Generalized linear models

3 Generalized linear models
- Generalisation of linear models
- Metropolis–Hastings algorithms
- The Probit Model
- The logit model
- Loglinear models

Generalisation of Linear Models

Linear models model connection between a response variable y and a set x of explanatory variables by a linear dependence relation with [approximately] normal perturbations.

Many instances where either of these assumptions not appropriate, e.g. when the support of y restricted to $\mathbb{R}_+$ or to $\mathbb{N}$.

bank

Four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones:

- x_1 length of the bill (in mm),
- x_2 width of the left edge (in mm),
- x_3 width of the right edge (in mm),
- x_4 bottom margin width (in mm).

Response variable y: status of the banknote [0 for genuine and 1 for counterfeit]

Probabilistic model that predicts counterfeiting based on the four measurements

The impossible linear model

Example of the influence of x_4 on y
Since y is binary,

$$y|x_4 \sim \mathscr{B}(p(x_4))\,,$$

ⓒ **Normal model is impossible**

Linear dependence in $p(x) = \mathbb{E}[y|x]$'s

$$p(x_{4i}) = \beta_0 + \beta_1 x_{4i}\,,$$

estimated [by MLE] as

$$\hat{p}_i = -2.02 + 0.268\, x_{i4}$$

which gives $\hat{p}_i = .12$ for $x_{i4} = 8$ and ... $\hat{p}_i = 1.19$ for $x_{i4} = 12!!!$
ⓒ **Linear dependence is impossible**

Generalisation of the linear dependence

Broader class of models to cover various dependence structures.

Class of *generalised linear models* (GLM) where

$$y|\mathbf{x}, \beta \sim f(y|\mathbf{x}^{\mathrm{T}}\beta)\,.$$

i.e., dependence of y on $\mathbf{x}$ partly *linear*

Notations

Same as in linear regression chapter, with n–sample

$$\mathbf{y} = (y_1, \ldots, y_n)$$

and corresponding explanatory variables/covariates

$$X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

Specifications of GLM's

Definition (GLM)

A GLM is a conditional model specified by two functions:

1. the density f of y given $\mathbf{x}$ parameterised by its expectation parameter $\mu = \mu(\mathbf{x})$ [and possibly its dispersion parameter $\varphi = \varphi(\mathbf{x})$]
2. the *link* g between the mean μ and the explanatory variables, written customarily as $g(\mu) = \mathbf{x}^{\mathrm{T}}\beta$ or, equivalently, $\mathbb{E}[y|\mathbf{x}, \beta] = g^{-1}(\mathbf{x}^{\mathrm{T}}\beta)$.

For identifiability reasons, g needs to be bijective.

Likelihood

Obvious representation of the likelihood

$$\ell(\beta, \varphi|\mathbf{y}, X) = \prod_{i=1}^{n} f\left(y_i|\mathbf{x}^{i\mathrm{T}}\beta, \varphi\right)$$

with parameters $\beta \in \mathbb{R}^k$ and $\varphi > 0$.

Examples

- Ordinary linear regression

  Case of GLM where

$$g(x) = x,\ \varphi = \sigma^2, \quad \text{and} \quad \mathbf{y}|X, \beta, \sigma^2 \sim \mathscr{N}_n(X\beta, \sigma^2).$$

Examples (2)

Case of binary and binomial data, when

$$y_i|\mathbf{x}^i \sim \mathscr{B}(n_i, p(\mathbf{x}^i))$$

with known n_i

- Logit [or logistic regression] model
  Link is *logit transform* on probability of success

  $$g(p_i) = \log(p_i/(1-p_i))\,,$$

  with likelihood

  $$\prod_{i=1}^n \binom{n_i}{y_i} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{y_i} \left(\frac{1}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{n_i-y_i}$$
  $$\propto \exp\left\{\sum_{i=1}^n y_i \mathbf{x}^{i\mathrm{T}}\beta\right\} \Big/ \prod_{i=1}^n \left(1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right)^{n_i-y_i}$$

Canonical link

Special link function g that appears in the natural exponential family representation of the density

$$g^{\star}(\mu) = \theta \quad \text{if} \quad f(y|\mu) \propto \exp\{T(y) \cdot \theta - \Psi(\theta)\}$$

Example

Logit link is canonical for the binomial model, since

$$f(y_i|p_i) = \binom{n_i}{y_i} \exp\left\{ y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1-p_i) \right\},$$

and thus

$$\theta_i = \log p_i/(1-p_i)$$

Examples (3)

Customary to use the canonical link, but only customary ...

