BAYESIAN ECONOMETRICS SPRING 2013

TAKE HOME MIDTERM EXAM **SOLUTION**

Let us continue in the context of `homework 1` and `homework 2`, where we modeled the relationship between per capita spending (y) on public schools as a linear function of per capita income (x). The data is in the file `spending.txt` and

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \sim (0, \sigma^2),$$

where we entertain three models:

$\mathcal{M}_1$: $\epsilon_i \sim N(0, \sigma^2)$ and $\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0)$

$\mathcal{M}_2$: $\epsilon_i \sim N(0, \sigma^2)$ and $\beta \sim N(b_0, B_0)$

$\mathcal{M}_3$: $\epsilon_i \sim t_\nu(0, \sigma^2)$ and $\beta|\sigma^2 \sim N(b_0, \sigma^2 B_0)$

with $\sigma^2 \sim IG(\eta_0/2, \eta_0 s_0^2/2)$ for $\mathcal{M}_i, i = 1, 2, 3$, and $\nu = 4.46$ for $\mathcal{M}_3$. The hyperparameters b_0, B_0, a and b are obtained such that, *a priori*, $E(\beta|\mathcal{M}_i) = (-70, 600)'$, $V(\beta|\mathcal{M}_i) = 10000 I_2$, $E(\sigma^2|\mathcal{M}_i) = 3750$ and $V(\sigma^2|\mathcal{M}_i) = 1562500$, for $i = 1, 2, 3$.

a) Compute Bayes factors B_{12}, B_{13} and B_{23}.

b) Draw $p(y_{new}|x_{new} = 8000, x, y, \mathcal{M}_i)$ for $i = 1, 2, 3$.

c) Draw $p(y_{new}|x_{new} = 8000, x, y)$.

d) Compute the DIC($\mathcal{M}_i$), for $i = 1, 2, 3$.

Note:

1) Recall that $B_{ij} = p(y|x, \mathcal{M}_i)/p(y|x, \mathcal{M}_j)$.

2) In a) $p(y|x, \mathcal{M}_1)$ has to be derived analytically.

3) In b) $p(y_{new}|x_{new} = 8000, x, y, \mathcal{M}_1)$ has to be derived analytically.

Solution

Prior distributions

For all models ($i = 1, 2, 3$), the prior of σ^2 is the same

$$
\begin{aligned}
E(\sigma^2|\mathcal{M}_i) &= 3750 = \frac{\eta_0 s_0^2/2}{\eta_0/2-1} = \frac{\eta_0}{\eta_0-2}s_0^2 \\
V(\sigma^2|\mathcal{M}_i) &= 1562500 = \frac{\{E(\sigma^2|\mathcal{M}_i)\}^2}{(\eta_0/2-2)} = \frac{7500}{\eta_0-4},
\end{aligned}
$$

so $\eta_0 = 4.0048$ and $s_0^2 = 1877$ (rounded to the nearest integer). The hyperparameter $b_0 = (-70, 600)'$ is the same for all models. However, B_0 varies with the model structure. For model $\mathcal{M}_2$, $V(\beta) = B_0 = 10000I_2$, while for models $\mathcal{M}_1$ and $\mathcal{M}_3$, $V(\beta) = \eta_0/(\eta_0-2)s_0^2 B_0$, so $B_0 = 2.667I_2$.

