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

- 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

Normal Linear Models

- The Bayesian procedure is unchanged

\[ \text{posterior} \propto \text{prior} \times \text{likelihood} \quad (\text{the “Bayesian mantra”}) \]

but for the special distribution

\[ N(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left( -\frac{1}{2\sigma^2} (x - \mu)^2 \right) \]

and some priors to be determined.

Normal Linear Models

- Assume \( x = \{x_1, x_2, \ldots, x_n\} \) are independent samples from some normal distribution (“\( \text{iid } N(\mu, \sigma^2) \)”)
- The likelihood function is

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

- 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

\[ \text{Cross terms cancel} \]

Show this!
Normal Linear Models

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

- 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

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

- 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
  \[ \bar{x} = 9.856 \]
  \[ S_\mu = 3.161 \text{ (not needed!)} \]
  \[ \mu \mid x \sim N(9.856, 1/10) \]

- 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:
      \[ \mu \mid x, \sigma^2 \sim N(\bar{x}, \sigma^2 / n) \]

- 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
  - 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 since $\sigma n^{1/2}=0.01\sigma$ implies $n=10^4$

Normal Linear Models

- Example 2 (more interesting): Prior information on $\mu$
  - Let the (conjugate) prior be $\mu \sim N(\mu_0, \sigma_0^2)$. Then
    \[
    pd \propto \frac{1}{\sigma_0^2} \exp \left[ - \frac{1}{2} \left( \frac{(\mu - \mu_0)^2}{\sigma_0^2} + \frac{S_x + n(\mu - \bar{x})^2}{\sigma^2} \right) \right] 
    \]
    \[
    \propto \exp \left[ - \frac{1}{2\sigma_0^2} \left( \sigma^2 (\mu - \mu_0)^2 + \sigma_0^2 (S_x + n(\mu - \bar{x})^2) \right) \right] 
    \]
    \[
    \propto \exp \left[ - \frac{1}{2\sigma_0^2} \left( \mu^2 (\sigma^2 + \sigma_0^2 n) - 2\mu (\sigma^2 \mu_0 + \sigma_0^2 n \bar{x}) \right) \right] 
    \]
    \[
    \propto \exp \left[ - \frac{\sigma^2 + \sigma_0^2 n}{2\sigma_0^2} \left( \mu^2 - 2\mu \frac{(\sigma^2 \mu_0 + \sigma_0^2 n \bar{x})}{(\sigma^2 + \sigma_0^2 n)} \right) \right] 
    \]

Normal Linear Models

- Completing the square,
  \[
  pd \propto \exp \left[ - \frac{\sigma^2 + \sigma_0^2 n}{2\sigma_0^2} \left( \mu^2 - 2\mu \frac{(\sigma^2 \mu_0 + \sigma_0^2 n \bar{x})}{(\sigma^2 + \sigma_0^2 n)} \right) \right] 
  \]
  \[
  \propto \exp \left[ - \frac{\sigma^2 + \sigma_0^2 n}{2\sigma_0^2} \left( \mu - \frac{(\sigma^2 \mu_0 + \sigma_0^2 n \bar{x})}{(\sigma^2 + \sigma_0^2 n)} \right)^2 \right] 
  \]
  Simplifying, by inspection this is normal with
  \[
  \text{mean} = m = w\mu_0 + (1-w)\bar{x}, \quad \text{variance} = \frac{1}{s^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}, \quad \text{where} \quad w = \frac{s^2}{\sigma_0^2} 
  \]
  \[
  \text{Precision} = \frac{1}{\text{variance}} = \frac{s^2}{\sigma_0^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}, \quad \text{where} \quad \text{Weight of prior} = w = \frac{s^2}{\sigma_0^2} 
  \]
  \[
  \text{Precision of result} = \text{precision of prior} + \text{precision of data} 
  \]

Weighted average (weighted by precision)
Normal Linear Models

- The most common case is one we treated before: unknown mean and variance
- Take flat prior for $\mu$ and Jeffreys prior for $\theta = \sigma^2$
- Then posterior $\propto$ prior $\times$ likelihood
  \[ pd \propto \frac{1}{\theta} \theta^{n/2} \exp\left( -\frac{1}{2\theta} (S_{xx} + n(\mu - \bar{x})^2) \right) \]
- The posterior is symmetric in $\mu$ but asymmetric in $\theta$ or $\sigma$
- Note that we could have used $\sigma$ instead of $\theta$ if we wished, but for the calculation that follows, $\theta$ is slightly more convenient.

