Stat515, Spring 2015: Takehome final
Due Wednesday, April 29, 2015 at midnight

1. Problem 1

The Metropolis-Hastings Algorithm uses and All-At-Once update based on calculating the probability of acceptance. We seek to maximize the likelihood of the errors coming from the EMG distribution. The distribution of the errors is given by

$$\varepsilon_i \sim EMG\left(Y_i - \beta_0 - \beta_1 X_i, \sigma_i, \lambda\right)$$

The posterior probability of β_1 is given by

$$P\left(\beta_1 \mid X, Y\right) \propto \prod_{i=1}^{i=n} EMG\left(Y_i - \beta_0 - \beta_1 X_i, \sigma_i, \lambda\right) \text{Prior}\left(\beta_1\right) \text{Prop}\left(\beta_1\right)$$

where $\text{Prior}\left(\beta_1\right)$ and $\text{Prop}\left(\beta_1\right)$ are the prior and proposal distributions of β_1 respectively. The log of the probability of acceptance is given by

$$\log\left(\text{Probability of Acceptance}\right) = \log\left(P\left(\beta_1 \mid X, Y\right)\right) - \log\left(P\left(\beta_1^* \mid X, Y\right)\right)$$

where β_1^* are the latest accepted values and β_1 are the proposed values.

(a) Procedure for the M-H simulation for Problem 1 is
  a. Declare eth maximum number of runs of the simulation
  b. Initialize the data storage matrix for the proposal.
  c. Initialize the proposal distribution for B1
  d. B1 is drawn from a t-distribution with 3 df.
  e. For the final run the initial mean = 3.3 and Var(B1)=1*(df)/(df-2)
  f. Declare the mean and variances for the prior distribution for B1~N(0, σ^2=100)
  g. Calculate the log likelihood of the proposal and posterior distributions*
  h. Randomly draw a value for B1 from the proposal distribution
  i. Calculate the log likelihood of the posterior distribution
  j. Calculate the probability of acceptance of the proposal.
  k. Update the mean values of the proposal distribution with the proposed value if accepted or the last accepted value.
  l. Store the latest value in the storage matrix.
  m. Return to step f and repeat process until the specified number of runs has been completed.

(b) Posterior mean table
(c) Posterior credible Intervals table

Table 1: The Posterior Mean, MCMCse, and Credible Interval

	Estimate (MCMCse)		Credible Interval	
			2.50%	97.50%
B1	3.2662	(0.00349)	2.594	3.897

