Max-Planck-Institut für demografische Forschung
Max Planck Institute for Demographic Research
Konrad-Zuse-Strasse 1 · D-18057 Rostock · GERMANY
Tel +49 (0) 3 81 20 81 - 0; Fax +49 (0) 3 81 20 81 - 202;
http://www.demogr.mpg.de

MPIDR WORKING PAPER WP 2008-022
SEPTEMBER 2008

A missing composite covariate in survival analysis: a case study of the Chinese Longitudinal Health and Longevity Survey

Francesco Lagona (lagona@uniroma3.it)
Zhen Zhang (zhang@demogr.mpg.de)

This working paper has been approved for release by: James W. Vaupel (jwv@demogr.mpg.de)
Head of the Laboratory of Survival and Longevity.

Working papers of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in working papers are attributable to the authors and do not necessarily reflect those of the Institute.

A missing composite covariate in survival analysis: a case study of the Chinese Longitudinal Health and Longevity Survey

Francesco Lagona
Department of Public Institutions, Economy and Society, University of Roma Tre, Rome, Italy.
Laboratory of Statistical Demography, Max Planck Institute for Demographic Research, Rostock, Germany.
email: lagona@uniroma3.it

and

Zhen Zhang
Laboratory of Survival and Longevity, Max Planck Institute for Demographic Research, Rostock, Germany
email: zhang@demogr.mpg.de

SUMMARY: We estimate a Cox proportional hazards model where one of the covariates measures the level of a subject's cognitive functioning by grading the total score obtained by the subject on the items of a questionnaire. A case study is presented where the sample includes partial respondents, who did not answer some or all of the questionnaire items. The total score takes hence the form of an interval-censored variable and, as a result, the level of cognitive functioning is missing on some subjects. We handle partial respondents by taking a likelihood-based approach where survival time is jointly modelled with the censored total score and the size of the censoring interval. Parameter estimates are obtained by an E-M-type algorithm that essentially reduces to the iterative maximization of three complete log-likelihood functions derived from two augmented datasets with case weights, alternated with weights updating. This methodology is exploited to assess the Mini Mental State Examination index as a prognostic factor of survival in a sample of Chinese older adults.

KEY WORDS: Beta-binomial regression; Composite index; Cox model; Data augmentation; E-M algorithm; Interval censoring; Non-ignorable missing covariate; Partial respondents.

1. Introduction

Composite covariates are indexes that summarize the values taken by several variables and are often exploited as predictors in regression modelling. In longevity studies, for example, the Mini-Mental State Examination (MMSE; Folstein *et al.*, 1975) index is frequently used to assess the cognitive mental status in older adults, and is often included as a covariate in a Cox (1972) proportional hazards model, to detect significant mortality differentials (Frisoni, *et al.*, 1999; Tilvis *et al.*, 2004; Lee *et al.*, 2006). The MMSE index is based on a questionnaire whose items are tests assessing orientation, attention, immediate and short-term recall, language, and the ability to follow simple commands. During the interview, each item receives a binary score, namely 1 for a correct answer and 0 if the answer is not correct. The cognitive mental status of a subject is typically assessed by comparing her/his score to a reference cut-off, which is chosen according to a specific definition of cognitive impairment (Lopez, *et al.*, 2005) or on the basis of population-based norms (as those reported, for example, by Crum, *et al.*, 1993), depending on the purpose of the analysis. Accordingly, the MMSE index grades the questionnaire total score in two levels, say 1 if the total score is greater than or equal to a cut-off d and 0 otherwise, clustering subjects into cognitively normal and impaired cases, respectively.

We present a case study of the Chinese Longitudinal Health and Longevity Survey (CLHLS), where cognitive functioning is assessed through a MMSE questionnaire, but the sample includes partial respondents, who did not answer some or all of the questionnaire items. The MMSE total score of partial respondents takes the form of an interval-censored variable, because the total score is only known to lie within a censoring interval. The lower extreme of this interval is equal to the partial score obtained by the subject on the observed part of the questionnaire. The size of the censoring interval is given by the number of the unanswered items. The MMSE index is a piece-wise constant function of the total score, with a jump at a cut-off point d. This index is hence missing when the censoring interval of the total score includes the cut-off point d (Figure 1). Given the cut-off d, the sample is therefore partitioned into three sub-samples that respectively include normal and impaired cases, and cases whose MMSE index level is unknown. These three sub-samples can be geometrically described by representing questionnaires as points whose coordinates are the number of missing items and the partial score obtained by the subject on the observed part of the

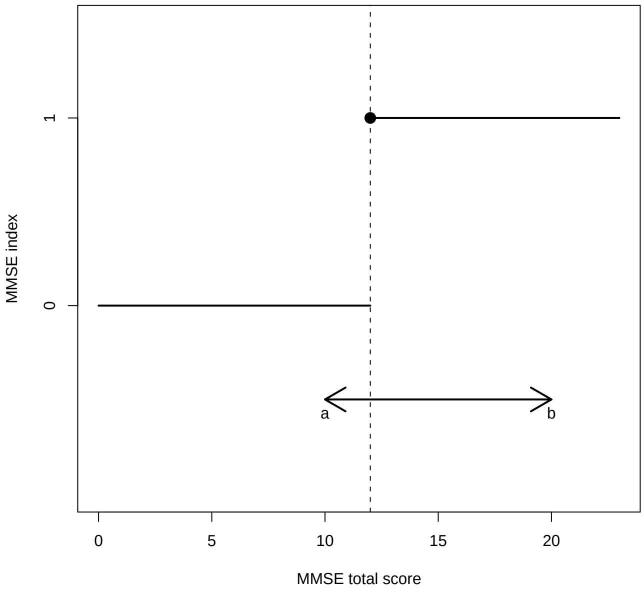

Figure 1. An example of a subject who correctly answers 10 questions, gives 3 wrong answers and leaves 10 items unanswered, in a MMSE questionnaire of 23 binary items. The subject's total score is interval-censored between $a = 10$ and $b = 20$. The piece-wise constant function describes the values taken by an MMSE index that grades the total score according to a cut-off $d = 12$. The index is missing because d is included in the censoring interval.