- To integrate out $\theta$ let
  \[ u = \frac{S_{xx} + n(\mu - \bar{x})^2}{2\theta} = \frac{A}{\theta} \]
  \[ p(\mu | X) \propto \int \frac{1}{\theta^{n/2+1}} \exp\left( -\frac{A}{\theta} \right) d\theta \]
  \[ \propto \int \frac{A}{u^2} \frac{u^{(n/2)+1}}{A^{(n/2)+1}} \exp(-u) du \]
  \[ \propto A^{-n/2} \]
  \[ \propto \left( 1 + \frac{n(\mu - \bar{x})^2}{S_{xx}} \right)^{-n/2} \]

- This is a student $t$ distribution on $v = n-1$ degrees of freedom, as can be verified by making the substitutions
  \[ v = n - 1, \quad t = \frac{\mu - \bar{x}}{s / \sqrt{n}}, \quad s^2 = \frac{S_{xx}}{v} \]
  \[ pd \propto \left( 1 + \frac{n(\mu - \bar{x})^2}{S_{xx}} \right)^{-n/2} \]
  \[ \propto \left( 1 + t^2 / v \right)^{-(v+1)/2} \]
  - Verify this!
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
- $\nu = \text{number of degrees of freedom}$
- For small $\nu$, the $t$ distribution is markedly different from the normal distribution

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

$\bar{x} = 9.856$

$S_{\bar{x}} = 3.613$

$n = 10, \; \nu = 9$

$s^2 = S_{\bar{x}} / \nu = 0.401$

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

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

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

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

To get the distribution of $\theta$, marginalize with respect to $\mu$

$pd \propto \int d\mu \frac{1}{\theta^{\nu/2+1}} \exp\left[-\frac{1}{2\theta} (S_{\bar{x}} + n(\mu - \bar{x})^2)\right]$

$\propto \frac{1}{\theta^{\nu/2+1}} \exp\left[-\frac{S_{\bar{x}}}{2\theta}\right] \int d\mu \exp\left[-\frac{n(\mu - \bar{x})^2}{2\theta}\right]$

$\propto \frac{1}{\theta^{(n-1)/2}} \exp\left[-\frac{S_{\bar{x}}}{2\theta}\right] \propto \text{InverseGamma}(\theta | \alpha, \beta)$

with $\alpha = (n-1)/2, \; \beta = S_{\bar{x}} / 2$

- This is an inverse gamma distribution (Albert, p. 188); the inverse chi-square distribution is a special case (Albert, p. 58)
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 $\sigma_0^2$, e.g.,
    $$p(\mu | \sigma_0) \propto \exp\left[-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right]$$
    results in a normal posterior

Regression

- Sometimes we have a general functional relationship between an observed $y$ and a precisely known $x$, e.g., $y = f(x; \beta_0, \beta_1, \ldots)$ would hold if there were no error in $y$
- Assume $y_k \sim N(f(x_k; \beta_0, \beta_1, \ldots), \sigma^2)$
- Likelihood is
  $$L \propto \prod_{k=1}^{N} \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2}(y_k - f(x_k; \beta_0, \beta_1, \ldots))^2\right]$$
  $$= \frac{1}{\sigma^N} \exp\left[-\frac{1}{2\sigma^2} \sum_{k=1}^{N}(y_k - f(x_k; \beta_0, \beta_1, \ldots))^2\right]$$
• The Bayesian mantra:

\[ \text{posterior} \propto \text{prior} \times \text{likelihood} \]

with Jeffreys prior on \( \sigma \) and flat prior on \( \beta_0, \beta_1, \ldots \) yields

\[
\text{pd} \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \sum_{k=1}^{N} (y_k - f(x_k; \beta_0, \beta_1, \ldots))^2 \right]
\]

• Generally this can’t be solved in closed form; but we can use MCMC simulation to provide a sample from the posterior distribution

