Part A: Introduction to Credibility

LIMITED FLUCTUATION CREDIBILITY

Prediction

The **u**pdated prediction, U, is a weighted average of D (**d**ata) and M (**m**anual 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	$n_0 = \lambda_F\left(\frac{\sigma_N^2}{\mu_N} + C_X^2\right)$	$n_0 = \lambda_F\left(1 + C_X^2\right) = \lambda_F\frac{\mathrm{E}(X^2)}{\mu_X^2}$

$Z = 1$ 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	$Z = \sqrt{\frac{\mu_N}{\lambda_F\left(\frac{\sigma_N^2}{\mu_N} + C_X^2\right)}}$	$Z = \sqrt{\frac{\mu_N}{\lambda_F(1+C_X^2)}}$

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)	$\dfrac{1}{\alpha(\alpha-2)}$	$\dfrac{(\alpha-1)^2}{\alpha(\alpha-2)}$
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, θ)	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 $n_0 = \lambda_F \left(1 + C_X^2\right)$ for a Poisson frequency distribution.

BÜHLMANN CREDIBILITY

Hypothetical mean	$\mu_X(\Theta) = \mathrm{E}(X\vert\Theta)$
Process variance	$\sigma_X^2(\Theta) = \mathrm{Var}(X\vert\Theta)$
Expected value of the hypothetical means (unconditional mean)	$\mu_X = \mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X\vert\Theta)] = \mathrm{E}[\mu_X(\Theta)]$
Expected value of the process variance (EPV)	$\mu_{\mathrm{PV}} = \mathrm{E}[\mathrm{Var}(X\vert\Theta)] = \mathrm{E}[\sigma_X^2(\Theta)]$
Variance of the hypothetical means (VHM)	$\sigma_{\mathrm{HM}}^2 = \mathrm{Var}[\mathrm{E}(X\vert\Theta)] = \mathrm{Var}[\mu_X(\Theta)]$
Total variance of X (unconditional variance)	$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X\vert\Theta)] + \mathrm{Var}[\mathrm{E}(X\vert\Theta)] = \mu_{\mathrm{PV}} + \sigma_{\mathrm{HM}}^2$
Bühlmann's k	$k = \dfrac{\mathrm{EPV}}{\mathrm{VHM}} = \dfrac{\mu_{\mathrm{PV}}}{\sigma_{\mathrm{HM}}^2}$
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 $\{X_1, \cdots, X_n, X_{n+1}\}$ are iid given the parameter θ. $\bar{X} = \sum_{i=1}^{n} X_i/n$ is the sample mean, and $\mu_X = \mathrm{E}(X)$ is the unconditional mean.

BÜHLMANN-STRAUB CREDIBILITY

Hypothetical mean	$\mathrm{E}(X_{ij}\mid\Theta)=\mu_X(\Theta)$
Process variance of X_{ij}	$\mathrm{Var}(X_{ij}\mid\Theta)=\sigma_X^2(\Theta)$
Expected value of the hypothetical means	$\mu_X=\mathrm{E}(X)=\mathrm{E}[\mathrm{E}(X\mid\Theta)]=\mathrm{E}[\mu_X(\Theta)]$
Expected value of the process variance (EPV)	$\mu_{\mathrm{PV}}=\mathrm{E}[\mathrm{Var}(X_{ij}\mid\Theta)]=\mathrm{E}[\sigma_X^2(\Theta)]$
Variance of the hypothetical mean (VHM)	$\sigma^2_{\mathrm{HM}}=\mathrm{Var}[\mathrm{E}(X_{ij}\mid\Theta)]=\mathrm{Var}[\mu_X(\Theta)]$
Total variance of X	$\mathrm{Var}(X)=\mathrm{E}[\mathrm{Var}(X\mid\Theta)]+\mathrm{Var}[\mathrm{E}(X\mid\Theta)]=\mu_{\mathrm{PV}}+\sigma^2_{\mathrm{HM}}$
Bühlmann's k	$k=\dfrac{\mathrm{EPV}}{\mathrm{VHM}}=\dfrac{\mu_{\mathrm{PV}}}{\sigma^2_{\mathrm{HM}}}$
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 jth insured in the ith year, $X_i=\frac{\sum_{j=1}^{m_i}X_{ij}}{m_i}$, $\bar{X}=\frac{1}{m}\sum_{i=1}^{n}X_i$ (sample mean) and $m=\sum_{i=1}^{n}m_i$, $j=1,\cdots,m_i$, $i=1,\cdots,n$.

BÜHLMANN PREDICTION FOR CONJUGATE PRIORS

Prior distribution	Conditional dist.	μ_{PV} (EPV)	σ^2_{HM} (VHM)	$k=\dfrac{\mu_{\mathrm{PV}}}{\sigma^2_{\mathrm{HM}}}$
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 $\Theta=\theta$	$f_{X_i\mid\Theta}(x_i\mid\theta)$

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

CONJUGATE DISTRIBUTION

Pair (Prior - Conditional)	Posterior dist.(1)	Bayesian prem. $\hat{\mu}_X(\boldsymbol{x})$	Predictive dist.(2)
Gamma $(\alpha,\, \theta)$ - Poisson (Λ)	$\alpha_* = \alpha + \sum x_i$ $\theta_* = \left(\theta^{-1} + n\right)^{-1}$	$\alpha_*\theta_*$	NB $(\theta_*,\, \alpha_*)$
Beta $(a,\, b)$ - geometric (Θ)(3)	$a_* = a + n$ $b_* = b + \sum x_i$	$\dfrac{b_*}{a_* - 1}$	
Beta $(a,\, b)$ - Bernoulli (Q)	$a_* = a + \sum x_i$ $b_* = b + n - \sum x_i$	$\dfrac{a_*}{a_* + b_*}$	
Beta $(a,\, b)$ - binomial (l, Q)	$a_* = a + \sum x_i$ $b_* = b + \ln - \sum x_i$	$(l)\dfrac{a_*}{a_* + b_*}$	
Gamma $(\alpha,\, \theta)$ - exponential (Λ)(4)	$\alpha_* = \alpha + n$ $\theta_* = \theta + \sum x_i$	$\dfrac{\theta_*}{\alpha_* - 1}$	Pareto $(\alpha_*,\, \theta_*)$
Inverse gamma $(\alpha,\, \theta)$ - exponential (Λ)	$\alpha_* = \alpha + n$ $\theta_* = \theta + \sum x_i$	$\dfrac{\theta_*}{\alpha_* - 1}$	Pareto $(\alpha_*,\, \theta_*)$
Normal $(\mu,\, a)$ - normal $(\Theta,\, v)$	$\mu_* = \dfrac{n\bar{x} + (v/a)\mu}{n + v/a}$ $a_* = \dfrac{v}{n + v/a}$	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 $\boldsymbol{x}$ given $\Theta = \theta_j$ $f(\boldsymbol{x}|\theta_j) = \prod_{i=1}^{n} f(x_i|\theta_j)$

