Coursework 6: STAT 570

Philip Pham

November 5, 2018

1. In this question, you will implement the algorithm you described in Question 4 of Exercises 5. The algorithm derived in 4(b) will now be implemented for the prostate cancer data. These data are available in the R package `lasso2` and are named `Prostate`. Take Y as log prostate specific antigen and x as log cancer volume. Implement the blocked Gibbs sampling algorithm using the prior given in the first equation of the aforementioned question, with $m_0 = m_1 = 0$, $v_{00} = v_{11} = 2$, $v_{01} = 0$, and $a = b = 0$. Run two chains, one with starting values corresponding to the unbiased estimates of the parameters and one starting from a point randomly generated from the prior $\pi(\beta_0, \beta_1)$. Report:

   (a) Histogram representations of the univariate marginal distributions $p(\beta_0 \mid y)$, $p(\beta_1 \mid y)$ and $p(\sigma \mid y)$, and scatterplots of the bivariate marginal distributions $p(\beta_0, \beta_1 \mid y)$, $p(\beta_0, \sigma \mid y)$, and $p(\beta_1, \sigma \mid y)$.

   Solution: The empirical univariate distributions from Gibbs sampling can be found in Figure 1. The joint distributions are found in Figures 2, 3, and 4.

   From the univariate distributions in Figure 1, we get samples close to the MLE estimates (see Table 5.3 of Wakefield's *Bayesian and Frequentist Regression Methods*). The dataset is rather large, so this is not surprising. The distributions for β_j are symmetrical. The distribution for σ seems to skew slightly to the right.

   From the bivariate distributions in Figures 3 and 4, we see that there is not much correlation between the β_j and σ, which is expected since σ is an ancillary statistic in the frequentist setting.

   From Figure 2, we see a negative correlation between β_0 and β_1, which is also expected since if we discount the effect of x on y, the estimate for β_0 must compensate.

   Code for the Gibbs sampler and plots can be found at `prostate.ipynb`.

   (b) The posterior means, standard deviations and 10%, 50%, 90% quantiles of β_0, β_1, and σ.

	Posterior mean	Standard deviation	10% quantile	50% quantile	90% quantile
σ	0.796154	0.056701	0.726527	0.793442	0.871269
β_0	1.495039	0.120843	1.341545	1.494805	1.652094
β_1	0.723290	0.068159	0.638009	0.724150	0.808873

Table 1: Summary statistics calculated from samples drawn with Gibbs sampling.