---

**Simple Linear Regression**

• An exception is when \( f(x; \beta_0, \beta_1, \ldots) = \beta_0 + \beta_1 x \) (linear case)

\[
\text{pd} \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \sum_{k=1}^{N} (y_k - \beta_0 - \beta_1 x_k)^2 \right]
\]

• This simplifies considerably in “centered variables”

\[ \alpha_0 = \beta_0 + \beta_1 \bar{x}, \quad \alpha_1 = \beta_1, \text{ where } \bar{x} \text{ is the sample mean of } x \]

\[ f = \alpha_0 + \alpha_1 (x - \bar{x}) \]

• Then

\[
\text{pd} \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \sum_{k=1}^{N} (y_k - \alpha_0 - \alpha_1 (x_k - \bar{x}))^2 \right]
\]

\[
= \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \sum_{k=1}^{N} ((y_k - \bar{y}) - (\alpha_0 - \bar{y}) - \alpha_1 (x_k - \bar{x}))^2 \right]
\]

---

**Regression**

• The Bayesian mantra:

\[ \text{posterior} \propto \text{prior} \times \text{likelihood} \]

with Jeffreys prior on \( \sigma \) and flat prior on \( \beta_0, \beta_1, \ldots \) yields

\[
\text{pd} \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \sum_{k=1}^{N} (y_k - f(x_k; \beta_0, \beta_1, \ldots))^2 \right]
\]

• Generally this can’t be solved in closed form; but we can use MCMC simulation to provide a sample from the posterior distribution

(Can be a vector!)

---

**Simple Linear Regression**

• With the definitions

\[
\hat{\alpha}_i = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{\sum (x_k - \bar{x})^2} \quad \text{“average slope”}
\]

\[
S_{xx} = \sum (y_k - \bar{y})^2 - \hat{\alpha}_i \sum (x_k - \bar{x})(y_k - \bar{y})
\]

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

we get

\[
\text{pd} \propto \frac{1}{\sigma^{N+1}} \exp \left[ -\frac{1}{2 \sigma^2} \left( S_{xx} + N(\alpha_0 - \bar{y})^2 + S_{xx}(\alpha_1 - \hat{\alpha}_1)^2 \right) \right]
\]
Simple Linear Regression

• Marginalization strategy: First marginalize out $\alpha_0$ or $\alpha_1$ respectively; then $\sigma$ to get the marginal distribution on $\alpha_1$ or $\alpha_0$ respectively. Marginalize out $\alpha_0$ and $\alpha_1$ to get the marginal distribution on $\sigma$. For example, integrating out $\alpha_1$ yields

$$pd(\alpha_0, \sigma \mid x,y) \propto \frac{1}{\sigma^2} \exp \left[ -\frac{1}{2\sigma^2} \left( S_{xx} + N(\alpha_0 - \bar{y})^2 \right) \right]$$

losing one power of $\sigma$, and then integrating out $\sigma$ yields

$$pd(\alpha_0 \mid x,y) \propto \left( 1 + \frac{N}{S_{xx}} (\alpha_0 - \bar{y})^2 \right)^{-\frac{N-1}{2}}$$

This is a "$t$" distribution with $N-2$ degrees of freedom.

• The sum in the exponential,

$$\sum \varepsilon_k^2 = \sum (y_k - \alpha_0 - \alpha_1 (x_k - \bar{x}))^2$$

$$= S_{xx} + N(\alpha_0 - \bar{y})^2 + S_{xx} (\alpha_1 - \hat{\alpha}_1)^2$$

is minimized by setting

$$\alpha_0 = \bar{y} = \hat{\alpha}_0, \quad \alpha_1 = \hat{\alpha}_1$$

• 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 flat-Jeffreys.

