Linear Models for Correlated Data: Examples

> We aim to develop a general linear model framework for longitudinal data, in which the inference we make about the parameters of interest recognize the likely correlation structure of the data.

There are two ways of achieving this:

1. to build explicit parametric models of the covariance structure
2. to use methods of inference which are robust to misspecification of the covariance structure

General Linear Model for Correlated Data: Examples

- Uniform correlation Model

  One sample repeated measures ANOVA
- Growth Model
- Exponential Correlation Model

  Autoregressive Model of order 1

A simple example

- $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, Y_{i3}),\ i = 1, \ldots, m,\ n_i = 3$
- $V_i = \sigma^2 \begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
- Y_{i1}, Y_{i2} highly correlated
- (Y_{i1}, Y_{i2}) independent of Y_{i3}
- same variance, correlation for all subjects

  $V_i = V,\ i = 1, \ldots, m$

Notation: balanced data

- $y_{ij},\ j = 1, \ldots, n,\ i = 1, \ldots, m$
- t_j corresponding times at which the measurements are taken on each unit
- $x_{ijk},\ k = 1, \ldots, p$ of explanatory variables

$$y_{ij} = \beta_1 x_{ij1} + \quad \beta_p x_{ijp} + \epsilon_{ij}$$

- If we have longitudinal data

$$cov(\epsilon_{ij}, \epsilon_{ik}) = v_{ijk} \neq 0$$

- v_{ijk} is the (j, k) element of the covariance matrix V_i

Notation: unbalanced data

- $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $\dim(\boldsymbol{Y}_i) = n_i \times 1$
- $\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{i1} \\ \vdots \\ \boldsymbol{x}_{in_i} \end{pmatrix}$, $\dim(\boldsymbol{X}_i) = n_i \times p$
- $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})$
- $\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i,\ \boldsymbol{\epsilon}_i \sim MVN(0, V_i)$
- $\text{var}(\boldsymbol{Y}_i) = V_i,\ \dim(V_i) = n_i \times n_i$

Notation

- $N = \sum_{i=1}^m n_i$ (total number of obs.)
- $\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_m \end{pmatrix}$, $\dim(\boldsymbol{Y}) = N \times 1$
- $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_m \end{pmatrix}$, $\dim(\boldsymbol{X}) = N \times p$
- $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$, $\dim(\boldsymbol{\beta}) = p \times 1$

Notation

- $V = \begin{pmatrix} var(\boldsymbol{Y}_1) & \\ Cov(\boldsymbol{Y}_j, \boldsymbol{Y}_k) & Var(\boldsymbol{Y}_m) \end{pmatrix}$
- $var(\boldsymbol{Y}_i) = V_i,\ i = 1, \ldots, m,\ Cov(\boldsymbol{Y}_j, \boldsymbol{Y}_k) = \boldsymbol{0}$
- $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim MVN(0, V)$
- $V = \begin{pmatrix} V_1 & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & V_2 & \ldots & \boldsymbol{0} \\ & & \ddots & \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & V_m \end{pmatrix}$ block diagonal
- if $n_i = n$ then $V_i = V_0$ for $i = 1, \ldots, m$

Covariates may be:

1. Cluster specific, time invariant, so that
$$x_{i1k} = x_{i2k} = \ldots = x_{in_ik}$$
for some $1 \le k \le p$. Examples include sex and race in a longitudinal study, and fixed experimental conditions in a longitudinal clinical trials.
2. Subject specific, time varying, or within-subject, i.e. covariate k varies with j so that $x_{ijk} \neq x_{ij'k}$. Examples include time since baseline, experimental condition in crossover or repeated measures designs, smoking status or height in a longitudinal study, or individual characteristics in a clustered sample survey. In some cases (pure repeated measures designs, or longitudinal studies with fixed time points) $x_{ijk} = x_{i'jk}$ for all j.
3. Fixed by design, e.g. treatment group indicator, time since baseline, or individual characteristics in a sample survey.
4. Stochastic, e.g. height, current smoking status, or pollution exposure in a longitudinal survey.

