Caterpillar Regression Example: Conjugate Priors, Conditional & Marginal Posteriors, Predictive Distribution, Variable Selection

Prof. Nicholas Zabaras
University of Notre Dame
Notre Dame, IN, USA

Email: nzabarast@gmail.com
URL: https://www.zabarash.com/

September 22, 2017
The Caterpillar Regression Problem, Linear Regression Model, MLE Estimator, Likelihood, MLE Estimates for our example, Conjugate Priors, Conditional and Marginal Posteriors, Ridge regression, Predictive Distribution, Implementation, Influence of the Conjugate Prior

Zellner’s G Prior, Marginal Posterior Mean and Variance, Predictive Modeling, Credible Intervals

Jeffreys’ non-informative Prior, Credible Intervals, Zellner’s G Prior Marginal Distribution of $y$, Point Null Hypothesis and Bayes Factors

Variable Selection, Model Competition, Variable Selection-Prior, Stochastic Search for the Most Probable Model, Gibb’s Sampling for Variable Selection, Implementation

Regression

- Regression refers to statistical analysis that deals with the representation of dependencies between several variables.

- In particular, we want to find a representation of the distribution $f(y|\theta,x)$ of an observable variable $y$ given a vector of observables $x$, using samples of $(x_i,y_i), i=1,\ldots,n$.

- Here, we will consider a particular example of modeling dependencies in pine processionary caterpillar colony size.
Linear Regression Models

- In linear regression, we analyze the linear influence of some variables on others.

- In our particular example of pine processionary caterpillar colonies,
  
  - **Response Variable**: \( y \) is the number of processionary caterpillar colonies
  
  - **Explanatory Variables**: Covariates \( x = (x_1, x_2, \ldots, x_k) \) as defined next (in general can be continuous, discrete or mixed type)
Caterpillar Regression Problem

The pine processionary caterpillar colony size \( (y=\text{number of nests}) \) is influenced by the following explanatory variables:

- \( x_1 \) is the altitude (in meters)
- \( x_2 \) is the slope (in degrees)
- \( x_3 \) is the number of pines in the square
- \( x_4 \) is the height (in meters) of the tree sampled at the center of the square
- \( x_5 \) is the diameter of the tree sampled at the center of the square
- \( x_6 \) is the index of the settlement density
- \( x_7 \) is the orientation of the square (from 1 if southbound to 2 otherwise)
- \( x_8 \) is the height (in meters) of the dominant tree
- \( x_9 \) is the number of vegetation strata
- \( x_{10} \) is the mix settlement index (from 1 if not mixed to 2 if mixed).
Caterpillar Regression Problem

Semilog-y plot of the data \((x_i, y), \ i=1,..,9\)

An implementation is available
MatLab, C++
The distribution of $y$ given $x$ is considered in the context of a set of experimental data, $i = 1, \ldots, n$, on which both $y$, and $x_{i1}, \ldots, x_{ik}$ are measured.

The dataset is made from the outcomes $y = (y_1, \ldots, y_n)$ and of the $n \times (k + 1)$ matrix of explanatory variables

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ 1 & x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$
The most common linear regression model is of the form:

\[ y \mid \beta, \sigma^2, X \sim \mathcal{N}_n(X \beta, \sigma^2 I_n) \]

From this, we conclude that:

\[
\mathbb{E}[y_i \mid \beta, X] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}
\]

\[
\text{Var}[y_i \mid \sigma^2, X] = \sigma^2
\]

Note the difference between finite valued regressors \( x_i \) (like in our caterpillar problem) and categorical variables which also take finite number of values but with range that has no numerical meaning.

Makes no sense to involve \( x \) directly in the regression: replace the single regressor \( x \) [belonging in \{1, \ldots, m\}, say] with \( m \) indicator (or dummy) variables.
Categorical Variables

- Note the difference between finite valued regressors $x_i$ (like in our caterpillar problem) and categorical variables which also take finite number of values but with range that has no numerical meaning.

- Makes no sense to use $x$ directly in the regression: replace the single regressor $x$ [in $\{1, \ldots, m\}$, say] with $m$ indicator variables

  $$ x_1 = \mathbb{I}_1(x), \quad x_2 = \mathbb{I}_2(x), \ldots, \quad x_m = \mathbb{I}_m(x) $$

- Use different $\beta_i$ for each class categorical variable value:

  $$ \mathbb{E}[y_i \mid \beta, X] = \ldots + \beta_1 \mathbb{I}_1(x) + \ldots + \beta_m \mathbb{I}_m(x) + \ldots $$

- Identifiability requires eliminating one of the classes (e.g. $\beta_1=0$) since

  $$ \sum_i \mathbb{I}_i(x) = 1 $$
Returning back to our linear regression model:

\[ y \mid \beta, \sigma^2, X \sim \mathcal{N}_n(X \beta, \sigma^2 I_n) \]

From this, we conclude that:

\[ \mathbb{E}[y_i \mid \beta, X] = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \]
\[ \text{Var}[y_i \mid \sigma^2, X] = \sigma^2 \]

Assume that \( k+1 < n \) and that \( X \) is of full rank: \( \text{rank}(X) = k + 1 \)
- \( X \) is of full rank if and only if \( X^T X \) is invertible

The likelihood is then:

\[ \ell(\beta, \sigma^2 \mid y, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T (y - X\beta)\right) \]
The MLE of $\beta$ is the solution of the least squares minimization problem:

$$
\min_{\beta} (y - X \beta)^T (y - X \beta) = \min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y
$$

Since $\hat{\beta} = (X^T X)^{-1} X^T y$ is a linear transform of $y \sim \mathcal{N}_n(X \beta, \sigma^2 I_n)$,

$$
\hat{\beta} \sim \mathcal{N}((X^T X)^{-1} X^T X \beta, (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1}) = \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})
$$

Thus $\hat{\beta}$ is an unbiased estimator and $\text{Var}(\hat{\beta}|\sigma^2, X) = \sigma^2 (X^T X)^{-1}$.

Also $\hat{\beta}$ is the best linear unbiased estimator of $\beta$:

$$
\alpha \in \mathbb{R}^{k+1}, \text{Var}(\alpha^T \hat{\beta}|\sigma^2, X) \leq \text{Var}(\alpha^T \tilde{\beta}|\sigma^2, X) \text{ where } \tilde{\beta} \text{ is any unbiased linear estimator of } \beta
$$
MLE Estimator

- The MLE of \( \sigma^2 \) is the solution of:

\[
\min_{\sigma^2} \frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) - \frac{n}{2}\ln \sigma |_{\beta=\hat{\beta}} = 0 \implies \sigma^2_{MLE} = \frac{(y - X\hat{\beta})^T (y - X\hat{\beta})}{n}
\]

- Let \( M = X(X^TX)^{-1}X^T \) be the projection \( n \times n \) matrix of the data \( y \) on the column space of \( X \)

\[
\mathbb{E}(\sigma^2_{MLE}) = \frac{\mathbb{E}[(y - X\hat{\beta})^T (y - X\hat{\beta})]}{n} = \frac{\mathbb{E}[(y - My)^T (y - My)]}{n} = \frac{\mathbb{E}[y^T (I - M)^T (I - M)y]}{n} = \frac{\mathbb{E}[y^T (I - M)y]}{n}
\]

- You can easily show that (see this slide):

For \( Y \) a \( n \)-dim random vector with \( \mathbb{E}[Y] = \mu, \text{Cov}[Y] = V \), and \( A \) a \( n \times n \) matrix then:

\[
\mathbb{E}[Y^T AY] = \text{tr}(AV) + \mu^T A \mu
\]
MLE Estimator

\[ \mathbb{E}\left(\sigma^2_{\text{MLE}}\right) = \frac{\mathbb{E}\left[y^T (I - M)y\right]}{n} \]

- You can easily show that (see this slide):

For \( Y \) a \( n \)-dim random vector with \( \mathbb{E}[Y] = \mu, \text{Cov}[Y] = V \), and \( A \) a \( n \times n \) matrix then:

\[ \mathbb{E}\left[Y^T AY\right] = \text{tr}(AV) + \mu^T A \mu \]

- Applying this we have:

\[ \mathbb{E}\left[y^T (I - M)y\right] = \text{tr}\left( (I - M) \text{Cov}(y) \right) + \mathbb{E}(y)^T (I - M) \mathbb{E}(y) = \]

\[ = \sigma^2 \text{rank}(I - M) + \beta^T X^T (I - M)X \beta = \sigma^2 \left(n - k - 1\right) \]

Note: \( I - M \) has 1 as its eigenvalue with multiplicity \( n - k - 1 \) and

\[ tr\left(I - M\right) = tr\left(I\right) - tr\left(M\right) = n - tr\left(X\left(X^TX\right)^{-1}X^T\right) = n - tr\left(\left(X^TX\right)^{-1}\right) = n - tr\left(I_{k+1}\right) = n - k - 1 \]
The expectation of the MLE of $\sigma^2$ is then given as:

$$\mathbb{E}(\sigma^2_{\text{MLE}}) = \frac{\sigma^2}{n} (n - k - 1)$$

The unbiased estimator of $\sigma^2$ is thus given:

$$\hat{\sigma}^2 = \frac{1}{n - k - 1} (y - X\hat{\beta})^T (y - X\hat{\beta}) = \frac{s^2}{n - k - 1}, \text{ where: } s^2 \triangleq (y - X\hat{\beta})^T (y - X\hat{\beta})$$

Indeed using the result from the earlier slide we have:

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n - k - 1} \mathbb{E} \left[(y - X\hat{\beta})^T (y - X\hat{\beta})\right] = \frac{\sigma^2(n - k - 1)}{n - k - 1} = \sigma^2$$

To approximate the covariance of $\hat{\beta}$, in $\hat{\beta} = \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$

use:

$$\text{Var}(\hat{\beta} | \sigma^2, X) = \hat{\sigma}^2 (X^T X)^{-1}$$
We can prove this theorem quite easily as follows:

$$E\left[ (Y - \mu)^T A (Y - \mu) \right] = E\left[ Y^T AY + \mu^T A \mu - \mu^T AY - Y^T A \mu \right] \Rightarrow$$

$$E\left[ (Y - \mu)^T A (Y - \mu) \right] = E\left[ Y^T AY \right] + \mu^T A \mu - \mu^T A \mu - \mu^T A \mu = E\left[ Y^T AY \right] - \mu^T A \mu$$

Using the identity $E\left[ tr(W) \right] = tr\left[ E(W) \right]$ for any random square matrix $W$, we can write the l.h.s. of the equation above as:

$$E\left[ (Y - \mu)^T A (Y - \mu) \right] = E\left[ tr\left( (Y - \mu)^T A (Y - \mu) \right) \right] = E\left[ tr\left( A (Y - \mu)(Y - \mu)^T \right) \right] =$$

$$tr\left[ E\left( A (Y - \mu)(Y - \mu)^T \right) \right] = tr\left( A E\left((Y - \mu)(Y - \mu)^T\right) \right) = tr(AV)$$

This proves the theorem.
T-Statistic

- We can define the standard t-statistic as follows:

\[ T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \sim T(n - k - 1, 0, 1), \text{ where: } \omega_{(i,i)} = (X^T X)^{-1}|_{(i,i)} \]

- This statistic can be used for hypothesis testing, e.g.

accept \( H_0 : \beta_i = 0 \) versus \( H_1 : \beta_i \neq 0 \) at the level \( \alpha \) if

\[ \frac{|\hat{\beta}_i|}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \leq F_{n-k-1}^{-1}(1 - a/2), \text{ the } (1 - a/2)\text{nd quantile of } T_{n-k-1} \]
The frequentist argument in using this bound is that there is significant evidence against $H_0$ if the p-value is smaller than $\alpha$.

$$p_i = P_{H_0} \left( |T_i| > |t_i| = \frac{|\hat{\beta}_i|}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \right) =$$

$$P_{H_0}(T_i < -|t_i|) + P_{H_0}(T_i > |t_i|) = F_{n-k-1}(-|t_i|) + (1 - F_{n-k-1}(|t_i|)) < \alpha$$

Finally the statistic $T_i$ can be used to derive the frequentist marginal confidence interval is:

$$\{ \beta_i: |\beta_i - \hat{\beta}_i| \leq \sqrt{\hat{\sigma}^2 \omega_{ii}} F_{n-k-1}^{-1}(1 - \alpha/2) \}$$
MLE (Least Squares) Estimates

\[
\hat{\beta}_i \quad \sqrt{\hat{\sigma}^2 \omega_{ii}} \quad t_i = \frac{\hat{\beta}_i}{\sqrt{\hat{\sigma}^2 \omega_{ii}}} \quad P_{H_0}\left(|T_i| > \frac{|\hat{\beta}_i|}{\sqrt{\hat{\sigma}^2 \omega_{ii}}}\right)
\]

| Coefficients | Estimate  | Std.Error | t-value | Pr(>|t|) |
|--------------|-----------|-----------|---------|----------|
| intercept    | 10.998412 | 3.060892  | 3.594   | 0.00161 ** |
| XV1          | -0.004431 | 0.001557  | -2.846  | 0.00939 ** |
| XV2          | -0.053830 | 0.021904  | -2.458  | 0.02232 *  |
| XV3          | 0.067939  | 0.099492  | 0.683   | 0.50174   |
| XV4          | -1.293636 | 0.563925  | -2.294  | 0.03168 *  |
| XV5          | 0.231637  | 0.104399  | 2.219   | 0.03709 *  |
| XV6          | -0.356800 | 1.566782  | -0.228  | 0.82193   |
| XV7          | -0.237469 | 1.006210  | -0.236  | 0.81558   |
| XV8          | 0.181060  | 0.236772  | 0.765   | 0.45248   |
| XV9          | -1.285316 | 0.865023  | -1.486  | 0.15142   |
| XV10         | -0.433106 | 0.735018  | -0.589  | 0.56162   |

