Bayesian linear regression

Bayesian Methodology in Biostatistics (BST 249)

Jeffrey W. Miller

Department of Biostatistics
Harvard T.H. Chan School of Public Health

Outline

Model, prior, and posterior

- Conditionally conjugate prior on β
- Conditionally conjugate prior on σ^2

Setting the hyperparameters

- Unit information prior
- Zellner's g-prior
- Computation with Zellner's g-prior

Outline

Model, prior, and posterior

- Conditionally conjugate prior on β
- Conditionally conjugate prior on σ^2

Setting the hyperparameters

- Unit information prior
- Zellner's g-prior
- Computation with Zellner's g-prior

Linear regression model

- Data: $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.
  - $x_i = (x_{i1}, \ldots, x_{ip})^{\mathtt{T}}$ = vector of covariates
  - y_i = outcome

- Model: $Y_i \sim \mathcal{N}(x_i^{\mathtt{T}}\beta,\, \sigma^2)$ independently for $i = 1, \ldots, n$.
- Throughout, we will treat $x_{1:n}$ as fixed and known.

- Equivalently, $Y \sim \mathcal{N}(X\beta,\, \sigma^2 I)$ where

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n \qquad X = \begin{bmatrix} x_1^{\mathtt{T}} \\ \vdots \\ x_n^{\mathtt{T}} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

and I is the $p \times p$ identity matrix.

Conditionally conjugate prior on β

- Assume a multivariate normal prior on β:

$$\boldsymbol{\beta} \sim \mathcal{N}(m_0, L_0^{-1}).$$

- For any fixed σ^2, this is a conjugate prior.

- The resulting posterior, with σ^2 fixed, is:

$$\boldsymbol{\beta}|y_{1:n} \sim \mathcal{N}(m_n, L_n^{-1})$$

where

$$L_n = L_0 + X^{\mathtt{T}} X/\sigma^2,$$
$$m_n = L_n^{-1}(L_0 m_0 + X^{\mathtt{T}} y/\sigma^2),$$

and $y = (y_1, \ldots, y_n)^{\mathtt{T}} \in \mathbb{R}^n$.

Connections with the MLE and ridge regression

- The *maximum a posteriori* (MAP) estimate is the mode of the posterior.
- Since the posterior is normal, the mode is equal to the mean:

$$\hat{\beta}_{\mathsf{MAP}} = m_n = (L_0 + X^{\mathtt{T}} X/\sigma^2)^{-1}(L_0 m_0 + X^{\mathtt{T}} y/\sigma^2).$$

- In the limit as $L_0 \to 0$, we have $L_n \to X^{\mathtt{T}} X/\sigma^2$ and

$$\hat{\beta}_{\mathsf{MAP}} = m_n \to (X^{\mathtt{T}} X)^{-1} X^{\mathtt{T}} y = \hat{\beta}_{\mathsf{MLE}}.$$

- Thus, in the limit of having "no prior information" about β,

$$\boldsymbol{\beta} | y_{1:n} \sim \mathcal{N}(\hat{\beta}_{\mathsf{MLE}},\ \sigma^2 (X^{\mathtt{T}} X)^{-1}).$$

- Note that the frequentist sampling distribution of the MLE is

$$\hat{\beta}_{\mathsf{MLE}} \sim \mathcal{N}(\beta,\ \sigma^2 (X^{\mathtt{T}} X)^{-1})$$

  when β is the true parameter, which exhibits perfect symmetry with the posterior above when we take $L_0 \to 0$.

Connections with the MLE and ridge regression

- Taking $L_0 \to 0$ is equivalent to using a *flat prior* $p(\beta) \propto 1$ and formally applying Bayes' theorem: $p(\beta|y_{1:n}) \propto p(y_{1:n}|\beta)$.
- $p(\beta) \propto 1$ is called an *improper* prior since it cannot be normalized to a probability density on $\mathbb{R}^p$.
- Even though the flat prior on β is improper, defining $p(\beta|y_{1:n}) \propto p(y_{1:n}|\beta)$ still results in a well-defined posterior distribution as long as $n \geq 1$.

