Resource Selection by Scientists

Navigating the model selection and multi-model landscape toward question-focused modeling

****

Considering the "Selection" of Model Averaging

Megan D. Higgs & Katharine Banner (Ph.D. student)

Department of Mathematical Sciences
Montana State University, Bozeman

The allures of model averaging

- Incorporate uncertainty in model choice into analysis
  - Leamer (1978) "presumably, ambiguity about model selection should dilute information about effect sizes and predictions" since "part of the evidence is spent to specify the model"
- Lessen desire to search through all models until find a large enough "effect"
- Improve predictive ability
- Try to overcome the bad press given to choosing a single model for inference
- It's the new cool statistical method that might help get work published!

The allures...cont.

Some quotes from (Montgomery & Nyhan, 2010) in *Political Analysis*

- "...BMA can help applied researchers to ensure that their estimates of the effects of key independent variables are robust to a wide range of possible model specifications."
- **not** using BMA "systematically understates the uncertainty of our results, generates fragile model specifications, and leads to the estimation of bloated models with too many control variables"
- "Basing inferences on a single model implicitly assumes that the probability that the reported model generated the data is 1, an assumption that is surely mistaken."

Reviewer comments

Anonymous reviewer comment (August 2014)

- "If possible, you could also conduct model selection and/or model averaging among competing models"
- And, suggested putting the explanatory variables that were designed to measure the same underlying trait into a single MLR model, because then "the coefficients are unbiased, so you can trust them"
  - Comments show lack of understanding of goals of model averaging and in general a lack of understanding of multiple linear regression and what we can and cannot accomplish by including variables together in the same model.

Rough progression of work

1. Model specification should be included in uncertainty idea
2. Posterior calculations - pre-MCMC
3. Gibbs sampling, Stochastic Search Variable Selection (SSVS)
4. Use of BIC to get model weights (made it accessible and easy) - then AIC too.
5. Priors to aid posterior calculations - conjugacy
6. Investigation of different priors and hyper-parameters (g-prior)
7. Automatic Reversible Jump MCMC
8. BMA, BMS, and BAS packages
9. Investigation of different priors, hyper-priors and hyper-parameters (extensions to g-priors)

Focus has been on computational and prior specification issues and improving prediction, but very little work on interpretation of coefficients post - BMA.

Let's restrict our context...

Multiple linear regression (MLR) - same model form, different covariates.

- Inferential goal: incorporate variables we are directly interested in AND variables we would like to "control" for.
    - Interpretation rather than prediction
    - Covariates do not have to be orthogonal (and rarely are in real life)
- This is the most common situation *we* have seen people suggest the use of model averaging.

For this context, what things should we consider to assess the appropriateness and/or usefulness of model averaging? When is it a smart thing to do?

Back to linear model basics...

Two things I remember from my first "methods" course:

1. In MLR, a regression parameter is actually called a "partial regression coefficient"
   - i.e. Meaning of a regression coefficient depends on the other variables in the model

2. Prediction and explanation are different modeling goals
   - Larry Wasserman (Gelman's blog, 06-02-2014): "Of course one should not the use the output of this (or any selection method) for inference. But for prediction it is great." ("I meant things like p-values after selecting variables by stepwise.")

Both of these are relevant when considering the "selection" of model averaging as inferential tool.

PARTIAL regression coefficients

What does the PARTIAL mean?

- The relationship between X and the mean of the response for *fixed values of the other explanatory variables*.
- The meaning of the regression coefficient associated with a particular variable **depends** on what other variables are in the model

****

$$\mu\{y|X1\} = \beta_0 + \beta_1 X1$$

$$\mu\{y|X1, X2\} = \beta_0 + \beta_1 X1 + \beta_2 X2$$

Does β_1 mean the same thing in both of these models if X_1 and X_2 are not orthogonal?

Sloppiness with regard to partial coefficients

- Sloppy notation:
  - Subscripts should denote which model the coefficient is part of:

$$\beta_{1.1} \text{ and } \beta_{1.12} \quad \text{instead of } \beta_1$$
$$\beta_{0.1} \text{ and } \beta_{0.12} \quad \text{instead of } \beta_0$$

- Sloppy interpretation:
  - The "coefficient of X1" is often referred to as "the effect of X1" regardless of model.
  - Often ignore the larger context of the model in interpretation

Transferring the sloppiness to model averaging?

What does it mean to average "the effect of X1" over many models?

- Assuming the estimated "coefficient of X1" is different across models and want to report an "effect" that falls somewhere in the middle?
- Assuming the estimated "coefficient of X1" is the same across different models (posterior distributions centered in same place) and want to increase reported uncertainty to account for uncertainty across models?

What are the implications of going after *one* posterior distribution (or point estimate and confidence interval) for the "coefficient of X1"? Are there hidden assumptions?

General advice regarding averaging...

"Only average things that are measuring the same thing"

- How do we reconcile this advice in the context of model averaging across models with non-orthogonal covariates?
  - When is it reasonable to average and when is it unreasonable?
  - Weighted averages - result of averaging depends on weights used

When are we in danger of trying to combine apples and oranges?

What do we get?

It depends....

- Consider interaction in a 2-way design. Do we average "effects" in the presence of an interaction?
- Consider a model by coefficient interaction:
  - Inference about the "effect of X1" *depends* on what other variables we account for
  - Are there practically meaningful differences in results among models?
- Should we average over models in the presence of such an interaction?
  - How close do individual model inferences have to be?
  - Does MA deal with this through weighted average via model weights?
  - Is it really possible to deal with this "automatically"?

Multicollinearity?

- Often given a bad rap (multicollinee-itis), but fear of multicollinearity should be balanced with what needs to be "controlled" for to best answer the research question and what methods are going to be used
  - Should decision about whether to include related variables be completely driven by automatic methods?
- Model averaging originally presented as a way to improve prediction
- Not enough attention has been given to the case where $\mathbf{X}$ is not orthogonal.

Quotes from paper selling MA

Montgomery & Nyhan (2010), *Political Analysis*

> "..., frequentist hypothesis testing offers no method for resolving conflicting findings across alternative specifications. What is one to infer if a variable is significant in some specifications but fails to pass traditional thresholds in others?"

> "..., it is possible to have variables that are "statistically significant" (i.e. their credible intervals do not overlap with zero) but that have low posterior probabilities of inclusion. Likewise, it is possible for a variable with a higher posterior probability of inclusion to have a model-averaged credible interval that overlaps with zero due to variation in sign and significance across models."

Is MA really a *fix* to these "problems" or just a cover-up?

How often does the second really happen in practice?

- If estimated coefficient changes dramatically under different models, do we really want to model average?
- The averaging depends on the **weights**:
  1. Both models have "large" weight
     - What does the coefficient mean? What does the increased uncertainty really mean?
  2. One model has much larger weight than the other
     - Essentially just reporting the results from a single model?

Insights from Simpson's Paradox discussions?

Can we glean wisdom from discussions of Simpson's Paradox and/or the Ecological Fallacy?

