General Linear Models for Longitudinal Data

Outline

- Review of multivariate normal distribution
- Correlation specifications
- Review of likelihood inference
- Ordinary least squares (OLS)
- Weighted least squares (WLS)
- Maximum likelihood (ML)
- Restricted maximum likelihood (REML)
- Robust estimation

Review of Multivariate Normal Distribution

The density function for a multivariate normal random vector $\boldsymbol{Y}$ is:

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(\sqrt{2\pi})^n |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \right\},$$

where $-\infty < y_j < \infty$, $j = 1, \ldots, n$.

- This distribution is completely specified by its first two moments, $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y})$ and $\Sigma = \mathrm{Var}(\boldsymbol{Y})$.
- Each Y_j has a marginal univariate normal distribution with mean μ_j and variance σ_{jj}^2.
- If we partition Σ as:

  $$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

  where Σ_{11} is a $n_1 \times n_1$ matrix, Σ_{22} is a $n_2 \times n_2$ matrix and $n_1 + n_2 = n$, then a subset of the Y_j's $\boldsymbol{Z}_1 = (Y_1, \ldots, Y_{n_1})$ also has a multivariate normal distribution with mean $\boldsymbol{\mu}_1 = (\mu_1, \ldots, \mu_{n_1})$ and variance Σ_{11}.
- Let $\boldsymbol{Z}_2 = (Y_{n_1+1}, \ldots, Y_n)$, then the conditional distribution of $\boldsymbol{Z}_1$ given $\boldsymbol{Z}_2$ is also normal with mean

  $$\boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{z}_2 - \boldsymbol{\mu}_2),$$

  and variance

  $$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

- If B is a $m \times n$ matrix, then $B\boldsymbol{Y}$ (a linear transformation) is also multivariate normal, with mean $B\boldsymbol{\mu}$ and variance $B\Sigma B^T$.
- Consider the MLE of the coefficients for a linear regression model,

$$\hat{\beta} = (X^T X)^{-1} X^T \boldsymbol{Y}.$$

If $\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I)$, then $\hat{\beta}$ also has a multivariate normal distribution with mean

$$\mathrm{E}(\hat{\boldsymbol{\beta}}) = (X^T X)^{-1} X^T X \boldsymbol{\beta} = \boldsymbol{\beta},$$

and variance:

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= (X^T X)^{-1} X^T \sigma^2 I \left[(X^T X)^{-1} X^T\right]^T \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}. \end{aligned}$$

- The random variable

$$U \equiv (\boldsymbol{Y} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{Y} - \boldsymbol{\mu}) \sim \chi^2_n.$$

- Correlations and independence
  - For a multivariate normal vector, being uncorrelated implies independent.
  - It is not true for two marginally normally distributed variables because their joint distribution may fail to be normal.

General Linear Models for Longitudinal Data

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 in 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 mis-specification of the covariance structure.

For the moment, we assume the observation times are common for all subjects, that is $t_{ij} = t_j$, $j = 1, \dots, n$ for all $i = 1, \dots, m$.

The general linear model assumes:

- All subjects are independent, that is, if X_i is stochastic, $(\boldsymbol{Y}_i, \mathsf{X}_i)$ are independent; if X_i is fixed by design, $\boldsymbol{Y}_i$ are independent.
- Given X_i,

$$\mathrm{E}(\boldsymbol{Y}_i \,|\, \mathsf{X}_i) = \mathsf{X}_i\boldsymbol{\beta} \tag{1}$$

$$\mathrm{Var}(\boldsymbol{Y}_i \,|\, \mathsf{X}_i) = \Sigma_i = \sigma^2 \mathsf{V}_i. \tag{2}$$

- We also assume $\boldsymbol{Y}_i \sim \mathcal{N}(\mathsf{X}_i\boldsymbol{\beta}, \Sigma_i)$ when the maximum likelihood method is used.

- This model implies:

  $$\mathrm{E}(Y_{ij} \,|\, \boldsymbol{X}_{i1}, \boldsymbol{X}_{i2}, \ldots, \boldsymbol{X}_{in}) = \mathrm{E}(Y_{ij} \,|\, \boldsymbol{X}_{ij}).$$

  When there are time-dependent covariates, this model may not be valid. For example, if Y_{ij} is a symptom measure, and X_{ij} is a drug treatment then past symptoms may influence current treatment. If

  $$f(X_{i,j+1} \,|\, Y_{ij}, X_{ij}) \neq f(X_{i,j+1} \,|\, X_{ij}),$$

  then

  $$f(Y_{ij} \,|\, X_{ij}, X_{ij+1}) \neq f(Y_{ij} \,|\, X_{ij}).$$

- The covariance Σ_i allows for dependence among measurement on the same subject. Covariance may vary with covariates, e.g., across treatment group, or the covariance may be a function of time.
- For the moment we will assume that the data is *balanced* (common set of t_j's) and *complete* (no missing data). The covariance of the response variable $\boldsymbol{Y}$ (rank $m \times n = N$) is:

  $$\mathrm{Cov}(\boldsymbol{Y}) = \Sigma = \begin{pmatrix} \Sigma_1 & 0 & \cdots & 0 \\ 0 & \Sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Sigma_m \end{pmatrix}.$$

Objective

Inference about $\boldsymbol{\beta}$ while accounting for the covariance ⇒ need know or estimate Σ.

