STAT 535: Chapter 9:
The Bayesian Linear Regression Model

David B. Hitchcock
E-Mail: hitchcock@stat.sc.edu

Spring 2022

Setup of Linear Regression Model

- We now consider the **regression model** in which a response variable Y is related to one or more **explanatory** or **predictor** variables $X_1, X_2, \ldots, X_{k-1}$.
- For a random sample of n individuals, our model is

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

Setup of Linear Regression Model

- This model can be written in matrix form as

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

where

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

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

Likelihood for Linear Regression Model

- Based on this normal model, the likelihood is:

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

- Note that the **least squares** estimates of $\boldsymbol{\beta}$ and σ^2 are:

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

Likelihood for Linear Regression Model

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

$$\propto \sigma^{-n} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})\}$$

$$= \sigma^{-n} \exp\Big\{-\frac{1}{2\sigma^2}\Big(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

$$- 2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]'\boldsymbol{X}'\boldsymbol{y} + 2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]'\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]\Big)\Big\}$$

Since $\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}$,

$$= \sigma^{-n} \exp\Big\{-\frac{1}{2\sigma^2}\Big(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

$$- 2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}]'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}$$

$$+ 2[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}]'\boldsymbol{X}'\boldsymbol{X}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}]\Big)\Big\}$$

Likelihood for Linear Regression Model

$$= \sigma^{-n} \exp\Big\{ -\tfrac{1}{2\sigma^2} \Big(\boldsymbol{y}'\boldsymbol{y} - 2\hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{y} + \hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} + 2\hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}$$

$$- \hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} - 2\hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} + 2\hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} - 2\beta'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta \Big) \Big\}$$

$$= \sigma^{-n} \exp\Big\{ -\tfrac{1}{2\sigma^2} [(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}}) + \hat{\boldsymbol{b}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}}$$

$$- 2\beta'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{b}} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta] \Big\}$$

$$= \sigma^{-n} \exp\Big\{ -\tfrac{1}{2\sigma^2} [\hat{\sigma}^2(n-k) + (\beta - \hat{\boldsymbol{b}})'\boldsymbol{X}'\boldsymbol{X}(\beta - \hat{\boldsymbol{b}})] \Big\}$$

Noninformative Priors for β and σ^2

Consider the independent vague priors

$$p(\boldsymbol{\beta}) \propto 1, \quad \boldsymbol{\beta} \in (-\infty, \infty)^k$$
$$\text{and } p(\sigma^2) = \frac{1}{\sigma}, \quad \sigma \in (0, \infty)$$

Then the joint posterior for $\boldsymbol{\beta}$ and σ^2 is:

$$p(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \boldsymbol{y}) \propto L(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{X}, \boldsymbol{y}) p(\boldsymbol{\beta}) p(\sigma^2)$$
$$\propto \sigma^{-n-1} \exp\{-\tfrac{1}{2\sigma^2}[\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{b}})' \boldsymbol{X}' \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{b}})]\}$$

Noninformative Priors for β and σ^2

- Using the transformation $s = \sigma^{-2}$ with Jacobian $|J| = \frac{1}{2}s^{-3/2}$:

$$
\begin{aligned}
p(\boldsymbol{\beta}, s|\boldsymbol{X}, \boldsymbol{y}) &\propto (s^{-1/2})^{-n-1} \exp\Big\{-\tfrac{1}{2}s[\hat{\sigma}^2(n-k) \\
&\qquad + (\boldsymbol{\beta} - \hat{\boldsymbol{b}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{b}})]\Big\}\Big(\tfrac{1}{2}s^{-3/2}\Big) \\
&\propto (s)^{\frac{n}{2}-1} \exp\{-\tfrac{1}{2}s[\hat{\sigma}^2(n-k) + (\boldsymbol{\beta} - \hat{\boldsymbol{b}})'\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{b}})]\}
\end{aligned}
$$

Noninformative Priors for β and σ^2

- To get the marginal posterior for β, integrate out s:

