Normal Linear Models

- Probably the most widely-used models are linear normal models
- Advantages
  - Simple analytical formulas
  - Good first cut, easy integrations
  - Normal errors often a good approximation
- Disadvantages
  - Because they are familiar and tractable, it is easy to be lured into using these models without having given sufficient thought to the problem
  - Assumptions can be inadequate or wrong

Normal Linear Models

- The Bayesian procedure is unchanged
  \[ \text{posterior} \propto \text{prior} \times \text{likelihood} \] (the “Bayesian mantra”)
  but for the special distribution
  \[ N(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi \sigma}} \exp\left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \]
  and some priors to be determined

Normal Linear Models

- Normal distributions, like some others, have many characteristics that make them easy to use
- They should not be used because they have these useful characteristics. Computers can do calculations, and model error can be important. It’s better to use the right model and expend computer time than to use the wrong one because it is easier to calculate

\[ p(x_1, x_2, \ldots, x_n | \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left( -\frac{1}{2\sigma^2} \sum (x_i - \mu)^2 \right) \]

- As we did before, we complete the square:
  \[ \sum (x_i - \mu)^2 = \sum (x_i - \bar{x} + \bar{x} - \mu)^2 \]
  \[ = \sum (x_i - \bar{x})^2 + n(\mu - \bar{x})^2 \]
  \[ = S_{xx} + n(\mu - \bar{x})^2 \]

\[ \bar{x} = \frac{\sum x_i}{n} \]
\[ S_{xx} = \sum (x_i - \bar{x})^2 \]

\{Sufficient statistics\}
Normal Linear Models

- The likelihood function is then
\[ p(x_1,x_2,\ldots,x_n \mid \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp\left( -\frac{1}{2\sigma^2} (S_{xx} + n(\mu - \bar{x})^2) \right) \]

- There are a number of interesting cases, depending upon what is known and what the priors are
  - May know \( \sigma \) but have no prior information on \( \mu \)
  - May know \( \mu \) and have prior information on \( \sigma \)
  - May have no prior information on either \( \sigma \) or \( \mu \)
  - Etc.

Normal Linear Models

- Case 1: \( \mu \) unknown with no prior information
  - Choose flat (improper) prior \( p(\mu) = 1 \)
  - Posterior
  \[ p(\mu \mid X, \sigma^2) \propto \exp\left( -\frac{1}{2\sigma^2} n(\mu - \bar{x})^2 \right) \]
  - As usual I’ve suppressed irrelevant factors that will cancel out (\( \sigma \) is a constant)
  - The posterior is Normal:
  \[ \mu \mid X, \sigma \sim N(\bar{x}, \sigma^2 / n) \]

Normal Linear Models

- Example
  \[ x = \{10.84,10.69,10.19,9.28,8.97,9.88,10.24,9.22,9.45,9.80\} \]
  \[ \sigma = 1 \]
  we calculate
  \[ \bar{x} = 9.856 \]
  \[ S_{xx} = 3.161 \text{ (not needed!)} \]
  \[ \mu \mid X \sim N(9.856,1/10) \]

Normal Linear Models

- Using R or a table we see that for the normal distribution, the fraction of the probability contained within \( \pm f \sigma \) of the center of the distribution is

  - 90% within 1.645\( \sigma \)
  - 95% within 1.960\( \sigma \)
  - 99% within 2.567\( \sigma \)
  - Check this out, using R!
  - The 95% credible interval (HDR) is therefore
  \[ (9.236,10.476) = 9.856 \pm 1.96 \times \frac{1}{\sqrt{10}} \]
  \[ = 9.856 \pm 0.620 \]
Normal Linear Models

- Assumptions made
  - Observations independently and identically distributed
  - Error normally distributed
  - Variance known

- Comments
  - Formally, the error decreases as \( n \) increases. In practice, other factors may prevent the error from going to zero. For example, if there are model errors the error may never go to zero
  - Even if this were not true, we would need an enormous number of observations to get very large reductions in standard error. An error of 1\% of the error of the data requires of order \( 10^4 \) observations.

Example 2 (more interesting): Prior information on \( \mu \)
- Let the (conjugate) prior be \( \mu \sim N(\mu_0, \sigma_0^2) \). Then

\[
\text{pd} \propto \frac{1}{\sigma^4 \sigma_0^2} \exp \left[ -\frac{1}{2} \left( \frac{(\mu - \mu_0)^2}{\sigma_0^2} + \frac{S_{xx} + n(\mu - \bar{x})^2}{\sigma^2} \right) \right] \]

\[
\propto \exp \left[ -\frac{1}{2\sigma_0^2 \sigma^2} \left\{ \sigma^2 (\mu - \mu_0)^2 + \sigma_0^2 (S_{xx} + n(\mu - \bar{x})^2) \right\} \right] \]

\[
\propto \exp \left[ -\frac{1}{2\sigma_0^2 \sigma^2} \left\{ \mu^2 (\sigma^2 + \sigma_0^2 n) - 2 \mu (\sigma^2 \mu_0 + \sigma_0^2 n \bar{x}) \right\} \right] \]

\[
\propto \exp \left[ -\frac{\sigma^2 + \sigma_0^2 n}{2 \sigma_0^2 \sigma^2} \left( \mu - \frac{\sigma^2 \mu_0 + \sigma_0^2 n \bar{x}}{\sigma^2 + \sigma_0^2 n} \right)^2 \right] \]