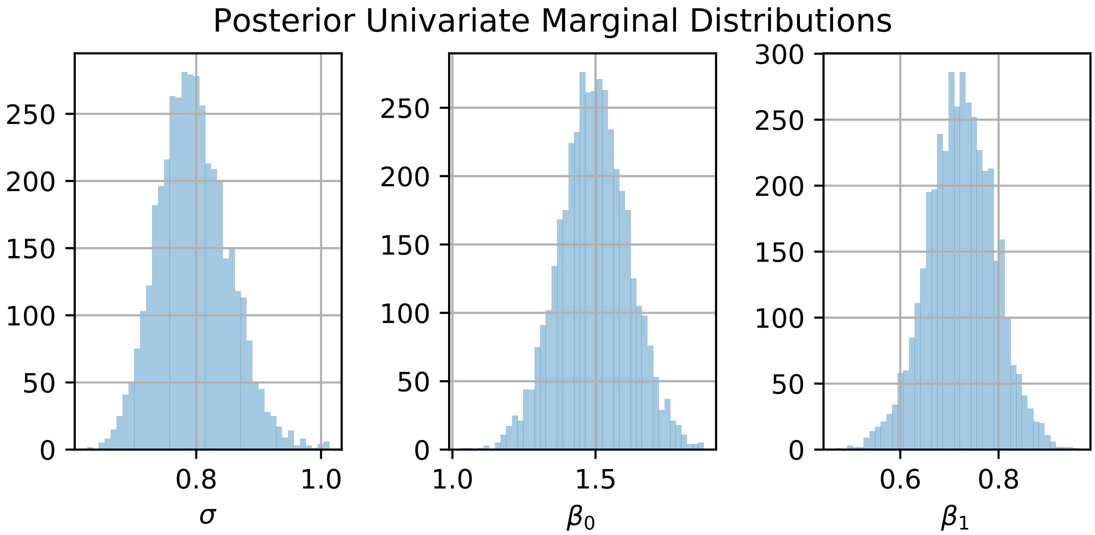


Figure 1: Emprical univariate distributions from Gibbs sampling.

Figure 2: Emprical joint distribution for $(\beta_0, \beta_1) \mid y$ from Gibbs sampling.

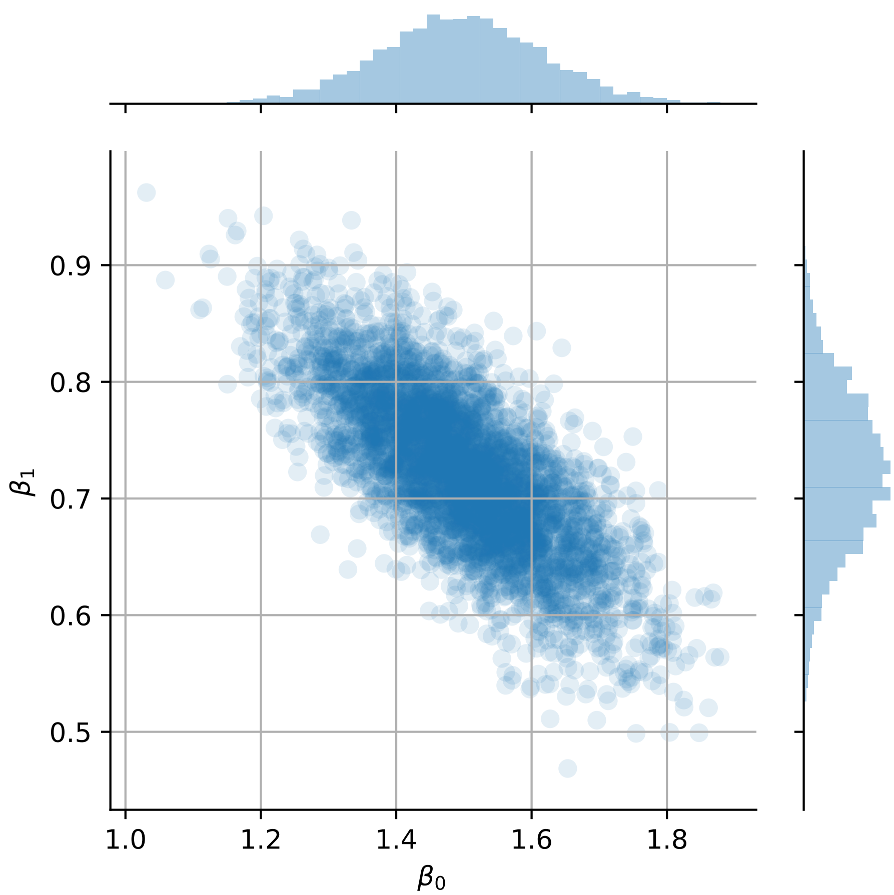


Figure 3: Emprical joint distribution for $(\beta_0, \sigma) \mid y$ from Gibbs sampling.