- This is an example of a *noninformative* prior.

- The MAP estimate also generalizes ridge regression.
- Specifically, if $m_0 = 0$ and $L_0 = \alpha I/\sigma^2$ then

$$\hat{\beta}_{\text{MAP}} = (X^{\mathtt{T}}X + \alpha I)^{-1}X^{\mathtt{T}}y = \hat{\beta}_{\text{ridge}}.$$

Outline

Model, prior, and posterior

- Conditionally conjugate prior on β
- Conditionally conjugate prior on σ^2

Setting the hyperparameters

- Unit information prior
- Zellner's g-prior
- Computation with Zellner's g-prior

Conjugate prior on σ^2

- Similar to the case of a univariate normal model, the Inverse Gamma distribution is a conjugate prior for σ^2.

- Suppose β is fixed, and define the prior on σ^2 as:

$$\boldsymbol{\sigma}^2 \sim \mathrm{InvGamma}\left(\tfrac{1}{2}\nu_0,\ \tfrac{1}{2}\nu_0\sigma_0^2\right).$$

- This parametrization facilitates interpretation:
  - $\sigma_0^2 =$ prior guess of σ^2 (since σ_0 is between the mean and the mode of this prior on $\boldsymbol{\sigma}^2$)
  - $\nu_0 =$ confidence in the choice of σ_0^2, in units of sample size — i.e., the strength of the prior is equivalent to ν_0 samples.

- The resulting posterior on σ^2 is then

$$\boldsymbol{\sigma}^2 | y_{1:n} \sim ???$$

(Whiteboard activity)

Conjugate prior on σ^2

- Similar to the case of a univariate normal model, the Inverse Gamma distribution is a conjugate prior for σ^2.

- Suppose β is fixed, and define the prior on σ^2 as:

$$\boldsymbol{\sigma}^2 \sim \mathrm{InvGamma}\big(\tfrac{1}{2}\nu_0,\ \tfrac{1}{2}\nu_0\sigma_0^2\big).$$

- This parametrization facilitates interpretation:
  - $\sigma_0^2 =$ prior guess of σ^2 (since σ_0 is between the mean and the mode of this prior on $\boldsymbol{\sigma}^2$)
  - $\nu_0 =$ confidence in the choice of σ_0^2, in units of sample size — i.e., the strength of the prior is equivalent to ν_0 samples.

- The resulting posterior on σ^2 is then

$$\boldsymbol{\sigma}^2 | y_{1:n} \sim \mathrm{InvGamma}\big(\tfrac{1}{2}(\nu_0 + n),\ \tfrac{1}{2}(\nu_0\sigma_0^2 + \mathrm{SSR}(\beta))\big)$$

where $\mathrm{SSR}(\beta) = \sum_{i=1}^{n}(y_i - x_i^{\mathtt{T}}\beta)^2$.

Inference for β and σ^2 jointly

- We've seen conjugate priors for β and σ^2, given the other.
- Thus, these define a conditionally conjugate prior, combined:

$$\boldsymbol{\beta} \sim \mathcal{N}(m_0, L_0^{-1}) \qquad \boldsymbol{\sigma}^2 \sim \mathrm{InvGamma}\big(\tfrac{1}{2}\nu_0,\ \tfrac{1}{2}\nu_0\sigma_0^2\big)$$

  independently.
- We know how to compute $\beta|\sigma^2, y_{1:n}$ and $\sigma^2|\beta, y_{1:n}$. How can we infer β and σ^2 jointly?

- One way: Initialize β, σ^2 and iteratively repeat the following:
  1. update β by sampling from $\beta|\sigma^2, y_{1:n}$
  2. update σ^2 by sampling from $\sigma^2|\beta, y_{1:n}$

- It turns out that this generates approximate samples from $\beta, \sigma^2|y_{1:n}$. This is an example of a *Gibbs sampler*, which is type of a Markov chain Monte Carlo algorithm.

