Model Choice

Hoff Chapter 9, Clyde & George “Model Uncertainty” StatSci,
Hoeting et al “BMA” StatSci

November 2, 2017

Topics

- Variable Selection / Model Choice
- Stepwise Methods
- Model Selection Criteria
- Model Averaging

Variable Selection

Reasons for reducing the number of variables in the model:

- Philosophical
  - Avoid the use of redundant variables (problems with interpretations)
  - KISS
  - Occam's Razor
- Practical
  - Inclusion of un-necessary terms yields less precise estimates, particularly if explanatory variables are highly correlated with each other
- it is too "expensive" to use all variables

Variable Selection Procedures

- Stepwise Regression: Forward, Stepwise, Backward – add/delete variables until all t-statistics are significant (easy, but not recommended)
- Select variables with non-zero coefficients from Lasso
- Select variables where shrinkage coefficient less than 0.5
- Use a Model Selection Criterion to pick the "best" model
  - R2 (picks largest model)
  - Adjusted R2
  - Mallow's Cp $C_p = (\text{SSE}/\hat{\sigma}^2_{Full}) + 2p_m - n$
  - AIC (Akaike Information Criterion) proportional to Cp for linear models
  - BIC(m) (Bayes Information Criterion) $\hat{\sigma}^2_m + \log(n)p_m$

Trade off model complexity (number of coefficients p_m) with goodness of fit ($\hat{\sigma}^2_m$)

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!

Winner's Curse

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 $\mathcal{M}_\gamma$:

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

  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.

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

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

  Marginal likelihod of a model is proportional to

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

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

$$\hat{Y}^* | \mathbf{Y} = \sum_j p(\mathcal{M}_j | \mathbf{Y}) \hat{Y}_{\mathcal{M}_j}$$

Prior Distributions

- Bayesian Model choice requires proper prior distributions on parameters that are not common across 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 the most popular.

Zellner's g-prior

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 $\mathbf{X}^c = (\mathbf{I} - \mathbf{P}_{\mathbf{1}_n})\mathbf{X}$

- $p(\alpha) \propto 1$
- $p(\sigma^2) \propto 1/\sigma^2$
- $\boldsymbol{\beta}_\gamma \mid \alpha, \sigma^2, \boldsymbol{\gamma} \sim \mathsf{N}(0, g\sigma^2(\mathbf{X}^{c\prime}\mathbf{X}^c)^{-1})$

which leads to marginal likelihood of $\mathcal{M}_\gamma$ that is proportional to

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

where R^2 is the usual R^2 for model $\mathcal{M}_\gamma$ and C is the marginal distribution under the null model.

Trade-off of model complexity versus goodness of fit

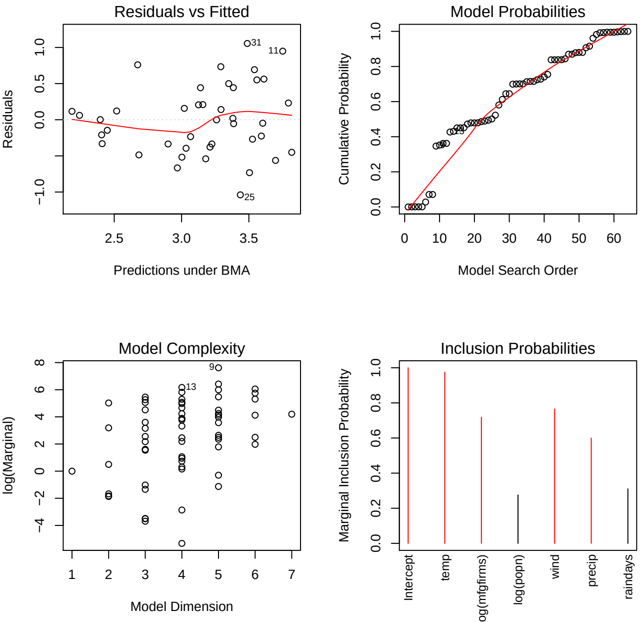
Lastly, assign prior distribution to space of models (Uniform, or Beta-binomial on model size)

USair Data

```
library(BAS)
poll.bma = bas.lm(log(SO2) ~ temp + log(mgfirms) +
                             log(popn) + wind +
                             precip+ raindays,
                  data=pollution,
                  prior="g-prior",
                  alpha=41,  # g = n
                  modelprior=uniform(),  # beta.binomial(1,
                  n.models=2^6,
                  update=50,
                  initprobs="Uniform")
```

```
par(mfrow=c(2,2))
plot(poll.bma, ask=F)
```

