BIOS 755: Covariance Pattern Analysis and the General Linear Model

Alexander McLain

Treatment of Lead-Exposed Chidren (TLC) Trial

- The methods we'll discuss today will add a covariance matrix to standard linear regression.
- For a covariance matrix to be meaningful it is easiest if the data have well-defined time-points.
- As a result, we'll reflect on the well worn TLC example.
- Part of the reason we'll do this is because the methods we will discuss are most useful for data that are (at least) planned to have **balanced** data.
- Randomized trial, 100 children randomized to placebo or Succimer, measures of blood lead level at baseline, 1, 4 and 6 weeks

General Linear Model

- For each observation, Y_{ij}, assume we have an associated set of covariates

$$\boldsymbol{X}_{ij} = \{1, X_{ij1}, X_{ij2}, \ldots, X_{ijp}\}$$

- Information about the time of the observations, treatment group, age, biomarkers, and other predictor variables can be expressed through a vector of covariates.
- The one represents the intercept.
- How we structure $\boldsymbol{X}$ will be discussed in later lectures.

General Linear Model

- The general linear model can be written as

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 X_{i11} + \beta_2 X_{i12} + \ldots + \beta_p X_{i1p} + e_{i1} \\ Y_{i2} &= \beta_0 + \beta_1 X_{i21} + \beta_2 X_{i22} + \ldots + \beta_p X_{i2p} + e_{i2} \\ \vdots &= \vdots \\ Y_{in} &= \beta_0 + \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + \ldots + \beta_p X_{in_ip} + e_{in} \end{aligned}$$

- We can summarize this to

$$Y_{ij} = \beta_0 + \sum_{k=1}^{p} X_{ijk}\beta_k + e_{ij} \quad \text{for} \quad j = 1, 2, \ldots, n$$

General Linear Model

- When we remove the error term e_{ij} we get the predicted values

$$E(Y_{ij}) = \hat{Y}_{ij} = \mu_{ij} = \beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j$$

- The difference between the predicted and observed values are the residuals or error

$$Y_{ij} - \hat{Y}_{ij} = e_{ij}$$

General Linear Model

- With longitudinal data, we expect the error terms, e_{ij}, to be correlated within individuals.
- For example, if an individual has a large positive error term in the first observation, i.e.,
$$Y_{i1} - \hat{Y}_{ij} = e_{i1} > 0 \quad \text{is large}$$
then what would you expect the error term of the second observation to be?
- So in longitudinal data we want to allow for
$$corr(e_{ij}, e_{ik}) \neq 0$$
for all j and k.

Covariance Matix

- This leads to a covariance matrix for $\boldsymbol{e}_i$

$$Cov(\boldsymbol{e}_i) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{pmatrix} = \Sigma$$

where $\text{cov}(e_{ij}, e_{ik}) = E(Y_j - \mu_j)(Y_k - \mu_k) = \sigma_{jk}$ with $\sigma_{jj} = \sigma_j^2$.

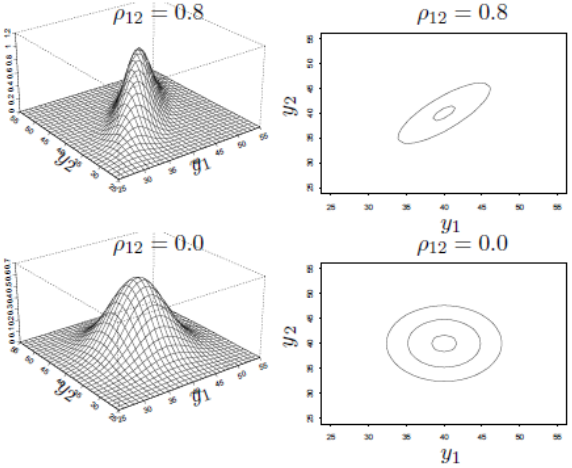
General Linear Model

- Yet *another* to write the general linear model is

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{e}_i$$

  where $\boldsymbol{X}_i = \{\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{in_i}\}$ and $\boldsymbol{e}_i \sim MVN(\mathbf{0}, \Sigma)$.
- Recall that Σ is a covariance matrix of the residual error terms.

Multivariate Normal Distribution



Covariance Structure

When choosing a covariance structure the important aspects to consider are:

- Balanced or unbalanced **time points.**
  - **For unbalanced time points our options are limited.**
- Homogeneity or heterogeneity (i.e., is the residual variance equal or not equal over time)?
- Are there simple forms we can use to represent the correlation?