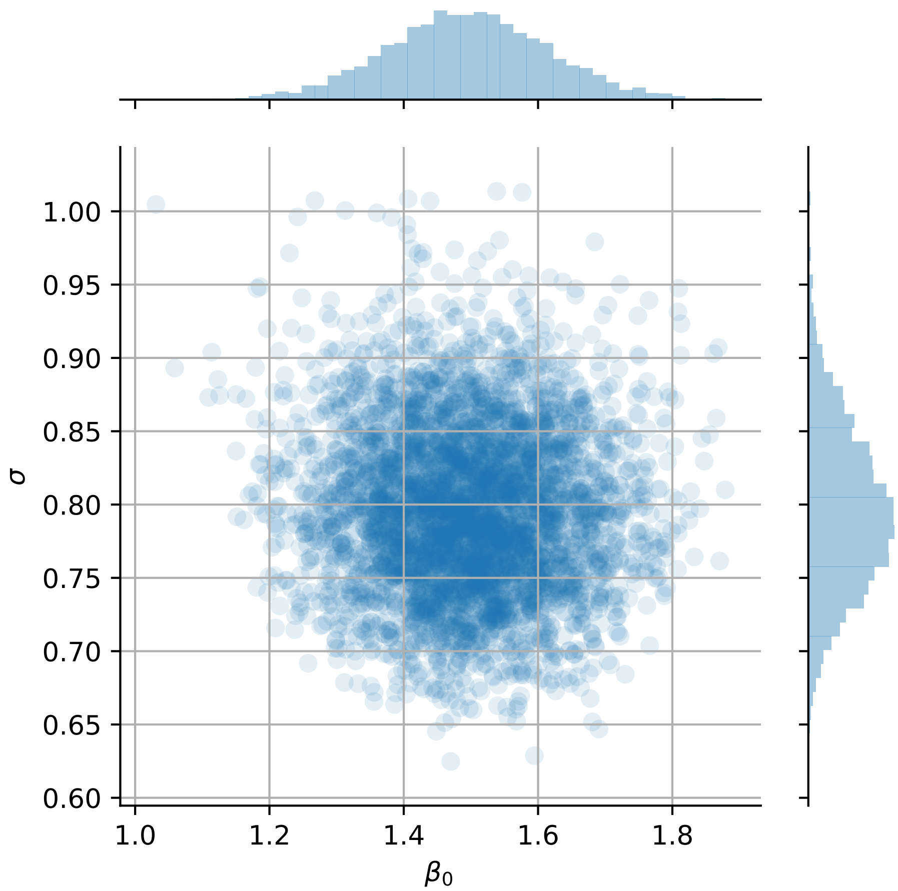
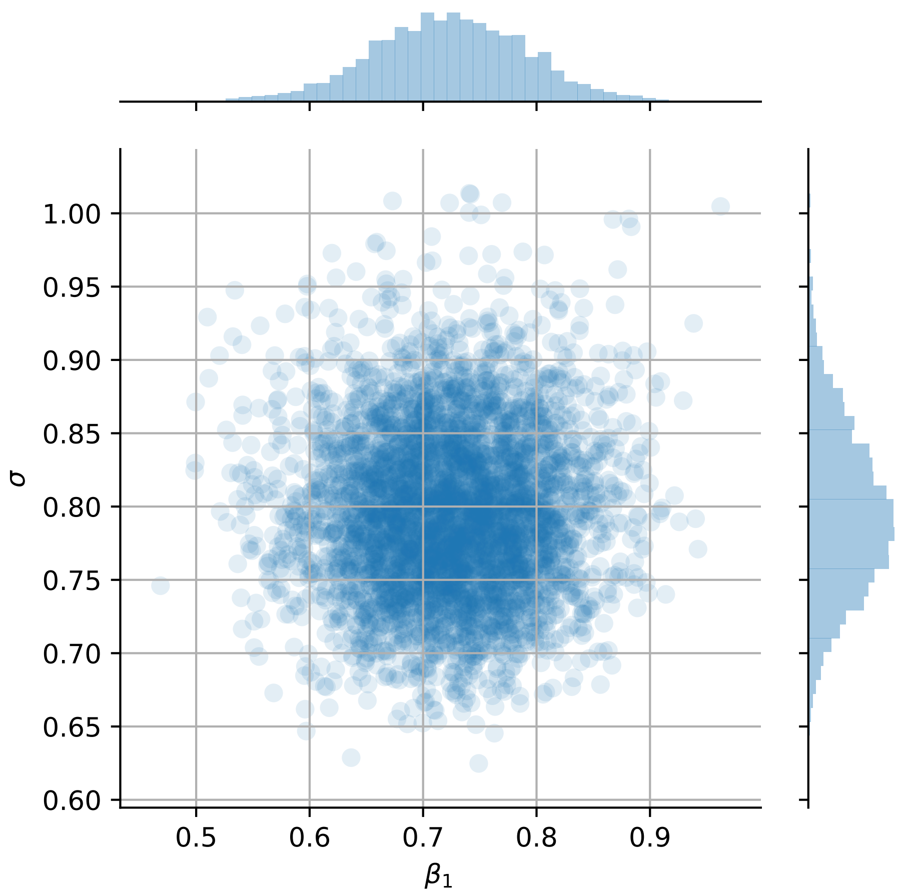


