The linear model

Padova, April 2011

Brunero Liseo

Sapienza Università di Roma

brunero.liseo@uniroma1.it

The base model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where

- $\mathbf{Y}$ is a vector of n independent observations,
- matrix $\mathbf{X}$ ($n \times p$) is known, (design matrix)
- $\boldsymbol{\beta}$ is a p-dimensional vector and represents the coefficients of the linear relation
- $\boldsymbol{\varepsilon}$ is a random vector.

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n).$$

Likelihood

The above assumption leads to

$$\mathbf{Y} \mid (\boldsymbol{\beta}, \sigma^2) \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 I_n\right), \tag{2}$$

and

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) \propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\,(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}. \tag{3}$$

If $\mathbf{X}$ has full rank the MLE of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

and the predicted values of $\mathbf{y}$ are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y},$$

where $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ is the projection matrix. We can then write

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &\propto \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\hat{\mathbf{y}})'\left(\mathbf{y}-\hat{\mathbf{y}}\right) + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\} \\ &= \frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(nS^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right)\right\}, \end{aligned}$$

with $nS^2 = (\mathbf{y}-\hat{\mathbf{y}})'(\mathbf{y}-\hat{\mathbf{y}}) = \sum_{j=1}^{n}(y_i - \hat{y}_i)^2$

Conjugate Bayesian Analysis

Bayesian inference implies the use of a subjective prior on $(\boldsymbol{\beta}, \sigma^2)$.
Practical considerations suggest to use a conjugate prior.

$$\pi(\boldsymbol{\beta}, \sigma^2) = \pi(\boldsymbol{\beta} \mid \sigma^2)\, \pi(\sigma^2) \tag{4}$$

with

$$\boldsymbol{\beta} \mid \sigma^2 \sim N_p(\beta_0, \sigma^2 V_0)$$

$\sigma^2 \sim GI(c_0/2, d_0/2)$. cioè

$$\pi(\sigma^2) \propto \exp\left\{-\frac{d_0}{2\sigma^2}\right\} \frac{1}{\sigma^{c_0+2}}$$

In other words, the parameter $(\boldsymbol{\beta}, \sigma^2)$ follows a Normal-Inverse Gamma distribution with hyperparameters

$$(\beta_0, V_0, c_0/2, d_0/2)$$

A posteriori,

$$\pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+\frac{c_0}{2}+\frac{p}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[nS^2 + d_0 + Q(\boldsymbol{\beta})\right]\right\}.$$

with

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'V_0^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

Working on the 2 quadratic forms in $Q(\boldsymbol{\beta})$,

$$Q(\boldsymbol{\beta}) = (\boldsymbol{\beta} - \boldsymbol{\beta}_\star)'V_\star^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_\star) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\left(\mathbf{X}'\mathbf{X}\right)V_\star V_0^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

with

$$V_\star = (\mathbf{X}'\mathbf{X} + V_0^{-1})^{-1}$$

and

$$\boldsymbol{\beta}_\star = V_\star(\mathbf{X}'\mathbf{y} + V_0^{-1}\boldsymbol{\beta}_0)$$

Setting $k = n - p$ e $nS^2 = k\tilde{S}^2$, the final distribution can be written as

$$\begin{aligned} \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \quad &\propto \quad \frac{1}{(\sigma^2)^{\frac{n+c_0}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[k\tilde{S}^2 + d_0 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\left(\mathbf{X}'\mathbf{X}\right)V_\star V_0^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]\right\} \\ &\times \quad \frac{1}{(\sigma^2)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)'V_\star^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_\star)\right]\right\}. \end{aligned}$$

It follows that the posterior of $(\boldsymbol{\beta}, \sigma^2)$ is still Normal-Inverse Gamma

$$NIG\left(\beta_\star, V_\star \frac{c_\star}{2}, \frac{d_\star}{2}\right) \tag{5}$$

with

$$c_\star = c_0 + n, \quad d_\star = d_0 + k\tilde{S}^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' \left[(\mathbf{X}'\mathbf{X})^{-1} + V_0\right]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

