STAT7630: Bayesian Statistics
Lecture Slides # 11

The Bayesian Linear Regression Model

Chapters 9-11 (Simple Normal Regression, Evaluating Regression Models, & Extending the Normal Regression Model

Elvan Ceyhan
Department of Mathematics & Statistics
Auburn University

Fall 2024,

Updated: November, 2024

Outline

Setup of Linear Regression Model

- **Model Framework:** We examine a regression model where the response variable Y is modeled as a function of $k-1$ predictor variables $X_1, X_2, \ldots, X_{k-1}$.
- **Model for n Observations:** For each observation $i = 1, 2, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{i,k-1} + \varepsilon_i, \quad \varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Setup of Linear Regression Model

- **Matrix Formulation:** The linear regression model can be expressed as:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

where

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1,k-1} \\ 1 & X_{21} & \cdots & X_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,k-1} \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

Likelihood for Linear Regression Model

- **Likelihood Function:** Based on the normality assumption, the likelihood is given by:

$$L(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y}) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right)$$

- **Least Squares Estimates:** The least squares estimators for $\boldsymbol{\beta}$ and σ^2 are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-k}$$

Likelihood for Linear Regression Model

- **Likelihood Derivation:**

$$
\begin{aligned}
&L(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) \\
&\propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \right) \right\} \\
&= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \right.\right. \\
&\qquad \left.\left. -2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'\mathbf{X}'\mathbf{y} + 2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]'\mathbf{X}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] \right) \right\}
\end{aligned}
$$

- **Simplification Using $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\beta}$:**

$$
\begin{aligned}
= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{X}\hat{\beta} + \beta'\mathbf{X}'\mathbf{X}\beta \right.\right. \\
- 2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta}]'\mathbf{X}'\mathbf{X}\hat{\beta} + \\
\left.\left. 2[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta}]'\mathbf{X}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\beta}] \right) \right\}
\end{aligned}
$$

Likelihood for Linear Regression Model

- **Likelihood Derivation:**

  $$L(\beta, \sigma^2 | \mathbf{X}, \mathbf{y}) \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \right) \right\}$$

  where:

  - $\mathbf{y}'\mathbf{y}$ represents the sum of squared outcomes.
  - $-2\beta'\mathbf{X}'\mathbf{y}$ involves the interaction between data and parameters.
  - $\beta'\mathbf{X}'\mathbf{X}\beta$ is the quadratic form involving the design matrix.

- **Simplification Using the Projection Matrix $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$:**

  $$\begin{aligned} &\sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\mathbf{y}'\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{y} \right) \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \left(\text{RSS}(\mathbf{y}, \hat{\mathbf{y}}) + \text{ESS}(\mathbf{X}, \hat{\beta}) \right) \right\} \end{aligned}$$

  where **RSS** (Residual Sum of Squares): Variance unexplained by the model, and **ESS** (Explained Sum of Squares): Variance explained by the model.

Likelihood for Linear Regression Model

- **Likelihood Expression:**

$$
\begin{aligned}
& L(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y}) \\
& \propto \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \Big(\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \right. \\
& \qquad \left. - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \Big) \right\} \\
& = \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \Big((\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \right. \\
& \qquad \quad \left. - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \Big) \right\} \\
& = \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \Big(\hat{\sigma}^2 (n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \Big) \right\}
\end{aligned}
$$

Outline

Outline

Noninformative Priors for β and σ^2

- **Independent Vague Priors:**

$$p(\boldsymbol{\beta}) \propto 1, \quad \boldsymbol{\beta} \in (-\infty, \infty)^k$$

$$p(\sigma^2) = \frac{1}{\sigma}, \quad \sigma \in (0, \infty)$$

- **Joint Posterior for β and σ^2:**

$$p(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y}) \propto p(\boldsymbol{\beta}) p(\sigma^2) L(\boldsymbol{\beta}, \sigma^2 | \mathbf{X}, \mathbf{y})$$

$$\propto \sigma^{-n-1} \exp\left\{ -\frac{1}{2\sigma^2} \left[\hat{\sigma}^2 (n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \right] \right\}$$

Noninformative Priors for β and σ^2

- **Transformation:** Let $s = \sigma^{-2}$ with Jacobian $|J| = \frac{1}{2}s^{-3/2}$.
- **Joint Posterior for $\boldsymbol{\beta}$ and s:**

$$p(\boldsymbol{\beta}, s|\mathbf{X}, \mathbf{y}) \propto (s^{-1/2})^{-n-1} \exp\left\{-\frac{1}{2}s\left[\hat{\sigma}^2(n-k)\right.\right. \\ \left.\left. +(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\} \cdot \frac{1}{2}s^{-3/2}$$

- **Simplified Joint Posterior:**

$$\propto s^{\frac{n}{2}-1} \exp\left\{-\frac{1}{2}s\left[\hat{\sigma}^2(n-k) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\}$$

