Bayesian Linear Model

Lin Zhang

Department of Biostatistics
School of Public Health
University of Minnesota

Linear Regression Model

- Suppose we have n independent observations of response $\mathbf{y} = (y_1, \ldots, y_n)$ and an $n \times p$ design matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$ (X is assumed to have been observed without error). The linear regression model relates the response $\mathbf{y}$ to the predictors X as
$$y_i = X_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$
- Or, equivalently
$$y_i \overset{ind}{\sim} N(X_i\boldsymbol{\beta}, \sigma^2)$$
Normal likelihood with different mean $\mu_i = X_i\boldsymbol{\beta}$ and common variance σ^2.
- Thus, we have joint distribution of $\mathbf{y}$:
$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \sigma^2 I_n),$$
with parameter set $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$. I_n: $n-$dimensional identity matrix.

Linear Regression Model

- The linear regression model is the most fundamental of all serious statistical models, encompassing ANOVA, regression, ANCOVA, random and mixed effect modelling, etc.

- Major problems of interest in regression:
  - Estimate the association between two variables while adjusting for other confounders
  - Select the set of variables that are associated with the an outcome.
  - Use one or more variables to predict an outcome.

Outline

- Frequentist Estimation
- Bayesian with Noninformative Prior
- Bayesian with Conjugate NIG Prior

Freqentist Estimation

- Recall from standard statistical analysis, the classical unbiased estimates of the parameters are

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T \mathbf{y}; \\ \hat{\sigma}^2 &= \frac{1}{n-p}(\mathbf{y} - X^T\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X^T\hat{\boldsymbol{\beta}}).\end{aligned}$$

  - $\hat{\boldsymbol{\beta}}$ is also the ordinary least square estimate of $\boldsymbol{\beta}$.
  - $\hat{\sigma}^2$ is just the sample variance s^2.

- The predicted value of $\mathbf{y}$ is given by

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = P_X\, \mathbf{y},$$

  $P_X = X(X^T X)^{-1} X^T$ is called the projector matrix of X. It is an operator that projects any vector to the space spanned by the columns of X.

Bayesian with Noninformative priors

- For the Bayesian analysis, we will need to specify priors for the unknown regression parameters $\boldsymbol{\beta}$ and variance σ^2.

- We consider the improper prior:

$$\pi(\boldsymbol{\beta}) = 1; \pi(\sigma^2) \propto \frac{1}{\sigma^2}, \quad \text{or equivalently } \pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}$$

- We thus have the hierarchical model

$$\begin{aligned} \mathbf{y} &\sim N(X\boldsymbol{\beta}, \sigma^2 I_n) \\ \pi(\boldsymbol{\beta}, \sigma^2) &\propto \frac{1}{\sigma^2} \end{aligned}$$

- The joint posterior of $(\boldsymbol{\beta}, \sigma^2)$ is

$$\begin{aligned} p(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) &\propto N(X\boldsymbol{\beta}, \sigma^2 I_n) \times \frac{1}{\sigma^2} \\ &\propto (\sigma^2)^{-\frac{n}{2}-1} \exp\left\{ -\frac{1}{2\sigma^2} (\mathbf{y} - X\boldsymbol{\beta})^T (\mathbf{y} - X\boldsymbol{\beta}) \right\} \end{aligned}$$

Posterior distributions

- The conditional posterior distribution of $\boldsymbol{\beta}$ (for a given σ^2) is

$$\begin{aligned} p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) &\propto f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta}|\sigma^2) \\ &\propto N(\mathbf{y}|X\boldsymbol{\beta}, \sigma^2 I_n) \times 1 \\ &\propto N\Big((X^T X)^{-1} X^T \mathbf{y},\ \sigma^2 (X^T X)^{-1}\Big). \end{aligned}$$

  That is, $p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) = N(\hat{\boldsymbol{\beta}}, \sigma^2 (X^T X)^{-1})$.

- Analogous to frequentist estimation given σ^2 is known.

- When σ^2 is unknown, we need the marginal posterior distribution of $\boldsymbol{\beta}$ for posterior inference of $\boldsymbol{\beta}$

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

  which requests the marginal posterior of σ^2.

Posterior distributions (cont'd)

- The marginal posterior of σ^2 can be derived by integrating the joint posterior over $\boldsymbol{\beta}$ space

$$
\begin{aligned}
p(\sigma^2|\mathbf{y}) &= \int p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\boldsymbol{\beta} \\
&\propto \int (\sigma^2)^{-\frac{n}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right\} d\boldsymbol{\beta} \\
&= (\sigma^2)^{-\frac{n}{2}-1}\sqrt{2\pi}(\sigma^2)^{\frac{p}{2}} \exp\left\{-\frac{(n-p)s^2}{2\sigma^2}\right\} \\
&\propto IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)
\end{aligned}
$$

Therefore, $p(\sigma^2|\mathbf{y}) = \mathrm{IG}(\ (n-p)/2,\ (n-p)s^2/2\ )$.

