

Bayesian Methods in Multilevel Regression



Joop Hox
MuLOG, 15 september 2000



What is Statistics?

- Statistics is about uncertainty

“To err is human, to forgive divine, but to include errors in your design is statistical”

Leslie Kish, 1977
Presidential address A.S.A.



Uncertainty in Classical Statistics

- Uncertainty = *sampling distribution*
 - Estimate population parameter θ by $\hat{\theta}$
 - Imagine drawing an infinity of samples
 - Distribution of $\hat{\theta}$ over samples
- We have only one sample
 - Estimate $\hat{\theta}$ and its sampling distribution
 - Estimate 95% confidence interval



Inference in Classical Statistics

- What does *95% confidence interval* actually mean?
 - Over an infinity of samples, 95% of these contain the true population value θ
 - But we have only one sample
 - We *never* know if our present estimate $\hat{\mu}$ and confidence interval is one of those 95% or not



Inference in Classical Statistics

- What does *95% confidence interval* NOT mean?
 - ~~We have a 95% probability that the true population value θ is within the limits of our confidence interval~~
 - We only have an aggregate assurance that in the long run 95% of our confidence intervals contain the true population value



Uncertainty in Bayesian Statistics

- Uncertainty = probability distribution for the population parameter
 - In classical statistics the population parameter θ has one single true value
 - Only we happen to not know it
 - In Bayesian statistics we imagine a distribution of possible values of population parameter θ
- Each unknown parameter must have an associated probability distribution



Uncertainty in Bayesian Statistics

- Each unknown parameter *must* have an associated probability distribution
 - Before we have data: prior distribution
 - After we have data:
posterior distribution = f (prior + data)
 - Posterior distribution used to find estimate for θ and confidence interval
 - $\hat{\theta}$ = mode
 - confidence interval = central 95% region



Inference in Bayesian Statistics

- Posterior distribution to estimate θ and confidence interval
 - Posterior = f (prior + data)
 - Prior distribution influences posterior
 - Bayesian statistical inference depends partly on the prior
 - Which does *not* depend on the data
 - (in *empirical Bayes* it does...)



Inference in Bayesian Statistics

- Bayesian statistical inference depends partly on the prior, so: which prior?
 - Technical considerations
 - Conjugate prior (posterior belongs to same distribution family)
 - Proper prior (real probability distribution)
 - Fundamental consideration
 - Informative prior or ignorance prior?
 - Total ignorance does not exist ... all priors add some information to the data



Computational Issues in Bayesian Statistics

- Posterior distribution used to find estimate for θ and confidence interval
 - $\hat{\theta} = \text{mode}$
 - confidence interval = central 95% region
- Assumes simple posterior distribution, so we can compute its characteristics
- In complex models, the posterior is often intractable



Computational Issues in Bayesian Statistics

- In complex models, the posterior is often intractable
- Solution: approximate posterior by simulation
 - Simulate many draws from posterior distribution
 - Compute mode, mean, 95% interval et cetera from simulated draws



Why Bayesian Statistics?

- Can do some things that cannot be done in classical statistics
- Valid in small samples
 - Maximum Likelihood is not
 - “Asymptotically we are all dead ...” (Novick)
- Always proper estimates
 - No negative variances



Why Bayesian Statistics?

- But prior information induces bias
 - Biased estimates
 - But hopefully more precise

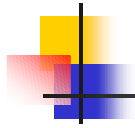
“In a corner of the forest, Dwells alone my Hiawatha
Permanently cogitating, On the normal law of error
Wondering in idle moments, Whether an increased precision
Might perhaps be rather better, *Even at the risk of bias*
If thereby one, now and then, Could register upon the target

Kendall, 1959, 'Hiawatha designs an experiment'
(Italics mine)



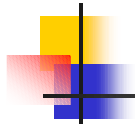
Simulating the Posterior Distribution

- Markov Chain Monte Carlo (MCMC)
 - Gibbs sampling
 - Metropolis-Hastings
- Given a draw from a specific probability distribution, MCMC produces a new pseudorandom draw from that distribution
 - Distributions typically multivariate



