Bayesian Methods for Data Analysis

ENAR Annual Meeting

Tampa, Florida – March 26, 2006

Course contents

- Introduction of Bayesian concepts using single-parameter models.
- Multiple-parameter models and hyerarchical models.
- Computation: approximations to the posterior, rejection and importance sampling and MCMC.
- Model checking, diagnostics, model fit.
- Linear hierarchical models: random effects and mixed models.
- Generalized linear models: logistic, multinomial, Poisson regression.
- Hierarchical models for spatially correlated data.

Course emphasis

- Notes draw heavily on the book by Gelman et al., *Bayesian Data Analysis* 2nd. ed., and many of the figures are 'borrowed' directly from that book.
- We focus on the implementation of Bayesian methods and interpretation of results.
- Little theory, but some is needed to understand methods.
- Lots of examples. Some are not directly drawn from biological problems, but still serve to illustrate methodology.
- Biggest idea to get across: inference by simulation.

- Software: R (or SPlus) early on, and WinBUGS for most of the examples after we discuss computational methods.
- All of the programs used to construct examples can be downloaded from www.public.iastate.edu/ $\sim$ alicia.
- On my website, go to Teaching and then to ENAR Short Course.

Single parameter models

- There is no real advantage to being a Bayesian in these simple models. We discuss them to introduce important concepts:
  - Priors: non-informative, conjugate, other informative
  - Computations
  - Summary of posterior, presentation of results
- Binomial, Poisson, Normal models as examples
- Some “real” examples

Binomial model

- Of historical importance: Bayes derived his theorem and Laplace provided computations for the Binomial model.
- Before Bayes, question was: given θ, what are the probabilities of the possible outcomes of y?
- Bayes asked: what is $Pr(\theta_1 < \theta < \theta_2|y)$?
- Using a uniform prior for θ, Bayes showed that

$$Pr(\theta_1 < \theta < \theta_2|y) \propto \frac{\int_{\theta_1}^{\theta_2} \theta^y(1-\theta)^{n-y}d\theta}{p(y)},$$

with $p(y) = 1/(n+1)$ uniform *a priori*.

- Laplace devised an approximation to the integral expression in numerator.
- First application by Laplace: estimate the probability that there were more female births in Paris from 1745 to 1770.
- Consider n exchangeable Bernoulli trials $y_1, ..., y_n$.
- $y_i = 1$ a “success”, $y_i = 0$ a “failure”.
- Exchangeability:summary is # of successes in n trials y.
- For θ the probability of success, $Y \sim B(n, \theta)$:

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

- MLE is sample proportion: $\hat{\theta} = y/n$.
- Prior on θ: uniform on [0,1] (for now)

$$p(\theta) \propto 1.$$

- Posterior:

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}.$$

- As a function of θ, posterior is proportional to a $Beta(y+1, n-y+1)$ density.
- We could also do the calculations to derive the posterior:

$$p(\theta|y) \quad = \quad \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}}{\int \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta}$$

$$
\begin{aligned}
&= (n+1)\frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \\
&= \frac{(n+1)!}{y!(n-y)!}\theta^y(1-\theta)^{n-y} \\
&= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{y+1-1}(1-\theta)^{n-y+1-1} \\
&= \text{Beta}(y+1, n-y+1)
\end{aligned}
$$

- Point estimation
  - Posterior mean $E(\theta|y) = \frac{y+1}{n+2}$
  - Note posterior mean is compromise between prior mean (1/2) and sample proportion y/n.
  - Posterior mode y/n
  - Posterior median θ^* such that $Pr(\theta \leq \theta^*|y) = 0.5$.
  - Best point estimator minimizes the expected loss (more later).

- Posterior variance

$$Var(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}.$$

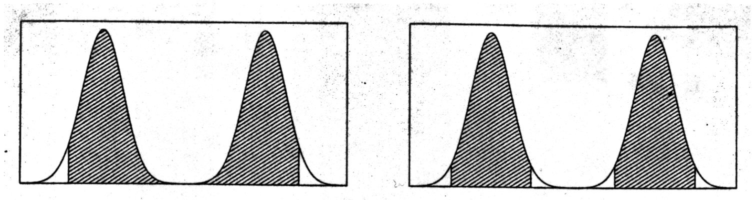
- Interval estimation
  - 95% credibe set (or central posterior interval) is (a, b) if:

$$\int_0^a p(\theta|y)d\theta = 0.025 \text{ and } \int_0^b p(\theta|y)d\theta = 0.975$$

  - A $100(1-\alpha)$% highest posterior density credible set is subset C of Θ such that
$$C = \{\theta \in \Theta : p(\theta|y) \leq k(\alpha)\}$$
where $k(\alpha)$ is largest constant such that $Pr(C|y) \leq 1-\alpha$.
  - For symmetric unimodal posteriors, credible sets and highest posterior density credible sets coincide.

- In other cases, HPD sets have smallest size.
- Interpretation (for either): probability that θ is in set is equal to $1-\alpha$. Which is what we *want* to say!.

Credible set and HPD set



- Inference by simulation:
  - Draw values from posterior: easy for closed form posteriors, can also do in other cases (later)
  - Monte Carlo estimates of point and interval estimates
  - Added MC error (due to "sampling")
  - Easy to get interval estimates and estimators for functions of parameters.

- Prediction:
  - Prior predictive distribution

$$p(y) = \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta = \frac{1}{n+1}.$$

  - A priori, all values of y are equally likely

- Posterior predictive distribution, to predict outcome of a new trial $\tilde{y}$ given y successes in previous n trials

$$\begin{aligned} Pr(\tilde{y}=1|y) &= \int_0^1 Pr(\tilde{y}=1|y,\theta)p(\theta|y)d\theta \\ &= \int_0^1 Pr(\tilde{y}=1|\theta)p(\theta|y)d\theta \\ &= \int_0^1 \theta p(\theta|y)d\theta = E(\theta|y) = \frac{y+1}{n+2} \end{aligned}$$

Binomial model: different priors

- How do we choose priors?
  - In a purely subjective manner (orthodox)
  - Using actual information (e.g., from literature or scientific knowledge)
  - Eliciting from experts
  - For mathematical convenience
  - To express ignorance

- Unless we have reliable prior information about θ, we prefer to let the data 'speak' for themselves.

- Asymptotic argument: as sample size increases, likelihood should dominate posterior

Conjugate prior

- Suppose that we choose a Beta prior for θ:

$$p(\theta|\alpha,\beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- Posterior is now

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$$

- Posterior is again proportional to a Beta:

$$p(\theta|y) = \text{Beta}(y+\alpha, n-y+\beta).$$

- For now, α, β considered fixed and known, but they can also get their own prior distribution (hierarchical model).

- Beta is the **conjugate** prior for binomial model: posterior is in the same form as the prior.

- To choose prior parameters, think as follows: observe α successes in $\alpha+\beta$ prior "trials". Prior "guess" for θ is $\alpha/(\alpha+\beta)$.

Conjugate priors

- Formal definition: F a class of sampling distributions and P a class of prior distributions. Then P is **conjugate** for F if $p(\theta) \in P$ and $p(y|\theta) \in F$ implies $p(\theta|y) \in P$.

- If F is *exponential family* then distributions in F have natural conjugate priors.

- A distribution in F has form

$$p(y_i|\theta) = f(y_i)g(\theta)\exp(\phi(\theta)^T u(y_i)).$$

- For an *iid* sequence, likelihood function is

$$p(y|\theta) \propto g(\theta)^n \exp(\phi(\theta)^T t(y)),$$

for $\phi(\theta)$ the natural parameter and $t(y)$ a sufficient statistic.

- Consider prior density

$$p(\theta) \propto g(\theta)^\eta \exp(\phi(\theta)^T \nu)$$

- Posterior also in exponential form

$$p(\theta|y) \propto g(\theta)^{n+\eta} \exp(\phi(\theta)^T (t(y) + \nu))$$

Conjugate prior for binomial proportion

- y is # of successes in n exchangeable trials, so sampling distribution is binomial with prob of success θ:

$$p(y|\theta) \propto \theta^y (1-\theta)^{n-y}$$

- Note that

$$\begin{aligned} p(y|\theta) &\propto \theta^y (1-\theta)^n (1-\theta)^{-y} \\ &\propto (1-\theta)^n \exp\{y[\log\theta - \log(1-\theta)]\} \end{aligned}$$

- Written in exponential family form

$$p(y|\theta) \propto (1-\theta)^n \exp\{y \log \frac{\theta}{1-\theta}\}$$

where $g(\theta)^n = (1-\theta)^n$, y is the sufficient statistic, and the logit $\log\theta/(1-\theta)$ is the natural parameter.

- Consider prior for θ: Beta(α, β):

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- If we let $\nu = \alpha - 1$ and $\eta = \beta + \alpha - 2$, can write

$$p(\theta) \propto \theta^\nu (1-\theta)^{\eta-\nu}$$

or

$$p(\theta) \propto (1-\theta)^\eta \exp\{\nu \log \frac{\theta}{1-\theta}\}$$

- Then posterior is in same form as prior:

$$p(\theta|y) \propto (1-\theta)^{n+\eta} \exp\{(y+\nu)\log\frac{\theta}{1-\theta}\}$$

- Since $p(y|\theta) \propto \theta^u(1-\theta)^{n-y}$ then prior Beta(α,β) suggests that a priori we believe in approximately α successes in $\alpha+\beta$ trials.

- Our prior guess for the probability of success is $\alpha/(\alpha+\beta)$

- By varying $(\alpha+\beta)$ (with $\alpha/(\alpha+\beta)$ fixed), we can incorporate more or less information into the prior: “prior sample size”

- Also note:

$$E(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n}$$

  is always between the prior mean $\alpha/(\alpha+\beta)$ and the MLE y/n.

- Posterior variance

$$\begin{aligned} Var(\theta|y) &= \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\ &= \frac{E(\theta|y)[1-E(\theta|y)]}{\alpha+\beta+n+1} \end{aligned}$$

- As y and $n-y$ get large:
  - $E(\theta|y) \rightarrow y/n$
  - $var(\theta|y) \rightarrow (y/n)[1-(y/n)]/n$ which approaches zero at rate $1/n$.
  - Prior parameters have diminishing influence in the posterior as $n \rightarrow \infty$.

Placenta previa example

- Condition in which placenta is implanted low in uterus, preventing normal delivery.

- In study in Germany, found that 437 out of 980 births with placenta previa were females so observed ratio is 0.446

- Suppose that in general population, proportion of female births is 0.485.

- Can we say that the proportion of female babies among placenta previa births is lower than in the general population?

- If $y =$ number of female births, θ is probability of a female birth, and $p(\theta)$ is uniform, then

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-1}$$

and therefore

$$\theta|y \sim Beta(y+1, n-y+1).$$

- Thus a posteriori, θ is Beta(438, 544) and

$$E(\theta|y) = \frac{438}{438+544}$$

$$Var(\theta|y) = \frac{438 \times 544}{(438+544)^2(438+544+1)}$$

- Check sensitivity to choice of priors: try several Beta priors with increasingly more “information” about θ.

- Fix prior mean at 0.485 and increase “prior sample size” $\alpha+\beta$.

Prior		Post.	Post 2.5th	Post 97.5th
mean	$\alpha+\beta$	median	pctile	pctile
0.5	2	0.446	0.415	0.477
0.485	2	0.446	0.415	0.477
0.485	5	0.446	0.415	0.477
0.485	10	0.446	0.415	0.477
0.485	20	0.447	0.416	0.478
0.485	100	0.450	0.420	0.479
0.485	200	0.453	0.424	0.481

Results robust to choice of prior, even very informative prior. Prior mean not in posterior 95% credible set.

Conjugate prior for Poisson rate

- $y \in (0, 1, ...)$ is counts, with rate $\theta > 0$. Sampling distribution is Poisson
$$p(y_i|\theta) = \frac{\theta^{y_i} \exp(\theta)}{y_i!}$$
- For exchangeable $(y_1, ..., y_n)$
$$p(y|\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} \exp(-\theta)}{y_i!} = \frac{\theta^{n\bar{y}} \exp(-n\theta)}{\prod_{i=1}^{n} y_i!}.$$
- Consider Gamma(α, β) as prior for θ: $p(\theta) \propto \theta^{\alpha-1} \exp(-\beta\theta)$
- Then posterior is also Gamma:
$$p(\theta|y) \propto \theta^{n\bar{y}} \exp(-n\theta)\theta^{\alpha-1} \exp(-\beta\theta) \propto \theta^{n\bar{y}+\alpha-1} \exp(-n\theta - \beta\theta)$$

- For $\eta = n\bar{y} + \alpha$ and $\nu = n + \beta$, posterior is Gamma(η, ν)
- Note:
  - Prior mean of θ is α/β
  - Posterior mean of θ is
$$E(\theta|y) = \frac{n\bar{y} + \alpha}{n + \beta}$$
- If sample size $n \to \infty$ then $E(\theta|y)$ approaches MLE of θ.
- If sample size goes to zero, then $E(\theta|y)$ approaches prior mean.

Mean of a normal distribution

- $y \sim N(\theta, \sigma^2)$ with σ^2 known
- For $\{y_1, ..., y_n\}$ an iid sample, the likelihood is:
$$p(y|\theta) = \Pi_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(y_i - \theta)^2)$$
- Viewed as function of θ, likelihood is exponential of a quadratic in θ:
$$p(y|\theta) \propto \exp(-\frac{1}{2\sigma^2} \sum_i (\theta^2 - 2y_i\theta + y_i^2))$$
- Conjugate prior for θ must belong to family of form
$$p(\theta) = \exp(A\theta^2 + B\theta + C)$$

that can be parameterized as

$$p(\theta) \propto \exp(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2)$$

and then $p(\theta)$ is $N(\mu_0, \tau_0^2)$

- (μ_0, τ_0^2) are *hyperparameters*. For now, consider known.
- If $p(\theta)$ is conjugate, then posterior $p(\theta|y)$ must also be normal with parameters (μ_n, τ_n^2).
- Recall that $p(\theta|y) \propto p(\theta)p(y|\theta)$
- Then

$$p(\theta|y) \propto \exp(-\frac{1}{2}[\frac{\sum_i (y_i - \theta)^2}{\sigma^2} + \frac{(\theta - \mu_0)^2}{\tau_0^2}])$$

- Expand squares, collect terms in θ^2 and in θ:

$$\begin{aligned} p(\theta|y) &\propto \exp(-\frac{1}{2}[\frac{\sum_i y_i^2 - 2\sum_i y_i\theta + \sum_i \theta^2}{\sigma^2} + \frac{\theta^2 - 2\mu_0\theta + \mu_0^2}{\tau_0^2}]) \\ &\propto \exp(-\frac{1}{2}[\frac{(\tau_0^2 + \sigma^2)n\theta^2 - 2(n\bar{y}\tau_0^2 + \mu_0\sigma^2)\theta}{\sigma^2\tau_0^2}]) \\ &\propto \exp(-\frac{1}{2}\frac{((\sigma^2/n) + \tau_0^2)}{(\sigma^2/n)\tau_0^2}[\theta^2 - 2(\frac{\bar{y}\tau_0^2 + \mu_0(\sigma^2/n)}{(\sigma^2/n) + \tau_0^2})\theta]) \end{aligned}$$

- Then $p(\theta|y)$ is normal with
  - Mean: $\mu_n = (\bar{y}\tau_0^2 + \mu_0(\sigma^2/n))/((\sigma^2/n) + \tau_0^2)$
  - Variance: $\tau_n^2 = ((\sigma^2/n)\tau_0^2)/((\sigma^2/n) + \tau_0^2)$

Posterior mean

- Note that

$$\mu_n = \frac{\bar{y}\tau_0^2 + \mu_0(\sigma^2/n)}{(\sigma^2/n) + \tau_0^2} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}$$

(by dividing numerator and denominator into $(\sigma^2/n)\tau_0^2$)

- Posterior mean is **weighted average** of prior mean and sample mean.
- Weights are given by **precisions** n/σ^2 and $1/\tau_0^2$.
- As data precision increases ($(\sigma^2/n) \to 0$) because $\sigma^2 \to 0$ or because $n \to \infty$, $\mu_n \to \bar{y}$.
- Also, note that

$$\mu_n = \frac{\bar{y}\tau_0^2 + \mu_0(\sigma^2/n)}{(\sigma^2/n) + \tau_0^2}$$

$$= \mu_0(\frac{\sigma^2/n}{(\sigma^2/n) + \tau_0^2}) + \bar{y}(\frac{\tau_0^2}{(\sigma^2/n) + \tau_0^2})$$

Add and subtract $\mu_0\tau_0^2/((\sigma^2/n) + \tau_0^2)$ to see that

$$\mu_n = \mu_0 + (\bar{y} - \mu_0)(\frac{\tau_0^2}{(\sigma^2/n) + \tau_0^2})$$

- Posterior mean is prior mean *shrunken* towards observed value. Amount of shrinkage depends on relative size of precisions

Posterior variance:

- Recall that

$$p(\theta|y) \propto \exp(-\frac{1}{2}\frac{((\sigma^2/n) + \tau_0^2)}{(\sigma^2/n)\tau_0^2}[\theta^2 - 2(\frac{\bar{y}\tau_0^2 + \mu_0(\sigma^2/n)}{(\sigma^2/n) + \tau_0^2})\theta])$$

- Then

$$\begin{aligned}\frac{1}{\tau_n^2} &= \frac{((\sigma^2/n)+\tau_0^2)}{(\sigma^2/n)\tau_0^2}\\ &= \frac{n}{\sigma^2}+\frac{1}{\tau_0^2}\end{aligned}$$

- Posterior precision = sum of prior and data precisions.

Posterior predictive distribution

- Want to predict next observation $\tilde{y}$.
- Recall $p(\tilde{y}|y) \propto \int p(\tilde{y}|\theta)p(\theta|y)d\theta$
- We know that

  - Given θ, $\tilde{y} \sim N(\theta, \sigma^2)$
  - $\theta|y \sim N(\mu_n, \tau_n^2)$

- Then

$$\begin{aligned}p(\tilde{y}|y) &\propto \int \exp\{-\frac{1}{2}\frac{(\tilde{y}-\theta)^2}{\sigma^2}\}\exp\{-\frac{1}{2}\frac{(\theta-\mu_n)^2}{\tau_n^2}\}d\theta\\ &\propto \int \exp\{-\frac{1}{2}[\frac{(\tilde{y}-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_n)^2}{\tau_n^2}]\}d\theta\end{aligned}$$

- Integrand is kernel of bivariate normal, so $(\tilde{y}, \theta)$ have bivariate normal joint posterior.
- Marginal $p(\tilde{y}|y)$ must be normal.

- Need $E(\tilde{y}|y)$ and $var(\tilde{y}|y)$:

$$E(\tilde{y}|y) = E[E(\tilde{y}|y,\theta)|y] = E(\theta|y) = \mu_n,$$

because $E(\tilde{y}|y,\theta) = E(\tilde{y}|\theta) = \theta$

$$\begin{aligned}var(\tilde{y}|y) &= E[var(\tilde{y}|y,\theta)|y] + var[E(\tilde{y}|y,\theta)|y]\\ &= E(\sigma^2|y) + var(\theta|y)\\ &= \sigma^2 + \tau_n^2\end{aligned}$$

because $var(\tilde{y}|y,\theta) = var(\tilde{y}|\theta) = \sigma^2$

- Variance includes additional term τ_n as a penalty for us not knowing the true value of θ.
- Recall $\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}}$

- In μ_n, prior precision $1/\tau_0^2$ and data precision n/σ^2 are "equivalent". Then:
  - For n large, $(\bar{y}, \sigma^2)$ determine posterior
  - For $\tau_0^2 = \sigma^2/n$, prior has the same weight as adding one more observation with value μ_0.
  - When $\tau_0^2 \rightarrow \infty$ with n fixed, or when $n \rightarrow \infty$ with τ_0^2 fixed:

$$p(\theta|\bar{y}) \rightarrow N(\bar{y}, \sigma^2/n)$$

  - Good approximation in practice when prior beliefs about θ are vague or when sample size is large.

Normal variance

- Example of a scale model
- Assume that $y \sim N(\theta, \sigma^2)$, θ known.
- For iid $y_1, ..., y_n$:

$$p(y|\sigma^2) \propto (\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2)$$

$$p(y|\sigma^2) \propto (\sigma^2)^{-n/2} \exp(-\frac{nv}{2\sigma^2})$$

for suffient statistic

$$v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$$

- Likelihood is in exponential family form with natural parameter $\phi(\sigma^2) = \sigma^{-2}$. Then, natural conjugate prior must be of form

$$p(\sigma^2) \propto (\sigma^2)^{-\eta} \exp(-\beta\phi(\sigma^2))$$

- Consider an inverse Gamma prior:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp(-\beta/\sigma^2)$$

- For ease of interpretation, reparameterize as a scaled-inverted χ^2 distribution, with ν_0 d.f. and σ_0^2 scale.
  - Scale σ_0^2 corresponds to prior "guess" for σ^2.
  - Large ν_0: lots of confidence in σ_0^2 as a good value for σ^2.

- If $\sigma^2 \sim inv - \chi^2(\nu_0, \sigma_0^2)$ then

$$p(\sigma^2) \propto (\frac{\sigma^2}{\sigma_0^2})^{-(\frac{\nu_0}{2}+1)} \exp(-\frac{\nu_0\sigma_0^2}{2\sigma^2})$$

- Corresponds to $Inv - Gamma(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2})$.
- Prior mean is $\sigma_0^2\nu_0/(\nu_0 - 2)$
- Prior variance behaves like σ_0^4/ν_0: large ν_0, large prior precision.
- Posterior

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto (\sigma^2)^{-n/2} \exp(-\frac{nv}{2\sigma^2})(\frac{\sigma^2}{\sigma_0^2})^{-(\frac{\nu_0}{2}+1)} \exp(-\frac{\nu_0\sigma_0^2}{2\sigma^2}) \end{aligned}$$

$$\propto \quad (\sigma^2)^{-(\frac{\nu_1}{2}+1)} \exp(-\frac{1}{2\sigma^2}\nu_1\sigma_1^2)$$

with $\nu_1 = \nu_0 + n$ and

$$\sigma_1^2 = \frac{nv + \nu_0\sigma_0^2}{n + \nu_0}$$

- $p(\sigma^2|y)$ is also a scaled inverted χ^2
- Posterior scale is weighted average of prior "guess" and data estimate
- Weights given by prior and sample degrees of freedom
- Prior provides information equivalent to ν_0 observations with average squared deviation equal to σ_0^2.
- As $n \to \infty$, $\sigma_1^2 \to v$.

Poisson model

- Appropriate for count data such as number of cases, number of accidents, etc. For a vector of n iid observations:

$$p(y|\theta) = \Pi_i \frac{\theta^{y_i} e^{-\theta}}{y_i!} = p(y|\theta) = \frac{\theta^{t(y)} e^{-n\theta}}{\prod_{i=1}^{n} y!},$$

where θ is the rate, $y = 0, 1, ...$ and $t(y) = \sum_{i=1}^{n} y_i$ the sufficient statistic for θ.

- We can write the model in the exponential family form:

$$p(y|\theta) \propto e^{-n\theta} e^{t(y)\log\theta}$$

where $\phi(\theta) = \log\theta$ is natural parameter.

- Natural conjugate prior must have form

$$p(\theta) \propto e^{-\eta\theta} e^{\nu \log\theta}$$

or $p(\theta) \propto e^{-\beta\theta}\theta^{\alpha-1}$ that looks like a Gamma(α, β).

- Posterior is Gamma$(\alpha + n\bar{y}, \beta + n)$

Poisson model - Digression

- In the conjugate case, can often derive $p(y)$ using

$$p(y) = \frac{p(\theta)p(y|\theta)}{p(\theta|y)}$$

- In case of Gamma-Poisson model for a single observation:

$$\begin{aligned} p(y) &= \frac{\text{Poisson}(\theta)\ \text{Gamma}(\alpha,\beta)}{\text{Gamma}(\alpha+y,\beta+1)} = \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{(\alpha+y)}} \\ &= (\alpha+y-1, y)(\frac{\beta}{\beta+1})^{\alpha}(\frac{1}{\beta+1})^{y} \end{aligned}$$

the density of a *negative binomial* with parameters (α, β).

- Thus we can interpret negative binomial distributions as arising from a mixture of Poissons with rate θ, where θ follows a Gamma distribution:

$$p(y) = \int \text{Poisson } (y|\theta) \text{ Gamma } (\theta|\alpha,\beta) d\theta$$

- Negative binomial is robust alternative to Poisson.

Poisson model (cont'd)

- Given rate, observations assumed to be exchangeable.
- Add an *exposure* to model: observations are exchangeable within small exposure intervals.
- Examples:
  - In a city of size 500,000, define rate of death from cancer per million people, then exposure is 0.5
  - Intersection in Ames with traffic of one million vehicles per year, consider number of traffic accidents per million vehicles per year. Exposure for intersection is 1.
  - Exposure is typically known and reflects the fact that in each unit, the number of persons or cars or plants or animals that are 'at risk' is different.

- Now
$$p(y|\theta) \propto \theta^{\sum y_i} \exp(-\theta \sum x_i)$$
for x_i the exposure of unit i.
- With prior $p(\theta) =$ Gamma (α, β), posterior is
$$p(\theta|y) = \text{ Gamma } (\alpha + \sum_i y_i, \beta + \sum_i x_i)$$

Non-informative prior distns

- A non-informative prior distribution
  - Has minimal impact on posterior
  - Lets the data "speak for themselves"
  - Also called vague, flat, diffuse
- So far, we have concentrated on conjugate family.
- Conjugate priors can also be almost non-informative.
- If $y \sim N(\theta, 1)$, natural conjugate for θ is $N(\mu_0, \tau_0^2)$, and posterior is $N(\mu_1, \tau_1^2)$, where
$$\mu_1 = \frac{\mu_0/\tau_0^2 + n\bar{y}/\sigma^2}{1/\tau_0^2 + n/\sigma^2}, \; \tau_1^2 = \frac{1}{1/\tau_0^2 + n/\sigma^2}$$

- For $\tau_0 \rightarrow \infty$:
  - $\mu_1 \rightarrow \bar{y}$
  - $\tau_1^2 \rightarrow \sigma^2/n$
- Same result could have been obtained using $p(\theta) \propto 1$.
- The uniform prior is a *natural non-informative prior* for location parameters (see later).
- It is *improper*:
$$\int p(\theta) d\theta = \int d\theta = \infty$$
yet leads to proper posterior for θ. This is not always the case.
- Poisson example: $y_1, ..., y_n$ iid Poisson(θ), consider $p(\theta) \propto \theta^{-1/2}$.
  - $\int_0^\infty \theta^{-1/2} d\theta = \infty$ improper

- Yet

$$\begin{aligned} p(\theta|y) &\propto \theta^{\sum_i y_i} e^{-n\theta}\theta^{-1/2} \\ &= \theta^{\sum_i y_i - 1/2} e^{-n\theta} \\ &= \theta^{\sum_i y_i + 1/2 - 1} e^{-n\theta} \end{aligned}$$

proportional to a Gamma($\frac{1}{2} + \sum_i y_i, n$), proper

- Uniform priors arise when we assign equal density to each $\theta \in \Theta$: no preferences for one value over the other.

- But they are not invariant under one-to-one transformations.

- Example:

  - $\eta = \exp\{\theta\}$ so that $\theta = log\{\eta\}$ is inverse transformation.

- $\frac{d\theta}{d\eta} = \frac{1}{\eta}$ is Jacobian
- Then, if $p(\theta)$ is prior for θ, $p^*(\eta) = \eta^{-1}p(\log\eta)$ is corresponding prior for transformation.
- For $p(\theta) \propto c$, $p^*(\eta) \propto \eta^{-1}$, informative.
- Informative prior is needed to arrive at same answer in both parameterizations.

Jeffreys prior

- In 1961, Jeffreys proposed a method for finding non-informative priors that are invariant to one-to-one transformations

- Jeffreys proposed

$$p(\theta) \propto [I(\theta)]^{1/2}$$

where $I(\theta)$ is the expected Fisher information:

$$I(\theta) = -E_\theta[\frac{d^2}{d\theta^2}\log p(y|\theta)]$$

- If θ is a vector, then $I(\theta)$ is a matrix with (i, j) element equal to

$$-E_\theta[\frac{d^2}{d\theta_i d\theta_j}\log p(y|\theta)]$$

and

$$p(\theta) \propto |I(\theta)|^{1/2}$$

- Theorem: Jeffreys prior is locally uniform and therefore non-informative.

Jeffreys prior for binomial proportion

- If $y \sim B(n, \theta)$ then $\log p(y|\theta) \propto y \log \theta + (n-y)\log(1-\theta)$ and

$$\frac{d^2}{d\theta^2} \log p(y|\theta) = -y\theta^{-2} - (n-y)(1-\theta)^{-2}$$

- Taking expectations:

$$\begin{aligned} I(\theta) &= -E_\theta[\frac{d^2}{d\theta^2} \log p(y|\theta)] = E(y)\theta^{-2} + (n - E(y))(1-\theta)^{-2} \\ &= n\theta\theta^{-2}(n - n\theta)(1-\theta)^{-2} = \frac{n}{\theta(1-\theta)}. \end{aligned}$$

- Then

$$p(\theta) \propto [I(\theta)]^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2} \propto \text{ Beta}(\frac{1}{2}, \frac{1}{2}).$$

Jeffreys prior for normal mean

- $y_1, ..., y_n$ iid $N(\theta, \sigma^2)$, σ^2 known.

$$\begin{aligned} p(y|\theta) &\propto \exp(-\frac{1}{2\sigma^2}\sum(y-\theta)^2) \\ \log p(y|\theta) &\propto -\frac{1}{2\sigma^2}\sum(y-\theta)^2 \\ \frac{d^2}{d\theta^2}\log p(y|\theta) &= -\frac{n}{\sigma^2} \end{aligned}$$

constant with respect to θ.

- Then $I(\theta)$ constant and $p(\theta) \propto$ constant.

Jeffreys prior for normal variance

- $y_1, ..., y_n$ iid $N(\theta, \sigma^2)$, θ known.

$$\begin{aligned} p(y|\sigma) &\propto \sigma^{-n}\exp(-\frac{1}{2\sigma^2}\sum(y-\theta)^2) \\ \log p(y|\sigma) &\propto -n\log\sigma - \frac{1}{2\sigma^2}\sum(y-\theta)^2 \\ \frac{d^2}{d\sigma^2}\log p(y|\sigma) &= \frac{n}{\sigma^2} - \frac{3}{2\sigma^4}\sum(y_i-\theta)^2 \end{aligned}$$

- Take negative of expectation so that:

$$I(\sigma) = -\frac{n}{\sigma^2} + \frac{3}{2\sigma^4}n\sigma^2 = \frac{n}{2\sigma^2}$$

and therefore the Jeffreys prior is $p(\sigma) \propto \sigma^{-1}$.

Invariance property of Jeffreys'

- The Jeffreys' prior is invariant to one-to-one transformations $\phi = h(\theta)$
- Then:

$$p(\phi) = p(\theta)|\frac{d\theta}{d\phi}| = p(\theta)|\frac{dh(\theta)}{d\theta}|^{-1}$$

- Under invariance, choose $p(\theta)$ such that $p(\phi)$ constructed as above would match what would be obtained directly.
- Consider $p(\theta) = [I(\theta)]^{1/2}$ and evaluate $I(\phi)$ at $\theta = h^{-1}(\phi)$:

$$\begin{aligned} I(\phi) &= -E[\frac{d^2}{d\phi^2}\log p(y|\phi)] \\ &= -E[\frac{d^2}{d\phi^2}\log p(y|\theta = h^{-1}(\phi))[\frac{d\theta}{d\phi}]^2] \end{aligned}$$

$$= \quad I(\theta)[\frac{d\theta}{d\phi}]^2$$

- Then $[I(\phi)]^{1/2} = [I(\theta)]^{1/2}\frac{d\theta}{d\phi}$ as required.

Multiparameter models - Intro

- Most realistic problems require models with more than one parameter
- Typically, we are interested in one or a few of those parameters
- Classical approach for estimation in multiparameter models:
  1. Maximize a joint likelihood: can get nasty when there are many parameters
  2. Proceed in steps
- Bayesian approach: base inference on the *marginal posterior distributions* of the parameters of interest.
- Parameters that are not of interest are called *nuisance parameters*.

Nuisance parameters

- Consider a model with two parameters (θ_1, θ_2) (e.g., a normal distribution with unknown mean and variance)
- We are interested in θ_1 so θ_2 is a nuisance parameter
- The marginal posterior distribution of interest is $p(\theta_1|y)$
- Can be obtained directly from the *joint posterior density*

$$p(\theta_1, \theta_2|y) \propto p(\theta_1, \theta_2)p(y|\theta_1, \theta_2)$$

by integrating with respect to θ_2:

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2$$

Nuisance parameters (cont'd)

- Note too that

$$p(\theta_1|y) = \int p(\theta_1, |\theta_2, y)p(\theta_2|y)d\theta_2$$

- The marginal of θ_1 is a **mixture of conditionals** on θ_2, or a **weighted average** of the conditional evaluated at different values of θ_2. Weights are given by marginal $p(\theta_2|y)$

Nuisance parameters (cont'd)

- Important difference with frequentists!
- By averaging conditional $p(\theta_1, |\theta_2, y)$ over possible values of θ_2, we explicitly recognize our uncertainty about θ_2.
- Two extreme cases:
  1. Almost certainty about the value of θ_2: If prior and sample are very informative about θ_2, marginal $p(\theta_2|y)$ will be concentrated around some value $\hat{\theta}_2$. In that case,

$$p(\theta_1|y) \approx p(\theta_1|\hat{\theta}_2, y)$$

  2. Lots of uncertainty about θ_2: Marginal $p(\theta_2|y)$ will assign relatively high probability to wide range of values of θ_2. Point estimate $\hat{\theta}_2$ no longer "reliable". Important to average over range of values of θ_2.

Nuisance parameters (cont'd)

- In most cases, integral not computed explicitly
- Instead, use a two-step simulation approach

1. Marginal simulation step: Draw value $\theta_2^{(k)}$ of θ_2 from $p(\theta_2|y)$ for $k = 1, 2, ...$
2. Conditional simulation step: For each $\theta_2^{(k)}$, draw a value of θ_1 from the conditional density $p(\theta_1|\theta_2^{(k)}, y)$

- Effective approach when marginal and conditional are of standard form.
- More sophisticated simulation approaches later.

Example: Normal model

- y_i *iid* from $N(\mu, \sigma^2)$, both unknown
- Non-informative prior for (μ, σ^2) assuming prior independence:

$$p(\mu, \sigma^2) \propto 1 \times \sigma^{-2}$$

- Joint posterior:

$$\begin{aligned} p(\mu, \sigma^2|y) &\propto p(\mu, \sigma^2)p(y|\mu, \sigma^2) \\ &\propto \sigma^{-n-2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \mu)^2) \end{aligned}$$

- Note that

$$\begin{aligned} \sum_{i=1}^n (y_i - \mu)^2 &= \sum_i (y_i^2 - 2\mu y_i + \mu^2) \\ &= \sum_i y_i^2 - 2\mu n\bar{y} + n\mu^2 \\ &= \sum_i (y_i - \bar{y})^2 + n(\bar{y} - \mu)^2 \end{aligned}$$

by adding and subtracting $2n\bar{y}^2$.

Example: Normal model (cont'd)

- Let

$$s^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

- Then can write posterior for (μ, σ^2) as

$$p(\mu, \sigma^2|y) \propto \sigma^{-n-2}\exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])$$

- Sufficient statistics are $(\bar{y}, s^2)$

Example: Conditional posterior $p(\mu|\sigma^2, y)$

- Conditional on σ^2:

$$p(\mu|\sigma^2, y) = N(\bar{y}, \sigma^2/n)$$

- We know this from earlier chapter (posterior of normal mean when variance is known)

- We can also see this by noting that, viewed as a function of μ only:

$$p(\mu|\sigma^2, y) \propto \exp(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2)$$

  that we recognize as the kernel of a $N(\bar{y}, \sigma^2/n)$

Example: Marginal posterior $p(\sigma^2|y)$

- To get $p(\sigma^2|y)$ we need to integrate $p(\mu, \sigma^2|y)$ over μ:

$$\begin{aligned} p(\sigma^2|y) &\propto \int \sigma^{-n-2} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])d\mu \\ &\propto \sigma^{-n-2} \exp(-\frac{(n-1)s^2}{2\sigma^2}) \int \exp(-\frac{n}{2\sigma^2}(\bar{y} - \mu)^2)d\mu \\ &\propto \sigma^{-n-2} \exp(-\frac{(n-1)s^2}{2\sigma^2})\sqrt{2\pi\sigma^2/n} \end{aligned}$$

- Then

$$p(\sigma^2|y) \propto (\sigma^2)^{-(n+1)/2} \exp(-\frac{(n-1)s^2}{2\sigma^2})$$

  which is proportional to a *scaled-inverse* χ^2 distribution with degrees of freedom $(n-1)$ and scale s^2.

Recall classical result: conditional on σ^2, the distribution of the scaled sufficient statistic $(n-1)s^2/\sigma^2$ is χ^2_{n-1}.

Normal model: analytical derivation

- For the normal model, we can derive the marginal $p(\mu|y)$ analytically:

$$\begin{aligned} p(\mu|y) &= \int p(\mu, \sigma^2|y)d\sigma^2 \\ &\propto \int (\frac{1}{2\sigma^2})^{n/2+1} \exp(-\frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2])d\sigma^2 \end{aligned}$$

- Use the transformation

$$z = \frac{A}{2\sigma^2}$$

  where $A = (n-1)s^2 + n(\bar{y} - \mu)^2$. Then

$$\frac{d\sigma^2}{dz} = -\frac{A}{2z^2}$$

and

$$
\begin{aligned}
p(\mu|y) &\propto \int_0^\infty (\frac{z}{A})^{\frac{n}{2}+1}\frac{A}{z^2}\exp(-z)dz \\
&\propto A^{-n/2}\int z^{\frac{n}{2}-1}\exp(-z)dz
\end{aligned}
$$

Normal model: analytical derivation

$$
p(\mu|y) \propto A^{-n/2}\int z^{\frac{n}{2}-1}\exp{(-z)}dz
$$

- Integrand is unnormalized Gamma(n/2, 1), so integral is constant w.r.t. μ
- Recall that $A = (n-1)s^2 + n(\bar{y}-\mu)^2$. Then

$$
\begin{aligned}
p(\mu|y) &\propto A^{-n/2} \\
&\propto [(n-1)s^2 + n(\bar{y}-\mu)^2]^{-n/2} \\
&\propto [1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}]^{-n/2}
\end{aligned}
$$

the kernel of a $t-$distribution with $n-1$ degrees of freedom, centered at $\bar{y}$ and with scale parameter s^2/n

- For the non-informative prior $p(\mu,\sigma^2) \propto \sigma^{-2}$, the posterior distribution of μ is a non-standard t. Then,

$$
p(\frac{\mu-\bar{y}}{s/\sqrt{n}}|y) = t_{n-1}
$$

the standard t distribution.

Normal model: analytical derivation

- We saw that

$$
p(\frac{\mu-\bar{y}}{s/\sqrt{n}}|y) = t_{n-1}
$$

- Notice similarity to classical result: for iid normal observations from $N(\mu,\sigma^2)$, given (μ,σ^2), the *pivotal quantity*

$$
\frac{\bar{y}-\mu}{s/\sqrt{n}}|\mu,\sigma^2 \sim t_{n-1}
$$

- A *pivot* is a non-trivial function of the data and the parameter(s) θ whose distribution, given θ, is independent of θ. Property deduced from *sampling distribution* as above.
- Baby example of pivot property: $y \sim N(\theta,1)$. Pivot is $x = y - \theta$. Given θ, $x \sim N(0,1)$, independent of θ.

Posterior predictive for future obs

- Posterior predictive distribution for future observation $\tilde{y}$ is a *mixture*:

$$p(\tilde{y}|y) = \int \int p(\tilde{y}|y, \sigma^2, \mu) p(\mu, \sigma^2|y) d\mu d\sigma^2$$

- First factor in integrand is just normal model, and it does not depend on y at all.