Noninformative Priors for β and σ^2

- **Marginal Posterior for β:** Integrate out s to obtain:

$$
\begin{aligned}
& p(\beta|\mathbf{X}, \mathbf{y}) \\
& \propto \int_0^\infty s^{\frac{n}{2}-1} \exp\left\{-\frac{1}{2}\left[\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right] s\right\} ds \\
& = \frac{\Gamma\left(\frac{n}{2}\right)}{\left(\frac{1}{2}\left[\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]\right)^{\frac{n}{2}}} \\
& \propto \left[(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\hat{\sigma}^{-2}\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right]^{-\frac{n}{2}}
\end{aligned}
$$

- $$\frac{(n-k)\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}}{n-k-2}$$

Noninformative Priors for β and σ^2

- **Marginal Posterior for σ^2:** Integrate out $\boldsymbol{\beta}$ from the joint posterior:

$$
\begin{aligned}
&p(\sigma^2|\mathbf{X},\mathbf{y}) \\
&\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\hat{\sigma}^2(n-k)\right) \\
&\quad \int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right)d\boldsymbol{\beta} \\
&\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\hat{\sigma}^2(n-k)\right)(2\pi\sigma^2)^{k/2} \\
&\propto (\sigma^2)^{-\frac{1}{2}(n-k-1)-1}\exp\left(-\frac{1}{2}\frac{\hat{\sigma}^2(n-k)}{\sigma^2}\right)
\end{aligned}
$$

- $$\sigma^2|\mathbf{X},\mathbf{y}\sim \mathrm{IG}\left(\frac{n-k-1}{2},\frac{\hat{\sigma}^2(n-k)}{2}\right)$$

- **Example:** Oxygen uptake data (available on Canvas)

Outline

Conjugate Analysis for the Linear Model

- Conjugate priors for linear regression are not actually recommended, because they are hard to elicit.
- Nonetheless, the mathematical results are elegant and hold historical and practical significance.
- Practical significance emerges in Bayesian nonparametric analysis involving Dirichlet process mixture models.
- If we have reliable prior information that can be quantified and used to specify priors for β and σ^2, then conjugate priors may be utilized.

Conjugate Analysis for the Linear Model

- **Conjugate Priors:** With strong prior knowledge, we can use conjugate priors for $\boldsymbol{\beta}$ and σ^2.
- **Prior on Error Precision τ:** Following the approach in BIDA by Christensen, Johnson, Branscum, and Hanson (2010), we specify a prior on the precision parameter $\tau = \frac{1}{\sigma^2}$:

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

  This is analogous to using an inverse-gamma prior for σ^2.
- **Prior on β (Conditional on τ):**

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

  where $\tau^{-1} = \sigma^2$.

Conjugate Analysis for the Linear Model

- **Hypothetical Observations:** Specify a set of k reasonable hypothetical observations with predictor vectors $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_k$. These, along with a column of 1's, form the rows of $\tilde{\mathbf{X}}$. Assume prior expected response values $\tilde{y}_1, \ldots, \tilde{y}_k$.
- **Prior on $\tilde{\mathbf{X}}\beta$:** The multivariate normal prior on $\boldsymbol{\beta}$ translates to a prior on $\tilde{\mathbf{X}}\boldsymbol{\beta}$:

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

- **Prior Mean and Weights:**
  - The prior mean of $\tilde{\mathbf{X}}\boldsymbol{\beta}$ is $\tilde{\mathbf{y}}$, so the prior mean δ of $\boldsymbol{\beta}$ is $\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{y}}$.
  - $\mathbf{D}^{-1}$ is a diagonal matrix with diagonal elements representing the weights of the hypothetical observations.
  - Intuitively, the prior has an equivalent "worth" of $\mathrm{tr}(\mathbf{D}^{-1})$ observations.

Conjugate Analysis for the Linear Model

- **Joint Posterior Density:**

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

- **Conditional Posterior for $\beta|\tau$:**

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

where

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

Conjugate Analysis for the Linear Model

- **Posterior for τ:**

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

where

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

- **Incorporation of Subjective Information:**
  - The estimate $\hat{\beta}$ incorporates prior knowledge through $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{y}}$.
  - s^* incorporates subjective parameters a and b, alongside $\hat{\beta}$.

Conjugate Analysis for the Linear Model

- **Marginal Posterior for $\boldsymbol{\beta}$:** The marginal posterior $p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y})$ is a (scaled) **noncentral multivariate t-distribution**, although the conditional posterior $p(\boldsymbol{\beta}|\tau, \mathbf{X}, \mathbf{y})$ is multivariate normal.
- **Inference on $\boldsymbol{\beta}$:** To simplify inference on $\boldsymbol{\beta}$, it is effective to use the conditional posterior $p(\boldsymbol{\beta}|\tau, \mathbf{X}, \mathbf{y})$.
- **Sampling Strategy:** Instead of basing inference on $p(\boldsymbol{\beta}|\hat{\tau}, \mathbf{X}, \mathbf{y}))$ by plugging in a posterior estimate of τ, it is preferable to:
  1. Sample random values $\tau^{[1]}, \ldots, \tau^{[J]}$ from the posterior of τ.
  2. For each $\tau^{[j]}$, sample from the conditional posterior $p(\boldsymbol{\beta}|\tau^{[j]}, \mathbf{X}, \mathbf{y}))$, $j = 1, \ldots, J$.
- **Estimation:** Posterior point estimates and interval estimates for $\boldsymbol{\beta}$ can then be based on these random draws.

