

# Joint Analysis of Longitudinal Latent Health Related Quality of Life and a Survival Process.

Mounir Mesbah<sup>1</sup>

*Laboratoire de Statistique Théorique et Appliquée, Université Pierre et Marie Curie, Paris6,  
mesbah@ccr.jussieu.fr*

**Abstract:** In this work, joint analysis of a longitudinal latent multidimensional variable and an event time is considered. The latent variable is measured through a Rasch model using data from questionnaires assessed at several time visits. The responses of our model are two way correlated. First, at a given visit, the responses to questions (items) of a single individual are correlated and second, they are repeated over the visits, they also become correlated. It is, however, well known that a full likelihood analysis for such mixed models is hampered by the need for numerical integrations. To overcome such integration problems, generalized estimating equations approach is used, following useful approximations. Fixed effects parameters and variance components are estimated consistently by asymptotical normal statistics. A second level of difficulty is the occurrence of death or missing response at dropout time. The likelihood of a global solution is derived, while a full statistical inference is given when the analysis is performed in two separate steps.

## 1- Introduction

### Part A: Some Latent Variable Model Results

## 2- The Rasch Model

## 3- Generalized Estimating Equations

- GEE1 Method
- GEE2 Method
- Approximation of the likelihood function
- Simulations results

## 4- Longitudinal Rasch Model

## 5- Multivariate Mixed Rasch Model for Quality of life scales

## 6- Estimation of latent parameters in Rasch model

### Part B: Joint analysis of a Longitudinal Latent HRQoL and a Survival

## 7- Joint analysis of a Longitudinal QoL variable and an event time

## 8- Joint analysis of a latent Longitudinal HRQoL variable and an event time

## 9- Conclusion

## 10- References

---

\*This work was done partially while the author was visiting the Institut for Mathematical Sciences, National University of Singapore in 2005/2006. The visit was supported by the Institute."

<sup>1</sup>University of Pierre et Marie Curie, Paris 6. 175 rue du Chevaleret, Bureau 8A25,75013, Paris, France

*Keywords and phrases:* Quality of Life, Cox Model, GEE, Graphical Model, Latent variable, Rasch model, Mixed model, Missing data, Joint analysis, Longitudinal analysis

## 1. Introduction

The relationship between a time-dependent covariate and a survival time process is usually assessed using the Cox model as a semiparametric specification of the distribution of the survival time **conditional** to the covariate. Time-dependent covariates are generally available as longitudinal data collected regularly or not during the course of the study, with, frequently, occurrence of missing covariate data. When the investigated covariate is **internal** (Kalbfleisch and Prentice, 1980), conditional analysis is insufficient or incorrect. Dupuy and Mesbah (2002) use a joint model for survival and the longitudinal covariate to estimate the parameters in the Cox model. Identifiability of this joint model, existence and consistency of nonparametric maximum likelihood estimators and asymptotic distribution of the estimators is obtained along with consistent estimator of the asymptotic variance (Dupuy, Grama and Mesbah (2006)).

In Health Related Quality of Life (HRQoL) studies, the longitudinal covariate is assessed by the way of a questionnaire consisting in set of questions (items). Responses to these questions are dichotomous or polytomous qualitative variables (the latter are generally ordinal). So, at each time (visit) the longitudinal covariate is a multivariate binary or ordinal polytomous variable. The usual way to handle such data is first to build few quantitative scores summarizing individual HRQoL data at each visit, then analyzing these scores by conditional or joint models similar to those indicated in previous paragraph (see also Awad, Zuber and Mesbah (2002) or Mesbah, Dupuy, Heutte and Awad (2004)).

Health Related Quality of Life is, mainly, a psycho-social concept. It is a multidimensional **latent** variable measured by an instrument, the HRQoL questionnaire. The observed multidimensional qualitative responses obtained by the questionnaire through items (questions) are **manifest noised variables** useful, only, to estimate the **true latent** HRQoL of main interest, i.e. building scores summarizing individual HRQoL. Psycho-sociometric models relating those latent variables to their manifest expression are basically models of unidimensionality. Most popular are classical parallel models for quantitative manifest items or modern item response theory models for ordinal polytomous responses. The Rasch model for binary outcomes (Rasch, (1960), Fisher and Molenaar, (1995)) or its natural extension to ordinal data, the Partial Credit model (Masters, (1982), Dorange, Chwalow et Mesbah(2003)) are the standard unidimensional models.

The Multivariate mixed Rasch model is used to analyze binary or ordinal responses of questionnaires assessed at several time visits, with possible death and censorship.

A first Multivariate level is handled to deal with multiple binary responses following a unidimensional latent model using the Rasch model.

A second Multivariate level is the longitudinal assessment of the questionnaires i.e. the same latent variable at different visits.

A third Multivariate level is due to the fact that the questionnaire include various dimensions (subscales) to identify and analyze, so, various dependent latent variables to identify and analyze longitudinally.

And finally, the last Multivariate level is given by the relationship of the **longitudinal HRQoL** with **survival**, and **Treatment Group**. The joint model of Dupuy and Mesbah (2002) Model is extended to the latent case.

This paper is organized in two parts. The first part is devoted to latent models for multi categorical data, with their extension to the longitudinal case and the multi latent case and use of marginal GEE methods to estimate faster their param-

eters. Then the joint model of Dupuy and Mesbah (2002) is presented followed by derivation of the likelihood of a new joint model for a longitudinal latent HRQoL and a Survival process.

PART A: SOME LATENT VARIABLE MODEL RESULTS

2. The Rasch Model

The Rasch model specify the conditional (to the latent value) probability response of a subject  $i$  to a question  $j$  by:

$$(1) \quad P(X_{ij} = x_{ij} | \theta_i, \beta_j) = f(x_{ij}, \theta_i, \beta_j) = \frac{e^{(\theta_i - \beta_j)x_{ij}}}{1 + e^{\theta_i - \beta_j}}$$

where :

$x_{ij} \in \{0, 1\}$  : the individual response  $i$  ( $1, \dots, K$ ) at item  $j$  ( $1, \dots, J$ );  $\beta_j$  and  $\theta_i$  : item (difficulty) **fixed** parameters and person **random** parameters;  $\theta_1, \theta_2, \dots, \theta_K$  iid follows a normal distribution  $\mathcal{N}(0, \sigma^2)$ .  $\sigma^2$  is the variance of the latent distribution.

Some Rasch model properties

1. Monotonicity of the response probability function is an important property for latent models. It is included in the Rasch model through the logistic link.
2. Mokken model (Molenaar and Sijstma, (1988)) does not assume the logistic link, but assumes a non parametric monotone link function: this is appealing for HRQoL field, but ...
3. Bur relaxing the logisting link, we loose the sufficiency property of the total individual score, which is the most interesting characteristic property of Rasch model in the HRQoL field.
4. Kreiner and Christensen (2002) focus on this sufficiency property and define a new class of non-parametric models: the Graphical Rasch Model
5. local independence (items are independent conditional to the latent) and non differential item functioning (items are independent from external variables conditional to the latent) are additional measurement properties often assumed

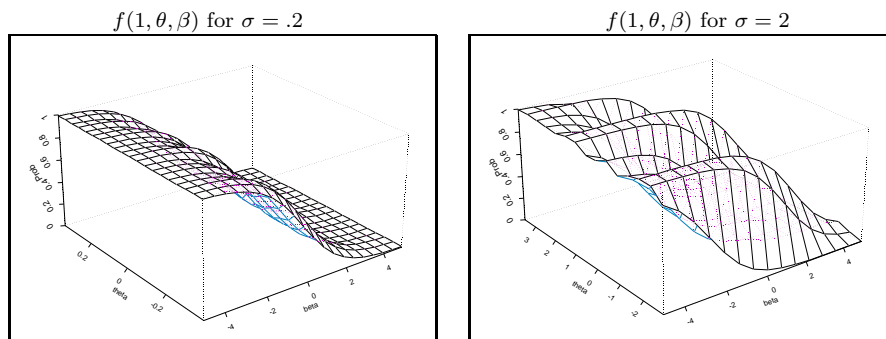


FIG 1. Estimation of  $(\beta, \sigma^2)$  : Likelihood function for  $x = 1$  and two values of  $\sigma$ .

