Bayesian Estimation in Linear Models

STA721 Linear Models Duke University

Merlise Clyde

September 22, 2016

Bayesian Estimation

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

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

$\phi = 1/\sigma^2$ is the *precision*.

In the Bayesian paradigm describe uncertainty about unknown parameters using probability distributions

- Prior Distribution $p(\boldsymbol{\beta}, \phi)$ describes uncertainty about parameters prior to seeing the data
- Posterior Distribution $p(\boldsymbol{\beta}, \phi \mid \mathbf{Y})$ describes uncertainty about the parameters after updating believes given the observed data
- updating rule is based on Bayes Theorem

$$p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) \propto \mathcal{L}(\boldsymbol{\beta}, \phi) p(\boldsymbol{\beta}, \phi)$$

reweight prior beliefs by likelihood of parameters under observed data

Posterior

Posterior is obtained by conditional distribution theory
Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)^T$

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \mathbf{Y}) &= \frac{p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, d\boldsymbol{\theta}} \\
&= \frac{p(\mathbf{Y}, \boldsymbol{\theta})}{p(\mathbf{Y})}
\end{aligned}
$$

$p(\mathbf{Y})$, the normalizing constant, is the marginal distribution of the data.

Easiest to work with Bayes Theorem in proportional form and then identify the normalizing constant.

Prior Distributions

Factor joint prior distribution

$$p(\boldsymbol{\beta}, \phi) = p(\boldsymbol{\beta} \mid \phi) p(\phi)$$

Convenient choice is to take

- $\boldsymbol{\beta} \mid \phi \sim \mathrm{N}(b_0, \Phi_0^{-1}/\phi)$ where b_0 is the prior mean and Φ^{-1}/ϕ is the prior covariance of $\boldsymbol{\beta}$
- $\phi \sim \mathbf{G}(\nu_0/2, \mathrm{SS}_0/2)$ with $\mathrm{E}(\sigma^2) = \mathrm{SS}_0/(\nu_0 - 2)$

$$p(\phi) = \frac{1}{\Gamma(\nu_0/2)} \left(\frac{\mathrm{SS}_0}{2}\right)^{\nu_0/2} \phi^{\nu_0/2-1} e^{-\phi \mathrm{SS}_0/2}$$

- $(\boldsymbol{\beta}, \phi)^T \sim \mathrm{NG}(\mathbf{b}_0, \Phi_0, \nu_0, \mathrm{SS}_0)$
- Conjugate "Normal-Gamma" family implies

$$(\boldsymbol{\beta}, \phi)^T \mid \mathbf{Y} \sim \mathrm{NG}(\mathbf{b}_n, \Phi_n, \nu_n, \mathrm{SS}_n)$$

Finding the Posterior Distribution

Express Likelihood: $\mathcal{L}(\beta, \phi) \propto \phi^{n/2} e^{-\phi \frac{\mathsf{SSE}}{2}} e^{-\frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\mathbf{X}^T \mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}$

$$
\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) \quad \propto \quad & \phi^{\frac{n+p+\nu_0}{2} - 1} e^{-\frac{\phi}{2}(\mathsf{SSE} + \mathsf{SS}_0)} \\
& e^{-\frac{\phi}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T (\mathbf{X}^T \mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})} e^{-\frac{\phi}{2}(\boldsymbol{\beta} - \mathbf{b}_0)^T \Phi (\boldsymbol{\beta} - \mathbf{b}_0)}
\end{aligned}
$$

Quadratic in Normal

$$
\exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta} - \mathbf{b})^T \Phi (\boldsymbol{\beta} - \mathbf{b})\right\} = \exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta}^T \Phi \boldsymbol{\beta} - 2\boldsymbol{\beta}^T \Phi \mathbf{b} + \mathbf{b}^T \Phi \mathbf{b})\right\}
$$

- Expand quadratics and regroup terms
- Read off posterior precision from Quadratic in $\boldsymbol{\beta}$
- Read off posterior mean from Linear term in $\boldsymbol{\beta}$
- will need to complete the quadratic in the posterior mean