Joint distribution of $\boldsymbol{X}$ and Θ $f(\theta_j, \boldsymbol{x}) = f(\boldsymbol{x}|\theta_j)\, \pi_j$

Marginal distribution of $\boldsymbol{X} = \boldsymbol{x}$ $f(\boldsymbol{x}) = \sum_{j=1}^{J} f(\boldsymbol{x}, \theta_j) = \sum_{j=1}^{J} f(\boldsymbol{x}|\theta_j)\, \pi_j$

Posterior pmf of $\Theta = \theta_j$ given $\boldsymbol{x}$ $f(\theta_j|\boldsymbol{x}) = \dfrac{f(\boldsymbol{x}, \theta_j)}{f(\boldsymbol{x})} = \dfrac{f(\boldsymbol{x}|\theta_j)\, \pi_j}{\sum_{k=1}^{J} f(\boldsymbol{x}|\theta_k)\, \pi_k} = \pi_j^*$

Predictive of X_{n+1} given $\boldsymbol{x}$ $f_{X_{n+1}|\boldsymbol{X}}(x_{n+1}|\boldsymbol{x}) = \mathrm{E}_{\Theta|\boldsymbol{x}}[f_{X_i|\Theta}(x_i|\theta)]$

Bayesian premium $\hat{\mu}_X(\boldsymbol{x}) = \mathrm{E}(X_{n+1}|\boldsymbol{x}) = \sum_{j=1}^{J} \mathrm{E}(X_{n+1}|\theta_j)\, \pi_j^*$

NON-PARAMETRIC MODEL IN BÜHLMANN-STRAUB'S CASE

Sample mean of the ith risk group $\bar{X}_i = \dfrac{1}{m_i} \sum_{j=1}^{n_i} m_{ij} X_{ij}$

Sample process variance of the ith risk group $s_i^2 = \dfrac{1}{n_i - 1} \sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2$

Unbiased estimate of μ_{PV} $\hat{\mu}_{\mathrm{PV}} = \dfrac{\sum_{i=1}^{r} (n_i - 1)s_i^2}{\sum_{i=1}^{r} (n_i - 1)}$

Overall sample mean $\hat{\mu}_X = \bar{X} = \dfrac{1}{m} \sum_{i=1}^{r} m_i \bar{X}_i$

Unbiased estimator of σ^2_{HM} $\hat{\sigma}^2_{\mathrm{HM}} = \dfrac{\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2 - (r-1)\hat{\mu}_{\mathrm{PV}}}{m - (\sum_{i=1}^{r} m_i^2)/m}$

Credibility premium of the ith risk group $\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\bar{X}$ where $\hat{Z}_i = \dfrac{m_i}{m_i + \hat{\mu}_{\mathrm{PV}}/\hat{\sigma}^2_{\mathrm{HM}}}$

Credibility premium for the ith 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 $\hat{\mu}_X = \dfrac{\sum_{i=1}^{r} \hat{Z}_i \bar{X}_i}{\sum_{i=1}^{r} \hat{Z}_i}$

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

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

m_i: The total number of exposures in the experience for risk i $\quad m = \sum_{i=1}^{r} m_i$

n_i: The number of experience periods for risk i $\quad n = \sum_{i=1}^{r} n_i$

$\hat{\sigma}^2_{\mathrm{HM}}$ may be negative in empirical applications. In this case, it maybe 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 ith risk group	$\bar{X}_i = \dfrac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$
Sample process variance of the ith risk group	$s_i^2 = \dfrac{\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2}{n_i-1}$
Unbiased estimate of μ_{PV}	$\tilde{\mu}_{\text{PV}} = \dfrac{\sum_{i=1}^{r}(n_i-1)s_i^2}{\sum_{i=1}^{r}(n_i-1)}$
Overall sample mean	$\bar{X} = \dfrac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_i} X_{ij}$
Unbiased estimator of σ^2_{HM}	$\tilde{\sigma}^2_{\text{HM}} = \dfrac{\sum_{i=1}^{r} n_i(\bar{X}_i-\bar{X})^2-(r-1)\tilde{\mu}_{\text{PV}}}{n-(\sum_{i=1}^{r} n_i^2)/n}$
Credibility premium of the ith risk group	$\tilde{Z}_i\bar{X}_i+(1-\tilde{Z}_i)\bar{X}$ where $\tilde{Z}_i = \dfrac{n_i}{n_i+\tilde{\mu}_{\text{PV}}/\tilde{\sigma}^2_{\text{HM}}}$
Credibility premium for the ith 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 $\hat{\mu}_X = \dfrac{\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i}{\sum_{i=1}^{r}\hat{Z}_i}$

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

$\tilde{\sigma}^2_{\text{HM}}$ 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 ith risk group	$\bar{X}_i = \dfrac{1}{n_*}\sum_{j=1}^{n_*} X_{ij}$
Sample process variance of the ith risk group	$s_i^2 = \dfrac{\sum_{j=1}^{n_*}(X_{ij}-\bar{X}_i)^2}{n_*-1}$
Unbiased estimate of μ_{PV}	$\tilde{\mu}_{\text{PV}} = \dfrac{\sum_{i=1}^{r} s_i^2}{r}$
Overall sample mean	$\bar{X} = \dfrac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{n_*} X_{ij}$
Unbiased estimator of σ^2_{HM}	$\tilde{\sigma}^2_{\text{HM}} = \dfrac{\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2}{r-1} - \dfrac{\tilde{\mu}_{\text{PV}}}{n_*}$
Credibility premium of the ith risk group	$\tilde{Z}_i\bar{X}_i+(1-\tilde{Z}_i)\bar{X}$ where $\tilde{Z}_i = \dfrac{n_*}{n_*+\tilde{\mu}_{\text{PV}}/\tilde{\sigma}^2_{\text{HM}}}$

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

i, $i = 1, \cdots, m$: 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

$$Y_i = \underbrace{X_i\,\boldsymbol{\beta}}_{\text{fixed}} + \underbrace{Z_i\,u_i}_{\text{random}} + \epsilon_i, \qquad u_i \sim N(0, D) \qquad \epsilon_i \sim N(0, R_i)$$

where

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

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

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.

$\boldsymbol{\beta}$: 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:

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

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

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

Implied Marginal Model

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

$$Y_i = X_i\,\boldsymbol{\beta} + \epsilon_i^*$$
$$\epsilon_i^* \sim N(0, V_i)$$
$$V_i = Z_i D Z_i' + R_i$$

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:

$$Y_i = X_i\,\boldsymbol{\beta} + \epsilon_i^*,$$
$$\epsilon_i^* \sim N(0, V_i),$$
$$V_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i$$

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

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

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 $\boldsymbol{\beta}$ using the **generalized least squares** (GLS) assuming the covariance parameters $\boldsymbol{\theta}$ are known.

