2. General Linear Models for Longitudinal Data

Review of Multivariate Normal Distribution

The density function of a multivariate normal random vector Y is

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

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

- This distribution is completely specified by its first two moments, $\mu = E(Y)$ and $\Sigma = var(Y)$.

- Each Y_j has a marginal univariate normal distribution with mean μ_j and variance σ_{jj}.

- 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, $Z_1 = (Y_1, \cdots, Y_{n_1})^T$ also has a multivariate normal distribution with mean $\mu_1 = (\mu_1, \cdots, \mu_{n_1})^T$ and variance Σ_{11}.

- Let $Z_2 = (Y_{n_1+1}, \cdots, Y_n)^T$. Then the conditional distribution of Z_1 given Z_2 is also normal with mean

$$\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(Z_2 - \mu_2),$$

and variance

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

- If B is an $m \times n$ matrix with $m \leq n$, then BY (a linear transformation) is also multivariate normal with mean $B\mu$ and variance $B\Sigma B^T$. If $Y \sim N(X\beta, \sigma^2 I)$, the MLE of the coefficients for a linear regression model,

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

$\hat{\beta}$ also has multivariate normal distribution with mean and variance respectively,

$$E(\hat{\beta}) = \beta, \qquad var(\hat{\beta}) = \sigma^2(X^T X)^{-1}.$$

The random variable

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

Weighted Estimation

Consider the situation where subjects report both the number of successes and the number of attempts (Y_i, N_i), for example, the number of live birth (Y_i) in a litter (N_i), the number of surgical errors in a hospital.
Question: How to combine these data from $i = 1, \cdots, m$ litters/hospitals to estimate a common rate (proportion) of successes?

$$\text{Proposal 1} : \hat{p}_1 = \sum_i Y_i / \sum_i N_i,$$

$$\text{Proposal 2} : \hat{p}_2 = \frac{1}{m} \sum_i Y_i / N_i.$$

A simple example with data $(1, 10)$, $(2, 100)$:

$$\hat{p}_1 = \frac{1+2}{10+100} = 0.03$$

$$\hat{p}_2 = \frac{1}{2}\left(\frac{1}{10} + \frac{2}{100}\right) = 0.06.$$

Each of these estimators, $\hat{p}_1$ and $\hat{p}_2$, can be viewed as weighted estimators

$$\hat{p}_w = \frac{\sum_i w_i \frac{Y_i}{N_i}}{\sum_i w_i}.$$

We obtain $\hat{p}_1$ by letting $w_i = N_i$, corresponding to equal weight given to each Bernoulli trial, Y_{ij}, $Y_i = \sum_{j=1}^{N_i} Y_{ij}$. We also obtain $\hat{p}_2$ by letting $w_i = 1$ corresponding to equal weight given to each litter/hospital.

What is optimal?
Answer: whatever weights are closet to inverse variance of Y_i/N_i (Gaussian-Markov).

- If the litters/hospitals are perfectly independent, then $var(Y_i) = N_i p(1-p)$ and $\hat{p}_1$ is best.

- If the litters/hospitals are perfectly dependent, then $var(Y_i) = N_i p(1-p)\{1 + (N_i - 1)\rho\}$. For example (where ρ is the within cluster correlation), and estimator closer to $\hat{p}_2$ is best.

Summary

- We need account for dependence in inference with clustered/correlated data.

- The choice of weighting depends on the variance of the variables.

General Linear Models

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.

The 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.

For the moment, we will assume the observation times are common for all subjects. i.e., $t_{ij} = t_j$, $j = 1, \cdots, n$ for all $i = 1, \cdots, m$. The general linear model assumes

- All subjects are independent. That is, if X_i is stochastic, $(Y_1, X_1), \cdots, (Y_m, X_m)$ are independent; if X_i is fixed by design, Y_i are independent.

- Given X_i,

$$E(Y_i|X_i) = X_i\beta, \tag{1}$$
$$var(Y_i|X_i) = \Sigma_i = \sigma^2 V. \tag{2}$$

- We also assume $Y_i \sim N(X_i\beta, \Sigma_i)$ when maximum likelihood is used.

Note

- This model implies

$$E(Y_{ij}|X_{i1}, \cdots, X_{in}) = E(Y_{ij}|X_{ij}).$$

  When there are time-varying 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 unit. 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 (comment set of t_j's) and complete (no missing data). The covariance of the response variable Y $(m \times n = N)$ is

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

Objective:
Inference about β while adjusting for Σ need know or estimate Σ.

