Some applications and perspectives in item response modelling

Brian W. Junker
Dept of Statistics
Carnegie Mellon University
Pittsburgh PA 15213
(412) 268-8873
brian@stat.cmu.edu

Thanks to Fienberg, Lerch, Patz, Sijtsma.

A range of item response applications

- *No items:*
  Capture-recapture population estimation

  With Fienberg (CMU/Statistics), Darroch & Glonek (Adeleide, AU).

- *A few items:*
  Human working memory capacity

  With Lerch (CMU/GSIA), Huguenard (Notre Dame, IN), Kass (CMU/Statistics), Patz (CMU/Statistics).

- *Many items:*
  Transitive reasoning

  With Sijtsma (Utrecht).
  (thanks to Verweij!)

No items:
Capture-recapture population estimation

- Goals: Model-fitting and prediction.

- Persons/items $\approx 200/3 \approx 66.67$

- Lots of subjects.

- There is information about multidimensionality here!

Capture-recapture population estimation

Motivation: U.S. Census

- Mandated by U.S. Constitution.

- Taken every 10 years (April 1) since 1790.

- Exhaustive count is the goal; huge and expensive undertaking.
- Allocation based on local counts.
  - * Representatives in U.S. Congress
  - * Tax money to 39,000 state and local governments
  - * *A fixed-sum game!*

- Undercount vs. Overcount.

Darroch, Fienberg, Glonek and Junker (1993, *JASA*).

"Dual-systems" (Capture-Recapture)

- Estimates undercount at local levels.

- Sources (systems, lists, traps).
  - * Census Enumeration (E). ("exhaustive")
  - * Post-Eumeration Survey (P). (165,000 households)

- Within a small region:

	n_{ep}	(P) 1	0
(E)	1	n_{11}	n_{10}
	0	n_{01}	$n_{00} =?$

# in original census (E):	$n_{11} + n_{10}$	
# added by P.E.S. (P):	n_{01}	Undercount:
# missed by both:	n_{00}	$n_{01} + n_{00}$
Total Population:	$n_{11} + n_{10} + n_{01} + n_{00}$	

How do we estimate n_{00}?

Standard Capture-Recapture

	$n_{y_e y_p}$	(P) 1	0
(E)	1	n_{11}	n_{10}
	0	n_{01}	n_{00} =?

Assume *independent samples*, so odds ratio=1:

$$1 = \frac{n_{11} n_{00}}{n_{01} n_{10}}$$

Then

$$\widehat{n}_{00} = \frac{n_{10} n_{01}}{n_{11}}$$

Equivalent to estimating the log-linear model with no interaction:

$$\log p_{y_e y_p} = \alpha + \beta_1 y_e + \beta_2 y_p$$

- y_e is 1 if you are in the Census (E); else 0.
- y_s is 1 if you are in the PES (P); else 0.
- If we model $\log n_{y_e y_p}$, then $\exp \widehat{\alpha} = \widehat{n}_{00}$.

What if the interaction is not zero?

- *Log-linear models:*

$$\begin{aligned}\log p_{y_1 y_2 y_3 \ldots} &= \alpha + \beta_1 y_1 + \beta_2 y_2 + \beta_3 y_3 + \ldots \\ &+ \beta_{12} y_1 y_2 + \beta_{23} y_2 y_3 + \ldots \\ &+ \ldots\end{aligned}$$

- *Heterogeneity* $\Rightarrow$ positive interactions.

- Need at least *3 lists* to estimate interactions
  US Census:

Admin Lists (A)

1

		PES (P) 1	PES (P) 0
Census (E)	1	n_{111}	n_{101}
	0	n_{011}	n_{001}

0

		PES (P) 1	PES (P) 0
Census (E)	1	n_{110}	n_{100}
	0	n_{010}	n_{000} =?