In particular, the conditional (on σ^2)posterior of $\boldsymbol{\beta}$ is still Gaussian

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim N\left(\boldsymbol{\beta}_\star, \sigma^2 V_\star\right)$$

and the marginal posterior of σ^2 is

$$\sigma^2 \sim GI(\frac{c_\star}{2}, \frac{d_\star}{2}).$$

The marginal posterior of the β vector is a multivariate Student

$$\boldsymbol{\beta} \sim St_p\left(c_\star, \boldsymbol{\beta}_\star, \frac{d_\star}{c_\star} V_\star\right)$$

Although not interesting here, we notice that the conditional distribution $\pi(\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y})$ will be useful in a MCMC approach. It is easy to see that

$$\sigma^2 \mid \boldsymbol{\beta}, \mathbf{y} \sim GI\left(\frac{n + c_0 + p}{2}, \frac{d_0 + k\tilde{S}^2 + Q(\boldsymbol{\beta})}{2}\right) \tag{6}$$

Noninformative analysis

There are several different criteria to select a default prior.
It is possible to show that Jeffreys' and reference priors can be respectively written as

$$\pi_J(\beta, \sigma^2) \propto \frac{1}{(\sigma^2)^{\frac{p+2}{2}}}, \qquad \pi_R(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

In what follows we use the symbol

$$\pi_\eta(\beta, \sigma^2) \propto 1/(\sigma^2)^{\eta+1};$$

η equal to $p/2$ produces a Jeffreys prior and $\eta = 0$ produces the reference prior.
The prior π_η is a limiting case of the conjugate Normal- Inverse Gamma prior;
it can be obtained by letting the prior variances go to infinity. that is setting

$$V_0^{-1} = \mathbf{0}, d_0 = 0 \text{ e } c_0 = 2\eta$$

Using the previous formulas, the noninformative posterior for (β, σ^2) is

$$\pi_\eta(\beta, \sigma^2 \mid \mathbf{y}) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+\eta+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[k\tilde{S}^2 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right]\right\},$$

It follows that

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{y} \sim N_p\left(\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right), \quad \sigma^2 \mid \mathbf{y} \sim GI(\eta + \frac{k}{2}, \frac{k\tilde{S}^2}{2}). \qquad (7)$$

Comments

the value of η plays a role only in the determination of the distribution of σ^2.
The marginal posterior of $\boldsymbol{\beta} \mid \mathbf{y}$ is always a multivariate Student t,
Only the number of degrees of freedom will depend on η.
Using Dickey's Theorem, it is easy to show that the marginal posterior of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \mathbf{y} \sim \mathrm{St}_p \left(2\eta + k, \hat{\boldsymbol{\beta}}, \frac{k\tilde{S}^2}{2\eta + k} (\mathbf{X}'\mathbf{X})^{-1} \right)$$

Comments

The number of degrees of freedom of the posterior marginal distribution of $\boldsymbol{\beta}$ is $2\eta + k$ (and not $2\eta + n$) as a direct use of the (**??**) would suggest This can be explained by the fact that, when using a flat prior on $\boldsymbol{\beta}$, the factor $1/(\sigma^2)^{p/2}$ does not appear.

The posterior marginal of $\boldsymbol{\beta}$ follows a t_p distribution also in the case of a Normal-Inverse Gamma prior: calculations are very similar to those already seen.

Remarks.

When $\eta = 0$, previous formulae remind to the classical solution. Also in this case the point estimate of $\boldsymbol{\beta}$ is given by the OLS or MLE estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X}))^{-1}\mathbf{X}'\mathbf{Y}$ and the posterior variance of $\boldsymbol{\beta}$ equals the frequentist well known quantity

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

The estimation of σ^2 gives

$$\mathbb{E}\left(\sigma^2 \mid \mathbf{y}\right) = \frac{k\tilde{S}^2}{2\eta + k - 2}, \qquad \mathrm{Var}\left(\sigma^2\right) = \frac{2k^2\tilde{S}^4}{(2\eta + k - 2)^2(2\eta + k - 4)}.$$

Example: Life Saving data

Description.
Data on the savings ratio 1960 - 1970.
Format: A data frame with 50 observations on 5 variables.