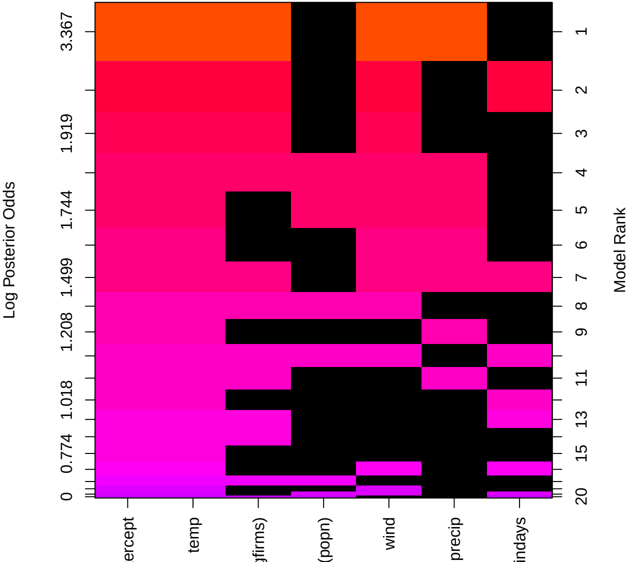
Plots



Residuals vs Fitted
Residuals
Predictions under BMA
31
11
25
Model Probabilities
Cumulative Probability
Model Search Order
Model Complexity
log(Marginal)
Model Dimension
9
13
Inclusion Probabilities
Marginal Inclusion Probability
Intercept
temp
log(mfgfirms)
log(popn)
wind
precip
raindays

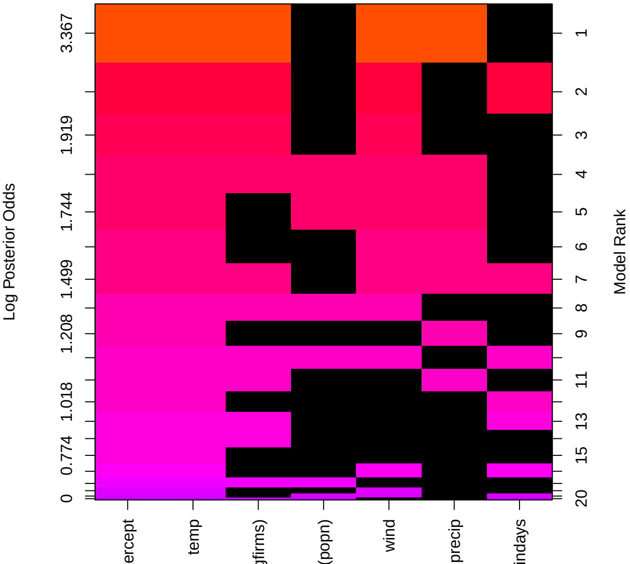
Posterior Distribution with Uniform Prior on Model Space

image(poll.bma)



Posterior Distribution with BB(1,p) Prior on Model Space

image(poll-bb.bma)



Jeffreys Scale of Evidence

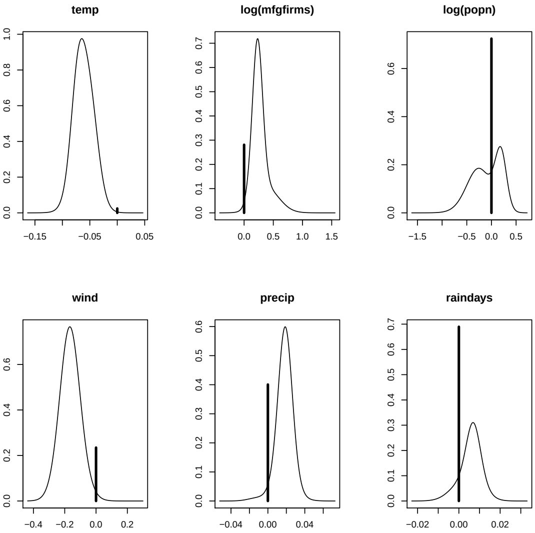
$B = BF[H_o : B_a]$

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

in context of testing one hypothesis with equal prior odds

Coefficients

```
beta = coef(poll.bma)
par(mfrow=c(2,3));  plot(beta, subset=2:7,ask=F)
```



Problem with g Prior

The Bayes factor for comparing $\mathcal{M}_\gamma$ to the null model:

$$BF(\mathcal{M}_\gamma : \mathcal{M}_0) = (1+g)^{(n-1-p_\gamma)/2}(1+g(1-\mathrm{R}^2))^{-(n-1)/2}$$

