An introduction to Bayesian modeling using R and JAGS

Instructors
Kent Holsinger
Xiaojing Wang

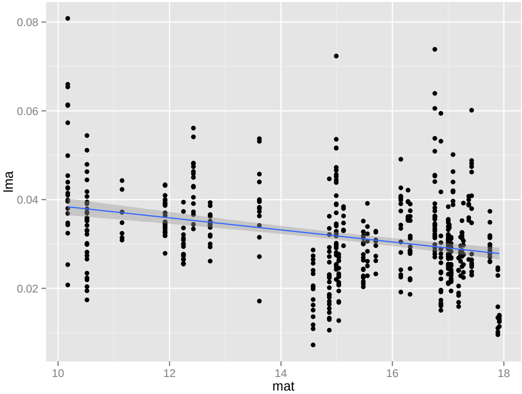
University of Connecticut

July 3, 2015

Linear Regression

One of the most common statistical procedures in ecology and evolution. For example,

- Data on LMA from 535 individuals in the genus *Protea* (42 species, 48 sites, 142 unique site/species combinations)
- Data on mean annual temperature for each of those sites



Linear Regression

In R

```
> summary(lm(lma ˜ mat, data=tmp))

Call:
lm(formula = lma ˜ mat, data = tmp)

Residuals:
      Min        1Q    Median        3Q       Max
-0.025126 -0.005781 -0.000785  0.004647  0.044444

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0521895  0.0027116  19.246  < 2e-16 ***
mat         -0.0013587  0.0001785  -7.611 1.24e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.009815 on 533 degrees of freedom
Multiple R-squared:  0.09803, Adjusted R-squared:  0.09634
F-statistic: 57.93 on 1 and 533 DF,  p-value: 1.241e-13
```

Linear Regression

Remember basic assumptions of simple linear regression

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ \epsilon_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

Here's another way to write that

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

The second way of writing the model will be more convenient for us, so that's the approach we'll use.

Statistical Analysis

Statistical inference is the process of learning about the general characteristics of a population from a sample.

- Characteristics often expressed in terms of parameters θ.
- Measurements on the subset of members given by numerical values Y.
- Before the data are observed, both Y and θ are unknown.
- A probability model is assumed for observed data if we knew θ is the truth.
- What if we have prior information about θ?

Bayesian Inference

Bayesian inference allows us to update prior beliefs with the observed data to quantify uncertainty about θ.

- Prior Distribution: $p(\theta)$
- Sampling Model (likelihood): $p(y \mid \theta)$
- Posterior Distribution

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

- Calculating $p(y)$ is typically very challenging. Use MCMC (implemented in JAGS) to estimate $p(\theta \mid y)$.

Linear Regression - as a Bayesian

We start with the sampling model $p(y \mid \theta)$[1]

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \quad , \end{aligned}$$

where x_i is the value of the covariate in individual i. Then we add prior distributions $p(\theta)$

$$\begin{aligned} \beta_0 &\sim \mathrm{N}(0, \tau) \\ \beta_1 &\sim \mathrm{N}(0, \tau) \\ \sigma^2 &= \frac{1}{\tau_{resid}} \\ \tau_{resid} &\sim \mathrm{Exponential}(\phi) \end{aligned}$$

[1] $\theta \in (\beta_0, \beta_1, \sigma^2)$

Linear Regression - in R+JAGS

- Rescale all variables to mean of 0, standard deviation of 1

```
Inference for Bugs model at "simple-linear-regression.jags", fit using jags,
 5 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 5000 iterations saved
            mu.vect sd.vect   2.5%    25%    50%    75%  97.5% Rhat n.eff
beta.0        0.000   0.041 -0.082 -0.028  0.000  0.028  0.080 1.001  5000
beta.mat     -0.313   0.041 -0.393 -0.341 -0.314 -0.285 -0.231 1.002  1900
sigma.resid   0.953   0.029  0.897  0.933  0.952  0.972  1.012 1.001  3700
```

- Compare with lm() results from R

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.614e-17  4.110e-02   0.000        1
mat         -3.131e-01  4.114e-02  -7.611 1.24e-13 ***

Residual standard error: 0.9506 on 533 degrees of freedom
```

Multiple linear regression

Simple generalization of what we've already seen

$$\begin{aligned} y_i &\sim \mathrm{N}(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} \quad , \end{aligned}$$

where x_{ik} is the value of the kth covariate in individual i. The priors are

$$\begin{aligned} \beta_i &\sim \mathrm{N}(0, \tau), \quad i = 0, \ldots, K \\ \sigma^2 &= \frac{1}{\tau_{resid}} \\ \tau_{resid} &\sim \mathrm{Exponential}(\phi) \end{aligned}$$

Multiple Linear Regression

From JAGS

```
Inference for Bugs model at "multiple-linear-regression.jags", fit using jags,
 5 chains, each with 10000 iterations (first 5000 discarded), n.thin = 5
 n.sims = 5000 iterations saved
             mu.vect sd.vect   2.5%    25%    50%    75%  97.5% Rhat n.eff
beta.0         0.000   0.040 -0.078 -0.027  0.000  0.028  0.078 1.001  4800
beta.cdd       0.106   0.073 -0.036  0.058  0.105  0.154  0.251 1.001  5000
beta.elev     -0.319   0.094 -0.500 -0.384 -0.319 -0.256 -0.136 1.001  5000
beta.inso      0.054   0.053 -0.048  0.018  0.054  0.091  0.157 1.001  5000
beta.map      -0.022   0.078 -0.178 -0.074 -0.020  0.031  0.129 1.001  5000
beta.mat      -0.463   0.114 -0.688 -0.541 -0.461 -0.385 -0.243 1.001  5000
beta.ratio    -0.016   0.079 -0.172 -0.069 -0.013  0.039  0.134 1.001  5000
sigma.resid    0.939   0.029  0.884  0.919  0.939  0.958  1.000 1.001  5000
```

Compare to lm() from R

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.520e-17  4.052e-02   0.000 1.000000
cdd          1.051e-01  7.315e-02   1.437 0.151291
elev        -3.184e-01  9.534e-02  -3.339 0.000899 ***
inso         5.288e-02  5.287e-02   1.000 0.317702
map         -2.309e-02  7.749e-02  -0.298 0.765836
mat         -4.629e-01  1.147e-01  -4.037  6.2e-05 ***
ratio       -1.481e-02  7.944e-02  -0.186 0.852218

Residual standard error: 0.9373 on 528 degrees of freedom
```

