

PART A

Introduction to Credibility

LIMITED FLUCTUATION CREDIBILITY

Prediction The updated prediction, U , is a weighted average of D (data) and M (manual rate):

$$U = Z D + (1 - Z) M$$

where Z , $0 \leq Z \leq 1$, is called the **credibility factor**.

STANDARDS FOR FULL-CREDIBILITY TO LIMIT THE FLUCTUATION AROUND:

Define $\lambda_F = \left(\frac{z_{1-\alpha/2}}{k}\right)^2$ and $C_X = \frac{\sigma_X}{\mu_X}$ the coefficient of variation of X .

| | Any frequency distribution | Poisson frequency distribution |
|--|---|--|
| Claim Frequency | $n_0 = \lambda_F \left(\frac{\sigma_N^2}{\mu_N}\right)$ | $n_0 = \lambda_F$ |
| Claim Severity | $n_0 = \lambda_F C_X^2$ | (same as any frequency) |
| Aggregate Losses and Pure Premium | $n_0 = \lambda_F \left(\frac{\sigma_N^2}{\mu_N} + C_X^2\right)$ | $n_0 = \lambda_F (1 + C_X^2) = \lambda_F \frac{E(X^2)}{\mu_X^2}$ |
| $Z = 1$ if the observed number of claims $> n_0$ | | |

PARTIAL CREDIBILITY FACTORS

| | Any frequency distribution | Poisson frequency distribution |
|---------------------------------|--|--|
| Claim Frequency | $Z = \sqrt{\frac{\mu_N}{\lambda_F \left(\frac{\sigma_N^2}{\mu_N}\right)}}$ | $Z = \sqrt{\frac{\mu_N}{\lambda_F}}$ |
| Claim Severity | $Z = \sqrt{\frac{N}{\lambda_F C_X^2}}$ | (same as any frequency) |
| Aggregate Loss and Pure Premium | $Z = \sqrt{\frac{\mu_N}{\lambda_F \left(\frac{\sigma_N^2}{\mu_N}\right) + C_X^2}}$ | $Z = \sqrt{\frac{\mu_N}{\lambda_F (1 + C_X^2)}}$ |

Within the square root, the denominator is the standard for full credibility of the corresponding risk measure. The numerator, μ_N or N , is observed from data, where μ_N is the expected number of claims coming from the data, and N is the observed number of claims. If μ_N can not be calculated from the data, then the observed number of claims can be used to calculate the partial credibility factor.

Note: If the ratio is greater than 1, then full credibility is attained and $Z = 1$.

 C_X^2 AND $(1 + C_X^2)$ FOR SOME COMMONLY USED SEVERITY DISTRIBUTIONS

| X | C_X^2 | $1 + C_X^2$ |
|--|--------------------------------|---|
| (Two-parameter) Pareto (α, θ) | $\alpha/(\alpha - 2)$ | $2(\alpha - 1)/(\alpha - 2)$ |
| Single-parameter Pareto (α, θ) | $\frac{1}{\alpha(\alpha - 2)}$ | $\frac{(\alpha - 1)^2}{\alpha(\alpha - 2)}$ |
| Gamma (α, θ) | $1/\alpha$ | $(\alpha + 1)/\alpha$ |
| Exponential (θ) | 1 | 2 |
| Inverse Gamma (α, θ) | $1/(\alpha - 2)$ | $(\alpha - 1)/(\alpha - 2)$ |
| Inverse Gaussian (μ, θ) | μ/θ | $(\theta + \mu)/\theta$ |
| Lognormal (μ, σ) | $e^{\sigma^2} - 1$ | e^{σ^2} |
| Uniform in $(0, \theta)$ | $1/3$ | $4/3$ |

Note: The standard for full-credibility for claim severity is $n_0 = \lambda_F C_X^2$, and the standard for full-credibility for aggregate losses and pure premium is $n_0 = \lambda_F (1 + C_X^2)$ for a Poisson frequency distribution.



BÜHLMANN CREDIBILITY

| | |
|---|--|
| Hypothetical mean | $\mu_X(\Theta) = E(X \Theta)$ |
| Process variance | $\sigma_X^2(\Theta) = \text{Var}(X \Theta)$ |
| Expected value of the hypothetical means (unconditional mean) | $\mu_X = E(X) = E[E(X \Theta)] = E[\mu_X(\Theta)]$ |
| Expected value of the process variance (EPV) | $\mu_{PV} = E[\text{Var}(X \Theta)] = E[\sigma_X^2(\Theta)]$ |
| Variance of the hypothetical means (VHM) | $\sigma_{HM}^2 = \text{Var}[E(X \Theta)] = \text{Var}[\mu_X(\Theta)]$ |
| Total variance of X (unconditional variance) | $\text{Var}(X) = E[\text{Var}(X \Theta)] + \text{Var}[E(X \Theta)] = \mu_{PV} + \sigma_{HM}^2$ |
| Bühlmann's k | $k = \frac{\text{EPV}}{\text{VHM}} = \frac{\mu_{PV}}{\sigma_{HM}^2}$ |
| Credibility factor | $Z = \frac{n}{n+k}$ where n represents the number of observations |
| Bühlmann premium | $\hat{X}_{n+1} = Z\bar{X} + (1-Z)\mu_X$ |

Note: X is a risk measure which may be **claim frequency**, **claim severity**, **aggregate loss**, or **pure premium**. Assume that $\{X_1, \dots, X_n, X_{n+1}\}$ are iid given the parameter θ . $\bar{X} = \sum_{i=1}^n X_i/n$ is the sample mean, and $\mu_X = E(X)$ is the unconditional mean.

BÜHLMANN-STRAUB CREDIBILITY

| | |
|--|--|
| Hypothetical mean | $E(X_{ij} \Theta) = \mu_X(\Theta)$ |
| Process variance of X_{ij} | $\text{Var}(X_{ij} \Theta) = \sigma_X^2(\Theta)$ |
| Expected value of the hypothetical means | $\mu_X = E(X) = E[E(X \Theta)] = E[\mu_X(\Theta)]$ |
| Expected value of the process variance (EPV) | $\mu_{PV} = E[\text{Var}(X_{ij} \Theta)] = E[\sigma_X^2(\Theta)]$ |
| Variance of the hypothetical mean (VHM) | $\sigma_{HM}^2 = \text{Var}[E(X_{ij} \Theta)] = \text{Var}[\mu_X(\Theta)]$ |
| Total variance of X | $\text{Var}(X) = E[\text{Var}(X \Theta)] + \text{Var}[E(X \Theta)] = \mu_{PV} + \sigma_{HM}^2$ |
| Bühlmann's k | $k = \frac{\text{EPV}}{\text{VHM}} = \frac{\mu_{PV}}{\sigma_{HM}^2}$ |
| Credibility factor | $Z = \frac{m}{m+k}$ where m represents the number of exposures |
| Bühlmann premium | $\hat{X}_{n+1} = Z\bar{X} + (1-Z)\mu_X$ |

Note: Denote X_{ij} the loss measure of the j th insured in the i th year, $X_i = \frac{\sum_j^m X_{ij}}{m_i}$, $\bar{X} = \frac{1}{m} \sum_{i=1}^n X_i$ (sample mean) and $m = \sum_{i=1}^n m_i$, $j = 1, \dots, m_i$, $i = 1, \dots, n$.

BÜHLMANN PREDICTION FOR CONJUGATE PRIORS

| Prior distribution | Conditional dist. | μ_{PV} (EPV) | σ_{HM}^2 (VHM) | $k = \frac{\mu_{PV}}{\sigma_{HM}^2}$ |
|------------------------------------|-----------------------------|---|---|--------------------------------------|
| Gamma (α, θ) | Poisson (Λ) | $a\theta$ | $a\theta^2$ | $1/\theta$ |
| Beta (a, b) | geometric (Θ)* | $\frac{b(a+b-1)}{(a-1)(a-2)}$ | $\frac{b(a+b-1)}{(a-1)^2(a-2)}$ | $a-1$ |
| Beta (a, b) | Bernoulli (Q) | $\frac{ab}{(a+b)(a+b+1)}$ | $\frac{ab}{(a+b)^2(a+b+1)}$ | $a+b$ |
| Gamma (α, θ) | exponential (Λ)** | $\frac{\theta^2}{(\alpha-1)(\alpha-2)}$ | $\frac{\theta^2}{(\alpha-1)^2(\alpha-2)}$ | $\alpha-1$ |
| Inverse gamma (α, θ) | exponential (Λ) | $\frac{\theta^2}{(\alpha-1)(\alpha-2)}$ | $\frac{\theta^2}{(\alpha-1)^2(\alpha-2)}$ | $\alpha-1$ |
| Normal (μ, a) | normal (Θ, v) | v | a | v/a |

(*) The pmf in [MAS-II Tables](#), $p_k = \beta^k/(1+\beta)^{k+1}$, is parameterized by $p_k = \theta(1-\theta)^k$ where $\theta = 1/(1+\beta)$.

(**) The pdf in [MAS-II Tables](#), $f(x) = (1/\theta) \exp(-x/\theta)$, is parameterized by $f(x) = \lambda \exp(-x\lambda)$ where $\lambda = 1/\theta$.

BAYESIAN INFERENCE AND ESTIMATION

| | |
|--|--|
| Prior probability density function (pdf) | $f_\Theta(\theta)$ |
| Conditional pdf of X_i , given parameter $\Theta = \theta$ | $f_{X_i \Theta}(x_i \theta)$ |
| Likelihood function of $\mathbf{x} = \{x_1, \dots, x_n\}$ | $f_{\mathbf{X} \Theta}(\mathbf{x} \theta) = \prod_{i=1}^n f_{X_i \Theta}(x_i \theta)$ |
| Joint pdf of \mathbf{X} and Θ | $f_{\Theta\mathbf{X}}(\theta, \mathbf{x}) = f_{\mathbf{X} \Theta}(\mathbf{x} \theta) \times f_\Theta(\theta)$ |
| Marginal pdf of \mathbf{X} | $f_{\mathbf{X}}(\mathbf{x}) = \int_\Theta f_{\Theta\mathbf{X}}(\theta, \mathbf{x}) d\theta = E_\Theta[f_{\mathbf{X} \Theta}(\mathbf{x} \theta)]$ |
| Posterior pdf | $f_{\Theta \mathbf{X}}(\theta \mathbf{x}) = \frac{f_{\Theta\mathbf{X}}(\theta, \mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}$ |
| Predictive pdf of X_{n+1} given \mathbf{x} | $f_{X_{n+1} \mathbf{X}}(x_{n+1} \mathbf{x}) = E_{\Theta \mathbf{x}}[f_{X_i \Theta}(x_i \theta)]$ |
| Bayesian premium | $\hat{\mu}_X(\mathbf{x}) = E(X_{n+1} \mathbf{x}) = E_{\Theta \mathbf{x}}[E(X_{n+1} \Theta) \mathbf{x}]$ |



CONJUGATE DISTRIBUTION

| Pair (Prior - Conditional) | Posterior dist. ⁽¹⁾ | Bayesian prem. $\hat{\mu}_X(\mathbf{x})$ | Predictive dist. ⁽²⁾ |
|--|--|--|---------------------------------|
| Gamma (α, θ) - Poisson (Λ) | $\alpha_* = \alpha + \sum x_i$ $\theta_* = (\theta^{-1} + n)^{-1}$ | $\alpha_* \theta_*$ | NB (θ_*, α_*) |
| Beta (a, b) - geometric $(\Theta)^{(3)}$ | $a_* = a + n$ $b_* = b + \sum x_i$ | $\frac{b_*}{a_* - 1}$ | |
| Beta (a, b) - Bernoulli (Q) | $a_* = a + \sum x_i$ $b_* = b + n - \sum x_i$ | $\frac{a_*}{a_* + b_*}$ | |
| Beta (a, b) - binomial (l, Q) | $a_* = a + \sum x_i$ $b_* = b + ln - \sum x_i$ | $(l) \frac{a_*}{a_* + b_*}$ | |
| Gamma (α, θ) - exponential $(\Lambda)^{(4)}$ | $\alpha_* = \alpha + n$ $\theta_* = \theta + \sum x_i$ | $\frac{\theta_*}{\alpha_* - 1}$ | Pareto (α_*, θ_*) |
| Inverse gamma (α, θ) - exponential (Λ) | $\alpha_* = \alpha + n$ $\theta_* = \theta + \sum x_i$ | $\frac{\theta_*}{\alpha_* - 1}$ | Pareto (α_*, θ_*) |
| Normal (μ, a) - normal (Θ, v) | $\mu_* = \frac{n\bar{x} + (v/a)\mu}{n + v/a}$ $a_* = \frac{v}{n + v/a}$ | μ_* | Normal $(\mu_*, a_* + v)$ |

(1) In each conjugate pair, the posterior distribution belongs to the same class as the prior distribution where “*” indicates the updated parameters. In Bühlmann-Straub model, replace “n” with “m” and “ $\sum x_i$ ” with “ $\sum \sum x_{ij}$ ”.

