Marginal and conditional distributions

With a flat prior on $\beta$ we obtain, after some algebra, the conditional posterior distribution:

$$p(\beta | \sigma^2, y) \propto N(\beta | (X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1})$$

The conditional posterior distribution of $\beta$ would have been the desired posterior distribution had $\sigma^2$ been known.

Since that is not the case, we need to obtain the marginal posterior distribution by integrating out $\sigma^2$ as:

$$p(\beta | y) = \int p(\beta | \sigma^2, y) p(\sigma^2 | y) d\sigma^2$$

Can we solve this integration using composition sampling? YES: if we can generate samples from $p(\sigma^2 | y)$!

bayesian regression models: a bayesian perspective

Linear regression models: a Bayesian perspective

For Bayesian analysis, we will need to specify priors for the unknown regression parameters $\beta$ and the variance $\sigma^2$.

Consider independent flat priors on $\beta$ and $\log \sigma^2$:

$$p(\beta) \propto 1 \quad \text{and} \quad p(\log(\sigma^2)) \propto 1$$

or equivalently

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?

It turns out that even if the priors are improper (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

Ingredients of a linear model include an $n \times 1$ response vector $y = (y_1, \ldots, y_n)^T$ and an $n \times p$ design matrix (e.g. including regressors) $X = [x_1, \ldots, x_n]$, assumed to have been observed without error. The linear model:

$$y = X\beta + \epsilon; \quad \epsilon \sim N(0, \sigma^2 I)$$

The linear model is the most fundamental of all serious statistical models encompassing:

- ANOVA: $y$ is continuous, $x_i$‘s are categorical
- REGRESSION: $y$ is continuous, $x_i$‘s are continuous
- ANCOVA: $y$ is continuous, some $x_i$‘s are continuous, some categorical.

Unknown parameters include the regression parameters $\beta$ and the variance $\sigma^2$. We assume $X$ is observed without error and all inference is conditional on $X$.

Bayesian regression with flat reference priors

For Bayesian analysis, we will need to specify priors for the unknown regression parameters $\beta$ and the variance $\sigma^2$.

Consider independent flat priors on $\beta$ and $\log \sigma^2$:

$$p(\beta) \propto 1 \quad \text{and} \quad p(\log(\sigma^2)) \propto 1$$

or equivalently

$$p(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

None of the above two “distributions” are valid probabilities (they do not integrate to any finite number). So why is it that we are even discussing them?

It turns out that even if the priors are improper (that’s what we call them), as long as the resulting posterior distributions are valid we can still conduct legitimate statistical inference on them.

Marginal and conditional distributions

With a flat prior on $\beta$ we obtain, after some algebra, the conditional posterior distribution:

$$p(\beta | \sigma^2, y) = \frac{1}{(\sigma^2)^{(n-p)/2+1}} \exp \left( \frac{(n-p)s^2}{2\sigma^2} \right)$$

where $s^2 = \hat{s}^2 = \frac{1}{n-p} y^T (I - P_X) y$.

This is known as an inverted Gamma distribution (also called a scaled chi-square distribution)

$$IG(s^2 | (n-p)/2, (n-p)s^2/2).$$

In other words: $s^2/\sigma^2 \sim \chi^2_{n-p}$ (with $n-p$ degrees of freedom). A striking similarity with the classical result: The distribution of $\hat{s}^2$ is also characterized as $(n-p)s^2/\sigma^2$ following a chi-square distribution.

Marginal and conditional distributions

So, we need to find the marginal posterior distribution of $\sigma^2$. With the choice of the flat prior we obtain:

$$p(\sigma^2 | y) \propto \frac{n-p}{2}^{(n-p)/2+1} \exp \left( \frac{(n-p)s^2}{2\sigma^2} \right)$$

where $s^2 = \hat{s}^2 = \frac{1}{n-p} y^T (I - P_X) y$.

This is known as an inverted Gamma distribution (also called a scaled chi-square distribution)

$$IG(s^2 | (n-p)/2, (n-p)s^2/2).$$

In other words: $s^2/\sigma^2 \sim \chi^2_{n-p}$ (with $n-p$ degrees of freedom). A striking similarity with the classical result: The distribution of $\hat{s}^2$ is also characterized as $(n-p)s^2/\sigma^2$ following a chi-square distribution.
The Gibbs sampler

More generally, if θ = (θ₁,...,θₚ) are the parameters in our model, we provide a set of initial values θ(0) = (θ₁(0),...,θₚ(0)) and then perform the j-th iteration, say for j = 1,...,M, by updating successively from the full conditional distributions:

θ(j)₁ ~ p(θ₁ |θ₂(j−1),...,θₚ(j−1), y)
θ(j)₂ ~ p(θ₂ |θ₁(j),θ₃(j),...,θₚ(j−1), y)

... (the generic j-th element)
θ(j)ₚ ~ p(θₚ |θ₁(j),...,θₚ−1(j), y)

The marginal distribution of each individual regression parameter βⱼ is a non-central univariate tₙ₋₀⁻ distribution. In fact,

\[ \beta_j - \hat{\beta}_j \sim t_{n-p-} \]

The 95% credible intervals for each βⱼ are constructed from the quantiles of the t-distribution. The credible intervals exactly coincide with the 95% classical confidence intervals, but the interpretation is direct: the probability of βⱼ falling in that interval, given the observed data, is 0.95.