Posterior distributions

Model $\mathcal{M}_1$. This is the textbook normal linear regression model with conjugate priors, which leads to

$$
\begin{aligned}
\beta|\sigma^2, y, x, \mathcal{M}_1 &\sim N(b_1, \sigma^2 B_1) \\
\sigma^2|y, x, \mathcal{M}_1 &\sim IG(\eta_1/2, \eta_1 s_1^2/2),
\end{aligned}
$$

where $\eta_1 = \eta_0 + n$,

$$
\begin{aligned}
B_1^{-1} &= B_0^{-1} + X'X \\
B_1^{-1}b_1 &= B_0^{-1}b_0 + X'y \\
\eta_1 s_1^2 &= \eta_0 s_0^2 + (y - Xb_1)'y + (b_0 - b_1)'B_0^{-1}b_0,
\end{aligned}
$$

and X is a $(n \times 2)$ matrix with ones in the first column and $x = (x_1, \ldots, x_n)'$ in the second column, and $y = (y_1, \ldots, y_n)'$. It also follows that $\beta|y, x, \mathcal{M}_1 \sim t_{\eta_1}(b_1, s_1^2 B_1)$.

Model $\mathcal{M}_2$. Model $\mathcal{M}_2$ is the normal linear regression model with conditionally conjugate priors, with full conditionals given by

$$
\begin{aligned}
\beta|\sigma^2, y, x, \mathcal{M}_2 &\sim N(b_1, B_1) \\
\sigma^2|\beta, y, x, \mathcal{M}_2 &\sim IG(\eta_1/2, \eta_1 s_1^2/2),
\end{aligned}
$$

where $\eta_1 = \eta_0 + n$ and

$$\begin{aligned} B_1^{-1} &= B_0^{-1} + X'X/\sigma^2 \\ B_1^{-1}b_1 &= B_0^{-1}b_0 + X'y/\sigma^2 \\ \eta_1 s_1^2 &= \eta_0 s_0^2 + (y - X\beta)'(y - X\beta)/\sigma^2. \end{aligned}$$

The Gibbs sampler iterates between these two full conditionals and, after convergence of the MCMC chain, produces draws from $p(\beta, \sigma^2|y, x, \mathcal{M}_2)$. Notice that $B_1 \equiv B_1(\sigma^2)$, $b_1 \equiv b_1(\sigma^2)$ and $s_1^2 \equiv s_1^2(\beta)$.

Model $\mathcal{M}_3$. We can use the same data augmentation argument applied to derived the Gibbs sampler for posterior inference in homework assignment 2. More precisely, the error term $\epsilon_i \sim t_\nu(0, \sigma^2)$ is replaced (by data augmentation) by the pair $\epsilon_i \sim N(0, \lambda_i\sigma^2)$ and $\lambda_i \sim IG(\nu/2, \nu/2)$. Then, it can be shown that

$$\begin{aligned} \beta|\sigma^2, y, x, \lambda, \mathcal{M}_3 &\sim N(b_1, \sigma^2 B_1) \\ \sigma^2|y, x, \lambda, \mathcal{M}_3 &\sim IG(\eta_1/2, \eta_1 s_1^2/2), \end{aligned}$$

where $\eta_1 = \eta_0 + n$,

$$\begin{aligned} B_1^{-1} &= B_0^{-1} + X'\Lambda^{-1}X \\ B_1^{-1}b_1 &= B_0^{-1}b_0 + X'\Lambda^{-1}y \\ \eta_1 s_1^2 &= \eta_0 s_0^2 + (y - Xb_1)'\Lambda^{-1}y + (b_0 - b_1)'B_0^{-1}b_0, \end{aligned}$$

$\lambda = (\lambda_1, \ldots, \lambda_n)'$ and $\Lambda = \text{diag}(\lambda)$. Therefore, the Gibbs sampler is such that β and σ^2 are sampled jointly and conditionally on λ. The full conditional distribution of λ_i, for $i = 1, \ldots, n$ is given by

$$\lambda_i|y_i, x_i, \beta, \sigma^2, \mathcal{M}_3 \sim IG\left(\frac{\nu + 1}{2}, \frac{\nu + (y_i - \beta_0 - \beta_1 x_i)^2/\sigma^2}{2}\right).$$

Notice that $b_1 \equiv b_1(\lambda)$, $B_1 \equiv B_1(\lambda)$ and $s_1^2 \equiv s_1^2(\lambda)$.