Multiple Linear Regression with Species Random Effect

$\gamma_i^{(s)}$ denotes the mean for species s to which inidividual i belongs

$$\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma^2_{resid}) \\
\mu_i &= \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \gamma_i^{(s)} \\
\beta_i &\sim \mathrm{N}(0, \tau), \quad i = 0, \ldots, K \\
\sigma^2_{resid} &= \frac{1}{\tau_{resid}} \\
\tau_{resid} &\sim \mathrm{Exponential}(\phi) \\
\gamma_i^{(s)} &\sim \mathrm{N}(0, \sigma^2_{species}) \\
\sigma^2_{species} &= \frac{1}{\tau_{species}} \\
\tau_{species} &\sim \mathrm{Exponential}(\phi)
\end{aligned}$$

Multiple Linear Regression with Species Random Effect

Alternatively

$$
\begin{aligned}
y_i &\sim \mathrm{N}(\mu_i, \sigma^2_{resid}) \\
\mu_i &= \beta^{(s)}_{0i} + \sum_{k=1}^{K} \beta_k x_{ik} \\
\sigma^2_{resid} &= \frac{1}{\tau_{resid}} \\
\tau_{resid} &\sim \mathrm{Exponential}(\phi) \\
\beta^{(s)}_{0i} &\sim \mathrm{N}(\beta_0, \sigma^2_{species}) \\
\sigma^2_{species} &= \frac{1}{\tau_{species}} \\
\tau_{species} &\sim \mathrm{Exponential}(\phi) \\
\beta_i &\sim \mathrm{N}(0, \tau), \quad i = 0, \ldots, K
\end{aligned}
$$

Multiple Linear Regression with Species Random Effect

From JAGS

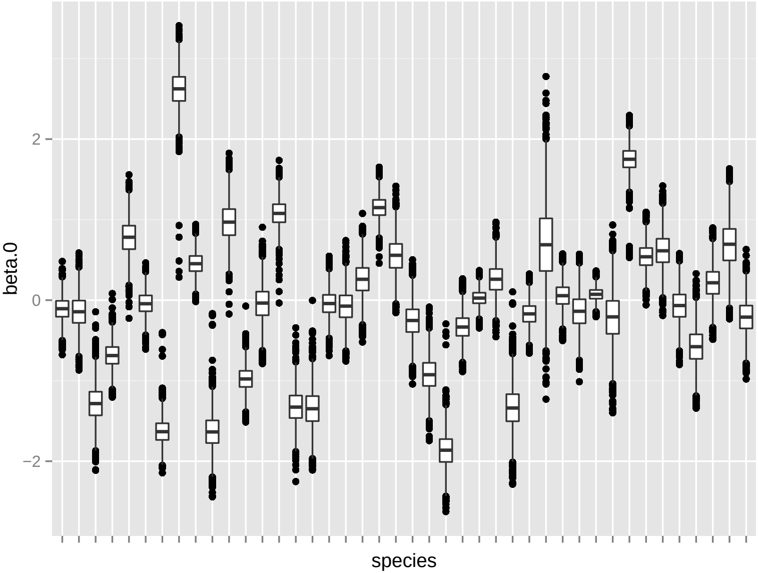
```
beta.cdd         0.081   0.055  -0.024   0.045   0.082   0.118   0.188 1.001  3100
beta.elev       -0.201   0.108  -0.408  -0.274  -0.201  -0.128   0.010 1.002  2400
beta.inso       -0.073   0.058  -0.187  -0.112  -0.073  -0.034   0.040 1.001  3400
beta.map        -0.418   0.083  -0.581  -0.474  -0.419  -0.362  -0.257 1.001  5000
beta.mat         0.093   0.110  -0.117   0.020   0.092   0.169   0.307 1.001  5000
beta.ratio       0.427   0.078   0.278   0.374   0.427   0.478   0.578 1.001  4800
beta.zero       -0.064   0.154  -0.369  -0.167  -0.062   0.037   0.241 1.001  5000
sigma.resid      0.545   0.023   0.511   0.532   0.544   0.556   0.581 1.001  3100
sigma.species    0.966   0.117   0.767   0.884   0.960   1.036   1.218 1.001  5000
```

Compare to lmer() From R

```
Random effects:
 Groups   Name        Variance Std.Dev.
 species  (Intercept) 0.8951   0.9461
 Residual             0.2924   0.5408
Number of obs: 535, groups:  species, 42

Fixed effects:
            Estimate Std. Error t value
(Intercept) -0.06289    0.14898  -0.422
cdd          0.07941    0.05576   1.424
elev        -0.19835    0.10946  -1.812
inso        -0.07259    0.05791  -1.254
map         -0.42009    0.08176  -5.138
mat          0.09811    0.10893   0.901
ratio        0.43022    0.07776   5.533
```

Multiple Linear Regression with Species Random Effect



Parametric Sampling Models

Assume

$$\begin{cases} Y_1, \cdots, Y_n \overset{i.i.d.}{\sim} p(y \mid \theta) \\ \theta \sim p(\theta) \end{cases}$$

Applicable if $Y_1, \cdots, Y_n$ are

- outcomes of a repeatable experiment;
- random sample from finite population with replacement;
- sampled from an infinite population w/out replacement;
- sampled from a finite population of size $N \gg n$ w/out replacement (approximate).