- Ecological Fallacy - Do group level relationships hold for individuals?
- When do we want individual level results (conditioning on appropriate variables to get there) and when do we want group level (no conditioning)?
  - Is it appropriate to average the two? What lies in between the group and the individual?
  - Can this be automatically decided for us?

American Statistician (2014)

Armistead (2014)
"Resurrecting the Third Variable: A Critique of Pearl's Causal Analysis of Simpson's Paradox" + **comments**

Four perspectives:

- ARMISTEAD: "Third" variables can have importance beyond causation
- CHRISTENSEN: Careful thought required - think about sampling design, distinction between causal and predictive inference
- PEARL: Causal diagrams = all that is needed
- LIU & MENG: "multi-resolution" considerations provide guidance for finding correct "unit" (like group or individual?)

American Statistician (2014) continued...

Can discussions such as these provide insight into more than Simpson's paradox?

- Appropriate use of MLR?
- Appropriate use of model averaging (and even automatic model selection)?
- Can we switch from a "problem to fix" mentality to accepting careful thought as a key component of model choice?
  - Dangers of automatic procedures that require little thought
  - Accept we must think and justify our way to appropriate models and methods - no single right answer.

Pearl (2014) quotes

- "Thus, it is hard, if not impossible, to explain the surprise part of Simpson's reversal without postulating that human intuition is governed by causal calculus together with a persistent tendency to attribute causal interpretation to statistical associations."

-"The idea that statistical data, however large, are insufficient for determining what is 'sensible,' and that it must be supplemented with extra-statistical knowledge to make sense was considered heresy in the 1950's."

Pearl (2014) quotes cont...

-"It is a reminder of how easy it is to fall into a web of paradoxical conclusions when relying solely on intuition, unaided by rigorous statistical methods."

M.D. Higgs quote: "We need reminders of how easy it is to fall into a web of paradoxical conclusions when relying soley on perceived rigorous (and automatic) statistical methods, unaided by intuition and careful thought."

Pearl's solution

- Pearl says "we now know which causal structures would support Simpson's reversals, we also know which structure places the correct answer with the aggregated data or with the disaggregated.
  - Pearl (1993) "back-door" graphical condition in the causal diagram
  - He does not argue that answer lies somewhere in between the aggregated and disaggregated data as MA might provide

Christensen's advice

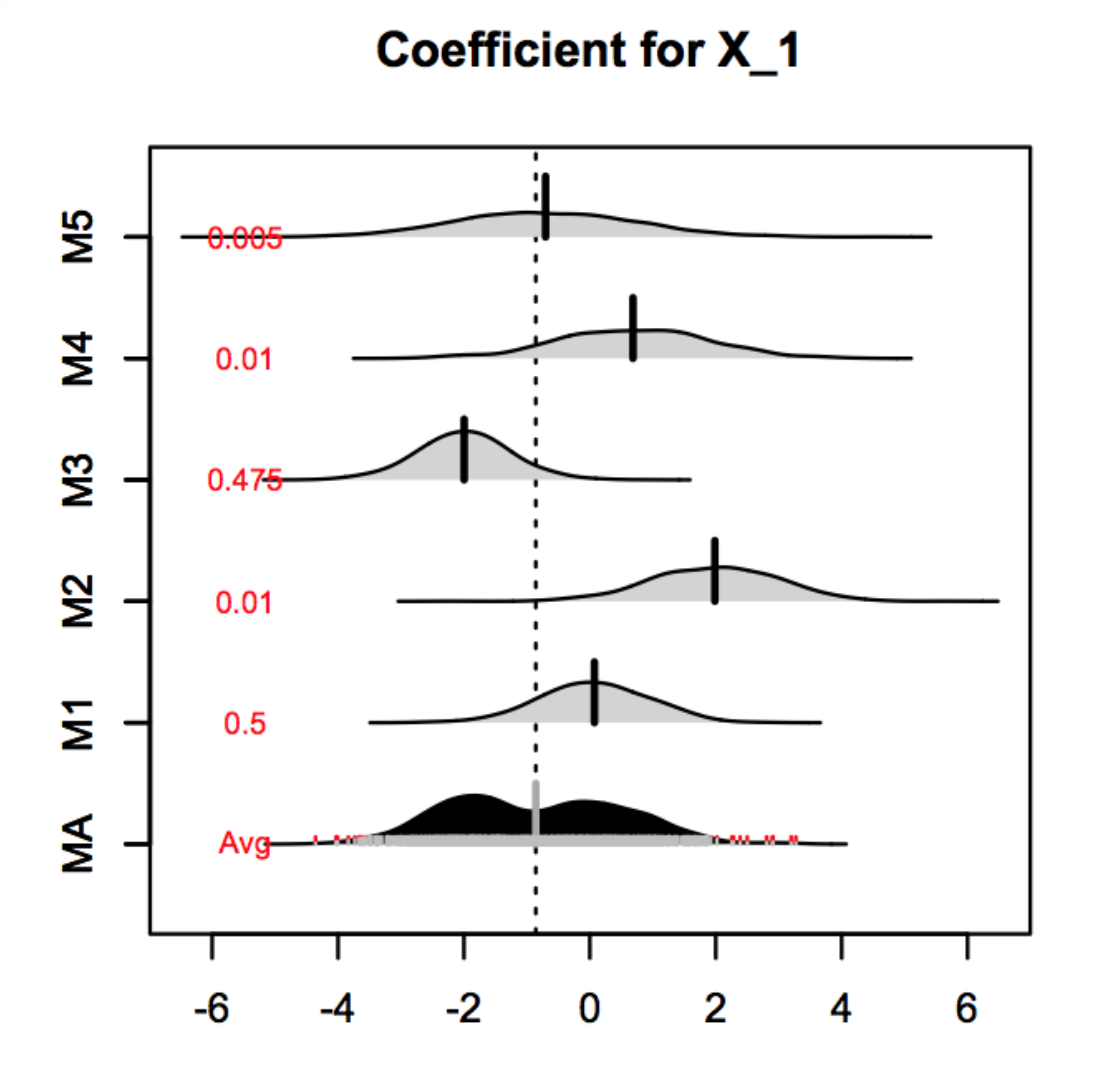
- Careful thought about making distinctions among goals of statistical inference and importance of study design.

- "While prediction is the ultimate goal of science, causation is the warm fuzzy. Causation can greatly simplify prediction and we like to think that good causative models provide the best predictions. But, in the end, getting predictions right is more important than *imagining* we understand why things happen they way they do."

Exploratory strategies when considering MA?

