The Bayesian Linear Model

Sudipto Banerjee

sudiptob@biostat.umn.edu

University of Minnesota

Linear Model Basics

- The linear model is the most fundamental of all serious statistical models, encompassing ANOVA, regression, ANCOVA, random and mixed effect modelling etc.
- Ingredients of a linear model include an $n \times 1$ response vector $\mathbf{y} = (y_1, \ldots, y_n)$ and an $n \times p$ design matrix (e.g. including regressors) $X = [\mathbf{x}_1, \ldots, \mathbf{x}_p]$, assumed to have been observed without error. The linear model:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon};\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$$

- Recall from standard statistical analysis, using calculus and linear algebra, the classical unbiased estimates of the regression parameter $\boldsymbol{\beta}$ and σ^2 are

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y};\ \hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}).$$

- The above estimate of $\boldsymbol{\beta}$ is also a least-squares estimate. The *predicted* value of $\mathbf{y}$ is given by

$$\hat{\mathbf{y}} = X\hat{\boldsymbol{\beta}} = P_X\mathbf{y} \text{ where } P_X = X(X^T X)^{-1} X^T.$$

- P_X is called the *projector* matrix of X. It is an operator that projects any vector to the space spanned by the columns of X. Note: $(\mathbf{y} - X\hat{\boldsymbol{\beta}})^T(\mathbf{y} - X\hat{\boldsymbol{\beta}}) = \mathbf{y}^T(I - P_X)\mathbf{y}$ is the model *residual*.

Non-informative priors

- For the Bayesian analysis, we will need to specify priors for the unknown regression parameters $\boldsymbol{\beta}$ and the variance σ^2.
- What are the "non-informative" priors that would make this Bayesian analysis equivalent to the classical distribution theory?
- We need to consider absolutely flat priors on $\boldsymbol{\beta}$ and $\log \sigma^2$. We have

$$P(\boldsymbol{\beta}) \propto 1;\ P(\sigma^2) \propto \frac{1}{\sigma^2} \text{ or equivalently } P(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

- Note that none of the above two "distributions" are valid probabilities (they do not integrate to any finite number, let alone 1). 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.
- With a flat prior on $\boldsymbol{\beta}$ we obtain, after some algebra, the *conditional posterior* distribution:

$$P(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}) = N((X^T X)^{-1} X^T \mathbf{y},\ (X^T X)^{-1} \sigma^2).$$

contd.

- The conditional posterior distribution of $\boldsymbol{\beta}$ would have been the desired posterior distribution had σ^2 been known. Since that is not the case, we need to obtain the *marginal posterior* distribution by integrating out σ^2 as:

$$P(\boldsymbol{\beta} \,|\, \mathbf{y}) = \int P(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{y}) P(\sigma^2 \,|\, \mathbf{y}) d\sigma^2$$

- So, we need to find the marginal posterior distribution of σ^2: $P(\sigma^2 \,|\, \mathbf{y})$. With the choice of the flat prior we obtain:

$$\begin{aligned} P(\sigma^2 \,|\, \mathbf{y}) &\propto \frac{1}{(\sigma^2)^{(n-p)/2+1}} \exp(-\frac{(n-p)s^2}{2\sigma^2}) \\ &\sim IG(\frac{n-p}{2}, \frac{(n-p)s^2}{2}), \end{aligned}$$

  where $s^2 = \hat{\sigma}^2 = \frac{1}{n-p}\mathbf{y}^T(I - P_X)\mathbf{y}$. This is a *scaled* inverse-chi-square distribution which is the same as an inverted Gamma distribution $IG((n-p)/2, (n-p)s^2/2)$.

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

contd.

- Returning to the marginal posterior distribution of $P(\boldsymbol{\beta} \,|\, \mathbf{y})$, we can perform further algebra (using the form of the Inverted Gamma density) to show that $[\boldsymbol{\beta} \,|\, \mathbf{y}]$ is a *non-central multivariate* t_{n-p} distribution with $n - p$ degrees of freedom and non-centrality parameter $\hat{\boldsymbol{\beta}}$.
- The above distribution is quite complicated but we rarely need to work with this. In order to carry out a non-informative Bayesian analysis, we use a simpler sampling based mechanism. For each $i = 1, \ldots, M$ first draw $\sigma^2_{(i)} \sim [\sigma^2 \,|\, \mathbf{y}]$ (which is Inverse-Gamma) followed by $\boldsymbol{\beta}_{(i)} \sim N\left((X^T X)^{-1} X^T \mathbf{y},\, (X^T X)^{-1} \sigma^2_{(i)}\right)$.
- The resulting samples $\left(\boldsymbol{\beta}_i, \sigma^2_{(i)}\right)_{i=1}^{M}$ are precisely samples from the *joint marginal posterior distribution* $P(\boldsymbol{\beta},\, \sigma^2 \,|\, \mathbf{y})$. *Automatically* the samples $\left(\boldsymbol{\beta}_{(i)}\right)_{i=1}^{M}$ are samples from marginal posterior distributions $P(\boldsymbol{\beta} \,|\, \mathbf{y})$ while the samples $\left(\sigma^2_{(i)}\right)_{i=1}^{M}$ are from the marginal posterior $P(\sigma^2 \,|\, \mathbf{y})$.

contd.