(2) The Bayesian premium is the expected value of the predictive distribution.

(3) The pmf in MAS-II Tables, $p_k = \beta^k / (1 + \beta)^{k+1}$, is parameterized by $p_k = \theta(1 - \theta)^k$ where $\theta = 1/(1 + \beta)$.

(4) The pdf in MAS-II Tables, $f(x) = (1/\theta) \exp(-x/\theta)$, is parameterized by $f(x) = \lambda \exp(-x\lambda)$ where $\lambda = 1/\theta$.

DISCRETE PRIOR DISTRIBUTION

| | |
|---|---|
| Prior probability mass function (pmf) | $\Pr(\Theta = \theta_j) = \pi_j$ |
| Likelihood function of \mathbf{x} given $\Theta = \theta_j$ | $f(\mathbf{x} \theta_j) = \prod_{i=1}^n f(x_i \theta_j)$ |
| Joint distribution of \mathbf{X} and Θ | $f(\theta_j, \mathbf{x}) = f(\mathbf{x} \theta_j) \pi_j$ |
| Marginal distribution of $\mathbf{X} = \mathbf{x}$ | $f(\mathbf{x}) = \sum_{j=1}^J f(\mathbf{x}, \theta_j) = \sum_{j=1}^J f(\mathbf{x} \theta_j) \pi_j$ |
| Posterior pmf of $\Theta = \theta_j$ given \mathbf{x} | $f(\theta_j \mathbf{x}) = \frac{f(\mathbf{x}, \theta_j)}{f(\mathbf{x})} = \frac{f(\mathbf{x} \theta_j) \pi_j}{\sum_{k=1}^J f(\mathbf{x} \theta_k) \pi_k} = \pi_j^*$ |
| Predictive of X_{n+1} given \mathbf{x} | $f_{X_{n+1} \mathbf{X}}(x_{n+1} \mathbf{x}) = E_{\Theta \mathbf{x}}[f_{X_i \Theta}(x_i \theta)]$ |
| Bayesian premium | $\hat{\mu}_X(\mathbf{x}) = E(X_{n+1} \mathbf{x}) = \sum_{j=1}^J E(X_{n+1} \theta_j) \pi_j^*$ |

NON-PARAMETRIC MODEL IN BÜHLMANN-STRAUB'S CASE

| | |
|--|---|
| Sample mean of the i th risk group | $\bar{X}_i = \frac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} X_{ij}$ |
| Sample process variance of the i th risk group | $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} m_{ij} (X_{ij} - \bar{X}_i)^2$ |
| Unbiased estimate of μ_{PV} | $\hat{\mu}_{PV} = \frac{\sum_{i=1}^r (n_i - 1) s_i^2}{\sum_{i=1}^r (n_i - 1)}$ |
| Overall sample mean | $\hat{\mu}_X = \bar{X} = \frac{1}{m} \sum_{i=1}^r m_i \bar{X}_i$ |
| Unbiased estimator of σ_{HM}^2 | $\hat{\sigma}_{HM}^2 = \frac{\sum_{i=1}^r m_i (\bar{X}_i - \bar{X})^2 - (r - 1) \hat{\mu}_{PV}}{m - (\sum_{i=1}^r m_i^2)/m}$ |
| Credibility premium of the i th risk group | $\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i) \bar{X}$ where $\hat{Z}_i = \frac{m_i}{m_i + \hat{\mu}_{PV}/\hat{\sigma}_{HM}^2}$ |
| Credibility premium for the i th risk group for balancing the total loss with the predicted loss | $\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i) \hat{\mu}_X$ where $\hat{\mu}_X = \frac{\sum_{i=1}^r \hat{Z}_i \bar{X}_i}{\sum_{i=1}^r \hat{Z}_i}$ |

X_{ij} : The observation per unit of exposure during the j th time period for risk i

m_{ij} : The number of exposures during the j th time period for risk i

m_i : The total number of exposures in the experience for risk i

$$m = \sum_{i=1}^r m_i$$

n_i : The number of experience periods for risk i

$$n = \sum_{i=1}^r n_i$$

$\hat{\sigma}_{HM}^2$ may be negative in empirical applications. In this case, it may be set to zero, which implies that \hat{Z}_i will be zero for all risk groups.



NON-PARAMETRIC MODEL IN BÜHLMANN'S CASE

Sample mean of the i th risk group

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

Sample process variance of the i th risk group

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2}{n_i - 1}$$

Unbiased estimate of μ_{PV}

$$\tilde{\mu}_{PV} = \frac{\sum_{i=1}^r (n_i - 1) s_i^2}{\sum_{i=1}^r (n_i - 1)}$$

Overall sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_i} X_{ij}$$

Unbiased estimator of σ_{HM}^2

$$\tilde{\sigma}_{HM}^2 = \frac{\sum_{i=1}^r n_i (\bar{X}_i - \bar{X})^2 - (r - 1) \tilde{\mu}_{PV}}{n - (\sum_{i=1}^r n_i^2)/n}$$

Credibility premium of the i th risk group

$$\tilde{Z}_i \bar{X}_i + (1 - \tilde{Z}_i) \bar{X} \text{ where } \tilde{Z}_i = \frac{n_i}{n_i + \tilde{\mu}_{PV}/\tilde{\sigma}_{HM}^2}$$

Credibility premium for the i th risk group for

$$\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i) \hat{\mu}_X \text{ where } \hat{\mu}_X = \frac{\sum_{i=1}^r \hat{Z}_i \bar{X}_i}{\sum_{i=1}^r \hat{Z}_i}$$

balancing the total loss with the predicted loss

Note: The Bühlmann-Straub model reduces to Bühlmann model when $m_{ij} = 1$ for all i and j . In this case, we have $\sum_{j=1}^{n_i} m_{ij} = m_i = n_i$ and $n = \sum_{i=1}^r n_i$. $\tilde{\sigma}_{HM}^2$ may be negative in empirical applications. In this case, it maybe be set to zero, which implies that \tilde{Z}_i will be zero for all risk groups.

NON-PARAMETRIC MODEL IN BÜHLMANN'S CASE (SAME SAMPLE SIZE IN ALL RISK GROUPS)

Sample mean of the i th risk group

$$\bar{X}_i = \frac{1}{n_*} \sum_{j=1}^{n_*} X_{ij}$$

Sample process variance of the i th risk group

$$s_i^2 = \frac{\sum_{j=1}^{n_*} (X_{ij} - \bar{X}_i)^2}{n_* - 1}$$

Unbiased estimate of μ_{PV}

$$\tilde{\mu}_{PV} = \frac{\sum_{i=1}^r s_i^2}{r}$$

Overall sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^{n_*} X_{ij}$$

Unbiased estimator of σ_{HM}^2

$$\tilde{\sigma}_{HM}^2 = \frac{\sum_{i=1}^r (\bar{X}_i - \bar{X})^2}{r - 1} - \frac{\tilde{\mu}_{PV}}{n_*}$$

Credibility premium of the i th risk group

$$\tilde{Z}_i \bar{X}_i + (1 - \tilde{Z}_i) \bar{X} \text{ where } \tilde{Z}_i = \frac{n_*}{n_* + \tilde{\mu}_{PV}/\tilde{\sigma}_{HM}^2}$$

Note: $n_i = n_*$ and $n = rn_*$

PART B

Linear Mixed Models

OVERVIEW

General linear mixed model (two-level)

$$Y_{ti} = \underbrace{\beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_{ti}^{(3)} + \cdots + \beta_p X_{ti}^{(p)}}_{\text{fixed}} + \underbrace{u_{1i} Z_{ti}^{(1)} + \cdots + u_{qi} Z_{ti}^{(q)}}_{\text{random}} + \epsilon_{ti}$$

$t, t = 1, \dots, n_i$: Time indexes for the n_i longitudinal observations of the dependent variable for a given subject.

$i, i = 1, \dots, m$: The i -th subject.

X : The **fixed factors** or **fixed covariates**, i.e., factors that represent conditions chosen specifically to meet the objectives of the study.

Z : The **random factors** or **random covariates**, i.e., the factors that may have an affect on the study but are not the explicit factors being studied.

Depending on the purpose of the study, a variable could be either fixed or random.

β : Coefficients on the fixed factors, i.e., the **fixed effects**.

u : Coefficients on the random factors, i.e., the **random effects**.

ϵ_{ti} : The residual for the t -th occasion of the i -th subject.

General Matrix Specification

$$Y_i = \underbrace{X_i \beta}_{\text{fixed}} + \underbrace{Z_i u_i}_{\text{random}} + \epsilon_i$$

$$u_i \sim N(0, D)$$

$$\epsilon_i \sim N(0, R_i)$$

where

$$Y_i = \begin{bmatrix} Y_{1i} \\ Y_{2i} \\ \vdots \\ Y_{n_i i} \end{bmatrix}, \quad X_i = \begin{bmatrix} X_{1i}^{(1)} & X_{1i}^{(2)} & \cdots & X_{1i}^{(p)} \\ X_{2i}^{(1)} & X_{2i}^{(2)} & \cdots & X_{2i}^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n_i i}^{(1)} & X_{n_i i}^{(2)} & \cdots & X_{n_i i}^{(p)} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix},$$

$$Z_i = \begin{bmatrix} Z_{1i}^{(1)} & Z_{1i}^{(2)} & \cdots & Z_{1i}^{(q)} \\ Z_{2i}^{(1)} & Z_{2i}^{(2)} & \cdots & Z_{2i}^{(q)} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n_i i}^{(1)} & Z_{n_i i}^{(2)} & \cdots & Z_{n_i i}^{(q)} \end{bmatrix}, \quad u_i = \begin{bmatrix} u_{1i} \\ u_{2i} \\ \vdots \\ u_{qi} \end{bmatrix}, \quad \epsilon_i = \begin{bmatrix} \epsilon_{1i} \\ \epsilon_{2i} \\ \vdots \\ \epsilon_{n_i i} \end{bmatrix},$$

Y_i : The response variable vector with n_i rows, one for each observation for subject i .

X_i : An $n_i \times p$ matrix with a row for every observation and a column for every fixed factor.

Z_i : An $n_i \times q$ matrix with a row for every observation and a column for every random factor.

β : The fixed effect vector with p rows (one for every fixed factor).

u_i : The random effect vector with q rows (one for every random factor).

ϵ_i : The vector of residuals with n_i rows, one for each observation for subject i .