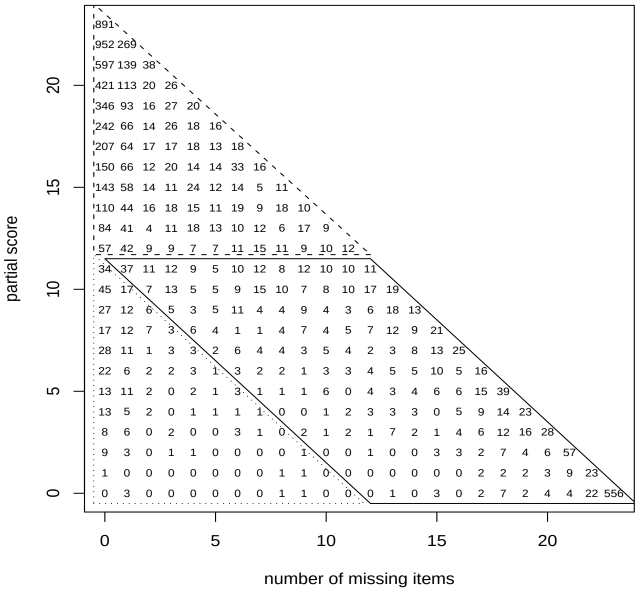

Figure 2. The number of questionnaires in the CLHLS dataset, clustered by the number of missing items and the partial score obtained by the subject on the observed part of the questionnaire. If the MMSE index grades the total score according to a cut-off $d = 12$, questionnaires in the lower (upper) triangle receive an index level 0 (1), while the index level is missing for all the questionnaires included in the parallelogram.

questionnaire. In this two-dimensional questionnaires space, subjects with a MMSE index level 0 (1) are included in a lower (upper) triangle, while subjects with a missing MMSE index are included in a parallelogram. The sizes of the three polygons depend on the cut-off d that have been chosen to specify the index. Figure 2 depicts the partitioning of the CLHLS questionnaires, showing whether a MMSE index is missing or takes a level 0/1, in an example where the cut-off d is equal to 12.

Because the MMSE index of partial respondents can be missing, we face with a missing value problem when this index is included in a Cox model as a covariate.

In gerontology studies that investigate the impact of the MMSE index on survival, two are the most popular approaches that are pursued to handle partial respondents. Referred to as complete cases (CC) analysis, a first approach is based on discarding subjects with a missing index from the study. Under a CC analysis, all the subjects with questionnaires in the parallelogram of the questionnaires space are discarded. The effect of the index is thus estimated by comparing subjects with questionnaires in the lower triangle and cases included in the upper triangle. Estimates provided by a CC analysis can be seriously biased, if the excluded subjects are not a random sample of the data, i.e., if the data are not missing completely at random (MCAR hypothesis; Rubin, 1976). In our case, it is difficult to motivate a MCAR assumption, because a missing MMSE index is the outcome of a disadvantageous combination of the cut-off chosen for grading the total scores, the number of missing items and the partial score of a questionnaire.

A second approach is based on counting missing answers as incorrect answers (missing-as-incorrect; MAI). By pursuing a MAI analysis, the lower triangle and the parallelogram of the questionnaire space are merged in one class of cognitively impaired cases. Results from a MAI analysis are difficult to interpret. Beside cognitive impairment, several can be the factors that lead to missing items in a questionnaire, including poor physical health, depression and anxiety. The effects of these factors are mixed with cognitive functioning when missing answers are counted as incorrect answers.

As a compromise between discarding partial respondents and including them as impaired cases, we work with a likelihood function where questionnaires contribute with different terms, according to the complete or partial information they provide. We essentially proceed along the general likelihood-based (LB) strategy that has been outlined by Herring *et al.* (2004), to estimate a Cox model with non-ignorable missing covariates. Within this methodological framework, we present a parsimonious model where, conditionally on the fully observed covariates x, the survival outcome t is jointly modelled with both the number m of missing items and the censored total score z obtained by a subject on the MMSE questionnarie. This joint distribution, say $p(t, m, z|x)$, is specified as the product of three one-dimensional conditional distributions. Specifically, a binomial regression model is exploited to specify the missing value mechanism, i.e., the conditional distribution of m, given the subject's profile (t, z, x). A Cox proportional hazards model is considered to specify a semi-parametric conditional distribution of the survival outcome t, given (z, x). A Beta-binomial regression is finally placed on the conditional distribution of the total score z, given the covariates x. Parameter estimation is carried out by a E-M-type algorithm, which essentially reduces to the iterative maximization of three complete log-likelihood functions on two augmented datasets with case weights, alternated with weights updating.

The rest of the paper is organized as follows. After reporting some details on the CLHLS data that stimulated this study (Section 2), modelling assumptions on the observed and missing data are outlined in Section 3. The practical implementation of the proposed LB approach is discussed in Section 4. In Section 5, we show the results provided by the proposed method on the CLHLS data and compare them with those obtained by CC- and MAI-based methods. Final comments are summarized in Section 6.

2. Data

The CLHLS data that motivated this article are drawn from the Study No. 3891 of the Inter-University Consortium for Political and Social Research (`www.icpsr.umich.edu`; Zeng *et al.*, 2002). The study was carried out on subjects aged 80 and above in 1998 and in two subsequent follow-up waves in 2000 and 2002. We have left-truncated, right censored survival data on 7908 subjects with a number of fully observed covariates, collected at the entry time: gender, type of residence (rural or urban), whether the subject is sedentary or active, and limits in activities of daily living (ADL; six activities including bathing, dressing, eating, indoor transferring, toileting and continence), categorized into three levels: no, one, two or more limits.

The covariate of main interest in this study is the MMSE index, which is computed from the total score obtained by a subject on the Chinese version of the MMSE questionnaire. We concentrate on the assessment of cognitive impairment as a prognostic factor and, accordingly, only the MMSE index obtained by subjects upon entry into the study is included in the analysis.

With respect to the standard 30-items MMSE (Folstein *et al.*, 1975), the 23-items Chinese MMSE adopts some appropriate adjustments to make the questions more understandable and answerable among ordinary oldest old Chinese, the majority of whom are illiterate (Zeng and Vaupel, 2002). For instance, the Chinese MMSE asks respondents to name as many foods as possible (in one minute) instead of writing a sentence, which is a quite difficult task for the elderly. Overall, respondents were asked 5 orientation related questions (naming the current time, animal year, season, festival, and county), one naming foods question, 6 word recall questions (3 words are mentioned and respondents are asked to repeat them two times), 4 calculations questions (respondents are asked to subtract 3 from 20, then 3 from the previous resulting, and so on), 3 language questions (repeating a sentence and naming simple items such as pen and watch that are shown to the respondents), 1 drawing question, and 3 comprehension questions (respondents are asked to take paper in their right hand, fold it, and then put it on the floor).