Figure 4: Emprical joint distribution for $(\beta_1, \sigma) \mid y$ from Gibbs sampling.



Length (mm)	0	1	2	3	4	5	6	7	8	9	10	11	12
1	2.247	2.640	2.842	2.908	3.099	3.126	3.245	3.328	3.355	3.383	3.572	3.581	3.681
10	1.901	2.132	2.203	2.228	2.257	2.350	2.361	2.396	2.397	2.445	2.454	2.454	2.474
20	1.312	1.314	1.479	1.552	1.700	1.803	1.861	1.865	1.944	1.958	1.966	1.997	2.006
50	1.339	1.434	1.549	1.574	1.589	1.613	1.746	1.753	1.764	1.807	1.812	1.840	1.852

Table 2: Failure stress data for four groups of fibers.

Solution: The summary statistics for the empirical posterior distributions can be found in Table 1. Samples from both chains were used for a total 4,096 samples.

Both the posterior mean and standard deviation agree closely with the MLE estimates in Table 5.3 of Wakefield's *Bayesian and Frequentist Regression Methods*. From the quantiles, we get an empirical estimate of the 80% credible interval, which appear to be symmetrical with respect to the median.

(c) $\mathbb{P}(\beta_1 > 0.5 \mid y)$

Solution: One way to interpret the significance of x's effect on y is to look at the distribution of β_1. The emprical estimate for $\mathbb{P}(\beta_1 > 0.5 \mid y)$ is $\boxed{0.9990}$, so the effect is likely significant, both statistically and in size.

This was calculated simply by taking the proportion of samples of β_1 that exceeded 0.5. See `prostate.ipynb` for the calculation.

(d) Justify your choice of *burn-in* period. For example, you may present the trace plots $\beta_0^{(t)}$, $\beta_1^{(t)}$, $(\log \sigma^2)^{(t)}$ versus t.

Solution: The trace plots for $\log \sigma^2$, β_0, and β_1 are shown in Figure 5. The first 128 results are plotted. The MLE chain immediately is stationary. The prior chain quickly becomes stationary in about 10 steps.

I specified a burn-in period of 128 steps for good measure. Then, I took 2,048 samples from each chain with no thinning.

2. Consider the data in Table 2 contain data on a typical reliability experiment and give the failure stresses (in GPa) of four samples of carbon fibers of lengths 1, 10, 20 and 50mm.

(a) Consider a Bayesian analysis with a Weibull likelihood and independent lognormal priors, $\eta \sim \text{LogNormal}(\mu_\eta, \sigma_\eta)$, $\alpha \sim \text{LogNormal}(\mu_\alpha, \sigma_\alpha)$. Choose μ_η, σ_η so that the prior probability that η lies between 0.5 and 30 is 0.9, and μ_α, σ_α so that the prior probability that Îś lies between 1 and 4 is 0.9.