An implementation is available [MatLab](#), [C++](#)
Note that the likelihood above can now be written in the following very useful form:

$$\ell(\beta, \sigma^2 | y, X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} (y - X\beta)^T (y - X\beta) \right)$$

\[= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left( -\frac{1}{2\sigma^2} s^2 - \frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) \right) \]

where: 
\[s^2 \triangleq (y - X\hat{\beta})^T (y - X\hat{\beta}),\]
\[\hat{\beta} = (X^T X)^{-1} X^T y\]
Observing the form of the likelihood suggests the following conjugate prior:

$$\beta | \sigma^2, X \sim \mathcal{N}_{k+1}(\bar{\beta}, \sigma^2 M^{-1}), \quad M \text{ a } (k + 1) \times (k + 1) \text{ pos. def. symm. matrix}$$

$$\sigma^2 | X \sim \text{InvGamma}(a, b), \quad a, b > 0$$

The posterior is then

$$\pi(\beta, \sigma^2 | \bar{\beta}, s^2, X) \propto \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} s^2 - \frac{1}{2\sigma^2} (\beta - \bar{\beta})^T X^T X (\beta - \bar{\beta}) \right)$$

$$\times \frac{1}{(2\pi \sigma^2)^{(k+1)/2}} \exp \left( -\frac{1}{2\sigma^2} (\beta - \bar{\beta})^T M (\beta - \bar{\beta}) \right) \times \left( \frac{1}{\sigma^2} \right)^{a+1} \exp \left( -\frac{1}{\sigma^2} b \right) =$$

$$= \sigma^{-k-1-2a-2-n} \exp \left( -\frac{1}{2\sigma^2} \left\{ s^2 + 2b + (\beta - \bar{\beta})^T X^T X (\beta - \bar{\beta}) + (\beta - \bar{\beta})^T M (\beta - \bar{\beta}) \right\} \right) =$$

$$= \sigma^{-k-n-2a-3} \exp \left( -\frac{1}{2\sigma^2} \left\{ s^2 + 2b + \bar{\beta}^T (M + X^T X) \beta - 2\beta^T (M\bar{\beta} + X^T X\bar{\beta}) + \bar{\beta}^T M \bar{\beta} + \bar{\beta}^T (X^T X) \bar{\beta} \right\} \right)$$
Conjugate Priors: Posterior of $\beta$ given $\sigma^2$

$\pi(\beta, \sigma^2 | \beta, s^2, X) \propto \sigma^{-k-n-2a-3} \exp \left( -\frac{1}{2\sigma^2} \left\{ s^2 + 2b + \beta^T (M + X^TX)\beta - 2\beta^T (M\bar{\beta} + X^TX\bar{\beta}) + \bar{\beta}^T M\bar{\beta} + \bar{\beta}^T (X^TX)\bar{\beta} \right\} \right) = \sigma^{-k-n-2a-3} \exp \left( -\frac{1}{2\sigma^2} \left\{ s^2 + 2b + \left( \beta - (M + X^TX)^{-1}(M\bar{\beta} + X^TX\bar{\beta}) \right)^T (M + X^TX) \left( \beta - (M + X^TX)^{-1}(M\bar{\beta} + X^TX\bar{\beta}) \right) + \bar{\beta}^T M\bar{\beta} + \bar{\beta}^T X^TX\bar{\beta} \right. \right. \\
\left. \left. \left. - (M\bar{\beta} + X^TX\bar{\beta})^T (M + X^TX)^{-T} (M + X^TX)(M + X^TX)^{-1}(M\bar{\beta} + X^TX\bar{\beta}) \right) \right\} \right)$

Based on this, the posterior of $\beta$ given $\sigma^2$ is:

$\pi(\beta | \sigma^2, y, X) \propto \exp \left( -\frac{1}{2\sigma^2} \left( \beta - \mathbb{E}[\beta | y, X] \right)^T (M + X^TX) \left( \beta - \mathbb{E}[\beta | y, X] \right) \right)$

$\beta | \sigma^2, y, X \sim \mathcal{N}_{k+1} \left( (M + X^TX)^{-1} \left\{ X^TX\bar{\beta} + M\bar{\beta} \right\}, \sigma^2 (M + X^TX)^{-1} \right)$

where

$\mathbb{E}[\beta | \sigma^2, y, X] = (M + X^TX)^{-1}(M\bar{\beta} + X^TX\bar{\beta})$

and

$\text{Var}[\beta | \sigma^2, y, X] = \sigma^2 (M + X^TX)^{-1}$
Conjugate Priors: Marginal posterior of $\sigma^2$

\[
\sigma^{-k-n-2a-3}\exp\left(-\frac{1}{2\sigma^2}\left\{s^2 + 2b + (\beta - \mathbb{E}[\beta|\sigma^2, y, X])^T(M + X^TX)(\beta - \mathbb{E}[\beta|\sigma^2, y, X]) + \tilde{\beta}^T M \tilde{\beta} + \tilde{\beta}^T X^TX \tilde{\beta} \right. \right. \\
\left. \left. - (M \tilde{\beta} + X^T X \tilde{\beta})^T(M + X^TX)^{-1}(M \tilde{\beta} + X^T X \tilde{\beta}) \right\} \right)
\]

- Integrating out $\beta$ gives the following marginal for $\sigma^2$:

\[
\pi(\sigma^2|\tilde{\beta}, s^2, X) \propto \sigma^{-n-2a-2}\exp\left(-\frac{1}{2\sigma^2}\left\{\tilde{\beta}^T M \tilde{\beta} + \tilde{\beta}^T X^TX \tilde{\beta} + s^2 \right. \right. \\
\left. \left. + 2b - (M \tilde{\beta} + X^T X \tilde{\beta})^T(M + X^TX)^{-1}(M \tilde{\beta} + X^T X \tilde{\beta}) \right\} \right)
\]

- To simplify, we will use the following two identities

\[
A: \left( M + X^TX \right)^{-1} = \left( X^TX \right)^{-1} - \left( X^TX \right)^{-1} \left( M^{-1} + \left( X^TX \right)^{-1} \right)^{-1} \left( X^TX \right)^{-1}
\]
\[
B: X^TX \left( M + X^TX \right)^{-1} M = \left( M^{-1} + \left( X^TX \right)^{-1} \right)^{-1}
\]
We can simplify the last term above using identities A & B:

\[
\begin{align*}
\left( M\tilde{\beta} + X^T X\tilde{\beta} \right)^T (M + X^T X)^{-1} \left( M\tilde{\beta} + X^T X\tilde{\beta} \right) &= (\text{expand}) \\
2\tilde{\beta}^T X^T X (M + X^T X)^{-1} M\tilde{\beta} + \tilde{\beta}^T M^T (M + X^T X)^{-1} M\tilde{\beta} + \tilde{\beta}^T X^T X (M + X^T X)^{-1} X^T X\tilde{\beta} = \\
&\quad (X^T X)^{-1} - (X^T X)^{-1} (M^{-1} + (X^T X)^{-1})^{-1} (X^T X)^{-1} \\
2\tilde{\beta}^T X^T X (M + X^T X)^{-1} M\tilde{\beta} + \tilde{\beta}^T M^T (M + X^T X)^{-1} M\tilde{\beta} + \tilde{\beta}^T X^T X\tilde{\beta} - \tilde{\beta}^T (M^{-1} + (X^T X)^{-1})^{-1} \tilde{\beta} = \\
&\quad (M^{-1} + (X^T X)^{-1})^{-1} \quad (M - X^T X (M + X^T X)^{-1} M = M - (M^{-1} + (X^T X)^{-1})^{-1} \\
2\tilde{\beta}^T (M^{-1} + (X^T X)^{-1})^{-1} \tilde{\beta} + \tilde{\beta}^T M\tilde{\beta} - \tilde{\beta}^T (M^{-1} + (X^T X)^{-1})^{-1} \tilde{\beta} + \tilde{\beta}^T X^T X\tilde{\beta} - \tilde{\beta}^T (M^{-1} + (X^T X)^{-1})^{-1} \tilde{\beta} = \\
&\quad \tilde{\beta}^T M\tilde{\beta} + \tilde{\beta}^T X^T X\tilde{\beta} - (\tilde{\beta} - \tilde{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\tilde{\beta} - \tilde{\beta})
\end{align*}
\]

Finally:

\[
\pi(\sigma^2|\tilde{\beta}, s^2, X) \propto \sigma^{-n-2a-2} \exp \left\{-\frac{1}{2\sigma^2} \left( (\tilde{\beta} - \tilde{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\tilde{\beta} - \tilde{\beta}) + s^2 + 2b \right) \right\}
\]
Conjugate Priors: Marginal Posterior of $\sigma^2$

$$\pi(\sigma^2|\beta, s^2, X) \propto \sigma^{-n-2a-2} \exp \left\{ -\frac{1}{2\sigma^2} \left( (\beta - \beta) (M^{-1} + (X^TX)^{-1})^{-1} (\beta - \beta) + s^2 + 2b \right) \right\}$$

- The marginal posterior is:

$$\sigma^2|y, X \sim \text{InvGamma} \left( \frac{n}{2} + a, b + \frac{\beta - \bar{\beta}}{2} \left( M^{-1} + (X^TX)^{-1} \right)^{-1} \left( \beta - \bar{\beta} \right) \right)$$

- The marginal posterior mean for $n \geq 2$ is then:

$$\mathbb{E}^{\pi} [\sigma^2|y, X] = \frac{2b + s^2 + (\beta - \bar{\beta}) (M^{-1} + (X^TX)^{-1})^{-1} (\beta - \bar{\beta})}{n + 2a - 2}$$

Inverse-gamma

\[ p(\theta) = \text{Inv-gamma}(\theta|\alpha, \beta) \]

- shape $\alpha > 0$
- scale $\beta > 0$

\[ p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \theta > 0 \]

\[ \mathbb{E}(\theta) = \frac{\beta}{\alpha-1}, \text{ for } \alpha > 1 \]

\[ \text{var}(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \alpha > 2 \]

\[ \text{mode}(\theta) = \frac{\beta}{\alpha+1} \]
Conjugate Priors: MLE, Posterior Mean and Ridge Estimator

- Note that setting $M = I_{k+1}/c$, $c > 0$ and $\tilde{\beta} = 0_{k+1}$ in the conditional posterior mean

\[
\mathbb{E}[\beta | \sigma^2, y, X] = (M + X^T X)^{-1}(M\tilde{\beta} + X^T X\tilde{\beta})
\]

we obtain the classical Ridge Regression estimate:

\[
\mathbb{E}[\beta | \sigma^2, y, X] = \left(\frac{1}{c} I_{k+1} + X^T X\right)^{-1} X^T X\tilde{\beta} = \left(\frac{1}{c} I_{k+1} + X^T X\right)^{-1} X^T y
\]

- The general estimator $\mathbb{E}[\beta | \sigma^2, y, X]$ can be seen as a weighted average of the prior mean and the MLE.
Conjugate Priors: Marginal Posterior of $\beta$

$$\pi(\beta | y, X) = \int \pi(\beta | \sigma^2, y, X) \pi(\sigma^2 | y, X) d\sigma^2$$

$$\pi(\beta | \sigma^2, y, X) \propto \mathcal{N}_{k+1} \left( (M + X^T X)^{-1} \{X^T \bar{\beta} + M \bar{\beta} \}, \sigma^2 (M + X^T X)^{-1} \right)$$

$$\pi(\sigma^2 | y, X) \propto \mathcal{IG} \left( \frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\bar{\beta} - \hat{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\bar{\beta} - \hat{\beta})}{2} \right)$$