By inspection, this is normal with

- Mean

\[
\text{mean} = \frac{\sigma^2 \mu_0 + \sigma_0^2 n \bar{x}}{\sigma^2 + \sigma_0^2 n} = \frac{\mu_0}{\sigma_0^2} + \frac{n \bar{x}}{\sigma^2}
\]

- Variance

\[
\text{variance} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}
\]
Bayesian Inference 10/4/07 14
Normal Linear Models

• To integrate out $\theta$ let

$$u = \frac{S_{xx} + n(\mu - \bar{x})^2}{2\theta} = A$$

$$p(\mu | X) \propto \int d\theta \frac{1}{\theta^{n/2+1}} \exp \left( -\frac{A}{\theta} \right)$$

$$\propto \int du \frac{A u^{(n/2+1)}}{u^2 A^{(n/2+1)}} \exp(-u)$$

$$\propto A^{-n/2}$$

$$\propto \left( 1 + \frac{n(\mu - \bar{x})^2}{S_{xx}} \right)^{-n/2}$$

Bayesian Inference 10/4/07 15
Normal Linear Models

• This is a student $t$ distribution on $\nu = n - 1$ degrees of freedom, as can be verified by making the substitutions

$$\nu = n - 1, \quad t = \frac{\mu - \bar{x}}{s\sqrt{n} / \nu}, \quad s^2 = \frac{S_{xx}}{\nu}$$

$$pd \propto \left( 1 + \frac{n(\mu - \bar{x})^2}{S_{xx}} \right)^{-n/2}$$

$$\propto \left( 1 + t^2 / \nu \right)^{-(\nu+1)/2}$$

• Verify this!

Bayesian Inference 10/4/07 16
Normal Linear Models

• The $t$ distributions are central, symmetric, resemble normal distributions, but have heavier “tails”

• As $\nu \to \infty$ they approach a standard normal distribution

• $\nu$ = number of degrees of freedom

• For small $\nu$, the $t$ distribution is markedly different from the normal distribution
Bayesian Inference 10/4/07 20

Normal Linear Models

- To get the distribution of \( \theta \), marginalize with respect to \( \mu \)

\[
\begin{align*}
    pd & \propto \int d\mu \frac{1}{\theta^{n/2+1}} \exp \left[ -\frac{1}{2\theta} (S_\mu + n(\mu - \bar{x})^2) \right] \\
    & \propto \frac{1}{\theta^{n/2+1}} \exp \left[ -\frac{S_\mu}{2\theta} \right] \int d\mu \exp \left[ -\frac{n(\mu - \bar{x})^2}{2\theta} \right] \\
    & \propto \frac{1}{\theta^{(n+3)/2}} \exp \left[ -\frac{S_\mu}{2\theta} \right] \propto \text{InverseGamma}(\theta | \alpha, \beta)
\end{align*}
\]

with \( \alpha = (n-1)/2 \), \( \beta = S_\mu/2 \)

- This is an inverse gamma distribution; the inverse chi-square distribution is a special case (Gelman et. al., Table A1 (pp. 574-5))

Bayesian Inference 10/4/07 18

Normal Linear Models

Comparison of Cauchy (t with \( \nu = 1 \)) and Normal Distributions

- Example: Same data as for Example 1, but now we don’t have prior knowledge of the variance

\[
\bar{x} = 9.856 \\
S_\bar{x} = 3.613 \\
n = 10, \quad \nu = 9
\]

\( S^2 = S_\bar{x} / \nu = 0.401 \)

To get the 95% credible interval (HDR) use R or a table of the \( t \) distribution. We get 2.27 for the standard \( t \) distribution (verify this) thus finding

\[
t = 0 \pm 2.27 = \frac{\mu - \bar{x}}{s/\sqrt{n}}
\]

\[
\mu = \bar{x} \pm s/\sqrt{n} \times 2.27 = 9.856 \pm 0.455
\]

Bayesian Inference 10/4/07 19

Normal Linear Models
Conjugate Priors for Normal Inference

- The idea is to choose a conjugate prior that is close enough to your actual prior so that the posterior distribution will be in the special family of distributions, thus making the mathematical problem simpler. Examples:
  - A normal prior on the mean with fixed variance $\sigma_0$, e.g.,
    \[ p(\mu \mid \sigma_0) \propto \exp \left[ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right] \]
    results in a normal posterior $p(\mu \mid y_1, \ldots, y_N, \sigma_0)$.