3. Modelling

3.1 *Likelihood-based analysis*

In the present study, the data are available for n subjects as vectors $(e_i, y_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{m}_i)$, $i = 1 \dots n$. For each subject i, e_i and y_i are respectively the entry and exit time, while δ_i is a failure indicator ($\delta_i = 1$ if a death occurred at y_i, and 0 otherwise) and $\boldsymbol{x}_i$ is a row profile of K fully observed covariates. Furthermore, $\boldsymbol{z}_i = (z_{i1} \dots z_{ij} \dots z_{iJ})$ is a row vector of binary covariates, some of which may be missing, where $z_{ij} = 1$ if subject i knows the correct answer to item j in the MMSE questionnaire, and 0 otherwise. Finally, $\boldsymbol{m}_i$ is a vector of missing indicators, say $m_{ij} = 1$, if z_{ij} is missing and 0 otherwise.

For each questionnaire i, we partition the items set $\{1 \dots J\}$ into the set $M(i) = \{j : m_{ij} = 1\}$ of the missing items and the set $O(i) = \{j : m_{ij} = 0\}$ of the observed items. Accordingly, $\boldsymbol{z}_{M(i)}$ denotes the vector of the $m_{i\cdot} = \sum_{j=1}^{J} m_{ij}$ missing scores, while $\boldsymbol{z}_{O(i)}$ indicates the vector of the $J - m_{i\cdot}$ observed scores. Furthermore,

$$z_{i\cdot}^{\text{obs}} = \sum_{j \in O(i)} z_{ij} = \sum_{j=1}^{J} z_{ij}(1 - m_{ij}), \text{ and}$$

$$z_{i\cdot}^{\text{mis}} = \sum_{j \in M(i)} z_{ij} = \sum_{j=1}^{J} z_{ij} m_{ij}$$

respectively denote the partial and the unobserved scores obtained by the ith subject. The total score $z_{i\cdot} = z_{i\cdot}^{\text{obs}} + z_{i\cdot}^{\text{mis}}$ is hence an interval-censored variable, with censoring interval $[z_{i\cdot}^{\text{obs}}, z_{i\cdot}^{\text{obs}} + m_{i\cdot}]$.

Information provided by $\boldsymbol{z}_i$ is summarized by the MMSE index

$$D(\boldsymbol{z}_i) = D(z_{i\cdot}) = \begin{cases} 1 & z_{i\cdot} \geqslant d \\ 0 & z_{i\cdot} < d \end{cases}$$

which is a piece-wise constant function of the total score $z_{i\cdot}$, with jump at the cut-off d. In the presence of missing items, only the partial score $z_{i\cdot}^{\text{obs}}$ is known and, as result, index D is equal to 1 if $z_{i\cdot}^{\text{obs}} \geqslant d$, it takes the value 0 if $z_{i\cdot}^{\text{obs}} + m_{i\cdot} < d$, and it is otherwise missing (Fig. 2).

Time up to death t is modelled by a semi-parametric Cox proportional hazards model. Precisely, we assume that the survival time of subject i is drawn from a positive random variable T with hazard function

$$\begin{aligned} h(t|\boldsymbol{x}_i, \boldsymbol{z}_i) =& h_0(t) \exp(\beta_0 D(\boldsymbol{z}_i) + \boldsymbol{x}_i \boldsymbol{\beta}_K) \\ =& h_0(t) r_i(\boldsymbol{\beta}) \end{aligned} \tag{1}$$

where $h_0(t)$ is a nonparametric baseline hazard function, $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_K)$ is a vector of fixed effects to be estimated and $r_i(\boldsymbol{\beta})$ is a parametric hazards ratio. Accordingly, the cumulative hazard is given by $H(t|\boldsymbol{x}_i, \boldsymbol{z}_i) = H_0(t) r_i(\boldsymbol{\beta})$ where $H_0(t) = \int_0^t h_0(\tau) d\tau$ is the baseline cumulative hazard.

In addition to the distribution of survival time, we introduce the parametric conditional distribution

$$p(\boldsymbol{m}_i|\boldsymbol{z}_i;\boldsymbol{\alpha}) = p(\boldsymbol{m}_i|e_i, y_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i;\boldsymbol{\alpha}) \tag{2}$$

of a subject's missingness pattern, known up to the parameters $\boldsymbol{\alpha}$. We also introduce the conditional distribution

$$p(\boldsymbol{z}_i|\boldsymbol{\gamma}) = p(\boldsymbol{z}_{O(i)}, \boldsymbol{z}_{M(i)}|\boldsymbol{x}_i;\boldsymbol{\gamma}) \tag{3}$$

of the subject's scores, some of which can be missing, driven by the unknown parameters $\boldsymbol{\gamma}$.

Following Herring and Ibrahim (2001), if the censoring time is non-informative and the censoring time distribution does not depend on the missing values, the likelihood contribution of subject i is proportional to

$$L_i(\boldsymbol{\theta}) = \sum_{\boldsymbol{z}_{M(i)}} p(\boldsymbol{m}_i|\boldsymbol{z}_i;\boldsymbol{\alpha}) L_i(\boldsymbol{\beta}|D(\boldsymbol{z}_i)) p(\boldsymbol{z}_i|\boldsymbol{\gamma}), \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, h_0)$ denotes all the parameters to be estimated and

$$L_i(\boldsymbol{\beta}|D(\boldsymbol{z}_i)) = (h_0(y_i) r_i(\boldsymbol{\beta}))^{\delta_i} \exp(H_0(e_i) - H_0(y_i))^{r_i(\boldsymbol{\beta})} \tag{5}$$

is the likelihood contribution of a left-truncated, right-censored subject, as derived from the Cox model (1).

In this study, the parameter of main interest is the effect β_0 of being cognitively normal, as measured by the MMSE index. Parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ are instead treated as nuisance parameters. Because parameter estimation may become too computationally intensive and unstable with many nuisance parameters, we need to employ strategies to reduce the number of nuisance parameters in the specification of both the missing data mechanism $p(\boldsymbol{m}_i|\boldsymbol{\alpha})$ and the score vector distribution $p(\boldsymbol{z}_i|\boldsymbol{\gamma})$. Following Ibrahim and Lipsitz (1996), the two joint distributions could be, for example, modelled as a product of one-dimensional conditional distributions, a strategy that greatly reduce the number of nuisance parameters. Unfortunately, this idea is still unpractical in our case, because of the large dimension of the MMSE questionnaire ($J = 23$). In the following, we present an alternative parsimonious specification of the two distributions, which is based on a binomial regression model for the missing data mechanism and a Beta-binomial regression model for the distribution of the vector of the questionnaire scores.