Expand and Regroup

Quadratic in Normal

$$\exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b})^T\Phi(\boldsymbol{\beta}-\mathbf{b})\right\} = \exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta}^T\Phi\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\Phi\mathbf{b} + \mathbf{b}^T\Phi\mathbf{b})\right\}$$

$$\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) \quad \propto \quad & \phi^{\frac{n+p+\nu_0}{2}-1} e^{-\frac{\phi}{2}(\mathsf{SSE}+\mathsf{SS}_0)} \\
& e^{-\frac{\phi}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})} e^{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^T\Phi_0(\boldsymbol{\beta}-\mathbf{b}_0)} \\
= \quad & \phi^{\frac{n+p+\nu_0}{2}-1} e^{-\frac{\phi}{2}(\mathsf{SSE}+\mathsf{SS}_0)} \\
& e^{-\frac{\phi}{2}\left(\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X}+\Phi_0)\boldsymbol{\beta}\right)} \\
& e^{-\frac{\phi}{2}\left(-2\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\Phi_0\mathbf{b}_0)\right)} \\
& e^{-\frac{\phi}{2}(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{b}_0^T\Phi_0\mathbf{b}_0)}
\end{aligned}$$

Identify Hyperparameters and Complete the Quadratic

Quadratic in Normal

$$\exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b})^T\Phi(\boldsymbol{\beta}-\mathbf{b})\right\} = \exp\left\{-\frac{\phi}{2}(\boldsymbol{\beta}^T\Phi\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\Phi\mathbf{b} + \mathbf{b}^T\Phi\mathbf{b})\right\}$$

Let $\Phi_n = \mathbf{X}^T\mathbf{X} + \Phi_0$

$$\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) \quad &\propto \quad \phi^{\frac{n+p+\nu_0}{2}-1} e^{-\frac{\phi}{2}(\mathrm{SSE}+\mathrm{SS}_0)} \\
&\quad\quad e^{-\frac{\phi}{2}\left(\boldsymbol{\beta}^T(\mathbf{X}^T\mathbf{X}+\Phi_0)\boldsymbol{\beta}\right)} \\
&\quad\quad e^{-\frac{\phi}{2}\left(-2\boldsymbol{\beta}^T\Phi_n\Phi_n^{-1}(\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\Phi_0\mathbf{b}_0)\right)} \\
&\quad\quad e^{-\frac{\phi}{2}(\mathbf{b}_n^T\Phi_n\mathbf{b}_n-\mathbf{b}_n^T\Phi_0\mathbf{b}_n)} \\
&\quad\quad e^{-\frac{\phi}{2}(\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{b}_0^T\Phi_0\mathbf{b}_0)} \\
&= \quad \phi^{\frac{n+p+\nu_0}{2}-1} e^{-\frac{\phi}{2}(\mathrm{SSE}+\mathrm{SS}_0+\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{b}_0^T\Phi_0\mathbf{b}_0-\mathbf{b}_n^T\Phi_n\mathbf{b}_n)} \\
&\quad\quad e^{-\frac{\phi}{2}\left(\boldsymbol{\beta}^T(\Phi_n)\boldsymbol{\beta}\right)} \\
&\quad\quad e^{-\frac{\phi}{2}\left(-2\boldsymbol{\beta}^T\Phi_n\Phi_n^{-1}(\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}+\Phi_0\mathbf{b}_0)\right)} \\
&\quad\quad e^{-\frac{\phi}{2}(\mathbf{b}_n^T\Phi_n\mathbf{b}_n)}
\end{aligned}$$

Posterior Distribution

$$
\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) \quad \propto \quad & \phi^{\frac{n+\nu_0}{2}-1} e^{-\frac{\phi}{2}(\mathrm{SSE}+\mathrm{SS}_0+\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}+\mathbf{b}_0^T \Phi_0 \mathbf{b}_0-\mathbf{b}_n^T \Phi_n \mathbf{b}_n)} \\
& \phi^{\frac{p}{2}} e^{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b}_n)^T \Phi_n(\boldsymbol{\beta}-\mathbf{b}_n)}
\end{aligned}
$$

$$
\begin{aligned}
\Phi_n &= \mathbf{X}^T \mathbf{X} + \Phi_0 \\
\mathbf{b}_n &= \Phi_n^{-1}(\mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} + \Phi_0 \mathbf{b}_0)
\end{aligned}
$$