- The second step is to obtain the estimates of $\boldsymbol{\theta}$ by optimizing the profile log-likelihood function. After obtaining the estimates of $\boldsymbol{\theta}$, we can then calculate the estimates of $\boldsymbol{\beta}$.

The estimator of $\boldsymbol{\beta}$ has the desirable statistical property of being the **best linear unbiased estimator** (EBLUE) of $\boldsymbol{\beta}$.

Restricted Maximum Likelihood (REML) Estimation

REML estimation maximizes the REML log-likelihood function:

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

The REML estimates of the covariance parameters ($\boldsymbol{\theta}$) 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 $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \left(\sum_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$$

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

$$\hat{\boldsymbol{\beta}} = \left(\sum_i \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i\right)^{-1} \sum_i \boldsymbol{X}_i' \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{y}_i$$

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:

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

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

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:

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

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

where $\boldsymbol{\beta}$ is a vector of p unknown fixed-effect parameters and L is a known matrix.

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

***F*-test for testing multiple fixed-effect parameters**

The null hypothesis:

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

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

The F-statistic is: $F = \frac{W}{\text{rank}(\boldsymbol{L})}$, which follows an approximate F distribution, with numerate 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:

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

where:

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

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

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

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

EBLUPs are also known as **shrinkage estimators** because they lend 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 $\sigma_i^2/(\sigma^2_{\text{random factor}} + \sigma_i^2)$,
- σ_i^2 is the variance for level i of the random factor, calculated by $\sigma^2_{\text{error}}/n_i$,
- $\sigma^2_{\text{random factor}}$ 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:

$$
\begin{aligned}
Y_i &= X_i\,\beta + u_i + \epsilon_i \\
u_j &\sim N(0, \sigma_u^2), \\
\epsilon_{ij} &\sim N(0, \sigma^2)
\end{aligned}
$$

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**

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

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

Control Treatment: $\epsilon_{ij} \sim N(0, \sigma^2_c)$

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

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \beta_1\,X_j^{(1)} + \beta_2\,X_j^{(2)} + \beta_3\,X_{ij}^{(3)} + \beta_4\,X_j^{(4)} + \beta_5\,X_{ij}^{(5)} + \beta_6\,X_{ij}^{(6)} + u_j + \epsilon_{ij}, \\
u_j &\sim N(0, \sigma^2_{l.e.}), \\
\epsilon_{ij} &\sim N(0, \sigma^2)
\end{aligned}
$$

Model 3.1A

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

The null and alternative hypotheses are:

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

$$H_{\mathrm{A}} : \sigma_{l.e.} > 0 \qquad \textbf{(Model 3.1)}$$

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

p-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:

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

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

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

The p-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

High Treatment	$\epsilon_{ij} \sim N(0, \sigma_h^2)$
Low Treatment	$\epsilon_{ij} \sim N(0, \sigma_l^2)$
Control Treatment	$\epsilon_{ij} \sim N(0, \sigma_c^2)$

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

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

The null and alternative hypotheses are:

$$H_0 : \sigma_h = \sigma_l \qquad \textbf{(Model 3.2B)}$$

$$H_\text{A} : \sigma_h \neq \sigma_l \qquad \textbf{(Model 3.2A)}$$

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

The p-value is $\Pr(\chi^2_1 > T)$

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

$$H_0 : \sigma^2_{h/l} = \sigma^2_c = \sigma^2 \qquad \textbf{(Model 3.1)}$$

$$H_\text{A} : \sigma^2_{h/l} \neq \sigma^2_c \qquad \textbf{(Model 3.2B)}$$

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

The p-value is $\Pr(\chi^2_1 > T)$.

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

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

The null and alternative hypotheses are

$$H_0 : \beta_5 = \beta_6 = 0 \qquad \textbf{(Model 3.3)}$$

$$H_\text{A} : \beta_5 \neq 0 \text{ or } \beta_6 \neq 0 \qquad \textbf{(Model 3.2B)}$$

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

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

$$H_0 : \beta_1 = \beta_2 = 0 \qquad \textbf{(Model 3.3A)}$$

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

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

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**:

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

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:

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

$$H_\text{A} : \sigma_{i:c}^2 > 0 \qquad \textbf{(Model 4.1)}$$

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

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**.

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:

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

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

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

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

$$H_0 : \beta_5 = 0 \quad \text{vs.} \quad H_{\mathrm{A}} : \beta_5 \neq 0$$

The test statistic is t-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. $H_{\mathrm{A}} : \beta_6 \neq 0$

The test statistic is t-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. $H_{\mathrm{A}} : \beta_7 \neq 0$.

The test statistic is t-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:

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

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

The null and alternative hypotheses are:

$$H_0 : \beta_8 = 0 \tag{Model 4.2}$$

$$H_{\mathrm{A}} : \beta_8 \neq 0 \tag{Model 4.4}$$

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

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

Hierarchical Model

Full model:

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

Level 1 Model (Student)

$$Y_{ijk} = b_{0j|k} + \beta_1 X_{ijk}^{(1)} + \beta_2 X_{ijk}^{(2)} + \beta_3 X_{ijk}^{(3)} + \beta_4 X_{ijk}^{(4)} + \epsilon_{ijk}$$

where $b_{0j|k}$ is the unobserved classroom-specific intercepts, $X_{ijk}^{(1)}$ to $X_{ijk}^{(4)}$ are the student level covariates, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

Level 2 Model (Classroom)

$$b_{0j|k} = b_{0k} + u_{j|k}$$

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.

Level 3 Model (School)

$$b_{0k} = \beta_0 + u_k$$

where $u_k \sim N(0, \sigma_{i:s}^2)$.

Intraclass Correlation Coefficients

The school-level ICC:

$$\text{ICC}_{school} = \frac{\sigma_{i:s}^2}{\sigma_{i:s}^2 + \sigma_{i:c}^2 + \sigma^2}$$

The classroom-level ICC:

$$\text{ICC}_{classroom} = \frac{\sigma_{i:s}^2 + \sigma_{i:c}^2}{\sigma_{i:s}^2 + \sigma_{i:c}^2 + \sigma^2}$$

Y_{ijk}: the dependent variable mathgain

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.

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

Level 3 covariates (School):

$X_{k}^{(8)}$ (housepov): Percentage of households in the neighborhood of the school below the poverty level

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.

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:

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

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

The null and alternative hypotheses are:

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

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

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

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

$$H_0 : \sigma_b^2 = \sigma_c^2 \qquad \textbf{(Model 5.2)}$$

$$H_{\text{A}} : \sigma_b^2 \neq \sigma_c^2 \qquad \textbf{(Model 5.3)}$$

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

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

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. $H_{\mathsf{A}} : \beta_4 \neq$ 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

$\mathsf{AIC} = (-2) \times \mathsf{logLik} + 2 \times p$

p is the number of parameters estimated in the model.