Labels carry no information.

Normal Model

Assume

$$Y_1, \cdots, Y_n \mid \mu, \sigma^2 \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2),$$

where μ and σ^2 are unknown parameters. From a Bayesian perspective, it is easier to work with the *precision*, ϕ, where $\phi = 1/\sigma^2$. Define $Y = (Y_1, \cdots, Y_n)$.

Likelihood

$$\begin{aligned} \mathcal{L}(\mu, \phi \mid Y) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \phi^{1/2} \exp\left\{-\frac{1}{2}\phi(Y_i - \mu)^2\right\} \\ &\propto \phi^{n/2} \exp\left\{-\frac{1}{2}\phi\sum_{i=1}^{n}(Y_i - \mu)^2\right\}. \end{aligned}$$

Likelihood Factorization

$$
\begin{aligned}
\mathcal{L}(\mu, \phi \mid Y) &\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_{i=1}^{n} (Y_i - \mu)^2 \right\} \\
&\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \sum_{i=1}^{n} \left[(Y_i - \overline{Y}) - (\mu - \overline{Y}) \right]^2 \right\} \\
&\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi \left[\sum_{i=1}^{n} (Y_i - \overline{Y})^2 + n(\mu - \overline{Y})^2 \right]^2 \right\} \\
&\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi s^2 (n-1) \right\} \exp\left\{ \frac{1}{2}\phi n(\mu - \overline{Y})^2 \right\} \\
&\propto \phi^{n/2} \exp\left\{ -\frac{1}{2}\phi SS \right\} \exp\left\{ \frac{1}{2}\phi n(\mu - \overline{Y})^2 \right\},
\end{aligned}
$$

where $\overline{Y} = \sum_{i=1}^{n} Y_i / n$ is the sample mean,
$s^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 / (n-1)$ is the sample variance and
$SS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ is the sample sum of squares.

Conjugate Priors

Consider a class of prior distributions, $p(\theta) \in \mathcal{P}$. We say that the class is conjugate for a sampling model $p(Y \mid \theta)$, if $p(\theta) \in \mathcal{P}$ implies that $p(\theta \mid y) \in \mathcal{P}$ for all $p(\theta) \in \mathcal{P}$ and data y.

Conjugate Normal-Gamma Prior for Normal Data

The conjugate prior distribution for (μ, ϕ) is Normal-Gamma, i.e.,

$$\begin{aligned} \mu \mid \phi &\sim \mathcal{N}(m_0, 1/(p_0\phi)), \\ \phi &\sim \mathcal{G}a(\nu_0/2, SS_0/2), \end{aligned}$$

which is written as

$$p(\mu, \phi) \propto \phi^{\nu_0/2-1} \exp\left\{-\phi \frac{SS_0}{2}\right\} \exp\left\{-\phi \frac{p_0}{2}(\mu - m_0)^2\right\}.$$

and further we can denote it is drawn from a Normal-Gamma family

$$\mu, \phi \sim \mathcal{NG}(m_0, p_0, \nu_0/2, SS_0).$$

Updating the Posterior Parameters

Under the Normal-Gamma prior distribution, we can derive the posterior distribution

$$\begin{aligned}
\mu \mid \phi, Y &\sim \mathcal{N}\left(m_n, \frac{1}{p_n \phi}\right), \\
\phi \mid Y &\sim \mathcal{G}a\left(\frac{\nu_n}{2}, \frac{SS_n}{2}\right), \\
\text{where } p_n &= p_0 + n, \\
m_n &= \frac{n\overline{Y} + p_0 m_0}{p_n}, \\
\nu_n &= \nu_0 + n, \\
SS_n &= SS_0 + SS + \frac{n p_0}{p_n}(\overline{Y} - m_0)^2.
\end{aligned}$$

The posterior distribution of (μ, ϕ) can be rewritten as a Normal-Gamma family

$$\mu, \phi \mid Y \sim \mathcal{NG}(m_n, p_n, \nu_n/2, SS_n/2).$$

Interpretation

- p_n indicates the precision for estimating μ after getting n observations.
- m_n is the expected value for μ after obtaining n observations, which can be viewed as a weighted average of sample mean and prior mean, i.e.,

$$m_n = \frac{n}{p_n}\overline{Y} + \frac{p_0}{p_n} m_0.$$

- ν_n is called the degrees of freedom, by noticing that

$$\phi \sim \mathcal{G}a(a/2, b/2) \Leftrightarrow \phi b \sim \chi^2_a \text{ with degrees of freedom } a.$$

- Denote $SS_n = SS_0 + SS + \frac{np_0}{p_n}(\overline{Y} - m_0)^2$ as the posterior variation, where the three terms can be explained as prior variation, observed variation (sum of squares) and variation between prior mean and sample mean, respectively.

Marginal Distribution for $\mu \mid Y$

$$
\begin{aligned}
p(\mu|Y) &\propto \int p(\mu,\phi|Y)d\phi \\
&= \int \phi^{\frac{\nu_n+1}{2}-1} \exp\left[-\phi\left\{\frac{SS_n + p_n(\mu - m_n)^2}{2}\right\}\right] d\phi
\end{aligned}
$$

This has the form of a Gamma integral with $a = (\nu + 1)/2$ and b equal to the mess multiplying ϕ in the exponential term, so that the result is $\propto b^{-a}$ (at least that is all that matters)

$$
\begin{aligned}
p(\mu \mid Y) &\propto \left\{SS_n + p_n(\mu - m_n)^2\right\}^{\frac{-(\nu_n+1)}{2}} \\
&\propto \left(\nu_n + \frac{(\mu - m_n)^2}{\frac{1}{p_n}\frac{SS_n}{\nu_n}}\right)^{-(\nu_n+1)/2},
\end{aligned}
$$

which is a Student $t_{\nu_n}(m_n, s_n^2)$ with location $m_n, \mathrm{d}f = \nu_n$, scale $s_n^2 = \frac{1}{p_n}\frac{SS_n}{\nu_n}$ and degrees of freedom being ν_n.

