Nuisance parameters and likelihood corrections

Richard P. Waterman
University of Pennsylvania

waterman@compstat.wharton.upenn.edu
http://compstat.wharton.upenn.edu:8001
/~waterman/homepage.html

March, 1995

Overview

- Motivation – Matched pairs and Rasch model.
- Generalization.
- Approaches to inference.
- Asymptotics.
- Computations.
- Output.
- Summary.

Motivation

Matched Pairs $(X_{i1}, X_{i2}), i = 1, \ldots, r$:

$$X_{i1} \sim Bernoulli\left(\frac{\exp(\lambda_i + \psi_1)}{1 + \exp(\lambda_i + \psi_1)}\right)$$
$$X_{i2} \sim Bernoulli\left(\frac{\exp(\lambda_i + \psi_2)}{1 + \exp(\lambda_i + \psi_2)}\right)$$

Take $\psi_1 = 0$ for identifiability.
Many λ's, but (ψ_1, ψ_2) are fixed.

- Maximum likelihood estimation for ψ is bad.
- Inconsistent estimates. Neyman-Scott problem.
- Conditional inference is an alternative
  $S_i = X_{i1} + X_{i2}$.

Rasch Model

Let x_{ij} be response of subject i for item j.

$$f(x_{ij};\psi_j,\lambda_i)=\frac{\exp\{(\psi_j+\lambda_i)y_{ij}\}}{1+\exp(\psi_j+\lambda_i)}$$

ψ_j: item difficulty parameter. (Parameter of interest).
λ_i: subject ability parameter. (Nuisance parameter).

Note: Multivariate parameter of interest, and multivariate nuisance parameter.

+ No interaction between subject and item parameters. Independence of responses across items.

CSS for the λ_i are the row totals, (number of questions an individual got correct).

Why conditional inference for item effects??

Like semi-parametric model with an unspecified mixing distribution on the λ_i.

Consequences of Rasch model

Linear regression on the logits of the success probabilities, p_{ij}.

$$logit(p_{ij}) = \psi_j + \lambda_i.$$

Lines with equal slopes but different intercepts. Lines do not cross.

Data: 678 Wharton MBA's. Statistics Multiple Choice Exam.

Select two items from the test:

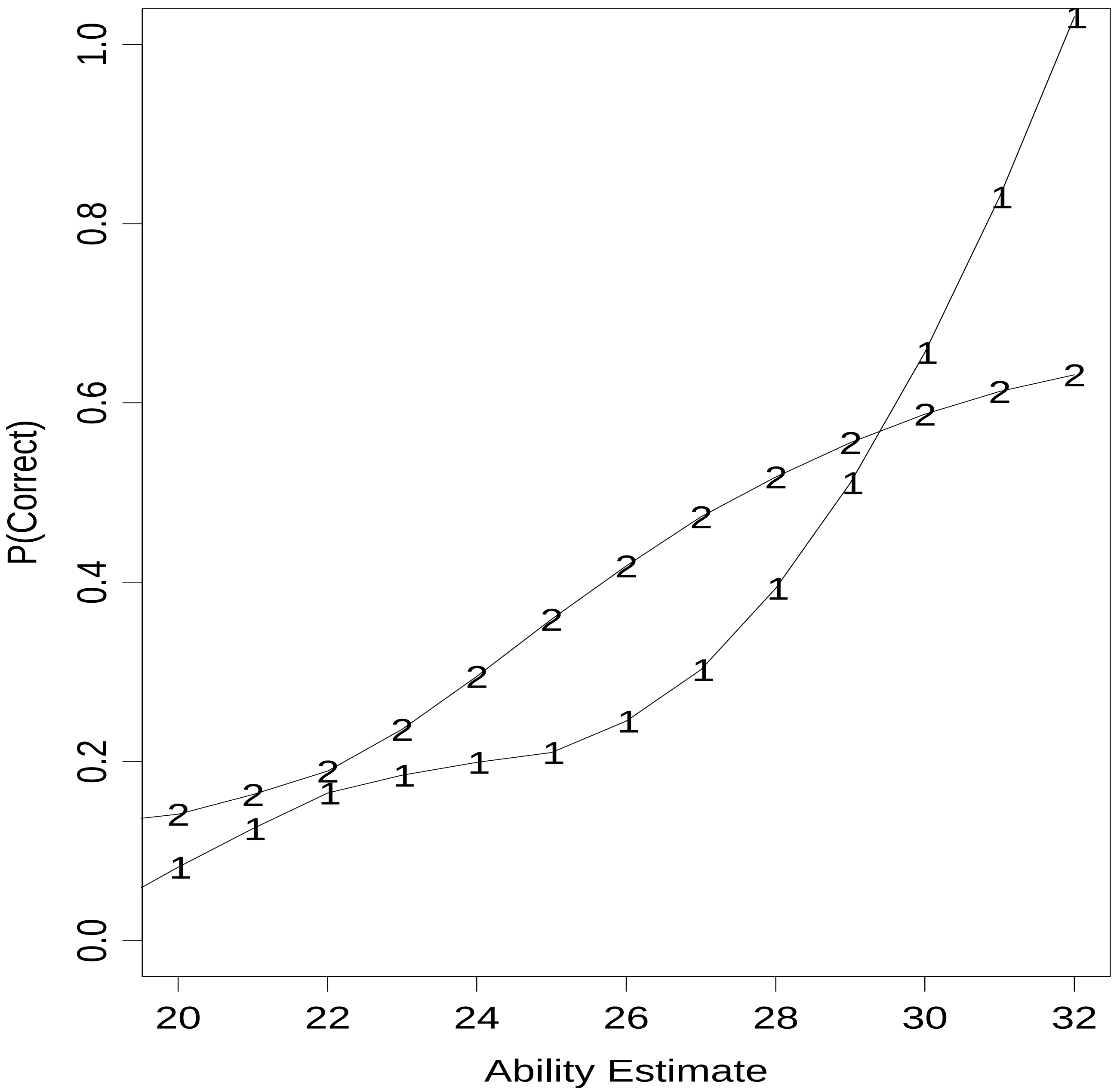

Figure 1: Smoothed response curves for two items on a statistics multiple choice test

Conclude

Rasch model may not be appropriate.

Alternative models lack conditional structure though.

Mimic conditional inference.

Alternatives for estimation

Likelihood

$$f(x) = \prod_{i=1}^{r} \int \prod_{j=1}^{c} p(u, \psi_j)^{x_{ij}} (1 - p(u, \psi_j))^{1-x_{ij}} dG_i(u).$$

How to treat G_i?

- $G_i \equiv N(\mu, \sigma)$. Mixed-model.
- $G_i \equiv G \in \mathcal{G}$. Semi-parametric.
- $G_i \equiv \delta_{\lambda_i}$. Fixed effects.

Also

- Population averaged models, GEE.
- Bayesian approaches.

Generalization

Cuboid and likelihood:

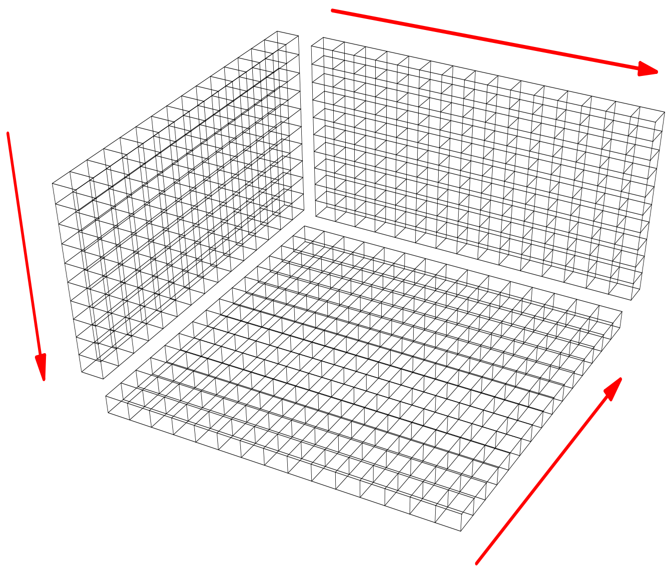

Figure 2: Cuboid data structure

Conditional on λ_i

$$f(x;\psi,\lambda) = \prod_{ijk} f(x_{ijk};\lambda_i,\psi_j).$$

Asymptotics

Let r, c, n go to infinity in different ways.

