Priors and Intro to Bayesian Variable Selection

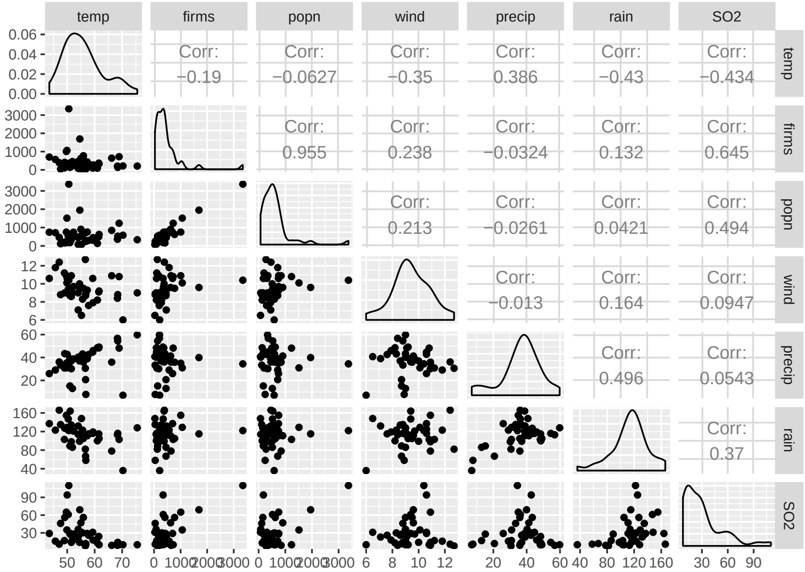
Hoff Chapter 9, Mixtures of g-Priors Liang et al JASA

October 18, 2017

Outline

- Priors in Bayesian Regression
- Model Selection

US Air Example



lm summary

```
lm(formula = log(SO2) ~ temp + log(firms) + log(popn) + wi
    precip + rain, data = usair)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.7142760  1.6475086   4.075 0.000261 ***
temp        -0.0649495  0.0227711  -2.852 0.007333 **
log(firms)   0.3698588  0.1934076   1.912 0.064289 .
log(popn)   -0.1771293  0.2335520  -0.758 0.453428
wind        -0.1738606  0.0656713  -2.647 0.012204 *
precip       0.0156032  0.0132718   1.176 0.247893
rain         0.0009153  0.0057335   0.160 0.874104
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.5108 on 34 degrees of freedom
Multiple R-squared:  0.5503,Adjusted R-squared:  0.471
F-statistic: 6.936 on 6 and 34 DF,  p-value: 7.12e-05
```

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$
- center prior at MLE and base covariance on the information in "1" observation

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$
- center prior at MLE and base covariance on the information in "1" observation
- Posterior mean

$$\frac{n}{1+n}\hat{\boldsymbol{\beta}} + \frac{1}{1+n}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$$

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$
- center prior at MLE and base covariance on the information in "1" observation
- Posterior mean

$$\frac{n}{1+n}\hat{\boldsymbol{\beta}} + \frac{1}{1+n}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$$

- Posterior Distribution

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathrm{N}\left(\hat{\boldsymbol{\beta}}, \frac{n}{1+n}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

Unit Information Prior

Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$
- center prior at MLE and base covariance on the information in "1" observation
- Posterior mean

$$\frac{n}{1+n}\hat{\boldsymbol{\beta}} + \frac{1}{1+n}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$$

- Posterior Distribution

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathrm{N}\left(\hat{\boldsymbol{\beta}}, \frac{n}{1+n}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

Cannot represent real prior beliefs; double use of data

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

- Treat $\boldsymbol{\beta}$ and ϕ independently ("orthogonal parameterization")

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

- Treat $\boldsymbol{\beta}$ and ϕ independently ("orthogonal parameterization")
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

- Treat $\boldsymbol{\beta}$ and ϕ independently ("orthogonal parameterization")
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$
- $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)\|^{1/2}$

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

- Treat $\boldsymbol{\beta}$ and ϕ independently ("orthogonal parameterization")
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$
- $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)\|^{1/2}$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Expected Fisher Information matrix

Jeffreys Prior

Jeffreys proposed a default procedure so that resulting prior would be invariant to model parameterization