The Likelihood function can be easily derived after marginalizing over the unobserved random parameter, the joint distribution of item responses and the latent variable and using local independence property:  $L(\beta, \sigma^2) =$

$$(2) \quad \frac{1}{(\sqrt{2\pi\sigma^2})^K} \prod_{i=1}^K \left\{ \int_{-\infty}^{+\infty} \prod_{j=1}^J \frac{\exp((\theta - \beta_j) x_{ij})}{1 + \exp(\theta - \beta_j)} \exp\left(\frac{-\theta^2}{2\sigma^2}\right) d\theta \right\}$$

Estimation of  $(\beta, \sigma^2)$  can be obtained by Maximum Likelihood method using Newton-Raphson and numerical integration techniques or EM algorithm followed by Gauss-Hermite quadrature (Hamon and Mesbah (2002), Fisher and Molenaar (1995)). Dorange, Chwalow and Mesbah (2003) or Hardouin J.B. and Mesbah, M. (2006) indicates solutions using SAS NLMixed Procedure. An alternative and faster way is the **Generalized Estimating Equation (GEE)** approach (Feddag, Grama and Mesbah (2003)).

### 3. Generalized Estimating Equations

Let us suppose the following assumptions:

- $Y_i = (y_{ij})_{j=1, \dots, n_i}, \quad X_i = (x_{i1}, \dots, x_{in_i})', \quad (i = 1, \dots, N),$
- Independence between subjects and correlation within subjects
- The joint density of  $Y_i = (y_{i1}, \dots, y_{in_i})'$  is not specified.
- For  $i = 1, \dots, N, j = 1, \dots, n_i, E(y_{ij}) = \mu_{ij} = g(x_{ij}\beta); \beta = (\beta_1, \dots, \beta_p)';$   
 $Var(y_{ij}) = V(\mu_{ij})\phi$

with the main goal, estimation of  $\beta$ . The Likelihood approach is not tractable: the GEE is an alternative approach. The GEE Approach was introduced by Liang and Zeger (1986) as an extension of Quasi Likelihood method to correlated data analysis using Generalized Linear Models (Mac Cullagh and Nelder (1989)). It can be considered as a semiparametric method because it consist on estimation without full specification of the joint distribution: only the marginal distribution at each point or the two first marginal moments are specified. The correlation is treated as nuisance. A Working covariance matrix for the repeated observations is introduced. The GEE approaches are based on first order expansion methods (Taylor series expansion) around  $\theta_i = \hat{\theta}_i$  or around  $\theta_i = 0$ . Taylor series expansion around  $\theta_i = \hat{\theta}_i$  (or Laplace approximation), also known as Partial Quasi Likelihood methods were used by Breslow and Clayton (1993), Breslow and Lin (1995), Vonesh (1996) or Vonesh et al (2002). The implementation is quite easy (GLIMMIX, NLINMIX, NLME) but the estimators are biased. Taylor series expansion around  $\theta_i = 0$  were used by Zeger, Liang and Albert (1988) or Breslow and Clayton (1993). They estimate only the fixed effects by GEE, the variance components were estimated by other methods.

#### 3.1. GEE1 Method

The method consist to estimate the  $\beta$  parameters by the solution of the following equations

$$(3) \quad U_1(\beta, \alpha) = \sum_{i=1}^N D'_{i,11} V_{i,11}^{-1} (y_i - \mu_i) = 0$$

$$\begin{aligned} \mu_i &= E(y_i) = (\mu_{i1}, \dots, \mu_{in_i})', \quad D_{i,11} = \frac{\partial \mu_i}{\partial \beta}, \\ V_{i,11}(\beta, \alpha) &= A_i^{1/2} R_i(\alpha) A_i^{1/2}, \quad A_i = \text{diag} \{V(\mu_{ij})\}_{j=1, \dots, n_i}, \\ R_i(\alpha) &= \text{Working correlation}, \quad \alpha = (\alpha_1, \dots, \alpha_s). \end{aligned}$$

Some known properties of the method are:

- GEE estimators belongs to the family of **M-estimators**,
- $\hat{\beta}$  solution of (3) is consistent and asymptotically normal, even with misspecification of  $R_i$
- Under assumption of independence ( $R_i = \mathbb{I}_{n_i}$ ) : GEE estimators are identical to Maximum Likelihood estimators,
- $\alpha$  is generally estimated by the Pearson residuals
- Several forms of the matrix  $R_i$  and the efficiency depend on the assumed  $R_i$

Prentice and Zhao (1991) generalize GEE to  $\alpha$  in order to improve its efficiency:

$$\begin{aligned} \bullet \quad S'_i &= (S_{i,jl})_{1 \leq j < l \leq n_i}, \quad S_{i,jl} = (y_{ij} - \mu_{ij})(y_{il} - \mu_{il}), \\ \eta_i &= E(S_i), \quad V_{i,22} = \text{Var}(S_i), \quad D_{i,22} = \frac{\partial \eta_i}{\partial \alpha}, \end{aligned}$$

$$(4) \quad U_2(\alpha) = \sum_{i=1}^N D'_{i,22} V_{i,22}^{-1} (S_i - \eta_i) = 0$$

The GEE1 method is known as the method producing estimators solution of equations (3) and (4).

### 3.2. GEE2 Method

The GEE2 method is given by incorporating a dependence between (??) and (5)

$$D_{i,21} = \frac{\partial \eta_i}{\partial \beta}, \quad V_{i,12} = \text{Cov}(Y_i, S_i)$$

$$(5) \quad U(\beta, \sigma^2) = D^t V^{-1} \sum_{i=1}^K \xi_i = 0,$$

where

$$D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}, \quad V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12}^t & V_{22} \end{pmatrix}, \quad \xi_i = \begin{pmatrix} X_i - \mu \\ S_i - \eta \end{pmatrix},$$

$$S'_i = (S_{i,jl})_{1 \leq j < l \leq J}, \quad S_{i,jl} = (X_{ij} - \mu_j)(X_{il} - \mu_l),$$

$$\mu = E(X_i), \quad \eta = E(S_i),$$

$$D_{11} = \frac{\partial \mu}{\partial \beta}, \quad D_{12} = \frac{\partial \mu}{\partial \sigma^2}, \quad D_{21} = \frac{\partial \eta}{\partial \beta}, \quad D_{22} = \frac{\partial \eta}{\partial \sigma^2},$$

$$V_{11} = \text{Var}(X_i) = (\sigma_{jl})_{1 \leq j < l \leq J}, \quad V_{12} = \text{Cov}(X_i, S_i), \quad V_{22} = \text{Var}(S_i).$$

### 3.3. Approximation of the likelihood function

Under the Rasch mixed model, the likelihood function is fully specified. So, there is no need to use a method such GEE devoted to partially specified models. It is, however, well known that a full likelihood analysis for such mixed models is hampered by the need for numerical integrations. To overcome such integration

problems, generalized estimating equations approach is used, following useful approximations. The likelihood as a function of the observed response items can be approximated as:

$$(6) \quad L(\beta, \sigma^2 | x) \simeq \prod_{i=1}^K \prod_{j=1}^J f_{ij}(\beta_j) \left( 1 + \frac{\sigma^2}{2}(A_i^2 - B_i) + \frac{\sigma^4}{8}Q_i \right),$$

with  $f_{ij}(\beta_j) = \exp(-x_{ij}\beta_j - a_{ij})$ ;  $a_{ij} = \ln(1 + e^{-\beta_j})$ ;  
 $A_i = \sum_{j=1}^J (x_{ij} - a_{ij}^{(1)})$ ;  $B_i = \sum_{j=1}^J a_{ij}^{(2)}$ ;  $C_i = \sum_{j=1}^J a_{ij}^{(3)}$ ;  $D_i = \sum_{j=1}^J a_{ij}^{(4)}$ , and  
 $Q_i = A_i^4 - 6A_i^2B_i - 4A_iC_i + 3B_i^2 - D_i$ .

The approximations of the joint moments up to order four of the previous approximate likelihood (6)

$$E(X_{ij}), E(X_{ij}X_{ik}), E(X_{ij}X_{ik}X_{il}), E(X_{ij}X_{ik}X_{il}X_{ih})$$

are easily derived. Then a GEE2 method is applied which lead to estimates which fine asymptotic properties.