MCMC Issues: Burn In

- Sequence of draws $Z^{(1)} \rightarrow Z^{(2)} \rightarrow \dots \rightarrow Z^{(t)}$
 - From target distribution $f(Z)$
 - Even if $Z^{(1)}$ not from $f(Z)$, the distribution of $Z^{(t)}$ is $f(Z)$, as $t \rightarrow \infty$
 - So, for arbitrary $Z^{(1)}$, if t is sufficiently large, $Z^{(t)}$ is from target distribution $f(Z)$
 - But having good starting values helps
- MCMC must run t iterations 'burn in' before we reach target distribution $f(Z)$



MCMC Issues: Burn In

- MCMC must run t iterations 'burn in' before we reach target distribution $f(Z)$
 - How many iterations are needed to converge on the target distribution?
- Diagnostics
 - examine graph of burn in
 - try different starting values



MCMC Issues: Monitoring

- How many iterations must be monitored?
 - Depends on required accuracy
 - Problem: successive draws are correlated
- Diagnostics
 - Graph successive draws
 - Compute autocorrelations
 - Raftery-Lewis: \hat{n} = minimum for quantile
 - Brooks-Draper: \hat{n} = minimum for mean



Bayesian Methods in Multilevel Regression

- Software

- *BUGS*

- Bayesian inference Using Gibbs Sampling
 - Very general, difficult to use

- *MLwiN*

- Special implementation for multilevel regression
 - Limitations
 - No complex 1st level variances
 - No multivariate models
 - No extrabinomial variation

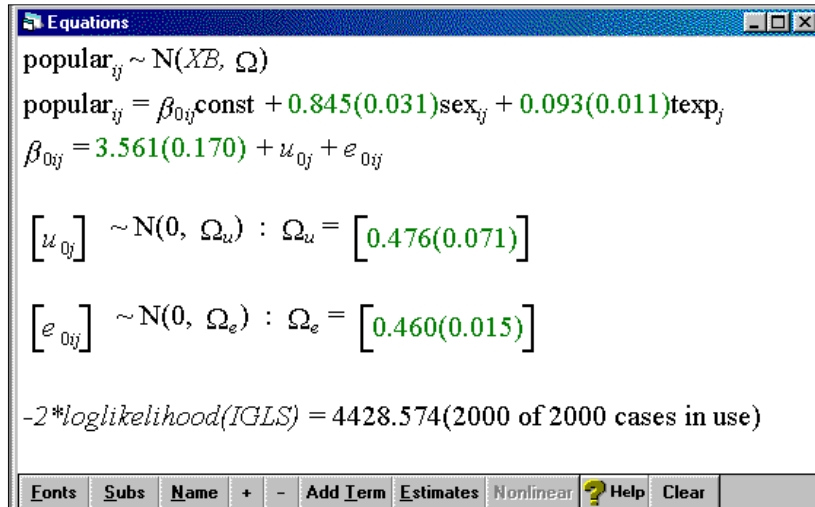


Example Data

- Popularity data
 - Pupil popularity (0 ... 10)
 - Pupil sex (0=boy, 1=girl)
 - Teacher experience (years)
- 2000 pupils, 100 classes

(artificial data)