- Probit model

  Probit link function given by

$$g(\mu_i) = \Phi^{-1}(\mu_i)$$

  where Φ standard normal cdf

  Likelihood

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{iT}\beta)^{y_i}(1 - \Phi(\mathbf{x}^{iT}\beta))^{n_i - y_i}.$$

Full processing

Log-linear models

Standard approach to describe associations between several *categorical* variables, i.e, variables with finite support

Sufficient statistic: *contingency table*, made of the cross-classified counts for the different categorical variables. Full entry to loglinear models

Example (Titanic survivor)

		Child		Adult	
Survivor	Class	Male	Female	Male	Female
No	1st	0	0	118	4
	2nd	0	0	154	13
	3rd	35	17	387	89
	Crew	0	0	670	3
Yes	1st	5	1	57	140
	2nd	11	13	14	80
	3rd	13	14	75	76
	Crew	0	0	192	20

Poisson regression model

1. Each count y_i is Poisson with mean $\mu_i = \mu(\mathbf{x}_i)$
2. Link function connecting $\mathbb{R}^+$ with $\mathbb{R}$, e.g. logarithm $g(\mu_i) = \log(\mu_i)$.

Corresponding likelihood

$$\ell(\beta|y, X) = \prod_{i=1}^{n} \left(\frac{1}{y_i!}\right) \exp\left\{y_i \mathbf{x}^{i\mathrm{T}}\beta - \exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\} .$$

Metropolis–Hastings algorithms

Posterior inference in GLMs harder than for linear models

© Working with a GLM requires specific numerical or simulation tools [E.g., GLIM in classical analyses]

Opportunity to introduce universal MCMC method: *Metropolis–Hastings* algorithm

Generic MCMC sampler

- Metropolis–Hastings algorithms are generic/down-the-shelf MCMC algorithms
- Only require likelihood up to a constant [difference with Gibbs sampler]
- can be tuned with a wide range of possibilities [difference with Gibbs sampler & blocking]
- natural extensions of standard simulation algorithms: based on the choice of a *proposal* distribution [difference in Markov proposal $q(x, y)$ and acceptance]

Why Metropolis?

Originally introduced by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller in a setup of optimization on a discrete state-space. All authors involved in Los Alamos during and after WWII:

- Physicist and mathematician, Nicholas Metropolis is considered (with Stanislaw Ulam) to be the father of Monte Carlo methods.
- Also a physicist, Marshall Rosenbluth worked on the development of the hydrogen (H) bomb
- Edward Teller was one of the first scientists to work on the Manhattan Project that led to the production of the A bomb. Also managed to design with Ulam the H bomb.

Generic Metropolis–Hastings sampler

For *target* π and proposal kernel $q(x, y)$

Initialization: Choose an arbitrary $x^{(0)}$

Iteration t**:**

1. Given $x^{(t-1)}$, generate $\tilde{x} \sim q(x^{(t-1)}, x)$
2. Calculate

$$\rho(x^{(t-1)}, \tilde{x}) = \min\left(\frac{\pi(\tilde{x})/q(x^{(t-1)}, \tilde{x})}{\pi(x^{(t-1)})/q(\tilde{x}, x^{(t-1)})}, 1\right)$$

3. With probability $\rho(x^{(t-1)}, \tilde{x})$ accept $\tilde{x}$ and set $x^{(t)} = \tilde{x}$; otherwise reject $\tilde{x}$ and set $x^{(t)} = x^{(t-1)}$.