Correlation Specifications

Exchangeable Correlation

Assume $\mathrm{Cor}(\epsilon_{ij}, \epsilon_{ik}) = \rho$ for $j \neq k$ (the same for all pairs of observations),

$$\boldsymbol{V}_0 = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{J} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \ldots & \rho & 1 \end{pmatrix},$$

where $\boldsymbol{V}_0$ is the correlation matrix and $\Sigma_i = \sigma^2 \boldsymbol{V}_0$, $\boldsymbol{I}$ is a $n \times n$ identity matrix; and $\boldsymbol{J}$ is $n \times n$ matrix of 1's.

This is called the **uniform**, **exchangeable**, or **compond symmetry** correlation model. It is equivalent to assume a **random effect** shared by repeated measures of the same subject:

$$Y_{ij} = X_{ij}\beta + U_i + Z_{ij}.$$

- random effect: U_i are mutually independent $N(0, \nu^2)$ r.v.
- measurement error: Z_{ij} are mutually independent $\sim N(0, \tau^2)$ r.v.
- U_i and Z_{ij} are independent.

Then, the correlation structure has $\rho = \nu^2/(\nu^2 + \tau^2)$ and $\sigma^2 = \nu^2 + \tau^2$.

Example: One-Sample Repeated Measures ANOVA

N subjects are measured repeatedly under n different experimental conditions. The goal is to quantify differences in experimental conditions. We can write the model as:

$$Y_{ij} = \mu_j + \alpha_i + \epsilon_{ij} \tag{3}$$

where μ_j, $j = 1, \ldots, n$ are the "treatment" effect (fixed), and α_i are the "subject" effect (random) and it is often assumed that $\text{Var}(\alpha_i) = \nu^2$, $\text{Var}(\epsilon_{ij}) = \tau^2$ and α_i, ϵ_{ij} are independent.

Then

$$\begin{aligned} \text{Cov}(Y_{ij}, Y_{ik}) &= \text{Cov}(\mu_j + \alpha_i + \epsilon_{ij}, \mu_k + \alpha_i + \epsilon_{ik}) \\ &= \text{Cov}(\alpha_i, \alpha_i) \\ &= \nu^2, \end{aligned}$$

$$\text{Var}(Y_{ij}) = \text{Var}(\mu_j + \alpha_i + \epsilon_{ij}) = \nu^2 + \tau^2.$$

Hence

$$\rho = \text{Cor}(Y_{ij}, Y_{ik}) = \frac{\nu^2}{\nu^2 + \tau^2}.$$

- ν^2 is the heterogeneity variance (between subjects).
- τ^2 is within subject variation.
- $\sigma^2 = \nu^2 + \tau^2$ total variance.

Model (3) can be written as:

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} \alpha_i \\ \alpha_i \\ \vdots \\ \alpha_i \end{pmatrix} + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in} \end{pmatrix}.$$

Using vector notation, it becomes:

$$\boldsymbol{Y}_i = \boldsymbol{\mu} + \alpha_i \mathbf{1} + \boldsymbol{\epsilon}_i, \tag{4}$$

where α_i and $\boldsymbol{\epsilon}_i$ are independent,

$$\begin{aligned} \alpha_i &\sim \mathcal{N}(0, \nu^2), \\ \boldsymbol{\epsilon}_i &\sim \mathcal{N}(\mathbf{0}, \tau^2 \mathsf{I}). \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(\boldsymbol{Y}_i) &= \boldsymbol{\mu} \\ \mathrm{Var}(\boldsymbol{Y}_i) &= \mathrm{Var}(\alpha_i \mathbf{1}) + \mathrm{Var}(\boldsymbol{\epsilon}_i) \\ &= \nu^2 \mathbf{1}\mathbf{1}^T + \tau^2 \mathsf{I}. \end{aligned}$$

Note:

$$\begin{aligned} \mathrm{Var}(\boldsymbol{b}Y) &= \boldsymbol{b}\boldsymbol{b}^T \,\mathrm{Var}(Y), \\ \mathrm{Var}(\mathsf{B}\boldsymbol{Y}) &= \mathsf{B}\,\mathrm{Var}(\boldsymbol{Y})\mathsf{B}^T. \end{aligned}$$

Exponential Correlation

A different model assumes the correlation of observations closer together in time is larger than that of observations farther apart.

$$\rho_{jk} = \exp(-\phi|t_j - t_k|), \quad \phi > 0. \tag{5}$$

If t_j are **equally spaced** and $t_{j+1} - t_j = d$, thus $|t_j - t_k| = d|j-k|$,

$$\rho_{jk} = \rho^{|j-k|},$$

where $\rho = \exp(-\phi d)$. This is equivalent to an **autoregressive model**:

$$\begin{aligned} Y_{ij} &= x_{ij}\beta + W_{ij} \\ W_{ij} &= \rho W_{ij-1} + \eta_{ij}, \quad |\rho| < 1 \\ \eta_{ij} &\sim N(0, \sigma^2(1-\rho^2)). \end{aligned} \tag{6}$$

Note: W_{ij} is called a discrete-time first-order autoregressive or AR(1) process.