Standard Student t

Standardize $X \sim t_\delta(l, S)$ by subtracting location and dividing by square root of the scale:

$$\frac{X - l}{\sqrt{S}} \sim t_\delta(0, 1)$$

(new location 0 and scale 1)

$$\Rightarrow \frac{\mu - m_n}{s_n} \sim t_{\nu_n}(0, 1)$$

$$\mu \overset{D}{=} m_n + t_{\nu_n} s_n,$$

where t_{ν_n} can be easily evaluated using rt, qt, pt, and dt in R.

Problem

- Data: Observe pairs (Y_i, X_i), $i = 1, \cdots, n$;
- Response or dependent variable Y;
- Predictor or independent variable X.

Goals

- Exploring $p(y \mid x)$ as a function of x;
- Understanding the mean (variability) in Y as a function of X;
- Special case: Linear regression (normal Y).

Simple Linear Regression Models

$$Y_i = \alpha + \beta X_i + \epsilon_i, \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

- Estimate parameters (α,β,σ^2);
- interpretation of parameters: β, α
- assess model fit – adequate? good? if inadequate, how?
- predict new ("future") response at new $x_{n+1}, \cdots$
- how much variability does x explain?

Conjugate Priors for Simple Linear Regression

Normal-Gamma Prior

The Normal-Gamma distribution is conjugate for α, β and precision $\phi \equiv 1/\sigma^2$.

- $(\alpha, \beta)|\phi \sim N((\alpha_0, \beta_0), \phi^{-1}\Sigma)$, where Σ is a 2×2 matrix of variances and covariance;
- $\phi \sim \text{Gamma}(\nu_0/2, \nu_0\sigma_0^2/2)$.

A Reference Prior for Regression

Limiting case of conjugate prior as the prior variances goes to infinity (information goes to zero)

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

Theory for Inference

$$\frac{\beta - \hat{\beta}}{\sqrt{s^2_{Y|X} \frac{1}{S_{xx}}}} \sim t_{n-2}(0, 1)$$

- $\hat{\beta}$ is OLS(MLE) estimate of β, $s^2_{Y|X} = \hat{\sigma}^2$ is the MSE.
- (marginal) posterior for β is a Student t distribution with $n - 2$ df.
- Sampling distribution of $\hat{\beta}$ given β is Student t distribution with $n - 2$ df.

Used for classical and Bayesian (Reference) analysis

Distributions Continued

- (marginal) posterior for α is a Student t distribution with $n-2$ df.
- Sampling distribution of $\hat{\alpha}$ is Student t distrbution with $n-2$ df.

$$\frac{\alpha - \hat{\alpha}}{\sqrt{s^2_{Y|X}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}(0,1)$$

Significance of Regression

Measuring the "explanatory power" of predictor X

- Credible Intervals (HPD or equal-tailed)

$$\hat{\beta} \pm t_{\alpha/2} S_{\beta}$$

where $t_{\alpha/2}$ is 100 $\alpha/2\%$ quantile of a standard t_{n-2} and $s_{\beta} = \sqrt{s^2_{Y|X}/S_{xx}}$

- Confidence Intervals:

$$\hat{\beta} \pm t_{\alpha/2} SE_{\hat{\beta}}$$

and $SE(\hat{\beta}) = \sqrt{s^2_{Y|X} 1/S_{xx}}$

Predictions

The (posterior) predictive distribution for a new case, $y_{n+1} = \alpha + \beta x_{n+1} + \epsilon_{n+1}$ is also a Student t distribution with $n-2$ df.

$$
\begin{aligned}
y_{n+1} | y_1, \cdots, y_n &\sim t_{n-2}(\hat{y}, s^2_{y_{n+1}}) \\
\hat{y} &= \hat{\alpha} + \hat{\beta} x_{n+1} \\
s^2_{y_{n+1}} &= s^2_{Y|X} \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{S_{xx}} \right)
\end{aligned}
$$

- posterior uncertainty about $\alpha + \beta x_{n+1}$
- depends on x_{n+1} spread is higher for x_{n+1} far from $\bar{x}$
- additional variability $+s^2_{Y|X}$ due to ϵ_{n+1}

Multiple Linear Regression Models

Let us assume the general linear model is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, \tau^{-1}),$$

where τ is called the error precision.

- Likelihood:

$$\mathcal{L}(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \tau) = \prod_{i=1}^{n} \sqrt{\frac{\tau}{2\pi}} \exp\left\{ -\frac{\tau}{2}(y_i - \mathbf{x}_i'\beta)^2 \right\}$$

- Normal-Gamma Prior:

$$\begin{aligned} \pi(\boldsymbol{\beta}, \tau) &= \pi(\boldsymbol{\beta} \mid \tau)\pi(\tau) = \mathcal{N}_p(\boldsymbol{\beta}; \boldsymbol{\beta}_0, \tau^{-1}\Sigma_0)\mathcal{G}(\tau; a, b). \\ \pi(\boldsymbol{\beta} \mid \tau) &= |2\pi\tau^{-1}\Sigma_0|^{-p/2} \exp\left\{ -\frac{\tau}{2}(\beta - \beta_0)'\Sigma_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\}, \end{aligned}$$

where $\boldsymbol{\beta}_0$ is the prior mean and Σ_0 is the prior covariance and

$$\pi(\tau) = \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\}.$$

- A non-informative prior: $\pi(\boldsymbol{\beta}, \tau) \propto 1/\tau$.

