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.

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

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:

\[
\pi(\beta, \phi \mid y) = \frac{f(y; \beta, \phi) \pi(\beta, \phi)}{\int f(y; \beta, \phi) \pi(\beta, \phi) d\beta d\phi} = \frac{f(y; \beta, \phi) \pi(\beta, \phi)}{\pi(y)}
\]

\[
\propto f(y; \beta, \phi) \pi(\beta, \phi)
\]

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

where \(\pi(y)\) 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.
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.
Normal Linear Model:

Suppose \( y_i \sim N(x'_i \beta, \phi^{-1}) \) and we choose the prior \( \pi(\beta, \phi) = \pi(\beta)\pi(\phi) \), with \( \pi(\beta) \overset{d}{=} N(\beta_0, \Sigma_0) \) and \( \pi(\phi) = G(a_0, b_0) \), then the posterior density of \( \beta \) is \( \pi(\beta \mid \phi, y, X) \)

\[
\propto N(\beta; \beta_0, \Sigma_0) \prod_{i=1}^{n} N(y_i; x'_i \beta, \phi^{-1}) \\
\propto \exp\left[ -\frac{1}{2}\left\{ (\beta - \beta_0)' \Sigma_0^{-1}(\beta - \beta_0) + \phi(y - X\beta)'(y - X\beta) \right\} \right] \\
\propto \exp\left[ -\frac{1}{2}\left\{ \beta' \Sigma_0^{-1} \beta - 2\beta' \Sigma_0^{-1} \beta_0 + \phi \beta' X' \beta - 2\phi \beta' X' y \right\} \right] \\
= \exp\left[ -\frac{1}{2}\left\{ \beta'(\Sigma_0^{-1} + \phi X' X) \beta - 2\beta'(\Sigma_0^{-1} \beta_0 + \phi X' y) \right\} \right] \\
= \exp\left[ -\frac{1}{2}\left\{ \beta' \Sigma_{\beta}^{-1} \beta - 2\beta' \Sigma_{\beta}^{-1} \hat{\beta} \right\} \right] \\
\propto N(\beta; \hat{\beta}, \Sigma_{\beta}),
\]

where \( \Sigma_{\beta} = (\Sigma_0^{-1} + \phi X' X)^{-1} \) is the posterior covariance and \( \hat{\beta} = \Sigma_{\beta}(\Sigma_0^{-1} \beta_0 + \phi X' y) \) is the posterior mean.
Homework Exercise (Turn in Next Monday):

1. Derive the posterior density $\pi(\phi | \beta, y, X)$.

2. Simulate $y_i \sim N(-1 + x_i, 1)$, for $i = 1, \ldots, 100$ and $x_i \sim N(0, 1)$.

3. Choose the priors, $(\beta_1, \beta_2) \sim N(0, diag(10, 10))$ and $\phi \sim G(0.01, 0.01)$.

4. Starting at the prior means, alternately sample from

   (a) $\pi(\beta | \phi, y, X)$

   (b) $\pi(\phi | \beta, y, X)$.

5. Plot iterations 1-1000 for $\beta_1, \beta_2, \phi$.

6. Comment on convergence & provide posterior summaries of the parameters.
Sampling-Based Approaches to Bayesian Estimation

- In most problems, posterior distribution is not available in closed form and standard integral approximations can perform poorly.

- Calculation of the posterior density typically involves high dimensional integration, with no analytic solution available.

- **Sampling approach:**

  1. Construct an algorithm for simulating a long chain of draws from the posterior distribution.

  2. Base inferences on posterior summaries of the parameters or functionals of the parameters calculated from the samples.
Markov chain Monte Carlo (MCMC) Algorithms

- **Gibbs Sampler** (see Casella and George, 1992):
  - Repeatedly samples each parameter from its full conditional posterior distribution given the current values of the other parameters.
  - Under some regularity conditions (Gelfand and Smith, 1990), samples converge to a stationary distribution that is the joint posterior distribution.
  - Requires algorithm for sampling from full conditional distributions.
• Metropolis-Hastings Algorithm (see Chib and Greenberg, 1995):

  – Sample a candidate for a parameter from a candidate generating density (e.g., normal centered on the previous value of the parameter).

  – Accept the candidate with probability equal to the minimum of one and the ratios of the posterior probabilities at the new and old values of the parameter multiplied by a correction for asymmetric candidate generating densities.

  – Repeat for all the parameters and for a large # of iterations.