So $p(\beta|\boldsymbol{X}, \boldsymbol{y})$

$$= \int_0^\infty (s)^{\frac{n}{2}-1} \exp\{-\tfrac{1}{2}[\hat{\sigma}^2(n-k) + (\beta - \hat{\boldsymbol{b}})^{'}\boldsymbol{X}^{'}\boldsymbol{X}(\beta - \hat{\boldsymbol{b}})]s\}\,\mathrm{d}s$$

$$= \frac{\Gamma(\frac{n}{2})}{\frac{1}{2}[\hat{\sigma}^2(n-k) + (\beta - \hat{\boldsymbol{b}})^{'}\boldsymbol{X}^{'}\boldsymbol{X}(\beta - \hat{\boldsymbol{b}})]^{\frac{n}{2}}}$$

$$\propto [(n-k) + (\beta - \hat{\boldsymbol{b}})^{'}\hat{\sigma}^{-2}\boldsymbol{X}^{'}\boldsymbol{X}(\beta - \hat{\boldsymbol{b}})]^{-\frac{n}{2}}$$

- This is the kernel of a multivariate t-distribution with $(n-k)$ degrees of freedom and covariance matrix

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

Noninformative Priors for β and σ^2

- Now we integrate β out of the joint posterior to get the marginal posterior for σ^2:

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

which is clearly an $\text{IG}(\frac{1}{2}(n-k-1), \frac{1}{2}\hat{\sigma}^2(n-k))$ posterior distribution.

- Example: Oxygen update data on course web page

Conjugate Analysis for the Linear Model

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

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

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

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

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

Conjugate Analysis for the Linear Model

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

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

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

Conjugate Analysis for the Linear Model

- The joint density is

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

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

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

where

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

Conjugate Analysis for the Linear Model

- And the posterior for τ is:

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

where

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

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

Conjugate Analysis for the Linear Model

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

Prior Specification for the Conjugate Analysis

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

Prior Specification for the Conjugate Analysis

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

Prior Specification for the Conjugate Analysis

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

Example of a Conjugate Analysis

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

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

Bayesian Regression with rstanarm

- The R package `rstanarm` allows for estimation of Bayesian regression model via simulation of parameter values from their posterior.
- This approach allows us to avoid having to derive the posterior explicitly.
- For the normal regression model, we already derived the posterior with our approach.
- But for regression models with non-normal responses, conjugate priors for the regression coefficients will not exist. So simulating from their posterior distributions is the only workable approach.
- The `rstanarm` package uses `rstan` behind the scenes to estimate several common Bayesian regression models.

Parts of the `stan_glm` function call

- The R function `stan_glm` in the `rstanarm` package estimates any of several Bayesian regression models via simulation.
- For a model for a normal response, we specify `method="gaussian"` in the call of the `stan_glm` function.
- We can also provide the hyperparameters of (typically) normal priors on the intercept β_0 and the model coefficients $\beta_1, \beta_2, \ldots$.
- We can put another prior on the unknown standard deviation σ of the response (the book suggests using an exponential prior for σ).
- Finally, we specify the details of the MCMC like the number of iterations, and the number of chains generated (for diagnostic purposes).

Output of the stan_glm function

- Various MCMC diagnostic functions in the rstanarm package give trace plots, autocorrelation function plots, density plots, etc., to gauge convergence of the MCMC algorithm.
- The tidy function presents a summary of the Bayesian posterior estimation of the regression coefficients.
- The posterior_predict function and the posterior_interval function give a point prediction of the response value and a posterior prediction interval of the response value, given a set of specified predictor value(s).
- We can also plot the density function of the posterior predictive model.
- See R example on the "cars" data set.

A Bayesian Approach to Model Selection

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

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

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

Example: Oxygen uptake example:

X_1 = group, X_2 = age, X_3 = group $\times$ age:

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

A Bayesian Approach to Model Selection

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

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

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

Example of Bayesian Model Selection

- Example in R with Oxygen Data Set
- We can consider all possible subsets of set of predictor variables:
- Result: The model with the interaction omitted has the highest posterior probability.
- We can consider only certain subsets (here, we only consider including the interaction term when both first-order terms appear):
- Result: Again, the model with the interaction omitted has the highest posterior probability (by a greater margin).

