

Bayesian Regression Notes

Jacob M. Schauer

An important aspect of imputation in a meta-regression, as proposed by our ongoing work, is that it relies on a Bayesian estimation of the meta-regression equation. This document summarizes a few useful results on Bayesian regression and Bayesian meta-regression.

Bayesian Regression

Suppose the regression equation is

$$Y_i = \mathbf{X}_i\beta + e_i, \quad i = 1, \dots, k$$

where $Y_i \in \mathbb{R}$, $\mathbf{X}_i = [1, X_{i1}, \dots, X_{ip}] \in \mathbb{R}^{p+1}$, $\beta = [\beta_0, \dots, \beta_p]$ and $e_i \stackrel{IID}{\sim} N(0, \sigma^2)$.

The parameters of the model are β, σ^2 . The fully Bayesian approach is that we set a probability model for the data X, Y : $p(X, Y|\beta, \sigma^2, \psi)$ such that ψ is part of joint distribution of X, Y .

We know that the regression model is given by

$$p(Y|X, \beta, \sigma^2, \psi) = (2\pi\sigma^2)^{-k/2} \exp \left\{ -\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \right\}$$

Denote the conditional distribution of $X|\beta, \sigma^2, \psi$ as $p(X|\beta, \sigma^2, \psi)$. Then, the full joint distribution is given by:

$$P(Y, X|\beta, \sigma^2, \psi) = p(Y|X, \beta, \sigma^2, \psi)p(X|\beta, \sigma^2, \psi)$$

The joint posterior is given by

$$p(\beta, \sigma^2, \psi|X, Y) \propto p(Y|X, \beta, \sigma^2, \psi)p(X|\beta, \sigma^2, \psi)p(\beta, \sigma^2, \psi)$$

where $p(\beta, \sigma^2, \psi)$ is the joint prior on the parameters.

If we assume that:

1. $X \perp \beta, \sigma^2 | \psi$, so that given ψ the parameters β, σ^2 tell us nothing about X
2. $\beta, \sigma^2 \perp \psi$ *a priori*

then we can rewrite the posterior as

$$\begin{aligned} p(\beta, \sigma^2, \psi|X, Y) &\propto p(Y|X, \beta, \sigma^2, \psi)p(X|\beta, \sigma^2, \psi)p(\beta, \sigma^2, \psi) \\ &= p(Y|X, \beta, \sigma^2)p(\beta, \sigma^2)p(X|\psi)p(\psi) \\ &\propto p(Y|X, \beta, \sigma^2)p(\beta, \sigma^2) \end{aligned}$$

Thus, inference about β, σ^2 does not depend on the probability distribution of $X|\psi$, and we can thus ignore it. Note that an alternative justification for this is that often in experiments the X are fixed, so that they are degenerate random variables, and hence their distribution can be ignored.

Given that we can just zero in on $p(Y|X, \beta, \sigma^2)p(\beta, \sigma^2)$ for inference on β, σ^2 , there are a few useful results. First, we can write the likelihood function $p(Y|X, \beta, \sigma^2)$ as:

$$p(Y|X, \beta, \sigma^2) \propto \sigma^{-k} \exp \left\{ -\frac{1}{2\sigma^2} \left((k-p-1)S^2 + (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right) \right\}$$

where $\hat{\beta} = (X^T X)^{-1} X^T Y$ and $S^2 = (Y - X\hat{\beta})^T (Y - X\hat{\beta}) / (k-p-1)$

Thus, we can write the posterior as:

$$p(\beta, \sigma^2 | X, Y) \propto \sigma^{-k} \exp \left\{ -\frac{(k-p-1)S^2}{2\sigma^2} \right\} \exp \left\{ -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right\} p(\beta, \sigma^2)$$

When $\beta \perp \sigma^2$ a priori and $p(\beta) \propto a \in \mathbb{R}$ and $p(\sigma^2) \propto \sigma^{-2}$ then

$$p(\beta, \sigma^2 | X, Y) \propto \underbrace{(\sigma^2)^{-k/2-1} \exp \left\{ -\frac{(k-p-1)S^2}{2\sigma^2} \right\}}_{\text{scaled inverse chi-square}} \underbrace{\exp \left\{ -\frac{1}{2\sigma^2} (\beta - \hat{\beta})^T X^T X (\beta - \hat{\beta}) \right\}}_{\text{normal}}$$

Thus, under that prior, we have:

$$\begin{aligned} p(\sigma^2 | X, Y) &= p(\sigma^2 | S^2) \sim (k-p-1)S^2 \chi_{k-p-1}^{-2} \\ p(\beta | \sigma^2 X, Y) &= N(\hat{\beta}, (X^T X)^{-1} \sigma^2) \end{aligned}$$

Using the method of composition, sampling from the joint posterior is pretty simple:

1. Draw $X \sim \chi_{k-p-1}^2$ and set $\sigma^{2(i)} = (k-p-1)S^2/X$
2. Draw $\beta^{(i)} \sim N(\hat{\beta}, (X^T X)^{-1} \sigma^{2(i)})$

Bayesian Meta-Regression

Assume

$$T_i = \mathbf{X}_i \beta + u_i + e_i$$

where $u_i \stackrel{IID}{\sim} N(0, \tau^2)$ and $e_i \stackrel{indep}{\sim} N(0, v_i)$ with $e_i \perp u_i$. Denote $\Sigma = \text{diag}(v_i)$ and $T = \tau^2 I$ and $\Omega = \Sigma + T$.

For the same reason we can ignore the probability distribution of X above, we can also ignore it in the meta-regression model, so that the posterior of interest is

$$p(\beta, \tau^2 | X, T, v) \propto p(T | X, v, \beta, \tau^2) p(\beta, \tau^2)$$

As above, we can factor the likelihood so that it can be expressed:

$$p(T | X, v, \beta, \tau^2) \propto |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2} (T - X\hat{\beta})^T \Omega^{-1} (T - X\hat{\beta}) \right\} \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})^T X^T \Omega^{-1} X (\beta - \hat{\beta}) \right\}$$

where $\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} T$

Assuming prior independence and $p(\beta) \propto a \in \mathbb{R}$, we can write the posterior as:

$$p(T | X, v, \beta, \tau^2) \propto |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2} (T - X\hat{\beta})^T \Omega^{-1} (T - X\hat{\beta}) \right\} p(\Omega) \exp \left\{ -\frac{1}{2} (\beta - \hat{\beta})^T X^T \Omega^{-1} X (\beta - \hat{\beta}) \right\}$$

Note that Ω is a function of τ^2 and the v_i , so we can write it as $\Omega(\tau^2, v)$. We can again factorize this as

$$p(\beta | T, X, v, \tau^2) \sim N(\beta, (X^T \Omega^{-1} X)^{-1})$$

However, unlike with the regular homoscedastic model, sampling is not quite so simple. This is because Ω is random, but it is the sum of something fixed Σ and random T . Thus we would need to draw T from some distribution in order to form draws of Ω . Moreover, the first exponent is a much more complex function of Ω since it is part of $\hat{\beta}$.

One potential idea for speed is to use importance sampling. We would need to figure out some convenient $f(\tau^2)$ that has the right properties and is easy to sample from. Then select a bootstrap sample of those τ^2 values with probability proportional to $p(\tau^2|S^2)/f(\tau^2)$.

Another idea is to use a different approximation, where we compute $\hat{\beta}(\tau^2 = 0)$ so that $T - X\hat{\beta} = T - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}T$. Then, Ω would be approximately inverse Wishart, so we could draw from that and replace the diagonals with v_i if they are smaller than v_i .