Handling Covariate Uncertainty in Environmental Epidemiology and Risk Assessment

Louise Ryan,
Distinguished Professor of Statistics
University of Technology Sydney

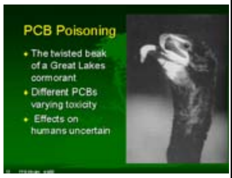
Drawing on the work of former Harvard PhD student, Melissa Whitney, now at Weil, Gotshall and Manges LLP, New York,

Presentation Outline

- Motivating application
- Challenges of model uncertainty
- How environmental risk assessment proceeds
- Limitations when applied to epidemiological data
- Bayesian model averaging as a tool for handling model selection uncertainty
- Comparison of some common approaches to BMA
- Discussion

Polychlorinated Biphenyls (PCBs)

- widely used early to mid 20th century in transformers, capacitors, and electric motors
- Health danger firmly established in 1970s with poisoning from contaminated rice oil, and evidence of carcinogenicity based on animal studies
- Though US banned production in 1979, PCBs considered a ubiquitous, persistent environmental pollutant.
- PCBs store in animal fat and bioaccumulate through food chain.
- 90% of current-day exposure via diet, especially dairy, meat, fatty fish



Impact of Chronic Low-level Exposure remains controversial

Well-designed studies yield different conclusions

- Methods to assess exposure/outcome differ
- Varying methods to analyze data and adjust for potential confounders (model selection)
- Levels of exposure vary across study
- Extreme observed exposures within a given study
- Concurrent exposures to other pollutants such as Mercury, Dioxins (possible effect modification)
- Actual chemical exposure different (209 PCB congeners)
- Beneficial effects of supportive environment may protect some population subgroups

Duesseldorf Cohort
Walkowiak et al (2001)

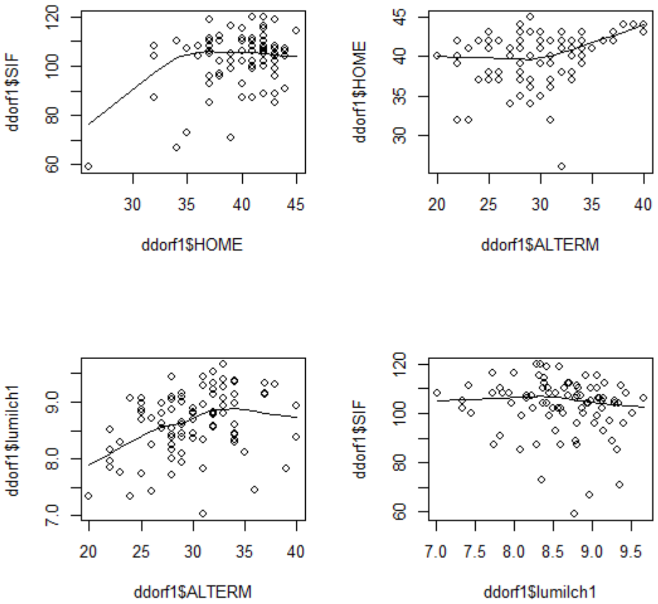
- Cohort study of 71 mother/infant pairs
- Our subset – 88 mothers who breastfed
- Exposure: log(sum of 3 main PCB congeners * duration of breastfeeding)
- Outcome: Kaufman Assessment Battery for Children (at 42 months old)
- Other possible covariates: maternal age, BMI, alcohol consumption, socioeconomic status, HOME score, and gender

Cohort characteristics

	mean	std dev
Kaufman Assessment Battery for Children (K-ABC)	102.86	11.286
log(PCB) in breast milk	8.625	0.566
HOME	39.91	3.327
BMI	24.81	4.563
SES	12.77	2.707
Maternal Age	29.92	4.297
Child's Gender: Male Female	$n = 56$ $n = 32$	$p = 0.64; 0.36$
Maternal Alcohol Consumption: Yes No	$n = 53$ $n = 35$	$p = 0.60; 0.40$