To integrate, we use the following transformations:

$$\tau = 1/\sigma^2, \quad d\tau = -\tau^2 d\sigma^2, \quad \hat{\mu} = \{M + X^T X\}^{-1} [X^T \bar{\beta} + M \bar{\beta}]$$

$$Z = \tau \frac{1}{2} \left[ (\beta - \hat{\mu})^T (M + X^T X) (\beta - \hat{\mu}) + 2b + s^2 + (\bar{\beta} - \hat{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\bar{\beta} - \hat{\beta}) \right]$$

$$d\tau = \frac{2}{G} dZ$$

| Inverse-gamma | $\theta \sim \text{Inv-gamma}(\alpha, \beta)$ | shape $\alpha > 0$ | $p(\theta) = \text{Inv-gamma}(\theta | \alpha, \beta)$ | scale $\beta > 0$ |
Conjugate Priors: Marginal Posterior of $\beta$

\[
\tau = 1/\sigma^2, \quad d\tau = -\tau^2 d\sigma^2, \quad \tilde{\mu} = \{M + X^T X\}^{-1}[X^T X\tilde{\beta} + M\tilde{\beta}]
\]

\[
Z = \tau \frac{1}{2} \left[ (\beta - \tilde{\mu})^T (M + X^T X)(\beta - \tilde{\mu}) + 2b + s^2 + (\tilde{\beta} - \tilde{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1}(\tilde{\beta} - \tilde{\beta}) \right]
\]

\[
d\tau = \frac{2}{G} dZ
\]

\[\square \text{ The marginal now takes the form:}\]

\[
\pi(\beta|y, X) \propto \int_0^\infty e^{-Z^2} \frac{(k+1)}{2} + \frac{n+a+1}{2} d\sigma^2 \sim \int_0^\infty e^{-Z^2} \frac{k}{\tau^2} + \frac{n+a+1}{2} d\tau
\]

\[
\sim G^{-\frac{k}{2} - \frac{n}{2} - a - \frac{1}{2}} \int_0^\infty e^{-Z^2} \frac{k}{Z^2} + \frac{n+a+1}{2} dZ \sim G^{-\frac{k}{2} - \frac{n}{2} - a - \frac{1}{2}}
\]

where

\[
G = (\beta - \tilde{\mu})^T (M + X^T X)(\beta - \tilde{\mu}) + 2b + s^2 + (\tilde{\beta} - \tilde{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1}(\tilde{\beta} - \tilde{\beta})
\]

\[\square \text{ The last integral above is a scalar quantity.}\]
Thus integrating out $\sigma^2$ leads to a multivariate Student’s t marginal posterior on $\beta$:

$$
\pi(\beta|y, X) \propto \left(\beta - \hat{\mu}\right)^T (M + X^T X)(\beta - \hat{\mu}) + 2b + s^2 + (\bar{\beta} - \hat{\beta})^T \left(M^{-1} + (X^T X)^{-1}\right)\left(M^{-1} + (X^T X)^{-1}\right)^{-1}(\bar{\beta} - \hat{\beta})
$$

where:

$$
\hat{\mu} = \{M + X^T X\}^{-1}[X^T X\hat{\beta} + M\bar{\beta}]
$$

Recall that the density of the multivariate $t_p(\nu, \theta, \Sigma)$ is:

$$
\mathcal{T}_p(\nu, \theta, \Sigma) = \frac{\Gamma\left(\frac{(\nu + p) / 2}{2}\right)/\Gamma\left(\nu / 2\right)}{\sqrt{\det(\Sigma)\nu\pi}} \left[1 + \frac{(t - \theta)^T \Sigma^{-1} (t - \theta)}{\nu}\right]^{-\frac{\nu + p}{2}}
$$
We can now see that our marginal
\[
\pi(\beta|y, X) \propto [(\beta - \hat{\mu})^T(M + X^TX)(\beta - \hat{\mu}) + 2b + s^2 + (\bar{\beta} - \bar{\beta})^T(M^{-1} + (X^TX)^{-1})^{-1}(\bar{\beta} - \bar{\beta})]^{-k - n - a - 1/2}
\]
can be written (can check with substitution of the Eqs below) in the form of
\[
\mathcal{I}_p (v, \theta, \Sigma) = \frac{\Gamma((v + p)/2) / \Gamma(v/2) [1 + (t - \theta)^T \Sigma^{-1} (t - \theta)]^{v + p}}{\sqrt{\det(\Sigma) v \pi}}
\]
as follows \((p=k+1, v=n+2a)\):

\[
\beta|y, X \sim T_{k+1}(n + 2a, \hat{\mu}, \hat{\Sigma})
\]
\[
\hat{\mu} = (M + X^TX)^{-1}((X^TX)\bar{\beta} + M\bar{\beta})
\]
\[
\hat{\Sigma} = \frac{2b + s^2 + (\bar{\beta} - \bar{\beta})^T(M^{-1} + (X^TX)^{-1})^{-1}(\bar{\beta} - \bar{\beta})}{n + 2a}(M + X^TX)^{-1}
\]
Recall the properties of the multivariate Student’s-\( \mathcal{T} \) distribution:

<table>
<thead>
<tr>
<th>Multivariate Student-( t )</th>
<th>( \theta \sim t_\nu(\mu, \Sigma) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(\theta) = t_\nu(\theta</td>
<td>\mu, \Sigma) )</td>
</tr>
<tr>
<td>degrees of freedom ( \nu &gt; 0 )</td>
<td>location ( \mu = (\mu_1, \ldots, \mu_d) )</td>
</tr>
<tr>
<td>symmetric, pos. definite</td>
<td>( d \times d ) scale matrix ( \Sigma )</td>
</tr>
</tbody>
</table>

\[
p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}/2} |\Sigma|^{-1/2} \times (1 + \frac{1}{\nu}(\theta - \mu)^T\Sigma^{-1}(\theta - \mu))^{-(\nu+d)/2}
\]

\[
E(\theta) = \mu, \text{ for } \nu > 1
\]
\[
\text{var}(\theta) = \frac{\nu}{\nu-2} \Sigma, \text{ for } \nu > 2
\]
\[
\text{mode}(\theta) = \mu
\]

*Note that in the notation of the tables above the degrees of freedom of the distribution rather than the dimensionality are shown as subscripts, e.g., \( t_\nu \) vs \( t_d \) in the earlier slide. Hopefully the notation will be clear from the discussion.*

Bayesian Data Analysis, A. Gelman, J. Carlin, H. Stern and D. Rubin, 2004
For a given \((m,k+1)\) explanatory matrix \(\tilde{X}\), the outcome \(\tilde{y}\) can be inferred through the predictive distribution*

\[
\pi(\tilde{y}|\sigma^2, y, X, \tilde{X})
\]

Since \(\pi(\tilde{y}|\tilde{X}, \beta, \sigma^2) \propto N(\tilde{X}\beta, \sigma^2 I_m)\),

and since the posterior of \(\beta\) conditional on \(\sigma^2\) is given as

\[
\beta|\sigma^2, y, X \sim N_{k+1}((M + X^T X)^{-1}\{X^T X\beta + M\tilde{\beta}\}, \sigma^2(M + X^T X)^{-1})
\]

we can see that:

\[
\pi(\tilde{y}|\sigma^2, y, X, \tilde{X}) \propto \int \pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X}) \pi(\beta|\sigma^2, y, X) d\beta
\]

is a Gaussian.

* We will later on integrate \(\sigma^2\) out to compute: \(\pi(\tilde{y}|y, X, \tilde{X})\)
We will make use next of the following conditional expectation equalities:

\[ \mathbb{E}[X|Z] = \mathbb{E}[\mathbb{E}[X|Y,Z]|Z] \]

\[ \text{Var}[X|Z] = \text{Var}[\mathbb{E}[X|Y,Z]|Z] + \mathbb{E}[\text{Var}[X|Y,Z]|Z] \]

The proof of these is quite straightforward, e.g.

\[ \mathbb{E}[X|Z] = \int_x x p(x|z)dx = \int_x x \int_y p(x,y|z)dy \, dx \]

\[ \int_x x \int_y p(x|y,z)p(y|z)dy \, dx = \int_y \left( \int_x x p(x|y,z)dx \right) p(y|z)dy = \mathbb{E}[\mathbb{E}[X|Y,Z]|Z] \]
We can compute the mean and the variance of this predictive distribution as follows:

\[
\mathbb{E}_\pi[\tilde{y}|\sigma^2, y, X, \tilde{X}] = \mathbb{E}_\pi(\mathbb{E}_\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}) = \\
\mathbb{E}_\pi(\tilde{X}\beta|\sigma^2, y, X, \tilde{X}) = \tilde{X}(M + X^TX)^{-1}(X^T\tilde{X} \hat{\beta} + M\tilde{\beta})
\]

and

\[
\text{Var}_\pi[\tilde{y}|\sigma^2, y, X, \tilde{X}] = \\
\mathbb{E}_\pi(\text{Var}_\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}) + \text{Var}_\pi(\mathbb{E}_\pi(\tilde{y}|\beta, \sigma^2, y, X, \tilde{X})|\sigma^2, y, X, \tilde{X}) = \\
\mathbb{E}_\pi(\sigma^2 I_m|\sigma^2, y, X, \tilde{X}) + \text{Var}_\pi(\tilde{X}\beta|\sigma^2, y, X, \tilde{X}) = \\
\sigma^2 I_m + \tilde{X}\sigma^2 (M + X^TX)^{-1}\tilde{X}^T = \sigma^2(I_m + \tilde{X}(M + X^TX)^{-1}\tilde{X}^T)
\]

In conclusion:

\[
\tilde{y}|\sigma^2, y, X, \tilde{X} \sim \mathcal{N}_m\left(\tilde{X}\mathbb{E}[\beta|\sigma^2, y, X], \sigma^2(I_m + \tilde{X}\text{Var}[\beta|\sigma^2, y, X]\tilde{X}^T)\right) \\
\mathbb{E}[\beta|\sigma^2, y, X] = (M + X^TX)^{-1}\{X^T\tilde{X} \hat{\beta} + M\tilde{\beta}\}
\]
We now integrate $\sigma^2$ against the posterior distribution to obtain:

$$\pi(\tilde{y}|y, X, \tilde{X}) \propto \int_0^\infty \mathcal{N}(\tilde{X} \mathbb{E}[\beta|\sigma^2, y, X], \sigma^2 I_m + \tilde{X} \text{Var}[\beta|\sigma^2, y, X] \tilde{X}^T) \times \mathcal{IG}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \bar{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\tilde{\beta} - \bar{\beta})}{2}\right) d\sigma^2$$
Conjugate Priors: Predictive Distribution

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \int_0^\infty \mathcal{N}(\tilde{X}E[\beta|\sigma^2, y, X], \sigma^2 I_m + \tilde{X}Var[\beta|\sigma^2, y, X]\tilde{X}^T) \\
\times J\mathcal{G}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{\beta - \tilde{\beta}}{2} (M^{-1} + (X^T X)^{-1})^{-1} (\beta - \tilde{\beta})\right) d\sigma^2
\]

- We substitute:

\[
\sigma^2 I_m + \tilde{X}Var[\beta|\sigma^2, y, X]\tilde{X}^T = \sigma^2 (I_m + \tilde{X}(M + X^T X)^{-1}\tilde{X}^T)
\]

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \int_0^\infty \mathcal{N}(\tilde{X}E[\beta|\sigma^2, y, X], \sigma^2 (I_m + \tilde{X}(M + X^T X)^{-1}\tilde{X}^T)) \\
\times J\mathcal{G}\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{\beta - \tilde{\beta}}{2} (M^{-1} + (X^T X)^{-1})^{-1} (\beta - \tilde{\beta})\right) d\sigma^2
\]

\[
\propto \int_0^\infty \sigma^{-m-n-2a-2} \exp\left(-\frac{1}{2\sigma^2} \left\{2b + s^2 + (\beta - \tilde{\beta})^T (M^{-1} + (X^T X)^{-1})^{-1} (\beta - \tilde{\beta}) +
\right\}

\left[(\tilde{y} - E[\tilde{y}|\sigma^2, y, X, \tilde{X}])^T (I_m + \tilde{X}(M + X^T X)^{-1}\tilde{X}^T)^{-1} (\tilde{y} - E[\tilde{y}|\sigma^2, y, X, \tilde{X}])\right\}\right) d\sigma^2
\]
Conjugate Priors: Predictive Distribution

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto
\int_{0}^{\infty} \sigma^{-m-n-2a-2} \exp\left(-\frac{1}{2\sigma^2} \{2b + s^2 + (\tilde{\beta} - \beta)^T (M^{-1} + (X^TX)^{-1})^{-1} (\tilde{\beta} - \beta) +
(\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])^T (I_m + \tilde{X}(M + X^TX)^{-1}\tilde{X}^T)^{-1} (\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])\} \right) d\sigma^2
\]

If we call the term inside \{.\} as G, then:

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \int_{0}^{\infty} \sigma^{-m-n-2a-2} e^{-\frac{1}{2\sigma^2} G} d\sigma^2 \Rightarrow
\]

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \int_{0}^{\infty} (\sigma^2)^{-\frac{m+n+2a+2}{2}} e^{-\frac{1}{2\sigma^2} G} d\sigma^2 \Rightarrow \pi(\tilde{y}|y, X, \tilde{X}) \propto G^{-\frac{m+n+2a+2}{2}} \int_{0}^{\infty} Z^{-\frac{m+n+2a+2}{2}} e^{-Z} \left(-\frac{1}{2} G \frac{1}{Z^2} dZ \right) \sim G^{-\frac{m+n+2a}{2}}
\]

