Dependence and sensitivity in regression models for longitudinal responses subject to dropout

Dipendenza e sensibilità nei modelli ad effetti casuali per dati longitudinali in presenza di dropout

Marco Alfò and Maria Francesca Marino

Abstract In longitudinal studies, subjects may be lost to follow-up and present incomplete response sequences. When the mechanism that leads to exit the study is non ignorable, a possible route is to define a model that accounts for potential dependence between the longitudinal and the dropout process. This model should have, at least, two major features: (*i*) it should (simply) reduce to an ignorable missing data model, when some conditions are met; (*ii*) the nested structure should give the way to measure sensitivity of parameter estimates to assumptions on non ignorability. In this work, we discuss random coefficient based dropout models and review measures of local sensitivity.

Abstract Negli studi longitudinali, alcuni soggetti abbandonano lo studio prima del suo completamento, presentando sequenze incomplete. Quando il meccanismo di generazione del dato mancante è non ignorabile, si può considerare un modello che descriva la dipendenza tra processo longitudinale e generazione del dato mancante stesso. Tale modello dovrebbe includere, come caso particolare, il modello per dati mancanti di tipo ignorabile, e permettere un'analisi di sensibilità delle stime rispetto alle ipotesi fatte circa il meccanismo di generazione dei dati mancanti stessi. In questo lavoro, si discutono i modelli a coefficienti casuali per l'analisi di studi longitudinali con dati mancanti non ignorabili e si confrontano misure di sensibilità locale.

Marco Alfò
Dipartimento di Scienze Statistiche, Sapienza Università di Roma, e-mail: marco.alfo@uniroma1.it

Maria Francesca Marino
Dipartimento di Statistica, Informatica, Applicazioni, Università degli Studi di Firenze, e-mail: mariafrancesca.marino@unifi.it

1 Introduction

Longitudinal studies entail repeated measurements from the same units over time. Often, units leave the study before the planned end, leading to *dropout* (also referred to as *attrition*) According to Rubin's taxonomy [19], if the probability of a missing response, conditional on observed data, does not depend on the responses that should have been observed, the data are said to be missing completely at random (MCAR), or missing at random (MAR). When, even after conditioning on observed data, the mechanism still depends on the unobserved responses, data are referred to as missing not at random (MNAR). In the context of likelihood inference, when either the parameters in the measurement and the missingness process are not distinct or the missing data process is MNAR, missing data are non ignorable (NI). In this case, some form of joint modeling of the longitudinal response and the missing data process is required [12].

Random Coefficient Based Dropout Models (RCBDMs, [11]) may be used as a quite general approach in this context. Here, two separate (conditional) models are built for the longitudinal response and the missingness indicator. Dependence arises due to models sharing common/dependent unit- and (possibly) outcome-specific random parameters. The model specification is completed by adopting an appropriate distribution for these random parameters, which can be either fully parametric [23, 8], or semi-parametric [2, 3]. This latter approaches have been introduced in the literature to avoid the impact that parametric assumptions may have on inference [20], especially in the case of short longitudinal sequences. More elaborated approaches are also available in the literature [5, 6, 4]. Besides the advantages of the semi-parametric approach, it presents the substantial drawback that dependence *within* outcomes can not be separated by dependence *between* outcomes. Starting from this drawback, we define a bi-dimensional finite mixture model for longitudinal data subject to dropout [21]. When the missing data mechanism is ignorable, such MNAR model directly reduces to its MAR counterpart. See also [14] for a dynamic extension of the model. Sensitivity of parameter estimates to assumptions on non ignorability of the dropout process can be explored by adopting either a global or a local perspective. Within the latter, we discuss the so-called *index of sensitivity to non ignorability* (ISNI) proposed by [22] and [13]. We show that, if the proposed model specification is employed, this approach to sensitivity analysis can be seen as a particular version of local influence diagnostics [10, 17, 18]. Obviously, a *global* influence approach could be adopted as well, for example by looking at the *mean score* approach by [25].