Predictives

Model $\mathcal{M}_1$. It can be shown (we've shown in class!) that the prior predictive density and that the posterior predictive for a new observation y_{n+1} are given by

$$\begin{aligned} p(y|x, \mathcal{M}_1) &= p_t(y; Xb_0, s_0^2(I_n + XB_0X'), \eta_0) \\ p(y_{n+1}|x_{n+1}, y, x, \mathcal{M}_1) &= p_t(y_{n+1}; \tilde{x}'b_1, s_1^2(1 + \tilde{x}'B_1\tilde{x}), \eta_1), \end{aligned}$$

respectively, where $\tilde{x} = (1, x_{n+1})'$ and $p_t(y; \mu, \sigma^2)$ is the density of a (univariate or multivariate) Student's t distribution with location μ and scale σ^2 evaluated at y.

Model $\mathcal{M}_2$. Prior and posterior draws of σ^2 can be used to approximate, by Monte Carlo integration, $p(y|x, \mathcal{M}_2)$ and $p(y_{n+1}|x_{n+1}, x, y, \mathcal{M}_2)$ (this is raoblackwellization in action!). More precisely, it is easy to see that

$$\begin{aligned} p(y|x, \sigma^2, \mathcal{M}_2) &= p_n(y; Xb_0, \sigma^2 I_n + XB_0X') \\ p(y_{n+1}|x_{n+1}, y, x, \mathcal{M}_2, \sigma^2) &\equiv p_n(y_{n+1}; \tilde{x}b_1, \sigma^2 + \tilde{x}'B_1\tilde{x}), \end{aligned}$$

where $\tilde{x} = (1, x_{n+1})'$. Here $p_n(y; \mu, \sigma^2)$ is the density of a (univariate or multivariate) normal distribution with mean μ and variance σ^2 evaluated at y. The MC approximations to the prior and posterior predictive densities are

$$\begin{aligned} p_{MC}(y|x, \mathcal{M}_2) &= \frac{1}{M}\sum_{i=1}^{M} p_n(y; Xb_0, \tilde{\sigma}^{2(i)} I_n + XB_0X') \\ p_{MC}(y_{n+1}|x_{n+1}, y, x, \mathcal{M}_2) &= \frac{1}{M}\sum_{i=1}^{M} p_n(y_{n+1}; \tilde{x}b_1^{(i)}, \sigma^{2(i)} + \tilde{x}'B_1^{(i)}\tilde{x}), \end{aligned}$$

respectively, where $\{\tilde{\sigma}^{2(i)}\}_{i=1}^M$ are draws from the prior $p(\sigma^2|\mathcal{M}_2)$, and $\{\sigma^{2(i)}\}_{i=1}^M$ are draws from the posterior $p(\sigma^2|y, x, \mathcal{M}_2)$ and the pairs $\{(b_1, B_1)^{(i)}\}_{i=1}^M$ are moments of the full conditional of $\beta|\sigma^2, \lambda$, all obtained via a Gibbs sampler (as in homework assignment 2). Finally, from standard matrix algebra, it can be shown that

$$\left(\sigma^2 I_n + XB_0X'\right)^{-1} = \sigma^{-2}\left(I_n + X(\sigma^2 B_0^{-1} + X'X)^{-1}X'\right),$$

with the left-hand side involving the inversion of a $n \times n$ matrix and the right-hand side involving the inversion of a much smaller 2×2 matrix. This will make the computation of $\hat{p}(y|x, \mathcal{M}_2)$ obviously faster.