- Fix r, c. $\quad n \to \infty$. Standard.
- Fix c, n. $\quad r \to \infty$. Neyman-Scott.
- Fix c. $\quad n, r \to \infty, r/n \to Constant$. eg $n = r$; Square array etc.

Methodology for mimicry

Projections.

Optimal Estimating Equation.

$$U_\infty = U_0 - \sum_{\alpha=1}^{\infty} \rho_{\alpha\infty} V_\alpha^*,$$

V_α^* orthogonalized version of V_α where $V_\alpha = f^{(\alpha)}/f$.

Optimal; maximizes information in class of functions satisfying

$$E_\lambda\{g(\psi_0, \lambda_0)\} = 0, \text{ for all } \lambda_0, \psi_0, \lambda.$$

If a conditional score exists, U_c then $U_c = U_\infty$.

Approximations: **truncate** the series.

Leads to ...

Maximum likelihood:

Solve

$$\begin{pmatrix} U_0 \\ V_1 \end{pmatrix} = 0, \tag{1}$$

U_2 score

$$\begin{pmatrix} U_0 - \rho_{12}V_1 - \rho_{22}V_2 \\ V_1 \end{pmatrix} = 0,$$

which is simplified by using the estimates $\hat{\lambda}_\psi$ for λ, and becomes

$$\begin{pmatrix} U_0 - \rho_{22}V_1^\lambda \\ V_1 \end{pmatrix} = 0, \tag{2}$$

Issues:

- How to construct it?
- How to solve it?

von-Mises

Consider the solution to (1) or (2) as $\psi(F) = \psi(F_1, \ldots, F_r)$. Then expand $\psi(F)$ evaluated at $\psi(\hat{F})$ about $\psi(F_0)$.

Technique: von-Mises expansion. For the i^{th} stratum define

$$F_i(\epsilon_i) = F_{i0} + \epsilon_i \sqrt{n_i}(\hat{F}_i - F_{i0}),$$

so that the derivative with respect to ϵ_i is

$$\sqrt{n_i}(\hat{F}_i - F_i) = \Delta_{in}.$$

The second order von-Mises expansion is defined by

$$\psi\left(\frac{1}{\sqrt{n_1}}, \ldots, \frac{1}{\sqrt{n_p}}\right) - \psi(0, \ldots, 0) = \sum_k \psi'_k \Big|_{\epsilon_k=0} \frac{1}{\sqrt{n_k}} +$$

$$\frac{1}{2} \sum_{kl} \psi''_{kl} \Big|_{\epsilon_k=\epsilon_l=0} \frac{1}{\sqrt{n_k}} \frac{1}{\sqrt{n_l}} + Rem(\hat{F} - F_0)$$

Key fact in square array $r/n \to const$

If use maximum likelihood (1), $\hat{\psi}$, then

$$\sum_{kl} \psi''_{kl}\Big|_{\epsilon_k=\epsilon_l=0} \xrightarrow{p} Bias.$$

If use projective (2), $\tilde{\psi}$, then

$$\sum_{kl} \psi''_{kl}\Big|_{\epsilon_k=\epsilon_l=0} \xrightarrow{p} 0.$$

Consequences

First order effects in this asymptotics.

$$\sqrt{n}\sqrt{r}(\hat{\psi} - \psi) \xrightarrow{\ell} N(Bias, Info),$$

but

$$\sqrt{n}\sqrt{r}(\tilde{\psi} - \psi) \xrightarrow{\ell} N(0, Info),$$

fully efficient, even though nuisance parameters head off to infinity.