- Treat $\boldsymbol{\beta}$ and ϕ independently ("orthogonal parameterization")
- $p_{IJ}(\boldsymbol{\beta}) \propto |\mathcal{I}(\boldsymbol{\beta})|^{1/2}$
- $p_{IJ}(\phi) \propto |\mathcal{I}(\phi)||^{1/2}$

where $\mathcal{I}(\boldsymbol{\theta})$ is the Expected Fisher Information matrix

$$\mathcal{I}(\theta) = -\mathsf{E}[\left[\frac{\partial^2 \log(\mathcal{L}(\boldsymbol{\theta}))}{\partial\theta_i \partial\theta_j}\right]]$$

Fisher Information Matrix

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2}\log(\phi) - \frac{\phi}{2}\text{SSE} - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

Fisher Information Matrix

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2}\log(\phi) - \frac{\phi}{2}\text{SSE} - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

$$\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & -(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X}) & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

Fisher Information Matrix

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2}\log(\phi) - \frac{\phi}{2}\text{SSE} - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & -(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X}) & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix} \\
\mathsf{E}[\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}] &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}
\end{aligned}$$

Fisher Information Matrix

Log Likelihood

$$\log(\mathcal{L}(\boldsymbol{\beta}, \phi)) = \frac{n}{2}\log(\phi) - \frac{\phi}{2}\text{SSE} - \frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

$$\begin{aligned}
\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & -(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \\ -(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X}) & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix} \\
\mathrm{E}[\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}] &= \begin{bmatrix} -\phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & -\frac{n}{2}\frac{1}{\phi^2} \end{bmatrix} \\
\mathcal{I}((\boldsymbol{\beta}, \phi)^T) &= \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}
\end{aligned}$$

Independent Jeffreys Prior "Reference Prior"

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

Independent Jeffreys Prior "Reference Prior"

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

$$p_{IJ}(\boldsymbol{\beta}) \propto |\phi \mathbf{X}^T\mathbf{X}|^{1/2} \propto 1$$

Independent Jeffreys Prior "Reference Prior"

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

$$p_{IJ}(\boldsymbol{\beta}) \propto |\phi \mathbf{X}^T\mathbf{X}|^{1/2} \propto 1$$

$$p_{IJ}(\phi) \propto \phi^{-1}$$

Independent Jeffreys Prior "Reference Prior"

$$\mathcal{I}((\boldsymbol{\beta}, \phi)^T) = \begin{bmatrix} \phi(\mathbf{X}^T\mathbf{X}) & \mathbf{0}_p \\ \mathbf{0}_p^T & \frac{n}{2}\frac{1}{\phi^2} \end{bmatrix}$$

$$p_{IJ}(\boldsymbol{\beta}) \propto |\phi \mathbf{X}^T\mathbf{X}|^{1/2} \propto 1$$

$$p_{IJ}(\phi) \propto \phi^{-1}$$

Independent Jeffreys Prior is

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

$$\boldsymbol{\beta} \mid \phi, \mathbf{Y} \quad \sim \quad \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1})$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

$$\begin{aligned} \boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\ \phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p)/2, \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2/2) \end{aligned}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

$$\begin{aligned} \boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\ \phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p)/2, \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2) \\ \boldsymbol{\beta} \mid \mathbf{Y} &\sim t_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}) \end{aligned}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

$$\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p)/2, \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2) \\
\boldsymbol{\beta} \mid \mathbf{Y} &\sim t_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})
\end{aligned}$$

Bayesian Credible Sets $p(\boldsymbol{\beta} \in C_\alpha) = 1 - \alpha$ correspond to frequentist Confidence Regions

$$\frac{\boldsymbol{\lambda}^T\boldsymbol{\beta} - \boldsymbol{\lambda}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\lambda}}} \sim t_{n-p}$$

Formal Posterior Distribution

With Independent Jeffreys Prior

$$p_{IJ}(\beta, \phi) \propto p_{IJ}(\boldsymbol{\beta}) p_{IJ}(\phi) = \phi^{-1}$$

Formal Posterior Distribution (Show!)

$$\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}, (\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p)/2, \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/2) \\
\boldsymbol{\beta} \mid \mathbf{Y} &\sim t_{n-p}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1})
\end{aligned}$$

Bayesian Credible Sets $p(\boldsymbol{\beta} \in C_\alpha) = 1 - \alpha$ correspond to frequentist Confidence Regions

$$\frac{\boldsymbol{\lambda}^T\boldsymbol{\beta} - \boldsymbol{\lambda}\hat{\boldsymbol{\beta}}}{\sqrt{\hat{\sigma}^2\boldsymbol{\lambda}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\lambda}}} \sim t_{n-p}$$

BUT Cannot be used for Model Selection

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean
- Invariance under linear transformations of X and Y

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathrm{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean
- Invariance under linear transformations of X and Y
- Avoids extra inverses beyond those in obtaining OLS estimates (computational)

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean
- Invariance under linear transformations of X and Y
- Avoids extra inverses beyond those in obtaining OLS estimates (computational)
- $\frac{g}{1+g}$ weight given to the data

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean
- Invariance under linear transformations of X and Y
- Avoids extra inverses beyond those in obtaining OLS estimates (computational)
- $\frac{g}{1+g}$ weight given to the data
- Choice of g?