- To simulate $\tilde{y}$ from posterior predictive distributions, do the following:

1. Draw σ^2 from Inv-$\chi^2(n-1, s^2)$
2. Draw μ from $N(\bar{y}, \sigma^2/n)$
3. Draw $\tilde{y}$ from $N(\mu, \sigma^2)$

- Can derive the posterior predictive distribution in analytic form. Note that

$$\begin{aligned} p(\tilde{y}|\sigma^2, y) &= \int p(\tilde{y}|\sigma^2, \mu) p(\mu|\sigma^2, y) d\mu \\ &\propto \int \exp(-\frac{1}{2\sigma^2}(\tilde{y}-\mu)^2) \exp(-\frac{n}{2\sigma^2}(\mu - \bar{y})^2) d\mu \end{aligned}$$

- After some algebra:

$$p(\tilde{y}|\sigma^2, y) = N(\bar{y}, (1+\frac{1}{n})\sigma^2)$$

- Using same approach as in deriving posterior distribution of μ, we find that

$$p(\tilde{y}|y) \propto t_{n-1}(\bar{y}, (1+\frac{1}{n})^{1/2} s)$$

Normal data and conjugate prior

- Recall that using a non-informative prior, we found that

$$\begin{aligned} p(\mu|\sigma^2, y) &\propto N(\bar{y}, \sigma^2/n) \\ p(\sigma^2|y) &\propto Inv-\chi^2(n-1, s^2) \end{aligned}$$

- Then, factoring $p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2)$ the conjugate prior for σ^2 would also be scaled inverse χ^2 and for μ (conditional on σ^2) would be normal. Consider

$$\begin{aligned} \mu|\sigma^2 &\sim \mathrm{N}(\mu_0, \sigma^2/\kappa_0) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- Jointly:

$$p(\mu, \sigma^2) \propto \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp(-\frac{1}{2\sigma^2}[\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2])$$

Normal data and conjugate prior

- Note that μ and σ^2 are not independent a priori.

- Posterior density for (μ, σ^2):
  - Multiply likelihood by N-Inv-$\chi^2(\mu_0, \sigma^2/\kappa_0; \nu_0, \sigma_0^2)$ prior
  - Expand the two squares in μ
  - Complete the square by adding and subtracting term depending on $\bar{y}$ and μ_0

  Then $p(\mu, \sigma^2|y) \propto$ N-Inv-$\chi^2(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2)$ where

$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \\ \kappa_n &= \kappa_0 + n, \quad \nu_n = \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu)^2 \end{aligned}$$

Normal data and conjugate prior

- Interpretation of posterior parameters:
  - μ_n a weighted average as earlier
  - $\nu_n \sigma_n^2$: sum of the sample sum of squares, prior sum of squares, and additional uncertainty due to difference between sample mean and prior mean.

Normal data and conjugate prior

- Conditional posterior of μ: As before $\mu | \sigma^2, y \sim$ N$(\mu_n, \sigma^2/\kappa_n)$.
- Marginal posterior of σ^2: As before $\sigma^2 | y \sim$ Inv-$\chi^2(\nu_n, \sigma_n^2)$.
- Marginal posterior of μ: As before $\mu | y \sim t_{\nu_n}(\mu | \mu_n, \sigma_n^2/\kappa_n)$.
- Two ways to sample from joint posterior distribution:

1. Sample μ from t and σ^2 from Inv-χ^2
2. Sample σ^2 from Inv-χ^2 and given σ^2, sample μ from N

Semi-conjugate prior for normal model

- Consider setting independent priors for μ and σ^2:

$$\begin{aligned} \mu &\sim \text{N}(\mu_0, \tau_0^2) \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2) \end{aligned}$$

- Example: mean weight of students, visual inspection shows weights between 100 and 200 pounds.
- $p(\mu, \sigma^2) = p(\mu) p(\sigma^2)$ is *not conjugate* and does not lead to posterior of known form
- Can factor as earlier:

$$\mu | \sigma^2, y \sim \text{N}(\mu_n, \tau_n^2)$$

with

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_2^2} + \frac{n}{\sigma^2}}, \quad \tau_n^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}$$

- NOTE: Even though μ and σ^2 are independent a priori, they are not independent in the posterior.

Semi-conjugate prior and $p(\sigma^2|y)$

- The marginal posterior $p(\sigma^2|y)$ can be obtained by integrating the joint $p(\mu, \sigma^2|y)$ w.r.t. μ:

$$p(\sigma^2|y) \propto \int \text{N}(\mu|\mu_0, \tau_0^2)\ \text{Inv}\chi^2(\sigma^2|\nu_0, \sigma_0^2)\Pi\ \text{N}(y_i|\mu, \sigma^2)d\mu$$

- Integration can be performed by noting that integrand as function of μ is proportional to normal density.

- Keeping track of normalizing constants that depend on σ^2 is messy. Easier to note that:

$$p(\sigma^2|y) = \frac{p(\mu, \sigma^2|y)}{p(\mu|\sigma^2, y)}$$

so that

$$p(\sigma^2|y) \propto \frac{\text{N}(\mu|\mu_0, \tau_0^2)\ \text{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2)\Pi\ \text{N}(y_i|\mu, \sigma^2)}{\text{N}(\mu|\mu_n, \tau_n^2)}$$

which is still a mess.

Semi-conjugate prior and $p(\sigma^2|y)$

- From earlier page:

$$p(\sigma^2|y) \propto \frac{\text{N}(\mu|\mu_0, \tau_0^2)\ \text{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2)\Pi\ \text{N}(y_i|\mu, \sigma^2)}{\text{N}(\mu|\mu_n, \tau_n^2)}$$

- Factors that depend on μ must cancel, and therefore we know that $p(\sigma^2|y)$ does not depend on μ in the sense that we can evaluate $p(\sigma^2|y)$ for a grid of values of σ^2 and *any arbitrary* value of μ.

- Choose $\mu = \mu_n$ and then denominator simplifies to something proportional to τ_n^{-1}. Then

$$p(\sigma^2|y) \propto \tau_n\ \text{N}(\mu|\mu_0, \tau_0^2)\ \text{Inv} - \chi^2(\sigma^2|\nu_0, \sigma_0^2)\Pi\ \text{N}(y_i|\mu, \sigma^2)$$

that can be evaluated for a grid of values of σ^2.

Implementing inverse CDF method

- We often need to draw values from distributions that do not have a "nice" standard form. Example is expression 3.14 in text, for $p(\sigma^2|y)$ in the normal model with semi-conjugate priors for μ and σ^2.

- One approach is the inverse cdf method (see earlier notes), implemented numerically.

- We assume that we can evaluate the pdf (even in unnormalized form) for a grid of values of θ in the appropriate range

Implementing inverse CDF method

- To generate 1000 draws from a distribution $p(\theta|y)$ with parameter θ, do

1. Evaluate $p(\theta|y)$ for a grid of m values of θ. Let the evaluations be denoted by $(p_1, p_2, ..., p_m)$
2. Compute the CDF over the grid: $(p_1, p_1 + p_2, ..., \sum_{i=1}^{m-1} p_i, 1)$ and denote those $(f_1, f_2, ..., f_m)$
3. Generate M uniform random variables u in $[0, 1]$.
4. If $u \in [f_{i-1}, f_i]$, draw θ_i.

Inverse CDF method - Example

θ	Prob(θ)	CDF (F)
1	0.03	0.03
2	0.04	0.07
3	0.08	0.15
4	0.15	0.30
5	0.20	0.50
6	0.30	0.80
7	0.10	0.90
8	0.05	0.95
9	0.03	0.98
10	0.02	1.00

- Consider a parameter θ with values in $(1, 10)$ with probability distribution and cumulative distribution functions as above. Under distribution, values of θ between 4 and 7 are more likely.

Inverse CDF method - Example

- To implement method do:

1. Draw $u \sim U(0, 1)$. For example, in first draw $u = 0.45$
2. For $u \in (f_{i-1}, f_i)$, draw θ_i.
3. For $u = 0.45 \in (0.3, 0.5)$, draw $\theta = 5$.
4. Alternative approach:
   (a) Flip another coin $v \sim U(0, 1)$.
   (b) Pick θ_{i-1} if $v \leq 0.5$ and pick θ_i if $v > 0.5$
5. In example, for $u = 0.45$, would either choose $\theta = 4$ or would choose $\theta = 4$ with probability 1/2.
6. Repeat many times M.

- If M very large, we expect that about 50% of our draws will be $\theta = 4, 5,$ or 6, about 2% will be 10, etc.

Example: Football point spreads

- Data d_i, $i = 1, ..., 672$ are differences between predicted outcome of football game and actual score.

- Normal model:

$$d_i \sim \text{ N } (\mu, \sigma^2)$$

- Priors:

$$\mu \sim \text{ N } (0, 2^2)$$
$$p(\sigma^2) \propto \sigma^{-2}$$

- To draw values of (μ, σ^2) from posterior, do
  - First draw σ^2 from $p(\sigma^2|y)$ using inverse cdf method
  - Then draw μ from $p(\mu|\sigma^2, y)$, normal.

- To draw σ^2, evaluate $p(\sigma^2|y)$ on the grid $[150, 250]$.

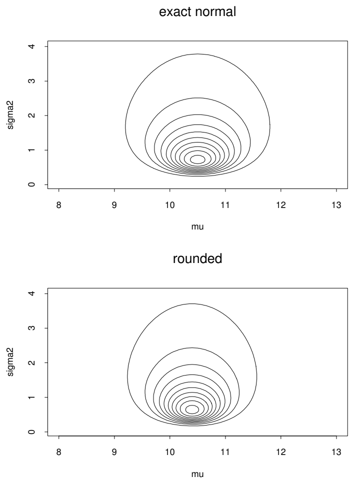
Example: Rounded measurements (prob. 3.5)

- Sometimes, measurements are rounded, and we do not observe "true" values.
- y_i are observed rounded values, and z_i are unobserved true measurements.
- If $z_i \sim$ N (μ, σ^2), then

$$y|\mu, \sigma^2 \sim \Pi_i \Phi \left(\frac{y_i + 0.5 - \mu}{\sigma} \right) - \Phi \left(\frac{y_i - 0.5 - \mu}{\sigma} \right)$$

- Prior: $p(\mu, \sigma^2) \propto \sigma^{-2}$
- We are interested in posterior inference about (μ, σ^2) and in differences between rounded and exact analysis.

Joint posterior contours



Multinomial model - Intro

- Generalization of binomial model, for the case where observations can have more than two possible values.
- Sampling distribution: multinomial with parameters $(\theta_1, ..., \theta_k)$, the probabilities associated to each of the k possible outcomes.
- Example: In a survey, respondents may: Strongly Agree, Agree, Disagree, Strongly Disagree, or have No Opinion when presented with a statement such as "Instructor for Stat 544 is spectacular".

Multinomial model - Sampling dist'n

- Formally:
  - $y = k \times 1$ vector of counts of $\#$s of observations in each outcome
  - θ_j: probability of jth outcome
  - $\sum_{j=1}^k \theta_j = 1$ and $\sum_{j=1}^k y_j = n$
- Sampling distribution:

$$p(y|\theta) \propto \Pi_{j=1}^k \theta_j^{y_j}$$

Multinomial model - Prior

- Conjugate prior for $(\theta_1, ..., \theta_k)$ is **Dirichlet** distribution, a multivariate generalization of the Beta:

$$p(\theta|\alpha) \propto \Pi_{j=1}^k \theta_j^{\alpha_j - 1}$$

with

$$\alpha_j > 0 \forall j, \quad \text{and} \quad \alpha_0 = \sum_{j=1}^k \alpha_j$$

$$\theta_j > 0 \forall j, \quad \text{and} \quad \sum_{j=1}^k \theta_j = 1$$

- The α_j can be thought of as "prior counts" associated to jth outcome, so that α_0 would then be a "prior sample size".

- For the Dirichlet:

$$E(\theta_j) = \frac{\alpha_j}{\alpha_0}, \qquad Var(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$$

$$Cov(\theta_i, \theta_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Dirichlet distribution

- The Dirichlet distribution is the conjugate prior for the parameters of the multinomial model.

- If $(\theta_1, \theta_2, ...\theta_K) \sim D(\alpha_1, \alpha_2, ..., \alpha_K)$ then

$$p(\theta_1, ...\theta_k) = \frac{\Gamma(\alpha_0)}{\Pi_j \Gamma(\alpha_j)} \Pi_j \theta_j^{\alpha_j - 1},$$

where $\theta_j \geq 0$, $\Sigma_j \theta_j = 1$, $\alpha_j \geq 0$ and $\alpha_0 = \Sigma_j \alpha_j$.

- Some properties:

$$E(\theta_j) = \frac{\alpha_j}{\alpha_0}$$

$$Var(\theta_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}$$

$$Cov(\theta_i, \theta_j) = -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}$$

Note that the θ_j are negatively correlated.

- Because of the sum to one restriction, the pdf of the K-dimensional random vector can be written in terms of $K - 1$ random variables.

- The marginal distributions can be shown to be Beta.

- Proof for $K = 3$:

$$p(\theta_1, \theta_2) = \frac{\Gamma(\alpha_1, \alpha_2, \alpha_3)}{\Pi_j \Gamma(\alpha_0)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} (1 - \theta_1 - \theta_2)^{\alpha_3 - 1}$$

- To get marginal for θ_1, integrate $p(\theta_1, \theta_2)$ with respect to θ_2 with limits of integration 0 and $1 - \theta_1$. Call normalizing constant Q and use change of variable:
$$v = \frac{\theta_2}{1 - \theta_1}$$
so that $\theta_2 = v(1 - \theta_1)$ and $d\theta_2 = (1 - \theta_1)dv$.

- The marginal is then:

$$\begin{aligned}
p(\theta_1) &= Q \int_0^1 \theta_1^{\alpha_1 - 1}(v(1-\theta_1))^{\alpha_2 - 1} \\
&\quad (1 - \theta_1 - v(1-\theta_1))^{\alpha_3 - 1}(1-\theta_1)dv \\
&= Q\theta_1^{\alpha_1 - 1}(1-\theta_1)^{\alpha_2 + \alpha_3 - 1} \int_0^1 v^{\alpha_2 - 1}(1-v)^{\alpha_3 - 1}dv \\
&= Q\theta_1^{\alpha_1 - 1}(1-\theta_1)^{\alpha_2 + \alpha_3 - 1}\frac{\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_2 + \alpha_3)}
\end{aligned}$$

- Then,
$$\theta_1 \sim Beta(\alpha_1, \alpha_2 + \alpha_3).$$

- This generalizes to what is called the *clumping property* of the Dirichlet. In general, if $(\theta_1, ..., \theta_k) \sim D(\alpha_1, ..., \alpha_K)$:

$$p(\theta_i) = Beta(\alpha_1, \alpha_0 - \alpha_i)$$

Multinomial model - Posterior

- Posterior must have Dirichlet form also:

$$\begin{aligned}
p(\theta|y) &\propto \Pi_{j=1}^k \theta_j^{\alpha_j - 1}\theta_j^{y_j} \\
&\propto \Pi_{j=1}^k \theta_j^{y_j + \alpha_j - 1}
\end{aligned}$$

- $\alpha_n = \sum_j(\alpha_j + y_j) = \alpha_0 + n$ is “total” number of “observations”.

- Posterior mean (a point estimate) of θ_j is:

$$\begin{aligned}
E(\theta_j|y) &= \frac{\alpha_j + y_j}{\alpha_0 + n} \\
&= \frac{\text{“\#” of obs. of jth outcome}}{\text{“total” \# of obs}}
\end{aligned}$$

- For $\alpha_j = 1 \ \forall j$: uniform non-informative prior on all vectors of θ_j such that $\sum_j \theta_j = 1$

- For $\alpha_j = 0 \ \forall j$: the uniform prior is on the $\log(\theta_j)$ with same restriction.

- In either case, posterior is proper if $y_j \geq 1 \ \forall j$.

Multinomial model - samples from $p(\theta|y)$

- Gamma method: Two steps for each θ_j:

1. Draw $x_1, x_2, ..., x_k$ from independent $Gamma(\delta, (\alpha_j + y_j))$ for any common δ
2. Set $\theta_j = x_j / \sum_{i=1}^{k} x_i$

- Beta method: Relies on properties of Dirichlet:
  - Marginal $p(\theta_j|y) = Beta(\alpha_j + y_j, \alpha_n - (\alpha_j + y_j))$
  - Conditional $p(\theta_j|\theta_{-j}, y) = Dirichlet$

Multinomial model - Example

- Pre-election polling in 1988:
  - n = 1,447 adults in the US
  - y_1 = 727 supported G. Bush (the elder)
  - y_2 = 583 supported M. Dukakis
  - y_3 = 137 supported other or had no opinion

- If no other information available, can assume that observations are exchangeable given θ. (However, if information on party affiliation is available, unreasonable to assume exchangeability)

- Polling is done under complex survey design. Ignore for now and assume simple random sampling. Then, $(y_1, y_2, y_3) \sim Mult(\theta_1, \theta_2, \theta_3)$

- Of interest: $\theta_1 - \theta_2$, the difference in the population in support of the two major candidates.

Multinomial model - Example

- Non-informative prior, for now: set $\alpha_1 = \alpha_2 = \alpha_3 = 1$

- Posterior distribution is Dirichlet(728, 584, 138). Then:

$$E(\theta_1|y) = 0.502, \quad E(\theta_2|y) = 0.403$$

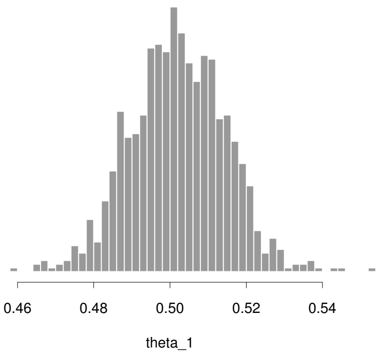
- Other quantities obtain by simulation (see next)

- To derive $p(\theta_1 - \theta_2|y)$ do:

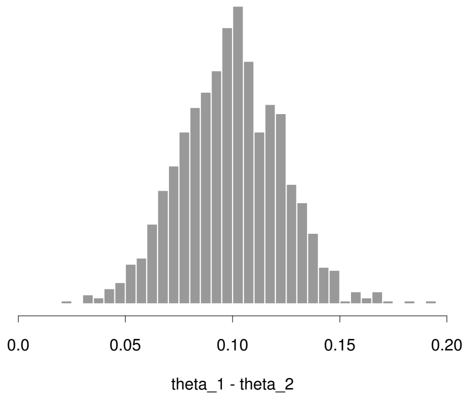
1. Draw m values of $(\theta_1, \theta_2, \theta_3)$ from posterior
2. For each draw, compute $\theta_1 - \theta_2$

- Results and program: see quantiles of posterior distributions of θ_1 and θ_2, credible set for $\theta_1 - \theta_2$ and Prob$(\theta_1 > \theta_2|y)$.

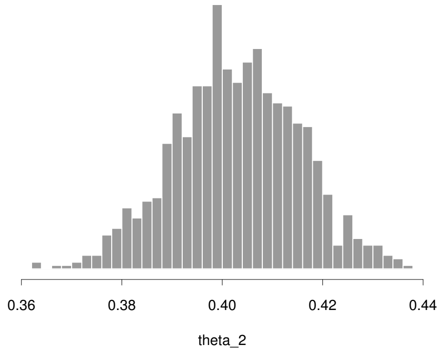
Proportion voting for Bush (the elder)



Difference between two candidates



Proportion voting for Dukakis



Example: Bioassay experiment

- Does mortality in lab animals increase with increased dose of some drug?
- Experiment: 20 animals randomly allocated to four doses ("treatments"), and number of dead animals within each dose recorded.

Dose x_i	n_i	No. of deaths y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

- Animals exchangeable *within dose*.

Bioassay example (cont'd)

- Model
$$y_i \sim \text{ Bin } (n_i, \theta_i)$$
- θ_i are not exchangeable because probability of death depends on dose.
- One way to model is with linear relationship:
$$\theta_i = \alpha + \beta x_i.$$
Not a good idea because $\theta_i \in (0, 1)$.
- Transform θ_i:
$$\text{logit } (\theta_i) = \log \left(\frac{\theta_i}{1 - \theta_i} \right)$$

- Since logit $(\theta_i) \in (-\infty, +\infty)$, can use linear model. (Logistic regression)

$$E[\text{logit } (\theta_i)] = \alpha + \beta x_i.$$

- Likelihood: if $y_i \sim$ Bin (n_i, θ_i), then

$$p(y_i|n_i, \theta_i) \propto (\theta_i)^{y_i}(1-\theta_i)^{n_i - y_i}.$$

- But recall that

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + \beta x_i,$$

so that

$$\theta_i = \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

- Then

$$p(y_i| \alpha, \beta, n_i, x_i) \propto \left[\frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right]^{y_i}\left[1 - \frac{\exp(\alpha+\beta x_i)}{1+\exp(\alpha+\beta x_i)}\right]^{n_i - y_i}$$

- Prior: $p(\alpha, \beta) \propto 1$.
- Posterior:

$$p(\alpha, \beta|y, n, x) \propto \Pi_{i=1}^{k} p(y_i|\alpha, \beta, n_i, x_i).$$

- We first evaluate $p(\alpha, \beta|y, n, x)$ on the grid

$$(\alpha, \beta) \in [-5, 10] \times [-10, 40]$$

and use inverse cdf method to sample from posterior.

Bioassay example (cont'd)

Posterior evaluated on 200×200 grid of values of (α, β).

	α_1	α_2	...	...	α_{200}	
β_1						$p(\beta_1\|y)$
β_2						$p(\beta_2\|y)$
$\vdots$						$\vdots$
$\vdots$						$\vdots$
β_{200}						$p(\beta_{200}\|y)$
	$p(\alpha_1\|y)$	$p(\alpha_2\|y)$	...	...	$p(\alpha_{200}\|y)$	

- Entry (j, i) in grid is $p(\alpha, \beta|\alpha = \alpha_i, \beta = \beta_j, y)$.
- Sum of column j entries is $p(\alpha_j|y)$ because:

$$p(\alpha_j|y) = \int p(\alpha_j, \beta|y) d\beta$$

$$\begin{aligned} &= \int p(\alpha_j|\beta, y) p(\beta|y) d\beta \\ &\approx \sum p(\alpha_j|\beta, y) \end{aligned}$$

- Sum of row i entries is $p(\beta_i|y)$.

Bioassay example (cont'd)

- We sample α from $p(\alpha|y)$, and then β from $p(\beta|\alpha, y)$ (or the other way around).
- To sample α from $p(\alpha|y)$:
  1. Obtain empirical $p(\alpha|\alpha = \alpha_j, y)$, $j = 1, ...200$ by summing over the β.
  2. Use inverse cdf method to sample from $p(\alpha|y)$.
- To sample β from $p(\beta|\alpha, y)$:
  1. Given a draw α^*, choose appropriate column in grid
  2. Use inverse cdf method on $p(\beta|\alpha = \alpha^*, y)$.

Bioassay example (cont'd)

LD_{50} is dose at which probability of death is 0.5, or

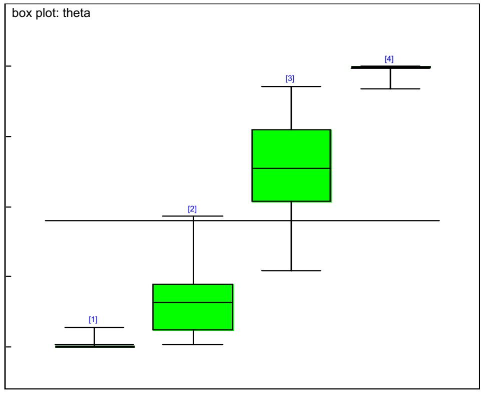
$$E\left(\frac{y_i}{n_i}\right) = \theta_i = 0.5$$

so that

$$\begin{aligned} 0.5 &= \text{logit}^{-1}(\alpha + \beta x_i) \\ \text{logit}(0.5) &= \alpha + \beta x_i \\ 0 &= \alpha + \beta x_i \end{aligned}$$

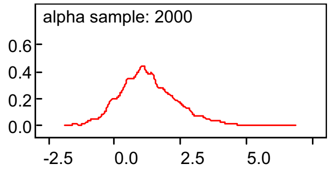
Then LD_{50} is computed as $x_i = -\alpha/\beta$.

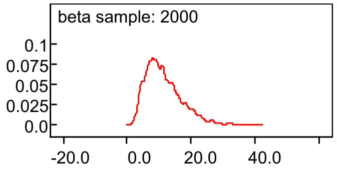
Posterior distribution of the probability of death for each dose



Posterior distributions of the regression parameters

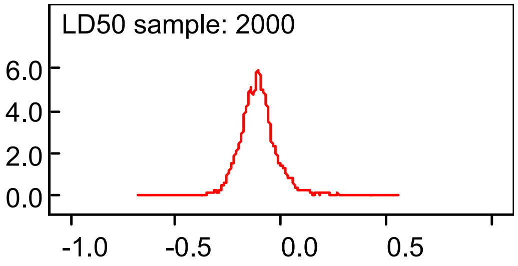
Alpha: posterior mean = 1.297, with 95% credible set [-0.365, 3.755]
Beta: posterior mean = 11.52, with 95% credible set [3.572, 25.79]





Posterior distribution of the LD50 in the log scale

Posterior mean: -0.1064
95% credible set: [-0.268, 0.1141]



Multivariate normal model - Intro

- Generalization of the univariate normal model, where now observations are $d \times 1$ vectors of measurements
- For the ith sample unit: $y_i = (y_{i_1}, y_{i_2}, ..., y_{i_d})$, and $y_i \sim N(\mu, \Sigma)$, with (μ, Σ) $d \times 1$ and $d \times d$ respectively
- For a sample of n iid observations:

$$\begin{aligned}
p(y|\mu,\Sigma) &\propto |\Sigma|^{-n/2}\exp[-\frac{1}{2}\sum_i (y_i-\mu)'\Sigma^{-1}(y_i-\mu)] \\
&\propto |\Sigma|^{-n/2}\exp[-\frac{1}{2}\text{tr}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)] \\
&\propto |\Sigma|^{-n/2}\exp[-\frac{1}{2}\text{tr}\Sigma^{-1}S]
\end{aligned}$$

with

$$S = \sum_i (y_i-\mu)(y_i-\mu)'$$

- S is the $d \times d$ matrix of sample squared deviations and cross-deviations from μ.

Multivariate normal model - Known Σ

- We want the posterior distribution of the mean $p(\mu|y)$ under the assumption that the variance matrix is known.
- Conjugate prior for μ is $p(\mu|\mu_0, \Lambda_0) = N(\mu_0, \Lambda_0)$, as in the univariate case
- Posterior for μ:

$$\begin{aligned}
p &\quad (\mu|y,\Sigma) \propto \\
\exp &\quad \left\{-\frac{1}{2}\left[(\mu-\mu_0)'\Lambda^{-1}(\mu-\mu_0) + \sum_i (y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right]\right\}
\end{aligned}$$

that is a quadratic function of μ. Completing the square in the exponent:

$$p(\mu|y,\Sigma) = N(\mu_n, \Lambda_n)$$

with

$$\begin{aligned}\mu_n &= (\Lambda^{-1}\mu_0 + n\Sigma^{-1}\bar{y})(\Lambda^{-1} + n\Sigma^{-1})^{-1} \\ \Lambda_n &= (\Lambda^{-1} + n\Sigma^{-1})^{-1}\end{aligned}$$

Multivariate normal model - Known Σ

- Posterior precision is equal to the sum of prior and sample precisions:

$$\Lambda_n^{-1} = \Lambda_0^{-1} + n\Sigma^{-1}$$

- Posterior mean is weighted average of prior and sample means:

$$\mu_n = (\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y})(\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

- From usual properties of multivariate normal distribution, can derive marginal posterior distribution of any subvector of μ or conditional posterior distribution of $\mu^{(1)}$ given $\mu^{(2)}$.

- Let $\mu' = (\mu^{(1)'}, \mu^{(2)'})'$ and correspondingly, let

$$\mu_n = \left[\begin{array}{c}\mu_n^{(1)} \\ \mu_n^{(2)}\end{array}\right]$$

and

$$\Lambda_n = \left[\begin{array}{c}\Lambda_n^{(1,1)}\Lambda_n^{(1,2)} \\ \Lambda_n^{(2,1)}\Lambda_n^{(2,2)}\end{array}\right]$$

Multivariate normal model - Known Σ

- Marginal of subvector $\mu^{(1)}$ is $p(\mu^{(1)}|\Sigma, y) = N(\mu_n^{(1)}, \Lambda_n^{(1,1)})$.
- Conditional of subvector $\mu^{(1)}$ given $\mu^{(2)}$ is

$$p(\mu^{(1)}|\mu^{(2)}, \Sigma, y) = N(\mu_n^{(1)} + \beta^{1|2}(\mu^{(2)} - \mu_n^{(2)}), \Lambda^{1|2})$$

where

$$\begin{aligned}\beta^{1|2} &= \Lambda_n^{(1,2)}\left(\Lambda_n^{(2,2)}\right)^{-1} \\ \Lambda^{1|2} &= \Lambda_n^{(1,1)} - \Lambda_n^{(1,2)}\left(\Lambda_n^{(2,2)}\right)^{-1}\Lambda_n^{(2,1)}\end{aligned}$$

- We recognize $\beta^{1|2}$ as the regression coefficient in a regression of $\mu^{(1)}$ on $\mu^{(2)}$

Multivariate normal - Posterior predictive

- We seek $p(\tilde{y}|y)$, and note that

$$p(\tilde{y}, \mu|y) = N(\tilde{y}; \mu, \Sigma)N(\mu; \mu_n, \Lambda_n)$$

- Exponential in $p(\tilde{y}, \mu|y)$ is a quadratic form in $(\tilde{y}, \mu)$, so $(\tilde{y}, \mu)$ have a normal joint posterior distribution.

- Marginal $p(\tilde{y}|y)$ is also normal with mean and variance:

$$\begin{aligned} \mathsf{E}(\tilde{y}|y) &= \mathsf{E}(\mathsf{E}(\tilde{y}|\mu, y)|y) = \mathsf{E}(\mu|y) = \mu_n \\ \mathsf{var}(\tilde{y}|y) &= \mathsf{E}(\mathsf{var}(\tilde{y}|\mu, y)|y) + \mathsf{var}(\mathsf{E}(\tilde{y}|\mu, y)|y) \\ &= \mathsf{E}(\Sigma|y) + \mathsf{var}(\mu|y) \\ &= \Sigma + \Lambda_n. \end{aligned}$$

Multivariate normal - Sampling from $p(\tilde{y}|y)$

- With Σ known, to draw a value $\tilde{y}$ from $p(\tilde{y}|y)$ note that

$$p(\tilde{y}|y) = \int p(\tilde{y}|\mu, y)p(\mu|y)d\mu$$

- Then:
  1. Draw μ from $p(\mu|y) = N(\mu_n, \Lambda_n)$
  2. Draw $\tilde{y}$ from $p(\tilde{y}|\mu, y) = N(\mu, \Sigma)$

- Alternatively (better), draw $\tilde{y}$ directly from

$$p(\tilde{y}|y) = N(\mu_n, \Sigma + \Lambda_n)$$

Sampling from a multivariate normal

- Want to sample y $d \times 1$ from $N(\mu, \Sigma)$.

- **Two common approaches**
  - Using Cholesky decomposition of Σ:
    1. Get A, a lower triangular matrix such that $AA' = \Sigma$
    2. Draw $(z_1, z_2, ..., z_d)$ iid $N(0, 1)$ and let $z = (z_1, ..., z_d)$
    3. Compute $y = \mu + Az$

Sampling from a multivariate normal

- Using sequential conditional sampling:

- Use fact that all conditionals in a multivariate normal are also normal.
  1. Draw y_1 from $N(\mu_1, \Sigma^{(11)})$
  2. Then draw $y_2|y_1$ from $N(\mu^{2|1}, \Sigma^{2|1})$
  3. Etc.

- For example, if $d = 2$,
  1. Draw y_1 from
  $$N(\mu^{(1)}, \sigma^{2(1)})$$
  2. Draw y_2 from
  $$N(\mu^{(2)} + \frac{\sigma^{(12)}}{\sigma^{2(2)}}(y_1 - \mu^{(1)}), \sigma^{2(2)} - \frac{(\sigma^{(12)})^2}{\sigma^{2(1)}})$$

Non-informative prior for μ

- A non-informative prior for μ is obtained by letting the prior precision Λ_0 go to zero.

- With the uniform prior, the posterior for μ is proportional to the likelihood.

- Posterior is proper only if $n > d$; otherwise, S is not of full column rank.

- If $n > d$,
$$p(\mu|\Sigma, y) = N(\bar{y}, \Sigma/n)$$

Multivariate normal - unknown μ, Σ

- The conjugate family of priors for (μ, Σ) is the normal-inverse Wishart family.

- The inverse Wishart is the multivariate generalization of the inverse χ^2 distribution

- If $\Sigma|\nu_0, \Lambda_0 \sim$ Inv-Wishart$_{\nu_0}(\Lambda_0^{-1})$ then
$$p(\Sigma|\nu_0, \Lambda_0) \propto |\Sigma|^{-(\nu_0+d+1)/2} \exp\left\{-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1})\right\}$$
for Σ positive definite and symmetric.

- For the Inv-Wishart with ν_0 degrees of freedom and scale Λ_0: $E(\Sigma) = (\nu_0 - d - 1)^{-1}\Lambda_0$

Multivariate normal - unknown μ, Σ

- Then conjuage prior is
$$\begin{aligned} \Sigma &\sim \text{Inv-Wishart}_{\nu_0}(\Lambda_0^{-1}) \\ \mu|\Sigma &\sim \text{N}(\mu_0, \Sigma/\kappa_0) \end{aligned}$$
which corresponds to
$$\begin{aligned} p(\mu, \Sigma) \propto& \ |\Sigma|^{-(\nu_0+d)/2+1} \\ & \exp\left(-\frac{1}{2}\text{tr}(\Lambda_0\Sigma^{-1}) - \frac{\kappa_0}{2}(\mu-\mu_0)'\Sigma^{-1}(\mu-\mu_0)\right) \end{aligned}$$

Multivariate normal - unknown μ, Σ

- Posterior must also be in the Normal-Inv-Wishart form

- Results from univariate normal generalize directly:
  - $\Sigma|y \sim$ Inv-Wishart$_{\nu_n}(\Lambda_n^{-1})$
  - $\mu|\Sigma, y \sim N(\mu_n, \Sigma/\kappa_n)$
  - $\mu|y \sim$ Mult-$t_{\nu_n-d+1}(\mu_n, \Lambda_n/(\kappa_n(\nu_n - d + 1)))$
  - $\tilde{y}|y \sim$ Mult-t

- Here:
$$\begin{aligned} \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \end{aligned}$$

$$\begin{aligned}\Lambda_n &= \Lambda_0 + S + \frac{n\kappa_0}{\kappa_0 + n}(\bar{y} - \mu_0)(\bar{y} - \mu_0)' \\ S &= \sum_i (y_i - \bar{y})(y_i - \bar{y})'\end{aligned}$$

Multivariate normal - sampling μ, Σ

- To sample from $p(\mu, \Sigma|y)$ do: (1) Sample Σ from $p(\Sigma|y)$ (2) Sample μ from $p(\mu|\Sigma, y)$
- To sample from $p(\tilde{y}|y)$, use drawn (μ, Σ) and obtain draw $\tilde{y}$ from $N(\mu, \Sigma)$
- Sampling from Wishart$_\nu(S)$:
  1. Draw ν independent $d \times 1$ vectors $\alpha_1, \alpha_2, ..., \alpha_\nu$ from a $N(0, S)$
  2. Let $Q = \sum_{i=1}^{\nu} \alpha_i \alpha_i'$
- Method works when $\nu > d$. If $Q \sim$ Wishart then $Q^{-1} = \Sigma \sim$ Inv-Wishart.
- We have already seen how to sample from a multivariate normal given mean vector and covariance matrix.

Multivariate normal - Jeffreys prior

- If we start from the conjugate normal-inv-Wishart prior and let :

$$\begin{aligned}\kappa_0 &\to 0 \\ \nu_0 &\to -1 \\ |\Lambda_0| &\to 0\end{aligned}$$

then resulting prior is Jeffreys prior for (μ, Σ):

$$p(\mu, \Sigma) \propto |\Sigma|^{-(d+1)/2}$$

Example: bullet lead concentrations

- In the US, four major manufacturers produce all bullets fired. One of them is Cascade.
- A sample of 200 round-nosed .38 caliber bullets with lead tips were obtained from Cascade.
- Concentration of five elements (antimony, copper, arsenic, bismuth, and silver) in the lead alloy was obtained. Data for for Cascade are stored in federal.data in the course's web site.
- Assuming that the 200 5×1 observation vectors for Cascade can be represented by a multivariate normal distribution (perhaps after transformation), we are interested in the posterior distribution of the mean vector and of functions of the covariance parameters.

- Particular quantities of interest for inference include:
  - The mean trace element concentrations $(\mu_1, ..., \mu_5)$
  - The correlations between trace element concentrations $(\rho_{11}, \rho_{12}, ..., \rho_{45})$.
  - The largest eigenvalue of the covariance matrix

- In the data file, columns correspond to antimony, copper, arsenic, bismuth, and silver, in that order.

- For this example, the **antimony concentration was divided by 100**

Bullet lead example - Model and prior

- We concentrate on the correlations that involve copper (four of them).

- Sampling distribution:

$$y|\mu, \Sigma \sim \mathsf{N}\ (\mu, \Sigma)$$

- We use the conjugate Normal-Inv-Wishart family of priors, and choose parameters for $p(\Sigma|\nu_0, \Lambda_0)$ first:

$$\begin{aligned}\Lambda_{0,ii} &= (100,\ 5000,\ 15000,\ 1000,\ 100)\\ \Lambda_{0,ij} &= 0\ \forall i, j\\ \nu_0 &= 7\end{aligned}$$

- For $p(\mu|\Sigma, mu_0, \kappa_0)$:

$$\begin{aligned}\mu_0 &= (200, 200, 200, 100, 50)\\ \kappa_0 &= 10\end{aligned}$$

- Low values for ν_0 and κ_0 suggest little confidence in prior guesses μ_0, Λ_0

- We set Λ_0 to be diagonal apriori: we have some information about the variances of the five element concentrations, but no information about their correlation.

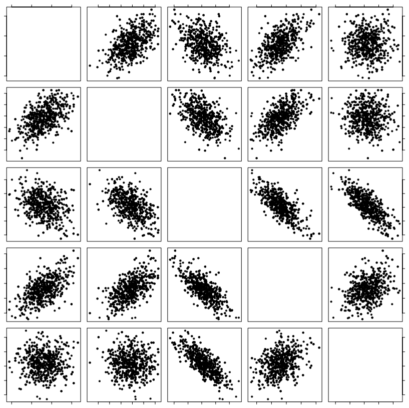
- Sampling from posterior distribution: We follow the usual approach:

1. Draw Σ from Inv-Wishart$_{\nu_n}(\Lambda_n^{-1})$
2. Draw μ from $N(\mu_n, \Sigma/\kappa_n)$

3. For each draw, compute:
   (a) the ratio between λ_1 (largest eigenvalue of Σ) and λ_2 (second largest eigenvalue of Σ)
   (b) the four correlations $\rho_{copper,j} = \sigma_{copper,j}/(\sigma_{copper}\sigma_j)$, for $j \in$ {antimony, arsenic, bismuth, silver}.

