# Bayesian Estimation in Linear Models


STA521 Linear Models Duke University

Merlise Clyde


February 19, 2017

# 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

- $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(b_0, \Phi_0^{-1}/\phi)$ where $b_0$ is the prior mean and $\Phi^{-1}/\phi$ is the prior covariance of $\boldsymbol{\beta}$

$$p(\boldsymbol{\beta} \mid \phi) = (2\pi)^{-p/2} \phi^{p/2} |\Phi_0|^{1/2} \exp\left(-\frac{\phi}{2}(\boldsymbol{\beta} - b_0)^T \Phi_0 (\boldsymbol{\beta} - b_0)\right)$$

- $\phi \sim \mathbf{G}(\nu_0/2, \mathsf{SS}_0/2)$ with $\mathsf{E}(\sigma^2) = \mathsf{SS}_0/(\nu_0 - 2)$

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

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

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

# Posterior Distribution

$$
\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) &\propto \mathcal{L}(\boldsymbol{\beta}, \phi) p(\boldsymbol{\beta} \mid \phi) p(\phi) \\
&= c_Y \, \phi^{n/2} e^{-\frac{\phi}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})} \\
& \quad c_{\boldsymbol{\beta}} \left|\phi \Phi_0\right|^{1/2} e^{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^T \Phi_0 (\boldsymbol{\beta}-\mathbf{b}_0)} \\
& \quad c_\phi \, \phi^{\frac{nu_0}{2}-1} e^{-\phi \mathrm{SS}_0/2}
\end{aligned}
$$

Write $\Phi_0 = \mathbf{X}_0^T \mathbf{X}_0$
(by Spectral Theorem $\Phi = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T$ take $\mathbf{X}_0 = \boldsymbol{\Lambda}^{1/2} \mathbf{U}^T$)

$$
\begin{aligned}
p(\boldsymbol{\beta} \mid \phi) &\propto \phi^p/2 e^{-\frac{\phi}{2}(\boldsymbol{\beta}-\mathbf{b}_0)^T \mathbf{X}_0^T \mathbf{X}_0 (\boldsymbol{\beta}-\mathbf{b}_0)} \\
&= \phi^{p/2} e^{-\frac{\phi}{2}(\mathbf{X}_0\boldsymbol{\beta}-\mathbf{X}_0\mathbf{b}_0)^T(\mathbf{X}_0\boldsymbol{\beta}-\mathbf{X}_0\mathbf{b}_0)}
\end{aligned}
$$

Looks like the likelihood from a prior sample of size $p$ with prior design $\mathbf{X}_0$ and $\mathbf{Y}_0 = \mathbf{X}_o \mathbf{b}_0$

# Prior Data

Let

$$\tilde{\mathbf{Y}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}_0 \end{bmatrix} \in \mathbb{R}^{n+p} \qquad \tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{Y} \\ \mathbf{X}_0 \end{bmatrix} (n + p \times p)$$

then

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

where $\tilde{\mathbf{Y}} \sim \mathrm{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \mathbf{I}_{n+p}/\phi)$

Likelihood proportional to sampling model of "augmented" data

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

# Decompose

Let $\mathrm{P}_{\tilde{\mathbf{X}}} = \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T$ - this is an orthogonal projection on $C(\tilde{\mathbf{X}})$

$$\tilde{\mathbf{Y}}^T\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}^T\mathrm{P}_{\tilde{\mathbf{X}}}\tilde{\mathbf{Y}} + \tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathrm{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}}$$

$$\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) &\propto \phi^{\frac{n+p}{2}} e^{-\frac{\phi}{2}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})} \phi^{\frac{nu_0}{2}-1} e^{-\phi\frac{\mathrm{SS}_0}{2}} \\
&= \phi^{\frac{n+p}{2}} e^{-\frac{\phi}{2}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T\mathrm{P}_{\tilde{\mathbf{X}}}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})} \\
&\quad e^{-\frac{\phi}{2}(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathbf{I}_{n+p}-\mathrm{P}_{\tilde{\mathbf{X}}})(\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})} \\
&\quad \phi^{\frac{nu_0}{2}-1} e^{-\phi\frac{\mathrm{SS}_0}{2}} \\
&= \phi^{\frac{p}{2}} e^{-\frac{\phi}{2}(\mathrm{P}_{\tilde{\mathbf{X}}}\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})^T(\mathrm{P}_{\tilde{\mathbf{X}}}\tilde{\mathbf{Y}}-\tilde{\mathbf{X}}\boldsymbol{\beta})} \\
&\quad \phi^{\frac{n+nu_0}{2}-1} e^{-\frac{\phi}{2}\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p}-\mathrm{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}}} e^{-\phi\frac{\mathrm{SS}_0}{2}}
\end{aligned}$$

# Continued

$\mathrm{P}_{\tilde{\mathbf{X}}}\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\tilde{\beta}$ Maximum a Posteriori (MAP) estimator

$$\begin{aligned}
p(\boldsymbol{\beta}, \phi \mid \mathbf{Y}) &\propto \phi^{\frac{p}{2}} e^{-\frac{\phi}{2}(\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\boldsymbol{\beta})^T(\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} - \tilde{\mathbf{X}}\boldsymbol{\beta})} \\
&\quad \phi^{\frac{n+\nu_0}{2}-1} e^{-\phi \frac{\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathrm{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}} + \mathrm{SS}_0}{2}} \\
&\propto \phi^{\frac{p}{2}} e^{-\frac{\phi}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})} \, \phi^{\frac{n+\nu_0}{2}-1} e^{-\phi \frac{\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathrm{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}} + \mathrm{SS}_0}{2}}
\end{aligned}$$