Posterior Distribution

$$
\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} \quad &\sim \quad \mathrm{N}(\mathbf{b}_n, (\phi \Phi_n)^{-1}) \\
\phi \mid \mathbf{Y} \quad &\sim \quad \mathbf{G}\left(\frac{n+\nu_0}{2}, \frac{\mathrm{SSE}+\mathrm{SS}_0+\hat{\boldsymbol{\beta}}^T \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}+\mathbf{b}_0^T \Phi_0 \mathbf{b}_0-\mathbf{b}_n^T \Phi_n \mathbf{b}_n}{2}\right)
\end{aligned}
$$

Marginal Distribution from Normal–Gamma

Theorem

Let $\boldsymbol{\theta} \mid \phi \sim N(m, \frac{1}{\phi}\Sigma)$ *and* $\phi \sim \mathbf{G}(\nu/2, \nu\hat{\sigma}^2/2)$. *Then* $\boldsymbol{\theta}$ $(p \times 1)$ *has a* p *dimensional multivariate* t *distribution*

$$\boldsymbol{\theta} \sim t_\nu(m, \hat{\sigma}^2\Sigma)$$

with density

$$p(\boldsymbol{\theta}) \propto \left[1 + \frac{1}{\nu}\frac{(\boldsymbol{\theta} - m)^T\Sigma^{-1}(\boldsymbol{\theta} - m)}{\hat{\sigma}^2}\right]^{-\frac{p+\nu}{2}}$$

Marginal density $p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \phi) p(\phi)\, d\phi$

$$
\begin{aligned}
p(\boldsymbol{\theta}) & \propto \int |\Sigma/\phi|^{-1/2} e^{-\frac{\phi}{2}(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)} \phi^{\nu/2-1} e^{-\phi\frac{\nu\hat{\sigma}^2}{2}}\, d\phi \\
& \propto \int \phi^{p/2} \phi^{\nu/2-1} e^{-\phi\frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)+\nu\hat{\sigma}^2}{2}}\, d\phi \\
& \propto \int \phi^{\frac{p+\nu}{2}-1} e^{-\phi\frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)+\nu\hat{\sigma}^2}{2}}\, d\phi \\
& = \Gamma((p+\nu)/2) \left(\frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)+\nu\hat{\sigma}^2}{2} \right)^{-\frac{p+\nu}{2}} \\
& \propto \left((\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)+\nu\hat{\sigma}^2 \right)^{-\frac{p+\nu}{2}} \\
& \propto \left(1 + \frac{1}{\nu} \frac{(\boldsymbol{\theta}-m)^T \Sigma^{-1}(\boldsymbol{\theta}-m)}{\hat{\sigma}^2} \right)^{-\frac{p+\nu}{2}}
\end{aligned}
$$

Marginal Posterior Distribution of $\boldsymbol{\beta}$

$$
\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathsf{N}(\mathbf{b}_n, \phi^{-1}\Phi_n^{-1}) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}\left(\frac{\nu_n}{2}, \frac{\mathsf{SS}_n}{2}\right)
\end{aligned}
$$

Let $\hat{\sigma}^2 = \mathsf{SS}_n/\nu_n$ (Bayesian MSE)
Then the marginal posterior distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \mathbf{Y} \sim t_{\nu_n}(\mathbf{b}_n, \hat{\sigma}^2\Phi_n^{-1})$$

Any linear combination $\lambda^T\boldsymbol{\beta}$

$$\lambda^T\boldsymbol{\beta} \mid \mathbf{Y} \sim t_{\nu_n}(\lambda^T\mathbf{b}_n, \hat{\sigma}^2\lambda^T\Phi_n^{-1}\lambda)$$

has a univariate t distribution with $\mathbf{v}_n$ degrees of freedom

Predictive Distribution

Suppose $\mathbf{Y}^* \mid \boldsymbol{\beta}, \phi \sim \mathsf{N}(\mathbf{X}^*\boldsymbol{\beta}, \mathbf{I}/\phi)$ and is conditionally independent of $\mathbf{Y}$ given $\boldsymbol{\beta}$ and ϕ

What is the predictive distribution of $\mathbf{Y}^* \mid \mathbf{Y}$?

