Conjugate Analysis for the Linear Model

- If we have good prior knowledge that can help us specify priors for $\boldsymbol{\beta}$ and σ^2, we can use conjugate priors.
- Following the procedure in Christensen, Johnson, Branscum, and Hanson (2010), we will actually specify a prior for the error **precision** parameter $\tau = \frac{1}{\sigma^2}$:

$$\tau \sim \text{gamma}(a, b)$$

- This is analogous to placing an **inverse gamma** prior on σ^2.
- Then our prior on $\boldsymbol{\beta}$ will depend on τ:

$$\boldsymbol{\beta}|\tau \sim MVN\Big(\boldsymbol{\delta}, \tau^{-1}[\tilde{\mathbf{X}}^{-1}\mathbf{D}(\tilde{\mathbf{X}}^{-1})']\Big)$$

(Note $\tau^{-1} = \sigma^2$)

Conjugate Analysis for the Linear Model

- We will specify a set of k *a priori* **reasonable** hypothetical observations having predictor vectors $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k$ (these — along with a column of 1's — will form the rows of $\tilde{\mathbf{X}}$) and prior expected response values $\tilde{\mathbf{y}}_1, \ldots, \tilde{\mathbf{y}}_k$.
- Our MVN prior on $\boldsymbol{\beta}$ is equivalent to a MVN prior on $\tilde{\mathbf{X}}\boldsymbol{\beta}$:

$$\tilde{\mathbf{X}}\boldsymbol{\beta}|\tau \sim MVN(\tilde{\mathbf{y}}, \tau^{-1}\mathbf{D})$$

- Hence prior mean of $\tilde{\mathbf{X}}\boldsymbol{\beta}$ is $\tilde{\mathbf{y}}$, implying that the prior mean $\boldsymbol{\delta}$ of $\boldsymbol{\beta}$ is $\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{y}}$.
- $\mathbf{D}^{-1}$ is a diagonal matrix whose diagonal elements represent the weights of the "hypothetical" observations.
- Intuitively, the prior has the same "worth" as $\mathrm{tr}(\mathbf{D}^{-1})$ observations.

Conjugate Analysis for the Linear Model

- The joint density is

$$\pi(\boldsymbol{\beta}, \tau, \mathbf{X}, \mathbf{y}) \propto \tau^{n/2}\tau^{n/2}|\mathbf{D}|^{-1/2}\tau^{a-1}e^{-b\tau}$$
$$\times \exp\left\{-\frac{1}{2}(\mathbf{X}\beta - \mathbf{y})^{'}(\tau^{-1}\mathbf{I})^{-1}(\mathbf{X}\beta - \mathbf{y})\right\}$$
$$\times \exp\left\{-\frac{1}{2}(\tilde{\mathbf{X}}\beta - \tilde{\mathbf{y}})^{'}(\tau^{-1}\mathbf{D})^{-1}(\tilde{\mathbf{X}}\beta - \tilde{\mathbf{y}})\right\}$$

- It can be shown that the conditional posterior for $\boldsymbol{\beta}|\tau$ is:

$$\boldsymbol{\beta}|\tau, \mathbf{X}, \mathbf{y} \sim MVN\left(\hat{\boldsymbol{\beta}}, \tau^{-1}(\mathbf{X}^{'}\mathbf{X} + \tilde{\mathbf{X}}^{'}\mathbf{D}^{-1}\tilde{\mathbf{X}})^{-1}\right)$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{'}\mathbf{X} + \tilde{\mathbf{X}}^{'}\mathbf{D}^{-1}\tilde{\mathbf{X}})^{-1}[\mathbf{X}^{'}\mathbf{y} + \tilde{\mathbf{X}}^{'}\mathbf{D}^{-1}\tilde{\mathbf{y}}]$$

Conjugate Analysis for the Linear Model

- And the posterior for τ is:

$$\tau|\mathbf{X}, \mathbf{y} \sim \text{gamma}\left(\frac{n+2a}{2}, \frac{n+2a}{2}s^*\right)$$

where

$$s^* = \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) + (\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta})'\mathbf{D}^{-1}(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\hat{\beta}) + 2b}{n+2a}$$

- The subjective information is incorporated via $\hat{\beta}$ (a function of $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$) and s^* (a function of $\hat{\beta}$, a, and b).

Conjugate Analysis for the Linear Model

- While the conditional posterior $\pi(\boldsymbol{\beta}|\tau, \mathbf{X}, \mathbf{y})$ is multivariate normal, the marginal posterior $\pi(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})$ is a (scaled) **noncentral multivariate t-distribution**.
- In making inference about $\boldsymbol{\beta}$, it is easier to use the conditional posterior for $\boldsymbol{\beta}|\tau$.
- Rather than basing inference on the posterior for $\boldsymbol{\beta}|\hat{\tau}$ (by plugging in a posterior estimate of τ), it is more appropriate to sample random values $\tau^{[1]}, \ldots, \tau^{[J]}$ from the posterior distribution of τ, and then randomly sample from the conditional posterior of $\boldsymbol{\beta}|\tau^{[j]}, j = 1, \ldots, J$.
- Posterior point estimates and interval estimates can then be based on those random draws.

Prior Specification for the Conjugate Analysis

- We will specify a matrix $\tilde{\mathbf{X}}$ of hypothetical predictor values.
- We also specify (via expert opinion or previous knowledge) a corresponding vector $\tilde{\mathbf{y}}$ of reasonable response values for such predictors.
- The number of such "hypothetical observations" we specify must be one more than the number of predictor variables in the regression.
- Our prior mean for β will be $\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{y}}$.

Prior Specification for the Conjugate Analysis

- We also must specify the shape parameter a and the rate parameter b for the gamma prior on τ.
- One strategy is to choose a first, based on the degree on confidence in our prior.
- For a given a, we can view the prior as being "worth" the same as $2a$ sample observations.
- A larger value of a indicates we are more confident in our prior.

Prior Specification for the Conjugate Analysis

- Here is one strategy for specifying b:
- Consider any of the “hypothetical observations” — take the first, for example.
- If $\tilde{\mathbf{y}}_1$ is the prior expected response for a hypothetical observation with predictors $\tilde{\mathbf{x}}_1$, then let $\tilde{\mathbf{y}}_{\text{max}}$ be the *a priori* **maximum reasonable response** for a hypothetical observation with predictors $\tilde{\mathbf{x}}_1$.
- Then (based on the normal distribution) let a prior guess for σ be $\dfrac{\tilde{\mathbf{y}}_{\text{max}} - \tilde{\mathbf{y}}_1}{1.645}$.
- Since $\tau = \dfrac{1}{\sigma^2}$, this gives us a reasonable guess for τ.
- Set this guess for τ equal to the mean $\dfrac{a}{b}$ of the gamma prior for τ.
- Since we have already specified a, we can solve for b.

Example of a Conjugate Analysis

- Example in `R` with Automobile Data Set
- We can get point and interval estimates for τ (and thus for σ^2).

- We can get point and interval estimates for the elements of $\boldsymbol{\beta}$ most easily by drawing from the posterior distributions of τ and then $\beta|\tau$..

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):