Bayesian Information Criterion

$\mathsf{BIC} = (-2) \times \mathsf{logLik} + \log(n) \times p$

n is the number of observation in the modeled dataset.

Y_{ti}: the dependent variable activate

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

$X_{ti}^{(3)} = \mathsf{TREATMENT}$: 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

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

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

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

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

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

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 × sicdegp interaction are equal to zero in **Model 6.2**.

The the null and alternative hypotheses are

$$H_0 : \beta_7 = \beta_8 = 0 \qquad \textbf{(Model 6.3)}$$

$$H_{\text{A}} : \beta_7 \neq \text{ or } \beta_8 \neq 0 \qquad \textbf{(Model 6.2)}$$

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

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 × 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

$$H_0 : \beta 5 = \beta_6 = 0 \qquad \textbf{(Model 6.4)}$$

$$H_A : \beta_5 \neq \text{ or } \beta_6 \neq 0 \qquad \textbf{(Model 6.3)}$$

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

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

The $X_i^{(3)}$: $\text{sicdegp1}_i = 1$ if sicdegp in level 1, 0 otherwise.

The $X_i^{(4)}$: $\text{sicdegp2}_i = 1$ 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

$$H_0 : \sigma^2_{t|p} = 0 \qquad \textbf{(Model 7.1A)}$$

$$H_A : \sigma^2_{t|p} > 0 \qquad \textbf{(Model 7.1)}$$

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

The p-value is $(0.5)\Pr(\chi^2_0 > T) + (0.5)\Pr(\chi^2_1 > T)$.

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

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \textbf{(Model 7.1)}$$

$$H_{\text{A}} : \sigma_1^2 \neq \sigma_2^2 \qquad \textbf{(Model 7.2C)}$$

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

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

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \qquad \textbf{(Model 7.3)}$$

$$H_{\text{A}} : \beta_5 \neq 0, \text{ or } \beta_6 \neq 0, \text{ or } \beta_7 \neq 0 \qquad \textbf{(Model 7.1)}$$

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

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

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.

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

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

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

$$H_{\text{A}} : \sigma_{t|p}^2 > 0 \qquad \textbf{(Model 7.1)}$$

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

The p-value is $p\text{-value} = (0.5)\Pr(\chi^2_0 > T) + (0.5)\Pr(\chi^2_1 > T) = (0.5)\Pr(\chi^2_1 > T)$.

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

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

The null and alternative hypotheses are

$$H_0 : \sigma^2_1 = \sigma^2_2 \qquad \textbf{(Model 7.1)}$$

$$H_{\text{A}} : \sigma^2_1 \neq \sigma^2_2 \qquad \textbf{(Model 7.2C)}$$

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

The p-value is $p\text{-value} = \Pr(\chi^2_1 > T)$.

Hierarchical Model

Full model:

$$\begin{aligned}\mathsf{GCF}_{tij} &= \beta_0 + \beta_1\,\mathsf{TIME}_t + \beta_2\,\mathsf{BAS\_GCP}_{ij} + \beta_3\,\mathsf{CDA}_{ij} + \beta_4\,\mathsf{AGE}_j \\ &\quad + u_{0j} + u_{1j}\,\mathsf{TIME}_t + u_{0i|j} + \epsilon_{tij}\end{aligned}$$

Level 1 Model (Time):

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

Level 2 Model (Tooth)

$$b_{0i|j} = b_{0j} + \beta_2\,\mathsf{BAS\_GCP}_{ij} + \beta_3\,\mathsf{CDA}_{ij} + u_{0i|j}$$

Level 3 Model (Patient)

At the Patient level indexed by j:

$$\begin{aligned}b_{0j} &= \beta_0 + \beta_4\,\mathsf{AGE}_j + u_{0j} \\ b_{1j} &= \beta_1 + u_{1j}\end{aligned}$$

Y_{tij}: dependent variable gcf_{tij}
$X^{(1)}_t = \mathsf{time}_t$
$X^{(2)}_{ij} = \mathsf{base\_gcf}_{ij}$
$X^{(3)}_{ij} = \mathsf{cda}_{ij}$
$X^{(4)}_j = \mathsf{age}_j$ 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:

$$H_0 : \sigma^2_{st} = 0 \qquad \textbf{(Model 8.2)}$$

$$H_{\mathrm{A}} : \sigma^2_{st} > 0 \qquad \textbf{(Model 8.1)}$$

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

The p-value is $(0.5)\Pr(\chi^2_0 > T) + (0.5)\Pr(\chi^2_1 > T) = (0.5)\Pr(\chi^2_1 > T)$.

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

$$H_0 : \sigma^2_{te} = 0 \qquad \textbf{(Model 8.3)}$$

$$H_{\mathrm{A}} : \sigma^2_{te} > 0 \qquad \textbf{(Model 8.1)}$$

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

The p-value is $p\text{-}\mathsf{value} = (0.5)\Pr(\chi^2_1 > T)$.

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:

$$H_0 : \beta_1 = 0$$

$$H_\text{A} : \beta_1 \neq 0 \qquad \textbf{(Model 8.1)}$$

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

In **Model 8.1**:

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

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

In **Model 8.2**:

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

In **Model 8.3**:

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

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

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

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

Part C: Statistical Learning

ASSESSING MODEL ACCURACY

Mean Squared Error $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$

Bias-Variance Trade-Off $\mathrm{E}[(y_0 - \hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon)$

CLASSIFICATION TREES

Training error rate $\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i), \qquad \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Test error rate $\mathrm{Average}\{I(y_0 \neq \hat{y}_0)\}$

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

Euclidean distance $\mathrm{E.d.}(X, Y) = \sqrt{\sum_{j=1}^{p}(x_j - y_j)^2}, \qquad X = (x_1, \cdots, x_p),\ Y = (y_1, \cdots, y_p)$

Conditional probability for class m in the KNN classifier $\Pr(Y = m|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = m) \qquad \text{for } m = 1, \cdots, M.$

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 mth region that are from the kth class

REGRESSION TREES

Residual sum of squares (RSS) $\mathrm{RSS} = \sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2 \quad \text{for regions } R_j,\ j = 1, \cdots, J$

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

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 jth 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 nth bootstrapped training set

$\hat{f}^{b}(x)$: the output of the bth 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 mth 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 mth 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 $\underset{C_1,\cdots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$

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 *trailing clusters*, 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 *inversions*, where two clusters are fused at a height lower than either individual cluster in the dendrogram.