Correlation Specifications

1. Exchangeable Correlation
   Assume $\rho_{jk} = cor(\epsilon_{ij}, \epsilon_{ik}) = \rho$ for $j \neq k$ (the same

for all pairs of observations),

$$V_0 = (1-\rho)I + \rho J = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

where V_0 is the correlation matrix and $\Sigma_i = \sigma^2 V_0$, I is a $n \times n$ identity matrix, and J is a $n \times n$ matrix of 1's.

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

$$y_{ij} = x_{ij}\beta + u_i + z_{ij},$$

where $u_i \sim N(0, v^2)$ are random effects and $z_{ij} \sim N(0, \tau^2)$ are measurement errors.

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

$$Y_{ij} = \mu_j + \alpha_i + \epsilon_{ij} \text{ for } i = 1, \cdots, N;\ j = 1, \cdots, n, \quad (3)$$

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

$$\begin{aligned} cov(Y_{ij}, Y_{ik}) &= cov(\mu_j + \alpha_i + \epsilon_{ij}, \mu_k + \alpha_i + \epsilon_{ik}) \\ &= cov(\alpha_i, \alpha_i) \\ &= v^2, \\ var(Y_{ij}) &= var(\mu_j + \alpha_i + \epsilon_{ij}) \\ &= v^2 + \tau^2. \end{aligned}$$

Hence,

$$\rho = cor(Y_{ij}, Y_{ik}) = \frac{v^2}{v^2 + \tau^2}.$$

Here, v^2 is the between subject variance, τ^2 is within subject variance, and $\sigma^2 = v^2 + \tau^2$ is total variance.

Model (3) can be written as

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

$$\leftrightarrow \quad Y_i = \mu + \alpha_i 1 + \epsilon_i, \tag{4}$$

where α_i and ϵ_i are independent,

$$\alpha_i \sim^{iid} N(0, v^2), \quad \epsilon_i \sim^{iid} N(0, \tau^2 I).$$

Therefore,

$$\begin{aligned} E(Y_i) &= \mu, \\ var(Y_i) &= var(\alpha_i 1) + var(\epsilon_i) \\ &= v^2 11^T + \tau^2 I. \end{aligned}$$

2. Exponential Correlation

A different model assumes the correlation of

observations closer together in time is larger than that of observations farther apart

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

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

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

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

3. Gaussian Correlation
   In exponential correlation, the lag 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 $\rho > 0$.

Review of Likelihood Inference

If Y has probability density function $f(y;\theta)$, then the likelihood function for θ is

$$L(\theta|y) = f(y;\theta).$$

For statistical correctness, we say the "likelihood of the parameter" and never the "likelihood of the data". The log-likelihood is

$$l(\theta|y) = \log L(\theta|y).$$

Note that the likelihood is only defined up to a multiplicative constant. That is, the likelihood is defined up to an additive constant.

Maximum likelihood estimation

The maximum likelihood estimator, $\hat{\theta}$, maximizes the likelihood (log-likelihood):

$$L(\hat{\theta}) = \sup_{\theta\in\Theta} L(\theta).$$

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 (Cramer-Rao lower bound) among asymptotically unbiased estimators.

Score equation and Information

The maximization is often achieved by solving the score equation:

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

Note that at the MLE,

$$\ddot{l}(\hat{\theta}) \equiv \frac{\partial S(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = \frac{\partial^2 \log L}{\partial \theta \partial \theta^T}|_{\theta=\hat{\theta}} < 0.$$

The function $\dot{l}(\theta)$ is called the score function and

$$E(\dot{l}(\theta)) = 0, \tag{6}$$

$$I(\theta) \equiv var(\dot{l}(\theta)) = E(\dot{l}(\theta)\dot{l}(\theta)^T) = -E(\ddot{l}(\theta)). \tag{7}$$

$I(\theta)$ is called the Fisher information or expected information for θ and $-\ddot{l}(\theta)$ is known as the observed information. The asymptotic variances of the MLE $\hat{\theta}$ is given by $I(\hat{\theta})^{-1}$.

Hypothesis Testing

- For testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, we can use

  1. The likelihood ratio test statistic

$$G \equiv 2\left\{l(\hat{\theta}) - l(\theta_0)\right\} \tag{8}$$

  2. Wald test statistic

$$W \equiv (\hat{\theta} - \theta_0)^T I(\hat{\theta})(\hat{\theta} - \theta_0), \tag{9}$$

  where $I(\hat{\theta})$ is the observed information $-\ddot{l}(\hat{\theta})$.
  - When testing the $H_0 : \beta = \beta_0$ for regression

models, $t - test$, or $F - test$ is often used for better small sample performance.