- Assume $y_k \sim N(f(x_k; a, b, \ldots), \sigma^2)$
- Likelihood is
  \[
  L \propto \prod_{k=1}^{N} \frac{1}{\sigma^N} \exp \left[ -\frac{1}{2\sigma^2} (y_k - f(x_k; a, b, \ldots))^2 \right] \\
  = \frac{1}{\sigma^N} \exp \left[ -\frac{1}{2\sigma^2} \sum_{k=1}^{N} (y_k - f(x_k; a, b, \ldots))^2 \right]
  \]

Conjugate Priors for Normal Inference

- The idea is to choose a conjugate prior that is close enough to your actual prior so that the posterior distribution will be in the special family of distributions, thus making the mathematical problem simpler. Examples:
  - A normal prior on the mean conditioned on the variance, with an inverse gamma prior on the variance, with parameters $n_0$, $\mu_0$, $S_{xx}$, and $\nu_0$, chosen so as to reproduce the shape of the desired prior, e.g.,
    \[ p(\mu, \sigma \mid \mu_0, S_{xx}, \nu_0) \propto \frac{1}{\alpha_{\nu_0+1}} \exp \left[ -\frac{1}{2\sigma^2} (S_{xx} + n_0(\mu - \mu_0)^2) \right] \]
    results in a posterior of the same form. Such a prior could, e.g., result from an earlier analysis.

Fitting Straight Lines

- Sometimes we have a general functional relationship between an observed $y$ and a precisely known $x$, e.g., $y = f(x; a, b, \ldots)$ would hold if there were no error in $y$
- Assume $y_k \sim N(f(x_k; a, b, \ldots), \sigma^2)$
- The Bayesian mantra:
  \[ \text{posterior} \propto \text{prior} \times \text{likelihood} \]
  with Jeffreys prior on $\sigma$ and flat prior on $a, b, \ldots$ yields
  \[ pd \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2\sigma^2} \sum_{k=1}^{N} (y_k - f(x_k; a, b, \ldots))^2 \right] \]
- Generally this can’t be solved in closed form; but we can use MCMC simulation to provide a sample from the posterior distribution
Bayesian Inference 10/4/07 28

Fitting Straight Lines

• Marginalization strategy: First marginalize out $\alpha$ or $\beta$ respectively; then $\sigma$ to get the marginal distribution on $\beta$ or $\alpha$ respectively. Marginalize out $\alpha$ and $\beta$ to get the marginal distribution on $\sigma$. For example, integrating out $\beta$ yields

$$pd(\alpha, \sigma \mid x, y) \propto \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2}\left(S_{\tau\tau} + N(\alpha - \bar{y})^2\right)\right]$$

losing one power of $\sigma$, and then getting rid of $\sigma$ yields

$$pd(\alpha \mid x, y) \propto \left(1 + \frac{N}{S_{\tau\tau}} (\alpha - \bar{y})^2\right)^{-\frac{N-1}{2}}$$

Bayesian Inference 10/4/07 26

Fitting Straight Lines

• An exception is when $f(x; a, b, \ldots) = a + bx$ (linear case)

$$pd \propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - a - bx_i)^2\right]$$

• This simplifies considerably if we write

$$a = \alpha - b\bar{x}, \quad b = \beta \quad \text{where} \quad \bar{x} \text{ is the sample mean of} \ x$$

$$f = \alpha + \beta(x - \bar{x})$$

• Then

$$pd \propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \alpha - \beta(x_i - \bar{x}))^2\right]$$

$$= \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - \bar{y} - (\alpha - \bar{y}) - \beta(x_i - \bar{x}))^2\right]$$

Bayesian Inference 10/4/07 27

Fitting Straight Lines

• With the definitions

$$\hat{\beta} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad \text{“average slope”}$$

$$S_{\tau\tau} = \sum (y_i - \bar{y})^2 - \hat{\beta} \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{\tau\tau} = \sum (x_i - \bar{x})^2$$

we get

$$pd \propto \frac{1}{\sigma^{N+1}} \exp\left[-\frac{1}{2\sigma^2}\left(S_{\tau\tau} + N(\alpha - \bar{y})^2 + S_{\tau\tau}(\beta - \hat{\beta})^2\right)\right]$$

Bayesian Inference 10/4/07 28

Fitting Straight Lines

• Likewise,

$$pd(\beta \mid x, y) \propto \left(1 + \frac{S_{\tau\tau}}{S_{\tau\tau}} (\beta - \hat{\beta})^2\right)^{-\frac{N-1}{2}}$$

and similarly,

$$pd(\sigma \mid x, y) \propto \frac{1}{\sigma^{N-1}} \exp\left(-\frac{S_{\tau\tau}}{2\sigma^2}\right)$$
Fitting Straight Lines

- The sum in the exponential,
  \[ \sum e_i^2 = \sum (y_i - \alpha - \beta(x_i - \bar{x}))^2 \]
  \[ = S_{\tau\tau} + N(\alpha - \bar{y})^2 + S_{\tau\tau}(\beta - \hat{\beta})^2 \]
  is minimized by setting
  \[ \alpha = \bar{y} = \hat{\alpha}, \quad \beta = \hat{\beta} \]

- The posterior is maximized at the same point. This is the familiar principle of least squares invented by Gauss, in the Bayesian context. The point that maximizes the posterior distribution is the so-called MAP (Maximum a posteriori) estimate, and it coincides with the least squares estimate when the data are normal and the priors are conventional.

