Bayesian Inference in GLMs

Frequentists typically base inferences on MLEs, asymptotic confidence limits, and log-likelihood ratio tests.

Bayesians base inferences on the posterior distribution of the unknowns of interest.

The posterior quantifies the current state of knowledge about the unknowns, incorporating subject matter knowledge, past data, and the current data.

Expert knowledge and past data are incorporated through a prior distribution, while the current data are incorporated through the likelihood function.
Bayesian inference in linear regression models

- **Likelihood:**

\[
L(y; x, \beta, \tau) = \prod_{i=1}^{n} (2\pi\tau^{-1})^{-1/2} \exp \left\{ -\frac{\tau}{2} (y_i - x_i'\beta)^2 \right\},
\]

where \( \tau = \sigma^{-2} \) is the error precision

- **Commonly-used prior:**

\[
\pi(\beta, \tau) = \pi(\beta)\pi(\tau) = N_p(\beta; \beta_0, \Sigma_0)G(\tau; a_\tau, b_\tau).
\]

Here, the \( p \times 1 \) vector of regression coefficients are assigned a multivariate normal prior:

\[
\pi(\beta) = |2\pi\Sigma_0|^{-p/2} \exp \left\{ -\frac{1}{2}(\beta - \beta_0)'\Sigma_0^{-1}(\beta - \beta_0) \right\},
\]

where \( \beta_0 \) is the prior mean and \( \Sigma_0 \) is the prior covariance
The *hyperparameters* $\beta_0, \Sigma_0$ quantify our state of knowledge about the regression parameters $\beta$ prior to observing the data from the current study.

In particular, $\beta_0$ is our best guess for $\beta$ before looking at the current data & $\Sigma_0$ expresses uncertainty in this guess.
The prior for the error precision follows the gamma density

\[ \pi(\tau) = \frac{b_\tau^{a_\tau}}{\Gamma(a_\tau)} \tau^{a_\tau-1} \exp(-b_\tau \tau), \]

which has expectation \( E(\tau) = a_\tau / b_\tau \) and variance \( V(\tau) = a_\tau / b_\tau^2 \).

Hyperparameters \( a_\tau, b_\tau \) are chosen to express knowledge about \( \tau \).

A non-informative prior is achieved by letting \( \pi(\beta, \tau) \propto \tau^{-1} \)

(in limit as prior variance \( \rightarrow \infty \))
Bayesian updating & posterior distributions

- After specifying the prior, we update the prior to incorporate information in the likelihood using Bayes rule.

- This updating process yields the posterior distribution:

\[
\pi(\beta, \tau | y, x) = \frac{\pi(\beta, \tau) L(y; x, \beta, \tau)}{\int \pi(\beta, \tau) L(y; x, \beta, \tau) d\beta d\tau} = \frac{\pi(\beta, \tau) L(y; x, \beta, \tau)}{\pi(y; x)}
\]

where \(\pi(y; x)\) is the marginal likelihood of the data (obtained by integrating the likelihood across the prior for the parameters).

- The expression for the posterior holds for any GLM, letting \(\tau\) denote the scale parameter and \(L(y; x, \beta, \tau)\) the exponential family likelihood.
• The primary difference between Bayesian inference in normal linear regression models and in other GLMs is computational

• In particular, for normal linear regression models with *conjugate priors*, the posterior distribution has a simple form.

• The posterior for the regression coefficients can be derived as follows: $\pi(\mathbf{\beta} \mid \mathbf{y}, \mathbf{x}, \tau)$

\[
\propto \pi(\mathbf{\beta}) L(\mathbf{y}; \mathbf{x}, \beta, \tau) \\
\propto \exp \left\{ -\frac{1}{2} (\mathbf{\beta} - \mathbf{\beta}_0)' \Sigma_0^{-1} (\mathbf{\beta} - \mathbf{\beta}_0) \right\} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} \tau (y_i - x'_i \mathbf{\beta})^2 \right\} \\
\propto \exp \left[ -\frac{1}{2} \left\{ \mathbf{\beta}' (\Sigma_0^{-1} + \tau \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}'_i) \mathbf{\beta} - 2 \mathbf{\beta}'(\mathbf{\beta}_0 + \tau \sum_{i=1}^{n} \mathbf{x}_i y_i) \right\} \right] \\
\propto \mathcal{N}_p(\mathbf{\beta}; \hat{\mathbf{\beta}}, \hat{\Sigma}_\beta),
\]

6
• Thus, the posterior distribution of $\beta$ given $\tau$ is multivariate normal

• The posterior mean is

$$\hat{\beta} = E(\beta \mid \tau, y, X) = \hat{\Sigma}_\beta (\Sigma_0^{-1} \beta_0 + \tau X'y)$$

• The posterior variance is

$$\hat{\Sigma}_\beta = V(\beta \mid \tau, y, X) = (\Sigma_0^{-1} + \tau X'X)^{-1}.$$  