3. The score or Rao test statistic

$$R \equiv \dot{l}(\theta_0)^T I(\theta_0)^{-1} \dot{l}(\theta_0). \tag{10}$$

Note that it is not necessary to find the MLE $\hat{\theta}$ 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 degree of freedom where p is the dimension of the parameter θ (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 statistic, however, do differ.

OLS for General Linear Model

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

$$(y - X\beta)^T (y - X\beta).$$

The OLS estimator is still unbiased

$$E(\hat{\beta}_{OLS}) = (X^T X)^{-1} X^T E(y) = \beta.$$

However,

$$\begin{aligned} var(\hat{\beta}_{OLS}) &= var\{(X^T X)^{-1} X^T y\} \\ &= \sigma^2 (X^T X)^{-1} X^T V X (X^T X)^{-1} \\ &\neq \sigma^2 (X^T X)^{-1}. \end{aligned} \tag{11}$$

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

Weighted Least Squares

In univariate regression, WLS yields estimates of β that minimizes the objective function

$$Q(\beta) = \sum_{i=1}^{m} W_i (Y_i - X_i \beta)^2. \tag{12}$$

Analogously, the multivariate version of WLS finds the value of the parameter $\beta(w)$ that minimizes

$$\begin{aligned} Q_w(\beta) &= (Y - X\beta)^T W (Y - X\beta) \\ &= \sum_{i=1}^{m} (Y_i - X_i \beta)^T W_i (Y_i - X_i \beta), \end{aligned} \tag{13}$$

where W_i is an $(n_i \times n_i)$ positive definite and symmetric matrix. It is straightforward to see that

$$\begin{aligned} U(\beta) &= \frac{\partial}{\partial \beta} Q_w(\beta) = -2X^T W (Y - X\beta) \\ &= -2 \sum_{i=1}^{m} X_i^T W_i (Y_i - X_i \beta). \end{aligned}$$

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

$$\begin{aligned} \hat{\beta}(W) &= (X^T W X)^{-1} X^T W Y \qquad (14) \\ &= (\sum_{i=1}^{m} X_i^T W_i X_i)^{-1} (\sum_{i=1}^{m} X_i^T W_i Y_i). \end{aligned}$$

- The OLS estimator corresponds to $W_i^{-1} = \sigma^2 I_i$ assuming observations are independent both within and between subjects.

- If $X_i = X$ and $W_i = W$ for all i (eg., complete and balanced design), then

$$\hat{\beta}(W) = (X^T W X)^{-1} X^T W \left(\frac{1}{m} \sum_{i=1}^{m} Y_i \right).$$

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

Weighted Least Squares

- $\hat{\beta}(W)$ is unbiased for any W.

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

- Variance

$$\begin{aligned} var(\hat{\beta}(W)) &= A^{-1} var(X^T W Y) A^{-1} \\ &= A^{-1} (X^T W var(Y) W X) A^{-1} \\ &= A^{-1} (X^T W \Sigma W X) A^{-1}, \end{aligned}$$

  where $A^{-1} = (X^T W X)^{-1}$.

- If $W = I$ (OLS),

$$var(\hat{\beta}(W)) = (X^T X)^{-1} (X^T \Sigma X) (X^T X)^{-1}, \quad (15)$$

- If $W = \Sigma^{-1}$,

$$var(\hat{\beta}(W)) = (X^T \Sigma^{-1} X)^{-1}.$$

- It can be shown that

$$var(\hat{\beta}(\Sigma^{-1})) \leq var(\hat{\beta}(W)). \tag{16}$$

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

The (relative) efficiency of estimator $\hat{\beta}(W)$ is measured by

$$Efficiency = \frac{var(\hat{\beta}(\Sigma^{-1}))}{var(\hat{\beta}(W))}.$$

- As noted in textbook (DHLZ, p. 60), the relative efficiency of OLS estimator is often quite good, sometimes, even fully efficient (efficiency=1).

- In the presence of positive autocorrelation, naive use of OLS can seriously over- or underestimate the variance of $\hat{\beta}(I)$ depending on the design matrix.

- In any case, the correct variance of the OLS estimator (15) should be used instead of the naive OLS variance (11).