Covariates tend to be quite correlated

Lots of noise, lots of correlation



Modeling Framework

Full model:

$$Y_i = \beta_0 + \beta_1[\log(PCB_i) * duration] + \beta_2(HOME_i) + \beta_3(BMI_i) + \beta_4(gender_i) + \beta_5(maternal\ alcohol_i) + \beta_6(SES_i) + \beta_7(maternal\ age_i) + \varepsilon_i$$

where: $\varepsilon_i \sim N(0, \sigma^2)$

P= 7 possible predictors → $2^7 = 128$ possible models, assuming no interactions or other functions of covariates

Traditional Approach

- Consider variety of plausible models, select one, and draw inferences from single, "final" model ignoring plausible alternatives
- Underestimates true variability and uncertainty due to model selection process
- Results in over-confident, risky decision-making (Draper, 1995)

Illustration with a simple case

Suppose we

1) Fit $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
2) Test $H_0 : \beta_1 = 0$
3) Only if we reject H_0, then calculate a confidence interval on β_1

What properties do we expect the confidence interval to have?

Illustration, continued

True coverage probability is

$$\Pr(\text{Confidence interval includes true } \beta \,|\, \text{rejected } H_0 : \beta = 0)$$
$$= \Pr(\hat{\beta} - se(\hat{\beta})Z_{1-\alpha/2} < \beta < \hat{\beta} + se(\hat{\beta})Z_{1-\alpha/2} \,\big|\, |\hat{\beta}| > se(\hat{\beta})Z_{1-\alpha/2})$$
$$= 1 - \alpha / Power(\beta)$$

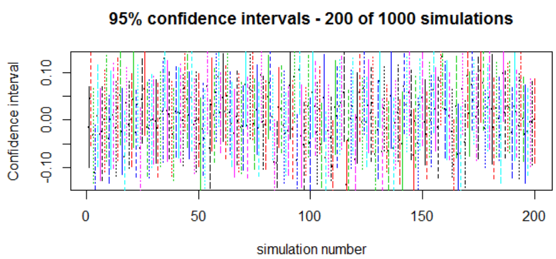
where *Power*(β) is Pr(reject H_0| β is true value)

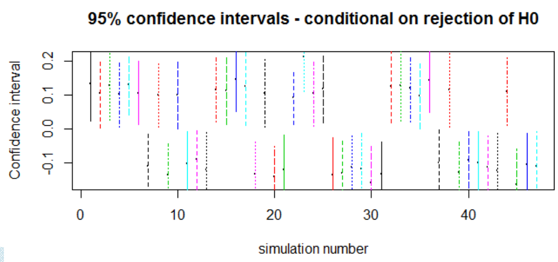
Note that coverage probability is 0 when β =0, and approaches α only as β grows large

Simulation for β=0

Unconditional confidence intervals look as expected

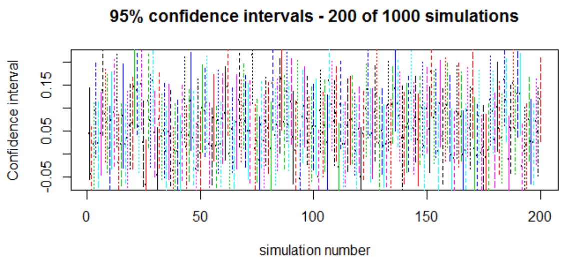
None of the conditional confidence intervals cover 0



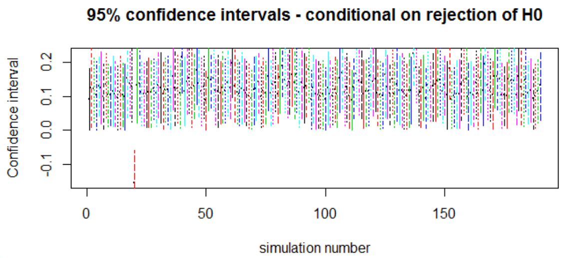


Simulation for medium β (.05)