Fitting Straight Lines

- The fact that \( \text{cov}(\alpha, \beta) = 0 \) is very special here, due to our having chosen to define \( \alpha \) and \( \beta \) as we did, and to the normal distribution of the data. We cannot expect this to happen in general.
- For example, if you substitute the definitions of \( \alpha \) and \( \beta \) to rewrite this in terms of \( a \) and \( b \), it is evident that \( \text{cov}(a, b) \neq 0 \)
  as the posterior distribution is
  \[ p(a, b \mid x, y) \propto (S_{\tau\tau} + N(a - \hat{a})^2 + 2N\tau(a - \hat{a})(b - \hat{b}) + S_{\tau\tau}(b - \hat{b})^2)^{-N/2} \]
  where we have set
  \[ a = \hat{a} + \hat{\beta}\tau, \quad b = \hat{b} \]

Fitting Straight Lines

- Returning to the posterior, if we marginalize out \( \sigma \) to get a posterior distribution on \( \alpha \) and \( \beta \), we find
  \[ p_d(\alpha, \beta \mid x, y) \propto \left( S_{\tau\tau} + N(\alpha - \hat{\alpha})^2 + S_{\tau\tau}(\beta - \hat{\beta})^2 \right)^{-N/2} \]
  - This cannot be written as a product
  \[ p(\alpha \mid \text{data}) \propto p(\beta \mid \text{data}) \]
  so \( \alpha \) and \( \beta \) are not independent; however, they are uncorrelated because their covariance is 0:
  \[ \text{cov}(\alpha, \beta) = E((\alpha - \hat{\alpha})(\beta - \hat{\beta})) = 0 \]

Fitting Straight Lines

- Generalizations
  - Variance in \( y \) varies from observation to observation, the so-called heteroskedastic case, or may be correlated
  - Error in \( x \) instead of \( y \)
  - Error in both \( x \) and \( y \), the errors-in-variables case, which requires a special treatment.
  - Significant prior information can be included without changing the mathematics much if you use multivariate normal priors on \( a \) and \( b \) and inverse gamma prior on \( \sigma \)
  - Anything more than that, and you will be forced to use MCMC or some other approximate method
Bayesian Inference 10/4/07 34

Heteroskedastic Case

- Assumptions must be made, e.g., assume
  \[ \sigma_i^2 = \sigma^2 / w_i \]
  where the \( w_i \) are fixed weights for each observation and \( \sigma^2 \)
  is an unknown variance of unit weight on which we can
  put a prior (e.g., a Jeffreys prior or an inverse-gamma
  prior). Then the likelihood is
  \[
p(x \mid a, b, \sigma) \propto \frac{1}{\sigma^n} \exp \left[ - \frac{1}{2\sigma^2} \sum w_i (y_i - a - bx_i)^2 \right]
  \]
- The analysis is not much different from the homoskedastic
  (equal weight) case
- The case of correlated observations is best handled by
  matrix methods (see below)

Bayesian Inference 10/4/07 35

Error in x instead of y

- Introduce latent variables \( X_i \), the “true” values of the
  observations \( x_i \). Then the likelihood is
  \[
p(x \mid a, b, \sigma) \propto \frac{1}{\sigma^n} \exp \left[ - \frac{1}{2\sigma^2} \sum (X_i - x_i)^2 \right]
  \]
  where \( X_i = (Y_i-a)/b \)
- An obvious way to handle this is to write \( \alpha = a/b, \beta = 1/b \)
  to get \( X_i = \alpha + \beta Y_i \), and compute the posterior on \( (\alpha, \beta) \);
  however, you have to be careful about your priors if you
  do this (i.e., if you thought the priors should have been flat
  on \( a \) and \( b \), they cannot be flat on \( \alpha \) and \( \beta \)).
- If you have priors specified on \( a \) and \( b \), a better approach
  might be to use MCMC to draw your sample from the
  posterior on \( (a, b) \) for inference.