Solution: $\log \eta \sim \mathcal{N}(\mu_\eta, \sigma_\eta)$ by definition of the lognormal distribution. Since log is a monotonic transformation,

$$\begin{aligned} 0.9 = \mathbb{P}(1/2 \leq \eta \leq 30) &= \mathbb{P}\left(\log \frac{1}{2} \leq \log \eta \leq \log 30\right) \\ &= \mathbb{P}\left(\Phi^{-1}(0.05) \leq \frac{\log \eta - \mu_\eta}{\sigma_\eta} \leq \Phi^{-1}(0.95)\right), \quad (1) \end{aligned}$$

where Φ is the cumulative distribution function of a standard normal.

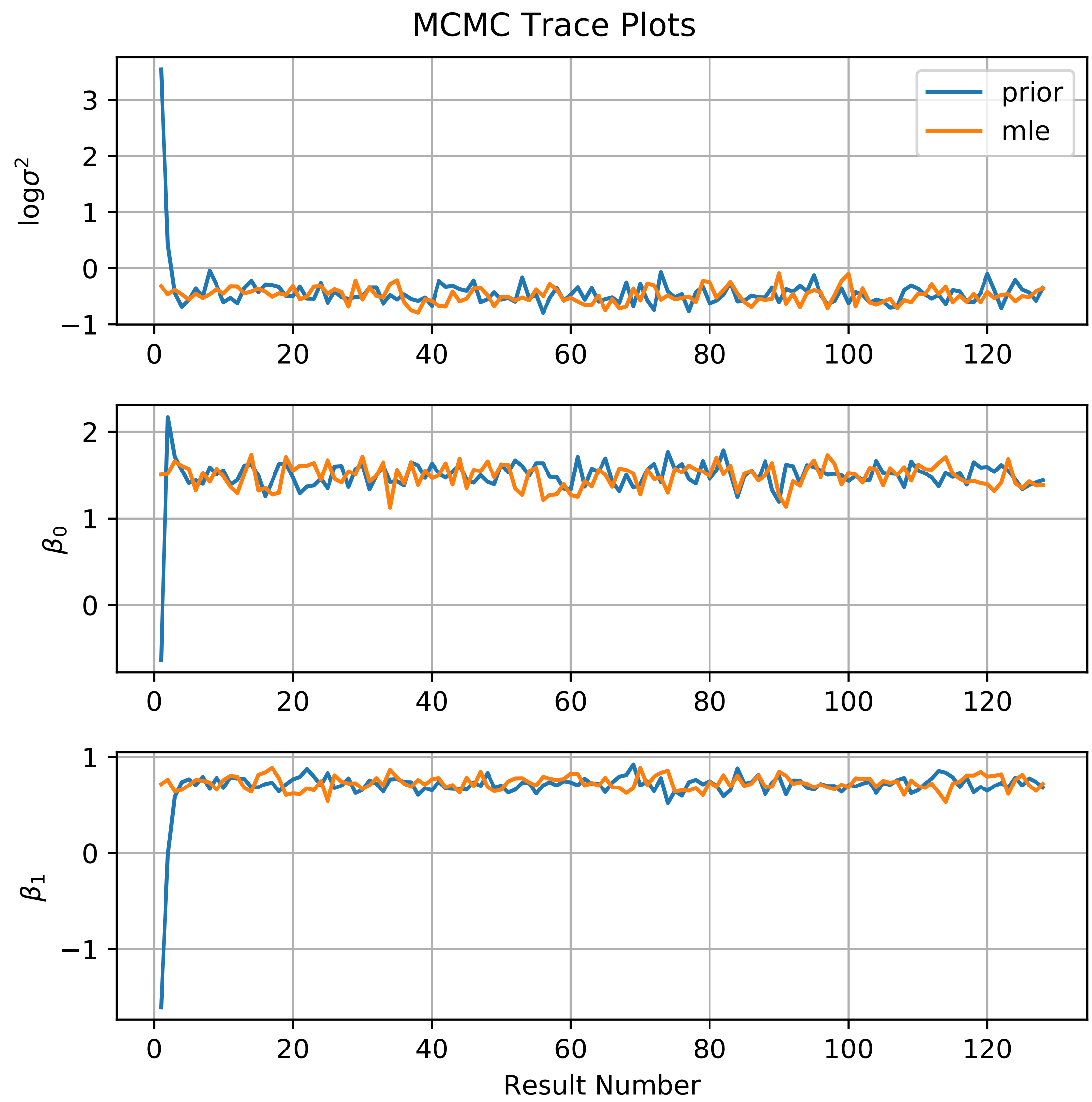


Figure 5: The trace plots show the sampled posterior parameters at each step in the MCMC chain.

i	time	drug concentration
1	2	1.63
2	4	1.01
3	6	0.73
4	8	0.55
5	10	0.41
6	24	0.01
7	28	0.06
8	32	0.02

Table 3: Concentrations of the drug Cadralazine (in mg/liter, y_i) as a function of time (in hours, x_i), for $i = 1, \ldots, 8$.

Equation 1 implies that

$$\begin{aligned} \frac{\log(1/2) - \mu_\eta}{\sigma_\eta} &= \Phi^{-1}(0.05) \\ \frac{\log 30 - \mu_\eta}{\sigma_\eta} &= \Phi^{-1}(0.95). \end{aligned}$$

Solving, we have that

$$\begin{aligned} \sigma_\eta &= \frac{\log 30 - \log \frac{1}{2}}{\Phi^{-1}(0.95) - \Phi^{-1}(0.05)} \approx 1.2446 \\ \mu_\eta &= \log 30 - \sigma_\eta \Phi^{-1}(0.95) = \log \frac{1}{2} - \sigma_\eta \Phi^{-1}(0.05) \approx 1.3540 \end{aligned}$$

Repeating the calculating for α, we have $\mu_\alpha \approx 0.6931$ and $\sigma_\alpha \approx 0.4214$. Calculations can be found in `failure_stresses.ipynb`.

