COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

STAT 4224/5224

Bayesian Statistics

Dobrin Marchev

Recall: Bayesian Regression

The likelihood function is:

$$f(\boldsymbol{y}\,|\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}SSE(\boldsymbol{\beta})}$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\boldsymbol{y'y}-2\boldsymbol{\beta' X' y}+\boldsymbol{\beta' X' X\beta})}$$

We will use a multivariate conjugate prior on $\boldsymbol{\beta} \sim \mathrm{N}_{p+1}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$

Then

$$f(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) \sim N_{p+1}(m, V)$$

where

$$V = Var(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{\boldsymbol{X'X}}{\sigma^2}\right)^{-1}$$

$$m = E(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\sigma^2) = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{X'X}{\sigma^2}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\beta_0 + \frac{X'y}{\sigma^2}\right)$$

Bayesian Estimation: posterior of σ^2

For σ^2 we will again use a semi-conjugate prior:

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Then:

$$f(\sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) \propto \pi(\sigma^2) f(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2)$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0}{2}-1} e^{-\frac{1}{\sigma^2}\frac{\nu_0\sigma_0^2}{2}} \times \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{\sigma^2}\frac{SSE(\boldsymbol{\beta})}{2}}$$

$$= \left(\frac{1}{\sigma^2}\right)^{\frac{\nu_0+n}{2}-1} e^{-\frac{1}{\sigma^2}\left(\frac{\nu_0\sigma_0^2}{2}+\frac{SSE(\boldsymbol{\beta})}{2}\right)}$$

$$\Rightarrow \sigma^2|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta} \sim \text{IG}\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSE(\boldsymbol{\beta})}{2}\right)$$

Gibbs Sampler

Constructing a Gibbs sampler to approximate the joint posterior distribution $f(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{X})$ is then straightforward: Given current values $\{\boldsymbol{\beta}^{(s)}, \sigma^{2(s)}\}$, new values can be generated by:

1. Updating $\boldsymbol{\beta}$:

a) Compute $V = Var\left(\boldsymbol{\beta}\middle|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}\right), m = E\left(\boldsymbol{\beta}\middle|\boldsymbol{y}, \boldsymbol{X}, \sigma^{2(s)}\right)$

b) Sample $\beta^{(s+1)} \sim N_{p+1}(m, V)$

2. Updating σ^2

a) Compute SSE($\boldsymbol{\beta}^{(s+1)}$)

b) Sample $\sigma^{2(s+1)} \sim IG\left(\frac{\nu_0+n}{2}, \frac{\nu_0\sigma_0^2+SSE\left(\boldsymbol{\beta}^{(s+1)}\right)}{2}\right)$

Weakly Informative Priors

A Bayesian analysis of a regression model requires specification of the prior parameters ($\boldsymbol{\beta}_0$, $\boldsymbol{\Sigma}_0$) and (ν_0, σ_0^2). Finding values of these parameters that represent actual prior information can be difficult.

Sometimes an analysis must be done in the absence of precise prior information, or information that is easily converted into the parameters of a conjugate prior distribution. One idea is that, if the prior distribution is not going to represent real prior information about the parameters, then it should be as minimally informative as possible. To some, such an analysis would give a "more objective" result than using an informative prior distribution. One type of weakly informative prior is the *unit information prior*. A unit information prior is one that contains the same amount of information as that would be contained in only a single observation. For example, the precision of $\widehat{\boldsymbol{\beta}}_{OLS}$ is the inverse of variance or $\frac{\boldsymbol{X'X}}{\sigma^2}$. Since this can be viewed as the amount of information in n observations, the amount of information in one observation should be $\frac{\boldsymbol{X'X}}{n\sigma^2}$. The unit information prior thus sets $\Sigma_0^{-1} = \frac{\boldsymbol{X'X}}{n\sigma^2}$. Further suggestion is to set $\boldsymbol{\beta_0} = \widehat{\boldsymbol{\beta}}_{OLS}$. For σ^2 prior we can choose $\nu_0 = 1$ and $\sigma_0^2 = \hat{\sigma}_{OLS}^2$.

