Bayesian Model Selection and Averaging

- Often there is uncertainty in one or more aspects of the GLM, which should be accounted for in the analysis.
- As motivation, we start with a simple example
- Suppose that data consist of $y_i \in \{0, 1\}$ Bernoulli distributed outcome variables, along with a single predictor $d_i = $ dose for subject $i$. 
• In this case, one could fit the generalized linear model:

\[ \eta_i = \beta_0 + \beta_1 d_i, \]

where \( \eta_i \) is the linear predictor

• However, suppose that we don’t know the link function, but have the following alternatives:

1. logistic \( (M = 1) \)
2. probit \( (M = 2) \)
3. complementary log-log \( (M = 3) \)

• \( M \in \{1, 2, 3\} \) is a model indicator
• If we don’t know whether \( M = 1 \), \( M = 2 \), or \( M = 3 \), how do we go about presenting results from the dose response analysis, such as an estimated dose response curve

• One ad hoc possibility:

  1. Fit each model in the list - in this case, rerun a maximum likelihood analysis for the logistic, probit and then complementary log-log models

  2. Estimate the dose response curves in each case:

     \[
     \hat{f}^{(1)}(d) = \left\{ 1 + \exp\left(-\hat{\beta}_0^{(1)} - \hat{\beta}_1^{(1)} d \right) \right\}^{-1}
     \]

     \[
     \hat{f}^{(2)}(d) = \Phi^{-1}(\hat{\beta}_0^{(2)} + \hat{\beta}_1^{(2)} d)
     \]

     \[
     \hat{f}^{(3)}(d) = 1 - \exp\left\{ - \exp(\hat{\beta}_0^{(3)} + \hat{\beta}_1^{(3)} d) \right\}
     \]

     with the \((m)\) superscript indexing model \( M = m \)

  3. Average the model-specific estimates:

     \[
     \tilde{f}(d) = \frac{1}{3} \hat{f}^{(1)}(d) + \frac{1}{3} \hat{f}^{(2)}(d) + \frac{1}{3} \hat{f}^{(3)}(d)
     \]

     to obtain a model-averaged estimator
Equal weights make sense?

1. What if we have prior information about which link functions are more plausible?

2. What if the data more strongly support one function than another?

3. A critical aspect of model averaging is choice of the weights.
Bayesian Model Averaging

- The model index $M$ is another unknown treated in a similar manner to the parameters $\beta$

- To complete a Bayesian specification, we need:
  1. A prior probability for each model in the list $\mathcal{M} = \{1, 2, 3\}$
  2. A prior for the parameters in each possible model, $\pi(\beta^{(m)})$, for $m = 1, 2, 3$.
  3. The likelihood for the data under each model, $L^{(m)}(y; x, \beta^{(m)}, M = m)$, for $m = 1, 2, 3$. 
• Prior for models:

\[ \pi(M) = \text{multinomial}([1, 2, 3]; \{\pi_1, \pi_2, \pi_3\}) \]

with \( \Pr(M = m) = \pi_m \) probability assigned to \( m \)th model

• Prior for parameters within models:

\[ \pi(\beta^{(m)}) = \text{N}(\theta_0^{(m)}, \Sigma_0^{(m)}), \quad \text{for } m = 1, 2, 3 \]

where \( \theta_0^{(m)}, \Sigma^{(m)} \) are the prior mean and variance for the regression parameters in the \( m \)th model

• Note that the normal prior is only one possible choice - if the prior information is better quantified using a different distribution, that is fine.
Bayesian updating

- Including the model indicator as an unknown (essentially another parameter, which can only take discrete values), the joint posterior is expressed as:

$$\pi(M, \beta \mid y, X) \propto \left\{ \prod_{m=1}^{3} \pi_1^{(M=m)} \pi(\beta^{(m)}) L^{(m)}(y; X, \beta^{(m)}, M = m) \right\}.$$  

- The posterior model probability is defined as follows:

$$\Pr(M = m \mid y, X) = \frac{\pi_m \int L^{(m)}(y; X, \beta^{(m)}, M = m) \pi(\beta^{(m)})}{\sum_{j=1}^{3} \pi_j \int L^{(j)}(y; X, \beta^{(j)}, M = j) \pi(\beta^{(j)})} = \frac{\pi_m L^{(m)}(y; X)}{\sum_{j=1}^{3} \pi_j L^{(j)}(y; X)},$$

where $L^{(m)}(y; X)$ is the marginal likelihood under model $M = m$. 

• Note that the marginal likelihood is obtained by integrating the likelihood under the $m$th model across the prior for the parameters in that model.

