Analysis of ordinal data via heteroscedastic threshold models

"Nurse, get on the internet, go to SURGERY.COM, scroll down and click on the 'Are you totally lost?' icon."

Example

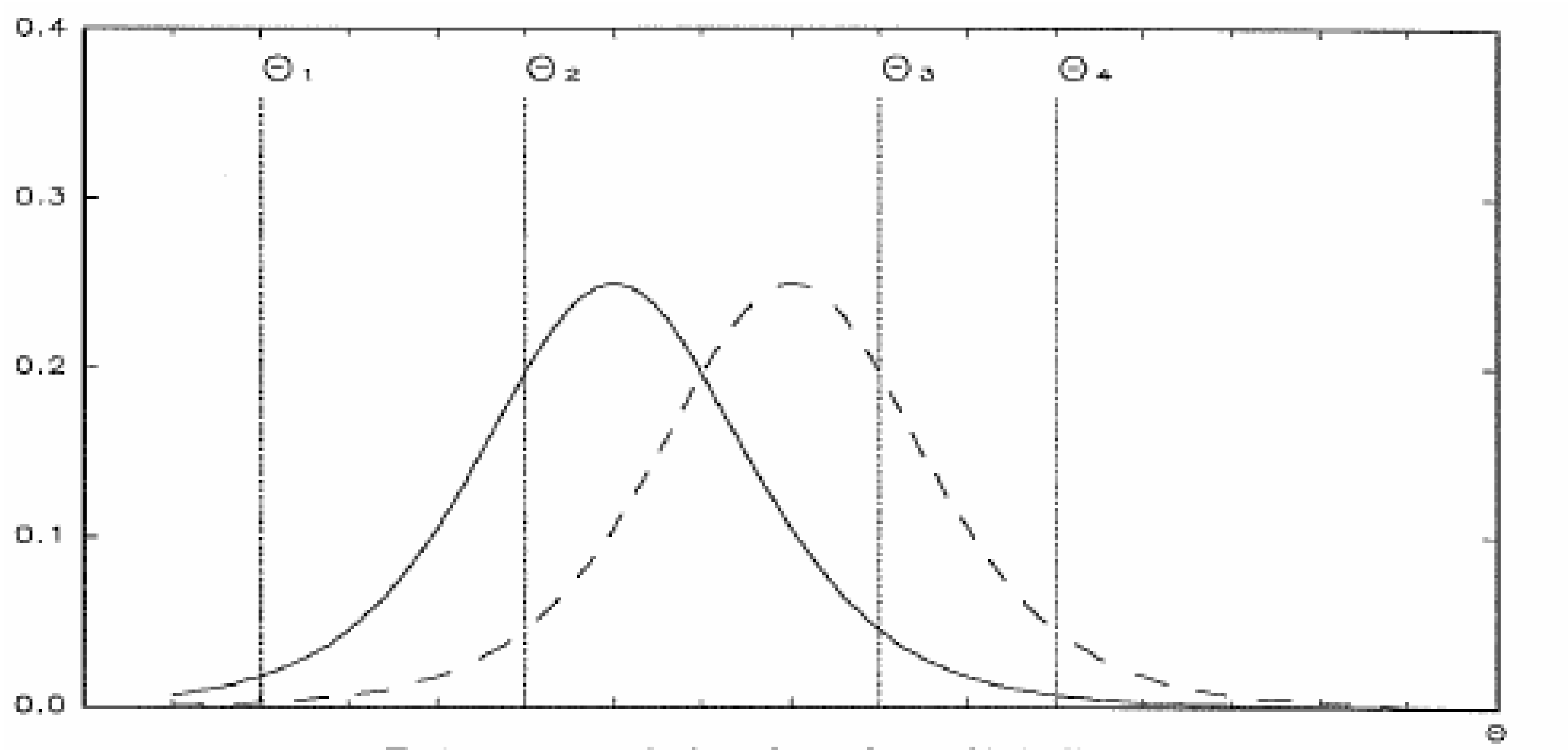
- Koch's 1990 data on a clinical trial for respiratory illness
- Treatment (A) vs Placebo (P)
- 111 patients (54 in A; 57 in P)
- Outcome: score from 0 (bad) to 4 (excellent)
- Explanatory variables
  - Center: 1,2
  - Treatment: A,P
  - Gender: M,F
  - Age: 3 classes
  - Visit: 4
  - Baseline: H,L