- Let g be a fixed constant and take n fixed.
- Let $F = \frac{R^2_\gamma/p_\gamma}{(1-R^2_\gamma)/(n-1-p_\gamma)}$ usual F statistic for comparing model $\mathcal{M}_\gamma$ to $\mathcal{M}_0$
- As $R^2_\gamma \to 1$, $F \to \infty$ LR test would reject H_0
- But BF remains bounded (contradiction)
- introduce prior on g - mixtures of g priors
- Jeffreys Zellner-Siow "JZS" Cauchy

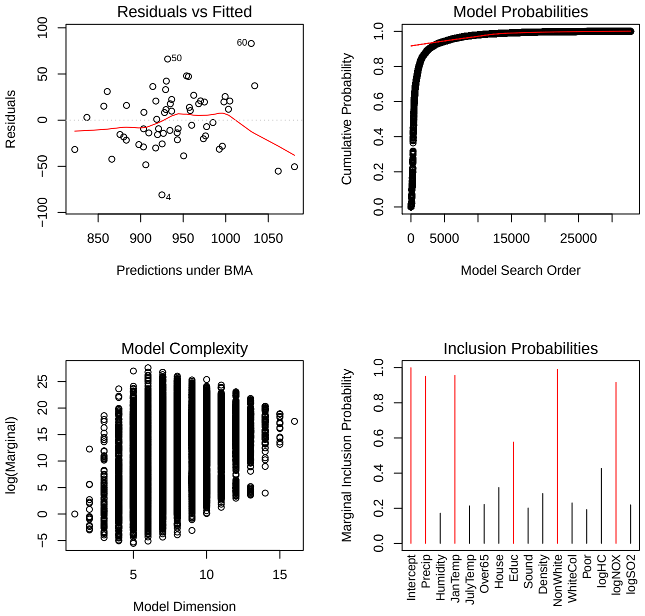
$$1/g \sim G(1/2, n/2)$$

Mortality & Pollution

- Data from Statistical Sleuth 12.17
- 60 cities
- response Mortality
- measures of HC, NOX, SO2
- Is pollution associated with mortality after adjusting for other socio-economic and meteorological factors?
- 15 predictor variables implies $2^{15} = 32,768$ possible models
- Use Zellner-Siow Cauchy prior $1/g \sim G(1/2, n/2)$

```
mort.bma = bas.lm(MORTALITY ~ ., data=mortality,
                  prior="ZS-null",
                  alpha=60,  n.models=2^15,
                  update=100, initprobs="eplogp")
```

Posterior Distributions



Posterior Probabilities

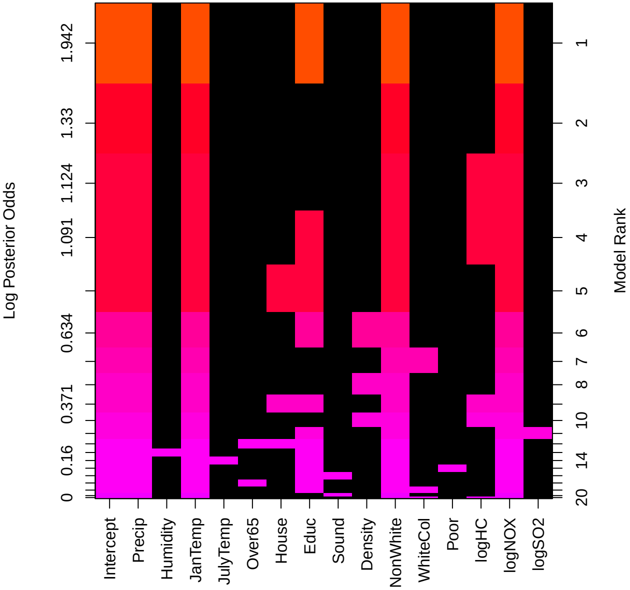
- What is the probability that there is no pollution effect?
- Sum posterior model probabilities over all models that include at least one pollution variable

```
> which.mat = list2matrix.which(mort.bma,1:(2^15))
> poll.in = (which.mat[, 14:16] %*% rep(1, 3)) > 0
> sum(poll.in * mort.bma$postprob)
[1] 0.9889641
```