(a) Approximate density plot

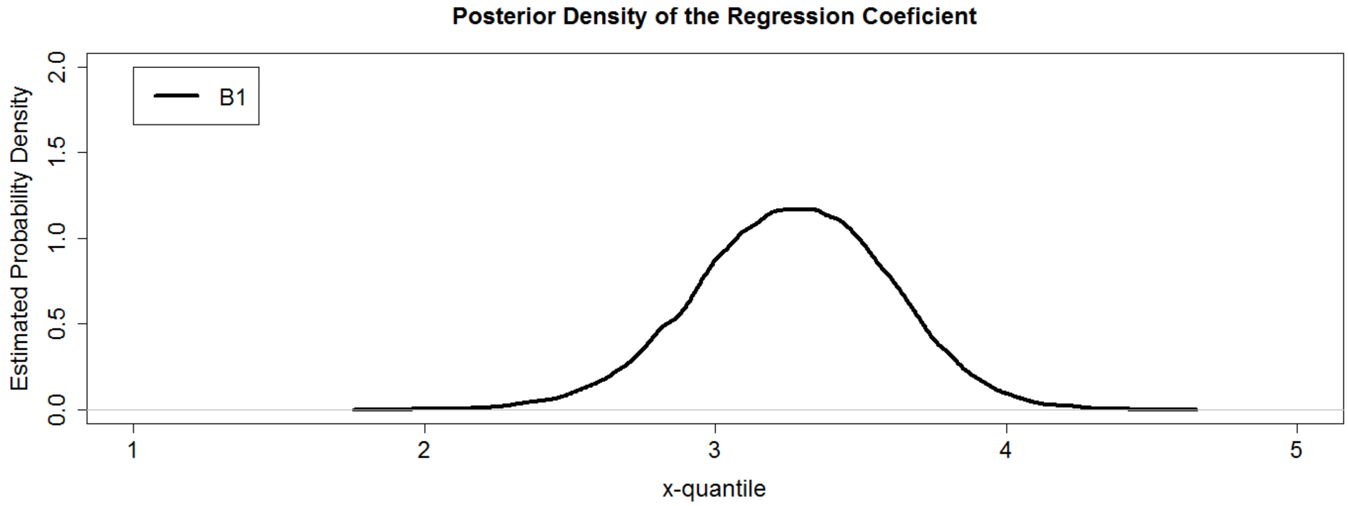


Figure 1: Posterior Density of the regression coefficients B1

(b) Process to determine reliability of algorithm
I ran several preliminary runs with 10^4 simulations and tracked the median of the estimated and the final ending points. All of the preliminary runs reached the same stationary distribution. I saved three of the preliminary runs for documentation. I then ran a final run with 10^5 simulations. The results of the preliminary and final simulations are shown in Table 1. Additional efforts to ensure reliability of the model included: making several plots of the autocorrelation, estimated batch means and MCMCse as the number of samples increased, and calculating the effective number of samples. These efforts ensured that the mean of B1 was properly converging and the MCMCse's were decreasing appropriately as sample size increased.

Table 2: Selected Results from Simulations.

Run Number	Type of Simulation	Starting Point B1	Var(B1)	Median B1	Ending points B1
10	Preliminary	10	1	3.273	3.763
11	Preliminary	0	1	3.281	3.788
12	Preliminary	-10	1	3.279	3.788
13	Final	3.3	1	3.273	3.443
Mean				3.276	3.696

2. Problem 2

The Metropolis-Hastings Algorithm uses and All-At-Once update based on calculating the probability of acceptance. We seek to maximize the likelihood of the errors coming from the EMG distribution. The distribution of the errors is given by

$$\varepsilon_i \sim EMG(Y_i - \beta_0 - \beta_1 X_i, \sigma_i, \lambda)$$

The joint posterior probability of $\beta_0, \beta_1, \lambda$ is given by

$$P(\beta_0, \beta_1, \lambda \mid X, Y) \propto \prod_{i=1}^{i=n} EMG(Y_i - \beta_0 - \beta_1 X_i, \sigma_i, \lambda)\text{Prior}(\beta_0)\text{Prior}(\beta_1)\text{Prior}(\lambda)\text{Prop}(\beta_0)\text{Prop}(\beta_1)\text{Prop}(\lambda)$$

where $\text{Prior}(\beta_0)$ and $\text{Prop}(\beta_0)$ are the prior and proposal distributions of β_0 respectively and likewise for β_1 and λ. The log of the probability of acceptance is given by

$$\log(\text{Probability of Acceptance}) = \log(P(\beta_0, \beta_1, \lambda \mid X, Y)) - \log(P(\beta_0^*, \beta_1^*, \lambda^* \mid X, Y))$$

where $\beta_0^*, \beta_1^*, \lambda^*$ are the latest accepted values and $\beta_0, \beta_1, \lambda$ are the proposed values.

(c) Procedure for the M-H simulation for Problem 1 is
   n. Declare the variances for the proposal and prior distribution for B0, B1 and lambda
   o. Initialize B0, B1 and lambda and calculate the log likelihood of the posterior distribution*
   p. Randomly draw a value for each of B0, B1 and lambda
   q. Calculate the log likelihood of the posterior distribution
   r. Calculate the probability acceptance of the proposed triple.
   s. Update the proposal distributions.
   t. Store the latest values if the thinning count == 0
   u. Return to step c and repeat process until the specified number of runs has been completed.