Prior Specification for the Conjugate Analysis

- **Hypothetical Predictor Values:** Specify a matrix $\tilde{\mathbf{X}}$ containing hypothetical predictor values.
- **Response Values:** Using expert opinion or prior knowledge, specify a corresponding vector $\tilde{\mathbf{y}}$ of reasonable response values for these predictors.
- **Requirement for Hypothetical Observations:** The number of hypothetical observations must be one more than the number of predictor variables in the model.
- **Prior Mean for β:** The prior mean for β is set to $\tilde{\mathbf{X}}^{-1}\tilde{\mathbf{y}}$.

Prior Specification for the Conjugate Analysis

- **Gamma Prior on τ:** We need to specify the shape parameter a and rate parameter b for the gamma prior on τ.
- **Choosing a:**
  - Start by selecting a based on the degree of confidence in the prior information.
  - For a given a, the prior can be viewed as having the equivalent informational "worth" of $2a$ sample observations.
- **Confidence Level:** A larger value of a reflects higher confidence in the prior (although variance tends to increase too), thus weighting prior information more heavily in the analysis.

Prior Specification for the Conjugate Analysis

- **Strategy for Specifying b:**
  - Select one of the hypothetical observations, say the first one.
  - Let $\tilde{y}_1$ be the prior expected response for this observation with predictor values $\tilde{\mathbf{x}}_1$.
  - Define $\tilde{y}_{\text{max}}$ as the maximum reasonable prior response for an observation with predictors $\tilde{\mathbf{x}}_1$.
- **Prior Estimate for σ:** - Based on a normal distribution, estimate σ as:

$$\sigma \approx \frac{\tilde{y}_{\text{max}} - \tilde{y}_1}{1.645}$$

  - Since $\tau = \frac{1}{\sigma^2}$, this provides a reasonable guess for τ.
- **Solving for b:**
  - Set this guess for τ equal to the mean a/b of the gamma prior for τ.
  - With a specified, solve for b.

- Given these results, it can (of BIDA) be shown that, ④

$$\frac{c'\vec{\beta} - c'\hat{\vec{\beta}}}{\sqrt{s^* c'(X'X + \tilde{X}'D^{-1}\tilde{X})^{-1} c}} \Big| \vec{y} \sim t(n+2a)$$

where $c' = (c_1, c_2, \ldots, c_k)$, e.g. for inference on $\beta_1 - \beta_2$, $\vec{c} = (1, -1, 0, \ldots, 0)$

- For predicting a new observation y_f with covariate vector x_f, we get the predictive distribution

$$\frac{y_f - x_f'\beta}{\sqrt{s^*\left(1 + x_f'(X'X + \tilde{X}'D^{-1}\tilde{X})^{-1} x_f\right)}} \Big| y \sim t(n+2a)$$

- Much of the analysis can be obtained by making minor changes to the output from a weighted least squares regression ~~code~~ output. Using partitioned matrices, and fitting,

$$\begin{pmatrix} y \\ \tilde{y} \end{pmatrix} = \begin{pmatrix} X \\ \tilde{X} \end{pmatrix}\beta + \begin{pmatrix} e \\ \tilde{e} \end{pmatrix}, \quad \begin{pmatrix} e \\ \tilde{e} \end{pmatrix} \sim MVN\left(\underset{\uparrow \atop n+k \text{ vector}}{0}, \tau^{-1}\begin{pmatrix} I_n & 0 \\ 0 & D \end{pmatrix}\right)$$

(one need do little more than modify the reported dfe and MSE to agree with $n+2a$ & s^*, but that also involves changes to all std errors.)

Benefits of BIDA Approach - I

- **Analytical Tractability:** Conjugate priors enable analytical solutions, reducing computational burden and allowing for quick updates.
- **Incorporation of Prior Knowledge:** Embeds expert knowledge through hypothetical predictor and response values, enhancing accuracy, especially with limited data.
- **Flexible Prior Influence:** Gamma prior parameters a and b adjust confidence in prior information, balancing reliance on prior vs. data.
- **Posterior Sampling Strategy:** Efficient sampling approach that accounts for uncertainty in τ without relying on point estimates.

Benefits of BIDA Approach - II

- **Interpretability:** Provides a meaningful prior mean for $\boldsymbol{\beta}$ and allows specifying observational "weights" to clarify model assumptions.
- **Robust Inference with t-Distributions:** The marginal posterior of $\boldsymbol{\beta}$ is a noncentral multivariate t-distribution, which is robust to outliers and effective under uncertainty in τ.
- **Overall Advantage:** Ideal for balancing prior knowledge with data-driven insights in an analytically manageable framework.