- Draw influence diagrams consistent with current state of knowledge (even for observational studies)
- Think carefully about relationships - what will be interpreted and what should be controlled for?
  - Are the coefficients measuring approximately the same thing across models?
- Graphical methods
  - Coded and cut scatter plots
  - Partial residual plots (component plus residual)
  - Individual model results compared to MA results (Katie's Plot)

Individual Model vs MA results (Katie's plot)



Gestation length and brain weight

Example used in *The Statistical Sleuth* (Ramsey & Schafer 2013) – Data from Sacher & Staffeldt (1974)

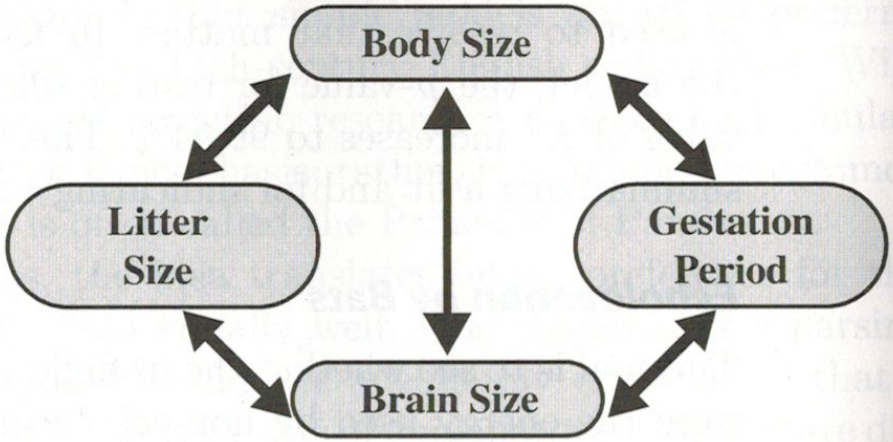
- Evolutionary biologists interested in relationship between mean brain weight (g) and gestation length (days), litter size, and body weight (kg)
- Average values constructed for 96 mammals
- Interested in investigating theoretical evolutionary costs that may be associated with large brained mammals after accounting for body size – longer pregnancies? smaller litter sizes?

Influence diagram from *The Statistical Sleuth*

DISPLAY 10.16 Associations of brain weight with gestation period and litter size: direct or indirect?

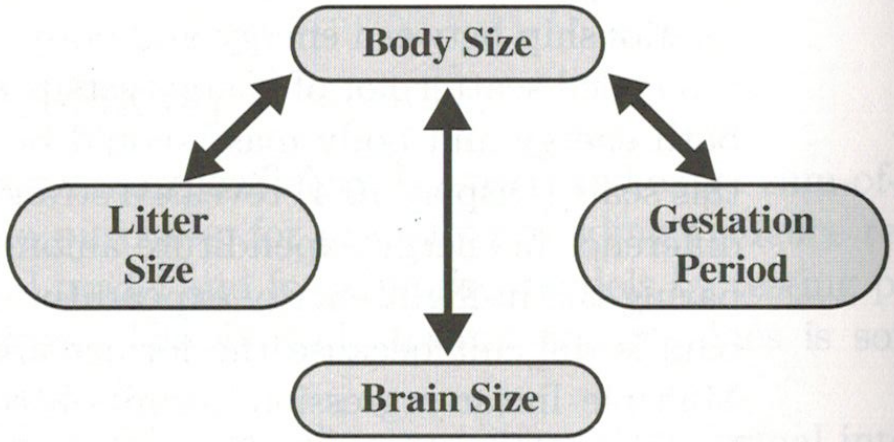
(a) DIRECT ASSOCIATION

Animals with about the same body size will have different brain sizes when their litter sizes and gestation periods are different.



(b) INDIRECT ASSOCIATION

Animals with about the same body size will have about the same brain size even if their gestation periods and litter sizes are different. Any apparent association between brain size and litter size, say, is a result of their both being related to body size.



Logic behind question (ignored litter size for now)

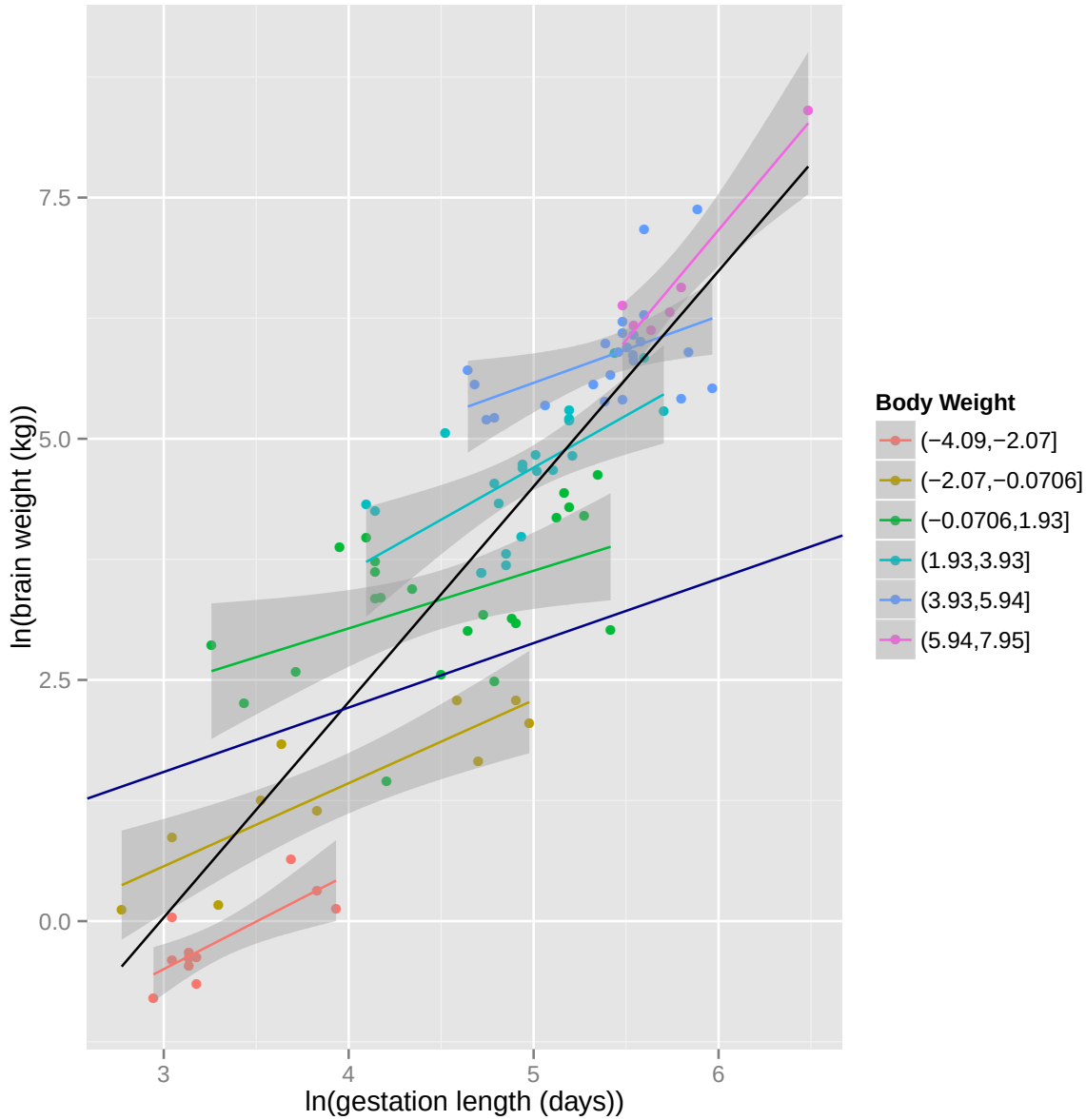
We expect relationships between

- brain weight and body size
- gestation length and body size
- gestation length and brain weight

