Multivariate Normal

Yingbo Li

Clemson University

MATH 9810

Reading Comprehension Example

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

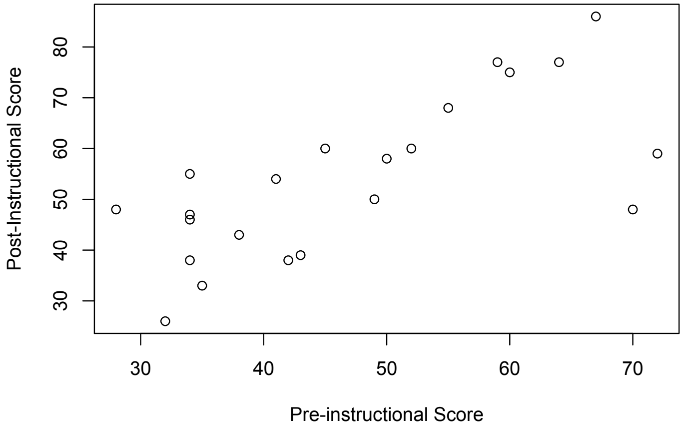
- $Y_{i,1}$: pre-instructional score for student i
- $Y_{i,2}$: post-instructional scorefor student i
- vector of observations for each student $\mathbf{Y}_i = (Y_{i,1}, Y_{i,2})'$

Questions:

- Do students improve in reading comprehesion on average?
- If so, by how much?
- Can we predict post-test score from pre-test score?

NOTE: CANNOT CLAIM THAT METHOD CAUSED ANY CHANGES BECAUSE NO CONTROL GROUP.

Scatter Plot

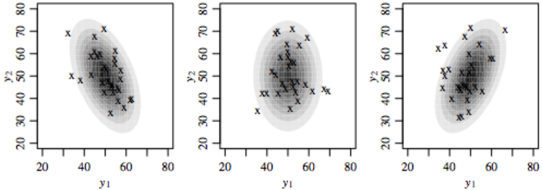


Bivariate Normal Model

Model the data as bivariate normal, $\mathbf{Y}_i \sim \mathrm{N}_2(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- $\boldsymbol{\mu} = (\mu_1, \mu_2)'$ where $\mu_1 = E(Y_1)$ and $\mu_2 = E(Y_2)$
- Covariance matrix:
$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}$$
where $\sigma_1^2 = Var(Y_1)$, $\sigma_2^2 = Var(Y_2)$, and σ_{12} is the covariance between Y_1 and Y_2, i.e., $\sigma_{21} = \sigma_{12} = E[(Y_1 - \mu_1)(Y_2 - \mu_2)]$
- Correlation between Y_1 and Y_2 is $\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$

Correlations: -0.5 (left) , 0 (middle), 0.5 (right)



General Form of Multivariate Normal

For $p \geq 2$ dimensions, we write $\mathbf{Y}_i \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- $\boldsymbol{\mu} = E(\mathbf{Y})$ is p dimensional vector of $E(Y_j)$ for $j = 1, \ldots, p$
- $\boldsymbol{\Sigma} = Cov(\mathbf{Y}) = E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})']$ is a $p \times p$ matrix with diagonal elements equal to the variances of Y_j and off-diagonal elements equal to the covariances $E((Y_j - \mu_j)(Y_k - \mu_k))$
- $\boldsymbol{\Sigma}$ has to be a positive definite matrix, i.e., for any $\mathbf{x} \neq 0$ in $\mathbb{R}^p$, $\mathbf{x}'\boldsymbol{\Sigma}\mathbf{x} > 0$.

