Bayesian Model Averaging in the GLM context

STAT 851

Zubia Mansoor, Louis Arsenault-Mahjoubi, Sonny Min

April 3, 2020

Overview

Motivating Example

Motivating Example : Cognitive Data

- Response :- Kid's cognitive score
- Predictors :-

  hs : whether or not the kid's mom attended high school
  iq : mom's iq
  work : whether or not the mom worked during the first three years of the kid's life
  age : mom's age

	kid_score	hs	iq	work	age
1	65	1	121.11753	1	27
2	98	1	89.36188	1	25
3	85	1	115.44316	1	27
4	83	1	99.44964	1	25
5	115	1	92.74571	1	27
6	98	0	107.90184	0	18
7	69	1	138.89311	1	20
8	106	1	125.14512	1	23
9	102	1	81.61953	0	24
10	95	1	95.07307	0	19
11	91	1	88.57700	0	23
12	58	1	94.85971	1	24
13	84	1	88.96280	1	27
14	78	1	114.11430	1	26
15	102	0	100.53407	1	24

Figure: Dataset

Motivating Example Contd. (1)

GOAL: Perform Model Selection using Backward Elimination

step	model	dropped	BIC
FULL	kid_score ~ hs + iq + work + age		2541.1
STEP 1	kid_score ~ hs + iq + work	[-age]	2535.4
STEP 2	kid_score ~ hs + iq	[-work]	**2530.6**
STEP 3	kid_score ~ hs + iq		2530.6
	kid_score ~ hs	[-iq]	2531.7
	kid_score ~ iq	[-hs]	2604.0

Best Model: kid_score ~ hs + iq

Motivating Example Contd. (1)

GOAL: Perform Model Selection using Backward Elimination

step	model	dropped	BIC
FULL	kid_score ~ hs + iq + work + age		2541.1
STEP 1	kid_score ~ hs + iq + work	[-age]	2535.4
STEP 2	kid_score ~ hs + iq	[-work]	**2530.6**
STEP 3	kid_score ~ hs + iq		2530.6
	kid_score ~ hs	[-iq]	2531.7
	kid_score ~ iq	[-hs]	2604.0

Best Model: kid_score $\sim$ hs + iq

Limitations of the standard methods

- Involves multiple testing of hypotheses
- Asymptotics break down for small samples
- Often rejects satisfactory model when the sample size is large
- **Ignores model uncertainty**

Limitations of the standard methods

- Involves multiple testing of hypotheses
- Asymptotics break down for small samples
- Often rejects satisfactory model when the sample size is large
- **Ignores model uncertainty**

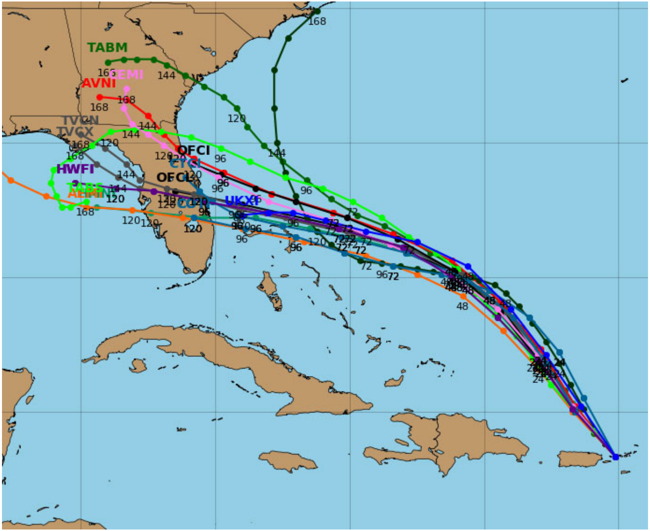
Limitations of the standard methods

- Involves multiple testing of hypotheses
- Asymptotics break down for small samples
- Often rejects satisfactory model when the sample size is large
- **Ignores model uncertainty**

Limitations of the standard methods

- Involves multiple testing of hypotheses
- Asymptotics break down for small samples
- Often rejects satisfactory model when the sample size is large
- **Ignores model uncertainty**

