If we have good prior knowledge that can help us specify priors for $\beta$ and $\sigma^2$, 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 $\sigma^2$.

Then our prior on $\beta$ will depend on $\tau$:

$$\beta | \tau \sim \text{MVN} \left( \delta, \tau^{-1} \left[ \tilde{X}^{-1} D (\tilde{X}^{-1})' \right] \right)$$

(Note $\tau^{-1} = \sigma^2$)
We will specify a set of \( k \) \textit{a priori reasonable} hypothetical observations having predictor vectors \( \tilde{x}_1, \ldots, \tilde{x}_k \) (these — along with a column of 1’s — will form the rows of \( \tilde{X} \)) and prior expected response values \( \tilde{y}_1, \ldots, \tilde{y}_k \).

Our MVN prior on \( \beta \) is equivalent to a MVN prior on \( \tilde{X}\beta \):

\[
\tilde{X}\beta|\tau \sim \text{MVN}(\tilde{y}, \tau^{-1}D)
\]

Hence prior mean of \( \tilde{X}\beta \) is \( \tilde{y} \), implying that the prior mean \( \delta \) of \( \beta \) is \( \tilde{X}^{-1}\tilde{y} \).

\( D^{-1} \) is a diagonal matrix whose diagonal elements represent the weights of the “hypothetical” observations.

Intuitively, the prior has the same “worth” as \( \text{tr}(D^{-1}) \) observations.
Conjugate Analysis for the Linear Model

The joint density is

$$
\pi(\beta, \tau, X, y) \propto \tau^{n/2} \tau^{n/2} |D|^{-1/2} \tau^{a-1} e^{-b\tau}
$$

\[ \times \exp\left\{-\frac{1}{2} (X\beta - y)' (\tau^{-1} I)^{-1} (X\beta - y)\right\} \]

\[ \times \exp\left\{-\frac{1}{2} (\tilde{X}\beta - \tilde{y})' (\tau^{-1} D)^{-1} (\tilde{X}\beta - \tilde{y})\right\} \]

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

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

where

$$
\hat{\beta} = (X'X + \tilde{X}' D^{-1} \tilde{X})^{-1}[X'y + \tilde{X}' D^{-1} \tilde{y}]
$$
And the posterior for $\tau$ 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 $\pi(\beta | \tau, X, y)$ is multivariate normal, the marginal posterior $\pi(\beta | X, y)$ is a (scaled) noncentral multivariate $t$-distribution.

In making inference about $\beta$, it is easier to use the conditional posterior for $\beta | \tau$.

Rather than basing inference on the posterior for $\beta | \hat{\tau}$ (by plugging in a posterior estimate of $\tau$), it is more appropriate to sample random values $\tau[1], \ldots, \tau[J]$ from the posterior distribution of $\tau$, and then randomly sample from the conditional posterior of $\beta | \tau[j], j = 1, \ldots, J$.

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 $\tilde{X}$ of hypothetical predictor values.
- We also specify (via expert opinion or previous knowledge) a corresponding vector $\tilde{y}$ 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 $\beta$ will be $\tilde{X}^{-1}\tilde{y}$. 
We also must specify the shape parameter $a$ and the rate parameter $b$ for the gamma prior on $\tau$.

One strategy is to choose $a$ first, based on the degree of 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.
Here is one strategy for specifying $b$:

Consider any of the “hypothetical observations” — take the first, for example.

If $\tilde{y}_1$ is the prior expected response for a hypothetical observation with predictors $\tilde{x}_1$, then let $\tilde{y}_{\text{max}}$ be the \textit{a priori maximum reasonable response} for a hypothetical observation with predictors $\tilde{x}_1$.

Then (based on the normal distribution) let a prior guess for $\sigma$ be $\frac{\tilde{y}_{\text{max}} - \tilde{y}_1}{1.645}$.

Since $\tau = \frac{1}{\sigma^2}$, this gives us a reasonable guess for $\tau$.

Set this guess for $\tau$ equal to the mean $\frac{a}{b}$ of the gamma prior for $\tau$.

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 $\tau$ (and thus for $\sigma^2$).

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

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

where any $z_j = 0$ indicates that this predictor variable does not belong in the model.
**Example:** Oxygen uptake example:

$X_1 = \text{group}$, $X_2 = \text{age}$, $X_3 = \text{group \times age}$:

| $z = (z_0, z_1, z_2, z_3)$ | True $E[Y|x, b, z]$ |
|-----------------------------|----------------------|
| $(1,0,0,0)$                 | $b_0$                |
| $(1,1,0,0)$                 | $b_0 + b_1 \text{ group}$ |
| $(1,0,1,0)$                 | $b_0 + b_2 \text{ age}$ |
| $(1,1,1,0)$                 | $b_0 + b_1 \text{ group} + b_2 \text{ age}$ |
| $(1,1,1,1)$                 | $b_0 + b_1 \text{ group} + b_2 \text{ age} + b_3 \text{ group \times age}$ |
For each possible value of the vector $z$, we calculate the posterior probability for that model:

For any particular $z^*$, say:

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

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):