Multiple-Recapture Census

- Two or more lists or captures
  (Analysis similar to 2-list case)

- Animal population estimation.
  * Burnham and Overton (1978)
  * Seber (1982)
  * Chao (1987)

- Visual scanning problems.
  * Sanathanan (1972,73)

- Survey and epidemiological work.
  * Sekar and Deming (1949)
  * The Diabetes Project (1994)

- US Census.
  * Kadane and Ericksen (1985); Fienberg (1991, ...)
  * Wolter (1986, 91); Freedman (1991)
  * Mulrey and Spencer (1991)
  * Zaslavsky and Wolfgang (1993)

Analogies with educational measurement

Multiple Recapture Census	*Measurement*
• Presence on a list	• Positive item response
• Capture effectiveness	• Item parameters
• Varying catchability	• Latent variable
• *Inference:* n_{000}	• *Inference: items/persons*
• Independent captures within persons, or strata, plus monotone catchability, implies positive dependence among captures (Kadane, Meyer and Tukey, 1992)	• LI, M, $d = 1$ implies conditional association (Holland and Rosenbaum, 1986)

An educational testing model

- Each person has a *capture record* $y = (y_1,$ y_2, $y_3)$, and a *catchability parameter* θ.
- Each list j [1=(E), 2=(P), 3=(A)] has an *effectiveness parameter* β_j
- For $P_j(\theta) = P[y_j = 1|\theta]$, assume

$$\log \frac{P_j(\theta)}{1 - P_j(\theta)} = \theta + \beta_j$$

- In sampling from the population under this model, probabilities in the 3-way table are

$$p(y_1, y_2, y_3) = \int_{-\infty}^{\infty} \prod_{j=1}^{3} P_j(\theta)^{y_j} (1 - P_j(\theta))^{1-y_j} dF(\theta)$$

- By Dutch Identity (Holland, 1990)

$$\log p(y_1, y_2, y_3) = \alpha + \sum_{j=1}^{3} \beta_j y_j + \gamma(y_1 + y_2 + y_3) \quad (DI)$$

where the $\gamma(k)$ "parametrize" the depenen-dence among the lists.

Model fitting and testing issues

- *No 3-way interaction?*
  For the log-linear estimate $\hat{n} = \exp\hat{\alpha}$ to be valid, must assume $y_1y_2y_3$ not in model.

  Impose on model, get (for 3 lists):

  $$\log p(y_1, y_2, y_3) = \alpha + \sum_{j=1}^{3} \beta_j y_j + \gamma \cdot (y_1 + y_2 + y_3)^2$$

- *Consistent with the mixture model?*
  Need $\gamma > 0$. Check $\hat{\gamma}$.

- *Rasch model tests?*
  $P[Y_i = 1 | Y_i + Y_j = 1, h(\text{other items}) = c]$ independent of c? (Rosenbaum, 1987)

(P vs A) $\coprod$ E?		In PES (P)	In A. Lists (A)
In Census (E)	Yes	n_{110}	n_{101}
	No	n_{010}	n_{001}

An example: Four "Post-strata" in a St. Louis (MO, USA) trial run [Own/Rent, 20's/30's].

Log-odds-ratio	O2	R2	O3	R3
(A vs E) $\coprod$ P	−2.20	−2.04	−2.48	−1.98
(P vs A) $\coprod$ E	2.34	2.00	2.32	2.89
(E vs P) $\coprod$ A	−0.14	0.05	0.15	−0.91

Chi-Squared	O2	R2	O3	R3
(A vs E) $\coprod$ P	66.50	37.35	58.30	38.36
(P vs A) $\coprod$ E	74.84	37.14	57.09	63.19
(E vs P) $\coprod$ A	0.15	0.01	0.11	3.44

- (E vs P) $\coprod$ A can be achieved by

$$\log \frac{P_j(\theta_1,\theta_2)}{1-P_j(\theta_1,\theta_2)} = \begin{cases} \theta_1+\beta_i, & i=1,2 \\ \theta_2+\beta_3, & i=3 \end{cases}$$

- Integrate out θ's, get:

$$\log p(y_1,y_2,y_3) = \alpha + y_1\beta_1 + y_2\beta_2 + y_3\beta_3 + \gamma(y_1+y_2,y_3)$$

where $\gamma(2,0) \geq 2\gamma(1,0)$ and $\gamma(2,1)+\gamma(0,1) \geq 2\gamma(1,1)$.

- Administrative lists (A) are "different".