- We are interested in the posterior distributions of the five means, the eigenvalues ratio, and the four correlation coefficients

Scatter plot of the posterior mean concentrations

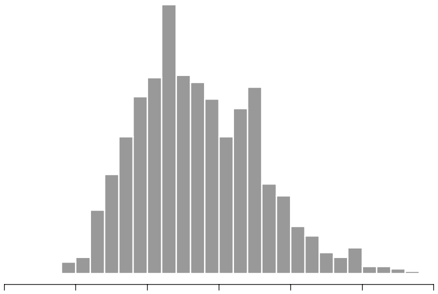


Results from R program

```
# antimony
c(mean(sample.mu[,1]),sqrt(var(sample.mu[,1])))
[1] 265.237302   1.218594
quantile(sample.mu[,1],probs=c(0.025,0.05,0.5,0.95,0.975))
    2.5%     5.0%    50.0%    95.0%    97.5%
 262.875 263.3408 265.1798 267.403 267.7005
# copper
c(mean(sample.mu[,2]),sqrt(var(sample.mu[,2])))
[1] 259.36335   5.20864
quantile(sample.mu[,2],probs=c(0.025,0.05,0.5,0.95,0.975))
     2.5%      5.0%     50.0%     95.0%     97.5%
 249.6674 251.1407 259.5157 268.0606 269.3144
# arsenic
c(mean(sample.mu[,3]),sqrt(var(sample.mu[,3])))
[1] 231.196891   8.390751
quantile(sample.mu[,3],probs=c(0.025,0.05,0.5,0.95,0.975))
```

```
     2.5%      5.0%     50.0%     95.0%     97.5%
 213.5079 216.7256 231.6443 243.5567 247.1686
# bismuth
c(mean(sample.mu[,4]),sqrt(var(sample.mu[,4])))
[1] 127.248741   1.553327
quantile(sample.mu[,4],probs=c(0.025,0.05,0.5,0.95,0.975))
     2.5%      5.0%     50.0%     95.0%    97.5%
 124.3257 124.6953 127.2468 129.7041 130.759
# silver
c(mean(sample.mu[,5]),sqrt(var(sample.mu[,5])))
[1] 38.2072916  0.7918199
quantile(sample.mu[,5],probs=c(0.025,0.05,0.5,0.95,0.975))
     2.5%      5.0%    50.0%     95.0%    97.5%
 36.70916 36.93737 38.1971 39.54205 39.73169
```

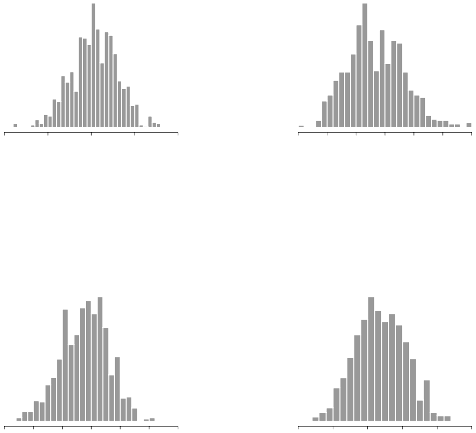
Posterior of λ_1/λ_2 of Σ



Summary statistics of posterior dist. of ratio

```
c(mean(ratio.l),sqrt(var(ratio.l)))
[1] 5.7183875 0.8522507
quantile(ratio.l,probs=c(0.025,0.05,0.5,0.95,0.975))
     2.5%     5.0%    50.0%    95.0%    97.5%
 4.336424 4.455404 5.606359 7.203735 7.619699
```

Correlation of copper with other elements



Summary statistics for correlations

```
# j = antimony
c(mean(sample.rho[,1]),sqrt(var(sample.rho[,1])))
[1] 0.50752063 0.05340247
quantile(sample.rho[,1],probs=c(0.025,0.05,0.5,0.95,0.975))
     2.5%      5.0%     50.0%     95.0%     97.5%
 0.4014274 0.4191201 0.5076544 0.5901489 0.6047564
# j = arsenic
c(mean(sample.rho[,3]),sqrt(var(sample.rho[,3])))
[1] -0.56623609  0.04896403
quantile(sample.rho[,3],probs=c(0.025,0.05,0.5,0.95,0.975))
      2.5%       5.0%      50.0%      95.0%     97.5%
 -0.6537461 -0.6448817 -0.570833 -0.4857224 -0.465808
# j = bismuth
c(mean(sample.rho[,4]),sqrt(var(sample.rho[,4])))
[1] 0.63909311 0.04087149
quantile(sample.rho[,4],probs=c(0.025,0.05,0.5,0.95,0.975))
```

```
     2.5%      5.0%     50.0%     95.0%     97.5%
 0.5560137 0.5685715 0.6429459 0.7013519 0.7135556
# j = silver
c(mean(sample.rho[,5]),sqrt(var(sample.rho[,5])))
[1] 0.03642082 0.07232010
quantile(sample.rho[,5],probs=c(0.025,0.05,0.5,0.95,0.975))
       2.5%       5.0%      50.0%     95.0%     97.5%
 -0.09464575 -0.0831816 0.03370379 0.164939 0.1765523
```

S-Plus code

```
# Cascade example
# We need to define two functions, one to generate random vectors from
# a multivariate normal distribution and the other to generate random
# matrices from a Wishart distribution.
# Note, that a function that generates random matrices from a
# inverse Wishart distribution is not necessary to be defined
# since if W ~ Wishart(S) then W^(-1) ~ Inv-Wishart(S^(-1))
#

# A function that generates random observations from a
# Multivariate Normal distribution:  rmnorm(a,b)
# the parameters of the function are
#             a = a column vector of kx1
#             b = a definite positive kxk matrix
#
rmnorm_function(a,b) {

k_nrow(b)
zz_t(chol(b))%*%matrix(rnorm(k),nrow=k)+a
zz                           }
# A function that generates random observations from a
# Wishart distribution:  rwishart(a,b)
# the parameters of the function are
#                   a = df
#                   b = a definite positive kxk matrix
# Note a must be > k
```

```
rwishart_function(a,b) {
        k_ncol(b)
        m_matrix(0,nrow=k,ncol=1)
cc_matrix(0,nrow=a,ncol=k)
for (i in 1:a) { cc[i,]_rmnorm(m,b) }
        w_t(cc)%*%cc
        w                                       }
#
#Read the data
#
y_scan(file="cascade.data")
y_matrix(y,ncol=5,byrow=T)
y[,1]_y[,1]/100
means_rep(0,5)
for (i in 1:5){ means[i]_mean(y[,i]) }
means_matrix(means,ncol=1,byrow=T)
n_nrow(y)
#
#Assign values to the prior parameters
#
v0_7
Delta0_diag(c(100,5000,15000,1000,100))
k0_10
mu0_matrix(c(200,200,200,100,50),ncol=1,byrow=T)
# Calculate the values of the parameters of the posterior
#
mu.n_(k0*mu0+n*means)/(k0+n)
v.n_v0+n
k.n_k0+n
```

```
ones_matrix(1,nrow=n,ncol=1)
S = t(y-ones%*%t(means))%*%(y-ones%*%t(means))
Delta.n_Delta0+S+(k0*n/(k0+n))*(means-mu0)%*%t(means-mu0)
#
# Draw Sigma and mu from their posteriors
#
samplesize_200
lambda.samp_matrix(0,nrow=samplesize,ncol=5) #this matrix will store
                                             #eigenvalues of Sigma
rho.samp_matrix(0,nrow=samplesize,ncol=5)
mu.samp_matrix(0,nrow=samplesize,ncol=5)
for (j in 1:samplesize) {

# Sigma
          SS = solve(Delta.n)
        # The following makes sure that SS is symmetric
          for (pp in 1:5) { for (jj in 1:5) {
             if(pp<jj){SS[pp,jj]=SS[jj,pp]} } }
  Sigma_solve(rwishart(v.n,SS))
        # The following makes sure that Sigma is symmetric
          for (pp in 1:5) { for (jj in 1:5) {
             if(pp<jj){Sigma[pp,jj]=Sigma[jj,pp]} } }
        # Eigenvalue of Sigma
  lambda.samp[j,]_eigen(Sigma)$values
        # Correlation coefficients
          for (pp in 1:5){
          rho.samp[j,pp]_Sigma[pp,2]/sqrt(Sigma[pp,pp]*Sigma[2,2]) }
        # mu
          mu.samp[j,]_rmnorm(mu.n,Sigma/k.n)
```

```
                                                }

# Graphics and summary statistics
sink("cascade.output")
# Calculate the ratio between max(eigenvalues of Sigma)
# and max({eigenvalues of Sigma}\{max(eigenvalues of Sigma)})
#
ratio.l_sample.l[,1]/sample.l[,2]
# "Histogram of the draws of the ratio"
postscript("cascade_lambda.eps",height=5,width=6)
hist(ratio.l,nclass=30,axes = F,
     xlab="eigenvalue1/eigenvalue2",xlim=c(3,9))
axis(1)
dev.off()
# Sumary statistics of the ratio of the eigenvalues
c(mean(ratio.l),sqrt(var(ratio.l)))
quantile(ratio.l,probs=c(0.025,0.05,0.5,0.95,0.975))
#
# correlations with copper
#
postscript("cascade_corr.eps",height=8,width=8)
par(mfrow=c(2,2))
# "Histogram of the draws of corr(antimony,copper)"
hist(sample.rho[,1],nclass=30,axes = F,xlab="corr(antimony,copper)",
     xlim=c(0.3,0.7))
axis(1)
# "Histogram of the draws of corr(arsenic,copper)"
hist(sample.rho[,3],nclass=30,axes = F,xlab="corr(arsenic,copper)")
axis(1)
```

```
# "Histogram of the draws of corr(bismuth,copper)"
hist(sample.rho[,4],nclass=30,axes = F,xlab="corr(bismuth,copper)",
     xlim=c(0.5,.8))
axis(1)
# "Histogram of the draws of corr(silver,copper)"
hist(sample.rho[,5],nclass=25,axes = F,xlab="corr(silver,copper)",
     xlim=c(-.2,.3))
axis(1)
dev.off()
# Summary statistics of the dsn. of correlation of copper with
# antimony
c(mean(sample.rho[,1]),sqrt(var(sample.rho[,1])))
quantile(sample.rho[,1],probs=c(0.025,0.05,0.5,0.95,0.975))
# arsenic
c(mean(sample.rho[,3]),sqrt(var(sample.rho[,3])))
quantile(sample.rho[,3],probs=c(0.025,0.05,0.5,0.95,0.975))
# bismuth
c(mean(sample.rho[,4]),sqrt(var(sample.rho[,4])))
quantile(sample.rho[,4],probs=c(0.025,0.05,0.5,0.95,0.975))
# silver
c(mean(sample.rho[,5]),sqrt(var(sample.rho[,5])))
quantile(sample.rho[,5],probs=c(0.025,0.05,0.5,0.95,0.975))
#
# Means
#
mulabels=c("Antimony","Copper","Arsenic","Bismuth","Silver")
postscript("cascade_means.eps",height=8,width=8)
pairs(sample.mu,labels=mulabels)
dev.off()
```

```
# Summary statistics of the dsn. of mean of
# antimony
c(mean(sample.mu[,1]),sqrt(var(sample.mu[,1])))
quantile(sample.mu[,1],probs=c(0.025,0.05,0.5,0.95,0.975))
# copper
c(mean(sample.mu[,2]),sqrt(var(sample.mu[,2])))
quantile(sample.mu[,2],probs=c(0.025,0.05,0.5,0.95,0.975))
# arsenic
c(mean(sample.mu[,3]),sqrt(var(sample.mu[,3])))
quantile(sample.mu[,3],probs=c(0.025,0.05,0.5,0.95,0.975))
# bismuth
c(mean(sample.mu[,4]),sqrt(var(sample.mu[,4])))
quantile(sample.mu[,4],probs=c(0.025,0.05,0.5,0.95,0.975))
# silver
c(mean(sample.mu[,5]),sqrt(var(sample.mu[,5])))
quantile(sample.mu[,5],probs=c(0.025,0.05,0.5,0.95,0.975))

q()
```

Advanced Computation

- Approximations based on posterior modes
- Simulation from posterior distributions
- Markov chain simulation
- Why do we need advanced computational methods?
- Except for simpler cases, computation is not possible with available methods:
  - Logistic regression with random effects
  - Normal-normal model with unknown sampling variances σ_j^2
  - Poisson-lognormal hierarchical model for counts.

Strategy for computation

- If possible, work in the log-posterior scale.
- Factor posterior distribution: $p(\gamma, \phi|y) = p(\gamma|\phi, y)p(\phi|y)$
  - Reduces to lower-dimensional problem
  - May be able to sample on a grid
  - Helps to identify parameters most influenced by prior
- Re-parametrizing sometimes helps:
  - Create parameters with easier interpretation
  - Permit normal approximation (e.g., log of variance or log of Poisson rate or logit of probability)

Normal approximation to the posterior

- It is often reasonable to approximate a complicated posterior distribution using a normal (or a mixture of normals) approximation.
- Approximation may be a good starting point for more sophisticated methods.
- Computational strategy:
  1. Find joint posterior mode or mode of marginal posterior distributions (better strategy if possible)
  2. Fit normal approximations at the mode (or use mixture of normals if posterior is multimodal)
- Notation:
  - $p(\theta|y)$: joint posterior of interest (target distribution)

  - $q(\theta|y)$: un-normalized density, typically $p(\theta)p(y|\theta)$.
  - $\theta = (\gamma, \phi)$, ϕ typically lower dimensional than γ.
- Practical advice: computations are often easier with log posteriors than with posteriors.

Finding posterior modes

- To find the mode of a density, we maximize the function with respect to the parameter(s). Optimization problem.
- Modes are not interesting per se, but provide the first step for analytically approximating a density.
- Modes are easier to find than means: no integration, and can work with un-normalized density
- Multi-modal posteriors pose a problem: only way to find multiple modes is to run mode-finding algorithms several times, starting from different locations in parameter space
- Whenever possible, shoot for finding the mode of marginal and

conditional posteriors. If $\theta = (\gamma, \phi)$, and

$$p(\theta, y) = p(\gamma|\phi, y)p(\phi|y)$$

then

1. Find mode $\hat{\phi}$
2. Then find $\hat{\gamma}$ by maximizing $p(\gamma|\phi = \hat{\phi}, y)$

- Many algorithms to find modes. Most popular include Newton-Raphson and Expectation-Maximization (EM).

Taylor approximation to $p(\theta|y)$

- Second order expansion of $\log p(\theta|y)$ around the mode $\hat{\theta}$ is

$$\log p(\theta|y) \approx \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})' \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}).$$

- Linear term in expansion vanishes because log posterior has a zero derivative at the mode.
- Considering $\log p(\theta|y)$ approximation as a function of θ, we see that:

1. $\log p(\hat{\theta}|y)$ is constant and

$$\log p(\theta|y) \approx \frac{1}{2}(\theta - \hat{\theta})' \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta})$$

is proportional to a log-normal density

- Then, for large n and $\hat{\theta}$ in interior of parameter space:

$$p(\theta|y) \approx \mathrm{N}(\hat{\theta}, [I(\hat{\theta})]^{-1})$$

with

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y),$$

the observed information matrix.

Normal and normal-mixture approximation

- For one mode, we saw that

$$p_{n-approx}(\theta|y) = \mathsf{N}(\hat{\theta}, V_\theta)$$

with

$$V_\theta = [-L''(\hat{\theta})]^{-1} \quad \text{or } [I(\hat{\theta})]^{-1}$$

- $L''(\hat{\theta})$ is called the *curvature* of the log posterior at the mode.
- Suppose now that $p(\theta|y)$ has K modes.
- Approximation to $p(\theta|y)$ is now a *mixture of normals*:

$$p_{n-approx}(\theta|y) \propto \sum_k \omega_k \; \mathsf{N}(\hat{\theta}_k, V_{\theta_k})$$

- The ω_k reflect the relative mass associated to each mode
- It can be shown that $\omega_k = q(\hat{\theta}_k|y)|V_{\theta_k}|^{1/2}$.
- The mixture-normal approximation is then

$$p_{n-approx}(\theta|y) \propto \sum_k q(\hat{\theta}_k|y) \exp\{-\frac{1}{2}(\theta - \hat{\theta}_k)' V_{\theta_k}^{-1}(\theta - \hat{\theta}_k)\}.$$

- A more robust approximation can be obtained by using a $t-$kernel instead of the normal kernel:

$$p_{t-approx}(\theta|y) \propto \sum_k q(\hat{\theta}_k|y)[\nu + (\theta - \hat{\theta}_k)' V_{\theta_k}^{-1}(\theta - \hat{\theta}_k)]^{-(d+\nu)/2},$$

with d the dimension of θ and ν relatively low. For most problems, $\nu = 4$ works well.

Sampling from a mixture approximation

- To sample from the normal-mixture approximation:
  1. First choose one of the K normal components, using relative probability masses ω_k as multinomial probabilities.
  2. Given a component, draw a value θ from either the normal or the t density.
- Reasons not to use normal approximation:
  - When mode of parameter is near edge of parameter space (e.g., τ in SAT example)
  - When even transformation of parameter makes normal approximation crazy.
  - Can do better than normal approximation using more advanced methods

Simulation from posterior - Rejection sampling

- Idea is to draw values of $p(\theta|y)$, perhaps by making use of an *instrumental* or *auxiliary* distribution $g(\theta|y)$ from which it is easier to sample.
- The *target* density $p(\theta|y)$ need only be known up to the normalizing constant.
- Let $q(\theta|y) = p(\theta)p(y;\theta)$ be the un-normalized posterior distribution so that

$$p(\theta|y) = \frac{q(\theta|y)}{\int p(\theta)p(y;\theta)d\theta}.$$

- Simpler notation:
  1. Target density: $f(x)$

2. Instrumental density: $g(x)$
3. Constant M such that $f(x) \leq Mg(x)$

- The following algorithm produces a variable Y that is distributed according to f:

1. Generate $X \sim g$ and $U \sim U_{[0,1]}$
2. Set $Y = X$ if $U \leq f(X)/Mg(X)$
3. Reject draw otherwise.

- <u>Proof:</u> The distribution function of Y is given by:

$$\begin{aligned} P(Y \leq y) &= P\left(X \leq y | U \leq \frac{f(X)}{Mg(X)}\right) \\ &= \frac{P\left(X \leq y, U \leq \frac{f(X)}{Mg(X)}\right)}{P\left(U \leq \frac{f(X)}{Mg(X)}\right)} \end{aligned}$$

Then

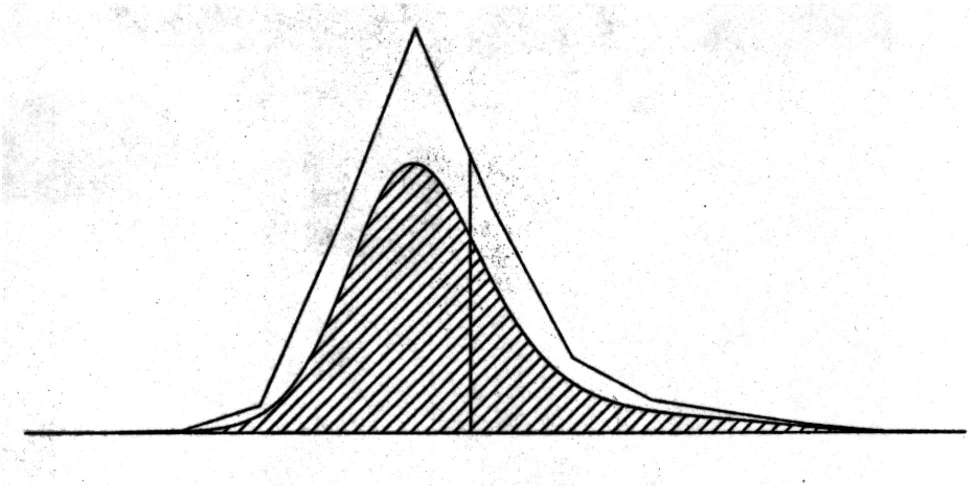
$$\begin{aligned} P(Y \leq y) &= \frac{\int_{-\infty}^{y} \int_{0}^{f(x)/Mg(x)} dug(x)dx}{\int_{-\infty}^{\infty} \int_{0}^{f(x)/Mg(x)} dug(x)dx} \\ &= \frac{M^{-1}\int_{-\infty}^{y} f(x)dx}{M^{-1}\int_{-\infty}^{\infty} f(x)dx}. \end{aligned}$$

- Since the last expression equals $\int_{-\infty}^{y} f(x)dx$, we have proven the result.
- In the Bayesian context, $q(\theta|y)$ (the un-normalized posterior) plays the role of $f(x)$ above.
- When both f and g are normalized densities:

1. The probability of accepting a draw is $1/M$, and expected number of draws until one is accepted is M.

2. For each f, there will be many instrumental densities $g_1, g_2, ...$. Choose the g that requires smallest bound M
3. M is necessarily larger than 1, and will approach minimum value 1 when g closely imitates f.

- In general, g needs to have thicker tails than f for f/g to remain bounded for all x.
- Cannot use a normal g to generate values from a Cauchy f.
- Can do the opposite, however.
- Rejection sampling can be used within other algorithms, such as the Gibbs sampler (see later).



Rejection sampling - Getting M

- Finding the bound M can be a problem, but consider the following implementation approach recalling that

$$q(\theta|y) = p(\theta)p(y;\theta).$$

  - Let $g(\theta) = p(\theta)$ and draw a value θ^* from the prior.
  - Draw $U \sim U(0,1)$.
  - Let $M = p(y;\hat{\theta})$, where $\hat{\theta}$ is the mode of $p(y;\theta)$.
  - Accept the draw θ^* if

$$u \leq \frac{q(\theta|y)}{Mp(\theta)} = \frac{p(\theta)p(y;\theta)}{p(y;\hat{\theta})p(\theta)} = \frac{p(y;\theta)}{p(y;\hat{\theta})}.$$

- Those θ from the prior that are likely under the likelihood are kept in the posterior sample.

Importance sampling

- Also known as SIR (Sampling Importance Resampling), the method is no more than a weighted bootstrap.

- Suppose we have a sample of draws $(\theta_1, \theta_2, ..., \theta_n)$ from the proposal distribution $g(\theta)$. Can we 'convert it' to a sample from $q(\theta|y)$?

- For each θ_i, compute

$$\begin{aligned} \phi_i &= q(\theta_i|y)/g(\theta_i) \\ w_i &= \phi_i / \sum_j \phi_j. \end{aligned}$$

- Draw θ^* from the discrete distribution over $\{\theta_1, ..., \theta_n\}$ with weight w_i on θ_i, without replacement.

- The sample of *re-sampled* θ's is approximately a sample from $q(\theta|y)$.

- Proof: suppose that θ is univariate (for convenience). Then:

$$\begin{aligned} \Pr(\theta^* \leq a) &= \sum_{i=1}^{n} w_i 1_{-\infty,a}(\theta_i) \\ &= \frac{n^{-1}\sum_i \phi_i 1_{-\infty,a}(\theta_i)}{n^{-1}\sum_i \phi_i} \\ &\rightarrow \frac{E_g \frac{q(\theta|y)}{g(\theta)} 1_{-\infty,a}(\theta_i)}{E_g \frac{q(\theta|y)}{g(\theta)}} \\ &\rightarrow \frac{\int_{-a}^{\infty} q(\theta|y)d\theta}{\int_{-\infty}^{\infty} q(\theta|y)d\theta} \\ &\rightarrow \int_{-\infty}^{a} p(\theta|y)d\theta. \end{aligned}$$

- The size of the re-sample can be as large as desired.

- The more g resembles q, the smaller the sample size that is needed for the re-sample to approximate well the target distribution $p(\theta|y)$.

- A consistent estimator of the normalizing constant is

$$n^{-1} \sum_i \phi_i.$$

Importance sampling in a different context

- Suppose that we wish to estimate $E[h(\theta)]$ for $\theta \sim p(\theta|y)$ (e.g., a posterior mean).

- If N values θ_i can be drawn from $p(\theta|y)$, can compute Monte Carlo estimate:
$$\hat{E}[h(\theta)] = \frac{1}{N}\sum h(\theta_i).$$

- Sampling from $p(\theta|y)$ may be difficult. But note:

$$E[h(\theta)|y] = \int \frac{h(\theta)p(\theta|y)}{g(\theta)} g(\theta) d\theta.$$

- Generate values from $g(\theta)$, and estimate

$$\hat{E}[h(\theta)] = \frac{1}{N}\sum h(\theta_i)\frac{p(\theta_i|y)}{g(\theta_i)}$$

- The $w(\theta_i) = p(\theta_i|y)/g(\theta_i)$ are the same *importance weights* from earlier.

- Will not work if tails of g are short relative to tails of p.

Importance sampling (cont'd)

- Importance sampling is most often used to improve normal approximation to posterior.

- Example: suppose that you have a normal (or t) approximation to the posterior and use it to draw a sample that you hope approximates a sample from $p(\theta|y)$.

- Importance sampling can be used to improve sample:

1. Obtain large sample of size L: $(\theta^1, ..., \theta^L)$ from the approximation $g(\theta)$.
2. Construct importance weights $w(\theta^l) = q(\theta^l|y)/g(\theta^l)$
3. Sample $k < L$ values from $(\theta^1, ..., \theta^L)$ with probability proportional to the weights, *without replacement.*

- Why without replacement? If weights are approximately constant, can re-sample with replacement. If some weights are large, re-sample would favor values of θ with large weights repeteadly unless we sample without replacement.

- Difficult to determine whether importance sampling draws approximate draws from posterior

- Diagnostic: monitor weights and look for outliers

- Note: draws from $g(\theta)$ can be reused for other $p(\theta|y)$!

- Given draws $(\theta^1, ..., \theta^m)$ from $g(\theta)$, can investigate sensitivity to prior by re-computing importance weights. For posteriors $p^1(\theta|y)$ and $p^2(\theta|y)$, compute (for the same draws of θ)

$$w^j(\theta_i) = \frac{p^j(\theta_i|y)}{g(\theta_i)}, \quad j = 1, 2.$$

Markov chain Monte Carlo

- Methods based on stochastic process theory, useful for approximating *target* posterior distributions.
- Iterative: must decide when convergence has happened
- Quite general, can be implemented where other methods fail
- Can be used even in high dimensional examples
- Three methods: Metropolis-Hastings, Metropolis and Gibbs sampler.
- M and GS special cases of M-H.

Markov chains

- A process X_t in discrete time $t = 0, 1, 2, ..., T$ where

  $$E(X_t|X_0, X_1, ..., X_{t-1}) = E(X_t|X_{t-1})$$

  is called a **Markov chain**.
- A Markov chain is *irreducible* if it is possible to reach all states from any other state:

  $$p^n(j|i) > 0, \quad p^m(i|j) > 0, \quad m, n > 0$$

  where $p^n(j|i)$ denotes probability of getting to state j from state i in n steps.

- A Markov chain is *periodic* with period d if $p^n(i|i) = 0$ unless $n = kd$ for some integer k. If $d = 2$, the chain returns to i in cycles of $2k$ steps.
- If $d = 1$, the chain is **aperiodic**.
- Theorem: Finite-state, irreducible, aperiodic Markov chains have a limiting distribution: $\lim_{n\to\infty} p^n(j|i) = \pi$, with $p^n(j|i)$ the probability that we reach j from i after n steps or transitions.

Properties of Markov chains (cont'd)

- **Ergodicity**: If a MC is irreducible and aperiodic and has stationary distribution π then we have *ergodicity*:

  $$\begin{aligned} \bar{a}_n &= \frac{1}{n}\sum_t a(X_t) \\ &\rightarrow E\{a(X)\} \text{ as } n \rightarrow \infty. \end{aligned}$$

- $\bar{a}_n$ is an ergodic average.
- Also, rate of convergence can be calculated and is geometric.

Numerical standard error

- Sequence $\{X_1, X_2, ..., X_n\}$ is not iid.
- The asymptotic standard error of a_n is approximately

$$\sqrt{\frac{\sigma_a^2}{n}\{1 + 2\sum_i \rho_i(a)\}}$$

where ρ_i is the ith lag autocorrelation in the function $a\{X_t\}$.

- First term σ_a^2/n is the usual sampling variance under iid sampling.
- Second term is a 'penalty' for the fact that sample is not iid and is usually bigger than 1.
- Often, the standard error is computed assuming a finite number of lags.

Markov chain simulation

- **Idea:** Suppose that sampling from $p(\theta|y)$ is hard, but that we can generate (somehow) a Markov chain $\{\theta(t), t \in T\}$ with stationary distribution $p(\theta|y)$.
- Situation is different from the usual stochastic process case:
  - Here we know the stationary distribution.
  - We seek an algorithm to transition from $\theta^{(t)}$ to $\theta^{(t+1)}$) and that will take us to the stationary distribution.
- Idea: start from some initial guess θ^0 and let the chain run for n steps (n large), so that it reaches its stationary distribution.
- After convergence, all additional steps in the chain are draws from the stationary distribution $p(\theta|y)$.

- MCMC methods all based on the same idea; difference is just in how the transitions in the MC are created.
- In MCMC simulation, we generate at least one MC for each parameter in the model. Often, more than one (independent) chain for each parameter

The Gibbs Sampler

- An iterative algorithm that produces Markov chains with joint stationary distribution $p(\theta|y)$ by cycling through all possible conditional posterior distributions.
- Example: suppose that $\theta = (\theta_1, \theta_2, \theta_3)$, and that the target distribution is $p(\theta|y)$. Steps in the Gibbs sampler are:

1. Start with a guess $(\theta_1^0, \theta_2^0, \theta_3^0)$
2. Draw θ_1^1 from $p(\theta_1|\theta_2 = \theta_2^0, \theta_3 = \theta_3^0, y)$
3. Draw θ_2^1 from $p(\theta_2|\theta_1 = \theta_1^1, \theta_3 = \theta_3^0, y)$
4. Draw θ_3^1 from $p(\theta_3|\theta_1 = \theta_1^1, \theta_2 = \theta_2^1, y)$

- Steps above complete **one iteration** of the GS
- Repeat the steps above n times, and after convergence (see later), draws $(\theta^{n+1}, \theta^{n+2}, ...)$ are sample from stationary distribution $p(\theta|y)$.

The Gibbs Sampler (cont'd)

- Baby example: $\theta = (\theta_1, \theta_2)$ are bivariate normal with mean $y = (y_1, y_2)$, variances $\sigma_1^2 = \sigma_2^2 = 1$ and covariance ρ. Then:

$$p(\theta_1|\theta_2, y) \propto N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

and

$$p(\theta_2|\theta_1, y) \propto N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$$

- In this case, would not need GS, just for illustration
- See figure: $y_1 = 0, y_2 = 0$, and $\rho = 0.8$.

Example: Gibbs sampling in the bivariate normal

- Just as illustration, we consider the trivial problem of sampling from the posterior distribution of a bivariate normal mean vector.
- Suppose that we have a single observation vector (y_1, y_2) where

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim \mathsf{N}\left(\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \qquad (1)$$

- With a uniform prior on (θ_1, θ_2), the joint posterior distribution is

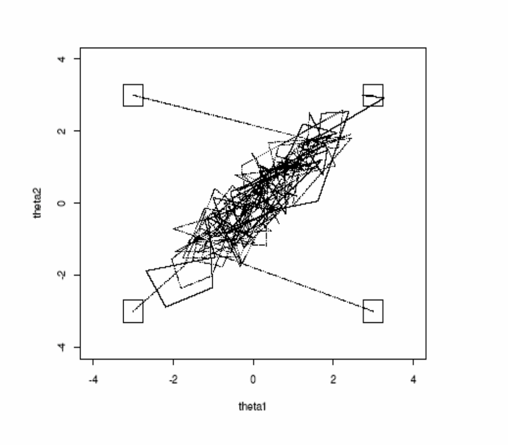
$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} |y \sim \mathsf{N}\left(\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right). \qquad (2)$$

- The conditional distributions are

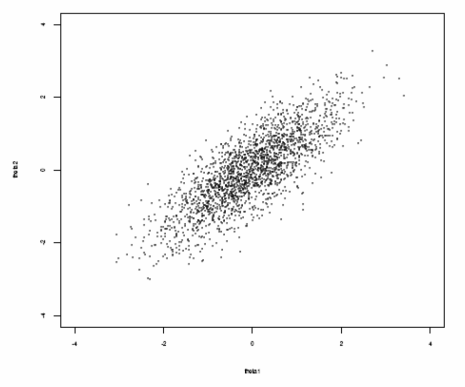
$$\begin{aligned} \theta_1|\theta_2, y &\sim \mathsf{N}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2) \\ \theta_2|\theta_1, y &\sim \mathsf{N}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2) \end{aligned}$$

- We generate four independent chains, starting from four different corners of the 2-dimensional parameter space.

After 50 iterations…..



After 1000 iterations, and eliminating the path lines….



The non-conjugate normal example

- Let $y \sim N(\mu, \sigma^2)$, with (μ, σ^2) unknown.
- Consider the semi-conjugate prior:

$$\begin{array}{rcl} \mu & \sim & N(\mu_0, \tau_0^2) \\ \sigma^2 & \sim & Inv-\chi^2(\nu_0, \sigma_0^2). \end{array}$$

- Joint posterior distribution is

$$p(\mu, \sigma^2) \propto (\sigma^2)^{\frac{n}{2}} e^{(-\frac{1}{2\sigma^2}\sum(y_i-\mu)^2)} e^{(-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2)} (\sigma^2)^{-(\frac{\nu_0}{2}+1))} e^{-(\frac{\nu_0\sigma_0^2}{2\sigma^2})}.$$

Non-conjugate example (cont'd)

- Derive full conditional distributions $p(\mu|\sigma^2, y)$ and $p(\sigma^2|\mu, y)$.
- For μ:

$$\begin{array}{rcl} p(\mu|\sigma^2, y) & \propto & \exp\{-\frac{1}{2}[\frac{1}{\sigma^2}\sum(y_i-\mu)^2 + \frac{1}{\tau_0^2}(\mu-\mu_0)^2]\} \\ & \propto & \exp\{-\frac{1}{2}[(n\tau_0^2+\sigma^2)\mu^2 - 2(n\tau_0^2\bar{y}+\sigma^2\mu_0)\mu]\} \\ & \propto & \exp\{-\frac{1}{2}\frac{n\tau_0^2+\sigma^2}{\sigma^2\tau_0^2}[\mu^2 - 2\left(\frac{n\tau_0^2\bar{y}+\sigma^2\mu_0}{n\tau_0^2+\sigma^2}\right)\mu]\} \\ & \propto & N(\mu_n, \tau_n^2), \end{array}$$

where

$$\mu_n = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \text{ and } \tau_n^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}}.$$

Non-conjugate normal example (cont'd)

- The full conditional for σ^2 is

$$\begin{aligned} p(\sigma^2|\mu, y) &\propto (\sigma^2)^{-(\frac{n+\nu_0}{2}+1)} \exp\{-\frac{1}{2}[\sum(y_i-\mu)^2 + \nu_0\sigma_0^2]\} \\ &\propto Inv-\chi^2(\nu_n, \sigma_n^2), \end{aligned}$$

where

$$\begin{aligned} \nu_n &= \nu_0 + n \\ \sigma_0^2 &= nS^2 + \nu_0\sigma_0^2, \end{aligned}$$

and

$$S = n^{-1}\sum(y_i-\mu)^2.$$

Non-conjugate normal example (cont'd)

- For the non-conjugate normal model, Gibbs sampling consists of following steps:

1. Start with a guess for σ^2, $\sigma^{2(0)}$.
2. Draw $\mu^{(1)}$ from a normal with mean and variance $(\mu_n(\sigma^{2(0)}), \tau_n^2(\sigma^{2(0)}))$, where

$$\mu_n(\sigma^{2(0)}) = \frac{\frac{n}{\sigma^{2(0)}}\bar{y} + \frac{1}{\tau_0^2}\mu_0}{\frac{n}{\sigma^{2(0)}} + \frac{1}{\tau_0^2}},$$

and

$$\tau_n^2(\sigma^{2(0)}) = \frac{1}{\frac{n}{\sigma^{2(0)}} + \frac{1}{\tau_0^2}}.$$

3. Next draw $\sigma^{2(1)}$ from $Inv-\chi^2(\nu_n, \sigma_n^2(\mu^{(1)}))$, where

$$\sigma_n^2(\mu^{(1)}) = nS^2(\mu^{(1)}) + \nu_0\sigma_0^2,$$

and

$$S^2(\mu^{(1)}) = n^{-1}\sum(y_i - \mu^{(1)})^2.$$

4. Repeat many times.

A more interesting example

- Poisson counts, with a change point: consider a sample of size n of counts $(y_1, ..., y_n)$ distributed as Poisson with some rate, and suppose that the rate changes, so that

$$\begin{aligned} \text{For } i = 1, ..., m \quad &, \quad y_i \sim \text{Poi}(\lambda) \\ \text{For } i = m+1, ..., n \quad &, \quad y_i \sim \text{Poi}(\phi) \end{aligned}$$

and m is unknown.

- Priors on (λ, ϕ, m):

$$\begin{aligned} \lambda &\sim \text{Ga}(\alpha, \beta) \\ \phi &\sim \text{Ga}(\gamma, \delta) \\ m &\sim U_{[1,n]} \end{aligned}$$

- Joint posterior distribution:

$$\begin{aligned} p(\lambda,\phi,m|y) &\propto \prod_{i=1}^{m} e^{-\lambda}\lambda^{y_i} \prod_{i=m+1}^{n} e^{-\phi}\phi^{y_i}\lambda^{\alpha-1}e^{-\lambda\beta}\phi^{\gamma-1}e^{-\phi\delta}n^{-1} \\ &\propto \lambda^{y_1^*+\alpha-1}e^{-\lambda(m+\beta)}\phi^{y_2^*+\gamma-1}e^{-\phi(n-m+\delta)} \end{aligned}$$

  with $y_1^* = \sum_{i=1}^{m} y_i$ and $y_2^* = \sum_{i=m+1}^{n} y_i$.

- Note: if we knew m, problem would be trivial to solve. Thus Gibbs, where we condition on all other parameters to determine Markov chains, appears to be the right approach.

- To implement Gibbs sampling, need full conditional distributions. Pick and choose pieces that depend on each parameter

- Conditional of λ:

$$\begin{aligned} p(\lambda|m,\phi,y) &\propto \lambda^{\sum_{i=1}^{m} y_i+\alpha-1}\exp(-\lambda(m+\beta)) \\ &\propto \text{Ga}(\alpha+\sum_{i=1}^{m} y_i, m+\beta) \end{aligned}$$

- Conditional of ϕ:

$$\begin{aligned} p(\phi|\lambda,m,y) &\propto \phi^{\sum_{i=m+1}^{n} y_i+\gamma-1}\exp(-\phi(n-m+\delta)) \\ &\propto \text{Ga}(\gamma+\sum_{i=m+1}^{n} y_i, n-m+\delta) \end{aligned}$$

- Conditional of $m = 1, 2, ..., n$:

$$p(m|\lambda,\phi,y) = c^{-1}q(m|\lambda,\phi)$$

$$= c^{-1}\lambda^{y_1^*+\alpha-1}\exp(-\lambda(m+\beta))\phi^{y_2^*+\gamma-1}\exp(-\phi(n-m+\delta)).$$

- Note that all terms in joint posterior depend on m, so conditional of m is proportional to joint posterior.

- Distribution does not look like any standard form so cannot sample directly. Need to obtain normalizing constant, easy to do for relatively small n and for this discrete case:

$$c = \sum_{k=1}^{n} q(k|\lambda,\phi,y).$$

- To sample from $p(m|\lambda,\phi,y)$ can use inverse cdf on a grid, or other methods.

Non-standard distributions

- It may happen that one or more of the full conditionals is not a standard distribution

- What to do then?
  - Try direct simulation: grid approximation, rejection sampling
  - Try approximation: normal or t approximation, need mode at each iteration (see later)
  - Try more general Markov chain algorithms: Metropolis or Metropolis-Hastings.

Metropolis-Hastings algorithm

- More flexible transition kernel: rather than requiring sampling from conditional distributions, M-H permits using many other "proposal" densities
- Idea: instead of drawing sequentially from conditionals as in Gibbs, M-H "jumps" around the parameter space
- The algorithm is the following:

1. Given a draw θ_t in iteration t, sample a *candidate* draw θ^* from a *proposal distribution* $J(\theta^*|\theta)$
2. Accept the draw with probability

$$r = \frac{p(\theta^*|y)/J(\theta^*|\theta)}{p(\theta|y)/J(\theta|\theta^*)}.$$

3. Stay in place (do not accept the draw) with probability $1 - r$, i.e., $\theta^{(t+1)} = \theta^{(t)}$.

- Remarkably, the proposal distribution (text calls it *jump distribution*) can have just about any form.
- When proposal distribution is *symmetric*, i.e.

$$J(\theta^*|\theta) = J(\theta|\theta^*),$$

Metropolis-Hastings acceptance probability is

$$r = \frac{p(\theta^*|y)}{p(\theta|y)}.$$

This is the **Metropolis** algorithm

Proposal distributions

- Convergence does not depend on J, but *rate of convergence* does.
- Optimal J is $p(\theta|y)$ in which case $r = 1$.
- Else, how do we choose J?

1. It is easy to get samples from J
2. It is easy to compute r
3. It leads to rapid convergence and *mixing*: jumps should be large enough to take us everywhere in the parameter space but not too large so that draw is accepted (see figure from Gilks et al. (1995)).

- Three main approaches: random walk M-H (most popular), independence sampler, and approximation M-H

Independence sampler

- Proposal distribution J_t does not depend on θ_t.
- Just find a distribution $g(\theta)$ and generate values from it
- Can work very well if $g(.)$ is a good approximation to $p(\theta|y)$ and g has heavier tails than p.
- Can work awfully bad otherwise
- Acceptance probability is

$$r = \frac{p(\theta^*|y)/g(\theta^*)}{p(\theta|y)/g(\theta)}$$

- With large samples (central limit theorem operating) proposal could be normal distribution centered at mode of $p(\theta|y)$ and with variance larger than inverse of Fisher information evaluated at mode.

Random walk M-H

- Most popular, easy to use.
- Idea: generate candidate using random walk.
- Proposal is often normal centered at current draw:

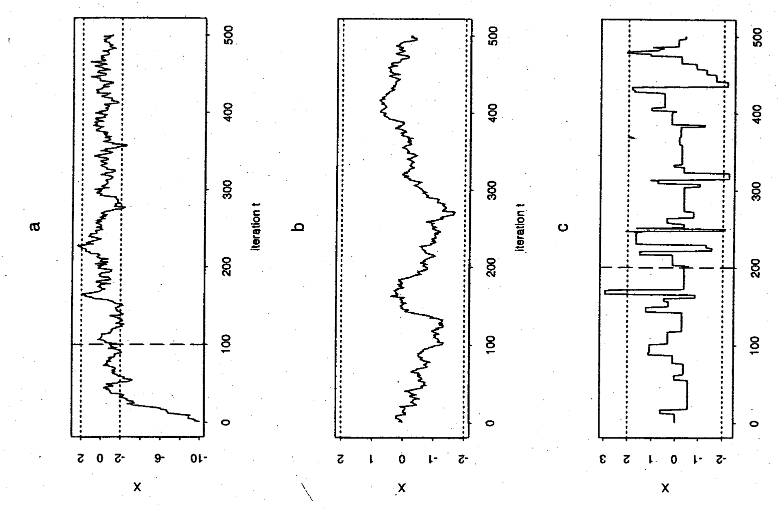
$$J(\theta|\theta^{(t-1)}) = N(\theta|\theta^{(t-1)}, V)$$

- Symmetric: think of drawing values $\theta^* - \theta^{(t-1)}$ from a $N(0, V)$. Thus, r simplifies.
- Difficult to choose V:

1. V too small: takes long to explore parameter space

2. V too large: jumps to extremes are less likely to be accepted. Stay in the same place too long
3. Ideal V: posterior variance. Unknown, so might do some trial and error runs

- Optimal acceptance rate (from some theory results) are between 25% and 50% for this type of proposal distribution. Gets lower with higher dimensional problem.