Universality

Algorithm only needs to simulate from

$$q$$

which can be chosen [almost!] arbitrarily, i.e. unrelated with π [q also called *instrumental* distribution]

Note: π and q known up to proportionality terms ok since proportionality constants cancel in ρ.

Validation

Markov chain theory

Target π is stationary distribution of Markov chain $(x^{(t)})_t$ because probability $\rho(x, y)$ satisfies *detailed balance equation*

$$\pi(x)q(x, y)\rho(x, y) = \pi(y)q(y, x)\rho(y, x)$$

[Integrate out x to see that π is stationary]

For convergence/ergodicity, Markov chain must be *irreducible*: q has positive probability of reaching all areas with positive π probability in a finite number of steps.

Choice of proposal

Theoretical guarantees of convergence very high, but choice of q is crucial in practice. Poor choice of q may result in

- very high rejection rates, with very few moves of the Markov chain $(x^{(t)})_t$ hardly moves, or in
- a myopic exploration of the support of π, that is, in a dependence on the starting value $x^{(0)}$, with the chain stuck in a neighbourhood mode to $x^{(0)}$.

Note: hybrid MCMC

Simultaneous use of different kernels valid *and* recommended

The independence sampler

Pick proposal q that is independent of its first argument,

$$q(x, y) = q(y)$$

ρ simplifies into

$$\rho(x, y) = \min\left(1, \frac{\pi(y)/q(y)}{\pi(x)/q(x)}\right) .$$

Special case: $q \propto \pi$

Reduces to $\rho(x, y) = 1$ and iid sampling

Analogy with Accept-Reject algorithm where $\max \pi/q$ replaced with the current value $\pi(x^{(t-1)})/q(x^{(t-1)})$ but sequence of accepted $x^{(t)}$'s not i.i.d.

Choice of q

Convergence properties highly dependent on q.

- q needs to be positive everywhere on the support of π
- for a good exploration of this support, π/q needs to be bounded.

Otherwise, the chain takes too long to reach regions with low q/π values.

The random walk sampler

Independence sampler requires too much global information about π: opt for a local gathering of information

Means exploration of the neighbourhood of the current value $x^{(t)}$ in search of other points of interest.

Simplest exploration device is based on random walk dynamics.

Random walks

Proposal is a symmetric transition density

$$q(x,y) = q_{RW}(y-x) = q_{RW}(x-y)$$

Acceptance probability $\rho(x,y)$ reduces to the simpler form

$$\rho(x,y) = \min\left(1, \frac{\pi(y)}{\pi(x)}\right) .$$

Only depends on the target π *[accepts all proposed values that increase π]*

Choice of q_{RW}

Considerable flexibility in the choice of q_{RW},

- tails: Normal versus Student's t
- scale: size of the neighbourhood

Can also be used for restricted support targets [with a waste of simulations near the boundary]

Can be tuned towards an acceptance probability of 0.234 at the *burnin* stage *[Magic number!]*

Convergence assessment

Capital question: How many iterations do we need to run???

- **Rule # 1** There is no absolute number of simulations, i.e. $1,000$ is neither large, nor small.
- **Rule # 2** It takes [much] longer to check for convergence than for the chain itself to converge.
- **Rule # 3** MCMC is a *"what-you-get-is-what-you-see"* algorithm: it fails to tell about unexplored parts of the space.
- **Rule # 4** When in doubt, run MCMC chains in parallel and check for consistency.

Many "quick-&-dirty" solutions in the literature, but not necessarily trustworthy.

Prohibited dynamic updating

$\lightning$ Tuning the proposal in terms of its past performances can only be implemented at *burnin*, because otherwise this cancels Markovian convergence properties.

Use of several MCMC proposals together within a single algorithm using circular or random design is ok. It almost always brings an improvement compared with its individual components (at the cost of increased simulation time)

Effective sample size

How many iid simulations from π are equivalent to N simulations from the MCMC algorithm?