3.2 *The missing value mechanism*

An outcome often reported by the literature (Herzog and Wallace, 1997; Hayward and Gorman, 2004; Zimmer, *et al.* 2002) is the observation that missing scores on tests of cognitive impairment occur more frequently among cognitively impaired and/or physically disable patients. Motivated by this, we consider a binomial regression model where the expected number of missing items depend on the observed data and on cognitive impairment, as measured by the MMSE total score.

For each subject i, we specifically assume that the missingness pattern $\boldsymbol{m}_i$ depends on the scores vector $\boldsymbol{z}_i$ through the individual total score $z_{i\cdot}$ and that, conditionally on the observed data, the coordinates of vector $\boldsymbol{m}_i$ are i.i.d., i.e.

$$\begin{aligned} p(\boldsymbol{m}_i|e_i, y_i, \delta_i, \boldsymbol{x}_i, \boldsymbol{z}_i;\boldsymbol{\alpha}) =& p(\boldsymbol{m}_i|e_i, y_i, \delta_i, \boldsymbol{x}_i, z_{i\cdot};\boldsymbol{\alpha}) \\ =& \prod_{j=1}^{J} p(m_{ij}|e_i, y_i, \delta_i, \boldsymbol{x}_i, z_{i\cdot};\boldsymbol{\alpha}) \\ =& p_i(\boldsymbol{\alpha})^{m_{i\cdot}} (1 - p_i(\boldsymbol{\alpha}))^{J - m_{i\cdot}}, \end{aligned} \tag{6}$$

where $p_i(\boldsymbol{\alpha}) = p(m_{ij} = 1|\boldsymbol{\alpha})$ is the subject-specific conditional probability of a missing item. We complete the specification of the binomial regression model by assuming a canonical link transformation

$$\text{logit} p_i(\boldsymbol{\alpha}) = \eta_i(\boldsymbol{\alpha}),$$

where $\eta_i(\boldsymbol{\alpha})$ is a linear predictor that is defined on the basis of the variables $(e_i, y_i, \delta_i, \boldsymbol{x}_i, z_{i\cdot})$.

3.3 *The MMSE score distribution*

Studies on the same CLHLS data considered in this paper have shown significant gender differentials in cognitive impairment (Zhang, 2006) and a strong link between cognitive functioning and limits in daily activities (Gu and Qiu, 2003). A Beta-binomial regression is a parsimonious model that allows us to include the effects of these covariates and simultaneously accounts for correlated scores. Specifically, we assume that the single MMSE binary scores obtained by the ith subject are exchangeable Bernoulli random variables, with common marginal correlation ρ_i and success probability drawn from a Beta density with mean π_i, where $\pi_i = \mathbb{E}z_{ij}, j = 1 \dots J$, is the marginal common mean of the J scores. We exploit a logit link for π and a Fisher's z-transform for ρ, as follows

$$\begin{aligned} \log\frac{\pi_i}{1-\pi_i} =& \gamma_{\mu,0} + \boldsymbol{x}_i \boldsymbol{\gamma}_{\mu,K} \\ \log\frac{1+\rho_i}{1-\rho_i} =& \gamma_\rho, \end{aligned} \tag{7}$$

where $\boldsymbol{\gamma} = (\gamma_{\mu,0}, \boldsymbol{\gamma}_{\mu,K}, \gamma_\rho)$ are parameters to be estimated. Under this setting, the distribution of the score vector obtained by subject i is given by:

$$\begin{aligned} p(\boldsymbol{z}_i|\boldsymbol{\gamma}) =& \int_0^1 p^{z_{i\cdot}} (1-p)^{J - z_{i\cdot}} b(p|\pi_i, \rho_i) dp \\ =& \frac{B(\frac{\pi_i(1-\rho_i)}{\rho_i} + z_{i\cdot}, \frac{(1-\pi_i)(1-\rho_i)}{\rho_i} + J - z_{i\cdot})}{B(\frac{\pi_i(1-\rho_i)}{\rho_i}, \frac{(1-\pi_i)(1-\rho_i)}{\rho_i})}, \end{aligned} \tag{8}$$

where

$$b(p|\pi, \rho) = \frac{p^{\psi(\pi,\rho)-1}(1-p)^{\xi(\pi,\rho)-1}}{B(\pi, \rho)}$$

is the Beta distribution with parameters

$$\begin{aligned} \psi(\pi, \rho) =& \frac{\pi(1-\rho)}{\rho} \\ \xi(\pi, \rho) =& \frac{(1-\pi)(1-\rho)}{\rho} \end{aligned}$$

and $B(\pi, \rho) = \int_0^1 b(p|\pi, \rho) dp$ is the Beta function.

4. Data augmentation and parameter estimation

Parameter $\boldsymbol{\theta}$ can be computed by the E-M-type algorithm suggested by Herring *et al.* (2004), following the approximation arguments provided by Herring and Ibrahim (2001). This algorithm reduces to the iterative evaluation of a set of weighted score equations, alternated with weights updating, and can be conveniently illustrated by using the counting process notation. Using this notation, the information contained in (e_i, y_i, δ_i) is represented by the bivariate process $(N_i(t), R_i(t))$, where the death process $N_i(t) = 1$ if the subject dies at or before time t and 0 otherwise, while the risk process $R_i(t) = 1$ if the subject is in the study at time t and 0 otherwise.

At each step of the iteration, the estimate $\hat{\boldsymbol{\theta}}$ available from the previous iteration is exploited to compute the predictive distribution of the missing scores in the ith case, namely

$$w(\boldsymbol{z}_{M(i)}|\hat{\boldsymbol{\theta}}) = \frac{p(\boldsymbol{m}_i|\boldsymbol{z}_i;\hat{\boldsymbol{\alpha}})L_i(\hat{\boldsymbol{\beta}}|D(\boldsymbol{z}_i))p(\boldsymbol{z}_i|\hat{\boldsymbol{\gamma}})}{\sum_{\boldsymbol{z}_{M(i)}} p(\boldsymbol{m}_i|\boldsymbol{z}_i;\hat{\boldsymbol{\alpha}})L_i(\hat{\boldsymbol{\beta}}|D(\boldsymbol{z}_i))p(\boldsymbol{z}_i|\hat{\boldsymbol{\gamma}})}. \tag{9}$$