- $(\hat{\beta}, \hat{\sigma}^2)$  converge to  $(\beta, \sigma^2)$
- $K^{1/2} \left\{ \begin{pmatrix} \hat{\beta} - \beta \\ \hat{\sigma}^2 - \sigma^2 \end{pmatrix} \right\} \xrightarrow{K \rightarrow \infty} N(0, V)$ ,

where  $\hat{V} = \lim_{K \rightarrow \infty} K \left( \hat{A}_1^{-1} \hat{A}_2 \hat{A}_1^{-1} \right)$ ,

$$\hat{A}_1 = K \left( \hat{D}^t \hat{V}^{-1} \hat{D} \right), \quad \hat{A}_2 = \hat{D}^t \hat{V}^{-1} \left( \sum_{i=1}^K \hat{\xi}_i \hat{\xi}_i^t \right) \hat{V}^{-1} \hat{D}.$$

Computation of  $(\hat{\beta}, \hat{\sigma}^2)$  is given by the iteration formula:

$$\begin{pmatrix} \hat{\beta}^{(j+1)} \\ \hat{\sigma}^{2(j+1)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{(j)} \\ \hat{\sigma}^{2(j)} \end{pmatrix} + \left( K \hat{D}^t \hat{V}^{-1} \hat{D} \right)^{-1} \left( \hat{D}^t \hat{V}^{-1} \sum_{i=1}^K \hat{\xi}_i \right).$$

### 3.4. Simulations results

1000 replications of  $K = 500$  individuals, with fixed number of items,  $J = 4$ , fixed item parameter values  $\beta = (-1, -0.5, 0.5, 1)$  and three successive values of  $\sigma^2 = 0.2, 0.4, 0.9$ , and various models for covariance structure were chosen:  $V_{22} = V_f, V_3, V_2$ ; where

$V_f : V_{22}$  completely specifically

$V_3 : V_{22}$  where  $Cov(S_{i,jl}, S_{i,km}) = 0$ ,

$V_2 : V_{22} = Diag\{Var(S_{i,jl})\}$ ,  $Cov(X_{ij}, S_{i,km}) = 0$

Table 1 : Simulation Results

$\sigma^2$	V	$B\hat{\beta}_1(SE)[PR]$	$B\hat{\beta}_2(SE)[PR]$	$B\hat{\beta}_3(SE)[PR]$	$B\hat{\beta}_4(SE)[PR]$	$B\hat{\sigma}^2(SE)[PR]$
0.2	$V_f$	-.002 (.104)[953]	.001 (.094)[949]	-.001 (.099)[947]	.000 (.106)[959]	.015 (.130)[967]
	$V_3$	-.003 (.106)[956]	.001 (.100)[949]	.004 (.096)[954]	.000 (.103)[951]	.014 (.106)[958]
	$V_2$	-.003 (.106)[956]	.001 (.100)[949]	.004 (.096)[954]	.000 (.103)[951]	.013 (.105)[957]
0.4	$V_f$	.009 (.107)[952]	.010 (.097)[945]	.009 (.098)[950]	.011 (.103)[948]	.043 (.238)[968]
	$V_3$	-.006 (.109)[955]	-.002 (.101)[951]	-.006 (.106)[955]	.000 (.108)[948]	.013 (.140)[951]
	$V_2$	.006 (.110)[955]	.005 (.104)[956]	.003 (.102)[950]	.000 (.109)[950]	.015 (.143)[947]
0.9	$V_f$	.023 (.114)[942]	.024 (.101)[944]	-.001 (.101)[952]	.015 (.109)[945]	.087 (.300)[923]
	$V_3$	-.001 (.118)[950]	.011 (.106)[949]	-.014 (.108)[937]	-.008 (.119)[950]	.097 (.272)[917]
	$V_2$	-.001 (.118)[953]	.010 (.104)[948]	-.013 (.109)[939]	-.009 (.118)[952]	.098 (.250)[927]

The following conclusions can be made about Table 1:

1. The specification of the third and fourth joint moments do not improve the estimates.
2. The bias of the estimators of  $\sigma^2$  tend to be large for large values of  $\sigma^2$
3. Computation less intensive with  $V_2$ .

#### 4. Longitudinal Rasch Model

Let  $X = (X_{ij}^t)_{\substack{i=1,\dots,K \\ j=1,\dots,J \\ t=1,\dots,T}}$ ,  $\underline{X}_i^t = (X_{i1}^t, \dots, X_{iJ}^t)^t$ ,  $\underline{X}_i = (\underline{X}_i^1, \dots, \underline{X}_i^T)^t$ .

With

$$\begin{aligned} P(\underline{X}_i = \underline{x}_i | \theta_i, \beta) &= \prod_{t=1}^T \prod_{j=1}^J P(X_{ij}^t = x_{ij}^t | \theta_{it}, \beta_j), \\ P(X_{ij}^t = x_{ij}^t | \theta_{it}, \beta_j) &= \exp[(\theta_{it} - \beta_j)x_{ij}^t - \ln(1 + \exp(\theta_{it} - \beta_j))], \\ \theta_1, \dots, \theta_K, \text{ iid} &\rightsquigarrow N(0, \Sigma) \text{ with } \Sigma = (\sigma_{jl})_{j,l=1,\dots,T}. \end{aligned}$$

As previously, the Likelihood function can be easily derived:

$$(7) \quad L(\beta, \alpha, \gamma | x) = \prod_{i=1}^K \int_{\mathcal{R}^T} \prod_{t=1}^T \prod_{j=1}^J \frac{\exp((\theta_{it} - \beta_j) x_{ij}^t)}{1 + \exp(\theta_{it} - \beta_j)} \phi(\theta_i, \alpha, \gamma) d\theta_i,$$

where  $\phi(\theta_i, \alpha, \gamma) = \text{d. f. of } N(0, \Sigma)$ ,  $\alpha = (\sigma_{jj})_{j=1,\dots,T}$ ,  $\gamma = (\sigma_{jl})_{1 \leq j < l \leq T}$ .

Approximation of this Likelihood function can also be obtained:

$$(8) \quad L(\beta, \alpha, \gamma | x) \simeq \prod_{i=1}^K \prod_{t=1}^T \prod_{j=1}^J g_{ij}^*(x_{ij}^t; \beta_j) (1 + P_T(\beta, \alpha, \gamma)),$$

where

$$\begin{aligned} P_T(\beta, \alpha, \gamma) &= \int_{\mathcal{R}^T} \sum_{t=1}^T \sum_{\substack{1 \leq t_1 < t_2 < \dots < t_T \leq T \\ \sum t_j = t}} \left( A_i^{t_1} \theta_{it_1} + \frac{1}{2} R_i^{t_1} \theta_{it_1}^2 + \frac{1}{6} P_i^{t_1} \theta_{it_1}^3 \right) \\ &\quad \times \dots \left( A_i^{t_T} \theta_{it_T} + \frac{1}{2} R_i^{t_T} \theta_{it_T}^2 + \frac{1}{6} P_i^{t_T} \theta_{it_T}^3 \right) d\theta_{i1} \dots d\theta_{iT}, \end{aligned}$$

with

$$(9) \quad g_{ij}^*(x_{ij}^t; \beta_j) = \exp(-x_{ij}^t \beta_j - \ln(1 + e^{-\beta_j})),$$

$$(10) \quad R_{i,j}^t = (A_{i,j}^t)^2 - B_{i,j}, \quad P_{i,j}^t = (A_{i,j}^t)^3 - 3A_{i,j}^t B_{i,j} - C_{i,j}, \quad A_{i,j}^t = x_{ij}^t - a_{i,j}^{(1)},$$

and

$$(11) \quad B_{ij} = a_{ij}^{(2)}, \quad C_{i,j} = a_{ij}^{(3)}, \quad A_i^t = \sum_{j=1}^J A_{i,j}^t, \quad R_i^t = \sum_{j=1}^J R_{i,j}^t, \quad P_i^t = \sum_{j=1}^J P_{i,j}^t.$$

Parameter estimations can be performed by resolving:

$$(12) \quad U(\beta, \alpha, \gamma) = D^t V^{-1} \sum_{i=1}^K (\xi_i - E(\xi_i)) = 0,$$

where

$$\xi_i = (\underline{X}_i, \underline{S}_i, \underline{W}_i), \quad S_i^t = \left( S_{i,jl}^t \right)_{1 \leq j < l \leq J}, \quad W_i^{th} = \left( W_{i,jl}^{th} \right)_{1 \leq j < l \leq J},$$

$$S_{i,jl}^t = (X_{ij}^t - \mu_j^t)(X_{il}^t - \mu_l^t), \quad W_{i,jl}^{th} = (X_{ij}^t - \mu_j^t)(X_{il}^h - \mu_l^h),$$

$$D = \frac{\partial E(\xi_i)}{\partial(\beta, \alpha, \gamma)} = (D_{ij})_{i,j=1,2,3}, \quad V = \text{Var}(\xi_i) = (V_{ij})_{i,j=1,2,3}.$$

The following asymptotic properties of the solutions are obtained:

- $(\hat{\beta}, \hat{\alpha}, \hat{\gamma})$  converge to  $(\beta, \alpha, \gamma)$ ,
- $K^{1/2} \left\{ (\hat{\beta} - \beta)^t, (\hat{\alpha} - \alpha)^t, (\hat{\gamma} - \gamma)^t \right\} \xrightarrow{K \rightarrow \infty} N(0, W)$ ,

with

$$\hat{W} = \lim_{K \rightarrow \infty} K \left( \hat{A}_1^{-1} \hat{A}_2 \hat{A}_1^{-1} \right),$$

and

$$\hat{A}_1 = K \left( \hat{D}^t \hat{V}^{-1} \hat{D} \right), \quad \hat{A}_2 = \hat{D}^t \hat{V}^{-1} \left( \sum_{i=1}^K \hat{\xi}_i \hat{\xi}_i^t \right) \hat{V}^{-1} \hat{D}.$$

Conclusion:

- For  $\sigma^2 < 1$ , The estimators obtained are consistent and asymptotically normal.
- The simulation results show that the computation of the third and fourth joint moments are not necessary.
- For more details, see (Feddag and Mesbah (2005)).