Likelihood ratio test:

$$\psi = \psi_0 \quad vs \quad \psi \neq \psi_0 \qquad \lambda \quad \text{unspecified..}$$

Using

$$l(\tilde{\psi}, \hat{\lambda}_{\tilde{\psi}}) - l(\psi_0, \hat{\lambda}_{\psi_0}),$$

is OK (ie χ^2 approximation) but using

$$l(\hat{\psi}, \hat{\lambda}_{\hat{\psi}}) - l(\psi_0, \hat{\lambda}_{\psi_0})$$

is not.

Correction is not optional.

When is correction zero?

Ans: If $U_1 \perp V_2^*$.

This happens for a canonical exponential family in each strata

$$b(y)exp\{s\eta_1 + t\eta_2 - a(\eta_1, \eta_2)\},$$

and we use **mean** parameter for ψ and **canonical** for λ.

$$\psi = \frac{\partial}{\partial \eta_1} a(\eta_1, \eta_2).$$

$$\eta_1 = \eta_1(\psi, \eta_2).$$

Since things are invariant under nuisance parameter transformations, OK as long as parameter of interest is a mean.

See Cox & Reid (1992).

Two things going for U_2

1. In square array asymptotics it does a precise correction.

2. At extreme, n = 1, it is approximating conditional inference

Computations

Find the ρ_{it}.

Find extended information matrix

$$M_2 = E\left\{\begin{pmatrix} U_0 \\ V_1 \\ V_2 \end{pmatrix}(U_0, V_1, V_2)\right\}.$$

If have **Closed form for MGF** then symbolic math package saves the day.

E.g. Complementary Log-Log link. $E(V_4 V_4)$.

```
C     Find other elements of E(V4 V4)
  o1=a+u
  o2=exp(2.d0*o1)
  o3=exp(o1)
  o4=exp(o3)
  o5=n(i)
  o6=exp(3.d0*o1)
  o7=exp(4.d0*o1)
  o8=exp(5.d0*o1)
  o9=exp(6.d0*o1)
  o10=a+o3+u
  o11=exp(2.d0*o10)
  o12=2.d0*a+o3+2.d0*u
  o13=exp(o12)
  o14=3.d0*a+o3+3.d0*u
  o15=exp(o14)
  o16=4.d0*a+o3+4.d0*u
  o17=exp(o16)
  o18=5.d0*a+o3+5.d0*u
  o19=exp(o18)
  o20=6.d0*a+o3+6.d0*u
  o21=exp(o20)
  o22=3.d0*a+2.d0*o3+3.d0*u
  o23=exp(o22)
  o24=4.d0*a+2.d0*o3+4.d0*u
  o25=exp(o24)
```

```
 o26=5.d0*a+2.d0*o3+5.d0*u
 o27=exp(o26)
 o28=6.d0*a+2.d0*o3+6.d0*u
 o29=exp(o28)
 o30=o5**2
 t1=o2/(-1.d0+o4)**4
 t2=-1.d0-2.81d2*o11+3.79d2*o13-1.266d3*o15+2.27d3*o17-1.908d
& 3*o19
 t3=-1.59d2*o2+6.31d2*o21+7.62d2*o23-9.94d2*o25+5.4d2*o27-1.0
& 1d2*o29
 t4=1.4d1*o3+2.16d2*o17*o30-4.32d2*o19*o30+2.16d2*o21*o30+3.d
& 0*o4+9.8d1*o11*o5
 t5=-1.96d2*o13*o5+1.008d3*o15*o5-2.336d3*o17*o5+2.304d3*o19*
& o5+9.8d1*o2*o5-8.44d2*o21*o5-5.04d2*o23*o5
 t6=8.44d2*o25*o5-5.04d2*o27*o5+9.8d1*o29*o5+5.9d2*o6-5.04d2*
& o5*o6-1.326d3*o7
 t7=-2.16d2*o30*o7+1.492d3*o5*o7+1.38d3*o8+4.32d2*o30*o8-1.8d
& 3*o5*o8-6.75d2*o9
 t8=-3.6d2*o30*o9+1.01d3*o5*o9+2.4d1*o5**3*o9
 t9=-4.2d1*exp(o10)-8.6d1*exp(3.d0*o10)-3.d0*exp(2.d0*o3)
 t10=exp(3.d0*o3)+4.2d1*exp(a+2.d0*o3+u)-1.4d1*exp(a+3.d0*o3+
& u)
 t11=6.1d1*exp(2.d0*a+3.d0*o3+2.d0*u)
 t12=5.d1*exp(4.d0*a+3.d0*o3+4.d0*u)
 t13=-1.2d1*exp(5.d0*a+3.d0*o3+5.d0*u)+exp(6.d0*a+3.d0*o3+6.d
& 0*u)
 EV4V4(j)=o5*t1*(t10+t11+t12+t13+t2+t3+t4+t5+t6+t7+t8+t9)
```

