A flexible IRT Model for health questionnaire: an application to HRQoL

Serena Broccoli
Giulia Cavrini

Department of Statistical Science, University of Bologna

19th International Conference on Computational Statistics
Paris - August 22-27, 2010
Statement of the problem

* Scalone et al, Value in Health, 2007
What you NEED

- To obtain a synthetic score of Quality of Life
- To relax the hypothesis of normal distributed latent variable
- To treat mixed responses
What you HAVE

Partial Credit Model
(Masters, 1982)

“A Skew item response model”
(Bazan, 2006)

“Latent Trait Model for mixed manifest variables”
(Moustaki, 1996)
A flexible IRT Model

- $r$ \rightarrow \text{continuous items}
- $s$ \rightarrow \text{dichotomous items}
- $c$ \rightarrow \text{ordered polytomous items}
- $s + r + c = q$ \rightarrow \text{total number of items}

Letting

- $w_{ij}$ with $j=1...r$ be the answer of subject $i$ to the continuous item $j$
- $v_{ij}$ with $j=r+1 \ldots r+s$ be the answer of subject $i$ to the dichotomous item $j$
- $t_{ij}$ with $j=r+s+1 \ldots r+s+c$ be the answer of subject $i$ to the ordered polytomous item $j$
Assumptions and constrains

- Items are independent conditionally on $\theta$
- $\theta_i \approx SN(\alpha, \beta, \delta) \quad i = 1...n$ (Azzalini, 1985)
- $E(\theta) = 0$ and $\text{Var}(\theta) = 1$

**SKEW NORMAL CENTERED PARAMETERIZED (SN_{cp})**

Given $Z \sim SN(0,1,\delta)$

\[
\theta \mid Z = \frac{Z - \left(\frac{2}{\pi}\right)^{1/2} \delta}{\sqrt{1 - \frac{2}{\pi} \delta^2}} \sim SN_{cp}(0,1,\delta)
\]

$\delta = -0.98$
A flexible IRT model

- The conditional joint density function $g(y_i | \theta)$ of the observed variables
  
  $y_i = \begin{pmatrix} w_i \\ v_i \\ t_i \end{pmatrix}$

  is

  $g(y_i | \theta_i) = \prod_{j=1}^{q} g(y_{ij} | \theta_i) = \prod_{j=1}^{r} h(w_{ij} | \theta_i) \prod_{j=r+1}^{r+s} k(v_{ij} | \theta_i) \prod_{j=r+s+1}^{r+s+c} l(t_{ij} | \theta_i)$

  where

  - $h(.)$ is the Normal density function of mean $\theta_i - b_j$ and variance $\sigma_j^2$
  - $k(.)$ is the Bernoulli probability function of parameter $\mu_{ij} = r^{-1}(\theta_i - b_j)$ and $r(.) = \text{logit link}$
  - $l(.)$ is the Multinomial probability function of parameters “PCM” and 1
The probability of subject $i$ scoring $x$ to item $j$ (item with $k_j+1$ levels of answer), given the latent variable $\theta$ is

$$p_{ijx}(\theta_i) = \frac{\exp \left( \sum_{t=1}^{x} (\theta_i - b_{jt}) \right)}{1 + \sum_{k=1}^{k_j} \exp \left( \sum_{t=1}^{k} (\theta_i - b_{jt}) \right)}, \text{ for } x = 1 \ldots k_j$$

$$p_{ijx}(\theta_i) = \frac{1}{1 + \sum_{k=1}^{k_j} \exp \left( \sum_{t=1}^{k} (\theta_i - b_{jt}) \right)}, \text{ for } x = 0$$
A flexible IRT model

- The log-likelihood for a random sample of $n$ individuals can be expressed as

$$\log L = \sum_{i=1}^{n} \log f(y_i) = \sum_{i=1}^{n} \log \int_{-\infty}^{+\infty} g(y_i \mid \theta) h(\theta) d\theta$$

where $h(\theta)$ is now the Skew Normal distribution function of mean 0 and variance 1.
Bayesian estimation of the parameters

- Joint Posterior distribution of the parameters of the model

\[
p(b_j, \theta_i \mid y_i) \propto \prod_{j=1}^{r} g(w_{ij} \mid b_j, \theta_i) h(\sigma_j) \prod_{j=r+1}^{r+s} g(v_{ij} \mid b_j, \theta_i) \prod_{j=r+s+1}^{r+s+c} g(t_{ij} \mid b_j, \theta_i) \prod_{j=1}^{q} h(b_j) h(\theta_i \mid \delta) r(\delta)
\]

where

\[
\theta_i \sim SN_{cp}(0,1,\delta) \\
b_j \sim N(0,100) \\
\sigma_j \sim \text{invgamma}(10,10) \text{  } j=1...r \\
\delta \sim U(-1,1)
\]
Bayesian estimation of the parameters

- Bayesian parameter estimates were obtained using Gibbs sampling algorithms as implemented in the computer program WinBUGS 1.4 (Spiegelhalter, Thomas, Best, & Lunn, 2003).

- The value taken as the MCMC estimate is the mean over iterations sampled starting with the first iteration following burn-in.

- The R-Package CODA (Best, Cowles, & Vines, 1995) was used to compute convergence Geweke’s diagnostic.
Results

- **Model 1**: Partial Credit Model
  - 10,000 iterations with the first 3,000 as burn-in

- **Model 2**: IRT model for mixed responses
  - 25,000 iterations with the first 10,000 as burn-in

- **Model 3**: IRT for mixed responses and skew latent variable
  - 15,000 iterations with the first 5,000 as burn-in
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>181.1</td>
</tr>
<tr>
<td>2</td>
<td>133.3</td>
</tr>
<tr>
<td>3</td>
<td>129.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HRQol</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>[43] 11111 VAS=100</td>
<td>Posterior mean</td>
<td>SD</td>
<td>MC error</td>
</tr>
<tr>
<td></td>
<td>1.162</td>
<td>1.347</td>
<td>0.028</td>
</tr>
<tr>
<td>[1]   11111 VAS=85</td>
<td>1.147</td>
<td>1.370</td>
<td>0.027</td>
</tr>
<tr>
<td>[6]   11111 VAS=50</td>
<td>1.164</td>
<td>1.369</td>
<td>0.025</td>
</tr>
<tr>
<td>[29]  31122 VAS=65</td>
<td>-2.752</td>
<td>1.176</td>
<td>0.041</td>
</tr>
</tbody>
</table>

**Note:** VAS = Visual Analog Scale
The procedure had a run length of 15,000 iterations with a burn-in period of 8,000 iterations. Every three states of the chain were included in the posterior estimates, to avoid autocorrelation.
A flexible IRT Model for health questionnaire: an application to HRQoL – S. Broccoli & G. Cavrini

Results
Results

- The HRQoL mean value is 0.06 (s.d. 0.80)
- The maximum value is 1.065 and the minimum is -4.25
- The right-skewed shape of the histogram is expected, as well as the mean centered on 0.
Some limits

- Long computational times
- Not user-friendly software

Further developments

- Generalized Partial Credit model
- Covariates
References