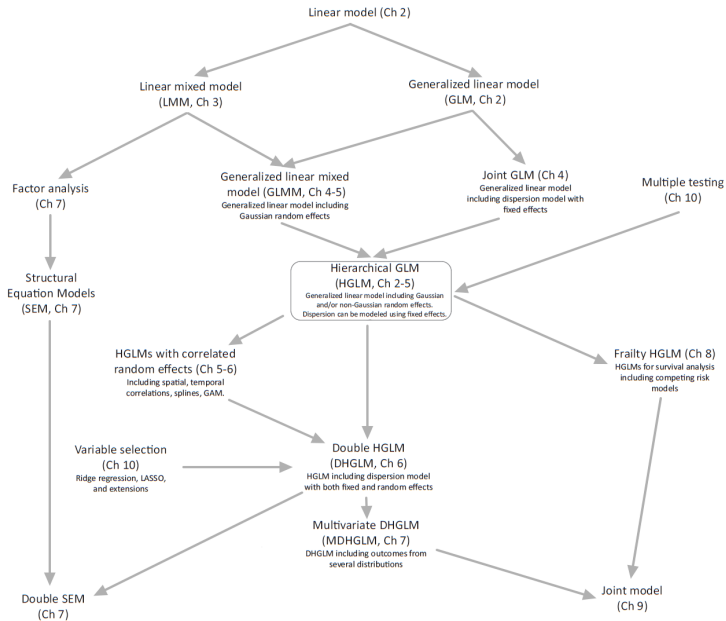
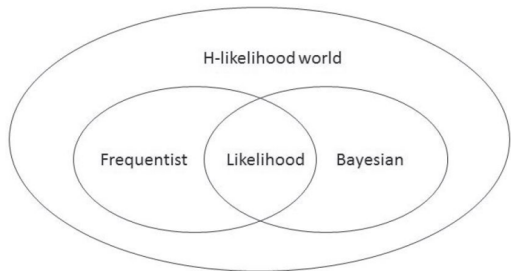
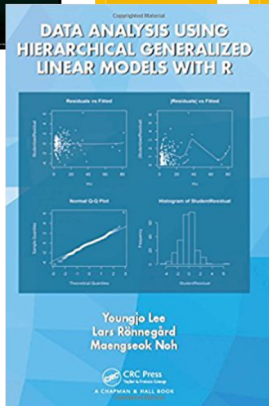
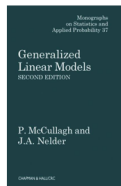
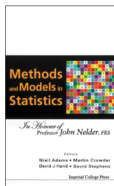
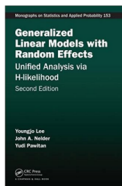
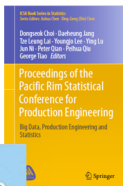
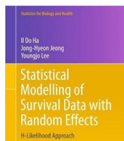
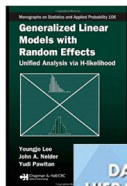


## Data Analysis using Albatross

2020





# Chapter 0. Basic Analysis

## Materials : Download data-sets and manual

Albatross Analytics

Data Import

Data Management ▾

Basic Analysis ▾

Regression ▾


Random Effect Model ▾

Survival Analysis ▾

Multiple Response Analysis ▾

Materials ▾

Download Dataset & Manual

 Download Data-sets

 Download Manual

Manual PDF Link

<http://cheolingsnu.ac.kr:3838/Manual/Manual.pdf>



# Chapter 1. Regression

## Linear Regression Model

### Components of Linear Regression

- the response  $Y$
- the linear predictor :  $\mu = E(Y) = X\beta$
- the distribution of  $y$  : Gaussian Distribution
- Variance :  $\text{Var}(Y) = \phi I$

### Gaussian Distribution

- Log-likelihood

$$\log L(\mu, \phi; y) = -\frac{(y - \mu)^2}{2\phi} - \frac{1}{2} \log(2\pi\phi)$$

## Normal Equation

- $(X^T W X) \hat{\beta} = X^T W y$  where  $W = \frac{1}{\phi} I$   
 $\iff (X^T X) \hat{\beta} = X^T y$
- $\hat{\beta} = (X^T X)^{-1} X^T y$ ,  $Var(\hat{\beta}) = (X^T W X)^{-1} = \phi (X^T X)^{-1}$

**Hat matrix**  $H = X(X^T X)^{-1} X^T$

**Leverage**  $q_i$  :  $i$ -th diagonal elements of  $H$

**Residual**  $\hat{e}_i = y_i - x_i \hat{\beta}$

**Studentized residual** :

$$\frac{\hat{e}_i}{\sqrt{\phi(1 - q_i)}}$$

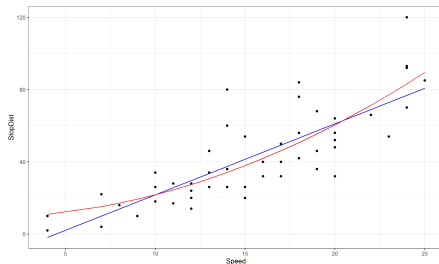
- The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s (Ezekiel, M., 1930).