Based on estimated k-th order auto-correlation,

$$\rho_k = \text{cov}\left(x^{(t)}, x^{(t+k)}\right) ,$$

effective sample size

$$N^{\text{ess}} = n\left(1 + 2\sum_{k=1}^{T_0} \hat{\rho}_k\right)^{-1/2} ,$$

↯ Only partial indicator that fails to signal chains stuck in one mode of the target

The Probit Model

Likelihood Recall Probit

$$\ell(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}^{iT}\beta)^{y_i}(1 - \Phi(\mathbf{x}^{iT}\beta))^{n_i - y_i}\,.$$

If no prior information available, resort to the flat prior $\pi(\beta) \propto 1$ and then obtain the posterior distribution

$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \Phi\left(\mathbf{x}^{iT}\beta\right)^{y_i}\left(1 - \Phi(\mathbf{x}^{iT}\beta)\right)^{n_i - y_i}\,,$$

nonstandard and simulated using MCMC techniques.

MCMC resolution

Metropolis–Hastings random walk sampler works well for binary regression problems with small number of predictors

Uses the maximum likelihood estimate $\hat{\beta}$ as starting value and asymptotic (Fisher) covariance matrix of the MLE, $\hat{\Sigma}$, as scale

MLE proposal

R function glm very useful to get the maximum likelihood estimate of β and its asymptotic covariance matrix $\hat{\Sigma}$.

Terminology used in R program

```
mod=summary(glm(y~X-1,family=binomial(link="probit")))
```

with mod$coeff[,1] denoting $\hat{\beta}$ and mod$cov.unscaled $\hat{\Sigma}$.

MCMC algorithm

Probit random-walk Metropolis-Hastings

Initialization: Set $\beta^{(0)} = \hat{\beta}$ and compute $\hat{\Sigma}$

Iteration t:

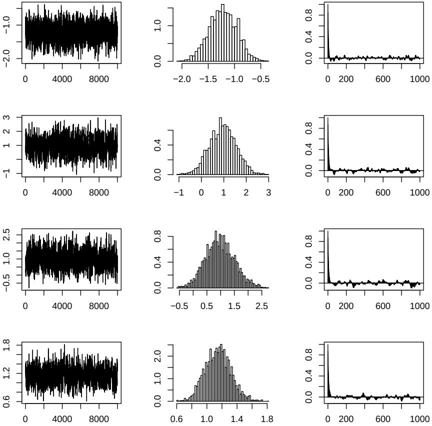
1. Generate $\tilde{\beta} \sim \mathscr{N}_{k+1}(\beta^{(t-1)}, \tau\hat{\Sigma})$
2. Compute

$$\rho(\beta^{(t-1)}, \tilde{\beta}) = \min\left(1, \frac{\pi(\tilde{\beta}|y)}{\pi(\beta^{(t-1)}|y)}\right)$$

3. With probability $\rho(\beta^{(t-1)}, \tilde{\beta})$ set $\beta^{(t)} = \tilde{\beta}$; otherwise set $\beta^{(t)} = \beta^{(t-1)}$.

bank

Probit modelling with no intercept over the four measurements. Three different scales $\tau = 1, 0.1, 10$: best mixing behavior is associated with $\tau = 1$. Average of the parameters over $9,000$ iterations gives plug-in estimate



$$\hat{p}_i = \Phi\left(-1.2193 x_{i1} + 0.9540 x_{i2} + 0.9795 x_{i3} + 1.1481 x_{i4}\right).$$

G-priors for probit models

Flat prior on β inappropriate for comparison purposes and Bayes factors.
Replace the flat prior with a hierarchical prior,

$$\beta|\sigma^2, X \sim \mathscr{N}_k\left(0_k, \sigma^2(X^{\mathrm{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2},$$

as in normal linear regression

Note

The matrix $X^{\mathrm{T}}X$ is *not* the Fisher information matrix

G-priors for testing

Same argument as before: while π is improper, use of the *same* variance factor σ^2 in both models means the normalising constant cancels in the Bayes factor.