Symbolic Numeric Environments –
automatic code generation.

An Efficient Symbolic-Numeric Environment
By Extending Mathematica's Format Rules.
Mark Sofroniou.

Solving the equations. Not so bad.

Useful: use ∇U_0, not $\nabla(U_0 - \rho_{22} V_1^\lambda)$.
Block structures in information matrices are exploited.

Examples: 2 parameter logistic. MBA data
Comparison between projective and mixed.

Parameters	Mixed		U_2	
	Estimates	S.D.	Estimates	S.D.
σ	0.8305	0.1871	—	—
μ	-0.9817	0.1076	—	—
ψ_{21}	0.0034	0.2083	-0.1353	0.1617
ψ_{22}	0.5561	0.2184	0.3709	0.1570
ψ_{31}	0.4939	0.2615	0.4707	0.1889
ψ_{32}	0.8389	0.2812	0.7978	0.1991
ψ_{41}	0.3918	0.2217	0.3895	0.1594
ψ_{42}	0.6506	0.2296	0.6358	0.1605
ψ_{51}	2.5022	0.3348	2.3919	0.2790
ψ_{52}	0.4932	0.2717	0.3872	0.2103
ψ_{61}	0.8861	0.2330	0.8766	0.1801
ψ_{62}	0.6077	0.2269	0.5855	0.1681
ψ_{71}	2.2969	0.3673	2.3652	0.3456
ψ_{72}	0.9379	0.3229	0.9616	0.2761
ψ_{81}	1.8281	0.3049	1.8717	0.2950
ψ_{82}	0.8079	0.2780	0.8207	0.2506
ψ_{91}	2.9364	0.4301	3.0487	0.4770
ψ_{92}	1.0880	0.3505	1.1343	0.3443
$\psi_{10\,1}$	3.0596	0.4418	2.9679	0.3875
$\psi_{10\,2}$	0.8182	0.3454	0.7401	0.2750

Notes

- This is a do-able procedure.
- Estimates not radically different.
- Some shrinkage on U_2.
- **Fixed/Mixed** information available, $\sim$20% increase in standard errors.
- U_2 settles down faster than mixed.

Simulation: 2 parameter logistic. 100 reps
Comparison between n = 1, n = 2, U_1, U_2.
$\sim$2500 $\lambda' s$, dim(ψ) = 18