Weakly Informative Priors

Another principle for constructing a prior distribution for $\boldsymbol{\beta}$ is based on the idea that the parameter estimation should be invariant to changes in the scale of the regressors. Suppose $\boldsymbol{X}$ is a given set of regressors and $\widetilde{\boldsymbol{X}} = \boldsymbol{XH}$ for some $p \times p$ matrix $\boldsymbol{H}$. If we obtain the posterior distribution of $\boldsymbol{\beta}$ from $\boldsymbol{y}$ and $\boldsymbol{X}$, and the posterior distribution of $\widetilde{\boldsymbol{\beta}}$ from $\boldsymbol{y}$ and $\widetilde{\boldsymbol{X}}$, then, according to this principle of invariance, the posterior distributions of $\boldsymbol{\beta}$ and $\boldsymbol{H}\widetilde{\boldsymbol{\beta}}$ should be the same. Linear algebra shows that this can happen if $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = k(\boldsymbol{X}'\boldsymbol{X})^{-1}$ for any $k > 0$. A popular specification of k is to relate it to the error variance σ^2, so that $k = g\sigma^2$ for some positive value g. These choices of prior parameters result in a version of the so-called "g-prior". Under it,

$$\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2 \sim N_{p+1}(\boldsymbol{m}, \boldsymbol{V})$$

where

$$\boldsymbol{V} = Var(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{g\sigma^2} + \frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\right)^{-1} = \frac{g}{g+1}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\boldsymbol{m} = E(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}, \sigma^2) = \left(\frac{\boldsymbol{X}'\boldsymbol{X}}{g\sigma^2} + \frac{\boldsymbol{X}'\boldsymbol{X}}{\sigma^2}\right)^{-1}\left(\frac{\boldsymbol{X}'\boldsymbol{y}}{\sigma^2}\right) = \frac{g}{g+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

Weakly Informative Priors

We choose again prior:

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Under the "*g*-prior" it can be shown that

$$\sigma^2|\boldsymbol{y}, \boldsymbol{X} \sim IG\left(\frac{\nu_0 + n}{2}, \frac{\nu_0\sigma_0^2 + SSE_g}{2}\right)$$

where

$$\boldsymbol{V} = \frac{g}{g+1}\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$\boldsymbol{m} = \frac{g}{g+1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$SSE_g = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{m}'\boldsymbol{V}^{-1}\boldsymbol{m} = \boldsymbol{y}'\left(\boldsymbol{I} - \frac{g}{g+1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y}$$

The effect of g is that it shrinks down the magnitude of the regression coefficients and can prevent overfitting of the data.

These results mean we don't need Gibbs sampler and can use Monte Carlo simulation to obtain estimate of the posterior (see p. 159).

Model Selection

Often in regression analysis we are faced with many possible regressor variables, even though we suspect that a majority of the regressors have no true relationship to the variable *Y*. In these situations, including all possible variables in a regression model can lead to poor statistical performance. Standard statistical advice is that we should include in our regression model only those variables for which there is substantial evidence of an association with *Y*. Doing so not only produces simpler, more aesthetically pleasing data analyses, but also generally provides models with better statistical properties in terms of prediction and estimation.

Model Selection: Example

Baseline data for ten variables $x_1, \ldots, x_{10}$ on a group of 442 diabetes patients were gathered, as well as a measure y of disease progression taken one year after the baseline measurements. From these data we hope to make a predictive model for y based on the baseline measurements. While a regression model with ten variables would not be overwhelmingly complex, it is suspected that the relationship between y and the x_j's may not be linear, and that including second-order terms like x_j^2 and $x_j x_k$ in the regression model might aid in prediction.

Therefore, we have 10 main effects, $\binom{10}{2} = 45$ interactions and 9 quadratic terms (since there is one binary variable). There are total of $p = 64$ potential predictors. To put the variables on a common scale all predictors and the y variable are standardized. Then the data are split into training and test parts.

Model Selection

One standard way to assess the evidence that the true value of a regression coefficient β_j is not zero is with a t-statistic. Consider the following procedure, aka backwards elimination:

1. Obtain the estimator $\widehat{\boldsymbol{\beta}}_{OLS}$ and all t-statistics.

2. If there is any j such that $|t_j| <$ tcutoff

a) Find the smallest $|tj|$ and remove the corresponding column from X.

b) Return to step 1.