(d) Posterior mean table

Table of the Posterior Means, MCMCse's, and Credible Intervals

	Estimate (MCMCse)		Credible Interval	
			2.50%	97.50%
B0	1.095	(0.00258)	0.736	1.448
B1	3.460	(0.00239)	3.069	3.849
lambda	0.8029	(0.00073)	0.704	0.914

(e) Correlation of B0 and B1

Table of the Correlation Estimates for B0 and B1

		95% Confidence Interval	
	Estimate	Lower	Upper
Corr(B0, B1)	-0.5636	-0.5678	-0.5594

(f) Approximate density plot

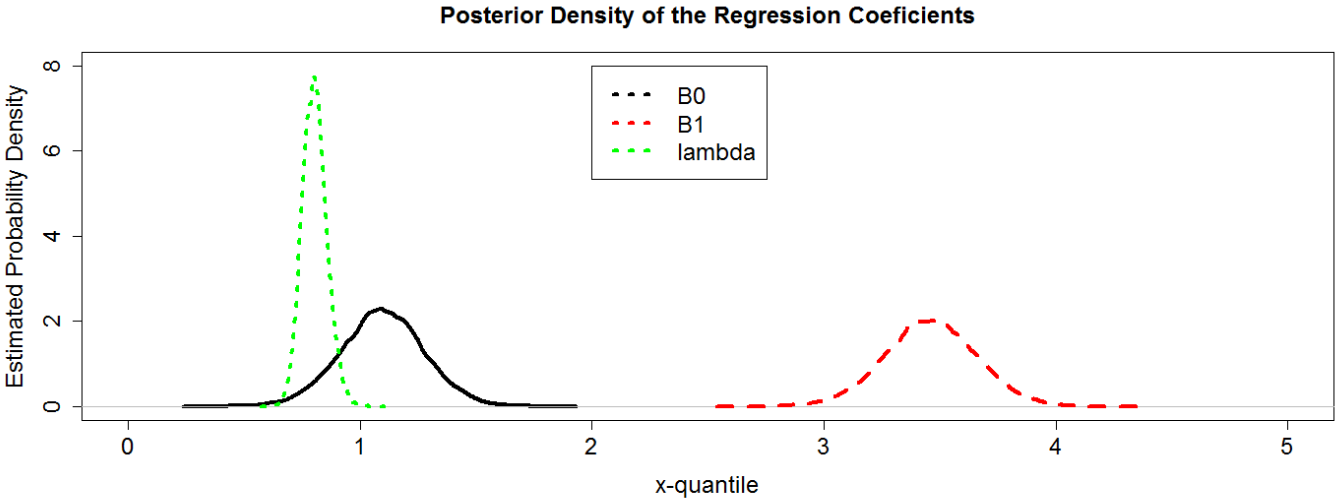


Figure 1: Posterior Density of the regression coefficients B0, B1, and lambda

(g) Process to determine reliability of algorithm

I ran several preliminary runs with 10^4 simulations and tracked the median of the estimated and the final ending points. I then ran some secondary longer runs with 10^5 simulations from three staring points with tuned variance parameters to test whether they converged to the same stationary distribution. They did in fact reach the same stationary distribution. So then I averages of all medians of the longer runs and used them as a starting point for the final run with 10^6 simulations. The results of the preliminary and secondary simulations are shown in Table 1. Additional efforts to ensure reliability included making plots of the autocorrelation, estimated batch means, and MCMCse as the number of samples increased, and calculating the effective number of samples. This ensured that the three variables were mixing properly, and that the means converging and the MCMCse's were decreasing appropriately as sample size increased.