Example of a Conjugate Analysis

- **Example in R:** Using the Automobile Data Set, we perform a conjugate analysis.
- **Estimates for τ and σ^2:** Obtain point and interval estimates for the precision parameter τ and, consequently, for σ^2.
- **Estimates for Elements of $\boldsymbol{\beta}$:** Draw samples from the posterior distributions of τ and then from the conditional posterior of $\boldsymbol{\beta}|\tau$ to obtain point and interval estimates for each element of $\boldsymbol{\beta}$.

Alternative Approach to Conjugate Analysis for the Linear Model

The Approach in Bayesian Methods (BaM) by Jeff Gill:

- For conjugate priors with a sampling distribution (data model):

$$\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n)$$

- Conditional distribution of $\boldsymbol{\beta}$ on σ^2 resembles the normal-normal model before:

$$p(\boldsymbol{\beta}|\sigma^2) = (2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{B})'\Sigma^{-1}(\boldsymbol{\beta} - \mathbf{B})\right)$$

- Prior for σ^2: $p(\sigma^2) \propto \sigma^{-(a-k)} \exp\left(-\frac{b}{\sigma^2}\right)$
- Joint prior as a product of conditionals:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2)$$

Conjugate Analysis for the Linear Model

Joint Posterior Derivation

- Combining the data likelihood with the prior specification yields the joint posterior:

$$
\begin{aligned}
&p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) \\
&\quad\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right)\right) \\
&\times (2\pi)^{-\frac{k}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \mathbf{B})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \mathbf{B})\right)\sigma^{-(a-k)} \exp\left(-\frac{b}{\sigma^2}\right) \\
&\propto \sigma^{-(n+a)} \exp\Bigg(-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) + \\
&\qquad\qquad 2b + (\boldsymbol{\beta} - \mathbf{B})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \mathbf{B})\Bigg)
\end{aligned}
$$

Conjugate Analysis for the Linear Model

Simplifying the Joint Posterior

- The form of the joint posterior can be simplified with a change of variables.
- Define:

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{\Sigma}^{-1} + \mathbf{X}'\mathbf{X})^{-1}(\boldsymbol{\Sigma}^{-1}\mathbf{B} + \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\tilde{s} = 2b + \hat{\sigma}^2(n-k) + (\mathbf{B} - \tilde{\boldsymbol{\beta}})'\Sigma^{-1}\mathbf{B} + (\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

- The joint posterior can now be re-expressed as:

$$p(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{y}) \propto$$
$$(\sigma^2)^{-\frac{n+a}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(\tilde{s} + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'(\boldsymbol{\Sigma}^{-1} + \mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right)\right)$$

Conjugate Analysis for the Linear Model

Posterior Distribution of $\beta|X, y$

- By applying the marginalization trick, we obtain the posterior distribution of $\boldsymbol{\beta}|\mathbf{X},\mathbf{y}$:

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) \propto \left(\tilde{s} + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})'(\boldsymbol{\Sigma}^{-1} + \mathbf{X}'\mathbf{X})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})\right)^{-\frac{n+a}{2}+1}$$

- This is the kernel of a multivariate-t distribution with $\nu = n + a - k - 2$ degrees of freedom.
- The mean and covariance of the posterior distribution for β are:

$$\mathbf{E}(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \tilde{\boldsymbol{\beta}}$$

$$\text{Cov}(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \frac{\tilde{s}(\boldsymbol{\Sigma}^{-1} + \mathbf{X}'\mathbf{X})^{-1}}{n + a - k - 3}$$

Marginal Distribution of σ^2

- The marginal distribution of σ^2 is derived similarly to the case with an uninformed prior:

$$p(\sigma^2|\mathbf{X},\mathbf{y}) \propto (\sigma^2)^{-\frac{n+a-k-1}{2}} \exp\left(-\frac{1}{2\sigma^2}\hat{\sigma}^2(n+a-k)\right)$$

- This corresponds to the kernel of an Inverse-Gamma distribution:

$$\text{IG}\left(\frac{n+a-k-2}{2}, \frac{1}{2}\hat{\sigma}^2(n+a-k)\right)$$

Conjugate Analysis for the Linear Model

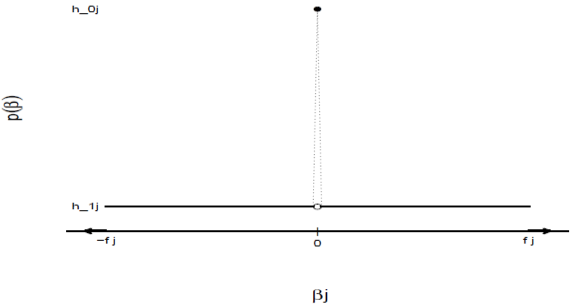
Comparison of Informative Conjugate and Noninformative Models