Variance-covariance matrix

The variance-covariance matrix for the random effects in u_i : D , also denoted as $\text{Var}(u_i)$. The main diagonal of D (the diagonal from the upper left corner to the lower right) represent the variances of each random effect. The off-diagonal entries are the random effect covariances, where the row and column determine which random effects.

The variance-covariance matrix for the residuals for subject i : $R_i = \text{Var}(\epsilon_i)$. The size of the matrix would be $n_i \times n_i$, because each observation would have its own residual.

The unique elements of the D and R matrices can be expressed in vectors θ_D and θ_R , respectively.

Common Covariance Structures for Residuals

Diagonal structure:

$$R_i = \text{Var}(\epsilon_i) = \sigma^2 \mathbf{I}_{n_i} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Parameter: $\theta_R = (\sigma^2)$

Compound symmetry structure:

$$R_i = \text{Var}(\epsilon_i) = \begin{bmatrix} \sigma^2 + \sigma_1 & \sigma_1 & \cdots & \sigma_1 \\ \sigma_1 & \sigma^2 + \sigma_1 & \cdots & \sigma_1 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_1 & \sigma_1 & \cdots & \sigma^2 + \sigma_1 \end{bmatrix}$$

Parameters: $\theta_R = (\sigma^2, \sigma_1)$

AR(1) structure:

$$R_i = \text{AR}(1) = \text{Var}(\epsilon_i) = \begin{bmatrix} \sigma^2 & \sigma^2 \rho & \cdots & \sigma^2 \rho^{n_i-1} \\ \sigma^2 \rho & \sigma^2 & \cdots & \sigma^2 \rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2 \rho^{n_i-1} & \sigma^2 \rho^{n_i-2} & \cdots & \sigma^2 \end{bmatrix}$$

Parameters: $\theta_R = (\sigma^2, \rho)$

Specification of the Marginal Model

$$Y_i = X_i \beta + \epsilon_i^*, \quad \epsilon_i^* \sim N(0, V_i^*)$$

Implied Marginal Model

$Y_i = X_i\beta + Z_iu_i + \epsilon_i$, where $u_i \sim N(0, D)$ and $\epsilon_i \sim N(0, R_i)$ can be reformulated as:

$$Y_i = X_i\beta + \epsilon_i^*$$

$$\epsilon_i^* \sim N(0, V_i)$$

$$V_i = Z_iDZ_i' + R_i$$

The covariance parameters θ or θ_V are the same as the parameters for θ_D and θ_R . For example, if θ_D followed the diagonal structure with parameter σ_D^2 , and θ_R followed the compound symmetry structure with parameters σ_R^2 and σ_1 , then $\theta_V = (\sigma_D^2, \sigma_R^2, \sigma_1)$.

Maximum Likelihood (ML) Estimation

Model for subject i :

$$Y_i = X_i\beta + \epsilon_i^*, \quad \epsilon_i^* \sim N(0, V_i)$$

$$V_i = Z_iDZ_i' + R_i$$

The joint log-likelihood function of (y_1, \dots, y_m) :

$$l(\beta, \theta) = -(n/2) \log(2\pi) - (1/2) \sum \log(\det(V_i)) - (1/2) \sum (y_i - X_i\beta)'(V_i)^{-1}(y_i - X_i\beta), \quad n = \sum_1^m n_i$$

The **maximum likelihood estimates** (MLEs) of the parameters are the values of the arguments that maximize the likelihood function. The ML estimation is a two-step procedure.

- The first step is to estimate the fixed-effect parameters β using the **generalized least squares** (GLS) assuming the covariance parameters θ are known.
- The second step is to obtain the estimates of θ by optimizing the profile log-likelihood function. After obtaining the estimates of θ , we can then calculate the estimates of β .

The estimator of β has the desirable statistical property of being the **best linear unbiased estimator** (BLUE) of β .

Restricted Maximum Likelihood (REML) Estimation

REML estimation maximizes the REML log-likelihood function:

$$l_{REML}(\beta, \theta) = -\left(\frac{n-p}{2}\right) \log(2\pi) - (1/2) \sum \log(\det(V_i)) - (1/2) \sum (y_i - X_i\beta)'(V_i)^{-1}(y_i - X_i\beta) - (1/2) \sum \log(\det(X_i'V_i^{-1}X_i)), \quad (n = \sum_1^m n_i)$$

The REML estimates of the covariance parameters (θ) are unbiased, whereas the ML estimates are biased. Both the ML and the REML estimates of the diagonal elements of $var(\beta)$ are downward biased.

Best Linear Unbiased Estimator (BLUE)

If θ is known, the BLUE of β is:

$$\hat{\beta} = \left(\sum_i X_i'V_i^{-1}X_i \right)^{-1} \sum_i X_i'V_i^{-1}y_i$$

If θ is unknown, estimate θ and then calculate \hat{D} , \hat{R}_i , $\hat{V}_i = Z_i\hat{D}Z_i' + \hat{R}_i$, and:

$$\hat{\beta} = \left(\sum_i X_i'\hat{V}_i^{-1}X_i \right)^{-1} \sum_i X_i'\hat{V}_i^{-1}y_i$$

Likelihood Ratio Tests (LRT)

Denote L_{nested} the value of the likelihood function evaluated at the ML or REML estimates of the parameters in the nested model M_0 (null hypothesis) and $L_{\text{reference}}$ the value in the reference model M_A (alternative hypothesis). The likelihood ratio test (LRT) statistics, or simply the likelihood ratio, is defined as:

$$T = -2 \ln \left(\frac{L_{\text{nested}}}{L_{\text{reference}}} \right).$$

T asymptotically follows a χ^2 distribution with degrees of freedom equal to the number of parameters in M_A subtracted by the number of parameters in M_0 .

When the LRT is performed on covariance parameters, with the null hypothesis lying on the boundary of the parameter space, the test statistics has an asymptotic null distribution that is a mixture of two χ^2 distributions.



t -test for testing single fixed-effect parameter

When testing a single fixed-effect parameter:

$$H_0 : \beta = 0 \quad (\text{nested model})$$

$$H_A : \beta \neq 0 \quad (\text{reference model})$$

The t -statistic is $T = \hat{\beta}/\text{se}(\hat{\beta})$.

T **does not** follow an exact t distribution in the context of an LMM. Instead, we use the standard normal distribution when the sample is large, which gives us a z -statistic $z = \hat{\beta}/\text{se}(\hat{\beta})$ and p value $p\text{-value} = (2) \Pr(Z > |z|)$.

Omnibus Wald test for testing multiple fixed-effect parameters

The hypothesis:

$$H_0 : L\beta = \mathbf{0} \quad (\text{nested model})$$

$$H_A : L\beta \neq \mathbf{0} \quad (\text{reference model})$$

where β is a vector of p unknown fixed-effect parameters and L is a known matrix.

The test statistic is $W = \hat{\beta}' L' \left(L \left(\sum_i \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} L' \right)^{-1} L \hat{\beta}$, which asymptotically follows a χ^2 with degrees of freedom equal to the rank of the L matrix.

F -test for testing multiple fixed-effect parameters

The null hypothesis:

$$H_0 : \beta = \mathbf{0} \quad (\text{nested model})$$

$$H_A : \beta \neq \mathbf{0} \quad (\text{reference model})$$

The F -statistic is: $F = \frac{W}{\text{rank}(\mathbf{L})}$, which follows an approximate F distribution, with numerator degrees of freedom equal to the rank of \mathbf{L} , and an approximated denominator degrees of freedom equal to $n - p$ where n is the sample size and p is the total number of fixed-effect parameter estimated.

Best Linear Unbiased Predictors (BLUPs)

The empirical BLUPs (EBLUPs) of u_i is:

$$\hat{u}_i = E(u_i | Y_i = y_i) = \hat{D} Z_i' \hat{V}_i^{-1} (y_i - X_i \hat{\beta})$$

where:

$$Y_i = X_i \beta + Z_i \mathbf{u}_i + \epsilon_i$$

$$\mathbf{u}_i \sim N(0, D)$$

$$\epsilon_i \sim N(0, \sigma^2)$$

$$V_i = Z_i D Z_i' + \sigma^2.$$

EBLUPs are also known as **shrinkage estimators** because they tend to be closer to 0 than the estimated effects if the random factors were treated as fixed effects.

BLUP mean from the study note
(Additional Notes on Shrinkage Means)

$$u_i = \alpha_i \times \mu + (1 - \alpha_i) \times \mu_i,$$

where

- u_i is the shrinkage mean for level i of the random factor,
- α_i is a weighting factor for level i , calculated by $\sigma_i^2 / (\sigma_{\text{random factor}}^2 + \sigma_i^2)$,
- σ_i^2 is the variance for level i of the random factor, calculated by $\sigma_{\text{error}}^2 / n_i$,
- $\sigma_{\text{random factor}}^2$ is the variance of the random effects associated with the random factor,
- μ is the overall mean of the response values,
- μ_i is the mean of the response values for level i of the random factor.

The formula above assumes no other fixed factors than the intercept.



Intraclass Correlation Coefficients

The ICC is defined as the proportion of the total random variation in the response that is due to the variance of the random effects. For example, given the model:

$$Y_i = X_i \beta + u_i + \epsilon_i$$

$$u_j \sim N(0, \sigma_u^2), \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

The ICC for the random effect u_i is:

$$\text{ICC}_{u_i} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}$$

TWO-LEVEL MODELS FOR CLUSTERED DATA

Best model for Rat Pup data

Model 3.3

$$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + u_j + \epsilon_{ij}$$

$$\text{High/Low Treatment:} \quad \epsilon_{ij} \sim N(0, \sigma_{h/l}^2)$$

$$\text{Control Treatment:} \quad \epsilon_{ij} \sim N(0, \sigma_c^2)$$

Hypothesis 3.1: Test whether the random effects, u_j , associated with the litter-specific intercepts can be omitted from **Model 3.1**.

Model 3.1

$$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + \beta_5 X_{ij}^{(5)} + \beta_6 X_{ij}^{(6)} + u_j + \epsilon_{ij},$$

$$u_j \sim N(0, \sigma_{l.e.}^2), \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

Model 3.1A

$$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + \beta_5 X_{ij}^{(5)} + \beta_6 X_{ij}^{(6)} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2)$$

The null and alternative hypotheses are:

$$H_0 : \sigma_{l.e.} = 0 \quad (\text{Model 3.1A})$$

$$H_A : \sigma_{l.e.} > 0 \quad (\text{Model 3.1})$$

Test statistic: $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

$$p\text{-value} = (0.5) \Pr(\chi_0^2 > T) + (0.5) \Pr(\chi_1^2 > T) = (0.5) \Pr(\chi_1^2 > T).$$

Decision: The p -value is less than 1%. Therefore, we have strong evidence to reject the null hypothesis, and retain the litter-specific random effects (**Model 3.1**).

Hypothesis 3.2: Test whether the variance of ϵ_{ij} is specific to treatment effects in **Model 3.1**.

Model 3.2A: Same as **Model 3.1** except

$$\epsilon_{ij} \sim N(0, \sigma_h^2) \text{ if high-dose treatment,}$$

$$\epsilon_{ij} \sim N(0, \sigma_l^2) \text{ if low-dose treatment,}$$

$$\epsilon_{ij} \sim N(0, \sigma_c^2) \text{ if control treatment.}$$

The null and alternative hypotheses are:

$$H_0 : \sigma_h^2 = \sigma_l^2 = \sigma_c^2 = \sigma^2 \quad (\text{Model 3.1})$$

$$H_A : \text{At least one pair of residual variances is not equal to each other} \quad (\text{Model 3.2A})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

$$\text{The } p\text{-value is } \Pr(\chi_2^2 > T).$$

Decision: The p -value is less than $< .0001$. We have strong evidence to reject the null hypothesis and choose the model with heterogeneous variance (**Model 3.2A**).