PDF of Multivariate Normal $\mathbf{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- Using Σ, the pdf is

$$p(\mathbf{Y}) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp(-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\boldsymbol{\mu})).$$

- Or, letting $\boldsymbol{\Phi} = \boldsymbol{\Sigma}^{-1}$ be the precision matrix, we have

$$p(\mathbf{Y}) = (2\pi)^{-p/2}|\boldsymbol{\Phi}|^{1/2}\exp(-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\boldsymbol{\Phi}(\mathbf{Y}-\boldsymbol{\mu}))$$

Suppose A is a $q \times p$ matrix, then

$$\mathbf{Y} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \Longrightarrow \mathbf{AY} \sim \mathsf{N}_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$$

Semi-conjugate Prior Distribution

Need prior distributions on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

- Semi-conjugate prior for $\boldsymbol{\mu}$ is $\mathsf{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$, where $\boldsymbol{\Lambda}_0$ is a $p \times p$ positive definite matrix.
- Semi-conjugate prior for $\boldsymbol{\Phi}$ is the Wishart distribution, a generalization of the Gamma distribution to higher dimensions.
- Semi-conjugate specification requires MCMC simulation (Gibbs sampler) for posterior inference

Wishart Distribution

- Suppose $\mathbf{X}_i$ is a $p \times 1$ vector of random variables such that $\mathbf{X}_i \sim \mathsf{N}_p(\mathbf{0}, \mathbf{\Lambda})$. Let
$$\mathbf{X} = [\mathbf{X}_1', \ldots, \mathbf{X}_n']'$$
be n iid samples from this distribution, i.e., the i-th row of $\mathbf{X}$ is $\mathbf{X}_i'$.
- Then the sum of squares
$$\mathbf{S} = \mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{X}_i\mathbf{X}_i' \sim \mathsf{Wishart}_p(n, \mathbf{\Lambda}),$$
with pdf
$$p(\mathbf{S}) = \frac{1}{C}|\mathbf{S}|^{\frac{n-p-1}{2}} \exp(-\frac{1}{2}\mathsf{tr}(\mathbf{S}\mathbf{\Lambda}^{-1}))$$
where $C = 2^{np/2}\pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma(n/2 + (1-j)/2)|\mathbf{\Lambda}|^{n/2}$
- n is the degrees of freedom, and $\mathbf{\Lambda}$ is the scale parameter.
- If $\mathbf{S} \sim \mathsf{Wishart}_p(n, \mathbf{\Lambda})$, then $E(\mathbf{S}) = n\mathbf{\Lambda}$
- When $p = 1$, this is just a Gamma distribution

Inverse Wishart

- Let $\mathbf{Z}$ be a random variable such that $\mathbf{Z}^{-1} \sim \text{Wishart}_p(n, \mathbf{\Psi}^{-1})$. Then, $\mathbf{Z} \sim$ inverse $\text{Wishart}_p(n, \mathbf{\Psi}^{-1})$
- If $\mathbf{Z} \sim$ inverse $\text{Wishart}_p(n, \mathbf{\Psi}^{-1})$ then
  - $E(\mathbf{Z}^{-1}) = n\mathbf{\Psi}^{-1}$
  - $E(\mathbf{Z}) = \frac{1}{n-p-1}\mathbf{\Psi}$
- Some people write distribution slightly differently: they use $\mathbf{\Psi}$ instead of $\mathbf{\Psi}^{-1}$. As long as you correctly interpret scale parameter, you get same answers.

Back to Priors for $\mathbf{\Phi}$ and $\mathbf{\Sigma}$

- Assume $\mathbf{\Phi} \sim \text{Wishart}_p(\nu_0, \mathbf{S}_0^{-1})$ or equivalently $\mathbf{\Sigma} \sim \text{inverse Wishart}_p(\nu_0, \mathbf{S}_0^{-1})$, so that
  - $E[\mathbf{\Phi}] = \nu_0 \mathbf{S}_0^{-1}$
  - $E[\mathbf{\Sigma}] = \frac{1}{\nu_0 - p - 1} \mathbf{S}_0$
- Setting hyperparameters for priors
  - Determine your prior best guess (i.e., prior mean) of $\mathbf{\Sigma}$, say $\mathbf{\Sigma}_0$. Set $\mathbf{S}_0 = (\nu_0 - p - 1)\mathbf{\Sigma}_0$.
  - Choose $\nu_0 = p + 2$ for vague prior beliefs that $\mathbf{\Sigma}$ is centered around $\mathbf{S}_0$
  - Choose ν_0 large for stronger prior beliefs that $\mathbf{\Sigma} \approx \mathbf{S}_0$
  - Note that $\nu_0 > p + 1$ for proper prior

