SYST 664 / CSI 674: Supplemental Exercises

These are exercises on material covered in the class notes but not covered in either homework or the midterm exam. Solutions will be provided.

1. A compound called estriol was measured over a 24 hour period in the blood of pregnant women. The babies' weights were then recorded at birth. The purpose of the study was to determine whether there was a relationship between estriol and birth weight.[1] The data from this study can be found at http://www.biostat.umn.edu/~lynn/iid/estriol.dat. Estriol was measured in milligrams per 24 hours and birthweight was measured in 100 gram units.
   a. Using the non-informative prior distribution $g(\eta, \beta, \rho) \propto \rho^{-1}$, find the joint posterior distribution for the transformed intercept η, the slope β, and the precision ρ. Find the joint posterior distribution for the intercept α, the slope β, and the precision ρ.
   b. Comment on your results, including whether the assumptions for normal linear regression are met.
   c. What is the predictive distribution for birthweight given milligrams of estriol?
   d. Use 1000 Monte Carlo samples to find a 90% predictive interval for the birthweight of a baby given that the 19 mg of estriol were measured in the mother's urine over 24 hours.

2. The following table, taken from Hand, et al. (1994), shows weight gains in grams for rats fed four different diets. The researchers who collected the data wanted to evaluate whether there are systematic differences in weight gain for the four diets. Assume the weight gains x_{gi} for the i^{th} rat in diet group g are independent normal random variables with an unknown diet-specific mean and a common standard deviation σ. Assume the unknown means θ_g, $g = 1,\ldots 4$ are iid normal random variables with mean μ and standard deviation τ.

Diet 1	Diet 2	Diet 3	Diet 4
90	73	107	98
76	102	95	74
90	118	97	56
64	104	80	111
86	81	98	95
51	107	74	88
72	100	74	82
90	87	67	77
95	117	89	86
78	111	58	92

   Assume that the weight gain for rats fed diet $j = 1, 2, 3, 4$ are independent normal random variables with unknown diet-specific mean M_j and common known precision ρ. Assume the means M_j are independent and identically distributed normal random variables with mean μ and standard deviation τ. We will use the following empirical Bayesian estimates for the hyperparameters:
   - The observation precision is estimated by $\rho = 0.00447$, the inverse of the average sample variance. Therefore, the standard deviation of the observations is $s = 1/\rho^{1/2} = 14.96$.
   - The prior mean is estimated by $\mu = 87.25$, the average of the sample means.

[1] Original source: Greene and Touchstone (1963). 'Urinary tract estriol: an index of placental function,' American Journal of Obstetrics and Gynecology, 85:1-9. Reprinted in Rosner (1982). Fundamentals of biostatistics, Duxbury Press.

- The precision of the prior means M_j is estimated by $k\rho$, where $k = 2.79$ is sample variance of the observations divided by the variance of the diet-specific sample means about the grand mean. Therefore, the standard deviation of M_j is $\tau = (k\rho)^{-1/2} = (2.79 \times 0.00447)^{-1/2} = 8.95$.

a. Find the marginal likelihood of the four sample means $\bar{y}_j$, $j = 1, \ldots, 4$, given ρ, μ, and τ.

b. Repeat part *a*, but assume that all four means are equal. That is, that the weight gain for rats fed diet $j = 1, 2, 3, 4$ are independent normal random variables with unknown common mean M and common known precision ρ. Assume the mean M is a normal random variables with mean μ and standard deviation τ. Assume that ρ, μ and τ are as given in Part *a*.

c. The rats fed Diet 2 seemed to gain more weight than rats fed the other diets. It seems reasonable to ask whether the first three diets produce the same weight gain on average, while weight gains for Diet 2 are systematically larger. To explore this question, repeat Part *a*, but assume that $M_1 = M_3 = M_4 \neq M_2$. That is, the weight gain for rats fed diet $j = 1, 3, 4$ are independent normal random variables with unknown common mean M and common known precision ρ. The mean M is a normal random variables with mean μ and standard deviation τ. The weight gains for rats fed diet 2 are independent and normal with mean M_2 and precision ρ, where M_2 is independent of M and normal with mean μ and standard deviation τ. Assume that ρ, μ and τ are as given in Part *a*.

d. Consider three hypotheses:

- H1: All M_j are different
- H2: All M_j are the same
- H3: $M_1 = M_3 = M_4 \neq M_2$

Assume these hypotheses are equally likely *a priori*. What is the posterior probability of the three hypotheses?

3. Respondents in a health and lifestyle survey gave subjective assessments of their personal health.[2] 954 respondents in one of the surveyed regions rated their health as *Good*; 444 rated their health as *Fairly Good*; and 78 rated their health as *Not Good*.

   a. Assuming a uniform prior distribution on the probabilities of the three categories, find the posterior distribution on the probabilities of the three categories. Find 95% credible intervals for the probabilities of each of the three categories. Make a triplot of the probability that a respondent assesses him or herself to be in good health.

   b. In a different region, 459 respondents in one of the surveyed regions rated their health as *Good*; 175 rated their health as *Fairly Good*; and 43 rated their health as *Not Good*. Repeat Problems 1 and 2 for the second region. Discuss your results with regard to whether the two regions differ in subjectively assessed ratings of personal health.

   c. Consider two hypotheses:

   - The probabilities of the three rating categories are the same in both regions, with a uniform distribution *a priori*;
   - The probabilities of the three rating categories are different in the two regions, with a uniform distribution *a priori* for each region. Assume these hypotheses are equally likely *a priori*.

   Find the posterior probabilities of the two hypotheses.

4. For the rat tumor problem of Unit 7, implement a Metropolis-Hastings sampler to estimate the joint posterior distribution for U, V and $\Theta_{1:71}$.

   a. To sample $\Theta_{1:71}$, use the Gibbs sampler, which, as described in Unit 9, is a special case of the Metropolis-Hastings sampler.

---

[2] Data set 459 from Hand et al. (1994). Original source: Turrall, K. (1992) *An Analysis of 5 Health and Lifestyle Surveys*. MSc dissertation, Southampton University, Faculty of Mathematics, Table 3.52, 43.

b. Sample $u^{(k)}$ from a uniform distribution on the interval $[u^{(k-1)} - 0.03, u^{(k-1)} + 0.03]$ centered at the previous estimate and having width 0.06.

c. Sample $v^{(k)}$ from a uniform distribution on the interval $[v^{(k-1)} - 7, v^{(k-1)} + 7]$ centered at the previous estimate and having width 14.

Note that by using a uniform distribution, the Hastings correction factor (ratio of proposal probabilities) cancels out and we can calculate the acceptance probability just from the ratio of proposed and old state likelihoods. Also note that the proposal distribution can generate values outside the legal range of $0 < U < 1$ and $V > 0$. These have likelihood zero and should be rejected. The widths of the proposal distributions were chosen by trial and error: if the width is too wide, then acceptance will be rare; if the width is too narrow, then the Markov chain will move very slowly through the space. Both of these cause high autocorrelations.

a. Generate 10000 samples with your Metropolis sample. Do a traceplot and plot the autocorrelation function for U and V. Calculate the expected sample size for U, V and $\Theta_{1:71}$. Discuss.

b. Find posterior credible intervals for $\Theta_{1:71}$. Produce a shrinkage diagram like the one on page 13 of the Unit 7 notes.

c. Compare your results with the Gibbs sampler and empirical Bayes model from Unit 7. Discuss.