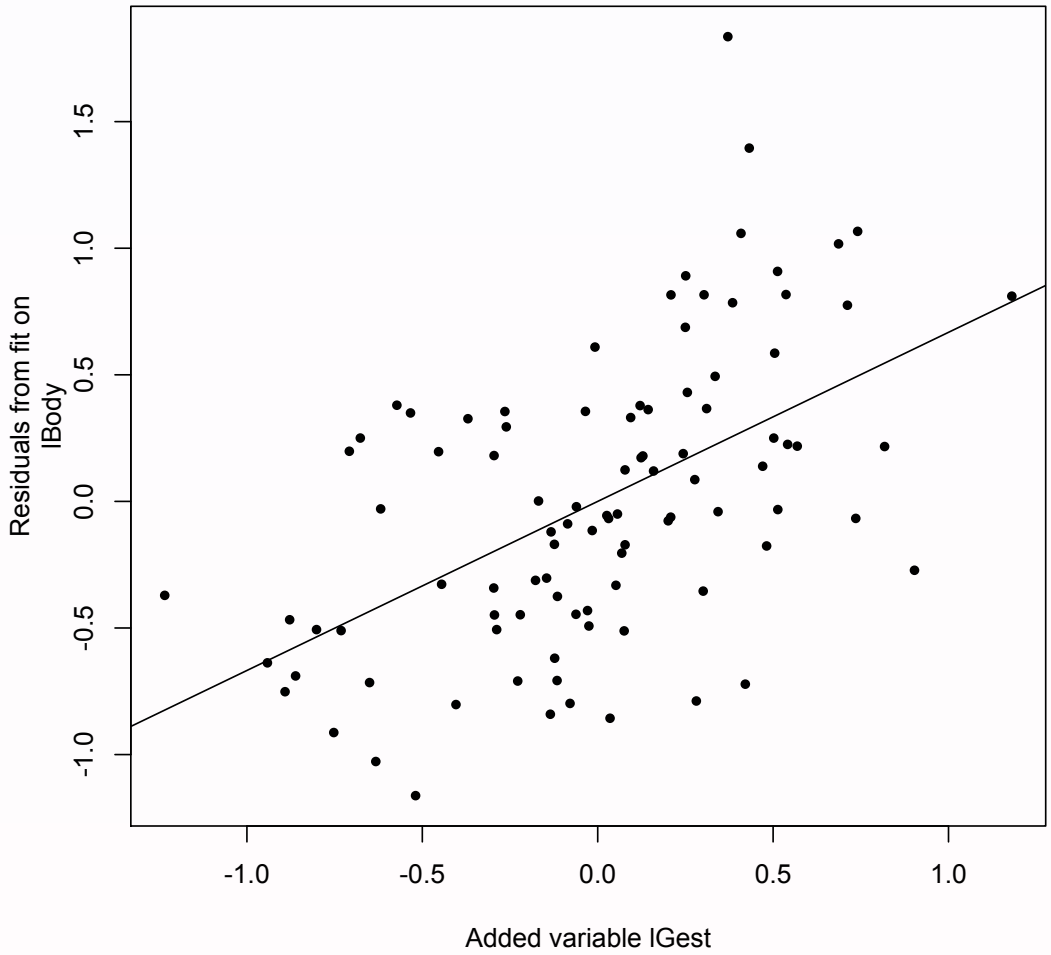
If we could hold body size constant, would we still see a relationship between brain weight and gestation length?

Practically interesting question that can be addressed using MLR?