- X_1 % growth rate of dpi
- X_2 % of population under 15
- X_3 % of population over 75
- X_4 real per-capita disposable income
- Y aggregate personal savings

Under the life-cycle savings hypothesis (Modigliani), the savings ratio, or aggregate personal saving divided by disposable income, is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the % of people under 15 and the % of people over 75.
The data are averaged over the decade 1960 - 1970 to remove the business cycle or other short-term fluctuations.

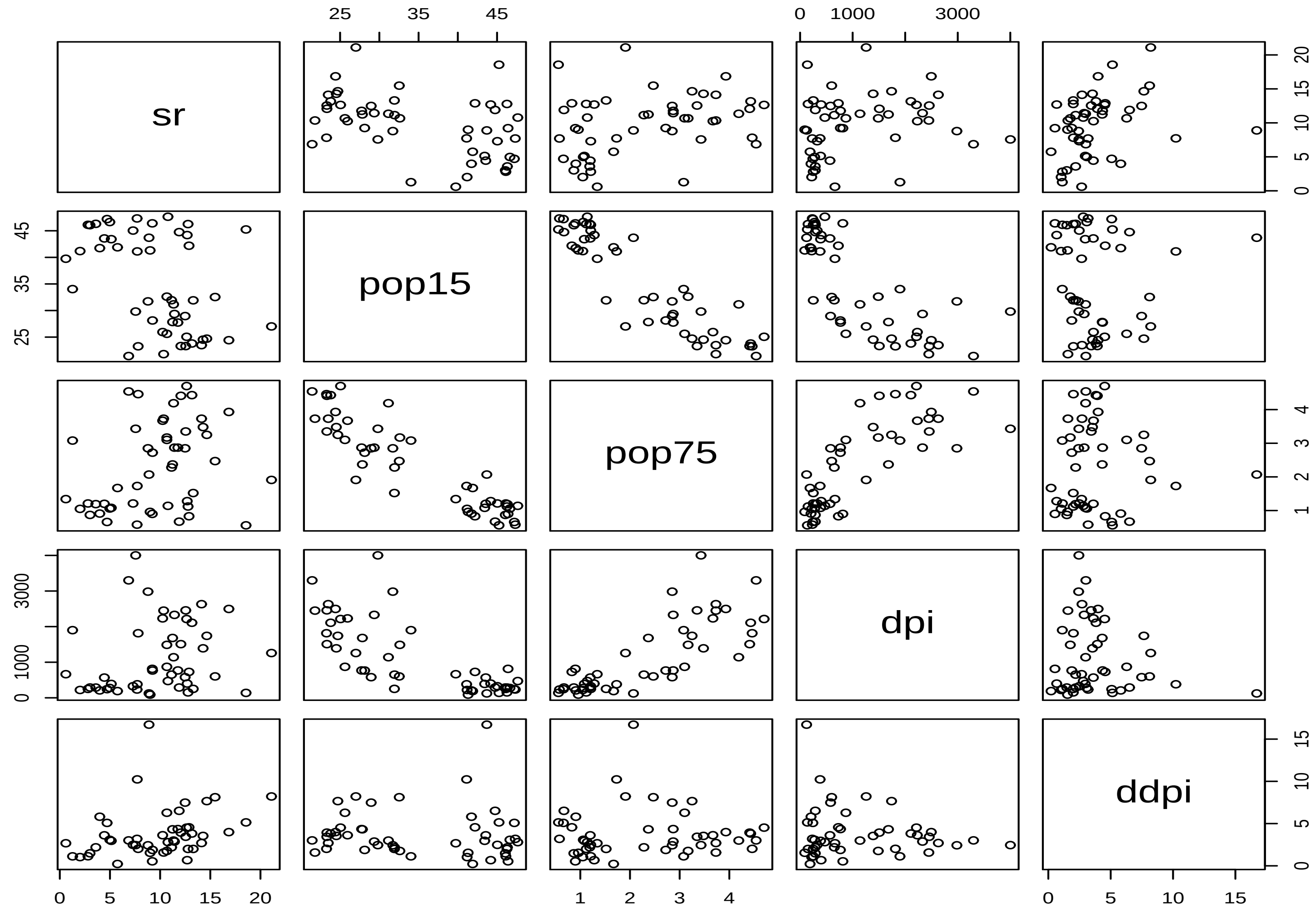
First lines of the dataset

	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56
Canada	8.79	31.72	2.85	2982.88	2.43
Chile	0.60	39.74	1.34	662.86	2.67
China	11.90	44.75	0.67	289.52	6.51
Colombia	4.98	46.64	1.06	276.65	3.08
Costa Rica	10.78	47.64	1.14	471.24	2.80
Denmark	16.85	24.42	3.93	2496.53	3.99
Ecuador	3.59	46.31	1.19	287.77	2.19
Finland	11.24	27.84	2.37	1681.25	4.32
France	12.64	25.06	4.70	2213.82	4.52
Germany	12.55	23.31	3.35	2457.12	3.44

Summary

```
      sr              pop15            pop75      
 Min.   : 0.600   Min.   :21.44   Min.   :0.560  
 1st Qu.: 6.970   1st Qu.:26.21   1st Qu.:1.125  
 Median :10.510   Median :32.58   Median :2.175  
 Mean   : 9.671   Mean   :35.09   Mean   :2.293  
 3rd Qu.:12.617   3rd Qu.:44.06   3rd Qu.:3.325  
 Max.   :21.100   Max.   :47.64   Max.   :4.700  
      dpi               ddpi       
 Min.   :  88.94   Min.   : 0.220  
 1st Qu.: 288.21   1st Qu.: 2.002  
 Median : 695.66   Median : 3.000  
 Mean   :1106.76   Mean   : 3.758  