SINGLE LAYER NEURAL NETWORK

Single layer neural network $$f(X)=\beta_0+\sum_{k=1}^{K}\beta_k g\left(\omega_{k0}+\sum_{j=1}^{p}\omega_{kj}X_j\right)$$

Activation $$A_k=h_k(X)=g\left(\omega_{k0}+\sum_{j=1}^{p}\omega_{kj}X_j\right)$$

Sigmoid activation $$g(z)=\frac{e^z}{1+e^z}=\frac{1}{1+e^{-z}}$$

ReLU (rectified linear unit) activation $$g(z)=(z)_+=0,\quad \text{if } z<0 \text{ and } =z,\quad \text{if } z\geq 0$$

MULTILAYER NEURAL NETWORKS

First hidden layer $$A_k^{(1)}=h_k^{(1)}(X)=g\left(\omega_{k0}^{(1)}+\sum_{j=1}^{p}\omega_{kj}^{(1)}X_j\right)$$

Second hidden layer $$A_\ell^{(2)}=h_\ell^{(2)}(X)=g\left(\omega_{k0}^{(2)}+\sum_{k=1}^{K_1}\omega_{\ell k}^{(2)}A_k^{(1)}\right)$$

Output layer $$Z_m=\beta_{m0}+\sum_{\ell=1}^{K_2}\beta_{m\ell}A_\ell^{(2)}$$

Softmax activation $$f_m(X)=\Pr(Y=m|X)=\frac{e^{Z_m}}{\sum_{\ell=0}^{9}e^{Z_\ell}}$$

Cross-entropy $$-\sum_{i=1}^{n}\sum_{m=0}^{9}y_{im}\log(f_m(x_i))$$

Nonlinear logistic regression $$\log\left(\frac{\Pr(Y=1|X)}{\Pr(Y=0|X)}\right)=Z_1-Z_0=(\beta_{10}-\beta_{00})+\sum_{\ell}^{K_2}(\beta_{1\ell}-\beta_{0\ell})A_\ell^{(2)}$$

RECURRENT NEURAL NETWORKS

Hidden Layers	$A_{\ell k} = g\left(\omega_{k0} + \sum_{j=1}^{p} \omega_{kj} X_{\ell j} + \sum_{s=1}^{K} u_{ks} A_{\ell-s,s}\right)$
Output layer	$O_\ell = \beta_0 + \sum_{s=1}^{K} \beta_k A_{\ell k}$
Sum of squared errors	$\sum_{i=1}^{n}(y_i - o_{iL})^2 = \sum_{i=1}^{n}\left(y_i - \left(\beta_0 + \sum_{s=1}^{K} \beta_k g(\omega_{k0} + \sum_{j=1}^{p} \omega_{kj} X_{iLj} + \sum_{s=1}^{K} u_{ks} a_{i,L-1,s})\right)\right)^2$

FITTING A NEURAL NETWORK

Fitting a neural network	$\min_{\{\omega_k\}_1^K,\beta} \frac{1}{2}\sum_{i=1}^{n}(y_i - f(x_i))^2$
Reformulation of the objective function	$R(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f_\theta(x_i))^2$
A single term	$R_i(\theta) = \frac{1}{2}\left(y_i - \beta_0 - \sum_{k=1}^{K} \beta_k g(\omega_{k0} + \sum_{j=1}^{p} \omega_{kj} x_{ij})\right)^2$
Gradient of $R(\theta)$ evaluated at $\theta = \theta^m$	$\nabla R(\theta^m) = \frac{\partial R(\theta)}{\partial \theta}\|_{\theta=\theta^m}$
Gradient descent	$\theta^{m+1} \leftarrow \theta^m - \rho \nabla R(\theta^m)$
The derivative of $R_i(\theta)$ with respect to β_k	$\frac{\partial R_i(\theta)}{\partial \beta_k} = \frac{\partial R_i(\theta)}{\partial f_\theta(x_i)} \frac{\partial f_\theta(x_i)}{\partial \beta_k} = -\left(y_i - f_\theta(x_i)\right) \cdot g(z_{ik})$
The derivative of $R_i(\theta)$ with respect to ω_{kj}	$\frac{\partial R_i(\theta)}{\partial \omega_{kj}} = \frac{\partial R_i(\theta)}{\partial f_\theta(x_i)} \frac{\partial f_\theta(x_i)}{\partial \omega_{kj}} = -\left(y_i - f_\theta(x_i)\right)) \beta_k g'(z_{ik}) x_{ij}.$
Objective function with a penalty term	$R(\theta;\lambda) = -\sum_{i=1}^{n}\sum_{m=0}^{9} y_{im} \log\left(f_m(x_i)\right) + \lambda \sum_j \theta_j^2$

Part D: Time Series

TREND AND SEASONALITY

Let m_t be the trend component, s_t be the seasonality component, and z_t be the remainder term.

Additive model: $x_t = m_t + s_t + z_t$

Multiplicative model: $x_t = m_t s_t + z_t$

Centered moving average for monthly data: $\hat{m}_t = \dfrac{0.5m_{t-6} + \cdots + m_{t-1} + m_t + m_{t+1} + \cdots + 0.5m_{t+6}}{12}$

Monthly additive effect: $\hat{s}_t = x_t - \hat{m}_t$ We can adjust $\hat{s}_t$ so that they average 0.

Monthly multiplicative effect: $\hat{s}_t = \dfrac{x_t}{\hat{m}_t}$ We can adjust $\hat{s}_t$ so that they average 1.

Additive model, seasonally adjusted data: $x_t - \hat{s}_t$

Multiplicative model, seasonally adjusted data: $\dfrac{x_t}{\hat{s}_t}$

STATIONARITY

Stationary: Properties like mean and variance do not depend on the time.

Second-order stationary: Stationary in mean and variance, and autocorrelation is a function only of the lag, not of the time.

Strictly stationary: No moments vary with t.

Time series with trends, or with seasonality, are not stationary.

Time series: x_t

Mean: $E\left[x_t\right] = \mu\left(t\right)$

Variance: $Var\left(x_t\right) = \sigma^2\left(t\right) = E\left[\left(x_t - \mu\left(t\right)\right)^2\right]$

If the series is stationary in the variance: $Var\left(x_t\right) = \sigma^2$

The sample variance is: $s^2 = \dfrac{1}{n-1}\sum_{t=1}^{n}\left(x_t - \bar{x}\right)^2$

AUTOCORRELATIONS

In this section, we assume that the time series x_t is second-order stationary.

Autocovariance: $\gamma_k = \gamma\left(x_{t+k}, x_t\right) = E\left[\left(x_{t+k} - \mu\right)\left(x_t - \mu\right)\right]$

Autocorrelation: $\rho_k = \rho(x_{t+k}, x_t) = \dfrac{\gamma(x_{t+k}, x_t)}{\sqrt{Var(x_{t+k})\,Var(x_t)}} = \dfrac{\gamma(x_{t+k}, x_t)}{\gamma(x_t, x_t)}$ $\quad Var(x_{t+k}) = Var(x_t)$

Sample autocovariance: $c_k = c(x_{t+k}, x_t) = \dfrac{1}{n}\sum_{t=1}^{n-k}(x_{t+k}-\bar{x})(x_t-\bar{x})$

Sample autocorrelation: $r_k = r(x_{t+k}, x_t) = \dfrac{c(x_{t+k}, x_t)}{\sqrt{c(x_{t+k}, x_{t+k})\,c(x_t, x_t)}} = \dfrac{c(x_{t+k}, x_t)}{c(x_t, x_t)}$ $\quad c(x_{t+k}, x_{t+k}) = c(x_t, x_t)$

Note that $c(x_t, x_t) \neq s^2$.