3. If $|t_j| > t_{cutoff}$ for all j, then stop.

Such procedures, in which a potentially large set of regressors is reduced to a smaller set, are called model selection procedures. A standard choice for t_{cutoff} is an upper quantile of a t or standard normal distribution.

Bayesian Model Comparison

The Bayesian solution to the model selection problem is conceptually straightforward: If we believe that many of the regression coefficients are potentially equal to zero, then we simply use a prior distribution that reflects this possibility.

This can be accomplished by specifying that each regression coefficient has some probability of being exactly zero.

A convenient way to represent this is to write the regression coefficient for variable j as $\beta_j = z_j \times b_j$, where $z_j \in \{0,1\}$ and b_j is a coefficient to be estimated.

Then different choices of z_j correspond to different subsets of predictors selected to be included in the model. That is, each value of $z = (z_1, \ldots, z_p)$ corresponds to a different model, or more specifically, a different collection of variables having non-zero regression coefficients.

Bayesian Model Comparison

Bayesian model selection proceeds by obtaining a posterior distribution for $\boldsymbol{z}$. Of course, doing so requires a joint prior distribution on $\{z, \beta, \sigma^2\}$. It turns out that a version of the *g*-prior described in the previous section allows us to evaluate $f(y|X, z)$ for each possible model z. Then with a prior $\pi(z)$, we have

$$p(z|y, X) = \frac{\pi(z) f(\boldsymbol{y}|X, z)}{\sum_z \pi(z) f(y|X, z)}$$

Alternatively, we can compare the evidence for any two models with the posterior odds:

$$odds(z_a, z_b|y, X) = \frac{p(\boldsymbol{z}_a|y, X)}{p(\boldsymbol{z}_b|y, X)} = \frac{\pi(z_a)}{\pi(z_b)} \times \frac{f(\boldsymbol{y}|X, z_a)}{f(\boldsymbol{y}|X, z_a)}$$

Posterior odds = prior odds × Bayes factor

See pp. 164-166 on how to compute these.

Gibbs sampling and model averaging

If we allow each of the p regression coefficients to be either zero or non-zero, then there are 2^p different models to consider. If p is large, then it will be impractical for us to compute the marginal probability of each model. In these situations, our data analysis goals become more modest: For example, we may be content with a decent estimate of $\boldsymbol{\beta}$ from which we can make predictions, or a list of relatively high-probability models. These items can be obtained with a Markov chain which searches through the space of models for values of $\boldsymbol{z}$ with high posterior probability. This can be done with a Gibbs sampler in which we iteratively sample each z_j from its full conditional distribution. Specifically, given a current value $\boldsymbol{z} = (z_1, \ldots, z_p)$, a new value of zj is generated by sampling from $p(z_j|\boldsymbol{y}, \boldsymbol{X}, z_{-j})$.

Gibbs sampling and model averaging

More precisely, generating values of $\{z^{(s+1)}, \sigma^{2(s+1)}, \beta^{(s+1)}\}$ from $\boldsymbol{z}^{(s)}$ is achieved with the following steps:

1. Set $z = \boldsymbol{z}^{(s)}$
2. For $j \in \{1,..., p\}$ in random order, replace z_j with a sample from $p(zj \,|z_{-j}, y, X)$;
3. Set $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}$
4. Sample $\sigma^{2(s+1)} \sim f(\sigma^2|\boldsymbol{z}^{(s+1)}\, y, X)$
5. Sample $\beta^{(s+1)} \sim f(\beta|\, \sigma^{2(s+1)}\,, \boldsymbol{z}^{(s+1)}\, y, X)$

Note that the entries of $z^{(s+1)}$ are not sampled from their full conditional distributions given $\sigma^{2(s)}$ and $\beta^{(s)}$. This is to ensure the distribution of $z^{(s)}$ converges to the target posterior distribution $p(z|y, X)$

Gibbs sampling and model averaging

The final estimate of β is

$$\hat{\beta}_{BMA} = \frac{1}{S}\sum_{s=1}^{S} \beta^{(s)}$$

This parameter estimate is sometimes called the (Bayesian) model averaged estimate of β, because it is an average of regression parameters from different values of z, i.e. over different regression models. This estimate, obtained by averaging the regression coefficients from several high-probability models, often performs better than the estimate of β obtained by considering only a single model. The predicted values have smaller MSE on the test data set.