A natural generalization of (6) is to replace W_{ij} with $W_i(t_j)$, where $W_i(t_j), j = 1, \ldots, n$ are realizations of independent, continuous-time, stationary Gaussian process $\{W_i(t), t \in R\}$ with a covariance structure, $\gamma(u) = \text{Cov}\{W_i(t), W_i(t-u)\}$.

Gaussian Correlation

In exponential correlation, the log correlation is linear in the distance. Alternatively, we can model faster decay of correlation using squared distance:

$$\rho_{jk} = \exp\{-\phi(t_j - t_k)^2\},$$

where $\phi > 0$.

- For continuous time, $\rho(u) = \exp\{-\phi u^2\}$.
- Difference between the Gaussian and exponential correlation functions.

Review of Likelihood Inference

If $\boldsymbol{Y}$ has probability density function $f(\boldsymbol{y};\boldsymbol{\theta})$ then the *likelihood function* for $\boldsymbol{\theta}$ is:

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) = f(\boldsymbol{y};\boldsymbol{\theta}).$$

- The log-likelihood is
$$\ell(\boldsymbol{\theta} \mid \boldsymbol{y}) = \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y}).$$
- Note that the likelihood is only *defined* up to a multiplicative constant, that is, the log-likelihood is defined up to an additive constant.

Maximum Likelihood Estimation

- The *maximum likelihood estimator*, $\hat{\boldsymbol{\theta}}$, maximizes the likelihood (log-likelihood):
$$\mathcal{L}(\hat{\boldsymbol{\theta}}) = \sup_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}).$$
- Under certain regularity conditions, the MLE is
  - asymptotically unbiased;
  - strongly consistent (converges almost surely to the true parameter);
  - asymptotically efficient, i.e., has the smallest asymptotic variance (Cramér-Rao Lower Bound) among asymptotically unbiased estimators.

Score Equation and Information

- The maximization is often achieved by solving the *score equation*:

  $$S(\boldsymbol{\theta}) \equiv \dot{\ell}(\boldsymbol{\theta}) \equiv \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}} = 0.$$

  Note that at the MLE,

  $$\ddot{\ell}(\hat{\boldsymbol{\theta}}) \equiv \left.\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = \left.\frac{\partial^2 \log \mathcal{L}}{\partial \boldsymbol{\theta}^2}\right|_{\hat{\boldsymbol{\theta}}} < 0.$$

- The function $\dot{\ell}(\boldsymbol{\theta})$ is called the *score function*, and

  $$\mathrm{E}\,\dot{\ell}(\boldsymbol{\theta}) = 0 \tag{7}$$

  $$\boldsymbol{I}(\boldsymbol{\theta}) \equiv \mathrm{Var}\,\dot{\ell}(\boldsymbol{\theta}) = \mathrm{E}(\dot{\ell}(\boldsymbol{\theta})\dot{\ell}(\boldsymbol{\theta})^T) = -\,\mathrm{E}\,\ddot{\ell}(\boldsymbol{\theta}). \tag{8}$$

  $\boldsymbol{I}(\boldsymbol{\theta})$ is called the *Fisher information* or *expected information* for $\boldsymbol{\theta}$ and $-\ddot{\ell}(\boldsymbol{\theta})$ is known as the *observed information.*

- The asymptotic variances of the MLE $\hat{\boldsymbol{\theta}}$ is given by: $\boldsymbol{I}(\hat{\boldsymbol{\theta}})^{-1}$.

Hypothesis Testing

For testing: $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, we can use:

- The likelihood ratio test statistic:
$$G \equiv 2 \left\{ \ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}_0) \right\}. \tag{9}$$
- The Wald test statistic:
$$W \equiv (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \boldsymbol{I}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \tag{10}$$
  - $\boldsymbol{I}(\boldsymbol{\theta}_0)$ is sometimes replaced by $\boldsymbol{I}(\hat{\boldsymbol{\theta}})$ or the observed information $-\ddot{\ell}(\hat{\boldsymbol{\theta}})$.
  - When testing the $H_0 : \beta = \beta_0$ for regression models, t-test or F-test is often used for better small sample performance.
- The score or Rao test statistic:
$$R \equiv \dot{\ell}(\boldsymbol{\theta}_0)^T \boldsymbol{I}(\boldsymbol{\theta}_0)^{-1} \dot{\ell}(\boldsymbol{\theta}_0). \tag{11}$$
  - Note that it is not necessary to find the MLE $\hat{\boldsymbol{\theta}}$ under the alternative distribution when using the score statistic.

- When the sample size is large, all three statistics have an asymptotic χ^2 distribution with p degrees of freedom, where p is the dimension of the parameter $\boldsymbol{\theta}$ (or the difference in the number of parameters of the two models, one nested in the other).

- Computationally, the Wald statistic is often the easiest one to compute, while the likelihood ratio statistic is the hardest.

- The small sample performance of the three statistics, however, do differ.