Setup	Prior	Posterior
Noninf.	$p(\boldsymbol{\beta}) \propto c$ on $(-\infty, \infty)$	$\boldsymbol{\beta}\vert\mathbf{X}, \mathbf{y} \sim MVt(n-k)$
	$p(\sigma^2) \propto \frac{1}{\sigma}$ on $(0, \infty)$	$\sigma^2\vert\mathbf{X}, \mathbf{y} \sim \text{IG}\left(\frac{n-k-1}{2}, \frac{\hat{\sigma}^2(n-k)}{2}\right)$
Conj.	$\boldsymbol{\beta}\vert\sigma^2 \sim MVN(\mathbf{B}, \sigma^2 I_n)$	$\boldsymbol{\beta}\vert\mathbf{X}, \mathbf{y} \sim MVt(n+a-k-2)$
	$\sigma^2 \sim \text{IG}\left(\frac{a-k-2}{2}, b\right)$	$\sigma^2\vert\mathbf{X}, \mathbf{y} \sim \text{IG}\left(\frac{n+a-k-2}{2}, \frac{\hat{\sigma}^2(n+a-k)}{2}\right)$

This table summarizes the priors and posterior distributions for both the vague and informative conjugate models in linear regression.

Spike-and-Slab Priors for Linear Models

- In regression, the priors on the regression coefficients are crucial.
- Whether or not $\beta_j = 0$ defines whether X_j is “important” in the regression.
- For any j, a useful prior for β_j is a “spike-and-slab” prior, which allows for a mixture of values concentrated around zero (spike) and a broader range (slab).



Spike-and-Slab Priors for Linear Models

- Here $P(\beta_j = 0) = h_{0j}$, which represents the prior probability that X_j is not needed in the model.
-
$$P(\beta_j \neq 0) = 1 - h_{0j} = h_{1j}(f_j - (-f_j)) = 2f_j h_{1j}$$
where $[-f_j, f_j]$ contains all "reasonable" values for β_j.
- To include X_j in the model with certainty, set $h_{0j} = 0$.
- To increase the doubt that X_j should be in the model, increase the ratio:
$$\frac{h_{0j}}{h_{1j}} = \frac{h_{0j}}{(1 - h_{0j})/2f_j} = 2f_j \frac{h_{0j}}{1 - h_{0j}}$$
- Recently, "nonparametric priors" have become popular, often involving a mixture of Dirichlet processes.

Outline

Bayesian Regression with `rstanarm`

- - The `rstanarm` package in R enables Bayesian regression modeling by simulating parameter values from their posterior distributions.
  - This approach circumvents the need to derive the posterior distribution explicitly.
- For normal regression models, we can derive the posterior analytically as shown in our approach.
- - For models with non-normal responses, conjugate priors for regression coefficients may not exist.
  - Simulating from the posterior is often the only viable method for estimation.
- `rstanarm` leverages `rstan` to estimate several standard Bayesian regression models efficiently.

Parts of the `stan_glm` Function Call

- **Overview of `stan_glm`:** The `stan_glm` function in the `rstanarm` package performs Bayesian regression model estimation via simulation.
- **Specifying the Model Type:** For normal responses, specify `method = "gaussian"` in the `stan_glm` call.
- **Priors on Model Parameters:** - Set hyperparameters for priors, typically normal priors on the intercept β_0 and coefficients $\beta_1, \beta_2, \ldots$.
  - An exponential prior is often recommended for the unknown standard deviation σ of the response.
- **MCMC Specifications:** Configure MCMC details, including the number of iterations and the number of chains, to ensure adequate diagnostic assessment.

Output of the `stan_glm` Function

- **MCMC Diagnostics:** Functions in `rstanarm` provide diagnostic plots, including trace plots, autocorrelation plots, and density plots, to assess MCMC convergence.
- **Summarizing Posterior Estimates:** The `tidy` function displays a summary of the Bayesian posterior estimates for the regression coefficients.
- **Prediction and Intervals:** - `posterior_predict` provides point predictions for the response, given specific predictor values.

  - `posterior_interval` generates posterior prediction intervals for the response.
- **Posterior Predictive Density:** Plot the density function of the posterior predictive distribution to visualize the model's predictive spread.
- **Example:** See the R example using the "cars" dataset.

Outline

A Bayesian Approach to Model Selection

- **Model Selection in Regression:** In exploratory regression, selecting the optimal subset of predictor variables is essential for identifying the "best model."
- **Bayesian Comparison of Models:** A Bayesian approach involves evaluating candidate models based on their posterior probabilities.
- **Inclusion of Predictor Variables:**
  - If the coefficient $\beta_j = 0$, the variable X_j is unnecessary in the model.
  - Define $\beta_j = z_j b_j$ for each j, where $z_j = 0$ or 1 and $b_j \in (-\infty, \infty)$.
- **Model Specification:**

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

where any $z_j = 0$ indicates that the corresponding predictor variable is excluded from the model.