  – Does not require closed form full conditionals but efficiency can be strongly dependent on choice of candidate generating density & tuning may be needed.
Bayesian Analyses of Binary Response Models

Define a binary regression model as

\[ p_i = \Pr(y_i = 1 \mid x_i, \beta) = h(x_i' \beta), \]

where \( h(\cdot) \) is a known cdf

Let \( \pi(\beta) \) be a prior density for the regression coefficients, \( \beta \).

Then the posterior density of \( \beta \) is given by

\[
\pi(\beta \mid \text{data}) = \frac{\pi(\beta) \Pi_{i=1}^{n} h(x_i' \beta)^{y_i} \{1 - h(x_i' \beta)\}^{1-y_i}}{\int \pi(\beta) \Pi_{i=1}^{n} h(x_i' \beta)^{y_i} \{1 - h(x_i' \beta)\}^{1-y_i} d\beta},
\]
Typically, the integral in the denominator of the above expression is intractable, but one can use an asymptotic approximation.

In particular, we can use the approximation that

\[ \pi(\beta \mid \text{data}) \approx d N(\hat{\beta}, I(\hat{\beta})^{-1}), \]

where \( \hat{\beta} \) is the posterior mode and \( I(\hat{\beta}) \) is the negative of the second derivative matrix evaluated at the mode.

When the improper uniform prior, \( \pi(\beta) \propto 1 \), is chosen, \( \hat{\beta} \) is the MLE and \( I() \) is the observed information matrix.

The normal approximation (which you may note is typically the basis for frequentist inference on binary response glms) is often biased for small samples.
Various Markov chain Monte Carlo (MCMC) approaches have been proposed for posterior computation of binary response GLMs.

The most commonly used algorithms are the Gibbs sampler, implemented using Adaptive Rejection Sampling, and (for probit models) the data augmentation Gibbs sampler of Albert and Chib (1993).

The Metropolis-Hastings algorithm can be used in general, but requires choice of a candidate-generating (i.e., proposal density) and possibly tuning of this density. The choice of density can greatly affect efficiency and is not necessarily straightforward.
Gibbs Sampling via Adaptive Rejection Sampling

(ARS; Gilks and Wild, 1992; Dellaportas and Smith, 1993)

ARS is a general algorithm for sampling from log-concave densities that do not have a closed form.

If the likelihood and prior are log-concave with respect to each of the regression parameters, then the ARS algorithm can be used to sample from the full conditional posteriors of each parameter given the other parameters and data.

Hence, in such cases, the ARS algorithm can be used to implement Gibbs sampling.

Most commonly-used glms have the log concavity property when log-concave priors are chosen (basis of WinBUGS software).
Algorithm:

1. Choose a prior density and initial values for $\beta$

2. For $j = 1, \ldots, p$, draw a value from the full conditional posterior density,

$$\pi(\beta_j | \beta_{(j)}, \text{data}),$$

where $\beta_{(j)} = \{\beta_k : k \neq j, k = 1, \ldots, p\}$. Although this conditional distribution is not available in closed form, except in special cases, we can sample from this distribution using ARS.

3. Repeat step 2 for a large number of iterations. Discard an initial burn-in period to allow convergence to a stationary distribution (which is the joint posterior density under mild regularity conditions). Calculate summaries of the posterior of $\beta$ based on a large number of additional draws.
Rejection Sampling

General method for sampling from a density $f(x)$, which is possibly unnormalized ($g(x) = cf(x)$).

1. Define an envelope function, $g_u(x)$ such that $g_u(x) \geq g(x)$ for all $x \in D$.

2. (Optionally) define a squeezing function $g_1(x)$ such that $g_1(x) \leq g(x)$ for all $x \in D$.

3. Repeat the following sampling step until $n$ samples have been accepted: $x^* \sim g_u$ and $w \sim U(0, 1)$. If you have a squeezing function, then accept $x^*$ if

$$w \leq g_1(x^*)/g_u(x^*).$$

If you don’t have a squeezing function, then accept $x^*$ if

$$w \leq g(x^*)/g_u(x^*).$$
Rejection sampling is useful when it is easy to sample from $g_u$, but not from $f$ directly.

Unfortunately, it can be difficult to find suitable envelope and squeezing functions in practice.
Adaptive Rejection Sampling

Reduces the number of evaluations of $g(x)$ by

1. Assumes log-concavity, for $h(x) = \log g(x)$,

   $$h'(x) = \frac{dh(x)}{dx} \text{ decreases with increasing } x \in D,$$

   to avoid the need to identify the $\sup\{g(x) : x \in D\}$.

2. After each rejection, the envelope and squeezing functions are updated to incorporate the new info about $g(x)$. 