Applied Bayesian Statistics, week 3

Robin J. Ryder

January 31, 2023

Linear regression

Outcome: continuous variable y
Explanatory variables $x_1, \ldots, x_p$

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \ldots + \beta_8 x_i^8 + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

Frequentist setting

Likelihood:

$$L(\beta, \sigma^2 | y, X) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right]$$

Least squares method

$$\hat{\beta} = (X^TX)^{-1}X^Ty$$

$$\hat{\sigma}^2 = \frac{1}{n}s^2 = \frac{1}{n}(y - X\hat{\beta})^T(y - X\hat{\beta})$$

Conjugate prior

If [conditional prior]

$$\beta|\sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim IG(a, b), \qquad a, b > 0,$$

Conjugate prior

If [conditional prior]

$$\beta|\sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

where M $(k+1, k+1)$ positive definite symmetric matrix, and

$$\sigma^2|X \sim IG(a, b), \qquad a, b > 0,$$

then

$$\beta|\sigma^2, y, X \sim N_{k+1}\left((M + X^T X)^{-1}\{(X^T X)\hat{\beta} + M\tilde{\beta}\}, \sigma^2(M + X^T X)^{-1}\right)$$

and

$$\sigma^2|y, X \sim IG\left(\frac{n}{2} + a, b + \frac{s^2}{2} + \frac{(\tilde{\beta} - \hat{\beta})^T \left(M^{-1} + (X^T X)^{-1}\right)^{-1} (\tilde{\beta} - \hat{\beta})}{2}\right)$$

Experimenter's dilemma

$$\beta|\sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

The choice of M, or of g if $M = I_{k+1}/g$, is problematic.

Experimenter's dilemma

$$\beta|\sigma^2, X \sim N_{k+1}(\tilde{\beta}, \sigma^2 M^{-1}),$$

The choice of M, or of g if $M = I_{k+1}/g$, is problematic.
Zellner's informative G-prior allows the experimenter to introduce information about the location parameter of the regression while bypassing the most difficult aspect of prior specification; the derivation of the prior correlation structure.

$$\begin{aligned}
\beta|\sigma^2, X &\sim N_{k+1}(\tilde{\beta}, g\sigma^2(X^T X)^{-1}) \\
\sigma^2 &\sim \pi(\sigma^2|X) \propto \sigma^{-2}.
\end{aligned}$$

Prior selection

We now just need to choose $\tilde{\beta}$ and g.
g can be interpreted as a measure of the amount of information available in the prior relative to the sample. For instance, setting $1/g = 0.5$ gives the prior the same weight as 50% of the sample. Setting $g = n$ gives the prior the same weight as one observation.

Posterior distribution

With this prior model, the posterior simplifies into

$$\begin{aligned}
\pi(\beta,\sigma^2|y,X) &\propto f(y|\beta,\sigma^2,X)\pi(\beta,\sigma^2|X)\\
&\propto \left(\sigma^2\right)^{-(n/2+1)}\exp\left[-\frac{1}{2\sigma^2}(y-X\hat\beta)^T(y-X\hat\beta)\right.\\
&\qquad \left.-\frac{1}{2\sigma^2}(\beta-\hat\beta)^TX^TX(\beta-\hat\beta)\right]\left(\sigma^2\right)^{-k/2}\\
&\qquad \times\exp\left[-\frac{1}{2g\sigma^2}(\beta-\tilde\beta)^TX^TX(\beta-\tilde\beta)\right],
\end{aligned}$$

because X^TX used in both prior and likelihood.

Posterior distribution

Therefore,

$$
\begin{aligned}
\beta|\sigma^2, y, X &\sim \mathcal{N}_{k+1}\left(\frac{g}{g+1}(\tilde\beta/g+\hat\beta), \frac{\sigma^2 g}{g+1}(X^TX)^{-1}\right)\\
\sigma^2|y, X &\sim \mathcal{IG}\left(\frac{n}{2}, \frac{s^2}{2}+\frac{1}{2(g+1)}(\tilde\beta-\hat\beta)^T X^T X(\tilde\beta-\hat\beta)\right)
\end{aligned}
$$

and

$$
\begin{aligned}
\beta|y, X &\sim \mathcal{T}_{k+1}\left(n, \frac{g}{g+1}\left(\frac{\tilde\beta}{g}+\hat\beta\right),\right.\\
&\qquad \left.\frac{g(s^2+(\tilde\beta-\hat\beta)^T X^T X(\tilde\beta-\hat\beta)/(g+1))}{n(g+1)}(X^TX)^{-1}\right).
\end{aligned}
$$

Null hypothesis

If a null hypothesis is $H_0 : R\beta = r$, the model under H_0 can be rewritten as

$$y|\beta^0, \sigma^2, X_0 \overset{H_0}{\sim} \mathcal{N}_n\left(X_0\beta^0, \sigma^2 I_n\right)$$

where β^0 is $(k+1-q)$ dimensional.

