36-720: The Rasch Model

Brian Junker

October 15, 2007

- Multivariate Binary Response Data
- Rasch Model
- Rasch Marginal Likelihood as a GLMM
- Rasch Marginal Likelihood as a Log-Linear Model
- Example

For more reading, see:

- Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* University of Chicago.
- DeBoeck, P. & Wilson, M. (2004). *Explanatory Item Response Models.* NY: Springer.
- van der Linden R. J. & Hambleton, R. K. (1997). *Handbook of Modern Item Response Theory.* NY: Springer.

Multivariate Binary Response Data

Ubiquitous in

- Education (standardized testing);
- Psychology (positive and negative responses to stimuli);
- Social Science & Marketing (opinion/attitude/preference data);
- and other areas.

For specificity, we use the language of educational testing:

For student i and question j on a particular exam, define

$$y_{ij} = \begin{cases} 1, & \text{if student } i \text{ got question } j \text{ correct} \\ 0, & \text{else} \end{cases}$$

say, for $i = 1, \ldots, N$ students and $j = 1, \ldots, J$ questions.

Viewing the data as a contingency table

- For a test of J questions, we construct a J-way table, with each dimension of the table corresponding to a single question, with two levels (0 = wrong; 1 = right):

$$\{n_{\underline{y}} : \text{ as } \underline{y} \text{ ranges over all } 2^J \text{ possible patterns } (y_1, \ldots, y_J)\}$$

is a 2^J table (J-way table with two levels each "way").

- Even if $N = \sum_{\underline{y}} n_{\underline{y}}$ is large, the 2^J table quickly becomes sparse: for example, with $N = 100$ and only $J = 8$ questions, there *must* be over 100 sampling zeros in the table *(why??)*.

- Thus, the usual hierarchical log-linear models for the 2^J table won't be of much use, because sampling zeros will frustrate many model fit and model comparison efforts.

However, there are log-linear models that are useful with $\{n_{\underline{y}}\}$ *and we will return to this representation later.*

Viewing the data as two-way ANOVA data

Instead of considering the table of counts $n_{\underline{y}}$ we may consider the rectangular array

$$\mathcal{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1J} \\ y_{21} & y_{22} & \cdots & y_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ y_{N1} & y_{N2} & \cdots & y_{NJ} \end{bmatrix}$$

- The i^{th} *row* corresponds to the correct & incorrect answers given by examinee i to all J questions, and
- The j^{th} *column* corresponds to the correct & incorrect answers given by all N examinees to the j^{th} question.

A logit analogue to the two-way additive ANOVA model for this array would be

$$\log \frac{p_{ij}}{1 - p_{ij}} = \theta_i - \beta_j \qquad (1)$$

where $p_{ij} = P[y_{ij} = 1 \mid \theta_i, \beta_j]$. θ_i is the row effect and β_j is the column effect.

Rasch Model

In the model $\log \frac{p_{ij}}{1-p_{ij}} = \theta_i - \beta_j$,

- As θ_i increases so does p_{ij}: θ_i represents examinee i's *proficiency*, regardless of question.
- As β_j increases, p_{ij} decreases: β_j represents the question's *difficulty*[a].

The model in (1) is called the *Rasch Model* (after Rasch's 1960 monograph); in logistic form it is written

$$p_{ij} = P[y_{ij} = 1 \mid \theta_i, \beta_j] = \frac{\exp\{\theta_i - \beta_j\}}{1 + \exp\{\theta_i - \beta_j\}} \qquad (2)$$

and is an example of an *item response theory (IRT)* model. ("item" = "survey or test question").

[a]The choice of sign here, i.e. $\theta_i - \beta_j$ instead of $\theta_i + \beta_j$, is just a convention, but leads to this nice interpretation for β_j.

The likelihood for the i^{th} examinee is a product of Bernoulli likelihoods for each y_{ij}:

$$P[y_{i1},\ldots,y_{iJ} \mid \theta_i;\beta_1,\ldots,\beta_J] = \prod_{j=1}^{J} p_{ij}^{y_{ij}}(1-p_{ij})^{y_{ij}} = \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\theta_i-\beta_j)\}}{1+\exp\{\theta_i-\beta_j\}} \tag{3}$$

We could formulate a joint likelihood for all examinees (and hence the entire array $\mathcal{Y}$ above) as

$$P[\mathcal{Y} \mid \theta_1,\ldots,\theta_N;\beta_1,\ldots,\beta_J] = \prod_{i=1}^{N}\prod_{j=1}^{J} \frac{\exp\{y_{ij}(\theta_i-\beta_j)\}}{1+\exp\{\theta_i-\beta_j\}} \tag{4}$$

and maximize over θ's and β's but it is well-known[a] that this will result in inconsistent estimates as N increases, since the number of θ_i parameters also increases.