Popularity Example: FML Estimates



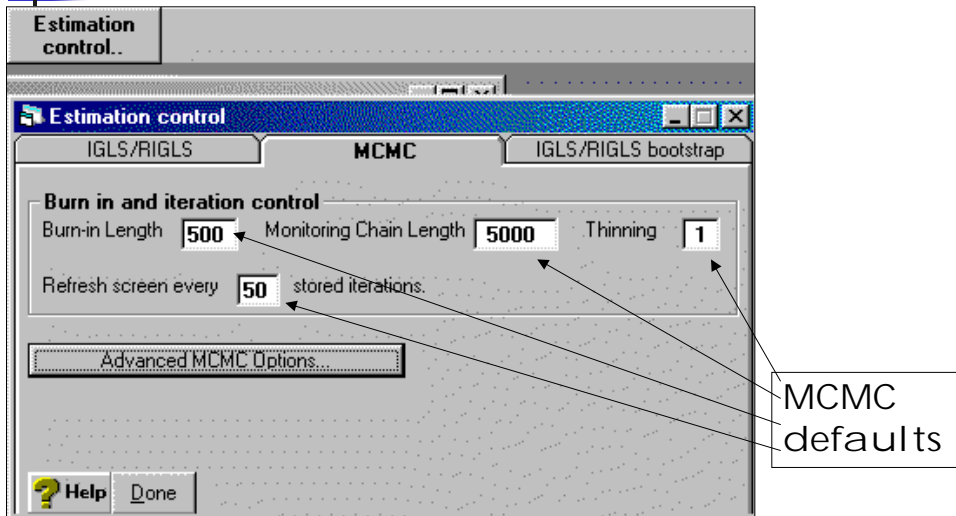
The screenshot shows a window titled "Equations" with the following content:

$$\text{popular}_{ij} \sim N(XB, \Omega)$$
$$\text{popular}_{ij} = \beta_{0ij} \text{const} + 0.845(0.031) \text{sex}_{ij} + 0.093(0.011) \text{texp}_j$$
$$\beta_{0ij} = 3.561(0.170) + u_{0ij} + e_{0ij}$$
$$\begin{bmatrix} u_{0ij} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.476(0.071) \end{bmatrix}$$
$$\begin{bmatrix} e_{0ij} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.460(0.015) \end{bmatrix}$$

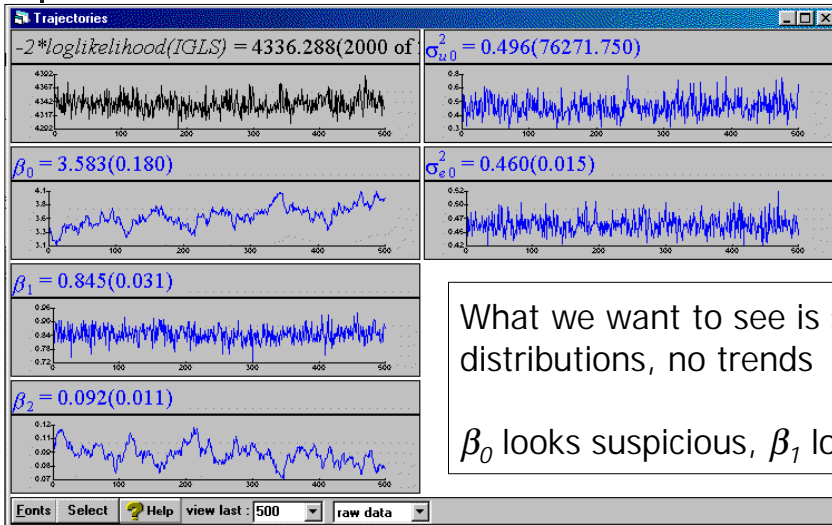
$-2 * \text{loglikelihood(IGLS)} = 4428.574(2000 \text{ of } 2000 \text{ cases in use})$

The window has a menu bar with the following items: **Fonts**, **Subs**, **Name**, **+**, **-**, **Add Item**, **Estimates**, **Nonlinear**, **? Help**, **Clear**.

