Correlated Data

Repeated Measurements

repeated observations of the response variable on the same individuals over multiple occasions or under different experimental conditions

Clustered Data

observations are grouped in clusters, e.g., schools, hospitals, villages, families

measurements taken on individuals within the same cluster may be correlated even after adjusting for known covariates

Spatially Correlated Data

observations associated with a specific location

e.g. epidemiological studies on incidence and prevalence of disease from region-specific counts

Multivariate Data

two or more response variables measured on each individual

Predictor Variables

Within-Unit / Time-Dependent Covariates

covariate that changes over time

e.g. age at visit, medical measurements

Between-Unit / Time-Independent Covariates

baseline characteristics, e.g. sex, race

Independence

two random variables, X and Y, are independent if their joint density function is the product of their two marginal density functions

$$f_{X,Y}(X,Y) = f_x(X) f_Y(Y)$$

or if the conditional distribution of Y given X doesn't depend on X

e.g. blood pressure is independent of age if the distribution of blood pressures are the same for every age group

$$f_Y(Y|X) = f_Y(Y)$$

Correlation

$$Cov(X,Y) = E(Y-\mu_Y)(X-\mu_X)$$
$$Var(Y) = E(Y-\mu_Y)(Y-\mu_Y)$$

two random variables, X and Y, are uncorrelated if $covariance(X,Y) = E(Y-\mu_Y)(X-\mu_X) = 0$
two random variables, X and Y, are correlated if $covariance(X,Y) = E(Y-\mu_Y)(X-\mu_X) \neq 0$
independent variables are uncorrelated, but variables can be uncorrelated without being independent
covariance can be any positive or negative value and unit depends on units of the variables
to make covariance independent of units, divide by standard deviations of the two variables

$$corr(X,Y) = \frac{E(Y-\mu_Y)(X-\mu_X)}{\sigma_Y \sigma_X}$$

correlation is between -1 and 1
repeated measures on the same individual or cluster are usually positively correlated

Variance-Covariance Matrix

$$Y_{ij} = j^{th}\ measure\ of\ i^{th} subject$$

vector of all p observations of i^{th} subject used to create a symmetric square variance-covariance matrix

$$\sum\nolimits_i = Cov\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \\ \vdots \\ Y_{ip} \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2p} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \cdots & \sigma_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \sigma_{p3} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$Cov(Y_{ir}, Y_{is}) = \sigma_{rs}$$

σ_{rs} = covariance between r and s repeated measures of the i^{th} subject

Model Estimation

Generalized Least Squares (GLS)

extension of ordinary least squares
tries to minimize a weighted sum of squared residuals
can accommodate heterogeneity and correlation
weights correspond to inverse of variance-covariance matrix
estimate parameters by minimizing objective function, $Q_{GLS}(\beta|\theta)$

$$Q_{GLS}(\beta|\theta) = \sum_{i=1}^{n} (Y_i - X_i\beta)' \Sigma_i(\theta)^{-1} (Y_i - X_i\beta)$$

$\Sigma_i(\theta) = Var(Y_i)$
θ = vector of variance-covariance parameters

- θ Known

$$\sum_{i=1}^{n} X_i' \Sigma_i(\theta)^{-1} (Y_i - X_i\beta) = 0$$

$$\hat{\beta}_{GLS} = \left(\sum_{i=1}^{n} X_i' \Sigma_i(\theta)^{-1} X_i\right)^{-1} \sum_{i=1}^{n} X_i' \Sigma_i(\theta)^{-1} Y_i$$

$$Var(\hat{\beta}_{GLS}) = \left(\sum_{i=1}^{n} X_i' \Sigma_i(\theta)^{-1} X_i\right)^{-1}$$

- θ Unknown
  replace θ with a consistent estimate in GLS formulas
  estimated $\hat{\theta}$ gets closer to true θ as sample size increases

Maximum Likelihood (ML)

estimate parameters based on what is the most probable given what has been observed
estimated iteratively by maximizing profile log-likelihood
in small samples, estimated variances are affected by small-sample bias and will underestimate true variance

Restricted Maximum Likelihood (REML)

accounts for small-sample bias
gives unbiased estimates for variances by maximizing restricted profile log-likelihood