**StopDist** : stopping distance (ft)

**Speed** : speed of car (mph)

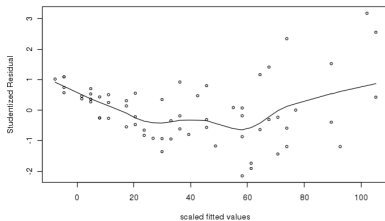
**Model 1** :  $StopDist = \alpha + \beta \text{ Speed}$

**Model 2** :  $StopDist = \alpha + \beta \text{ Speed}^2$

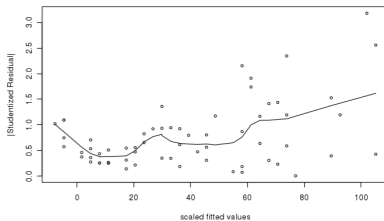


## ex. Carstopping - Model Checking Plot

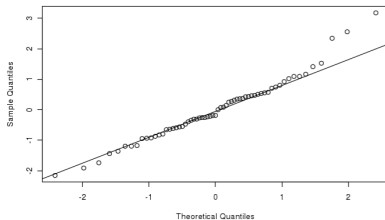
Residuals vs Fitted



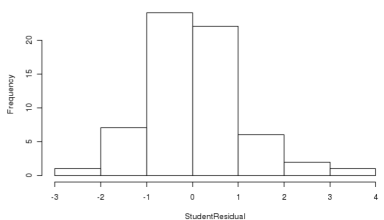
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



- Ozone data measured for 330 days in 1976. All measurements are in the area of Upland, CA, east of Los Angeles (Breiman and Friedman, 1985).

**TempSandburg** : Sandburg Air Force Base temperature ( $^{\circ}\text{C}$ )

**InvHeight** : inversion base height (ft)

**DaggettPressure** : Daggett pressure gradient (mmhg)

**PresHeight** : Vandenburg 500 millibar height (m)

**Visibility** : visibility (miles)

**Humidity** : humidity (%)

**Wind** : wind speed (mph)

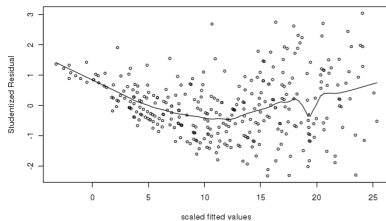
**Day** : day of the year

**Ozone** : upland ozone concentration (ppm)

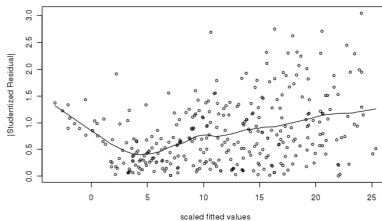
$$\text{Model : Ozone} = \beta_0 + \beta_1 \text{TempSandburg} + \beta_2 \text{InvHeight} + \beta_3 \text{DaggettPressure} + \beta_4 \text{PresHeight}$$

# ex. Ozone - Model Checking Plot

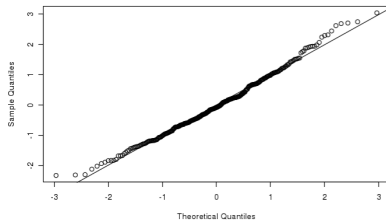
Residuals vs Fitted



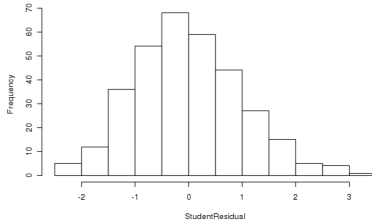
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



## ex. UC Berkeley Admission - UCBA Admission2.csv

- Aggregate data on 4,526 applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex (Bickel et al., 1975).

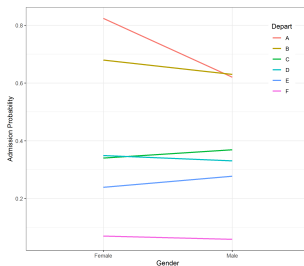
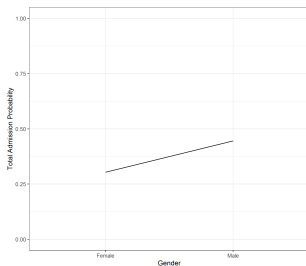
Gender : Male, Female

Department : A, B, C, D, E, F

Admit : 1(Admit), 0(Reject)

Model 1 :  $\text{Admit} = \beta_0 + \beta_1 \text{ Gender}$

Model 2 :  $\text{Admit} = \beta_0 + \beta_1 \text{ Gender} + \beta_2 \text{ Department}$



# Chapter 2. GLMs

## Five components of GLM

- the response  $Y$
- the linear predictor  $\eta = X\beta$
- the distribution of  $y$  (exponential dispersion family)
- the link function  $g(\mu) = \eta$  with  $\mu = E(Y)$
- a prior weight  $1/\phi$

## Likelihood Principle(Birnbaum, 1962)

- The classical likelihood function contains all the information in the observed data about the fixed parameter, provided that the assumed stochastic model is right. Thus, if the model is correct, likelihood captures all the information in the data for analysis.
- Model checking is possible.
- All necessary inferential tools can be derived from the likelihood.
- In GLMs, the likelihood inference can proceed via IWLS equations. Also, the least square methods in regression becomes the ML procedure in GLMs.