- Note: This result parallels the classical inference:
  $\Rightarrow$ $(n-p)s^2/\sigma^2$ follows a chi-square distribution.

Posterior distributions (cont'd)

- The marginal posterior of $\boldsymbol{\beta}$ is then obtained as

$$\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y}) &= \int p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) p(\sigma^2|\mathbf{y}) d\sigma^2 \\
&= \int N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \sigma^2 (X^T X)^{-1}) IG\left(\sigma^2 | \frac{n-p}{2}, \frac{(n-p)s^2}{2}\right) d\sigma^2 \\
&\propto \int (\sigma^2)^{-\frac{n}{2}-1} \exp\left\{ -\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (X^T X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + (n-p)s^2}{2\sigma^2} \right\} d\sigma^2 \\
&= \Gamma(n/2) \left[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (X^T X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/2 + (n-p)s^2/2 \right]^{-n/2} \\
&\propto \left[1 + \frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T (X^T X)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{(n-p)s^2} \right]^{-n/2}
\end{aligned}$$

- This is a multivariate student-t density:

$$MVSt_{\nu}(\boldsymbol{\mu}, \Sigma) = \frac{\Gamma[(\nu+p)/2]}{(\pi\nu)^{p/2}\Gamma(\nu/2)|\Sigma|^{1/2}} \left[1 + \frac{(\boldsymbol{\beta}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{\beta}-\boldsymbol{\mu})}{\nu} \right]^{-(\nu+p)/2}$$

with $\nu = n - p$, $\boldsymbol{\mu} = \hat{\boldsymbol{\beta}}$, $\Sigma = s^2 (X^T X)^{-1}$.

Implication

- We have derived that the marginal posterior of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\mathbf{y}) = MVSt_{n-p}(\hat{\boldsymbol{\beta}}, s^2(X^TX)^{-1})$$

- By properties of MVSt distribution, the marginal distribution of each individual regression parameter β_j is a univariate student-t with the same degree-of-freedom, i.e.

$$\frac{\beta_j - \hat{\beta}_j}{s\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}.$$

- With the noninformative prior, the inference results of the Bayesian method are thesame as the frequentist regression.

Sampling-based approximation

- Again, we can use a simpler sampling based mechanism to approximate the posterior distribution.
- For each $i = 1, \ldots, M$,
  1. draw $\sigma^2_{(i)} \sim p(\sigma^2|\mathbf{y}) = IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)$
  2. draw $\boldsymbol{\beta}_{(i)} \sim p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) = N\left(\hat{\boldsymbol{\beta}}, \sigma^2_{(i)}(X^T X)^{-1}\right)$.
- The resulting samples can be used to approximate the joint as well as marginal posteriors.

Prediction from Bayesian Linear Models

- Prediction for a new $m \times p$ covariance matrix $\tilde{X}$ relies on the posterior predictive distribution

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = \int p(\tilde{\mathbf{y}}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) d\boldsymbol{\beta} d\sigma^2.$$

- This is another multivariate student-t distribution

$$p(\tilde{\mathbf{y}}|\mathbf{y}) = MVSt_{n-p}\left(\tilde{X}\hat{\boldsymbol{\beta}}, s^2(I_m + \tilde{X}(X^T X)^{-1}\tilde{X})\right)$$

- Yet another way: sampling-based approximation
  1. draw $\sigma^2_{(i)} \sim p(\sigma^2|\mathbf{y}) = IG\left(\frac{n-p}{2}, \frac{(n-p)s^2}{2}\right)$
  2. draw $\boldsymbol{\beta}_{(i)} \sim p(\boldsymbol{\beta}|\sigma^2, \mathbf{y}) = N\left(\hat{\boldsymbol{\beta}}, \sigma^2_{(i)}(X^T X)^{-1}\right)$.
  3. draw $\tilde{\mathbf{y}}_{(i)}$ from $N\left(\tilde{X}\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)} I\right)$

The NIG conjugate prior

- The Normal-Inverse-Gamma (NIG) prior is conjugate for the regression parameters $(\boldsymbol{\beta}, \sigma^2)$

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim N(\boldsymbol{\mu}_\beta, \sigma^2 V_\beta) \\ \sigma^2 &\sim IG(a, b) \end{aligned}$$

denoted as $NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b)$.