Bayesian Regression with Non-informative Priors

- The non-informative prior are invalid probabilities (it does not integrate to any finite number). So why is it that we are even discussing them?
- It turns out that even if the priors are improper (that's what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.
- When the non-informative prior is used, after some algebra, the joint posterior distribution of $(\boldsymbol{\beta}, \tau)$ is

$$\begin{aligned} & \pi(\boldsymbol{\beta}, \tau \mid \mathbf{X}, \mathbf{y}) \\ \propto \quad & \mathcal{L}(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}, \tau) \\ \propto \quad & (\tau)^{n/2-1} \exp\left(-\frac{\tau}{2}(\hat{\sigma}^2(n-p) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right), \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma}^2 = \frac{(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p}$, which are the classical unbiased estimates of the regression parameter $\boldsymbol{\beta}$ and $\sigma^2 = \tau^{-1}$.

Bayesian Regression with Non-informative Priors (Cont.)

- The marginal posterior distribution of τ is actually a Gamma distribution, i.e., $\tau \mid \mathbf{X}, \mathbf{y} \sim \mathcal{G}((n-p)/2, (n-p)\hat{\sigma}^2/2)$. In other words, σ^2 follow a inverse Gamma distribution, i.e., $\sigma^2 \mid \mathbf{X}, \mathbf{y} \sim \mathcal{IG}((n-p)/2, (n-p)\hat{\sigma}^2/2)$. That implies $(n-p)\hat{\sigma}^2/\sigma^2 \mid \mathbf{X}, \mathbf{y} \sim \chi^2_{n-p}$.
- A striking similarity with the classic result: The distribution of $\hat{\sigma}^2$ is also characterized as $(n-p)\hat{\sigma}^2/\sigma^2$ following a chi-square distribution with $n-p$ degrees of freedom.
- The marginal posterior distribution of $\boldsymbol{\beta}$ follows a multivariate t distribution, i.e.,

$$\begin{aligned}\pi(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) &= \frac{\Gamma(n/2)}{\pi^{p/2}(n-p)^{p/2}\Gamma((n-p)/2)s^p|(\mathbf{X}'\mathbf{X})|^{-1/2}} \\ &\times \left[1 + \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{(n-p)\hat{\sigma}^2}\right]^{-n/2},\end{aligned}$$

which we denote as $\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y} \sim \mathbf{t}_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})$.

Interpretation

Under the non-informative prior $\pi(\boldsymbol{\beta}, \tau) \propto 1/\tau$,

- Marginally $\beta_j \mid \mathbf{X}, \mathbf{y} \sim t_{n-p}(\hat{\beta}_j, s^2_{\beta_j})$, where $s^2_{\beta_j}$ is the jth element on the diagonal of $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.
- HPD interval $\hat{\beta}_j \pm t_{n-p,\alpha/2} s_{\beta_j}$.

Question: How probable is $\beta_j = 0$ under the posterior?

- Informal "test": Probability in tails = signifcance level = (Bayesian) p-value

$$p\text{-value} = \mathrm{P}(|t| > |\hat{\beta}_j / s_{\beta_j}|).$$

- Lindley's Method: Lindley suggested rejecting the hypothesis that $\beta_j = 0$ at the α level of significance if the $100(1-\alpha)\%$ HPD region does not include 0, i.e.,

$$0 \notin (\hat{\beta}_j - t_{n-p,\alpha/2} s_{\beta_j}, \hat{\beta}_j + t_{n-p,\alpha/2} s_{\beta_j}).$$

  which is equivalent to comparing the p-value to α and concluding that the regression is significant if the p-value is less than α.

Bayes Factor

Testing $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$.

- Assign prior probabilities to H_0 and H_a.
- Find $\mathrm{P}(H_i \mid \mathbf{y})$ via Bayes Theorem.

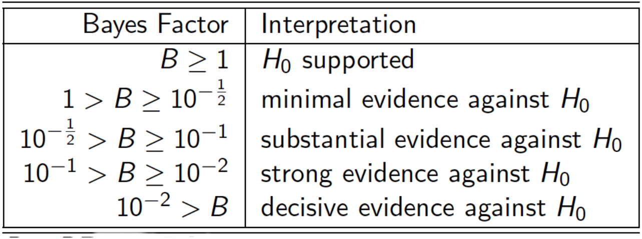
Bayes Factor for comparing evidence in favor of H_0

$$BF[H_0 : H_a] = \frac{P(H_0 \mid \mathbf{y})/p(H_0)}{P(H_a \mid \mathbf{y})/p(H_a)}.$$

Often difficult to calculate, instead use lower bound based on p-values (Berger, Sellke and Bayarri, 2001)

$$BF[H_0 : H_a] = -ep \log(p).$$

Jeffreys Scale of Evidence



Bayes Factor	Interpretation
$B \geq 1$	H_0 supported
$1 > B \geq 10^{-\frac{1}{2}}$	minimal evidence against H_0
$10^{-\frac{1}{2}} > B \geq 10^{-1}$	substantial evidence against H_0
$10^{-1} > B \geq 10^{-2}$	strong evidence against H_0
$10^{-2} > B$	decisive evidence against H_0

Here $B = BF[H_0 : H_a]$.

Problems with Multicollinearity

- Variables may appear to be unimportant (when they are).
- Coefficient estimates are unstable and hard to interpret (can estimate combinations of coefficients but not individual coefficients).

Alternative Bayesian solutions:

- Independent Prior Distributions
- Variable Selection

Hierarichal Model with Independent Priors

Hierarchical Model:

$$\begin{aligned} \beta_j \mid \lambda_j, \sigma^2 &\sim N(0, \sigma^2/\lambda_j) \\ \lambda_j \mid \sigma^2 &\sim G(1/2, 1/2) \\ 1/\sigma^2 &\sim G(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

- leads to nice conjugate updates for all full conditionals.
- Easy to code in WinBUGS (JAGS?)
- Allows each parameter to have own precision with mean 1
- Usually re-scale X so that columns have mean 0 and standard deviation 1.