Example /Data

Table 1: Respiratory data (extract)

center	id	treatment	gender	age	baseline	visit	respstatus
1	53	A	F	32	1	1	2
1	53	A	F	32	1	2	2
1	53	A	F	32	1	3	4
1	53	A	F	32	1	4	2
1	18	A	F	47	2	1	2
1	18	A	F	47	2	2	3
1	18	A	F	47	2	3	4
1	18	A	F	47	2	4	4
1	54	A	M	11	4	1	4
1	54	A	M	11	4	2	4
1	54	A	M	11	4	3	4
1	54	A	M	11	4	4	2
1	12	A	M	14	2	1	3
1	12	A	M	14	2	2	3
1	12	A	M	14	2	3	3
1	12	A	M	14	2	4	2
1	51	A	M	15	0	1	2
1	51	A	M	15	0	2	3
1	51	A	M	15	0	3	3
1	51	A	M	15	0	4	3
1	20	A	M	20	3	1	3
1	20	A	M	20	3	2	2
1	20	A	M	20	3	3	3
1	20	A	M	20	3	4	1

Threshold concept



Threshold model

l_{ir} : observation r in stratum i

One assumes the existence of a latent continuous variable $L_{ir} \sim \mathcal{N}\left(\mu_i, \sigma_i^2\right)$

with thresholds: $\tau_1, \tau_2, ..., \tau_j, ..., \tau_{J\text{-}1}$

$$\kappa_{i1} = \pi_{i1} = \Pr(L_{ir} \leq \tau_1) = \Pr(\underbrace{\frac{L_{ir} - \mu_i}{\sigma_i}}_{\mathcal{N}(0,1)} \leq \frac{\tau_1 - \mu_i}{\sigma_i}) = \Phi\left(\frac{\tau_1 - \mu_i}{\sigma_i}\right)$$

$$\kappa_{i2} = \pi_{i1} + \pi_{i2} = \Pr(L_{ir} \leq \tau_2) = \Phi\left(\frac{\tau_2 - \mu_i}{\sigma_i}\right)$$

Threshold model/continued

1) Equivalence with $l_{ir} = \mathbf{x}_i^{'}\boldsymbol{\beta} + \mathbf{z}_i^{'}\mathbf{u} + \sigma_i e_{ir}^{*}$ and $y_{ijr} = 1 \Leftrightarrow \tau_{j-1} < l_{ir} \leq \tau_j$

2) Usually, one assumes $\sigma_i = \sigma = 1$, $\boxed{\kappa_{ij} = \Phi\left(\tau_j - \mu_i\right)}$

3) Φ may be replaced by other CDF's: logit, studit, gompit $\left[\log(-\log(1-x)\right]$

Note: probit, logit, studit: palindromic invariance

Models for scaling parameters

Model for scale factors: $\ln(\sigma_i) = \mathbf{p}_i^{'}\boldsymbol{\delta}$

$\mathbf{p}_\mathrm{i}$:vector of covariates with coefficients $\boldsymbol{\delta}$ (Mc Cullagh, 1980)

Extension to include random effects $\boxed{\ln(\sigma_i) = \mathbf{p}_i^{'}\boldsymbol{\delta} + \mathbf{q}_i^{'}\mathbf{v}}$

(Foulley et al,1992 for continuous data; Foulley & Gianola, 1996)

Statistical Inference

Full Bayesian Inference

$$[\boldsymbol{\beta},\mathbf{u},\boldsymbol{\delta},\mathbf{v},\boldsymbol{\tau},\mathbf{G},\boldsymbol{\Lambda} \mid \mathbf{y}] \propto$$

$$[\mathbf{y} \mid \boldsymbol{\beta},\mathbf{u},\boldsymbol{\delta},\mathbf{v},\boldsymbol{\tau}]: \text{Product Multinomial}$$