 3rd Qu.:1795.62   3rd Qu.: 4.478  
 Max.   :4001.89   Max.   :16.710  
```

Graphics



MCMCpack

One can use the **R** suite `MCMCpack`

```
library(MCMCpack)
fm1.bayes <- MCMCregress(sr ~ pop15 + pop75 + dpi + ddpi,
data = LifeCycleSavings)
summary(fm1.bayes)
plot(fm1.bayes)
```

MCMCregress

```
Usage

MCMCregress(formula, data = NULL, burnin = 1000, mcmc = 10000,
   thin = 1, verbose = 0, seed = NA, beta.start = NA,
   b0 = 0, B0 = 0, c0 = 0.001, d0 = 0.001,
   marginal.likelihood = c("none", "Laplace", "Chib95"), ...)
Arguments
```

output

```
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

                 Mean        SD  Naive SE Time-series SE
(Intercept) 28.535956 7.4890169 7.489e-02      8.503e-02
pop15       -0.460344 0.1471324 1.471e-03      1.673e-03
pop75       -1.685494 1.1176515 1.118e-02      1.169e-02
dpi         -0.000336 0.0009593 9.593e-06      1.013e-05
ddpi         0.405780 0.2002496 2.002e-03      2.001e-03
sigma2      15.137668 3.3598233 3.360e-02      3.385e-02
```

output

```
2. Quantiles for each variable:

                 2.5%        25%        50%        75%     97.5%
(Intercept) 13.806656 23.5205246 28.5775909 33.4014431 43.394205
pop15       -0.753139 -0.5572213 -0.4609449 -0.3617458 -0.169847
pop75       -3.923511 -2.4273591 -1.6802012 -0.9367293  0.480088
dpi         -0.002168 -0.0009834 -0.0003408  0.0003106  0.001555
ddpi         0.010956  0.2732593  0.4079713  0.5391184  0.803095
sigma2       9.912009 12.7346800 14.6652476 16.9943656 22.976855
```