Cut Scatterplot - "accounting for" body weight

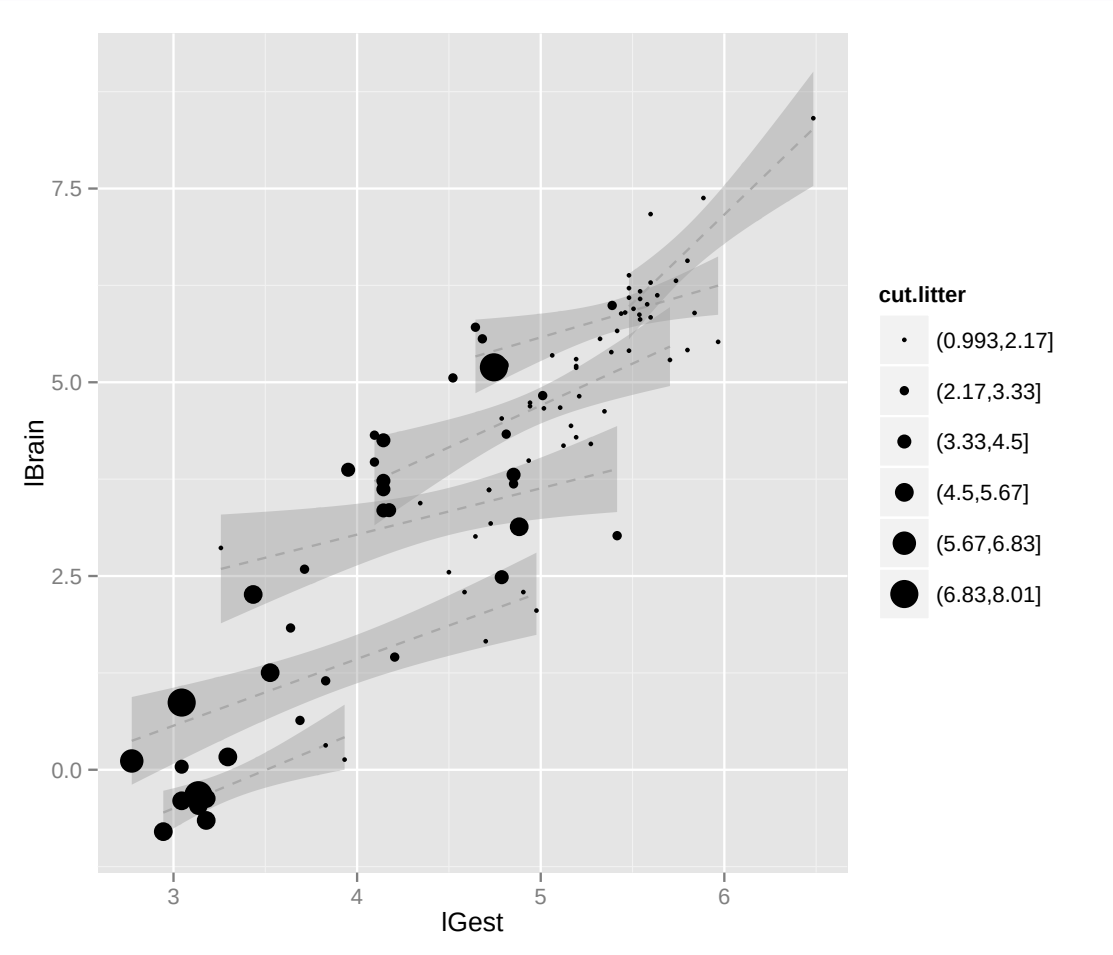


Partial residual plot

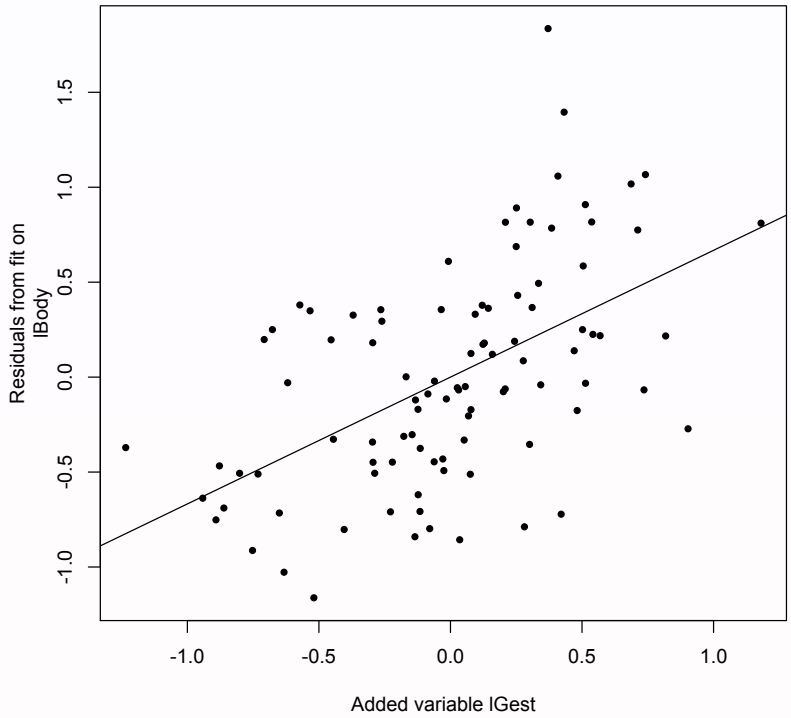


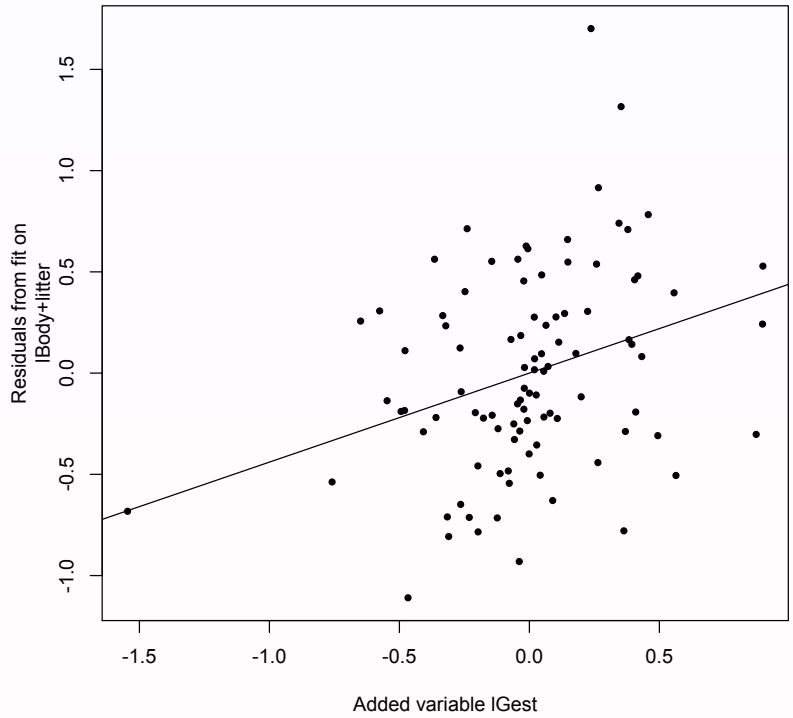
Cut and Coded Scatterplot

Relationship after "accounting for" body weight and visualizing litter size simultaneously