Simple Linear Regression

• Likewise,

$$pd(\alpha_1 \mid x,y) \propto \left( 1 + \frac{S_{yy}}{S_{xx}} (\alpha_1 - \hat{\alpha}_1)^2 \right)^{-\frac{N-1}{2}}$$

and similarly,

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

• Returning to the posterior, if we marginalize out $\sigma$ to get a posterior distribution on $\alpha_0$ and $\alpha_1$, we find

$$pd(\alpha_0, \alpha_1 \mid x,y) \propto \left( S_{xx} + N(\alpha_0 - \hat{\alpha}_0)^2 + S_{xx} (\alpha_1 - \hat{\alpha}_1)^2 \right)^{-N/2}$$

• This cannot be written as a product

$$p(\alpha_0 \mid \text{data}) \times p(\alpha_1 \mid \text{data})$$

so $\alpha_0$ and $\alpha_1$ are not independent; however, they are uncorrelated because their covariance is 0:

$$\text{cov}(\alpha_0, \alpha_1) = E((\alpha_0 - \hat{\alpha}_0)(\alpha_1 - \hat{\alpha}_1)) = 0$$
Simple Linear Regression

- The fact that \( \text{cov}(\alpha_0, \alpha_1) = 0 \) is very special here, due to our having chosen to define \( \alpha_0 \) and \( \alpha_1 \) 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 \( \alpha_0 \) and \( \alpha_1 \) to rewrite this in terms of \( \beta_0 \) and \( \beta_1 \), it is evident that \( \text{cov}(\beta_0, \beta_1) \neq 0 \) as the posterior distribution is

\[
pd(\beta_0, \beta_1 | x, y) \propto \left( S_w + N(\beta_0 - \hat{\beta}_0)^2 + 2N\hat{\beta}_0(\beta_1 - \hat{\beta}_1) + S_w(\beta_1 - \hat{\beta}_1)^2 \right)^{-N/2}
\]

where we have set 
\[
\hat{\beta}_0 = \hat{\alpha}_0 + \hat{\alpha}_1 x, \quad \hat{\beta}_1 = \hat{\alpha}_1
\]

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

\[
p(x | a, b, \sigma) \propto \frac{1}{\sigma^2} \exp \left[ -\frac{1}{2\sigma^2} \sum w_i (y_i - \alpha_0 - \alpha_1 x_i)^2 \right]
\]

- 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)

Error in \( x \) instead of \( y \)

- If we have error in \( x \) instead of \( y \), we could write

\[
x = \xi + \delta
\]

\[
y = \alpha_0 + \alpha_1 \xi
\]

where we have introduced a \textit{vector of latent variables} \( \xi \), the components of which are the (unknown) \textit{true} values of the corresponding components of the data vector \( x \), and \( \delta \) is the \textit{vector of (normal) errors} for each observation.
- Therefore, we can write the likelihood as follows:

\[
p(x | \alpha_0, \alpha_1, \sigma) \propto \frac{1}{\sigma^2} \exp \left[ -\frac{1}{\sigma^2} \sum (x_i - \xi_i)^2 \right]
\]

where \( \xi_i = (y - \alpha_0) / \alpha_1 \)
Error in $x$ instead of $y$

- So the posterior on flat-Jeffreys would be
  \[ p(x \mid \alpha_0, \alpha_1, \sigma) \propto \frac{1}{\sigma} \exp \left[ -\frac{1}{\sigma^2} \sum (x_i - \xi_i)^2 \right] \]
  with \( \xi = (y - \alpha_0)/\alpha_1 \)
- It is tempting to write \( \beta_0 = \alpha_0/\alpha_1, \beta_1 = 1/\alpha_1 \) to get \( \xi = \beta_0 + \beta_1 y \), and compute the posterior on \((\beta_0, \beta_1)\); but this may be problematic, since flat priors on the \( \alpha' \)s do not translate into flat priors on the \( \beta' \)s.
- If you have priors specified on \( \alpha_0 \) and \( \alpha_1 \), a better approach for inference might be to use MCMC to draw your sample from the posterior on \((\alpha_0, \alpha_1)\).

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

- Analogously to the case where the error is in $x$ instead of $y$, we can write
  \[ x = \xi + \delta \]
  \[ y = \eta + \epsilon = \alpha_0 + \alpha_1 \xi + \epsilon \]
  where now \( \epsilon \) is a vector of the errors in the $y$'s, and \( \eta \) is a vector of the true values of the $y$'s; the \( \eta ' \)s are a new set of latent variables.