Posterior distribution of β

$$\begin{aligned}\pi(\beta|\mathbf{y},X) \quad &\propto \quad |X^{\mathrm{T}}X|^{1/2}\Gamma((2k-1)/4)\left(\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\beta\right)^{-(2k-1)/4}\pi^{-k/2} \\ &\quad \times\prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathrm{T}}\beta)\right]^{1-y_i}\end{aligned}$$

[where k matters!]

Marginal approximation

Marginal

$$
\begin{aligned}
f(\mathbf{y}|X) \quad \propto \quad & |X^{\mathrm{T}}X|^{1/2}\,\pi^{-k/2}\Gamma\{(2k-1)/4\}\int \left(\beta^{\mathrm{T}}(X^{\mathrm{T}}X)\beta\right)^{-(2k-1)/4} \\
& \times \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathrm{T}}\beta)^{y_i}\left[1-(\Phi(\mathbf{x}^{i\mathrm{T}}\beta)\right]^{1-y_i} d\beta\,,
\end{aligned}
$$

approximated by

$$
\begin{aligned}
\frac{|X^{\mathrm{T}}X|^{1/2}}{\pi^{k/2}M} \sum_{m=1}^{M} \left|\left|X\beta^{(m)}\right|\right|^{-(2k-1)/2} \prod_{i=1}^{n} \Phi(\mathbf{x}^{i\mathrm{T}}\beta^{(m)})^{y_i}\left[1-\Phi(\mathbf{x}^{i\mathrm{T}}\beta^{(m)})\right]^{1-y_i} \\
\times\, \Gamma\{(2k-1)/4\}\,|\widehat{V}|^{1/2}(4\pi)^{k/2}\,e^{(\beta^{(m)}-\widehat{\beta})^{\mathrm{T}}\widehat{V}^{-1}(\beta^{(m)}-\widehat{\beta})/4}\,,
\end{aligned}
$$

where

$$
\beta^{(m)} \sim \mathcal{N}_k(\widehat{\beta}, 2\,\widehat{V})
$$

with $\widehat{\beta}$ MCMC approximation of $\mathbb{E}^{\pi}[\beta|\mathbf{y}, X]$ and $\widehat{V}$ MCMC approximation of $\mathbb{V}(\beta|\mathbf{y}, X)$.

Linear hypothesis

Linear restriction on β

$$H_0 : R\beta = r$$

($r \in \mathbb{R}^q$, R $q \times k$ matrix) where β^0 is $(k-q)$ dimensional and X_0 and $\mathbf{x}_0$ are linear transforms of X and of $\mathbf{x}$ of dimensions $(n, k-q)$ and $(k-q)$.

Likelihood

$$\ell(\beta^0|\mathbf{y}, X_0) \propto \prod_{i=1}^{n} \Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)^{y_i} \left[1 - \Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)\right]^{1-y_i} ,$$

Linear test

Associated [projected] G-prior

$$\beta^0|\sigma^2, X_0 \sim \mathscr{N}_{k-q}\left(0_{k-q}, \sigma^2(X_0^{\mathrm{T}}X_0)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X_0) \propto \sigma^{-3/2},$$

Marginal distribution of $\mathbf{y}$ of the same type

$$\begin{aligned} f(\mathbf{y}|X_0) \quad &\propto \quad |X_0^{\mathrm{T}}X_0|^{1/2}\pi^{-(k-q)/2}\Gamma\left\{\frac{(2(k-q)-1)}{4}\right\}\int \left|\left|X\beta^0\right|\right|^{-(2(k-q)-1)/2} \\ &\prod_{i=1}^{n}\Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)^{y_i}\left[1-(\Phi(\mathbf{x}_0^{i\mathrm{T}}\beta^0)\right]^{1-y_i}\mathrm{d}\beta^0\,. \end{aligned}$$

banknote

For $H_0 : \beta_1 = \beta_2 = 0$, $B^{\pi}_{10} = 157.73$ [against H_0]

Generic regression-like output:

```
  Estimate  Post. var.  log10(BF)

X1          -1.1552   0.0631        4.5844 (****)
X2           0.9200   0.3299       -0.2875
X3           0.9121   0.2595       -0.0972
X4           1.0820   0.0287       15.6765 (****)

evidence against H0: (****) decisive, (***) strong,
(**) subtantial, (*) poor
```

Informative settings

If prior information available on $p(\mathbf{x})$, transform into prior distribution on β by technique of *imaginary observations:*

Start with k different values of the covariate vector, $\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^k$
For each of these values, the practitioner specifies

(i) a prior guess g_i at the probability p_i associated with $\mathbf{x}^i$;

(ii) an assessment of (un)certainty about that guess given by a number K_i of equivalent "prior observations".