## 5. Multivariate Mixed Rasch Model for Quality of life scales

In Health Related Quality of life setting, we often have a questionnaire with  $q$  subscales with each subscale  $l$  consisting in  $J_l$  items ( $l = 1, \dots, q$ ). One of the main interest is estimation of the correlation between the subscales.

The model is

$$Y_i = (y_{ij}^l)_{\substack{j=1, \dots, J_l \\ l=1, \dots, q}}, \quad Y_i^l = (y_{ij}^l)_{j=1, \dots, J_l}, \quad i = 1, \dots, N$$

where  $y_{i1}^1, \dots, y_{iJ_1}^1, y_{i1}^2, \dots, y_{iJ_q}^q \mid \theta_i$  are independent with

$$(13) \quad f(y_{ij}^l \mid \theta_{il}, \beta_j^l) = \frac{e^{(\theta_{il} - \beta_j^l) y_{ij}^l}}{1 + e^{\theta_{il} - \beta_j^l}}, \quad y_{ij}^l \in \{0, 1\},$$

Let  $\theta_i = (\theta_{i1}, \dots, \theta_{iq})'$  the multidimensional latent vector and  $\beta_j^l$  the difficulty parameter of item  $j$  of dimension  $l$ . Suppose  $\theta_1, \dots, \theta_N$  *i.i.d.*  $\rightsquigarrow N_q(0, \Sigma)$

The main interest is to estimate  $(\beta, \Sigma)$ . The Multivariate Mixed Rasch Model is a Generalized Linear Mixed Model **GLMM** with link function  $h(x) = \text{logit}(x)$  and variance function  $v(x) = x(1 - x)$ . The marginal likelihood of  $Y$  is

$$(14) \quad L(\beta, \alpha, \gamma \mid y) \propto \prod_{i=1}^N \int_{\mathbb{R}^q} \prod_{l=1}^q \prod_{j=1}^{J_l} \frac{\exp[(\theta_{il} - \beta_j^l) y_{ij}^l]}{1 + \exp(\theta_{il} - \beta_j^l)} \exp\left(-\frac{1}{2} \theta_i' \Sigma^{-1} \theta_i\right) d\theta_i,$$

With

$$\Sigma = (\sigma_{jl})_{j,l=1,\dots,q}, \quad \alpha = (\sigma_{jj})_{j=1,\dots,q}, \quad \gamma = (\sigma_{jl})_{1 \leq j < l \leq q}$$

Once more, there is no closed form solutions to the marginal likelihood and joint moments cannot be derived analytically, so, as previously approximations of the likelihood are necessary, if we want to use the GEE method instead of the classical maximum likelihood method which is very slow in time computing as  $q$  increase (when  $q > 2$ ). The slowness is due to MCMC techniques (based on Gibbs sampling). Feddag and Mesbah (2005) using taylor series expansion around  $\theta_i = 0$ , generalize Sutradhar and Rao approximations (Sutradhar and Rao (2001)). Under the assumption that  $E(\|\theta_i\|^r) = o(f_r(\Sigma))$  for  $r \geq 6$ , they obtain approximations of the marginal likelihood (??) and of the joint moments up to order four. More precisely:

**i- Approximations of the joint moments:**

$$E(Y_{ij}^l) = \mu_j^l = \frac{1}{1 + e^{\beta_j^l}} + \frac{\sigma_{tt}}{2} \frac{e^{\beta_j^l}(e^{\beta_j^l} - 1)}{(1 + e^{\beta_j^l})^3} + \frac{\sigma_{ll}^2}{8} \frac{e^{\beta_j^l}(e^{3\beta_j^l} - 11e^{2\beta_j^l} + 11e^{\beta_j^l} - 1)}{(1 + e^{\beta_j^l})^5},$$

$$Cov(Y_{ij}^l, Y_{ik}^l) = \frac{\sigma_{ll}}{2} \frac{1}{(1 + e^{\beta_j^l})^2(1 + e^{\beta_k^l})^2} \times \left[ 2e^{\beta_j^l}e^{\beta_k^l} + \sigma_{ll} \frac{(e^{3\beta_j^l} - 4e^{2\beta_j^l} + e^{\beta_j^l})e^{\beta_k^l}}{(1 + e^{\beta_j^l})^2} + \frac{e^{\beta_j^l}(e^{\beta_j^l} - 1)e^{\beta_k^l}(e^{\beta_k^l} - 1)}{(1 + e^{\beta_j^l})(1 + e^{\beta_k^l})} + \frac{e^{\beta_j^l}(e^{3\beta_k^l} - 4e^{2\beta_k^l} + e^{\beta_k^l})}{(1 + e^{\beta_k^l})^2} \right],$$

$$Cov(Y_{ij}^l, Y_{ik}^h) = \frac{1}{2} \frac{\sigma_{lh}}{(1 + e^{\beta_j^l})^2(1 + e^{\beta_k^h})^2} \times \left[ 2e^{\beta_j^l}e^{\beta_k^h} + \sigma_{ll} \frac{(e^{3\beta_j^l} - 4e^{2\beta_j^l} + e^{\beta_j^l})e^{\beta_k^h}}{(1 + e^{\beta_j^l})^2} + \sigma_{lh} \frac{e^{\beta_j^l}(e^{\beta_j^l} - 1)e^{\beta_k^h}(e^{\beta_k^h} - 1)}{(1 + e^{\beta_j^l})(1 + e^{\beta_k^h})} + \sigma_{hh} \frac{e^{\beta_j^l}(e^{3\beta_k^h} - 4e^{2\beta_k^h} + e^{\beta_k^h})}{(1 + e^{\beta_k^h})^2} \right].$$

**ii- Estimating Equations:**

$$U(\beta, \alpha, \gamma) = D'V^{-1} \sum_{i=1}^N (\xi_i - \eta) = 0,$$

$$\xi_i' = (Y_i', S_i', W_i'), \quad \eta = E(\xi_i),$$

$$S_i' = (S_{i,jk}^l)_{\substack{l=1,\dots,q \\ 1 \leq j < k \leq J_l}}, \quad S_{i,jk}^l = (Y_{ij}^l - \mu_j^l)(Y_{ik}^l - \mu_k^l),$$

$$W_i' = (W_{i,jk}^{lh})_{\substack{1 \leq l < h \leq q \\ 1 \leq j < k \leq J_l, 1 \leq k < \leq J_h}}, \quad W_{i,jk}^{lh} = (Y_{ij}^l - \mu_j^l)(Y_{ik}^h - \mu_k^h),$$

$$D = \frac{\partial \eta}{\partial(\beta, \alpha, \gamma)} = \begin{pmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ D_{31} & D_{32} & D_{33} \end{pmatrix}, \quad V = Var(\xi_i) = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_{31} & V_{32} & V_{33} \end{pmatrix}.$$

### iii- Asymptotic properties of the parameter estimators:

Let

$$(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) : U(\hat{\beta}, \hat{\alpha}, \hat{\gamma}) = 0$$

Then:

- $(\hat{\beta}, \hat{\alpha}, \hat{\gamma})$  converge to  $(\beta, \alpha, \gamma)$ ,
- $N^{1/2} \left\{ (\hat{\beta} - \beta)', (\hat{\alpha} - \alpha)', (\hat{\gamma} - \gamma)' \right\} \xrightarrow{N \rightarrow \infty} N(0, W)$ ,

$$\hat{W} = \lim_{N \rightarrow \infty} N \left( \hat{A}_1^{-1} \hat{A}_2 \hat{A}_1^{-1} \right),$$

$$\hat{A}_1 = N \left( \hat{D}' \hat{V}^{-1} \hat{D} \right), \quad \hat{A}_2 = \hat{D}' \hat{V}^{-1} \left( \sum_{i=1}^N (\xi_i - \hat{\eta}) (\xi_i - \hat{\eta})' \right) \hat{V}^{-1} \hat{D}.$$

### iv- Iterative computation formula

$$\begin{aligned} \left( \hat{\beta}'^{[t+1]}, \hat{\alpha}'^{[t+1]}, \hat{\gamma}'^{[t+1]} \right)' &= \left( \hat{\beta}'^{[t]}, \hat{\alpha}'^{[t]}, \hat{\gamma}'^{[t]} \right)' \\ &+ \frac{1}{N} \left( \hat{D}' \hat{V}^{-1} \hat{D} \right)_{[t]}^{-1} \left( \hat{D}' \hat{V}^{-1} \sum_{i=1}^n \hat{\xi}_i \right)_{[t]}. \end{aligned}$$

The obtained estimators are asymptotically unbiased and normal. The GEE approach is computationally less intensive than classical methods. Nevertheless, specification of the third and fourth joint moments do not improve the estimates, and for large  $J_l$ , the method performs well with  $V = \text{diag}(V_{jj})_{j=1,2,3}$  and  $D = \text{diag}(D_{jj})_{j=1,2,3}$ . Finally, performances are less when  $N$  is small.

## 6. Estimation of latent parameter in Rasch model

Estimation of item parameters is generally the main interest in psychometrical area. Calibration of the HRQoL is the preliminary goal. When item parameters are known (or assumed as fixed and known) estimation of the latent parameter is straight forward. One easy method is just to maximize classical joint likelihood method assuming latent parameter fixed. As item parameter is supposed known there is no inconsistency problem. Another popular estimator of latent parameter is the Bayes estimator, given by the posterior mean of the latent distribution (Admane and Mesbah (2006)). Other estimators can be obtained. Mislevy (1984) propose a non-parametric bayesian estimator for latent distribution the Rasch model. Martynov and Mesbah (2006) gives a nonparametric estimator of the latent distribution of a Mixed Rasch Model.

The posterior distribution of the latent parameter is defined as:

$$(15) \quad P(\theta_i/x_i, \beta) = \frac{P(X_i = x_i/\theta_i, \beta) g(\theta_i)}{\int P(X_i = x_i/\theta_i, \beta) g(\theta_i) d\theta_i}$$

The Bayesian modal estimator is  $\hat{\theta}_i$ , the value of  $\theta_i$  which maximize the posterior distribution, while the Bayes estimator is given by:

$$(16) \quad \hat{\theta}_i = \int \theta_i P(\theta_i/x_i, \beta) g(\theta_i) d\theta_i$$

PART B: JOINT ANALYSIS OF A LONGITUDINAL LATENT HRQoL AND A SURVIVAL

7. Joint analysis of a Longitudinal QoL variable and an event time

Motivations of the following models is a HRQoL clinical trial involving analysis of a longitudinal HRQoL variable and an event time. In such clinical trial, the longitudinal HRQoL variable is often **unobserved** at **dropout time**. The model proposed by Dupuy and Mesbah(DM Model) (Dupuy and Mesbah (2002) works when the longitudinal HRQoL is directly observed at each time visit except of course at dropout time. We propose to extent the DM model to the **latent** context case, i.e. when the HRQoL variable is obtained through a questionnaire.

Let  $T$  be a random time to some event of interest, and  $Z$  be the HRQoL longitudinally measured. Let  $C$  be a random right-censoring time. Let  $X = T \wedge C$  and  $\Delta = 1_{\{T \leq C\}}$ . Suppose that  $T$  and  $C$  are independent conditionally on  $Z$

Following, the Cox model, the hazard function of  $T$  has the form

$$(17) \quad \lambda(t|Z) = \lambda(t) \exp(\beta^T Z(t)),$$

The observations are:  $[X_i, \Delta_i, Z_i(u), 0 \leq u \leq X_i]_{1 \leq i \leq n}$ . The unknown parameters are:  $\beta$  and  $\Lambda(t) = \int_0^t \lambda(u) du$ . Let us assume that  $C$  is non informative for  $\beta$  and  $\lambda$ . Dupuy and Mesbah (2002) suggest a method that suppose a non ignorable missing process, take into account the unobserved value of the longitudinal HRQoL variable at dropout time and use a joint modeling approach of event-time and longitudinal variable.

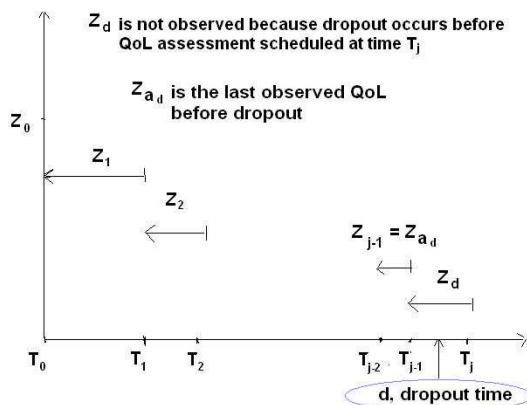


FIG 2. QoL assessments:  $t_0 = 0 < \dots < t_{j-1} < t_j < \dots < \infty$ .  $Z$ : takes value  $Z(t)$   $t$  time  $t$  and constant values  $Z_j$  in the intervals  $(t_{j-1}, t_j]$ .  $Z_j$  is unobserved until  $t_j$ .

Dupuy and Mesbah’s model assume that:

$$(18) \quad \lambda(t|Z) = \lambda(t) \exp(W(t)\beta_0 Z_{a,d} + \beta_1 Z_d)$$

with

- $Z$  has a density satisfying a Markov property:  $f_Z(z_j|z_{j-1}, \dots, z_0; \alpha) = f_Z(z_j|z_{j-1}; \alpha)$ ,  $\alpha \in \mathbb{R}^p$ ,
- $C$  is non informative for  $\alpha$  and does not depend on  $Z(t)$ .

Let  $W(t) = (Z_{a_d}, Z_d)^T$  and  $\beta^T = (\beta_0, \beta_1)$ .

The observations are  $Y_i = (X_i, \Delta_i, Z_{i,0}, \dots, Z_{i,a_d})_{1 \leq i \leq n}$ . The unknown parameters of the model are  $\tau = (\alpha, \beta, \Lambda)$ . There are hidden variables in the model, the **missing values of  $\mathbf{Z}$**  at dropout time,  $Z_{i,a_d}$ . The objective is to estimate  $\tau$  from  $n$  independent vectors of observations  $Y_i$ .

The likelihood for one observation  $y_i$  ( $1 \leq i \leq n$ ) is obtained as:

$$\begin{aligned} L^{(i)}(\tau) &= \int \lambda(x_i)^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \int_0^{x_i} \lambda(u) e^{\beta^T w_i(u)} du \right] \\ &\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_d \\ &= \int l(y_i, z_d, \tau) dz_d, \end{aligned}$$

The parameter  $\tau$  is identifiable. First, suppose that the functional parameter  $\tau$  is a step function  $\Lambda_n(t)$  with jumps at event times  $X_i$  and taking unknown values  $\Lambda_n(X_i) = \Lambda_{n,i}$ , then rewrite the Likelihood and estimate  $\alpha, \beta$  and  $\Lambda_{n,i}$ . The contribution of  $y_i$  to the likelihood obtained is now taken to be:

$$\begin{aligned} L^{(i)}(\tau) &= \int \Delta \Lambda_{n,i}^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \sum_{k=1}^{p(n)} \Delta \Lambda_{n,k} e^{\beta^T w_i(x_k)} 1_{\{x_k \leq x_i\}} \right] \\ &\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_d, \end{aligned}$$

where  $\Delta \Lambda_{n,k} = \Delta \Lambda_n(X_k) = \Lambda_{n,k} - \Lambda_{n,k-1}$ ,  $\Delta \Lambda_{n,1} = \Lambda_{n,1}$  and  $X_1 < \dots < X_{p(n)}$  ( $p(n) \leq n$ ) are the increasingly ordered event times. The maximiser  $\hat{\tau}_n$  of  $\sum_{i=1}^n \log L^{(i)}(\tau)$  over  $\tau \in \Theta_n$  satisfies:

$$\sum_{i=1}^n \frac{\partial}{\partial \tau} \left[ L_{\hat{\tau}_n}^{(i)}(\tau) \right]_{\tau=\hat{\tau}_n} = 0.$$

where  $L_{\hat{\tau}_n}^{(i)}(\tau) = E_{\hat{\tau}_n} [\log l(Y, Z; \tau) | y_i]$ .