Group activity: Statistics trivia challenge!

Go to breakout rooms and work together to answer these questions:
https://forms.gle/agm9P7PpLTWNUmDr5

(Three people per room, randomly assigned. 5 minutes.)

Outline

Model, prior, and posterior

- Conditionally conjugate prior on β
- Conditionally conjugate prior on σ^2

Setting the hyperparameters

- Unit information prior
- Zellner's g-prior
- Computation with Zellner's g-prior

Setting hyperparameters: Data-dependent, Unit info

- The subjective Bayesian approach is to set hyperparameters based solely on prior beliefs. However, sometimes we prefer to use *default priors* that don't require subjective input.

- Data-dependent priors are a useful way of creating default priors, even though, strictly speaking, they violate the principle of not using the data twice.

- Kass & Wasserman (1995) propose the following settings:

$$m_0 = \hat{\beta}_{\text{MLE}} \qquad L_0 = X^{\mathtt{T}} X / (n\sigma^2).$$

- Similarly, we can define a data-dependent prior on σ^2:

$$\sigma_0^2 = \hat{\sigma}^2_{\text{MLE}} \qquad \nu_0 = 1$$

where $\hat{\sigma}^2_{\text{MLE}}$ is the maximum likelihood estimate of σ^2.

- These are both *unit information priors* — the strength of the prior is equivalent to one sample.

Setting hyperparameters: Zellner's g-prior

- A compelling property of $\hat{\beta}_{\text{MLE}}$ is that the scale of the predictors doesn't matter, in the following sense:
  - Suppose one of the predictors is age in years.
  - If we change the units of age to months, then entry of $\hat{\beta}_{\text{MLE}}$ corresponding to age is scaled accordingly by $1/12$.
  - Consequently, $X\hat{\beta}_{\text{MLE}}$ is invariant to the choice of units.

- Mathematically, changing the units means using the model $Y \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I)$ with $\tilde{X} = XH$ in place of X, where H is a diagonal matrix with $H_{jj} > 0$ for all j. Observe that

$$\tilde{X}\tilde{\hat{\beta}}_{\text{MLE}} = ??? = X\hat{\beta}_{\text{MLE}}.$$

(Whiteboard activity)

Setting hyperparameters: Zellner's g-prior

- A compelling property of $\hat{\beta}_{\mathsf{MLE}}$ is that the scale of the predictors doesn't matter, in the following sense:
  - Suppose one of the predictors is age in years.
  - If we change the units of age to months, then entry of $\hat{\beta}_{\mathsf{MLE}}$ corresponding to age is scaled accordingly by $1/12$.
  - Consequently, $X\hat{\beta}_{\mathsf{MLE}}$ is invariant to the choice of units.

- Mathematically, changing the units means using the model $Y \sim \mathcal{N}(\tilde{X}\beta, \sigma^2 I)$ with $\tilde{X} = XH$ in place of X, where H is a diagonal matrix with $H_{jj} > 0$ for all j. Observe that

$$\begin{aligned}
\tilde{X}\tilde{\hat{\beta}}_{\mathsf{MLE}} &= \tilde{X}(\tilde{X}^{\mathtt{T}}\tilde{X})^{-1}\tilde{X}^{\mathtt{T}}y \\
&= XH(H^{\mathtt{T}}X^{\mathtt{T}}XH)^{-1}H^{\mathtt{T}}X^{\mathtt{T}}y \\
&= XHH^{-1}(X^{\mathtt{T}}X)^{-1}H^{-\mathtt{T}}H^{\mathtt{T}}X^{\mathtt{T}}y \\
&= X(X^{\mathtt{T}}X)^{-1}X^{\mathtt{T}}y = X\hat{\beta}_{\mathsf{MLE}}.
\end{aligned}$$