Approximation M-H

- Idea is to improve approximation of J to $p(\theta|y)$ as we know more about θ.
- E.g., in random walk M-H, can perhaps increase acceptance rate by considering
$$J(\theta|\theta^{(t-1)}) = N(\theta|\theta^{(t-1)}, V_{\theta^{(t-1)}})$$
variance also depends on current draw
- Proposals here typically not symmetric, requires full r expression.

Starting values

- If chain is irreducible, choice of θ^0 will not affect convergence.
- With multiple chains (see later), can choose over-dispersed starting values for each chain. Possible algorithm:
  1. Find posterior modes
  2. Create over-dispersed approximation to posterior at mode (e.g., t_4)
  3. Sample values from that distribution
- Not much research on this topic.

How many chains?

- Even after convergence, draws from stationary distribution are correlated. Sample is not $i.i.d.$
- An iid sample of size n can be obtained by keeping only last draw from each of n independent chains. Too inefficient.
- Compromise: If autocorrelation at lag k and larger is negligible, then can generate an almost $i.i.d.$ sample by keeping only every kth draw in the chain after convergence. Need a very long chain.
- To check convergence (see later) multiple chains can be generated independently (in parallel).
- Burn-in: iterations required to reach stationary distribution (approximately). Not used for inference.

Convergence

- Impossible to decide whether chain has converged, can only monitor behavior.
- Easiest approach: graphical displays (trace plots in WinBUGS).
- A bit more formal (and most popular in terms of use): $\sqrt{(\hat{R})}$ of Gelman and Rubin (1992).
- To use the G-R diagnostic, must generate multiple independent chains for each parameter.
- The G-R diagnostic compares the within-chain sd to the between-chain sd.

- Before convergence, the within-chain sd is an underestimate of the $sd(\theta|y)$, and between-chain sd is overestimate.

- If chains converge, all draws are from stationary distribution, so within and between chain variances should be similar.

- Consider scalar parameter η and let η_{ij} be jth draw in ith chain, with $i = 1, ..., n$ and $j = 1, ..., J$

- Between-chain variance:

$$B = \frac{n}{J-1} \sum (\bar{\eta}_j - \bar{\eta}_{..})^2, \quad \bar{\eta}_j = \frac{1}{n} \sum \eta_{ij}, \quad \bar{\eta}_{..} = \frac{1}{J} \bar{\eta}_j$$

- Within-chain variance:

$$W = \frac{1}{J(n-1)} \sum (\eta_{ij} - \bar{\eta}_j)^2.$$

- An $unbiased$ estimate of $var(\eta|y)$ is weighted average:

$$v\hat{a}r(\eta|y) = \frac{n-1}{n} W + \frac{1}{n} B$$

- Early in iterations, $v\hat{a}r(\eta|y)$ overestimates true posterior variance

- Diagnostic measures potential reduction in the scale if iterations are continued:

$$\sqrt{(\hat{R})} = \sqrt{(\frac{v\hat{a}r(\eta|y)}{W})}$$

which goes to 1 as $n \to \infty$.

Hierarchical models - Introduction

- Consider the following study:
  - J counties are sampled from the population of 99 counties in the state of Iowa.
  - Within each county, n_j farms are selected to participate in a fertilizer trial
  - Outcome y_{ij} corresponds to ith farm in jth county, and is modeled as $p(y_{ij}|\theta_j)$.
  - We are interested in estimating the θ_j.
- Two obvious approaches:
  1. Conduct J separate analyses and obtain $p(\theta_j|y_j)$: all θ's are independent
  2. Get one posterior $p(\theta|y_1, y_2, ..., y_J)$: point estimate is same for all θ's.

- Neither approach appealing.
- Hierarchical approach: view the θ's as a sample from a common *population distribution* indexed by a parameter ϕ.
- Observed data y_{ij} can be used to draw inferences about θ's even though the θ_j are not observed.
- A simple hierarchical model:

$$p(y, \theta, \phi) = p(y|\theta)p(\theta|\phi)p(\phi).$$

$$\begin{aligned} p(y|\theta) &= \text{Usual sampling distribution} \\ p(\theta|\phi) &= \text{Population dist. for } \theta \text{ or prior} \\ p(\phi) &= \text{Hyperprior} \end{aligned}$$

Hierarchical models - Rat tumor example

- Imagine a single toxicity experiment performed on rats.
- θ is probability that a rat receiving no treatment develops a tumor.
- Data: $n = 14$, and $y = 4$ develop a tumor.
- From the sample: tumor rate is 4/14 = 0.286.
- Simple analysis:

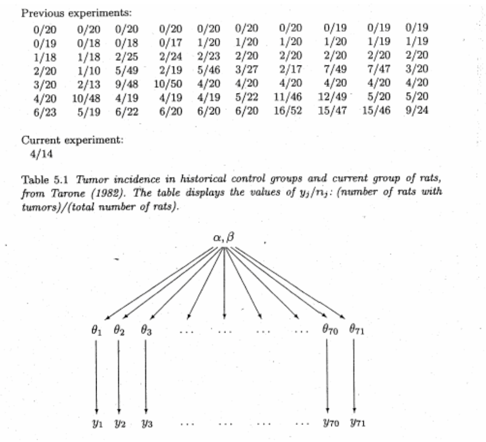
$$\begin{aligned} y|\theta &\sim \text{Bin}(n, \theta) \\ \theta|\alpha, \beta &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

- Posterior for θ is

$$p(\theta|y) = \text{Beta}(\alpha + 4, \beta + 10)$$

- Where can we get good "guesses" for α and β?
- One possibility: from the literature, look at data from similar experiments. In this example, information from 70 other experiments is available.
- In jth study, y_j is the number of rats with tumors and n_j is the sample size, $j = 1, .., 70$.
- See Table 5.1 and Figure 5.1 in Gelman et al..
- Model the y_j as independent binomial data given the study-specific θ_j and sample size n_j

From Gelman et al.: hierarchical representation of tumor experiments



Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of y_j/n_j: (number of rats with tumors)/(total number of rats).*

α, β
θ_1 θ_2 θ_3 … … … … θ_{70} θ_{71}
y_1 y_2 y_3 … … … … y_{70} y_{71}

- Representation of hierarchical model in Figure 5.1 assumes that the $Beta(\alpha, \beta)$ is a good population distribution for the θ_j.

- Empirical Bayes as a first step to estimating mean tumor rate in experiment number 71:

  1. Get point estimates of α, β from earlier 70 experiments as follows:
  2. Mean tumor rate is $\bar{r} = (70)^{-1} \sum_{j=1}^{70} y_j/n_j$ and standard deviation is $[(69)^{-1} \sum_j (y_j/n_j - \bar{r})^2]^{1/2}$. Values are 0.136 and 0.103 respectively.

- Using method of moments:

$$\frac{\alpha}{\alpha+\beta} = 0.136, \quad \longrightarrow \quad \alpha + \beta = \alpha/0.136$$

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = 0.103^2$$

- Resulting estimate for (α, β) is (1.4, 8.6).

- Then posterior is

$$p(\theta|y) = \text{Beta}(5.4, 18.6)$$

  with posterior mean 0.223, lower than the sample mean, and posterior standard deviation 0.083.

- Posterior point estimate is lower than crude estimate; indicates that in current experiment, number of tumors was unusually high.

- Why not go back now and use the prior to obtain better estimates for tumor rates in earlier 70 experiments?

- Should not do that because:

  1. Can't use the data twice: we use the historical data to get the prior, and cannot now combine that prior with the data from the experiments for inference
  2. Using a point estimate for (α, β) suggests that there is no uncertainty about (α, β): not true

3. If the prior $Beta(\alpha, \beta)$ is the appropriate prior, shouldn't we know the values of the parameters prior to observing any data?

- Approach: place a hyperprior on the tumor rates θ_j with parameters (α, β).

- Can still use all the data to estimate the hyperparameters.

- Idea: Bayesian analysis on the joint distribution of all parameters $(\theta_1, ..., \theta_{71}, \alpha, \beta|y)$

Hierarchical models - Exchangeability

- J experiments. In experiment j, $y_j \sim p(y_j|\theta_j)$.

- To create joint probability model, use idea of *exchangeability*

- Recall: a set of random variables $(z_1, ..., z_K)$ are exchangeable if their joint distribution $p(z_1, ..., z_K)$ is invariant to permutations of their labels.

- In our set-up, we have two opportunities for assuming exchangeability:

1. At the level of data: conditional on θ_j, the y_{ij}'s are exchangeable if we cannot "distinguish" between them
2. At the level of the parameters: unless something (other than the data) distinguishes the θ's, we assume that they are exchangeable as well.

- In rat tumor example, we have no information to distinguish between the 71 experiments, except sample size. Since n_j is presumably not related to θ_j, we choose an exchangeable model for the θ_j.

- Simplest exchangeable distribution for the θ_j:

$$p(\theta|\phi) = \Pi_{j=1}^{J} p(\theta_j|\phi).$$

- The hyperparameter ϕ is typically unknown, so marginal for θ must be obtained by averaging:

$$p(\theta) = \int \Pi_{j=1}^{J} p(\theta_j|\phi)p(\phi)d\phi.$$

- Exchangeable distribution for $(\theta_1, ..., \theta_J)$ is written in the form of a *mixture distribution*

- Mixture model characterizes parameters $(\theta_1, ..., \theta_J)$ as independent draws from a *superpopulation* that is determined by unknown parameters ϕ

- Exchangeability does not hold when we have additional information on covariates x_j to distinguish between the θ_j.

- Can still model exchangeability with covariates:

$$p(\theta_1, ..., \theta_J|x_1, ..., x_J) = \int \Pi_{j=1}^{J} p(\theta_j|x_j, \phi)p(\phi|x)d\phi,$$

  with $x = (x_1, ..., x_J)$. In Iowa example, we might know that different counties have very different soil quality.

- Exchangeability does not imply that θ_j are all the same; just that they can be assumed to be draws from some common superpopulation distribution.

Hierarchical models - Bayesian treatment

- Since ϕ is unknown, posterior is now $p(\theta, \phi|y)$.

- Model formulation:

$$\begin{aligned} p(\phi, \theta) &= p(\phi)p(\theta|\phi) = \text{ joint prior} \\ p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\phi, \theta) = p(\phi, \theta)p(y|\theta) \end{aligned}$$

- The hyperparameter ϕ gets its own prior distribution.

- Important to check whether posterior is proper when using improper priors in hierarchical models!

Posterior predictive distributions

- There are two posterior predictive distributions potentially of interest:

1. Distribution of future observations $\tilde{y}$ corresponding to an existing θ_j. Draw $\tilde{y}$ from the posterior predictive given existing draws for θ_j.
2. Distribution of new observations $\tilde{y}$ corresponding to new $\tilde{\theta}_j$'s drawn from the same superpopulation. First draw ϕ from its posterior, then draw $\tilde{\theta}$ for a new experiment, and then draw $\tilde{y}$ from the posterior predictive given the simulated $\tilde{\theta}$.

- In rat tumor example:

1. More rats from experiment #71, for example
2. Experiment #72 and then rats from experiment #72

Hierarchical models - Computation

- Harder than before because we have more parameters
- Easiest when population distribution $p(\theta|\phi)$ is conjugate to the likelihood $p(\theta|y)$.
- In non-conjugate models, must use more advanced computation.
- Usual steps:

1. Write $p(\phi, \theta|y) \propto p(\phi)p(\theta|\phi)p(y|\theta)$
2. Analytically determine $p(\theta|y, \phi)$. Easy for conjugate models.
3. Derive the marginal $p(\phi|y)$ by

$$p(\phi|y) = \int p(\phi, \theta|y)d\theta,$$

or, if convenient, by $p(\phi|y) = \frac{p(\phi,\theta|y)}{p(\theta|y,\phi)}$.

- Normalizing "constant" in denominator may depend on ϕ as well as y.
- Steps for computation in hierarchical models are:

1. Draw vector ϕ from $p(\phi|y)$. If ϕ is low-dimensional, can use inverse cdf method as before. Else, need more advanced methods
2. Draw θ from conditional $p(\theta|\phi, y)$. If the θ_j are conditionally independent, $p(\theta|\phi, y)$ factors as

$$p(\theta|\phi, y) = \Pi_j p(\theta_j|\phi, y)$$

so components of θ can be drawn one at a time.
3. Draw $\tilde{y}$ from appropriate posterior predictive distribution.

Rat tumor example

- Sampling distribution for data from experiments $j = 1, ..., 71$:

$$y_j \sim \text{ Bin}(n_j, \theta_j)$$

- Tumor rates θ_j assumed to be independent draws from Beta:

$$\theta_j \sim \text{ Beta}(\alpha, \beta)$$

- Choose a non-informative prior for (α, β) to indicate prior ignorance.
- Since hyperprior will be non-informative, and perhaps improper, must check integrability of posterior.

- Defer choice of $p(\alpha, \beta)$ until a bit later.

- Joint posterior distribution:

$$\begin{aligned} p(\theta, \alpha, \beta|y) &\propto p(\alpha, \beta)p(\theta|\alpha, \beta)p(y|\theta, \alpha, \beta) \\ \propto p(\alpha, \beta) \quad \Pi_j &\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\Pi_j\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}. \end{aligned}$$

- Conditional of θ: notice that given (α, β), the θ_j are independent, with Beta distributions:

$$p(\theta_j|\alpha, \beta, y) = \frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}(1-\theta_j)^{\beta+n_j-y_j-1}.$$

- Marginal posterior distribution of hyperparameters is obtained using

$$p(\alpha, \beta|y) = \frac{p(\alpha, \beta, \theta|y)}{p(\theta|y, \alpha, \beta)}$$

- Substituting into expression above, we get

$$p(\alpha, \beta|y) \propto p(\alpha, \beta)\Pi_j\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}{\Gamma(\alpha+\beta+n_j)}$$

- Not a standard distribution, but easy to evaluate and only in two dimensions.

- Can evaluate $p(\alpha, \beta|y)$ over a grid of values of (α, β), and then use inverse cdf method to draw α from marginal and β from conditional.

- But what is $p(\alpha, \beta)$? As it turns out, many of the 'obvious' choices for the prior lead to improper posterior. (See solution to Exercise 5.7, on the course web site.)

- Some obvious choices such as $p(\alpha, \beta|y) \propto 1$ lead to non-integrable posterior.

- Integrability can be checked analytically by evaluating the behavior of the posterior as α, β (or functions) go to ∞.

- An empirical assessment can be made by looking at the contour plot of $p(\alpha, \beta|y)$ over a grid. Significant mass extending towards infinity suggests that the posterior will not integrate to a constant.

- Other choices leading to non-integrability of $p(\alpha, \beta|y)$ include:

  - A flat prior on prior guess and "degrees of freedom"

$$p(\frac{\alpha}{\alpha+\beta}, \alpha+\beta) \propto 1.$$

  - A flat prior on the log(mean) and log(degrees of freedom)

$$p(\log(\alpha/\beta), \log(\alpha+\beta)) \propto 1.$$

- A reasonable choice for the prior is a flat distribution on the prior mean and square root of inverse of the degrees of freedom:

$$p(\frac{\alpha}{\alpha+\beta}, (\alpha+\beta)^{-1/2}) \propto 1.$$

- This is equivalent to

$$p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}.$$

- Also equivalent to

$$p(\log(\alpha/\beta), \log(\alpha+\beta) \propto \alpha\beta(\alpha+\beta)^{-5/2}.$$

- We use the parameterization $(\log(\alpha/\beta), \log(\alpha+\beta))$.

- Idea for drawing values of α and β from their posterior is the usual: evaluate $p(\log(\alpha/\beta), \log(\alpha+\beta)|y)$ over grid of values of (α, β), and then use inverse cdf method.

- For this problem, easier to evaluate the log posterior and then exponentiate.

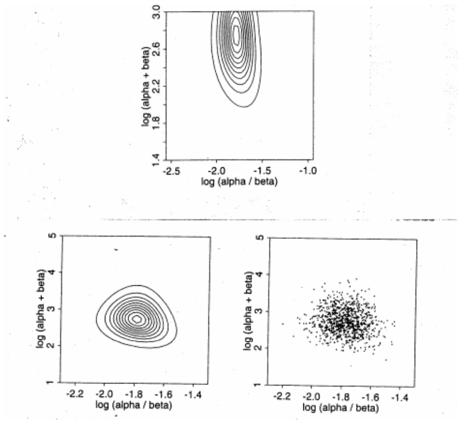
- Grid: From earlier estimates, potential centers for the grid are $\alpha = 1.4, \beta = 8.6$. In the new parameterization, this translates into $u = \log(\alpha/\beta) = -1.8, v = \log(\alpha+\beta) = 2.3$.

- Grid too narrow leaves mass outside. Try $[-2.3, -1.3] \times [1, 5]$.

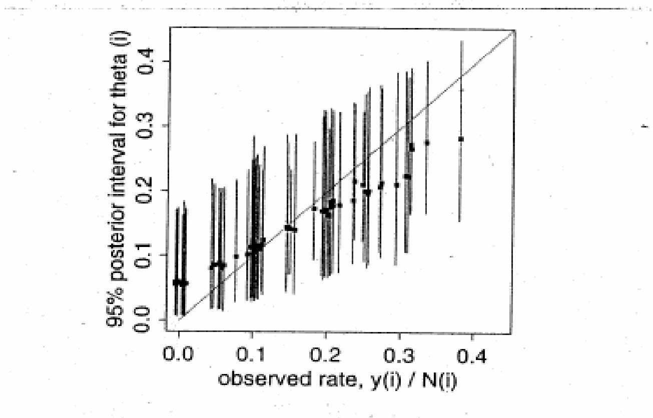
- Steps for computation:

1. Given draws (u, v), transform back to (α, β).
2. Finally, sample θ_j from Beta$(\alpha + y_j, \beta + n_j - y_j)$

From Gelman et al.:
Contours of joint posterior distribution of alpha and beta in reparametrized scale



Posterior means and 95% credible sets for θ_i



Normal hierarchical model

- What we know as one-way random effects models
- Set-up: J independent experiments, in each want to estimate a mean θ_j.
- Sampling model:
$$y_{ij}|\theta_j, \sigma^2 \sim \text{ N}(\theta_j, \sigma^2)$$
for $i = 1, ..., n_j$, and $j = 1, ..., J$.
- Assume for now that σ^2 is known.
- If $\bar{y}_{.j} = n_j^{-1} \sum_i y_{ij}$, then
$$\bar{y}_{.j}|\theta_j \sim \text{ N}(\theta_j, \sigma_j^2)$$

with $\sigma_j^2 = \sigma^2/n_j$.

- Sampling model for $\bar{y}_{.j}$ is quite general. For n_j large, $\bar{y}_{.j}$ is normal even if y_{ij} is not.
- Need now to think of priors for the θ_j.
- What type of posterior estimates for θ_j might be reasonable?
  1. $\hat{\theta}_j = \bar{y}_{.j}$: reasonable if n_j large
  2. $\hat{\theta}_j = \bar{y}_{..} = (\sum_j \sigma^{-2})^{-1} \sum_j \sigma_j^{-2} \bar{y}_{.j}$: pooled estimate, reasonable if we believe that all means are the same.
- To decide between those two choices, do an F-test for differences between groups.
- ANOVA approach (for $n_j = n$):

Source	df	SS	MS	E(MS)
Between groups	J-1	SSB	SSB / (J-1)	$n\tau^2 + \sigma^2$
Within groups	J(n-1)	SSE	SSE / J(n-1)	σ^2

where τ^2, the variance of the group means can be estimated as

$$\hat{\tau}^2 = \frac{MSB - MSE}{n}$$

- If $MSB >>> MSE$ then $\tau^2 > 0$ and F-statistic is significant: do not pool and use $\hat{\theta}_j = \bar{y}_{.j}$.
- Else, F-test cannot reject $H_0 : \tau^2 = 0$, and must pool.
- Alternative: why not a more general estimator:
$$\theta_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j) \bar{y}_{..}$$

for $\lambda_j \in [0, 1]$. Factor $(1 - \lambda_j)$ is a *shrinkage factor*.

- All three estimates have a Bayesian justification:
  1. $\hat{\theta}_j = \bar{y}_{.j}$ is posterior mean of θ_j if sample means are normal and $p(\theta_j) \propto 1$
  2. $\hat{\theta}_j = \bar{y}_{..}$ is posterior mean if $\theta_1 = ... = \theta_J$ and $p(\theta) \propto 1$.
  3. $\theta_j = \lambda_j \bar{y}_{.j} + (1 - \lambda_j) \bar{y}_{..}$ is posterior mean if $\theta_j \sim \text{ N}(\mu, \tau^2)$ independent of other θ's and sampling distribution for $\bar{y}_{.j}$ is normal
- Latter is called the Normal-Normal model.

Normal-normal model

- Set-up (for σ^2 known):

$$\begin{aligned} \bar{y}_{.j}|\theta_j &\sim \mathsf{N}(\theta_j,\sigma_j^2)\\ \theta_1,...,\theta_J|\mu,\tau &\sim \prod_j \mathsf{N}(\mu,\tau^2)\\ p(\mu,\tau^2) & \end{aligned}$$

- It follows that

$$p(\theta_1,...,\theta_J)=\int\prod_j \mathsf{N}(\theta_j;\mu,\tau^2)p(\mu,\tau^2)d\mu d\tau^2$$

- The joint prior can be written as $p(\mu,\tau)=p(\mu|\tau)p(\tau)$. For μ, we will

consider a contidional flat prior, so that

$$p(\mu,\tau)\propto p(\tau)$$

- Joint posterior

$$\begin{aligned} p(\theta,\mu,\tau|y) &\propto p(\mu,\tau)p(\theta|\mu,\tau)p(y|\theta)\\ &\propto p(\mu,\tau)\prod_j \mathsf{N}(\theta_j;\mu,\tau)\prod_j \mathsf{N}(\bar{y}_{.j};\theta_j,\sigma_j^2) \end{aligned}$$

Conditional distribution of group means

- For $p(\mu|\tau)\propto 1$, the conditional distributions of the θ_j given $\mu,\tau,\bar{y}_{.j}$ are independent, and

$$p(\theta_j|\mu,\tau,\bar{y}_{.j}) = \mathsf{N}(\hat{\theta}_j,V_j)$$

with

$$\hat{\theta}_j=\frac{\sigma_j^2\mu+\tau^2\bar{y}_{.j}}{\sigma_j^2+\tau^2},\quad V_j^{-1}=\frac{1}{\sigma_j^2}+\frac{1}{\tau^2}$$

Marginal distribution of hyperparameters

- To get $p(\mu,\tau|y)$ we would typically either integrate $p(\mu,\tau,\theta|y)$ with respect to $\theta_1,...,\theta_J$, or would use the algebraic approach

$$p(\mu,\tau|y)=\frac{p(\mu,\tau,\theta|y)}{p(\theta|\mu,\tau,y)}$$

- In the normal-normal model, data provide information about μ,τ, as shown below.

- Consider the marginal posterior:

$$p(\mu,\tau|y)\propto p(\mu,\tau)p(y|\mu,\tau)$$

where $p(y|\mu,\tau)$ is the *marginal likelihood*:

$$\begin{aligned} p(y|\mu,\tau) &= \int p(y|\theta,\mu,\tau)p(\theta|\mu,\tau)d\theta\\ &= \int\prod_j \mathsf{N}(\bar{y}_{.j};\theta_j,\sigma_j^2)\prod_j \mathsf{N}(\theta_j;\mu,\tau)d\theta_1,...,d\theta_J \end{aligned}$$

- Integrand above is the product of quadratic functions in $\bar{y}_{.j}$ and θ_j, so they are jointly normal.

- Then the $\bar{y}_{.j}|\mu,\tau$ are also normal, with mean and variance:

$$\begin{aligned} E(\bar{y}_{.j}|\mu,\tau) &= E[E(\bar{y}_{.j}|\theta_j,\mu,\tau)|\mu,\tau]\\ &= E(\theta_j|\mu,\tau)=\mu\\ var(\bar{y}_{.j}|\mu,\tau) &= E[var(\bar{y}_{.j}|\theta_j,\mu,\tau)|\mu,\tau] \end{aligned}$$

$$
\begin{aligned}
&\quad +var[E(\bar{y}_{.j}|\theta_j,\mu,\tau)|\mu,\tau] \\
&= E(\sigma_j^2|\mu,\tau) + var(\theta_j|\mu,\tau) \\
&= \sigma_j^2 + \tau^2
\end{aligned}
$$

- Therefore
$$p(\mu,\tau|y) \propto p(\tau)\prod_j \mathsf{N}(\bar{y}_{.j};\mu,\sigma_j^2+\tau^2)$$

- We know that
$$
\begin{aligned}
p(\mu,\tau|y) &\propto p(\tau)p(\mu|\tau) \\
&\times \prod_j (\sigma_j^2+\tau^2)^{-1/2}\exp\left[-\frac{1}{2(\sigma_j^2+\tau^2)}(\bar{y}_{.j}-\mu)^2\right]
\end{aligned}
$$

- Now fix τ, and think of $p(\bar{y}_{.j}|\mu)$ as the "likelihood" in a problem in

which μ is the only parameter.

- Recall that $p(\mu|\tau) \propto 1$.

- From earlier, we know that $p(\mu|y,\tau) = \mathsf{N}(\hat{\mu}, V_\mu)$ where
$$\hat{\mu} = \frac{\sum_j \bar{y}_{.j}/(\sigma_j^2+\tau^2)}{\sum_j 1/(\sigma_j^2+\tau^2)}, \quad V_\mu = \sum_j \frac{1}{\sigma_j^2+\tau^2}$$

- To get $p(\tau|y)$ we use the old trick:
$$
\begin{aligned}
p(\tau|y) &= \frac{p(\mu,\tau|y)}{p(\mu|\tau,y)} \\
&\propto \frac{p(\tau)\prod_j \mathsf{N}(\bar{y}_{.j};\mu,\sigma_j^2+\tau^2)}{\mathsf{N}(\mu;\hat{\mu},V_\mu)}
\end{aligned}
$$

- Expression holds for any μ, so set $\mu = \hat{\mu}$. Denominator is then just $V_\mu^{-1/2}$.

- Then
$$p(\tau|y) \propto p(\tau)V_\mu^{1/2}\prod_j \mathsf{N}(\bar{y}_{.j};\hat{\mu},\sigma_j^2+\tau^2).$$

- What do we use for $p(\tau)$?

- Safe choice is always a proper prior. For example, consider
$$p(\tau) = \mathsf{Inv}-\chi^2(\nu_0,\tau_0^2),$$
where
  - τ_0^2 can be "best guess" for τ
  - ν_0 can be small to reflect prior uncertainty.

- Non-informative (and improper) choice:
  - Beware. Improper prior in hierarchical model can easily lead to non-integrable posterior.
  - In normal-normal model, "natural" non-informative $p(\tau) \propto \tau^{-1}$ or equivalently $p(\log(\tau)) \propto 1$ results in improper posterior.
  - $p(\tau) \propto 1$ leads to proper posterior $p(\tau|y)$.

Normal-normal model - Computation

1. Evaluate the one-dimensional $p(\tau|y)$ on a grid.

2. Sample τ from $p(\tau|y)$ using inverse cdf method.

3. Sample μ from N$(\hat{\mu}, V_\mu)$

4. Sample θ_j from N$(\hat{\theta}_j, V_j)$

Normal-normal model - Prediction

- Predicting future data $\tilde{y}$ from the current experiments with means $\theta = (\theta_1, ..., \theta_J)$:
  1. Obtain draws of $(\tau, \mu, \theta_1, ..., \theta_J)$
  2. Draw $\tilde{y}$ from N(θ_j, σ_j^2)
- Predicting future data $\tilde{y}$ from a future experiment with mean $\tilde{\theta}$ and sample size $\tilde{n}$:
  1. Draw μ, τ from their posterior
  2. Draw $\tilde{\theta}$ from the population distribution $p(\theta|\mu, \tau)$ (also known as the prior for θ)
  3. Draw $\tilde{y}$ from N$(\tilde{\theta}, \tilde{\sigma}^2)$, where $\tilde{\sigma}^2 = \sigma^2/\tilde{n}$

Example: effect of diet on coagulation times

- Normal hierarchical model for coagulation times of 24 animals randomized to four diets. Model:
  - Observations: y_{ij} with $i = 1, ..., n_j$, $j = 1, ..., J$.
  - Given θ, coagulation times y_{ij} are exchangeable
  - Treatment means are normal $N(\mu, \tau^2)$, and variance σ^2 is constant across treatments
  - Prior for $(\mu, \log\sigma, \log\tau) \propto \tau$. A uniform prior on $\log\tau$ leads to improper posterior.

$$
\begin{aligned}
p(\theta, \mu, \log\sigma, \log\tau|y) \quad \propto \quad & \tau \Pi_j \text{ N}(\theta_j|\mu, \tau^2) \\
& \Pi_j \Pi_i \text{ N}(y_{ij}|\theta_j, \sigma^2)
\end{aligned}
$$

- Crude initial estimates: $\hat{\theta}_j = n_j^{-1}\sum y_{ij} = \bar{y}_{.j}$, $\hat{\mu} = J^{-1}\sum \bar{y}_{.j} = \bar{y}_{..}$,

$\hat{\sigma}_j^2 = (n_j - 1)^{-1}\sum(y_{ij} - \bar{y}_{.j})^2$, $\hat{\sigma}^2 = J^{-1}\sum \hat{\sigma}_j^2$, and $\hat{\tau}^2 = (J - 1)^{-1}\sum(\bar{y}_{.j} - \bar{y}_{..})^2$.

Data

The following table contains data that represents coagulation time in seconds for blood drawn from 24 animals randomly allocated to four different diets. Different treatments have different numbers of observations because the randomization was unrestricted.

Diet	Measurements
A	62,60,63,59
B	63,67,71,64,65,66
C	68,66,71,67,68,68
D	56,62,60,61,63,64,63,59

Conditional maximization for joint mode

- Conjugacy makes conditional maximization easy.

- Conditional modes of treatment means are

$$\theta_j | \mu, \sigma, \tau, y \sim \text{N}(\hat{\theta}_j, V_j)$$

  with

$$\hat{\theta}_j = \frac{\mu/\tau^2 + n_j \bar{y}_{.j}/\sigma^2}{1/\tau^2 + n_j/\sigma^2},$$

  and $V_j^{-1} = 1/\tau^2 + n_j/\sigma^2$.

- For $j = 1, ..., J$, maximize conditional posteriors by using $\hat{\theta}_j$ in place of current estimate θ_j.

- Conditional mode of μ:

$$\mu | \theta, \sigma, \tau, y \sim \text{N}(\hat{\mu}, \tau^2/J), \quad \hat{\mu} = J^{-1} \sum \theta_j.$$

- Conditional maximization: replace current estimate μ with $\hat{\mu}$.

Conditional maximization

- Conditional mode of $\log \sigma$: first derive conditional posterior for σ^2:

$$\sigma^2 | \theta, \mu, \tau, y \sim \text{Inv} - \chi^2(n, \hat{\sigma}^2),$$

  with $\hat{\sigma}^2 = n^{-1} \sum \sum (y_{ij} - \theta_j)^2$.

- The mode of the $\text{Inv} - \chi^2$ is $\hat{\sigma}^2 n/(n+2)$. To get mode for $\log \sigma$, use transformation. Term $n/(n+2)$ disappears with Jacobian, so conditional mode of $\log \sigma$ is $\log \hat{\sigma}$.

- Conditional mode of $\log \tau$: same reasoning. Note that

$$\tau^2 | \theta, \mu, \sigma^2, y \sim \text{Inv} - \chi^2(J - 1, \hat{\tau}^2),$$

  with $\hat{\tau}^2 = (J-1)^{-1} \sum (\theta_j - \mu)^2$. After accounting for Jacobian of transformation, conditional mode of $\log \tau$ is $\log \hat{\tau}$.

Conditional maximization

- Starting from crude estimates, conditional maximization required only three iterations to converge approximately (see table)

- Log posterior increased at each step

- Values in final iteration is approximate joint mode.

- When J is large relative to the n_j, joint mode may not provide good summary of posterior. Try to get marginal modes.

- In this problem, factor:

$$p(\theta, \mu, \log \sigma, \log \tau | y) = p(\theta | \mu, \log \sigma, \log \tau, y) p(\mu, \log \sigma, \log \tau | y).$$

- Marginal of $\mu, \log \sigma, \log \tau$ is three-dimensional regardless of J and n_j.

Conditional maximization results

Table shows iterations for conditional maximization.

Parameter	Crude estimate	Stepwise ascent First iteration	 Second iteration	 Third iteration	 Fourth iteration
θ_1	61.000	61.282	61.288	61.290	61.290
θ_2	66.000	65.871	65.869	65.868	65.868
θ_3	68.000	67.742	67.737	67.736	67.736
θ_4	61.000	61.148	61.152	61.152	61.152
μ	64.000	64.010	64.011	64.011	64.011
σ	2.291	2.160	2.160	2.160	2.160
τ	3.559	3.318	3.318	3.312	3.312
$\log p(\text{params.}\vert y)$	-61.604	-61.420	-61.420	-61.420	-61.420

Marginal maximization

- Recall algebraic trick:

$$\begin{aligned} p(\mu, \log\sigma, \log\tau|y) &= \frac{p(\theta, \mu, \log\sigma, \log\tau|y)}{p(\theta|\mu, \log\sigma, \log\tau, y)} \\ &\propto \frac{\tau \Pi_j \ \mathsf{N}(\theta_j|\mu, \tau^2)\Pi_j\Pi_i \ \mathsf{N}(y_{ij}|\theta_j, \sigma^2)}{\Pi_j \ \mathsf{N}(\theta_j|\hat{\theta}_j, V_j)} \end{aligned}$$

- Using $\hat{\theta}$ in place of θ:

$$p(\mu, \log\sigma, \log\tau|y) \propto \tau\Pi_j\mathsf{N}(\theta_j|\mu, \tau^2)\Pi_j\Pi_i\mathsf{N}(y_{ij}|\theta_j, \sigma^2)\Pi_j V_j^{1/2}$$

with $\hat{\theta}$ and V_j as earlier.

- Can maximize $p(\mu, \log\sigma, \log\tau|y)$ using EM. Here, (θ_j) are the "missing data".
- Steps in EM:

1. Average over missing data θ in E-step
2. Maximize over $(\mu, \log\sigma, \log\tau)$ in M-step

EM algorithm for marginal mode

Log joint posterior:

$$\begin{aligned} \log p(\theta, \mu, \log\sigma, \log\tau|y) &\propto -n\log\sigma - (J-1)\log\tau \\ -\frac{1}{2\tau^2}\sum_j(\theta_j-\mu)^2 &- \frac{1}{2\sigma^2}\sum_j\sum_i(y_{ij}-\theta_j)^2 \end{aligned}$$

- **E-step**: Average over θ using conditioning trick (and $p(\theta|$ rest). Need two expectations:

$$\begin{aligned} \mathsf{E}_{old}[(\theta_j-\mu)^2] &= \mathsf{E}[(\theta_j-\mu)^2|\mu^{old}, \sigma^{old}, \tau^{old}, y] \\ &= [\mathsf{E}_{old}(\theta_j-\mu)]^2 + \ \mathsf{var}_{old}(\theta_j) \\ &= (\hat{\theta}_j-\mu)^2 + V_j \end{aligned}$$

Similarly:

$$\mathsf{E}_{old}[(y_{ij}-\theta_j)^2] = (y_{ij}-\hat{\theta}_j)^2 + V_j.$$

EM algorithm for marginal mode (cont'd)

- **M-step:** Maximize the expected log posterior (expectations just taken in E-step) with respect to $(\mu, \log\sigma, \log\tau)$. Differentiate and equate to zero to get maximizing values $(\mu^{new}, \log\sigma^{new}, \log\tau^{new})$. Expressions are:

$$\mu^{new} = \frac{1}{J}\sum_j \hat{\theta}_j.$$

$$\sigma^{new} = \left(\frac{1}{n}\sum_j\sum_i[(y_{ij} - \hat{\theta}_j)^2 + V_j]\right)^{1/2}$$

and

$$\tau^{new} = \left(\frac{1}{J-1}\sum_j[(\hat{\theta}_j - \mu^{new})^2 + V_j]\right)^{1/2}.$$

EM algorithm for marginal mode (cont'd)

- Beginning from joint mode, algorithm converged in three iterations.

- Important to check that log posterior increases at each step. Else, programming or formulation error!

- Results:

Parameter	Value at joint mode	First iteration	Second iteration	Third iteration
μ	64.01	64.01	64.01	64.01
σ	2.17	2.33	2.36	2.36
τ	3.31	3.46	3.47	3.47

Using EM results in simulation

- In a problem with standard form such as normal-normal, can do the following:
  1. Use EM to find marginal and conditional modes
  2. Approximate posterior at the mode using N or t approximation
  3. Draw values of parameters from p_{approx}, and act as if they were from p.

- Importance re-sampling can be used to improve accuracy of draws. For each draw, compute importance weights, and re-sample draws (without replacement) using probability proportional to weight.

- In diet and coagulation example, approximate approach as above likely to produce quite reasonable results.

- But must be careful, specially with scale parameters.

Gibbs sampling in normal-normal case

- The Gibbs sampler can be easily implemented in the normal-normal example because all posterior conditionals are of standard form.

- Refer back to Gibbs sampler discussion, and see derivation of conditionals in hierarchical normal example.

- Full conditionals are:
  - $\theta_j|$all $\sim N$ (J of them)
  - $\mu|$all $\sim N$
  - $\sigma^2|$all $\sim$ Inv $- \chi^2$
  - $\tau^2|$all $\sim$ Inv $- \chi^2$.

- Starting values for the Gibbs sampler can be drawn from, e.g., a t_4 approximation to the marginal and conditional mode.

- Multiple parallel chains for each parameter permit monitoring convergence using potential scale reduction statistic.

Results from Gibbs sampler

- Summary of posterior distributions in coagulation example.
- Posterior quantiles and estimated potential scale reductions computed from the second halves of then Gibbs sampler sequences, each of length 1000.
- Potential scale reductions for τ and σ were computed on the log scale.
- The hierarchical variance τ^2, is estimated less precisely than the unit-level variance, σ^2, as is typical in hierarchical models with a small number of batches.

Posterior quantiles from Gibbs sampler

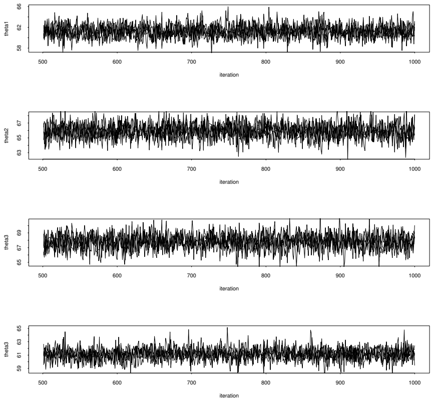
Burn-in $=$ 50% of chain length.