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \{2b + s^2 + (\tilde{\beta} - \beta)^T (M^{-1} + (X^TX)^{-1})^{-1} (\tilde{\beta} - \beta) +
(\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])^T (I_m + \tilde{X}(M + X^TX)^{-1}\tilde{X}^T)^{-1} (\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])\}^{-\frac{(m+n+2a)}{2}}
\]
Conjugate Priors: Predictive Distribution

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \{2b + s^2 + (\tilde{\beta} - \beta)^T (M^{-1} + (X^T X)^{-1})^{-1} (\tilde{\beta} - \beta) + \\
(\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])^T (I_m + \tilde{X}(M + X^T X)^{-1} \tilde{X}^T)^{-1} (\tilde{y} - \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}])\}^{-(m+n+2a)/2}
\]

This corresponds to a Student’s t distribution:

\[
\mathcal{I}_p(\nu, \theta, \Sigma) = \frac{\Gamma((\nu + p)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\det(\Sigma)^{\nu\pi}}} \left[1 + \frac{(t - \theta)^T \Sigma^{-1} (t - \theta)}{\nu}\right]^{-\frac{\nu + p}{2}}
\]

with

\[
\tilde{y}|y, X, \tilde{X} \sim \mathcal{I}_m(n + 2a, \tilde{\mu}, \tilde{\Sigma}) \\
\tilde{\mu} = \mathbb{E}[\tilde{y}|\sigma^2, y, X, \tilde{X}] = \tilde{X}(M + X^T X)^{-1}(X^T X\tilde{\beta} + M\tilde{\beta}) \\
\tilde{\Sigma} = \left\{2b + s^2 + (\tilde{\beta} - \beta)^T (M^{-1} + (X^T X)^{-1})^{-1} (\tilde{\beta} - \beta)\right\} (I_m + \tilde{X}(M + X^T X)^{-1} \tilde{X}^T) / (n + 2a)
\]
Implementation of the Conjugate Prior

Our Prior:
\[ \beta | \sigma^2, X \sim \mathcal{N}_{k+1}(\bar{\beta}, \sigma^2 M^{-1}), \quad M = I_{k+1}/c, \quad a(k + 1) \times (k + 1) \text{ pos. def. symm. matrix} \]
\[ \sigma^2 | X \sim IG(a, b), \quad a, b > 0 \]

- We assume that there is no precise information available about \( \bar{\beta}, M, a, b \).

- Let us choose as \( M = I_{k+1}/c \) and \( \bar{\beta} = 0_{k+1} \).

- Let us take \( a = 2.1 \) and \( b = 2 \), i.e. prior mean and prior variance of \( \sigma^2 \) equal to 1.82 and 33.06.

- The intuition is that if \( c \) is large, the prior on \( \beta \) should be more diffuse and have less bearing on the outcome.

- However, this turns out not to be the case. There is lasting influence of \( c \) on the posterior means of \( \sigma^2 \) and \( \beta_0 \).
## Influence of the Prior Scale $c$ on Bayesian Estimates of $\beta$

### Conjugate Priors

| $c$  | $\mathbb{E}[\sigma^2 | y, X]$ | $\mathbb{E}[\beta_0 | y, X]$ | $\text{Var}[\beta_0 | y, X]$ |
|------|------------------------------|-----------------------------|-----------------------------|
| 0.1  | 1.0044                       | 0.1251                      | 0.0988                      |
| 1.0  | 0.8541                       | 0.9031                      | 0.7733                      |
| 10.0 | 0.6976                       | 4.7299                      | 3.8991                      |
| 100.0| 0.5746                       | 9.6626                      | 6.8355                      |
| 1000.0| 0.5470                      | 10.8476                     | 7.3419                      |

An implementation is available [MatLab, C++](#).

\[
\mathbb{E}_\pi[\sigma^2 | y, X] = \frac{2b + s^2 + (\tilde{\beta} - \hat{\beta})^T (M^{-1} + (X^TX)^{-1})^{-1}(\tilde{\beta} - \hat{\beta})}{n + 2a - 2}
\]

\[
\mathbb{E}_\pi[\beta | y, X] = (M + X^TX)^{-1} \left((X^TX)\hat{\beta} + M\tilde{\beta}\right)
\]

\[
\text{Var}[\beta | y, X] = \frac{2b + s^2 + (\tilde{\beta} - \hat{\beta})^T (M^{-1} + (X^TX)^{-1})^{-1}(\tilde{\beta} - \hat{\beta})}{n + 2a} (M + X^TX)^{-1}
\]
### Bayesian Estimates of $\beta$ for $c=100$

#### Conjugate Priors

Bayes estimates of beta for $c=100$

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>$E[\beta_i \mid y, X]$</th>
<th>$V[\beta_i \mid y, X]$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>9.6626</td>
<td>6.8355</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.0040</td>
<td>0.0000</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.0516</td>
<td>0.0004</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.0418</td>
<td>0.0077</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-1.2633</td>
<td>0.2615</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.2307</td>
<td>0.0090</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>-0.0832</td>
<td>1.9310</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>-0.1917</td>
<td>0.8254</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>0.1608</td>
<td>0.0462</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>-1.2069</td>
<td>0.6127</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>-0.2567</td>
<td>0.4267</td>
</tr>
</tbody>
</table>

An implementation is available in [MatLab, C++]
The value of $c$ thus has a significant influence on the estimators and even more on the posterior variance.

The Bayes estimates stabilize for very large values of $c$. Thus the prior associated with a particular $c$ should not be considered as a weak or pseudo-noninformative prior but, on the opposite, associated with a specific proper prior information.

The dependence on $(a, b)$ is equally strong.

Considering these limitations of conjugate priors on at least the posterior variance, a more sophisticated noninformative strategy is needed.

We first look a middle-ground perspective which settles the problem of the choice of $M$. 
We start with a middle ground solution by introducing only information about the location parameter of the regression but bypassing the selection of the prior correlation structure.

\[
\beta | \sigma^2, X \sim \mathcal{N}_{k+1}(\tilde{\beta}, c\sigma^2(X^TX)^{-1}) \\
\sigma^2 \sim \pi(\sigma^2|X) \propto \sigma^{-2} 
\]

We start with a middle ground solution by introducing only information about the location parameter of the regression but bypassing the selection of the prior correlation structure.

\begin{align*}
\beta | \sigma^2, X & \sim \mathcal{N}_{k+1}(\bar{\beta}, c \sigma^2 (X^T X)^{-1}) \\
\sigma^2 & \sim \pi(\sigma^2 | X) \propto \sigma^{-2} \text{ improper Jeffrey's prior}
\end{align*}

The prior determination is restricted to the choices of \( \bar{\beta} \) and constant \( c \). \( c \) can be interpreted as a measure of the information available in the prior relative to the sample.

- E.g., we will see that setting \( 1/c = 0.5 \) gives the prior the same weight as 50% of the sample.

There is still strong influence of \( c \).
The joint posterior now takes the form (note $X^TX$ is used in both likelihood and prior):

$$
\pi(\beta, \sigma^2 | y, X) \propto (\sigma^2)^{-\frac{n}{2}} \exp \left( -\frac{1}{2\sigma^2} (y - X\hat{\beta})^T (y - X\hat{\beta}) - \frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) (\sigma^2)^{-\frac{(k+1)}{2}} \\
\times \exp \left( -\frac{1}{2c\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) (\sigma^2)^{-1}
$$

We can now compute the following:

$$
\beta | \sigma^2, y, X \sim \exp \left( -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) - \frac{1}{2c\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) \\
\sim \exp \left( -\frac{1}{2} \left( \beta - \frac{c}{c+1} \left( \frac{\beta}{c} + \hat{\beta} \right) \right)^T \frac{c+1}{c\sigma^2} X^T X \left( \beta - \frac{c}{c+1} \left( \frac{\beta}{c} + \hat{\beta} \right) \right) \right) \Rightarrow \\
\beta | \sigma^2, y, X \sim \mathcal{N}_{k+1} \left( \frac{c}{c+1} \left( \frac{\beta}{c} + \hat{\beta} \right), \frac{c\sigma^2}{c+1} (X^T X)^{-1} \right)
$$
Starting again with the posterior

$$\pi(\beta, \sigma^2|y, X) \propto (\sigma^2)^{-(n+k+3/2)} \exp \left(-\frac{1}{2\sigma^2} s^2 - \frac{1}{2\sigma^2} (\beta - \beta)^T X^T X(\beta - \beta) \right)$$

$$\times \exp \left(-\frac{1}{2c\sigma^2} (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) \right) =$$

$$\sim (\sigma^2)^{-(k+1)/2} \exp \left(-\frac{1}{2} \left(\beta - \frac{c}{c+1} \left(\frac{\hat{\beta}}{c} + \hat{\beta}\right) \right)^T \frac{c+1}{c\sigma^2} X^T X \left(\beta - \frac{c}{c+1} \left(\frac{\hat{\beta}}{c} + \hat{\beta}\right) \right) \right)$$

$$\times (\sigma^2)^{-(n/2+1)} \exp \left(-\frac{1}{2\sigma^2} \left(s^2 + \frac{1}{c+1} (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) \right) \right)$$

We can now compute the following posterior marginal with integration in $\beta$:

$$\sigma^2|y, X \sim \text{InvGamma} \left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(c+1)} (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}) \right)$$

$$\mathbb{E}[\sigma^2|y, X] = \frac{s^2 + \frac{1}{(c+1)} (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})}{n-2}$$
Zellner's Informative G-Prior: Posterior Marginal of $\beta$

- Starting again with the posterior

$$
\pi(\beta, \sigma^2 | y, X) \propto (\sigma^2)^{-(n+k+3/2)} \exp\{- \frac{1}{2\sigma^2} \left[ \left( \beta - \frac{c}{c+1} (\bar{\beta} + \hat{\beta}) \right)^T \frac{c+1}{c} X^T X \left( \beta - \frac{c}{c+1} (\bar{\beta} + \hat{\beta}) \right) + s^2 + \frac{1}{c+1} (\bar{\beta} - \hat{\beta})^T X^T X (\bar{\beta} - \hat{\beta}) \right] \}
$$

- We can now compute the following posterior marginal with integration in $\sigma^2$:

$$
\beta | y, X \sim \mathcal{T}_{k+1} \left( \left( \beta - \frac{c}{c+1} (\bar{\beta} + \hat{\beta}) \right)^T \frac{c+1}{c} X^T X \left( \beta - \frac{c}{c+1} (\bar{\beta} + \hat{\beta}) \right) + s^2 + \frac{1}{c+1} (\bar{\beta} - \hat{\beta})^T X^T X (\bar{\beta} - \hat{\beta}), n, \frac{c}{c+1} (\bar{\beta} + \hat{\beta}), \frac{c \left( s^2 + (\bar{\beta} - \hat{\beta})^T X^T X (\bar{\beta} - \hat{\beta})/(c+1) \right)}{n(c+1)} \right)^{-1} X^T X (X^T X)^{-1}
$$

<table>
<thead>
<tr>
<th>Multivariate Student-t</th>
<th>$\theta \sim t_\nu(\mu, \Sigma)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(\theta) = t_\nu(\theta</td>
<td>\mu, \Sigma)$</td>
</tr>
<tr>
<td>location $\mu = (\mu_1, \ldots, \mu_d)$</td>
<td>symmetric, pos. definite</td>
</tr>
<tr>
<td>$d \times d$ scale matrix $\Sigma$</td>
<td></td>
</tr>
</tbody>
</table>

- $p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}|\Sigma|^{-1/2}} \times (1 + \frac{1}{\nu}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu))^{-(\nu+d)/2}$

- $E(\theta) = \mu$, for $\nu > 1$

- $\text{var}(\theta) = \frac{1}{\nu-2} \Sigma$, for $\nu > 2$

- $\text{mode}(\theta) = \mu$
The Bayes estimate of $\beta$ can be derived from:

$$\beta | y, X \sim T_{k+1} \left( n, \frac{c}{c+1} \left( \bar{\beta} + \hat{\beta} \right), \frac{c \left( s^2 + (\bar{\beta} - \hat{\beta})^T X^T X (\bar{\beta} - \hat{\beta}) / (c + 1) \right)}{n(c+1)} \right) \left( X^T X \right)^{-1}$$

$$\mathbb{E}[\beta | y, X] = \frac{1}{c+1} (\bar{\beta} + c\hat{\beta})$$

The posterior variance of $\beta$ is also given from above as:

$$V_{\pi}[\beta | y, X] = \frac{c}{c+1} \frac{\left( s^2 + (\bar{\beta} - \hat{\beta})^T X^T X (\bar{\beta} - \hat{\beta}) / (c + 1) \right)}{n-2} \left( X^T X \right)^{-1}$$

$p(\theta) = \frac{\Gamma((\nu+d)/2)}{\Gamma(\nu/2)\nu^{d/2}\pi^{d/2}} |\Sigma|^{-1/2} \times (1 + \frac{1}{\nu}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu))^{-(\nu+d)/2}$

$\mathbb{E}(\theta) = \mu$, for $\nu > 1$

$\text{var}(\theta) = \frac{\nu}{\nu-2} \Sigma$, for $\nu > 2$

$\text{mode}(\theta) = \mu$
If \( c = 1 \), you can see from the above equation that it is like putting the same weight on the prior information and on the sample:

\[
E[\beta | y, X] = \frac{1}{c + 1} (\bar{\beta} + c\hat{\beta})
\]

which is the average of the prior mean and the MLE estimator.