Let refer  $\sum_{i=1}^n L_{\hat{\tau}_n}^{(i)}(\tau)$  to as the EM-loglikelihood.

An EM algorithm used to solve the maximization problem is described by Dupuy and Mesbah (*Lifetime Data Analysis*, 2002). A maximiser  $\hat{\tau}_n = (\hat{\alpha}_n, \hat{\beta}_n, \hat{\Lambda}_n)$  of  $\sum_{i=1}^n \log L^{(i)}(\tau)$  over  $\tau \in \Theta_n$  exists and under some additional conditions,

$$(\sqrt{n}(\hat{\alpha}_n - \alpha_t), \sqrt{n}(\hat{\beta}_n - \beta_t), \sqrt{n}(\hat{\Lambda}_n - \Lambda_t)) \sim G,$$

where  $G$  is a tight Gaussian process in  $l^\infty(H)$  with zero mean and a covariance process  $\text{cov}[G(g), G(g^*)]$ .

From this we deduce for instance:

1.  $\sqrt{n}(\hat{\beta}_n - \beta_t)$  converges in distribution to a bivariate normal distribution with mean 0 and variance-covariance matrix  $\Sigma_{\tau_t}^{-1}$ ,
2. consistent estimate of  $\Sigma_{\tau_t}$  is obtained,

and similar results for  $\hat{\alpha}_n$  and  $\hat{\Lambda}_n$  are obtained.

**Example**

Quality of life was assessed among subjects involved in a cancer clinical trial. Quantitative scores were obtained via a HRQoL instrument by auto-evaluation. There was two treatment groups and a nonignorable dropout analysis were performed. Results are indicated in Table 2 below.

Table 2: HRQoL analysis in a cancer clinical trial

	Arm	A	Arm	B	Test	Statistics
	Random	NI	Random	NI	Random	NI
$\hat{\beta}_0$	-0.164	0.128	-0.167	0.089	0.03277	0.34679
SE( $\hat{\beta}_0$ )	0.078	0.081	0.078	0.080	-	-
$\hat{\beta}_1$		-0.362		-0.316	-	-0.37464
SE( $\hat{\beta}_1$ )		0.086		0.087	-	-
$\hat{\alpha}$	0.959	0.952	0.955	0.948	0.35086	0.32268
SE( $\hat{\alpha}$ )	0.008	0.009	0.009	0.010	-	-
$\hat{\sigma}_e^2$	0.704	0.710	0.571	0.576	2.19565	2.18126
SE( $\hat{\sigma}_e^2$ )	0.046	0.047	0.039	0.0401	-	-
loglikelihood	-963.928	-896.4	-927.2	-857.2	-	-

In this example HRQoL, excepted for its value at dropout time, was just considered as an observed continuous score,  $Z$ . But in fact, HRQoL is an unobserved latent variable. In practice, HRQoL data consist always in a multidimensional binary or categorical observed variable named Quality of Life Scale used to measure the true unobserved latent variable HRQoL. From the Quality of Life Scale, we can derive HRQoL scores, i.e. Statistics. These scores surrogate of the true unobserved latent variable HRQoL. In the next section, we will extent the previous DM model to the latent case.

**8. Joint analysis of a latent Longitudinal HRQoL variable and an event time**

When the HRQoL variable  $z$  was observed (excepted for the last unobserved dropout value  $z_d$ ), the likelihood for one observation  $y_i (1 \leq i \leq n)$  was:

$$\begin{aligned}
 L^{(i)}(\tau) &= \int \lambda(x_i)^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \int_0^{x_i} \lambda(u) e^{\beta^T w_i(u)} du \right] \\
 &\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_d \\
 &= \int l(y_i, z_{i_d}, \tau) dz_{i_d},
 \end{aligned}$$

where

$$\begin{aligned}
 y_i &= (x_i, \delta_i, z_{i_0}, \dots, z_{i_{a_d}}) \\
 &= (x_i, \delta_i, z_{i_{obs}})
 \end{aligned}$$

and, all the previous statistical inference, based on the likelihood

$$(19) \quad L^{(i)}(\tau) = \int l(x_i, \delta_i, z_{i_{obs}}, z_d, \tau) dz_d$$

is highly validated by theoretical asymptotic results and well working computer algorithms. In the latent variable context,  $z_{i_{obs}}$  is in fact not directly observed. The  $k$  item responses  $Q_{ij}$  of a subject  $i$  (response or raw vector  $Q_i$ ) are observed and must be used to recover the latent HRQoL values  $z_i$  through a measurement model. The obvious choice in our context is the **Rasch model**, which is for binary responses:

$$(20) \quad P(Q_{ij} = q_{ij} \mid z_i, \zeta_j) = f(q_{ij}, z_i, \zeta_j) = \frac{e^{(z_i - \zeta_j)q_{ij}}}{1 + e^{z_i - \zeta_j}}$$

So, currently, observations are  $Y_i = (X_i, \Delta_i, Q_{i_0}, \dots, Q_{i_{a_D}})_{1 \leq i \leq n}$ ; with  $Q_i = (Q_{i_1}, \dots, Q_{i_p})$  for an unidimensional scale of  $p$  items. Unknown parameters of the model are  $\tau = (\alpha, \beta, \Lambda)$  and nuisance parameters,  $\zeta$ . The objective is now to estimate  $\tau$  from  $n$  independent vectors of observations  $Y_i$ . Let us suppose the following two assumptions hold:

1. The DM analysis Model hold for the true unobserved QoL  $Z$  and Dropout  $D$  or Survival  $T$
2. The Rasch measurement model relate the observed response items  $Q$  to QoL  $Z$

First, we have two main issues:

- Specification of a model for the data and the true Latent QoL
- Choice of a Method of estimation

1) The statistics:  $S_i = \sum_{j=1}^p Q_{ij}$ , i.e., the total person score (or the raw score) is a sufficient statistic for the person parameter  $z_i$  :

$$(21) \quad Prob(Q_{i1} = q_{i1}, \dots, Q_{ip} = q_{ip}) \mid S_i = s_i$$

does not depend on  $z_i$ . This is a characteristic property of Rasch Model, very nice and useful in HRQoL

2) Local independence is a classical assumption in latent variable models:

$$Prob(Q_{i1} = q_{i1}, \dots, Q_{ip} = q_{ip} \mid z_i) = Prob(Q_{i1} = q_{i1} \mid z_i) \times \dots \times Prob(Q_{ip} = q_{ip} \mid z_i)$$

Items are independent conditionally to the true latent value.

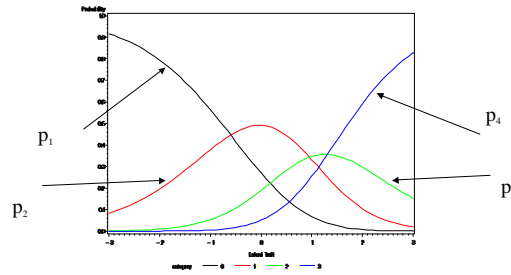


FIG 3. *Partial Credit Model*

Similar to **Rasch model**, for categorical ordinal responses (with number of levels  $m_j$  different per item), the Partial Credit model:

$$(22) \quad p_c = P(Q_{ij} = c \mid z_i, \zeta_j) = \frac{e^{(cz_i - \sum_{l=1}^c \zeta_{jl})}}{\sum_{c=0}^{m_j} e^{(cz_i - \sum_{l=1}^c \zeta_{jl})}}$$

The joint distribution of Q (items), Z (latent), D(Time to death or dropout) and T (treatment) can be derived, using only the conditional independence property:

$$(23) \quad f(Q, Z, D, T/Z) = \frac{f(Q, Z, D, T)}{f(Z)} = \frac{f(Q, Z)}{f(Z)} \times \frac{f(Z, D, T)}{f(Z)}$$

so, we have:

$$(24) \quad f(Q, Z, D, T/Z) = f(Q/Z) \times f(D, T/Z)$$

Then, without any other assumption, we can specify two models:

- First model:

$$(25) \quad f(Q, Z, D, T) = f(Q/Z) \times f(D/Z, T) \times f(Z/T) \times f(T)$$

- Second model:

$$(26) \quad f(Q, Z, D, T) = f(Q/Z) \times f(Z/D, T) \times f(D/T) \times f(T)$$

The right likelihood must be based on the probability function of the observations, i.e., currently,  $Y_i = (X_i, \Delta_i, Q_{i_0}, \dots, Q_{i_{a_D}})_{1 \leq i \leq n}$ . The parameters of the model are  $\tau = (\alpha, \beta, \Lambda)$  and the nuisance difficulty parameters of the HRQoL questionnaire,  $\zeta$ . There are non observed (hidden) variables in the model (latent Z, missing Q):  $(Z_{i_0}, \dots, Z_{i_{a_D}}, Z_{i_d}, Q_d)_{1 \leq i \leq n}$ .