Elicitation problems: the g-priors

In practical applications, not easy to translate prior information into a value for the hyperparameters of the prior.
Particularly true for the covariance matrix $\mathbf{V}_0$.
Sometimes one can use a flat prior. Some other times we suspect some form of association among the β coefficients and the flat prior would not work.
A very popular solution was proposed by Zellner (1986): it is based on an empirical Bayesian idea: one can use

- the usual default prior for σ^2
- a Normal proper prior for $\boldsymbol{\beta}$

$$\boldsymbol{\beta} \sim N_p\left(\boldsymbol{\beta}_0, c\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right).$$

$\boldsymbol{\beta}_0$ is usually set equal to $\mathbf{0}$.

The role of c: The choice of c is done in terms of the relative weight of the prior information with respect to the sample evidence.
For example, $c = 5$ implies that the prior weight is one fifth of the likelihood weight.
This way we avoid to elicit the entire prior covariance matrix which is set equal to a multiple of the empirical covariance matrix
The analytical treatment of the linear model with g-priors is a particular case of the above.
For example, one can easily seen that the marginal posterior of $\boldsymbol{\beta}$ is

$$St_p\left(k, \boldsymbol{\beta}_\star, \frac{c}{1+c}(\mathbf{X}'\mathbf{X})^{-1}\right).$$

Hypothesis Testing

Hypothesis testing is performed in terms of Bayes factor between the competing hypotheses or models.
In this sense, we need to compute the marginal distribution of the data $\mathbf{y}$ under the two models.

For example, testing $\beta_1 = 0$ vs. $\beta_1 \neq 0$ would reduce to compare models

$$M_0 : Y = \beta_0 + \beta_2 X_2 + \cdots + \varepsilon$$

versus

$$M_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \varepsilon$$

The marginal distribution of $\mathbf{y}$

Crucial saper esprimere in forma analitica la quantità

$$p(\mathbf{y}) = \int_{\mathbb{R}^p} \int_0^\infty p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta} \mid \sigma^2)\pi(\sigma^2) d\boldsymbol{\beta} d\sigma^2.$$

A simple technique to get this distribution when the prior is Inverse Gamma Normal, i.e.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n), \quad \boldsymbol{\beta}|\sigma^2 \sim N_p(\boldsymbol{\beta}_0, \sigma^2 \mathbf{V}_0), \quad \sigma^2 \sim GI(\frac{c_0}{2}, \frac{d_0}{2}),$$

is the following: since

$$\mathbf{X}\,\boldsymbol{\beta} \mid \sigma^2 \sim N_n(\mathbf{X}\beta_0, \sigma^2 \mathbf{X}\mathbf{V}_0\mathbf{X}')$$

using the multidimensional version of the Lemma,

$$\mathbf{Y} \mid \sigma^2 \sim N_n\left(X\boldsymbol{\beta}_0, \sigma^2(\mathbf{I}_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')\right).$$

Using Dickey's Thm, $\mathbf{Y}$ has a

$$St_n(c_0, \mathbf{X}\boldsymbol{\beta}_0, \frac{d_0}{c_0}\left(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}'\right),$$

that is

$$m(\mathbf{y}) = \frac{d_0^{c_0/2}\Gamma((c_0+n)/2)/\Gamma(c_0/2)}{(\pi)^{n/2} \mid I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}' \mid^{1/2}} \left[d_0 + G(\mathbf{y}, \boldsymbol{\beta}_0, \mathbf{X})\right]^{-\frac{n+c_0}{2}}, \tag{8}$$

where

$$G(\mathbf{y}, \boldsymbol{\beta}_0, \mathbf{X}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)'(I_n + \mathbf{X}\mathbf{V}_0\mathbf{X}')^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0). \tag{9}$$