Illustration: Hurricane Dorian



The Bayesian Model Averaging Framework

Posterior model probabilities

- Assign each model a prior probability $p(M_k)$
- Bayes Theorem $\rightarrow$ Posterior Model probabilities

$$p(M_k|data) = \frac{p(data|M_k) \times p(M_k)}{\sum_{l=1}^{K} p(data|M_l) \times p(M_l)}$$

where,

$$p(data|M_k) = \int p(data|\theta_k, M_k) p(\theta_k|M_k) d\theta_k$$

Posterior model probabilities

- Assign each model a prior probability $p(M_k)$
- Bayes Theorem $\rightarrow$ Posterior Model probabilities

$$p(M_k|data) = \frac{p(data|M_k) \times p(M_k)}{\sum_{l=1}^{K} p(data|M_l) \times p(M_l)}$$

where,

$$p(data|M_k) = \int p(data|\theta_k, M_k)p(\theta_k|M_k)d\theta_k$$

Posterior model probabilities

- Assign each model a prior probability $p(M_k)$
- Bayes Theorem $\rightarrow$ Posterior Model probabilities

$$p(M_k|data) = \frac{p(data|M_k) \times p(M_k)}{\sum_{l=1}^{K} p(data|M_l) \times p(M_l)}$$

where,

$$p(data|M_k) = \int p(data|\theta_k, M_k) p(\theta_k|M_k) d\theta_k$$

Posterior model probabilities

- Assign each model a prior probability $p(M_k)$
- Bayes Theorem $\rightarrow$ Posterior Model probabilities

$$p(M_k|data) = \frac{p(data|M_k) \times p(M_k)}{\sum_{l=1}^{K} p(data|M_l) \times p(M_l)}$$

where,

$$p(data|M_k) = \int p(data|\theta_k, M_k) p(\theta_k|M_k) d\theta_k$$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \tag{1}$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \tag{2}$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \tag{1}$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \tag{2}$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \quad (1)$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \quad (2)$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

'

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \quad (1)$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \quad (2)$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \quad (1)$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \quad (2)$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

‘

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

Bayesian Model Averaging

- Let Δ be the quantity of interest
  Y^*, β_j, γ_j indicator variable that variable j is included, $p(\beta_j|data)$

$$p(\Delta|data) = \sum_{k=1}^{K} p(\Delta|M_k, data) \times p(M_k|data) \tag{1}$$

$$E(\Delta|data) = \sum_{k=1}^{K} E(\Delta|M_k, data) \times p(M_k|data) \tag{2}$$

- **NOTE**
  - The space of possible models K can be very large
  - The integrals implicit in (1) can be difficult to compute

‘

- weighted average of model-specific quantities
- BMA predictions $\hat{Y}^* = \sum_{k=1}^{K} \hat{Y}_k^* \times p(M_k|data)$

A Brief History of Model Averaging

- 1963, Barnard: First mention of model combination in the statistical literature.
- 1965, Roberts: Suggests the use of a weighted averages of two models' posteriors
- 1969, Bates & Granger: Combined predictions from different forecast models in economics.
- 1978, Leamer: Introduced the basic BMA paradigm.
- Mid-late 1990's: Computational development enables the implementation of BMA and it is used in the context of decision under model uncertainty (Draper (1995), Chatfield(1995), Kass and Raftery (1995), George (1999)).

Motivating Example Contd. (2)

kid_score $\sim$ hs + iq + work +age

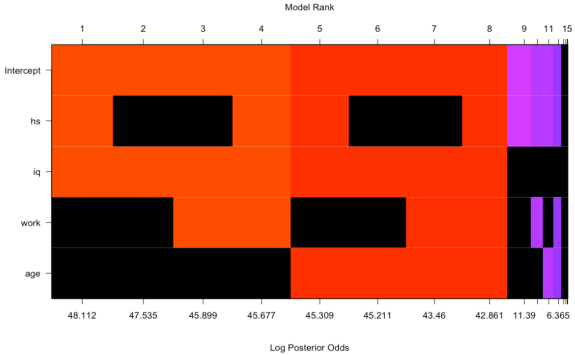
- p predictors (4)
- 2^p possible models (16)

	P(B != 0 | Y)	model 1	model 2	model 3	model 4	model 5
Intercept	1.000	1.000	1.000	1.000	1.000	1.000
hs	0.611	1.000	0.000	0.000	1.000	1.000
iq	1.000	1.000	1.000	1.000	1.000	1.000
work	0.112	0.000	0.000	1.000	1.000	0.000
age	0.069	0.000	0.000	0.000	0.000	1.000
BF	NA	1.000	0.562	0.109	0.088	0.061
PostProbs	NA	0.529	0.297	0.058	0.046	0.032
R2	NA	0.214	0.201	0.206	0.216	0.215
dim	NA	3.000	2.000	3.000	4.000	4.000
logmarg	NA	-2583.135	-2583.712	-2585.349	-2585.570	-2585.939

83%

Figure: Summary of top models

Visualising Model Uncertainty



- Models arranged in decreasing order of their posterior probability.
- Model 1 (highest posterior probability) includes the predictors high school and IQ, but not age or work.
- *iq* is in all of the top eight models.