Maximum Likelihood forGeneral Linear Model

Assuming model $Y \sim N(X\beta, \sigma^2 V)$, the log-likelihood function is

$$\begin{aligned} l(\beta, \sigma^2, V_0) &= -\frac{nm}{2}\log(\sigma^2) - \frac{m}{2}\log|V_0| \\ &+ \sigma^{-2}\left\{-\frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta)\right\}, \end{aligned} \tag{17}$$

where $V = diag\{V_0, \cdots, V_0\}$.

- If V_0 is known, the score functions with respect to β and σ^2 are

$$\frac{\partial}{\partial\beta} l(\beta, \sigma^2, V_0) = \sigma^{-2} X^T V^{-1}(Y - X\beta), \tag{18}$$

$$\frac{\partial}{\partial\sigma^2} l(\beta, \sigma^2, V_0) = -\frac{nm}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\left\{(Y - X\beta)^T V^{-1}(Y - X\beta)\right\} \tag{19}$$

Setting (18) to 0, we get the MLE

$$\hat{\beta}(V_0) = (X^T V^{-1} X)^{-1} X^T V^{-1} Y, \tag{20}$$

which is just the efficient WLS estimator $\hat{\beta}(\Sigma^{-1})$. Note that the absence of a scaling factor does not

affect the estimate, implying that we only need use a weight matrix $W \propto \Sigma^{-1}$.

- In general, V_0 is not known. Let

$$RSS(V_0) = (Y - X\hat{\beta}(V_0))^T V^{-1} (Y - X\hat{\beta}(V_0)). \quad (21)$$

Substitute $\hat{\beta}(V_0)$ into (19), set it to 0, and solve for σ^2. Then we get

$$\hat{\sigma}^2 = \frac{RSS(V_0)}{nm}. \quad (22)$$

If the correlation matrix V_0 is parameterized by α, then substitute (20) and (22) into (17). We have

$$\begin{aligned} l(V_0(\alpha)) &= -\frac{nm}{2} \log\{RSS(V_0)\} - \frac{m}{2} \log |V_0(\alpha)| \\ &\quad -\frac{nm}{2}. \end{aligned} \quad (23)$$

  - By maximizing (23), we can get the MLE for α and then the MLEs for β and σ^2 follow.

- The maximization with regard to α in general does not have a closed form and requires numerical optimization techniques.
- The ML estimator for σ^2 and α are biased. The bias is substantial when the dimension of β is large.
- If the design matrix X is not correctly specified, this simultaneous estimation of β, σ^2 and α will not work because $\hat{\sigma}^2$ and $\hat{\alpha}$ may not even be consistent.
- A possible strategy is to use an over-elaborated or saturated model for X, get a consistent estimator for the variance structure, and then refit a more economical design matrix. However, this may exaggerate the bias.

Restricted Maximum Likelihood (REML)

- Consider the simple example where $y = (y_1, \cdots, y_n)^T$ are iid random sample from $N(\mu, \sigma^2)$. The MLE for μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}. \tag{24}$$

  The MLE for σ^2 is

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2. \tag{25}$$

  However, $\hat{\sigma^2}$ is biased:

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

One can simple "adjust" for the bias and get the usual estimator

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

For general models, a more general approach is needed to correct the bias of the ML estimators and that is REML.

- For general linear model, $Y \sim N(X\beta, \Sigma)$, the REML estimator is defined as a maximum likelihood estimator based on a linear transformed set of data $Y^* = AY$ such that the distribution of Y^* does not depend on β.
  - A remarkable property is that the resulted estimators for σ^2 and α dose not depend on the choice of A.
  - The transformation needs not be explicit.

Restricted Maximum Likelihood (REML)

Combining the variance component parameters α and σ^2, the log-likelihood function for the general linear model can be written as

$$l(\beta, \Sigma) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left\{(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)\right\}. \quad (27)$$

For fixed Σ, (27) is maximized over β by

$$\tilde{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y = GY. \quad (28)$$

Substitute (28) into (27). Then we obtain the profile log-likelihood for Σ.

$$\begin{aligned} l_p(\Sigma) &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left\{(Y - X\tilde{\beta})^T\Sigma^{-1}(Y - X\tilde{\beta})\right\} \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left\{Y^T\Sigma^{-1}Y - (X\tilde{\beta})^T\Sigma^{-1}(X\tilde{\beta})\right\} \\ &= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}\left\{Y^T\Sigma^{-1}\left[I - X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right]Y\right\} \quad (29) \end{aligned}$$

REML estimator of Σ maximizes the restricted log-likelihood

$$l_R(\Sigma) = l_p(\Sigma) - \frac{1}{2}\log|X^T\Sigma^{-1}X|. \quad (30)$$

- REML takes into account the loss of degrees of freedom for estimating β and is less biased for small sample sizes (relative to p).