	Posterior quantiles				
Parameter	2.5%	25.0%	50.0%	75.0%	97.5%
$\theta 1$	58.92	60.44	61.23	62.08	63.69
$\theta 2$	63.96	65.26	65.91	66.57	67.94
$\theta 3$	65.72	67.11	67.77	68.44	69.75
$\theta 4$	59.39	60.56	61.14	61.71	62.89
μ	55.64	62.32	63.99	65.69	73.00
σ	1.83	2.17	2.41	2.7	3.47
τ	1.98	3.45	4.97	7.76	24.60
$\log p(\mu, \log\sigma, \log\tau \mid y)$	-70.79	-66.87	-65.36	-64.20	-62.71
$\log p(\theta, \mu, \log\sigma, \log\tau \mid y)$	-71.07	-66.88	-65.25	-64.00	-62.42

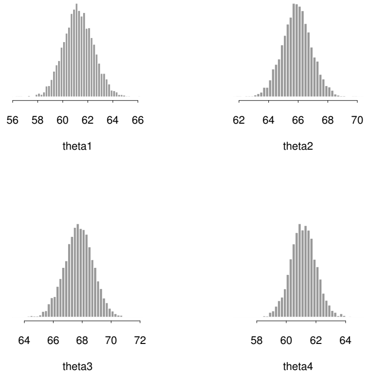
Checking convergence

Parameter	$\hat{R}$	upper
$\theta 1$	1.001	1.003
$\theta 2$	1.001	1.003
$\theta 3$	1.000	1.002
$\theta 4$	1.000	1.001
μ	1.005	1.005
σ	1.000	1.001
τ	1.012	1.013
$\log p(\mu, \log\sigma, \log\tau \mid y)$	1.000	1.000
$\log p(\theta, \mu, \log\sigma, \log\tau \mid y)$	1.000	1.001

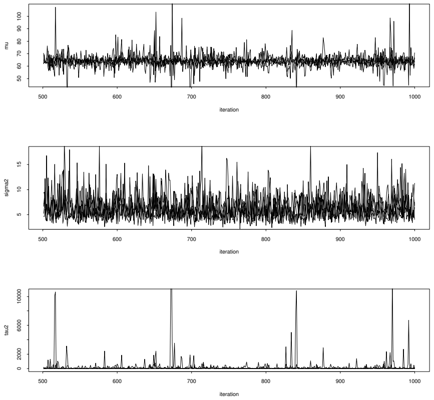
Chains for diet means



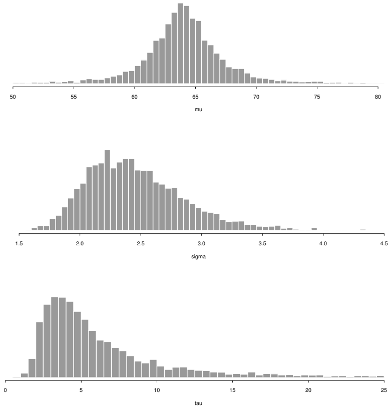
Posteriors for diet means



Chains for μ, σ, τ



Posteriors for μ, σ, τ



Hierarchical modeling for meta-analysis

- Idea: summarize and integrate the results of research studies in a specific area.
- Example 5.6 in Gelman et al.: 22 clinical trials conducted to study the effect of beta-blockers on reducing mortality after cardiac infarction.
- Considering studies separately, no obvious effect of beta-blockers.
- Data are 22 2×2 tables: in jth study, n_{0j} and n_{1j} are numbers of individuals assigned to control and treatment groups, respectively, and y_{0j} and y_{1j} are number of deaths in each group.
- Sampling model for jth experiment: two independent Binomials, with probabilities of death p_{0j} and p_{1j}.

- Possible quantities of interest:
  1. difference $p_{1j} - p_{0j}$
  2. probability ratio p_{1j}/p_{0j}
  3. odds ratio
     $\rho_j = [p_{1j}/(1-p_{1j})]/[p_{0j}/(1-p_{0j})]$
- We parametrize in terms of the log odds ratios: $\theta_j = \log \rho_j$ because the posterior is almost normal even for small samples.

Normal approximation to the likelihood

- Consider estimating θ_j by the empirical logits:

$$y_j = \log\left(\frac{y_{1j}}{n_{1j}-y_{1j}}\right) - \log\left(\frac{y_{0j}}{n_{0j}-y_{0j}}\right)$$

- Approximate sampling variance (e.g., using a Taylor expansion):

$$\sigma_j^2 = \frac{1}{y_{1j}} + \frac{1}{n_{1j}-y_{1j}} + \frac{1}{y_{0j}} + \frac{1}{n_{0j}-y_{0j}}$$

- See Table 5.4 for the estimated log-odds ratios and their estimated standard errors.

	Raw data				Log-	
Study	Control		Treated		odds,	sd,
j	deaths	total	deaths	total	y_j	σ_j
1	3	39	3	38	0.0282	0.8503
2	14	116	7	114	-0.7410	0.4832
3	11	93	5	69	-0.5406	0.5646
4	127	1520	102	1533	-0.2461	0.1382
5	27	365	28	355	0.0695	0.2807
6	6	52	4	59	-0.5842	0.6757
7	152	939	98	945	-0.5124	0.1387
8	48	471	60	632	-0.0786	0.2040
9	37	282	25	278	-0.4242	0.2740
10	188	1921	138	1916	-0.3348	0.1171
11	52	583	64	873	-0.2134	0.1949
12	47	266	45	263	-0.0389	0.2295
13	16	293	9	291	-0.5933	0.4252
14	45	883	57	858	0.2815	0.2054
15	31	147	25	154	-0.3213	0.2977
16	38	213	33	207	-0.1353	0.2609
17	12	122	28	251	0.1406	0.3642
18	6	154	8	151	0.3220	0.5526
19	3	134	6	174	0.4444	0.7166
20	40	218	32	209	-0.2175	0.2598
21	43	364	27	391	-0.5911	0.2572
22	39	674	22	680	-0.6081	0.2724

Goals of analysis

- Goal 1: if studies can be assumed to be exchangeable, we wish to estimate the mean of the distribution of effect sizes, or "overall average effect".
- Goal 2: the average effect size in each of the exchangeable studies
- Goal 3: the effect size that could be expected if a new, exchangeable study, were to be conducted.

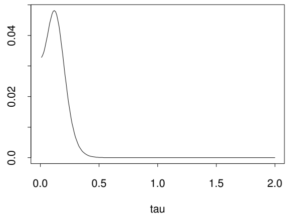
Normal-normal model for meta-analysis

- First level: sampling distribution $y_j|\theta_j, \sigma_j^2 \sim$ N(θ_j, σ_j^2)
- Second level: population distribution $\theta_j|\mu, \tau \sim$ N(μ, τ^2)
- Third level: prior for hyperparameters μ, τ $p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto 1$ mostly for convenience. Can incorporate information if available.

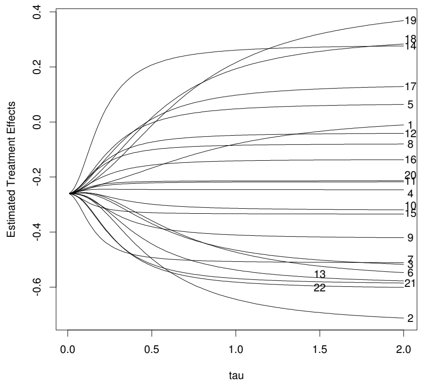
Study	Posterior quantiles of effect θ_j normal approx. (on log-odds scale)				
j	2.5%	25.0%	50.0%	75.0%	97.5%
1	-0.58	-0.32	-0.24	-0.15	0.13
2	-0.63	-0.36	-0.28	-0.20	-0.03
3	-0.64	-0.34	-0.26	-0.18	0.06
4	-0.44	-0.31	-0.25	-0.18	-0.04
5	-0.44	-0.28	-0.21	-0.12	0.13
6	-0.62	-0.36	-0.27	-0.19	0.04
7	-0.62	-0.44	-0.36	-0.27	-0.16
8	-0.43	-0.28	-0.20	-0.13	0.08
9	-0.56	-0.36	-0.27	-0.20	-0.05
10	-0.48	-0.35	-0.29	-0.23	-0.12
11	-0.47	-0.31	-0.24	-0.17	-0.01
12	-0.42	-0.28	-0.21	-0.12	0.09
13	-0.65	-0.37	-0.27	-0.20	0.02
14	-0.34	-0.22	-0.12	0.00	0.30
15	-0.54	-0.32	-0.26	-0.17	0.00
16	-0.50	-0.30	-0.23	-0.15	0.06
17	-0.46	-0.28	-0.21	-0.11	0.15
18	-0.53	-0.30	-0.22	-0.13	0.15
19	-0.51	-0.31	-0.22	-0.13	0.17
20	-0.51	-0.32	-0.24	-0.17	0.04
21	-0.67	-0.40	-0.30	-0.23	-0.09
22	-0.69	-0.40	-0.30	-0.22	-0.07

	Posterior quantiles				
Estimand	2.5%	25.0%	50.0%	75.0%	97.5%
Mean, μ	-0.38	-0.29	-0.25	-0.21	-0.12
Standard deviation, τ	0.01	0.08	0.13	0.18	0.32
Predicted effect, $\tilde{\theta}_j$	-0.57	-0.33	-0.25	-0.17	0.08

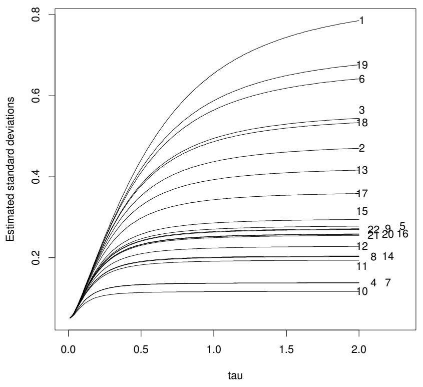
Marginal posterior density $p(\tau|y)$



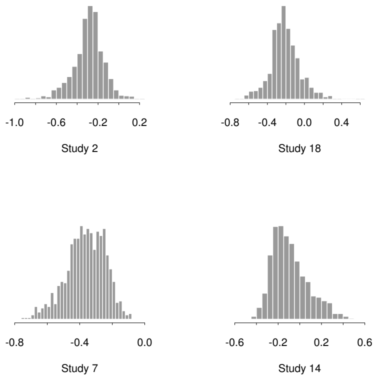
Conditional posterior means $E(\theta_j|\tau, y)$



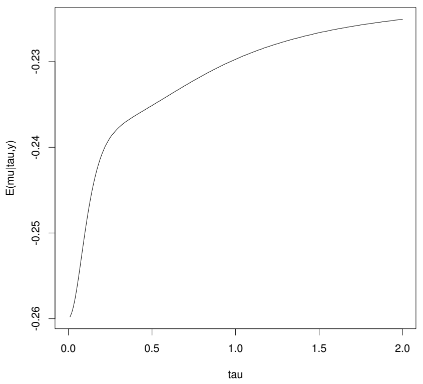
Conditional posterior standard deviations $sd(\theta_j|\tau, y)$



Histogram of 1000 simulations of θ_2, θ_{18}, θ_7, and θ_{14}



Overall mean effect, $E(\mu|\tau, y)$



Example: Mixed model analysis

- Sheffield Food Company produces dairy products.
- Government cited company because the actual content of fat in yogurt sold by Sheffield appers to be higher than the label amount.
- Sheffield believes that the discrepancy is due to the method employed by the goverment to measure fat content, and conducted a multi-laboratory study to investigate.
- Four laboratories in the United States were randomly chosen.
- Each laboratory received 12 carefully mixed samples of yogurt, with instructions to analyze 6 using the government method and 6 using Sheffield's method.

- Fat content in all samples was known to be very close to 3%.
- Because of technical difficulties, none of the labs managed to analyze all six samples using the government method within the alloted time.

Method	Lab 1	Lab 2	Lab 3	Lab 4
Government	5.19	4.09	4.62	3.71
	5.09	3.0	4.32	3.86
		3.75	4.35	3.79
		4.04	4.59	3.63
		4.06		
Sheffield	3.26	3.02	3.08	2.98
	3.38	3.32	2.95	2.89
	3.24	2.83	2.98	2.75
	3.41	2.96	2.74	3.04
	3.35	3.23	3.07	2.88
	3.04	3.07	2.70	3.20

- We fitted a mixed linear model to the data, where
  - Method is a fixed effect with two levels
  - Laboratory is a random effect with four levels
  - Method by laboratory interaction is random with six levels

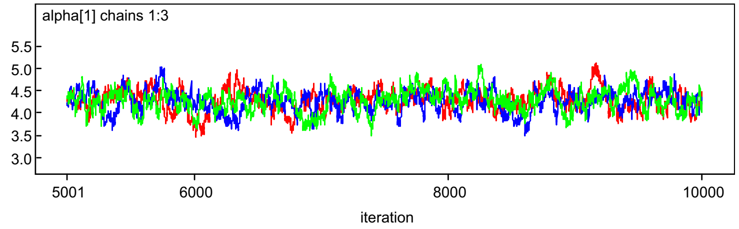
$$y_{ijk} = \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk},$$

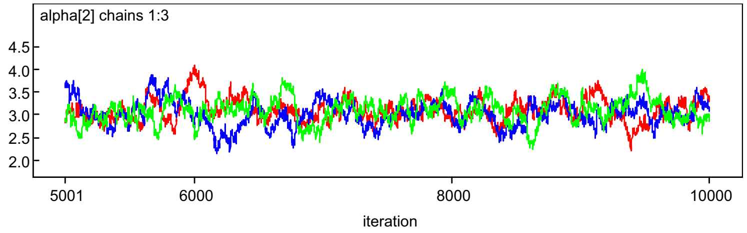
  and $\alpha_i = \mu + \gamma_i$ and $e_{ijk} \sim \mathsf{N}(0, \sigma^2)$.

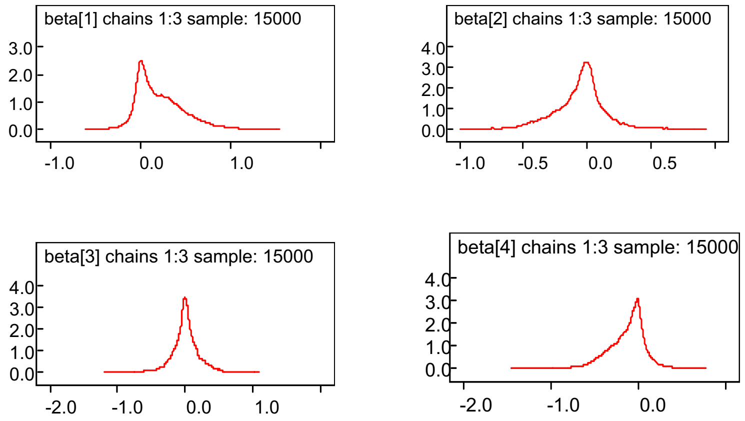
- Priors:

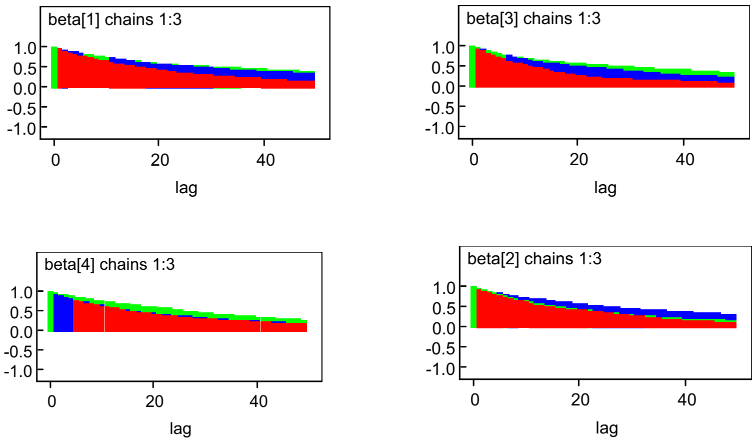
$$\begin{aligned} p(\alpha_i) &\propto 1 \\ p(\beta_j|\sigma^2_\beta|\sigma^2_\beta) &\propto \mathsf{N}(0, \sigma^2_\beta) \\ p((\alpha\beta)_{ij}|\sigma^2_{\alpha\beta}) &\propto \mathsf{N}(0, \sigma^2_{\alpha\beta}) \end{aligned}$$

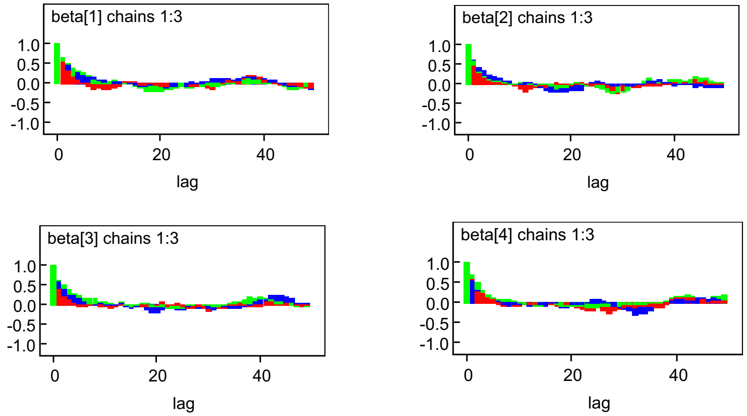
- The three variance components were assigned diffuse inverted gamma priors.

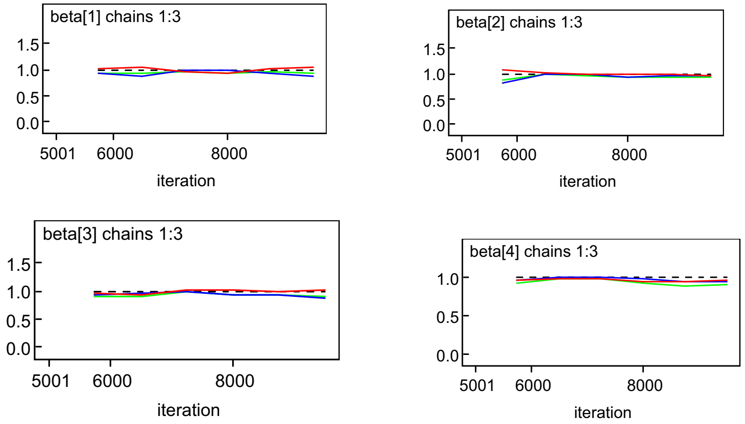


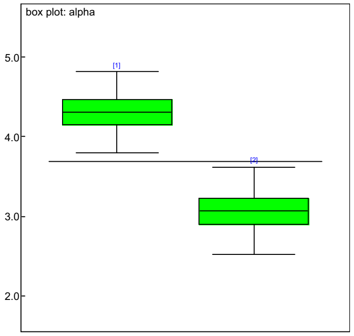


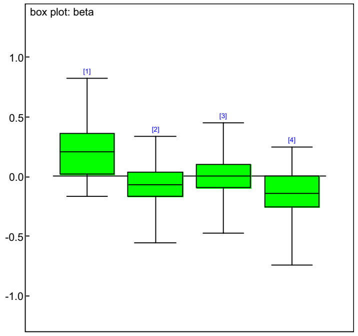

beta[1] chains 1:3 sample: 15000
3.0
2.0
1.0
0.0
-1.0
0.0
1.0
beta[2] chains 1:3 sample: 15000
4.0
3.0
2.0
1.0
0.0
-1.0
-0.5
0.0
0.5
beta[3] chains 1:3 sample: 15000
4.0
3.0
2.0
1.0
0.0
-2.0
-1.0
0.0
1.0
beta[4] chains 1:3 sample: 15000
4.0
3.0
2.0
1.0
0.0
-2.0
-1.0
0.0


beta[1] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[3] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[4] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[2] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta[1] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[2] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[3] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag
beta[4] chains 1:3
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta[1] chains 1:3
1.5
1.0
0.5
0.0
5001
6000
8000
iteration
beta[2] chains 1:3
1.5
1.0
0.5
0.0
5001
6000
8000
iteration
beta[3] chains 1:3
1.5
1.0
0.5
0.0
5001
6000
8000
iteration
beta[4] chains 1:3
1.0
0.5
0.0
5001
6000
8000
iteration


box plot: alpha
5.0
4.0
3.0
2.0
[1]
[2]


box plot: beta
1.0
0.5
0.0
-0.5
-1.0
[1]
[2]
[3]
[4]

Model checking

- What do we need to check?
  - Model fit: does the model fit the data?
  - Sensitivity to prior and other assumptions
  - Model selection: is this the best model?
  - Robustness: do conclusions change if we change data?
- Remember: models are never $true$; they may just fit the data well and allow for useful inference.
- Model checking strategies must address various parts of models:
  - priors
  - sampling distribution
  - hierarchical structure

  - other model characteristics such as covariates, form of dependence between response variable and covariates, etc.
- Classical approaches to model checking:
  - Do parameter estimates make sense
  - Does the model generate data like observed sample
  - Are predictions reasonable
  - Is model "best" in some sense (e.g., AIC, likelihood ratio, etc.)
- Bayesian approach to model checking
  - Does the posterior distribution of parameters correspond with what we know from subject-matter knowledge?
  - Predictive distribution for future data must also be consistent with substantive knowledge
  - Future data generated by predictive distribution compared to current sample

  - Sensitivity to prior and other model components
- Most popular model checking approach has a frequentist flavor: we generate replications of the sample from posterior and observe the behavior of sample summaries over repeated sampling.

Posterior predictive model checking

- Basic idea is that data generated from the model must look like observed data.
- Posterior predictive model checks based on replicated data y^{rep} generated from the posterior predictive distribution:

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta.$$

- y^{rep} is not the same as $\tilde{y}$. Predictive outcomes $\tilde{y}$ can be anything (e.g., a regression prediction using different covariates $\tilde{x}$). But y^{rep} is a replication of y.
- y^{rep} are data that could be observed if we repeated the exact same experiment again tomorrow (if in fact the θs in our analysis gave rise to the data y).

- Definition of a replicate in a hierarchical model:

$$\begin{aligned} p(\phi|y) \rightarrow p(\theta|\phi, y) &\rightarrow p(y^{rep}|\theta): \text{ rep from same units} \\ p(\phi|y) \rightarrow p(\theta|\phi) &\rightarrow p(y^{rep}|\theta): \text{ rep from new units} \end{aligned}$$

- Test quantity: $T(y, \theta)$ is a *discrepancy statistic* used as a standard.
- We use $T(y, \theta)$ to determine discrepancy between model and data on some specific aspect we wish to check.
- Example of $T(y, \theta)$: proportion of standardized residuals outside of $(-3, 3)$ in a regression model to check for outliers
- In classical statistics, use $T(y)$ a test-statistic that depends only on the data. Special case of Bayesian $T(y, \theta)$.
- For model checking:

  - Determine appropriate $T(y, \theta)$
  - Compare posterior predictive distribution of $T(y^{rep}, \theta)$ to posterior distribution of $T(y, \theta)$

Bayes p-values

- p-values attempt to measure tail-area probabilities.
- Classical definition:

$$\text{class } p-\text{value} = \Pr(T(y^{rep}) \geq T(y)|\theta)$$

  - Probability is taken over distribution of y^{rep} with θ fixed
  - Point estimate $\hat{\theta}$ typically used to compute the $p-$value.

- Posterior predictive p-values:

$$\text{Bayes } p-\text{value} = \Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$$

- Probability taken over joint posterior distribution of (θ, y^{rep}):

$$\text{Bayes } p-\text{value} = \int\int I_{T(y^{rep},\theta)\geq T(y,\theta)} p(\theta|y) p(y^{rep}|\theta) d\theta dy^{rep}$$

Relation between p-values

- Small example:
  - Sample $y_1, ..., y_n \sim N(\mu, \sigma^2)$
  - Fit $N(0, \sigma^2)$ and check whether $\mu = 0$ is good fit
- In real life, would fit more general $N(\mu, \sigma^2)$ and would decide whether $\mu = 0$ is plausible.
- Classical approach:
  - Test statistic is sample mean $T(y) = \bar{y}$

$$\begin{aligned} p-\text{value} &= Pr(T(y^{rep}) \geq T(y)|\sigma^2) \\ &= Pr(\bar{y}^{rep} \geq \bar{y}|\sigma^2) \end{aligned}$$

$$\begin{aligned} &= Pr\left(\frac{\sqrt{n}\bar{y}^{rep}}{S} \geq \frac{\sqrt{n}\bar{y}}{S}|\sigma^2\right) \\ &= P\left(t_{n-1} \geq \frac{\sqrt{n}\bar{y}}{S}\right) \end{aligned}$$

- This is special case: not always possible to get rid of nuisance parameters.
- Bayes approach:

$$\begin{aligned} p-\text{value} &= Pr(T(y^{rep}) \geq T(y)|y) \\ &= \int\int I_{T(y^{rep}) \geq T(y)} p(y^{rep}|\sigma^2) p(\sigma^2|y) dy^{rep} d\sigma^2 \end{aligned}$$

- Note that

$$I_{T(y^{rep}) \geq T(y)} p(y^{rep}|\sigma^2) = P(T(y^{rep}) \geq T(y)|\sigma^2)$$

$$= \text{classical p-value}$$

- Then:

  $$\text{Bayes } p-\text{value} = E\{p-\text{value}_{\text{class}}|y\}$$

  where the expectation is taken with respect to $p(\sigma^2|y)$.
- In this example, classical $p-$value and Bayes $p-$value are the same.
- In general, Bayes can handle easily nuisance parameters.

Interpreting posterior predictive $p-$values

- We look for tail-area probabilities that are not too small or too large.
- The ideal posterior predictive p-value is 0.5: the test quantity falls right in the middle of its posterior predictive.
- Posterior predictive p-values are actual posterior probabilities
- Wrong interpretation: Pr(model is true |data).

Example: Independence of Bernoulli trials

- Sequence of binary outcomes $y_1, ..., y_n$ modeled as iid Bernoulli trials with probability of success θ.

- Uniform prior on θ leads to $p(\theta|y) \propto \theta^s(1-\theta)^{n-s}$ with $s = \sum y_i$.

- Sample:

$$1, 1, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0$$

- Is the assumption of independence warranted?

- Consider discrepancy statistic

$$T(y, \theta) = T(y) = \text{ number of switches between 0 and 1.}$$

- In sample, $T(y) = 3$.

- Posterior is Beta(8,14).

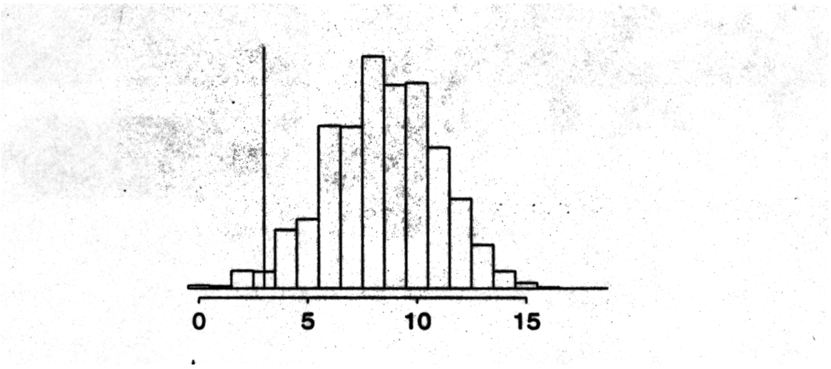
- To test assumption of independence, do:

1. For $j = 1, ..., M$ draw θ^j from Beta(8,14).
2. Draw $\{y_1^{\text{rep},j}, ..., y_{20}^{\text{rep},j}\}$ independent Bernoulli variables with probability θ^j.
3. In each of the M replicate samples, compute $T(y)$, the number of switches between 0 and 1.

$$p_B = \text{ Prob}(T(y^{\text{rep}}) \geq T(y)|y) = 0.98.$$

- If observed trials were independent, expected number of switches in 20 trials is approximately 8 and 3 is very unlikely.

Posterior predictive dist. of number of switches



Example: Newcomb's speed of light

- Newcomb obtained 66 measurements of the speed of light. (Chapter 3). One measurement of -44 is a potential outlier under a normal model.

- From data: $\bar{y} = 26.21$ and $s = 10.75$.

- Model: $y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, $(\mu, \sigma^2) \sim \sigma^{-2}$.

- Posteriors

$$p(\sigma^2|y) = \text{ Inv} - \chi^2(65, s^2)$$
$$p(\mu|\sigma^2, y) = \text{ N}(\bar{y}, \sigma^2/66)$$

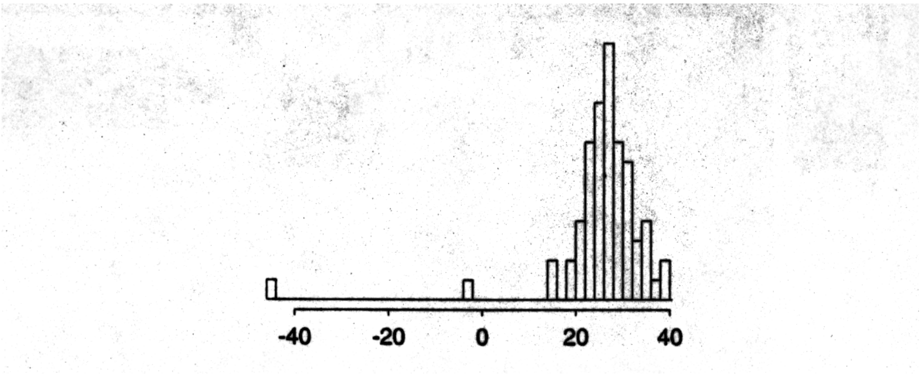
or

$$p(\mu|y) = t_{65}(\bar{y}, s^2/66).$$

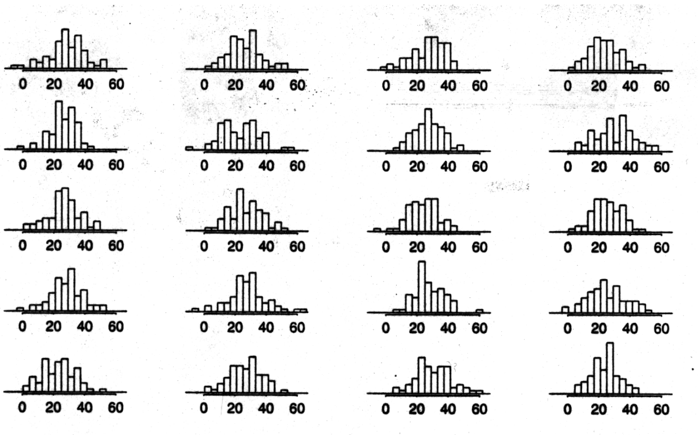
- For posterior predictive checks, do for $i = 1, ..., M$

1. Generate $(\mu^{(i)}, \sigma^{2(i)})$ from $p(\mu, \sigma^2|y)$
2. Generate $y_1^{rep(i)}, ..., y_{66}^{rep(i)}$ from $N(\mu^{(i)}, \sigma^{2(i)})$

Example: Newcomb's data

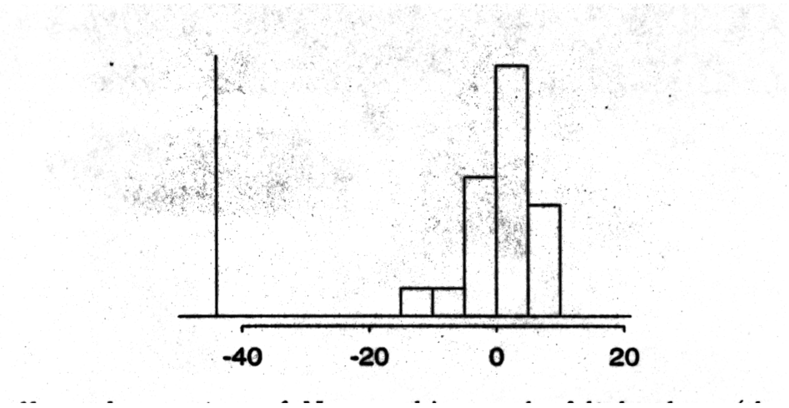


Example: Replicated datasets



- Is observed minimum value -44 consistent with model?
  - Define $T(y) = \min\{y_i\}$
  - get distribution of $T(y^{rep})$.
  - Large negative value of -44 inconsistent with distribution of $T(y^{rep}) = \min\{y_i^{rep}\}$ because observed $T(y)$ very unlikely under distribution of $T(y^{rep})$.
  - Model inadequate, must account for long tail to the left.

Posterior predictive of smallest observation

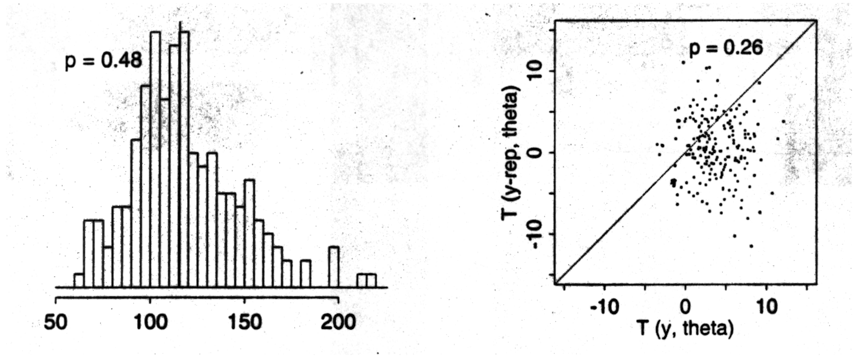


- Is sampling variance captured by model?
  - Define
$$T(y) = s_y^2 = \frac{1}{65}\Sigma_i(y_i - \bar{y})^2$$
  - obtain distribution of $T(y^{rep})$.
  - Observed variance of 115 very consistent with distribution (across reps) of sample variances
  - Probability of observing a larger sample variance is 0.48: observed variance in middle of distribution of $T(y^{rep})$
  - But this is meaningless: $T(y)$ is sufficient statistic for σ^2, so model MUST have fit it well.

- Symmetry in center of distribution
  - Look at difference in the distance of 10th and 90th percentile from the mean.
  - Approx 10th percentile is 6th order statistic $y_{(6)}$, and approx 90th percentile is $y_{(61)}$.

- Define $T(y, \theta) = |y_{(61)} - \theta| - |y_{(6)} - \theta|$
- Need joint distribution of $T(y, \theta)$ and of $T(y^{rep}, \theta)$.
- Model accounts for symmetry in center of distribution
- Joint distribution of $T(y, \theta)$ and $T(y^{rep}, \theta)$ about evenly distributed along line $T(y, \theta) = T(y^{rep}, \theta)$.
- Probability that $T(y^{rep}, \theta) > T(y, \theta)$ about 0.26, plausible under sampling variability.

Posterior predictive of variance and symmetry



Choice of discrepancy measure

- Often, choose more than one measure, to investigate different attributes of the model.
  - Can smoking behavior in adolescents be predicted using information on covariates? (Example on page 172 of Gelman et al.)
  - Data: six observations of smoking habits in 2,000 adolescents every six months
- Two models:
  - Logistic regression model
  - Latent-class model: propensity to smoke modeled as nonlinear function of predictors. Conditional on smoking, fit same logistic regression model as above.
- Three different test statistics chosen:

  - % who never smoked
  - % who always smoked
  - % of “incident smokers”: began not smoking but then switched to smoking and continued doing so.
- Generate replicate datasets from both models.
- Compute the three test statistics in each of the replicated datasets
- Compare the posterior predictive distributions of each of the three statistics with the value of the statistic in the observed dataset
- Results: Model 2 slightly better than 1 at predicting % of always smokers, but both models fail to predict % of incident smokers.

		95% CI		95% CI	
Test variable	$T(y)$	for $T(y^{rep})$	p-value	for $T(y^{rep})$	p-value
% never smokers	77.3	[75.5, 78.2]	0.27	[74.8, 79.9]	0.53
% always-smokers	5.1	[5.0, 6.5]	0.95	[3.8, 6.3]	0.44
% incident smokers	8.4	[5.3, 7.9]	0.005	[4.9, 7.8]	0.004

Omnibus tests

- It is often useful to also consider summary statistics:

  χ^2-discrepancy: $T(y,\theta) = \sum \frac{(y_i - E(y_i|\theta))^2}{var(y_i|\theta)}$

  Deviance: $T(y,\theta) = -2\log p(y|\theta)$
- Deviance is proportional to mean squared error if model is normal and with constant variance.
- In classical approach, insert θ_{null} or θ_{mle} in place of θ in $T(y,\theta)$ and compare test statistic to reference distribution derived under asymptotic arguments.
- In Bayesian approach, reference distribution for test statistic is automatically calculated from posterior predictive simulations.

- A Bayesian χ^2 test is carried out as follows:
  - Compute the distribution of $T(y,\theta)$ for many draws of θ from posterior and for observed data.
  - Compute the distribution of $T(y^{rep},\theta)$ for replicated datasets and posterior draws of θ
  - Compute Bayesian p-value: probability of observing a more extreme value of $T(y^{rep},\theta)$.
  - Calculate p-value empirically from simulations over θ and y^{rep}.

Criticisms of posterior predictive checks

- Too conservative because data get used twice
- Posterior predictive p-values are not really p-values: distribution under null hypothesis is not uniform, as it should be
- What is high/low if pp p-values not uniform?
- Some Bayesians object to frequentist slant of using unobserved data for anything
- Lots of work recently on improving posterior predictive p-values: Bayarri and Berger, $JASA$, 2000, is good reference
- In spite of criticisms, pp checks are easy to use in very general cases, and intuitively appealing.

Pps can be conservative

- One criticism for pp model checks is that they tend to reject models only in the face of extreme evidence.
- Example (from Stern):
  - $y \sim N(\mu, 1)$ and $\mu \sim N(0, 9)$.
  - Observation: $y_{obs} = 10$
  - Posterior $p(\mu|y) = N(0.9y_{obs}, 0.9) = N(9, 0.9)$
  - Posterior predictive dist: $N(9, 1.9)$
  - Posterior predictive p-value is 0.23, we would not reject the model
- Effect of prior is minimized because $9 > 1$ and posterior predictive mean is "close" to observed datapoint.
- To reject, we would need to observe $y \geq 23$.

Graphical posterior predictive checks

- Idea is to display the data together with simulated data.
- Three kinds of checks:
  - Direct data display
  - Display of data summaries or parameter estimates
  - Graphs of residuals or other measures of discrepancy

Model comparison

- Which model fits the data best?
- Often, models are $nested$: a model with parameters θ is nested within a model with parameters (θ, ϕ).
- Comparison involves deciding whether adding ϕ to the model improves its fit. Improvement in fit may not justify additional complexity.
- It is also possible to compare non-nested models.
- We focus on predictive performance and on model posterior probabilities to compare models.

Expected deviance

- We compare the observed data to several models to see which predicts more accurately.
- We summarize model fit using the deviance

$$D(y, \theta) = -2 \log p(y|\theta).$$

- It can be shown that the model with the lowest expected deviance is best in the sense of minimizing the (Kullback-Leibler) distance between the model $p(y|\theta)$ and the true distribution of y, $f(y)$.
- We compute the expected deviance by simulation:

$$D_{\text{avg}}(y) = E_{\theta|y}(D(y, \theta)|y).$$

- An estimate is

$$\hat{D}_{\text{avg}}(y) = \frac{1}{M} \sum_j D(y, \theta^j).$$

Deviance information criterion - DIC

- Idea is to estimate the error that would be expected when applying the model to future data.
- Expected mean square predictive error:

$$D_{\text{avg}}^{\text{pred}} = E\left[\frac{1}{n} \sum_i (y_i^{\text{rep}} - E(y_i^{\text{rep}}|y))^2\right].$$

- A model that minimizes the expected predictive deviance is best in the sense of out-of-sample predictive power.
- An approximation to $D_{\text{avg}}^{\text{pred}}$ is the Deviance Information Criterion or DIC

$$DIC = \hat{D}_{\text{avg}}^{\text{pred}} = 2\hat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y),$$

where $D_{\hat{\theta}}(y) = D(y, \hat{\theta}(y))$ and $\hat{\theta}(y)$ is a point estimator of θ such as the posterior mean.

Example: SAT study

- We compare three models: no pooling, complete pooling and hierarchical model (partial pooling).
- Deviance is

$$\begin{aligned} D(y,\theta) &= -2\sum_j \log\ \mathsf{N}(y_j|\theta_j,\sigma_j^2) \\ &= \log(2\pi J\sigma^2) + \sum_j((y_j-\theta_j)^2/\sigma_j^2) \end{aligned}$$

- The DIC for the three models were: 70.3 (no pooling), 61.5 (complete pooling), 63.4 (hierarchical model).
- Based on DIC, we would pick the complete pooling model.

- We still prefer the hierarchical model because assumption that all schools have identical mean is strong.

Bayes Factors

- Suppose that we wish to decide between two models M_1 and M_2 (different prior, sampling distribution, parameters).
- Priors on models are $p(M_1)$ and $p(M_2) = 1 - p(M_1)$.
- The *posterior odds* favoring M_1 over M_2 are

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)}\frac{p(M_1)}{p(M_2)}.$$

- The ratio $p(y|M_1)/p(y|M_2)$ is called a *Bayes factor*.
- It tells us how much the data support one model over the other.

- The BF_{12} is computed as

$$BF_{12} = \frac{\int p(y|\theta_1, M_1)p(\theta_1|M_1)d\theta_1}{\int p(y|\theta_2, M_2)p(\theta_2|M_2)d\theta_2}.$$

- Note that we need $p(y)$ under each model to compute $p(\theta_i|M_i)$. Therefore, the BF is only defined when the marginal distribution of the data $p(y)$ is proper, and therefore, when the prior distribution in each model is proper.
- For example, if $y \sim N(\theta, 1)$ with $p(\theta) \propto 1$, we get

$$p(y) \propto (2\pi)^{1/2}\int \exp\{-\frac{1}{2}(y-\theta)^2\}d\theta = 1,$$

  which is constant for any y and thus is not a proper distribution. This creates an indeterminacy because we can increase or decrease the BF simply by using a different constant in the numerator and denominator.

Computation of Bayes Factors

- Difficulty lies in the computation of $p(y)$

$$p(y) = \int p(y|\theta)p(\theta)d\theta.$$

- The simplest approach is to draw many values of θ from $p(\theta)$ and get a Monte Carlo approximation to the integral.

- May not work well because the θ's from the prior may not come from the parameter space region where the sampling distribution has most mass.

- A better Monte Carlo approximation is similar to importance sampling.

Note that

$$\begin{aligned} p(y)^{-1} &= \int \frac{h(\theta)}{p(y)} d\theta \\ &= \int \frac{h(\theta)}{p(y|\theta)p(\theta)} p(\theta|y) d\theta. \end{aligned}$$

- $h(\theta)$ can be anything (e.g., a normal approximation to the posterior).