• Note that in the limiting case as the prior variance increases, 

$$\hat{\beta} \to (X'X)^{-1}X'y$$, which is simply the least squares estimator or MLE

• Hence, the posterior mean is shrunk back towards the prior mean $\beta_0$ to a degree dependent on the prior variance.
• We can similarly derive the posterior distribution of $\tau$: $\pi(\tau | \mathbf{y}, \mathbf{X}, \mathbf{\beta})$:

$$
\propto \pi(\tau) L(\mathbf{y}; \mathbf{X}, \mathbf{\beta}, \tau)
$$

$$
\propto \tau^{a_\tau - 1} \exp(-b_\tau \tau) (\tau^{-1})^{-n/2} \exp \left\{ -\frac{\tau}{2} (y_i - x_i' \mathbf{\beta})^2 \right\}
$$

$$
\propto \tau^{a_\tau + n/2 - 1} \exp \left[ -\tau \left\{ b_\tau + \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i' \mathbf{\beta})^2 \right\} \right]
$$

$$
\propto \mathcal{G}(\tau; a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i' \mathbf{\beta})^2).
$$

• Thus, the error precision has a gamma posterior distribution
Example: Bayesian updating

- Suppose for example that we have a simple linear regression model

\[ y_i = \beta_0 + \beta_1 \text{dose}_i + \epsilon_i, \epsilon_i \sim N(0, 1) \]

- We simulate data under the true model: \( \beta = (-1, 2), n = 25, \text{dose}_i \sim \text{Uniform}(0, 1) \)

- We consider priors \( \pi(\beta) \propto 1 & \pi(\beta) = N(0, I_2) \)
slope parameter (beta_2)
density
−2 0 2 4
0.0 0.2 0.4 0.6
N(0,1) prior
Posterior, uniform prior
Posterior, N(0,1) prior
Some comments

- For a non-informative, uniform prior posterior is centered on the least squares estimator (specific to normal linear models)

- For an informative prior, posterior mean is shrunk back towards prior mean and posterior variance decreases

- As sample size increases, the contribution of the prior is swamped out by the likelihood
• Hence, as $n \to \infty$, the posterior will be centered on the MLE regardless of the prior & frequentist/Bayes inferences will be similar

• However, for finite samples, there can be substantial differences

• Choosing a $N(0,I)$ prior results in a type of *shrinkage estimator*
• Since the result of Stein (1955) and James & Stein (1960) (MLE is inadmissible for $p \geq 3$), shrinkage estimators have been very popular

• Choosing a $N(0, \kappa I_p)$ prior for $\beta$, results in a \textit{ridge regression} (Hoerl and Kennard, 1970) estimator

• Hence, priors for the regression parameters having diagonal covariance are commonly referred to as \textit{ridge regression priors}.

• For a recent article on shrinkage estimators and properties, refer to Maruyama & Strawderman (2005, \textit{Annals of Statistics} 33, 1753-1770).
• This and other articles have shown excellent properties for Bayes and empirical Bayes estimators

• Schools of thought: subjective Bayesians argue that the prior should be chosen based on one’s own belief and state of knowledge about the unknowns

• Objective Bayesians argue that priors should be chosen (at least in cases in which real substantive information is not available) to yield good frequentist and Bayesian properties

• empirical Bayes instead uses the data in estimating the prior distribution - this would be cheating for a fully Bayes procedure!
Bayesian inference in GLMs

Suppose, we have a GLM: $\eta_i = g(\mu_i) = x_i' \beta$

To complete a Bayesian specification of the GLM, we need to choose a prior density for the parameters $(\beta, \phi)$, $\pi(\beta, \phi)$. 
The posterior density is then expressed as:

$$
\pi(\beta, \phi \mid y) = \frac{L(y; X, \beta, \phi) \pi(\beta, \phi)}{\int L(y; X, \beta, \phi) \pi(\beta, \phi) d\beta d\phi} = \frac{L(y; X, \beta, \phi) \pi(\beta, \phi)}{\pi(y; X)} \propto f(y; X, \beta, \phi) \pi(\beta, \phi)
$$

$$
= \exp \left[ \sum_{i=1}^{n} \{ y_i \theta_i - b(\theta_i) \} / a(\phi) + c(y_i, \phi) \right] \pi(\beta, \phi)
$$

where $$\pi(y; X)$$ is the marginal likelihood of the data, obtained by integrating the likelihood conditional on the unknown regression coefficients $$\beta$$ and dispersion parameter $$\phi$$ across the prior density.
Unlike in the normal linear regression case, there is typically no simple form for the posterior distribution of $\beta, \phi$

Hence, Bayesian inference has historically relied on approximations

For example, a large sample approximation would replace the exact, exponential family likelihood with a normal approximation

