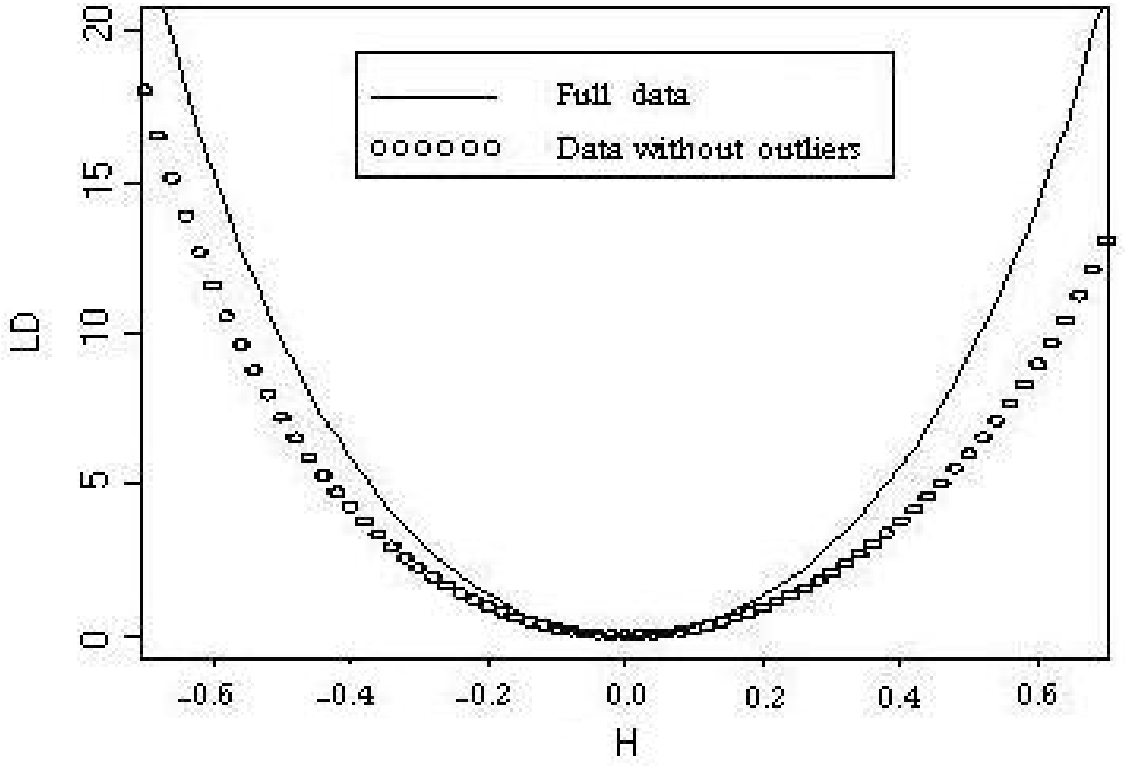

Fig. 1. Likelihood displacement against values of H

Figure 1 shows that there is no major difference between LD for full data and LD for data without outliers. This suggests that only some outliers are the cause of ID in these data.

5. CONCLUSIONS

We have presented an approach for assessing the influence of the modification of covariance structure of GHM from RD in the direction of ID. For cross-sectional and two-period longitudinal data (when the first response is fully observed), this approach gives a global sensitivity analysis. For longitudinal data with more periods, normal curvature can be used to find more influential elements of covariance structure.

Acknowledgments- The authors are grateful to Prof. M. R. Meshkani for his comments on an earlier draft and the referees for useful comments which led to a much improved paper.

REFERENCES

1. Little, R. J. A. & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York, Wiley.
2. Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*, 153-161.

3. Diggle, P. J. & Kenward, M. G. (1994). Informative Drop-out in longitudinal data analysis. *Appl. Statist.*, *43*, 49-93.
4. Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated measures studies. *J. Amer. Statist. Assoc.*, *90*, 1112-1121.
5. Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc.*, *94*, 10961146.
6. Verbeke, G. & Molenberghs, G. (1997) *Linear Mixed Models in Practice: A SAS-Oriented Approach*. Lecture Notes in Statistics 126. New York, Springer-Verlag.
7. Verbeke, G. & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York, Springer-Verlag..
8. Molenberghs, G., Goetghebeur, E. J. T., Lipsitz, S. R. & Kenward, M. G. (1999). Non-random missingness in categorical data: strengths and limitations. *Amer. Statist.*, *53*, 110-118.
9. Kenward, M. G. (1998). Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Statist. Medic., 17*, 2723-2732.
10. Molenberghs, G., Thijs, H., Lesaffre, E. & Kenward, M. G. (2001). Influence analysis to assess sensitivity of the dropout process, *Com. Statist. Data Anal.*, *37*, 93-113.
11. Cook, R. D. (1986). Assessment of local influence. *J. Roy. Statist. Soc. Ser. B*, *48*, 133-169.
12. Crouchley, R. & Ganjali, M. (2002). The common structure of several models for nonignorable dropout. *Statistical modelling*, *2*, 39-62.
13. Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.
14. Jansen, I., Molenberghs, G., Aerts, M., Thijs, H. & Steen, K. V. (2003). A local influence approach applied to binary data from a Psychiatric study. *Biometrics*, *59*, 410-419.
15. NAG (1996). *Numerical Algorithms Group Manual*. Oxford. U. K., Mark 16.