OLS for General Linear Model

Consider the general linear model specified in (1) and (2), the ordinary least squares estimator is $\hat{\boldsymbol{\beta}}_{\text{OLS}} = (X^T X)^{-1} X^T \boldsymbol{Y}$, which minimizes:

$$(\boldsymbol{Y} - X\boldsymbol{\beta})^T(\boldsymbol{Y} - X\boldsymbol{\beta}).$$

The OLS estimator is still unbiased:

$$\begin{aligned} \mathrm{E}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= (X^T X)^{-1} X^T \,\mathrm{E}(\boldsymbol{Y}) \\ &= (X^T X)^{-1} X^T X \boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

However the variance of $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is not the usual formula for the variance of OLS estimator:

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}_{\text{OLS}}) &= \mathrm{Var}\left\{(X^T X)^{-1} X^T \boldsymbol{Y}\right\} \\ &= (X^T X)^{-1} X^T \,\mathrm{Var}(\boldsymbol{Y}) \left\{(X^T X)^{-1} X^T\right\}^T \\ &= \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1} \\ &\neq \sigma^2 (X^T X)^{-1} \end{aligned} \tag{12}$$

Therefore, even though $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is unbiased, inference based on the naive OLS estimate of the variance of $\hat{\boldsymbol{\beta}}$ is wrong.

Weighted Least Squares

In univariate regression, WLS yields estimates of $\boldsymbol{\beta}$ that minimize the objective function:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{N} w_i (Y_i - \boldsymbol{X}_i \boldsymbol{\beta})^2. \tag{13}$$

Analogously, the multivariate version of WLS finds the value of the parameter $\boldsymbol{\beta}(\mathsf{W})$ that minimizes:

$$\begin{aligned} Q_W(\boldsymbol{\beta}) &= (\boldsymbol{Y} - \mathsf{X}\boldsymbol{\beta})^T \mathsf{W} (\boldsymbol{Y} - \mathsf{X}\boldsymbol{\beta}) \qquad (14) \\ &= \sum_{i=1}^{m} (\boldsymbol{Y}_i - \mathsf{X}_i\boldsymbol{\beta})^T \mathsf{W}_i (\boldsymbol{Y}_i - \mathsf{X}_i\boldsymbol{\beta}) \end{aligned}$$

where W is a symmetric weight matrix. Note that the second equality holds if W is a block-diagonal matrix with non-zero blocks W_i, a $(n_i \times n_i)$ matrix. We assume W has this form for the rest of this lecture.

It is straightforward to see that:

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} Q_W(\boldsymbol{\beta}) &= -2\mathsf{X}^T \mathsf{W} (\boldsymbol{Y} - \mathsf{X}\boldsymbol{\beta}) \\ &= -2 \sum_{i=1}^{m} \mathsf{X}_i^T \mathsf{W}_i (\boldsymbol{Y}_i - \mathsf{X}_i\boldsymbol{\beta}). \end{aligned}$$

The solution to the minimization solves $\boldsymbol{U}(\boldsymbol{\beta}) = 0$ and yields,

$$\begin{aligned}\hat{\boldsymbol{\beta}}(W) &= \left(X^T W X\right)^{-1} X^T W \boldsymbol{Y} \\ &= \left(\sum_{i=1}^{m} X_i^T W_i X_i\right)^{-1} \left(\sum_{i=1}^{m} X_i^T W_i \boldsymbol{Y}_i\right)\end{aligned} \qquad (15)$$

- The OLS estimator corresponds to $W_i^{-1} = \sigma^2 I_i$, assuming observations are independent both within and between subjects.
- If $X_i = X_1$ and $W_i = W_1$ for all i, (e.g., complete and balanced design), then

$$\hat{\boldsymbol{\beta}}(W) = \left(X_1^T W_1 X_1\right)^{-1} X_1^T W_1 \left(\frac{1}{m} \sum_i \boldsymbol{Y}_i\right).$$

This implies that $\hat{\boldsymbol{\beta}}$ is the regression of the **averages**.

Properties of $\hat{\boldsymbol{\beta}}(W)$

- $\hat{\boldsymbol{\beta}}(W)$ is unbiased, for any W:

$$\begin{aligned}\mathrm{E}[\hat{\boldsymbol{\beta}}(W)] &= \left(X^T W X\right)^{-1} X^T W\, \mathrm{E}(\boldsymbol{Y}) \\ &= \left(X^T W X\right)^{-1} X^T W X \boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

- Variance

$$\begin{aligned}\mathrm{Var}[\hat{\boldsymbol{\beta}}(W)] &= A^{-1}\,\mathrm{Var}(X^T W \boldsymbol{Y}) A^{-1} \\ &= A^{-1}\left(X^T W\,\mathrm{Var}(\boldsymbol{Y}) W X\right) A^{-1} \\ &= A^{-1}\left(X^T W \Sigma W X\right) A^{-1}\end{aligned}$$

where

$$A^{-1} = \left(X^T W X\right)^{-1}.$$

- If $W = I$ (OLS)

$$\mathrm{Var}[\hat{\boldsymbol{\beta}}(I)] = \left(X^T X\right)^{-1}\left(X^T \Sigma X\right)\left(X^T X\right)^{-1}. \qquad (16)$$

- If $\mathsf{W} = \Sigma^{-1}$,
$$\mathrm{Var}[\hat{\boldsymbol{\beta}}(\Sigma^{-1})] = \left(\mathsf{X}^T\Sigma^{-1}\mathsf{X}\right)^{-1}.$$

- It can be shown that
$$\mathrm{Var}\,\hat{\boldsymbol{\beta}}(\Sigma^{-1}) \leq \mathrm{Var}\,\hat{\boldsymbol{\beta}}(\mathsf{W}). \tag{17}$$

Conclusion: any choice of W (including I) yields unbiased estimator for $\boldsymbol{\beta}$ but using $\mathsf{W} = \Sigma^{-1}$ is the most efficient.