Unconditional confidence intervals look as expected

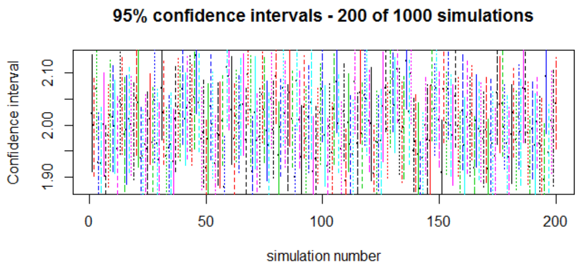


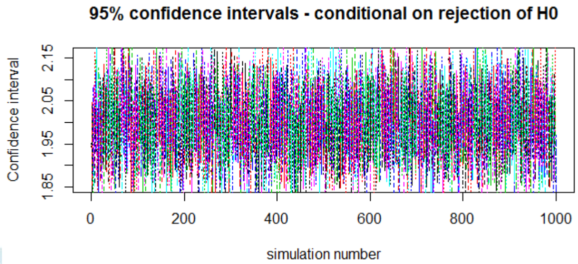
Most of the conditional confidence intervals cover β=.05, but are clearly skewed



Simulation for large β(2)

Unconditional and conditional confidence intervals are very similar





Implications in practice

Traditional practice of picking a best model then reporting confidence intervals for coefficients may be biased, especially in settings where there is a high degree of noise.

Implications particularly problematic in environmental risk assessment where we use estimated coefficient of exposure of interest to predict a "safe" dose. Lets take a brief diversion to see how this works.

Environmental Risk Assessment – estimating a Benchmark Dose (BMD)

Step 1: Establish that exposure of interest has an adverse effect

Step 2: Benchmark Dose (BMD) solves

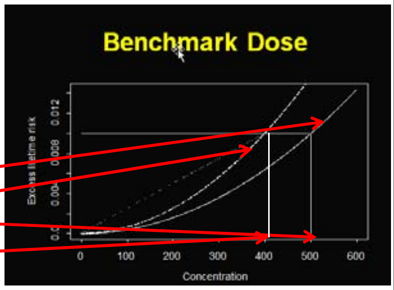
$$P(d) - P(0) = BMR$$

where P(d) is probability of an adverse effect and BMR (benchmark response) = 0.01, 0.05 or 0.1

Step 3: Compute lower confidence limit, BMDL

Step 4: Linearly extrapolate

Estimated dose response curve
Upper limit on dose response curve
BMD
BMDL



Benchmark Dose estimation for continuous outcomes

Need to specify a threshold level of Y that can be considered adverse. E.g.

- Pick a cutoff that has some clinical meaning
  - Scoring below 75 on an IQ test
  - Having a BMI above 25 (or 30?)
  - Scoring below 85 on K-ABC assessment
- Pick a cutoff that corresponds to a lower (or upper) percentile (usually 1% or 5%) of general population

To do analysis, can dichotomize outcomes (inefficient) or do regression analysis and then compute the tail probabilities

BMDs derived from linear regression

Suppose $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

Where X_1 is exposure of interest and X_2 is a confounder. If lower values of outcome are "adverse", then

$$P(c) = P(Y < c | X_1 = x_1, X_2 = x_2) = \Phi\left(\frac{c - \beta_0 - \beta_1 x_1 - \beta_2 x_2}{\sigma}\right)$$

And solution to *P(c)-P(0)=BMR* is:

$$BMD = \frac{Q\sigma}{\beta_1}$$

where $Q = \Phi^{-1}(P(0)) - \Phi^{-1}(P(0) + BMR)$

And solution is the same for all values of the confounder, X_2

Lower Bound: BMDL

Where lower outcome are adverse, Budtz-Jorgensen (2001) provides an approximation to BMDL that accounts for fact that variance must be estimated:

$$BMDL = \frac{Q\sigma}{\hat{\beta}_{\log(PCB)} + u_{05}\widehat{SE}(\hat{\beta}_{\log(PCB)})\sqrt{1+(t^2-u_{05}^2)/2df}}$$

Where $t = \frac{(\hat{\beta}_{\log(PCB)})}{\widehat{SE}(\hat{\beta}_{\log(PCB)})/\sqrt{df}}$ and $u_{05} = -1.645$

BMD estimation in the context of model uncertainty

Consider

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

where X_{i1} is exposure of interest and rest are confounders. Estimated coefficient (and hence BMD) sensitive to model choice.