Interpreting coefficient summaries

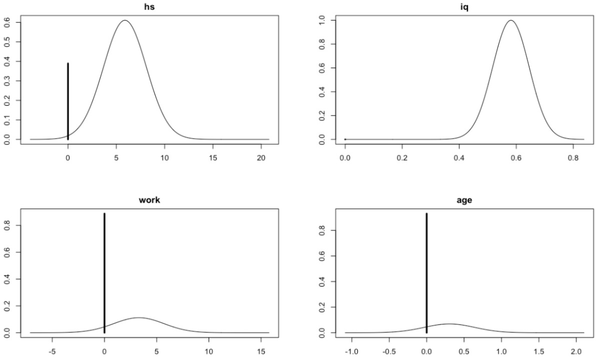
```
Marginal Posterior Summaries of Coefficients:

Using  BMA

Based on the top  16 models
           post mean  post SD   post p(B != 0)
Intercept  86.79724    0.87287   1.00000
hs          3.59494    3.35643   0.61064
iq          0.58101    0.06363   1.00000
work        0.36696    1.30939   0.11210
age         0.02089    0.11738   0.06898
```

- BMA estimate $\hat{\beta}_j =$ post Mean
- Posterior inclusion probabilities = post $p(\beta! = 0)$
- *iq* has posterior inclusion probability 1 $\rightarrow$ very likely that *iq* should be included in the model
- *hs* also has a high posterior inclusion probability of about 0.61
- *work* and *age* $\rightarrow$ relatively small compared to *hs* and *iq*

Posterior Density Plots



- Vertical bar → Posterior probability that the coefficient is 0
- Bell-shaped curve → Density of plausible values from all the models where the coefficient was non-zero
- *iq* → probability that the coefficient is non-zero is quite small
- *age* → Much higher probability of being 0

Key Takeaways

Benefits:

- takes into account model uncertainty and results in improved predictions
- updates its estimates as the data accumulates and the resulting model weights are continually adjusted
- relatively robust to model misspecification

Drawbacks:

- Number of models can be large rendering exhaustive summation infeasible
- The integrals implicit in (1) can be hard to compute
- Specification of $p(M_k)$ is challenging

Key Takeaways

Benefits:

- takes into account model uncertainty and results in improved predictions
- updates its estimates as the data accumulates and the resulting model weights are continually adjusted
- relatively robust to model misspecification

Drawbacks:

- Number of models can be large rendering exhaustive summation infeasible
- The integrals implicit in (1) can be hard to compute
- Specification of $p(M_k)$ is challenging

Implementation of the BMA framework

Dealing with the Summation

- How to choose the set of models of interest, M?
- What if M is large making the summation is difficult to compute?
- 2 approaches:
  - 1) Occam's Window
  - 2) Monte Carlo Methods (MC^3)

Occam's Window

- Occam's Window is based on two standard practices of the scientific method.
- First, if a model performs too poorly, it is considered discredited.

Mathematically, we will consider models in the set

$$A' = \left\{ M_k : \frac{max_l\{p(M_l \mid D)\}}{p(M_k \mid D)} \leq C \right\}$$

Occam's Window

- The second principle is Occam's razor:
- "Plurality must never be posited without necessity" - William of Occam (circa 1287–1347)

We then want to remove models in the set

$$B = \left\{ M_k : \exists M_l \in A', M_l \subset M_k : \frac{p(M_l \mid D)}{p(M_k \mid D)} > 1 \right\}$$

And are left to consider only

$$A = A' \setminus B$$

Occam's Window

- Madigan & Raftery 1994 propose a search algorithm to find the models in A using
$$\frac{P(M_0 \mid D)}{P(M_1 \mid D)}, M_0 \subset M_1$$

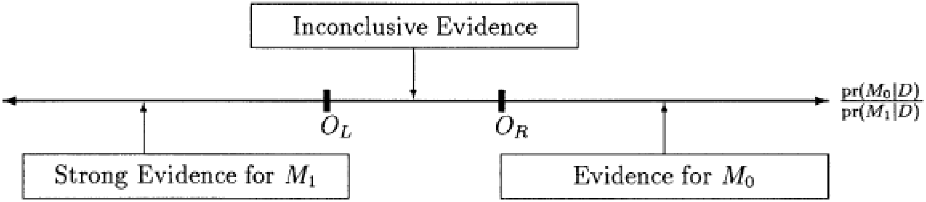


FIG. 1. *Occam's window: interpreting the posterior odds.*

Occam's Window

- If $\frac{P(M_0|D)}{P(M_1|D)} < O_L$, reject M_0 and the models nested within it.
- If $\frac{P(M_0|D)}{P(M_1|D)} > O_R$ reject M_1
- If $O_L \leq \frac{P(M_0|D)}{P(M_1|D)} \leq O_R$, we have inconclusive evidence.