On how many imaginary observations did you build this guess?

Informative prior

$$\pi(p_1,\ldots,p_k) \propto \prod_{i=1}^{k} p_i^{K_i g_i - 1}(1-p_i)^{K_i(1-g_i)-1}$$

translates into *[Jacobian rule]*

$$\pi(\beta) \propto \prod_{i=1}^{k} \Phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)^{K_i g_i - 1}\left[1-\Phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)\right]^{K_i(1-g_i)-1}\phi(\tilde{\mathbf{x}}^{i\mathrm{T}}\beta)$$

[Almost] equivalent to using the G-prior

$$\beta \sim \mathscr{N}_k\left(0_k, \left[\sum_{j=1}^{k}\tilde{\mathbf{x}}^{j}\tilde{\mathbf{x}}^{j\mathrm{T}}\right]^{-1}\right)$$

The logit model

Recall that [for $n_i = 1$]

$$y_i|\mu_i \sim \mathscr{B}(1, \mu_i), \quad \varphi = 1 \quad \text{and} \quad g(\mu_i) = \left(\frac{\exp(\mu_i)}{1+\exp(\mu_i)}\right).$$

Thus

$$\mathbb{P}(y_i = 1|\beta) = \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}$$

with likelihood

$$\ell(\beta|\mathbf{y}, X) = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{1-y_i}$$

Links with probit

- usual vague prior for β, $\pi(\beta) \propto 1$
- Posterior given by

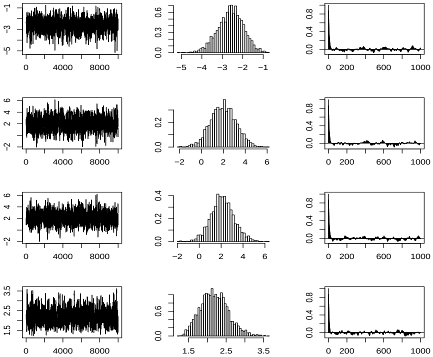
$$\pi(\beta|\mathbf{y}, X) \propto \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{y_i} \left(1 - \frac{\exp(\mathbf{x}^{i\mathrm{T}}\beta)}{1+\exp(\mathbf{x}^{i\mathrm{T}}\beta)}\right)^{1-y_i}$$

[intractable]

- Same Metropolis–Hastings sampler

bank

Same scale factor equal to $\tau = 1$: slight increase in the skewness of the histograms of the β_i's.



Plug-in estimate of predictive probability of a counterfeit

$$\hat{p}_i = \frac{\exp\left(-2.5888 x_{i1} + 1.9967 x_{i2} + 2.1260 x_{i3} + 2.1879 x_{i4}\right)}{1 + \exp\left(-2.5888 x_{i1} + 1.9967 x_{i2} + 2.1260 x_{i3} + 2.1879 x_{i4}\right)}.$$

G-priors for logit models

Same story: Flat prior on β inappropriate for Bayes factors, to be replaced with hierarchical prior,

$$\beta|\sigma^2, X \sim \mathscr{N}_k\left(0_k, \sigma^2(X^{\mathrm{T}}X)^{-1}\right) \quad \text{and} \quad \pi(\sigma^2|X) \propto \sigma^{-3/2}$$

Example (bank)

	Estimate	Post. var.	log10(BF)	
X1	-2.3970	0.3286	4.8084	(****)
X2	1.6978	1.2220	-0.2453	
X3	2.1197	1.0094	-0.1529	
X4	2.0230	0.1132	15.9530	(****)

evidence against H0: (****) decisive, (***) strong, (**) subtantial, (*) poor

Loglinear models

Introduction to loglinear models

Example (airquality)