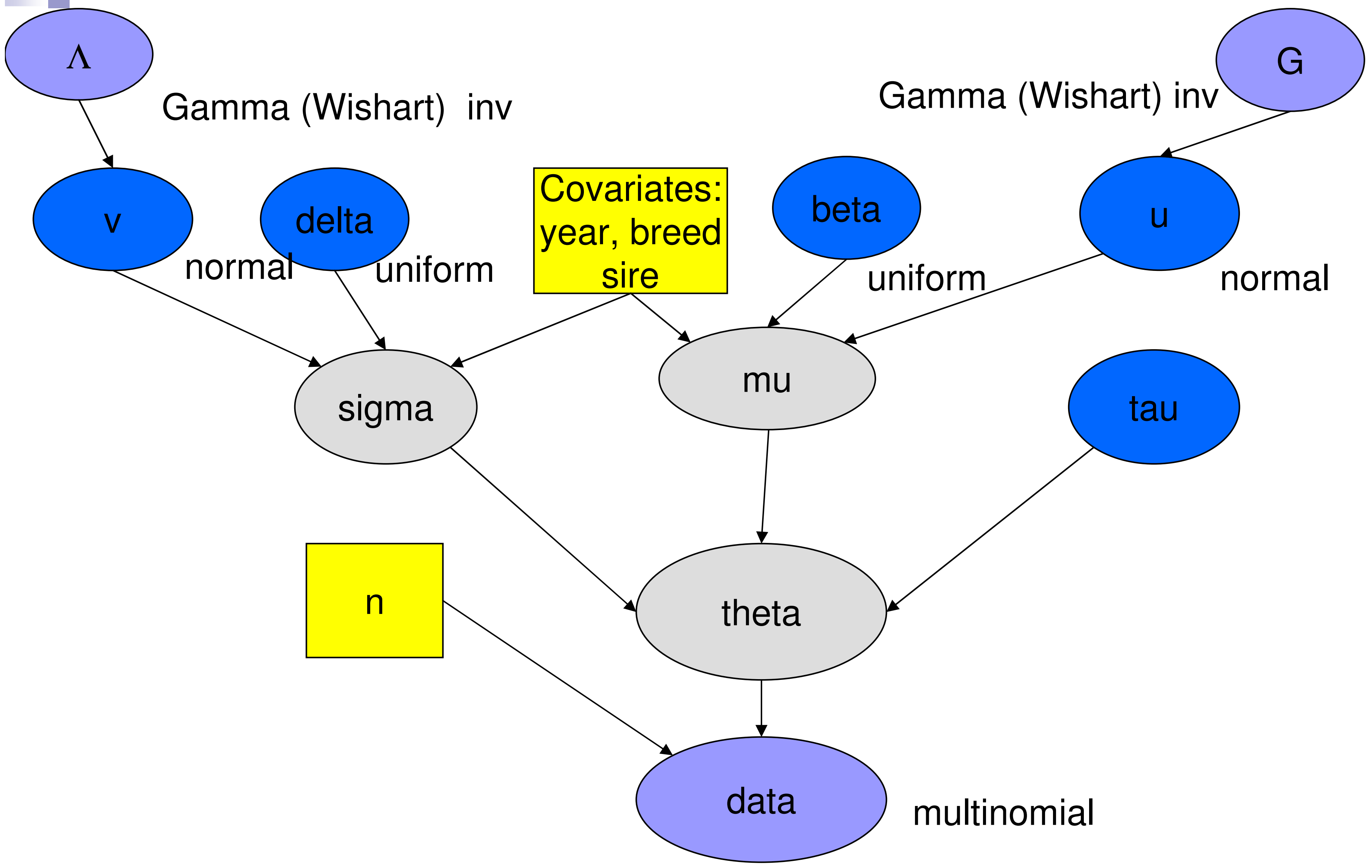
$$[\mathbf{u} \mid \mathbf{G}][\mathbf{v} \mid \boldsymbol{\Lambda}]: \text{Gaussian}$$