Parameter estimates are then updated by solving the following set of weighted score equations

$$\bar{\boldsymbol{u}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\sum_{\boldsymbol{z}_{M(i)}} \begin{pmatrix} u_i(\beta_0|D(\boldsymbol{z}_i)) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_K|D(\boldsymbol{z}_i)) \\ u_i(h_0|D(\boldsymbol{z}_i)) \\ \boldsymbol{u}_i(\boldsymbol{\alpha}|\boldsymbol{z}_i) \\ \boldsymbol{u}_i(\boldsymbol{\gamma}|\boldsymbol{z}_i) \end{pmatrix} w(\boldsymbol{z}_{M(i)}|\hat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{10}$$

where

$$\begin{aligned} u_i(\beta_0|D(\boldsymbol{z}_i)) &= \int_0^\infty \left(D(\boldsymbol{z}_i) - \bar{D}_w(\boldsymbol{\beta}, u)\right) dN_i(u) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_K|D(\boldsymbol{z}_i)) &= \int_0^\infty \left(\boldsymbol{x}_i - \bar{X}_w(\boldsymbol{\beta}, u)\right) dN_i(u) \\ u_i(h_0|D(\boldsymbol{z}_i)) &= dN_i(t) - h_0(t)r_i(\hat{\boldsymbol{\beta}})R_i(t) \\ \boldsymbol{u}_i(\boldsymbol{\alpha}|\boldsymbol{z}_i;\hat{\boldsymbol{\theta}}) &= \frac{\partial}{\partial\boldsymbol{\alpha}} \log p(\boldsymbol{m}_i|\boldsymbol{z}_i;\boldsymbol{\alpha}) \\ \boldsymbol{u}_i(\boldsymbol{\gamma}|\boldsymbol{z}_i;\hat{\boldsymbol{\theta}}) &= \frac{\partial}{\partial\boldsymbol{\gamma}} \log p(\boldsymbol{z}_i|\boldsymbol{\gamma}) \end{aligned}$$

while

$$\begin{aligned} \bar{D}_w(\boldsymbol{\beta}, u) &= \frac{\sum_{i=1}^n \sum_{\boldsymbol{z}_{M(i)}} w(\boldsymbol{z}_{\boldsymbol{M(i)}}|\hat{\boldsymbol{\theta}})D(\boldsymbol{z}_i)R_i(u)r_i(\boldsymbol{\beta})}{\sum_{i=1}^n \sum_{\boldsymbol{z}_{M(i)}} w(\boldsymbol{z}_{\boldsymbol{M(i)}}|\hat{\boldsymbol{\theta}})R_i(u)r_i(\boldsymbol{\beta})} \\ \bar{X}_W(\boldsymbol{\beta}, u) &= \frac{\sum_{i=1}^n \sum_{\boldsymbol{z}_{M(i)}} w(\boldsymbol{z}_{\boldsymbol{M(i)}}|\hat{\boldsymbol{\theta}})\boldsymbol{x}_i R_i(u)r_i(\boldsymbol{\beta})}{\sum_{i=1}^n \sum_{\boldsymbol{z}_{M(i)}} w(\boldsymbol{z}_{\boldsymbol{M(i)}}|\hat{\boldsymbol{\theta}})R_i(u)r_i(\boldsymbol{\beta})}. \end{aligned}$$

The first two components of (10) are the $K+1$ score equations suggested by Herring and Ibrahim (2001) to update the parameters of a Cox model with missing covariates. These equations provide an updated estimate $\tilde{\boldsymbol{\beta}}$ that can be exploited in $u_i(h_0|D(\boldsymbol{z}_i))$ to update the baseline hazard and computing a new Breslow's (1974) estimate $\tilde{H}_0$ of the cumulative hazard. The last two components of (10) separately provide us with the updated estimates $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$. Estimate $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}, \tilde{H}_0, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ is then used to update the weighting schemes (9). The algorithm is iterated up to convergence of the estimates.

Significant simplifications arise in the practical implementation of this algorithm, under the parsimonious specification considered in this paper. Under the modelling assumption (5),(6) and (8), the predictive weights depend on $\boldsymbol{z}_i$ only through the totals $z_{i\cdot}^{\text{obs}}$ and $z_{i\cdot}^{\text{mis}}$, as follows

$$w(\boldsymbol{z}_{M(i)}|\hat{\boldsymbol{\theta}}) = w(z_{i\cdot}^{\text{mis}}|z_{i\cdot}^{\text{obs}};\hat{\boldsymbol{\theta}}) =$$

$$\frac{p(\boldsymbol{m}_i|z_{i\cdot}^{\text{obs}}+z_{i\cdot}^{\text{mis}};\hat{\boldsymbol{\alpha}})L_i(\hat{\boldsymbol{\beta}}|D(z_{i\cdot}^{\text{obs}}+z_{i\cdot}^{\text{mis}}))p(z_{i\cdot}^{\text{obs}}+z_{i\cdot}^{\text{mis}}|\hat{\boldsymbol{\gamma}})}{\sum_{j=0}^{m_{i\cdot}} \binom{m_{i\cdot}}{j} p(\boldsymbol{m}_i|z_{i\cdot}^{\text{obs}}+j;\hat{\boldsymbol{\alpha}})L_i(\hat{\boldsymbol{\beta}}|D(z_{i\cdot}^{\text{obs}}+j))p(z_{i\cdot}^{\text{obs}}+j|\hat{\boldsymbol{\gamma}})}.$$

As a result, the last two components of the score vector (10) reduce to

$$\bar{\boldsymbol{u}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\sum_{j=0}^{m_{i\cdot}} \begin{pmatrix} \boldsymbol{u}_i(\boldsymbol{\alpha}|\boldsymbol{z}_i) \\ \boldsymbol{u}_i(\boldsymbol{\gamma}|\boldsymbol{z}_i) \end{pmatrix} w_{ij}(\hat{\boldsymbol{\theta}}) = \mathbf{0}, \tag{11}$$

where $w_{ij}(\hat{\boldsymbol{\theta}}) = w(z_{i\cdot}^{\text{mis}} = j|z_{i\cdot}^{\text{obs}};\hat{\boldsymbol{\theta}})$. The roots of the equation above can be computed by separately fitting a binomial regression model and a Beta-binomial regression model on an augmented dataset D1, obtained by including all the subjects with no missing items, each weighted by $w = 1$, and replacing each partial respondent i with $m_{i\cdot} + 1$ pseudo-profiles, each given a total MMSE score $z_{i\cdot}^{\text{obs}} + j$ and a case weight $w_{ij}(\hat{\boldsymbol{\theta}})$.

