Bayesian Analysis of the Multiple Linear Regression (MLR) Model (STT465)

Gustavo de los Campos
(gustavoc@msu.edu)

1. Setting the stage

In a MLR model a quantitative outcome (y_i) is described as the sum of a linear function of covariates x_{ij} $(j = 1, \dots, p)$ plus an error term (ε_i)

$$y_i = \sum_{j=1}^{p} x_{ij} b_j + \varepsilon_i \qquad (i = 1, \dots, n). \qquad [1]$$

Typically, the first covariate is used to accommodate an intercept, that is $x_{ij} = 1$.

The *simple linear regression* is a special case of [1] with *p=2*, $y_i = b_1 + x_i b_2 + \varepsilon_i$, here b_1 is the 'y-intercept' and b_2 is the slope of the regression.

2. Matrix representation

The regression function can be also written as $\sum_{j=1}^{p} x_{ij} b_j = \boldsymbol{x}_i' \boldsymbol{b}$, where $\boldsymbol{x}_i = \left[x_{i1}, \dots, x_{ip}\right]'$ is a (column) vector containing the covariate values for the ith observation and $\boldsymbol{b} = \left[b_1, b_2, \dots, b_p\right]'$ is a vector of regression coefficients. Thus, $y_i = \boldsymbol{x}_i' \boldsymbol{b} + \varepsilon_i$. This expression gives the data-equation for the i^{th} observation. Sacking all the equations, we obtain the following system

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Or, in a more compact form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{\varepsilon} \qquad [2]$$

where: $\boldsymbol{y} = [y_1, \dots, y_n]'$ is an n-dimensional vector with the response, $\boldsymbol{X} = \{x_{ij}\}$ is an nxp matrix (rows corresponding to individuals and columns correspond to predictors), and $\boldsymbol{\varepsilon} = [\varepsilon_1, \dots, \varepsilon_n]'$ is a vector of model residuals. Expression [2] is the standard matrix representation of a multiple linear regression model.

3. Ordinary Least Squares (OLS)

OLS estimates are obtained by minimizing the residual sum of squares, that is

$$\widehat{\boldsymbol{b}}_{OLS} \underset{=}{argmin} \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2 \qquad [3]$$

To solution to the above optimization problem has a closed-form, if f **X** is full-column rank we have that $\widehat{\boldsymbol{b}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. The entry OLS.md in our repository demonstrates how to obtain OLS estimates in R.

4. Bayesian model

A Bayesian model is defined by the likelihood function and the prior distribution of the unknowns; each of these elements of the Bayesian MLR model are described next.

4.1. Likelihood with IID Gaussian errors

If the error terms are independent and identically distributed, each following a normal distribution with null mean and variance σ_ε^2, denoted as $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$, then, it follows from [1] and from the properties of the normal distribution that

$$p(y_i|\boldsymbol{b}, \sigma_\varepsilon^2) = N\left(y_i|\sum_{j=1}^{p} x_{ij}b_j, \sigma_\varepsilon^2\right) = (2\pi\sigma_\varepsilon^2)^{-\frac{1}{2}} e^{-\frac{\left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2}{2\sigma_\varepsilon^2}} \qquad [4]$$

From the independence assumption of the error terms we have that the joint distribution of the data given the parameters (i.e., the likelihood function) is

$$\begin{aligned} p(y_1, y_2, \ldots, y_n|\boldsymbol{b}, \sigma_\varepsilon^2) &= p(y_1|\boldsymbol{b}, \sigma_\varepsilon^2)p(y_2|\boldsymbol{b}, \sigma_\varepsilon^2) \times \ldots \times p(y_n|\boldsymbol{b}, \sigma_\varepsilon^2) \\ &= \prod_{i=1}^{n} (2\pi\sigma_\varepsilon^2)^{-\frac{1}{2}} e^{-\frac{\left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2}{2\sigma_\varepsilon^2}} \\ &= (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2} \\ &= (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}RSS(\boldsymbol{y},\boldsymbol{X},\boldsymbol{b})} \end{aligned} \qquad [5]$$

where $RSS(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{b}) = \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}b_j\right)^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})$ is the residual sum of squares which is a function of the response (**y**), the incidence matrix for effects (**X**) and the vector of regression coefficients (**b**).