$$[\boldsymbol{\beta}][\boldsymbol{\delta}]: \text{Flat}$$

$$[\boldsymbol{\tau}]: \text{Product uniform on the } \Delta\tau_k\text{'}s$$

$$[\mathbf{G},\boldsymbol{\Lambda}]: \text{Inverse Wishart (Gamma)}$$

Graph of the model



Estimation « Fixed Model »/SAS logistic vs Winbugs

Parameters		SAS-Logistic		Winbugs	
		Estimate	SE	Estimate	SE
Intercept	4	-1.420	0.172	-1.429	0.171
	3	-0.723	0.165	-0.727	0.163
	2	0.213	0.164	0.212	0.164
	1	0.709	0.170	0.715	0.170
Center	2-1	0.095	0.157	0.098	0.158
Treatment	T-P	1.084	0.164	1.092	0.164
Gender	F-M	0.329	0.152	0.329	0.154
Age	1-4	0.299	0.156	0.297	0.158
	2-4	-0.085	0.157	-0.089	0.157
	3-4	-0.355	0.146	-0.357	0.147
Baseline	H-L	1.030	0.116	1.037	0.116
Center*Treat		-0.601	0.215	-0.605	0.214

Model comparison

Table 2: Model comparison for respiratory data

Model	Location							Scale		Comparison			
	C	T	G	A	B	CT	S	B	S	No	DIC	Pd	PPP
1	X	X	X	X	X	X				12	882	12	0.000
2	X	X	X	X	X	X	X			13	653	94	0.003
3		X			X		X			7	657	95	0.003
4		X			X		X	X	X	9	615	133	0.304
5		X			X		X		X	8	614	133	0.300

C: Center; T: Treatment; G: Gender; A: Age; B: Baseline;

CT: Center by Treatment Interaction; S: Subject as random

No: Number of parameters; DIC: Deviance Information Criterion; Pd: Complexity; PPP: Posterior predictive p-value

Estimation: Standard TM/Bugs vs Glimmix

Table : Openbugs vs SAS-Glimmix outputs for a mixed Standard TM

		Openbugs	**SAS-Glimmix**
		Estimate	Estimate
Intercept	4	-2.055±0.261	-1.949±0.224
	3	-0.994±0.245	-0.923±0.214
	2	0.437±0.243	0.434±0.213
	1	1.290±0.256	1.244±0.225
Fixed Effects	Treatment	1.074±0.288	0.987±0.249
	Base	1.643±0.298	1.498±0.253
Random Effects	Variance	1.798±0.388	1.337±0.253