General Linear model with correlated errors
Balanced Data

- $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m),\ N = m \times n$
- X with row $(x_{ij1}, \ldots, x_{ijp})$
- V block diagonal matrix with non-zero $n \times n$ block V_0, each representing the variance matrix for the vector of measurement on a single subject
- we consider the following model

$$\boldsymbol{Y} \sim MVN(X\boldsymbol{\beta}, V)$$

- two special cases

1. the uniform correlation model
2. the exponential correlation model

Uniform Correlation Model:
Parametric form of the covariance matrix

- When measurements are equally spaced, and the data are balanced, one assumption is that the correlation between any pair of measurements is always the same

$$V_0 = \sigma^2 \begin{pmatrix} 1 & \rho & \rho & \ldots & \rho \\ \rho & 1 & \rho & \ldots & \rho \\ \rho & \rho & 1 & \ldots & \rho \\ & & \vdots & & \\ \rho & \rho & \rho & \ldots & 1 \end{pmatrix} \qquad \textit{compound symmetry}$$

- This implies

$$\begin{aligned} \text{var}(Y_{ij}) &= \sigma^2 \\ \text{corr}(Y_{ij}, Y_{i,j+k}) &= \rho, \text{ for every } j, k \end{aligned}$$

Example: Weight of Pigs

- Body weights of 48 pigs in 9 successive weeks of follow-up (Table 3.1 DLZ)
- The response is measures at n different times, or under n different conditions. In the guinea pigs example the time of measurement is referred to as a "within-units" factor. For the pigs $n = 6$
- Although the pigs example considers a single treatment factor, it is straightforward to extend the situation to one where the groups are formed as the results of a factorial design (for example, if the pigs were separated into males and female and then allocated to the diet groups)

Example: Weight of Pigs

- For this type of repeated measures study we recognize two sources of random variation

1. **Between:** There is heterogeneity between pigs, due for example to natural biological (genetic?) variation
2. **Within:** There is random variation in the measurement process for a particular unit at any given time. For example, on any given day a particular guinea pig may yield different weight measurements due to differences in scale (equipment) and/or small fluctuations in weight during a day.

Marginal Model

- **model for the mean**

$$E[\boldsymbol{Y}_i] = \boldsymbol{\mu}$$

- **model for the covariance**

$$\text{cov}[\boldsymbol{Y}_i] = \sigma^2 \left[(1-\rho)I + \rho \mathbf{1}\mathbf{1}^{\mathbf{T}}\right]$$

- σ^2 is the total variance (within + between)
- ρ is the correlation between any two measurements

Random Effect Model

- **model for the mean given** U_i

$$\begin{aligned} Y_{ij} &= \mu_j + U_i + Z_{ij} \\ E[\boldsymbol{Y}_i \mid U_i] &= \boldsymbol{\mu} + U_i\mathbf{1} \end{aligned}$$

- $\text{cov}[\boldsymbol{Y}_i \mid U_i] = \tau^2 I$
- $U_i \sim N(0, v^2)$ U_i is a random effect
- v^2 is the heterogeneity variance (between)
- Z_{ij} is the measurement error
- τ^2 is the measurement error variance (within)
- $\sigma^2 = v^2 + \tau^2$ total variance
- $\rho = v^2/(v^2+\tau^2)$