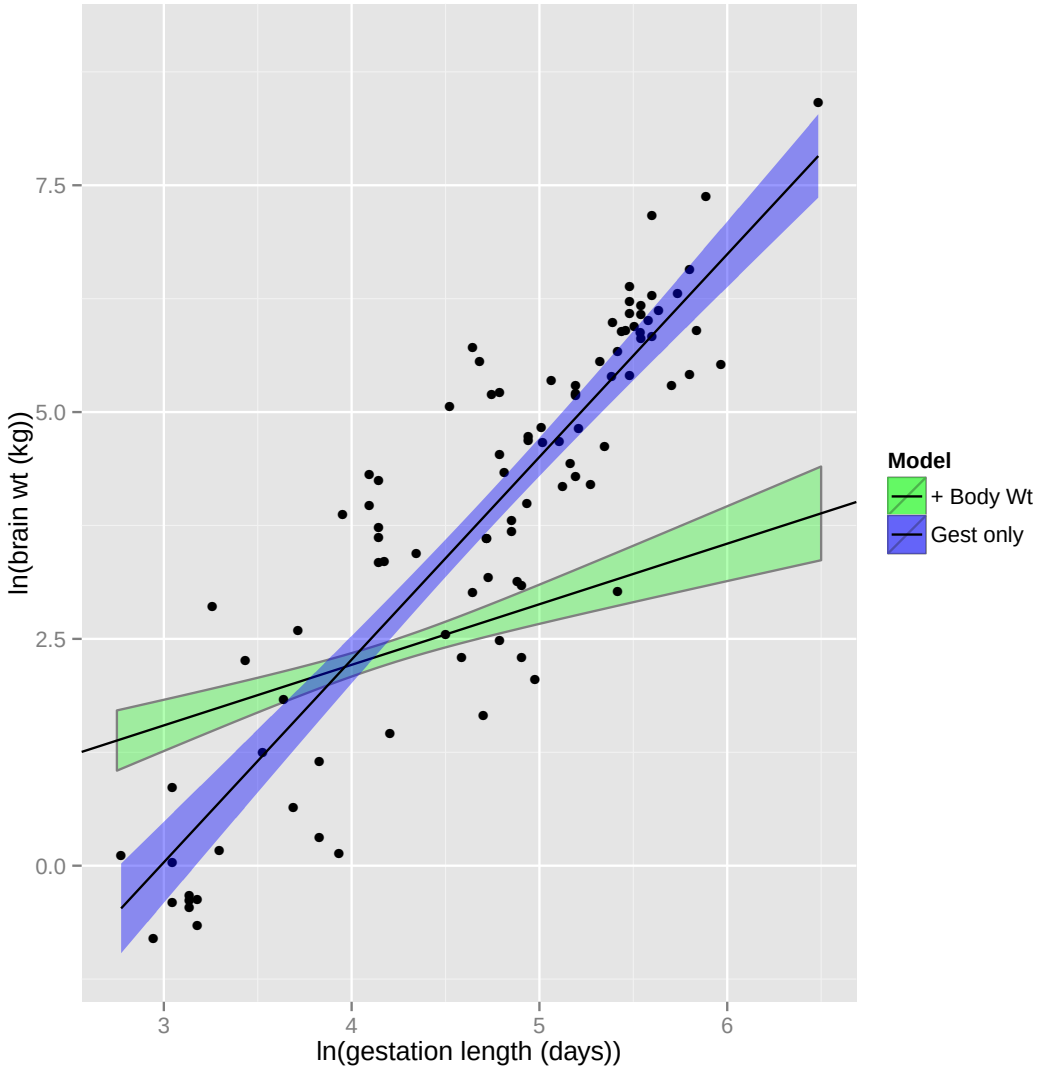


Partial residual plot





Marginal vs. Conditional fitted lines



Aligning the question and the model?

What questions align with the following model(s)?

M1 $\mu\{\text{brain}|\text{gest}\} = \beta_{0.g} + \beta_{g.g}\text{gest}$

M2 $\mu\{\text{brain}|\text{g, body}\} = \beta_{0.gb} + \beta_{g.gb}\text{gest} + \beta_{b.gb}\text{body}$

M3 $\mu\{\text{brain}|\text{g, b, litter}\} =$
$\beta_{0.gbl} + \beta_{g.gbl}\text{gest} + \beta_{b.gbl}\text{body} + \beta_{l.gbl}\text{litter}$

Is $\beta_{g.g} = \beta_{g.gb} = \beta_{g.gbl}$?

MA posterior distribution

Just consider the three models above for simplicity.

$$\begin{aligned} p(\beta_{g.MA}|y, \mathcal{M}_3) &= p(\beta_{g.g}|y, M_g)p(M_g|y) + \\ & \quad p(\beta_{g.gb}|y, M_{gb})p(M_{gb}|y) + \\ & \quad p(\beta_{g.gbl}|y, M_{gbl})p(M_{gbl}|y) \end{aligned}$$

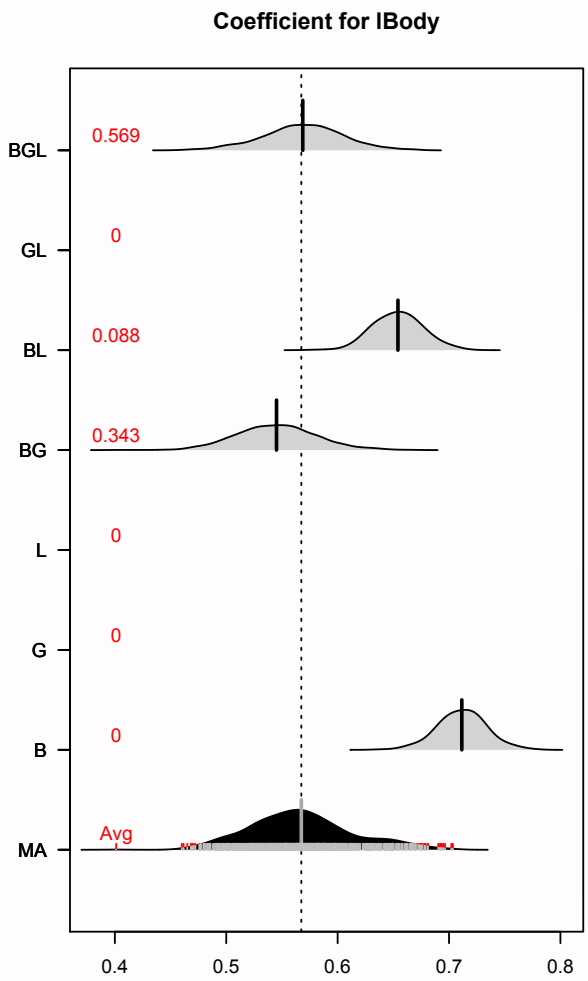
What is this $\beta_{g.MA}$?

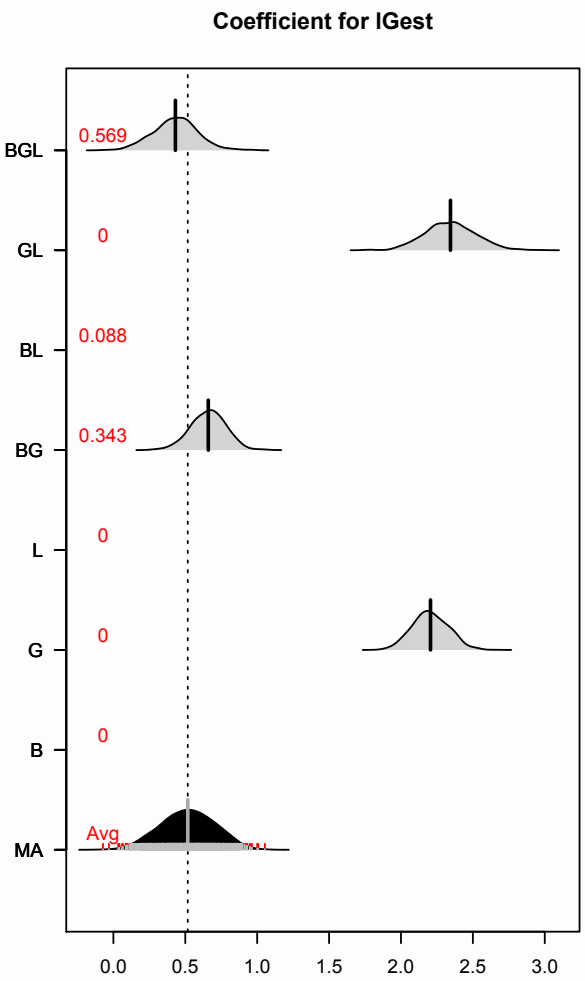
- What is it capturing and does it make sense?
- Actual posterior distribution depends on posterior model weights