Benchmark in R

```
> air=data(airquality)
```

Repeated measurements over 111 consecutive days of ozone u (in parts per billion) and maximum daily temperature v discretized into dichotomous variables

```
             month  5  6  7  8  9
ozone      temp
[1,31]     [57,79] 17  4  2  5 18
           (79,97]  0  2  3  3  2
(31,168]   [57,79]  6  1  0  3  1
           (79,97]  1  2 21 12  8
```

Contingency table with $5 \times 2 \times 2 = 20$ entries

Poisson regression

Observations/counts $\mathbf{y} = (y_1, \ldots, y_n)$ are integers, so we can choose

$$y_i \sim \mathscr{P}(\mu_i)$$

Saturated likelihood

$$\ell(\mu|\mathbf{y}) = \prod_{i=1}^{n} \frac{1}{\mu_i!} \mu_i^{y_i} \exp(-\mu_i)$$

GLM constraint via log-linear link

$$\log(\mu_i) = \mathbf{x}^{iT}\beta\,, \quad y_i|\mathbf{x}^i \sim \mathscr{P}\left(e^{\mathbf{x}^{iT}\beta}\right)$$

Categorical variables

Special feature

Incidence matrix $X = (\mathbf{x}^i)$ such that its elements are all zeros or ones, i.e. covariates are all indicators/dummy variables!

Several types of (sub)models are possible depending on relations between categorical variables.

Re-special feature

Variable selection problem of a specific kind, in the sense that all indicators related with the *same* association must either remain or vanish at once. Thus much fewer submodels than in a regular variable selection problem.

Parameterisations

Example of three variables $1 \leq u \leq I$, $1 \leq v \leq j$ and $1 \leq w \leq K$.

Simplest non-constant model is

$$\log(\mu_\tau) = \sum_{b=1}^{I} \beta_b^u \mathbb{I}_b(u_\tau) + \sum_{b=1}^{J} \beta_b^v \mathbb{I}_b(v_\tau) + \sum_{b=1}^{K} \beta_b^w \mathbb{I}_b(w_\tau)\,,$$

that is,

$$\log(\mu_{l(i,j,k)}) = \beta_i^u + \beta_j^v + \beta_k^w\,,$$

where index $l(i,j,k)$ corresponds to $u = i$, $v = j$ and $w = k$.
Saturated model is

$$\log(\mu_{l(i,j,k)}) = \beta_{ijk}^{uvw}$$

Log-linear model (over-)parameterisation

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw}\,,$$

as in Anova models.

- λ appears as the overall or reference average effect
- λ_i^u appears as the marginal discrepancy (against the reference effect λ) when $u = i$,
- λ_{ij}^{uv} as the interaction discrepancy (against the added effects $\lambda + \lambda_i^u + \lambda_j^v$) when $(u, v) = (i, j)$

and so on...

Example of submodels

1. if both v and w are irrelevant, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u \,,$$
2. if all three categorical variables are mutually independent, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w \,,$$
3. if u and v are associated but are both independent of w, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} \,,$$
4. if u and v are conditionally independent given w, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} \,,$$
5. if there is no three-factor interaction, then
$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw}$$
[the most complete submodel]

Identifiability

Representation

$$\log(\mu_{l(i,j,k)}) = \lambda + \lambda_i^u + \lambda_j^v + \lambda_k^w + \lambda_{ij}^{uv} + \lambda_{ik}^{uw} + \lambda_{jk}^{vw} + \lambda_{ijk}^{uvw} \,,$$

not identifiable but Bayesian approach handles non-identifiable settings and still estimate properly identifiable quantities. Customary to impose identifiability constraints on the parameters: set to 0 parameters corresponding to the first category of each variable, i.e. remove the indicator of the first category.