Model selection a big topic in stats literature

- Traditional (stepwise etc)
- Penalty-based approaches such as the Lasso
- Bayesian approaches that place a mixture prior on each β
- Model averaging

Bayesian model averaging

Let $\boldsymbol{M} = \{M_{1,} M_2, \ldots M_K\}$ be the family of models over which we will average. For PCB data, $K = 2^7 = 128$

$\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_{\beta_k})$ parameters characterizing k^th (multiple) linear regression model

$p(M_k)$ prior probability of the k^th model

$p(\beta_k | \sigma^2, M_k)$ prior probability model for the regression coefficients of M_k, where $\boldsymbol{\beta_k}$ is $p \times 1$ matrix, and $p = \dim(M_k)$

We'll come back presently to details (e.g. hyperpriors)

BMA, Continued

Given priors and normal data likelihood model, for models k = 1,2,…K, posterior model probability for M_k is:

$$p(M_k \mid Data) = \frac{p(Data \mid M_k) p(M_k)}{\sum_{l=1}^{K} p(Data \mid M_l) p(M_l)}$$

where $p(Data \mid M_k) = \int p(Data \mid \boldsymbol{\theta_k}, M_k) p(\boldsymbol{\theta_k} \mid M_k) d\boldsymbol{\theta_k}$

Last piece (posterior marginal likelihood) can be hard to compute.

We'll discuss approaches in a moment.

We'll compare three approaches to estimation of a BMD and BMDL

- 1. Traditional Method: 2-step process:
  - Fit Model, Test hypothesis
  - If reject, then calculate BMD
- 2. Compute Single, Model-Averaged BMD and BMDL via Formula: $\Sigma\, \Delta_k p(M_k|\text{Data})$
- 3. Using Posterior Model Distribution: to summarize empirical Distribution of BMD/BMDL

Use Posterior Model Probabilities to Estimate Averaged Quantities of Interest

After calculating posterior probabilities and using classical procedures to estimate any quantity for a given model, Δ_k, use model weights to obtain the BMA estimate of its *unconditional* expectation, variance

$$\hat{\Delta}_{BMA} = \sum_k \hat{\Delta}_k \hat{p}(M_k \mid D)$$

$$\hat{V}_{BMA}(\hat{\Delta}) = \underbrace{\sum_k \hat{V}[\hat{\Delta} \mid M_k]\hat{p}(M_k \mid D)}_{\text{Estimated via a classical procedure or bootstrap}} + \overbrace{\sum_k (\hat{\Delta}_k - \hat{\Delta}_{BMA})^2 \hat{p}(M_k \mid D)}^{\text{Model selection uncertainty component}}$$

Empirical Distribution of BMD/BMDL

- Provides helpful “Big picture” view, integrating estimation and testing into one step
- No need to conclude with a single model/estimate (even a model-averaged one)
- Depending on approach, can use MCMC samples or simulate data based on posterior model probabilities and parameter estimates to explore entire risk distribution

First we need to discuss Approaches to BMA

- Closed Form Solution (rarely exists in practice)
- Approximations to Posterior Model Probability
  - E.g. BIC approximation (Raftery, 1996)
- MCMC Methods, such as:
  - Carlin and Chib Approach
  - Reversible Jump MCMC (RJ MCMC)
  - Stochastic Search Variable Selection (SSVS)
  - Gibbs Variable Selection (GVS)

Assumed Model and Priors

$p(\sigma^2 | M_k)$: prior variance

$p(Data \mid \boldsymbol{\theta}_k, M_k)$: likelihood for model $k = 1,2, \dots K$

Priors (adopting those of Hoeting et al 1999):