Cauchy Prior

First two equations imply that $\beta_j|\sigma^2 \sim C(0, \sigma^2)$

$$p(\beta) = \frac{1}{\pi\sigma}\left(1 + \frac{\beta^2}{\sigma^2}\right)^{-1}$$

leading to a collapsed model

$$\begin{aligned} \mathbf{Y}|\beta, \sigma^2 &\sim N(\mathbf{X}\beta, \sigma^2\mathbf{I_n}) \\ \beta_j|\sigma^2 &\sim C(0, \sigma^2) \\ 1/\sigma^2 &\sim G(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

No nice full conditional for β_j.

Metropolis-Hastings Algorithm

For θ_j

- propose a new $\theta_j^* \sim q(\cdot|\theta_j^t)$
- Sampling from the wrong distribution!
- Need to balance acceptance of proposed values so that we have samples that represent the posterior
- Metropolis-Hastings ratio

$$\alpha = \frac{p(\mathbf{Y}|\theta^*)p(\theta^*)/q(\theta^*|\theta^{(t)})}{p(\mathbf{Y}|\theta^{(t)})p(\theta^{(t)})/q(\theta^{(t)}|\theta^*)}$$

- if $\alpha < 1$ set

$$\theta^{(t+1)} = \begin{cases} \theta^* & \text{with probability } \alpha \\ \theta^{(t)} & \text{with probability } 1 - \alpha \end{cases}$$

Otherwiese if $\alpha > 1$ set $\theta^{(t+1)} = \theta^*$

Metropolis Algorithm

Simplest version is called Random-Walk Metropolis

- Take $q(\theta|\theta(t))$ to be a symmetric density
- Normal density centered at $\theta^{(t)}$ with variance c^2
- ratio of proposals cancels!

```
thetastar = rnorm(1, theta[t], c)
alpha = min(1, ((logL(thetastar) + logprior(thetastar)
-(logL(theta[t]) +logprior(theta[t])))
u = log(runif(1))
theta[t+1] = theta[t]
if (u < alpha) theta[t+1] = thetastar
```

Can repeat for each θ_j.

Model Selection

Selection of a single model has the following problems

- When the criteria suggest that several models are equally good, what should we report? Still pick only one model?
- What do we report for our uncertainty after selecting a model?

Typical analysis ignores model uncertainty!

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{x}_1, \cdots, \mathbf{x}_p$ variables.
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)'$ where $\gamma_j \in \{0, 1\}$ is is an indicator for whether variable $\mathbf{x}_j$ should be included in the model $M_{\boldsymbol{\gamma}}$. Notice $\gamma_j = 0 \Leftrightarrow \beta_j = 0$.
- Each value of $\boldsymbol{\gamma}$ represents one of the 2^p models.
- Under model $M_{\boldsymbol{\gamma}}$:

$$\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \tau \sim \mathcal{N}(\mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1} \mathbf{I})$$

where $\mathbf{X}_{\boldsymbol{\gamma}}$ is design matrix using the columns in $\mathbf{X}$ where $\gamma_j = 1$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the subset of $\boldsymbol{\beta}$ that are non-zero.

Bayesian Model Averaging

Rather than use a single model, BMA uses all (or potentially a lot) models, but weights model predictions by their posterior probabilities (measure of how much each model is supported by the data).

- Posterior model probabilities

$$\mathrm{P}(M_j \mid \mathbf{y}) = \frac{\mathrm{P}(\mathbf{y} \mid M_j)\mathrm{P}(M_j)}{\sum_j \mathrm{P}(\mathbf{y} \mid M_j)\mathrm{P}(M_j)},$$

  Marginal likelihod of a model is

$$\mathrm{P}(\mathbf{y} \mid M_\gamma) = \int\int \mathrm{P}(\mathbf{y} \mid \boldsymbol{\beta}_\gamma, \tau)\mathrm{P}(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \tau)\mathrm{P}(\tau \mid \boldsymbol{\gamma}) d\boldsymbol{\beta}_\gamma d\tau.$$

- Probability $\beta_j \neq 0$: $\sum_{M_j:\beta_j\neq 0} \mathrm{P}(M_j \mid \mathbf{y})$.

Bayesian Model Averaging (Continued)

- Predictions

$$\mathrm{P}(\mathbf{y}^{new} \mid \mathbf{y}) = \sum_{j} \mathrm{P}(\mathbf{y}^{new} \mid \mathbf{y}, M_j)\mathrm{P}(M_j \mid \mathbf{y}),$$

  where

$$\mathrm{P}(\mathbf{y}^{new} \mid \mathbf{y}, M_\gamma) = \int \mathrm{P}(\mathbf{y}^{new} \mid \mathbf{y}, \boldsymbol{\beta}_\gamma, \tau)\mathrm{P}(\boldsymbol{\beta}_\gamma, \tau \mid \mathbf{y}) d\boldsymbol{\beta}_\gamma d\tau.$$

Prior Distributions

- Bayesian Model choice requires proper prior distributions on regression coefficients.
- Vague but proper priors may lead to paradoxes!
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods, Zellner's g-prior is the most popular.

Zellner's g-prior

Centered model:

$$\mathbf{y} = \mathbf{1}_n\alpha + \tilde{\mathbf{X}}_c\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\tilde{\mathbf{X}}_c = (\mathbf{I}_n - 1/n\mathbf{J}_n)\mathbf{X}$ is the centered design matrix where all variables have had their mean subtracted.

- $\pi(\alpha) \propto 1$;
- $\pi(\tau) \propto 1/\tau$;
- $\boldsymbol{\beta} \mid \tau, \boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, g\tau^{-1}(\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1})$;
- take $g = n$.

which leads to marginal likelihood of M_γ that is proportional to

$$\mathrm{P}(\mathbf{y} \mid M_\gamma, g) \propto \frac{(1+g)^{(n-p_\gamma-1)/2}}{(1+g(1-R_\gamma^2))^{(n-1)/2}},$$

where R_γ^2 is the ordinary coefficient of determination of regression model M_γ.