Results

Strata	O2	R2	O3	R3
$\widehat{x}_{000}$: Saturated	290.06	670.47	496.83	825 .97
S.E. (delta method)	118.96	335.01	247.31	449.62
Fit / d.f.	—/0	—/0	—/0	—/0
$\widehat{x}_{000}$: 2-dim Rasch	291.55	669.50	489.84	767.80
S.E. (delta method)	119.46	334.28	242.12	402.86
Fit / d.f.	0.15/1	0.01/1	0.11/1	3.45/1
$\widehat{x}_{000}$: Rasch	46.69	102.06	42.33	81.91
S.E. (delta method)	14.40	35.76	13.41	28.72
Fit / d.f.	94.04/2	52.99/2	67.89/2	73.01/2
$\widehat{x}_{000}$: Independent Lists	44.64	48.26	27.51	34.98
S.E. (delta method)	5.69	6.78	3.80	5.06
Fit / d.f.	94.06/3	58.34/3	70.15/3	80.06/3

Related work (!)

- Stegelman (1983, *Psy*): Multidimensional Rasch.
- Fischer (1983, *Psy*; ...): LLTM (stratum variables).
- Kelderman (1992, *Psy*; ...): Log-Linear Rasch. Also Lindsay, Clogg and Grego (1991, *JASA*)
- Fienberg (1992, *Statistics Canada*): Survey of literature in Census, Animal and Epidemiology settings.
- Diabetes Project (1994, in progress).
  Yip, Bruno, Tajima, Seber, Buckland, Cormack, Unwin, Fienberg, Junker, Chang, LaPorte, Libman, McCarty.

Some conclusions and questions

- The log-linear item response model has useful applications outside of IRT.

- Not enough is known about meeting the "moment restrictions" on the interactions.

- There is information in the data about multidimensionality.

- It helps to be in the context of a parametric model.

A few items:
Human working memory capacity

- Goals: Evaluate a small-scale designed experiment in psychology.

- Persons/items $\approx 87/27 \approx 3.22$ (at best)

- Moderate number of subjects must be spread among many more "virtual items".

- Model selection: "think about it."

Human working memory capacity

Human facility for maintaining and manipulating information during complex cognitive activities such as learning, comprehension, and reasoning. (Baddeley, 1992)

Phone-Based Interaction (PBI)

Examples: Bank by phone; campus information system

PBI creates demands for both

- Storage (remembering) and
- Processing (pattern matching)

Patz, Junker, Lerch, Huguenard (1994, subm.).

The Experiment

- 87 University students

- Administered a reading span (WM) test (Daneman and Carpenter, 1980)

- Randomized to one of two menus
  Deep ($3 \times 3 \times 3 \times 3$)
  Broad (9×9)

- (27 tasks) $\times$ (3 WM levels) $\times$ (2 Menus)
  $=$ 162 "virtual items"
  (between-subjects factors—Fischer...)

Experimental factors

- Menu structure (MENU) at 2 levels:
  Deep, Broad
- Reading span (WM) at 3 levels:
  Low, Medium, High
- Modifier type (MODT) at 3 levels:
  None, Verbal, Numeric
- Task order (TORD) at 3 levels:
  Single, 1^{st} in pair, 2^{nd} in pair
- Number of options (NOPT) at 3 levels:
  Few, Moderate, Many

Response vector

- Navigational error (backtracking in menu)
- Rehearsal error
  (forgetting: action, object or modifier)
- Both navigational and rehearsal errors
- Task failure (failure to complete)
- Perfect response

A polytomous Rasch model

Logits relative to the "perfect" category, M:

$$\log \frac{P_{jm}(\theta_i)}{P_{jM}(\theta_i)} = \theta_i - \beta_{jm}$$

- $i = 1, 2, \ldots, N$ subjects; $(N = 87)$
- $j = 1, 2, \ldots, J$ items $(J = 6 \times 27 = 162)$
- $m = 1, 2, \ldots, M - 1$ resp. categories $(M = 5)$

Local independence yields the mixture model for subject i

$$p(Y_i = y_i|\beta) = \int \prod_{j=1}^{J} \frac{\prod_{m=1}^{M-1} \exp[y_{ijm}(\theta - \beta_{jm})]}{1 + \sum_{k=1}^{M-1} \exp[\theta - \beta_{jk}]} dF(\theta).$$

For experimental factors we decompose β into K factors:

$$\beta = G\psi$$

Each item parameter may be expressed as

$$\beta_{jm} = \psi_{WM,m} + \psi_{MENU,m} + \psi_{MODT,m} + \psi_{TORD,m} + \psi_{NOPT,m}$$

The model and effects of interest

• Decomposition of β_{jm} accounts for within-subjects factors and known between-subjects factors

• θ accounts for dependence due to unknown subject effects (e.g. individual strategies)

• Main focus of inference is on the decomposed β_{jm}'s

• Omitting θ inflates between-subjects effects

• Similar structure to large scale educational assessments(!)