The first two components of (10) reduce to

$$\bar{\boldsymbol{u}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n}\sum_{l=0}^{1} \begin{pmatrix} u_i(\beta_0|D=l) \\ \boldsymbol{u}_i(\boldsymbol{\beta}_K|D=l) \end{pmatrix} W_{il}(\hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{12}$$

where

$$W_{il}(\hat{\boldsymbol{\theta}}) = \begin{cases} \sum_{j:z_{i\cdot}^{\text{obs}}+j<d} w_{ij}(\hat{\boldsymbol{\theta}}) & l=0 \\ \sum_{j:z_{i\cdot}^{\text{obs}}+j\geqslant d} w_{ij}(\hat{\boldsymbol{\theta}}) & l=1. \end{cases}.$$

Estimates $\tilde{\boldsymbol{\beta}}$ can be hence found by exploiting an augmented dataset D2, obtained by including all the subjects with an observed MMSE index D, each weighted by 1, and replacing each subject i with a missing MMSE index with two pseudo-subjects with composite index equal to 0 and 1, respectively weighted by $W_{i0}(\hat{\boldsymbol{\theta}})$ and $W_{i1}(\hat{\boldsymbol{\theta}})$. Finally, the baseline hazard can be updated as follows:

$$\tilde{h}(t) = \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n \sum_{l=0}^1 R_i(t)r_i(\tilde{\boldsymbol{\beta}})W_{il}(\hat{\boldsymbol{\theta}})}.$$

As pointed out by Herring and Ibrahim (2004), variance estimation of the parameters of interest in this algorithm is complicated because of the large dimension of the vector of the hazard estimates. Following Goetghebeur and Ryan (2000), they suggest to impute missing data by sampling values of the missing variable, to obtain naive point estimates and variance estimates of the parameter of interest. Then the variance of the EM estimator is obtained as a weighted sum of the mean of the imputation variances and the empirical variance of the imputation point estimates, with weights 1 and m, where m is the number of imputation used. In our case, values of D can be imputed by sampling values of the total latent score $z_{i\cdot}^{\text{mis}}$ from the predictive distribution

$$\begin{aligned} &p(z_{i\cdot}^{\text{mis}} = j|z_{i\cdot}^{\text{obs}}, \boldsymbol{x}_i, \hat{\gamma}) = \\ &\binom{m_{i\cdot}}{j} \frac{B(j+z_{i\cdot}^{\text{obs}}+\hat{\pi}_i\frac{1-\hat{\rho}_i}{\hat{\rho}_i}, J-j-z_{i\cdot}^{\text{obs}}+(1-\hat{\pi}_i)\frac{1-\hat{\rho}_i}{\hat{\rho}_i})}{B(z_{i\cdot}^{\text{obs}}+\hat{\pi}_i\frac{1-\hat{\rho}_i}{\hat{\rho}_i}, J-m_{i\cdot}-z_{i\cdot}^{\text{obs}}+(1-\hat{\pi}_i)\frac{1-\hat{\rho}_i}{\hat{\rho}_i})}, \end{aligned} \tag{13}$$

where $\hat{\gamma}$ is the point estimate of γ that is obtained at last iteration of the algorithm.

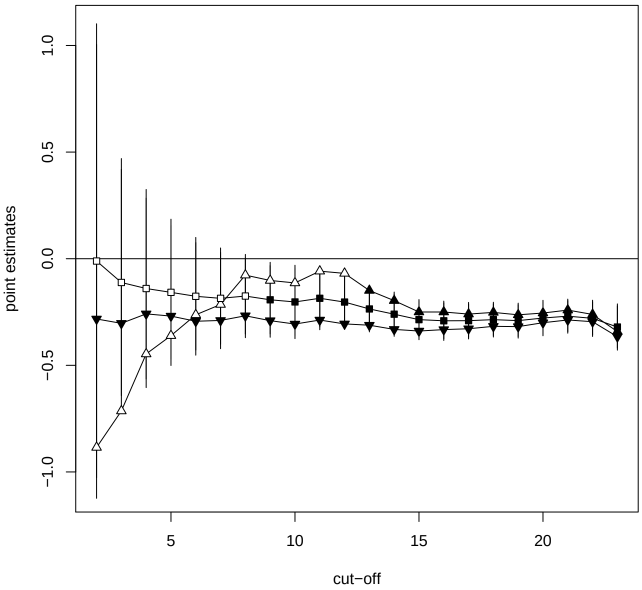

Figure 3. CC (point-up triangles), MAI (point-down triangles) and LB (squares) estimates of the effect of being cognitively normal, estimated by a battery of Cox models that include different definitions of the MMSE index, according to a sequence of cut-off points. Black (white) symbols indicate significant (not significant) estimates at a 95% confidence level. Segments represent the 95% confidence intervals of LB estimates.

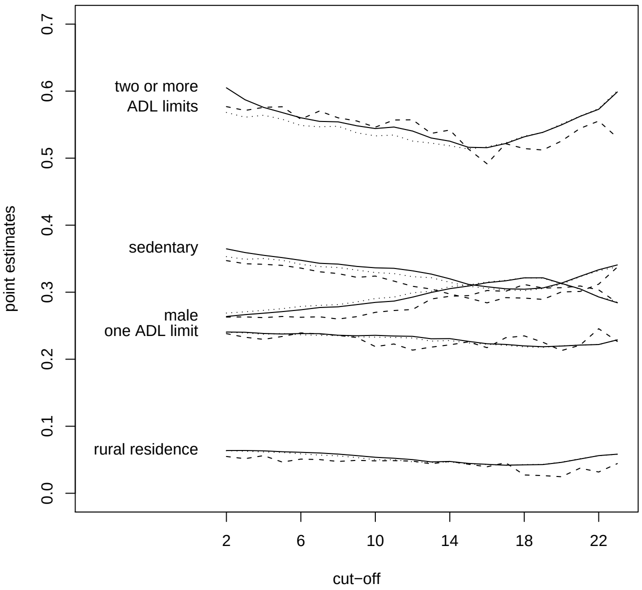

Figure 4. The effect of a number of covariates on survival, as estimated by a battery of Cox models that include different definitions of the MMSE index, according to a sequence of cut-off points. Results under a MAI approach (dotted lines), a CC analysis (dashed lines) and a LB strategy (solid lines).

5. Results

The structure of the missing value problem considered in this paper depends on the choice of the cut-off point d that specifies the MMSE index, as illustrated in the Introduction. Specifically, the partitioning of the questionnaire space (Figure 2) in subsamples of normal, impaired and missing cases depends on d. The outcomes of alternative estimation strategies can be therefore conveniently compared by repeating the analysis for a sequence of different cut-offs. We have thus considered a battery of 22 MMSE indexes, as defined by a sequence of cut-off points ($d = 2, 3, \ldots 23$). These indexes have been then separately included among the covariates of 22 Cox models. The resulting battery of Cox models have been estimated under a CC, MAI and LB analysis. For each cut-off d, log-hazards differences between normal and impaired subjects are depicted in Figure 3, while Figure 4 shows the point estimates of the effects of the additional covariates included in the Cox model. Regardless of the estimation method and the MMSE cut-off point, the estimated effect of the MMSE index is never positive, indicating that, overall, the mortality risk among normal subjects is not higher than the risk experienced by impaired subjects, even after adjusting for gender, type of residence, physical disabilities and life style.