The **relative efficiency** of estimator $\hat{\boldsymbol{\beta}}_k(\mathsf{W})$ is measured by:
$$\text{Relative Efficiency} = \frac{\mathrm{Var}\,\hat{\boldsymbol{\beta}}_k(\Sigma^{-1})}{\mathrm{Var}\,\hat{\boldsymbol{\beta}}_k(\mathsf{W})}.$$

- As noted in DHLZ (p. 60), the relative efficiency of OLS estimator is often quite good, sometimes, even fully efficient (relative efficiency = 1).
- In any case, the correct variance of the OLS estimator (16) should be used, instead of the naive OLS variance (12).

Maximum Likelihood Estimation under Normal Assumption

Assuming model $\boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma^2 V\right)$, where V is a block-diagonal matrix with common non-zero blocks V_0, the log-likelihood function is:

$$\ell(\boldsymbol{\beta}, \sigma^2, V_0) \quad = \quad -\frac{nm}{2}\log(\sigma^2) \quad - \quad \frac{m}{2}\log(|V_0|) \quad + \quad \sigma^{-2}\left\{-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\} \tag{18}$$

If V_0 is known, the score functions with respect to $\boldsymbol{\beta}$ and σ^2 are:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, \sigma^2, V_0) = \sigma^{-2}X^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta}) \tag{19}$$

$$\frac{\partial}{\partial\sigma^2}\ell(\boldsymbol{\beta}, \sigma^2, V_0) = -\frac{nm}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left\{(\boldsymbol{Y} - X\boldsymbol{\beta})^T V^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\} \tag{20}$$

Setting (19) to 0, we get the MLE:

$$\hat{\boldsymbol{\beta}}(V_0) = \left(X^T V^{-1} X\right)^{-1} X^T V^{-1}\boldsymbol{y}, \tag{21}$$

which is just the most efficient WLS estimator $\hat{\boldsymbol{\beta}}(\Sigma^{-1})$.
Note that the absence of a scaling factor does not affect the estimate, implying that we only need use a weight matrix $W \propto \Sigma^{-1}$.

Estimation of the Variance Components

Let

$$\mathrm{RSS}(V_0) = (\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}(V_0))^T V^{-1}(\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}(V_0)). \tag{22}$$

Substitute $\hat{\boldsymbol{\beta}}(V_0)$ into (20), set it to 0 and solve for σ^2, we get:

$$\hat{\sigma}^2(V_0) = \frac{\mathrm{RSS}(V_0)}{nm}. \tag{23}$$

If the correlation matrix V_0 is parameterized by $\boldsymbol{\alpha}$, then substitute (21) and (23) into (18), we get

$$\ell(V_0(\boldsymbol{\alpha})) = -\frac{nm}{2}\log\{\mathrm{RSS}(V_0(\boldsymbol{\alpha}))\} - \frac{m}{2}\log(|V_0(\boldsymbol{\alpha})|). \tag{24}$$

- By maximizing (24) we can get the MLE for $\boldsymbol{\alpha}$ and then the MLEs for $\boldsymbol{\beta}$ and σ^2 follow.
- The maximization with regard to $\boldsymbol{\alpha}$ in general does not have a close form and requires numerical optimization techniques.
- The ML estimators for σ^2 and $\boldsymbol{\alpha}$ are biased. The bias is substantial when the dimension of $\boldsymbol{\beta}$ is large.
- If the design matrix X is not correctly specified, this simultaneous estimation of $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\alpha}$ will not work because $\hat{\sigma}^2$ and $\hat{\boldsymbol{\alpha}}$ may not even be consistent.
- A sensible strategy is to use an over-elaborate or *saturated* model for X, get a **consistent** estimator for the variance structure and then refit a more economical design matrix. Problems of this approach ...

Restricted Maximum Likelihood (REML)

Why REML

Consider the simple example where $\boldsymbol{y} = (y_1, \ldots, y_n)$ are i.i.d. random samples from $\mathcal{N}\left(\mu, \sigma^2\right)$. The MLE for μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i}^{n} y_i = \bar{y}. \tag{25}$$

The MLE for σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i}^{n} y_i^2 - \bar{y}^2. \tag{26}$$

However, it is biased:

$$\begin{aligned}
\mathrm{E}(\hat{\sigma}^2) &= \mathrm{E}\left(\frac{1}{n}\sum_{i}^{n} y_i^2 - \bar{y}^2\right) \\
&= (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{n} + \mu^2\right) \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

One can simply "adjust" for the bias and get the unbiased estimator

$$\frac{1}{n-1}\sum_{i}^{n}(y_i - \bar{y})^2. \tag{27}$$

In the case of the GLM with independent errors, $\boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma^2 I\right)$, the MLE for σ^2 is $\hat{\sigma}^2 = RSS/(nm)$, and the unbiased estimator is $RSS/(nm - p)$.