• Fischer and Ponocny, 1993;
• Glas and Verhelst, 1989;
• Mislevy, 1988;
• Pirolli and Wilson, 1992;
• others...

E-M Estimation

Likelihood for the complete matrix $\underline{\underline{Y}}$ is

$$L(\beta, \sigma) = \prod_{i=1}^{N} P(Y_i = y_i|\beta) = \int P(\underline{\underline{Y}} = \underline{\underline{y}}|\beta, \theta) f(\theta|\sigma) d\theta.$$

E-M yields

$$(\sigma^{\nu+1})^2 = \frac{1}{N} \sum_i \int \theta_i^2 g(\theta_i|y_i, \beta^\nu, \sigma^\nu) d\theta_i.$$

and (solve for $\beta_{jm}^{\nu+1}$)

$$\sum_i Y_{ijm} = \sum_i \int \frac{1}{1 + \exp(\theta_i - \beta_{jm}^{\nu+1})} g(\theta_i|y_i, \beta^\nu, \sigma^\nu) d\theta_i.$$

where

$$g(\theta_i|y_i, \beta^\nu, \sigma^\nu) = \frac{P(Y_i = y_i|\beta^\nu, \theta_i) f(\theta_i|\sigma^\nu)}{\int P(Y_i = y_i|\beta^\nu, \theta_i) f(\theta_i|\sigma^\nu) d\theta_i}.$$

- Starting from "independence": $\sigma^2 \equiv 0$;
- Rigdon and Tsutakawa's (1983) CMLF;
- Simple-minded Ramsay (1975) accelleration.

Focusing parameters on experimental effects

• Marginal table for one "virtual item," the first item for low WM, deep MENU subjects:

$Nav.Error$ $(m=1)$	$Rhse.Error$ $(m=2)$	$Nav+Rhse.$ $(m=3)$	$Failure$ $(m=4)$	$Correct$ $(m=5)$
1	1	2	0	6

Requires $162 \times 4 = 648$ item parameters β_{jm}.

• Marginal table for experimental factor, *second in pair*:

$Nav.Error$ $(m=1)$	$Rhse.Error$ $(m=2)$	$Nav+Rhse.$ $(m=3)$	$Failure$ $(m=4)$	$Correct$ $(m=5)$
482	90	87	24	100

Only $10 \times 4 = 40$ task factor parameters ψ_{km}

• Re-using responses for different inferences

Testing and inference

- Likelihood ratio χ^2 statistics for the main effects and interactions; Bonferroni corrections.

- Other hypotheses stated in terms of margins

$$P(\text{RHSE or BOTH} \mid \text{paired task}) \\ - P(\text{RHSE or BOTH} \mid \text{singelton}) > 0$$

For these we used a mixture of the Delta Method and Monte Carlo testing, based on

$$Z = \frac{\text{fitted difference}}{\text{SE of fitted difference}} \sim N(0,1)$$

Three Sensitivity Issues

- Have achieved the global maximum?
  * Use a range of starting values
  * Perturb fitted parameters and reconverge

- Stable when the data is perturbed slightly?
  * Similar results with simulated data

- Do asymptotic distributions hold for tests?
  * Null-model simulations support χ^2

Specific results not "generalizable"—but methodology generalizes to your data!

Results

- Menu especially important for navigation
- Working memory especially important for "total failure"
(Surprisingly, not for "rehearsal errors")
- Modifier type important for rehearsal-related errors
- Task Order important for rehearsal-related errors
- Interactions: Menu $\times$ Modifier Type and Menu $\times$ Task Order
- Number of Options important
(Nonmonotonicity of Few/Medium/Many $\Rightarrow$ surrogate for item ambiguity)

Conclusions and Questions

- "Moderately many" items diluted by sensible experimental designs

- Small sample IRT possible with careful parametrization and hard work at fitting

- Structure similar to educational assessment surveys

- Hidden: Adapting Cognitive Science models to statistical framework

- How do we know model is right?
  von Maanen, Been and Sijtsma (1989)

Many items:
Transitive reasoning

- Goals: Scaling! What items “hang together?” Why?