Popularity Example: MLwiN MCMC Window



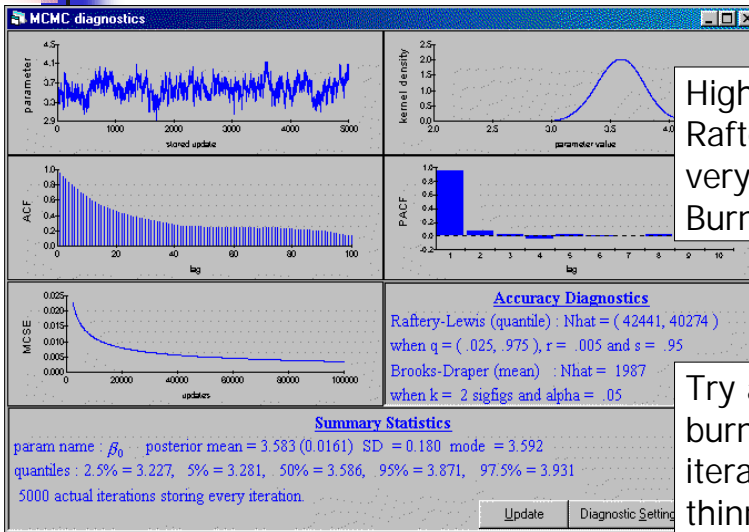
Popularity Example: Last 500 Iterations



What we want to see is stable distributions, no trends

β_0 looks suspicious, β_1 looks fine

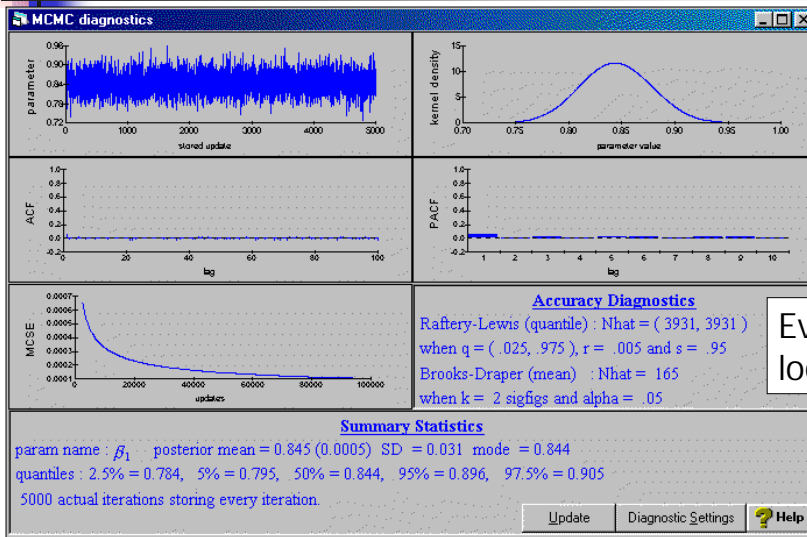
Popularity Example: Diagnostics for Posterior of β_0



High autocorrelation,
Raftery-Lewis *nhat*
very high
Burn in too short?

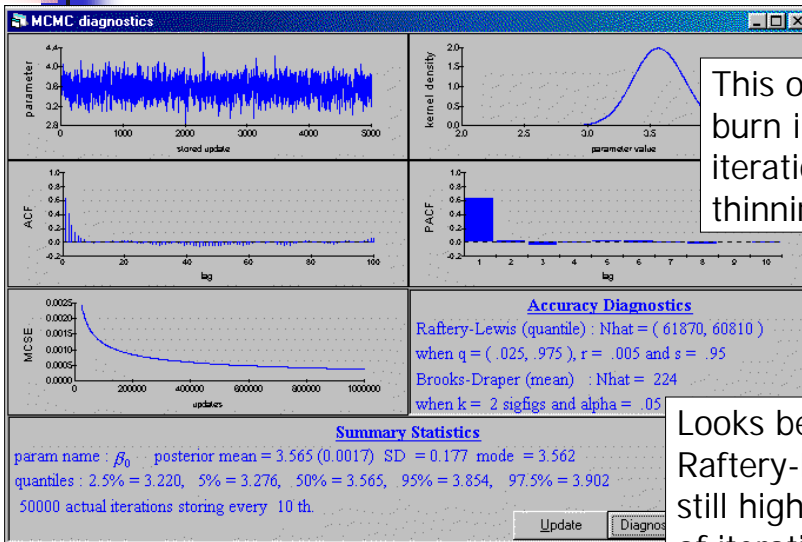
Try again:
burn in = 5000
iterations = 50000
thinning = 10

Popularity Example: Diagnostics for Posterior of β_1



Everything looks fine!