If \( c = 100 \), the prior weights 1\% of the sample.
The Bayes estimate of $\sigma^2$ (use the mean of InvGamma) can be derived from

$$
\sigma^2 | y, X \sim \mathcal{IG} \left( \frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(c+1)} (\tilde{\beta} - \bar{\beta})^T X^T X (\tilde{\beta} - \bar{\beta}) \right) \\
$$

as

$$
\mathbb{E}[\sigma^2 | y, X] = \frac{s^2 + (\tilde{\beta} - \bar{\beta})^T X^T X (\tilde{\beta} - \bar{\beta})}{(c+1)} \\
\frac{n-2}{(c+1)}
$$

Note that only when $c$ goes to infinity the influence of the prior vanishes.
### Zellner's Informative G-Prior: Marginal Posterior Mean and Variance of β

**Zellner's G-Prior**

Posterior mean and variance of beta for $c=100$

$$\mathbb{E}[\beta | y, X] = \frac{1}{c + 1}(\hat{\beta} + c\bar{\beta})$$
$$V[\beta | y, X] = \frac{c}{c + 1} \left( s^2 + (\hat{\beta} - \bar{\beta})^T X^T X (\hat{\beta} - \bar{\beta}) / (c + 1) \right) \left( X^T X \right)^{-1}$$

| beta_i | $\mathbb{E}[\text{beta}_i | y, X]$ | $V[\text{beta}_i | y, X]$ | $\log_{10}(BF)$ |
|--------|-----------------|-----------------|-----------------|
| (Intercept) | 10.8895 | 6.8229 | 2.1873 (****) |
| X1 | -0.0044 | 0.0000 | 1.1571 (***) |
| X2 | -0.0533 | 0.0003 | 0.6667 (**) |
| X3 | 0.0673 | 0.0072 | -0.8585 |
| X4 | -1.2808 | 0.2316 | 0.4726 (*) |
| X5 | 0.2293 | 0.0079 | 0.3861 (*) |
| X6 | -0.3533 | 1.7877 | -0.9860 |
| X7 | -0.2351 | 0.7373 | -0.9849 |
| X8 | 0.1793 | 0.0408 | -0.8225 |
| X9 | -1.2726 | 0.5449 | -0.3461 |
| X10 | -0.4288 | 0.3934 | -0.8949 |

An implementation is available in MatLab, C++

Evidence against $H_0$:

<table>
<thead>
<tr>
<th>Evidence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(****)</td>
<td>Decisive</td>
</tr>
<tr>
<td>(***)</td>
<td>Strong</td>
</tr>
<tr>
<td>(**)</td>
<td>Substantial</td>
</tr>
<tr>
<td>(*)</td>
<td>Poor</td>
</tr>
</tbody>
</table>

See here for BF calculation
Zellner's G-Prior: Marginal Posterior Mean and Variance of $\beta$

Posterior mean and variance of beta for $c=1000$

| beta_i | $E[\beta_i|y, X]$ | $V[\beta_i|y, X]$ | log10(BF) |
|--------|--------------------|--------------------|-----------|
| (Intercept) | 10.9874 | 6.6644 | 1.7973 (***) |
| X1 | -0.0044 | 0.0000 | 0.7375 (**) |
| X2 | -0.0538 | 0.0003 | 0.2313 (*) |
| X3 | 0.0679 | 0.0070 | -1.3506 |
| X4 | -1.2923 | 0.2262 | 0.0307 |
| X5 | 0.2314 | 0.0078 | -0.0588 |
| X6 | -0.3564 | 1.7461 | -1.4834 |
| X7 | -0.2372 | 0.7202 | -1.4822 |
| X8 | 0.1809 | 0.0399 | -1.3130 |
| X9 | -1.2840 | 0.5323 | -0.8177 |
| X10 | -0.4327 | 0.3843 | -1.3885 |

An implementation is available

MatLab, C++

Evidence against $H_0$:

- (***) decisive
- (*** ) strong
- (**) substantial
- (*) poor

$E[\beta|y, X] = \frac{1}{c + 1} (\bar{\beta} + c\bar{\beta})$

$V[\beta|y, X] = \frac{c}{c + 1} \left( s^2 + (\bar{\beta} - \beta)^T X^T X (\bar{\beta} - \beta)/(c + 1) \right) (X^T X)^{-1}$

See here for BF calculation
We want to predict \((m \geq 1)\) future observations when the explanatory variables \(\tilde{x}\) but not the outcome variables have been observed.

Predictive distribution on \(\tilde{y}\) defined as marginal of the joint posterior distribution on \((\tilde{y}, \beta, \sigma^2)\). Can be computed analytically by

\[
\pi(\tilde{y}|y, X, \tilde{X}) \propto \int \pi(\tilde{y}|\sigma^2, y, X, \tilde{X})\pi(\sigma^2 | y, X, \tilde{X})d\sigma^2
\]
Conditional on $\sigma^2$ the future vector of observations has a Gaussian distribution with

$$
\beta | \sigma^2, y, X \sim N_{k+1} \left( \frac{c}{c+1} \left( \frac{\tilde{\beta}}{c} + \tilde{\beta} \right), \frac{c\sigma^2}{c+1} (X^T X)^{-1} \right)
$$

This representation is quite intuitive, being the product of the matrix of explanatory variables $\tilde{X}$ by the Bayes estimate of $\beta$. 

$$
\mathbb{E}_\pi(\tilde{y} | \sigma^2, y, X, \tilde{X}) = \mathbb{E}_\pi[\mathbb{E}_\pi(\tilde{y} | \beta, \sigma^2, y, X, \tilde{X}) | \sigma^2, y, X, \tilde{X}] = \mathbb{E}_\pi[\tilde{X} \beta | \sigma^2, y, X, \tilde{X}] = \tilde{X} \mathbb{E}_\pi[\beta | \sigma^2, y, X, \tilde{X}] \Rightarrow
$$

$$
\mathbb{E}_\pi(\tilde{y} | \sigma^2, y, X, \tilde{X}) = \tilde{X} \frac{\tilde{\beta} + c\tilde{\beta}}{c + 1} \quad (independent \ of \ \sigma^2)
$$
Zellner's Informative G-Prior: Predictive Modeling

\[ \beta | \sigma^2, y, X \sim \mathcal{N}_{k+1} \left( \frac{c}{c+1} \left( \frac{\hat{\beta}}{c} + \hat{\beta} \right), \frac{c\sigma^2}{c+1} (X^T X)^{-1} \right) \]

- Similarly, we can compute:

\[
V_\pi (\tilde{y} | \sigma^2, y, X, \tilde{X}) = \mathbb{E}_\pi \left[ V (\tilde{y} | \beta, \sigma^2, y, X, \tilde{X}) | \sigma^2, y, X, \tilde{X} \right] \\
+ V_\pi \left[ \mathbb{E}_\pi (\tilde{y} | \beta, \sigma^2, y, X, \tilde{X}) | \sigma^2, y, X, \tilde{X} \right] = \\
= \mathbb{E}_\pi [\sigma^2 I_m | \sigma^2, y, X, \tilde{X}] + V_\pi [\tilde{X} \beta | \sigma^2, y, X, \tilde{X}]
\]

\[
V_\pi (\tilde{y} | \sigma^2, y, X, \tilde{X}) = \sigma^2 \left( I_m + \frac{c}{c+1} \tilde{X} (X^T X)^{-1} \tilde{X}^T \right)
\]
Here, we are interested on the highest posterior density (HPD) regions on subvectors of the parameter $\beta$ derived from the marginal posterior distribution of $\beta$.

For a single parameter,

$$\beta_i | y, X \sim T_1 \left( n, \frac{c}{c + 1} \left( \widehat{\beta}_i + \beta_i \right), \frac{c \left( s^2 + (\widehat{\beta} - \beta)^T X^T X (\widehat{\beta} - \beta) / (c + 1) \right)}{n(c + 1)} \omega_{ii} \right)$$

where $\omega_{ii}$ is the $(i,i)$ element of $(X^T X)^{-1}$.
Let us define

$$\tau = \frac{\tilde{\beta} + c\hat{\beta}}{c + 1}$$

Also

$$K = \frac{c \left( s^2 + (\tilde{\beta} - \hat{\beta})^T X^T X (\tilde{\beta} - \hat{\beta}) / (c+1) \right)}{n(c+1)} (X^T X)^{-1} = (K_{ij})$$

The variable

$$\zeta_i = \frac{\beta_i - \tau_i}{\sqrt{K_{ii}}}$$

has a $t$-distribution with $n$ degrees of freedom.
A 1 − \( \alpha \) HPD interval on \( \beta_i \) is thus given by (see also here)

\[
\left[ \tau_i - \sqrt{K_{ii}}F_n^{-1}(1 - \alpha / 2), \tau_i + \sqrt{K_{ii}}F_n^{-1}(1 - \alpha / 2) \right]
\]

where \( F_n \) is the CDF of \( \mathcal{T}(\nu=n) \).
Zellner’s Informative G-Prior: High Posterior Density (HPD) Intervals for $\beta_i$’s

Note that these HPD are different from the frequentist confidence intervals defined earlier as:

$$
\beta_i | \mathbf{y}, \mathbf{X} \sim \mathcal{T} (n - k - 1, \hat{\beta}_i, \omega_{(i,i)} s^2 / (n - k - 1))
$$

$$
\frac{\beta_i - \hat{\beta}_i}{\omega_{(i,i)} s^2 / (n - k - 1)} \sim \mathcal{T}(\nu = n - k - 1, 0, 1), \text{ where: } \omega_{(i,i)} = (\mathbf{X}^T \mathbf{X})^{-1}|_{(i,i)}
$$

$$
\left\{ \beta_i: |\beta_i - \hat{\beta}_i| \leq F_{n-k-1}^{-1}(1 - \alpha / 2) \sqrt{\omega_{(i,i)} s^2 / (n - k - 1)} \right\}
$$

$$
s^2 \triangleq (\mathbf{y} - \mathbf{X} \hat{\beta})^T (\mathbf{y} - \mathbf{X} \hat{\beta})
$$
Zellner's Informative G-Prior: 95% High Posterior Density (HPD) Intervals for $\beta_i$'s

<table>
<thead>
<tr>
<th>$\beta_i$</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>[4.6518, 17.3450]</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>[-0.0077, -0.0012]</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>[-0.0992, -0.0084]</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>[-0.1384, 0.2742]</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>[-2.4629, -0.1244]</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>[0.0152, 0.4481]</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>[-3.6054, 2.8918]</td>
</tr>
<tr>
<td>$\beta_7$</td>
<td>[-2.3238, 1.8489]</td>
</tr>
<tr>
<td>$\beta_8$</td>
<td>[-0.3099, 0.6720]</td>
</tr>
<tr>
<td>$\beta_9$</td>
<td>[-3.0789, 0.5083]</td>
</tr>
<tr>
<td>$\beta_{10}$</td>
<td>[-1.9571, 1.0909]</td>
</tr>
</tbody>
</table>

An implementation is available in **MatLab, C++**.

\[ \left\{ \beta_i : |\beta_i - \hat{\beta}_i| \leq F_{n-k-1}^{-1}(1 - a/2) \sqrt{\omega_{(i,i)} s^2 / (n - k - 1)} \right\} \]
Zellner’s Informative $G$-Prior: 90% High Posterior Density (HPD) Intervals for $\beta_i$’s

<table>
<thead>
<tr>
<th>beta_i</th>
<th>HPD Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>beta_0</td>
<td>[5.7435, 16.2533]</td>
</tr>
<tr>
<td>beta_1</td>
<td>[-0.0071, -0.0018]</td>
</tr>
<tr>
<td>beta_2</td>
<td>[-0.0914, -0.0162]</td>
</tr>
<tr>
<td>beta_3</td>
<td>[-0.1029, 0.2387]</td>
</tr>
<tr>
<td>beta_4</td>
<td>[-2.2618, -0.3255]</td>
</tr>
<tr>
<td>beta_5</td>
<td>[0.0524, 0.4109]</td>
</tr>
<tr>
<td>beta_6</td>
<td>[-3.0466, 2.3330]</td>
</tr>
<tr>
<td>beta_7</td>
<td>[-1.9649, 1.4900]</td>
</tr>
<tr>
<td>beta_8</td>
<td>[-0.2254, 0.5875]</td>
</tr>
<tr>
<td>beta_9</td>
<td>[-2.7704, 0.1998]</td>
</tr>
<tr>
<td>beta_10</td>
<td>[-1.6950, 0.8288]</td>
</tr>
</tbody>
</table>

\[
\left\{ \beta_i : |\beta_i - \hat{\beta_i}| \right\}
\]

Note: The results given in “C. P. Robert, *The Bayesian Core*, Springer, 2nd edition, chapter 3” refer to 90% HPD intervals rather than 95% as posted. These results agree with what is shown above.
The marginal distribution of \( y \) (evidence) is a multivariate \( T \).