## Exponential Dispersion Family

- Log-likelihood

$$\log L(\theta, \phi; y) = \frac{\theta^T y - b(\theta)}{\phi} + c(y, \phi)$$

- $b(\theta)$  is cumulant generating function.
- Mean :  $\mu = E(Y) = b'(\theta)$
- Variance :  $\text{Var}(Y) = \phi b''(\theta)$

Distribution	$E(Y)$	$\theta$	$\phi$	$V(\mu)$	$\text{Var}(Y)$	$b(\theta)$
$N(\mu, \sigma^2)$	$\mu$	$\mu$	$\sigma^2$	1	$\sigma^2$	$\theta^2/2$
$Poi(\mu)$	$\mu$	$\log \mu$	1	$\mu$	$\mu$	$\exp(\theta)$
$Bin(n, p)$	$\mu = np$	$\log \frac{\mu}{n-\mu}$	1	$\frac{\mu(n-\mu)}{n}$	$\frac{\mu(n-\mu)}{n}$	$n \log(1 + \exp(\theta))$
$Gamma(\alpha, \beta)$	$\mu = \frac{\alpha}{\beta}$	$-1/\mu$	$\phi = \frac{1}{\alpha}$	$\mu^2$	$\phi \mu^2 = \frac{\alpha}{\beta^2}$	$-\log(-\theta)$

1. Specify a starting value for  $\beta$ , say  $\beta^{(0)}$  ( $k = 0$ )
2. Compute adjusted linear predictor

$$\eta^{(k+1)} = X\beta^{(k)} \text{ and } \mu^{(k+1)} = g^{-1}(\eta^{(k+1)})$$

3. Compute adjusted dependent variable

$$s_i = \eta_i^{(k+1)} + \frac{\partial \eta_i^{(k+1)}}{\partial \mu_i^{(k+1)}} \left( Y_i - \mu_i^{(k+1)} \right)$$

4. Fit the weighted linear regression  $s = X\beta + \epsilon$  with  $\epsilon \sim (0, W^{(k+1)})$  where

$$W^{(k+1)} = \text{diag} \left( \left( \frac{\partial \eta_i^{(k+1)}}{\partial \mu_i^{(k+1)}} \right)^2 \text{Var}(Y_i) \right).$$

5. Solve  $(X^T W^{(k+1)} X) \hat{\beta} = X^T W^{(k+1)} s$ .

6. Put estimated coefficient as

$$\beta^{(k+1)} = (X^T W^{(k+1)} X)^{-1} X^T W^{(k+1)} s$$

7. Repeat step 2~6 for  $k = 0, 1, 2, \dots$  until convergence.
8. After convergence, report  $\hat{\beta}$  and  $\text{Var}(\hat{\beta}) = (X^T W X)^{-1}$  where

$$W = \text{diag} \left( \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \text{Var}(Y_i) \right).$$

→ IWLS is the extension of least squares method to GLMs!

Homework : You may derive IWLS from the likelihood.

## Residual

- Unscaled deviance :  $D = 2\phi(\ell(y; y) - \ell(\mu; y)) = \sum d_i$
- Unscaled deviance components :  $d_i$

Distribution	Deviance component $d_i$
Normal	$(y_i - \hat{\mu}_i)^2$
Poisson	$2[y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$
Binomial	$2\left[y_i \log(y_i/\hat{\mu}_i) - (m_i - y_i) \log \frac{m_i - y_i}{m_i - \mu_i}\right]$
Gamma	$2[-\log(y_i/\hat{\mu}_i) + (y_i - \hat{\mu}_i)/\hat{\mu}_i]$

- Standardized deviance residuals :  $r_{D,i} = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i/\phi}$
- Pearson residuals :  $r_{P,i} = (y_i - \hat{\mu}_i) / \sqrt{\phi V(\hat{\mu}_i)}$
- $D^* = \sum r_{D,i}^2$  is the log likelihood ratio statistic.
- $P^* = \sum r_{P,i}^2$  is the Pearson chi-squared statistic.

## Hat Values

- GLMs have the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}$$

where  $\mathbf{W} = \text{diag} \left( \left( \frac{\partial \eta_i}{\partial \mu_i} \right)^2 \text{Var}(Y_i) \right)$ .