These two model specifications are equivalent

$$\begin{aligned}
\boldsymbol{Y}_i &= \boldsymbol{\mu} + U_i\mathbf{1} + \boldsymbol{Z}_{\mathbf{i}} \\
E[\boldsymbol{Y}_i] &= E[E[\boldsymbol{Y}_i \mid U_i]] = E[\boldsymbol{\mu} + U_i\mathbf{1}] = \boldsymbol{\mu} \\
\text{cov}(\boldsymbol{Y}_i) &= \text{cov}[E(\boldsymbol{Y}_i \mid U_i)] + E[\text{cov}(\boldsymbol{Y}_i \mid U_i)] \\
&= \text{cov}[U_i\mathbf{1}] + \boldsymbol{\tau^2 I_n} \\
\text{cov}(\boldsymbol{Y}_i) &= v^2\mathbf{1}\mathbf{1}^{\mathbf{T}} + \boldsymbol{\tau^2 I_n} \\
&= \begin{pmatrix} v^2+\tau^2 & v^2 & \ldots & v^2 \\ v^2 & v^2+\tau^2 & \ldots & v^2 \\ & & \vdots & \\ v^2 & v^2 & \ldots & v^2+\tau^2 \end{pmatrix} \\
&= (v^2+\tau^2)\left[\frac{v^2}{(v^2+\tau^2)}\mathbf{1}\mathbf{1}^{\mathbf{T}} + \frac{\boldsymbol{\tau^2}}{\boldsymbol{\tau^2+v^2}}\boldsymbol{I_n}\right] \\
&= \sigma^2\left[\rho\mathbf{1}\mathbf{1}^{\mathbf{T}} + (\mathbf{1}-\rho)\boldsymbol{I_n}\right]
\end{aligned}$$

Uniform Correlation Model

- Let ρ be correlation between any two measurements on the same subject, then $V_0 = \sigma^2 \left[(1-\rho)I + \rho\mathbf{1}\mathbf{1}^{\mathbf{T}}\right]$
- Let's consider this model:

$$Y_{ij} = \mu_j + U_i + Z_{ij}$$

- $U_i \sim N(0, v^2),\ Z_{ij} \sim N(0, \tau^2)$
- U_i and Z_{ij} are independent
- $\rho = v^2/\sigma^2$ and $\sigma^2 = v^2 + \tau^2$
- *the uniform correlation model can be interpreted as a linear model for the mean response with a random intercept U_i with a variance v^2 between subjects.*

One sample repeated measures ANOVA

- N subjects are measures repeatedly under n different experimental conditions. The goals is to quantify differences in experimental conditions.

$$E(\boldsymbol{Y}_i) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mu_1 + U_i \\ \mu_2 + U_i \\ \vdots \\ \mu_n + U_i \end{pmatrix}$$

- Here $\boldsymbol{X}_i = \boldsymbol{I}_n$, $\boldsymbol{\beta} = \boldsymbol{\mu}$ and $p = n$.

$$Y_{ij} = \mu_j + U_i + Z_{ij}$$

where the U_is and the Z_{ij}s are independent of each other, with var$(U_i) = v^2$, var$(Z_{ij}) = \tau^2$, var$(Y_{ij}) = \tau^2 + v^2$, and $\rho = v^2/(v^2 + \tau^2)$

One group polynomial growth curve model

- N subjects are observed at the same times $t_1,\ t_2,\ ,\dots,\ t_n$; for example a group of children from the same cohort is observed yearly at ages $6,\ 7,\ ,8,\dots,12$. A linear model of the average response can be expressed as a polynomial in t_j

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_j \\ E(\boldsymbol{Y}_i) &= \begin{pmatrix} 1 & t_1 \\ 1 & t_2 \\ & \vdots \\ 1 & t_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \end{aligned}$$

- **Pigs data**
- $E[Y_{ij}] = U_i + \beta_0 + \beta_1 t_j$ (random intercept + linear trend)

One group polynomial growth curve model

- Similarly, if you want to fit a quadratic curve

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_j + \beta_2 t_j^2 \\ E(\boldsymbol{Y}_i) &= \begin{pmatrix} 1 & t_1 & t_1^2 \\ 1 & t_2 & t_2^2 \\ & \dots & \\ 1 & t_n & t_n^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \end{aligned}$$