Hypothesis 3.3: Test whether $\sigma_h = \sigma_l$

Model 3.2A: Same as **Model 3.1** except

$$\begin{array}{ll} \text{High Treatment} & \epsilon_{ij} \sim N(0, \sigma_h^2) \\ \text{Low Treatment} & \epsilon_{ij} \sim N(0, \sigma_l^2) \\ \text{Control Treatment} & \epsilon_{ij} \sim N(0, \sigma_c^2) \end{array}$$

Model 3.2B: Same as **Model 3.1** except

$$\begin{array}{ll} \text{High/Low Treatment:} & \epsilon_{ij} \sim N(0, \sigma_{h/l}^2) \\ \text{Control Treatment:} & \epsilon_{ij} \sim N(0, \sigma_c^2) \end{array}$$

The null and alternative hypotheses are:

$$H_0 : \sigma_h = \sigma_l \quad (\text{Model 3.2B})$$

$$H_A : \sigma_h \neq \sigma_l \quad (\text{Model 3.2A})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_1^2 > T)$

Decision: The p -value is greater than 5%. We fail to reject the null hypothesis. We should select **Model 3.2B** under the null hypothesis.

Hypothesis 3.4: Test whether the residual variance for the combined high-/low treatment group is equal to the residual variance for the control group.

The null and alternative hypotheses are

$$H_0 : \sigma_{h/l}^2 = \sigma_c^2 = \sigma^2 \quad (\text{Model 3.1})$$

$$H_A : \sigma_{h/l}^2 \neq \sigma_c^2 \quad (\text{Model 3.2B})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_1^2 > T)$.

Decision: The p -value is less than 0.0001. We should reject the null hypothesis and choose **Model 3.2B** under the alternative hypothesis as our preferred model at this stage.

Hypothesis 3.5: The fixed effects associated with the treatment by sex interaction are equal to zero in **Model 3.2B**.

Model 3.3: Same as **Model 3.2B** except $\beta_5 = \beta_6 = 0$

Model 3.2B: Same as **Model 3.1** except

$$\begin{array}{ll} \text{High/Low Treatment:} & \epsilon_{ij} \sim N(0, \sigma_{h/l}^2) \\ \text{Control Treatment:} & \epsilon_{ij} \sim N(0, \sigma_c^2) \end{array}$$

The null and alternative hypotheses are

$$H_0 : \beta_5 = \beta_6 = 0 \quad (\text{Model 3.3})$$

$$H_A : \beta_5 \neq 0 \text{ or } \beta_6 \neq 0 \quad (\text{Model 3.2B})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_2^2 > T)$.

Decision: The p -value is 0.7255 and we do NOT reject the null hypothesis. We choose the nested model **Model 3.3** under the null hypothesis as our preferred model.



Hypothesis 3.6: The fixed effects associated with the treatment are equal to zero in **Model 3.3**.

Model 3.3A: Same as **Model 3.3** except $\beta_1 = \beta_2 = 0$.

Model 3.3: Same as **Model 3.2B** except $\beta_5 = \beta_6 = 0$

The null and alternative hypotheses are

$$H_0 : \beta_1 = \beta_2 = 0 \quad (\text{Model 3.3A})$$

$$H_A : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \quad (\text{Model 3.3})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_2^2 > T)$.

Decision: The p -value is 0.0001. We reject the null hypothesis and choose **Model 3.3** under the alternative hypothesis as **our final model**.

Hierarchical Specification

Full model:

$$Y_{ij} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + u_j + \epsilon_{ij},$$

The **Level 1 Model** reflects the variation between pups within a given litter:

$$Y_{ij} = b_{0j} + \beta_3 X_{ij}^{(3)} + \epsilon_{ij}.$$

The **Level 2 Model** reflects the variation between litters:

$$b_{0j} = \beta_0 + \beta_1 X_j^{(1)} + \beta_2 X_j^{(2)} + \beta_4 X_j^{(4)} + u_j$$

Intraclass Correlation Coefficients

Let $\epsilon_{ij} \sim N(0, \sigma^2)$ in **Model 3.1**:

$$\text{ICC}_{litter} = \frac{\sigma_{l.e.}^2}{\sigma_{l.e.}^2 + \sigma^2}$$

Y_{ij} : Birth weight observation on rat pup i within the j -th litter

$X_j^{(2)}$: Indicator variable for the low-dose treatment

$X_{ij}^{(3)}$: The indicator for female rat pups

$X_{ij}^{(4)}$: The litter size

u_j : The random effect associated with the intercept for litter j

ϵ_{ij} : Residuals.

THREE-LEVEL MODELS FOR CLUSTERED DATA

The best model to fit classroom data

Model 4.2

$$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + u_k + u_{j|k} + \epsilon_{ijk}$$

where $u_k \sim N(0, \sigma_{i:s}^2)$, $u_{j|k} \sim N(0, \sigma_{i:c}^2)$, $\epsilon_{ijk} \sim N(0, \sigma^2)$. u_k , $u_{j|k}$, and ϵ_{ijk} are all mutually independent.

Hypothesis 4.1: Test whether the random effects associated with the intercepts for classroom nested within schools can be omitted

Model 4.1: $Y_{ijk} = \beta_0 + u_k + u_{j|k} + \epsilon_{ijk}$

Model 4.1A: $Y_{ijk} = \beta_0 + u_k + \epsilon_{ijk}$,

The null and alternative hypotheses are:

$$H_0 : \sigma_{i:c}^2 = 0 \quad (\text{Model 4.1A})$$

$$H_A : \sigma_{i:c}^2 > 0 \quad (\text{Model 4.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $(0.5) \Pr(\chi_0^2 > T) + (0.5) \Pr(\chi_1^2 > T) = (0.5) \Pr(\chi_1^2 > T)$.

Decision: The p -value is less than 1% which shows strong evidence to reject the null hypothesis. Therefore, we choose the model under the alternative hypothesis (**Model 4.1**) which retains the nested random classroom effects.

Hypothesis 4.2: Test whether the fixed effects associated with the four student-level covariates (mathkind, sex, minority, and ses) should be added to **Model 4.1**.

$$\textbf{Model 4.1: } Y_{ijk} = \beta_0 + u_k + u_{j|k} + \epsilon_{ijk}$$

Model 4.2:

$$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + u_k + u_{j|k} + \epsilon_{ijk}$$

The null and alternative hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad (\textbf{Model 4.1})$$

$$H_A : \text{At least one fixed effect is not equal to zero} \quad (\textbf{Model 4.2})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_4^2 > T)$

Decision: The p -value is less than 0.5% and we conclude that at least one of the fixed effects associated with the **Level 1** covariates is significant. Therefore, we proceed with **Model 4.2** as our preferred model.

Hypothesis 4.3: The fixed effect associated with the classroom-level covariate yearstea ($X_{jk}^{(5)}$) should be retained in **Model 4.3**.

Model 4.3:

$$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + \beta_5 X_{jk}^{(5)} + \beta_6 X_{jk}^{(6)} + \beta_7 X_{jk}^{(7)} + u_k + u_{j|k} + \epsilon_{ijk}$$

The null and alternative hypotheses are

$$H_0 : \beta_5 = 0 \quad \text{vs.} \quad H_A : \beta_5 \neq 0$$

The test statistic is $t\text{-value} = \hat{\beta}_5 / se(\hat{\beta}_5)$

The p -value is $2 \Pr(t \geq |t\text{-value}|)$.

Hypothesis 4.4: The fixed effect associated with the classroom-level covariate mathprep ($X_{jk}^{(6)}$) should be retained in **Model 4.3**.

The null and alternative hypotheses are $H_0 : \beta_6 = 0$ vs. $H_A : \beta_6 \neq 0$

The test statistic is $t\text{-value} = \hat{\beta}_6 / se(\hat{\beta}_6)$

The p -value is $2 \Pr(t \geq |t\text{-value}|)$

Hypothesis 4.5: The fixed effect associated with the classroom-level covariate mathknow ($X_{jk}^{(7)}$) should be retained in **Model 4.3**.

The null and alternative hypotheses are $H_0 : \beta_7 = 0$ vs. $H_A : \beta_7 \neq 0$.

The test statistic is $t\text{-value} = \hat{\beta}_7 / se(\hat{\beta}_7)$

The p -value is $2 \Pr(t \geq |t\text{-value}|)$

Hypothesis 4.6: Test whether the fixed effect associated with the school-level covariate housepov (β_8) should be added to **Model 4.2**.

Models:

$$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + u_k + u_{j|k} + \epsilon_{ijk} \quad (\textbf{Model 4.2})$$

$$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + \beta_8 X_k^{(8)} + u_k + u_{j|k} + \epsilon_{ijk} \quad (\textbf{Model 4.4})$$

The null and alternative hypotheses are:

$$H_0 : \beta_8 = 0 \quad (\textbf{Model 4.2})$$

$$H_A : \beta_8 \neq 0 \quad (\textbf{Model 4.4})$$

The test statistic is $t\text{-value} = \hat{\beta}_8 / se(\hat{\beta}_8)$

The p -value is $2 \Pr(t \geq |t\text{-value}|)$

| | |
|---|---|
| Hierarchical Model | <div>Full model:$Y_{ijk} = \beta_0 + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + u_k + u_{j k} + \epsilon_{ijk}$</div> <div>Level 1 Model (Student)$Y_{ijk} = b_{0j k} + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + \epsilon_{ijk}$<p>where $b_{0j k}$ is the unobserved classroom-specific intercepts, $X_{ijk}^{(1)}$ to $X_{ijk}^{(4)}$ are the student level covariates, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.</p></div> <div>Level 2 Model (Classroom)$b_{0j k} = b_{0k} + u_{j k}$<p>where b_{0k} is the unobserved intercept, specific to the k-th school, and $u_{j k}$ is random effect associated with classroom j within school k.</p></div> <div>Level 3 Model (School)$b_{0k} = \beta_0 + u_k$<p>where $u_k \sim N(0, \sigma_{i:s}^2)$.</p></div> |
| Intraclass Correlation Coefficients | <div>The school-level ICC:$ICC_{school} = \frac{\sigma_{i:s}^2}{\sigma_{i:s}^2 + \sigma_{i:c}^2 + \sigma^2}$</div> <div>The classroom-level ICC:$ICC_{classroom} = \frac{\sigma_{i:s}^2 + \sigma_{i:c}^2}{\sigma_{i:s}^2 + \sigma_{i:c}^2 + \sigma^2}$</div> |
| <div>Y_{ijk}: the dependent variable mathgain</div> <div>Level 1 covariates (Student):$X_{ijk}^{(1)}$ (mathkind): Student's math score in the kindergarten year$X_{ijk}^{(2)}$ (sex): Indicator variable (0 = boy, 1 = girl)$X_{ijk}^{(3)}$ (minority): Indicator variable (0 = non-minority, 1 = minority)$X_{ijk}^{(4)}$ (ses): Student socioeconomic status.</div> <div>Level 2 covariates (Classroom):$X_{jk}^{(5)}$ (yearstea): First-grade teacher's years of teaching experience$X_{jk}^{(6)}$ (mathprep): First-grade teacher's math preparations$X_{jk}^{(7)}$ (mahtknow): First-grade teacher's math content knowledge</div> <div>Level 3 covariates (School):$X_k^{(8)}$ (housepov): Percentage of households in the neighborhood of the school below the poverty level</div> <div>u_k: the random effects associated with the intercept for school k, $u_{j k}$: the random effect associated with the intercept for classroom j within school k, and ϵ_{ijk}: the residuals, for the ith student, in the jth classroom, within the kth school.</div> | |



MODELS FOR REPEATED-MEASURES DATA

The best model to fit Rat Brain data

Model 5.2

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_{ti}^{(3)} + \beta_4 X_{ti}^{(4)} + \beta_5 X_{ti}^{(5)} + u_{0i} + u_{3i} X_{ti}^{(3)} + \epsilon_{ti}$$

Hypothesis 5.1: Test whether the random treatment effect associated with animal i , u_{3i} , can be omitted from **Model 5.2**

Models:

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_{ti}^{(3)} + \beta_4 X_{ti}^{(4)} + \beta_5 X_{ti}^{(5)} + u_{0i} + \epsilon_{ti} \quad (\text{Model 5.1})$$

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_{ti}^{(3)} + \beta_4 X_{ti}^{(4)} + \beta_5 X_{ti}^{(5)} + u_{0i} + u_{3i} X_{ti}^{(3)} + \epsilon_{ti} \quad (\text{Model 5.2})$$

The null and alternative hypotheses are:

$$H_0 : D = \begin{bmatrix} \sigma_{in}^2 & 0 \\ 0 & 0 \end{bmatrix} \quad (\text{Model 5.1})$$

$$H_A : D = \begin{bmatrix} \sigma_{in}^2 & \sigma_{i,t} \\ \sigma_{i,t} & \sigma_{tr}^2 \end{bmatrix} \quad (\text{Model 5.2})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $(0.5) \Pr(\chi_1^2 > T) + (0.5) \Pr(\chi_2^2 > T)$

Decision: The p -value for testing Hypothesis 5.1 is less than 1%. We have strong evidence to reject the null hypothesis and select the model under the alternative hypothesis **Model 5.2** which is our preferred model at this stage.