- The diagonal elements of  $\mathbf{H}$  are the hat values here denoted by  $q_i$ .
- Studentized residuals adjust for the hat values and are obtained as

$$\frac{r_i}{\sqrt{1 - q_i}}.$$

- We can use the unscaled deviance to estimate the dispersion parameter

$$\hat{\phi} = \frac{\sum d_i}{\sum (1 - q_i)} = \frac{D}{n - p}$$

- Crack-growth data from experiment where crack length in inches are measured on a compact tension steel test (CT test) operated in different laboratories (Hudak et al., 1978).

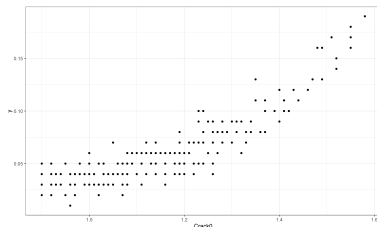
$y$  : increment of crack length (inch)

$crack0$  : initial value of crack length (inch)

$cycle$  : number of cumulative loading cycles ( $10^6$  cycle)

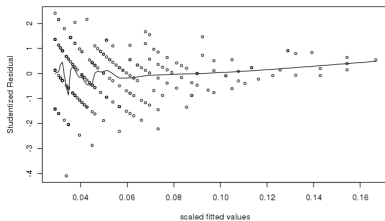
$specimen$  : 21 metallic specimens

**Model** :  $\eta = \log \mu = \alpha + \beta \text{ crack0}$

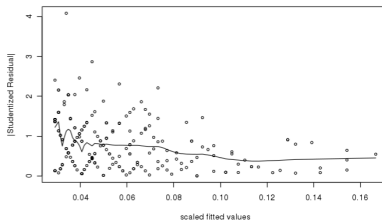


## ex. Crackgrowth - Model Checking Plot

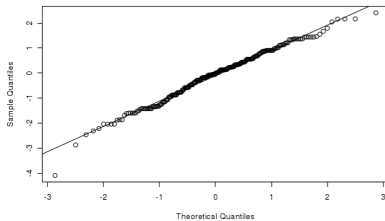
Residuals vs Fitted



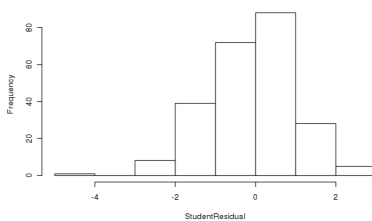
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



- Train-related accidents data in the UK between 1975 and 2003 (Agresti, 2007).

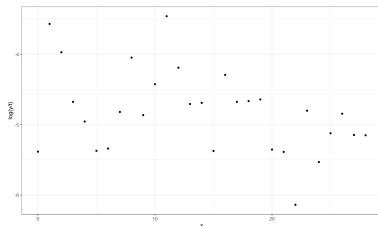
$x$  : number of years since 1975

$y$  : number of accidents between trains and road vehicles

$t$  : distance of train travel (million kilometer)

$\log t$  : logarithm of  $t$

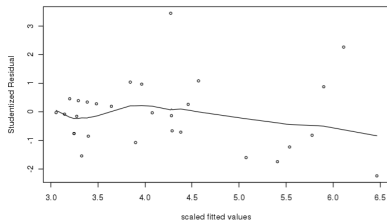
**Model** :  $\eta = \log(\mu) = \log(t) + \alpha + \beta x$



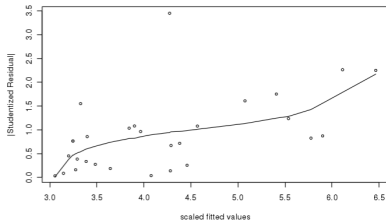


## ex. Train - Model Checking Plot

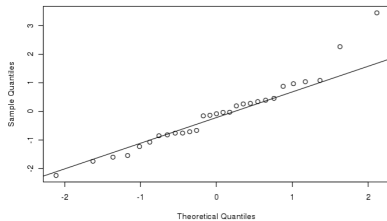
Residuals vs Fitted



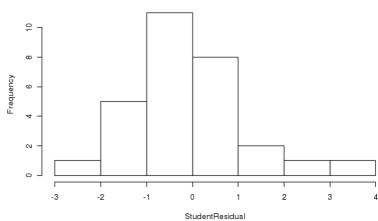
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



- Data from a study of nesting horseshoe crabs, which investigated factors that affect whether the female crab had any other males, called satellites, residing near her (Jane Brockmann, 1996).