- $p(M_k) = K^{-1}$ for all k
- $\boldsymbol{\beta}_k \mid \sigma^2,\ M_k \sim N_p(\boldsymbol{\mu}, \boldsymbol{V}\sigma^2)$
- $\sigma^2 | \ M_k \sim Inv - Gamma(\nu,\ \lambda)$

Closed Form Solution

- Assuming normal likelihood and restricting priors to certain conjugate distributions:
  - Normal priors on Betas
  - Variance prior: inverse-chi-squared distribution
- Results in marginal posterior distribution of the data Pr(Data|M_k) following an n-dimensional non-central Student's t distribution

BIC Approximation

$$BIC = 2\log pr(D \mid \hat{\theta}_k, M_k) - (p_k/2)\log(n)$$
$$p(M_k \mid D) = \exp(0.5BIC_k) / \sum \exp(0.5BIC_i)$$

where:
- p_k is dimension of model k
- n = sample size
- Pr(D|model, estimates) is maximized likelihood for M_k

GVS Method

Family of multiple linear regression models can be written:

$$Y = \sum_{j=1}^{p} g(j) X_j \beta_j + \varepsilon$$

where $g(j) = 1$ if j^{th} variable is included in model

- Introducing variable indicator function, g, reduces framework to one of fixed dimensionality
- Now, can utilize standard simulation techniques to estimate g and other parameters

GVS (continued)

- Framework implemented in WinBUGS (via R plug-in R2Winbugs)

$g(j) \sim Bernoulli(0.5)$ for $j = 1,2, \dots P$

$\beta_j \sim N(0, \sigma^2 V)$

$\sigma^2 \sim Inv - Gamma(\nu, \lambda)$

And likelihood: $Y \sim N(\sum_{j=1}^{P} g(j) X_j \beta_j ,\ \sigma^2 I)$

Reversible Jump MCMC

Basic idea/process:

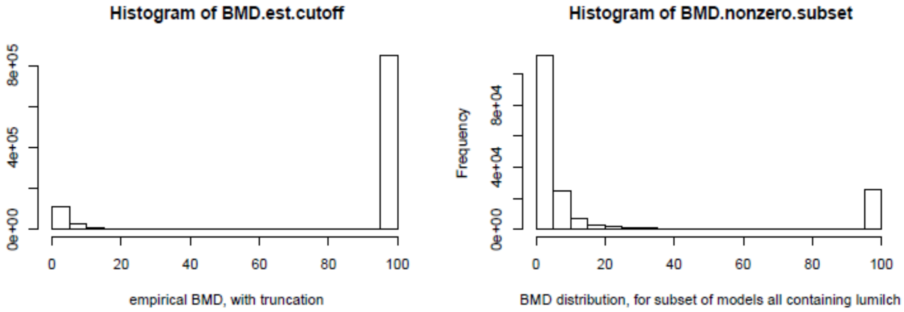
- Given starting model M, propose jump to new model M* that differs by adding/deleting 1 variable
- In this case, used jump probability j(M|M*) = j(M*|M) = 1/P for all models
- Generate series of 1-to-1 deterministic functions that allow us to jump between model spaces of differing dimensions (merely tool/construct so that MCMC theory principles hold)
- Accept move with probability somewhat proportional to ratio of marginal likelihoods of the data under M* vs. M