Bayesian Inference 10/4/07 36

Error in both x and y (Errors-in-Variables)

- Introduce \( X_i \), the “true” values of the observations \( x_i \). We
  won’t be able to eliminate these latent variables as we did
  in the error-in-x example. The likelihood is
  \[
p(x, y \mid a, b, \sigma_x, \sigma_y, X) \propto \frac{1}{\sigma_x \sigma_y^n} \exp \left[ - \frac{1}{2\sigma_x^2} \sum (X_i - x_i)^2 \right]\]
  \[
  \times \exp \left[ - \frac{1}{2\sigma_y^2} \sum (Y_i - y_i)^2 \right]
  \]
  where \( Y_i = a + b X_i \)
- The latent variables are a complete new set of variables
  and each needs a prior, along with priors on \( a, b, \sigma_x, \sigma_y \).
  We will sample all of the variables, including \( X_i \). Here’s a
  case where MCMC will be useful.

Bayesian Inference 10/4/07 37

Error in both x and y (Errors-in-Variables)

- If we put Jeffreys priors on \( \sigma_x, \sigma_y \) and flat priors on \( X_i, a \)
  and \( b \) in the problem as stated we will run into difficulty.
  The reason is that the posterior distribution of \( \phi = \sigma_y^2/\sigma_x^2 \)
  is improper due to the fact that the data give no
  information about \( \sigma_x \) given \( \sigma_y \)
- This is partly due to the fact that as we increase the
  amount of data, we are also increasing the number of
  parameters (the \( X_i \)) in step. As a consequence, there is no
  unique maximum of the posterior distribution.
  Statisticians would say that \( \sigma_x \) and \( \sigma_y \) are
  unidentified.
Error in both $x$ and $y$ (Errors-in-Variables)

- This situation can be remedied in a number of ways, e.g.,
  - Taking informative priors on $\sigma_x$, $\sigma_y$
  - Fixing the ratio $\varphi = \sigma_y^2/\sigma_x^2$ at some particular value
  - Choosing informative priors on the $X_i$
  - Some other way consistent with prior knowledge
- See Arnold Zellner, *An Introduction to Bayesian Inference in Econometrics*, Wiley (1971), Chapter V.

MCMC Approaches

- Even though the problem of fitting straight lines to normal data on $y$ (and similar linear problems) can be solved exactly, it frequently happens that such fits are embedded in more complicated problems for which simulation would be necessary in any case. Thus, we need to study strategies for including either Gibbs or Metropolis-Hastings steps on a normal linear subproblem within a larger MCMC simulation
- This means studying how to simulate the normal problem efficiently

MCMC Approaches

- Consider for example a target distribution of the form
  
  $$pd \propto \exp \left[ -\frac{1}{2} (ax^2 + 2bxy + cy^2) \right]$$

  where $a$, $b$, and $c$ are fixed numbers, presumably functions of the data
- One might have gotten this by “completing the square” for a particular data set—owing to the existence of sufficient statistics for normal problems
- (Difficult): Suppose $a=c=5.25$, $b=-5.24$. Then the level curves of the equation $pd=\text{constant}$ will be very elongated ellipses oriented at a $45^\circ$ angle to the axes. Let us look at a Metropolis-Hastings approach to simulating this

Here’s a program to do this. First, the log posterior

$$
\begin{align*}
  a &= 5.25 \\
  c &= a \\
  b &= -5.24 \ # \text{ Set values required} \\
  \text{logpost} &= \text{function}(x,y) \{ \\
  &\quad -(a*x^2+2*b*x*y+c*y^2)/2 \\
  \}
\end{align*}
$$
MCMC Approaches

- We propose a step centered on our current position with a bivariate uniform distribution of a specified width

```r
samplexy = function(x,y,dx,dy) {
xstar = x + runif(1,-dx,dx)
ystar = y + runif(1,-dy,dy)
logalpha = logpost(xstar,ystar)-logpost(x,y)
accept = (log(runif(1,0,1)) < logalpha)
x = xstar*accept + x*(1-accept)
y = ystar*accept + y*(1-accept)
list(x=x,y=y,accept=accept)
}
```

MCMC Approaches

- The approach given is not ideal. A better idea would be to do a Gibbs step that attempts to sample from the actual posterior conditional distribution on the \(x,y\) variables. This would take advantage of what we know about the solution of the linear problem to sample along the major axis of the posterior distribution.
- This would give more rapid convergence to the stationary distribution and better sampling of the posterior distribution.
- We would need to sample from the multivariate normal distribution implied by \(a, b, c\).

```r
Q = a b
c b
```

MCMC Approaches

- We can write the matrix implied by \(a, b, c\) as follows:

```r
Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix}
```

