Statistics 135 – Assignment 3
Due: Wednesday, November 9, 2005

For this assignment, please submit your answers typeset in some package, preferably in LaTeX.

1. This question will investigate two possible ways of simulating random vectors from a multivariate normal distribution. The multivariate normal distribution is defined by two parameters, μ, a vector of length p, and Σ, the $p \times p$, variance-covariance matrix. The density of this distribution is

$$f(\mathbf{x}|\mu, \sigma) = (2\pi)^{-p/2}|\Sigma|^{-p/2} \exp(-0.5(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu))$$

While one of the most important distributions in statistics, there is no random number generator built into the base of **S-Plus** or **R** for this distribution (there is one if you attach the package `MASS`). What is desired is a function that will generate realizations from this distribution where the call to the function `mrnorm(n, mu, sigma)` will return a $n \times p$ matrix, where each row is a realization from the p dimensional multivariate normal.

(a) Matrix approach

Suppose that $\mathbf{z} = [z_1 z_2 \dots z_p]^T$ is a vector of p independent standard normal random variables. Then $\mu + R^T\mathbf{z}$ is a realization from $N(\mu, \Sigma)$ where R is a matrix satisfying $R^T R = \Sigma$. The are number of ways of producing R with the most common based on the Choleski decomposition or the eigenvalue/eigenvector decomposition. R for the Choleski decomposition can gotten with `R = chol(Sigma)`. Note that the form of the matrix R doesn't matter for the distributional result to hold.

Write the function for generating from the multivariate normal based on this Choleski decomposition idea.

(b) Gibbs sampler

The Gibbs sampler is an approach that allows one to sample from complicated distributions. The Gibbs sampler, and Markov Chain Monte Carlo (MCMC) methods in general have opened up many areas of statistics, for example Bayesian statistics, and have made them tractible.

Suppose you wanted to generate samples from the joint density $f(x, y, z)$, but f is complicated. A scheme that will generate (dependent) samples (asymptotically) is

```
initialize x, y, and z as x(0), y(0), and z(0)
for i = 1 to n {
  draw x(i) from f(x|y(i-1), z(i-1))
  draw y(i) from f(y|x(i), z(i-1))
  draw z(i) from f(z|x(i), y(i))
}
```

The realizations `(x(i), y(i), z(i))` form a Markov Chain with a stationary distribution having density $f(x, y, z)$. Note that the realizations from the chain are not initially

from the desired distribution, as the result is asymptotic. Due to this fact, the initial part of the chain is usually thrown away (known as burn-in). Also the realizations are dependent, as the the realization at step i depends on the previous realizations. The Matrix approach mentioned above does not have this problem.

Write a function for implementing the Gibbs sampler to draw from a bivariate normal (assume p is 2). As part of the function, allow for a burn-in figure to be given, and for the user to be able to specify the starting state of the chain. However for the starting state, have the mean of the distribution be used as the default. The conditional distributions you need to implement this sampler are

$$X|Y = y \sim N\left(\mu_x + \frac{\sigma_{xy}}{\sigma_y^2}(y - \mu_y), \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}\right)$$

$$Y|X = x \sim N\left(\mu_y + \frac{\sigma_{xy}}{\sigma_x^2}(x - \mu_x), \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2}\right)$$

where σ_x^2 and σ_y^2 are the variances of the two components and σ_{xy} is the covariance.

(c) Generate 1000 realizations with both functions, with `mu = c(10,5)` and `Sigma = matrix(c(10,5,5,5),ncol=2)` (this corresponds to the correlation ρ being 0.707. For the Gibbs sampler sample, use a burnin of 100 iterations.

(d) Calculate the sample mean vector and sample variance-covariance matrix for both samples to see if they are close to the desired values. When calculating the mean vectors, do it in a single operation without looping.

(e) Calulate the lag one autocorrations for both variables for both samples. You can do this along the lines of `cor(mvn[-1,1], mvn[-1000,1])`, where `mvn` is a matrix containing the draws from one of the functions. Also for both samples, plot x_i against x_{i-1} and y_i against y_{i-1}.

(f) Generate histograms for both variables and both samplers, superimposing the normal density curves with the matching mean and variance and a kernel density estimate. Is there any evidence or non-normality in any of the plots.

2. Binomial Regression - Bottle Return

A carefully controlled experiment was conducted to study the effect of the size of the deposit level on the likelihood that a returnable one-litre soft-drink bottle will be returned. The following data show the number of bottles returned (y_i) out of 500 sold (n_i) at each of 6 deposit levels (x_i, in cents)

Observation i:	1	2	3	4	5	6
Deposit level x_i:	2	5	10	20	25	30
Number sold n_i:	500	500	500	500	500	500
Number returned y_i:	72	103	170	296	406	449

(a) Fit a logistic regression model to the above data. What is the fitted response function.

(b) Obtain an estimate of e^{β_1} and calculate a 95% confidence interval for the quantity.

(c) Fit a probit regression model to the same data.

(d) What is the estimated probability that a bottle will be returned when the deposit is 15 cents under both models. Similarly for 50 cents.

(e) Plot the estimated return probabilities $p_i = y_i/n_i$ against x_i. Superimpose on this plot curves of the estimated probabilities of return for the two models. Based on the information seen so far, is there any reason to prefer one model fit over the other?

(f) For the logistic regression model, estimate the deposit level x where 75% of the bottles are expected to be returned.