In the case of GLM with dependent errors,$\boldsymbol{Y} \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma^2 V\right)$, a more general approach is needed to correct the bias of the ML estimators and that is REML.

The Restricted Likelihood

Let's reasbosrb the variance component parameters $\boldsymbol{\alpha}$ and σ^2 back into Σ. The log-likelihood function for the general linear model can be written as:

$$\ell(\boldsymbol{\beta}, \Sigma) = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\left\{(\boldsymbol{Y} - X\boldsymbol{\beta})^T \Sigma^{-1} (\boldsymbol{Y} - X\boldsymbol{\beta})\right\}. \tag{28}$$

For fixed Σ, (28) is maximized over $\boldsymbol{\beta}$ by:

$$\tilde{\boldsymbol{\beta}} = \left(X^T\Sigma^{-1}X\right)^{-1} X^T\Sigma^{-1}\boldsymbol{Y} = G\boldsymbol{Y}. \tag{29}$$

Substitute (29) into (28), we obtain the **profile log-likelihood** for Σ:

$$\begin{aligned}\ell_p(\Sigma) &= -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\left\{\left(\boldsymbol{Y} - X\tilde{\boldsymbol{\beta}}\right)^T \Sigma^{-1} \left(\boldsymbol{Y} - X\tilde{\boldsymbol{\beta}}\right)\right\} \\ &= -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\left\{\boldsymbol{Y}^T\Sigma^{-1}\boldsymbol{Y} - (X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(X\tilde{\boldsymbol{\beta}})\right\}\end{aligned}$$

$$
\begin{aligned}
(X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(X\tilde{\boldsymbol{\beta}}) &= \left[X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\boldsymbol{Y}\right]^T\Sigma^{-1} \\
&\quad \left[X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\boldsymbol{Y}\right] \\
&= \boldsymbol{Y}^T\Sigma^{-1}X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1} \\
&\quad X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\boldsymbol{Y} \\
&= \boldsymbol{Y}^T\Sigma^{-1}X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\boldsymbol{Y}
\end{aligned}
$$

Therefore

$$
\ell_p(\Sigma) = -\frac{1}{2}\log(|\Sigma|) - \frac{1}{2}\left\{\boldsymbol{Y}^T\Sigma^{-1}\left[I - X\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\right]\boldsymbol{Y}\right\} \tag{30}
$$

REML estimators of Σ maximizes the **restricted log-likelihood**:

$$\ell_R(\Sigma) = \ell_p(\Sigma) - \frac{1}{2}\log\left|X^T\Sigma^{-1}X\right|. \tag{31}$$

- What is $-\frac{1}{2}\log|X^T\Sigma^{-1}X|$?

- Need numerical maximization of (31) to obtain the estimate of Σ.
- REML takes into account the loss of degrees of freedom for estimating $\boldsymbol{\beta}$ and is less biased, for small sample sizes (relative to p).
- Once Σ is estimated, it is plugged back to (29) to get the "REML estimate" of $\boldsymbol{\beta}$ even though strictly speaking, REML only refers to the variance components.

Derivation of REML Method

For general linear model, $\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \Sigma)$, the **REML estimator** is defined as a maximum likelihood estimator based on a linear transformed set of data $\boldsymbol{Y}^* = A\boldsymbol{Y}$ such that the distribution of $\boldsymbol{Y}^*$ does not depend on $\boldsymbol{\beta}$.

- The resulted estimators for σ^2 and $\boldsymbol{\alpha}$ does not depend on the choice of A.
- The transformation needs not be explicit.

The main idea behind REML is to separate that part of the data used for estimation of Σ_i from that used for estimation of $\boldsymbol{\beta}$.

Let

$$\mathsf{A} = \mathsf{I} - \mathsf{X}(\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T, \tag{32}$$

and B be a $nm \times (nm - p)$ matrix defined by

$$\mathsf{B}\mathsf{B}^T = \mathsf{A}_{nm\times nm} \text{ and } \mathsf{B}^T\mathsf{B} = \mathsf{I}_{(nm-p)\times(nm-p)}, \tag{33}$$

and $\boldsymbol{Z} = \mathsf{B}^T\boldsymbol{Y}$ (an $nm - p$ vector), then

$$\begin{aligned} \mathrm{E}(\boldsymbol{Z}) &= \mathsf{B}^T\,\mathrm{E}(\boldsymbol{Y}) = \mathsf{B}^T\mathsf{X}\boldsymbol{\beta} \\ &= \mathsf{B}^T\mathsf{B}\mathsf{B}^T\mathsf{X}\boldsymbol{\beta} = \mathsf{B}^T\mathsf{A}\mathsf{X}\boldsymbol{\beta}. \end{aligned}$$