- Persons/items $\approx 425/10 = 42.5$

- Good setting to think about nonparametric IRT.

- Unidimensionality well-developed; hard to think about multidimensionality!

Transitive reasoning

- 10 Transitive Reasoning Tasks;

  Pairwise Comparisons ⇒ Judgement of Total Order

Item	Object	Concept	Correct Judgement
1.	Balls	Weight	$X_A > X_B = X_C$
2.	Tubes	Length	$X_A = X_B = X_C = X_D$
3.	Disks	Area	$X_A > X_B > X_C$
4.	Cubes	Weight	$X_A = X_B = X_C = X_D$
5.	Sticks	Length	$X_A > X_B = X_C$
6.	Sticks	Length	$X_A > X_B > X_C$
8.	Tubes	Weight	$X_A > X_B > X_C$
9.	Balls	Weight	$X_A < X_B < X_C$
10.	Sticks	Length	$X_A = X_B < X_C = X_D$
11.	Balls	Weight	$X_A = X_B < X_C = X_D$

- 425 School children.

- Items scored three ways:

  * Judgement only

  * Judgement plus reason

  * 0/1/2 (Wrong/Judgement/Reason)

Sijtsma and Junker (1994, subm.)—Builds on Sijtsma's work with Verweij.

Developmental scale

- Order the items so each is a step in a developmental journey. (Guttman!)

- The locally independent "marginal" model:

$$p(y_1, \ldots, y_J) = \int_{-\infty}^{\infty} \prod_{j=1}^{J} P_j(\theta)^{y_j} (1 - P_j(\theta))^{1-y_j} \, dF(\theta).$$

- Mokken models:

  Monotone Homogeneity (MH):

  dim(θ) = 1; $P_j(\theta)$ nondecreasing.

  (LI, M, d=1).

  Invariant item ordering (IO):

  dim(θ) = 1; $P_j(\theta)$'s do not cross.

  Double Monotonicity (DM):

  MH plus IO.

Tools

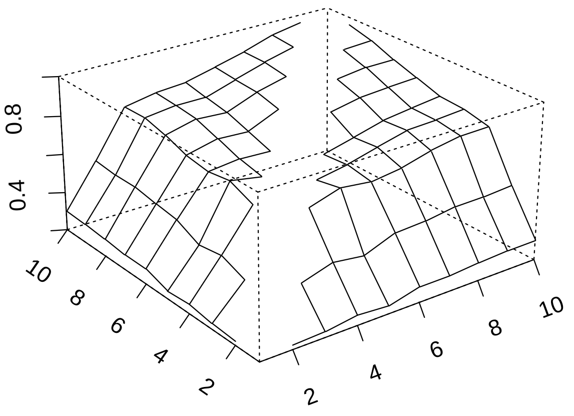
- Global association measures
  - * Loevinger's H coefficient
  - * Sijtsma/Meijer's H^T coefficient
- Pairwise measures [e.g. $P(++)$, $P(--)$]
  - * Most (old and new) follow from:

$$\begin{aligned} &Y_i \text{ uniformly harder than } Y_j \Rightarrow \\ &\qquad P[Y_i = 1 | h(\text{other items}) = c] \\ &\qquad\quad \leq P[Y_j = 1 | h(\text{other items}) = c] \end{aligned}$$

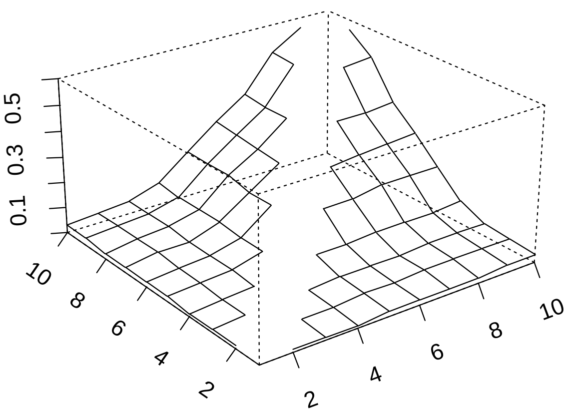
    (Rosenbaum, 1987).