Popularity Example: Diagnostics for Posterior of β_0



This one with
burn in = 5000
iterations = 50000
thinning = 10

Looks better,
Raftery-Lewis *nhat*
still higher than nr
of iterations

Comparison of Popularity Example Estimates

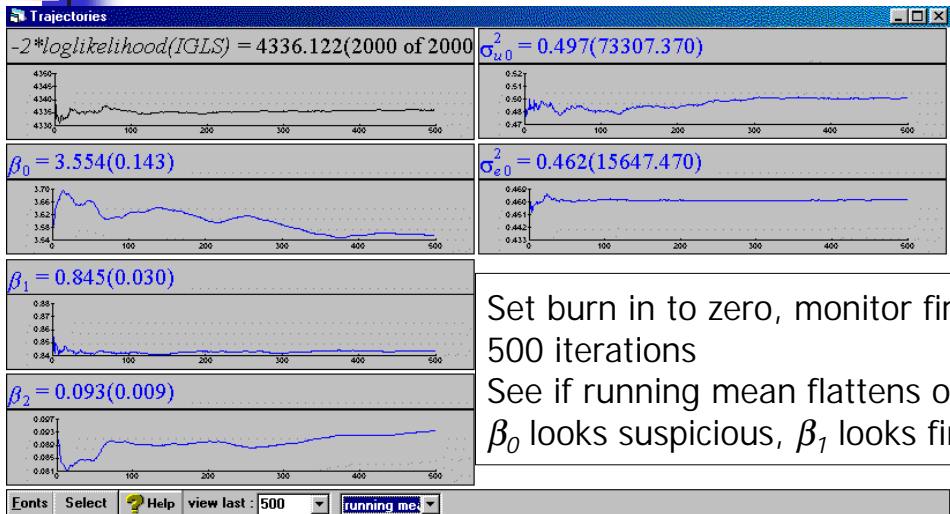


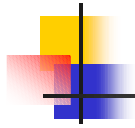
Different estimates popularity example

Model:	FML	MCMC, 50000/5000	MCMC, 5000/500
Fixed part Predictor	coefficient	posterior mode s.d.	posterior mode s.d.
intercept	3.56 .17	3.57 .18	3.58 .18
pupil sex	0.85 .03	0.85 .03	0.85 .03
teacher exp.	0.09 .01	0.09 .01	0.09 .01
Random part			
intercept ₁	0.46 .02	0.46 .02	0.46 .02
intercept ₂	0.48 .07	0.50 .08	0.50 .08