Hypothesis 5.2: Test whether residual variances should differ for each level of treatment

In **Model 5.3**, we allow the residual variances to differ for each level of treatment, by including separate residual variances (σ_b^2 and σ_c^2) for the basal and carbachol treatments.

The null and alternative hypotheses are

$$H_0 : \sigma_b^2 = \sigma_c^2 \quad (\text{Model 5.2})$$

$$H_A : \sigma_b^2 \neq \sigma_c^2 \quad (\text{Model 5.3})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $p\text{-value} = \Pr(\chi_1^2 > T)$.

The p -value is 0.6965 showing lack of evidence to reject the null hypothesis. We should choose the model under the null hypothesis, **Model 5.2**, and keep the model as our preferred model at this stage.

Hypothesis 5.3: The fixed effects associated with the region by treatment interaction can be omitted from **Model 5.2**

The null and alternative hypotheses are $H_0 : \beta_4 = \beta_5 = 0$ vs. $H_A : \beta_4 \neq 0$ or $\beta_5 \neq 0$.

We test Hypothesis 5.3 using Type III F -test in **R**, where the test statistic follows a F distribution with degrees of freedom (2,20).

Akaike Information Criterion

$$\text{AIC} = (-2) \times \log\text{Lik} + 2 \times p$$

p is the number of parameters estimated in the model.

Bayesian Information Criterion

$$\text{BIC} = (-2) \times \log\text{Lik} + \log(n) \times p$$

n is the number of observation in the modeled dataset.

Y_{ti} : the dependent variable activate

$X_{ti}^{(1)} = \text{REGION1}$ and $X_{ti}^{(2)} = \text{REGION2}$: indicator variables

$X_{ti}^{(3)} = \text{TREATMENT}$: indicator variable, 1 for Carbachol and 0 for Basal treatment

u_{0i} : the random intercept

u_{3i} : the random treatment effect associated with animal i

ϵ_{ti} : the residuals



RANDOM COEFFICIENT MODELS FOR LONGITUDINAL DATA

The best model to fit autism data

Model 6.3

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_i^{(3)} + \beta_4 X_i^{(4)} + \beta_5 X_{ti}^{(1)} X_{ti}^{(3)} + \beta_6 X_{ti}^{(1)} X_{ti}^{(4)} + u_{1i} X_{ti}^{(1)} + u_{2i} X_{ti}^{(2)} + \epsilon_{ti}$$

Hypothesis 6.1: Test whether the random effects (u_{1i}) associated with the quadratic effect of age can be omitted from the model **Model 6.2**

Model 6.2

$$Y_{ti} = \beta_0 + \beta_1 X_{ti}^{(1)} + \beta_2 X_{ti}^{(2)} + \beta_3 X_i^{(3)} + \beta_4 X_i^{(4)} + \beta_5 X_{ti}^{(5)} + \beta_6 X_{ti}^{(6)} + \beta_7 X_{ti}^{(7)} + \beta_8 X_{ti}^{(8)} + u_{1i} X_{ti}^{(1)} + u_{2i} X_{ti}^{(2)} + \epsilon_{ti}$$

The null and alternative hypotheses are

$$H_0 : D = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & 0 \end{bmatrix} \quad (\text{Model 6.2A})$$

$$H_A : D = \begin{bmatrix} \sigma_a^2 & \rho_{a,as} \sigma_a \sigma_{as} \\ \rho_{a,as} \sigma_a \sigma_{as} & \sigma_{as}^2 \end{bmatrix} \quad (\text{Model 6.2})$$

where D is the variance-covariance matrix of u_{1i} and u_{2i} .

The **test statistic** is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The **p value** is $(0.5) \Pr(\chi_1^2 > T) + (0.5) \Pr(\chi_2^2 > T)$.

Decision: The p -value for testing Hypothesis 6.1 is less than 1%. We have strong evidence to reject the null hypothesis and select the model under the alternative hypothesis **Model 6.2**. The random coefficients associated with the quadratic, as well as linear effects of age should be included in **Model 6.2**.

Hypothesis 6.2: Test whether the fixed effects associated with the age-squared \times sicdegp interaction are equal to zero in **Model 6.2**.

The the null and alternative hypotheses are

$$H_0 : \beta_7 = \beta_8 = 0 \quad (\text{Model 6.3})$$

$$H_A : \beta_7 \neq \text{ or } \beta_8 \neq 0 \quad (\text{Model 6.2})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The **p value** is $\Pr(\chi_2^2 > T)$.

Decision: The p -value is 0.3926 which shows no evidence to reject the null hypothesis. We exclude the fixed effects associated with the age-squared \times sicdegp interaction and choose **Model 6.3**.

Hypothesis 6.3: The fixed effects associated with the age \times sicdegp interaction are equal to zero in **Model 6.3**.

The null and alternative hypotheses are

$$H_0 : \beta_5 = \beta_6 = 0 \quad (\text{Model 6.4})$$

$$H_1 : \beta_5 \neq \text{ or } \beta_6 \neq 0 \quad (\text{Model 6.3})$$

We test Hypothesis 6.3 using Type I F -test, where the test statistic follows a F distribution with degrees of freedom (2, 448).

Decision: The p -value is less than 0.0001 showing strong evidence to reject the null hypothesis. We include the fixed effects associated with the age \times sicdegp interaction and choose **Model 6.3** as our final model.

Y_{ti} : the dependent variable activate

The $X_i^{(1)}$: (age.2) variable represents the value of age minus 2.

The $X_i^{(2)}$: (age.2sq) variable represents age.2 squared.

The $X_i^{(3)}$: sicdegp1_{*i*} = 1 if sicdegp in level 1, 0 otherwise.

The $X_i^{(4)}$: sicdegp2_{*i*} = 1 if sicdegp in level 2, 0 otherwise.

MODELS FOR CLUSTERED LONGITUDINAL DATA

The best model to fit veneer data

Model 7.3

$$Y_{tij} = \beta_0 + \beta_1 X_t^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + u_{0j} + u_{1j} X_t^{(1)} + u_{0i|j} + \epsilon_{tij}$$

Hypothesis 7.1: The nested random effects $u_{0i|j}$ associated with teeth within the same patient can be omitted from **Model 7.1**.

Model 7.1:

$$Y_{tij} = \beta_0 + \beta_1 X_t^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + \beta_5 X_{tij}^{(5)} + \beta_6 X_{tij}^{(6)} + \beta_7 X_{tj}^{(7)} + u_{0j} + u_{1j} X_t^{(1)} + u_{0i|j} + \epsilon_{tij}$$

The null and alternative hypotheses are

$$H_0 : \sigma_{t|p}^2 = 0 \quad (\text{Model 7.1A})$$

$$H_A : \sigma_{t|p}^2 > 0 \quad (\text{Model 7.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $(0.5) \Pr(\chi_0^2 > T) + (0.5) \Pr(\chi_1^2 > T)$.

Decision: The p -value is less than 1%, showing strong evidence to reject the null hypothesis. Therefore, we choose the model under the alternative hypothesis (**Model 7.1**) which retains the rested random tooth effects.

Hypothesis 7.2: The variance of the residuals is constant (homogeneous) across the time points in **Model 7.2C**.

Model 7.2C is similar to **Model 7.1** except that

$$\epsilon_{tij} \sim N(0, \sigma_t^2), \quad t = 1, 2.$$

The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad (\text{Model 7.1})$$

$$H_A : \sigma_1^2 \neq \sigma_2^2 \quad (\text{Model 7.2C})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $\Pr(\chi_1^2 > T)$.

Decision: The p -value is 0.3289. We do NOT reject the null hypothesis at $\alpha = 1\%$. Therefore, we choose the model under the null hypothesis **Model 7.1** (homogeneous variance).

Hypothesis 7.3: Test whether the fixed effects associated with the two-way interactions between time and the patient- and tooth-level covariates can be omitted from **Model 7.1**.

Model 7.1:

$$Y_{tij} = \beta_0 + \beta_1 X_t^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + \beta_5 X_{tij}^{(5)} + \beta_6 X_{tij}^{(6)} + \beta_7 X_{tj}^{(7)} + u_{0j} + u_{1j} X_t^{(1)} + u_{0i|j} + \epsilon_{tij}$$

The null and alternative hypotheses are

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \quad (\text{Model 7.3})$$

$$H_A : \beta_5 \neq 0, \text{ or } \beta_6 \neq 0, \text{ or } \beta_7 \neq 0 \quad (\text{Model 7.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $p\text{-value} = \Pr(\chi_3^2 > T)$.

Decision: The p -value is 0.606 for testing Hypothesis 7.3. We DO NOT reject the null hypothesis and we choose **Model 7.3** as our final model.

Model 7.1A:

$$Y_{tij} = \beta_0 + \beta_1 X_t^{(1)} + \beta_2 X_{ij}^{(2)} + \beta_3 X_{ij}^{(3)} + \beta_4 X_j^{(4)} + \beta_5 X_{tij}^{(5)} + \beta_6 X_{tij}^{(6)} + \beta_7 X_{tj}^{(7)} + u_{0j} + u_{1j} X_t^{(1)} + \epsilon_{tij}$$

The null and alternative hypotheses are

$$H_0 : \sigma_{t|p}^2 = 0 \quad (\text{Model 7.1A})$$

$$H_A : \sigma_{t|p}^2 > 0 \quad (\text{Model 7.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $p\text{-value} = (0.5) \Pr(\chi_0^2 > T) + (0.5) \Pr(\chi_1^2 > T) = (0.5) \Pr(\chi_1^2 > T)$.

Test whether the nested random effects $u_{0i|j}$ associated with teeth within the same patient can be omitted from **Model 7.1**



The variance of the residuals is constant (homogeneous) across the time points in **Model 7.2C**

Model 7.2c is similar to **Model 7.1** except that

$$\epsilon_{tij} \sim N(0, \sigma_t^2), \quad t = 1, 2.$$

The null and alternative hypotheses are

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad (\text{Model 7.1})$$

$$H_A : \sigma_1^2 \neq \sigma_2^2 \quad (\text{Model 7.2C})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $p\text{-value} = \Pr(\chi_1^2 > T)$.