4.2. *Maximum Likelihood estimation*

ML estimates are obtained by maximizing expression [5] with respect to the unknown parameters $\boldsymbol{\theta} = \{\boldsymbol{b}, \sigma_\varepsilon^2\}$. Equivalently, MLE can be obtained by minimizing the negative log-likelihood that is

$$\left\{\widehat{\boldsymbol{b}}, \hat{\sigma}_\varepsilon^2\right\} \underset{argmax}{=} (2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2} RSS(\boldsymbol{y},\boldsymbol{X},\boldsymbol{b})}$$
$$\underset{argmin}{=} \frac{n}{2} log(\sigma_\varepsilon^2) \quad + \frac{1}{2\sigma_\varepsilon^2} RSS(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{b})$$

It can be seen that the MLE of **b** is the argument that minimizes the RSS; therefore, in this model MLE(**b**)= $\widehat{\boldsymbol{b}}_{OLS}$. This solution does not depend on σ_ε^2. Differentiating with respect to the error variance renders the following first order condition $\hat{\sigma}_\varepsilon^2 = \frac{RSS(\boldsymbol{y},\boldsymbol{X},\hat{\boldsymbol{b}}_{OLS})}{n}$.

4.3. Prior distribution-I

The parameters entering in the likelihood (expression [5]) include the vector of regression coefficients and the error variance. The Bayesian model is completed by specifying a prior for these unknown parameters.

A standard approach consists of assigning IID normal priors to the regression coefficients, $b_j \underset{\sim}{iid} N\left(b_j|b_0, \sigma_b^2\right)$, and an independent scaled-inverse Chi-square prior for the error variance, $p(\sigma_\varepsilon^2) = \chi^{-2}(\sigma_\varepsilon^2|df_0, S_0)$; therefore,

$$p(\boldsymbol{b}, \sigma_\varepsilon^2) = \left\{\prod_{j=1}^p N\left(b_j|b_0, \sigma_b^2\right)\right\}\chi^{-2}(\sigma_\varepsilon^2|df_0, S_0) =$$
$$\left\{\prod_{j=1}^p (2\pi\sigma_b^2)^{-\frac{1}{2}} e^{-\frac{\left(b_j - b_0\right)^2}{2\sigma_b^2}}\right\} \frac{\left(\frac{S_0}{2}\right)^{\frac{df}{2}}}{\Gamma\left(\frac{df}{2}\right)} (\sigma_\varepsilon^2)^{-\left(1+\frac{df}{2}\right)} e^{-\frac{S_0}{2\sigma_\varepsilon^2}} \qquad [6]$$

4.4. Controlling the influence of the prior on inferences

An important consideration in Bayesian analyses is how much inferences will be influenced by the prior distribution. In some cases, we may want to use weekly informative priors. In other cases (e.g., regressions involving large number of coefficients) informative priors will yield more precise estimates. The hyper-parameters are the parameters that index the prior. These need to be specified by the analyst (we will discuss extensions later on where some may be inferred from data), in this case the hyper-parameters include

$$\boldsymbol{hyper - parameters}: \{b_0, \sigma_b^2, df, S_0\}$$

Since the prior on regression coefficient is normal, the amount of information provided by this prior can be controlled by specifying the variance. Choosing a very large σ_b^2 makes the prior for

effects "flat", in this case inferences about effects will be largely driven by the information provided by the likelihood and Bayesian estimates of regression coefficients will be very close to MLE(**b**).

The expected value, mean and mode of the scaled-inverse chi-square are $E[\sigma_\varepsilon^2] = \frac{S_0}{df-2}$ and $Mode[\sigma_\varepsilon^2] = \frac{S_0}{df+2}$, respectively. The mean is defined for $df > 2$. One possibility is to choose df to be small but greater than 2 (e.g., $df = 4$) and then use either the equation for the man or the model to solve for the scale as a function of df and the expected error variance, for instance, using the equation for the mode, we can use $Var(y_i) \times (1 - R^2) \times (df + 2) = S_0$. Here, $Var(y_i)$ is the sample variance of the data and R^2 our prior guess about the proportion of variance that will be explained by the model. If the df is chosen to be small, these rules will lead to a relatively weak prior (although the scaled-inverse chi-square cannot be made strictly flat).

4.*5. Joint Posterior Distribution-I*

According to Bayes theorem the joint posterior distribution is proportional to the product of the likelihood times the prior

$$p(\boldsymbol{b}, \sigma_\varepsilon^2 | y_1, y_2, \dots, y_n) = \frac{p(y_1, y_2, \dots, y_n | \boldsymbol{b}, \sigma_\varepsilon^2)p(\boldsymbol{b}, \sigma_\varepsilon^2)}{p(y_1, y_2, \dots, y_n)} \propto p(y_1, y_2, \dots, y_n | \boldsymbol{b}, \sigma_\varepsilon^2)p(\boldsymbol{b}, \sigma_\varepsilon^2)$$

Therefore,

$$p(\boldsymbol{b}, \sigma_\varepsilon^2 | y_1, y_2, \dots, y_n) \propto$$
$$(2\pi\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij} b_j\right)^2} \left\{\prod_{j=1}^{p}(2\pi\sigma_b^2)^{-\frac{1}{2}} e^{-\frac{\left(b_j - b_0\right)^2}{2\sigma_b^2}}\right\} \frac{\left(\frac{S_0}{2}\right)^{\frac{df}{2}}}{\Gamma\left(\frac{df}{2}\right)} (\sigma_\varepsilon^2)^{-\left(1+\frac{df}{2}\right)} e^{-\frac{S_0}{2\sigma_\varepsilon^2}} \quad [7]$$

Inferences on this model are often carried out using Monte Carlo (MC) methods. There are several approaches that can be followed, here we focus on the Gibbs Sampler.