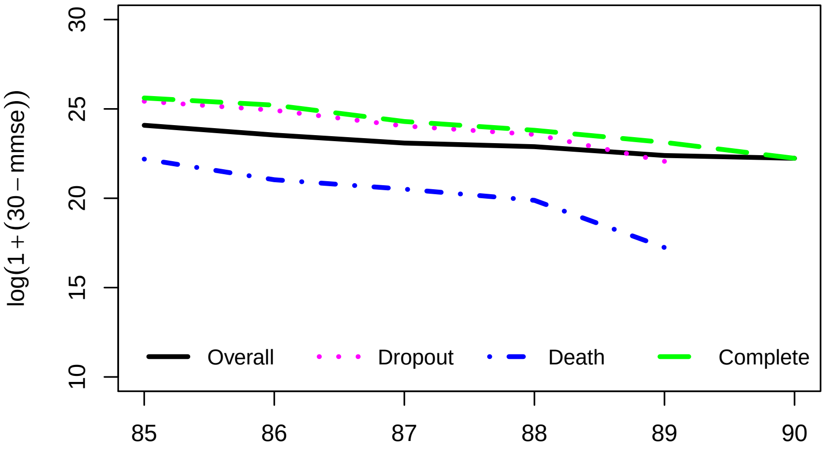
The structure of the paper follows. In section 2 we introduce the motivating application, the Leiden 85+ study, entailing the dynamics of cognitive functioning in the elderly. Section 3 discusses general random coefficient based dropout models, while sensitivity analysis is described in section 4.

2 Motivating example: Leiden 85+ data

To discuss our proposal, we consider data from the Leiden 85+ study, a retrospective study on 705 Leiden (Netherlands) inhabitants, who reached the age of 85 years between September 1997 and September 1999. The study aim was at identifying demographic and genetic determinants of cognitive functioning dynamics in the elderly. The following covariates were collected at the beginning of the study: gender, educational status (primary/higher education), plasma Apolipoprotein E (APOE) genotype (22-23, 24, 33, 34-44). Only 541 subjects present complete covariate information and will be considered in the following. Study participants were visited at their place of residence once a year until the age of 90; orientation, attention, language skills and ability to perform simple actions were assessed through a 30-items questionnaire. The Mini Mental State Examination index (MMSE, [7]), is obtained by summing the binary scores on such 30 items.

We report in Figure 1 the evolution of the mean response over time, stratified by participation.

Fig. 1: Mean *MMSE* value over time stratified by subjects' participation to the study.



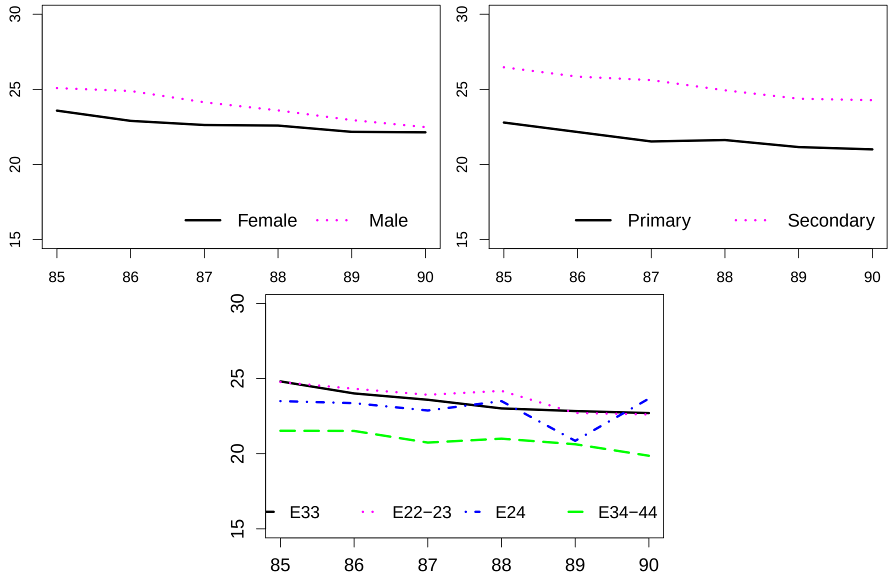
By looking at this figure, we may observe that, while the decline over time in the MMSE mean is (at least approximately) constant across the groups defined by patterns of dropouts, the differential participation in the study leads to a different slope for the overall mean score. Such a finding highlights a potential dependence between the evolution of the response over time and the dropout process, which may bias parameter estimates and corresponding inference. We report in Table 1 the distribution of the observed covariates by pattern of participation. This suggests a differential participation in the study by gender and educational level, while differences can be observed only for $APOE_{34-44}$ group.