Intercept (k)=mu-tau(k)
100(Glimmix-Openbugs)/Openbugs=-25.6

Estimation: Standard vs Heteroskedastic

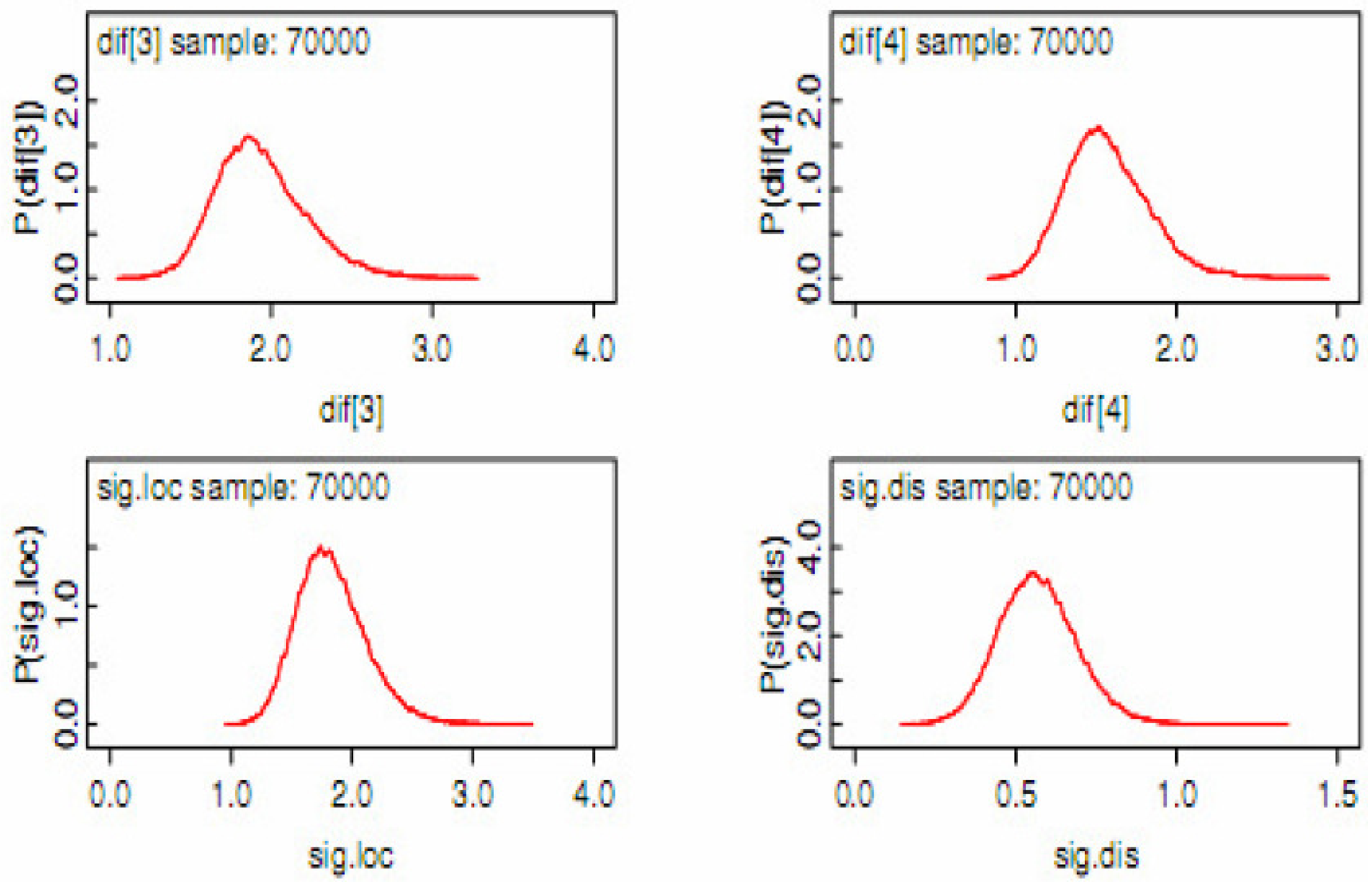
		Standard model		Heteroscedastic model	
		Estimate[6]	95% CS[7]	Estimate	95% CS
Threshold	$\tau_2 - \tau_1$ [1]	0.852	0.003:0.846	1.284	0.854:1.855
	$\tau_3 - \tau_2$	1.429	1.205:1.676	1.935	1.469:2.551
	$\tau_4 - \tau_3$	1.064	0.876:1.060	1.570	1.137:2.131
Location	$\mu_R - \tau_1$ [2]	1.290	0.794:1.795	1.870	1.092:2.800
	$T_2 - T_1$ [3]	1.074	0.502:1.636	1.448	0.653:2.336
	$B_2 - B_1$ [4]	1.643	1.077:2.236	2.466	1.555:3.590
	σ_S [5]	1.333	1.074:1.635	1.552	1.176:2.039
Scale	$B_2 - B_1$ [4']			0.419	0.051:0.798
	σ_S [5']			0.569	0.347:0.821
DIC		657		615	
FT[8]		542		422	

Predictions

Table 4: Observed vs Expected responses for two patients under the standard (S, No. 3) and heteroscedastic (H, No. 4) models for respiratory data