Posterior Model Probabilities, Calculation Methods Compared

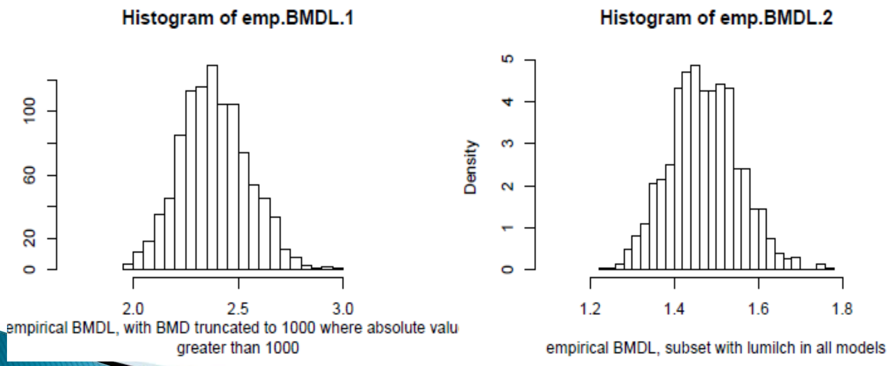
M_k	$\Pr(M_k \mid Data)$			
	Exact	BIC Approx	GVS	RJ MCMC
(int, HOME)	0.2453	0.2561	0.506	0.2652
(int)	0.1887	0.0152	0.104	0.1997
(int, log(PCB), HOME)	0.0648	0.1809	0.136	0.0587
(int, HOME, maternal.age)	0.0464	0.0842	0.056	0.0441
(int, gender)	0.0384	0.0058	0.017	0.0421
(int, HOME, gender)	0.0368	0.0538	0.037	0.0330
(int, log(PCB))	0.0295	0.0035	0.009	0.0256
(int, HOME, maternal.alcohol)	0.0281	0.0310	0.017	0.0428
(int, HOME, SES)	0.0277	0.0274	0.015	0.0227
(int, HOME, BMI)	0.0275	0.03033	0.0208	0.0237
total posterior prob of top 10	0.7331	0.6872	0.925	0.7180

Empirical BMD Distributions

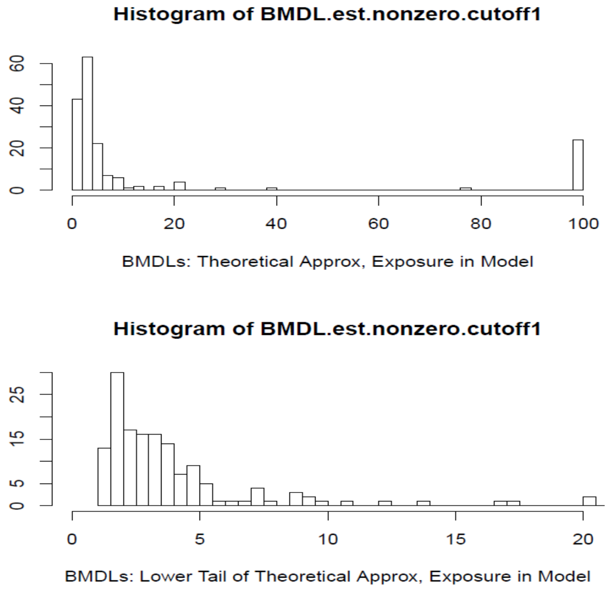
Left hand panel: All Models
Right hand panel: Only models that include exposure Covariate Inclusion in Model (Infinite/Extreme Values Truncated at 100)



Empirical BMDL Distributions: (1) unconditional model posterior distribution, and (2) conditional on exposure being included in the model



BMDLs based on Theoretical Approximation (Budtz-Jorgensen)



Posterior inclusion probabilities

- Posterior probability for risk factor inclusion (i.e. probability of a nonzero effect size or "weight" for the jth covariate):

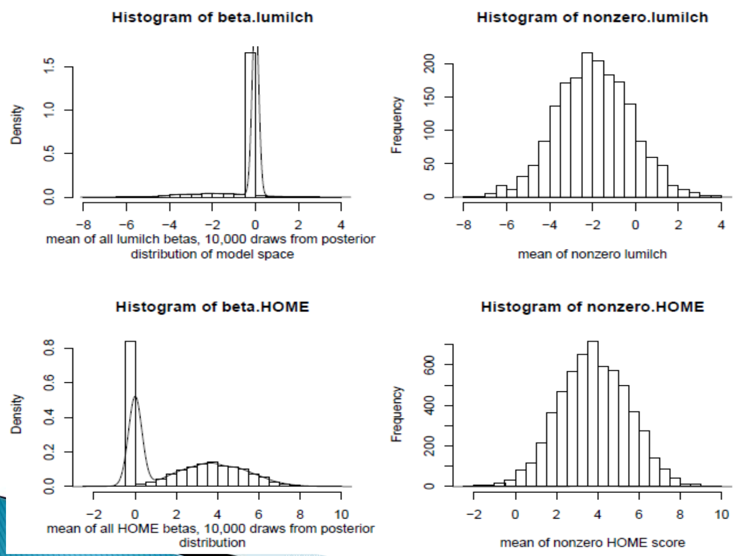
$$\Pr(\beta_j \neq 0 \,|\, Data) = \sum_{M_i : j \in M_i} p(M_i \,|\, Data)$$