Table 1: Leiden 85+ Study: demographic and genetic characteristics of participants

Variable	Total	Completed (%)	Did not complete (%)	(Row) Total
Gender				
Male	180 (33.27)	74 (41.11)	106 (58.89)	100
Female	361 (66.73)	192 (53.19)	169 (46.81)	100
Education				
Primary	351 (64.88)	166 (47.29)	185 (52.71)	100
Secondary	190 (35.12)	100 (52.63)	90 (47.37)	100
APO-E				
22-23	96 (17.74)	54 (56.25)	42 (43.75)	100
24	12 (2.22)	6 (50.00)	6 (50.00)	100
33	319 (58.96)	162 (50.78)	157 (49.22)	100
34-44	114 (21.08)	44 (38.60)	70 (61.40)	100
Total	541 (100)	266 (49.17)	275 (50.83)	100

Figure 2 depicts the dynamics of mean MMSE score over time by available covariates. From this figure, it is evident that cognitive impairment is lower for males than females, even if the difference seems to decrease with age, maybe due to a differential dropout by gender. No further interaction with age can be evinced, as the dynamics seem to be consistent for both levels of education, and for all the 4 levels of APOE genotype, but for the one with a very reduced sample size ($APOE_{24}$).

Fig. 2: Leiden 85+ Study: mean of MMSE score stratified by age and gender, educational level, APOE



3 Random coefficient-based dropout models

Let Y_{it} represent a response recorded on $i = 1,\dots,n$, subjects at time occasions $t = 1,\dots,T$, and let $\mathbf{x}_{it} = (x_{it1},\dots,x_{itp})'$ be a vector of observed covariates. We assume that, conditional on a q-dimensional set of individual-specific random coefficients $\mathbf{b}_i$, the observed responses are independent draws from a distribution in the Exponential Family with canonical parameter defined by

$$\theta_{it} = \eta_{it}^{Y} = \mathbf{x}_{it}'\beta + \mathbf{z}_{it}'\mathbf{b}_i.$$

The terms $\mathbf{b}_i$, $i = 1,\dots,n$, describe unobserved, individual-specific, heterogeneity (which may also be time-varying), while β is a vector of fixed parameters. Usually, $\mathbf{z}_{it} = (z_{it1},\dots,z_{itq})'$ represents a subset of $\mathbf{x}_{it}$. For identifiability purposes, standard assumptions on the random coefficient vector are introduced: $\mathrm{E}(\mathbf{b}_i) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{b}_i) = \mathbf{D}$ for $i = 1\dots,n$.

Let $\mathbf{R}_i$ denote the vector of missing data indicators, with generic element $R_{it} = 1$ if the i-th unit drops-out at any point in the window $(t-1,t)$, $R_{it} = 0$ else. As we focus on dropouts, we have $R_{it'} = 1, \forall t' > t$, so that $T_i \leq T$ measures are available for each study participant. We consider studies in discrete time; however, most of the following arguments may apply, with a limited number of changes, to (continuous time) survival process as well. To describe potential dependence between the longitudinal and the dropout process, we introduce an explicit model for the latter, conditional on a set of covariates, say $\mathbf{w}_i$, and a subset of the random coefficients in the longitudinal model. That is, we assume that, conditional on $\mathbf{b}_i^* = \mathbf{C}\mathbf{b}_i$, $i = 1,\dots,n$, $\mathbf{C} \in \mathscr{M}_{\mathbf{q},\mathbf{q_R}}$, dropout indicators are independent and follow a Bernoulli distribution with probability ϕ_{it} defined by:

$$\mathrm{logit}(\phi_{it}) = \eta_{it}^{R} = \mathbf{w}_{it}'\gamma + \mathbf{v}_{it}'\mathbf{b}_i^*. \tag{1}$$

Previous equations define a so-called shared (random) coefficient model [27, 26]. The assumption is that the longitudinal response and the dropout indicator are independent conditional on the individual-specific random coefficients:

$$f_{Y,R}(\mathbf{y}_i,\mathbf{r}_i \mid \mathbf{X}_i,\mathbf{W}_i) = \int \left[\prod_{t=1}^{T_i} f_Y(y_{it} \mid \mathbf{x}_{it},\mathbf{b}_i) \prod_{t=1}^{\min(T,T_i+1)} f_R(r_{it} \mid \mathbf{w}_{it},\mathbf{b}_i)\right] dG(\mathbf{b}_i), \tag{2}$$