Hierarchical Model

Full model:

$$\text{GCF}_{tij} = \beta_0 + \beta_1 \text{TIME}_t + \beta_2 \text{BAS_GCP}_{ij} + \beta_3 \text{CDA}_{ij} + \beta_4 \text{AGE}_j + u_{0j} + u_{1j} \text{TIME}_t + u_{0i|j} + \epsilon_{tij}$$

Level 1 Model (Time):

$$\text{GCF}_{tij} = b_{0i|j} + b_{1j} \text{TIME}_t + \epsilon_{tij}$$

Level 2 Model (Tooth)

$$b_{0i|j} = b_{0j} + \beta_2 \text{BAS_GCP}_{ij} + \beta_3 \text{CDA}_{ij} + u_{0i|j}$$

Level 3 Model (Patient)

At the Patient level indexed by j :

$$b_{0j} = \beta_0 + \beta_4 \text{AGE}_j + u_{0j}$$

$$b_{1j} = \beta_1 + u_{1j}$$

Y_{tij} : dependent variable gcf_{tij}

$X_t^{(1)} = \text{time}_t$

$X_{ij}^{(2)} = \text{base_gcf}_{ij}$

$X_{ij}^{(3)} = \text{cda}_{ij}$

$X_j^{(4)} = \text{age}_j$ at visit t on tooth i nested within patient j

u_{0j} : the patient-specific random **intercept**

u_{1j} : the patient-specific random **coefficient** associated with the time slope

$u_{0i|j}$: the random effect associated with a tooth nested within a patient

MODELS FOR DATA WITH CROSSED RANDOM FACTORS

The best model to fit sat data

Model 8.1

$$Y_{tij} = \beta_0 + \beta_1 X_{tij} + u_i + v_j + \epsilon_{tij}$$

Hypothesis 8.1: Test whether the random effects u_i associated with the students can be omitted from **Model 8.1**

Model 8.2

$$Y_{tij} = \beta_0 + \beta_1 \times X_{tij} + v_j + \epsilon_{tij}$$

The null and alternative hypotheses are:

$$H_0 : \sigma_{st}^2 = 0 \quad (\text{Model 8.2})$$

$$H_A : \sigma_{st}^2 > 0 \quad (\text{Model 8.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $(0.5) \Pr(\chi_0^2 > T) + (0.5) \Pr(\chi_1^2 > T) = (0.5) \Pr(\chi_1^2 > T)$.

Decision: The p -value is less than 1% for testing Hypothesis 8.1, which shows strong evidence to reject the null hypothesis. Therefore, we should retain the random student effects and choice **Model 8.1**.



Hypothesis 8.2: Test whether the random effects v_j associated with the teachers can be omitted from **Model 8.1**

Model 8.3:

$$Y_{tij} = \beta_0 + \beta_1 \times X_{tij} + u_i + \epsilon_{tij}$$

The null and alternative hypotheses are

$$H_0 : \sigma_{te}^2 = 0 \quad (\text{Model 8.3})$$

$$H_A : \sigma_{te}^2 > 0 \quad (\text{Model 8.1})$$

The test statistic is $T = 2 \times \{\log\text{Lik}(\text{reference}) - \log\text{Lik}(\text{nested})\}$

The p -value is $p\text{-value} = (0.5) \Pr(\chi_1^2 > T)$.

Decision: The p -value for testing Hypothesis 8.2 is less than 1% and thus we have strong evidence to reject the null hypothesis. We should retain the random teacher effects (**Model 8.1**).

Hypothesis 8.3: Test whether the fixed effects associated with the year variable can be omitted in **Model 8.1**

The null and alternative hypotheses are:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0 \quad (\text{Model 8.1})$$

The test statistic is $T = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$

The p -value is $\Pr(Z > T)$ using the normal approximation.

Decision: The p -value for testing Hypothesis 8.3 is less than 1% and we should reject the null hypothesis. Therefore, we choose **Model 8.1** as our final model.

Empirical Best Linear Unbiased Predictors (EBLUPs)

General formula: $\hat{u}_i = \hat{D}Z_i'\hat{V}_i^{-1}(y_i - \mathbf{X}_i\hat{\beta})$

In **Model 8.1:**

$$\hat{u}_i = \frac{\hat{\sigma}_{st}^2}{\hat{\sigma}_{st}^2 + \hat{\sigma}_{te}^2 + \hat{\sigma}^2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$\hat{v}_i = \frac{\hat{\sigma}_{te}^2}{\hat{\sigma}_{st}^2 + \hat{\sigma}_{te}^2 + \hat{\sigma}^2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

In **Model 8.2:**

$$\hat{v}_i = \frac{\hat{\sigma}_{te}^2}{\hat{\sigma}_{te}^2 + \hat{\sigma}^2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

In **Model 8.3:**

$$\hat{u}_i = \frac{\hat{\sigma}_{st}^2}{\hat{\sigma}_{st}^2 + \hat{\sigma}^2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i).$$

Y_{tij} : dependent variable math_{tij}

X_{tij} : year_{tij} measured in t -th year, i -th student being instructed by the j -th teacher

$u_i \sim N(0, \sigma_{st}^2)$ and $v_j \sim N(0, \sigma_{te}^2)$: the two random effects

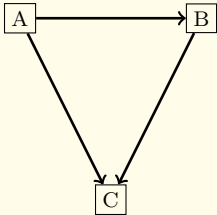
$\epsilon_{tij} \sim N(0, \sigma^2)$: residuals



Bayesian Analysis and Markov Chain Monte Carlo

OVERVIEW AND THE BAYES' FORMULA

| | |
|---|--|
| Binomial distribution with parameters n and p | $\Pr(w n, p) = \binom{n}{w} p^w (1-p)^{n-w} = \frac{n!}{w!(n-w)!} p^w (1-p)^{n-w}$ |
| Likelihood | The likelihood of a model is a mathematical function specifying the plausibility of the data. In the binomial distribution above, the likelihood is $\Pr(w n, p)$. |
| Parameters | A likelihood usually consists of data and parameter(s). In the binomial distribution above, w and n are data and p is an unknown parameter. |
| Prior | <p>The prior is an initial set of plausibility assigned to each possible value of the parameter. In the binomial distribution above, the prior of p is $\Pr(p)$.</p> <p>Informative prior means a prior that expresses specific, definite information about a variable, while a non-informative prior expresses vague or general information about a variable.</p> <p>A prior distribution is a proper prior if the sum (or integral) of probabilities over all possible values of the parameter is 1. When this is not the case, we call it the improper prior.</p> |
| Posterior | <p>Once we have observed data (likelihood), we can update our initial plausibility (prior) conditional on the data. The resulting estimates are known as the posterior distribution.</p> <p>In the binomial distribution above, the posterior of p is $\Pr(p w)$.</p> |
| Bayes' theorem (continuous) | $\Pr(p w) = \frac{\Pr(w p) \Pr(p)}{\Pr(w)}, \quad \Pr(w) = \int \Pr(w p) \Pr(p) dp = E[\Pr(w p)]$ <p>In word form:</p> $\text{Posterior} = \frac{\text{Probability of the data} \times \text{Prior}}{\text{Average Probability of the data}}$ |
| Grid approximation | <p>For a continuous prior distribution, the grid approximation approximates the continuous posterior distribution by calculating the posterior densities for a finite number, say m, of parameter values, and then connecting the dots (i.e. linear interpolating). The discrete counterpart of the Baye's theorem is:</p> $\Pr(p_j w) = \frac{\Pr(w p_j) \Pr(p_j)}{\Pr(w)}, \quad j = 1, \dots, m,$ $\Pr(w) = \sum_{j=1}^m \Pr(w p_j) \Pr(p_j)$ <p>Shortcomings of grid approximation: While the grid approximation could give accurate approximation of continuous posterior distributions, it scales very poorly with model complexity.</p> |
| Quadratic approximation | Quadratic approximation approximates the region near the peak of the posterior distribution by a normal distribution. |
| Markov chain Monte Carlo (MCMC) | MCMC methods draw samples from the posterior distribution without knowing the exact posterior distribution explicitly. From the samples, we get estimates of parameter values, and apply the empirical distribution to get the MCMC estimate of the posterior distribution. |
| Confidence interval | Confidence interval is an interval of defined probability mass. Credible interval is an interval of posterior probability. |
| Percentile Interval (PI) | Percentile Interval gives an interval with equal probability mass to each tail of a probability distribution. |

| | |
|---|---|
| Highest Posterior Density Interval (HPDI) | Highest Posterior Density Interval gives an interval with the highest posterior density, or with the narrowest interval containing the specified probability mass. |
| Absolute loss function | The absolute loss function of the point estimate of parameter \hat{p} is $ \hat{p} - p $. The estimate minimizing the expected loss using the absolute loss function $ \hat{p} - p $ is the posterior median . |
| Squared-error loss function | The squared loss function of the point estimate of parameter \hat{p} is $(\hat{p} - p)^2$. The estimate minimizing the expected loss using the square-error loss function $(\hat{p} - p)^2$ is the posterior mean . |
| Directed Acyclic Graph (DAG) | <p>A way of describing the qualitative casual relationship between variables. It tells you the consequence of changing one variable (if it is correct). Below is an example of DAG. A, B, and C are the observed variables. The arrows show directions of influence. What the DAG says is:</p> <p>A directly influences B; A directly influences C; B directly influences C.</p>  |

LINEAR AND MULTIVARIATE REGRESSION MODELS

| | |
|-----------------------------|---|
| Simple Regression Model | $h_i \sim \text{Normal}(\mu_i, \sigma)$ <div>(Model m4.3)</div> $\mu_i = \alpha + \beta(x_i - \bar{x})$ $\alpha \sim \text{Normal}(178, 20)$ $\beta \sim \text{Normal}(0, 1)$ $\sigma \sim \text{Uniform}(0, 50)$ |
| Polynomial regression model | $h_i \sim \text{Normal}(\mu_i, \sigma)$ <div>(Model m4.5)</div> $\mu_i = \alpha + \beta_1 x_i + \beta_2 x_i^2$ $\alpha \sim \text{Normal}(178, 20)$ $\beta_1 \sim \text{Log-Normal}(0, 1)$ $\beta_2 \sim \text{Normal}(0, 1)$ $\sigma \sim \text{Uniform}(0, 50)$ |
| B-spline model | $D_i \sim \text{Normal}(\mu_i, \sigma)$ <div>(Model m4.7)</div> $\mu_i = \alpha + \sum_{k=1}^K w_k B_{k,i}$ $\alpha \sim \text{Normal}(100, 10)$ $w_j \sim \text{Normal}(0, 10)$ $\sigma \sim \text{Exponential}(1)$ |

| | |
|--|---|
| Predictor Residual Plots | First model MedAgeMar.s in terms of Mar.s: <div>$R_i \sim \text{Normal}(\mu_i, \sigma)$$\mu_i = \alpha + \beta A_i$$\alpha \sim \text{Normal}(0, 0.2)$$\beta \sim \text{Normal}(0, 0.5)$$\sigma \sim \text{Exponential}(1)$</div> Then plot the residual against the divorce rate. <div>(Model m5.4)</div> |
| Masked relationship | <div>$k_i \sim \text{Normal}(\mu_i, \sigma)$$\mu_i = \alpha + \beta_n n_i + \beta_m \log(m_i)$$\alpha \sim \text{Normal}(0, 0.2)$$\beta_n \sim \text{Normal}(0, 0.5)$$\beta_m \sim \text{Normal}(0, 0.5)$$\sigma \sim \text{Exponential}(1)$</div> <div>(Model m5.7)</div> |
| Categorical variables (two categories) | <div>$h_i \sim \text{Normal}(\mu_i, \sigma)$$\mu_i = \alpha_{\text{SEX}[i]}$$\alpha_j \sim \text{Normal}(178, 20), \text{ for } j = 1, 2.$$\sigma \sim \text{Uniform}(0, 50)$</div> <div>(Model m5.8)</div> |
| Categorical variables (many categories) | <div>$K_i \sim \text{Normal}(\mu_i, \sigma)$$\mu_i = \alpha_{\text{CLADE}[i]}$$\alpha_j \sim \text{Normal}(0, 0.5), \text{ for } j = 1, \dots, 4$$\sigma \sim \text{Exponential}(1)$</div> <div>(Model m5.9)</div> |
| Interaction terms | <div>$\log(y_i) \sim \text{Normal}(\mu_i, \sigma)$$\mu_i = \alpha_{CID[i]} + \beta_{CID[i]}(r_i - \bar{r})$$\alpha_{CID[i]} \sim \text{Normal}(1, 0.1),$$\beta_{CID[i]} \sim \text{Normal}(0, 0.3),$$\sigma \sim \text{Exponential}(1)$</div> <div>(Model m8.3)</div> |