Since

$$\begin{aligned} \mathsf{A}\mathsf{X} &= \left\{\mathsf{I} - \mathsf{X}(\mathsf{X}^T\mathsf{X})^{-1}\mathsf{X}^T\right\}\mathsf{X} \\ &= \mathsf{X} - \mathsf{X} = 0, \end{aligned}$$

we have

$$\mathrm{E}(\boldsymbol{Z}) = \mathbf{0}.$$

In addition, the covariance of $\boldsymbol{Z}$ with $\tilde{\boldsymbol{\beta}}$ from (29), which we know is unbiased for $\boldsymbol{\beta}$, is:

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Z}, \tilde{\boldsymbol{\beta}}) &= \mathrm{E}\left\{\boldsymbol{Z}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\right\} \\
&= \mathrm{E}\left\{B^T\boldsymbol{Y}(\boldsymbol{Y}^T G^T - \boldsymbol{\beta}^T)\right\} \\
&= B^T\,\mathrm{E}(\boldsymbol{Y}\boldsymbol{Y}^T)G^T - B^T\,\mathrm{E}(\boldsymbol{Y})\boldsymbol{\beta}^T \\
&= B^T\left\{\mathrm{Var}(\boldsymbol{Y}) + \mathrm{E}(\boldsymbol{Y})\,\mathrm{E}(\boldsymbol{Y})^T\right\}G^T - B^T\,\mathrm{E}(\boldsymbol{Y})\boldsymbol{\beta}^T
\end{aligned}$$

Substituting in $\mathrm{Var}(\boldsymbol{Y}) = \Sigma$ and $\mathrm{E}(\boldsymbol{Y}) = X\boldsymbol{\beta}$ gives

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{Z}, \tilde{\boldsymbol{\beta}}) &= B^T\Sigma G^T + B^T X\boldsymbol{\beta}\boldsymbol{\beta}^T X^T G^T - B^T X\boldsymbol{\beta}\boldsymbol{\beta}^T \\
&= B^T\Sigma\left\{\left(X^T\Sigma^{-1}X\right)^{-1}X^T\Sigma^{-1}\right\}^T \\
&= B^T\Sigma\Sigma^{-1}X\left(X^T\Sigma^{-1}X\right)^{-1} = 0.
\end{aligned}$$

Therefore $\boldsymbol{Z}$ and $\tilde{\boldsymbol{\beta}}$ are independent, because they have a joint multivariate normal distribution with zero covariance.

We now show that the distribution of $\boldsymbol{Z}$ does not depend on $\boldsymbol{\beta}$ or the choice of B. Note that $\binom{\boldsymbol{Z}}{\tilde{\boldsymbol{\beta}}} = \binom{B^T}{G}\boldsymbol{Y}$ is a linear transformation of $\boldsymbol{Y}$ with density:

$$f(\boldsymbol{z}, \tilde{\boldsymbol{\beta}}) = \frac{1}{|J|} f(\boldsymbol{y}) = f(\boldsymbol{z}) g(\tilde{\boldsymbol{\beta}}),$$

where J is the Jacobian. Therefore

$$f(\boldsymbol{z}) = \frac{1}{|J|}\frac{f(\boldsymbol{y})}{g(\tilde{\boldsymbol{\beta}})}.$$

To obtain the explicit form of $f(\boldsymbol{z})$ we need the following result:

$$\begin{aligned}
&(\boldsymbol{y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\\
&= \left\{\boldsymbol{y} - X\tilde{\boldsymbol{\beta}} + X(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\}^T \Sigma^{-1} \left\{\boldsymbol{y} - X\tilde{\boldsymbol{\beta}} + X(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\}\\
&= (\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}) + 0 + 0\\
&\qquad + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Sigma^{-1}X(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\\
&= (\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}) + (\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T X^T\Sigma^{-1}X(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})
\end{aligned}$$

The second equality is due to the fact,

$$(X^T\Sigma^{-1}X)\tilde{\boldsymbol{\beta}} = X^T\Sigma^{-1}\boldsymbol{y}.$$

The density function of $\boldsymbol{y}$ is:

$$f(\boldsymbol{y}) = \frac{1}{|\Sigma|^{1/2}(2\pi)^{nm/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{\beta})^T\Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})\right\}. \tag{34}$$

and the density function of $\tilde{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, (X^T\Sigma^{-1}X)^{-1}\right)$ is:

$$g(\tilde{\boldsymbol{\beta}}) = \frac{1}{|\left(X^T\Sigma^{-1}X\right)^{-1}|^{1/2}(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})^T\left(X^T\Sigma^{-1}X\right)(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})\right\}. \tag{35}$$

Thus

$$f(\boldsymbol{z}) = \frac{1}{|J|}\frac{1}{|\Sigma|^{1/2}|X^T\Sigma^{-1}X|^{1/2}(2\pi)^{(nm-p)/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})\right\}, \tag{36}$$

which does not depend on $\boldsymbol{\beta}$. It can be shown that the Jacobian term does not depend on any parameters and can therefore can be ignored for making inferences about $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$.

Also note that the RHS of (36) is independent of A, and the same result would therefore hold for any $\boldsymbol{Z}$ s.t. $E[\boldsymbol{Z}] = 0$ and $\text{Cov}(\boldsymbol{Z}, \hat{\boldsymbol{\beta}}) = 0$.

