A Bayesian Approach to Model Selection

- In exploratory regression problems, we often must select which subset of our potential predictor variables produces the "best model."
- A Bayesian may consider the possible models and compare them based on their posterior probabilities.
- Note that if the value of coefficient β_j is 0, then variable X_j is not needed in the model.
- Let $\beta_j = z_j b_j$ for each j, where $z_j = 0$ or 1 and $b_j \in (-\infty, \infty)$.
- Then our model is

$$Y_i = z_0 b_0 + z_1 b_1 X_{i1} + z_2 b_2 X_{i2} + \cdots + z_{k-1} b_{k-1} X_{i,k-1} + \epsilon_i, \ i = 1, \ldots, n$$

where any $z_j = 0$ indicates that this predictor variable does not belong in the model.

A Bayesian Approach to Model Selection

Example: Oxygen uptake example:
X_1 = group, X_2 = age, X_3 = group $\times$ age:

$\mathbf{z} = (z_0, z_1, z_2, z_3)$	True $E[Y \mid \mathbf{x}, \mathbf{b}, \mathbf{z}]$
(1,0,0,0)	b_0
(1,1,0,0)	$b_0 + b_1$ group
(1,0,1,0)	$b_0 + b_2$ age
(1,1,1,0)	$b_0 + b_1$ group $+ b_2$ age
(1,1,1,1)	$b_0 + b_1$ group $+ b_2$ age $+ b_3$ group $\times$ age

A Bayesian Approach to Model Selection

- For each possible value of the vector $\mathbf{z}$, we calculate the posterior probability for that model:
- For any particular $\mathbf{z}^*$, say:

$$\pi(\mathbf{z}^*|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{z}^*)p(\mathbf{y}|\mathbf{X}, \mathbf{z}^*)}{\sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{y}|\mathbf{X}, \mathbf{z})}$$

- This involves a prior $p(\cdot)$ on each possible model — a noninformative approach would be to let all these prior probabilities be equal.
- If there are a large number of potential predictors, we would use a method called **Gibbs sampling**) (more on this later) to search over the many models.

Example of Bayesian Model Selection

- Example in `R` with Oxygen Data Set
- We can consider all possible subsets of set of predictor variables:
- We can consider only certain subsets (here, we only consider including the interaction term when both first-order terms appear):

The Posterior Predictive Distribution of the Data

- Suppose we have built our Bayesian regression model using response data $\mathbf{y}$ and explanatory data matrix $\mathbf{X}$.
- Suppose we consider future observations whose explanatory variable values are in the matrix $\mathbf{X}^*$.
- What is the marginal distribution of the corresponding future response values $\mathbf{y}^*$?
- This is the **posterior predictive distribution**

$$\pi(\mathbf{y}^* | \mathbf{y}, \mathbf{X}^*, \mathbf{X}).$$

- We will use this later as a tool for checking the fit of our regression model.

The Posterior Predictive Distribution of the Data

- In our analysis with the noninformative priors, note that

$$\pi(\mathbf{y}^*, \boldsymbol{\beta}, \sigma^2 | \mathbf{y}, \mathbf{X}^*, \mathbf{X}) = \pi(\mathbf{y}^* | \boldsymbol{\beta}, \sigma^2, \mathbf{X}^*) \pi(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y})$$

- Then integrating out $\boldsymbol{\beta}$ and σ^2, it can be shown that the posterior predictive distribution of $\mathbf{y}^*$ is multivariate-t with $(n-k)$ degrees of freedom so that

$$E(\mathbf{y}^*) = \mathbf{X}^* \hat{\mathbf{b}} \text{ and}$$

$$\text{covariance matrix } = \frac{(n-k)\hat{\sigma}^2}{n-k-2}[\mathbf{I} + \mathbf{X}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}^{*\prime}]$$

- **Intuition**: Our original data are multivariate normal, given the model.
- Our future predictions are multivariate-t (reflects added uncertainty about the model).