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

- If we put Jeffreys priors on \( \alpha_x, \sigma_x \) and flat priors on \( \xi, \alpha_0 \) and \( \alpha_1 \) in the problem as stated we will run into difficulty. The reason is that the posterior distribution of \( \phi = \sigma^2_x/\sigma^2_y \) 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 \( \xi \)) 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 $\phi = \sigma_y^2/\sigma_x^2$ at some particular value
  - Choosing informative priors on the $\xi_i$
  - 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
  \[ pd \propto \exp\left[-\frac{1}{2}(a\xi^2 + 2b\xi\eta + c\eta^2)\right] \]
  where $a$, $b$, and $c$ are fixed numbers, presumably functions of the data, and $\xi$, $\eta$ are the quantities we wish to sample.
- 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 $a=c=5.25$, $b=-5.2499$. Then the level curves of the equation $pd=\text{constant}$ will be very elongated ellipses oriented at a 45° 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

\[
\begin{align*}
a &= 5.25 \\
c &= a \\
b &= -5.2499 \ # Set values required \\
\text{logpost} &= \text{function}(\text{xi,et}) \{ \\
\quad &- (a*\text{xi}^2 + 2*b*\text{xi}*\text{et} + c*\text{et}^2)/2 \\
\}
\end{align*}
\]
MCMC Approaches

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

```r
sample1 = function(xi, et, dxi, det) {
  xist = xi + runif(1, -dxi, dxi)
  etst = et + runif(1, -det, det)
  lnalph = logpost(xist, etst) - logpost(xi, et)
  accept = (log(runif(1,0,1)) < lnalph)
  xi = xist*accept + xi*(1-accept)
  et = etst*accept + et*(1-accept)
  list(xi=xi, et=et, accept=accept)
}
```

MCMC Approaches

• Here we run \( n \) samples with variable starting information

```r
sammen = function(n=3000, dxi=2, det=2, xi=0, et=0) {
  xis = rep(NaN, n)
  ets = rep(NaN, n)
  accepts = 0
  for(i in 1:n){
    z = sample1(xi, et, dxi, det)
    xis[i] = xi = z$xi
    ets[i] = yet = z$et
    accepts = accepts + z$accept
  }
  invisible(list(xi=xis, et=ets, acc=accepts))
}
```

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 \( \xi, \eta \) 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 \( a, b, c \).

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

\[
Q = \begin{bmatrix} a & b \\ b & c \end{bmatrix}
\]

• We want to find an upper triangular matrix \( R \) with the property that \((RR^t) = Q\). Then if \( z \) is a sample from a standard bivariate normal distribution, then \( u = R z \) (or \( u = z R \), 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 superscript \(^t\) means matrix transpose.
MCMC Approaches

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

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

\[ \text{sample1} = \text{function}(R) \{ \]
\[ \quad u = \text{rnorm}(2) \quad \%\% \quad R \]
\[ \quad \text{return}(u) \]
\[ \} \]

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

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

\[ \text{sample1} = \text{function}(R) \{ \]
\[ \quad u = \text{rnorm}(2) \quad \%\% \quad R \]
\[ \quad \text{return}(u) \]
\[ \} \]

MCMC Approaches

• Here we sample our posterior distribution:

\[ \text{samplen} = \text{function}(n=3000,Q) \{ \]
\[ \quad \text{xis} = \text{rep}(\text{NaN},n) \]
\[ \quad \text{ets} = \text{rep}(\text{NaN},n) \]
\[ \quad R = \text{chol}(\text{solve}(Q)) \]
\[ \quad \text{for}(i \text{ in } 1:n)\{ \]
\[ \quad \quad z = \text{sample1}(R) \]
\[ \quad \quad \text{xis}[i] = z[1] \]
\[ \quad \quad \text{ets}[i] = z[2] \]
\[ \quad \} \]
\[ \quad \text{invisible}(\text{list}(\text{xi} = \text{xis}, \text{et} = \text{ets})) \]
\[ \} \]

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 the parameters \( \alpha \) through nonlinear relations like \( y = f(x; \alpha) \) 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

\[ p(y | \alpha, x) \propto \frac{1}{\sigma^2} \exp\left[ -\frac{1}{2\sigma^2} \sum (y_i - f(x_i; \alpha_0, \alpha_1, \ldots))^2 \right] \]

to compute the transition matrix

\[ R \]
\( \text{opt} \)
\( \text{chol} \)
\( \text{solve} \)
\( \text{mask} \)