Covariance Structure

- The most flexible is the unrestricted or unstructured covariance matrix (**heterogenous** and **no assumptions on correlation**):

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix} \quad \Gamma = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{21} & 1 & \rho_{23} & \rho_{24} \\ \rho_{31} & \rho_{32} & 1 & \rho_{34} \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{pmatrix}$$

- The most restrictive covariance matrix is the Independence matrix (**homogeneous** and **no correlation allowed**):

$$\Sigma = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix} \quad \Gamma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Compound Symmetry, Exchangeable

- Two popular covariance models with this correlation matrix are the:
  - compound symmetric, and
  - heterogeneous compound symmetric structure
- The difference in these structures is whether or not we assume the variance is homogeneous or heterogeneous across time points.

(Homogeneous) Compound Symmetric

- The compound symmetric structure:

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \dots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \dots & \rho\sigma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & \dots & \sigma^2 \end{pmatrix} \quad \Gamma = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

also called an exchangeable structure.

Heterogeneous Compound Symmetric

- The heterogeneous compound symmetric structure:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \dots & \rho\sigma_1\sigma_k \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho\sigma_2\sigma_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho\sigma_1\sigma_k & \rho\sigma_2\sigma_k & \dots & \sigma_k^2 \end{pmatrix} \quad \Gamma = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{pmatrix}$$

- What is error? Would it change?

Autoregressive Structure of Order 1 (AR(1))

- Autoregressive correlation matrix:

$$\mathbf{\Gamma} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

- Since ρ is less than one, as we take higher powers of it the results gets closer and closer to zero.
- As observations get further away (in terms of number of observations) there correlation gets smaller.

Visualize the Correlation Structures

$$R = \begin{array}{c|cccc} & \text{Visit 1} & \text{Visit 2} & \text{Visit 3} & \text{Visit 4} \\ \hline \text{Visit 1} & 1 & & & \\ \text{Visit 2} & 0.7 & 1 & & \\ \text{Visit 3} & 0.7^2 & 0.7 & 1 & \\ \text{Visit 4} & 0.7^3 & 0.7^2 & 0.7 & 1 \end{array}$$

Exponential Structure

- The Exponential Structure is one that uses the time between points in calculating the correlation.
- The correlation between two points Y_{ij} and Y_{ik} is equal to

$$\rho_{jk} = \exp\left\{-\frac{|t_{ij} - t_{ik}|}{\theta}\right\}$$

  recall that t_{ij} is the time of observation Y_{ij}, similarly for t_{ik}.
- The parameter θ is estimated, the larger the value of θ the smaller the correlation.

Fitting in SAS

- SAS can be used to fit many, many covariance structures.
- Click here for a full list of covariance matrices.

Covariance Structure

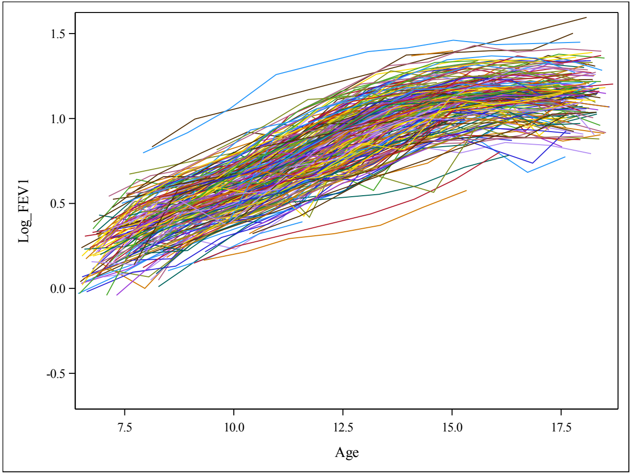
- How important is it to take account of the correlation among repeated measures?
- We can address that question by analyzing the TLC data under the assumption of independence and comparing the results to those analyzed with an unstructured covariance matrix.

GO TO EXAMPLE

Balanced or Unbalanced time points

- One big factor in choosing a covariance matrix is if the time points are balanced or not.
- In the unstructured correlation matrix $\rho_{jk} = Corr(Y_{ij}, Y_{ik})$ for all i.
- Does this make sense if t_{ij} and t_{ik} are different for all i?

Air pollution example



Unbalanced covariance structures

- Of the covariance matrices we've discussed, only the homogeneous compound symmetric and homogeneous exponential make sense for unbalanced time points.

Unbalanced covariance structures

- Of the covariance matrices we've discussed, only the homogeneous compound symmetric and homogeneous exponential make sense for unbalanced time points.
- The exponential can also be used to model the covariate over space.
- For example, say d_{1k} and d_{2k} are the latitude and longitude of the kth measurement.
- The correlation between two points Y_j and Y_k is equal to

$$\rho_{jk} = \exp\left\{-\frac{\sqrt{(d_{1j} - d_{1k})^2 + (d_{2j} - d_{2k})^2}}{\theta}\right\}$$

which is a spatial covariance matrix.