Bayesian Estimation in Linear Models

STA 721: Lecture 8

Merlise Clyde (clyde@duke.edu)
Duke University

Outline

Readings:

- Christensen Chapter 2.9 and Chapter 15
- Seber & Lee Chapter 3.12

Bayes Estimation

Model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \textsf{N}(\mathbf{0}_n, \sigma^2\mathbf{I}_n)$ is equivalent to

$$\mathbf{Y} \sim \textsf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n/\phi)$$

- $\phi = 1/\sigma^2$ is the **precision** of the data.
- we might expect $\boldsymbol{\beta}$ to be close to some vector $\mathbf{b}_0$
- represent this *a priori* with a **Prior Distribution** for $\boldsymbol{\beta}$, e.g.

$$\boldsymbol{\beta} \sim \textsf{N}(\mathbf{b}_0, \boldsymbol{\Phi}_0^{-1})$$

- $\mathbf{b}_0$ is the prior mean and $\boldsymbol{\Phi}_0$ is the **prior precision** of $\boldsymbol{\beta}$ that captures how close $\boldsymbol{\beta}$ is to $\mathbf{b}_0$
- Similarly, we could represent prior uncertainty about σ, σ^2 or equivalently ϕ with a probability distribution
- for now treat ϕ as fixed

Bayesian Inference

- once we see data $\mathbf{Y}$, Bayesian inference proceeds by updating prior beliefs
- represented by the **posterior distribution** of $\boldsymbol{\beta}$ which is the conditional distribution of $\boldsymbol{\beta}$ given the data $\mathbf{Y}$ (and ϕ for now)
- Posterior $p(\boldsymbol{\beta} \mid \mathbf{Y}, \phi)$

$$p(\boldsymbol{\beta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta} \mid \phi)}{c}$$

- c is a constant so that the posterior density integrates to 1

$$c = \int_{\mathbb{R}^p} p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta} \mid \phi) d\,\boldsymbol{\beta} \equiv p(\mathbf{Y})$$

- since c is a constant that doesn't depend on $\boldsymbol{\beta}$ just ignore
- work with density up to constant of proportionality

Posterior Density

Posterior for $\boldsymbol{\beta}$ is $p(\boldsymbol{\beta} \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta} \mid \phi)$

- Likelihood for $\boldsymbol{\beta}$ is proportional to $p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi)$

$$
\begin{aligned}
p(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) &= (2\pi)^{-n/2} |\mathbf{I}_n/\phi|^{-1/2} \exp\left\{-\frac{1}{2}\left((\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \phi \mathbf{I}_n (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\phi \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T \phi \mathbf{X}^T\mathbf{Y} + \phi \boldsymbol{\beta} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta}\right)\right\}
\end{aligned}
$$

- similarly expand prior

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \phi) &= (2\pi)^{-p/2} |\boldsymbol{\Phi}_0^{-1}|^{-1/2} \exp\left\{-\frac{1}{2}\left((\boldsymbol{\beta} - \mathbf{b}_0)^T \boldsymbol{\Phi}_0 (\boldsymbol{\beta} - \mathbf{b}_0)\right)\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left(\mathbf{b}_0^T \boldsymbol{\Phi}_0 \mathbf{b}_0 - 2\boldsymbol{\beta}^T \boldsymbol{\Phi}_0 \mathbf{b}_0 + \boldsymbol{\beta} \boldsymbol{\Phi}_0 \boldsymbol{\beta}\right)\right\}
\end{aligned}
$$

Posterior Steps

- Expand quadratics and regroup terms

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{Y}, \phi) &\propto e^{\left\{-\frac{1}{2}\left(\phi \boldsymbol{\beta} \mathbf{X}^T \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta} \boldsymbol{\Phi}_0 \boldsymbol{\beta} - 2(\phi \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{Y} + \boldsymbol{\beta}^T \boldsymbol{\Phi}_0 \mathbf{b}_0) + \phi \mathbf{Y}^T \mathbf{Y} + \mathbf{b}_0^T \boldsymbol{\Phi}_0 \mathbf{b}_0\right)\right\}} \\
&\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}(\phi \mathbf{X}^T \mathbf{X} + \boldsymbol{\Phi}_0)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\phi \mathbf{X}^T \mathbf{Y} + \boldsymbol{\Phi}_0 \mathbf{b}_0)\right)\right\}
\end{aligned}
$$

Kernel of a Multivariate Normal

- Read off posterior precision from Quadratic in $\boldsymbol{\beta}$
- Read off posterior precision $\times$ posterior mean from Linear term in $\boldsymbol{\beta}$
- will need to complete the quadratic in the posterior mean†