MCMC Approaches

• 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 \text{t} distribution rather than a multivariate normal distribution, e.g., something like this:

\[ \text{sample1} = \text{function}(R) \{ \]
\[ \quad u = c * \text{rt}(2, \text{dof}=4) \quad \%\% \quad R + \text{alphahat} \]
\[ \quad \ldots \quad \# \quad \text{Accept/reject step code here} \]
\[ \quad \ldots \}

where \( R \) is the matrix arising from the solution of the linearized least squares problem, \text{alphahat} is the desired mean and \text{c} a constant we choose for efficient sampling. \text{rt} samples from a \text{t} distribution.
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 \( y \sim N(\theta, \sigma^2 I) \). 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
  \[
  L \propto \frac{1}{\sigma^n} \exp\left[ -\frac{1}{2\sigma^2} (y - A\theta)'(y - A\theta) \right] = \frac{1}{\sigma^n} \exp\left[ -\frac{S}{2\sigma^2} \right]
  \]

- Define the ordinary least squares estimator of \( \theta \) by
  \[
  \hat{\theta} = (A'A)^{-1} A' y
  \]

- Then a simple calculation shows that
  \[
  S = (\hat{\theta} - \theta)'A'(A\theta - \theta) + (y - A\hat{\theta})' (y - A\hat{\theta}) = S_e(\theta) + S_e
  \]

- 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

- With a Jeffreys prior on \( \sigma \), the posterior distribution is seen to be
  \[
  p(\theta, \sigma | y) \propto \sigma^{-(n+1)} \exp\left( -\frac{1}{2\sigma^2} (S_e(\theta) + S_e) \right)
  \]

- As before, the marginalization strategy is to integrate out every \( \theta_j \) except the one we are interested in, and then to integrate out \( \sigma \). Each of the \((k-1)\) integrations with respect to \( \theta_j \) loses one power of \( \sigma \) in the posterior. After integrating out \( \sigma \) we get a \( t \) distribution with \( \nu = n-k \) degrees of freedom where
  \[
  t = \frac{\theta_j - \hat{\theta}_j}{s \sqrt{m_{kk}}}, \quad s^2 = S_e / \nu
  \]
  and \( m_{kk} \) is the \((kk)\) element of \( M=(A'A)^{-1} \)

- 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

\[
p(\theta \mid y) \propto (\nu s^2 + S_i(\theta))^{-n/2} \exp \left( -\frac{1 + k/\nu}{2s^2} S_i(\theta) \right)
\]

How do we get this?

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

\[
S_e = \frac{1}{n} \sum (y - A\hat{\theta})' (y - A\hat{\theta})
\]


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 \( S_e, S_i(\theta) \) and \( \hat{\theta} \).
• See if you can get the following posterior!

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

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

\[
A' A = \left( \sum x_i \sum x_i^2 \right)
\]
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 $(\theta - \theta)$ can be made to vanish by choosing
  $$\hat{\theta} = \left[ \frac{1}{\sigma^2} A'A + Q^{-1} \right]^{-1} \left[ A'y \sigma^2 + Q^{-1} \theta_0 \right]$$
  This is a kind of average between prior and likelihood, related to “ridge regression.”

Posterior Prediction

- Often it is necessary to make predictions of future observations, based on our best inferences on parameters determined through observations already made.
- 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 $x_{n+1}$
- We can make predictions using probability theory, e.g., suppose we observe the value of $y$ at a given $x$ as
  $$y = \alpha_0 + \alpha_1 x + \epsilon = \eta + \epsilon$$
  where $\eta$ is the true value and $\epsilon$ the observation error. Then
  $$p(d(\alpha_0, \alpha_1, \sigma | D)) \propto p(D | \alpha_0, \alpha_1, \sigma)p(\alpha_0, \alpha_1, \sigma)$$
  $$p(\eta | \alpha_0, \alpha_1, \sigma, x) = \delta(\eta - \alpha_0 - \alpha_1 x)$$
  $$p(\eta | D, x) \propto \iint \delta(\eta - \alpha_0 - \alpha_1 x) p(d\alpha_0, d\alpha_1, d\sigma)$$