Model $\mathcal{M}_3$

Conditional on λ and λ_{n+1}, it follows from the derivations under model $\mathcal{M}_1$ that:

$$\begin{aligned} p(y|x, \lambda, \mathcal{M}_3) &= p_t(y; Xb_0, s_0^2(\Lambda + XB_0X'), \eta_0) \\ p(y_{n+1}|x_{n+1}, y, x, \lambda, \lambda_{n+1}, \mathcal{M}_3) &= p_t(y_{n+1}; \tilde{x}'b_1, s_1^2(\lambda_{n+1} + \tilde{x}'B_1\tilde{x}), \eta_1), \end{aligned}$$

where, again, $\tilde{x} = (1, x_{n+1})'$. Therefore, the MC approximations to the prior and posterior predictive densities are

$$\begin{aligned} p_{MC}(y|x, \mathcal{M}_3) &= \frac{1}{M}\sum_{i=1}^{M} p_t(y; Xb_0, s_0^2(\tilde{\Lambda}^{(i)} + XB_0X'), \eta_0) \\ p_{MC}(y_{n+1}|x_{n+1}, y, x, \mathcal{M}_3) &= \frac{1}{M}\sum_{i=1}^{M} p_t(y_{n+1}; \tilde{x}'b_1^{(i)}, s_1^{2(i)}(\lambda_{n+1}^{(i)} + \tilde{x}'B_1^{(i)}\tilde{x}), \eta_1^{(i)}), \end{aligned}$$

respectively, where $\{\tilde{\lambda}^{(i)}\}_{i=1}^M$ and $\{\lambda_{n+1}^{(i)}\}_{i=1}^M$ are i.i.d. draws from $IG(\nu/2, \nu/2)$. The quantities $\{(\eta_1, s_1^2, b_1, B_1)^{(i)}\}_{i=1}^M$ are full conditional sufficient statistics obtained via Gibbs sampler and all functions of posterior draws $\{\lambda^{(i)}\}_{i=1}^M$.

DIC

Recall that the deviance information criterion is defined as

$$DIC(\mathcal{M}) = -4E_{\theta|y,x,\mathcal{M}}\{\log p(y|x,\theta,\mathcal{M})\} + 2\log p(y|x,\hat{\theta},\mathcal{M})$$

where $\hat{\theta} = E(\theta|x,y,\mathcal{M})$. For $\mathcal{M}_1$, it follows that $\hat{\theta} = b_1$. For $\mathcal{M}_3$, $\hat{\theta}$ can be approximated via MC integration by

$$\frac{1}{M}\sum_{i=1}^M E(\theta|x,y,\lambda^{(i)},\mathcal{M}_3) = \frac{1}{M}\sum_{i=1}^M b_1(\lambda^{(i)}),$$

where $\lambda^{(i)}$, for $i = 1,\ldots,M$, are draws from $p(\lambda|y,x,\mathcal{M}_3)$, obtained via a Gibbs sampler (as in homework assignment 2). For $\mathcal{M}_2$, $\hat{\theta}$ can be approximated, also via MC integration, by

$$\frac{1}{M}\sum_{i=1}^M E(\theta|x,y,\sigma^{(i)},\mathcal{M}_2) = \frac{1}{M}\sum_{i=1}^M b_1(\sigma^{2(i)})$$

where $\sigma^{2(i)}$, for $i = 1,\ldots,M$, are draws from $p(\sigma^2|y,x,\mathcal{M}_2)$, obtained via a Gibbs sampler (as in homework assignment 2).

The log-likelihood densities are

$$L_1 = c_1 - 0.5n\log\sigma^2 - 0.5\sum_{i=1}^n \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}$$

for models $\mathcal{M}_1$ and $\mathcal{M}_2$ and $c_1 = -0.5n\log 2\pi$, and

$$L_3 = c_3 - 0.5n\log\sigma^2 - 0.5(\nu+1)\sum_{i=1}^n \log\left(1 + \frac{1}{\nu}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}\right)$$

for model $\mathcal{M}_3$ and $c_3 = n(\log\Gamma((\nu+1)/2) - \log\Gamma(\nu/2)) - 0.5n\log\pi\nu$. Therefore, the posterior expectation of L_1 (for $i = 1, 2$) is