5. Gibbs Sampler-I

In a Gibbs sampler samples from the posterior distribution are collected by sampling from the fully conditional distributions. An outline of the algorithm is as follows

Box 1: Outline of Gibbs Sampler for multiple linear regression model with fix variance of effects.

```
for( i in (1:nIter)){
  for(j in 1:p){
       sample the jth regression coefficient
       from p(b_j|ELSE) (see expression [8] below)
  }
       sample the error variance
       from p(σ_ε^2|ELSE) (see expression [9] below)
}
```

Above, `nIter` is an algorithm-control-variable that determines the number of samples to be collected.

Before we present an implementation of the Gibbs sampler we need to derive the fully conditional distributions.

5.1. Fully-conditional distributions

To derive the fully-conditionals: (i) remove for posterior any proportionality constant that does not involve the unknown parameter and (ii) combine terms in search for a closed form. Since we are using conjugate priors all the fully conditionals will have closed forms.

Regression coefficients

After removing proportionality constants that do not involve b_j we get

$$p(b_k|ELSE) \propto e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}b_j\right)^2} \; e^{-\frac{\left(b_j-b_0\right)^2}{2\sigma_b^2}}$$

We now need to combine the two exponentials. The term $y_i - \sum_{j=1}^{p} x_{ij}b_j$ can be written as $y_i - \sum_{j\neq k}^{p} x_{ij}b_j - x_{ik}b_k = \tilde{y}_i - x_{ik}b_k$ where $\tilde{y}_i = y_i - \sum_{j\neq k}^{p} x_{ij}b_j$ is an 'off-set' formed by substracting from the data the contribution to the regression function of all the terms that do not involve the kth regression coefficient. Since we are sampling from $p(b_k|ELSE)$ all the regression coefficients, except the kth one, can be treated as known. Therefore, $\tilde{y}_i$ can be treated as data, thus

$$p(b_k|ELSE) \propto e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(\tilde{y}_i-x_{ik}b_k)^2} \; e^{-\frac{(b_k-b_0)^2}{2\sigma_b^2}}$$

The RSS $\sum_{i=1}^{n}(\tilde{y}_i - x_{ik}b_k)^2$ can be written as $\sum_{i=1}^{n}\tilde{y}_i^2 + b_k^2\sum_{i=1}^{n}x_{ik}{}^2 - 2b_k\sum_{i=1}^{n}x_{ik}^2\tilde{y}_i$. Likewise, the quadratic form entering in the second to the right exponential can be written as $(b_k - b_0)^2 = b_k^2 + b_0^2 - 2b_kb_0$; therefore,

$$p(b_k|ELSE) \propto e^{-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}\tilde{y}_i^2}\, e^{-\frac{\left(b_k^2\sum_{i=1}^{n}x_{ik}{}^2 - 2b_k\sum_{i=1}^{n}x_{ik}^2\tilde{y}_i\right)}{2\sigma_\varepsilon^2}}\, e^{-\frac{b_k^2-2b_kb_0}{2\sigma_b^2}}\, e^{-\frac{b_0^2}{2\sigma_b^2}}$$

The first and third exponentials do not involve the unknown parameter, thus

$$p(b_k|ELSE) \propto e^{-\frac{\left(b_k^2\sum_{i=1}^{n}x_{ik}{}^2 - 2b_k\sum_{i=1}^{n}x_{ik}\tilde{y}_i\right)}{2\sigma_\varepsilon^2}}\, e^{-\frac{b_k^2-2b_kb_0}{2\sigma_b^2}}$$

or

$$p(b_k|ELSE) \propto e^{-\frac{1}{2}\left[\frac{\left(b_k^2\sum_{i=1}^{n}x_{ik}{}^2 - 2b_k\sum_{i=1}^{n}x_{ik}\tilde{y}_i\right)}{\sigma_\varepsilon^2} + \frac{b_k^2-2b_kb_0}{\sigma_b^2}\right]}$$