Par.	True Vals.	n = 1, U_2 Mean	n = 1, U_2 S.D.	n = 2, U_1 Mean	n = 2, U_1 S.D.	n = 2, U_2 Mean	n = 2, U_2 S.D.
ψ_{12}	-0.5000	-0.5249	0.0412	-0.5081	0.0346	-0.5008	0.0349
ψ_{22}	0.5000	0.3462	0.1228	0.4347	0.0554	0.4878	0.0532
ψ_{13}	-0.5000	-0.5088	0.0667	-0.5173	0.0497	-0.4954	0.0477
ψ_{23}	1.0000	1.0119	0.1131	1.0053	0.0815	1.0093	0.0680
ψ_{14}	-0.5000	-0.5230	0.0972	-0.5645	0.0692	-0.5026	0.0582
ψ_{24}	1.5000	1.6086	0.1811	1.7132	0.1422	1.5159	0.0970
ψ_{15}	0.5000	0.4810	0.0507	0.5012	0.0333	0.4979	0.0332
ψ_{25}	0.5000	0.3985	0.0952	0.4413	0.0402	0.4890	0.0387
ψ_{16}	0.5000	0.5037	0.0664	0.5191	0.0521	0.4970	0.0491
ψ_{26}	1.0000	1.0246	0.1136	0.9989	0.0864	1.0009	0.0720
ψ_{17}	0.5000	0.5070	0.0898	0.5680	0.0768	0.5070	0.0654
ψ_{27}	1.5000	1.5847	0.2008	1.6938	0.1414	1.5054	0.0957
ψ_{18}	-1.0000	-1.0149	0.1015	-1.1153	0.0740	-0.9967	0.0569
ψ_{28}	1.5000	1.5953	0.1779	1.6919	0.1482	1.5108	0.0969
ψ_{19}	1.0000	0.9943	0.0809	1.0393	0.0473	0.9980	0.0444
ψ_{29}	1.0000	1.0063	0.1272	0.9826	0.0816	0.9977	0.0692
$\psi_{1\,10}$	2.0000	1.9721	0.0958	2.0708	0.0648	2.0060	0.0600
$\psi_{2\,10}$	1.0000	0.9662	0.1202	0.9360	0.0788	0.9961	0.0705

Notes

- Harder to estimate ψ_{2j} than ψ_{1j}
- U_1, ie $m.l.e$ for $n = 1$ is unstable.
- U_2 is quite successful in reducing bias, even for r = 2500, n = 2.

Non-canonical links, and higher order scores.
Comparison between n = 1, n = 2, n = 5, U_1, U_2.
$\sim$2500 $\lambda's$, dim(ψ) = 4, reps = 100

n = 1	U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.0743	0.0610	1.0372	0.0551	1.0951	0.0644
ψ_3	1.9431	0.0707	2.0818	0.0743	1.9697	0.0635
ψ_4	-1.0714	0.0661	-0.9261	0.0548	-0.9809	0.0645
ψ_5	-2.4439	0.1075	-2.0018	0.0780	-1.9176	0.0847

n = 2	U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.0195	0.0386	1.0163	0.0378	1.0145	0.0377
ψ_3	1.9607	0.0516	2.0154	0.0560	1.9810	0.0529
ψ_4	-1.0229	0.0398	-0.9989	0.0385	-1.0231	0.0384
ψ_5	-2.0869	0.0465	-1.9682	0.0428	-1.9699	0.0418

n = 5	U_1		U_2		U_3		U_4	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
ψ_2	1.0912	0.0268	1.0022	0.0243	1.0028	0.0243	0.9996	0.0242
ψ_3	2.2063	0.0401	1.9932	0.0347	2.0046	0.0351	1.9960	0.0348
ψ_4	-1.0797	0.0244	-1.0097	0.0229	-1.0053	0.0227	-1.0067	0.0228
ψ_5	-2.1391	0.0283	-2.0114	0.0264	-1.9898	0.0260	-1.9967	0.0262

Notes

- U_1 still is unstable. Stable at $n = 5$
- Some serious bias in U_2 for $n = 1$
- Moving to U_3 dampens this.
- Correction is accurate, see $n = 5$, $U_1 \quad vs \quad U_2$.
- Tougher than 2 parameter-logistic

Summary

- Maximum likelihood fails in square array asymptotics
- U_2 sorts out the problem
- This projective approach is implementable for class of problems described here
- At present must have symbolic-numeric interface
- Theoretical insight
- Practical insight – **Fixed/Mixed tradeoff**