**sat** : number of satellites

**y** : indicator of whether a female crab has any satellites

**weight** : weight (kg)

**width** : shell width (cm)

**color** : 1(medium light), 2(medium), 3(medium dark), 4(dark)

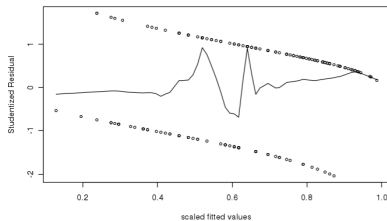
**spine** : 1(both good), 2(one broken), 3(both broken)

When  $p = \text{Prob}(Y=1)$ ,

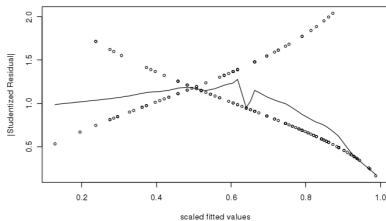
$$\text{Model} : \eta = \log \left( \frac{p}{1-p} \right) = \alpha + \beta \text{ width}$$

## ex. Crabs - Model Checking Plot

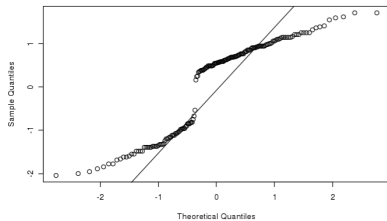
Residuals vs Fitted



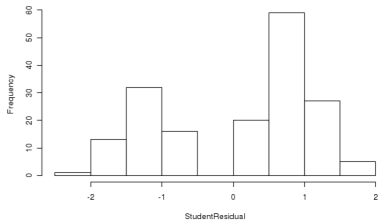
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



- Data based on an epidemiological survey to investigate snoring as a possible risk factor for heart disease (P.G. Norton and E.V. Dunn, 1985).

yes : number of people who have heart disease

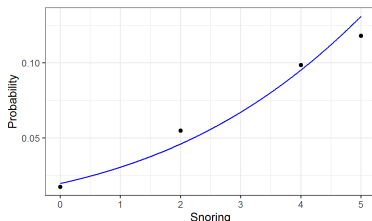
no : number of people who don't have heart disease

x : snoring level. 0(Never), 2(Occasional), 4(Nearly every night), 5(Every night)

n : yes + no

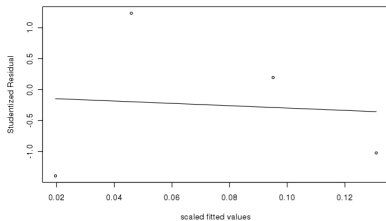
When  $p = \text{Prob}(Y=1)$ ,

$$\text{Model} : \eta = \text{probit}(p) = \Phi^{-1}(p) = \alpha + \beta x$$

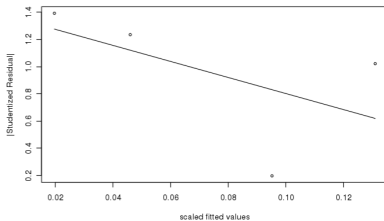


## ex. Snoring - Model Checking Plot

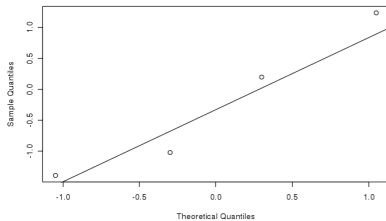
Residuals vs Fitted



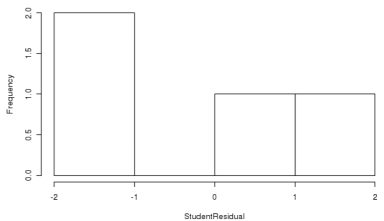
|Residuals| vs Fitted



Normal Probability Plot



Histogram of Student Residual



- Results of matches among five professional tennis players between January 2014 and January 2018 (Agresti, 2019).
- The fitted model provides a ranking of the players.
- It also estimates the probabilities of win and of loss for matches between each pair of players.

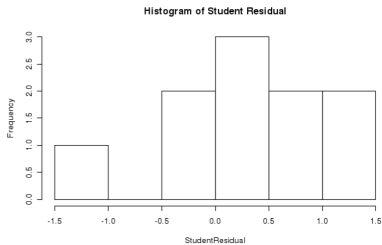
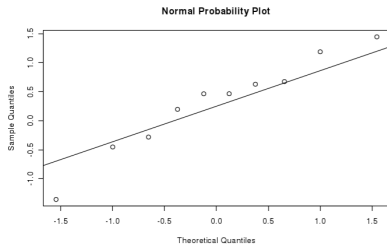
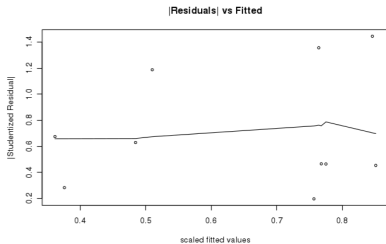
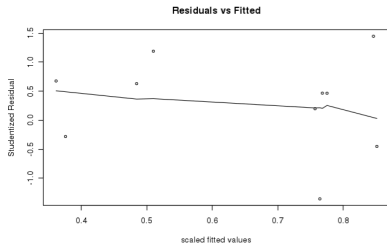
$\pi_{ij}$  : probability that player  $i$  is the victor when  $i$  and  $j$  play

$\pi_{ji} = 1 - \pi_{ij}$  (ties cannot occur)

**Model** :  $\log \left( \frac{\pi_{ij}}{\pi_{ji}} \right) = \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_i - \beta_j$