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathrm{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Zellner proposed informative choice for the prior mean
- Invariance under linear transformations of X and Y
- Avoids extra inverses beyond those in obtaining OLS estimates (computational)
- $\frac{g}{1+g}$ weight given to the data
- Choice of g?
- Same g for intercept and other coefficients

Zellner's g-prior II

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c \boldsymbol{\beta} + \epsilon$$

where $\mathbf{X}^c$ is the centered design matrix where all variables have had their mean subtracted

Zellner's g-prior II

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c \boldsymbol{\beta} + \epsilon$$

where $\mathbf{X}^c$ is the centered design matrix where all variables have had their mean subtracted

- $p(\phi) \propto 1/\phi$

Zellner's g-prior II

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c \boldsymbol{\beta} + \epsilon$$

where $\mathbf{X}^c$ is the centered design matrix where all variables have had their mean subtracted

- $p(\phi) \propto 1/\phi$
- $p(\alpha \mid \phi) \propto 1$

Zellner's g-prior II

Centered model:

$$\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}^c\boldsymbol{\beta} + \epsilon$$

where $\mathbf{X}^c$ is the centered design matrix where all variables have had their mean subtracted

- $p(\phi) \propto 1/\phi$
- $p(\alpha \mid \phi) \propto 1$
- $\boldsymbol{\beta} \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}^{c\prime}\mathbf{X}^c)^{-1})$

Zellner's g-prior II

Centered model:

$$\mathbf{Y} = \mathbf{1}_n\alpha + \mathbf{X}^c\boldsymbol{\beta} + \epsilon$$

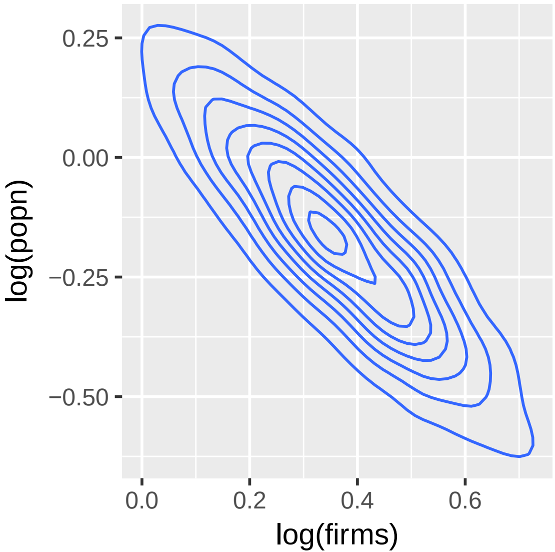
where $\mathbf{X}^c$ is the centered design matrix where all variables have had their mean subtracted

- $p(\phi) \propto 1/\phi$
- $p(\alpha \mid \phi) \propto 1$
- $\boldsymbol{\beta} \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathrm{N}(0, g\phi^{-1}(\mathbf{X}^{c\prime}\mathbf{X}^c)^{-1})$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \alpha, \phi \sim N\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}}, \phi^{-1}\frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\right)$$

$$\phi \mid \mathbf{Y} \sim \mathrm{Gamma}\left(\frac{n-1}{2}, \frac{\mathrm{SSE} + \frac{1}{1+g}\hat{\boldsymbol{\beta}}^T(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}}}{2}\right)$$

joint posterior draws of beta's



Bayesian Variable Selection

- Avoid the use of redundant variables (problems with interpretations)

Bayesian Variable Selection

- Avoid the use of redundant variables (problems with interpretations)
- Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other

Bayesian Variable Selection

- Avoid the use of redundant variables (problems with interpretations)
- Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other
- reduced MSE: reduced variance but possibly higher bias

Bayesian Variable Selection

- Avoid the use of redundant variables (problems with interpretations)
- Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other
- reduced MSE: reduced variance but possibly higher bias
- it is too "expensive" to use all variables

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables

Bayesian Model Choice

- Models for the variable selection problem are based on a subset of the $\mathbf{X}_1, \ldots \mathbf{X}_p$ variables
- Encode models with a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_p)$ where $\gamma_j \in \{0, 1\}$ is an indicator for whether variable $\mathbf{X}_j$ should be included in the model $\mathcal{M}_\gamma$. $\gamma_j = 0 \Leftrightarrow \beta_j = 0$

Bayesian Model Choice

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

Bayesian Model Choice

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

$$\mathbf{Y} \mid \alpha, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{1}\alpha + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 \mathbf{I})$$

  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.