- Draw values of θ from the posterior and evaluate the integral numerically.

- Computations can get tricky because denominator can be small.

- As $n \to \infty$,

$$\log(BF) \approx \log(p(y|\hat{\theta}_2, M_2) - \log(p(y|\hat{\theta}_1, M_1) - \frac{1}{2}(d_1 - d_2)\log(n),$$

where $\hat{\theta}_i$ is the posterior mode under model M_i and d_i is the dimension of θ_i.

- Notice that the criterion penalizes a model for additional parameters.

- Using $\log(BF)$ is equivalent to ranking models using the BIC criterion, given by

$$\text{BIC} = -\log(p(y|\hat{\theta}, M) + \frac{1}{2}d\log(n).$$

Ordinary Linear Regression - Introduction

- Question of interest: how does an outcome y vary as a function of a vector of covariates X.
- We want the conditional distribution $p(y|\theta, x)$, where observations $(y, x)_i$ are assumed exchangeable.
- Covariates $(x_1, x_2, ..., x_k)$ can be discrete or continuous, and typically $x_1 = 1$ for all n units. The matrix X, $n \times k$ is the model matrix
- Most common version of the model is the normal linear model

$$E(y_i|\beta, X) = \sum_{j=1}^{k} \beta_j x_{ij}$$

Ordinary Linear Regression - Model

- In ordinary linear regression models, the conditional variance is equal across observations: $var(y_i|\theta, X) = \sigma^2$. Thus, $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$.
- Likelihood: For the ordinary normal linear regression model, we have

$$p(y|X, \beta, \sigma^2) = N(X\beta, \sigma^2 I)$$

with I_n and $n \times n$ identity matrix.

- Priors: A non-informative prior distribution for (β, σ^2) is

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

(more on informative priors later)

- Joint posterior:

$$p(\beta, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}-1} \exp\left[-\frac{1}{2}\sigma^{-2}(y - X\beta)'(y - X\beta)\right]$$

- Joint posterior

$$p(\beta, \sigma^2|y) \propto (\sigma^2)^{-\frac{n}{2}-1} \exp\left[-\frac{1}{2}\sigma^{-2}(y - X\beta)'(y - X\beta)\right]$$

- Consider

$$p(\beta, \sigma^2|y) = p(\beta|\sigma^2, y)p(\sigma^2|y)$$

- Conditional posterior for β: Expand and complete the squares in

$p(\beta|\sigma^2, y)$ (viewed as function of β). We get:

$$\begin{aligned} p(\beta|\sigma^2, y) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left[\beta' X' X\beta - 2\hat{\beta}' X' X\beta\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}(\beta - \hat{\beta})' X' X(\beta - \hat{\beta})\right\} \end{aligned}$$

for

$$\hat{\beta} = (X'X)^{-1}X'y$$

- Then:

$$\beta|\sigma^2, y \sim \text{N}\ (\hat{\beta}, \sigma^2(X'X)^{-1})$$

Ordinary Linear Regression - $p(\sigma^2|y)$

- The marginal posterior distribution for σ^2 is obtained by integrating the joint posterior with respect to β:

$$\begin{aligned} p(\sigma^2|y) &= \int p(\beta,\sigma^2|y)d\beta \\ &\propto (\sigma^2)^{-\frac{n}{2}-1}\int \exp\left[-\sigma^{-2}\frac{1}{2}(y-X\beta)'(y-X\beta)\right]d\beta \end{aligned}$$

- By expanding square in integrand, and adding and subtracting $2y'X(X'X)^{-1}X'y$, we can write:
arge

$$p(\sigma^2|y) = (\sigma^2)^{-\frac{n}{2}-1}\times$$

$$\int \exp\left\{-\frac{1}{2\sigma^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta})+(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]\right\}d\beta$$

$$\propto (\sigma^2)^{-\frac{n}{2}-1}\exp\left[-\frac{1}{2\sigma^2}(n-k)S^2\right]\int \exp\left[-\frac{1}{2\sigma^2}(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})\right]d\beta$$

- Integrand is kernel of $k-$ dimensional normal, so result of integration is proportional to $(\sigma^2)^{k/2}$. Then

$$p(\sigma^2|y) \propto (\sigma^2)^{-\frac{n-k}{2}+1}\exp\left[-\sigma^{-2}(n-k)S^2\right]$$

proportional to an Inv-$\chi^2(n-k,S^2)$.

Ordinary Linear Regression - Notes

- Note that $\hat{\beta}$ and S^2 are the MLEs of β and σ^2
- When is the posterior proper? $p(\beta,\sigma^2|y)$ is proper if
  1. $n > k$
  2. rank$(X) = k$: columns of X are linearly independent or $|X'X| \neq 0$
- To sample from the joint posterior:
  1. Draw σ^2 from Inv-$\chi^2(n-k,S^2)$
  2. Given σ^2, draw vector β from N$(\hat{\beta},\sigma^2(X'X)^{-1})$
- For efficiency, compute $\hat{\beta}, S^2, (X'X)^{-1}$ once, before starting repeated drawing.

Regression - Posterior predictive distribution

- In regression, we often want to predict the outcome $\tilde{y}$ for a new set of covariates $\tilde{x}$. Thus, we wish to draw values from $p(\tilde{y}|y,\tilde{X})$
- By simulation:
  1. Draw σ^2 from Inv-$\chi^2(n-k,S^2)$
  2. Draw β from N$(\hat{\beta},\sigma^2(X'X)^{-1})$
  3. Draw $\tilde{y}_i$ for $i = 1,...,m$ from N$(\tilde{x}_i'\beta,\sigma^2)$

Regression - Posterior predictive distribution

- We can derive $p(\tilde{y}|y)$ in steps, first considering $p(\tilde{y}|y,\sigma^2)$.
- Note that
$$p(\tilde{y}|y,\sigma^2) = \int p(\tilde{y}|\beta,\sigma^2)p(\beta|\sigma^2,y)d\beta$$
is normal because exponential in integrand is quadratic function in $(\beta,\tilde{y})$.
- To get mean and variance use conditioning trick:
$$\begin{aligned} E(\tilde{y}|\sigma^2,y) &= E\left[E(\tilde{y}|\beta,\sigma^2,y)|\sigma^2,y\right] \\ &= E(\tilde{X}\beta|\sigma^2,y) \\ &= \tilde{X}\hat{\beta} \end{aligned}$$

where inner expectation averages over $\tilde{y}$ conditional on β and outer averages over β.

$$\begin{aligned} var(\tilde{y}|\sigma^2,y) &= E\left[var(\tilde{y}|\beta,\sigma^2,y)|\sigma^2,y\right] + var\left[E(\tilde{y}|\beta,\sigma^2,y)|\sigma^2,y\right] \\ &= E[\sigma^2 I|\sigma^2,y] + var[\tilde{X}\beta|\sigma^2,y] \\ &= \sigma^2(I + \tilde{X}(X'X)^{-1}\tilde{X}') \end{aligned}$$

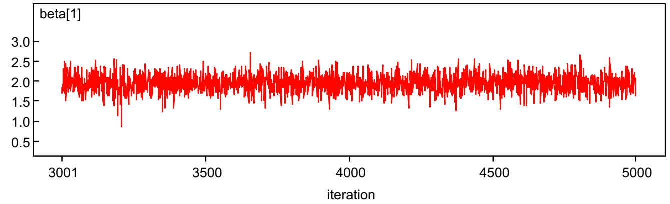
- Var has two terms: $\sigma^2 I$ is sampling variation, and $\sigma^2\tilde{X}(X'X)^{-1}\tilde{X}'$ is due to uncertainty about β
- To complete specification of $p(\tilde{y}|y)$ must integrate $p(\tilde{y}|y,\sigma^2)$ with respect to marginal posterior distribution of σ^2.
- Result is:
$$p(\tilde{y}|y) = t_{n-k}(\tilde{X}\hat{\beta}, S^2[I + \tilde{X}(X'X)^{-1}\tilde{X}])$$

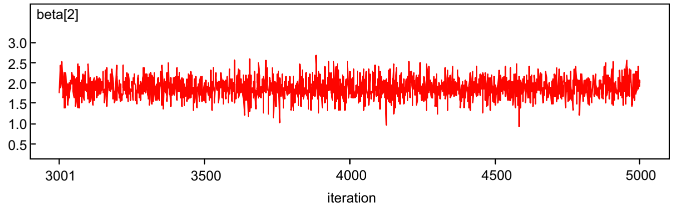
Regression example: radon measurements in Minnesota

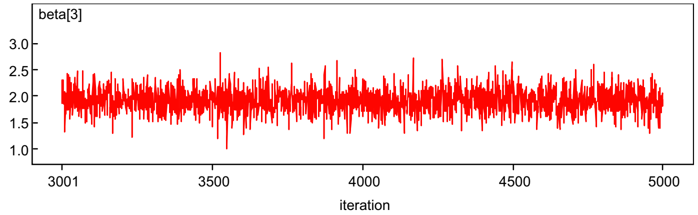
- Radon measurements y_i were taken in three counties in Minnesota: Blue Earth, Clay and Goodhue.
- 14 houses in each of Blue Earth and Clay county and 13 houses in Goodhue were sampled.
- Measurements were taken in the basement and in the first floor.
- We fit an ordinary regression model to the log radon measurements, without an intercept.
- We define dummy variables as follows: $X_1 = 1$ if county is Blue Earth and is 0 otherwise. Similarly, X_2 and X_3 are dummies for Clay and Goodhue counties, respectively.
- $X_4 = 1$ if measurement was taken in the first floor.
- The model is
$$\log(y_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + e_i,$$
with $e \sim \text{N}(0,\sigma^2)$.
- Thus
$$\begin{aligned} E(y|\text{Blue Earth, basement}) &= \exp(\beta_1) \\ E(y|\text{Blue Earth, first floor}) &= \exp(\beta_1+\beta_4) \\ E(y|\text{Clay, basement}) &= \exp(\beta_2) \end{aligned}$$

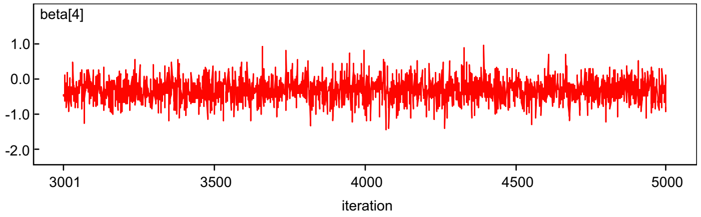
$$\begin{aligned} E(y|\text{Clay, first floor}) &= \exp(\beta_2+\beta_4) \\ E(y|\text{Goodhue, basement}) &= \exp(\beta_3) \\ E(y|\text{Goodhue, first floor}) &= \exp(\beta_3+\beta_4) \end{aligned}$$

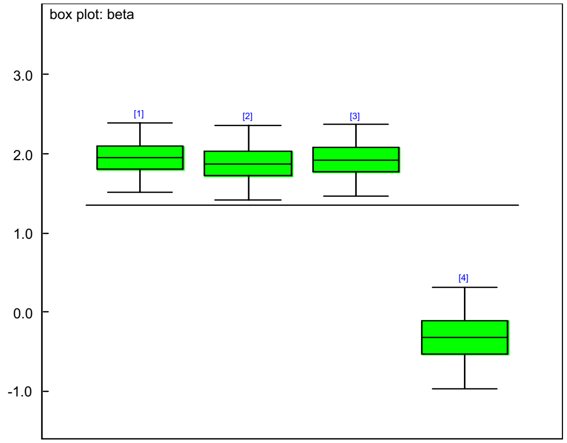
- We used noninformative priors N$(0, 1000)$ for the regression coefficients and a noninformative prior Gamma$(0.01, 0.01)$ for the error variance.

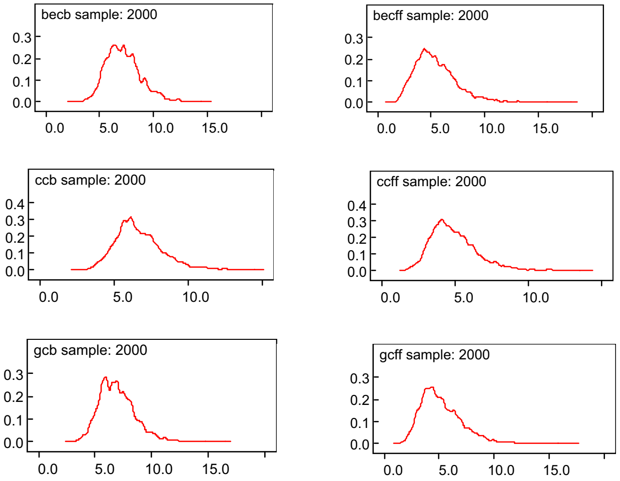












Regression - Posterior predictive checks

- For regression models, there are well-known methods for checking model and assumptions using estimated residuals. Residuals:

$$\epsilon_i = y_i - x_i'\beta$$

- Two useful test statistics are:
  - Proportion of outliers among residuals
  - Correlation between squared residuals and fitted values $\hat{y}$

- Posterior predictive distributions of both statistics can be obtained via simulation

- Define a standardized residual as

$$e_i = (y_i - x_i'\beta)/\sigma$$

- If normal model is correct, standardized residuals should be about $N(0,1)$ and therefore, $|e|_i > 3$ suggests that ith observation may be an outlier.

- To derive the posterior predictive distribution for the proportion of outliers q and for ρ, the correlation between $(e^2, \hat{y})$, do:

1. Draw (σ^2, β) from the joint posterior distribution
2. Draw y^{rep} from $N(X\beta, \sigma^2 I)$ given the existing X
3. Run the regression of y^{rep} on X, and save residuals
4. Compute ρ
5. Compute proportion of "large" standardized residuals q
6. Repeat for another y^{rep}

- Approach is *frequentist* in nature: we act as if we could repeat the experiment many times.

- Inspection of posterior predictive distribution of ρ and of q provides

information about model adequacy.

- Results of posterior predictive checks suggest that there are no outliers and that the correlation between the squared residuals and the fitted values is negligible.
- The 95% credible set for the proportion of absolute residuals (in the 41 observations) above 3 was $(0, 4.88)$ with a mean of 0.59% and a median of 0.
- The 95% credible set for ρ was $(-0.31, 0.32)$ with a mean of 0.011.
- Notice that the posterior distribution of the proportion of outliers is skewed. This sometimes happens when the quantity of interest is bounded and most of the mass of its distribution is close to the boundary.

Regression with unequal variances

- Consider now the case where $y \sim N(X\beta, \Sigma_y)$, with $\Sigma_y \neq \sigma^2 I$.
- Covariance matrix Σ_y is $n \times n$ and has $n(n+1)/2$ distinct parameters, and cannot be estimated from n observations. Must either specify Σ_y or it must be assigned an informative prior distribution.
- Typically, some structure is imposed on Σ_y to reduce the number of free parameters.

Regression - known Σ_y

- As before, let $p(\beta) \propto 1$ be the non-informative prior
- Since Σ_y is positive definite and symmetric, it has an upper-triangular square root matrix (Cholesky factor) $\Sigma_y^{1/2}$ such that $\Sigma_y^{1/2}\Sigma_y^{1/2'} = \Sigma_y$ so that if:

$$y = X\beta + e, \quad e \sim N(0, \Sigma_y)$$

then

$$\Sigma_y^{-1/2}y = \Sigma_y^{-1/2}X\beta + \Sigma_y^{-1/2}e, \quad \Sigma_y^{-1/2}e \sim N(0, I)$$

- With Σ_y known, just proceed as in ordinary linear regression, but use the transformed y and X as above and fix $\sigma^2 = 1$. Algebraically, this

is equivalent to computing

$$\begin{aligned} \hat{\beta} &= (X'\Sigma_y^{-1}X)^{-1}X\Sigma_y^{-1}y \\ V_\beta &= (X'\Sigma_y^{-1}X)^{-1} \end{aligned}$$

- Note: By using the Cholesky factor you avoid computing the $n \times n$ inverse Σ_y^{-1}.

Prediction of new $\tilde{y}$ with known Σ_y

- Even if we know Σ_y, prediction of new observations is more complicated: must know covariance matrix of old and new data.
- Example: heights of children from same family are correlated. To predict height of a new child when a brother is in the old dataset, we must include that correlation in the prediction.
- $\tilde{y}$ are $\tilde{n}$ new observations given a $\tilde{n} \times k$ matrix of regressors $\tilde{X}$.
- Joint distribution of $\tilde{y}, y$ is:

$$\begin{array}{rcl} \begin{array}{c} y \\ \tilde{y} \end{array} |X, \tilde{X}, \theta & \sim & N\left(\left[\begin{array}{c} X\beta \\ \tilde{X}\beta \end{array} \right], \left[\begin{array}{cc} \Sigma_y & \Sigma_{y\tilde{y}} \\ \Sigma_{\tilde{y}y} & \Sigma_{\tilde{y}\tilde{y}} \end{array} \right] \right) \\ \tilde{y}|y, \beta, \Sigma_y & \sim & N(\mu_{\tilde{y}}, V_{\tilde{y}}) \end{array}$$

where

$$\begin{array}{rcl} \mu_{\tilde{y}} & = & \tilde{X}\beta + \Sigma_{y\tilde{y}}\Sigma_{yy}^{-1}(y - X\beta) \\ V_{\tilde{y}} & = & \Sigma_{\tilde{y}\tilde{y}} - \Sigma_{y\tilde{y}}\Sigma_{yy}^{-1}\Sigma_{\tilde{y}y} \end{array}$$

Regression - unknown Σ_y

- To draw inferences about (β, Σ_y), we proceed in steps: derive $p(\beta|\Sigma_y, y)$ and then $p(\Sigma_y|y)$.
- Let $p(\beta) \propto 1$ as before
- We know that

$$\begin{array}{rcl} p(\Sigma_y|y) & = & \dfrac{p(\beta, \Sigma_y|y)}{p(\beta|\Sigma_y, y)} \propto \dfrac{p(\Sigma_y)p(y|\beta, \Sigma_y)}{p(\beta|\Sigma_y, y)} \\ & \propto & \dfrac{p(\Sigma_y)\ \mathrm{N}(y|\beta, \Sigma_y)}{\mathrm{N}(\beta|\hat{\beta}, V_\beta)}, \end{array}$$

where $(\hat{\beta}, V_\beta)$ depend on Σ_y.

- Expression for $p(\Sigma_y|y)$ must hold for any β, so we set $\beta = \hat{\beta}$.
- Note that

$$p(\beta|\Sigma_y, y) = \mathrm{N}(\hat{\beta}, V_\beta) \propto |V_\beta|^{1/2}$$

for $\beta = \hat{\beta}$. Then

$$p(\Sigma_y|y) \propto p(\Sigma_y)|V_\beta|^{1/2}|\Sigma_y|^{-1/2}\exp[-\frac{1}{2}(y - X\hat{\beta})'\Sigma_y^{-1}(y - X\hat{\beta})]$$

- In principle, $p(\Sigma_y|y)$ *could* be evaluated for a range of values of Σ_y. However:

1. It is difficult to determine a prior for an $n \times n$ unstructured matrix.
2. $\hat{\beta}, V_\beta$ depend on Σ_y and it is very difficult to draw values of Σ_y from $p(\Sigma_y|y)$.

- Need to put some structure on Σ_y

Regression - $\Sigma_y = \sigma^2 Q_y$

- Suppose that we know Σ_y upto a scalar factor σ^2.
- Non-informative prior on β, σ^2 is $p(\beta, \sigma^2) \propto \sigma^{-2}$
- Results follow directly from ordinary linear regression, by using transformation $Q_y^{-1/2}y$ and $Q_y^{-1/2}X$, which is equivalent to computing

$$\begin{aligned}
\hat{\beta} &= (X'Q_y^{-1}X)^{-1}X'Q_y^{-1}y \\
V_\beta &= (X'Q_y^{-1}X)^{-1} \\
s^2 &= (n-k)^{-1}(y - X\hat{\beta})'Q_y^{-1}(y - X\hat{\beta})
\end{aligned}$$

- To estimate the joint posterior distribution $p(\beta, \sigma^2|y)$ do

1. Draw σ^2 from Inv-$\chi^2(n-k, s^2)$
2. Draw β from n$(\hat{\beta}, V_\beta \sigma^2)$

- In large datasets, use Cholesky factor transformation and unweighted regression to avoid computation of Q_y^{-1}.

Regression - other covariance structures

Weighted regression

- In some applications, $\Sigma_y = \text{diag}(\sigma^2/w_i)$, for known weights and σ^2 unknown.
- Inference is the same as before, but now $Q_y^{-1} = \text{diag}(w_i)$

Parametric model for unequal variances:

- Variances may depend on weights in non-linear fashion: $\Sigma_{ii} = \sigma^2 v(w_i, \phi)$ for unknown parameter $\phi \in (0, 1)$ and known function v such as $v = w_i^{-\phi}$
- Note that for that function v:

$$\phi = 0 \quad \longrightarrow \quad \Sigma_{ii} = \sigma^2 \;\; \forall i$$

$$\phi = 1 \quad \longrightarrow \quad \Sigma_{ii} = \sigma^2/w_i \;\; \forall i$$

- In practice, to uncouple ϕ from the scale of the weights, we multiply weights by a factor so that their product equals 1.

Parametric model for unequal variances

- Three parameters to estimate: β, σ^2, ϕ.
- Possible prior for ϕ is uniform in $[0, 1]$
- Non-informative prior for β, σ^2 is $\propto \sigma^{-2}$
- Joint posterior distribution:

$$p(\beta, \sigma^2, \phi|y) \propto p(\phi)p(\beta, \sigma^2)\Pi_{i=1}^n \mathsf{N}(y_i; (X\beta)_i, \sigma^2 v(w_i, \phi))$$

- For a given ϕ, we are back in the earlier weighted regression case with

$$Q_y^{-1} = \text{ diag}(v(w_1, \phi), ..., v(w_n, \phi))$$

- This suggests the following scheme
  1. Draw ϕ from marginal posterior $p(\phi|y)$
  2. Compute $Q_y^{-1/2}y$ and $Q_y^{-1/2}X$
  3. Draw σ^2 from $p(\sigma^2|\phi, y)$ as in ordinary regression
  4. Draw β from $p(\beta|\sigma^2, \phi, y)$ as in ordinary regression
- Marginal posterior distribution of ϕ:

$$\begin{aligned} p(\phi|y) &= \frac{p(\beta, \sigma^2, \phi|y)}{p(\beta, \sigma^2|\phi, y)} \\ &= \frac{p(\beta, \sigma^2, \phi|y)}{p(\beta|\sigma^2, \phi, y)p(\sigma^2|\phi, y)} \\ &\propto \frac{p(\phi)\sigma^{-2}\Pi_i \mathsf{N}(y_i|(X\beta)_i, \sigma^2 v(w_i, \phi))}{\text{Inv-}\chi^2(n-k, s^2)\ \mathsf{N}(\hat{\beta}, V_\beta)} \end{aligned}$$

- Expression must hold for any (β, σ^2), so we set $\beta = \hat{\beta}$ and $\sigma^2 = s^2$.

- Recall that $\hat{\beta}$ and s^2 depend on ϕ.
- Also recall that weights are scaled to have product equal to 1.
- Then

$$p(\phi|y) \propto p(\phi)|V_\beta|^{1/2}(s^2)^{-(n-k)/2}$$

- **To carry out computation**
  - First sample ϕ from $p(\phi|y)$:
    1. Evaluate $p(\phi|y)$ for a range of values of $\phi \in [0, 1]$
    2. Use inverse cdf method to sample ϕ
  - Given ϕ, compute $y^* = Q_y^{-1/2}y$ and $X^* = Q_y^{-1/2}X$
  - Compute

$$\begin{aligned} \hat{\beta} &= (X^{*'}X^*)^{-1}X^{*'}y \\ V_\beta &= (X^{*'}X^*)^{-1} \end{aligned}$$

$$s^2 = (n-k)^{-1}(y^* - X^*\hat{\beta})'(y^* - X^*\hat{\beta})$$

- Draw σ^2 from Inv-$\chi^2(n-k, s^2)$
- Draw β from $\mathsf{N}(\hat{\beta}, \sigma^2 V_\beta)$

Including prior information about β

- Suppose we wish to add prior information about a single regression coefficient β_j of the form:

$$\beta_j \sim \mathsf{N}(\beta_{j0}, \sigma^2_{\beta_j}),$$

  with $\beta_{j0}, \sigma^2_{\beta_j}$ known.

- Prior information can be added in the form of an additional 'data point'.

- An ordinary observation y is normal with mean $x\beta$ and variance σ^2.

- As a function of β_j, the prior can be viewed as an 'observation' with
  - 0 on all x's except x_j

  - variance $\sigma^2_{\beta_j}$.

- To include prior information, do the following:

  1. Append one more 'data point' to vector y with value β_{j0}.
  2. Add one row to X with zeroes except in jth column.
  3. Add a diagonal element with value $\sigma^2_{\beta_j}$ to Σ_y.

- Now apply computational methods for non-informative prior.

- Given Σ_y, posterior for β obtained by weighted linear regression.

Adding prior information for several βs

- Suppose that for the entire vector β,

$$\beta \sim \mathsf{N}(\beta_0, \Sigma_\beta).$$

- Proceed as before: add k 'data points' and draw posterior inference by weighted linear regression applied to 'observations' y_*, explanatory variables X_* and variance matrix Σ_*:

$$y_* = \begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \quad X_* = \begin{bmatrix} X \\ I_k \end{bmatrix}, \quad \Sigma_* = \begin{bmatrix} \Sigma_y & 0 \\ 0 & \Sigma_\beta \end{bmatrix}$$

- Computation can be carried out conditional on Σ_* first and then inverse cdf for Σ_* or using the Gibbs sampling.

Prior information about σ^2

- Typically we do not wish to include prior information about σ^2.

- If we do, we can use the conjugate prior

$$\sigma^2 \sim \text{Inv-}\chi^2(n_0, \sigma_0^2).$$

- The marginal posterior of σ^2 is

$$\sigma^2|y \sim \text{Inv-}\chi^2(n_0 + n, \frac{n_0\sigma_0^2 + nS^2}{n_0 + n}).$$

- If prior information on β is also incorporated, S^2 is replaced by corresponding value from regression of y_* on X_* and Σ_*, and n is replaced by length of y_*.

Inequality constraints on β

- Sometimes we wish to impose inequality constraints such as

$$\beta_1 > 0$$

  or

$$\beta_2 < \beta_3 < \beta_4.$$

- The easiest way is to ignore the constraint until the end:
  - Simulate (β, σ^2) from posterior
  - discard all the draws that do not satisfy the constraint.

- Typically a reasonably efficient way to proceed unless constraint eliminates a large portion of unconstrained posterior distribution.

- If so, data tend to contradict the model.

Generalized Linear Models

- Generalized linear models are an extension of linear models to the case where relationship between $E(y|X)$ and X is not linear or normal assumption is not appropriate.

- Sometimes a transformation suffices to return to the linear setup. Consider the multiplicative model

$$y_i = x_{i1}^{b1} x_{i2}^{b2} x_{i3}^{b3} \epsilon_i$$

  A simple log transformation leads to

$$\log(y_i) = b_1 \log(x_{i1}) + b_2 \log(x_{i2}) + b_3 \log(x_{i3}) + e_i$$

- When simple approaches do not work, we use GLIMs.

- There are three main components in the model:

  1. Linear predictor $\eta = X\beta$
  2. Link function $g(.)$ relating linear predictor to mean of outcome variable: $E(y|X) = \mu = g^{-1}(\eta) = g^{-1}(X\beta)$
  3. Distribution of outcome variable y with mean $\mu = E(y|X)$. Distribution can also depend on a *dispersion parameter* ϕ:

$$p(y|X,\beta,\phi) = \Pi_{i=1}^n p(y_i|(X\beta)_i, \phi)$$

- In standard GLIMs for Poisson and binomial data, $\phi = 1$.

- In many applications, however, excess dispersion is present.

Some standard GLIMs

- **Linear model**:
  - Simplest GLIM, with identity link function $g(\mu) = \mu$.

- **Poisson model**:
  - Mean and variance μ and link function $\log(\mu) = X\beta$, so that

$$\mu = \exp(X\beta) = \exp(\eta)$$

  - For $y = (y_1, ..., y_n)$:

$$p(y|\beta) = \Pi_{i=1}^n \frac{1}{y!} \exp(-\exp(\eta_i))(\exp(\eta_i))^{y_i}$$

    with $\eta_i = (X\beta)_i$.

- **Binomial model**: Suppose that $y_i \sim \text{Bin}(n_i, \mu_i)$, n_i known. Standard link function is logit of probability of success μ:

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = (X\beta)_i = \eta_i$$

- For a vector of data y:

$$p(y|\beta) = \Pi_{i=1}^n \binom{n_i}{y_i} \left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right)^{y_i} \left(\frac{1}{1+\exp(\eta_i)}\right)^{n_i - y_i}$$

- Another link used in econometrics is the *probit* link:

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

  with $\Phi(.)$ the normal cdf.

- In practice, inference from logit and probit models is almost the same, except in extremes of the tails of the distribution.

Overdispersion

- In many applications, the model can be formulated to allow for extra variability or *overdispersion*.
- E.g. in Poisson model, the variance is constrained to be equal to the mean.
- As an example, suppose that data are the number of fatal car accidents at K intersections over T years. Covariates might include intersection characteristics and traffic control devices (stop lights, etc).
- To accommodate overdispersion we model the (log)rate as a linear combination of covariates and add a random effect for intersection with its own population distribution.

Setting up GLIMs

- **Canonical link functions:** Canonical link is function of mean that appears in exponent of exponential family form of sampling distribution.
- All links discussed so far are canonical except for the probit.
- **Offset:** Arises when counts are obtained from different population sizes or volumes or time periods and we need to use an exposure. Offset is a covariate with a known coefficient.
- Example: Number of incidents in a given exposure time T are Poisson with rate μ per unit of time. Mean number of incidents is μT.
- Link function would be $\log(\mu) = \eta_i$, but here mean of y is not μ but μT.

- To apply the Poisson GLIM, add a column to X with values $\log(T)$ and fix the coefficient to 1. This is an offset.

Interpreting GLIMs

- In linear models, β_j represents the change in the outcome when x_j is changed by one unit.
- Here, β_j reflects changes in $g(\mu)$ when x_j is changed.
- The effect of changing x_j depends of current value of x.
- To translate effects into the scale of y, measure changes relative to a baseline
$$y_0 = g^{-1}(x_0\beta).$$
- A change in x of Δx takes outcome from y_0 to y where
$$g(y_0) = x_0\beta \longrightarrow y_0 = g^{-1}(x_0\beta)$$

and
$$y = g^{-1}(g(y_0) + (\Delta x)\beta)$$

Priors in GLIM

- We focus on priors for β although sometimes ϕ is present and has its own prior.
- *Non-informative prior for* β:
  - With $p(\beta) \propto 1$, posterior mode = MLE for β
  - Approximate posterior inference can be based on normal approximation to posterior at mode.
- *Conjugate prior for β:*
  - As in regression, express prior information about β in terms of hypothetical data obtained under same model.
  - Augment data vector and model matrix with y_0 hypothetical observations and $X_{0_{n_0 \times k}}$ hypothetical predictors.

  - Non-informative prior for β in augmented model.
- *Non-conjugate priors:*
  - Often more natural to model $p(\beta|\beta_0, \Sigma_0) = N(\beta_0, \Sigma_0)$ with (β_0, Σ_0) known.
  - Approximate computation based on normal approximation (see next) particularly suitable.
- *Hierarchical GLIM*:
  - Same approach as in linear models.
  - Model some of the β as exchangeable with common population distribution with unknown parameters. Hyperpriors for parameters.

Computation

- Posterior distributions of parameters can be estimated using MCMC methods in WinBUGS or other software.
- Metropolis within Gibbs will often be necessary: in GLIM, most often full conditionals do not have standard form.
- An alternative is to **approximate** the sampling distribution with a **cleverly chosen** approximation.
- **Idea**:
  - Find mode of likelihood $(\hat{\beta}, \hat{\phi})$ perhaps conditional on hyperparameters
  - Create *pseudo-data* with their *pseudo-variances* (see later)
  - Model pseudo-data as normal with known (pseudo-)variances.

Normal approximation to likelihood

- Objective: find z_i and σ_i^2 such that normal likelihood

$$N(z_i|(X\beta)_i, \sigma_i^2)$$

is good approximation to GLIM likelihood $p(y_i|(X\beta)_i, \phi)$.

- Let $(\hat{\beta}, \hat{\phi})$ be mode of (β, ϕ) so that $\hat{\eta}_i$ is the mode of η_i.
- For L the loglikelihood, write

$$\begin{aligned} p(y_1, ..., y_n) &= \Pi_i p(y_i|\eta_i, \phi) \\ &= \Pi_i \exp(L(y_i|\eta_i, \phi)) \end{aligned}$$

- Approximate factor in exponent by normal density in η_i:

$$L(y_i|\eta_i, \phi) \approx -\frac{1}{2\sigma_i^2}(z_i - \eta_i)^2,$$

where (z_i, σ_i^2) depend on (y_i, η_i, ϕ).

- Now need to find expressions for (z_i, σ_i^2).
- To get (z_i, σ_i^2), match first and second order terms in Taylor approx around $\hat{\eta}_i$ to (η_i, σ_i^2) and solve for z_i and for σ_i^2.
- Let $L' = \delta L/\delta\eta_i$:

$$L' = \frac{1}{\sigma_i^2}(z_i - \eta_i)$$

- Let $L'' = \delta^2 L/\delta\eta_i^2$:

$$L'' = -\frac{1}{\sigma_i^2}$$

- Then

$$\begin{aligned} z_i &= \hat{\eta}_i - \frac{L'(y_i|\hat{\eta}_i, \hat{\phi})}{L''(y_i|\hat{\eta}_i, \hat{\phi})} \\ \sigma_i^2 &= -\frac{1}{L''(y_i|\hat{\eta}_i, \hat{\phi})} \end{aligned}$$

- Example: binomial model with logit link:

$$\begin{aligned} L(y_i, |\eta_i) &= y_i \log\left(\frac{\exp(\eta_i)}{1+\exp(\eta_i)}\right) \\ &\quad +(n_i - y_i)\log\left(\frac{1}{1+\exp(\eta_i)}\right) \end{aligned}$$

$$= \quad y_i\eta_i - n_i \log(1+\exp(\eta_i))$$

- Then

$$\begin{aligned} L' &= y_i - n_i \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \\ L'' &= -n_i \frac{\exp(\eta_i)}{(1+\exp(\eta_i))^2} \end{aligned}$$

- Pseudo-data and pseudo-variances:

$$\begin{aligned} z_i &= \hat{\eta}_i + \frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \left(\frac{y_i}{n_i} - \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)} \right) \\ \sigma_i^2 &= \frac{1}{n_i} \frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \end{aligned}$$

Models for multinomial responses

- Multinomial data: outcomes $y = (y_i, ..., y_K)$ are counts in K categories.
- Examples:
  - Number of students receiving grades A, B, C, D or F
  - Number of alligators that prefer to eat reptiles, birds, fish, invertebrate animals, or other (see example later)
  - Number of survey respondents who prefer Coke, Pepsi or tap water.
- In Chapter 3, we saw non-hierarchical multinomial models:

$$p(y|\alpha) \propto \Pi_{j=1}^{k} \alpha_j^{y_j}$$

  with α_j: probability of jth outcome and $\sum_{j=1}^{k} \alpha_j = 1$ and $\sum_{j=1}^{k} y_j = n$.

- Here: we model α_j as a function of covariates (or predictors) X with corresponding regression coefficients β_j.
- For full hierarchical structure, the β_j are modeled as exchangeable with some common population distribution $p(\beta|\mu,\tau)$.
- Model can be developed as extension of either binomial or Poisson models.

Logit model for multinomial data

- Here $i = 1, ..., I$ is number of covariate patters. E.g., in alligator example, 2 sizes × four lakes = 8 covariate categories.
- Let y_i be a multinomial random variable with sample size n_i and k possible outcomes. Then

$$y_i \sim \text{ Mult } (n_i; \alpha_{i1}, ..., \alpha_{ik})$$

  with $\sum_i y_i = n_i$, and $\sum_j^k \alpha_{ij} = 1$.
- α_{ij} is the probability of jth outcome for ith covariate combination.
- Standard parametrization: log of the probability of jth outcome relative

to baseline category $j = 1$:

$$\log\left(\frac{\alpha_{ij}}{\alpha_{i1}}\right) = \eta_{ij} = (X\beta_j)_i,$$

with β_j a vector of regression coefficients for jth category.

- Sampling distribution:

$$p(y|\beta) \propto \Pi_{i=1}^I \Pi_{j=1}^k \left(\frac{\exp(\eta_{ij})}{\sum_{l=1}^k \exp(\eta_{il})}\right)^{y_{ij}}.$$

- For identifiability, $\beta_1 = 0$ and thus $\eta_{i1} = 0$ for all i.
- β_j is effect of changing X on probability of category j relative to category 1.

- Typically, indicators for each outcome category are added to predictors to indicate relative frequency of each category when $X = 0$. Then

$$\eta_{ij} = \delta_j + (X\beta_j)_i$$

with $\delta_1 = \beta_1 = 0$ typically.