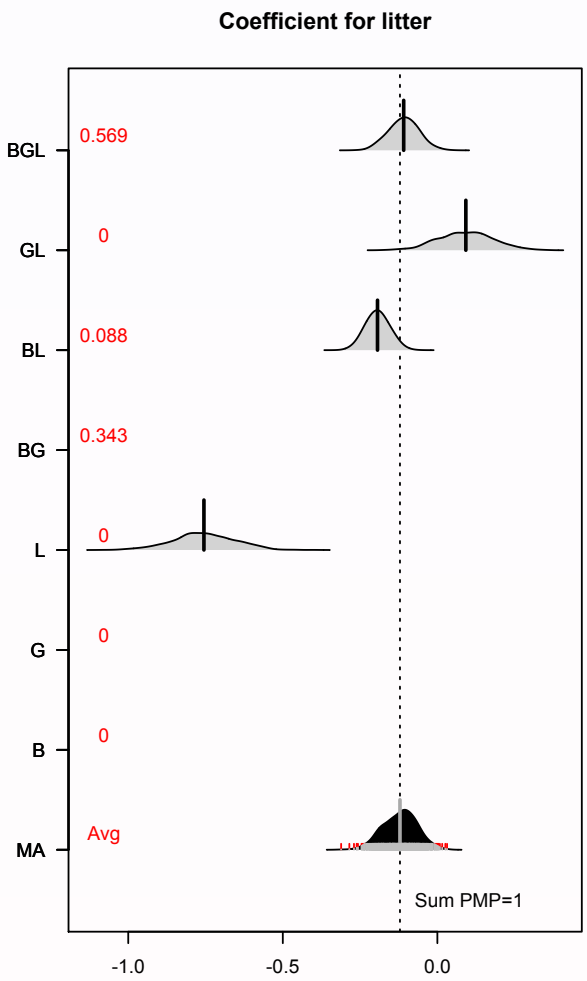
****

Now, let's actually consider 7 models: intercept-only model + one-covariate models + two-covariate models

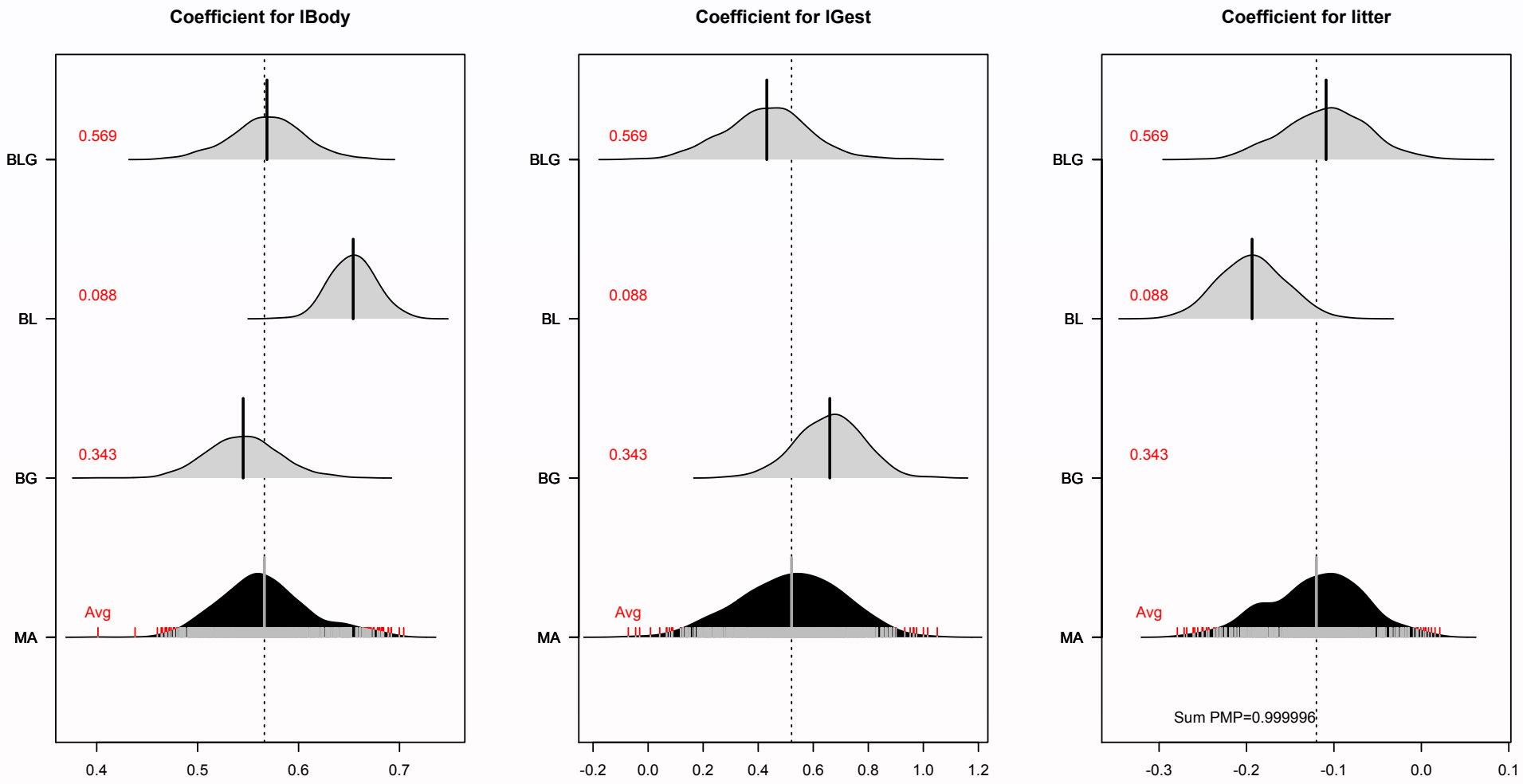
Katie's plot - all 7 models







Katie's plot - 3 models with largest weights



"Should I model average?"

- Recognize different questions need different models
  - It might not always be appropriate to average over them.
  - Prediction vs. explanation (consider goals of inference)
- MA results may be nearly identical to the results of one model – this should be acknowledged.
- Remind ourselves that MLR not designed to discover hidden laws of nature and neither is MA
- What uncertainty are we trying to account for?
  - that coming from not knowing which model is "right"?
  - that coming from not knowing which model to use?

My worries...

– "I don't know which model I should use, and by using model averaging I don't have justify my choice in model(s) *and* and at the same time I look sophisticated and cutting edge!"

- Will people use it to avoid the hard decisions/justifications, rather than thinking carefully about context, advantages, and disadvantages?
- The allure of "automaticity"
  - Proceed under a guise of "objectivity"
  - Usual dangers – encourages a lack of thought while increasing confidence in the results

Did a similar thing happen with AIC and AIC weights despite the good intentions of original proponents?

Other more technical issues to address further

- Supermodel effect - default MA often concentrates a lot of posterior mass on a few models — continue to investigate priors to avoid this if considered "bad" behavior.
- Interactions and higher order terms – Chipman has done some work
- Incorporation of design variables
- Sensitivity of MA to violations of MLR assumptions

Closing thought

Research questions should drive modeling decisions, rather than modeling methods driving research questions.

****

QUESTIONS?