Setting hyperparameters: Zellner's g-prior

- This is nice since, intuitively, the units shouldn't matter.
- For the Bayesian model, the invariance property we seek is:
  - Suppose $\tilde{X} = XH$ for some invertible $H \in \mathbb{R}^{p \times p}$.
  - Let $\boldsymbol{\beta}$ be distributed according to the posterior using X.
  - Let $\tilde{\boldsymbol{\beta}}$ be distributed according to the posterior using $\tilde{X}$.
  - We would like $X\boldsymbol{\beta}$ and $\tilde{X}\tilde{\boldsymbol{\beta}}$ to have the same distribution.
- Unfortunately, for most choices of prior, the posterior doesn't have this invariance property.
- However, if we can make it so that $\boldsymbol{\beta}$ and $H\tilde{\boldsymbol{\beta}}$ have the same distribution, then this will work, since then $X\boldsymbol{\beta}$ has the same distribution as $XH\tilde{\boldsymbol{\beta}} = \tilde{X}\tilde{\boldsymbol{\beta}}$.

Setting hyperparameters: Zellner's g-prior

- This invariance property is satisfied by Zellner's g-*prior*:

  $\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}(m_0, L_0^{-1})$ where $m_0 = 0$ and $L_0 = X^\mathtt{T} X/(g\sigma^2)$.

- Here, $g > 0$ is a free parameter.

- Given σ^2, the posterior on β under a g-prior simplifies to:

$$\boldsymbol{\beta} \mid \sigma^2, y_{1:n} \sim \mathcal{N}(m_n, L_n^{-1})$$

  where

$$L_n = L_0 + X^\mathtt{T} X/\sigma^2 = \frac{g+1}{g\sigma^2} X^\mathtt{T} X,$$
$$m_n = L_n^{-1}(L_0 m_0 + X^\mathtt{T} y/\sigma^2) = \frac{g}{g+1}(X^\mathtt{T} X)^{-1} X^\mathtt{T} y.$$

Setting hyperparameters: Zellner's g-prior

- To check that the invariance property is satisfied, suppose:

$$\tilde{m}_0 = 0 \qquad \tilde{L}_0 = (XH)^{\mathtt{T}}(XH)/(g\sigma^2).$$

- Then the posterior is $\tilde{\boldsymbol{\beta}} \mid \sigma^2, y_{1:n} \sim \mathcal{N}(\tilde{m}_n, \tilde{L}_n^{-1})$ where

$$\tilde{L}_n = ???$$
$$\tilde{m}_n = ???.$$

- Since $H\tilde{\boldsymbol{\beta}} \mid \sigma^2, y_{1:n} \sim \mathcal{N}(???, ???)$, the invariance property requires that

$$H\tilde{m}_n = m_n \quad \text{and} \quad H\tilde{L}_n^{-1}H^{\mathtt{T}} = L_n^{-1}.$$

- **Group activity: Go to breakout rooms and work together to (1) fill in the blanks above and (2) check that the equations above in red hold. (15 minutes)**

Outline

Model, prior, and posterior

Conditionally conjugate prior on β

Conditionally conjugate prior on σ^2

Setting the hyperparameters

Unit information prior

Zellner's g-prior

Computation with Zellner's g-prior

Posterior computation with Zellner's g-prior

- The g-prior also simplifies posterior inference for (β, σ^2).
- Assume a g-prior on $\beta|\sigma^2$ and an InvGamma prior on σ^2:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim \mathcal{N}\big(0,\, g\sigma^2(X^{\mathtt{T}}X)^{-1}\big) \\ \boldsymbol{\sigma^2} &\sim \mathrm{InvGamma}\left(\tfrac{1}{2}\nu_0,\, \tfrac{1}{2}\nu_0\sigma_0^2\right). \end{aligned}$$

- It turns out that we can generate i.i.d. samples from the posterior on (β, σ^2), so we don't need to use MCMC.
- These samples can be used for Monte Carlo approximation of posterior expectations, e.g., for any integrable function h,

$$\mathrm{E}(h(\boldsymbol{\beta}, \boldsymbol{\sigma^2}) \mid y_{1:n}) \approx \frac{1}{T}\sum_{t=1}^{T} h(\beta_t, \sigma_t^2).$$

- Monte Carlo has various advantages over MCMC: simplicity, efficiency, and reliable quantification of approximation error.