Subject		Category				
		0	1	2	3	4
13	Observed	0	0	4	0	0
	Expected-S	0.270	0.643	1.900	0.903	0.284
	Expected-H	0.048	0.377	3.014	0.518	0.044
55	Observed	0	0	4	0	0
	Expected-S	0.351	0.747	1.908	0.781	0.214
	Expected-H	0.062	0.445	3.019	0.439	0.035

Posteriors



dif[3] sample: 70000
P(dif[3])
0.0
1.0
2.0
1.0
2.0
3.0
4.0
dif[3]
dif[4] sample: 70000
P(dif[4])
0.0
1.0
2.0
0.0
1.0
2.0
3.0
dif[4]
sig.loc sample: 70000
P(sig.loc)
0.0
1.0
0.0
1.0
2.0
3.0
4.0
sig.loc
sig.dis sample: 70000
P(sig.dis)
0.0
2.0
4.0
0.0
0.5
1.0
1.5
sig.dis

Priors for dispersion parameters

1) $\boxed{\sigma^2 \sim \mathrm{U}\left(0,\Delta_{\mathrm{v}}\right)}$

2) $\boxed{\sigma \sim I_{(\sigma\geq 0)} A * \mathrm{C}\left(0,1\right)\ A \sim \mathcal{U}\left(0,\Delta_A\right)}$

3) $\boxed{\sigma \sim \mathcal{U}(0,\Delta_{\mathrm{s}})}$ si A $\rightarrow +\infty$ Gelman, 2006

4) $\boxed{\log\sigma \sim \mathcal{N}\left(0,\Delta_{\mathrm{L}}\right)}$

5) $\boxed{\sigma^{-2} \sim \mathcal{G}\left(\tfrac{1}{2}\eta, \tfrac{1}{2}\eta\underline{\sigma}^2\right)}$

η et $\underline{\sigma}^2$ connus en particulier η petit ie 2

6) $\boxed{\sigma^{-2} \sim \mathcal{G}\left(\varepsilon,\varepsilon\right)} \Leftrightarrow \left| \begin{array}{l} \ln\sigma^2 \sim \mathcal{U}\left[-\infty,+\infty\right] \text{ si } \varepsilon \rightarrow 0 \\ \pi\left(\sigma^2\right) \propto 1/\sigma^2\,(\textit{Jeffreys}) \end{array} \right.$

ε petit à calibrer en fonction de $\sum e^2 \left(\sum u^2\right)$

Conclusion

- Better efficiency of H-TM vs S-TM
- Large flexibility for scale models
  - Fixed and random effects
- Inference
  - Feasibility with Bayes via MCMC

References

Derquenne C (1995) Heteroskedastic Logit Model, 5oth Session of the ISI, Being, China.

Foulley JL, San Cristobal M, Gianola D, Im S (1990) Marginal likelihood and Bayesian approaches to the analysis of heterogeneous residual variances in mixed linear Gaussian models. *Computational Statistics & Data Analysis*, 13, 291-305.

Foulley JL, Gianola D (1996) Statistical analysis of ordered categorical data via a structural heteroskedastic threshold model. *Genetics Selection Evolution*, 28, 249-273.

Foulley JL, Jaffrézic F (2009) Modelling and estimating heterogeneous variances in threshold models for ordinal discrete data via Winbugs/Openbugs. *Computer Methods and Programs in Biomedicine*, in print

Jaffrézic F, Robert-Granié C, Foulley JL (1999) A quasi-score approach to the analysis of ordered categorical data via a mixed heteroskedastic threshold model. *Genetics Selection Evolution*, 31, 301-318.

Lee Y, Nelder JA (2006) Double hierarchical generalized linear models. *Applied Statistics*, 55, 139-185.

McCullagh P (1980) Regression models for ordinal data. *JR Statistical Society*, B 42, 109-142.

Liu I, Agresti A (2005) The analysis of ordered categorical data: an overview and a survey of recent developments. Societas de Estadistica e Investigacion Operativa, 14, 1-73.

Meza C, Jaffrézic F, Foulley JL (2009) Estimation in the probit model for binary outcomes using the SAEM algorithm. *Computational Statistics & Data Analysis*, 53,1350-1360