Posterior Precision and Covariance

$$p(\boldsymbol{\beta} \mid \mathbf{Y}, \phi) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}(\phi\mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\phi\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0)\right)\right\}$$

- Posterior Precision

$$\boldsymbol{\Phi}_n \equiv \phi\mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0$$

- sum of data precision and prior precision
- posterior Covariance

$$\textsf{Cov}[\boldsymbol{\beta} \mid \mathbf{Y}, \phi] = \boldsymbol{\Phi}_n^{-1} = (\phi\mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)^{-1}$$

- if $\boldsymbol{\Phi}_0$ is full rank, then Cov$[\boldsymbol{\beta} \mid \mathbf{Y}, \phi]$ is full rank even if $\mathbf{X}^T\mathbf{X}$ is not

Posterior Mean Updating

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \mathbf{Y}, \phi) &\propto \exp\left\{ \frac{1}{2} (\boldsymbol{\beta}(\phi \mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T(\phi\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0)) \right\} \\
&\propto \exp\left\{ \frac{1}{2} (\boldsymbol{\beta}(\phi \mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^{-1}(\phi\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0)) \right\}
\end{aligned}
$$

- posterior mean $\mathbf{b}_n$

$$
\begin{aligned}
\mathbf{b}_n &\equiv \boldsymbol{\Phi}_n^{-1}(\phi\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0) \\
&= (\phi\mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)^{-1}\left(\phi(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} + \boldsymbol{\Phi}_0\mathbf{b}_0\right) \\
&= (\phi\mathbf{X}^T\mathbf{X} + \boldsymbol{\Phi}_0)^{-1}\left(\phi(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} + \boldsymbol{\Phi}_0\mathbf{b}_0\right)
\end{aligned}
$$

- a precision weighted linear combination of MLE and prior mean
- first expression useful if $\mathbf{X}$ is not full rank!

Notes

Posterior is a Multivariate Normal $p(\boldsymbol{\beta} \mid \mathbf{Y}, \phi) \sim \mathsf{N}(\mathbf{b}_n, \boldsymbol{\Phi}_n^{-1})$

- posterior mean: $\mathbf{b}_n = \boldsymbol{\Phi}_n^{-1}(\phi \mathbf{X}^T \mathbf{Y} + \boldsymbol{\Phi}_0 \mathbf{b}_0)$
- posterior precision: $\boldsymbol{\Phi}_n = \phi \mathbf{X}^T \mathbf{X} + \boldsymbol{\Phi}_0$
- the posterior precision (inverse posterior variance) is the sum of the prior precision and the data precision.
- the posterior mean is a linear combination of MLE/OLS and prior mean
- if the prior precision $\boldsymbol{\Phi}_n$ is very large compared to the data precision $\phi \mathbf{X}^T \mathbf{X}$, the posterior mean will be close to the prior mean $\mathbf{b}_0$.
- if the prior precision $\boldsymbol{\Phi}_n$ is very small compared to the data precision $\phi \mathbf{X}^T \mathbf{X}$, the posterior mean will be close to the MLE/OLS estimator.
- data precision will generally be increasing with sample size

Bayes Estimators

A Bayes estimator is a potential value of β that is obtained from the posterior distribution in some principled way.

- Standard estimators include
  - the posterior mean estimator, which is the minimizer of the Bayes risk under squared error loss
  - the maximum a posteriori (MAP) estimator, the value β that maximizes the posterior density (or log posterior density)
- The first estimator is based on principles from classical decision theory, whereas the second can be related to penalized likelihood estimation.
- in the case of linear regression they turn out to be the same estimator!

Bayes Estimator under Squared Error Loss

- the Frequentist Risk $R(\beta, \delta) \equiv \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\|\delta(\mathbf{Y}) - \boldsymbol{\beta}\|^2]$ is the expected loss of decision δ for a given $\boldsymbol{\beta}$

▼ **Definition:** Bayes Rule and Bayes Risk

The Bayes rule under squared error loss is the function of $\mathbf{Y}$, $\delta^*(\mathbf{Y})$, that minimizes the **Bayes risk** $B(p_{\boldsymbol{\beta}}, \delta)$

$$\delta^*(\mathbf{Y}) = \arg\min_{\delta \in \mathcal{D}} B(p_{\boldsymbol{\beta}}, \delta)$$

$$B(p_{\boldsymbol{\beta}}, \delta) = \mathsf{E}_{\boldsymbol{\beta}} R(\boldsymbol{\beta}, \delta) = \mathsf{E}_{\boldsymbol{\beta}} \mathsf{E}_{\mathbf{Y}|\boldsymbol{\beta}}[\|\delta(\mathbf{Y}) - \boldsymbol{\beta}\|^2]$$

where the expectation is with respect to the prior distribution, $p_{\boldsymbol{\beta}}$, over $\boldsymbol{\beta}$ and the conditional distribution of $\mathbf{Y}$ given $\boldsymbol{\beta}$