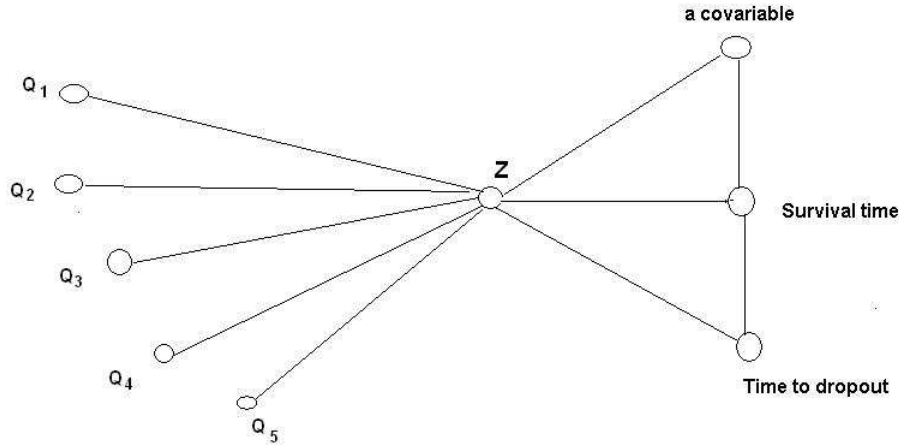


FIG 4. Non Differential Functioning of Items

Straightly followed from the graph of the DMq model, factorization rules of the joint distribution function of the observations ( $Y_i$ ), the latent HRQoL ( $Z$ ) and the missing questionnaire  $Q_d$  can now be specified, and then, integrating through the hidden variables, one gets the likelihood:

$$\begin{aligned}
L^{(i)}(\tau) &= \int \prod_{j=1}^p \left[ \frac{e^{(cz_{i_0} - \sum_{l=1}^c \zeta_{jl})}}{\sum_{h=0}^{m_j} e^{(cz_{i_0} - \sum_{l=1}^c \zeta_{jl})}} \right. \\
&\times \dots \times \\
&\times \frac{e^{(cz_{i_{a_d}} - \sum_{l=1}^c \zeta_{jl})}}{\sum_{h=0}^{m_j} e^{(cz_{i_{a_d}} - \sum_{l=1}^c \zeta_{jl})}} \\
&\times \left. \sum_{c=1}^p \frac{e^{(cz_{i_d} - \sum_{l=1}^c \zeta_{jl})}}{\sum_{h=0}^{m_j} e^{(cz_{i_d} - \sum_{l=1}^c \zeta_{jl})}} \right] \\
&\times \Delta \Lambda_{n,i}^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \sum_{k=1}^{p(n)} \Delta \Lambda_{n,k} e^{\beta^T w_i(x_k)} \mathbf{1}_{\{x_k \leq x_i\}} \right] \\
&\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_{i_0}, \dots, z_{i_{a_d}}, z_d
\end{aligned}$$

a slightly different likelihood, integrating the total Score can be derived:

$$\begin{aligned}
L^{(i)}(\tau) &= \int \prod_{j=1}^p [P(Q_{i_0,j} = c | S_{i_0}, \zeta_j) P(S_{i_0} = s_{i_0} | z_{i_0}, \zeta_j) \\
&\times \dots \times \\
&\times P(Q_{i_{a_d},j} = c | S_{i_{a_d}}, \zeta_j) P(S_{i_{a_d}} = s_{i_{a_d}} | z_{i_{a_d}}, \zeta_j) \\
&\times \sum_{c=1}^p P(Q_{i_d,j} = c | S_{i_d}, \zeta_j) P(S_{i_d} = s_{i_d} | z_{i_d}, \zeta_j)] \\
&\times \Delta \Lambda_{n,i}^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \sum_{k=1}^{p(n)} \Delta \Lambda_{n,k} e^{\beta^T w_i(x_k)} \mathbf{1}_{\{x_k \leq x_i\}} \right] \\
&\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_{i_0}, \dots, z_{i_{a_d}}, z_d
\end{aligned}$$

and using the sufficiency property, we get the likelihood:

$$\begin{aligned}
L^{(i)}(\tau) &= \int P(S_{i_0} = s_{i_0} | z_{i_0}) \times \dots \times P(S_{i_{a_d}} = s_{i_{a_d}} | z_{i_{a_d}}) P(S_{i_d} = s_{i_d} | z_{i_d}) \\
&\times \prod_{j=1}^p [P(Q_{i_0,j} = c | S_{i_0}, \zeta_j) \\
&\times \dots \times \\
&\times P(Q_{i_{a_d},j} = c | S_{i_{a_d}}, \zeta_j) \\
&\times \sum_{c=1}^p P(Q_{i_d,j} = c | S_{i_d}, \zeta_j)] \\
&\times \Delta \Lambda_{n,i}^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \sum_{k=1}^{p(n)} \Delta \Lambda_{n,k} e^{\beta^T w_i(x_k)} \mathbf{1}_{\{x_k \leq x_i\}} \right] \\
&\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_{i_0}, \dots, z_{i_{a_d}}, z_d
\end{aligned}$$

and:

$$\begin{aligned}
 L^{(i)}(\tau) &= \prod_{j=1}^p [P(Q_{i_0,j} = c \mid S_{i_0}, \zeta_j)] \\
 &\times \dots \\
 &\times P(Q_{i_{a_d},j} = c \mid S_{i_{a_d}}, \zeta_j) \\
 &\times \sum_{c=1}^p P(Q_{i_d,j} = c \mid S_{i_d}, \zeta_j) \\
 &\times \int P(S_{i_0} = s_{i_0} \mid z_{i_0}) \times \dots \times P(S_{i_{a_d}} = s_{i_{a_d}} \mid z_{i_{a_d}}) P(S_{i_d} = s_{i_d} \mid z_{i_d}) \\
 &\times \Delta \Lambda_{n,i}^{\delta_i} \exp \left[ \delta_i \beta^T w_i(x_i) - \sum_{k=1}^{p(n)} \Delta \Lambda_{n,k} e^{\beta^T w_i(x_k)} 1_{\{x_k \leq x_i\}} \right] \\
 &\times f(z_{i_0}, \dots, z_{i_{a_d}}, z_d; \alpha) dz_{i_0}, \dots, z_{i_{a_d}}, z_{i_d}
 \end{aligned}$$

## 9. Conclusion

article J.B. Hardouin et M.Mesbah : ANAQOL,

The definition (or construction) of variables and indicators, and the analysis of the evolution of their joint distribution between various populations, times and areas are generally two different, well separated steps of the work for a statistician in the field of Health Related Quality of Life. The first step generally deals with calibration and metrology of variables. Key words are measurement or scoring, depending on the area of application. Most of the time, the statistical methods used are exploratory. The kinds of models specified are generally structural: classical factorial analysis models or modern item response theory models. The second step is certainly more known by most inferential statisticians. Linear, generalized linear, time series and survival models are very useful models in this step, where the variables constructed in the first step are incorporated and their joint distribution with the other analysis variables (treatment group, time, duration of life, etc ...) is investigated. Mesbah (2004) compared the simple strategy of separating the two steps with the global strategy of defining and analyzing a global model including both the measurement and the analysis step. If, with a real data set, one find a significant association between a built (from items) score and an external covariate, then the true association, i.e., the one between that external covariate and the true latent, is probably larger. So, if the scientific goal is to show an association between the true and the covariate, one don't need to use a global model: just use the model with the surrogate built score instead of the true latent. Conclusions with the built score also stand for the true. But, one gets no significant association between built score and the covariate, then the true association could be anything (and perhaps larger ... ). So, one have to consider a global model, even if one don't need to build new scores or to valid the measurement model. Building a global model taking into account the latent trait parameter in a one step way, i.e. without separation between measurement and analysis is a promising latent regression approach (Christensen et al (2004), Sebille et Mesbah (2005)) allowed by the increasing performance of computers. In HRQoL field, most of papers are devoted to two steps approach, where the HRQoL scores are used instead of the original item responses data.