Winner	Loser				
	Djokovic	Federer	Murray	Nadal	Wawrinka
Djokovic	-	9	14	9	4
Federer	6	-	5	5	7
Murray	3	0	-	2	2
Nadal	2	1	4	-	4
Wawrinka	3	2	2	3	-

## ex. Tennis - Model Checking Plot



# Chapter 3. H-likelihood

## Introduction

- The h-likelihood method can fit rather complex models in an elegant manner.
- In contrast, classical likelihood software may not be as flexible, whereas Bayesian MCMC approaches allow fitting these models but at the expense of more computation time and requires to assume priors for fixed parameters.
- In this chapter we define the h-likelihood and provide insight to inference and predictions based on the h-likelihood. We introduce the *extended likelihood principle* underlying the h-likelihood framework and show how it is related both to classical likelihood and Bayesian inference.



Five important points are made:

- Inference about random effects can be made using the h-likelihood, whilst classical likelihood cannot give any information about the random effects,
- H-likelihood inference of random effects takes into account the uncertainty in estimating the fixed effects, whereas empirical Bayes (EB) estimation of random effects assumes known values of the fixed effects,
- Model checking is possible for all parts of the model,
- All necessary inferential tools can be derived from the h-likelihood, and
- The h-likelihood can be used for predictions of unobserved random variables such as future outcomes.

## Example for prediction of future outcome

- Suppose that we have the number of epileptic seizures in an individual for five weeks,  $\mathbf{y} = (3, 2, 5, 0, 4)$ .
- Suppose also that these counts are i.i.d. from a Poisson distribution with mean  $\theta$ .
- Here,  $\hat{\theta} = (3 + 2 + 5 + 0 + 4)/5 = 2.8$  is the maximum likelihood estimator of  $\theta$ , which maximizing the Fisher likelihood  $f_{\theta}(\mathbf{y})$ . The inferences about  $\theta$  can be made by using the likelihood.
- Now we want to have a predictive probability function for the seizure counts for the next week  $v$ .
- Then, because  $f_{\theta}(v = i | \mathbf{y}) = f_{\theta}(v = i)$ , the plug-in technique gives the predictive distribution for the seizure count  $v$  of the next week:

$$f_{\hat{\theta}}(v = i | \mathbf{y}) = f_{\hat{\theta}}(v = i) = \exp(-2.8) 2.8^i / i!$$

## Example for prediction of future outcome

- Pearson (1920) pointed out the limitation of this Fisher likelihood using the plug-in method because it cannot account for uncertainty in estimating  $\theta$ .
- This plug-in technique is a kind of empirical Bayes method. With Jeffreys' prior,  $\pi(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$ , the resulting marginal posterior

$$p(v|y) = \int f_{\theta}(v|y)\pi(\theta)d\theta$$

gives a predictive probability with higher probabilities for larger  $y$ . This Bayesian procedure handles uncertainty caused by estimating  $\theta$ .

- However, it depends upon the choice of a prior and it might be difficult to justify why the choice of Jeffreys' prior is the right choice.

## Example for prediction of future outcome

- Here the h-likelihood including  $v$  is proportional to

$$f_{\theta}(3, 2, 5, 0, 4, v) = \exp(-6\theta)\theta^{3+2+5+0+4+v}/(3!2!5!0!4!v!)$$

- Now,  $\hat{\theta}(v) = (3 + 2 + 5 + 0 + 4 + v)/6$  is the potential ML estimate if  $v$  is observed.
- Then, the normalized profile likelihood  $f_{\hat{\theta}(v)}(3, 2, 5, 0, 4, v)$  gives the predictive probability  $p(\mathbf{v}|\mathbf{y})$ , almost identical to Pearson's but without assuming a prior on  $\theta$ .
- This is a method to eliminate  $\theta$  from the predictive probability  $f_{\theta}(\mathbf{v}|\mathbf{y})$ .
- This example shows that standard methods for likelihood inferences can be used for the prediction problem by using the h-likelihood without assuming a prior on  $\theta$ .

## Example for prediction of future outcome

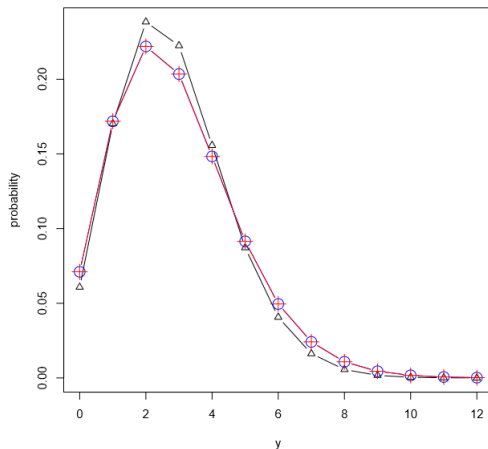


Figure 3.3 Predictive density of the number of seizure counts: Plug-in method ( $\Delta$ ), Bayesian method ( $\circ$ ) and h-likelihood method (+).