| OVERFITTING, REGULARIZATION, AND INFORMATION CRITERIA | | |
|---|--|--|
| Coefficient of determination | $R^2 = \frac{\text{var}(\text{outcome}) - \text{var}(\text{residuals})}{\text{var}(\text{outcome})} = 1 - \frac{\text{var}(\text{residuals})}{\text{var}(\text{outcome})}$ | |
| Information entropy | $H(p) = -\text{E}[\log(p_i)] = -\sum_{i=1}^n p_i \log(p_i)$ | |
| Cross entropy | $H(p, q) = -\sum_1^n p_i \log(q_i)$ | |



| | |
|---|--|
| Divergence | $D_{\text{KL}}(p, q) = H(p, q) - H(p) = \sum_1^n p_i [\log(p_i) - \log(q_i)] = \sum_1^n p_i \log\left(\frac{p_i}{q_i}\right)$ |
| Deviance | $D(q) = (-2) \sum_i \log(q_i)$ |
| Log-pointwise-predictive-density | $\text{lppd}_{\text{CV}} = \sum_{i=1}^N \frac{1}{S} \sum_{s=1}^S \log \Pr(y_i \theta_{-i,s})$ |
| Akaike information criterion | $\text{AIC} = (-2) \times \log \text{Lik} + 2p = D_{\text{train}} + 2p$ D_{train} : deviance using the training sample (in-sample) |
| Widely Applicable Information Criterion | $\text{WAIC} = (-2) (\text{lppd}) + 2 \times p_{\text{WAIC}}$ |
| Weight for model i (model comparison) | $w_i = \frac{\exp(-\frac{1}{2} d\text{WAIC}_i)}{\sum_{j=1}^m \exp(-\frac{1}{2} d\text{WAIC}_j)}$ |
| Pareto-smoothed importance sampling cross-validation (PSIS) | <p>PSIS uses importance weights approach, which assigns weight to each observation based on their “importance” — some observations have a larger impact on the posterior distribution, while some observations will change the posterior distribution less if they were left out — to compute an approximate cross-validation score of the model, without actually doing any cross-validation.</p> |

MARKOV CHAIN MONTE CARLO

| | |
|-------------------------------|--|
| Metropolis algorithm | $p_{\text{acceptance}} = \min \left\{ \frac{h(\theta_{\text{candidate}})}{h(\theta^{(k)})}, 1 \right\}$ |
| Metropolis-Hastings algorithm | $p_{\text{acceptance}} = \min \left\{ \frac{h(\theta_{\text{candidate}})}{h(\theta^{(k)})} \cdot \frac{g(\theta^{(k)} \theta_{\text{candidate}})}{g(\theta_{\text{candidate}} \theta^{(k)})}, 1 \right\}$ |
| Gibbs algorithm | $p_{\text{acceptance}} = 1$ |
| Hamiltonian Monte Carlo | <p>Starting from an initial value of θ, which we refer to as $\theta_{\text{candidate}}$, a HMC involves L steps where each step has the following two sub-steps:</p> <ol style="list-style-type: none"> 1. Use the gradient of the potential function of θ to make a step of ϕ: $\phi \leftarrow \phi - \epsilon \frac{\partial U(\theta y)}{\partial \theta}$ 2. Use the momentum to update the position of θ: $\theta \leftarrow \theta + \epsilon \frac{\phi}{s}$ <p> L: the number of steps. ϵ: the step size, i.e., how big each update can be. ϕ: the value of the momentum, which follows a normal distribution with standard deviation s. $U(\theta y)$: the potential (posterior) function of θ. </p> <p>HMC performs these sub-steps L number of times, and the final value of θ is the proposed parameter, θ_{current}. The probability of accepting θ_{current} is:</p> $p_{\text{acceptance}} = \min \left\{ \frac{h(\theta_{\text{candidate}})}{h(\theta_{\text{current}})} \cdot \frac{p(\phi_{\text{candidate}})}{p(\phi_{\text{current}})}, 1 \right\}$ |
| Effective number of samples | $n_{\text{eff}} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} \rho_k}$ |



Gelman-Rubin convergence diagnostic

The Gelman-Rubin convergence diagnostic is calculated as follows:

- Suppose we have c chains for a posterior parameter θ , and each chain has length $2n$ after removing the samples in the warmup period.
- Split each chain in half with equal size n so we have a total of $m = 2c$ subchains of length n .
- Let θ_{ij} be the i th value in the j th subchain, $i = 1, \dots, n$, and $j = 1, \dots, m$.
- Let $\bar{\theta}_j = \frac{1}{n} \sum_{i=1}^n \theta_{ij}$ be the average of the j th subchain.
- Let $\bar{\theta} = \frac{1}{m} \sum_{j=1}^m \bar{\theta}_j$ be the overall average across all subchains.
- Compute the between-chain variance: $B = \frac{1}{m-1} \sum_{j=1}^m (\bar{\theta}_j - \bar{\theta})^2$
- Compute the within-chain variance: $W = \frac{1}{m} \sum_{j=1}^m \frac{1}{n-1} \sum_{i=1}^n (\theta_{ij} - \bar{\theta}_j)^2$
- Calculate an estimate of the total chain variance: $V = \frac{n-1}{n} W + B$
- Calculate the Gelman-Rubin convergence diagnostic: $\hat{R} = \sqrt{V/W}$

$h(\theta)$: a function proportional to the desired target probability

$g(\theta|\theta^{(k)})$: the proposal density

$p(\phi)$: momentum density

V : total chain variance

W : within-chain variance

GENERALIZED LINEAR MODELS

Logit link

$$y_i = \text{Binomial}(n, p_i), \quad \text{logit}(p_i) = \log \frac{p_i}{1-p_i} = \alpha + \beta x_i$$

Log link

$$y_i = \text{Normal}(\mu, \sigma_i), \quad \log(\sigma_i) = \alpha + \beta x_i$$

Binomial regression

$$L_i \sim \text{Binomial}(1, p_i)$$

Model m11.4

$$\text{logit}(p_i) = \alpha_{\text{marital}[i]} + \beta_{\text{charac}[i]}$$

$$\alpha_{\text{marital}} \sim \text{Normal}(0, 1.5)$$

$$\beta_{\text{charac}[i]} \sim \text{Normal}(0, 0.5), \quad \text{charac} = 1, 2, 3, 4.$$

Proportional odds change

$$\frac{\text{odds}(\text{charac}_i=1) - \text{odds}(\text{charac}_i=0)}{\text{odds}(\text{charac}_i=0)}$$

Binomial regression with unequal trials

$$y_i \sim \text{Binomial}(n_i, p_i)$$

Model m11.8

$$\text{logit}(p_i) = \alpha_{\text{group}[i]} + \beta_m m_i$$

$$\alpha_{\text{group}} \sim \text{Normal}(0, 10)$$

$$\beta_m \sim \text{Normal}(0, 10)$$

Poisson regression

$$y_i \sim \text{Poisson}(\lambda_i), \quad \log(\lambda_i) = \alpha + \beta x_i$$

Poisson regression with different exposures

$$y_i \sim \text{Poisson}(\lambda_i), \quad \log(\mu_i) = \log(\tau_i) + \alpha + \beta x_i, \quad \log(\tau_i) \text{ is the offset.}$$

Multinomial logit formula (softmax function)

$$\Pr(k|s_1, s_2, \dots, s_K) = \frac{\exp(s_k)}{\sum_{i=1}^K \exp(s_i)}$$

Beta-Binomial regression

$$y_i = \text{BetaBinomial}(N_i, \bar{p}_i, \beta)$$

Model m12.1

$$\text{logit}(\bar{p}_i) = a_{\text{GID}[i]}$$

$$a_j \sim \text{Normal}(0, 1.5)$$

$$\beta = \phi + 2$$

$$\phi \sim \text{Exp}(1)$$



| | | |
|----------------------------------|--|--------------------|
| Gamma-Poisson regression | $y_i = \text{Gamma-Poisson}(\lambda_i, \phi)$ $\lambda_i = e^{a_{\text{CID}[i]} P^{\beta_{\text{CID}[i]}} / \gamma}$ $a_j \sim \text{Normal}(0, 1)$ $\beta_j \sim \text{Exp}(1)$ $\gamma \sim \text{Exp}(1)$ $\phi \sim \text{Exp}(1)$ | Model m12.2 |
| Zero-inflated Poisson | $\Pr(N = n) = (1 - p) \frac{\lambda^n e^{-\lambda}}{n!}, \quad \Pr(n = 0) = p + (1 - p)e^{-\lambda}$ | |
| Zero-inflated Poisson regression | $y_i \sim \text{ZIPoisson}(p_i, \lambda_i)$ $\text{logit}(\pi_i) = \alpha_p$ $\log(\lambda_i) = \alpha_\lambda$ $a_p \sim \text{Normal}(-1.5, 1)$ $a_\lambda \sim \text{Normal}(1, 0.5)$ | Model m12.3 |
| Ordered logit regression | $R_i = \text{Ordered-logit}(\phi_i, \alpha)$ $\phi_i = \beta_E \sum_{j=0}^{E_i-1} \delta_j + \beta_A A_I + \beta_I I_i + \beta_C C_i$ $\alpha_k \sim \text{Normal}(0, 1.5)$ $\beta_A, \beta_I, \beta_C, \beta_E \sim \text{Normal}(0, 1)$ $\delta \sim \text{Dirichlet}(\alpha)$ | ordered likelihood |
| The log-cumulative-odds | $\log\left(\frac{\Pr(y_i \leq k)}{1 - \Pr(y_i \leq k)}\right) = \alpha_k - \phi_i$ $\phi_i = \beta_E \sum_{j=0}^{E_i-1} \delta_j + \beta_A A_I + \beta_I I_i + \beta_C C_i$ | |

MULTILEVEL MODELS AND COVARIANCE

| | | |
|-----------------------------------|--|---------------|
| Complete pooling | $s_i \sim \text{Binomial}(n_i, p_i)$ $\text{logit}(p_i) = \alpha_{\text{TANK}[i]}$ $\alpha_{\text{TANK}} \sim \text{Normal}(0, 1.5)$ | Model m13.1 |
| Partial pooling | $s_i \sim \text{Binomial}(n_i, p_i)$ $\text{logit}(p_i) = \alpha_{\text{TANK}[i]}$ $\alpha_{\text{TANK}} \sim \text{Normal}(\alpha, \sigma)$ $\alpha \sim \text{Normal}(0, 1.5)$ $\sigma \sim \text{Exponential}(1)$ | Model m13.2 |
| No pooling | $s_i \sim \text{Binomial}(n_i, p_i)$ $\text{logit}(p_i) = \alpha_{\text{TANK}[i]}$ $\alpha_{\text{TANK}[i]} \sim \text{Normal}(0, 1.5)$ | Model m13.3np |
| Multiple-cluster multilevel model | $L_i \sim \text{Binomial}(1, p_i)$ $\text{logit}(p_i) = \alpha_{\text{actor}[i]} + \gamma_{\text{block}[i]} + \beta_{\text{treatment}[i]}$ $\beta_j \sim \text{Normal}(0, 0.5), j = 1, \dots, 4$ $\alpha_j \sim \text{Normal}(\bar{\alpha}, \sigma_\alpha), j = 1, \dots, 7$ $\gamma_j \sim \text{Normal}(0, \sigma_\gamma), j = 1, \dots, 6$ $\alpha \sim \text{Normal}(0, 1.5)$ $\sigma_\alpha \sim \text{Exponential}(1)$ $\sigma_\gamma \sim \text{Exponential}(1)$ | Model m13.4 |