- Once Σ is estimated, it is plugged back to (28) to get the "REML estimate" of β even though strictly speaking, REML only refer to the variance components.

Derivation of REML Method

Let

$$A = I - X(X^T X)^{-1} X^T \tag{31}$$

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

$$BB^T = A_{nm \times nm} \text{ and } B^T B = I_{(nm-p)\times(nm-p)}, \tag{32}$$

and $Z = B^T Y$ (an nm-p vector), then

$$\begin{aligned} E(Z) &= B^T E(Y) = B^T X\beta \\ &= B^T BB^T X\beta = B^T AX\beta. \end{aligned}$$

Since

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

we have

$$E(Z) = 0.$$

In addition, the covariance of Z with $\tilde{\beta}$ from (28), which is unbiased for β, is

$$\begin{aligned} cov(Z, \tilde{\beta}) &= E\left\{Z(\tilde{\beta} - \beta)^T\right\} \\ &= E\left\{B^TY(Y^TG^T - \beta^T)\right\} \\ &= B^TE(YY^T)G^T - B^TE(Y)\beta^T \\ &= B^T\left\{var(Y) + E(Y)E(Y)^T\right\}G^T - B^TE(Y)\beta^T \\ &= B^T\Sigma\left\{(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\right\}^T \\ &= B^TX(X^T\Sigma^{-1}X)^{-1} = 0. \end{aligned}$$

Therefore, Z and $\tilde{\beta}$ are independent because they have a joint multivariate normal distribution with zero covariance. We now show that the distribution of Z does not depend on β and the choice of B. Note that

$\begin{pmatrix} Z \\ \tilde{\beta} \end{pmatrix} = \begin{pmatrix} B^T \\ G \end{pmatrix} Y$ is a linear transformation of Y with density

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

where J is the Jacobian. Therefore,

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

To obtain the explicit form of $f(z)$, we use the following result.

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

Note:

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

The density function of Y is

$$f(y) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{nm/2}} \exp \left\{ -\frac{1}{2} (y - X\beta)^T \Sigma^{-1} (y - X\beta) \right\} \quad (33)$$

and the density of $\tilde{\beta}$ is normal with mean β and covariance $(X^T\Sigma^{-1}X)^{-1}$.

$$g(\tilde{\beta}) = \frac{1}{|(X^T\Sigma^{-1}X)^{-1}|^{1/2}(2\pi)^{p/2}} \exp\left\{-\frac{1}{2}(\tilde{\beta}-\beta)^T(X^T\Sigma^{-1}X)(\tilde{\beta}-\beta)\right\}. \quad (34)$$

Thus,

$$\begin{aligned} f(z) &= \frac{1}{|J|}\frac{1}{|\Sigma|^{1/2}|(X^T\Sigma^{-1}X)^{-1}|^{-1/2}(2\pi)^{(nm-p)/2}} \\ &\quad \exp\left\{-\frac{1}{2}(y-X\tilde{\beta})^T\Sigma^{-1}(y-X\tilde{\beta})\right\} \quad (35) \end{aligned}$$

which does not depend on β. It can be shown that the Jacobian term does not depend on any parameters. We now have

$$\log f(z) = l_R(\Sigma).$$

- REML estimator for σ^2 is

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

- The REML estimator for V_0 maximize the reduced log-likelihood

$$L^*(V_0) = -\frac{1}{2}m\left\{n\log RSS(V_0) + \log|V_0|\right\} - \frac{1}{2}\log\left(|X^T\Sigma^{-1}X|\right).$$

  Then find the REML estimators $\tilde{\beta} = \tilde{\beta}(\tilde{V}_0)$ and $\tilde{\sigma^2} = \tilde{\sigma^2}(\tilde{V}_0)$.

- The REML estimator is asymptotically equivalent to MLE when p is fixed and $nm \to \infty$.

- When $p \to \infty$, REML is preferable for estimating variance parameters.

- The estimates of β do differ between ML and REML, but often not substantially.

- Justification of REML method: in the absence of information on β, no information is lost about Σ by using Z (marginal sufficiency). From a Bayesian perspective, it corresponds to using a uniform prior on β and integrate it out.

- More about REML in discussing linear mixed models.