Occam's Window: Choosing O_L & O_R

- Madigan and Raftery (1994) used $O_L = 1/20 \;\; O_R = 1$
- Raftery, Madigan and Volinsky (1996) found improved predictive performance with $O_L = 1/20 \;\; O_R = 20$
- Note: setting $O_L = O_R^{-1}$ is equivalent to using only the first principle.

Monte Carlo Markov Chain Model Composition (MC^3)

- Madigan and York (1995) use a Markov chain (MC) to target

$$p(\Delta \mid Data) = \sum_{k=1}^{K} p(\Delta \mid M_k, Data) p(M_k \mid Data)$$

- The MC, $\{M(t)\}, t = 1, 2, ...,$ has the models under consideration as its state space

Monte Carlo Markov Chain Model Composition (MC^3)

- We will to define a neighborhood to model M, $nbd(M)$, from which we can propose M' at each step according to a transition matrix (also to be determined by the user).
- We can then use the Metropolis-Hastings algorithm with acceptance probability

$$min\left\{1, \frac{p(M' \mid D)}{p(M \mid D)}\right\}$$

- to get N observations $M(1)$,..., $M(N)$ under the posterior

Monte Carlo Markov Chain Model Composition (MC^3)

- Thanks to MCMC results we have that if

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} g(M(t))$$

then,

$$\hat{G} \to E(g(M)) \ as \ N \to \infty$$

- set $g(M) = p(\Delta \mid M, D)$ and we get the desired summation.

Computing the integrals for BMA

- We want to compute the integrals

$$p(D \mid M) = \int p(D \mid \theta, M) p(\theta \mid M) d\theta$$

- as they are used in (1)
- Two potential methods:
  - The Laplace Method (Tierney and Kadane 1986)
  - MLE approximation (Taplin 1993)

Prior specification

- When we have a parameter associated with each predictor in the model, we can set

$$p(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{(1-\delta_{ij})}$$

- with $\pi_j \in [0, 1]$ is the prior probability that $\beta_j \neq 0$ and δ_{ij} is an indicator that variable j is in M_i.

BMA GLM Example: Baseball data

Baseball Data

- Salary vs. performance measures
- Y_{ij} = 1 if the salary of an individual i is greater than 1 million dollars, 0 otherwise.
- $Y_{ij} \sim$ Binomial(π_{ij}), Y_{ij}'s are independent.
  i=1,...,336, j= 1,2(Free agent eligibility no/yes)
  Where π_{ij} = probability of a player i in group j earning more than 1 million dollars.

Baseball Data - Predictors

- 13 Predictors: 12 continuous, 1 indicator, 336 players.

avg bat : Batting average (cont.)
OBP : On-base percentage (OBP) (cont.)
runs : Number of runs (cont.)
hits : Number of hits (cont.)
doubles : Number of doubles (cont.)
triples : Number of triples (cont.)
homeruns : Number of home runs (cont.)
RBI : Number of runs batted in (RBI) (cont.)
walks : Number of walks (cont.)
s outs : Number of strike-outs (cont.)
stolen : Number of stolen bases (cont.)
errors : Number of errors (cont.)
FA : "free agency eligibility" (indicator)

Baseball Data - Dataset

y	avg_bat	OBP	runs	hits	doubles	triples	homeruns	RBI	walks	s_outs	stolen	errors
1	272	302	69	153	21	4	31	104	22	80	4	3
1	269	335	58	111	17	2	18	66	39	69	0	3
1	249	337	54	115	15	1	17	73	63	116	6	5
1	26	292	59	128	22	7	12	50	23	64	21	21
1	273	346	87	169	28	5	8	58	70	53	3	8
1	291	379	104	170	32	2	26	100	87	89	22	4
0	258	37	34	86	14	1	14	38	15	45	0	10
0	228	279	16	38	7	2	3	21	11	32	2	3
0	25	327	40	61	11	0	1	18	24	26	14	2
0	203	24	39	64	10	1	10	33	14	96	13	6
0	262	283	7	38	5	0	0	10	5	18	2	7
0	222	307	21	45	9	0	6	22	19	56	3	3
0	227	28	4	5	2	0	1	3	2	1	0	0
0	261	37	1	6	0	0	0	2	4	3	0	0
1	3	368	69	141	22	3	19	75	53	64	31	7