[a]E.g. Haberman, S.J. (1977). Maximum likelihood estimates in exponential response models., *The Annals of Statistics*, *5*, 815–841.

Rasch Marginal Likelihood as a GLMM

A way around this is to think of θ_i as a random effect, so that the likelihood for one examinee is really a mixture over the random effect,

$$P[y_{i1}, \ldots, y_{iJ} \mid \beta_1, \ldots, \beta_J; \sigma] = \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\theta_i - \beta_j)\}}{1 + \exp\{\theta_i - \beta_j\}} f(\theta_i \mid \sigma)\, d\theta_i \quad (5)$$

and the joint likelihood for all examinees is

$$P[\mathcal{Y} \mid \beta_1, \ldots, \beta_J; \sigma] = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp\{y_{ij}(\theta_i - \beta_j)\}}{1 + \exp\{\theta_i - \beta_j\}} f(\theta_i \mid \sigma)\, d\theta_i \quad (6)$$

Often $f(\theta \mid \sigma)$ is taken to be a normal density with mean 0 and variance σ^2 but in fact any parametric family $f(\theta \mid \sigma)$ would do.

This is essentially the likelihood that is maximized when we fit the Rasch model as a GLMM with `lmer()` in R (or other software).

One can use (6) in several different ways, e.g.:

- MLE's $\hat{\beta}_j$ and $\hat{\sigma}$ are useful in calibrating how easy or difficult the question are. For fixed J as N grows, the $\hat{\beta}_j$'s and $\hat{\sigma}$ are consistent and efficient estimators of the β_j's and σ.
- Given $\hat{\beta}_j$'s and $\hat{\sigma}$ we can produce predictors $\hat{\theta}_i$ of θ_i's (e.g. conditional MLE's, empirical Bayes posterior modes, etc.), e.g. to rank examinees, compare examinees' performance on different tests (given the right experimental design), etc.
- Fully Bayesian versions could be obtained by assigning priors to the β_j's and to σ, and obtain a joint posterior distribution for $\theta_1, \ldots, \theta_N$, $\beta_1, \ldots, \beta_J$, σ, providing similar information to the MLE's and predictors above.

Rasch Marginal Likelihood as a Log-Linear Model

We can view the probability

$$p_{\underline{y}} = P[y_1, \ldots, y_J \mid \beta_1, \ldots, \beta_J; \sigma]$$

in equation (5) as a cell probability in a multinomial model for the 2^J table $n_{\underline{y}}$. This turns out to be a certain log-linear model:

$$\begin{aligned} p(y_1, \ldots, y_J) &= \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp\{y_j(\theta - \beta_j)\}}{1 + \exp\{\theta - \beta_j\}} f(\theta|\sigma) d\theta \\ &= \left(\prod_{j=1}^{J} \exp\{-\beta_j y_j\} \right) \int_{-\infty}^{\infty} \prod_{j=1}^{J} \frac{\exp\{y_j \theta\}}{1 + \exp\{\theta - \beta_j\}} f(\theta|\sigma) d\theta \\ &= \left(\prod_{j=1}^{J} \exp\{-\beta_j y_j\} \right) \int_{-\infty}^{\infty} \frac{\exp\{\theta y_+\}}{\prod_{j=1}^{J} 1 + \exp\{\theta - \beta_j\}} f(\theta|\sigma) d\theta \end{aligned}$$

Therefore, $\log p(y_1, \ldots, y_J) = -\sum_{j=1}^{J} \beta_j y_j + \sum_{k=1}^{J} \gamma_k I_{\{y_+=k\}}$.

To maintain the hierarchy principal, we incorporate an intercept term, writing

$$\log p_{\underline{y}} = \alpha - \sum_{j=1}^{J} \beta_j y_j + \sum_{k=0}^{J} \gamma_k 1_{\{y_+ = k\}} \qquad (*)$$

where we define $y_+ = \sum_{j=1}^{J} y_j$, and $1_{\{y_+=k\}}$ is a dummy variable that equals 1 when $y_+ = k$ and equals 0 otherwise. Note that

- The β_j in $(*)$ are exactly the item difficulties in the Rasch model;
- The γ_k can be written as:

  $$\gamma_k = E[(e^{\theta})^k | \underline{y} = (0, 0, \ldots, 0)]$$

  i.e. they are moments of a positive random variable.
- The γ_k's are constrained by the β_j's in a complicated way, but as a first approximation the model can be fit, ignoring these constraints, as a straightforward log-linear model.

Cressie & Holland (1981, *Pmka*); Holland (1990; *Pmka*).