A Bayesian Approach to Model Selection: Example

- **Oxygen Uptake Example:** Consider predictor variables $X_1 = \text{group}$, $X_2 = \text{age}$, and $X_3 = \text{group} \times \text{age}$.
- **Indicator Vector for Model Inclusion:** Define $\mathbf{z} = (z_0, z_1, z_2, z_3)$ to specify the inclusion of each variable in the model. The true conditional expectation $\mathbf{E}[Y|\mathbf{x}, \mathbf{b}, \mathbf{z}]$ for each configuration of $\mathbf{z}$ is:

$\mathbf{z}$	**True** $\mathbf{E}[Y\|\mathbf{x}, \mathbf{b}, \mathbf{z}]$
$(1, 0, 0, 0)$	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

- **Calculating Posterior Probabilities:**
  - For each possible configuration of the vector $\mathbf{z}$, calculate the posterior probability for that model.
  - For a specific configuration $\mathbf{z}^*$:

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

- **Model Priors:**
  - A prior $p(\cdot)$ is assigned to each potential model.
  - For a noninformative approach, assign equal prior probabilities across all models.
- **Handling Many Predictors:** With a large number of predictors, employ Gibbs sampling to efficiently search across the model space.

Example of Bayesian Model Selection

- **Example in R:** Analyze the Oxygen Data Set to perform Bayesian model selection.
- **Exploring Subsets of Predictors:**
  - Consider all possible subsets of the predictor variables.
  - **Result:** The model excluding the interaction term has the highest posterior probability.
- **Restricted Subset Consideration:**
  - Restrict to certain subsets, such as only including the interaction term when both first-order terms are present.
  - **Result:** The model without the interaction term again shows the highest posterior probability, with an even greater margin.

A Bayesian Approach to Model Selection - Technical Aside for the Details

- Bayesian model selection uses posterior probabilities to evaluate model configurations.
- Here, we assess the likelihood of observing the data $\mathbf{y}$ given the design matrix $\mathbf{X}$ for various subsets of predictors.
- Each configuration of predictors, represented by $\mathbf{z}$, is treated as a potential model with a specific probability.

Model Prior Specification

- **Prior Probability on Models $p(\mathbf{z})$:**
  - A prior $p(\mathbf{z})$ is assigned to each subset configuration $\mathbf{z}$, which indicates which predictors are included in the model.
  - In a noninformative setting, each model can have equal prior probability, i.e., $p(\mathbf{z}) = \frac{1}{M}$ for M possible models.
- **Parameter Priors:**
  - Hyperparameters are used: g controls variance scaling, and ν_0 influences prior degrees of freedom.
  - s_0^2 represents a prior guess for the residual variance, often calculated from an initial ordinary least squares (OLS) model.

Model Prior Specification: Prior on Configurations $p(\mathbf{z})$

- Each subset configuration $\mathbf{z}$ represents a unique model by specifying which predictors are included.
- **Noninformative Prior:** Assign equal probability to each model configuration:
$$p(\mathbf{z}) = \frac{1}{M}$$
where M is the total number of possible configurations.
- **Informative Prior:** If we have prior knowledge or prefer simpler models, we can assign higher probabilities to specific configurations (e.g., those with fewer predictors).

Model Prior Specification: Prior on Coefficients β

- For each model configuration $\mathbf{z}$, we define a multivariate normal prior on the coefficients $\boldsymbol{\beta}$ of the included predictors:

$$\boldsymbol{\beta} \sim MVN\left(\tilde{\boldsymbol{\beta}}, \sigma^2 (g\mathbf{X}'\mathbf{X})^{-1}\right)$$

- **Components of the Prior:**
  - $\tilde{\boldsymbol{\beta}}$: Prior mean, often calculated as $\tilde{\boldsymbol{\beta}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$, based on OLS solution of hypothetical predictor values $\tilde{\mathbf{X}}$ and responses $\tilde{\mathbf{y}}$.
  - σ^2: Residual variance, representing uncertainty in predictions.
  - g: Variance scaling factor; larger g reduces the influence of the prior mean.

Model Prior Specification: Prior on Variance σ^2

- The residual variance σ^2 has a conjugate gamma prior on its precision $\tau = \sigma^{-2}$:

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

- **Parameters of the Gamma Prior:**
  - Shape a and rate b parameters are selected to reflect prior beliefs on variance.
  - **Mean:** $\mathbf{E}[\tau] = \frac{a}{b}$ **Variance:** $\text{Var}(\tau) = \frac{a}{b^2}$
- A typical choice for a and b is:

$$a = \frac{\nu_0}{2}, \quad b = \frac{\nu_0 s_0^2}{2}$$

where ν_0 is the prior degrees of freedom and s_0^2 is a prior estimate of residual variance.

Data Model and Likelihood

Likelihood of Observing y Given X:

- The goal of Bayesian model selection is to calculate the probability of observing the data $\mathbf{y}$ given a particular model configuration, represented by a subset of predictors in $\mathbf{X}$.
- For each subset model, the function `log_Py_x` in R code calculates the marginal log-likelihood $\log p(\mathbf{y}|\mathbf{X}, \mathbf{z})$, which measures the fit of the data under that model.
- This marginal likelihood incorporates a projection of $\mathbf{y}$ onto the predictor space defined by $\mathbf{X}$, which is captured by the "hat matrix" Hg, which is defined as:

$$Hg = \frac{g}{g+1}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

- This matrix projects $\mathbf{y}$ onto the subspace spanned by $\mathbf{X}$ and scales it by g, a hyperparameter that controls the variance scaling.