The Posterior Predictive Distribution of the Data

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

$$p(\boldsymbol{y}^* | \boldsymbol{y}, \boldsymbol{X}^*, \boldsymbol{X}).$$

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

The Posterior Predictive Distribution of the Data

- In our analysis with the noninformative priors, note that

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

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

$$E(\boldsymbol{y}^*) = \boldsymbol{X}^* \hat{\boldsymbol{\beta}} \text{ and}$$

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

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

Posterior Prediction of Response Values in Regression

Example 3: In the regression setting, we have shown that the posterior predictive distribution for a new response vector $\boldsymbol{y}^*$ is multivariate-t.

- To check model fit, we can generate samples from the posterior predictive distribution (letting $\boldsymbol{X}^* =$ the observed sample $\boldsymbol{X}$) and plot the values against the y-values from the original sample.
- If an observed y_i falls far from the center of the posterior predictive distribution, this i-th observation is an outlier.
- If this occurs for many y-values, we would doubt the adequacy of the model.
- See `R` example (small automobile data set).

Posterior Prediction Intervals in Regression

- We can also make predictions and "prediction intervals" for new responses with specified predictor values.
- For example, consider a new observation with predictor variable values in the vector $\boldsymbol{x}^* = (1, x_1^*, x_2^*, \ldots, x_{k-1}^*)$ (or the predictor values for several new observations could be contained in the matrix $\boldsymbol{X}^*$).
- We can generate the posterior predictive distribution with $\boldsymbol{X}^*$ and compute the posterior median (for a point prediction) or posterior quantiles (for a prediction interval).
- See R example.

Posterior Prediction Using bayesrules Package

- The `bayesrules` package has some nice functions to do posterior predictions and diagnostics for models fit using the `stan_glm` function.
- The `ppc_intervals` function gives prediction intervals corresponding to the observations in the sample (or to hypothetical future observations).
- If we do 95% prediction intervals for observations in the sample, we could assess model fit by checking how many observed y values in the sample fall within their corresponding 95% prediction interval (hopefully around 95% of them do).

Measures of Predictive Accuracy

- The `prediction_summary` function gives several numerical measures of predictive accuracy.
- **median absolute error (MAE)**: measures the typical difference between the observed responses and their posterior predictive means
- **scaled median absolute error**: measures the typical number of std deviations that the observed responses fall from their posterior predictive mean
- **within_50 statistic**: measures the proportion of observed response values that fall within their 50% posterior prediction interval.
- **within_95 statistic**: measures the proportion of observed response values that fall within their 95% posterior prediction interval.

Concerns with Measures of Predictive Accuracy

- However, these are measures of how well the model predicts observations that are within the sample (the observations that were used to fit the model).
- These measures may overstate how well the model would predict the response value of an observation that is **outside the sample**.

Measures of Out-of-Sample Predictive Accuracy

- To assess the prediction of out-of-sample data, we use an approach called **cross-validation**.
- We split the data into subsets, and we use some of the subsets to "train" the model (i.e., estimate the parameters).
- Then we call the held-out observations the "test" data and we use the fitted model to predict the response values of the "test" observations.
- Since we know the actual response values of the held-out observations, we can compare the predicted values to the actual values to assess the predictive accuracy.
- The cross-validation MAE, scaled MAE, etc., can be calculated for a set of models under consideration, and we might choose the model that has a low cross-validation MAE.

Expected Log Predictive Density (ELPD)

- Another tool to compare Bayesian regression models is the expected log-predictive density (ELPD).
- If the value of the posterior predictive density at y_{new} is large, this means that the new data value y_{new} is compatible with the predictive model for the responses.
- The ELPD is $E(\log f(y_{new}|\boldsymbol{y}))$, the value of the log posterior predictive density at y_{new}, averaged across all possible values of y_{new}.
- A model with a higher ELPD has greater posterior predictive accuracy when using the model to predict new data points.
- *BIC* is another very common tool for model selection (review the end of the Chapter 8 notes to see the relationship between the *BIC* and Bayes Factors).