- Problem ends up "feeling" very similar to diagnostics for linear regression:
  - * Multiple inference means everything should be viewed as "index"
  - * Many indices measure approximately the same thing

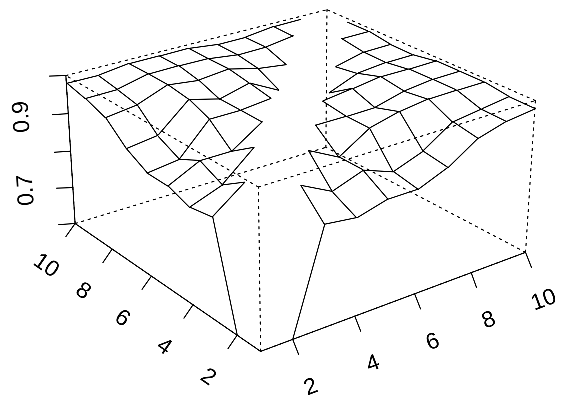
Graphical aids help alot!



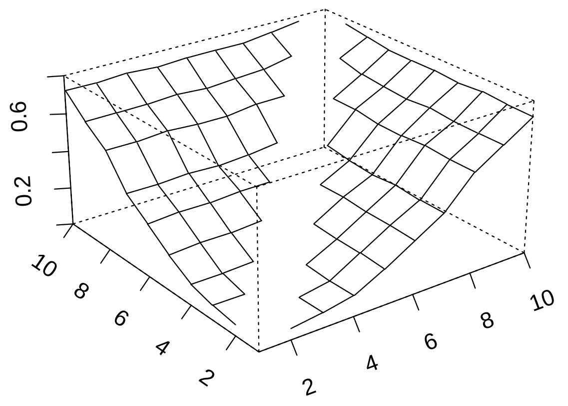
Off-diagonal entries of P(++) matrix (judgement only)



Off-diagonal entries of P(++) matrix (judgement plus)



Off-diagonal entries of 1-P(--) matrix (judgement only)



Off-diagonal entries of 1-P(--) matrix (judgement plus)

Results

Item	Object	Concept	Correct Judgement
1.	Balls	Weight	$X_A > X_B = X_C$
2.*	Tubes	Length	$X_A = X_B = X_C = X_D$
3.	Disks	Area	$X_A > X_B > X_C$
4.*	Cubes	Weight	$X_A = X_B = X_C = X_D$
5.+	Sticks	Length	$X_A > X_B = X_C$
6.	Sticks	Length	$X_A > X_B > X_C$
8.+	Tubes	Weight	$X_A > X_B > X_C$
9.−	Balls	Weight	$X_A < X_B < X_C$
10.	Sticks	Length	$X_A = X_B < X_C = X_D$
11.	Balls	Weight	$X_A = X_B < X_C = X_D$

- Judgement plus reason: great!
- Judgement only:

  $\ast$ Could drop 2 and 4.

  $+$ Or could drop 5 and 8.

  $-$ After that, maybe drop 9.

Here we rely on "expert judgement"—situation not that different from DIF analysis for example!

Some issues

- This kind of “item-sifting” leads to “unidimensional pools”—simple structure.

- What do we do with the “dropped” items? (e.g. 2,4).

- Multidimensional nonparametric IRT?

- Polytomous nonparametric IRT?

- How do we know when we’ve arrived at enough conditions to test?

Summary thoughts

- *No items: Census*
  - * Item response models useful outside of IRT!
  - * Many interesting IRT style questions still arise.
  - * Few items $\Rightarrow$ parametric models.
- *A few items: Working Memory*
  - * Even moderately many items can be diluted by perfectly sensible experimental designs.
  - * Parametrize carefully, work hard at fitting $\Rightarrow$ progress!
  - * Many applications like this waiting to be done!
  - * Similar in structure to large scale educational assessments.

• *Many items: Scaling (Model Identification), Inference about Abilities*

* X_+ a sensible stand-in for θ.

* Nonparametrics possible and useful.

* As number of items grows, must have lots of persons too!

Next 3 talks

• A view of nonparametric item response theory.

• Inference, robustness, and essential unidimensionality.

• A characterization of monotone unidimensional latent variable models.