$$\begin{aligned} c_1 \quad &- \quad 0.5n\left\{E(\log\sigma^2|x,y,\mathcal{M}_i) + E(\beta_0\sigma^{-2}|x,y,\mathcal{M}_i)\right\} \\ &- \quad 0.5\left(\sum_{i=1}^n y_i\right)E(\sigma^{-2}|x,y,\mathcal{M}_i) - 0.5\left(\sum_{i=1}^n x_i\right)E(\beta_1\sigma^{-2}|x,y,\mathcal{M}_i), \end{aligned}$$

an the posterior expectation of L_3 is

$$c_3 - 0.5nE(\log\sigma^2|x,y,\mathcal{M}_3) - 0.5(\nu+1)\sum_{i=1}^{n} E\left\{\log\left(1+\frac{1}{\nu}\frac{(y_i-\beta_0-\beta_1 x_i)^2}{\sigma^2}\right) \mid y,x,\mathcal{M}_3\right\}.$$

Results

Summary of the posterior distributions for each model appear in Table 1, while the posterior predictive densities for y_{new} given $x_{new} = 8000$, i.e. $p(y_{new}|x_{new} = 8000, x, y, \mathcal{M}_i)$ for $i = 1, 2, 3$, appear in Figure 1. The Bayes factors are $B_{31} = 2955806$, $B_{32} = 2.870907e^{192}$ and $B_{12} = 9.712772e^{185}$. Following Jeffreys (1961) recommendations, there is decisive evidence against models $\mathcal{M}_1$ and $\mathcal{M}_2$. In addition, when prior model probabilities are uniform, i.e. $Pr(\mathcal{M}_i) = 1/3$, for $i = 1, 2, 3$, then posterior model probabilities are

$$Pr(\mathcal{M}_i|y,x) = \frac{1}{\sum_{j=1}^{3} B_{ji}} \qquad i = 1, 2, 3,$$

or $Pr(\mathcal{M}_3|y,x) = 0.9999996616829 = 1 - Pr(\mathcal{M}_1|y,x)$. The DICs for models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ are, respectively, -336.6, -328.2 and 551.0030, suggesting that model $\mathcal{M}_1$ is the best of the three models.

Model $\mathcal{M}_1$ - $\log p(y\|x, \mathcal{M}_1) = 145.7119$					
Parameter	Mean	Standard deviation	Percentiles 2.5%	Median	97.5%
β_0	-112.9	43.9	-200.9	-112.9	-24.9
β_1	639.4	57.0	525.1	639.4	753.7
σ^2	3675.8	735.1	2508.8	3583.9	5371.2
Model $\mathcal{M}_2$ - $\log p(y\|x, \mathcal{M}_1) = -282.5398$					
Parameter	Mean	Standard deviation	Percentiles 2.5%	Median	97.5%
β_0	-113.0	43.9	-199.1	-113.4	-31.5
β_1	639.7	56.7	533.1	640.7	750.5
σ^2	3787.4	787.9	2573.0	3665.6	5562.4
Model $\mathcal{M}_3$ - $\log p(y\|x, \mathcal{M}_1) = 160.6112$					
Parameter	Mean	Standard deviation	Percentiles 2.5%	Median	97.5%
β_0	-78.1	37.2	-150.0	-77.7	-7.4
β_1	585.0	49.7	489.7	583.6	684.8
σ^2	2131.0	560.9	1250.9	2038.6	3457.8

Table 1: Posterior summaries for the three models. The Gibbs samplers for models $\mathcal{M}_2$ and $\mathcal{M}_3$ are run for a total of 11,000 draws, with the first 1,000 discarded (burn-in) and keeping every 10th after that (1000 draws from the posteriors). OLS estimates are used as initial values for the MCMC schemes. See Figures 2 and 3.