Prediction

After calibrating the regression model, one might need to use the model to predict a new value $\mathbf{Y}_0$ based on some covariates values $\mathbf{X}_0$. From a formal perspective one needs to obtain the distribution of $\mathbf{Y}_0$ given ALL the information on the parameters $p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0)$. One can write

$$p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) = \int_{\mathbb{R}^p} \int_0^\infty p(\mathbf{y}_0, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) d\boldsymbol{\beta} d\sigma^2$$

$$= \int_{\mathbb{R}^p} \int_0^\infty p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0, \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{X}_0, \mathbf{y}) d\boldsymbol{\beta} d\sigma^2$$

- First term: likelihood related to new values $\mathbf{y}_0$ (there is no conditioning on $\mathbf{y}$ and $\mathbf{X}$;
- Second term: posterior distribution of the parameters of the previous model (no conditioning upon $\mathbf{X}_0$).

This problem is identical to that of the derivation of the marginal distribution of $\mathbf{y}$, already discussed.
One can show that

The predictive distribution of $\mathbf{Y}_0$ is

$$St_r\left(2c_\star, \mathbf{X}_0'\boldsymbol{\beta}_\star, \frac{d_\star}{c\star}(\mathbf{I}_r + \mathbf{X}_0(\mathbf{X}'\mathbf{X} + \mathbf{V}_0^{-1})^{-1}\mathbf{X}_0')\right),$$

The above model was simple enough to allow a closed form analysis.
It is sufficient to modify some assumption (e.g. heteroscedasticity, serial correlation, panel data, etc) to be bound to abandon the analytical road and use Bayesian computation.
Here we describe the steps to obtain a posterior sample from the distribution of $\mathbf{Y}_0$ under the previous hypotheses: more complex modes will require specific adjustements.
However the philosophy of the approach will not change.

For $t = 1, \cdots, M$,

- draw $\boldsymbol{\beta}_t \sim \pi(\boldsymbol{\beta} \mid \sigma_{(t-1)}, \mathbf{y})$ (Gaussian)
- draw $\sigma_t \sim \pi(\sigma \mid \mathbf{y})$ (Inverse Gamma)
- draw $\mathbf{y}_t \sim \pi(\mathbf{y} \mid \boldsymbol{\beta}_t, \sigma_t)$ (Gaussian)

Variable selection in a linear model

it is not always reasonable to include all the covariates into the regression model

- overfitting (n must be $>> p$)
- multicollinearity

Formalisation of the problem

We have a vector of observations Y and p covariates $(\mathbf{x}_1, \ldots, \mathbf{x}_p)$.
Each single covariate may be or may be not included in the model - we will always include the intercept - the total number of possible models is 2^{p-1}.
The generic model $M_{\boldsymbol{\gamma}}$ is

$$M_{\boldsymbol{\gamma}} : \mathbf{Y} = \mathbf{X}_{\boldsymbol{\gamma}} \beta_{\boldsymbol{\gamma}} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$$

and

$$\gamma_1 = 1, \qquad \gamma_j = 0, 1 \qquad\qquad j = 2, \ldots, p$$

Each model is then identified by the included variables, that is by the value of the parameter γ, which is a vector of 0's and 1's.
The Bayesian analyis of this problem needs then the elicitation of

- a probability distribution for γ.
- given γ, a probabitity distribution for the parameters of model $M_{\boldsymbol{\gamma}}$, that is $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ e σ^2

Notice that σ^2 is considered “equal" across different models: it lost its meaning of residual variance, and it will be only the variance of the random component

Prior for γ

Usually one uses

$$\Pr(M_{\boldsymbol{\gamma}}) = \frac{1}{2^{p-1}}, \qquad \forall \boldsymbol{\gamma}$$

(all models have the same prior probability) or, alternatively

$$\Pr(M_{\boldsymbol{\gamma}}) \propto \frac{1}{q(\boldsymbol{\gamma})}, \qquad \text{dove } q(\boldsymbol{\gamma}) = \#1 \text{ in } \boldsymbol{\gamma}$$