Since \( \beta|\sigma^2, X \sim N_{k+1}(\bar{\beta}, c\sigma^2(X^TX)^{-1}) \), the linear transform of \( \beta \) satisfies:

\[
X\beta|\sigma^2, X \sim N(X\bar{\beta}, c\sigma^2X(X^TX)^{-1}X^T)
\]

which implies

\[
y|\sigma^2, X \sim N_n(X\bar{\beta}, \sigma^2(I_n + cX(X^TX)^{-1}X^T))
\]

Integration in \( \sigma^2 \) with \( \pi(\sigma^2) = 1/\sigma^2 \) gives (see Appendices):

\[
f(y|X, c) = \frac{1}{(2\pi)^{n/2} (c + 1)^{(k+1)/2}} \int_0^\infty \frac{1}{(2\pi)^{n/2} (c + 1)^{(k+1)/2}} (\sigma^2)^{-n/2-1} e^{-\frac{1}{2\sigma^2}(y-X\bar{\beta})^T(I_n+cX(X^TX)^{-1}X^T)^{-1}(y-X\bar{\beta})} d\sigma^2
\]

\[
f(y|X, c) = (c + 1)^{-(k+1)/2} \pi^{-n/2} \Gamma\left(\frac{n}{2}\right) \left[ y^T y - \frac{c}{c + 1} y^T X (X^TX)^{-1} X^T y + \frac{1}{c + 1} \bar{\beta}^T X^T X \bar{\beta} - \frac{2}{c + 1} y^T X \bar{\beta} \right]^{-n/2}
\]
It can be easily shown that
\[
\left( I_n + cX \left( X^T X \right)^{-1} X^T \right)^{-1} = I_n - \frac{c}{c+1} X \left( X^T X \right)^{-1} X^T
\]

Indeed:
\[
\left( I_n - \frac{c}{c+1} X \left( X^T X \right)^{-1} X^T \right) \left( I_n + cX \left( X^T X \right)^{-1} X^T \right) = I_n - \frac{c}{c+1} X \left( X^T X \right)^{-1} X^T + cX \left( X^T X \right)^{-1} X^T
\]
\[
- \frac{c^2}{c+1} X \left( X^T X \right)^{-1} X^T X \left( X^T X \right)^{-1} X^T = I_n + \left( \frac{c}{c+1} - \frac{c^2}{c+1} \right) X \left( X^T X \right)^{-1} X^T = I_n
\]

This can simplify the term in the exponential in the earlier slide:
\[
(y - X\tilde{\beta})^T \left( I_n + cX \left( X^T X \right)^{-1} X^T \right)^{-1} (y - X\tilde{\beta}) = (y - X\tilde{\beta})^T \left( I_n - \frac{c}{c+1} X \left( X^T X \right)^{-1} X^T \right) (y - X\tilde{\beta}) =
\]
\[
y^T y - \frac{c}{c+1} y^T X \left( X^T X \right)^{-1} X^T y - y^T X\tilde{\beta} + \frac{c}{c+1} y^T X \left( X^T X \right)^{-1} X^T X\tilde{\beta}
\]
\[
- \tilde{\beta}^T X^T y + \tilde{\beta}^T X^T X\tilde{\beta} + \frac{c}{c+1} \tilde{\beta}^T X^T \left( X^T X \right)^{-1} X^T y - \frac{c}{c+1} \tilde{\beta}^T X^T X \left( X^T X \right)^{-1} X^T X\tilde{\beta} =
\]
\[
= y^T y - \frac{c}{c+1} y^T X \left( X^T X \right)^{-1} X^T y + \frac{1}{c+1} \tilde{\beta}^T X^T X\tilde{\beta} - \frac{2}{c+1} y^T X\tilde{\beta}
\]
Let us revisit the matrix \( I_n + cX(X^T X)^{-1} X^T \)

Note that for any \( \beta \in \mathbb{R}^{k+1} \):

\[
\left[ I_n + cX(X^T X)^{-1} X^T \right] X \beta = X \beta + cX \beta = (1+c)X \beta
\]

which implies that \( X \beta \) is an eigenvector with eigenvalue \((1+c)\). There are \((k+1)\) of those.

Finally note that for any \( z \) in the null space of \( X^T \), \( X^T z = 0 \),

\[
\left[ I_n + cX(X^T X)^{-1} X^T \right] z = z + cX(X^T X)^{-1} X^T z = z
\]

which implies that these \( z \) \((n-k-1)\) in number\) are eigenvectors of our matrix with \(1\) as the eigenvalues.

The determinant of the matrix is then \( (1+c)^{k+1}1^{n-k-1} = (1+c)^{k+1} \). This explains the derivation in the earlier slide.
Zellner’s Informative $G$-Prior: Point Null Hypothesis

- If a null hypothesis is $H_0: R\beta = r$, $R$ being a $q \times (k+1)$ matrix, the model under $H_0$ can be rewritten as

$$y | \beta^0, \sigma^2, X_0 \sim \mathcal{N}_n (X_0 \beta^0, \sigma^2 I_n)$$

where $\beta^0$ is $(k + 1 - q)$ dimensional.

- Under the prior

$$\beta^0 | \sigma^2, X_0 \sim \mathcal{N}_{k+1-q} (\tilde{\beta}^0, c_0 \sigma^2 (X_0^T X_0)^{-1})$$

- The marginal distribution of $y$ under $H_0$ is:

$$f (y|X_0, H_0) = (c_0 + 1)^{-(k+1-q)/2} \pi^{-n/2} \Gamma \left( \frac{n}{2} \right)$$

$$\times \left[ y^T y - \frac{c_0}{c_0 + 1} y^T X_0 (X_0^T X_0)^{-1} X_0^T y + \frac{1}{c_0 + 1} \tilde{\beta}_0^T X_0^T X_0 \tilde{\beta}_0 - \frac{2}{c_0 + 1} y^T X_0 \tilde{\beta}_0 \right]^{-n/2}$$
The Bayes factor is then given in analytical form as:

\[
B_{10}^\pi = \frac{f(y|X)}{f(y|X_0, H_0)} = \frac{(c_0 + 1)^{(k+1-q)/2}}{(c + 1)^{(k+1)/2}} \times \left[ y^T y - \frac{c_0}{c_0 + 1} y^T X_0 (X_0^T X_0)^{-1} X_0^T y + \frac{1}{c_0 + 1} \beta_0^T X_0^T X_0 \beta_0 - \frac{2}{c_0 + 1} y^T X_0 \beta_0 \\
\frac{y^T y - \frac{c}{c + 1} y^T X (X^T X)^{-1} X^T y + \frac{1}{c + 1} \beta^T X^T X \beta - \frac{2}{c + 1} y^T X \beta}{n/2}
\right]
\]

Note that we use the same \( \sigma^2 \) in both models.

The Bayes factor depends on \( c_0 \) and \( c \).

This calculation can be used to evaluate the inclusion or not of each of the terms \( \beta_i, i = 1, \ldots, k + 1 \) in our regression model.

Note: \( X_0 = X(:, \text{setdiff}(1:k + 1, j)) \)
Noninformative Prior Analysis: Jeffreys’ Prior

- Considering the robustness issues of the two priors examined earlier in the case of a complete lack of prior information, we consider now the non-informative Jeffreys’ prior.

- The Jeffreys’ prior is a flat prior on \((\beta, \log \sigma^2)\)

\[
\pi^J(\beta, \sigma^2 | X) \propto \sigma^{-2}
\]

- The posterior is then given as:

\[
\pi^J(\beta, \sigma^2 | y, X) \propto (\sigma^{-2})^{n/2} \exp \left( -\frac{1}{2\sigma^2} (y - X\hat{\beta})^T (y - X\hat{\beta}) - \frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) \times \sigma^{-2}
\]

\[
= (\sigma^{-2})^{(k+1)/2} \exp \left( -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) \times
\]

\[
(\sigma^{-2})^{(n-k-1)/2+1} \exp \left( -\frac{1}{2\sigma^2} s^2 \right)
\]
Noninformative Prior Analysis: Jeffreys’ Prior

\[ \pi^J(\beta, \sigma^2 | y, X) \propto (\sigma^{-2})^{(k+1)/2} \exp\left(-\frac{1}{2\sigma^2} (\beta - \bar{\beta})^T X^T X (\beta - \bar{\beta}) \right) \times \]

\[ (\sigma^{-2})^{(n-k-1)/2+1} \exp\left(-\frac{1}{2\sigma^2} s^2 \right) \]

- From this joint posterior, we can immediately evaluate the following conditional and marginal posteriors:

\[ \pi^J(\sigma^2 | y, X) \propto (\sigma^2)^{-\frac{(n-k-1)/2-1}{2}} \exp\left(-\frac{1}{2\sigma^2} s^2 \right) = \mathcal{IG}\left(\frac{(n-k-1)}{2}, \frac{s^2}{2}\right) \]

\[ \pi^J(\beta | \sigma^2, y, X) \propto (\sigma^{-2})^{(k+1)/2} \exp\left(-\frac{1}{2\sigma^2} (\beta - \bar{\beta})^T X^T X (\beta - \bar{\beta}) \right) \]

\[ = \mathcal{N}_{k+1}(\bar{\beta}, \sigma^2 (X^T X)^{-1}) \]

- From the 1st of these eqs., the Bayes’ estimate of \( \sigma^2 \) is:

\[ \mathbb{E}(\sigma^2 | y, X) = \frac{s^2}{n-k-3} \]

\[ p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{-(\alpha+1)} e^{-\beta/\theta}, \ \theta > 0 \]

E(\theta) = \frac{\beta}{\alpha-1}, \text{ for } \alpha > 1 \]

\[ \text{var}(\theta) = \frac{\beta^2}{(\alpha-1)(\alpha-2)}, \alpha > 2 \]

\[ \text{mode}(\theta) = \frac{\beta}{\alpha+1} \]

- This estimate is larger (more pessimistic) than earlier estimates
Noninformative Prior Analysis: Jeffreys' Prior

\[
\pi(\beta, \sigma^2 | y, X) \propto (\sigma^{-2})^{n/2+1} \exp \left( - \frac{s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}{2\sigma^2} \right)
\]

- To compute the marginal posterior of \( \beta \), we need to integrate in \( \sigma^2 \). Using symbolic integrator:

\[
\pi(\beta | y, X) \propto \int (\sigma^2)^{-n/2-1} \exp \left( - \frac{s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}{2\sigma^2} \right) d\sigma^2
\]

\[
\propto \left( s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right)^{-n/2} = \left( 1 + \frac{(\beta - \hat{\beta})^T \left( \frac{s^2 (X^T X)^{-1}}{n-k-1} \right)^{-1} (\beta - \hat{\beta})}{n-k-1} \right)^{-n/2}
\]

- This is a Students' \( \mathcal{T} \) distribution:

\[
\pi(\beta | y, X) \propto \mathcal{T}_{k+1}(\nu = n - k - 1, \theta, \Sigma), \text{ with: } \theta = \hat{\beta}, \Sigma = \frac{s^2 (X^T X)^{-1}}{n - k - 1}
\]

- Thus the Bayes' estimate of \( \beta \) is:

\[
\mathbb{E}^{\pi}(\beta | y, X) = \hat{\beta}
\]

- The HPD intervals coincide with the frequentist confidence intervals.
Zellner's Non-Informative G-Prior

- The main difference with informative G-prior setup is that we now consider $c$ as unknown.
- We use the same G-prior distribution with $\tilde{\beta} = 0_{k+1}$ conditional on $c$, and introduce a diffuse prior on $c$,

$$\pi(c) = c^{-1}I_{N^*}(c)$$

- The corresponding marginal posterior is given as:

$$\pi(\beta, \sigma^2 \mid y, X) = \int \pi(\beta, \sigma^2 \mid y, X, c) \pi(c \mid y, X) dc$$

$$\propto \sum_{c=1}^{\infty} \pi(\beta, \sigma^2 \mid y, X, c) \frac{f(y \mid X, c) \pi(c)}{\pi(y \mid X)}$$

$$\propto \sum_{c=1}^{\infty} \pi(\beta, \sigma^2 \mid y, X, c) f(y \mid X, c) c^{-1}$$
\[ \pi(\beta, \sigma^2 | y, X) \propto \sum_{c=1}^{\infty} \pi(\beta, \sigma^2 | y, X, c) f(y | X, c) c^{-1} \]

- **We have proved earlier** (for constant c) that:

\[
f(y|X, c) = (c + 1)^{-(k+1)/2} \pi^{-n/2} \Gamma \left( \frac{n}{2} \right)
\left[ y^T y - \frac{c}{c + 1} y^T X (X^T X)^{-1} X^T y + \frac{1}{c + 1} \tilde{\beta}^T X^T X \tilde{\beta} - \frac{2}{c + 1} y^T X \tilde{\beta} \right]^{-n/2}
\]