Lastly, assign uniform distribution to space of models.

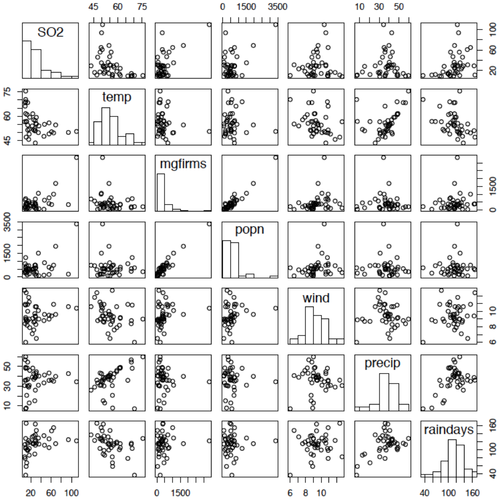
Air Pollution Data (KENT: REPLACE WITH YOUR EXAMPLE?)

- Response SO_2 measurements in 41 metropolitan areas.
- Predictors:
  - temp
  - mfgfirms
  - popn
  - wind
  - precip
  - raindays

Model for SO_2 as a function of the other variables?

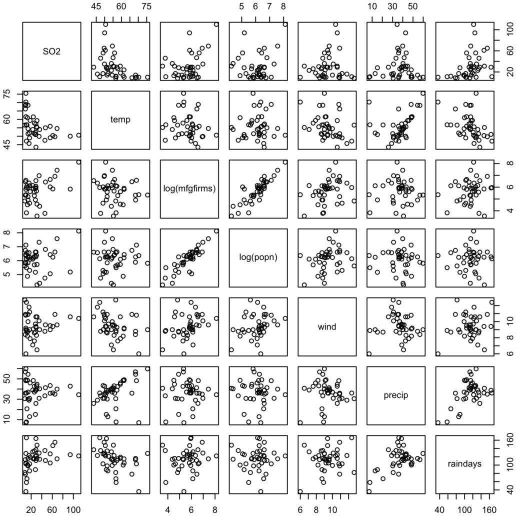
Scatterplot Matrix

Original Variables

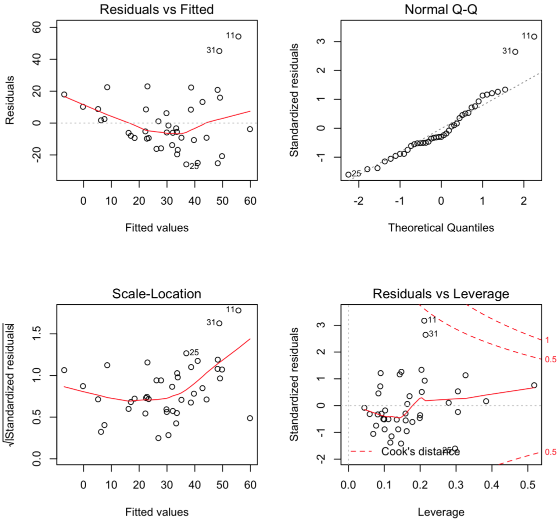


Scatterplot Matrix

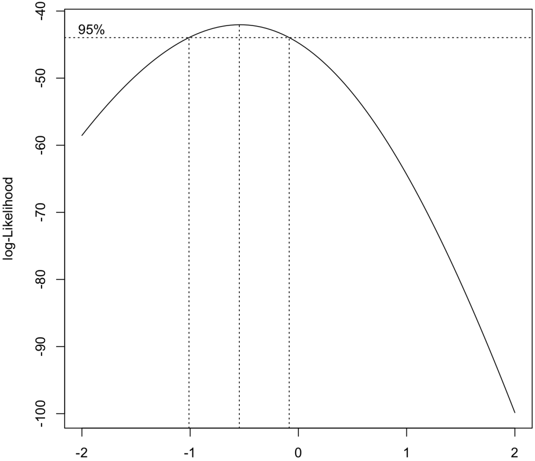
Transformed Predictor Variables



Residual Plots



BoxCox Profile Likelihood



$\lambda \approx -0.5.$

Residual Plots

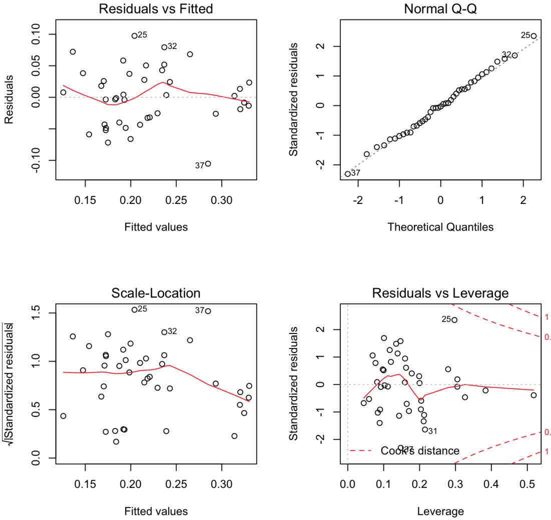


Table: Least Squares Estimates of the Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2287	0.1594	-1.43	0.1605
temp	0.0076	0.0022	3.46	0.0015
log(mfgfirms)	-0.0284	0.0187	-1.52	0.1389
log(popn)	0.0097	0.0226	0.43	0.6713
wind	0.0216	0.0064	3.39	0.0018
precip	-0.0018	0.0013	-1.39	0.1746
raindays	-0.0001	0.0006	-0.21	0.8327