E.g., if $u \in \{1,2\}$ and $v \in \{1,2\}$, constraint could be

$$\lambda_1^u = \lambda_1^v = \lambda_{11}^{uv} = \lambda_{12}^{uv} = \lambda_{21}^{uv} = 0\,.$$

Inference under a flat prior

Noninformative prior $\pi(\beta) \propto 1$ gives posterior distribution

$$\begin{aligned}\pi(\beta|\mathbf{y},X) &\propto \prod_{i=1}^{n} \left\{\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}^{y_i} \exp\{-\exp(\mathbf{x}^{i\mathrm{T}}\beta)\} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\,\mathbf{x}^{i\mathrm{T}}\beta - \sum_{i=1}^{n}\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}\end{aligned}$$

Use of same random walk M-H algorithm as in probit and logit cases, starting with MLE evaluation

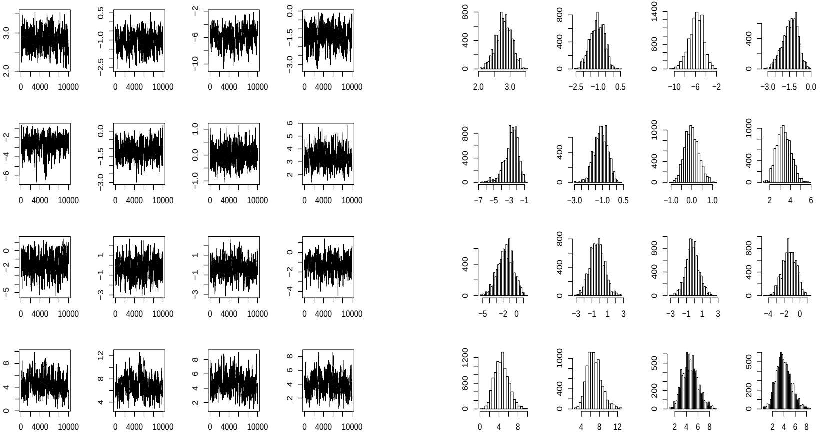
```
> mod=summary(glm(y~-1+X,family=poisson()))
```

airquality

Identifiable non-saturated model involves 16 parameters
Obtained with $10,000$ MCMC iterations with scale factor $\tau^2 = 0.5$

Effect	Post. mean	Post. var.
λ	2.8041	0.0612
λ_2^u	-1.0684	0.2176
λ_2^v	-5.8652	1.7141
λ_2^w	-1.4401	0.2735
λ_3^w	-2.7178	0.7915
λ_4^w	-1.1031	0.2295
λ_5^w	-0.0036	0.1127
λ_{22}^{uv}	3.3559	0.4490
λ_{22}^{uw}	-1.6242	1.2869
λ_{23}^{uw}	- 0.3456	0.8432
λ_{24}^{uw}	-0.2473	0.6658
λ_{25}^{uw}	-1.3335	0.7115
λ_{22}^{vw}	4.5493	2.1997
λ_{23}^{vw}	6.8479	2.5881
λ_{24}^{vw}	4.6557	1.7201
λ_{25}^{vw}	3.9558	1.7128

airquality: MCMC output



Model choice with G-prior

G-prior alternative used for probit and logit models still available:

$$\begin{aligned}\pi(\beta|\mathbf{y},X) &\propto |X^{\mathrm{T}}X|^{1/2}\Gamma\left\{\frac{(2k-1)}{4}\right\}||X\beta||^{-(2k-1)/2}\,\pi^{-k/2} \\ &\times\exp\left\{\left(\sum_{i=1}^{n}y_i\,\mathbf{x}^i\right)^{\mathrm{T}}\beta-\sum_{i=1}^{n}\exp(\mathbf{x}^{i\mathrm{T}}\beta)\right\}\end{aligned}$$

Same MCMC implementation and similar estimates for airquality

airquality

Bayes factors once more approximated by importance sampling based on normal importance functions

Anova-like output

```
Effect log10(BF)

u:v     6.0983 (****)
u:w    -0.5732
v:w     6.0802 (****)

evidence against H0: (****) decisive, (***) strong,
(**) subtantial, (*) poor
```