Significant differences however appear between LB, MAI and CC estimates, when the cut-off takes lower values, say $2 \leqslant d \leqslant 12$. These differences decrease as the cut-off point increases. Partial respondents play a major role in the interpretation of these differences. Under a MAI analysis, partial respondents are treated as impaired cases if their partial score is less than the threshold d. For lower values of the cut-off point, most of these partial respondents are discarded by a CC analysis. When these subjects are excluded from the analysis, the significant effect of cognitive impairment, as detected by a MAI analysis, becomes not significant under a CC analysis. This implies that lower scores obtained on questionnaires with many missing items (the parallelogram of the questionnaire space) are associated with higher mortality risks than those experienced by subjects who obtained lower scores on complete questionnaires (the lower triangle in the questionnaire space). For lower values of the cut-off point, in summary, MAI estimates are significant because they are based on a overestimated number of impaired cases, while CC estimates are not significant because they are based on the exclusion of subjects with a high mortality risk. As d increases, partial respondents with lower scores are progressively replaced in the sample by a CC analysis, CC estimates become significant and differences between MAI and CC estimates decrease.

LB estimates appear as a reasonable compromise between the outcomes of the MAI and CC analysis. Estimates smoothly decrease as the cut-off increases, indicating that a value of the MMSE total score in the neighborhood of 10 can be already used as a prognostic cut-off to detect significant risk differentials. The critical cut-off $d = 8$, after which the LB-based effect is significant, is lower than that estimated by a CC analysis ($d = 12$), because the LB analysis appropriately includes those subjects that are excluded by a CC analysis.

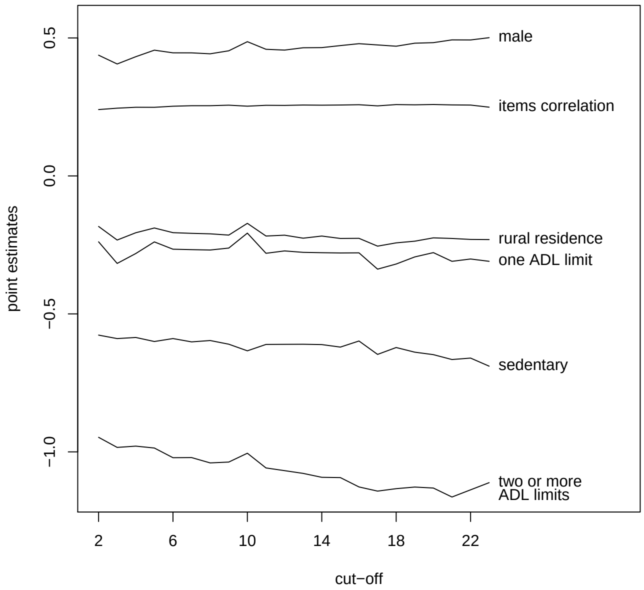

Figure 5. Effects of the fully observed covariates on the MMSE total score, as estimated by a Beta-binomial regression model, including pairwise items correlation, for each MMSE cut-off point, under a LB estimation strategy.

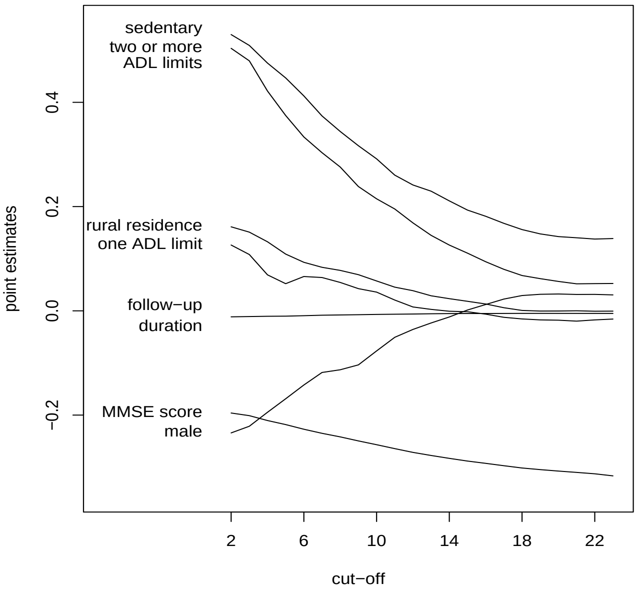

Figure 6. Effects of the fully observed covariates, the MMSE total score and the follow-up duration on a subject's probability to leave a questionnaire item unanswered, as estimated by a binomial regression model, for each MMSE cut-off point, under a LB estimation strategy.

Simultaneously, the LB-based effect of the MMSE index is less strong than than predicted by a MAI procedure, where all the partial respondents are treated as impaired cases.

Nevertheless, the LB estimation method essentially confirms the effects of gender, type of residence, physical disabilities and life style, as estimated by pursuing a CC or a MAI procedure (Figure 4). The individual physical status, as measured by ADL limits and life style, has a stronger impact on survival than the type of residence. As expected, males have a higher mortality risk than females. Viewed as functions of the threshold d, LB and MAI estimates appear smoother than those resulting from a CC analysis. It is possible that cut-off-specific exclusions and replacements of partial respondents affect the pattern of CC estimates.

Figures 5 and 6 show the estimates of the nuisance parameters that are exploited by the LB method to weight partial respondents, for each cut-off d.

Figure 5 shows, in particular, the estimates of the Beta-binomial regression model that has been assumed for the distribution of the MMSE total score. After adjusting for the items correlation within questionnaires, males in the sample are less cognitively impaired than females, confirming the results on gender differentials found by Zhang (2006) on the same CLHLS data used in this paper. In keeping with the outcomes reported by Gu and Qiu (2003), urban residents are less cognitively impaired than rural residents, while physical disabilities and life styles negatively influence a subject's cognitive functioning.

Figure 6 shows the estimates of the parameters that drive the missing value mechanism (2), as specified by a binomial regression model whose covariates include the survival outcome (as measured by the follow-up duration) and the MMSE score, in addition to the fully observed covariates. Although follow-up duration was expected to be an important factor for the probability of coping with MMSE items, it does not seem to play a significant role in our case study. On the contrary, the negative effect of the MMSE score indicates that missing answers occur more frequently among cognitively impaired subjects, as expected. Finally, the influence of the fully observed covariates on the weighting schemes of partial respondents decreases as the cut-off value increases, reflecting the convergence of the LB, MAI and CC outcomes, shown by Figure 3.