- We want to find an upper triangular matrix \(R\) with the property that \((R \ R^\top)^{-1} = Q\). Then if \(z\) is a sample from a standard bivariate normal distribution, then \(u=Rz\) (or \(u=zR\), which is faster, using row vectors for \(u\) and \(z\)) is a sample from a bivariate normal distribution with mean 0 and variance-covariance matrix \(Q^{-1}\). This will be our desired multivariate Gibbs sample.
- The prime (´) means matrix transpose

```r
samplen = function(n=3000,dx=2,dy=2,x=0,y=0) {
  xs = rep(NaN,n)
y = rep(NaN,n)
  accepts = 0
  for(i in 1:n){
    z = samplexy(x,y,dx,dy)
x[i] = x = z$x
y[i] = y = z$y
    accepts = accepts + z$accept
  }
invisible(list(x=xs,y=ys,acc=accepts))
}
```
MCMC Approaches

• In R, we can accomplish this in the following way:

\[
Q = c(a, b, c) \\
\text{dim}(Q) = c(2, 2) \# \text{form matrix} \\
R = \text{chol}(\text{solve}(Q))
\]

\[
samplexy = \text{function}(R) \{
    u = \text{rnorm}(2) \%\% R \\
    \text{return}(u)
\}
\]

• `t` transposes a matrix and `solve` inverts a matrix (`solve` can also solve linear systems, type `?solve` to learn more). `chol` produces the upper triangular matrix we desire. It performs a *Cholesky decomposition* of the symmetric, positive definite square matrix \( Q \).

---

MCMC Approaches

• Often it happens that the linear problem is an approximation to a nonlinear problem, e.g., the observations \( y \) may be related to \( a, b \) through nonlinear relations like \( y = f(x; a, b, \ldots) \) so that the linearized equations can be solved but not these nonlinear equations. In such a case it would be convenient to use the solution of the linear problem (essentially the Laplace approximation) as our proposal distribution for a Metropolis-Hastings sampler, but use the exact likelihood

\[
p(y | a, b, x) \propto \frac{1}{\sigma^n} \exp \left[ -\frac{1}{2\sigma^2} \sum (y_i - f(x_i; a, b, \ldots))^2 \right]
\]

to compute the transition matrix

\[
\text{samplexy} = \text{function}(R) \{
    u = c * \text{rt}(2, \text{dof}=4) \%\% R + \text{alphahat} \\
    \ldots \# \text{Accept/reject step code here} \\
    \ldots
\}
\]

In this nonlinear case (and in general), it is important to make sure that the “tails” of the posterior distribution are sampled adequately. This can be accomplished by using a fat-tailed proposal distribution like a multivariate Cauchy or \( t \) distribution rather than a multivariate normal distribution, e.g., something like

\[
\text{samplexy} = \text{function}(R) \{
    u = c * \text{rt}(2, \text{dof}=4) \%\% R + \text{alphahat} \\
    \ldots \# \text{Accept/reject step code here} \\
    \ldots
\}
\]

where \( R \) is the matrix arising from the solution of the linearized least squares problem, \( \text{alphahat} \) is the desired mean and \( c \) a constant we choose for efficient sampling.
Multivariate Linear Models

• Suppose we have \( n \) observations \( y=(y_1, y_2, \ldots, y_n) \) and a set of \( k \) parameters \( \theta=(\theta_1, \theta_2, \ldots, \theta_k) \), which would in the absence of error be related by \( y=A\theta \), where \( A \) is a matrix of known, fixed numbers. We presume that \( y \sim N(\theta, \sigma^2I) \). The matrix \( A \) is called the design matrix of the problem. In general the elements of \( A \) could be complicated functions of some other variables, e.g., \( x \) or time, but that is not necessary.

• The likelihood function in the iid case is

\[
L \propto \frac{1}{\sigma^n} \exp \left[ -\frac{1}{2\sigma^2} (y - A\theta)'(y - A\theta) \right] = \frac{1}{\sigma^n} \exp \left[ -\frac{S}{2\sigma^2} \right]
\]

Multivariate Linear Models

• With a Jeffreys prior on \( \sigma \), the posterior distribution is seen to be

\[
p(\theta, \sigma \mid y) \propto \sigma^{-(n+1)} \exp \left( -\frac{1}{2\sigma^2} (S_\theta + S_e) \right)
\]

• As before, the marginalization strategy is to integrate out every \( \hat{\theta}_j \) except the one we are interested in, and then to integrate out \( \sigma \). Each of the \( (k-1) \) integrations with respect to \( \hat{\theta}_j \) loses one power of \( \sigma \) in the posterior. After integrating out \( \sigma \) we get a \( t \) distribution with \( v=n-k \) degrees of freedom where

\[
t = \frac{\theta_1 - \hat{\theta}_1}{s \sqrt{m_{kk}}}, \quad s^2 = S_e / v
\]

and \( m_{kk} \) is the \((kk)\) element of \( M=(A' A)^{-1} \)

Multivariate Linear Models

• Define the ordinary least squares estimator of \( \theta \) by

\[
\hat{\theta} = (A' A)^{-1} A'y
\]

• Then a simple calculation shows that

\[
S = (\hat{\theta} - \theta)' A' A (\hat{\theta} - \theta) + (y - A\hat{\theta})'(y - A\hat{\theta})
\]

\[
= S_\theta + S_e
\]

• Be prepared to show this in class!