In particular, utilizing the CAN property of the MLE for $\beta$, we could use $\beta \overset{asy}{\sim} N(\hat{\beta}, I(\hat{\beta})^{-1})$ [refer to page 25 of lecture 3 notes] [here $\hat{\beta}$ is the MLE]
We then could obtain the following approximation to the posterior assuming a $\text{N}(\beta_0, \Sigma_0)$ prior for $\beta$:

$$
\pi(\beta \mid y, X) \approx \exp \left\{ -\frac{1}{2}(\beta - \beta_0)\Sigma_0^{-1}(\beta - \beta_0) \right\} \\
\times \exp \left\{ -\frac{1}{2}(\beta - \hat{\beta})^\prime I(\hat{\beta})(\beta - \hat{\beta})' \right\} \\
\propto \text{N}(\beta; \tilde{\beta}, \tilde{\Sigma}_\beta)
$$

where the approximate posterior variance is

$$
\tilde{\Sigma}_\beta = V(\beta \mid y, X) = (\Sigma_0^{-1} + I(\hat{\beta}))^{-1}
$$

In addition, the approximate posterior mean is

$$
\tilde{\beta} = E(\beta \mid y, X) = \tilde{\Sigma}(\Sigma_0^{-1}\beta_0 + I(\hat{\beta})\hat{\beta}).
$$
Recall, that frequentist inferences were based on the asymptotic result $\beta - \hat{\beta} \overset{asy}{\sim} N(0, I(\hat{\beta})^{-1})$

Thus, we can clearly see the role of the prior here & that Bayes/frequentist inferences tend to be very similar in large samples

Note that by using simulation, we will not need to rely on asymptotic justifications
Some Advantages of the Bayesian Approach:

1. Confidence limits and posterior probabilities are more intuitive:
   
   • **P-value**: probability under $H_0$ of data at least as extreme as that actually observed (What?).
   
   • **Posterior probability**: probability of $H_0$ given the current data and outside information.
• **95% Confidence Interval**: in repeated sampling the interval will contain the true parameter approximately 95% of time.

• **95% Credible Interval**: ranges from 2.5th to 97.5th percentile of posterior density, so that the true parameter falls within this interval with 95% probability
2. Provides natural framework for formalizing the process of learning from the current data to obtain updated beliefs.

3. Flexible in the incorporation of historical data and outside information (e.g., order restrictions, knowledge of plausible range for parameter, etc).
4. Exact posterior distributions can be estimated using Markov chain Monte Carlo (MCMC) - does not rely on asymptotic normality as does MLE-based inferences.

5. Since MCMC methods are so general, more realistic models can be formulated without as many computational problems.
Homework Exercise (Turn in Next Tuesday):

1. Simulate data from $x_i \sim Unif(0, 1)$ and $y_i \sim N(-1 + 2x_i, 1)$, for $i = 1, \ldots, n = 20$.

2. Fit the data using maximum likelihood estimation and do hypothesis tests for whether dose is important.

3. Alternately sample from the conditional posterior of $\beta$ given $\tau$ and $\tau$ given $\beta$ for a large number of iterations.

4. Plot the samples, discard the first 100 iterations, and calculate summary statistics.

5. How do these statistics vary from the MLEs?
Example: Predictors of length of gestation

- Pregnancy, Infection and Nutrition (PIN) study is an epidemiologic study of predictors of pregnancy outcomes.
- One of the primary outcomes is preterm birth ($y_i = 1$ if delivery prior to 37 weeks, $y_i = 0$ otherwise).
- Loss of information in dichotomizing - we have information on the length of gestation in weeks.
- Instead let $y_i$ denote the number of weeks of gestation for woman $i$. 


Empirical gestational age distribution in PIN study
Some Comments

• We are interested in assessing how this distribution changes with predictors, such as bleeding, maternal age, exposures, stress during pregnancy, etc.

• How to analyze with a frequentist generalized linear model?

• Should we log transform $y_i$ and use a linear regression model?

• What if some of the women drop out of the study prior to delivery, but we have information that their gestational age was at least $t_i$?
Discrete Time to Event Data

• Gestational length is a type of time to event (or survival) data

• Because we do not know the exact times at delivery, but instead have the week number, the data are discrete

• Discrete hazard (letting $t_i \in \{1, \ldots, T\}$ denote event time):

\[
\Pr(t_i = t \mid t_i \geq t) = \lambda_t
\]

$\lambda_t \in [0, 1] = \text{discrete hazard of event at time } t$
Note that we can get back to the probability mass function as follows:

\[
Pr(t_i = t) = \left\{ \prod_{s=1}^{t-1} (1 - Pr(t_i = s \mid t_i \geq s)) \right\} \Pr(t_i = t \mid t_i \geq t) \\
= \left\{ \prod_{s=1}^{t-1} (1 - \lambda_s) \right\} \lambda_t
\]

where the term in \{\cdot\} is the probability of surviving up to \(t\).