Plan

1. Fit standard GLM(logit, probit)
2. Fit BMA GLM(logit, probit)
3. Compare the coefficients
4. Look at some of the main features of BMA
5. Measure predictive performances

Fit GLM

- logit link:

$$log(\frac{\pi_{ij}}{1-\pi_{ij}}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ...\beta_{13} x_{13i}$$

- probit link:

$$\Phi^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ...\beta_{13} x_{13i}$$

Fit GLM: Coefficients

- Coefficients and p-values:

	Estimate	SE	z	p-value
(Intercept)	-0.596	1.717	-0.347	0.728
avg_bat	-1.820	17.927	-0.102	0.919
OBP	-12.408	14.592	-0.850	0.395
runs	0.019	0.025	0.771	0.441
hits	0.017	0.016	1.072	0.284
doubles	0.019	0.037	0.509	0.611
triples	0.044	0.098	0.449	0.654
homeruns	0.073	0.055	1.345	0.179
RBI	0.020	0.022	0.938	0.348
walks	0.011	0.023	0.469	0.639
s_outs	-0.031	0.010	-3.149	0.002
stolen	0.022	0.020	1.126	0.260
errors	-0.003	0.032	-0.090	0.928
fa1	2.87	0.40	7.12	0.00

GLM: Variable Selection

- Variable selection using BIC with backward elimination.
- Bayesian Information Criterion
- BIC = -2ln($\hat{likelihood}$) $+ (p+1)\ln(n)$, p = Number of parameters
  where $\hat{likelihood} = pr(D|\hat{\beta}, M) = L(\hat{\beta}, M)$
  BIC = $n\ln(1-R^2) + (p+1)\ln(n)$
- Smaller the BIC the better

GLM: Model Selection

- GLM with the smallest BIC:

	Estimate	SE	z	p-value
(Intercept)	-0.969	1.406	-0.689	0.491
OBP	-12.264	4.921	-2.492	0.013
runs	0.058	0.012	4.665	0.000
RBI	0.042	0.012	3.562	0.000
s_outs	-0.025	0.008	-3.214	0.001
fa1	2.88	0.38	7.61	0.00

- BIC(full) = 304.45, BIC(red) = 262.37

Fit BMA GLM

- BMA GLM: Binomial("logit")
  $log\frac{\pi_{ij}}{1-\pi_{ij}} = \sum_{k=0}^{2^p}[(\beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2i} + ... + \beta_{13k}x_{13i}) \times pr(M_k|D)]$
- BMA GLM: Binomial("probit")
  $\Phi^{-1}(\pi_{ij}) = \sum_{k=0}^{2^p}[(\beta_{0k} + \beta_{1k}x_{1i} + \beta_{2k}x_{2i} + ... + \beta_{13k}x_{13i}) \times pr(M_k|D)]$
- Uniform prior model probability (i.e. $p(M_m) = \frac{1}{2^{13}}$ for all m)

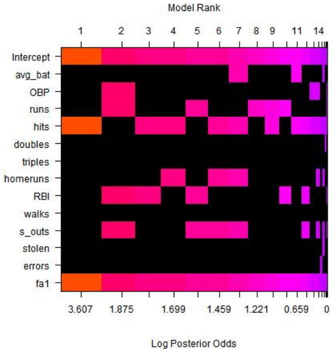
BMA GLM: top 10 models

- BMA result(top 10 models):