Bayes Estimators

▼ **Definition:** Bayes Action

The Bayes Action is the action $a \in \mathcal{A}$ that minimizes the posterior expected loss:

$$\delta^*_B(\mathbf{Y}) = \arg\min_{\delta \in \mathcal{D}} E_{\boldsymbol{\beta}|\mathbf{Y}}[\|\delta - \boldsymbol{\beta}\|^2]$$

Prior Choice

One of the most common priors for the normal linear model is the **g-prior** of Zellner (1986) where $\mathbf{\Phi}_0 = \frac{\phi}{g}\mathbf{X}^T\mathbf{X}$

$$\boldsymbol{\beta} \mid \phi, g \sim \mathsf{N}(\mathbf{0}, g/\phi(\mathbf{X}^T\mathbf{X})^{-1})$$

$$\begin{aligned}
\mathbf{b}_n &= \left(\mathbf{X}^T\mathbf{X} + \frac{\phi}{g}\frac{\mathbf{X}^T\mathbf{X}}{\phi}\right)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \left(\mathbf{X}^T\mathbf{X} + \frac{1}{g}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \left(\frac{1+g}{g}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \\
&= \frac{g}{1+g}\hat{\boldsymbol{\beta}}
\end{aligned}$$

Another Common Choice

- another common choice is the independent prior

$$\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{0}, \boldsymbol{\Phi}_0^{-1})$$

  where $\boldsymbol{\Phi}_0 = \phi\kappa\mathbf{I}_b$ for some $\kappa > 0$

- the posterior mean is

$$\begin{aligned}\boldsymbol{\beta}_n &= (\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{X}^T\mathbf{X} + \kappa\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

- this is also a shrinkage estimator but the amount of shrinkage is different for the different components of $\mathbf{b}_n$ depending on the eigenvalues of $\mathbf{X}^T\mathbf{X}$
- easiest to see this via an orthogonal rotation of the model

Rotated Regression

- Use the singular value decomposition of $\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$ and multiply thru by $\mathbf{U}^T$ to get the rotated model

$$\mathbf{U}^T\mathbf{Y} = \boldsymbol{\Lambda}\mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon}$$
$$\tilde{\mathbf{Y}} = \boldsymbol{\Lambda}\boldsymbol{\alpha} + \tilde{\boldsymbol{\epsilon}}$$

  where $\boldsymbol{\alpha} = \mathbf{V}^T\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\epsilon}} = \mathbf{U}^T\boldsymbol{\epsilon}$
- the induced prior is still $\boldsymbol{\alpha} \mid \phi \sim \mathsf{N}(\mathbf{0}, (\phi\kappa)^{-1}\mathbf{I})$
- the posterior mean of $\boldsymbol{\alpha}$ is

$$\mathbf{a} = (\boldsymbol{\Lambda}^2 + \kappa\mathbf{I})^{-1}\boldsymbol{\Lambda}^2\hat{\boldsymbol{\alpha}}$$
$$a_j = \frac{\lambda_j^2}{\lambda_j^2 + \kappa}\hat{\alpha}_j$$

- sets to zero the components of the OLS solution where eigenvalues are zero!

Connections to Frequentist Estimators

- The posterior mean under this independent prior is the same as the classic **ridge regression** estimator of Hoerl and
- the variance of $\hat{\alpha}_j$ is σ^2/λ_j^2 while the variance of a_j is $\sigma^2/(\lambda_j^2 + \kappa)$
- clearly components of $\boldsymbol{\alpha}$ with small eigenvalues will have large variances
- ridge regression keeps those components from "blowing up" by shrinking them towards zero and having a finite variance
- rotate back to get the ridge estimator for $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_R = \mathbf{V}\mathbf{a}$
- ridge regression applies a high degree of shrinkage to the "parts" (linear combinations) of $\boldsymbol{\beta}$ that have high variability, and a low degree of shrinkage to the parts that are well-estimated.
- turns out there always exists a value of κ that will improve over OLS!
- Unfortunately no closed form solution except in orthogonal regression and then it depends on the unknown $||\boldsymbol{\beta}||^2$!

Next Class

- Frequentist risk of Bayes estimators
- Bayes and penalized loss functions