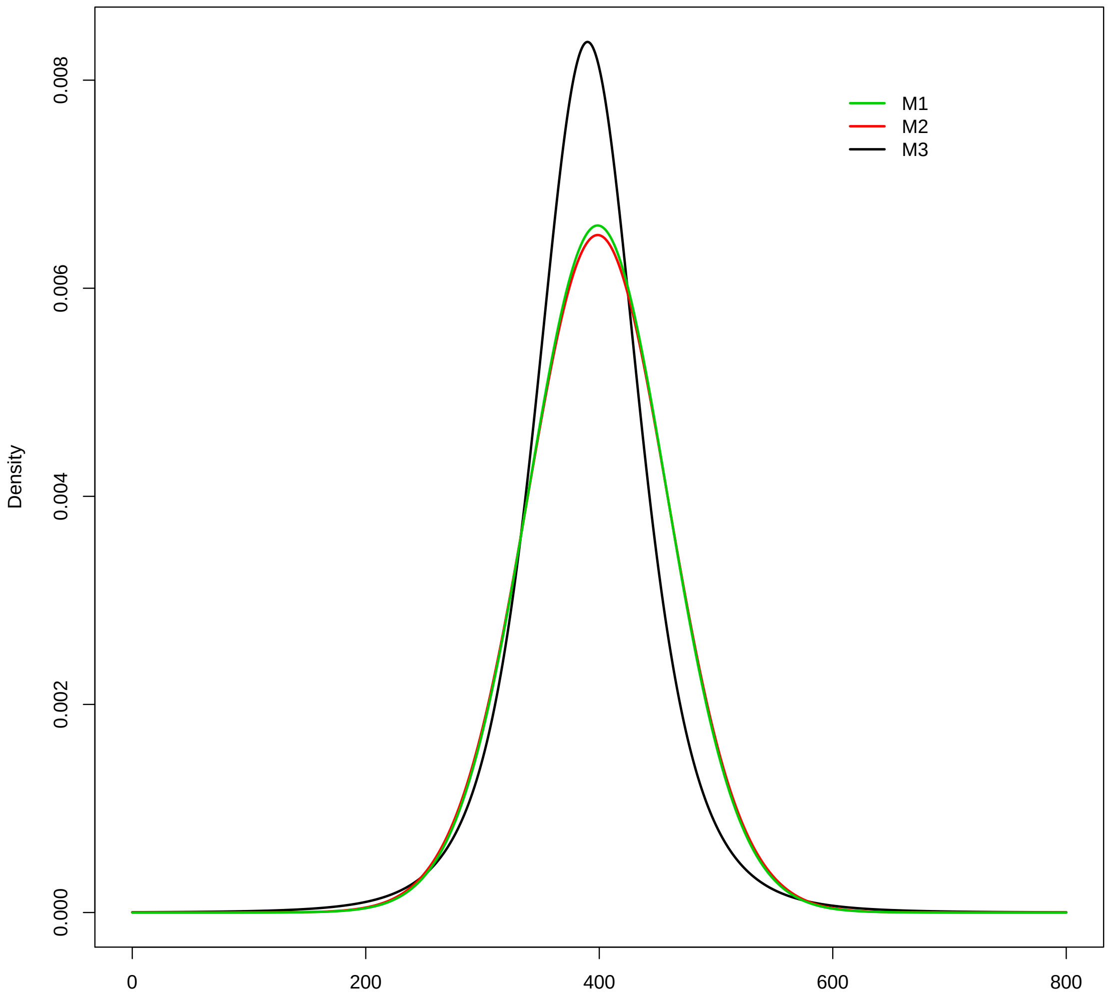

Figure 1: Posterior predictives $p(y_{new}|x_{new} = 8000, x, y, \mathcal{M}_i)$ for $i = 1, 2, 3$.