- The resulting joint posterior is $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)$

where

$$\begin{aligned} \boldsymbol{\mu}^* &= (V_\beta^{-1} + X^T X)^{-1}(V_\beta \boldsymbol{\mu}_\beta + X^T \mathbf{y}) \\ V^* &= (V_\beta^{-1} + X^T X)^{-1} \\ a^* &= a + n/2 \\ b^* &= b + \left(\boldsymbol{\mu}_\beta^T V_\beta^{-1} \boldsymbol{\mu}_\beta + \mathbf{y}^T \mathbf{y} - (\boldsymbol{\mu}^*)^T (V^*)^{-1} \boldsymbol{\mu}^*\right)/2 \end{aligned}$$

- The marginal posterior is

$$p(\boldsymbol{\beta}|\mathbf{y}) = MVSt_{2a^*}(\boldsymbol{\mu}^*, \frac{b^*}{a^*} V^*), \quad p(\sigma^2|\mathbf{y}) = IG(a^*, b^*)$$

The NIG conjugate prior

- The noninformative prior $\pi(\boldsymbol{\beta}, \sigma^2) = 1/\sigma^2$ can be considered as the limit of an NIG prior with
  - $V_\beta^{-1} \to 0$ (i.e. the null matrix)
  - $a \to -p/2$
  - $b \to 0$

  and results in the posterior parameters

$$\boldsymbol{\mu}^* = \hat{\boldsymbol{\beta}},\ V^* = (X^T X)^{-1},\ a^* = \frac{n-p}{2},\ b^* = \frac{(n-p)s^2}{2}$$

Bayesian Prediction

- The posterior predictive distribution is obtained as

$$
\begin{aligned}
p(\tilde{\mathbf{y}}|\mathbf{y}) &= \int p(\tilde{\mathbf{y}}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta},\sigma^2|\mathbf{y})d\boldsymbol{\beta}d\sigma^2 \\
&= \int N(\tilde{\mathbf{y}}|\tilde{X}\boldsymbol{\beta},\sigma^2 I_m) \times NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)d\boldsymbol{\beta}d\sigma^2 \\
&= MVSt_{2a^*}\left(\tilde{X}\boldsymbol{\mu}^*, \frac{b^*}{a^*}(I_m + \tilde{X}V^*\tilde{X}^T)\right)
\end{aligned}
$$

- Note: There are two sources of uncertainty in the posterior predictive distribution
  (1) the variability in the model due to residual errors
  (2) the posterior uncertainty in $\boldsymbol{\beta}$ and σ^2 estimation

  As the sample size $n \to \infty$, the variance due to estimation uncertainty *disappears*, but the predictive uncertainty *remains*.

- Similarly, we can use multi-stage sampling algorithm to approximate the posterior predictive distribution.

Marginal distribution $m(\mathbf{y})$

- With a proper NIG prior, we can obtain the marginal likelihood $m(\mathbf{y})$

$$
\begin{aligned}
m(\mathbf{y}) &= \int f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2) d\boldsymbol{\beta} d\sigma^2 \\
&= \int N(\mathbf{y}|X\boldsymbol{\beta}, \sigma^2 I_n) \times NIG(\boldsymbol{\mu}, V, a, b) d\boldsymbol{\beta} d\sigma^2 \\
&= MVSt_{2a}\left(X\boldsymbol{\mu}_\beta, \frac{b}{a}(I_n + XV_\beta X^T)\right)
\end{aligned}
$$

- The closed-form marginal likelihood allows for straightforward model selection using Bayes Factor!

Key Summaries

- Bayesian analysis with Noninformative priors parallels the classical results.

- Using the NIG conjugate prior $NIG(\boldsymbol{\mu}_\beta, V_\beta, a, b)$ for the linear model
  - The joint posterior $p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$ is again an NIG distribution $NIG(\boldsymbol{\mu}^*, V^*, a^*, b^*)$.
  - The marginal posterior $p(\sigma^2)$ is $IG(a^*, b^*)$; and the marginal posterior $p(\boldsymbol{\beta}|\mathbf{y})$ is $MVSt_{2a^*}(\boldsymbol{\mu}^*, \frac{b^*}{a^*}V^*)$.
  - The marginal distribution $m(\mathbf{y})$ is $MVSt_{2a}(X\boldsymbol{\mu}_\beta, \frac{b}{a}(I_n + XV_\beta X^T))$.
  - The posterior predictive distribution $p(\tilde{y}|\mathbf{y}) = MVSt_{2a^*}(\tilde{X}\boldsymbol{\mu}^*, \frac{b^*}{a^*}(I_m + \tilde{X}V^*\tilde{X}^T))$.
  - All these distributions can be well approximated using a sampling-based algorithm!

- The noninformative prior $\pi(\boldsymbol{\beta}, \sigma^2) = 1/\sigma^2$ is the limit of an NIG prior with $V_\beta^{-1} \to 0$ (i.e. the null matrix), and $a \to -p/2,\ b \to 0$.