	P(B!=0|D)	Post Mean	SD	model.1	model.2	model.3	model.4	model.5	model.6	model.7	model.8	model.9	model.10
Intercept	100.00	-3.497	1.947	-5.019	-0.969	-4.343	-4.512	-5.021	-0.993	-4.990	-4.983	-4.688	-4.658
avg_bat	13.00	-1.730	5.422	.	.	.	.	.	-16.120	.	.	.	.
OBP	26.30	-2.864	5.532	.	-12.260	.	.	.	.	.	.	.	.
runs	42.90	0.020	0.026	.	0.058	0.044	.	.	.	.	0.025	.	0.039
hits	65.10	0.020	0.017	0.034	.	.	0.034	0.025	0.044	0.030	0.022	0.024	.
doubles	2.60	0.001	0.008	.	.	.	.	.	.	.	.	.	.
triples	3.60	0.002	0.021	.	.	.	.	.	.	.	.	.	.
homeruns	26.70	0.022	0.042	.	.	.	0.088	.	0.096	0.042	.	.	.
RBI	47.90	0.018	0.021	.	0.042	0.040	.	0.021	.	.	.	0.034	0.025
walks	3.00	0.000	0.003	.	.	.	.	.	.	.	.	.	.
s_outs	57.70	-0.013	0.013	.	-0.025	-0.020	-0.020	.	-0.028	.	.	-0.015	.
stolen	8.60	0.002	0.009	.	.	.	.	.	.	.	.	.	.
errors	1.70	0.000	0.005	.	.	.	.	.	.	.	.	.	.
fa	100.00	2.777	0.371	2.773	2.876	2.729	2.703	2.751	2.685	2.724	2.760	2.766	2.684
nVar				2	5	4	4	3	5	3	3	4	3
BIC				-1,692	-1,692	-1,691	-1,691	-1,691	-1,691	-1,690	-1,689	-1,689	-1,689
PMP				0.119	0.112	0.075	0.060	0.056	0.051	0.048	0.028	0.027	0.024

- PMP of top 10 models = 0.6

Uncertainty Visualization



- Log Posterior odds: $\ln(PO[M_m : M_0])$ $= \ln(BF[M_m : M_0] \times O[M_m : M_0])$
- $BF[M_m : M_0] = \frac{pr(D|M_m)}{pr(D|M_0)}$
- $O[M_m : M_2] = \frac{pr(M_1)}{pr(M_0)}$

Bayes Factor

- Prior odds: $O[M_1 : M_2] = \frac{pr(M_1)}{pr(M_2)}$
- Posterior odds: $PO[M_1 : M_2] = \frac{pr(M_1|D)}{pr(M_2|D)}$
- Using the Bayes' rule:

$$\begin{aligned} PO[M_1 : M_2] &= \frac{pr(M_1|D)}{pr(M_2|D)} \\ &= \frac{(pr(D|M_1) \times pr(M_1))/pr(D)}{(pr(D|M_2) \times pr(M_2))/pr(D)} \\ &= \frac{pr(D|M_1)}{pr(D|M_2)} \times \frac{pr(M_1)}{pr(M_2)} \\ &= \text{Bayes Factor} \times O[M_1 : M_2] \end{aligned} \tag{3}$$

- $BF[M_1 : M_2] = \frac{pr(D|M_1)}{pr(D|M_2)} = \frac{PO[M_1:M_2]}{O[M_1:M_2]}$

Bayes Factor

- Prior odds: $O[M_1 : M_2] = \frac{pr(M_1)}{pr(M_2)}$
- Posterior odds: $PO[M_1 : M_2] = \frac{pr(M_1|D)}{pr(M_2|D)}$
- Using the Bayes' rule:

$$\begin{aligned} PO[M_1 : M_2] &= \frac{pr(M_1|D)}{pr(M_2|D)} \\ &= \frac{(pr(D|M_1) \times pr(M_1))/pr(D)}{(pr(D|M_2) \times pr(M_2))/pr(D)} \\ &= \frac{pr(D|M_1)}{pr(D|M_2)} \times \frac{pr(M_1)}{pr(M_2)} \\ &= \text{Bayes Factor} \times O[M_1 : M_2] \end{aligned} \tag{3}$$

- $BF[M_1 : M_2] = \frac{pr(D|M_1)}{pr(D|M_2)} = \frac{PO[M_1:M_2]}{O[M_1:M_2]}$

Interpreting Bayes factor

	P(B!=0|data)	model 1	model 2	model 3	model 4	model 5	model.6	model.7	model.8	model.9	model.10
BF		1.000	0.887	0.226	0.278	0.276	0.222	0.110	0.202	0.218	0.214
PostProbs		0.134	0.127	0.085	0.018	0.018	0.016	0.016	0.015	0.015	0.015
R2		0.665	0.664	0.648	0.669	0.669	0.668	0.655	0.668	0.668	0.668
dim		5.000	5.000	4.000	6.000	6.000	6.000	5.000	6.000	6.000	6.000
logmarg		-89.008	-89.128	-90.495	-90.288	-90.296	-90.511	-91.218	-90.606	-90.529	-90.548

Table: Jeffreys' Scale (1961)