Robust Estimation

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

$$var(\tilde{\beta}(W)) = A^{-1}BA^{-1} \tag{36}$$

where $A = X^T W X$ and $B = X^T W \Sigma W X$.

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

$$\hat{var}(\hat{\beta}(W)) = A^{-1}\hat{B}A^{-1} \tag{37}$$

where $\hat{B} = X^T W \hat{\Sigma} W X$. (37) is 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 $\hat{\beta}(W)$ and $\hat{var}(\hat{\beta}(W))$.
- The choice of W may affect the efficiency (larger variances).

- One estimator for Σ is

$$var(y_i) = E(y_i - \mu_i)(y_i - \mu_i)^T$$

  where μ_i corresponds to the fitted value from a correctly specified, sometimes over-elaborated or saturated model.

- The corresponding estimator for the variance of $\hat{\beta}(W)$ is often referred to as the sandwich or empirical estimator.

Comments on the sandwich estimator

- For testing $H_0 : Q\beta = 0$ where Q is a full rank ($q \times p$) matrix for some $q < p$, we have (approximately)

$$Q\hat{\beta}(W) \sim N\left(Q\beta, Q\hat{var}(\hat{\beta}(W))Q^T\right).$$

  The Wald test statistic can be used

$$(Q\hat{\beta})^T \left\{Q\hat{var}(\hat{\beta}(W))Q^T\right\}^{-1} (Q\hat{\beta})$$

which has an asymptotically χ^2 distribution with q degrees of freedom under the null hypothesis.

- The key property of this estimator is consistency - requires large number of subjects (m).

- A special case of the Generalized Estimating Equation (GEE) method.

- This approach is semi-parametric in the sense the estimation and inference for parameter β only require specification of the mean.

- When the observational times are largely unique for each subject, some smoothing may be required to use the sandwich estimator.

More about Weighted Least Square / GEE

- For a fixed W, under the only assumption $E(Y) = X\beta$, the WLS estimator $\hat{\beta}(W)$ is

  1. Unbiased.
  2. Consistent and asymptotically normal.
  3. Efficient if a consistent estimator of $var(Y_i)$ is available.

- If the $W(\alpha)$ depends on the data, i.e., α has to be estimated from the data. The WLS estimator $\hat{\beta}(\hat{W})$ solves:

$$X^T \hat{W}(Y - X\beta) = 0,$$

  and is not necessarily unbiased.

- α can be estimated using simple methods of moments or ML or REML (even without normality assumption).

- However, under mild assumptions, $\hat{\beta}(\hat{W})$ is still consistent and asymptotically normal and has the

same asymptotic variance as $\hat{\beta}(W)$ if $\hat{W}$ converges to W. So again, using a consistent estimator of $var(Y_i)$ ensures asymptotic efficiency.

- α is not parameter of interest and nothing is said about them. There are extensions of GEE that models variance parameters by specifying higher moments (Leave to later).

- GEE method trades off some efficiency with consistency, depending upon whether the correlation structure is correctly specified.

- Using a reasonable working correlation matrix can improve efficiency.

- When there is missing data, or highly unbalanced design, the robust approach becomes problematic.

- The GEE method is most appropriate when the number of subjects is much larger than the number of observations per subject, and complete balanced design.

Miscellaneous

Derivation of REML in Bayesian Framework

Assuming an noninformative (flat) prior on β, the marginal posterior distribution of Σ is obtained by integrating over β:

$$f_R(\Sigma) \propto \int L(\beta, \Sigma) d\beta. \tag{38}$$

First write,

$$\begin{aligned} L(\beta, \Sigma) \quad \propto \quad & |\Sigma|^{-1/2} \exp\left\{ -\frac{1}{2}(Y - X\hat{\beta})^T \Sigma^{-1} (Y - X\hat{\beta}) \right\} \\ & \times \exp\left\{ -\frac{1}{2}(\hat{\beta} - \beta)^T \left(X^T \Sigma^{-1} X \right) (\hat{\beta} - \beta) \right\}, \end{aligned} \tag{39}$$

where $\hat{\beta} = \hat{\beta}(\Sigma^{-1})$. Note that

$$\int \exp\left\{ -\frac{1}{2} t^T \Omega^{-1} t \right\} dt = (2\pi)^{1/2} |\Omega|^{1/2}$$

we have the restricted likelihood function

$$f_R(\Sigma) \quad \propto \quad |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(Y - X\hat{\beta})^T \Sigma^{-1} (Y - X\hat{\beta})\right\} \\ |X^T \Sigma^{-1} X|^{-1/2}. \tag{40}$$