In General

One model for V_i arises from assuming that each subject has his or her own growth curve with parameter $\boldsymbol{\beta}_i$:

$$\begin{aligned} E(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i) &= \boldsymbol{X}\boldsymbol{\beta}_i \\ \text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i) &= \sigma^2 \boldsymbol{I}_n \end{aligned}$$

and

$$\begin{aligned} E(\boldsymbol{\beta}_i) &= \boldsymbol{\beta} \\ \text{cov}(\boldsymbol{\beta}_i) &= D \end{aligned}$$

then

$$\begin{aligned} E(\boldsymbol{Y}_i) &= \boldsymbol{X}\boldsymbol{\beta} \\ \text{cov}(\boldsymbol{Y}_i) &= \text{cov}[E(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i)] + E[\text{cov}(\boldsymbol{Y}_i \mid \boldsymbol{\beta}_i)] \\ &= \underbrace{\boldsymbol{X}D\boldsymbol{X}^t}_{\text{between}} + \underbrace{\tau^2 \boldsymbol{I}_n}_{\text{within}} \end{aligned}$$

Exponential Correlation/Autoregressive Model

- The correlation between a pair of measurements on the same unit decays towards zero as the time separation between the measurement increases, i.e.
  $Cov(Y_{ij}, Y_{ik}) = [V_0]_{jk} = v_{jk} = \sigma^2 e^{(-\phi|t_j - t_k|)}$
  (**unequally-spaced data**)
- If $t_{j-1} - t_j = d$ for all j, then $v_{jk} = \sigma^2\rho$, where $\rho = \exp(-\phi d)$ is the correlation between successive observations on the same subject.
  (**equally-spaced data**)
- Let $Y_{ij} = \mu_{ij} + W_{ij}$ where
  $W_{ij} = \rho W_{ij-1} + Z_{ij}$
  $Z_{ij} \sim N(0, \sigma^2(1-\rho^2))$
  $var(Y_{ij}) = var(W_{ij}) = \sigma^2$

Autoregressive

- For data that arise over time it is often reasonable to assume that correlation between measurements that are close in time is greater than the correlation of measurements that are widely separated in time, One model is (here for unit spaced observations):

$$V_0 = \sigma^2 \begin{pmatrix} 1 & \rho^1 & \rho^2 & \dots & \rho^{n-1} \\ \rho^1 & 1 & \rho^1 & \dots & \rho^{n-1} \\ \rho^2 & \rho^1 & 1 & \dots & \rho^{n-3} \\ & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \dots 1 & \end{pmatrix}$$

- One construction is given by

$$\begin{aligned} \mathsf{var}(Y_{i1}) &= \sigma^2 \\ \mathsf{E}(Y_{ij} \mid Y_{ij-1}) &= \rho Y_{ij-1} \end{aligned}$$

Marginal Model

- **model for the mean**

$$E[\boldsymbol{Y}_i] = \boldsymbol{\mu}$$

- **model for the covariance**

$$\mathsf{cov}[\boldsymbol{Y}_i] = \text{AR-1 structure}$$
$$\mathsf{cov}(Y_{ij}, Y_{ik}) = \sigma^2\rho^{|j-k|}$$

Important Points

- Modelling the correlation in longitudinal data is important to be able to obtain correct inferences on regression coefficients β
- These are **marginal models** because the interpretation of the regression coefficients is the same as that in cross-sectional data
- Correlation can be formulated in terms of **subject-specific** models and/or **transition** models
  - Exchangeable correlation model: subject-specific formulation
  - Exponential correlation model: transition model formulation
- Three basic elements of correlation structure:
  - random effects
  - autocorrelation or serial dependence
  - noise, measurement error
- Incorporating correlation into estimation of regression models is achieved via **weighted least squares**. We will see how this works

- There are **many** ways of estimating correlation parameters. We will study some of thes
- Correlation models can be **approximate**
  - We will call these **working correlation models** ("our best shot")
  - Regression coefficients estimates will still be correct
  - We will see how to "fix up" standard errors to account for inaccuracies in correlation models