Twenty-two children are given a reading comprehension test before and after receiving a particular instruction method.

- $Y_{i,1}$: pre-instructional score for student $i$
- $Y_{i,2}$: post-instructional score for student $i$
- vector of observations for each student $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2})'$

Questions:

- Does the instructional method lead to improvements in reading comprehension (on average)?
- If so, by how much?
- Can improvements be predicted based on the first test?
Plot

Pre-instructional Score

Post-instructional Score

Pre-instructional Score
Parameters

We could model the data as bivariate normal $Y_i \overset{iid}{\sim} N_2(\mu, \Sigma)$. Normal distributions are characterized by their

- mean vector $\mu = (\mu_1, \mu_2)'$ where $\mu_1 = E[Y_{i,1}]$ and $\mu_2 = E[Y_{i,2}]$
- variance-covariance matrix

$$
\Sigma = \begin{bmatrix}
\sigma_1^2 & \sigma_{12} \\
\sigma_{21} & \sigma_2^2
\end{bmatrix}
$$

where $\sigma_1^2$ is the variance of $Y_{i,1}$, $\sigma_2^2$ is the variance of $Y_{i,2}$, and $\sigma_{12}$ is the covariance between $Y_{i,1}$ and $Y_{i,2}$, $\sigma_{21} = \sigma_{12} = E[(Y_{i,1} - \mu_1)(Y_{i,2} - \mu_2)]$

- Correlation between pre and post test scores is $\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}$, measure of the strength of association. $-1 \leq \rho < 1$
Multivariate Normal

For more than 2 dimensions, we write \( Y \sim N_d(\mu, \Sigma) \)

- \( \mathbb{E}[Y] = \mu \): \( d \) dimensional vector with means \( \mathbb{E}[Y_j] \)
- \( \text{Cov}[Y] = \Sigma \): \( d \times d \) matrix with diagonal elements that are the variances of \( Y_j \) and off diagonal elements that are the covariances \( \mathbb{E}[(Y_j - \mu_j)(Y_k - \mu_k)] \)
- if \( \Sigma \) is positive definite \( x'\Sigma x > 0 \) for any \( x \neq 0 \) in \( \mathbb{R}^d \) then there is a density

\[
p(Y) = (2\pi)^{-d/2} |\Sigma|^{-d/2} \exp\left(-\frac{1}{2}(Y - \mu)^T \Sigma^{-1}(Y - \mu)\right)
\]

or

\[
p(Y) = (2\pi)^{-d/2} |\Phi|^{d/2} \exp\left(-\frac{1}{2}(Y - \mu)^T \Phi(Y - \mu)\right)
\]

where \( \Phi \) is the precision matrix, \( \Phi = \Sigma^{-1} \).
Bayesian Inference

Need to specify prior distributions, then use Bayes Theorem to obtain posterior distributions.

- Normal \( \mathcal{N}_d(\mu_0, P_0^{-1}) \) is the semi-conjugate prior for the mean (conjugate given the covariance matrix \( \Sigma \)).

- Given the mean vector, the semi-conjugate prior for \( \Phi \) is the Wishart distribution, a generalization of the gamma distribution to higher dimensions. If \( \Phi \sim \text{Wishart}_d(\nu_0, \Phi_0) \) then it has density

\[
p(\Phi) = \frac{1}{C} |\Phi|^{(\nu_0-d-1)/2} \exp\left(-\frac{1}{2} \text{tr}\Phi S_0\right)
\]

where

\[
C = 2^{\nu d/2} \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma(\nu/2 + (1 - j)/2) |\Phi_0|^{-\nu/2}
\]

- If \( \Phi \sim \text{Wishart}_d(\nu_0, \Phi_0) \) then

\[
\text{E}[\Phi] = \nu_0 \Phi_0 \\
\text{E}[\Sigma] = \frac{1}{\nu_0-d-1} \Phi_0^{-1}
\]

Choose \( \nu_0 = d + 2 \) so that \( \text{E}[\Sigma] = \Phi_0^{-1} \).
Priors Distributions

- Prior for $\mu$:
  \[
  \mu \sim N_2 \left( \begin{pmatrix} 50 \\ 50 \end{pmatrix}, \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix} \right)
  \]

- Test designed to have a mean of 50

- True mean constrained to be between 0 and 100; $\mu \pm 2\sigma = (0, 100)$ implies $\sigma^2 = (50/2)^2 = 625$.