Example from WinBUGS - Alligators

- Agresti (1990) analyzes feeding choices of 221 alligators.
- Response is one of five categories: fish, invertebrate, reptile, bird, other.
- Two covariates: length of alligator (less than 2.3 meters or larger than 2.3 meters) and lake (Hancock, Oklawaha, Trafford, George).
- $2 \times 4 = 8$ covariate combinations (see data)
- For i, j a combination of size and lake, we have counts in five possible categories $y_{ij} = (y_{ij1}, ..., y_{ij5})$.
- Model

$$p(y_{ij}|\alpha_{ij}, n_{ij}) = \text{ Mult } (y_{ij}|n_{ij}, \theta_{ij1}, ..., \theta_{ij5})$$

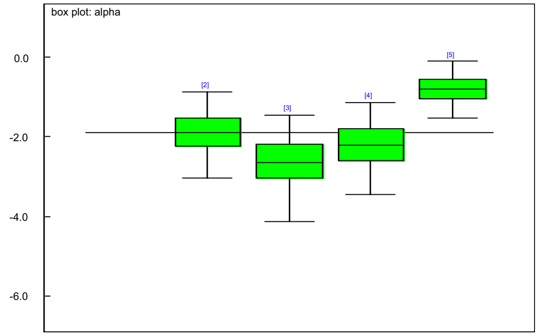
with

$$\theta_{ijk} = \frac{\exp(\eta_{ijk})}{\sum_{l=1}^k \exp(\eta_{ijl})},$$

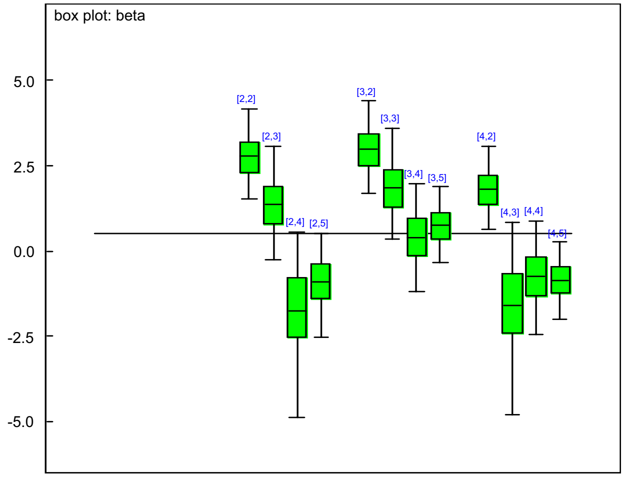
and

$$\eta_{ijk} = \delta_k + \beta_{ik} + \gamma_{jk}.$$

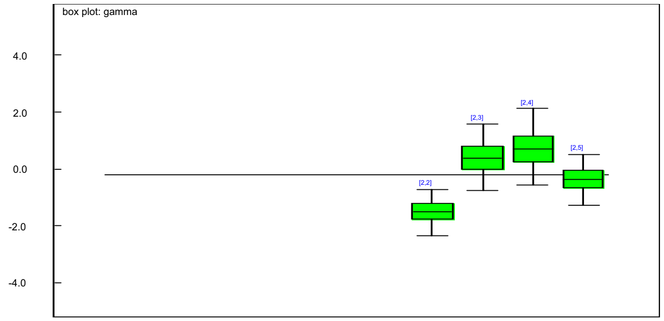
- Here,
  - δ_k is baseline indicator for category k
  - β_{ik} is coefficient for indicator for lake
  - γjk is coefficient for indicator for size



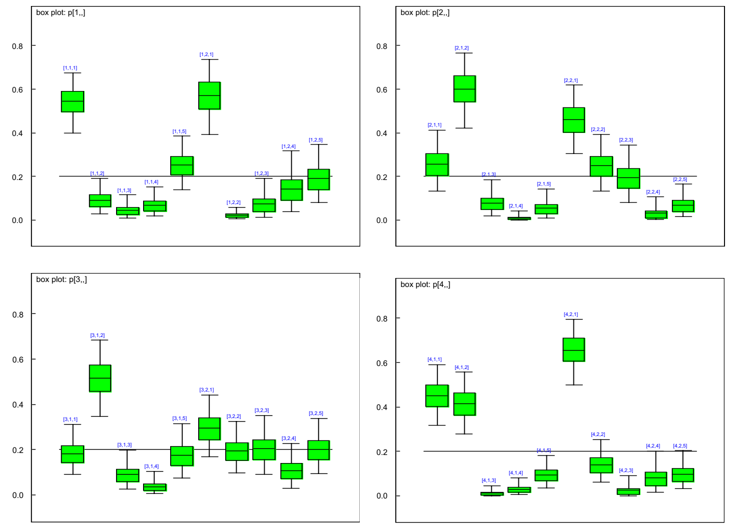
	Mean	Std	2.5%	Median	97.5%
alpha[2]	-1.838	0.5278	-2.935	-1.823	-0.8267
alpha[3]	-2.655	0.706	-4.261	-2.593	-1.419
alpha[4]	-2.188	0.58	-3.382	-2.153	-1.172
alpha[5]	-0.7844	0.3691	-1.531	-0.7687	-0.1168



	Mean	**Std**	**2.5%**	**Median**	**97.5%**
beta[2,2]	2.706	0.6431	1.51	2.659	4.008
beta[2,3]	1.398	0.8571	-0.2491	1.367	3.171
beta[2,4]	-1.799	1.413	-5.1	-1.693	0.5832
beta[2,5]	-0.9353	0.7692	-2.555	-0.867	0.4413
beta[3,2]	2.932	0.6864	1.638	2.922	4.297
beta[3,3]	1.935	0.8461	0.37	1.886	3.842
beta[3,4]	0.3768	0.7936	-1.139	0.398	1.931
beta[3,5]	0.7328	0.5679	-0.3256	0.7069	1.849
beta[4,2]	1.75	0.6116	0.636	1.73	3.023
beta[4,3]	-1.595	1.447	-4.847	-1.433	0.9117
beta[4,4]	-0.7617	0.8026	-2.348	-0.7536	0.746
beta[4,5]	-0.843	0.5717	-1.97	-0.8492	0.281



	Mean	**Std**	**2.5%**	**Median**	**97.5%**
gamma[2,2]	-1.523	0.4101	-2.378	-1.523	-0.7253
gamma[2,3]	0.342	0.5842	-0.788	0.3351	1.476
gamma[2,4]	0.7098	0.6808	-0.651	0.7028	2.035
gamma[2,5]	-0.353	0.4679	-1.216	-0.3461	0.518



p[1,1,1]	0.5392	0.07161	0.4046	0.5384	0.676
p[1,1,2]	0.09435	0.04288	0.03016	0.08735	0.2067
p[1,1,3]	0.04585	0.02844	0.007941	0.04043	0.1213
p[1,1,4]	0.06837	0.03418	0.01923	0.06237	0.1457
p[1,1,5]	0.2523	0.06411	0.1389	0.2496	0.3808
p[1,2,1]	0.5679	0.09029	0.3959	0.5684	0.7342
p[1,2,2]	0.02322	0.01428	0.005301	0.02013	0.06186
p[1,2,3]	0.06937	0.04539	0.01191	0.05918	0.1768
p[1,2,4]	0.1469	0.07437	0.03647	0.1344	0.3182
p[1,2,5]	0.1927	0.07171	0.0793	0.1852	0.349
p[2,1,1]	0.2578	0.07217	0.1367	0.2496	0.4265
p[2,1,2]	0.5968	0.08652	0.4159	0.6	0.7551
p[2,1,3]	0.08137	0.0427	0.01939	0.0728	0.1841
p[2,1,4]	0.0095	0.01186	0.00533	0.04105	0.143
p[2,1,5]	0.05445	0.03517	0.009716	0.0476	0.1398
p[2,2,1]	0.4612	0.08211	0.3055	0.4632	0.6175
p[2,2,2]	0.2455	0.06879	0.1263	0.2426	0.3867
p[2,2,3]	0.1947	0.07042	0.07776	0.189	0.3502
p[2,2,4]	0.0302	0.02921	7.653E-4	0.02009	0.1122
p[2,2,5]	0.06833	0.03984	0.01357	0.06097	0.1643
p[3,1,1]	0.1794	0.05581	0.08555	0.1732	0.296
p[3,1,2]	0.5178	0.09136	0.3334	0.5185	0.6933
p[3,1,3]	0.09403	0.04716	0.02652	0.08517	0.2219
p[3,1,4]	0.03554	0.02504	0.006063	0.02996	0.1055
p[3,1,5]	0.1732	0.06253	0.07053	0.167	0.3261
p[3,2,1]	0.2937	0.07225	0.1618	0.2901	0.4469

Interpretation of results

Because we set to zero several model parameters, interpreting results is tricky. For example:

- Beta[2,2] ha posterior mean 2.714. This is the effect of lake Oklawaha (relative to Hancock) on the alligator's preference for invertebrates relative to fish.

  Since beta[2,2] > 0, we conclude that alligators in Oklawaha eat more invertebrates than do alligators in Hancock (even though both may prefer fish!).

- Gamma[2,2] is the effect of size 2 relative to size on the relative preference for invertebrates. Since gamma[2,2] < 0, we conclude that large alligators prefer fish more than do small alligators.

- The alpha are baseline counts for each type of food relative to fish.

Hierarchical Poisson model

- Count data are often modeled using a Poisson model.

- If $y \sim$ Poisson(μ) then $E(y) = var(y) = \mu$.

- When counts are assumed exchangeable given μ and the rates μ can also be assumed to be exchangeable, a Gamma population model for the rates is often chosen.

- The hierarchical model is then

$$\begin{aligned} y_i &\sim \text{Poisson}(\mu_i) \\ \mu_i &\sim \text{Gamma}(\alpha, \beta). \end{aligned}$$

- Priors for the hyperparameters are often taken to be Gamma (or exponential):

$$\begin{aligned}\alpha &\sim \text{Gamma}(a,b)\\ \beta &\sim \text{Gamma}(c,d),\end{aligned}$$

  with (a,b,c,d) known.

- The joint posterior distribution is

$$\begin{aligned}p(\mu,\alpha,\beta|y) \;\propto\; & \Pi_i \mu_i^{y_i}\exp\{-\mu_i\}\mu_i^{\alpha-1}\exp\{-\mu_i\beta\}\\ & \alpha^{a-1}\exp\{-\alpha b\}\beta^{c-1}\exp\{-\beta d\}\end{aligned}$$

- To carry out Gibbs sampling we need to find the full conditional distributions.

- Conditional for μ_i is

$$p(\mu_i|\text{ all}) \propto \mu_i^{y_i+\alpha-1}\exp\{-\mu_i(\beta+1)\},$$

  which is proportional to a Gamma with parameters $(y_i+\alpha, \beta+1)$.

- The full conditional for α is

$$p(\alpha|\text{ all}) \propto \Pi_i \mu_i^{\alpha-1}\alpha^{a-1}\exp\{\alpha b\}.$$

- The conditional for α does not have a standard form.

- For β:

$$\begin{aligned}p(\beta|\text{ all}) &\propto \Pi_i \exp\{-\beta\mu_i\}\beta^{c-1}\exp\{-\beta d\}\\ &\propto \beta^{c-1}\exp\{-\beta(\sum_i \mu_i + d)\},\end{aligned}$$

which is proportional to a Gamma with parameters $(c, \sum_i \mu_i + d)$.

- Computation:
  - Given α, β, draw each μ_i from the corresponding Gamma conditional.
  - Draw α using a Metropolis step or rejection sampling or inverse cdf method.
  - Draw β from the Gamma conditional.

- See Italian marriages example.

Italian marriages – Example

Gill (2002) collected data on the number of marriages per 1,000 people in Italy during 1936-1951.

Question: did the number of marriages decrease during WWII years? (1939 – 1945).

Model:

Number of marriages y_i are Poisson with year-specific means λ_i.

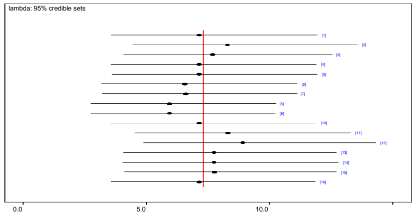
Assuming that rates of marriages are exchangeable across years, we model the λ_i as Gamma(α, β).

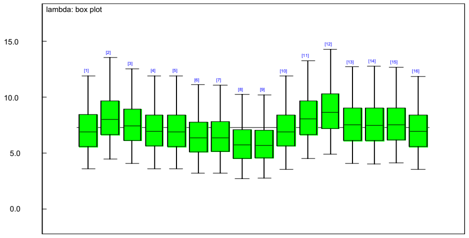
To complete model specification, place independent Gamma priors on (α, β), with known hyper-parameter values.

WinBUGS code:

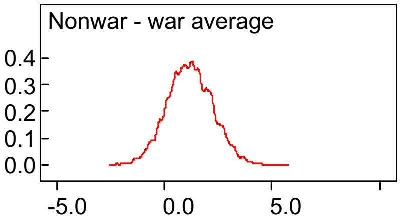
```
model {
  for (i in 1:16) {
     y[i] ~ dpois(l[i])
     l[i] ~ dgamma(alpha, beta)
          }

alpha ~ dgamma(1,1)
beta ~ dgamma(1,1)
warave <- (l[4]+l[5]+ l[6]+l[7]+l[8]+l[9]+l[10]) / 7
nonwarave<- (l[1]+l[2]+l[3]+l[11]+l[12]+l[13]+l[14]+l[15]+l[16]) / 9
diff <- nonwarave - warave
}

list(y = c(7,9,8,7,7,6,6,5,5,7,9,10,8,8,8,7))
```

Results





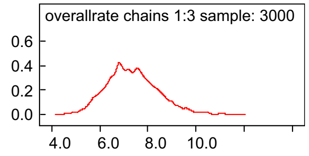
Difference between non-war and war years marriage rate

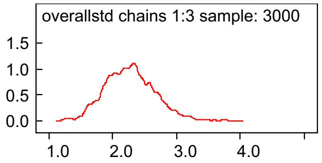


Overall marriage rate:

If $\lambda_i \sim \text{Gamma}(\alpha, \beta)$, then $E(\lambda_i \mid y) = \alpha / \beta$.

	mean	sd	2.5%	median	97.5%
overallrate	7.362	1.068	5.508	7.285	9.665
overallstd	2.281	0.394	1.59	2.266	3.125





Poisson regression

- When rates are not exchangeable, we need to incorporate covariates into the model. Often we are interested in the association between one or more covariate and the outcome.

- It is possible (but not easy) to incorporate covariates into the Poisson-Gamma model.

- Christiansen and Morris (1997, JASA) propose the following model:

- Sampling distribution, where e_i is a known exposure:

$$y_i|\lambda_i \sim \text{ Poisson}(\lambda_i e_i).$$

  Under model, $E(y_i/e_i) = \lambda_i$.

- Population distribution for the rates:

$$\lambda_i|\alpha \sim \text{ Gamma}(\zeta, \zeta/\mu_i),$$

  with $\log(\mu_i) = x_i'\beta$, and $\alpha = (\beta_0, \beta_1, ..., \beta_{k-1}, \zeta)$.

- ζ is thought of as an unboserved prior count.

- Under population model,

$$\begin{aligned} E(\lambda_i) &= \frac{\zeta}{\zeta/\mu_i} \\ &= \mu_i \\ \text{CV}^2(\lambda_i) &= \frac{\mu_i^2}{\zeta}\frac{1}{\mu_i^2} \\ &= \frac{1}{\zeta} \end{aligned}$$

- For $k = 0$, μ_i is known. For $k = 1$, μ_i are exchangeable. For $k \geq 2$, μ_i are (unconditionally) nonexchangeable.

- In all cases, standardized rates λ_i/μ_i are Gamma(ζ, ζ), are exchangeable, and have expectation 1.

- The covariates can include random effects.

- To complete specification of model, we need priors on α.

- Christensen and Morris (1997) suggest:
  - β and ζ independent a priori.
  - Non-informative prior on β's associated to 'fixed' effects.
  - For ζ a proper prior of the form:

$$p(\zeta|y_0) \propto \frac{y_0}{(\zeta + y_0)^2},$$

where y_0 is the prior guess for the median of ζ.

- Small values of y_0 (for example, $y_0 < \hat{\zeta}$ and $\hat{\zeta}$ the MLE of ζ) provide less information.

- When the rates cannot be assumed to be exchangeable, it is common to choose a generalized linear model of the form:

$$\begin{aligned} p(y|\beta) &\propto \Pi_i \exp\{-\lambda_i\}\lambda_i^{y_i} \\ &\propto \Pi_i \exp\{-\exp(\eta_i)\}[\exp(\eta_i)]^{y_i}, \end{aligned}$$

  for $\eta_i = x_i'\beta$ and $\log(\lambda_i) = \eta_i$.

- The vector of covariates can include one or more random effects to accommodate additional dispersion (see epylepsy example).

- The second-level distribution for the β's will typically be flat (if covariate is a 'fixed' effect) or normal

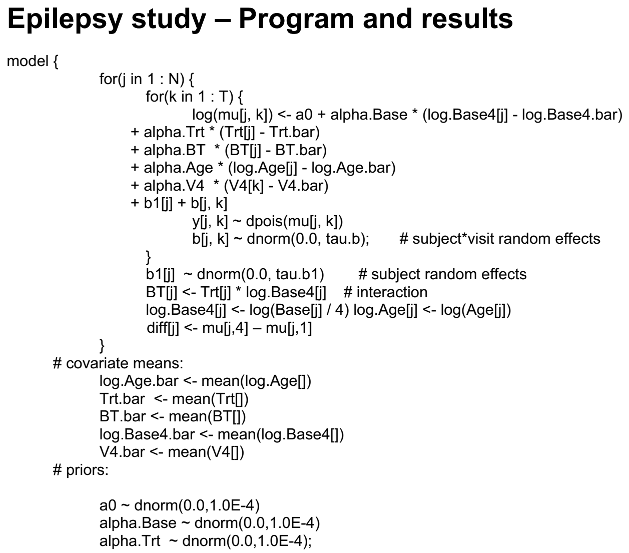
$$\beta_j \sim \text{Normal}(\beta_{j0}, \sigma^2_{\beta_j})$$

  if jth covariate is a random effect. The variance $\sigma^2_{\beta_j}$ represents the between 'batch' variability.

Epilepsy example

- From Breslow and Clayton, 1993, JASA.
- Fifty nine epilectic patients in a clinical trial were randomized to a new drug: T = 1 is the drug and T = 0 is the placebo.
- Covariates included:
  - Baseline data: number of seizures during eight weeks preceding trial
  - Age in years.
- Outcomes: number of seizures during the two weeks preceding each of four clinical visits.
- Data suggest that number of seizures was significantly lower prior to fourth visit, so an indicator was used for V4 versus the others.

- Two random effects in the model:
  - A patient-level effect to introduce between patient variability.
  - A patients by visit effect to introduce between visit within patient dispersion.



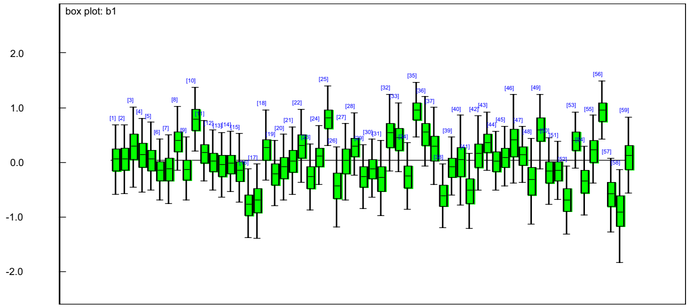
Epilepsy study – Program and results

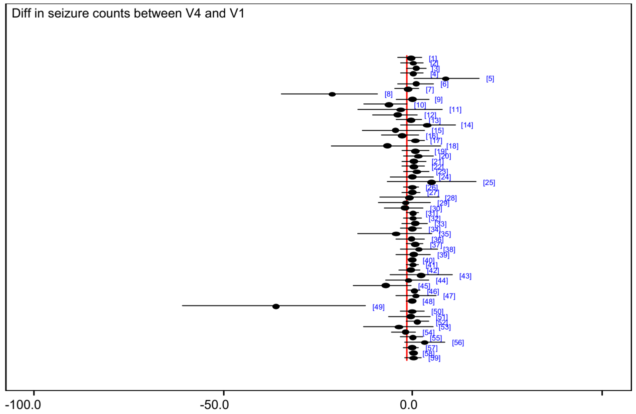
```
model {
                for(j in 1 : N) {
                        for(k in 1 : T) {
                                log(mu[j, k]) <- a0 + alpha.Base * (log.Base4[j] - log.Base4.bar)
                     + alpha.Trt * (Trt[j] - Trt.bar)
                     + alpha.BT  * (BT[j] - BT.bar)
                     + alpha.Age * (log.Age[j] - log.Age.bar)
                     + alpha.V4  * (V4[k] - V4.bar)
                     + b1[j] + b[j, k]
                                y[j, k] ~ dpois(mu[j, k])
                                b[j, k] ~ dnorm(0.0, tau.b);       # subject*visit random effects
                        }
                        b1[j]  ~ dnorm(0.0, tau.b1)        # subject random effects
                        BT[j] <- Trt[j] * log.Base4[j]    # interaction
                        log.Base4[j] <- log(Base[j] / 4) log.Age[j] <- log(Age[j])
                        diff[j] <- mu[j,4] – mu[j,1]
                }
        # covariate means:
                log.Age.bar <- mean(log.Age[])
                Trt.bar  <- mean(Trt[])
                BT.bar <- mean(BT[])
                log.Base4.bar <- mean(log.Base4[])
                V4.bar <- mean(V4[])
        # priors:

                a0 ~ dnorm(0.0,1.0E-4)
                alpha.Base ~ dnorm(0.0,1.0E-4)
                alpha.Trt  ~ dnorm(0.0,1.0E-4);
        alpha.BT   ~ dnorm(0.0,1.0E-4)
        alpha.Age  ~ dnorm(0.0,1.0E-4)
        alpha.V4   ~ dnorm(0.0,1.0E-4)
        tau.b1     ~ dgamma(1.0E-3,1.0E-3); sigma.b1 <- 1.0 / sqrt(tau.b1)
        tau.b      ~ dgamma(1.0E-3,1.0E-3); sigma.b  <- 1.0/  sqrt(tau.b)

# re-calculate intercept on original scale:
        alpha0 <- a0 - alpha.Base * log.Base4.bar - alpha.Trt * Trt.bar
        - alpha.BT * BT.bar - alpha.Age * log.Age.bar - alpha.V4 * V4.bar
}
```

Results

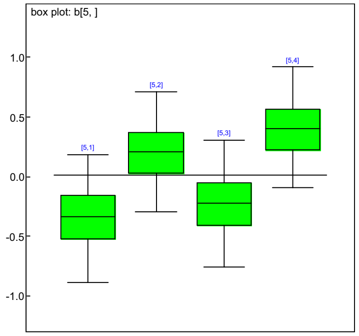
Parameter	Mean	Std	2.5th	Median	97.5th
alpha.Age	0.4677	0.3557	-0.2407	0.4744	1.172
alpha.Base	0.8815	0.1459	0.5908	0.8849	1.165
alpha.Trt	-0.9587	0.4557	-1.794	-0.9637	-0.06769
alpha.V4	-0.1013	0.08818	-0.273	-0.09978	0.07268
alpha.BT	0.3778	0.2427	-0.1478	0.3904	0.7886
sigma.b1	0.4983	0.07189	0.3704	0.4931	0.6579
sigma.b	0.3641	0.04349	0.2871	0.362	0.4552

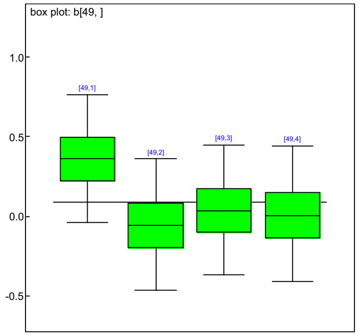
Individual random effects





Patients 5 and 49 are ‘different’. For #5, the number of events increases from visits 1 through 4. For #49, there is a significant decrease.



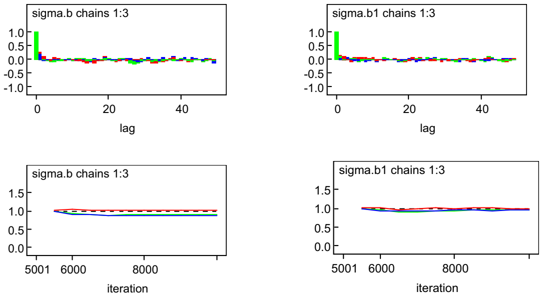


Convergence and autocorrelation

We ran three parallel chains for 10,000 iterations

Discarded the first 5,000 iterations from each chain as burn-in

Thinned by 10, so there are 1,500 draws for inference.



Example: hierarchical Poisson model

- The USDA collects data on the number of farmers who adopt conservation tillage practices.
- In a recent survey, 10 counties in a midwestern state were randomly sampled, and within county, a random number of farmers were interviewed and asked whether they had adopted conservation practices.
- The number of farmers interviewed in each county varied from a low of 2 to a high of 100.
- Of interest to USDA is the estimation of the overall adoption rate in the state.

- We fitted the following model:

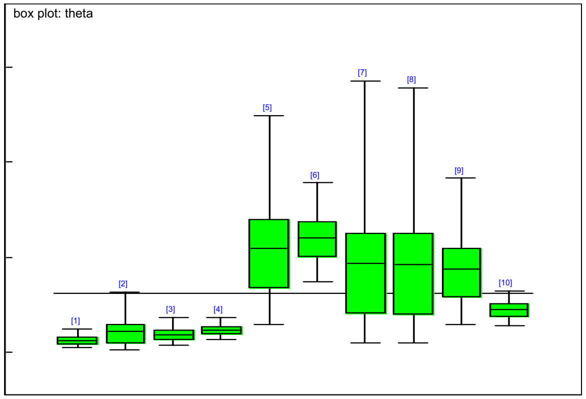
$$\begin{aligned} y_i &\sim \text{Poisson}(\lambda_i) \\ \lambda_i &= \theta_i E_i \\ \theta_i &\sim \text{Gamma}(\alpha, \beta), \end{aligned}$$

  where y_i is the number of adopters in the ith county, E_i is the number of farmers interviewed, θ_i is the expected adoption rate, and the expected number of adopters in a county can be obtained by multiplying θ_i by the number of farms in the county.
- The hierarchical structure establishes that even though counties may vary significantly in terms of the rate of adoption, they are still exchangeable, so the rates are generated from a common population distribution.
- Information about the *overall rate of adoption* across the state is contained in the posterior distribution of (α, β).
- We fitted the model using WinBUGS and chose non-informative priors for the hyperparameters.

Observed data:

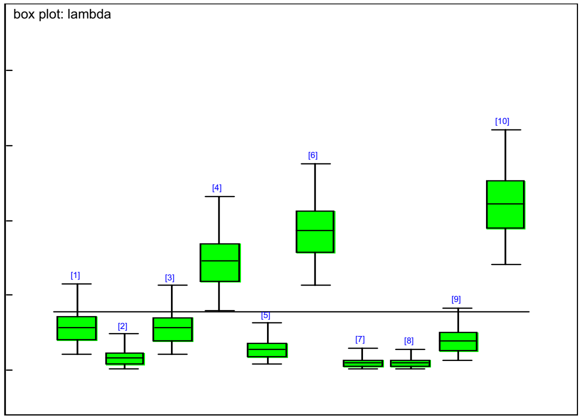
County	E_i	y_i	County	E_i	y_i
1	94	5	6	31	19
2	15	1	7	2	1
3	62	5	8	2	1
4	126	14	9	9	4
5	5	3	10	100	22

Posterior distribution of rate of adoption in each county



node	mean	sd	2.5%	median	97.5%
theta[1]	0.06033	0.02414	0.02127	0.05765	0.1147
theta[2]	0.1074	0.08086	0.009868	0.08742	0.306
theta[3]	0.09093	0.03734	0.03248	0.08618	0.1767
theta[4]	0.1158	0.03085	0.06335	0.1132	0.1831
theta[5]	0.5482	0.2859	0.131	0.4954	1.242
theta[6]	0.6028	0.1342	0.3651	0.5962	0.8932
theta[7]	0.4586	0.3549	0.0482	0.3646	1.343
theta[8]	0.4658	0.3689	0.04501	0.3805	1.41
theta[9]	0.4378	0.2055	0.1352	0.3988	0.9255
theta[10]	0.2238	0.0486	0.1386	0.2211	0.33

Posterior distribution of number of adopters in each county, given exposures



node	mean	sd	2.5%	median	97.5%
lambda[1]	5.671	2.269	2.0	5.419	10.78
lambda[2]	1.611	1.213	0.148	1.311	4.59
lambda[3]	5.638	2.315	2.014	5.343	10.95
lambda[4]	14.59	3.887	7.982	14.26	23.07
lambda[5]	2.741	1.43	0.6551	2.477	6.21
lambda[6]	18.69	4.16	11.32	18.48	27.69
lambda[7]	0.9171	0.7097	0.09641	0.7292	2.687
lambda[8]	0.9317	0.7377	0.09002	0.761	2.819
lambda[9]	3.94	1.85	1.217	3.589	8.33
lambda[10]	22.38	0.1028	13.86	22.11	33.0

node	mean	sd	2.5%	median	97.5%
alpha	0.8336	0.3181	0.3342	0.7989	1.582
beta	2.094	1.123	0.4235	1.897	4.835
mean	0.4772	0.2536	0.1988	0.4227	1.14

Poisson model for small area deaths

Taken from Bayesian Statistical Modeling, Peter Congdon, 2001

- Congdon considers the incidence of heart disease mortality in 758 electoral wards in the Greater London area over three years (1990-92). These small areas are grouped administratively into 33 boroughs.

- Regressors:
  at ward level: x_{ij}, index of socio-economic deprivation.
  at borough level: w_j, where

$$w_i = \begin{cases} 1 & \text{for inner London boroughs} \\ 0 & \text{for outer suburban boroughs} \end{cases}$$

- We assume borough level variation in the intercepts and in the impacts of deprivation; this variation is linked to the category of borough (inner vs outer).

Model

- **First level of the hierarchy**

$$O_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$$

$$\log(\mu_{ij}) = \log(E_{ij}) + \beta_{1j} + \beta_{2j}(x_{ij} - \bar{x}) + \delta_{ij}$$

$$\delta_{ij} \sim N(0, \sigma_\delta^2)$$

  - Death counts O_{ij} are Poisson with means μ_{ij}.
  - $\log(E_{ij})$ is an offset, i.e. an explanatory variable with known coefficient (equal to 1).

  - $\beta_j = (\beta_{1j}, \beta_{2j})'$ are random coefficients for the intercepts and the impacts of deprivation at the borough level.
  - δ_{ij} is a random error for Poisson *over-dispersion.* We fitted two models: one without and another with this random error term.

Model (cont'd)

- **Second level of the hierarchy**

$$\beta_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \end{pmatrix} \sim N_2(\mu_{\beta_j}, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{bmatrix}$$
$$\mu_{\beta_{1j}} = \gamma_{11} + \gamma_{12} w_j$$
$$\mu_{\beta_{2j}} = \gamma_{21} + \gamma_{22} w_j$$

  - γ_{11}, γ_{12}, γ_{21}, and γ_{22} are the **population coefficients** for the intercepts, the impact of borough category, the impact of deprivations, and, respectively, the interaction impact of the level-2 regressors and level-1 regressors.

- **Hyperparameters**

$$\Sigma^{-1} \sim \text{Wishart}\left((\rho R)^{-1}, \rho\right)$$
$$\gamma_{ij} \sim N(0, 0.1) \quad i, j \in \{1, 2\}$$
$$\sigma_\delta^2 \sim \text{Inv-Gamma}(a, b)$$

Computation of E_{ij}

Suppose that p^* is an overall disease rate. Then, $E_{ij} = n_{ij}p^*$ and $\mu_{ij} = p_{ij}/p^*$.

The μ's are said:

1. externally standardized if p^* is obtained from another data source (such as a standard reference table);
2. internally standardized if p^* is obtained from the given dataset, e.g.

$$p^* = \frac{\sum\limits_{ij} O_{ij}}{\sum\limits_{ij} n_{ij}}$$

- In our example we rely on the latter approach.
- Under (a), the joint distribution of the O_{ij} is a product Poisson; under (b) is multinomial.
- However, since likelihood inference is unaffected by whether we condition on $\sum_{ij} O_{ij}$, the product Poisson likelihood is commonly retained.

Model 1: without overdispersion term

(i.e. without δ_{ij})

		std.	Posterior quantiles		
node	mean	dev.	2.5%	median	97.5%
γ_{11}	-0.075	0.074	-0.224	-0.075	0.070
γ_{12}	0.078	0.106	-0.128	0.078	0.290
γ_{21}	0.621	0.138	0.354	0.620	0.896
γ_{22}	0.104	0.197	-0.282	0.103	0.486
σ_{β_1}	0.294	0.038	0.231	0.290	0.380
σ_{β_2}	0.445	0.077	0.318	0.437	0.618
Deviance	945.800	11.320	925.000	945.300	969.400

Model 2: with overdispersion term

		std.	Posterior quantiles		
node	mean	dev.	2.5%	median	97.5%
γ_{11}	-0.069	0.074	-0.218	-0.069	0.076
γ_{12}	0.068	0.106	-0.141	0.068	0.278
γ_{21}	0.616	0.141	0.335	0.615	0.892
γ_{22}	0.105	0.198	-0.276	0.104	0.499
σ_{β_1}	0.292	0.037	0.228	0.289	0.376
σ_{β_2}	0.431	0.075	0.306	0.423	0.600
Deviance	802.400	40.290	726.600	800.900	883.500

- Including ward-level random variability δ_{ij} reduces the average Poisson GLM deviance to 802, with a 95% credible interval from 726 to 883. This is in line with the expected value of the GLM deviance for $N = 758$ areas if the Poisson Model is appropriate.

Hierarchical models for spatial data

Based on the book by Banerjee, Carlin and Gelfand *Hierarchical Modeling and Analysis for Spatial Data*, 2004. We focus on Chapters 1, 2 and 5.

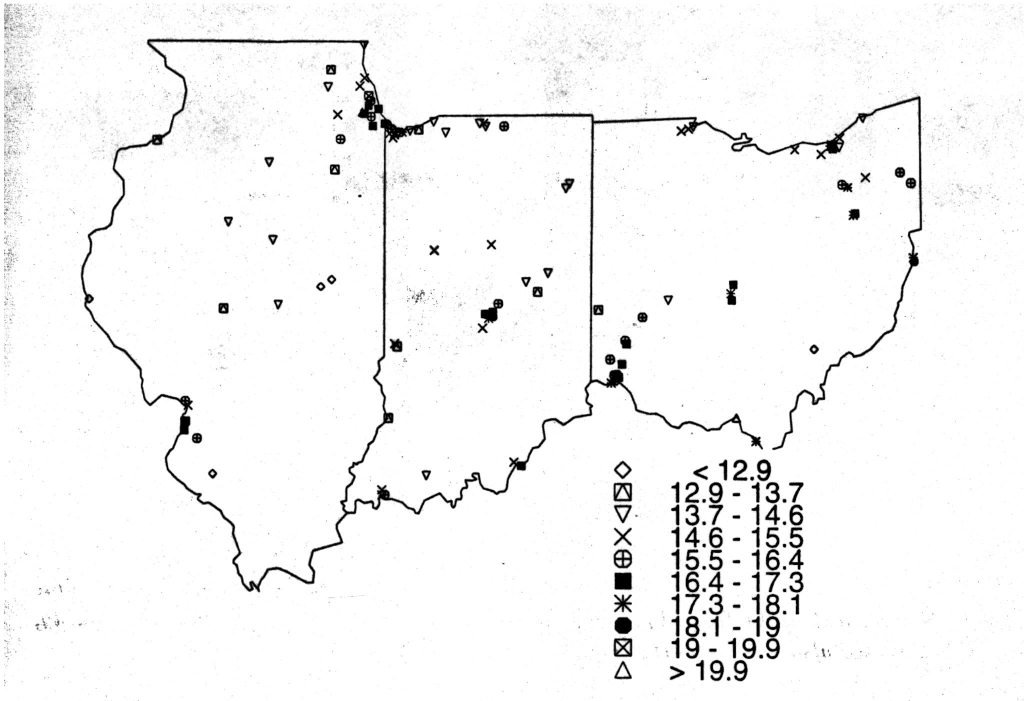
- Geo-referenced data arise in agriculture, climatology, economics, epidemiology, transportation and many other areas.
- What does geo-referenced mean? In a nutshell, we know the geographic location at which an observation was collected.
- Why does it matter? Sometimes, relative location can provide information about an outcome beyond that provided by covariates.
- Example: infant mortality is typically higher in high poverty areas. Even after incorporating poverty as a covariate, residuals may still be spatially correlated due to other factors such as nearness to pollution sites, distance to pre and post-natal care centers, etc.

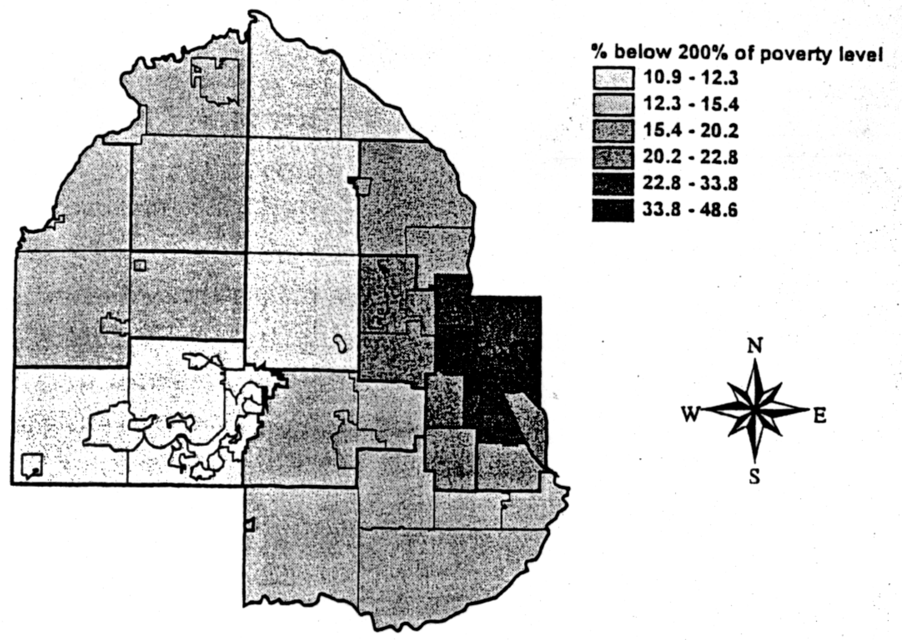
- Often data are also collected over time, so models that include spatial and temporal correlations of outcomes are needed.
- We focus on spatial, rather than spatio-temporal models.
- We also focus on models for univariate rather than multivariate outcomes.

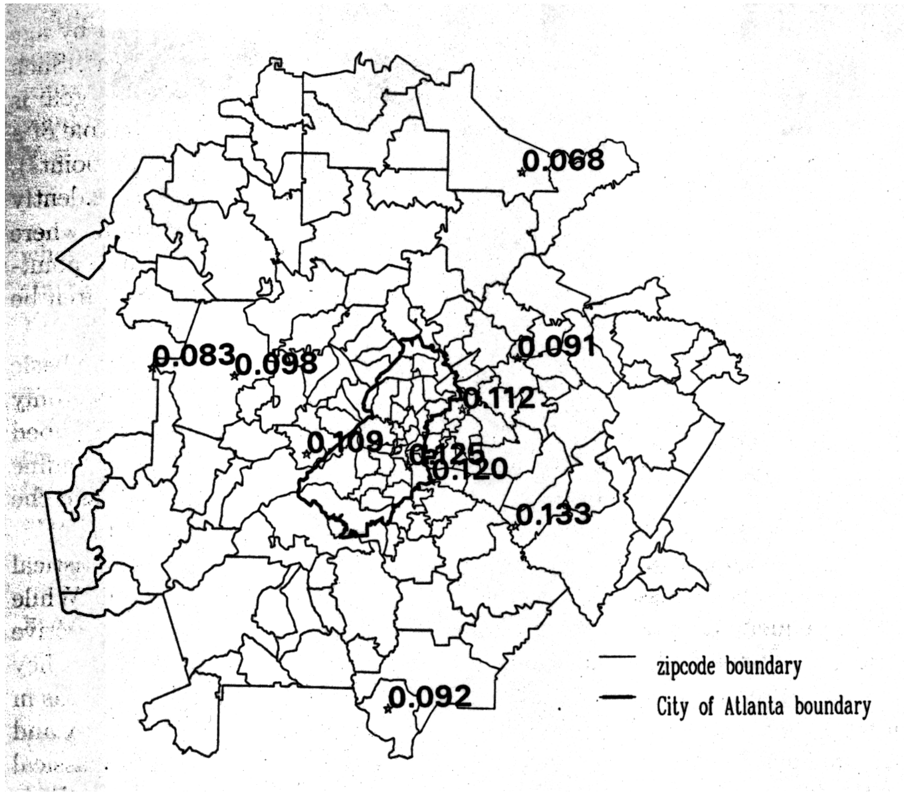
Types of spatial data

- Point-referenced data: $Y(s)$ a random outcome (perhaps vector-valued) at location s, where s varies continuously over some region D. The location s is typically two-dimensional (latitude and longitude) but may also include altitude. Known as geostatistical data.
- Areal data: outcome Y_i is an aggregate value over an areal unit with well-defined boundaries. Here, D is divided into a finite collection of areal units. Known as lattice data even though lattices can be irregular.
- Point-pattern data: Outcome $Y(s)$ is the occurrence or not of an event and locations s are random. Example: locations of trees of a species in a forest or addresseses of persons with a particular disease. Interest is often in deciding whether points occur independently in space or whether there is clustering.

- Marked point process data: If covariate information is available we talk about a marked point process. Covariate value at each site marks the site as belonging to a certain covariate batch or group.
- Combinations: e.g. ozone daily levels collected in monitoring stations with precise location, and number of children in a zip code reporting to the ER on that day. Require data re-alignment to combine outcomes and covariates.







Models for point-level data

The basics

- Location index s varies continuously over region D.
- We often assume that the covariance between two observations at locations s_i and s_j depends only on the distance d_{ij} between the points.
- The spatial covariance is often modeled as exponential:

$$\text{Cov } (Y(s_i), Y(s_{i'})) = C(d_{ii'}) = \sigma^2 e^{-\phi d_{ii'}},$$

  where $(\sigma^2, \phi) > 0$ are the partial sill and decay parameters, respectively.
- Covariogram: a plot of $C(d_{ii'})$ against $d_{ii'}$.

- For $i = i'$, $d_{ii'} = 0$ and $C(d_{ii'}) = var(Y(s_i))$.
- Sometimes, $var(Y(s_i)) = \tau^2 + \sigma^2$, for τ^2 the nugget effect and $\tau^2 + \sigma^2$ the sill.

Models for point-level data (cont'd)

Covariance structure

- Suppose that outcomes are normally distributed and that we choose an exponential model for the covariance matrix. Then:

$$Y|\mu, \theta \sim \text{N}(\mu, \Sigma(\theta)),$$

with

$$\begin{aligned} Y &= \{Y(s_1), Y(s_2), ..., Y(s_n)\} \\ \Sigma(\theta)_{ii'} &= \text{cov}(Y(s_i), Y(s_{i'})) \\ \theta &= (\tau^2, \sigma^2, \phi). \end{aligned}$$

- Then

$$\Sigma(\theta)_{ii'} = \sigma^2 \exp(-\phi d_{ii'}) + \tau^2 I_{i=i'},$$

with $(\tau^2, \sigma^2, \phi) > 0$.

- This is an example of an *isotropic* covariance function: the spatial correlation is only a function of d.

Models for point-level data, details

- Basic model:

$$Y(s) = \mu(s) + w(s) + e(s),$$

where $\mu(s) = x'(s)\beta$ and the residual is divided into two components:

$w(s)$ is a realization of a zero-centered stationary Gaussian process and $e(s)$ is uncorrelated pure error.

- The $w(s)$ are functions of the partial sill σ^2 and decay ϕ parameters.
- The $e(s)$ introduces the nugget effect τ^2.
- τ^2 interpreted as pure sampling variability or as *microscale* variability, i.e., spatial variability at distances smaller than the distance between two outcomes: the $e(s)$ are sometimes viewed as spatial processes with rapid decay.