6. Discussion

Motivated by a specific case study, we have presented a likelihood-based strategy to estimate the Cox model when one of the covariates is a piece-wise constant function of the total score obtained by a subject on a questionnaire, but some of the questionnaires in the sample are partially observed. We have shown that this particular missing value problem can be naturally handled by a likelihood-based approach where the survival outcome is jointly modelled with the missing value mechanism and the total score distribution. A parsimonious specification of the latter two models greatly reduces the number of nuisance parameters and the computational complexity of the estimation algorithm, through an appropriate augmentation of the observed data. The proposed LB approach enhances the outcomes that are obtained when subjects with missing values are removed from the analysis or when missing answers are counted as incorrect answers. The signs of the nuisance parameters are in keeping with the findings reported by the literature about both the relationship between cognitive impairment and physical disabilities, and the factors that are influential in the occurrence of unanswered items in a MMSE questionnaire.

Although encouraging, the outcomes of a LB strategy strongly depend on the assumptions that have been made on the distributions of both the missing and observed data.

Through this paper, we have essentially assumed that the items of a MMSE questionnaire are homogeneous. Specifically, we have assumed that the scores on the single items are exchangeable, placing a Beta-binomial distribution on the questionnaire total score. The inclusion of a Beta-distributed random effect simultaneously allows for unobserved heterogeneity between subjects and correlated scores within the questionnaire of each subject, strategically compensating for unobserved covariates (e.g., educational level) that could be influential in the measurement of cognitive functioning. On the other side, this model explicitly assumes that the successful coping with a questionnaire item does not depend on the item. Because more flexible models would typically involve a greater number of nuisance parameters and our data did not show evidence of heterogeneity in items difficulties, we have taken a Beta-binomial regression as a reasonable compromise between realism and parsimony.

Items homogeneity was also assumed in the specification of the missing value mechanism. Conditionally on cognitive impairment, we have assumed that the pattern of the unanswered items is drawn at random by a Binomial distribution. This can be a shortcoming when factors such as fatigue or anxiety are responsible for a dependence structure between answered and unanswered items. The availability of detailed information on the MMSE interview would allow to check the independence assumption and perhaps to try more complex models. On the basis of the available data, a binomial regression model parsimoniously captures the relationship between the number of missing items and the cognitive impairment of a subject, as measured by the MMSE total score.

Although with these limitations, a LB analysis allows for a sharper validation of the MMSE index as a prognostic factor, compared to popular protocols that are based upon either the exclusion or the deterministic classification of partial respondents.

Acknowledgements

The authors are grateful to Dr. Heiner Maier and Dr. Li Qiang for their valuable suggestions. The 1998, 2000, and 2002 waves of the Chinese Longitudinal Healthy Longevity Survey were supported by NIA/NIH grant P01 AG 08761.

References

Breslow, N. (1974). Covariance Analysis of Censored Survival Data, *Biometrics*, 30, 89-99.

Cox, D. R. (1972). Regression Models and Life-Tables (with discusstion). *Journal of the Royal Statistical Society*, Ser. B, 34, 187-220.

Crum, R.M, Anthony J.C, Bassett S.S. and Folstein M.F (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Journal of the American Medical Association*, 269(18), 2386-2391.

Folstein, M. F., Folstein, S. E., and McHugh, P. R. (1975). 'Mini-mental state': A practical method for grading the cognitive state of patients for the clinician, *Journal of Psychiatry Research*, 12, 189-198.

Frisoni, G. B., Fratiglioni, L.; Fastbom, J., Viitanen, M., Winblad, B. (1999). Mortality in Nondemented Subjects with Cognitive Impairment: The Influence of Health-related Factors. *American Journal of Epidemiology*, 150(10), 1031-1044.

Goetghebeur, E. and Ryan, L. (2000). Semiparametric regression analysis of interval-cenosred data. *Biometrics*, 56, 1139-1144.

Gu, D. and Qiu, L. (2003). Cognitive functioning and its determinants among the oldest-old in China, *Journal of Nanjng College for Population and Management*, 2, 3-9.

Hayward, M.D. and Gorman, B.K. (2004). The Long arm of childhood: the influence of early-life social conditions on men's mortality. *Demography*, 41(1), 87107.

Herring, A. H., Ibrahim, J. G. and Lipsitz, S.R. (2004). Nonignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial , *Applied Statistics*, 53, 293-310.

Herzog, A.R and Wallace, R.B. (1997). Measures of cognitive functioning in the AHEAD Study, *Journals of Gerontology: Psycological Sciences and Social Sciences*, 52B (Special Issue), 37-48.

Ibrahim, J.G., Chen, M.-H., Lipsits, S.R. and Herring, A. (2005). Missing-Data Methods for Generalized Linear Models: A Comparative Review, *Journal of the American Statistical Association*, 100, 332-346.

Ibrahim, J. G. and Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, 52, 1071-1078.

Lee, H. B., Kasper, J. D., Shore, A. D., Yokley, J. L. , Black, B. S. and Rabins, P. V. (2006). Level of Cognitive Impairment Predicts Mortality in High-Risk Community Samples: The Memory and Medical Care Study, *Journal of Neuropsychiatry and Clinical Neurosciences*, 18, 543-546.

Lopez, M. N., Charter, R.A., Mostafavi, B., Nibut, L.P. and Smith, W.E. (2005). Psychometric Properties of the Folstein Mini-Mental State Examination, *Assessment*, 12(2), 137-144.

Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, 63, 81-92.

Tilvis, R. S., Khnen-Vre, M. H., Jolkkonen, J., Valvanne, J., Pitkala, K. H. and Strandberg T. E. (2004). Predictors of Cognitive Decline and Mortality of Aged People Over a 10-Year Period, *The Journals of Gerontology*, A, 59, M268-M274.

Zeng, Y. and J. W. Vaupel (2002). Functional Capacity and Self-Evaluation of Health and Life of Oldest Old in China. *Journal of Social Issues*, 58, 733-748.

Zeng, Y. et al. (2002). Sociodemographic and Health Profiles of the Oldest Old in China, *Population and Developement Review*, 28, 251-273.

Zhang, Z. (2006). Gender differentials in cognitive impairment and decline in the oldest old in China, *Journal of Gerontology: Social Sciences*, 61B(2), S107-S115.

Zimmer, Z., Martin, L. G. and Chang, M.-C. (2002). Changes in functional limitation and survival among older Taiwanese, 1993, 1996, and 1999, *Population Studies*, 56(3), 265 276.