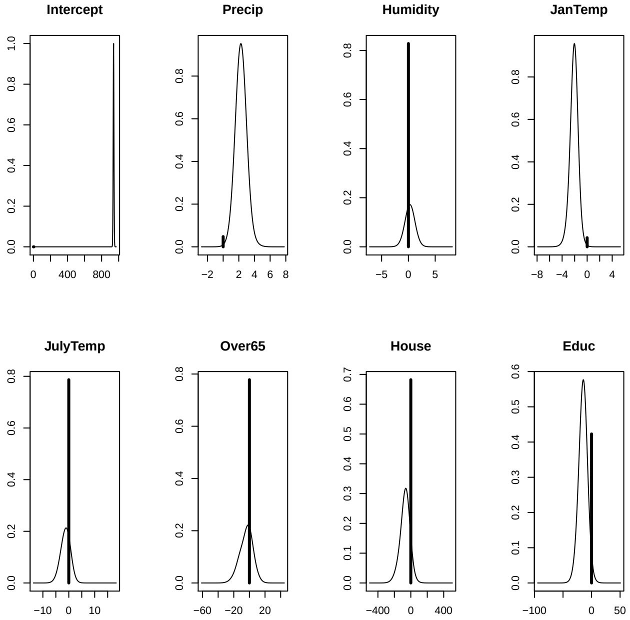
- Posterior probability no effect is 0.011
- Odds that there is an effect $(1 - .011)/(.011) = 89.9$
- Prior Odds $7 = (1 - .5^3)/.5^3$
- Bayes Factor for a pollution effect $89.9/7 = 12.8$
- Bayes Factor for NOXEffect based on marginal inclusion probability $0.917/(1 - 0.917) = 11.0$
- Marginal inclusion probability for logHC $= 0.4271$; BF $= 0.75$
- Marginal inclusion probability for logSO2 $= 0.2189$; BF $= 0.28$

Note Bayes Factors are not additive!

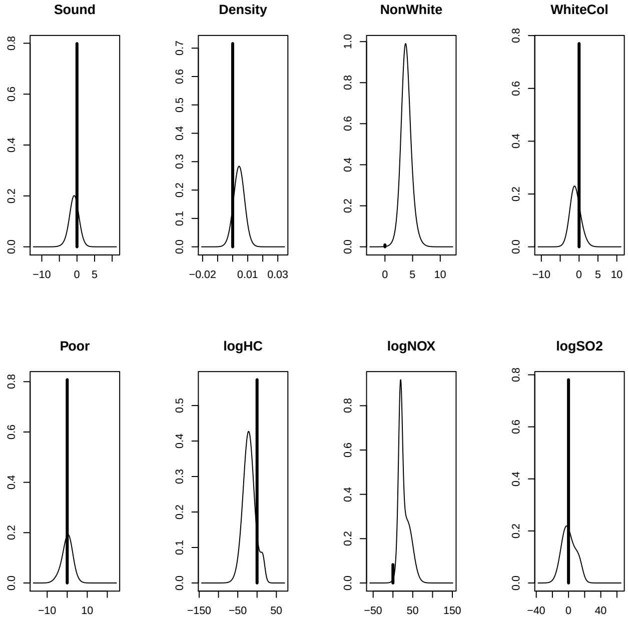
Model Space


Log Posterior Odds
1.942
1.33
1.124
1.091
0.634
0.371
0.16
0
Model Rank
1
2
3
4
5
6
7
8
10
14
20
Intercept
Precip
Humidity
JanTemp
JulyTemp
Over65
House
Educ
Sound
Density
NonWhite
WhiteCol
Poor
logHC
logNOX
logSO2

Coefficients



Coefficients



Effect Estimation

- Coefficients in each model are adjusted for other variables in the model
- OLS: leave out a predictor with a non-zero coefficient then estimates are biased!
- Model Selection in the presence of high correlation, may leave out "redundant" variables;
- improved MSE for prediction (Bias-variance tradeoff)
- Bayes is biased anyway so should we care?
- What is meaning of $\sum_{\boldsymbol{\gamma}} \beta_{j\boldsymbol{\gamma}} \gamma_j P(\mathcal{M}_\gamma \mid \mathbf{Y})$

Problem with confounding! Need to change prior?

Challenges

- Computational if $p > 35$ enumeration is difficult
  - Gibbs sampler or Random-Walk algorithm on γ
  - slow convergence/mixing with high correlations
  - Metropolis Hastings algorithms more flexibility
  - "Stochastic Search" (no guarantee samples represent posterior)
  - in BMA all variables are included, but coefficients are shrunk to 0; alternative is to use Shrinkage methods
- Prior Choice: Choice of prior distributions on $\boldsymbol{\beta}$ and on γ

Model averaging versus Model Selection – what are objectives?