A few important theoretical consequences are noted below:

- The multivariate t density is given by:

$$P(\boldsymbol{\beta}\,|\,\mathbf{y}) = \frac{\Gamma(n/2)}{(\pi(n-p))^{p/2}\Gamma((n-p)/2)|s^2(X^TX)^{-1}|}\left[1+\frac{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(X^TX)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})}{(n-p)s^2}\right]^{-n/2}$$

- The marginal distribution of each individual regression parameter β_j is a non-central univariate t_{n-p} distribution. In fact,

$$\frac{\beta_j-\hat{\beta}_j}{s\sqrt{(X^TX)^{-1}_{jj}}} \sim t_{n-p}.$$

- In particular, with a simple univariate $N(\mu,\sigma^2)$ likelihood, we have

$$\frac{\mu-\bar{y}}{s/\sqrt{n}} \sim t_{n-1}.$$

- The promised reproduction of the classical case is now complete.

Predicting from Linear Models

- Now we want to apply our regression analysis to a new set of data, where we have observed the new covariate matrix $\tilde{X}$, and we wish to predict the corresponding outcome $\tilde{\mathbf{y}}$
- If β and σ^2 were known, then $\tilde{\mathbf{y}}$ would have a $N(\tilde{X}\beta, \sigma^2 I)$ distribution.
- When parameters unknown: all predictions for the data must follow from the *posterior predictive* distribution:

$$p(\tilde{\mathbf{y}} \mid \mathbf{y}) = \int p(\tilde{\mathbf{y}} \mid \boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) d\boldsymbol{\beta} d\sigma^2.$$

- For each posterior draw of $(\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)})_{i=1}^M$, draw $\tilde{\mathbf{y}}_{(i)}$ from $N(\tilde{X}\boldsymbol{\beta}_{(i)}, \sigma^2_{(i)} I)$. The resulting sample $(\tilde{\mathbf{y}}_{(i)})_{i=1}^M$ represents the predictive distribution.
- Theoretical Mean and Variance (conditional upon σ^2):

$$E(\tilde{y}|\sigma^2, y) = \tilde{X}\hat{\boldsymbol{\beta}}$$
$$var(\tilde{y}|\sigma^2, y) = (I + \tilde{X}(X^T X)^{-1}\tilde{X}^T)\sigma^2.$$

- Theoretical unconditional predictive distribution, $p(\tilde{y}|y)$, is a *multivariate t* distribution, $t_{n-p}(\tilde{X}\hat{\boldsymbol{\beta}}, s^2(I + \tilde{X}(X^T X)^{-1}\tilde{X}^T))$.

Lindley and Smith (1972)

- Hierarchical Linear Models in their full general form:

$$\mathbf{y} \,|\, \boldsymbol{\beta}_1, \Sigma_1 \sim N(X_1\boldsymbol{\beta}_1, \Sigma_1)$$
$$\boldsymbol{\beta}_1 \,|\, \boldsymbol{\beta}_2, \Sigma_2 \sim N(X_2\boldsymbol{\beta}_2, \Sigma_2).$$

  It is assumed that X_1, X_2, Σ_1 and Σ_2 are known matrices, with the latter two being positive definite covariance matrices.

- The marginal distribution of **y** is given by:

$$\mathbf{y} \sim N(X_1X_2\boldsymbol{\beta}_2, \Sigma_1 + X_1\Sigma_2X_1^T).$$

- The conditional distribution of $\boldsymbol{\beta}_1$ given **y** is:

$$\boldsymbol{\beta}_1 \,|\, \mathbf{y} \sim N(\Sigma^*\boldsymbol{\mu}, \Sigma^*), \text{ where}$$
$$\Sigma^* = \left(X_1^T\Sigma^{-1}X_1 + \Sigma_2^{-1}\right)^{-1}, \text{ and}$$
$$\boldsymbol{\mu} = X_1^T\Sigma_1^{-1}\mathbf{y} + \Sigma_2^{-1}X_2\boldsymbol{\beta}_2.$$

The Proof: a sketch

- To derive the marginal distribution of **y**, we first rewrite the system

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \mathbf{u}, \text{ where } \mathbf{u} \sim N(\mathbf{0}, \Sigma_1);$$
$$\boldsymbol{\beta}_1 = X_2\boldsymbol{\beta}_2 + \mathbf{v}, \text{ where } \mathbf{v} \sim N(\mathbf{0}, \Sigma_2).$$

  Also, **u** and **v** are independent of each other. It then follows that:

$$\mathbf{y} = X_1X_2\boldsymbol{\beta}_2 + X_1\mathbf{v} + \mathbf{u}.$$

  This shows that the marginal distribution is normal. The final result now follows by simple calculations of the mean and variance of **y**.

- For the conditional distribution $\boldsymbol{\beta}_1 \mid \mathbf{y}$, we see from Bayes Theorem that $p(\boldsymbol{\beta}_1 \mid \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\beta}_1)p(\boldsymbol{\beta}_1)$. By substituting expressions for the respective normal densities in the product on the right hand side and by completing the square in the quadratic form in the exponent, we find $p(\boldsymbol{\beta}_1 \mid \mathbf{y}) \propto \exp(-Q/2)$, where

$$Q = (\boldsymbol{\beta}_1 - \Sigma^*\boldsymbol{\mu})^T \Sigma^{*-1} (\boldsymbol{\beta}_1 - \Sigma^*\boldsymbol{\mu}) + [\mathbf{y}^T\Sigma_1^{-1}\mathbf{y} + \boldsymbol{\beta}_2^T X_2^T \Sigma_2^{-1} X_2\boldsymbol{\beta}_2 - \boldsymbol{\mu}^T\Sigma^*\boldsymbol{\mu}].$$