Hint: In the original equation for \( S \), write \( \theta = \hat{\theta} + (\theta - \hat{\theta}) \). Expand this out, and choose \( \hat{\theta} \) so that the coefficient of the linear term in \( (\theta - \hat{\theta}) \) vanishes. This is the analog of “completing the square” for this matrix formula

Multivariate Linear Models

• The posterior marginal distribution for \( \sigma^2 \) is inverse gamma.

• Also, \( \theta = \hat{\theta} \) minimizes the sum-of-squares of the residuals and maximizes the posterior; this is the MAP solution of the problem, and is also the least-squares solution. So, in the case of a multivariate linear normal model with these priors, the Bayesian MAP solution and the ordinary least-squares solution coincide (numerically).
Multivariate Linear Models

- Finally, by evaluating at \( \theta = \hat{\theta} \) we can expand about the maximum, obtaining the Laplace approximation

\[
p(\theta | \mathbf{y}) \propto (\nu s^2 + S_j(\theta))^{-n/2}
\]

\[
\propto \left(1 + \frac{S_j(\theta)}{\nu s^2}\right)^{-n/2}
\]

How do we get this?

- This is a multivariate normal distribution in \( \theta \) with mean \( \hat{\theta} \) and covariance matrix

\[
E((\theta - \hat{\theta})(\theta - \hat{\theta})') = s^2 (A'A)^{-1}
\]

Heteroskedastic Case

- If the errors vary from observation to observation there is a minor complication
- Let \( E((y - A\theta)(y - A\theta)') = \sigma^2 W^{-1} \) be the covariance matrix of the observations so that \( y \sim N(A\theta, \sigma^2 W^{-1}) \). Then the likelihood is slightly modified:

\[
L \propto \frac{1}{\sigma^n} \exp \left[ -\frac{1}{2\sigma^2} (y - A\theta)' W(y - A\theta) \right]
\]

- The solution is obtained by slightly modifying the definitions of \( S_c, S_j(\theta) \) and \( \hat{\theta} \).
- See if you can get the following posterior!

\[
p(\theta, \sigma | \mathbf{y}) \propto \sigma^{-(n+1)} \exp \left[ -\frac{1}{2\sigma^2} \left( \hat{S}_c + \hat{S}_j(\theta) \right) \right]
\]

Correlated Errors

- This is handled by noting that \( W \) does not have to be diagonal. Formally the solution is the same as in the heteroskedastic case
Informative Prior on the Parameters

- We can put an informative prior on the parameters for a multivariate normal prior by writing,
  \[ p(\theta) \propto \exp\left[ -\frac{1}{2}(\theta - \theta_0)^T Q^{-1}(\theta - \theta_0) \right] \]
  where \( Q \) is a positive-definite symmetric matrix and \( \theta_0 \) is the most probable prior value of \( \theta \)
- Show by completing the square that the linear terms in \( (\theta - \hat{\theta}) \) can be made to vanish by choosing
  \[
  \hat{\theta} = \left[ \frac{1}{\sigma^2} A' A + Q^{-1} \right]^{-1} \left[ A'y + Q^{-1} \theta_0 \right]
  \]
- This is a kind of average between prior and likelihood, related to "ridge regression."

Posterior Prediction

- Often it is necessary to make predictions of quantities to be observed in the future, based on our best inferences on parameters determined through observations.
- For example, we might observe a phenomenon \( y \) at times \( \{x_1, x_2, \ldots, x_n\} \) and wish to predict it at a later time \( x \)
- In this case it is possible to make the predictions using probability theory, e.g., suppose we can estimate the true value of \( y \) at a given \( x \) by
  \[
  y = a + bx
  \]
  \[
  pd(a, b, \sigma \mid D) \propto p(D \mid a, b, \sigma) p(a, b, \sigma)
  \]
  \[
  p(y \mid a, b, \sigma, x) = \delta(y - a - bx)
  \]
  \[
  p(y \mid D, x) = \iint \delta(y - a - bx) p(D \mid a, b, \sigma) p(a, b, \sigma) da db d\sigma
  \]
Posterior Prediction

• We see that as a function of $x$, the predicted values of $y$ lie along a straight line; but the standard deviation increases asymptotically linearly in $x$

$$y = \hat{a} + \hat{b}x$$

• Suppose we are interested in predicting $Y$, where $Y$ is observed with error. This can also be done easily using MCMC. Use the results of the previous chart to provide a sample from $N(y, \sigma^2)$ where $(y, \sigma^2)$ are provided by our MCMC sample.

• This is the posterior predictive model for $Y$.

Posterior Prediction

• Suppose we have used MCMC to produce a sample from the posterior distribution $p(\theta \mid D)$, where $\theta$ are our parameters (including the variance of the observations if we sample that). We wish to predict the distribution of new observations $Y$ where if there were no error we would have $y = f(\theta)$; we presume that $p(Y \mid y, \theta)$ is known.