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iint p(\mathbf{Y} \mid \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \boldsymbol{\gamma}, \sigma^2)p(\sigma^2 \mid \boldsymbol{\gamma})d\boldsymbol{\beta}\, d\sigma^2$$

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iint p(\mathbf{Y} \mid \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \gamma, \sigma^2)p(\sigma^2 \mid \gamma)d\boldsymbol{\beta}\, d\sigma^2$$

- Bayes Factor $BF[i:j]$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds = Bayes Factor $\times$ Prior odds

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iint p(\mathbf{Y} \mid \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \gamma, \sigma^2)p(\sigma^2 \mid \gamma)d\boldsymbol{\beta}\, d\sigma^2$$

- Bayes Factor $BF[i:j]$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds $=$ Bayes Factor $\times$ Prior odds

- Probability $\beta_j \neq 0$: $\sum_{\mathcal{M}_j : \beta_j \neq 0} p(\mathcal{M}_j \mid \mathbf{Y})$ (marginal posterior inclusion probability)

Posterior Probabilities of Models

- Posterior model probabilities

$$p(\mathcal{M}_j \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}{\sum_j p(\mathbf{Y} \mid \mathcal{M}_j)p(\mathcal{M}_j)}$$

  Marginal likelihod of a model is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = \iint p(\mathbf{Y} \mid \boldsymbol{\beta}_\gamma, \sigma^2)p(\boldsymbol{\beta}_\gamma \mid \gamma, \sigma^2)p(\sigma^2 \mid \gamma)d\boldsymbol{\beta}\, d\sigma^2$$

- Bayes Factor $BF[i : j]$

$$\frac{P(\mathcal{M}_i \mid \mathbf{Y})}{P(\mathcal{M}_j \mid \mathbf{Y})} = \frac{p(\mathbf{Y} \mid \mathcal{M}_i)}{p(\mathbf{Y} \mid \mathcal{M}_j)} \times \frac{P(\mathcal{M}_i)}{P(\mathcal{M}_j)}$$

  Posterior Odds = Bayes Factor × Prior odds

- Probability $\beta_j \neq 0$: $\sum_{\mathcal{M}_j : \beta_j \neq 0} p(\mathcal{M}_j \mid \mathbf{Y})$ (marginal posterior inclusion probability)

Prior Distributions

- Bayesian Model choice requires proper prior distributions on regression coefficients (exception parameters that are included in all models)

Prior Distributions

- Bayesian Model choice requires proper prior distributions on regression coefficients (exception parameters that are included in all models)
- Vague but proper priors may lead to paradoxes!

Prior Distributions

- Bayesian Model choice requires proper prior distributions on regression coefficients (exception parameters that are included in all models)
- Vague but proper priors may lead to paradoxes!
- Conjugate Normal-Gammas lead to closed form expressions for marginal likelihoods, Zellner's g-prior is one of the most popular.

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_\gamma \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}_\gamma^c \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}_\gamma^{c\prime} \mathbf{X}_\gamma^c)^{-1})$$

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\gamma} \boldsymbol{\beta}_{\gamma} + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_{\gamma} \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}^{c\,\prime}_{\gamma} \mathbf{X}^c_{\gamma})^{-1})$$

- Marginal likelihood of $\mathcal{M}_{\gamma}$ is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_{\gamma}) = C(1+g)^{\frac{n-p-1}{2}}(1 + g(1 - R^2_{\gamma}))^{-\frac{(n-1)}{2}}$$

where R^2_{γ} is the usual R^2 for model $\mathcal{M}_{\gamma}$ and C is a constant that is $p(\mathbf{Y} \mid \mathcal{M}_0)$ (model with intercept alone)

Zellner's g-prior within Models

Centered model:

$$\mathbf{Y} = \mathbf{1}_n \alpha + \mathbf{X}^c_{\boldsymbol{\gamma}} \boldsymbol{\beta}_\gamma + \epsilon$$