A correlogram plots the autocorrelation function (ACF):

ACF slowly decreases from 1 $\rightarrow$ Sign of trend

ACF shows an oscillation $\rightarrow$ Sign of seasonality

Cross-covariance: $\gamma_k(x, y) = \gamma(x_{t+k}, y_t) = E[(x_{t+k}-\mu_x)(y_t-\mu_y)]$ $\quad$ x lags y by k periods

Cross-correlation:

$$\rho_k(x, y) = \rho(x_{t+k}, y_t) = \frac{\gamma(x_{t+k}, y_t)}{\sqrt{Var(x_{t+k})\,Var(y_t)}} = \frac{\gamma(x_{t+k}, y_t)}{\sqrt{Var(x_t)\,Var(y_t)}}$$

$Var(x_{t+k}) = Var(x_t)$

Sample cross-covariance: $c_k(x, y) = c(x_{t+k}, y_t) = \dfrac{1}{n}\sum_{t=1}^{n-k}(x_{t+k}-\bar{x})(y_t-\bar{y})$

Sample cross-correlation:

$$[t]r_k(x, y) = r(x_{t+k}, y_t) = \frac{c(x_{t+k}, y_t)}{\sqrt{c(x_{t+k}, x_{t+k})\,c(y_t, y_t)}} = \frac{c(x_{t+k}, y_t)}{\sqrt{c(x_t, x_t)\,c(y_t, y_t)}}$$

$c(x_{t+k}, x_{t+k}) = c(x_t, x_t)$

WHITE NOISE

A **white noise** time series w_t is a stationary time series: $\quad w_t \overset{iid}{\sim} N\left(0,\ \sigma_w^2\right)$

Mean: $E[w_t] = 0$

Variance: $Var(w_t) = \sigma_w^2$

Autocovariance: $\gamma(w_{t+k}, w_t) = 0 \qquad k \geq 1$

Autocorrelation: $\rho(w_{t+k}, w_t) = 0 \qquad k \geq 1$

Correlogram: The ACF is close to 0 for $k \geq 1$.

RANDOM WALK

A **random walk** is a nonstationary time series: $\quad x_1 - \mu = w_1$

$$x_t - \mu = (x_{t-1} - \mu) + w_t \quad \rightarrow \quad x_t = \mu + (w_1 + \cdots + w_t)$$

Mean: $E[x_t] = \mu$

Variance: $Var(x_t) = \sigma_w^2 t$

Autocovariance: $\gamma(x_{t+k}, x_t) = \sigma_w^2 t$

Autocorrelation: $\rho(x_{t+k}, x_t) = \dfrac{\gamma(x_{t+k}, x_t)}{\sqrt{Var(x_{t+k})\, Var(x_t)}} = \dfrac{\sigma_w^2 t}{\sqrt{\sigma_w^2 (t+k)\, \sigma_w^2 t}} = \dfrac{t}{\sqrt{t(t+k)}}$

Correlogram: The ACF will slowly decrease from 1 to 0. Note that the difference of a random walk, $y_t = x_t - x_{t-1} = w_t$ is a white noise.

AUTOREGRESSIVE MODELS

***AR*(*p*) model:** $x_t - \mu = \alpha_1 (x_{t-1} - \mu) + \alpha_2 (x_{t-2} - \mu) + \cdots + \alpha_p (x_{t-p} - \mu) + w_t$

Example of AR models:

$x_t = w_t$ $\rightarrow$ $AR(0)$ with $\mu = 0$ is a white noise.

$x_t = x_{t-1} + w_t$ $\rightarrow$ $AR(1)$ with $\mu = 0$ and $\alpha_1 = 1$ is a random walk.

Mean: $E[x_t] = \mu$

Correlogram: The PACF cuts off after lag p.

For an $AR(p)$ model, the PACF at lag p is just α_p.

Fitting *AR*(*p*) model: Coefficients of an $AR(p)$ can be estimated by linear regression, regressing the series on itself with various lags.

R estimates the best AR model using maximum likelihood and the AIC.

Backward shift operator for *AR*(*p*) model:

Define: $B^k (x_t - \mu) = (x_{t-k} - \mu)$ B is the backward shift operator.

We can write: $\alpha_p(B)\ (x_t - \mu) = w_t$ where $\alpha_p(B) = 1 - \alpha_1 B - \alpha_2 B^2 - \cdots - \alpha_p B^p$.

Stationary *AR*(*p*) model: An $AR(p)$ is stationary if the roots of $\alpha_p(B) = 0$ exceed 1 in absolute value.

For $AR(1)$, the root is greater than 1 in absolute value if: $|\alpha_1| < 1$

For $AR(2)$, the roots are greater than 1 in absolute value if:

$\alpha_2 - \alpha_1 < 1$

$\alpha_2 + \alpha_1 < 1$

$|\alpha_2| < 1$

Invertible $AR(p)$ model: An $AR(p)$ is always invertible.

We can write: $x_t - \mu = \alpha_p^{-1}(B)\, w_t \quad \rightarrow \quad x_t - \mu = w_t + \psi_1 w_{t-1} + \cdots + \psi_\infty w_{t-\infty}$

This is an $MA(\infty)$ series.

The variance, covariance, etc., of an $AP(p)$ can be derived using the resulting $MA(\infty)$ series.

For details, refer to Stationary **$MA(q)$ model** in section **K7**.

Stationary $AR(1)$ model:

The $AR(1)$ **model** is: $x_t = \mu + \alpha(x_{t-1} - \mu) + w_t$ An $AR(1)$ is stationary if $|\alpha| < 1$.

We have $E[x_{t+k}] = E[x_t]$ and $Var(x_{t+k}) = Var(x_t)$, and the following:

Mean: $E[x_t] = \mu + \alpha(E[x_{t-1}] - \mu) \quad \rightarrow \quad E[x_t] = \mu$

Variance: $Var(x_t) = \alpha^2 Var(x_{t-1}) + \sigma_w^2 \quad \rightarrow \quad Var(x_t) = \dfrac{\sigma_w^2}{1-\alpha^2}$

Autocovariance: $\gamma(x_{t+k}, x_t) = E[x_t x_{t+k}] = \alpha^k Var(x_t) \quad \rightarrow \quad \gamma(x_{t+k}, x_t) = \dfrac{\alpha^k \sigma_w^2}{1-\alpha^2}$

Autocorrelation: $\rho(x_{t+k}, x_t) = \dfrac{\alpha^k Var(x_t)}{\sqrt{Var(x_{t+k})\, Var(x_t)}} \quad \rightarrow \quad \rho(x_{t+k}, x_t) = \alpha^k$

Partial autocorrelation: $\rho(x_{t+k}, x_t) = \begin{cases} \alpha, & k = 1 \\ 0, & k \geq 2 \end{cases}$

Correlogram: The ACF exponentially decays from 1 to 0.