Dependence between the measurement and the missigness, if any, is completely accounted for by the latent effects which are also used to describe unobserved, individual-specific, heterogeneity in each of the two (univariate) profiles. This class of models has been further extended by [28] to *joint* models where a continuous time setting and a survival data model are considered:

$$h_i(t) = h_0(t)\exp(\mathbf{w}_{it}'\gamma + \alpha\eta_{it}^{Y}). \tag{3}$$

As an alternative, we may consider equation-specific random coefficients [1]. In this respect, let $\mathbf{b}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2})$ denote an individual- and outcome-specific random coefficient. Introducing a local independence assumption, the joint density for the couple $(\mathbf{Y}_i, \mathbf{R}_i)$ can be written as follows:

$$f_{Y,R}(\mathbf{y}_i, \mathbf{r}_i \mid \mathbf{X}_i, \mathbf{W}_i) = \int \left[\prod_{t=1}^{T_i} f_Y(y_{it} | \mathbf{x}_{it}, \mathbf{b}_{i1}) \prod_{t=1}^{\min(T, T_i+1)} f_R(r_{it} | \mathbf{w}_{it}, \mathbf{b}_{i2}) \right] dG(\mathbf{b}_{i1}, \mathbf{b}_{i2}). \tag{4}$$

A further approach is that proposed by [6], where common, partially shared and independent (outcome-specific) random coefficients are considered in the measurement and the dropout process. For example, in the current context, we may write

$$\mathbf{b}_{i1} = \mathbf{b}_i + \varepsilon_{i1} \mathbf{b}_{i2} = \mathbf{b}_i + \varepsilon_{i2}, \qquad \varepsilon_{i1} \perp \varepsilon_{i2}.$$

This can be further extended to consider partially shared effects.

4 Sensitivity analysis: definition of the index

As highlighted by [15], for every MNAR model we may define a MAR counterpart that produces exactly the same fit to the observed data. Two issues are worth to be noticed. First, the MNAR model is fitted to the observed data only, assuming that the distribution of the missing responses is identical to that of the observed ones. Second, the structure describing dependence between the longitudinal responses (observed and missing) and the dropout indicators is just one out of several possible choices. Therefore, we may be interested in evaluating how much maximum likelihood estimates are influenced by hypotheses on the dropout mechanism.

Looking at *local* sensitivity, [22] defined the index of local sensitivity to non ignorability (ISNI) using a first-order Taylor expansion of the log-likelihood function. The aim was at describing the behaviour of parameter estimates in a neighbourhood of the MAR solution. The index was further extended by [9] by considering a second-order Taylor expansion; more general settings and different metrics were also considered [13, 31, 29, 30, 24].

To specify the index of local sensitivity, let $\lambda = (\lambda_{11}, \ldots, \lambda_{K_1 K_2})$ denote the vector of non ignorability parameters, with $\lambda = \mathbf{0}$ corresponding to the MAR model. Furthermore, let $\hat{\Phi}(\lambda)$ denote the ML estimates obtained conditional on a given value of λ. The ISNI may be written as

$$ISNI_{\Phi} = \left.\frac{\partial \hat{\Phi}(\lambda)}{\partial \lambda}\right|_{\Phi(0)} \simeq - \left(\left.\frac{\partial^2 \ell(\Phi, \Psi, \pi)}{\partial \Phi \Phi'}\right|_{\Phi(0)} \right)^{-1} \left.\frac{\partial^2 \ell(\Phi, \Psi, \pi)}{\partial \Phi \lambda}\right|_{\Phi(0)} \tag{5}$$

It measures the displacement of model parameter estimates from their MAR counterpart, in the direction of λ. Following [29], the following equation holds:

$$\hat{\Phi}(\lambda) = \hat{\Phi}(\mathbf{0}) + ISNI_{\Phi}\lambda;$$

The ISNI may be also interpreted as the linear impact that λ has on $\hat{\Phi}$. By using the proposed bi-dimensional model specification, we may show that the sensitivity analysis based on $ISNI_{\Phi}$ can be linked to local influence diagnostics developed for regression models to check for influential observations by perturbing individual-specific weights [10, 17, 18]. Here, we perturb weights associated to groups of subjects, rather than individual observations, See e.g. [16] for a comparison between multiple imputation and perturbation schemes in the more general setting of masking individual microdata.