Read off distributions!

$$\begin{aligned}
\boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}(\tilde{\boldsymbol{\beta}}, (\phi \tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1}) \\
\phi \mid \mathbf{Y} &\sim \mathbf{G}\left( \frac{n+\nu_0}{2}, \frac{\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathrm{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}} + \mathrm{SS}_0}{2} \right)
\end{aligned}$$

# Simplify Distribution of $\boldsymbol{\beta}$

Posterior Precision $\phi\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$

$$\begin{aligned}
\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} &= \mathbf{X}^T\mathbf{X} + \mathbf{X}_0^T\mathbf{X}_0 \\
&= \mathbf{X}^T\mathbf{X} + \Phi_0 \\
\Phi_n &= \mathbf{X}^T\mathbf{X} + \Phi_0
\end{aligned}$$

sum of precision from data plus prior precision

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

Conditional Posterior for $\boldsymbol{\beta} \mid \phi, \mathbf{Y} \sim \mathsf{N}(\mathbf{b}_n, \phi^{-1}\Phi_n^{-1})$

# Simplify Posterior Distribution for $\phi$

$$\phi \mid \mathbf{Y} \sim \mathbf{G}\left(\frac{n+\nu_0}{2}, \frac{\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathsf{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}} + \mathsf{SS}_0}{2}\right)$$

$$\begin{aligned}
\tilde{\mathbf{Y}}^T(\mathbf{I}_{n+p} - \mathsf{P}_{\tilde{\mathbf{X}}})\tilde{\mathbf{Y}} &= \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\mathbf{b}_n\|^2 \\
&= \|\mathbf{Y} - \mathbf{X}\mathbf{b}_n\|^2 + \|\mathbf{Y}_0 - \mathbf{X}_0\mathbf{b}_n\|^2 \\
&= \|\mathbf{Y} - \mathbf{X}\mathbf{b}_n\|^2 + (\mathbf{X}_0\mathbf{b}_0 - \mathbf{X}_0\mathbf{b}_n)^T(\mathbf{X}_0\mathbf{b}_0 - \mathbf{X}_0\mathbf{b}_n) \\
&= \|\mathbf{Y} - \mathbf{X}\mathbf{b}_n\|^2 + (\mathbf{b}_0 - \mathbf{b}_n)^T(\mathbf{X}_o^T\mathbf{X}_0)(\mathbf{b}_0 - \mathbf{b}_n) \\
&= \|\mathbf{Y} - \mathbf{X}\mathbf{b}_n\|^2 + (\mathbf{b}_0 - \mathbf{b}_n)^T\Phi_o(\mathbf{b}_0 - \mathbf{b}_n) \\
\nu_n &= n + \nu_0
\end{aligned}$$

Posterior for $\phi$

$$\phi \mid \mathbf{Y} \sim \mathbf{G}\left(\frac{\nu_n}{2}, \frac{\|\mathbf{Y} - \mathbf{X}\mathbf{b}_n\|^2 + (\mathbf{b}_0 - \mathbf{b}_n)^T\Phi_o(\mathbf{b}_0 - \mathbf{b}_n) + \mathsf{SS}_0}{2}\right)$$

## 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*

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

# Derivation

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 $\boldsymbol{\beta}$

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

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

# Posterior Distribution of $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$

- Use a limiting prior distribution: $p(\boldsymbol{\beta}, \phi) \propto \phi^{-1}$

$$b_0 = 0, \mathrm{SS}_0 = 0, \Phi_0 = 0, \nu_0 = -(p+1)$$

- joint posterior distribution depends on MLE's

$$\begin{aligned} \boldsymbol{\beta} \mid \phi, \mathbf{Y} &\sim \mathrm{N}\left(\hat{\boldsymbol{\beta}}, \phi^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\right) \\ \phi \mid \mathbf{Y} &\sim \mathbf{G}((n-p-1)/2, (n-p-1)\hat{\sigma}^2/2) \\ \mu_i \mid \mathbf{Y} &\sim t_{n-p-1}(\mathbf{x}_i^T\hat{\boldsymbol{\beta}}, \hat{\sigma}^2\mathbf{x}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}_i) \end{aligned}$$

- Same distribution as frequentists:

$$P(\mu \in (\mathbf{x}_i^T\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}^2 h_{ii}) \mid \mathbf{Y}) = 1 - \alpha$$

# Predictive Distribution

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

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 $\phi$

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

Under limiting prior same as frequentist predictive interval

# 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

## Some Default Choices

- Independent Jeffreys Prior $p(\boldsymbol{\beta}, \phi) \propto 1/\phi$
- Unit information prior $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\hat{\boldsymbol{\beta}}, n(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$
- Zellner's g-prior(s) $\boldsymbol{\beta} \mid \phi \sim \mathsf{N}(\mathbf{b}_0, g(\mathbf{X}^T\mathbf{X})^{-1}/\phi)$

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

- Independent Priors on $\beta_j$

$$\beta_j \mid \phi, \lambda_j \overset{\text{ind}}{\sim} N(0, s_j^2/(\phi\lambda_j))$$

# Disadvantages

Disadvantages:

- Results may have be sensitive to prior “outliers” due to linear updating
- Cannot capture all possible prior beliefs

# Mixtures of Conjugate Priors

## Theorem (Diaconis & Ylivisaker 1985)

*Given a sampling model $p(y \mid \boldsymbol{\theta})$ from an exponential family, any prior distribution can be expressed as a mixture of conjugate prior distributions*

- Prior $p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta} \mid \omega) p(\omega)\, d\omega$
- Posterior

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

# Next Class After Midterm

Review Hoff Chapters 1-9, Gelman-Hill Chapter 18