- We consider extended statistical models that consist of three types of objects, data  $\mathbf{y}$ , parameter (fixed unknowns)  $\theta$  and unobservables (random unknowns)  $\mathbf{v}$ . Then statistical inferences need to be made for both unknowns  $\theta$  and  $\mathbf{v}$ , based upon the observed data  $\mathbf{y}$ .
- Consider a linear mixed model for  $i = 1, \dots, m$  and  $j = 1, \dots, n_i$

$$y_{ij} = x_{ij}\beta + v_i + e_{ij}, \quad (1)$$

where  $\beta$  is the vector of fixed effects and  $v_i \sim \mathcal{N}(\mathbf{0}, \lambda)$  are i.i.d. random effects,  $e_{ij} \sim \mathcal{N}(0, \phi)$  is an i.i.d. random error.

- In this model, there are two types of unknowns; fixed unknowns  $\theta = (\beta, \phi, \lambda)$  and random unknowns  $\mathbf{v} = (v_1, \dots, v_m)^T$ .

- The linear mixed model (1) may be written in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}. \quad (2)$$

- In the classical likelihood setting the model for the data generation process  $f_{\theta}(y)$  is given by the density function of a multivariate normal distribution

$$N(\mathbf{X}\boldsymbol{\beta}, \lambda\mathbf{Z}\mathbf{Z}^T + \phi\mathbf{I})$$

with the corresponding marginal likelihood

$$L(\boldsymbol{\beta}, \lambda, \phi; \mathbf{y}) = (2\pi|\mathbf{V}|)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \quad (3)$$

where  $\mathbf{V} = \mathbf{Z}\mathbf{Z}^T\lambda + \mathbf{I}\phi$ .

- This marginal likelihood can be used to estimate and make inference about the fixed parameters  $\boldsymbol{\beta}$ ,  $\lambda$  and  $\phi$ . However, the random effect  $\mathbf{v}$  is not included so that the classical likelihood does not directly give inference about the random effects.

- Lee and Nelder (1996) proposed the use of the hierarchical likelihood

$$H(\theta, \mathbf{v}; \mathbf{y}) = f_{\theta}(\mathbf{y}|\mathbf{v})f_{\theta}(\mathbf{v}) = f_{\theta}(\mathbf{v}, \mathbf{y}), \quad (4)$$

where  $f_{\theta}(\mathbf{v}, \mathbf{y})$  is the joint density of  $\mathbf{v}$  and  $\mathbf{y}$ .

- It is related to the conditional distribution of  $\mathbf{v}$  given  $\mathbf{y}$  as

$$H(\theta, \mathbf{v}; \mathbf{y}) = f_{\theta}(\mathbf{v}, \mathbf{y}) = f_{\theta}(\mathbf{y})f_{\theta}(\mathbf{v}|\mathbf{y}). \quad (5)$$

- Bjørnstad introduced the extended likelihood principle where all information in the observed data for parameters  $\theta$  and unobservables  $\mathbf{v}$  are in the extended likelihood, such as the hierarchical likelihood.
- Lee and Nelder (1996) found that the scale of  $\mathbf{v}$  is important for meaningful statistical inference; they called the extended likelihood in a particular scale the *hierarchical likelihood* and its logarithm is referred to as the *h-likelihood*.



- For the h-likelihood, there is a close connection both to classical frequentist inference and Bayesian inference.
  - In the absence of random effects, the hierarchical likelihood is the same as the classical likelihood, i.e.  $H = f_{\theta}(\mathbf{y})$ .
  - In the absence of fixed parameters  $\theta$ ,

$$H(\mathbf{v}; \mathbf{y}) = f(\mathbf{y})f(\mathbf{v}|\mathbf{y}). \quad (6)$$

which is proportional to the posterior  $f(\mathbf{v}|\mathbf{y})$  used for inference in Bayesian statistics where  $f(\mathbf{v})$  is a prior.

- However, in hierarchical models such as linear mixed models,  $\mathbf{v}$  is random and  $f(\mathbf{v})$  is part of the model. To make this distinction clear, we call  $f(\mathbf{v}|\mathbf{y})$  the predictive density (or predictive probability) for random effect  $\mathbf{v}$ .
- In this book, the conditional likelihood  $f_{\theta}(\mathbf{v}|\mathbf{y})$  is called the predictive probability to highlight its probability property

$$\int f_{\theta}(\mathbf{v}|\mathbf{y})d\mathbf{v} = 1$$

- Lee and Nelder (1996) proposed that the random effects could be estimated by finding the mode of the joint density  $f_{\theta}(\mathbf{y}, \mathbf{v})$ .
- Using the mode of  $H$  can simplify the computations drastically compared to MCMC. However, it requires an appropriate scale of  $\mathbf{v}$  because the joint density will depend upon the transformation of  $\mathbf{v}$ .
- For example, the mode of the joint likelihood is not invariant to transformation of  $\mathbf{v}$  and different conclusions will be drawn depending on the scale of  $\mathbf{v}$  chosen when the mode is used for inference about the random effects.
- The novelty of Lee and Nelder's method (1996) is to limit the possible joint likelihoods to a given scale of  $\mathbf{v}$ , resolving the invariance problem.