We now combine the terms involving b_k^2 and $-2b_k$

$$p(b_k|ELSE) \propto e^{-\frac{1}{2}\left[b_k^2\left(\frac{\sum_{i=1}^{n}x_{ik}{}^2}{\sigma_\varepsilon^2}+\frac{1}{\sigma_b^2}\right)-2b_k\left(\frac{\sum_{i=1}^{n}x_{ik}\tilde{y}_i}{\sigma_\varepsilon^2}+\frac{b_0}{\sigma_b^2}\right)\right]}$$

or

$$p(b_k|ELSE) \propto e^{-\frac{1}{2}C\left[b_k^2-2b_k\tilde{b}_k\right]}$$

where $\tilde{b}_k = \frac{\left(\frac{\sum_{i=1}^{n}x_{ik}\tilde{y}_i}{\sigma_\varepsilon^2}+\frac{b_0}{\sigma_b^2}\right)}{\left(\frac{\sum_{i=1}^{n}x_{ik}{}^2}{\sigma_\varepsilon^2}+\frac{1}{\sigma_b^2}\right)}$ and $C = \left(\frac{\sum_{i=1}^{n}x_{ik}{}^2}{\sigma_\varepsilon^2}+\frac{1}{\sigma_b^2}\right)$. The expression above has a form very similar to the kernel of a normal distribution for the random variable b_k^2 with mean $\tilde{b}_k$ and variance C^{-1}. The term being missing in the exponential is $\tilde{b}_k^2$. This term does not involve the unknown coefficient; therefore we can write

$$p(b_k|ELSE) \propto e^{-\frac{1}{2}C\left[b_k^2-2b_k\tilde{b}_k\right]}\, e^{-\frac{1}{2}\tilde{b}_k^2}$$

Combining the two exponentials renders

$$p(b_k|ELSE) \propto e^{-\frac{1}{2}C\left[b_k^2-2b_k\tilde{b}_k+\tilde{b}_k^2\right]} \propto e^{-\frac{C}{2}\left(b_k^2-\tilde{b}_k^2\right)^2} \propto (2\pi C)^{-\frac{1}{2}}e^{-\frac{C}{2}\left(b_k^2-\tilde{b}_k\right)^2}$$

Therefore, we conclude that the fully-conditional density is normal, specifically

$$p(b_k|ELSE) = N\left(b_k|\tilde{b}_k, C^{-1}\right) \qquad [8]$$

where, as stated before $\tilde{b}_k = \frac{\left(\frac{\sum_{i=1}^{n}x_{ik}\tilde{y}_i}{\sigma_\varepsilon^2}+\frac{b_0}{\sigma_b^2}\right)}{\left(\frac{\sum_{i=1}^{n}x_{ik}{}^2}{\sigma_\varepsilon^2}+\frac{1}{\sigma_b^2}\right)}$ and $C = \left(\frac{\sum_{i=1}^{n}x_{ik}{}^2}{\sigma_\varepsilon^2}+\frac{1}{\sigma_b^2}\right)$.

Error variance

Removing from the joint posterior distribution (expression [7]) the terms that do not involve the error variance we get

$$p(\sigma_\varepsilon^2|ELSE) \propto (\sigma_\varepsilon^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_\varepsilon^2}RSS} (\sigma_\varepsilon^2)^{-\left(1+\frac{df}{2}\right)} e^{-\frac{S_0}{2\sigma_\varepsilon^2}}$$

Note that because we are sampling from the fully conditional distribution of the error variance, all the regression coefficients can be treated as known; thus, RSS can be treated as known constant.

Combining the exponentials yields

$$p(\sigma_\varepsilon^2|ELSE) \propto (\sigma_\varepsilon^2)^{-1+\frac{df+n}{2}} e^{-\frac{RSS+S_0}{2\sigma_\varepsilon^2}} = \chi^{-2}(\sigma_\varepsilon^2|df+n_0, S_0+RSS) \qquad [9]$$

The above expression is proportional to the kernel of a Scaled-inverse Chi-squared density with scale parameter $\frac{RSS+S_0}{2\sigma_\varepsilon^2}$ and degree of freedom $df + n$.

5.2. Implementation

The algorithm described in Box 1 is implemented in the script provided in gibbsMLR.md. This implementation uses a few 'computational tricks' that are summarized below

- Some quantities required to sample from expressions [8] and [9] do not vary across iterations of the sampler; therefore, we compute this quantities in advance and do not re-compute them during the execution of the sampler (an example of this is the sum of squares of each of each of the predictors).
- The mixing of the algorithm can be improved by making each of the predictors orthogonal to the incidence vector for the intercept, this is achieved by centering each predictor around its mean.
- Computing the offset $\tilde{y}_i = y_i - \sum_{j\neq k}^{p} x_{ij}b_j$ is computationally demanding. We note that $\tilde{y}_i = y_i - \sum_{j\neq k}^{p} x_{ij}b_j = \varepsilon_i + x_{ik}b_k$; therefore, we use this formulation to compute the offset needed in the computation of the mean of the fully conditional distribution of effects (expression [8]).

Note: the algorithm implemented in gibbsMLR.md is far from being optimized, we opted for leaving the algorithm as it is to maintain the connection with the code and the formulas derived here simpler. For more optimal implementations you may want to check the BGLR R-package.