Default Prior Distribution

- An improper, default prior distribution is the Jeffrey's prior,

$$p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2}$$

- This results in posterior distributions that are computed without the need for MCMC. Let sample cross-products matrix be $\mathbf{S} = \sum_{i=1}^{n} (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})'$. Then,

$$\begin{aligned} \boldsymbol{\Sigma} \mid \mathbf{Y} &\sim \text{inverse Wishart}_p(n-1, \mathbf{S}^{-1}) \\ \boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathbf{Y} &\sim \mathrm{N}_p(\bar{\mathbf{Y}}, \boldsymbol{\Sigma}/n) \end{aligned}$$

- In general, be careful with noninformative priors in high dimensions, because they can easily lead to improper posterior distributions.

Prior Distribution for $\boldsymbol{\mu} \sim \mathrm{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0)$

- Both tests designed to have a mean of 50. Bivariate normal prior makes sense, since we expect values of $\boldsymbol{\mu}$ to be decreasingly plausible as we move away from 50.
- Set $\boldsymbol{\mu}_0 = (50, 50)'$.
- Suppose we a priori believe that true means are not likely to be less than 25 or more than 75. To reflect this belief, we set $\boldsymbol{\Lambda}_0$ so that there is small chance of being outside that range. For any one test, $50 \pm 2\lambda_0 = (25, 75)$ implies $\lambda_0^2 = (25/2)^2 \approx 156$.
- Suppose we think the tests are measuring similar concepts, so we make a reasonably strong prior correlation between average test scores of $.50$.

As a result, we have a prior distribution for $\boldsymbol{\mu}$:

$$\boldsymbol{\mu} \sim N_2\left(\begin{pmatrix} 50 \\ 50 \end{pmatrix}, \begin{pmatrix} 156 & 78 \\ 78 & 156 \end{pmatrix}\right)$$

Priors for Φ (or Σ)

- Because individual scores constrained to [0, 100], we set prior for Σ such that values outside that range are unlikely. For any test $50 \pm 2\sigma_0 = (0, 100)$ implies $\sigma_0^2 = (50/2)^2 = 625$.
- Assume prior correlation between individual's test scores of $.50$.
- Vague beliefs about $\mathbf{\Sigma}$: set $\nu_0 = p + 2 = 4$
- Hence, we have $\mathbf{\Phi} \sim \text{Wishart}_2(4, \mathbf{S}_0^{-1})$,

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

- Equivalently, $p(\mathbf{\Sigma})$ is inverse-Wishart$_2$ using $\nu_0 = 4$ and the same $\mathbf{S}_0^{-1}$

Full Conditionals

For Gibbs sampler, we need full conditional distributions

$$\boldsymbol{\mu} \mid \boldsymbol{\Phi}, \mathbf{Y} \sim \mathsf{N}_p((\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Phi})^{-1}(n\boldsymbol{\Phi}\bar{\mathbf{Y}} + \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0), (\boldsymbol{\Lambda}_0^{-1} + n\boldsymbol{\Phi})^{-1})$$

$$\boldsymbol{\Phi} \mid \boldsymbol{\mu}, \mathbf{Y} \sim \mathsf{Wishart}_p(n + \nu_0, (\mathbf{S}_0 + \sum_i (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})')^{-1})$$

Equivalently,

$$\boldsymbol{\Sigma} \mid \boldsymbol{\mu}, \mathbf{Y} \sim \mathsf{inverse\text{-}Wishart}_p(n + \nu_0, (\mathbf{S}_0 + \sum_i (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})')^{-1})$$

We'll do the Gibbs sampler and posterior inference in class.

Answering Questions of Interest

Questions:

- Do students improve in reading comprehesion on average?

  Want $Pr(\mu_2 > \mu_1 \mid \mathbf{Y})$.

- If so, by how much?

  Want posterior inference for $\mu_2 - \mu_1$.

- Can we predict post-test score from pre-test score?

  Best approach is regression (Ch. 9). Related is inference for σ_{12}, e.g., $Pr(\sigma_{12} > 0 \mid \mathbf{Y})$