- **This for our problem with** \( \tilde{\beta} = 0_{k+1} \)

\[
f(y | X, c) \propto (c + 1)^{-(k+1)/2} \left[ y^T y - \frac{c}{c + 1} y^T X (X^T X)^{-1} X^T y \right]^{-n/2}
\]
Recall that

\[
y, X \sim \mathcal{T}_{k+1}\left(n, \frac{c}{c+1} \left(\bar{\beta} + \tilde{\beta}\right), \frac{c \left(s^2 + (\bar{\beta} - \tilde{\beta})^T X^T X (\bar{\beta} - \tilde{\beta})/(c + 1)\right)}{n(c + 1)}\right)(X^T X)^{-1}
\]

The Bayes estimates of \( \beta \) for \( \tilde{\beta} = 0_{k+1} \) is now given by:

\[
\mathbb{E}_\pi(\beta | y, X) = \mathbb{E}_\pi[\mathbb{E}_\pi(\beta | y, X, c) | y, X] = \mathbb{E}_\pi \left[ \frac{c}{c+1} \tilde{\beta} | y, X \right] = \frac{\sum_{c=1}^{\infty} \frac{c}{c+1} \pi(c | y, X)}{\sum_{c=1}^{\infty} \pi(c | y, X)} \tilde{\beta} \Rightarrow
\]

\[
\mathbb{E}_\pi(\beta | y, X) = \frac{\sum_{c=1}^{\infty} \frac{c}{c+1} f(y | X, c)c^{-1}}{\sum_{c=1}^{\infty} f(y | X, c)c^{-1}} \tilde{\beta}
\]
Recall that

$$\sigma^2|y, X \sim IG \left( \frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(c + 1)} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) \Rightarrow$$

The Bayes estimate of $\sigma^2$ is given similarly by:

$$\mathbb{E}[\sigma^2|y, X] = \frac{s^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta})}{n - 2}$$

Both Bayes estimates involve infinite summations on $c$. The denominator in both cases is the normalizing constant of the posterior $\sum_{c=1}^{\infty} f(y|X, c) c^{-1}$.
Zellner's Non-Informative G-Prior: Posterior Variance of $\beta$

$$\beta|y, X \sim T_{k+1}\left(n, \frac{c}{c+1}\left(\frac{\hat{\beta}}{c} + \bar{\beta}\right), \frac{c\left(s^2 + (\hat{\beta} - \bar{\beta})^T X^T X (\hat{\beta} - \bar{\beta})/(c + 1)\right)}{n(c + 1)}\right) (X^T X)^{-1}$$

$$V_\pi(\beta|y, X) = \mathbb{E}_\pi[V_\pi(\beta|y, X, c)|y, X] + V_\pi[\mathbb{E}_\pi(\beta|y, X, c)|y, X] =$$

$$\mathbb{E}_\pi\left[\frac{c}{(n-2)(c+1)}\left(s^2 + \hat{\beta}^T X^T X \hat{\beta}/(c + 1)\right)(X^T X)^{-1}\right] + V_\pi\left[\frac{c}{(c+1)}\hat{\beta}|y, X\right] =$$

$$\sum_{c=1}^{\infty} \frac{f(y|X, c)}{(n-2)(c+1)} \left( s^2 + \hat{\beta}^T X^T X \hat{\beta}/(c + 1) \right) \left( \frac{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}} \right) (X^T X)^{-1} +$$

$$\hat{\beta} \left[ \sum_{c=1}^{\infty} \left( \frac{c}{(c+1)} - \mathbb{E}_\pi \left( \frac{c}{(c+1)} | y, X \right) \right)^2 f(y|X, c)c^{-1} \right] \left( \frac{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}} \right) \hat{\beta}^T$$

See [earlier slide](#) for the term $f(y|X, c)$
The marginal distribution of the dataset is available in closed form:

\[
f(y|X) = \sum_{c=1}^{\infty} f(y|X,c)c^{-1} \propto \sum_{c=1}^{\infty} c^{-1}(c+1)^{-1/2} \left[ y^T y - \frac{c}{c+1} y^T X \left( X^T X \right)^{-1} X^T y \right]^{-n/2}
\]

The $\mathcal{G}$-shape means that we can also compute the normalizing constant!
### Zellner's Informative G-Prior: Posterior Mean and Variance

| Beta_i | $\mathbb{E}[\beta_i | y, X]$ | $V[\beta_i | y, X]$ |
|--------|-----------------|-----------------|
| (Intercept) | 9.2714 | 9.6424 |
| X1    | -0.0037 | 0.0000 |
| X2    | -0.0454 | 0.0005 |
| X3    | 0.0573  | 0.0092 |
| X4    | -1.0905 | 0.3079 |
| X5    | 0.1953  | 0.0105 |
| X6    | -0.3008 | 2.2750 |
| X7    | -0.2002 | 0.9383 |
| X8    | 0.1526  | 0.0522 |
| X9    | -1.0835 | 0.7063 |
| X10   | -0.3651 | 0.5020 |

The posterior mean $\mathbb{E}(\beta | y, X)$ and variance $V(\beta | y, X)$ can be calculated using the following formulas:

\[
\mathbb{E}(\beta | y, X) = \frac{\sum_{c=1}^{\infty} \frac{c}{c+1} f(y|X, c)c^{-1} \beta}{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}}
\]

\[
V(\beta | y, X) = \left(\sum_{c=1}^{\infty} \frac{f(y|X, c)}{(n-2)(c+1)(s^2 + \hat{\beta}^T X^T X \hat{\beta} / (c+1))} \right)^{-1} + \frac{\left(\sum_{c=1}^{\infty} \frac{c}{c+1} - \mathbb{E}(\beta | y, X) \right)^2 f(y|X, c)c^{-1}}{\sum_{c=1}^{\infty} f(y|X, c)c^{-1}} \hat{\beta}^T
\]

A Matlab implementation can be downloaded [here](#).

(We use $\tilde{\beta} = 0_{11}$, $c = 100$)
Zellner's Non-Informative $G$-Prior: Point Null Hypothesis

- If a null hypothesis is $H_0 : R\beta = r$, the model under $H_0$ can be rewritten as

  \[ y \mid \beta^0, \sigma^2, X_0 \sim \mathcal{N}_n(X_0\beta^0, \sigma^2 I_n) \]

  where $\beta^0$ is $(k + 1 - q)$ dimensional.

- Under $\pi(c) = c^{-1}$ and the prior

  \[ \beta^0 \mid \sigma^2, X_0, c \sim \mathcal{N}_{k+1-q} \left(0_{k+1-q}, c\sigma^2 \left(X_0^T X_0 \right)^{-1}\right) \]

  the marginal distribution of $y$ under $H_0$ is:

  \[ f(y \mid X_0, H_0) \propto \sum_{c=1}^{\infty} c^{-1} (c+1)^{-(k+1-q)/2} \left[ y^T y - \frac{c}{c+1} y^T X_0 \left(X_0^T X_0 \right)^{-1} X_0^T y \right]^{-n/2} \]

- The Bayes factor $B_{10}^\pi = f(y \mid X) / f(y \mid X_0, H_0)$ can now be computed.
### Zellner's Non-Informative G-Prior: Posterior Mean and Variance

For $H_0: \beta_7 = \beta_8 = 0$, $\log_{10}(B^\pi_{10}) = -0.7884$ (We use $\beta = 0_{11}$)

| beta_i | $E[\beta_i | y, X]$ | $V[\beta_i | y, X]$ | $\log_{10}(BF)$ |
|--------|---------------------|-------------------|-----------------|
| (Intercept) | 9.2714 | 9.6424 | 1.4205 (****) |
| X1 | -0.0037 | 0.0000 | 0.8503 (**) |
| X2 | -0.0454 | 0.0005 | 0.5664 (**) |
| X3 | 0.0573 | 0.0092 | -0.3609 |
| X4 | -1.0905 | 0.3079 | 0.4520 (*) |
| X5 | 0.1953 | 0.0105 | 0.4007 (*) |
| X6 | -0.3008 | 2.2750 | -0.4411 |
| X7 | -0.2002 | 0.9383 | -0.4404 |
| X8 | 0.1526 | 0.0522 | -0.3383 |
| X9 | -1.0835 | 0.7063 | -0.0424 |
| X10 | -0.3651 | 0.5020 | -0.3838 |

A Matlab implementation can be downloaded [here](#).

Evidence against $H_0$:
- (****) decisive
- (*** ) strong
- (**) substantial
- (*) poor

**Statistical Computing, University of Notre Dame, Notre Dame, IN, USA (Fall 2017, N. Zabaras)**
Let us return to our regression model with one dependent random variable $y$ and a set of $k \{x_1, x_2, ..., x_k\}$ explanatory variables.

Are all the $x_i$’s needed in the regression?

We assume that every $q$-subset $\{i_1, i_2, ..., i_q\}$, $0 \leq q \leq k$, of the explanatory variables,

$$\left\{1, x_{i_1}, x_{i_2}, ..., x_{i_q}\right\}$$

is a proper set of explanatory variables for the regression of $y$ (as before, the intercept is included in all models).

We have a total of $2^k$ models to select from!
Following earlier notation, we denote: \( X = [\mathbf{1}_n \; x_1 \; x_2 \ldots x_k] \) as the matrix that contains \( \mathbf{1}_n \) and the \( k \) potential predictor variables.

Each model \( M_\gamma \) is associated with binary indicator vector
\[
\gamma \in \Gamma = \{0, 1\}^k
\]
where \( \gamma_i = 1 \) means that the variable \( x_i \) is included in the model \( M_\gamma \) and \( \gamma_i = 0 \) that it is not.

The number of variables included in the model \( M_\gamma \) is:
\[
q_\gamma = \mathbf{1}^T \gamma
\]

The indices of the variables included in the model and not included in the model are denoted, respectively, as: \( t_1(\gamma), t_0(\gamma) \).
Variable Selection - Models in Competition

- For $\beta \in \mathbb{R}^{k+1}$ and $X$, we define $\beta_\gamma$ as the sub-vector
  
  $\beta_\gamma = \left( \beta_0, (\beta_i)_{i \in t_1(\gamma)} \right)$

- Let $X_\gamma$ be the submatrix of $X$ where only the column $1_n$ and the columns in $t_1(\gamma)$ have been left.

- The model $M_\gamma$ is then defined as:

  \[ y \mid \gamma, \beta_\gamma, \sigma^2, X \sim \mathcal{N}_n \left( X_\gamma \beta_\gamma, \sigma^2 I_n \right) \]

  where $\beta_\gamma \in \mathbb{R}^{q_\gamma+1}, \sigma^2 \in \mathbb{R}^*_+$ are the unknown parameters.

- The $\sigma^2$ is common to all models and we use the same prior for all models.
Variable Selection - Models in Competition

• We have a high number $2^k$ of models in competition.

• We cannot specify a prior on every $M_\gamma$ in a completely subjective and autonomous manner.

• We derive all priors from a single global prior associated with the full model that corresponds to $\gamma = (1, \ldots, 1)$. 
For the full model that corresponds to $\gamma = (1, \ldots, 1)$, we use the Zellner’s informative G-prior:

$$
\beta | \sigma^2, X \sim \mathcal{N}_{k+1}(\bar{\beta}, c\sigma^2(X^TX)^{-1})
$$

$$
\sigma^2 \sim \pi(\sigma^2 | X) \propto \sigma^{-2} \text{ improper Jeffreys prior}
$$

For each model $M_\gamma$, the prior distribution of $\beta_\gamma$ conditional on $\sigma^2$ is fixed as:

$$
\beta_\gamma | \gamma, \sigma^2 \sim \mathcal{N}_{q_\gamma+1}(\bar{\beta}_\gamma, c\sigma^2(X_\gamma^TX_\gamma)^{-1})
$$

where

$$
\bar{\beta}_\gamma = (X_\gamma^TX_\gamma)^{-1}X_\gamma^T\bar{\beta}
$$

and same prior on $\sigma^2$. 
The joint prior for model $M_\gamma$ is the improper prior

$$
\pi(\beta_\gamma, \sigma^2 | \mathbf{y}) \propto (\sigma^2)^{-(q_\gamma+1)/2-1} \exp \left[ -\frac{1}{2(c\sigma^2)} (\beta_\gamma - \tilde{\beta}_\gamma)^T (X_\gamma^T X_\gamma) (\beta_\gamma - \tilde{\beta}_\gamma) \right]
$$

Infinitely many ways of defining a prior on the model index $\gamma$:

Our choice is a uniform prior $p(\gamma | \mathbf{X}) = 2^{-k}$.