- This problem can be solved exactly. For simplicity I will consider the case when $\sigma$ is known. Then using centered variables with $\beta_0 = \alpha_0 + \alpha_1 x$, $\beta_1 = \alpha_1$, $\xi = x - \bar{x}$ we get
  $$p(\beta_0 | D) \propto \exp \left[ -\frac{1}{2} \left( \frac{(\beta_0 - \hat{\beta}_0)^2}{\sigma_{\beta_0}^2} + \frac{(\beta_1 - \hat{\beta}_1)^2}{\sigma_{\beta_1}^2} \right) \right]$$
  $$p(\eta | D) \propto \iint \delta(\eta - \beta_0 - \beta_1 \xi) p(\beta_0, \beta_1 | D) d\beta_0 d\beta_1$$
  $$\propto \int \exp \left[ -\frac{1}{2} \left( \frac{(\eta - \hat{\beta}_0 - \hat{\beta}_1 \xi)^2}{\sigma_{\beta_0}^2 + \xi^2 \sigma_{\beta_1}^2} \right) \right] d\beta_1$$
  where $A, B$ are messy constant expressions. The integral just gives us another constant, so the result is of the form
  $$p(\eta | D, \xi) \propto \exp \left[ -\frac{1}{2} \left( \frac{(\eta - \hat{\beta}_0 - \hat{\beta}_1 \xi)^2}{\sigma_{\beta_0}^2 + \xi^2 \sigma_{\beta_1}^2} \right) \right]$$
  $$\eta | D, \xi \sim N(\hat{\beta}_0 + \hat{\beta}_1 \xi, \sigma_{\beta_0}^2 + \xi^2 \sigma_{\beta_1}^2)$$
  $$\sim N(\hat{\alpha}_0 + \hat{\alpha}_1 x, \sigma_{\alpha_0}^2 + (x - \bar{x})^2 \sigma_{\alpha_1}^2)$$
Posterior Prediction

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

\[
\eta = \hat{\alpha}_0 + \hat{\alpha}_1 x
\]

- This calculation doesn't include the additional error due to the observational error $\varepsilon$ in $y$. It includes only the error in our estimates of the parameters $\alpha$ and $\sigma$. It is the posterior probability of the true quantity $\eta$, not that of the future observations that we might make.

- We observe $y = \eta + \varepsilon$, not $\eta$

- We still have to think about the observational error.

- It is reasonable to presume that the observational error in the predicted observations will be independent of the error in $\eta$. So that error ($\varepsilon$) will be additive.

- More generally, suppose we have a sample from the posterior distribution $p(\theta \mid D)$, where $\theta$ are our parameters [e.g., on previous charts, $\theta = (\alpha, \sigma)$ or $(\beta, \sigma)$]. We wish to predict the distribution of new observations $y$, where if there were no error we would have $y = \eta + \varepsilon = f(\theta) + \varepsilon$; we presume that $p(y \mid \eta, \theta)$ is known.

- It is easy to use our sample to produce a sample from $p(\eta \mid D, x)$: Simply make draws from $p(\eta \mid \theta)$ for the given value of $x$ over the MCMC samples of $\theta$.

- Then we can compute expectations of $\eta | D, x$ or variances of $\eta$, or credible intervals of $\eta$, etc., from the sample.

- Then we can predict the distribution of the observations $y$ by resampling from the distribution $p(y \mid \eta, \theta)$ for each sample of $\eta$.

- Supposed we are interested in predicting $y$, where $y$ is observed with error. This can also be done easily using the results on the previous charts.

- Our MCMC sample is a collection of values of $(\beta, \alpha)$ that we've generated by our random walk.

- For each sample, we can calculate a value $\eta$ since there is a functional relationship between $(\beta, \alpha)$ and $\eta$ given by some function $\eta(\beta, \alpha, x)$ (see previous charts for details).

- We get a sample of $y$ by drawing a random variable from $N(\eta(\beta, \alpha, x), \sigma^2)$ for the particular $(\beta, \alpha)$ in our sample.

- Do this for every $(\beta, \alpha)$ in the MCMC sample.