Monitoring Burn In

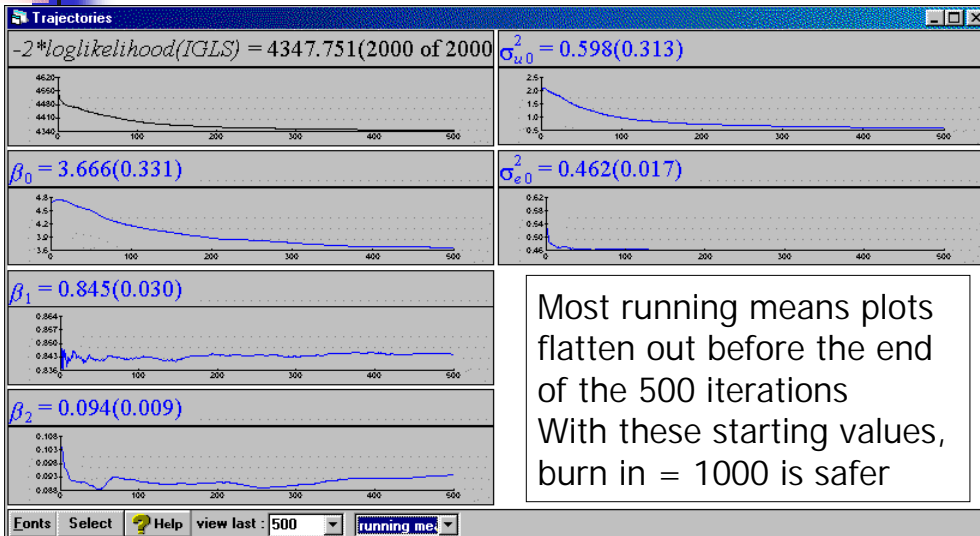




Monitoring Convergence

- Different starting values for the chain
 - MLwiN: try IGLS & RIGLS
 - Run both, should give same results
- Illustration: enter starting values manually
 - Popularity example
 - Intercept 5, sex 0.5, experience 0.1
 - Variance school level 0.2, pupil level 0.8
 - Then set burn in to zero, monitor 500 iterations

Manual Starting Values (poor estimates)



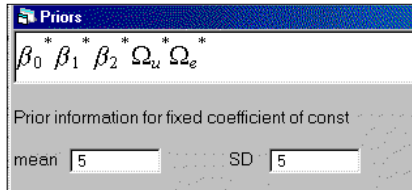


Priors in MLwiN

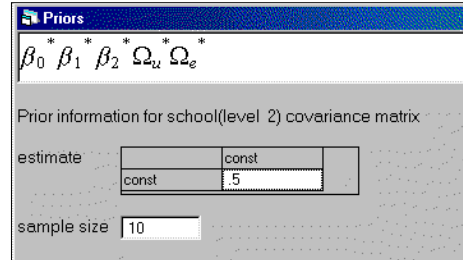
- Default: uninformative priors
 - Flat, diffuse priors, ignorance prior
 - Fixed coefficients:
 - Uniform prior: $P(\beta) \propto 1$
 - Single variance
 - Diffuse: $P(1/\sigma^2) \sim \text{Gamma}(\epsilon, \epsilon)$, ϵ small
 - Which is close to uniform prior for $\log(\sigma^2)$
 - Covariance matrix
 - $P(\Omega^{-1}) \sim \text{Wishart}(\hat{\Omega})$
 - Which is somewhat informative ($N = \#$ variances)

Example with Informative Priors

- For fixed coefficients the prior is $N(\mu, \sigma^2)$



The screenshot shows a dialog box titled "Priors" with a list of parameters: β_0^* , β_1^* , β_2^* , Ω_u^* , and Ω_e^* . Below the list, it says "Prior information for fixed coefficient of const". There are two input fields: "mean" with the value "5" and "SD" with the value "5".



The screenshot shows a dialog box titled "Priors" with the same parameter list as the first dialog. Below the list, it says "Prior information for school(level 2) covariance matrix". There is an "estimate" section with a table:

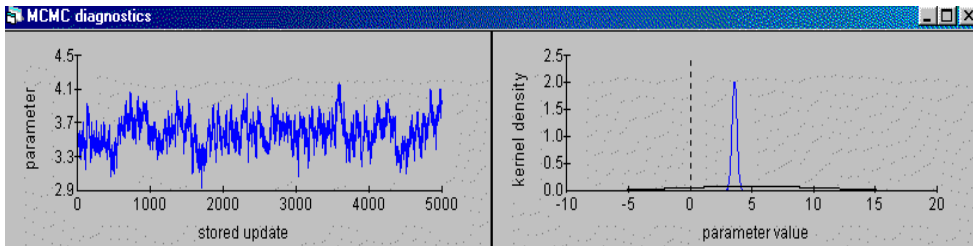
	const
const	5

Below the table, there is a "sample size" input field with the value "10".

- For variances enter a value and the number of observations it is worth

Results with Informative Priors

- For fixed coefficients the prior is $N(\mu, \sigma^2)$
 - Intercept $N(5, 5)$, sex $N(.2, 5)$, age $N(.1, 5)$
 - No prior on variances
- Diagnostics for β_0 :





Copies on

<http://www.fss.uu.nl/ms/jh>