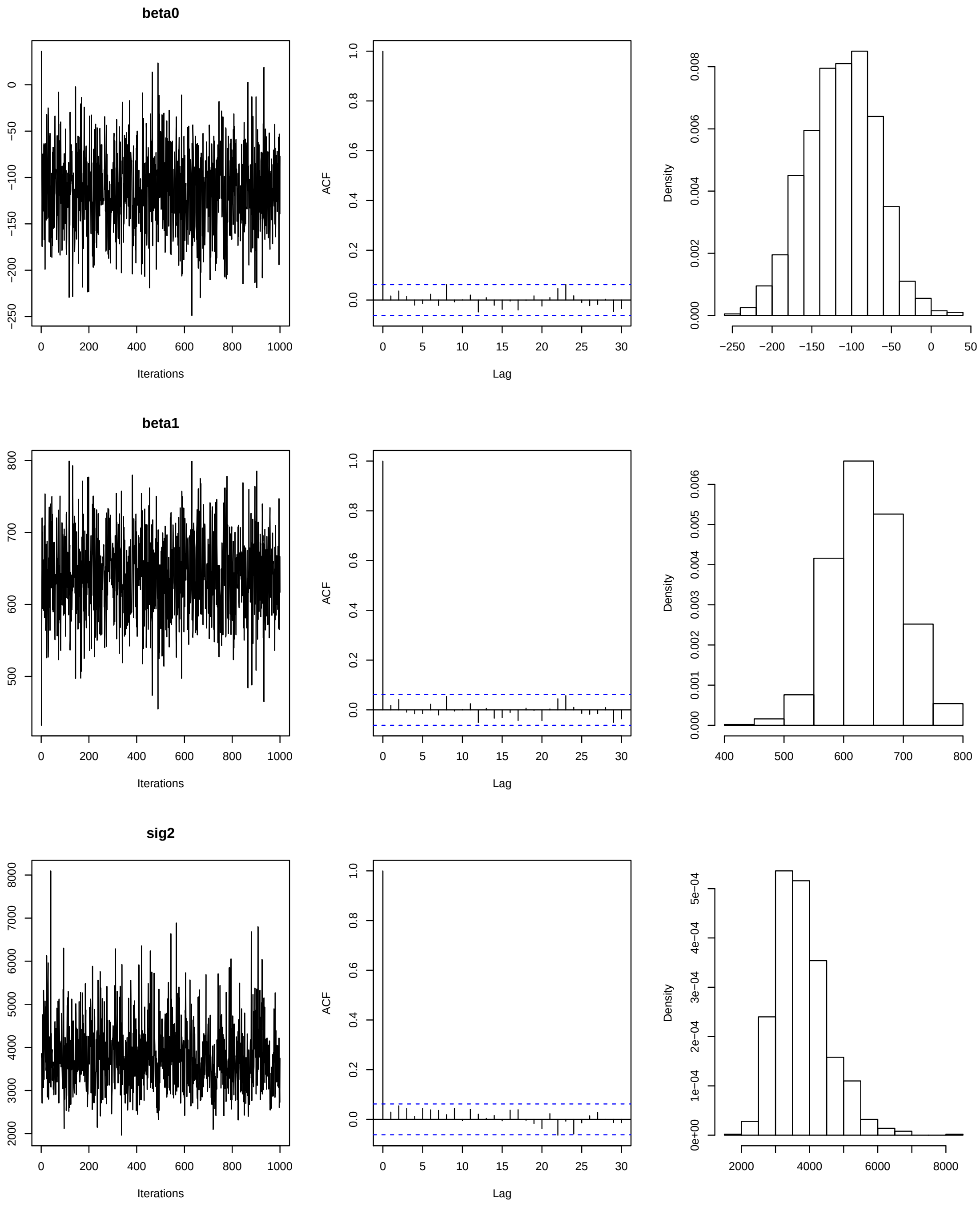

Figure 2: Model $\mathcal{M}_2$: Gibbs sampler outputs.