- This is the posterior predictive model for $y$. 
Bayesian Inference

### Gibbs Sampling \( \alpha \)

- Here’s our EIV model posterior probability \((\alpha_\xi = \alpha_y = \alpha)\):

\[
p(\alpha_0, \alpha_1, \alpha, \xi \mid x, y) \propto \frac{1}{\sigma^{2n+1}} \exp\left[-\frac{1}{2\sigma^2} \sum (\xi_j - x_i)^2\right]
\]

\[
x \exp\left[-\frac{1}{2\sigma^2} \sum (\alpha_0 + \alpha, \xi_j - y_j)^2\right]
\]

- It’s simplest to use matrices. Let

\[
\begin{bmatrix}
\xi_1 \\
\xi_2 \\
\vdots \\
\xi_n \\
x_1 \\
x_2 \\
\vdots \\
x_n \\
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
= \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}
= \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
= \begin{bmatrix}
1 \\
1 \\
\vdots \\
1
\end{bmatrix}
\]

\[
\xi = \begin{bmatrix}
\xi_1 \\
\xi_2 \\
\vdots \\
\xi_n \\
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}, \quad y = \begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}, \quad \Sigma = \begin{bmatrix}
\alpha_0 \\
\alpha_1
\end{bmatrix}, \quad \alpha = \begin{bmatrix}
\alpha_0 \\
\alpha_1
\end{bmatrix}
\]

### Gibbs Sampling \( \alpha \)

- Then the quadratic in the exponential becomes:

\[
U = \frac{1}{\sigma^2} [(\xi - x)'(\xi - x) + (\Sigma \alpha - y)'(\Sigma \alpha - y)]
\]

- Define \( \alpha = \hat{\alpha} + \Delta \alpha \) where \( \hat{\alpha} \) is to be determined by completing the square. We get

\[
U = \frac{1}{\sigma^2} [(\xi - x)'(\xi - x) + (\Sigma \hat{\alpha} - y)'(\Sigma \hat{\alpha} - y)]
\]

\[
= \frac{1}{\sigma^2} [(\xi - x)'(\xi - x) + (\Sigma \hat{\alpha} - y)'(\Sigma \hat{\alpha} - y) + \Delta \alpha' (\Sigma \Sigma) \Delta \alpha + 2 \Delta \alpha' \Sigma (\Sigma \hat{\alpha} - y)]
\]

- Comment: We see that

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

- The R code to choose a sample follows. The function `cbind`, which stands for “bind columns,” is used to create the matrix \( \Sigma \) out of vectors. Do `?cbind` for more information. `rep(1, n)` creates a vector of \( n \) 1’s, and `t(X)` transposes the matrix.

\[
\begin{align*}
\text{XI} &= \text{cbind}(\text{rep}(1, n), \text{xi}) \\
W &= \text{solve}(\text{t(X)} \ %*\ % \ \text{XI}) \ # \text{t(x)=transpose} \\
R &= \text{chol}(W) * \Sigma \\
\text{alphahat} &= y \ %*\ % \ \text{XI} \ %*\ % \ W \\
\text{alpha} &= \text{rnorm}(2, 0, 1) \ %*\ % \ R + \text{alphahat}
\end{align*}
\]
Sampling on the Latent Variables

• Recall that $\xi_i$, the “true” values of the observations $x_i$, are not fixed numbers, but have a probability distribution. We will have to sample over these as well. The posterior is

$$p(\alpha_0, \alpha_1, \sigma, \xi | x, y) \propto \frac{1}{\sigma^{2n+1}} \exp \left[ -\frac{1}{2\sigma^2} \sum (\xi_i - x_i)^2 \right] \times \exp \left[ -\frac{1}{2\sigma^2} \sum (\alpha_0 + \alpha_i \xi_i - y_i)^2 \right]$$

• Complete the square. We get the result on the next page:

Sampling on the Latent Variables

• Sample as follows:

$$\xi_i \sim N(u_i, \tau^2)$$

where

$$u_i = \frac{x_i + \alpha_1 (y_i - \alpha_0)}{1 + \alpha_i^2}$$

$$\tau^2 = \frac{\sigma^2}{1 + \alpha_i^2}$$

Note that one can sample on all the $\xi_i$ at once, since they are independent of each other (although dependent on the current values of $\alpha_0$ and $\alpha_1$). This is one line in R.

• Similar and only slightly more complicated formulas result if $\alpha_i \neq \sigma_i$.