$BF[M_1 : M_2]$	**Evidence against M2**
1 to 0.33	No evidence
0.33 to 0.03	Positive
0.03 to 0.01	Strong
< 0.01	Very Strong

Interpreting Bayes factor

P(B!=0|data)	model 1	model 2	model 3	model 4	model 5	model.6	model.7	model.8	model.9	model.10
BF	1.000	0.887	0.226	0.278	0.276	0.222	0.110	0.202	0.218	0.214
PostProbs	0.134	0.127	0.085	0.018	0.018	0.016	0.016	0.015	0.015	0.015
R2	0.665	0.664	0.648	0.669	0.669	0.668	0.655	0.668	0.668	0.668
dim	5.000	5.000	4.000	6.000	6.000	6.000	5.000	6.000	6.000	6.000
logmarg	-89.008	-89.128	-90.495	-90.288	-90.296	-90.511	-91.218	-90.606	-90.529	-90.548

Table: Jeffreys' Scale (1961)

$BF[M_1 : M_2]$	**Evidence against M2**
1 to 0.33	No evidence
0.33 to 0.03	Positive
0.03 to 0.01	Strong
< 0.01	Very Strong

Posterior model probabilities

- Suppose we have models $M_m, m = 0, 1, ..., 2^p$.
- The posterior probability of each model given data:

$$pr(M_m|D) = \frac{\text{marginal likelihood of } M_m \times pr(M_m)}{\sum_{j=0}^{2^p} \text{marginal likelihood of } M_j \times pr(M_j)}$$

$$= \frac{pr(\mathrm{D}|M_m)pr(M_m)}{\sum_{j=0}^{2^p} pr(D|M_j)pr(M_j)}$$

$$= \frac{pr(D|M_m) \times pr(M_m)/(pr(D|M_b) \times pr(M_b))}{\sum_{j=1}^{2^p}[pr(D|M_j) \times pr(M_j)/(pr(D|M_b) \times pr(M_b))]}$$

$$= \frac{BF[M_m : M_b] \times O[M_m : M_b]}{\sum_{j=1}^{2^p} BF[M_j : M_b] \times O[M_j : M_b]}$$

Posterior model probabilities

- Suppose we have models $M_m, m = 0, 1, ..., 2^p$.
- The posterior probability of each model given data:

$$\begin{aligned} pr(M_m|D) &= \frac{\text{marginal likelihood of } M_m \times pr(M_m)}{\sum_{j=0}^{2^p} \text{marginal likelihood of } M_j \times pr(M_j)} \\ &= \frac{pr(\mathrm{D}|M_m)pr(M_m)}{\sum_{j=0}^{2^p} pr(D|M_j)pr(M_j)} \\ &= \frac{pr(D|M_m) \times pr(M_m)/(pr(D|M_b) \times pr(M_b))}{\sum_{j=1}^{2^p} [pr(D|M_j) \times pr(M_j)/(pr(D|M_b) \times pr(M_b))]} \\ &= \frac{BF[M_m : M_b] \times O[M_m : M_b]}{\sum_{j=1}^{2^p} BF[M_j : M_b] \times O[M_j : M_b]} \end{aligned}$$

BMA GLM: Coefficients

	Post Mean	P(B!=0|D)	GLM coefs	p-values
Intercept	-3.497	100.00	-0.596	0.728
fa	2.777	100.00	2.87	0.00
hits	0.020	65.10	0.017	0.284
s_outs	-0.013	57.70	-0.031	0.002
RBI	0.018	47.90	0.020	0.348
runs	0.020	42.90	0.019	0.441
homeruns	0.022	26.70	0.073	0.179
OBP	-2.864	26.30	-12.408	0.395
avg_bat	-1.730	13.00	-1.820	0.919
stolen	0.002	8.60	0.022	0.260
triples	0.002	3.60	0.044	0.654
walks	0.000	3.00	0.011	0.639
doubles	0.001	2.60	0.019	0.611
errors	0.000	1.70	-0.003	0.928

- p-values don't tell us how 'important' each variable is

Baseball Data - BMA GLM coefficients

	Post Mean	P(B!=0|D)	Step coefs	p-values
Intercept	-3.497	100.00	-0.969	0.491
fa	2.777	100.00	2.88	0.00
hits	0.020	65.10		
s_outs	-0.013	57.70	-0.025	0.001
RBI	0.018	47.90	0.042	0.000
runs	0.020	42.90	0.058	0.000
homeruns	0.022	26.70		
OBP	-2.864	26.30	-12.264	0.013
avg_bat	-1.730	13.00		
stolen	0.002	8.60		
triples	0.002	3.60		
walks	0.000	3.00		
doubles	0.001	2.60		
errors	0.000	1.70		