The variogram and semivariogram

- A spatial process is said to be:
  - Strictly stationary if distributions of $Y(s)$ and $Y(s+h)$ are equal, for h the distance.
  - Weakly stationary if $\mu(s) = \mu$ and $Cov(Y(s), Y(s+h)) = C(h)$.
  - Instrinsically stationary if

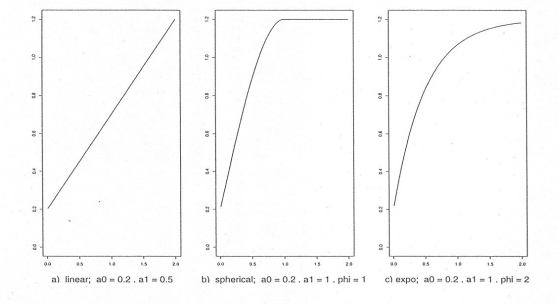
$$\begin{aligned} E[Y(s+h) - Y(s)] &= 0, \text{ and} \\ E[Y(s+h) - Y(s)]^2 &= Var[Y(s+h) - Y(s)] \\ &= 2\gamma(h), \end{aligned}$$

  defined for *differences* and depending only on distance.

- $2\gamma(h)$ is the variogram and $\gamma(h)$ is the *semivariogram*

Examples of semi-variograms

Semi-variograms for the linear, spherical and exponential models.



a) linear; a0 = 0.2 . a1 = 0.5 b) spherical; a0 = 0.2 . a1 = 1 . phi = 1 c) expo; a0 = 0.2 . a1 = 1 . phi = 2

Stationarity

- Strict stationarity implies weak stationarity but the converse is not true except in Gaussian processes.
- Weak stationarity implies intrinsec stationarity, but the converse is not true in general.
- Notice that intrinsec stationarity is defined on the differences between outcomes at two locations and thus says nothing about the joint distribution of outcomes.

Semivariogram (cont'd)

- If $\gamma(h)$ depends on h only through its length $||h||$, then the spatial process is *isotropic*. Else it is *anisotropic*.
- There are many choices for isotropic models. The *exponential* model is popular and has good properties. For $t = ||h||$:

$$\begin{aligned} \gamma(t) &= \tau^2 + \sigma^2(1 - \exp(-\phi t)) \text{ if } t > 0, \\ &= 0 \text{ otherwise.} \end{aligned}$$

- See figures, page 24.
- The *powered* exponential model has an extra parameter for smoothness:

$$\gamma(t) = \tau^2 + \sigma^2(1 - \exp(-\phi t^\kappa)) \text{ if } t > 0$$

- Another popular choice is the Gaussian variogram model, equal to the exponential except for the exponent term, that is $\exp(-\phi^2 t^2)$).
- Fitting of the variogram has been traditionally done ”by eye”:
  - Plot an empirical estimate of the variogram akin to the sample variance estimate or the autocorrelation function in time series
  - Choose a theoretical functional form to fit the empirical γ
  - Choose values for (τ^2, σ^2, ϕ) that fit the data well.
- If a distribution for the outcomes is assumed and a functional form for the variogram is chosen, parameter estimates can be estimated via some likelihood-based method.
- Of course, we can also be Bayesians.

Point-level data (cont’d)

- For point-referenced data, frequentists focus on spatial prediction using *kriging*.
- Problem: given observations $\{Y(s_1), ..., Y(s_n)\}$, how do we predict $Y(s_o)$ at a new site s_o?
- Consider the model

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim \text{ N}(0, \Sigma),$$

and where

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I.$$

Here, $H(\phi)_{ii'} = \rho(\phi, d_{ii'})$.

- Kriging consists in finding a function $f(y)$ of the observations that minimizes the MSE of prediction

$$Q = E[(Y(s_o) - f(y))^2 | y].$$

Classical kriging (cont’d)

- (Not a surprising!) Result: $f(y)$ that minimizes Q is the conditional mean of $Y(s_0)$ given observations y (see pages 50-52 for proof):

$$\begin{aligned} E[Y(s_o)|y] &= x_o'\hat{\beta} + \hat{\gamma}'\hat{\Sigma}^{-1}(y - X\hat{\beta}) \\ Var[Y(s_o)|y] &= \hat{\sigma}^2 + \hat{\tau}^2 - \hat{\gamma}'\hat{\Sigma}^{-1}\hat{\gamma}, \end{aligned}$$

where

$$\begin{aligned} \hat{\gamma} &= (\hat{\sigma}^2 \rho(\hat{\phi}, d_{o1}), ..., \hat{\sigma}^2 \rho(\hat{\phi}, d_{on})) \\ \hat{\beta} &= (X'\hat{\Sigma}^{-1}X)^{-1} X'\hat{\Sigma}^{-1} y \\ \hat{\Sigma} &= \hat{\sigma}^2 H(\hat{\phi}). \end{aligned}$$

- Solution assumes that we have observed the covariates x_o at the new site.

- If not,in the classical framework $Y(s_o), x_o$ are jointly estimated using an EM-type iterative algorithm.

Bayesian methods for estimation

- The Gaussian isotropic kriging model is just a general linear model similar to those in Chapter 15 of textbook.
- Just need to define the appropriate covariance structure.
- For an exponential covariance structure with a nugget effect, parameters to be estimated are $\theta = (\beta, \sigma^2, \tau^2, \phi)$.
- Steps:
  - Choose priors and define sampling distribution
  - Obtain posterior for all parameters $p(\theta|y)$
  - Bayesian kriging: get posterior predictive distribution for outcome at new location $p(y_o|y, X, x_o)$.

- Sampling distribution (marginal data model)

$$y|\theta \sim \text{ N}(X\beta, \sigma^2 H(\phi) + \tau^2 I)$$

- Priors: typically chosen so that parameters are independent a priori.
- As in the linear model:
  - Non-informative prior for β is uniform or can use a normal prior too.
  - Conjugate priors for variances σ^2, τ^2 are inverse gamma priors.
- For ϕ, appropriate prior depends on covariance model. For simple exponential where

$$\rho(s_i - s_j; \phi) = \exp(-\phi||s_i - s_j||),$$

a Gamma prior can be a good choice.

- Be cautious with improper priors for anything but β.

Hierarchical representation of model

- Hierarchical model representation: first condition on the spatial random effects $W = \{w(s_1), ..., w(s_n)\}$:

$$\begin{aligned} y|\theta, W &\sim \mathsf{N}(X\beta + W, \tau^2 I) \\ W|\phi, \sigma^2 &\sim \mathsf{N}(0, \sigma^2 H(\phi)). \end{aligned}$$

- Model specification is then completed by choosing priors for β, τ^2 and for ϕ, σ^2 (hyperparameters).

- Note that hierarchical model has n more parameters (the $w(s_i)$) than the marginal model.

- Computation with the marginal model preferable because $\sigma^2 H(\phi) + \tau^2 I$ tends to be better behaved than $\sigma^2 H(\phi)$ at small distances.

Estimation of spatial surface $W|y$

- Interest is sometimes on estimating the spatial surface using $p(W|y)$.

- If marginal model is fitted, we can still get marginal posterior for W as

$$p(W|y) = \int p(W|\sigma^2, \phi) p(\sigma^2, \phi|y) d\sigma^2 d\phi.$$

- Given draws $(\sigma^{2(g)}, \phi^{(g)})$ from the Gibbs sampler on the marginal model, we can generate W from

$$p(W|\sigma^{2(g)}, \phi^{(g)}) = \mathsf{N}(0, \sigma^{2(g)} H(\phi^{(g)})).$$

- Analytical marginalization over W is possible only if model has Gaussian form.

Bayesian kriging

- Let $Y_o = Y(s_o)$ and $x_o = x(s_o)$. Kriging is accomplished by obtaining the posterior predictive distribution

$$\begin{aligned} p(y_o|x_o, X, y) &= \int p(y_o, \theta|y, X, x_o) d\theta \\ &= \int p(y_o|\theta, y, x_o) p(\theta|y, X) d\theta. \end{aligned}$$

- Since (Y_o, Y) are jointly multivariate normal (see expressions 2.18 and 2.19 on page 51), then $p(y_o|\theta, y, x_o)$ is a conditional normal distribution.

- Given MCMC draws of the parameters $(\theta^{(1)}, ..., \theta^{(G)})$ from the posterior distribution $p(\theta|y, X)$, we draw values $y_o^{(g)}$ for each $\theta^{(g)}$ as

$$y_o^{(g)} \sim p(y_o|\theta^{(g)}, y, x_o).$$

- Draws $\{y_o^{(1)}, y_o^{(2)}, ..., y_o^{(G)}, \}$ are a sample from the posterior predictive distribution of the outcome at the new location s_o.

- To predict Y at a set of m new locations $s_{o1}, ..., s_{om}$, it is best to do joint prediction to be able to estimate the posterior association among m predictions.

- Beware of joint prediction at many new locations with WinBUGS. It can take forever.

Kriging example from WinBugs

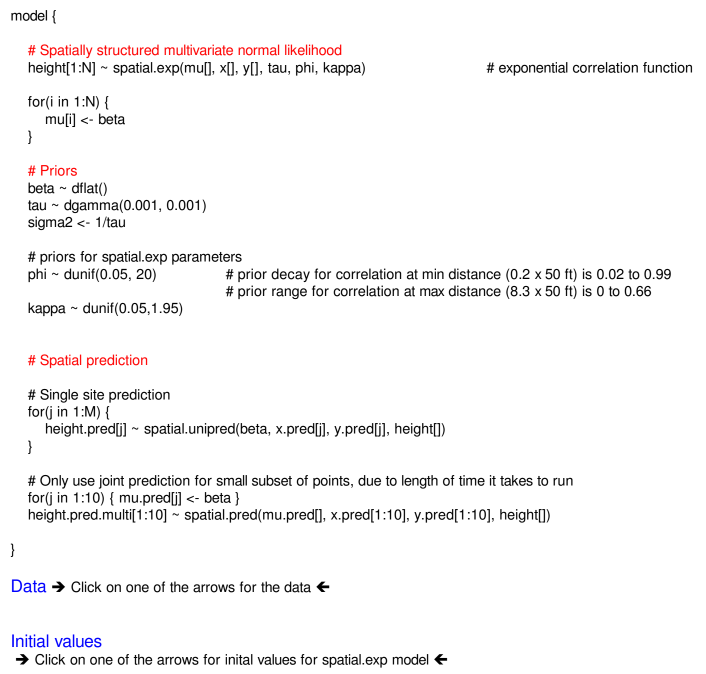
- Data were first published by Davis (1973) and consist of heights at 52 locations in a 310-foot square area.
- We have 52 $s = (x, y)$ coordinates and outcomes (heights).
- Unit of distance is 50 feet and unit of elevation is 10 feet.
- The model is

$$\text{height} = \beta + \epsilon, \text{ where } \epsilon \sim \text{N}(0, \Sigma),$$

and where

$$\Sigma = \sigma^2 H(\phi).$$

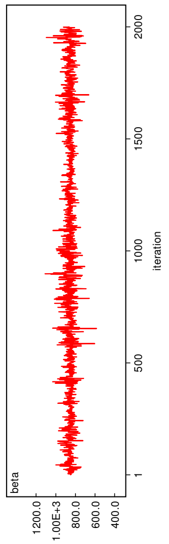
- Here, $H(\phi)_{ij} = \rho(s_i - s_j; \phi) = \exp(-\phi||s_i - s_j||^{\kappa})$.
- Priors on (β, ϕ, κ).
- We predict elevations at 225 new locations.

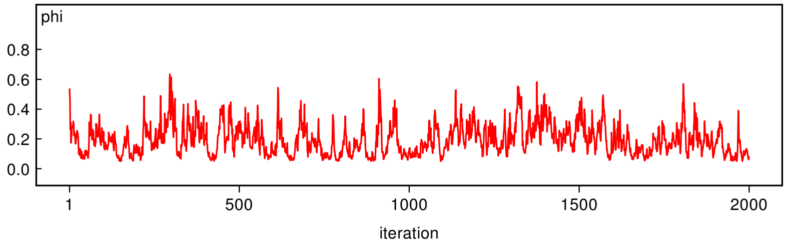


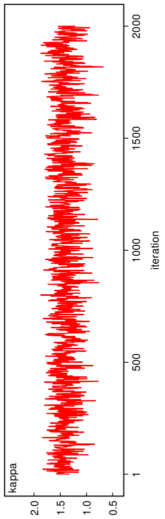
```
model {

   # Spatially structured multivariate normal likelihood
   height[1:N] ~ spatial.exp(mu[], x[], y[], tau, phi, kappa)                    # exponential correlation function

   for(i in 1:N) {
      mu[i] <- beta
   }

   # Priors
   beta ~ dflat()
   tau ~ dgamma(0.001, 0.001)
   sigma2 <- 1/tau

   # priors for spatial.exp parameters
   phi ~ dunif(0.05, 20)             # prior decay for correlation at min distance (0.2 x 50 ft) is 0.02 to 0.99
                                     # prior range for correlation at max distance (8.3 x 50 ft) is 0 to 0.66
   kappa ~ dunif(0.05,1.95)


   # Spatial prediction

   # Single site prediction
   for(j in 1:M) {
      height.pred[j] ~ spatial.unipred(beta, x.pred[j], y.pred[j], height[])
   }

   # Only use joint prediction for small subset of points, due to length of time it takes to run
   for(j in 1:10) { mu.pred[j] <- beta }
   height.pred.multi[1:10] ~ spatial.pred(mu.pred[], x.pred[1:10], y.pred[1:10], height[])

}
```

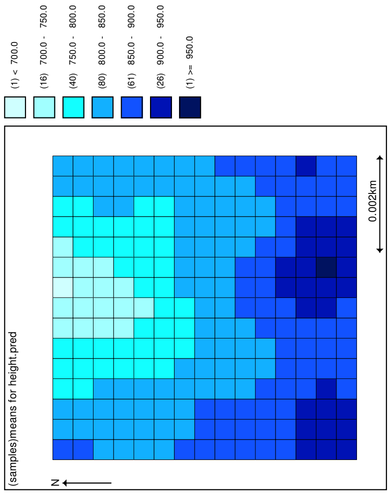
Data ➔ Click on one of the arrows for the data 🡸

Initial values
➔ Click on one of the arrows for inital values for spatial.exp model 🡸









Hierarchical models for areal data

- Areal data: often aggregate outcomes over a well-defined area. Example: number of cancer cases per county or proportion of people living in poverty in a set of census tracks.

- What are the inferential issues?

1. Identifying a spatial pattern and its strength. If data are spatially correlated, measurements from areas that are 'close' will be more alike.
2. Smoothing and to what degree. Observed measurements often present extreme values due to small samples in small areas. Maximal smoothing: substitute observed measurements by the overall mean in the region. Something less extreme is what we discuss later.
3. Prediction: for a new area, how would we predict Y given measurements in other areas?

Defining neighbors

- A *proximity matrix* W with entries w_{ij} spatially connects areas i and j in some fashion.

- Typically, $w_{ii} = 0$.

- There are many choices for the w_{ij}:
  - Binary: $w_{ij} = 1$ if areas i, j share a common boundary, and is 0 otherwise.
  - Continuous: decreasing function of intercentroidal distance.
  - Combo: $w_{ij} = 1$ if areas are within a certain distance.

- W need not be symmetric.

- Entries are often standardized by dividing into $\sum_j w_{ij} = w_{i+}$. If entries are standardized, the W will often be asymmetric.

- The w_{ij} can be thought of as weights and provide a means to introduce spatial structure into the model.

- Areas that are 'closer by' in some sense are more alike.

- For any problem, we can define first, second, third, etc order neighbors. For distance bins $(0, d_1], (d_1, d_2], (d_2, d_3], ...$ we can define

  1. $W^{(1)}$, the first-order proximity matrix with $w_{ij}^{(1)} = 1$ if distance between i abd j is less than d_1.
  2. $W^{(2)}$, the second-order proximity matrix with $w_{ij}^{(2)} = 1$ if distance between i abd j is more than d_1 but less than d_2.

Areal data models

- Because these models are used mostly in epidemiology, we begin with an application in *disease mapping* to introduce concepts.

- Typical data:

  Y_i observed number of cases in area i, $i = 1, ..., I$

  E_i expected number of cases in area i.

- The Y's are assumed to be random and the Es are assumed to be known and to depend on the number of persons n_i at risk.

- An *internal* standardized estimate of E_i is

$$E_i = n_i \bar{r} = n_i \left(\frac{\sum_i y_i}{\sum_i n_i} \right),$$

corresponding to a constant disease rate across areas.

- An *external* standardized estimate is

$$E_i = \sum_j n_{ij} r_j,$$

where r_j is the risk for persons of age group j (from some existing table of risks by age) and n_{ij} is the number of persons of age j in area i.

Standard frequentist approach

- For small E_i,
$$Y_i|\eta_i \sim \text{ Poisson}(E_i\eta_i),$$
with η_i the true relative risk in area i.

- The MLE is the *standard mortality ratio*
$$\hat{\eta}_i = SMR_i = \frac{Y_i}{E_i}.$$

- The variance of the SMR_i is
$$var(SMR_i) = \frac{var(Y_i)}{E_i^2} = \frac{\eta_i}{E_i},$$

estimated by plugging $\hat{\eta}_i$ to obtain
$$var(SMR_i) = \frac{Y_i}{E_i^2}.$$

- To get a confidence interval for η_i, first assume that $\log(SMR_i)$ is approximately normal.

- From a Taylor expansion:
$$\begin{aligned} Var[\log(SMR_i)] &\approx \frac{1}{SMR_i^2} Var(SMR_i) \\ &= \frac{E_i^2}{Y_i^2} \times \frac{Y_i}{E_i^2} = \frac{1}{Y_i}. \end{aligned}$$

- An approximate 95% CI for $\log(\eta_i)$ is
$$SMR_i \pm 1.96/(Y_i)^{1/2}.$$

- Transforming back, an approximate 95% CI for η_i is
$$(SMR_i \exp(-1.96/(Y_i)^{1/2}),\ \ SMR_i \exp(1.96/(Y_i)^{1/2})).$$

- Suppose we wish to test whether risk in area i is high relative to other areas. Then test
$$H_0 : \eta_i = 1 \text{ versus } H_a : \eta_i > 1.$$

- This is a one-sided test.

- Under H_0, $Y_i \sim$ Poisson(E_i) so the $p-$value for the test is
$$\begin{aligned} p &= \text{ Prob}(X \geq Y_i|E_i) \\ &= 1 - \text{ Prob}(X \geq Y_i|E_i) \\ &= 1 - \sum_{x=0}^{Y_i-1} \frac{\exp(-E_i)E_i^x}{x!}. \end{aligned}$$

- If $p < 0.05$ we reject H_0.

Hierarchical models for areal data

- To estimate and map underlying relative risks, might wish to fit a random effects model.
- Assumption: true risks come from a common underlying distribution.
- Random effects models permit *borrowing strength* across areas to obtain better area-level estimates.
- Alas, models may be complex:
  - High-dimensional: one random effect for each area.
  - Non-normal if data are counts or binomial proportions.
- We have already discussed hierarchical Poisson models, so material in the next few transparencies is a review.

Poisson-Gamma model

- Consider

$$\begin{aligned} Y_i|\eta_i &\sim \text{Poisson}(E_i\eta_i) \\ \eta_i|a,b &\sim \text{Gamma}(a,b). \end{aligned}$$

- Since $E(\eta_i) = a/b$ and $Var(\eta_i) = a/b^2$, we can fix a, b as follows:
  - A priori, let $E(\eta_i) = 1$, the *null* value.
  - Let $Var(\eta_i) = (0.5)^2$, large on that scale.
- Resulting prior is Gamma$(4, 4)$.
- Posterior is also Gamma:

$$p(\eta_i|y_i) = \text{Gamma}(y_i + a, E_i + b).$$

- A point estimate of η_i is

$$\begin{aligned} E(\eta_i|y) = E(\eta_i|y_i) &= \frac{y_i + a}{E_i + b} \\ &= \frac{y_i + \frac{[E(\eta_i)]^2}{Var(\eta_i)}}{E_i + \frac{E(\eta_i)}{Var(\eta_i)}} \\ &= \frac{E_i(\frac{y_i}{E_i})}{E_i + \frac{E(\eta_i)}{Var(\eta_i)}} + \frac{\frac{E(\eta_i)}{Var(\eta_i)}E(\eta_i)}{E_i + \frac{E(\eta_i)}{Var(\eta_i)}} \\ &= w_i SMR_i + (1 - w_i)E(\eta_i), \end{aligned}$$

  where $w_i = E_i/[E_i + (E(\eta_i)/Var(\eta_i))]$.
- Bayesian point estimate is a weighted average of the data-based SMR_i and the prior mean $E(\eta_i)$.

Poisson-lognormal models with spatial errors

- The Poisson-Gamma model does not allow (easily) for spatial correlation among the η_i.
- Instead, consider the Poisson-lognormal model, where in the second stage we model the log-relative risks $\log(\eta_i) = \psi_i$:

$$\begin{aligned} Y_i|\psi_i &\sim \text{Poisson}(E_i \exp(\psi_i)) \\ \psi_i &= x_i'\beta + \theta_i + \phi_i, \end{aligned}$$

  where x_i are area-level covariates.
- The θ_i are assumed to be exchangeable and model between-area variability:

$$\theta_i \sim \text{N}(0, 1/\tau_h).$$

- The θ_i incorporate *global* extra-Poisson variability in the log-relative risks (across the entire region).
- The ϕ_i are the 'spatial' parameters; they capture regional *clustering*.
- They model extra-Poisson variability in the log-relative risks at the *local* level so that 'neighboring' areas have similar risks.
- One way to model the ϕ_i is to proceed as in the point-referenced data case. For $\phi = (\phi_1, ..., \phi_I)$, consider

$$\phi|\mu, \lambda \sim \text{ N}_I(\mu, H(\lambda)),$$

and $H(\lambda)_{ii'} = cov(\phi_i, \phi_{i'})$ with hyperparameters λ.

- Possible models for $H(\lambda)$ include the exponential, the powered exponential, etc.

- While sensible, this model is difficult to fit because
  - Lots of matrix inversions required
  - Distance between ϕ_i and $\phi_{i'}$ may not be obvious.

CAR model

- More reasonable to think of a neighbor-based proximity measure and consider a *conditionally autoregressive* model for ϕ:

$$\phi_i \sim \text{ N}(\bar{\phi}_i, 1/(\tau_c m_i)),$$

where

$$\bar{\phi}_i = \sum_{i \neq j} w_{ij}(\phi_i - \phi_j),$$

and m_i is the number of neighbors of area i. Earlier we called this w_{i+}.

- The weights w_{ij} are (typically) 0 if areas i and j are *not* neighbors and 1 if they are.
- CAR models lend themselves to the Gibbs sampler. Each ϕ_i can be sampled from its conditional distribution so no matrix inversion is

needed:

$$p(\phi_i|\text{ all}) \propto \text{ Poi}(y_i|E_i e^{x_i\beta+\theta_i+\phi_i}) \text{ N}(\phi_i|\bar{\phi}_i, \frac{1}{m_i\tau_c}).$$

Difficulties with CAR model

- The CAR prior is improper. Prior is a pairwise difference prior identified only up to a constant.
- The posterior will still be proper, but to identify an intercept β_0 for the log-relative risks, we need to impose a constraint: $\sum_i \phi_i = 0$.
- In simulation, constraint is imposed numerically by recentering each vector ϕ around its own mean.
- τ_h and τ_c cannot be too large because θ_i and ϕ_i become $unidentifiable$. We observe only one Y_i in each area yet we try to fit two random effects. Very little data information.
- Hyperpriors for τ_h, τ_c need to be chosen carefully.

- Consider

$$\tau_h \sim \text{Gamma}(a_h, b_h), \quad \tau_c \sim \text{Gamma}(a_c, b_c).$$

- To place equal emphasis on heterogeneity and spatial clustering, it is tempting to make $a_h = a_c$ and $b_h = b_c$. This is not correct because

1. The τ_h prior is defined marginally, where the τ_c prior is conditional.
2. The conditional prior precision is $\tau_c m_i$. Thus, a scale that satisfies

$$sd(\theta_i) = \frac{1}{\sqrt{\tau_h)}} \approx \frac{1}{0.7\sqrt{\bar{m}\tau_c}} \approx sd(\phi_i)$$

with $\bar{m}$ the average number of neighbors is more 'fair' (Bernardinelli et al. 1995, *Statistics in Medicine*).

Example: Incidence of lip cancer in 56 areas in Scotland

Data on the number of lip cancer cases in 56 counties in Scotland were obtained. Expected lip cancer counts E_i were available.

Covariate was the proportion of the population in each county working in agriculture.

Model included only one random effect b_i to introduce spatial association between counties.

Two counties are neighbors if they have a common border (i.e., they are adjacent).

Three counties that are islands have no neighbors and for those, WinBUGS sets the random spatial effect to be 0. The relative risk for the islands is thus based on the baseline rate α_0 and on the value of the covariate x_i.

We wish to smooth and map the relative risks RR.

Model

```
model {
    # Likelihood
    for (i in 1 : N) {
        O[i]  ~ dpois(mu[i])
        log(mu[i]) <- log(E[i]) + alpha0 + alpha1 * X[i]/10 + b[i]
        RR[i] <- exp(alpha0 + alpha1 * X[i]/10 + b[i])                # Area-specific
relative risk (for maps)
    }

    # CAR prior distribution for random effects:
    b[1:N] ~ car.normal(adj[], weights[], num[], tau)
    for(k in 1:sumNumNeigh) {
        weights[k] <- 1
    }

    # Other priors:
    alpha0  ~ dflat()
    alpha1 ~ dnorm(0.0, 1.0E-5)
    tau  ~ dgamma(0.5, 0.0005)              # prior on precision
    sigma <- sqrt(1 / tau)              # standard deviation
}
```

Data

```
list(N = 56,
O   = c(    9,   39,   11,    9,   15,    8,   26,    7,    6,   20,
           13,    5,    3,    8,   17,    9,    2,    7,    9,    7,
           16,   31,   11,    7,   19,   15,    7,   10,   16,   11,
            5,    3,    7,    8,   11,    9,   11,    8,    6,    4,
           10,    8,    2,    6,   19,    3,    2,    3,   28,    6,
            1,    1,    1,    1,    0,    0),
E = c( 1.4, 8.7, 3.0, 2.5, 4.3, 2.4, 8.1, 2.3, 2.0, 6.6,
        4.4, 1.8, 1.1, 3.3, 7.8, 4.6, 1.1, 4.2, 5.5, 4.4,
       10.5,22.7, 8.8, 5.6,15.5,12.5, 6.0, 9.0,14.4,10.2,
        4.8, 2.9, 7.0, 8.5,12.3,10.1,12.7, 9.4, 7.2, 5.3,
       18.8,15.8, 4.3,14.6,50.7, 8.2, 5.6, 9.3,88.7,19.6,
        3.4, 3.6, 5.7, 7.0, 4.2, 1.8),
X = c(16,16,10,24,10,24,10, 7, 7,16,
         7,16,10,24, 7,16,10, 7, 7,10,
         7,16,10, 7, 1, 1, 7, 7,10,10,
         7,24,10, 7, 7, 0,10, 1,16, 0,
         1,16,16, 0, 1, 7, 1, 1, 0, 1,
         1, 0, 1, 1,16,10),
num = c(3, 2, 1, 3, 3, 0, 5, 0, 5, 4,
0, 2, 3, 3, 2, 6, 6, 6, 5, 3,
3, 2, 4, 8, 3, 3, 4, 4, 11, 6,
7, 3, 4, 9, 4, 2, 4, 6, 3, 4,
5, 5, 4, 5, 4, 6, 6, 4, 9, 2,
```

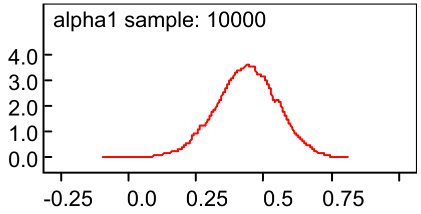
```
4, 4, 4, 5, 6, 5),
adj = c(
19, 9, 5,
10, 7,
12,
28, 20, 18,
19, 12, 1,
17, 16, 13, 10, 2,
29, 23, 19, 17, 1,
22, 16, 7, 2,
5, 3,
19, 17, 7,
35, 32, 31,
29, 25,
29, 22, 21, 17, 10, 7,
29, 19, 16, 13, 9, 7,
56, 55, 33, 28, 20, 4,
17, 13, 9, 5, 1,
56, 18, 4,
50, 29, 16,
16, 10,
39, 34, 29, 9,
56, 55, 48, 47, 44, 31, 30, 27,
29, 26, 15,
43, 29, 25,
56, 32, 31, 24,
```

```
45, 33, 18, 4,
50, 43, 34, 26, 25, 23, 21, 17, 16, 15, 9,
55, 45, 44, 42, 38, 24,
47, 46, 35, 32, 27, 24, 14,
31, 27, 14,
55, 45, 28, 18,
54, 52, 51, 43, 42, 40, 39, 29, 23,
46, 37, 31, 14,
41, 37,
46, 41, 36, 35,
54, 51, 49, 44, 42, 30,
40, 34, 23,
52, 49, 39, 34,
53, 49, 46, 37, 36,
51, 43, 38, 34, 30,
42, 34, 29, 26,
49, 48, 38, 30, 24,
55, 33, 30, 28,
53, 47, 41, 37, 35, 31,
53, 49, 48, 46, 31, 24,
49, 47, 44, 24,
54, 53, 52, 48, 47, 44, 41, 40, 38,
29, 21,
54, 42, 38, 34,
54, 49, 40, 34,
49, 47, 46, 41,
```

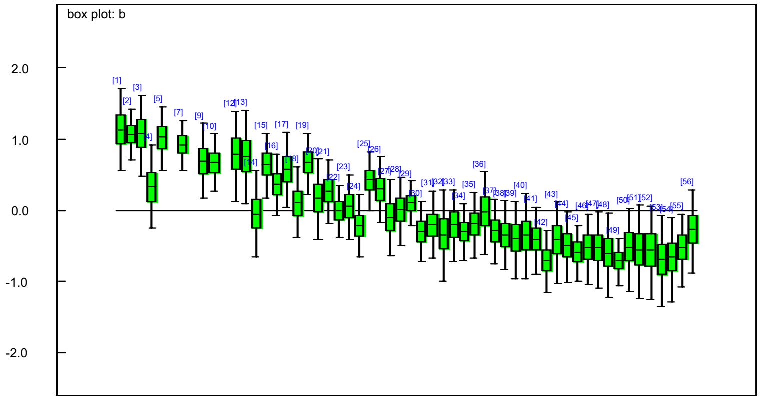
```
52, 51, 49, 38, 34,
56, 45, 33, 30, 24, 18,
55, 27, 24, 20, 18),
sumNumNeigh = 234)
```

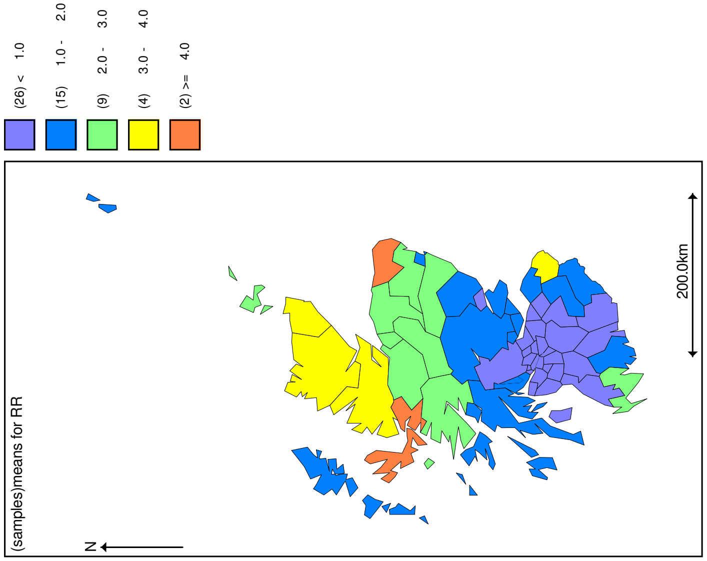
Results

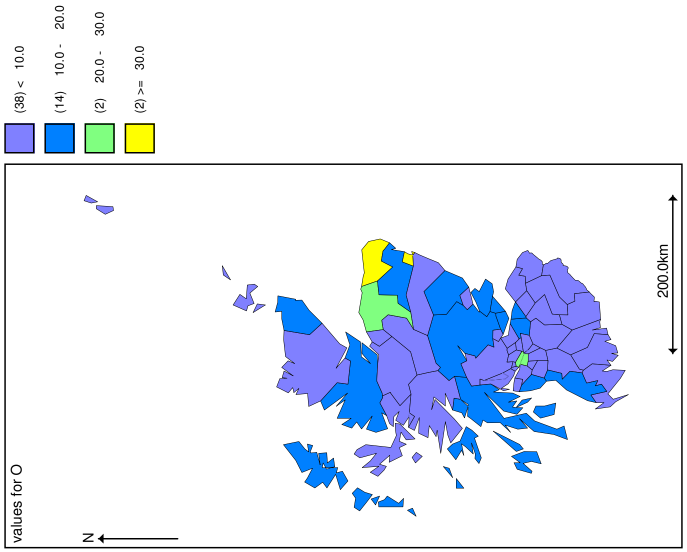
The proportion of individuals working in agriculture appears to be associated to the incidence of cancer



There appears to be spatial correlation among counties:







Example: Lung cancer in London

Data were obtained from an annual report by the London Health Authority in which the association between health and poverty was investigated.

Population under study are men 65 years of age and older.

Available were:

- Observed and expected lung cancer counts in 44 census ward districts
- Ward-level index of socio-economic deprivation.

A model with two random effects was fitted to the data. One random effect introduces region-wide heterogeneity (between ward variability) and the other one introduces regional clustering.

The priors for the two components of random variability, are sometimes known as *convolution priors*.

Interest is in:

- Determining association between health outcomes and poverty.
- Smoothing and mapping relative risks.
- Assessing the degree of spatial clustering in these data.

Model

```
model {

  for (i in 1 : N) {
    # Likelihood
    O[i]  ~ dpois(mu[i])
    log(mu[i]) <- log(E[i]) + alpha + beta * depriv[i] + b[i] + h[i]
    RR[i] <- exp(alpha + beta * depriv[i] + b[i] + h[i])                    # Area-specific
relative risk (for maps)

    # Exchangeable prior on unstructured random effects
    h[i] ~ dnorm(0, tau.h)
  }

  # CAR prior distribution for spatial random effects:
  b[1:N] ~ car.normal(adj[], weights[], num[], tau.b)
  for(k in 1:sumNumNeigh) {
    weights[k] <- 1
  }

  # Other priors:
  alpha  ~ dflat()
  beta ~ dnorm(0.0, 1.0E-5)
  tau.b  ~ dgamma(0.5, 0.0005)
```

```
  sigma.b <- sqrt(1 / tau.b)
  tau.h  ~ dgamma(0.5, 0.0005)
  sigma.h <- sqrt(1 / tau.h)
  propstd <- sigma.b / (sigma.b + sigma.h)

}
```

Data   click on one of the arrows to open data

Inits   click on one of the arrows to open initial values

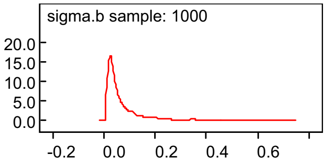
Note that the priors on the precisions for the exchangeable and the spatial random effects are Gamma(0.5, 0.0005).

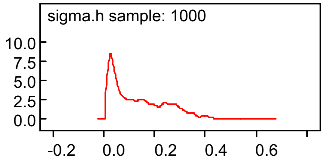
That means that a priori, the expected value of the standard deviations is approximately 0.03 with a relatively large prior standard deviation.

This is not a “fair” prior as discussed in class. The average number of neighbors is 4.8 and this is not taken into account in the choice of priors.

Results

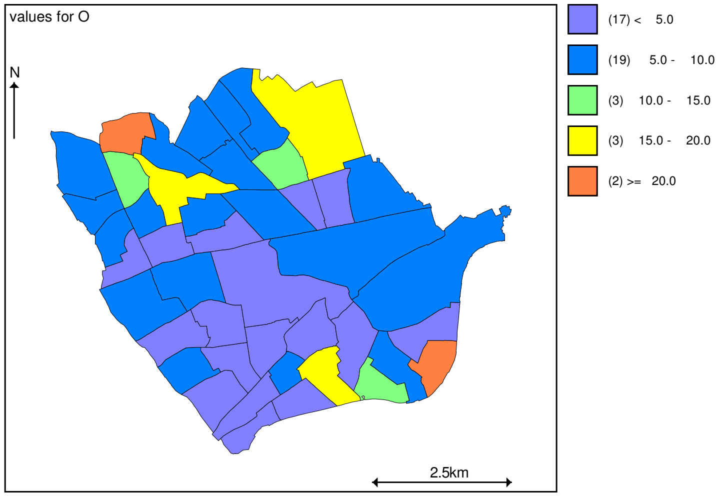
Parameter	Mean	Std	2.5th perc.	97.5th perc.
Alpha	-0.208	0.1	-0.408	-0.024
Beta	0.0474	0.0179	0.0133	0.0838
Relative size of spatial std	0.358	0.243	0.052	0.874

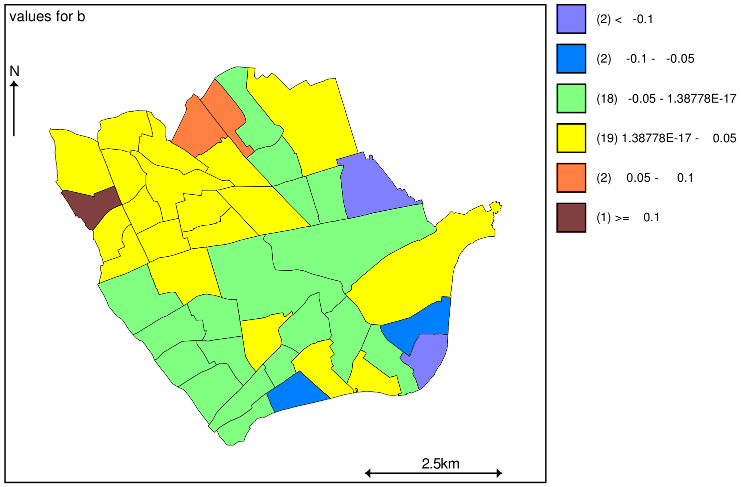


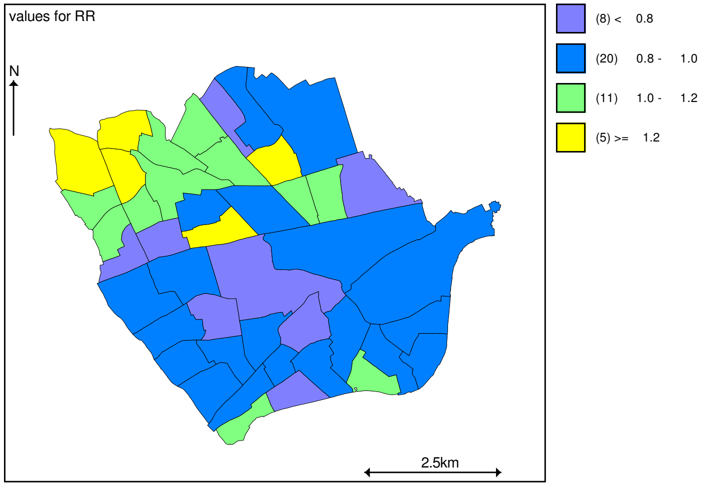


Posterior distribution of RR for first 15 wards

node	mean	sd	2.5%	median	97.5%
RR[1]	0.828	0.1539	0.51	0.8286	1.148
RR[2]	1.18	0.2061	0.7593	1.188	1.6
RR[3]	0.8641	0.1695	0.5621	0.8508	1.247
RR[4]	0.8024	0.1522	0.5239	0.7873	1.154
RR[5]	0.7116	0.1519	0.379	0.7206	0.9914
RR[6]	1.05	0.2171	0.7149	1.015	1.621
RR[7]	1.122	0.1955	0.7556	1.116	1.589
RR[8]	0.821	0.1581	0.4911	0.8284	1.132
RR[9]	1.112	0.2167	0.7951	1.07	1.702
RR[10]	1.546	0.2823	1.072	1.505	2.146
RR[11]	0.7697	0.1425	0.4859	0.7788	1.066
RR[12]	0.865	0.16	0.6027	0.8464	1.26
RR[13]	1.237	0.3539	0.8743	1.117	2.172
RR[14]	0.8359	0.1807	0.5279	0.8084	1.3
RR[15]	0.7876	0.1563	0.489	0.7869	1.141


values for O
N
2.5km
(17) < 5.0
(19) 5.0 - 10.0
(3) 10.0 - 15.0
(3) 15.0 - 20.0
(2) >= 20.0


values for b
N
2.5km
(2) < -0.1
(2) -0.1 - -0.05
(18) -0.05 - 1.38778E-17
(19) 1.38778E-17 - 0.05
(2) 0.05 - 0.1
(1) >= 0.1


values for RR
N
2.5km
(8) < 0.8
(20) 0.8 - 1.0
(11) 1.0 - 1.2
(5) >= 1.2