- Common parameters

$$p(\alpha, \phi) \propto \phi^{-1}$$

- Model Specific parameters

$$\boldsymbol{\beta}_\gamma \mid \alpha, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\phi^{-1}(\mathbf{X}^{c\,\prime}_{\boldsymbol{\gamma}} \mathbf{X}^c_{\boldsymbol{\gamma}})^{-1})$$

- Marginal likelihood of $\mathcal{M}_\gamma$ is proportional to

$$p(\mathbf{Y} \mid \mathcal{M}_\gamma) = C(1+g)^{\frac{n-p-1}{2}}(1+g(1-R^2_\gamma))^{-\frac{(n-1)}{2}}$$

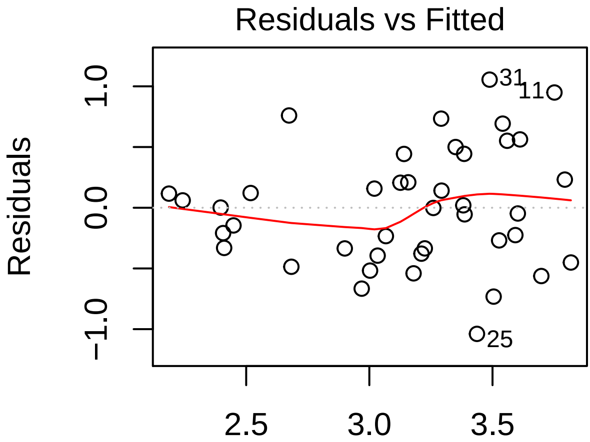
  where R^2_γ is the usual R^2 for model $\mathcal{M}_\gamma$ and C is a constant that is $p(\mathbf{Y} \mid \mathcal{M}_0)$ (model with intercept alone)
- uniform distribution over space of models $p(\mathcal{M}_\gamma) = 1/(2^p)$

USair Data: Enumeration of All Models

```
library(devtools)
suppressMessages(install_github("merliseclyde/BAS"))  # cu
library(BAS)
poll.bma = bas.lm(log(SO2) ~ temp + log(firms) +
                              log(popn) + wind +
                              precip+ rain,
                  data=usair,
                  prior="g-prior",
                  alpha=41,     # g = n
                  n.models=2^7,# enumerate (can omit)
                  modelprior=uniform(),
                  method="deterministic")   # fast enumera
```

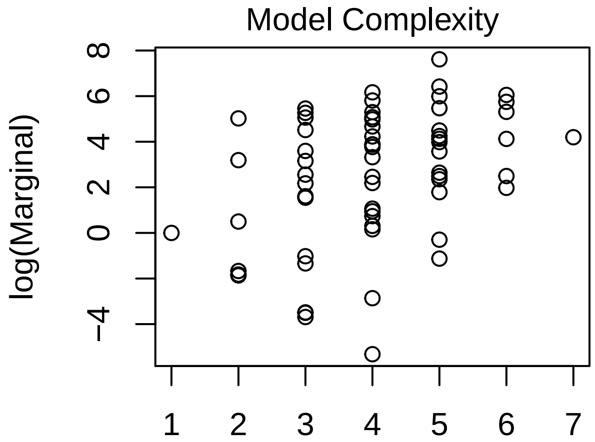
residual plot)

```
plot(poll.bma, which=1)
```



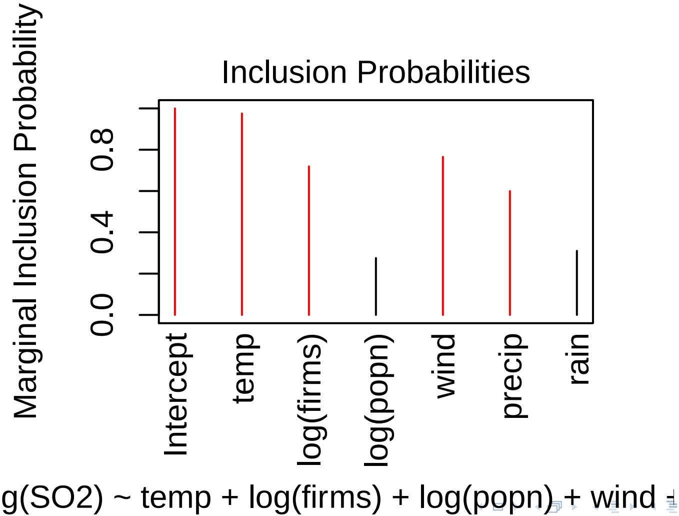
Model Complexity)

```
plot(poll.bma, which=3)
```