Joint analysis of a longitudinal variable and an events time is nowadays a very active fields. Vonesh, Greene and Schluchter (2006), Cowling (2006), Chi and Ibrahim (2006) are few recent papers indicating that " *Joint modeling of longitudinal and survival data is becoming increasingly essential in most cancer and AIDS clinical trials.*" Mainly due to the complexity of the computing programs, there is unfortunately no papers considering a joint model between a longitudinal **latent** trait and an event time..

Another very popular method used in the nineteenth is the Q-TWIST (**Q**uality adjusted **T**ime **W**ithout **S**ymptoms of **T**oxicity) approach(Gelber et al (1996)), where duration of life was just divided in different categories corresponding to various state of health with given utilities. So, it was a weighted (by utility weights or Quality of Life weights) survival analysis... It was a two step approach, but main criticisms comes more about the fact that used utility values, were, in practice with very poor measurement properties.

Our approach can be considered as in the framework of mixed models with a clear interpretation of the random factor by a latent trait previously validated in an measurement step. Items are repeated measurement of such true latent trait. Computer programs are nowadays available even in general softwares (Hardouin et Mesbah (2007)) which allows building and estimating models with nonlinear random effects models.

## 10. References

1. ADMANE, O. and MESBAH, M. (2006) "Estimation des paramètres du modèle de Rasch dichotomique". Annales de L'ISUP. 2006.
2. AWAD, L., ZUBER,E. and MESBAH,M. (2002) Applying survival data methodology to analyze longitudinal Quality of Life Data, in "Statistical Methods for Quality of Life Studies, Design, Measurement and Analysis". Editeurs: Mesbah, M., Cole, B.F., Lee, M.L.T. Kluwer Academic Publishing, Boston. pp 231-243.
3. BRESLOW, N. E. and CLAYTON, N. E. (1993), Approximate Inference in Generalized Linear Mixed Models, Journal of the American Statistical Association 88 (1993), 9-25.
4. BRESLOW, N. E. and LIN, N. (1995). Bias Correction in Generalized Linear Mixed Models with a Single Components of Dispersion, Biometrika 82 (1995), 81-91.
5. CHI, Y.-Y., IBRAHIM, J.G. (2006) Joint Models for Multivariate Longitudinal and Multivariate Survival Data. Biometrics 62, 432-445.
6. CHRISTENSEN, K.B., BJORNER, J.B., KREINER, S., PETERSEN, J.H. (2004). Latent regression in loglinear Rasch Models. Communication in Statistics. Theory and Methods. 33 (6), pp. 1295-1313.
7. COWLING, B.J., HUTTON, J.L. and SHAW, J. E. H. (2006) Joint modelling of event counts and survival times. Appl. Statist. 55, Part 1, pp. 31-39
8. DUPUY, J.F. GRAMA,I. and Mesbah, M. (2006). Asymptotic Theory for the Cox Model With Missing Time Dependent Covariate. Annals of Statistics. Vol. 34, N°2, April 2006
9. DUPUY, J.-F. and MESBAH, M. (2002). Joint modeling of event time and nonignorable missing longitudinal data. *Lifetime Data Analysis* 8 99-115.
10. DORANGE, C., CHWALOW, J. and MESBAH, M. (2003.) Analyzing Quality of Life data with the ordinal Rasch model and NLMixed SAS procedure. *In Proceedings of the International conference on Advance in Statistical Inferential Methods ASIM2003*. Editeur: KIMEP Ed, Almaty. ISBN: 9965-07-253-1. pp 41-73.

11. FEDDAG, M. L. and GRAMA, I. and MESBAH, M. (2003). *GEE to the Logistic Mixed Models* (Comm. In Sta. -Theory and Methods, V. 32, 851-874)
12. FEDDAG, M. L. and MESBAH, M. (2005). *Generalized Estimating Equations for longitudinal mixed Rasch model* Journal of Statistical Planning and Inference (JSPI), 129. 159-179.
13. FISHER, G.H. and MOLENAAR, I.W. (1995) *Rasch models, Foundations, recent Developments and Applications*, Springer-Verlag, New-York.
14. GELBER, R.D., GOLDBIRSH, A., COLE, B.F., WIEAND, H.S., SCHROEDER, G. and KROOK, G.E. (1996). A quality-adjusted time without symptoms or toxicity (Q-TWIST) analysis of adjuvant radiation therapy and chemotherapy for resectable rectal cancer. J. Natl. Cancer Inst., 88: 1039-45
15. HAMON, A. and MESBAH, M. (2002) Questionnaire Reliability under the Rasch Model, in "Statistical Methods for Quality of Life Studies. Design, Measurement and Analysis". Editeurs: Mesbah, M., Cole, B.F., Lee, M.L.T. Kluwer Academic Publishing, Boston. pp 155-168.
16. HARDOUIN, J.-B. and MESBAH, M. (2007) The SAS Macro-Program %ANAQOL to Estimate the Parameters of Item Responses Theory Models. In Press. Communication in Statistics. Simulation and Computation. 36 (2), 2007
17. KALBFLEISCH, J. D. and PRENTICE, R. L. (1980) *The Statistical Analysis of Failure Time Data*. Wiley: New-York, 1980.
18. KREINER, S. and Christensen, K.B. (2002) Graphical Rasch Models, in "Statistical Methods for Quality of Life Studies. Design, Measurement and Analysis". Editeurs: Mesbah, M., Cole, B.F., Lee, M.L.T. Kluwer Academic Publishing, Boston. pp 155-168.
19. LIANG, K. Y. and ZEGER, S. L. (1986). *Longitudinal Data Analysis Using Generalised Linear Models*. Biometrika, 73, 1, pp. 121-130.
20. MAC CULLAGH, P. and NELDER, J. *Generalized Linear Models*. Chapman and Hall, London (1989)
21. MASTERS, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.
22. MARTYNOV, G. and MESBAH, M. (2006) Goodness of Fit Test and Latent Distribution Estimation in the Mixed Rasch Model. Communication in Statistics. Theory and Methods. 35 (5), pp., 2006
23. MESBAH, M. (2004). Measurement and Analysis of Health Related Quality of Life and Environmental Data. Environmetrics. 2004; 15: 471-481
24. MESBAH, M., DUPUY, J. F., HEUTTE, N. and AWAD., L. (2004) Joint analysis of longitudinal quality of life and survival processes. in "Handbook of Statistics. Vol22, Advances in Survival Analysis." Editors : N. Balakrishnan et C.R.Rao. North Holland; Amsterdam 2004.
25. MISLEVY, R.J. (1984). Estimating Latent Distribution. Psychometrika-VOL. 49, NO. 3, 359-381
26. MOLENAAR, I. W. and SIJSTMA, K. (1988). Mokken's approach to reliability estimation extended to multicategory items. Kwantitatieve Methoden 9(28), 115-126.
27. PRENTICE, R. L. and ZHAO, L. P. (1991). Estimating Equation for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses, Biometrics 47 (1991), 825-839.
28. RASCH, G. (1960). *Probabilistic models for some intelligence and attainment tests*, Danmarks Paedagogiske Institut, Copenhagen.

29. SÉBILLE, V. and MESBAH, M. (2005). Sequential Analysis of Quality of Life Rasch Measurements. In "Probability Statistics and Modelling in Public Health" in honor of Marvin Zelen, Editors: Nikouline, M., Commenges, D. and Huber, C. Springer, New York, 2005. ISBN 0387260226. pp 421-439.
30. SUTRADHAR, B. C. and RAO, R. P. (2001). *On Marginal Quasi-Likelihood Inference in GLMM*. Journal of Multivariate Analysis, 76, pp. 1-34.
31. VONESH E. F. (1996). A note on the Use of Laplace's Approximation for Nonlinear Mixed-Effects Models, Biometrika 83 (1996), 447-452. [91] R. W. M.
32. VONESH, E.F., WANG, H., NIE L., MAJUMDAR (2002), Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. Journal of the American Statistical Association (2002), [97] pp 271-283.
33. VONESH E.F., GREENE, T. and SCHLUCHTER, M.D. (2006) Shared parameter models for the joint analysis of longitudinal data and event times. Statist. Med. 2006; 25:143-163.
34. ZEGER, S. L., LIANG, K. Y. and ALBERT, P. S (1988). A Generalized Estimating Equation Approach, Biometrics 44 (1988), 1049-1060.