- Prior correlation is 0.50
Priors Continued

- Prior for the precision matrix

\[ \Phi \sim \text{Wishart}_2 \left( 4, \begin{pmatrix} 625 & 312.5 \\ 312.5 & 625 \end{pmatrix}^{-1} \right) \]

- loosely centered at the same covariance as in the normal mean prior distribution.

- \( \nu_0 \) is a prior degrees of freedom \( \nu_0 = d + 2 \) leads to

\[ E[\Sigma] = \Phi_0^{-1} \equiv S_0 \]

For the full conditional distributions, we just need to find the updated hyperparameters (see Hoff)
Full Conditionals

\[ \mu \mid \Phi, Y \sim N_d((P_0 + n\Phi)^{-1}(n\Phi\bar{y} + P_0 m_0), (P_0 + n\Phi)^{-1}) \]

\[ \Phi \mid \mu, Y \sim \text{Wishart}_d(n + \nu_0, S_0 + \sum(y_i - \mu)(y_i - \mu)) \]

Limiting Case: \( m_0 = 0, P_0 = 0, S_0 = 0, \nu_0 = \text{Jeffrey’s prior} \)
multmodel=function() {
  for( i in 1 : N ) {
    Y[i,1:2] ~ dmnorm(mu[], Phi[,] )
  }
  mu[1:2] ~ dmnorm(mu0[], prec[,] )
  Phi[1:2, 1:2] ~ dwish(Phi0[,] , nu0 )
  Sigma[1:2,1:2] <- inverse(Phi[,])
  rho <- Sigma[1,2]/sqrt(Sigma[1,1]*Sigma[2,2])
  Ynew[1:2] ~ dmnorm(mu[], Phi[,] )
}
Joint Distribution of $\mu$
Predictive Distribution of a new student $Y$
Rao-Blackwellization

Monte Carlo estimates are noisy! Can do better by using the idea of “Rao-Blackwellization”

- $E_{MC}[g(\theta_1) \mid Y] = \frac{1}{T} \sum_t g(\theta_1^{(t)})$

- Iterated Expectations:

$$E[g(\theta_1) \mid Y] = E_{\theta_2}E_{\theta_1 \mid \theta_2}[g(\theta_1) \mid Y, \theta_2] = E_{\theta_2}[\tilde{g}(\theta_2) \mid Y]$$

$$\approx \frac{1}{T} \sum_t \tilde{g}(\theta_2^{(t)}) \equiv E_{RB}[g(\theta_1) \mid Y]$$

- Calculate inner conditional expectation analytically
- May reduce variance over MCMC average (Liu et al 1996)
- Motivated by the classical Rao Blackwell Theorem that says that taking any estimator and conditioning on a sufficient statistic will reduce its mean squared error. (proof in Stat 215)
Density estimation

Rather than use a histogram estimate of the density, use a Rao-Blackwell estimate. We know the posterior density of $\mu$ given $\Phi$ and $Y$, $p(\mu \mid \Sigma, Y)$

$$p(\mu \mid Y, \Phi) = \frac{1}{2\pi} \left| P_n \right|^{1/2} \exp \left\{ -\frac{1}{2} (\mu - \mu_n)' P_n (\mu - \mu_n) \right\}$$

where

- prior precision $P_0 = S_0^{-1}$
- posterior precision $P_n = P_0 + n\Phi$
- posterior mean $\mu_n = P_n^{-1}(P_0\mu_0 + \Phi nY)$

Use MC average over draws of $\Phi^{(1)}, \ldots, \Phi^{(T)}$ to integrate out $\Phi$ to obtain the marginal

$$\hat{p}(\mu \mid Y) = \frac{1}{T} \sum_t p(\mu \mid Y, \Phi^{(t)})$$
Alternative Priors on $\Sigma$

No reason to restrict attention to semi-conjugate prior distributions, as long as the prior on $\Sigma$ lead to positive definite matrices.

See Dongchu Sun and Jim Berger’s paper for an interesting discussion of priors for the Multivariate normal model.


May require Metropolis-Hastings algorithms as full conditionals distributions are not recognizable. Automatic in WinBugs.