The PACF cuts off after lag 1.

MOVING AVERAGE MODELS

$MA(q)$ model: $x_t - \mu = w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}$

Mean: $E[x_t] = \mu$

Correlogram: The ACF **cuts off** after lag q.

Fitting $MA(q)$ model: R estimates the parameters by minimizing the conditional sum of squared residuals, $\sum w_t^2$.

Model: $x_{t+1} = w_{t+1} + \beta_1 w_t + \beta_2 w_{t-1} + \cdots + \beta_q w_{t-q+1}$

Forecast: $\hat{x}_{t+1|t} = 0 + \beta_1 w_t + \beta_2 w_{t-1} + \cdots + \beta_q w_{t-q+1}$

Residual: $\hat{w}_{t+1} = x_{t+1} - \hat{x}_{t+1|t}$ where x_{t+1} is the actual value.

Backward shift operator for $MA(q)$:

Define: $B^k w_t = w_{t-k}$ B is the backward shift operator.

We can write: $x_t - \mu = \beta_q(B)\, w_t$ where $\beta_q(B) = 1 + \beta_1 B + \cdots + \beta_q B^p$.

Stationary $MA(q)$ model:

The $MA(q)$ model is: $x_t = \mu + w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}$ An $MA(q)$ is always stationary.

We have $E[x_{t+k}] = E[x_t]$ and $Var(x_{t+k}) = Var(x_t)$, and the following:

Mean: $E[x_t] = \mu$

Variance: $Var(x_t) = \sigma_w^2\left(\beta_0 + \beta_1^2 + \cdots + \beta_q^2\right)$ $\beta_0 = 1$

Autocovariance: $\gamma(x_{t+k}, x_t) = \sigma_w^2\left(\beta_0\beta_k + \beta_1\beta_{1+k} + \cdots + \beta_{q-k}\beta_q\right)$

Autocorrelation: $\rho(x_{t+k}, x_t) = \dfrac{\beta_0\beta_k + \beta_1\beta_{1+k} + \cdots + \beta_{q-k}\beta_q}{\beta_0 + \beta_1^2 + \cdots + \beta_q^2}$ $\rho_k = 0$ if $k > q$

Correlogram: The ACF cuts off after lag q.

Invertible $MA(q)$ model:

We can write: $\beta_q^{-1}(B)(x_t - \mu) = w_t \quad \rightarrow \quad (x_t - \mu) + \phi_1(w_{t-1} - \mu) + \cdots + \phi_\infty(x_{t-\infty} - \mu) = w_t$

This is an $AR(\infty)$ series.

For $MA(1)$, the root is greater than 1 in absolute value if: $|\beta_1| < 1$

For $MA(2)$, the roots are greater than 1 in absolute value if:

$-\beta_2 + \beta_1 < 1$

$-\beta_2 - \beta_1 < 1$

$|\beta_2| < 1$

ARMA MODELS

***ARMA*(p, q) model:** $x_t - \mu = \alpha_1(x_{t-1} - \mu) + \alpha_2(x_{t-2} - \mu) + \cdots + \alpha_p(x_{t-p} - \mu) + w_t + \beta_1 w_{t-1} + \cdots + \beta_q w_{t-q}$

Backward shift operator for $ARMA(p, q)$:

We can write: $\alpha_p(B)\; x_t = \beta_q(B)\; w_t$

We may invert: $x_t - \mu = \alpha_p^{-1}(B)\,\beta_q(B)\,w_t$ This is an $MA(\infty)$ series.

$\beta_q^{-1}(B)\,\alpha_p(B)\,(\,x_t - \mu) = w_t$ This is an $AR(\infty)$ series.

Stationary & Invertible $\boldsymbol{ARMA\,(p,q)}$ model:

An $ARMA\,(p,q)$ is **stationary** if the roots of $\alpha_p\,(B)=0$ exceed 1 in absolute value.

An $ARMA\,(p,q)$ is **invertible** if the roots of $\beta_q\,(B)=0$ exceed 1 in absolute value.

We can write: $x_t-\mu=\alpha_p^{-1}\,(B)\;\beta_q\,(B)\;w_t \quad\rightarrow\quad x_t-\mu=w_t+\psi_1 w_{t-1}+\cdots+\psi_\infty w_{t-\infty}$

This is an $MA\,(\infty)$ series.

The variance, covariance, etc., of an $ARMA\,(p,q)$ can be derived using the resulting $MA\,(\infty)$ series.

For details, refer to Stationary $\boldsymbol{MA\,(q)}$ **model** in section K7.

Simplifying $\boldsymbol{ARMA\,(p,q)}$ model:

If an $ARMA\,(p,q)$ has redundant parameters, we can simplify the model by cancelling common factors.

Stationary $\boldsymbol{ARMA\,(1,1)}$ model:

The $ARMA\,(1,1)$ model is: $x_t-\mu=\alpha\,(x_{t-1}-\mu)+w_t+\beta w_{t-1}$ An $ARMA\,(1,1)$ is stationary if $|\alpha|<1$.

We have $E\,[x_{t+k}]=E\,[x_t]$ and $Var\,(x_{t+k})=Var\,(x_t)$, and the following:

Mean: $E\,[x_t]=\mu$

Variance: $Var\,(x_t)=\dfrac{\left(1+2\alpha\beta+\beta^2\right)\sigma_w^2}{1-\alpha^2}$ This can be easily derived.

Autocovariance: $\gamma\,(x_{t+k},x_t)=\dfrac{\alpha^{k-1}\,(\alpha+\beta)\,(1+\alpha\beta)\,\sigma_w^2}{1-\alpha^2}$ It's hard to derive this. You may just memorize the formula.

Autocorrelation: $\rho\,(x_{t+k},x_t)=\dfrac{\alpha^{k-1}\,(\alpha+\beta)\,(1+\alpha\beta)}{1+2\alpha\beta+\beta^2}$ It's hard to derive this. You may just memorize the formula.

Note that: $\rho_k=\alpha\rho_{k-1} \quad\rightarrow\quad \rho_k=\alpha^{k-1}\rho_1$

ARIMA AND SARIMA MODELS

ARIMA $\boldsymbol{(p,d,q)}$ model: $\alpha_p\,(B)\,(1-B)^d\,(x_t-\mu)=\beta_q\,(B)\;w_t$

ARIMA $\boldsymbol{(p,d,q)(P,D,Q)_s}$ model: $\alpha_P\,(B^s)\,\alpha_p\,(B)\,(1-B^s)^D\,(1-B)^d\;(x_t-\mu)=\beta_Q\,(B^s)\,\beta_q\,(B)\,w_t$

Note that: $(1-B)^2\;x_t=(x_t-x_{t-1})-(x_{t-1}-x_{t-2})$

But: $\left(1-B^2\right)\;x_t=x_t-x_{t-2}$

FORECASTING

n-step ahead forecast: $\hat{x}_{t+n|t} = E\left[x_{t+n}|x_t, x_{t-1}, \ldots, x_0\right]$

The concepts and formulas in the following example can be generalized for other AR, MA, ARMA, ARIMA and SARIMA models.