- p-values don't tell us how 'important' each variable is

Model Comparison - Models

- Models:
  - Full GLM : 13 predictors, BIC = 304.45
  - Stepwise GLM (BIC, Backward elimination) : 5 predictors, BIC = 262.37
  - BMA with Occam's window(OW) with $O_R = 20$, $O_L = \frac{1}{20}$: averaging over 53 models
  - BMA with no Occam's window(No-OW) : averaging over $2^{13} = 8,192$ models

Performance measure: Partial Performance Score (PPS)

- Partial Predictive Scores (PPS)
  1. Randomly divide the dataset into two: D^{train}, D^{test}
  2. Train model M^{train} using D^{train}
  3. Estimate the responses on the subjects in D^{test} with M^{train}
  4. Calculate PPS: $-\sum_{d \in D^{test}} \ln[\sum_{M \in A} pr(d|M^{train}, D^{train}) pr(M^{train}|D^{train})]$
- Smaller the PPS the better

Model Comparison - Result

model	PPS
Full glm(logit)	63.279
Stepwise glm(logit)	59.410
BMA OW(logit)	58.236
BMA No-OW(logit)	58.039

- PPS(Full GLM(logit)) - PPS(BMA(logit, No-OW)) = 5.24, PPS(Step GLM(logit)) - PPS(BMA(logit, No-OW)) = 1.381
- exp(5.24/168)= 1.032, exp(1.381/168)=1.008
- Applying Occam's window increases PPS

Summary and Conclusions

- Illustrated how standard model selection procedures ignore model uncertainty
- Introduced Bayesian model averaging
- Gave an overview of the practical solutions required for its implementation
- Applied Bayesian Model averaging to a GLM example

Extensions

- BMA in Meta-Analysis
- BMA in Network Analysis

References

- Hoeting, Jennifer A., et al. "Bayesian model averaging: a tutorial." Statistical science (1999): 382-401.
- Madigan, David, and Adrian E. Raftery. "Model selection and accounting for model uncertainty in graphical models using Occam's window." Journal of the American Statistical Association 89.428 (1994): 1535-1546.
- Max Hinne et al. "A Conceptual Introduction to Bayesian Model Averaging" Preprint (2019)
- Raftery et al. "Package 'BMA'" R document (2020)
- Clyde et al. "Package 'BAS'" R document (2020)


THE PRIOR MODEL DEMONS ALL SHOUT ...
'YOU NEED ONLY THIS PREDICTOR!'
'INCLUDE ALL PREDICTORS!'
... BUT AFTER THEY GORGE ON DATA ...
'OMNOMNOM'
'OMNOMNOM'
'OMNOMNOM'
... SOME POSTERIOR MODEL DEMONS BECOME POWERFUL, AND OTHERS WITHER AWAY ...
'I MUST FLEE!'
'THE DATA BE CURSED!'
'WHAT HAPPENED?'
'I TOO SHALL BE HEARD!'
'I HAVE PROVEN WORTHY! BEHOLD MY HYPOTHESIS!'
... YET ALL KEEP SHOUTING, TO SOME EXTENT.

Questions?

Appendix. Effect of link function on PPS

model	PPS
Full glm(logit)	63.279
Stepwise glm(logit)	59.410
BMA OW(logit)	58.236
BMA No-OW(logit)	58.039

model	PPS
Full glm(probit)	63.533
Stepwise glm(probit)	59.582
BMA OW(probit)	58.678
BMA No-OW(probit)	58.321

- Link function does have impacts on PPS.

Interpretation

- "BMA can be viewed as standard Bayesian inference for just one model, the full model in which all variables are included. The twist is that the prior allows for the possibility that some of the coefficients might be equal to zero(or, essentially equivalently close to zero).....Once we recast the way we think of BMA in this way, in terms of just one model, the "apples and oranges" problem disappears." - **Hoeting, Jennifer A., et al. "Bayesian model averaging: a tutorial." Statistical science (1999): 382-401.**

Interpretation Contd. (1)

- Draper argues that Δ needs to have the same meaning in all models.
- The authors put forth that a sufficient condition would seem to be that Δ be an observable quantity that could be predicted.
- Model coefficients could be arguably observed asymptotically.
- However, BMA's validity for nonlinear effects and interactions becomes problematic.