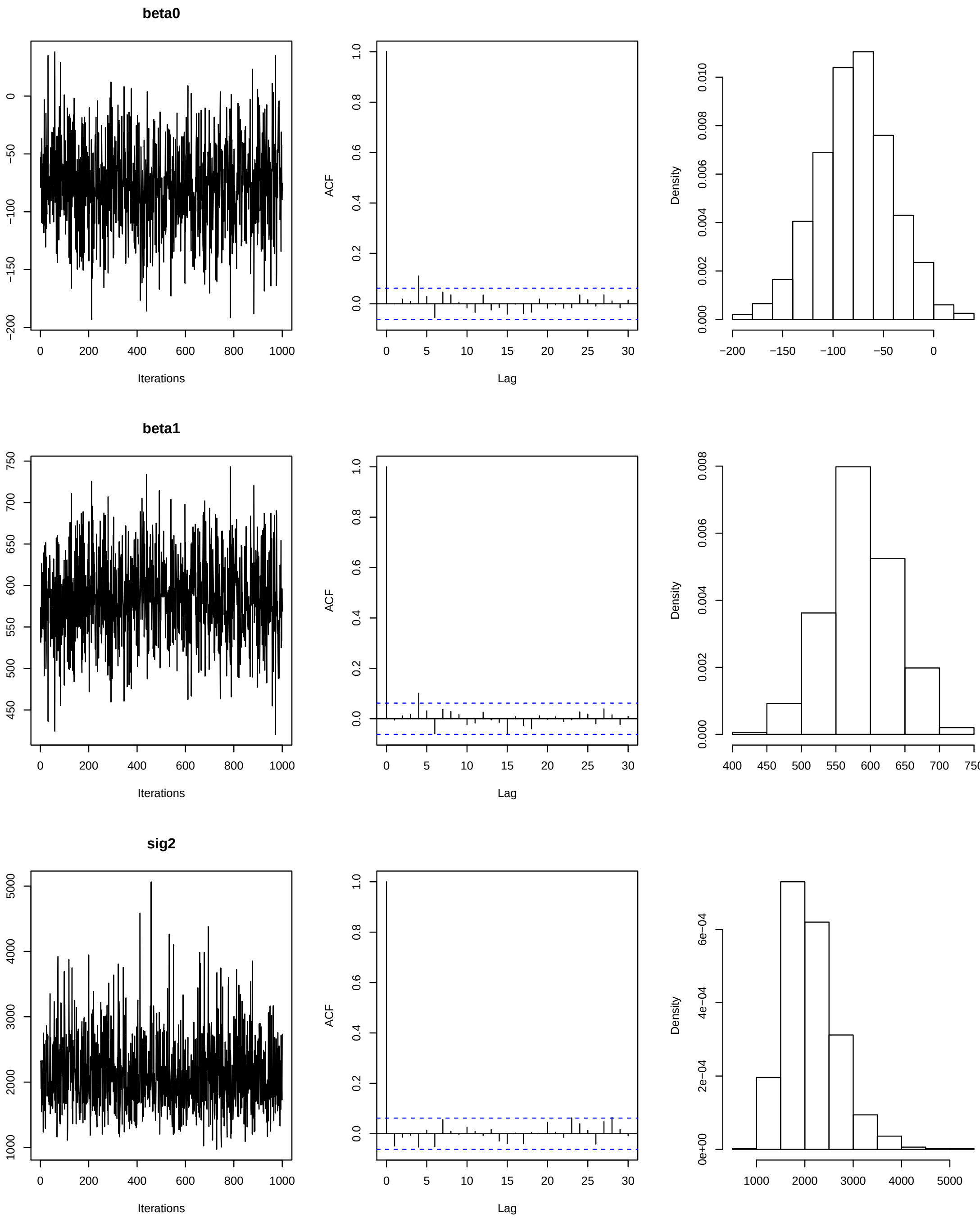

Figure 3: Model $\mathcal{M}_3$: Gibbs sampler outputs.