Time line of some research

1960's: Roberts (1965), Bates & Granger (1969)

1970's: Leamer (1978)

1980's: Zellner & Siow (1980), Zellner (1984), Stewart & Davis (1986), Hodges (1987), Mitchell & Beauchamp (1988)

1990's: Gelfand & Smith (1990), Besag & Green (1993), Smith & Roberts (1993), George & McCulloch (1993), Geweke (1994), Madigan & Raftery (1994), Kass & Raftery (1995),Carlin & Chib (1995), Draper (1995), Chatfield (1995), Green (1995), Chipman (1996), Clyde, DeSimone, & Parmigiani (1996), George & McCulloch (1997), Raftery, Madigan, and Hoeting (1997), Kuo & Mallick (1998), Hoeting, Madigan, Raftery, & Volinsky (1999), George (1999)

2000's: Chipman, George, McCulloch (2001), Green (2003), Clyde & George (2004), Andrieu, Doucet, & Robert (2004), Liang, Paulo, Molina, Clyde, & Berger (2008) , Feldkircher & Zeugner (2009)

BMA Software in R

{BMS} - Bayesian Model Selection - Feldkircher & Zeugner (2009)

- `bms()`:
- Exploits properties of Normal - Conjugate linear model
- Uses MC^3 to search through model space with two options for the jumping distribution within the MH algorithm for models with more than 14 covariates.
- Zellner's g-prior as prior for coefficients
- Many choices for specification of g including a hyper-g option Liang et. al. (2008)
- Choice of prior on model space including customizable prior (only package with the customizable prior on $\mathcal{M}$

BMA Software in R

{BAS}- Bayesian Adaptive Sampling - Clyde (2012)

- `bas.lm()`
- Exploits properties of Normal - Conjugate linear model
- Uses stochastic or deterministic sampling without replacement from posterior distributions for models with more than 25 covariates.
- Uses Zellner's g-prior or mixtures of g-priors corresponding to Zellner-Siow Cauchy priors as prior for coefficients
- Many choices for specification of g including hyper-g option Liang et. al. (2008)
- AIC/BIC approximations available
- Has three choices for prior on model space (uniform, binomial, beta-binomial (hyper-prior on inclusion probs)), no customizable prior for $\mathcal{M}$

Amini & Parmeter (2012) note that Adaptive sampling works

BMA Software in R

- {BMA} - Bayesian Model Averaging - Raftery, Hoeting, Volinsky, Painter & Yeung (2014)
  - `bic.reg()`
  - BIC approx for priors on coefficients
  - When # of covariates exceeds 30: uses Occam's window algorithm to shave the model set down - again using BIC approximations for posterior model weights.
  - Uses uniform distribution on model priors (only choice)
- It's important to note that differences among these packages were highlighted only for the functions that work on a normal error liner model. Generalizations are available in the BMA package only.
- Consult Amini & Parmeter (2012) for the specifics.

Include litter size as well?

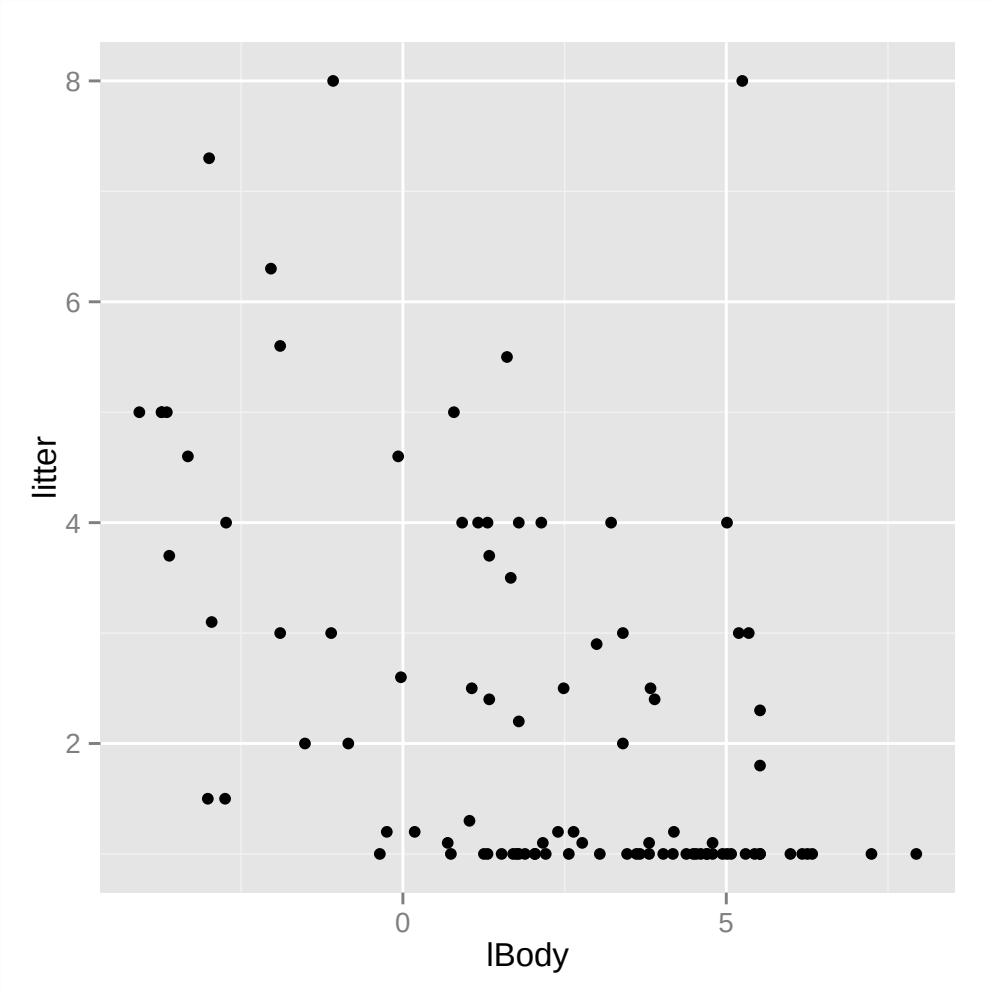


Table 1

Model	BMA	BIC	BMS	BAS
lBody	0.0000	0.0000	0.0000	0.0000
lGest	0.0000	0.0000	0.0000	0.0000
litter	0.0000	0.0000	0.0000	0.0000
lBody + lGest	0.2289	0.2972	0.3433	0.3433
lBody + litter	0.0438	0.0000	0.0877	0.0877
lGest + litter	0.0000	0.0000	0.0000	0.0000
lBody+lGest+litter	0.7273	0.7028	0.5690	0.5690

Estimates/Posterior means for MA coefficients

	lBody	lGest	litter
MLR	0.57	0.44	-0.11
BIC	0.57	0.47	-0.09
BMA	0.57	0.47	-0.09
BAS	0.57	0.47	-0.08
BMS	0.57	0.47	-0.08

Standard errors of the coefficients

	lBody	lGest	litter
MLR	0.033	0.137	0.042
BIC	0.036	0.153	0.039
BMA	0.038	0.189	0.063
BAS	0.042	0.217	0.070
BMS	0.045	0.224	0.072