Note: an intercept only linear model reduces to the simple univariate N(y | μ, σ²/n) likelihood, for which the marginal posterior of μ is:

\[ \mu - \bar{y} \sim t_{n-1} \]

Composition sampling for linear regression

Now we are ready to carry out composition sampling from \( p(\beta, \sigma^2 | y) \) as follows:

\[ \Gamma(n/2) \left[ \beta T X (X T X)^{-1} \beta + (\beta - \hat{\beta}) T (X T X)(\beta - \hat{\beta}) \right]^{-n/2} \]

Bayesian predictions from the linear model

Suppose we have observed the new predictors \( \bar{X} \), and we wish to predict the outcome \( y \). We specify \( p(y | \bar{X}, \theta) \) to be a normal distribution:

\( \left( \begin{array}{c} y \\ \bar{X} \end{array} \right) \sim N( \left( \begin{array}{c} X \\ \bar{X} \end{array} \right) \beta, \sigma^2 I) \)

The posterior predictive distribution:

\[ p(y | \bar{X}, \theta, \sigma^2) = p(\bar{y} | \beta, \sigma^2) = N(\bar{y} | \bar{X} \beta, \sigma^2 I) . \]

More generally, if \( \theta = (\theta_1, ..., \theta_p) \) are the parameters in our model, we provide a set of initial values \( \theta(0) = (\theta_1(0), ..., \theta_p(0)) \) and then perform the j-th iteration, say for \( j = 1, ..., M \), by updating successively from the full conditional distributions:

\[ \theta(j)₁ \sim p(\theta₁ |\theta₂(j−1),...,\thetaₚ(j−1), y) \]
\[ \theta(j)₂ \sim p(\theta₂ |\theta₁(j),\theta₃(j),...,\thetaₚ(j−1), y) \]

... (the generic j-th element)
\[ \theta(j)ₚ \sim p(\thetaₚ |\theta₁(j),...,\thetaₚ−1(j), y) \]

Note: an intercept only linear model reduces to the simple univariate N(y | μ, σ²/n) likelihood, for which the marginal posterior of μ is:

\[ \mu - \bar{y} \sim t_{n-1} \]

The Gibbs sampler

Suppose that θ = (θ₁, θ₂) and we seek the posterior distribution \( p(θ₁, θ₂ | y) \).

For many interesting hierarchical models, we have access to full conditional distributions \( p(θ₁ | θ₂, y) \) and \( p(θ₂ | θ₁, y) \).

The Gibbs sampler proposes the following sampling scheme. Set starting values \( θ(0) = (θ₁(0), θ₂(0)) \) For \( j = 1, ..., M \):

Draw \( θ₁(j) \sim p(θ₁ | θ₂(j−1), y) \)
Draw \( θ₂(j) \sim p(θ₂ | θ₁(j), y) \)

This constructs a Markov Chain and, after an initial “burn-in” period when the chains are trying to find their way, the above algorithm guarantees that \( \{θ₁(j), θ₂(j)\}_{j=M_0+1}^M \) will be samples from \( p(θ₁, θ₂ | y) \), where \( M_0 \) is the burn-in period.

Example: Consider the linear model. Suppose we set \( p(σ^2) = IG(σ² | a, b) \) and \( p(θ₁) \propto 1 \).

The full conditional distributions are:

\[ p(θ₁ | y, σ²) = N(θ₁ | X T X)^{-1}X T y, σ²(X T X)^{-1}) \]
\[ p(σ² | y, β) = IG \left( σ² | a + n/2, b + \frac{1}{2}(y - Xβ T)(y - Xβ) \right) . \]

Thus, the Gibbs sampler will initialize \( (θ₁(0), θ₂(0)) \) and draw, for \( j = 1, ..., M \):

Draw \( θ₁(j) \sim N((X T X)^{-1}X T y, θ₂(j−1)(X T X)^{-1}) \)
Draw \( θ₂(j) \sim IG \left( a + n/2, b + \frac{1}{2}(y - Xθ₁(j T)(y - Xθ₁(j)) \right) \)
The Metropolis-Hastings algorithm

- In principle, the Gibbs sampler will work for extremely complex hierarchical models. The only issue is sampling from the full conditionals. They may not be amenable to easy sampling – when these are not in closed form. A more general and extremely powerful - and often easier to code algorithm is the Metropolis-Hastings (MH) algorithm.

- This algorithm also constructs a Markov Chain, but does not necessarily care about full conditionals.

Example: For the linear model, our parameters are $(\beta, \sigma^2)$. We write $\theta = (\beta, \log(\sigma^2))$. The Metropolis-Hastings algorithm: Start with an initial value for $\theta = \theta(0)$. Select a candidate or proposal distribution from which to propose a value of $\theta$ at the $j$-th iteration: $\theta(j) \sim q(\theta(j-1))$. For example, $q(\theta(j-1)) = N(\theta(j-1), \nu)$ with $\nu$ fixed.

- Compute
  \[
  r = \frac{p(\theta^*) | y}{p(\theta(j-1)) | y} q(\theta(j-1)) \quad \text{for } j = 1, \ldots, M
  \]
  where $\theta^*$ is a sample from $q(\theta(j-1))$.

- If $r \geq 1$ then set $\theta(j) = \theta^*$. If $r \leq 1$ then draw $U \sim (0, 1)$. If $U \leq r$ then $\theta(j) = \theta^*$. Otherwise, $\theta(j) = \theta(j-1)$.

- Repeat for $j = 1, \ldots, M$. This yields $\theta(1), \ldots, \theta(M)$, which, after a burn-in period, will be samples from the true posterior distribution. It is important to monitor the acceptance ratio $r$ of the sampler through the iterations. Rough recommendations: for vector updates $r \approx 20\%$, for scalar updates $r \approx 40\%$. This can be controlled by “tuning” $\nu$.

- Popular approach: Embed Metropolis steps within Gibbs to draw from full conditionals that are not accessible to directly generate from.