Likelihood of Observing y Given X:

- The fit of the model is evaluated by calculating the sum of squared residuals $SSRg$, which measures the unexplained variation in $\mathbf{y}$ after projection:

$$SSRg = \mathbf{y}'(I - Hg)\mathbf{y}$$

- Here, $I - Hg$ is a matrix that projects $\mathbf{y}$ onto the orthogonal complement of the space spanned by $\mathbf{X}$, capturing the residuals that are not explained by the model.

Data Model and Likelihood

Marginal Likelihood Computation:

- The marginal likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{z})$ is a key element in Bayesian model selection, as it indicates the probability of observing $\mathbf{y}$ for a given model configuration $\mathbf{z}$.
- This likelihood combines the residual sum of squares $SSRg$ with prior parameters:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{z}) = -\frac{1}{2} \times \Bigg(n \log(\pi) + p \log(1 + g) + \\ (\nu_0 + n) \log(\nu_0 s_0^2 + SSRg) - \nu_0 \log(\nu_0 s_0^2) \Bigg)$$

Data Model and Likelihood

In the above expression:

- $n \log(\pi)$: A normalization term, adjusting for the dimensionality of $\mathbf{y}$.
- $p \log(1+g)$: Adjusts for the number of predictors p included in the model, scaled by g, impacting how model complexity is penalized.
- $(\nu_0 + n) \log(\nu_0 s_0^2 + SSRg)$: Combines the prior information (through ν_0 and s_0^2) with the residual variance $SSRg$.
- $-\nu_0 \log(\nu_0 s_0^2)$: A prior adjustment term, providing a reference for the variance under the prior alone.

- The value of $\log p(\mathbf{y}|\mathbf{X}, \mathbf{z})$ provides a measure of how well each subset model explains the data, balancing model fit and complexity.
- Models with higher marginal likelihood values are considered better explanations of the data.

- **Posterior Probability for Model z:**
  - Given prior $p(\mathbf{z})$ and marginal likelihood $p(\mathbf{y}|\mathbf{X}, \mathbf{z})$, the posterior for model $\mathbf{z}^*$ is:

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

  - The numerator captures the joint probability of $\mathbf{z}^*$ and data given $\mathbf{z}^*$, while the denominator sums this over all model configurations.

Gibbs Sampling for Model Space Exploration

- **Purpose:** With a large number of predictors, direct computation of posteriors for all subsets is computationally expensive.
- **Sampling Approach:** Gibbs sampling iteratively samples predictor inclusion/exclusion, toggling each predictor in/out of the model.
- **Sampling Probability:** For each predictor:
  - Calculate posterior difference for inclusion vs. exclusion.
  - Accept inclusion/exclusion based on a probability proportional to the calculated difference.

Model Comparison and Bayes Factors

- **Bayes Factors:** For comparing two models $\mathbf{z}_1$ and $\mathbf{z}_2$:
  Compute the ratio of marginal likelihoods (likelihood of data under each model):

$$BF_{12} = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_1)}{p(\mathbf{y}|\mathbf{X}, \mathbf{z}_2)}$$

- **Interpretation:**
  - $BF > 1$ suggests model $\mathbf{z}_1$ is more supported by the data than $\mathbf{z}_2$.
  - Posterior probabilities also incorporate these Bayes factors, favoring models with higher likelihood.

Posterior Summary for Model Selection

- The final output ranks model configurations by posterior probability.
- Constraints can be applied (e.g., include interaction terms only when main effects are present).
- Gibbs sampling results are used to estimate probabilities for each model, selecting the model with the highest posterior.

Outline

Outline

The Posterior Predictive Distribution of the Data

- **Bayesian Model Setup:** We have built a Bayesian regression model using response data $\mathbf{y}$ and explanatory data matrix $\mathbf{X}$.
- **Future Observations:**
  - Consider future observations with explanatory variable values in matrix $\mathbf{X}^*$.
  - **The question:** What is the marginal distribution of the corresponding future response values $\mathbf{Y}^*$?
- **Posterior Predictive Distribution:**
  The distribution $p(\mathbf{y}^*|\mathbf{y}, \mathbf{X}^*, \mathbf{X})$ represents the posterior predictive distribution of $\mathbf{y}^*$.
- **Application:** This distribution serves as a tool for assessing the fit of our regression model, allowing for model validation with future data.