(overfitting penalization)

Priors for the single models

Given the model $M_{\boldsymbol{\gamma}}$, we assume for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2$ a Zellenr g-prior, that is

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2 \sim N_{q(\gamma)}\left(\mathbf{0}, c\sigma^2(\mathbf{X}'_\gamma \mathbf{X}_{\boldsymbol{\gamma}})^{-1}\right); \pi(\sigma^2) \propto \frac{1}{\sigma^2},$$

where $q(\gamma)$ is the number of covariates in the model $M_{\boldsymbol{\gamma}}$.
This choice avoids computational difficulties and it allows to provide a simple interpretation of the quantities. This is quite important in the presence of thousands of models.
The choice of an improper prior for σ^2 è is allowed by the fact that σ^2 is present in ALL the competing models.

Posterior probability of single models

Simple calculations shows that

$$\Pr\left(M_{\boldsymbol{\gamma}}|\mathbf{y}\right) = \frac{\Pr\left(M_{\gamma}\right)\ m_{\gamma}(\mathbf{y})}{\sum_{\delta\in\Gamma}\Pr\left(M_{\delta}\right)\ m_{\delta}(\mathbf{y})}$$

where

$$m_{\gamma}(\mathbf{y}) = \int_{\mathbb{R}^{q(\gamma)}}\int_0^{\infty} p(\mathbf{y} \mid M_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid M_{\boldsymbol{\gamma}}, \sigma^2)\pi(\sigma^2 \mid M_{\boldsymbol{\gamma}})d\boldsymbol{\beta}_{\boldsymbol{\gamma}}d\sigma^2.$$

Then, the marginal distribution of $\mathbf{y}$ under each single model M_{γ} is crucial.

Variabil selection with g-priors

Using g-priors will correspond to set, for each possible model $M_{\boldsymbol{\gamma}}$, $\boldsymbol{\gamma} \in \Gamma$,

$$\boldsymbol{\beta}_0^{(\boldsymbol{\gamma})} = \mathbf{0}, \quad V_0^{(\boldsymbol{\gamma})} = c(\mathbf{X}'_{(\boldsymbol{\gamma})}\mathbf{X}_{(\boldsymbol{\gamma})})^{-1}, \quad c_0 = d_0 = 0.$$

Then

$$m_{\boldsymbol{\gamma}}(y) = \int_0^\infty N_n\left(\mathbf{0}, \sigma^2(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}})\right) \frac{1}{\sigma^2} d\sigma^2,$$

where

$$\mathbf{H}_{\boldsymbol{\gamma}} = \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}'_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}'_{\boldsymbol{\gamma}}$$

Since

$$(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}})^{-1} = \mathbf{I}_n - \frac{c}{c+1}\mathbf{H}_{\boldsymbol{\gamma}}$$

e

$$\det(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}}) = (c+1)^p$$

we will obtain that $m_{\boldsymbol{\gamma}}(y) =$

$$\int_0^\infty \frac{|\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\mathbf{y}'\left(\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}}\right)^{-1}\mathbf{y}\right\}\frac{1}{\sigma^2}d\sigma^2$$

$$= \int_0^\infty \frac{|\mathbf{I}_n + c\mathbf{H}_{\boldsymbol{\gamma}}|^{-\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - \mathbf{y}'\frac{c}{c+1}\mathbf{H}_\gamma\mathbf{y}\right\}\frac{1}{\sigma^2}d\sigma^2$$

$$= \frac{\Gamma(n/2)}{\pi^{n/2}(c+1)^{\frac{p}{2}}}\left[\mathbf{y}'\mathbf{y} - \frac{c}{c+1}\mathbf{y}'\mathbf{H}_\gamma\mathbf{y}\right]^{-\frac{n}{2}}$$

This way all the marginal densities of the data $\mathbf{y}$, under all possible models, are available in closed form.
It is then easy to evaluate the posterior probability of the single $M_{\boldsymbol{\gamma}}$.