References

1. Aitkin, M., Alfò, M.: Variance component models for longitudinal count data with baseline information: epilepsy data revisited. Statistics and Computing **16**, 231–238 (2006)
2. Alfò, M., Aitkin, M.: Random coefficient models for binary longitudinal responses with attrition. Statistics and Computing **10**, 279–287 (2000)
3. Alfò, M., Maruotti, A.: A selection model for longitudinal binary responses subject to non-ignorable attrition. Statistics in Medicine **28**, 2435–2450 (2009)
4. Bartolucci, F., Farcomeni, A.: A discrete time event-history approach to informative drop-out in mixed latent markov models with covariates. Biometrics **71**, 80–89 (2015)
5. Beunckens, C., Molenberghs, G., Verbeke, G., Mallinckrodt, C.: A latent-class mixture model for incomplete longitudinal gaussian data. Biometrics **64**, 96–105 (2008)
6. Creemers, A., Hens, N., Aerts, M., Molenberghs, G., Verbeke, G., Kenward, M.: Generalized shared-parameter models and missingness at random. Statistical Modelling **11**, 279–310 (2011)
7. Folstein, M., Folstein, S., McHig, P.: Mini-mental state: a pratical method for grading the cognitive state of patients for the clinician. Journal of Psychiatry Research **12**, 189–198 (1975)
8. Gao, S.: A shared random effect parameter approach for longitudinal dementia data with non-ignorable missing data. Statistics in Medicine **23**, 211–219 (2004)
9. Gao, W., Hedeker, D., Mermelstein, R., Xie, H.: A scalable approach to measuring the impact of nonignorable nonresponse with an ema application. Statistics in Medicine **35**, 5579–5602 (2016)
10. Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., Van Steen, K.: A local influence approach applied to binary data from a psychiatric study. Biometrics **59**, 410–419 (2003)
11. Little, R.: Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association **90**, 1112–1121 (1995)
12. Little, R., Rubin, D.: Statistical analysis with missing data, 2nd edition. Wiley (2002)
13. Ma, G., Troxel, A., Heitjan, D.: An index of local sensitivity to non-ignorability in longitudinal modeling. Statistics in Medicine **24**, 2129–2150 (2005)
14. Marino, M., Alfò, M.: A non-homogeneous hidden markov model for partially observed longitudinal responses. arXiv p. arXiv:1803.08255 (2018)
15. Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M.: Every missing not at random model has got a missing at random counterpart with equal fit. Journal of the Royal Statistical Society, Series B **70**, 371–388 (2008)
16. Muralidhar, K., Sarathy, R.: A comparison of multiple imputation and data perturbation for masking numerical variables. Journal of Official Statistics **22**, 507–524 (2006)

17. Rakhmawati, T., Molenberghs, G., Verbeke, G., Faes, C.: Local influence diagnostics for hierarchical count data models with overdispersion and excess zeros. Biometrical journal **58**, 1390–1408 (2016)
18. Rakhmawati, T., Molenberghs, G., Verbeke, G., Faes, C.: Local influence diagnostics for generalized linear mixed models with overdispersion. Journal of Applied Statistics **44**, 620–641 (2017)
19. Rubin, D.: Inference and missing data. Biometrika **63**, 581–592 (1975)
20. Scharfstein, D., Rotnitzky, A., Robins, J.: Adjusting for nonignorable drop-out using semiparametric nonresponse models. Journal of the American Statistical Association **94**, 1096–1120 (1999)
21. Spagnoli, A., Marino, M., Alfò, M.: A bidimensional finite mixture model for longitudinal data subject to dropout. Statistics in medicine **DOI:10.1002/sim.7698**, to appear (2018)
22. Troxel, A., Ma, G., Heitjan, D.: An index of local sensitivity to non-ignorability. Statistica Sinica **14**, 1221–1237 (2004)
23. Verzilli, C., Carpenter, J.: A monte carlo em algorithm for random-coefficient-based dropout models. Journal of Applied Statistics **29**, 1011–1021 (2002)
24. Viviani, S., Rizopoulos, D., Alfò, M.: Local sensitivity to non-ignorability in joint models. Statistical Modelling **14**, 205–228 (2014)
25. White, I., Carpenter, J., Horton, N.: A mean score method for sensitivity analysis to departures from the missing at random assumption in randomised trials. Statistica Sinica **to appear** (2017)
26. Wu, M., Bailey, K.: Estimation and comparison of changes in the presence of informative right censoring: conditional linear models. Biometrics **45**, 939–955 (1989)
27. Wu, M., Carroll, R.: Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. Biometrics **44**, 175–188 (1988)
28. Wulfsohn, M., Tsiatis, A.: A joint model for survival and longitudinal data measured with error. Biometrics **53**, 330–339 (1997)
29. Xie, H.: A local sensitivity analysis approach to longitudinal non-gaussian data with non-ignorable dropout. Statistics in Medicine **27**, 3155–3177 (2008)
30. Xie, H.: Analyzing longitudinal clinical trial data with nonignorable missingness and unknown missingness reasons. Computational Statistics and Data Analysis **56**, 1287–1300 (2012)
31. Xie, H., Heitjan, D.: Sensitivity analysis of causal inference in a clinical trial subject to crossover. Clinical Trials **1**, 21–30 (2004)