Non-centered version of Model m13.4

Model m13.4nc

$$\begin{aligned}
 L_i &\sim \text{Binomial}(1, p_i) \\
 \text{logit}(p_i) &= \underbrace{\bar{\alpha} + z_{\text{actor}[i]} \sigma_{\alpha}}_{\alpha_{\text{actor}[i]}} + \underbrace{x_{\text{block}[i]} \sigma_{\gamma}}_{\gamma_{\text{actor}[i]}} + \beta_{\text{treatment}[i]} \\
 \beta_j &\sim \text{Normal}(0, 0.5), j = 1, \dots, 4 \\
 z_j &\sim \text{Normal}(0, 1), j = 1, \dots, 7 \\
 \chi_j &\sim \text{Normal}(0, 1), j = 1, \dots, 6 \\
 \bar{\alpha} &\sim \text{Normal}(0, 1.5) \\
 \sigma_{\alpha} &\sim \text{Exponential}(1) \\
 \sigma_{\gamma} &\sim \text{Exponential}(1)
 \end{aligned}$$

Standardized actor intercepts

Standardized block intercepts

Varying slopes multilevel model

Model m14.1a

$$\begin{aligned}
 W_i &\sim \text{Normal}(\mu_i, \sigma) \\
 \mu_i &= \alpha_{\text{group}[i]} + \beta_{\text{group}[i]} A_i \\
 \begin{bmatrix} \alpha_{\text{group}} \\ \beta_{\text{group}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, S \right), \quad S = \begin{bmatrix} \sigma_{\alpha} & 0 \\ 0 & \sigma_{\beta} \end{bmatrix} R \begin{bmatrix} \sigma_{\alpha} & 0 \\ 0 & \sigma_{\beta} \end{bmatrix} \\
 \sigma &\sim \text{Exponential}(0, 1) \\
 \alpha &\sim \text{Normal}(0, 10) \\
 \beta &\sim \text{Normal}(0, 10) \\
 \sigma_{\alpha} &\sim \text{Exponential}(0, 1) \\
 \sigma_{\beta} &\sim \text{Exponential}(0, 1) \\
 R &= \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \sim \text{LKJcorr}(2)
 \end{aligned}$$

Multilevel cross-classified model

Model m14.2

$$\begin{aligned}
 L_i &\sim \text{Binomial}(1, p_i) \\
 \text{logit}(p_i) &= A_i + (B_{p,i} + B_{pc,i} C_i) P_i \\
 A_i &= \alpha + \alpha_{\text{actor}[i]} + \beta_{\text{block}[i]} \\
 B_{p,i} &= \beta_p + \beta_{p,\text{actor}[i]} + \beta_{p,\text{block}[i]} \\
 B_{pc,i} &= \beta_{pc} + \beta_{pc,\text{actor}[i]} + \beta_{pc,\text{block}[i]} \\
 \begin{bmatrix} \alpha_{\text{actor}} \\ \beta_{p,\text{actor}} \\ \beta_{pc,\text{actor}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{a1} & 0 & 0 \\ 0 & \sigma_{a2} & 0 \\ 0 & 0 & \sigma_{a3} \end{bmatrix} \times R_{\text{actor}} \times \begin{bmatrix} \sigma_{a1} & 0 & 0 \\ 0 & \sigma_{a2} & 0 \\ 0 & 0 & \sigma_{a3} \end{bmatrix} \right)
 \end{aligned}$$

Multilevel cross-classified model (continued)

$$\begin{aligned}
 \begin{bmatrix} \alpha_{\text{block}} \\ \beta_{p,\text{block}} \\ \beta_{pc,\text{block}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{b1} & 0 & 0 \\ 0 & \sigma_{b2} & 0 \\ 0 & 0 & \sigma_{b3} \end{bmatrix} \times R_{\text{block}} \times \begin{bmatrix} \sigma_{b1} & 0 & 0 \\ 0 & \sigma_{b2} & 0 \\ 0 & 0 & \sigma_{b3} \end{bmatrix} \right) \\
 (\alpha, \beta_p, \beta_{pc}) &\sim \text{Normal}(0, 1) \\
 (\sigma_{a1}, \sigma_{a2}, \sigma_{a3}) &\sim \text{Exponential}(1) \\
 (\sigma_{b1}, \sigma_{b2}, \sigma_{b3}) &\sim \text{Exponential}(1) \\
 R_{\text{actor}} &\sim \text{LKJcorr}(4) \\
 R_{\text{block}} &\sim \text{LKJcorr}(4)
 \end{aligned}$$

Model m14.2 continued

Non-centered version of Model m14.2

$$\begin{aligned}
 A_i &= \alpha + \alpha_{\text{actor}[i]} \sigma_{a1} + \beta_{\text{block}[i]} \sigma_{b1} \\
 B_{p,i} &= \beta_p + \beta_{p,\text{actor}[i]} \sigma_{a2} + \beta_{p,\text{block}[i]} \sigma_{b2} \\
 B_{pc,i} &= \beta_{pc} + \beta_{pc,\text{actor}[i]} \sigma_{a3} + \beta_{pc,\text{block}[i]} \sigma_{b3} \\
 \begin{bmatrix} \alpha_{\text{actor}} \\ \beta_{p,\text{actor}} \\ \beta_{pc,\text{actor}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, R_{\text{actor}} \right) \\
 \begin{bmatrix} \alpha_{\text{block}} \\ \beta_{p,\text{block}} \\ \beta_{pc,\text{block}} \end{bmatrix} &\sim \text{MVNormal} \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, R_{\text{block}} \right)
 \end{aligned}$$

Model m14.3

Continuous categories multilevel model (Gaus- $T_i \sim \text{Poisson}(\lambda_i)$
sian process regression)

Model m14.8

$$\begin{aligned}
 \log(\lambda_i) &= \alpha + \gamma_{\text{society}[i]} + \beta_p \log p_i \\
 \gamma &\sim \text{MVNormal} \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}, K \right), \text{ where } K_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}(0.01) \\
 \alpha &\sim \text{Normal}(0, 10) \\
 \beta_p &\sim \text{Normal}(0, 1) \\
 \eta^2 &\sim \text{Exponential}(2) \\
 \rho^2 &\sim \text{Exponential}(0.5)
 \end{aligned}$$



PART D

Statistical Learning

CLASSIFICATION TREES

Training error rate

$$\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i), \quad \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

Test error rate

$$\text{Average}\{I(y_0 \neq \hat{y}_0)\}$$

Bayes error rate

$$E[1 - \max_j \Pr(Y = j|X = x_0)] \approx 1 - \frac{\sum_{i=1}^m \max_j \Pr(Y_i = j|X_i)}{m}$$

Euclidean distance

$$\text{E.d.}(X, Y) = \sqrt{\sum_{j=1}^p (x_j - y_j)^2}, \quad X = (x_1, \dots, x_p), Y = (y_1, \dots, y_p)$$

Conditional probability for class m in the KNN classifier

$$\Pr(Y = m|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = m) \quad \text{for } m = 1, \dots, M.$$

Classification error rate

$$E_m = 1 - \max_k (\hat{p}_{mk})$$

Gini index

$$G_m = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$$

Cross-entropy

$$D_m = - \sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk})$$

 \hat{y}_0 : the predicted class label that results from applying the classifier to the test observation with predictor x_0 \hat{p}_{mk} : the proportion of training observations in the m th region that are from the k th class

REGRESSION TREES

Residual sum of squares (RSS)

$$\text{RSS} = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2 \quad \text{for regions } R_j, j = 1, \dots, J$$

Cost complexity pruning (weakest link pruning)

$$\text{minimize}_T \sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

α controls the trade-off between the subtree's complexity and its fit to the training data. As α increases, the subtree will end up with fewer terminal nodes.

 \hat{y}_{R_j} : the mean response for the training observations within the j th region \hat{y}_{R_m} : the mean of the training observations in R_m T : a decision tree $|T|$: the number of terminal nodes of the tree T

BAGGING AND BOOSTING

Bagging decision tree prediction

$$\hat{f}_{bag}(x) = \frac{1}{N} \sum_{n=1}^N \hat{f}^{*n}(x)$$

Boosted decision tree prediction

$$\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x)$$

 $\hat{f}^{*n}(x)$: the output of the decision tree fitted to the n th bootstrapped training set $\hat{f}^b(x)$: the output of the b th tree fitted to the residuals from the first $b - 1$ trees λ : the shrinkage parameter

PRINCIPAL COMPONENTS ANALYSIS

| | |
|--|--|
| The set of the first principal components | $Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \cdots \phi_{p1}X_p$ |
| Loading vector of the first principal components | $\phi_1 = (\phi_{11}, \cdots, \phi_{p1})^T$ |
| Scores of the first principal component. | z_{11}, \cdots, z_{n1} |
| The first principal component of observation i | $z_{i1} = \sum_{j=1}^p \phi_{j1}x_{ij} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \cdots \phi_{p1}x_{ip}$ |
| The second principal component of observation i | $z_{i2} = \sum_{j=1}^p \phi_{j2}x_{ij} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \cdots \phi_{p2}x_{ip}$ |
| Proportion of Variance Explained (PVE) | $\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$ |
| Variance explained by the m th principal component | $\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n (\sum_{j=1}^p \phi_{jm}x_{ij})^2$ |
| PVE of the m th principal component | $\text{PVE}_m = \frac{\frac{1}{n} \sum_{i=1}^n z_{im}^2}{\sum_{j=1}^p \text{Var}(X_j)} = \frac{\sum_{i=1}^n z_{im}^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$ |

K-MEANS CLUSTERING

| | |
|--|---|
| Minimize total within-cluster variation | $\text{minimize}_{C_1, \cdots, C_K} \left\{ \sum_{k=1}^K W(C_k) \right\}$ |
| Within-cluster variation estimated using squared Euclidean distance | $W(C_k) = 2 \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$, where $\bar{x}_{kj} = \frac{1}{ C_k } \sum_{i \in C_k} x_{ij}$ |
| Alternative formula of the variation | $W(C_k) = \frac{1}{ C_k } \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$ |
| C_1, \cdots, C_K : sets containing the indices of the observations in those clusters | |
| $W(C_k)$: within-cluster variations, $k = 1, \cdots, K$ | |

HIERARCHICAL CLUSTERING

| | |
|------------------|---|
| Complete linkage | Calculate all pairwise Euclidean distance between the observations in cluster A and the observations in cluster B, and record the largest of these distances. |
| Single linkage | Calculate all pairwise Euclidean distance between the observations in cluster A and the observations in cluster B, and record the smallest of these distances. Single linkage can result in <i>trailing clusters</i> , in which single observations are fused one-at-a-time. |
| Average linkage | Calculate all pairwise Euclidean distance between the observations in cluster A and the observations in cluster B, and record the average of these distances. |
| Centroid linkage | Calculate the two centroids and record the Euclidean distance of these two centroids. Centroid linkage can lead to <i>inversions</i> , where two clusters are fused at a height lower than either individual cluster in the dendrogram. |