If we match up the terms in the model $(*)$

$$\log p_{\underline{y}} = \alpha - \sum_{j=1}^{J} \beta_j y_j + \sum_{k=0}^{J} \gamma_k 1_{\{y_+ = k\}}$$

with the non-redundant u-terms in the usual hierarchical log-linear model

$$\log p_{\underline{y}} = u_0 + \sum_j u_{j1} + \sum\sum_{j<k} u_{jk11} + \sum\sum\sum_{j<k<\ell} u_{jk\ell 111} + \cdots + u_{j_1 j_2 \cdots j_J}$$

we can see that

- $u_0 = \alpha$ and $u_{j1} = -\beta_j, \forall j$;
- $u_{jk11} \equiv \gamma_2, \forall j, k$: the two-way interactions are *symmetric*;
- $u_{jk\ell 111} \equiv \gamma_3, \forall j, k, \ell$: the three-way interactions are *symmetric*;
- etc. etc., i.e. each set of s-way interactions is *symmetric*.

For these reasons, $(*)$ is sometimes called the model of *quasi-symmetry*.

The model of *symmetry* would also have symmetric main effects (all u_{j1} equal to each other); and is equivalent to asserting that the y_j's are *exchangeable* random variables.

Example

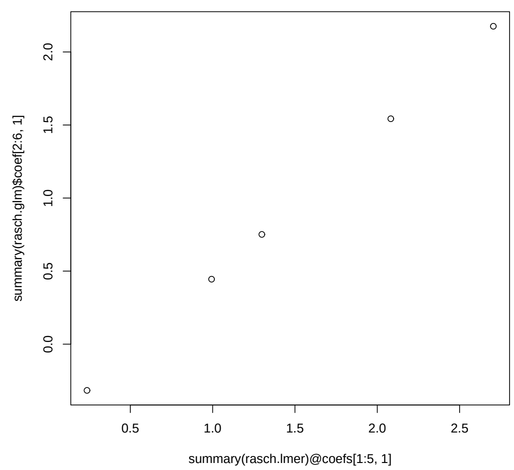
We return to the LSAT example that we used to illustrate GLMM fits of the Rasch model last time.

We can directly compare estimates of the fixed effects, β_j:

```
> rasch.lmer <- lmer(y ~ j-1 + (1|i),data=lsat,
+          family=binomial,method="Laplace")
> summary(rasch.lmer)@coefs
    Estimate Std. Error   z value     Pr(>|z|)
j1 2.7047288 0.12862039 21.028772 3.577920e-98
j2 0.9936196 0.07493543 13.259678 3.965962e-40
j3 0.2371917 0.06842979  3.466205 5.278602e-04
j4 1.2988310 0.08008535 16.218084 3.757348e-59
j5 2.0818837 0.10134168 20.543214 8.850748e-94
```

```
> rasch.glm <- glm(n ~ .,data=lsat.table,family=poisson)
> summary(rasch.glm)$coef
              Estimate Std. Error   z value     Pr(>|z|)
(Intercept)  1.0986123  0.5773503  1.902852 5.705982e-02
Y.1          2.1758247  0.1561053 13.938183 3.712707e-44
Y.2          0.4447902  0.1354840  3.282972 1.027189e-03
Y.3         -0.3162879  0.1349001 -2.344608 1.904710e-02
Y.4          0.7512857  0.1368979  5.487926 4.066813e-08
Y.5          1.5428685  0.1443920 10.685280 1.192882e-26
Yplus1      -0.9874669  0.5148710 -1.917892 5.512471e-02
Yplus2      -1.3256284  0.3676781 -3.605405 3.116666e-04
Yplus3      -1.2098123  0.2472272 -4.893524 9.904602e-07
Yplus4      -0.8532626  0.1388691 -6.144367 8.028322e-10
```

How are these related?




```
> plot(summary(rasch.lmer)@coefs[1:5,1],
+  summary(rasch.glm)$coef[2:6,1])
> lm(summary(rasch.glm)$coef[2:6,1] ~
+  summary(rasch.lmer)@coefs[1:5,1])
Coefficients:
       (Intercept)  summary(rasch.lmer)@coefs[1:5, 1]
            -0.558                                1.010
```

Almost perfectly:

$$\log \frac{p_{ij}}{1-p_{ij}} = \theta_i - \beta_j = a\Big([\theta_i - c)/a - [\beta_j - c]/a\Big)$$

The regression result above suggests $a \approx 1$ and $c = 0.558$, so that the random effects distribution implied by the log-linear fit has the same scale but is shifted down from the random effects distribution estimated by lmer. This is another change in parametrization that does not affect the fit.