• The posterior distribution of $\beta^{(m)}$ given that $M = m$ is simply the posterior distribution of the parameters in the $m$th model assuming that $m$ is the true model:

$$
\pi(\beta^{(m)} | M = m, y, X) = \frac{\pi(\beta^{(m)})L^{(m)}(y : X, \beta^{(m)}, M = m)}{L^{(m)}(y; X)}.
$$

we can obtain samples from this posterior using a standard MCMC algorithm - implemented exactly as if we were fitting one GLM under model $M = m$

• Note that the parameters under the different models are not on a comparable scale, so it does not make sense to average parameters across different models
• However, quantities such as the posterior probability that the slope is positive, can be averaged:

\[ \text{Pr}(\text{Increasing dose response} \mid \text{data}) = \sum_{m=1}^{3} \text{Pr}(\text{Increasing dose response} \mid M = m, \text{data}) \text{Pr}(M = m \mid \text{data}) \]

\[ = \sum_{m=1}^{3} \text{Pr}(\beta_1^{(m)} > 0 \mid \text{data}) \text{Pr}(M = m \mid \text{data}) \]

• This approach is very useful for performing inferences about predictor effects that accommodate model uncertainty

• The posterior probability \( \text{Pr}(\beta_1^{(m)} \mid \text{data}) \) can be estimated by running an MCMC algorithm under model \( M = m \), and using

\[ \bar{\text{Pr}}(\beta_1^{(m)} \mid \text{data}) = \frac{1}{T} \sum_{t=1}^{T} 1(\beta_1^{(m)t} > 0), \]

where \( t = 1, \ldots, T \) indexes the MCMC iteration starting at \( t = 1 \) after discarding a burn-in to allow convergence.
• It remains to calculate $Pr(M = m \mid \text{data})$, which is a function of the known prior probabilities for each model and the marginal likelihoods under each model

• The marginal likelihoods are difficult to compute and standard MCMC algorithms cannot be used directly

• There is a rich literature on approximating Bayes factors using the Laplace method.
• A rough approximation is provided by Clyde (2000):

\[
\Pr(M = m | \text{data}) \approx \hat{\pi}_m = \frac{\exp\{1/2(D_m - p_m \log c)\}}{\sum_{j=1}^{3} \exp\{1/2(D_j - p_j \log c)\}},
\]

• \(D_m = -2 \times \) maximized log-likelihood under null model - maximized log likelihood under model \(M = m\)

• \(p_m\)=number of parameters in model \(m\)

• \(\log c\)=calibration constant (for BIC \(\log c = \log n\)).

• This choice corresponds to the BIC-approximation to the posterior model probability
• Better approximations will be considered later

• This approximation to the posterior model probability can be calculated by fitting each model in the list using maximum likelihood

• We then apply these weights to the output from an MCMC algorithm applied for each model in the list to obtain the model-averaged posterior probabilities of interest (e.g., of an increasing dose-response)
• Potentially, to avoid the MCMC step one can instead use a normal approximation to the posterior of $\beta^{(m)}$.

• However, with some computational expense, better results can be obtained.

• We can apply this approach not only for posterior probabilities, but also for functionals of interest such as

1. The estimated dose-response curve $f(d)$:

$$\hat{f}(d) = \sum_{m=1}^{3} \bar{P}r(M = m \mid \text{data}) \hat{f}^{(m)}(d)$$

$$= \sum_{m=1}^{3} \hat{\pi}_m \hat{f}^{(m)}(d),$$

$$\hat{f}^{(m)}(d) = \frac{1}{T} \sum_{t=1}^{T} h_m(\beta_0^{(m)t} + \beta_1^{(m)t} d),$$

where $h_m(\cdot)$ is the link function mapping from $\mathbb{R} \rightarrow [0, 1]$ under model $M = m$

2. The predicted probability of response for a new subject at a given dose (note this extends automatically to multiple predictors) can be calculated using the same approach
• One of the advantages of model averaging is that one more appropriately quantifies uncertainty in predictions.

• In addition to model-averaged posterior probabilities, dose-response functions, and predicted response probabilities, we can obtain model-averaged credible intervals.

• Suppose we rely on the BIC-approximation to the posterior model probabilities and run a separate MCMC algorithm under each model.

• Then, we can use the following algorithm to estimate \( 100(1-\alpha)\% \) credible intervals for a functional \( \theta = f^{(m)}(\beta^{(m)}) \):

1. Sample \( M^{(t)} \sim \text{multinomial}(\{1, 2, 3\}, \{\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3\}) \), for \( t = 1, \ldots, T \).

2. Run an MCMC algorithm under each model, discard a burn-in, and collect draws \( t = 1, \ldots, T \).

3. Let \( \theta^{(t)} = \sum_{m=1}^{3} 1(M^{(t)} = m)f^{(m)}(\beta^{(m)t}) \), for \( t = 1, \ldots, T \).

4. Calculate the \( \alpha/2 \) and \( 1 - \alpha/2 \) percentile of the empirical distribution formed by the samples \( \{\theta^{(1)}, \ldots, \theta^{(T)}\} \).
Homework Assignment:

1. Download the DDE and preterm birth data available at the website

2. Using preterm birth as a 0/1 outcome, fit probit model using maximum likelihood estimation. Use DDE dose and the important confounding variables as predictors. Summarize results - is DDE an important predictor? [due Thursday]

3. Repeat this analysis using a Bayesian approach implement by a data augmentation Gibbs sampler, choosing priors as you see fit. Report results - are results different from MLE results? Describe convergence.