$\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ and $\boldsymbol{\epsilon}^*$ is independent of $\mathbf{Y}$ given ϕ

$$
\begin{aligned}
\mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^* \mid \phi, \mathbf{Y} &\sim \mathsf{N}(\mathbf{X}^*\mathbf{b}_n, (\mathbf{X}^*\Phi_n^{-1}\mathbf{X}^{*T} + \mathbf{I})/\phi) \\
\mathbf{Y}^* \mid \phi, \mathbf{Y} &\sim \mathsf{N}(\mathbf{X}^*\mathbf{b}_n, (\mathbf{X}^*\Phi_n^{-1}\mathbf{X}^{*T} + \mathbf{I})/\phi) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}\left(\frac{\nu_n}{2}, \frac{\hat{\sigma}^2\nu_n}{2}\right) \\
\mathbf{Y}^* \mid \mathbf{Y} &\sim t_{\nu_n}(\mathbf{X}^*\mathbf{b}_n, \hat{\sigma}^2(\mathbf{I} + \mathbf{X}^*\Phi_n^{-1}\mathbf{X}^T))
\end{aligned}
$$

Alternative Derivation

Conditional Distribution:

$$
\begin{aligned}
f(\mathbf{Y}^* \mid \mathbf{Y}) &= \frac{f(\mathbf{Y}^*, \mathbf{Y})}{f(\mathbf{Y})} \\
&= \frac{\iint f(\mathbf{Y}^*, \mathbf{Y} \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta}, \phi) d\boldsymbol{\beta}\, d\phi}{f(\mathbf{Y})} \\
&= \frac{\iint f(\mathbf{Y}^* \mid \boldsymbol{\beta}, \phi) f(\mathbf{Y} \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta}, \phi) d\boldsymbol{\beta}\, d\phi}{f(\mathbf{Y})} \\
&= \iint f(\mathbf{Y}^* \mid \boldsymbol{\beta}, \phi) p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) d\boldsymbol{\beta}\, d\phi
\end{aligned}
$$

$$
\mathbf{Y}^* = \mathbf{X}^* \beta + \boldsymbol{\epsilon}^* \mid \mathbf{Y}, \phi \sim \mathrm{N}(\mathbf{X}^* \mathbf{b}_n, \phi^{-1}(\mathbf{I} + \mathbf{X}^* \Phi_n \mathbf{X}^{*T}))
$$

Use result about Marginals of Normal-Gamma family to integrate out ϕ

Conjugate Priors

Definition

A class of prior distributions $\mathcal{P}$ for $\boldsymbol{\theta}$ is conjugate for a sampling model $p(y \mid \boldsymbol{\theta})$ if for every $p(\boldsymbol{\theta}) \in \mathcal{P}$, $p(\boldsymbol{\theta} \mid \mathbf{Y}) \in \mathcal{P}$.

Advantages:

- Closed form distributions for most quantities; bypass MCMC for calculations
- Simple updating in terms of sufficient statistics "weighted average"
- Interpretation as prior samples - prior sample size
- Elicitation of prior through imaginary or historical data
- limiting "non-proper" form recovers MLEs

Choice of conjugate prior?

Unit Information Prior

Unit information prior $\beta \mid \phi \sim \mathsf{N}(\hat{\beta}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

- Fisher Information is $\phi\mathbf{X}^T\mathbf{X}$ based on a sample of n observations
- Inverse Fisher information is covariance matrix of MLE
- "average information" in one observation is $\phi\mathbf{X}^T\mathbf{X}/n$
- center prior at MLE and base covariance on the information in "1" observation
- Posterior mean

$$\frac{n}{1+n}\hat{\beta} + \frac{1}{1+n}\hat{\beta} = \hat{\beta}$$

- Posterior Distribution

$$\beta \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\hat{\beta}, \frac{n}{1+n}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

Cannot represent real prior beliefs; double use of data

Zellner's g-prior

Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

$$\boldsymbol{\beta} \mid \mathbf{Y}, \phi \sim \mathsf{N}\left(\frac{g}{1+g}\hat{\boldsymbol{\beta}} + \frac{1}{1+g}\mathbf{b}_0, \frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\phi^{-1}\right)$$

- Invariance: Require posterior of $\mathbf{X}\boldsymbol{\beta}$ equal the posterior of $\mathbf{X}\mathbf{H}\boldsymbol{\alpha}$ ($\mathbf{a}_0 = \mathbf{H}^{-1}\mathbf{b}_0$) (take $\mathbf{b}_0 = \mathbf{0}$)
- Choice of g?
- $\frac{g}{1+g}$ weight given to the data
- Fixed g effect does not vanish as $n \to \infty$
- Use $g = n$ or place a prior diistribution on g

Shrinkage

Posterior mean under g-prior with $\mathbf{b}_0 = 0$ $\frac{g}{1+g}\hat{\beta}$