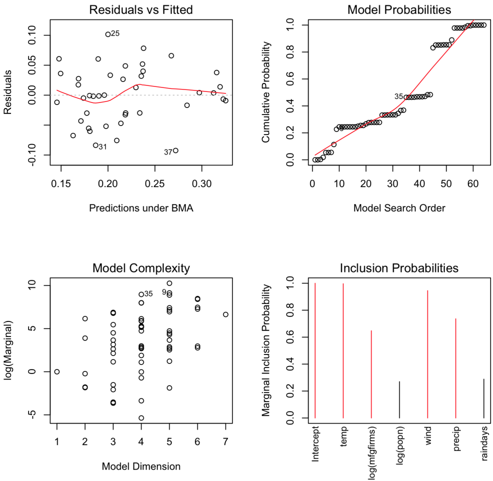
Pollution Example

- Temperature is a significant predictor of SO_2 according to the p-value.
- Lower bound on Bayes Factor

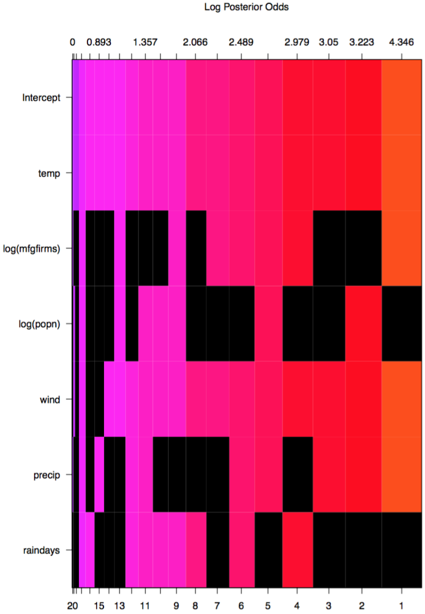
$$BF[H_0 : H_a] = -ep\log(p) = 0.027$$

  Here p is the p-value.
- Strong evidence against that the ccoefficient of Temperature is zero.

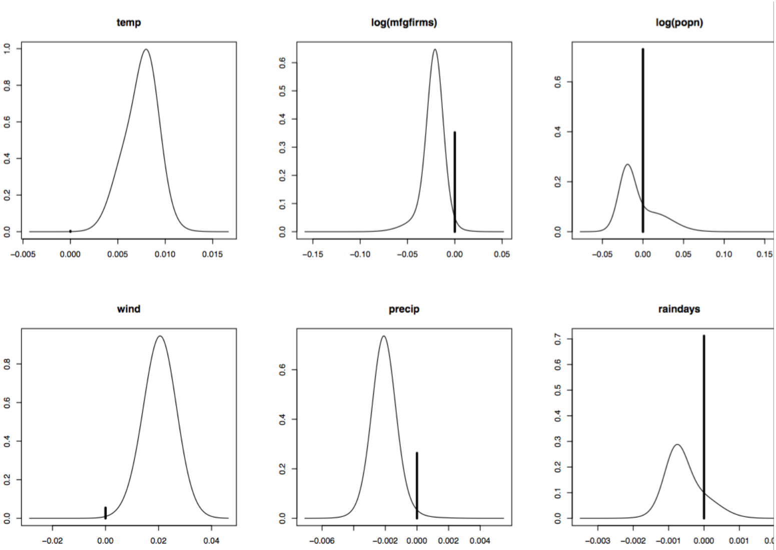
USair Data


Residuals vs Fitted
Residuals
Predictions under BMA
0.15
0.20
0.25
0.30
-0.10
0.00
0.05
0.10
25
31
37
Model Probabilities
Cumulative Probability
Model Search Order
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
35
Model Complexity
log(Marginal)
Model Dimension
1
2
3
4
5
6
7
-5
0
5
10
35
9
Inclusion Probabilities
Marginal Inclusion Probability
Intercept
temp
log(mfgfirms)
log(popn)
wind
precip
raindays

Model Space


Log Posterior Odds
0
0.893
1.357
2.066
2.489
2.979
3.05
3.223
4.346
Intercept
temp
log(mfgfirms)
log(popn)
wind
precip
raindays
20
15
13
11
9
8
7
6
5
4
3
2
1

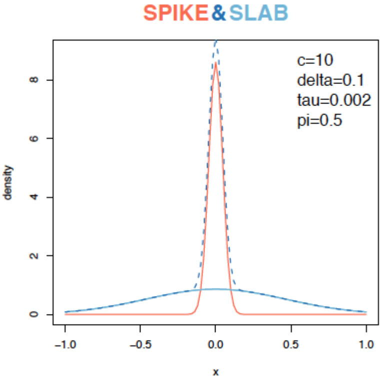
Coefficient



Stochastic Search Variable Selection

The Spike-and-Slab prior:

$$\begin{aligned}\beta_j \mid \gamma_j, c, \tau_j^2 &\sim (1-\gamma_j)\mathcal{N}(0, \tau_j^2) + \gamma_j \mathcal{N}(0, \tau_j^2 c^2) \\ \gamma_j \mid \pi_j &\sim Bernoulli(\pi_j)\end{aligned}$$



- $\gamma_j = 0$: Variable not in the model;
- $\gamma_j = 1$: Variable in the model;
- Calibration of hyper-parameters c, τ_j^2 needed.

Inference for Variable Selection

- Highest posterior model (HPM): Select a model that has been visited most often.
- Median probability model (MPM): Select variables that appear at least in 50% of visited models.

Alternative spike and slab models

- Popular approach in genomic research;
- Variants:
  - Conjugate version:

$$\beta_j \mid \gamma_j, c, \tau_j^2 \sim (1-\gamma_j)\mathcal{N}(0, \sigma^2\tau_j^2) + \gamma_j\mathcal{N}(0, \sigma^2\tau_j^2 c^2).$$

  - Replace the spike normal in Spike-and-Slab prior by Dirac, i.e.,

$$\beta_j \mid \gamma_j, \tau_j^2 \sim (1-\gamma_j)\delta_0 + \gamma_j\mathcal{N}(0, \tau_j^2).$$