Example:Life saving data

We have four covariates, that is 2^4 models. Setting $c = 5$ and assuming an improper prior for σ^2 one can easily get the posterior probability of the 16 models

mod	c0	c12	c13	14	c15	c123	c124	c125
prob	0.0364	0.0522	0.0573	0.0441	0.0576	0.0716	0.0527	0.0713

mod	c134	c135	c145	c1234	c1235	c1245	c1345	ctutte
prob	0.0713	0.0578	0.0694	0.0718	0.0718	0.0717	0.0713	0.0718

This approach has some limitations.
Suppose to compare the “full" model, M_1, including all the covariates, versus the model without covariates, M_0 (i.e. $\boldsymbol{\beta} = \mathbf{0}$).
One can show that, in this case, as the OLS estimate goes to infinity (that is when the evidence against the null hypothesis is overwhelming), the Bayes factor B_{01} will converge to the constant

$$(1 + c)^{\frac{p-n}{2}}.$$

The essence of this result is related to the well known Lindley’s paradox.
Alternative choices for the priors in model selection are given by Intrinsic or Fractional priors: see the book by O’Hagan and Forster (2004).

Large values of k

Stochastic search for the most likely model

- When k gets large, impossible to compute the posterior probabilities of the 2^k models.
- Need of a tailored algorithm that samples from $\pi(\gamma|y, X)$ and selects the most likely models.
- Can be done by Gibbs sampling, given the availability of the full conditional posterior probabilities of the γ_i's.

Let $\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_k)$ ($1 \leq i \leq k$)

$$\pi(\gamma_i|\mathbf{y}, \gamma_{-i}, \mathbf{X}) \propto \pi(\gamma|\mathbf{y}, \mathbf{X})$$

(to be evaluated in both $\gamma_i = 0$ and $\gamma_i = 1$)

Gibbs sampling for variable selection

First note that each γ_i is a Bernoulli random variable with probability of success

$$\frac{\pi(\gamma_i = 1|\mathbf{y}, \gamma_{-i}, \mathbf{X})}{\pi(\gamma_i = 0|\mathbf{y}, \gamma_{-i}, \mathbf{X}) + \pi(\gamma_i = 1|\mathbf{y}, \gamma_{-i}, \mathbf{X})}$$

Initialization: Draw γ^0 from the uniform distribution on Γ

Iteration t: Given $(\gamma_1^{(t-1)}, \ldots, \gamma_k^{(t-1)})$, generate

1. $\gamma_1^{(t)}$ according to $\pi(\gamma_1|y, \gamma_2^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$

2. $\gamma_2^{(t)}$ according to $\pi(\gamma_2|y, \gamma_1^{(t)}, \gamma_3^{(t-1)}, \ldots, \gamma_k^{(t-1)}, X)$

$\vdots$

p. $\gamma_k^{(t)}$ according to $\pi(\gamma_k|y, \gamma_1^{(t)}, \ldots, \gamma_{k-1}^{(t)}, X)$

MCMC interpretation

After $T \gg 1$ MCMC iterations, we have, as output of the model, a posterior simulation from the distribution of γ in Γ.
It can be used for many purposes:

- To approximate the posterior probabilities $\pi(\gamma|y, X)$ by empirical averages

  $$\widehat{\pi}(\gamma|y, X) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma^{(t)}=\gamma}.$$

  (the T_0 first values are eliminated as *burnin*).

- to approximate the probability to include i-th variable in the model,

  $$\widehat{P}^{\pi}(\gamma_i = 1|y, X) = \left(\frac{1}{T - T_0 + 1}\right) \sum_{t=T_0}^{T} \mathbb{I}_{\gamma_i^{(t)}=1}.$$

Comments

- The algorithm can be easily implemented.
- Since the number of points in Γ can nbe even larger than T, many models will be never visited by the Markov chain.
- We are confident that most likely models are those more often visited by the algorithm
- This approach is more coherent than backward or forward or stepwise approaches. These other methods are able to find a local maximum rather than a global one.