(b) Run MCMC for summarizing the posterior $p(\eta, \alpha \mid y)$, and implement this algorithm for each of the groups in Table 2. Report the posterior medians and 90% credible intervals for η and α and give histograms representations of the posterior margins for η and α, and a scatterplot representation of $p(\eta, \alpha \mid y)$.

Solution:

3. The data in Table 3, taken from Wakefield et al. (1994), were collected following the administration of a single 30mg dose of the drug Cadralazine to a cardiac failure patient. The response y_i represents the drug concentration at time x_i, $i = 1, \ldots, 8$. The most straightforward model for these data is to assume

$$\log y_i = \mu(\beta) + \epsilon_i = \log\left[\frac{D}{V}\exp(-k_e x_i)\right] + \epsilon_i \tag{2}$$

where $\epsilon_i \mid \sigma^2 \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$, $\beta = [V, k_e]$ and the dose is $D = 30$. The parameters are the volume of distribution $V > 0$ and the elimination rate k_e.

(a) For this model obtain expressions for:

i. The log-likelihood function $L(\beta, \sigma^2)$.

ii. The score function $S\left(\beta, \sigma^2\right)$.
iii. The expected information matrix $I\left(\beta, \sigma^2\right)$.

(b) Obtain the MLE, and give an asymptotic 95% confidence interval for each element of β.

(c) Plot the data, along with the fitted curve.

(d) Using residuals, examine the appropriateness of the assumptions of the above model. Does the model seem reasonable for these data?

(e) The clearance $Cl = V \times k_e$ and elimination half-life $x_{1/2} = \log 2/k_e$ are parameters of interest in this experiment. Find the MLEs of these parameters along with asymptotic 95% confidence intervals.