- To extent that priors convey a pre-data sense of uncertainty as to inclusion/exclusion of a covariate, these reflect post-data uncertainty

Covariate Posterior Probabilities

Covariate	Method			
	Exact	BIC Approx	GVS	RJ MCMC
log(PCB)	0.1797	0.3867	0.181	0.1799
HOME	0.5867	0.9569	0.837	0.5888
BMI	0.1044	0.1155	0.042	0.1088
gender	0.1473	0.1772	0.072	0.1508
maternal alcohol	0.1022	0.1196	0.036	0.1060
SES	0.1137	0.1056	0.036	0.1202
maternal age	0.1344	0.2100	0.087	0.1370

Examining the whole posterior distribution



Use Posterior Model Probabilities to Estimate Averaged Quantities of Interest

Bayesian Model-Averaged Estimates of Relationship between log(PCB) exposure (standardized) and test score, all 4 methods

BMA Technique	$\widehat{\beta}_{\log(PCB)}(BMA)$	$var(\widehat{\beta}_{\log(PCB)}(BMA))$	$\Pr(\beta_{\log(PCB)} \neq 0)$
Exact	-0.3553	1.1067	0.1797
BIC Approx	-0.8538	1.7759	0.3867
GVS	-0.4005	2.9205	0.181
RJ MCMC	-0.3639	1.1414	0.1799

Single, Model-Averaged BMD/BMDL

$$\hat{\Delta}_{BMA} = \sum_k \hat{\Delta}_k \hat{p}(M_k \mid D)$$

$$\hat{V}_{BMA}(\hat{\Delta}) = \sum_k \hat{V}[\hat{\Delta} \mid M_k]\hat{p}(M_k \mid D) + \sum_k (\hat{\Delta}_k - \hat{\Delta}_{BMA})^2 \hat{p}(M_k \mid D)$$

BMA Technique	$\widehat{\beta}_{\log(PCB)}(BMA)$	$var(\widehat{\beta}_{\log(PCB)}(BMA))$	BMD	BMDL
Exact	-0.3553	1.1067	11.5406	1.9458
BIC Approx	-0.8538	1.7759	4.8023	1.3348
GVS	-0.4005	2.9205	10.2382	0.7793
RJ MCMC	-0.3639	1.1414	11.2678	1.9142

Concluding remarks

- BMA can be used to find the full empirical distribution of BMDs, BMDLs or other quantities, which captures (1) model uncertainty and (2) parameter uncertainty
- Lots of interesting questions
  - Enlarging model space
  - Sensitivity to model space specification
  - Better approximate solutions
  - Improving the MCMC performance
  - Theoretical properties of BMD, BMDL – does it solve the two-stage problem of the traditional approach?

Closed Form Solution (Raftery et al. 1997)

assuming a likelihood of the form:

$$\mathbf{Y_{n\times 1}} = \mathbf{X}\boldsymbol{\beta}_{\mathbf{p\times 1}} + \boldsymbol{\varepsilon}_{\mathbf{n\times 1}} \text{ with } \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_{n\times n})$$

and restricting priors to certain conjugate distributions:

$\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}, \sigma^2 V)$ and

$\frac{\upsilon\lambda}{\sigma^2} \sim \chi^2_\upsilon$

Closed Form Solution (Continued)

results in marginal posterior distribution of the data,

$p(Data \mid M_k) = \int p(Data \mid \boldsymbol{\theta}_k, M_k)p(\boldsymbol{\theta}_k \mid M_k)d\boldsymbol{\theta}_k,$

following an n-dimensional, non-central Student's t distribution with υ degrees of freedom,

mean $X\mu$ and variance $\upsilon/(\upsilon-2)\lambda(I + X_k V_k X_k')$ [Hoeting et al. 1999]

Closed Form Solution (Continued)

i.e.

$$p(Data \mid M_k) = \frac{\Gamma(\frac{\upsilon+n}{2})(\upsilon\lambda)^{0.5\upsilon}}{\pi^{0.5n}\Gamma(\frac{\upsilon}{2}) \mid I + X_k V_k X_k' \mid^{0.5}} \times [\lambda\upsilon + (Y - X_k\mu_k)'(I + X_k V_k X_k')^{-1}(Y - X_k\mu_k)]^{-0.5(\upsilon+n)} \quad (3)$$

Reversible Jump MCMC [Green 1995]

General Algorithm for RJ MCMC [Gelman et al. 2004]

- Given starting model M, propose new model M^* that differs by adding/deleting 1 variable
- Generate a vector $\mathbf{u}$ from a continuous proposal density $q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)$
- Generate a series of one-to-one deterministic functions with:
  - $(\boldsymbol{\beta}^*_{M^*}, \mathbf{u}^*) = g_{M,M^*}(\boldsymbol{\beta}_M, \mathbf{u})$

RJ MCMC (Continued)

- $p_M + \text{dim}(\mathbf{u}) = p_{M^*} + \text{dim}(\mathbf{u}^*)$
- functions merely tool to match dimensions when jumping from M to M^*
- note M, M^*differ in dimension by one (1)

RJ MCMC (Continued)

- Accept proposed move to M^* with acceptance probability:

$$\alpha = min\{1, \frac{\Pr(Y \mid M^*) \Pr(\boldsymbol{\beta}^*_{M^*} \mid M^*) \Pr(M^*) j(M \mid M^*) q(\mathbf{u}^* \mid \boldsymbol{\beta}^*_{M^*}, M^*, M)}{\Pr(Y \mid M) \Pr(\boldsymbol{\beta}_M \mid M) \Pr(M) j(M^* \mid M) q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)} \times \det(\frac{\partial g_{M,M^*}(\boldsymbol{\beta}_M, \mathbf{u})}{\partial(\boldsymbol{\beta}_M, \mathbf{u})})\}$$

RJ MCMC (Continued)

Further Simplification of Algorithm [Suggested by Clyde in commentary for Hoeting et al. 1999]

- Given the normal prior on $\boldsymbol{\beta}_M$ for all 128 models, the posterior distribution $\boldsymbol{\beta}^*_{M^*} \mid M^*$ is available in closed form
- take $q(\mathbf{u} \mid \boldsymbol{\beta}_M, M, M^*)$ to be the posterior distribution of $\boldsymbol{\beta}^*_{M^*} \mid M^*$
- acceptance probability for jump from M to M^* simplifies to:

$$\alpha = min\{1, \frac{\Pr(Y|M^*) \Pr(M^*) j(M|M^*)}{\Pr(Y|M) \Pr(M) j(M^*|M)} \times \det(\frac{\partial g_{M,M^*}(\boldsymbol{\beta}_M, \mathbf{u})}{\partial(\boldsymbol{\beta}_M, \mathbf{u})})\}$$

Ways to Quantify/Depict BMD and BMDL

- 1. Traditional Method: 2-step process:
  - Fit Model, Test hypothesis
  - If reject, then calculate BMD
- 2. Compute Single, Model-Averaged BMD
  - Formula: $\Sigma\ \Delta_k p(M_k|\text{Data})$
- 3. Using Posterior Model Distribution:
  - Simulate Data,
  - Fit Model,
  - Estimate BMD, and
  - Repeat to Examine Empirical Distribution of BMD/BMDL