We now have

$$\log f(\boldsymbol{z}) = \ell_R(\Sigma).$$

The REML estimator, $\tilde{\Sigma}$, maximizes $\log f(\boldsymbol{z})$,

$$-\frac{1}{2}\log|\Sigma| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}),$$

whereas the MLE of Σ maximizes

$$-\frac{1}{2}\log|\Sigma| - \frac{1}{2}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{y} - X\tilde{\boldsymbol{\beta}}), \text{ see (20).}$$

Hence, REML only need a simple modification to the MLE algorithm.

- The REML estimator for $\boldsymbol{\beta}$ is:

$$\tilde{\boldsymbol{\beta}}(\tilde{\Sigma}) = (X^T\tilde{\Sigma}^{-1}X)^{-1}X^T\tilde{\Sigma}^{-1}\boldsymbol{y}.$$

- The REML estimator for σ^2 is:

$$\tilde{\sigma}^2 = \frac{\text{RSS}(\tilde{V}_0)}{nm - p}.$$

- The REML estimator is asymptotically equivalent to MLE when p is fixed and $mn \longrightarrow \infty$.
- When $p \longrightarrow \infty$, REML is preferable for estimating variance parameters.
- The estimates of $\boldsymbol{\beta}$ do differ between ML and REML, but often not substantially.
- Justification of REML method: in the absence of information on $\boldsymbol{\beta}$, no information is lost about Σ by using $\boldsymbol{Z}$ (marginal sufficiency). From a Bayesian perspective, it corresponds to using a uniform prior on $\boldsymbol{\beta}$ and integrating it out.
- More about REML when discussing linear mixed models.

Robust Estimation of Standard Errors

Recall the variance for the WLS estimator $\hat{\beta}(W)$ is

$$\operatorname{Var}\hat{\beta}(W) = A^{-1}BA^{-1} \tag{37}$$

where

$$A = X^T WX, \tag{38}$$
$$B = X^T W\Sigma WX. \tag{39}$$

Σ is often unknown. If we can get a *consistent* estimator of Σ, $\hat{\Sigma}$, then we can use

$$\hat{\text{Var}}[\hat{\beta}(W)] = A^{-1}\hat{B}A^{-1}$$

where $\hat{B} = X^T W\hat{\Sigma}WX$, as the estimated variance of $\hat{\beta}(W)$ which will converge to the correct variance asymptotically.

- W^{-1} is often called the *working variance matrix*.
- The choice of W will not affect the *validity* of the inference based on $\hat{\beta}(\mathsf{W})$ and $\hat{\text{Var}}\hat{\beta}(\mathsf{W})$.
- The choice of W may affect the *efficiency* (larger variances).
- One estimator for Σ is,
$$\text{Var}(\boldsymbol{Y}_i) = (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^T,$$
where $\boldsymbol{\mu}_i$ corresponds to the fitted value from a *correctly specified*, sometimes over-elaborated or saturated model.
- The corresponding estimator for the variance of $\hat{\beta}(\mathsf{W})$ is often referred to as the *sandwich* or empirical estimator.
- For testing $H_0 : \mathsf{Q}\boldsymbol{\beta} = 0$, where Q is a full rank $(q \times p)$ matrix for some $q < p$, we have (approximately)
$$\mathsf{Q}\hat{\beta}(\mathsf{W}) \sim \mathcal{N}\left(\mathsf{Q}\boldsymbol{\beta}, \mathsf{Q}\hat{\text{Var}}\hat{\beta}(\mathsf{W})\mathsf{Q}^T\right).$$
The Wald test statistic can be used:
$$\left\{\mathsf{Q}\hat{\beta}(\mathsf{W})\right\}^T \left\{\mathsf{Q}\hat{\text{Var}}\hat{\beta}(\mathsf{W})\mathsf{Q}^T\right\}^{-1} \left\{\mathsf{Q}\hat{\beta}(\mathsf{W})\right\}$$
which has an asymptotically χ^2 distribution with q degrees of freedom under the null hypothesis.

Comments on the Sandwich Estimator

- A special case of the Generalized Estimation Equation (GEE) method.
- This approach is *semi-parameteric* in the sense that the estimation and inference for parameter $\boldsymbol{\beta}$ only require specification of the mean.
  - The robust variance estimator $\hat{\text{Var}}[\hat{\beta}(W)] = A^{-1}\hat{B}A^{-1}$ can be shown to be consistent (consistency requires large number of subjects, m.) as long as the mean is correctly specified, no matter the covariance model is correctly specified or not.
  - This is not true for the model-based (naive) estimator $\left(X^T WX\right)^{-1}$.
  - Efficiency can be gained if an oppropriate covariance model can be specified.
- In the case of missing observations, the GEE method only provides valid inferences for the fixed effects under strict assumptions about the the underlying missingness process (more on this later).
- When the observational times are largely unique for each subject, some smoothing may be required to use the sandwich estimator.

Further Reading

- Chapters 4 (before Example 4.1), Appendix A.2-A.4, and 6.4 of the textbook (Diggle et al 2002).