Table 1: Results of the Preliminary and Secondary Runs of the regression algorithm

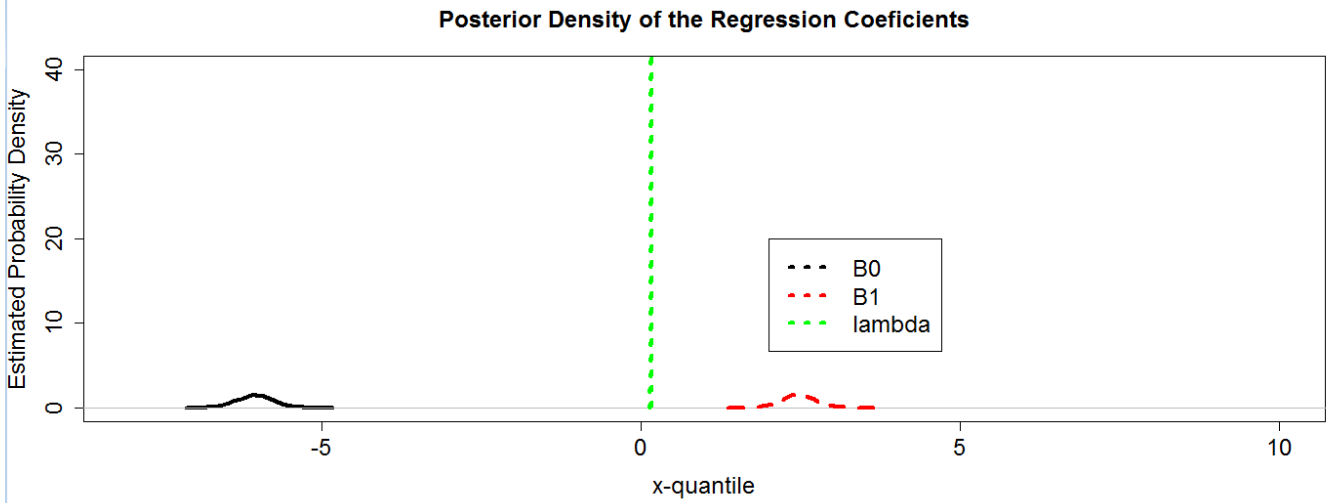
	Type of	Starting Point			Variance Parameters			Median			Ending points		
Run Number	Simulation	B0	B1	lambda	Var(B0)	Var(B1)	Var(lambda	B0	B1	lambda	B0	B1	lambda
10	Secondary	1	3.5	0.8	0.1	0.1	0.1	1.12	3.45	0.81	1.20	3.48	0.81
11	Secondary	1	3.5	0.8	0.1	0.1	0.01	1.10	3.46	0.80	1.32	3.25	0.85
12	Final	1	3.5	0.8	0.1	0.1	0.01	1.10	3.46	0.80	1.06	3.47	0.78
20	Preliminary	10	3.5	10	2	2	2	1.14	3.34	0.76	0.98	3.34	0.72
30	Preliminary	0.1	3.5	0.1	2	2	2	1.13	3.56	0.80	0.98	3.53	0.76
40	Preliminary	10	0	10	2	2	2	1.24	3.43	0.83	1.09	3.43	0.80
50	Preliminary	0.5	2	0.3	2	2	2	1.15	3.48	0.81	1.01	3.48	0.77
60	Secondary	0	0	0.001	2	2	2	1.17	3.50	0.85	0.94	3.80	0.83
61	Secondary	0	0	0.001	0.1	0.1	0.01	1.09	3.46	0.80	1.42	3.08	0.88
70	Preliminary	0	-10	0.001	2	2	2	1.11	3.56	0.79	1.11	3.44	0.78
80	Preliminary	0	-10	10	2	2	2	0.99	3.58	0.77	0.78	3.63	0.74
81	Secondary	0	-10	10	0.1	0.1	0.01	1.17	3.41	0.82	1.39	3.13	0.87
Mean								1.13	3.47	0.80	1.11	3.42	0.80

3. Problem 3

(a) Posterior mean table

	Estimate (MCMCse)		Credible Interval	
			2.50%	97.50%
B0	-6.049	(0.01038)	-6.580	-5.523
B1	2.466	(0.00916)	1.934	2.984
lambda	0.1614	(0.000158)	0.152	0.172

(b) Approximate density plots



(c) How algorithm was modified for Problem 3

The amount of thinning was increased to 50 for the final run to reduce the autocorrelation.