Example (Forecasting an ARMA model with $\boldsymbol{\mu = 0}$):

Model: $x_{t+1} = \alpha_1 x_t + \alpha_2 x_{t-1} + \alpha_3 x_{t-2} + \cdots + \alpha_p x_{t-p+1} + w_{t+1} + \beta_1 w_t + \beta_2 w_{t-1} + \cdots + \beta_q w_{t-q+1}$

1-step ahead forecast: $\hat{x}_{t+1|t} = \alpha_1 x_t + \alpha_2 x_{t-1} + \alpha_3 x_{t-2} + \cdots + \alpha_p x_{t-p+1} + 0 + \beta_1 w_t + \beta_2 w_{t-1} + \cdots + \beta_q w_{t-q+1}$

Model: $x_{t+2} = \alpha_1 x_{t+1} + \alpha_2 x_t + \alpha_3 x_{t-1} + \cdots + \alpha_p x_{t-p+2} + w_{t+2} + \beta_1 w_{t+1} + \beta_2 w_t + \beta_3 w_{t-1} + \cdots + \beta_q w_{t-q+2}$

2-step ahead forecast: $\hat{x}_{t+2|t} = \alpha_1 \hat{x}_{t+1|t} + \alpha_2 x_t + \alpha_3 x_{t-1} + \cdots + \alpha_p x_{t-p+2} + 0 + 0 + \beta_2 w_t + \beta_3 w_{t-1} + \cdots + \beta_q w_{t-q+2}$

Model: $x_{t+3} = \alpha_1 x_{t+2} + \alpha_2 x_{t+1} + \alpha_3 x_t + \cdots + \alpha_p x_{t-p+3} + w_{t+3} + \beta_1 w_{t+2} + \beta_2 w_{t+1} + \beta_3 w_t + \cdots + \beta_q w_{t-q+3}$

3-step ahead forecast: $\hat{x}_{t+3|t} = \alpha_1 \hat{x}_{t+2|t} + \alpha_2 \hat{x}_{t+1|t} + \alpha_3 x_t + \cdots + \alpha_p x_{t-p+3} + 0 + 0 + 0 + \beta_3 w_t + \cdots + \beta_q w_{t-q+3}$

And so on...

Example (3-step ahead forecast standard error):

We can write: $x_t = \alpha_p^{-1}(B)\ \beta_q(B)\ w_t \quad \rightarrow \quad x_t = w_t + \psi_1 w_{t-1} + \cdots + \psi_\infty w_{t-\infty}$

This is an $MA(\infty)$ series.

Model: $x_{t+3} = w_{t+3} + \psi_1 w_{t+2} + \psi_2 w_{t+1} + \psi_3 w_t + \psi_4 w_{t-1} + \cdots$

Forecast: $\hat{x}_{t+3|t} = 0 + 0 + 0 + \psi_3 w_t + \psi_4 w_{t-1} + \cdots$

Forecast error: $x_{t+3} - \hat{x}_{t+3|t} = w_{t+3} + \psi_1 w_{t+2} + \psi_2 w_{t+1} \quad \rightarrow \quad Var\left(x_{t+3} - \hat{x}_{t+3|t}\right) = \left(1 + \psi_1^2 + \psi_2^2\right)\sigma_w^2$

$SE\left(x_{t+3} - \hat{x}_{t+3|t}\right) = \sqrt{\left(1 + \psi_1^2 + \psi_2^2\right)\sigma_w^2}$

CAS calls this forecast standard error.

95% PI for x_{t+3}: $\hat{x}_{t+3|t} + 1.96\sqrt{\left(1 + \psi_1^2 + \psi_2^2\right)\sigma_w^2}$

TIME SERIES REGRESSION

Differencing:

Example of time series with **stochastic trend:** $x_t = \alpha x_{t-1} + w_t$ which is an $AR(1)$.

Example of time series with **deterministic trend:** $x_t = \alpha_0 + \alpha_1 t + z_t$ which depends on t.

For example: $x_t = \alpha + \beta t + w_t$ is not stationary.

But: $x_t - x_{t-1} = \beta + w_t - w_{t-1}$ is stationary.

Correcting for autocorrelations:

Suppose for a series $x_1 \dots x_n$ with variance σ^2 for each term.

The **autocorrelation** is: $\rho(x_{t+k}, x_t) = \rho_k$

The **variance of sample mean** is: $Var(\bar{x}) = \frac{\sigma^2}{n}\left(1 + \sum_{k=1}^{n-1} 2\left(1 - \frac{k}{n}\right)\rho_k\right)$

The residuals of a linear model are often correlated. In the presence of autocorrelation, the standard errors of coefficients of a regression are unreliable.

To correct for autocorrelation, use **Generalized Least Squares:**

1. Run a linear model, plot the ACF of residuals. Are the autocorrelations significant?
2. Use the autocorrelations from the ACF plot as input to GLS.
3. Run GLS using ML, and obtain GLS coefficient estimates.

Seasonality: To model **seasonal effects**, suppose there are s seasons, one could include s indicator variables, each one equal to 1 if the period is in the season and 0 otherwise, to the model, and remove the intercept from the model.
So $s - 1$ Boolean variables would be added. These variables are called factors.

$x_t = m_t + s_t + z_t$ $\rightarrow$ Trend m_t is usually continuous.

$\rightarrow$ Seasonality s_t is usually categorical/indicator.

Harmonic seasonal model: $x_t = m_t + \sum_{i=1}^{[s/2]} s_i \sin\left(2\pi \frac{it}{s}\right) + c_i \cos\left(2\pi \frac{it}{s}\right) + z_t$ Here, π=180.

Logarithmic transformations:

For **multiplicative model**: $x_t = e^{\alpha + \beta t + z_t}$ where $z_t \sim N(0, \sigma^2)$.

We can apply logarithm: $y_t = \log x_t = \alpha + \beta t + z_t$ This is a linear model.

Lognormal forecast correction factor: $e^{\frac{\sigma^2}{2}}$ σ^2 can be estimated by the s^2=MSE=$\frac{\sum \hat{\varepsilon}_i^2}{n-p}$.

Empirical forecast correction factor: $\frac{1}{n}\sum_{t=1}^{n} e^{\hat{\varepsilon}_i}$ where $\hat{\varepsilon}_i = x_i - \hat{x}_i$ are the residuals.

Forecast: $\hat{x}_t = e^{\hat{y}_t} \times$Forecast Correction Factor