Point null marginal

Under the prior

$$\beta^0 | X_0, \sigma^2 \sim \mathcal{N}_{k+1-q}\left(\tilde{\beta}^0, g_0\sigma^2 (X_0^T X_0)^{-1}\right),$$

the marginal distribution of y under H_0 is

$$\begin{aligned} f(y|X_0, H_0) &= (g+1)^{-(k+1-q)/2} \pi^{-n/2} \Gamma(n/2) \\ &\times \left[y^T y - \frac{g_0}{g_0+1} y^T X_0 (X_0^T X_0)^{-1} X_0^T y \right. \\ &\left. - \frac{1}{g_0+1} \tilde{\beta}_0^T X_0^T X_0 \tilde{\beta}_0 \right]^{-n/2} . \end{aligned}$$

Therefore the Bayes factor is closed form:

$$B_{10}^{\pi} = \frac{f(y|X,H_1)}{f(y|X_0,H_0)} = \frac{(g_0+1)^{(k+1-q)/2}}{(g+1)^{(k+1)/2}}$$

$$\left[\frac{y^Ty - \frac{g_0}{g_0+1}y^TX_0(X_0^TX_0)^{-1}X_0^Ty - \frac{1}{g_0+1}\tilde{\beta}_0^TX_0^TX_0\tilde{\beta}_0}{y^Ty - \frac{g}{g+1}y^TX(X^TX)^{-1}X^Ty - \frac{1}{g+1}\tilde{\beta}^TX^TX\tilde{\beta}}\right]^{n/2}$$

Variable selection

We now consider all possible models: since there are p explanatory variables, there are 2^p possible models. Each model γ is associated with a posterior probability $\pi(\gamma|y)$. For a large number of models, it is not practical to compute all marginal likelihoods. We therefore propose instead a method to sample from π.

MCMC

We shall build a Markov chain (Z_t) on the state space of models, such that the stationary distribution of Z is the posterior π. That way, when Z reaches stationarity, it produces samples from π.

MCMC

We shall build a Markov chain (Z_t) on the state space of models, such that the stationary distribution of Z is the posterior π. That way, when Z reaches stationarity, it produces samples from π.

There are two issues here:

1. Convergence: we need to ensure that Z has reached stationarity.
2. Mixing: the samples produced by Z are not independent, so we need to check that the chain moves around in the distribution well enough.

A Gibb's sampler is a simple form of MCMC which relies on the conditional probabilities. To sample γ, follow this algorithm for t between 1 and N_{iter}.

A Gibb's sampler is a simple form of MCMC which relies on the conditional probabilities. To sample γ, follow this algorithm for t between 1 and N_{iter}.

Initialization: Start with arbitrary value $\gamma^{(0)} = (\gamma_1^{(0)}, \gamma_2^{(0)}, \dots, \gamma_p^{(0)})$

Iteration t: Given $\gamma^{(t-1)} = (\gamma_1^{(t-1)}, \dots, \gamma_p^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi_1(\gamma_1|\gamma_2^{(t-1)}, \gamma_3^{(t-1)}, \dots, \gamma_p^{(t-1)}, y)$
2. $\gamma_2^{(t)}$ according to $\pi_2(\gamma_2|\gamma_1^{(t)}, \gamma_3^{(t-1)} \dots, \gamma_p^{(t-1)}, y)$
3. $\gamma_3^{(t)}$ according to $\pi_3(\gamma_3|\gamma_1^{(t)}, \gamma_2^{(t)}, \gamma_4^{(t-1)}, \dots, \gamma_p^{(t-1)}, y)$

...

p. $\gamma_p^{(t)}$ according to $\pi_p(\gamma_p|\gamma_1^{(t)}, \gamma_2^{(t)} \dots, \gamma_8^{(t)}, y)$

Conditional distribution

In our case, γ_1 can take only 2 values, so $\pi_1(\gamma_1|\gamma_2^{(t-1)}, \gamma_3^{(t-1)}, \dots, \gamma_p^{(t-1)}, y)$ is a Bernoulli distribution. The probabilities that γ_1 equal 0 or 1 are proportional to the corresponding marginal likelihoods of the entire vector γ and can thus be computed.

$$
\begin{aligned}
\pi_1(\gamma_1 = 0|\gamma_2^{(t-1)}, \dots, \gamma_p^{(t-1)}, y) &\propto \pi_1(0, \gamma_2^{(t-1)}, \dots, \gamma_p^{(t-1)}|y) \\
\pi_1(\gamma_1 = 1|\gamma_2^{(t-1)}, \dots, \gamma_p^{(t-1)}, y) &\propto \pi_1(1, \gamma_2^{(t-1)}, \dots, \gamma_p^{(t-1)}|y)
\end{aligned}
$$

The same holds for $\gamma_2, \dots, \gamma_p$.