Posterior distribution of $\gamma$ is central to variable selection since it is proportional to marginal density of $\mathbf{y}$ on $M_\gamma$ (or evidence of $M_\gamma$)
Posterior distribution of $\gamma$ is proportional to the marginal density of $y$ on $M_\gamma$ (so it can also be used to compute Bayes factors)

$$\pi(\gamma \mid y, X) \propto f(y \mid \gamma, X) \pi(\gamma \mid X) \propto f(y \mid \gamma, X)$$

$$= \int \left( \int f(y \mid \gamma, \beta, \sigma^2, X) \pi(\beta \mid \gamma, \sigma^2, X) d\beta \right) \pi(\sigma^2 \mid X) d\sigma^2$$

where (see earlier derivation)

$$f(y \mid \gamma, \sigma^2, X) = \int f(y \mid \gamma, \beta, \sigma^2) \pi(\beta \mid \gamma, \sigma^2) d\beta =$$

$$= (c + 1)^{-(q + 1)/2} (2\pi)^{-n/2} (\sigma^2)^{-n/2} \times$$

$$\exp\left(-\frac{1}{2\sigma^2} y^T y + \frac{1}{2\sigma^2(c + 1)} \left\{ cy^T X_\gamma (X_\gamma^T X_\gamma)^{-1} X_\gamma^T y - \tilde{\beta}_\gamma^T X_\gamma^T X_\gamma \tilde{\beta}_\gamma + 2 y^T X_\gamma \tilde{\beta}_\gamma \right\}\right)$$
Posterior distribution of $\gamma$ is then given as:

$$\pi(\gamma | y, X) \propto \int f(y | \gamma, \sigma^2, X) \pi(\sigma^2 | X) d\sigma^2 = \int f(y | \gamma, \sigma^2, X) \frac{1}{\sigma^2} d\sigma^2$$

We already have seen this distribution earlier for a fixed $c$:

$$f(\gamma | y, X) \propto (c + 1)^{- (q\gamma + 1)/2} \times \left( y^T y - \frac{c}{(c + 1)} y^T X \gamma (X^T \gamma X)^{-1} X^T \gamma y + \frac{1}{c + 1} \beta^T \gamma X^T \gamma X \gamma \beta \gamma - \frac{2}{c + 1} y^T \gamma X \gamma \beta \gamma \right)^{-n/2}$$
Most likely models ordered by decreasing posterior probabilities using Zellner’s informative G-prior with $c=100$.

<table>
<thead>
<tr>
<th>$t_1(\gamma)$</th>
<th>$\pi(\gamma \mid y, X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 4 5</td>
<td>0.231543</td>
</tr>
<tr>
<td>0 1 2 4 5 9</td>
<td>0.037358</td>
</tr>
<tr>
<td>0 1 9</td>
<td>0.034435</td>
</tr>
<tr>
<td>0 1 2 4 5 10</td>
<td>0.032975</td>
</tr>
<tr>
<td>0 1 4 5</td>
<td>0.030606</td>
</tr>
<tr>
<td>0 1 2 9</td>
<td>0.025016</td>
</tr>
<tr>
<td>0 1 2 4 5 7</td>
<td>0.024144</td>
</tr>
<tr>
<td>0 1 2 4 5 8</td>
<td>0.023784</td>
</tr>
<tr>
<td>0 1 2 4 5 6</td>
<td>0.023735</td>
</tr>
<tr>
<td>0 1 2 3 4 5</td>
<td>0.023207</td>
</tr>
<tr>
<td>0 1 6 9</td>
<td>0.014587</td>
</tr>
<tr>
<td>0 1 2 3 9</td>
<td>0.014491</td>
</tr>
<tr>
<td>0 9</td>
<td>0.014281</td>
</tr>
<tr>
<td>0 1 2 6 9</td>
<td>0.013551</td>
</tr>
<tr>
<td>0 1 4 5 9</td>
<td>0.012761</td>
</tr>
<tr>
<td>0 1 3 9</td>
<td>0.011712</td>
</tr>
<tr>
<td>0 1 2 8</td>
<td>0.011477</td>
</tr>
<tr>
<td>0 1 8</td>
<td>0.009519</td>
</tr>
<tr>
<td>0 1 2 3 4 5 9</td>
<td>0.009036</td>
</tr>
<tr>
<td>0 1 2 4 5 6 9</td>
<td>0.009031</td>
</tr>
</tbody>
</table>

An implementation is available: [MatLab](https://www.mathworks.com/), [C++]
Model Selection

- Model $M_1$ with the highest posterior probability is $t_1(\gamma) = (1, 2, 4, 5)$, which corresponds to the variables
  - altitude,
  - slope,
  - height of the tree sampled in the center of the area, and
  - diameter of the tree sampled in the center of the area.
Model Selection

- For the Zellner’s non-informative prior with $\pi(c) = 1/c$, we have ($\tilde{\beta} = 0_{k+1}$): 

\[
\pi(\gamma | y, X) = \sum_{c=1}^{\infty} c^{-1} (c + 1)^{-(q_y+1)/2} \left[ y^T y - \frac{c}{c+1} y^T X \gamma \left( X^T X \gamma \right)^{-1} X^T y \right]^{-n/2} 
\]

- Again we have seen this before as

\[
f(y | X) = \sum_{c=1}^{\infty} f(y | X, c) c^{-1} \propto \\
\sum_{c=1}^{\infty} c^{-1} (c + 1)^{-(k+1)/2} \left[ y^T y - \frac{c}{c+1} y^T X (X^T X)^{-1} X^T y \right]^{-n/2} 
\]
Model Selection

- Most likely models ordered by decreasing posterior probabilities using Zellner’s non-informative G-prior.

\[ t_1(\gamma) \quad \pi(\gamma | y, X) \]

| t1_gamma | pi(gamma | y, X) |
|-----------------|-----------------|
| 0 1 2 4 5 | 0.092914 |
| 0 1 2 4 5 9 | 0.032553 |
| 0 1 2 4 5 10 | 0.029512 |
| 0 1 2 4 5 7 | 0.023114 |
| 0 1 2 4 5 8 | 0.022843 |
| 0 1 2 4 5 6 | 0.022807 |
| 0 1 2 3 4 5 | 0.022409 |
| 0 1 2 3 4 5 9 | 0.016733 |
| 0 1 2 4 5 6 9 | 0.016725 |
| 0 1 2 4 5 8 9 | 0.013726 |
| 0 1 4 5 | 0.011031 |
| 0 1 2 4 5 9 10 | 0.009933 |
| 0 1 2 3 9 | 0.009698 |
| 0 1 2 9 | 0.009316 |
| 0 1 2 4 5 7 9 | 0.009253 |
| 0 1 2 6 9 | 0.009189 |
| 0 1 4 5 9 | 0.008756 |
| 0 1 2 3 4 5 10 | 0.007933 |
| 0 1 2 4 5 8 10 | 0.007901 |
| 0 1 2 4 5 7 10 | 0.007896 |

An implementation is available in **MatLab**, **C++**
When $k$ is large, it becomes computationally intractable to compute the posterior probabilities of the $2^k$ models.

Need of a tailored algorithm that samples from $\pi(\gamma|y,X)$ and selects the most likely models.

Can be done by Gibbs sampling*, given the availability of the full conditional posterior probabilities of the $\gamma_i$’s. If

$$\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k) \ (1 \leq i \leq k)$$

then

$$\pi(\gamma_i | y, \gamma_{-i}, X) \propto \pi(\gamma | y, X)$$

(to be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$)

* Gibbs and other sampling algorithms will be introduced and discussed in detail in forthcoming lectures.
Gibbs Sampling for Variable Selection

Initialization: Draw $\gamma^0$ from the uniform distribution on $\Gamma$

Iteration $t$: Given $\left( \gamma_1^{(t-1)}, ..., \gamma_k^{(t-1)} \right)$, generate

- $\gamma_1^{(t)}$ according to $\pi\left( \gamma_1 \mid y, \gamma_2^{(t-1)}, ..., \gamma_k^{(t-1)}, X \right)$
- $\gamma_2^{(t)}$ according to $\pi\left( \gamma_2 \mid y, \gamma_1^{(t)}, \gamma_3^{(t-1)}, ..., \gamma_k^{(t-1)}, X \right)$
- ..
- ..
- $\gamma_k^{(t)}$ according to $\pi\left( \gamma_k \mid y, \gamma_1^{(t)}, \gamma_2^{(t)}, ..., \gamma_{k-1}^{(t)}, X \right)$
Gibbs Sampling for Variable Selection

Question: How to sample $\gamma_1^t$ according to $\pi(\gamma_1 \mid y, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$

1. The conditional distribution $\pi(\gamma_1 \mid y, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$ is proportional to $\pi(\gamma \mid y, X)$

2. Since $\gamma_i^t$ only has two possible values which are 0 and 1, we get

$$p_0 \propto \pi(\gamma_1 = 0, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid y, X)$$

$$p_1 \propto \pi(\gamma_1 = 1, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)} \mid y, X)$$

So the probability that $\gamma_i^t = 1$ is $p_1$, then we can use Gibbs Sampling to approximate the distribution of $\{\gamma_i^t\}$
After $T \gg 1$ MCMC iterations, we approximate the posterior probabilities $p(\gamma|y, X)$ by empirical averages

$$\hat{\pi}(\gamma|y, X) = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \mathbb{I}_{\gamma(t) = \gamma}$$

The $T_0$ first values (burn in) in the MCMC chain are eliminated.
First level **Informative G-prior model** with \((\tilde{\beta} = 0_{11}, c=100)\) compared with the Gibbs estimates of the top ten posterior probabilities

\[
t_{1}(\gamma) \quad \pi(\gamma | y, X) \quad \hat{\pi}(\gamma | y, X)
\]

| t1_gamma         | pi(gamma | y, X) | pi_hat(gamma | y, X) |
|------------------|------------|-------------|
| 0 1 2 4 5        | 0.231543   | 0.239276    |
| 0 1 2 4 5 9      | 0.037358   | 0.034397    |
| 0 1 9            | 0.034435   | 0.032397    |
| 0 1 2 4 5 10     | 0.032975   | 0.030097    |
| 0 1 4 5          | 0.030606   | 0.029397    |
| 0 1 2 9          | 0.025016   | 0.025297    |
| 0 1 2 4 5 7      | 0.024144   | 0.022498    |
| 0 1 2 4 5 8      | 0.023784   | 0.024898    |
| 0 1 2 4 5 6      | 0.023735   | 0.023598    |
| 0 1 2 3 4 5      | 0.023207   | 0.022998    |

A **MatLab** implementation is available
**Model Choice Comparison: Gibbs Estimates**

Non-informative G-prior variable model choice compared with the Gibbs estimates of the top ten posterior probabilities

<table>
<thead>
<tr>
<th>$t_1(\gamma)$</th>
<th>$\pi(\gamma \mid y, X)$</th>
<th>$\hat{\pi}(\gamma \mid y, X)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>t1_gamma</td>
<td>pi(gamma</td>
<td>y, X)</td>
</tr>
<tr>
<td>0 1 2 4 5</td>
<td>0.092914</td>
<td>0.093391</td>
</tr>
<tr>
<td>0 1 2 4 5 9</td>
<td>0.032553</td>
<td>0.033097</td>
</tr>
<tr>
<td>0 1 2 4 5 10</td>
<td>0.029512</td>
<td>0.032597</td>
</tr>
<tr>
<td>0 1 2 4 5 7</td>
<td>0.023114</td>
<td>0.025097</td>
</tr>
<tr>
<td>0 1 2 4 5 8</td>
<td>0.022843</td>
<td>0.023098</td>
</tr>
<tr>
<td>0 1 2 4 5 6</td>
<td>0.022807</td>
<td>0.022498</td>
</tr>
<tr>
<td>0 1 2 3 4 5</td>
<td>0.022409</td>
<td>0.021698</td>
</tr>
<tr>
<td>0 1 2 3 4 5 9</td>
<td>0.016733</td>
<td>0.015998</td>
</tr>
<tr>
<td>0 1 2 4 5 6 9</td>
<td>0.016725</td>
<td>0.014899</td>
</tr>
<tr>
<td>0 1 2 4 5 8 9</td>
<td>0.013726</td>
<td>0.013399</td>
</tr>
</tbody>
</table>

A MatLab implementation is available
Gibbs Sampling: Probabilities of Inclusion

An approximation of the probability to include the i-th variable:

\[ \hat{P}^\pi(y_i = 1|y, X) = \frac{1}{T - T_0 + 1} \sum_{t=T_0}^{T} \mathbb{I}_{y_i^{(t)}=1} \]
Informative ($\bar{\beta} = 0_{11}$, $c=100$) and non-informative G-prior variable inclusion estimates (based on the same Gibbs output as in the earlier two tables)

| $\gamma_i$  | $\hat{P}^\pi(\gamma_i = 1|y, X)$ | $\hat{P}^\pi(\gamma_i = 1|y, X)$ |
|-------------|----------------------------------|----------------------------------|
| gamma_1     | 0.8733                           | 0.8806                           |
| gamma_2     | 0.7100                           | 0.7789                           |
| gamma_3     | 0.1515                           | 0.2958                           |
| gamma_4     | 0.6842                           | 0.7422                           |
| gamma_5     | 0.6635                           | 0.7234                           |
| gamma_6     | 0.1659                           | 0.2992                           |
| gamma_7     | 0.1343                           | 0.2812                           |
| gamma_8     | 0.1478                           | 0.2740                           |
| gamma_9     | 0.3942                           | 0.5015                           |
| gamma_10    | 0.1135                           | 0.2556                           |

A MatLab implementation is available