• If we have a sample from the posterior distribution on parameters $\theta$ and $\sigma$, it is easy to use the sample to produce a sample from $p(y \mid \theta, \sigma)$. Simply make a draw from $p(y \mid \theta, \sigma)$ for the given value of $x$ over the MCMC samples of $(\theta, \sigma)$.

• Then we can compute expectations of $y \mid D, x$, variances, credible intervals, etc., from the sample.

Gibbs Sampling $a$ and $b$

• Here’s our EIV model posterior probability ($\alpha_i = \alpha, \sigma$):

$$p(a, b, \sigma \mid x, y) \propto \frac{1}{\sigma^{2n+1}} \exp \left[ -\frac{1}{2\sigma^2} \sum (X_i - x) \right]$$

$$\times \exp \left[ -\frac{1}{2\sigma^2} \sum (a + bx_i - y_i) \right]$$

• It’s simplest to use matrices. Introduce matrices

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad Z = \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_n \end{bmatrix}, \quad \alpha = \begin{bmatrix} a \\ b \end{bmatrix}$$
Gibbs Sampling \(a\) and \(b\)

- Then the quadratic in the exponential becomes:
  \[
  U = \frac{1}{\sigma^2}[(X - x)'(X - x) + (Z\alpha - y)'(Z\alpha - y)]
  \]
- Define \(\alpha = \hat{\alpha} + \Delta\alpha\) where \(\hat{\alpha}\) is to be determined by completing the square. We get
  \[
  U = \frac{1}{\sigma^2}[(X - x)'(X - x) + (Z\hat{\alpha} - y + Z\Delta\alpha)'(Z\hat{\alpha} - y + Z\Delta\alpha)]
  \]
  \[
  = \frac{1}{\sigma^2}[(X - x)'(X - x) + (Z\hat{\alpha} - y)'(Z\hat{\alpha} - y) + \Delta\alpha'(ZZ)\Delta\alpha + 2\Delta\alpha Z'(Z\hat{\alpha} - y)]
  \]
  \(C\): Constant with respect to \(\Delta\alpha\)

Comment: We see that

\[
Q = \sigma^{-2}(Z'Z) = \frac{(Z/\sigma)'(Z/\sigma)}{\sigma^2}
\]

- The \(R\) code to generate the matrix \(R\) that we need to do our Gibbs sampling is as follows. The new function \(\text{cbind}\), which stands for “bind columns,” is used to create the matrix \(Z\) out of vectors. Do \(\text{help(cbind)}\) for more information. \(\text{rep(1, n)}\) creates a vector of \(n\) 1’s

\[
\begin{align*}
Z &= \text{cbind}(\text{rep(1, n)}, X) \\
W &= \text{solve}(t(Z) \%\% Z) \\
R &= \text{chol}(W)\*\text{sigma} \\
\text{alphahat} &= y \%\% Z \%\% W \\
\text{alpha} &= \text{rnorm}(2, 0, 1) \%\% R + \text{alphahat}
\end{align*}
\]

Gibbs Sampling \(a\) and \(b\)

- Imposing the condition \(Z'(Z\hat{\alpha} - y) = 0\) completes the square and eliminates the “mixed” term in the quadratic form. This leaves us with
  \[
  \hat{\alpha} = (Z'Z)^{-1}Z'y
  \]
  \[
  U = \frac{1}{\sigma^2}[C + \Delta\alpha'(Z'Z)\Delta\alpha]
  \]
- Thus we accomplish our Gibbs step for \(\alpha\) by sampling from a multivariate normal distribution with mean \(\hat{\alpha}\) and precision matrix \(Q = \sigma^{-2}(Z'Z)\)

Sampling on the Latent Variables

- Recall the \(X\), the “true” values of the observations \(x\). We will have to sample over these as well. The posterior is
  \[
  p(a, b, \sigma, X | x, y) \propto \frac{1}{\sigma^{2n+1}} \exp \left[-\frac{1}{2\sigma^2} \sum (X_i - x)^2 \right] \times \exp \left[-\frac{1}{2\sigma^2} \sum (a + bX_i - y)^2 \right]
  \]
- Complete the square. We get the result on the next page:
Sampling on the Latent Variables

- Sample as follows:
  
  \( X_i \sim N(u_i, \tau^2) \)

  where

  \[
  u_i = \frac{x_i + b(y_i - a)}{1 + b^2}
  \]

  \[
  \tau^2 = \frac{\sigma_x^2}{1 + b^2}
  \]

  Note that one can sample on all the \( X_i \) at once, since they are independent of each other (although dependent on the current values of \( a \) and \( b \)). This can be in a single line in R.

- Similar and only slightly more complicated formulas result if \( \sigma_x \neq \sigma_y \).