The Posterior Predictive Distribution of the Data

- **Joint Posterior Distribution:** With noninformative priors, the joint distribution is:

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

- **Posterior Predictive Distribution:** Integrating out $\boldsymbol{\beta}$ and σ^2, the posterior predictive distribution of $\mathbf{Y}^*$ is multivariate-t with $(n-k)$ degrees of freedom:

$$\mathbf{E}(\mathbf{Y}^* | \mathbf{y}, \mathbf{X}^*, \mathbf{X}) = \mathbf{X}^* \hat{\boldsymbol{\beta}}$$

$$\text{Cov}(\mathbf{Y}^* | \mathbf{y}, \mathbf{X}^*, \mathbf{X}) = \frac{(n-k)\hat{\sigma}^2}{n-k-2} \left(\mathbf{I} + \mathbf{X}^* (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}^{*\prime} \right)$$

- **Intuition:**
  - Given the model, our original data are multivariate normal.
  - Future predictions follow a multivariate-t distribution, which accounts for additional uncertainty about the model.

Posterior Prediction of Response Values in Regression

Ex 3: Posterior Predictive Distribution in Regression:

- **Model Fit Check:**
  - Generate samples from the posterior predictive distribution, using $\mathbf{X}^* = \mathbf{X}$ (the observed sample predictors).
  - Plot the predicted values against the actual y-values from the original sample.
- **Identifying Outliers:**
  - If an observed y_i lies far from the center of the posterior predictive distribution, then this i-th observation may be an outlier.
  - A high number of outliers would indicate a potential misfit of the model.
- See `R` example with a small automobile dataset.

Posterior Prediction Intervals in Regression

- **Prediction for New Responses:**
  Make predictions and construct "prediction intervals" for new responses given specified predictor values.
- **Example Setup:**
  - For a new observation with predictor values $\mathbf{x}^* = (1, x_1^*, x_2^*, \ldots, x_{k-1}^*)$.
  - Alternatively, predictor values for multiple new observations can be stored in matrix $\mathbf{X}^*$.
- **Posterior Predictive Distribution:**
  - Generate the posterior predictive distribution using $\mathbf{X}^*$.
  - Use the posterior median for point predictions and posterior quantiles to create prediction intervals.
- See `R` example for implementation.

Posterior Prediction Using `bayesrules` Package

- **Overview of `bayesrules` Package:**
  The `bayesrules` package provides useful functions for posterior predictions and diagnostics for models fitted with `stan_glm`.
- `ppc_intervals` **Function:**
  The `ppc_intervals` function generates prediction intervals for observations in the sample or for hypothetical future observations.
- **Model Fit Assessment:**
  - For 95% prediction intervals on sample observations, model fit can be checked by counting how many observed y-values fall within their 95% prediction intervals.
  - Ideally, around 95% of the sample y-values should lie within their respective intervals.

Outline

Measures of Predictive Accuracy

- **Prediction Summary Function:**
  Provides several numerical measures to assess predictive accuracy.
- **Key Measures:**
  - **Median Absolute Error (MAE):** Reflects the typical difference between observed responses and their posterior predictive means.
  - **Scaled Median Absolute Error:** Indicates the typical number of standard deviations by which observed responses deviate from their posterior predictive means.
  - **Within 50 Statistic:** Proportion of observed responses that lie within their 50% posterior prediction interval.
  - **Within 95 Statistic:** Proportion of observed responses that lie within their 95% posterior prediction interval.

Concerns with Measures of Predictive Accuracy

- **Sample-Based Prediction Accuracy:**
  These measures evaluate how accurately the model predicts observations within the sample (i.e., those used for model fitting).
- **Potential Overestimation:**
  Predictive accuracy measures based on sample data may overstate the model's performance for predicting response values of new, out-of-sample observations.

Measures of Out-of-Sample Predictive Accuracy

- **Cross-Validation for Out-of-Sample Prediction:** To evaluate predictive accuracy on out-of-sample data, we use cross-validation.
- **Cross-Validation Process:** - Split the data into subsets. - Use a portion of these subsets as "training" data to fit the model (estimate parameters).
  - The remaining data are "test" data, held out to assess the model's predictive performance.
- **Predictive Accuracy Assessment:** - Using the fitted model, predict the response values for the "test" data.
  - Since true response values of held-out observations are known, we can directly compare predictions to actual values.
- **Evaluating Models:** - Compute cross-validation metrics, such as MAE and scaled MAE, for each model.
  - Select the model with a lower cross-validation MAE to ensure robust out-of-sample performance.

Expected Log Predictive Density (ELPD)

- **ELPD for Model Comparison:** The expected log-predictive density (ELPD) is a tool for comparing Bayesian regression models based on predictive performance.
- **Interpretation of Posterior Predictive Density:** A high posterior predictive density value at Y_{new} indicates that the new data point y_{new} aligns well with the model.
- **Definition of ELPD:** The ELPD is defined as $\mathbf{E}(\log f(Y_{\text{new}}|\mathbf{X}, \mathbf{y}))$, the log posterior predictive density at Y_{new}, averaged over all possible values of Y_{new}.
- **Model Selection:** - A model with a *higher ELPD indicates better posterior predictive accuracy* for new data points.
  - The Bayesian Information Criterion (BIC) is another common tool for model selection, related to Bayes Factors (see Chapter 8 notes).