Inclusion Probabilities)

```
plot(poll.bma, which=4)
```



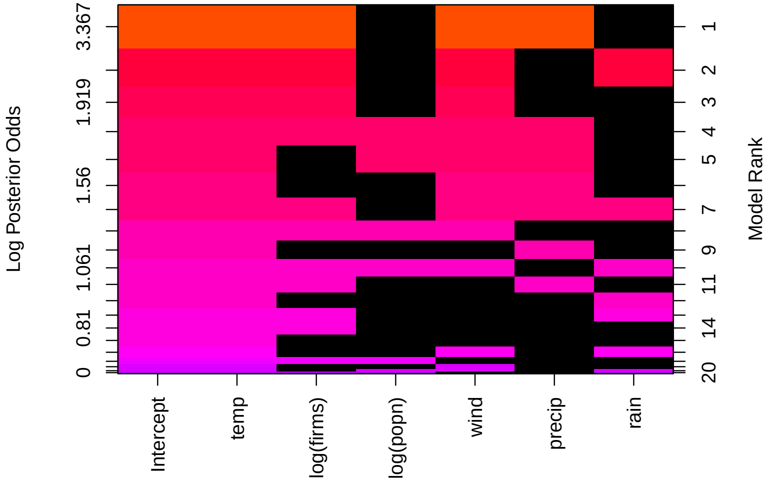
Model Space

```
summary(poll.bma)

##                  P(B != 0 | Y)  model 1   model 2   model 3
## Intercept            1.0000000 1.000000 1.0000000 1.0000000 1.
## temp                 0.9755041 1.000000 1.0000000 1.0000000 1.
## log(firms)           0.7190313 1.000000 1.0000000 1.0000000 1.
## log(popn)            0.2756811 0.000000 0.0000000 0.0000000 1.
## wind                 0.7654485 1.000000 1.0000000 1.0000000 1.
## precip               0.5993801 1.000000 0.0000000 0.0000000 1.
## rain                 0.3103574 0.000000 1.0000000 0.0000000 0.
## BF                          NA 1.000000 0.3022674 0.2349056 0.
## PostProbs                   NA 0.275800 0.0834000 0.0648000 0.
## R2                          NA 0.542700 0.5130000 0.4558000 0.
## dim                         NA 5.000000 5.0000000 4.0000000 6.
## logmarg                     NA 7.616228 6.4197847 6.1676565 6.
```

Summary

```
image(poll.bma)
```



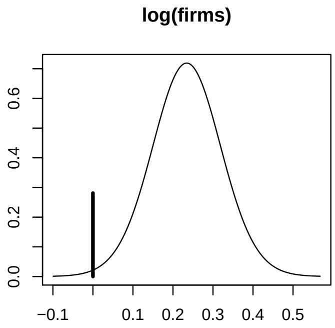
Coefficients

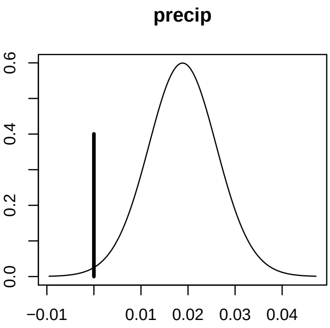
```
beta = coef(poll.bma, n.models=1)
beta

##
##  Marginal Posterior Summaries of Coefficients:
##
##  Using  BMA
##
##  Based on the top  1 models
##               post mean  post SD   post p(B != 0)
## Intercept     3.15300    0.07818   1.00000
## temp         -0.07130    0.01268   0.97550
## log(firms)    0.23428    0.08573   0.71903
## log(popn)     0.00000    0.00000   0.27568
## wind         -0.17998    0.06128   0.76545
## precip        0.01884    0.00729   0.59938
## rain          0.00000    0.00000   0.31036
```

Coefficients

```
par(mfrow=c(2,2));  plot(beta, subset=c(3, 6))
```





Bayesian Confidence Intervals

```r
confint(beta)
```

```
##                      2.5%       97.5%        beta
## Intercept     2.994993257  3.31101398  3.15300362
## temp         -0.096926645 -0.04567203 -0.07129934
## log(firms)    0.061014518  0.40753936  0.23427694
## log(popn)     0.000000000  0.00000000  0.00000000
## wind         -0.303835463 -0.05612195 -0.17997871
## precip        0.004105874  0.03357242  0.01883915
## rain          0.000000000  0.00000000  0.00000000
## attr(,"Probability")
## [1] 0.95
## attr(,"class")
## [1] "confint.bas"
```