Posterior computation with Zellner's g-prior

- Under the prior on the previous slide, it can be shown that the posterior is (Hoff, 2009):

$$\boldsymbol{\sigma^2} \mid y_{1:n} \sim \text{InvGamma}\left(\tfrac{1}{2}(\nu_0 + n),\ \tfrac{1}{2}(\nu_0\sigma_0^2 + \text{SSR}_g)\right)$$

$$\boldsymbol{\beta} \mid \sigma^2, y_{1:n} \sim \mathcal{N}\left(\frac{g}{g+1}\hat{\beta}_{\mathsf{MLE}},\ \frac{g}{g+1}\sigma^2(X^\mathtt{T}X)^{-1}\right)$$

  where

$$\text{SSR}_g = y^\mathtt{T}y - \frac{g}{g+1}y^\mathtt{T}X(X^\mathtt{T}X)^{-1}X^\mathtt{T}y.$$

- Thus, we can generate i.i.d. samples from the posterior by:
  1. sampling $\sigma^2 \mid y_{1:n}$,
  2. sampling $\beta \mid \sigma^2, y_{1:n}$.

References and supplements

- Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. Springer, New York.
- Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, 90(431), 928-934.
- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian inference and Decision Techniques, vol 6, North-Holland, Amsterdam, 233-243.

Setting hyperparameters: Zellner's g-prior (SOLUTIONS)

- To check that the invariance property is satisfied, suppose:

$$\tilde{m}_0 = 0 \qquad \tilde{L}_0 = (XH)^{\mathtt{T}}(XH)/(g\sigma^2).$$

- Then the posterior is $\tilde{\boldsymbol{\beta}} \mid \sigma^2, y_{1:n} \sim \mathcal{N}(\tilde{m}_n, \tilde{L}_n^{-1})$ where

$$\begin{aligned} \tilde{L}_n &= \frac{g+1}{g\sigma^2}\, H^{\mathtt{T}} X^{\mathtt{T}} X H \\ \tilde{m}_n &= \frac{g}{g+1} (H^{\mathtt{T}} X^{\mathtt{T}} X H)^{-1} H^{\mathtt{T}} X^{\mathtt{T}} y. \end{aligned}$$

- Since $H\tilde{\boldsymbol{\beta}} \mid \sigma^2, y_{1:n} \sim \mathcal{N}(H\tilde{m}_n,\, H\tilde{L}^{-1}H^{\mathtt{T}})$, the invariance property requires that

$$H\tilde{m}_n = m_n \quad \text{and} \quad H\tilde{L}_n^{-1}H^{\mathtt{T}} = L_n^{-1}.$$

Setting hyperparameters: Zellner's g-prior (SOLUTIONS)

These equations hold since:

$$\begin{aligned} H\tilde{m}_n &= \frac{g}{g+1} H(H^\mathsf{T} X^\mathsf{T} X H)^{-1} H^\mathsf{T} X^\mathsf{T} y \\ &= \frac{g}{g+1} H H^{-1} (X^\mathsf{T} X)^{-1} H^{-\mathsf{T}} H^\mathsf{T} X^\mathsf{T} y \\ &= \frac{g}{g+1} (X^\mathsf{T} X)^{-1} X^\mathsf{T} y = m_n \end{aligned}$$

and

$$\begin{aligned} H\tilde{L}_n^{-1} H^\mathsf{T} &= H\Big(\frac{g+1}{g\sigma^2}\, H^\mathsf{T} X^\mathsf{T} X H\Big)^{-1} H^\mathsf{T} \\ &= H H^{-1}\Big(\frac{g+1}{g\sigma^2}\, X^\mathsf{T} X\Big)^{-1} H^{-\mathsf{T}} H^\mathsf{T} \\ &= \Big(\frac{g+1}{g\sigma^2}\, X^\mathsf{T} X\Big)^{-1} = L_n^{-1}. \end{aligned}$$