- For inference about fixed parameters, we use the marginal likelihood derived from  $f_{\theta}(\mathbf{y}, \mathbf{v})$  by integrating out the random effects

$$f_{\theta}(\mathbf{y}) = \int f_{\theta}(\mathbf{y}, \mathbf{v}) d\mathbf{v}. \quad (7)$$

This is a classical Fisher likelihood, so we can obtain the ML estimator for  $\theta$  by maximizing  $f_{\theta}(\mathbf{y})$ .

- For estimating variance components, Patterson and Thompson (1971) suggested a REML approach to improve the estimation properties with reduced bias. REML for linear models can be extended to GLMs through a more general specification as a conditional likelihood  $f_{\theta}(\mathbf{y}|\hat{\beta})$  where  $\hat{\beta}$  is the estimator for the mean parameters. (Smyth and Verbyla, 1996; Lee and Nelder 2001).

- Recall the linear mixed model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

with  $\mathbf{v} \sim N(0, \lambda \mathbf{I})$  and  $\mathbf{e} \sim N(0, \phi \mathbf{I})$ .

- The marginal likelihood is given by

$$\log(f_{\theta}(\mathbf{y})) = \log \int H(\theta, \mathbf{v}; \mathbf{y}) d\mathbf{v} = -\frac{1}{2} \log(\det(2\pi \mathbf{V})) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

where  $\mathbf{V} = \mathbf{Z}\mathbf{Z}^T \lambda + \mathbf{I}\phi$ .

- From the marginal likelihood, we can obtain the ML estimator

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

- REML estimator equations for variance components is obtained from

$$\begin{aligned} \log(f_{\theta}(\mathbf{y}|\hat{\boldsymbol{\beta}})) &= -\frac{1}{2} \log(\det(2\pi \mathbf{V})) - \frac{1}{2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &\quad - \frac{1}{2} \log(\det(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})) \end{aligned}$$

- The h-likelihood given by

$$\begin{aligned}\log(f_{\theta}(\mathbf{y}, \mathbf{v})) &= \log(f_{\theta}(\mathbf{y}|\mathbf{v})) + \log(f_{\theta}(\mathbf{v})) \\ &= -\frac{n}{2} \log(2\pi\phi) - \frac{1}{2\phi} (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{v})^T (\mathbf{y} - \mathbf{X}\beta - \mathbf{Z}\mathbf{v}) \\ &\quad - \frac{m}{2} \log(2\pi\lambda) - \frac{\mathbf{v}^T \mathbf{v}}{2\lambda}\end{aligned}$$

where  $n$  is the number of observations and  $m$  is the length of  $\mathbf{v}$ .

- The joint maximization for  $\beta$  and  $\mathbf{v}$  gives Henderson's mixed model equation

$$\begin{pmatrix} \frac{1}{\phi} \mathbf{X}^T \mathbf{X} & \frac{1}{\phi} \mathbf{X}^T \mathbf{Z} \\ \frac{1}{\phi} \mathbf{Z}^T \mathbf{X} & \frac{1}{\phi} \mathbf{Z}^T \mathbf{Z} + \mathbf{I} \frac{1}{\lambda} \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\mathbf{v}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\phi} \mathbf{X}^T \mathbf{y} \\ \frac{1}{\phi} \mathbf{Z}^T \mathbf{y} \end{pmatrix},$$

which gives the BLUP for  $\mathbf{v}$  and the ML estimator for  $\beta$ .

## Deriving Sample Variance from the REML Likelihood

- Suppose  $y_1, y_2, \dots, y_n$  are i.i.d. observation from  $N(\mu, \sigma^2)$  where both parameters are unknown.
- The ML estimator for  $\mu$  is the sample mean  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i \sim N(\mu, \sigma^2/n)$ , whereas direct maximization of  $\log L$  gives the biased estimator  $\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$ .
- So we consider the REML likelihood

$$\begin{aligned} f(\mathbf{y}|\hat{\mu}) &= \frac{f(\mathbf{y})}{f(\hat{\mu})} = \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2\right)}{\frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\left(-\frac{1}{2(\sigma^2/n)} \left(\frac{1}{n} \sum_{i=1}^n y_i - \mu\right)^2\right)} \\ &= \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2} \left[ \left(\sum_{i=1}^n (y_i - \mu)^2\right) - \frac{1}{n} \left(\sum_{i=1}^n y_i - n\mu\right)^2 \right]\right) \\ &= \frac{1}{\sqrt{n}} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right]\right) \end{aligned}$$

## Deriving Sample Variance from the REML Likelihood

- Ignoring constant terms, the REML log-likelihood becomes

$$\log L_{REML} = -\frac{n-1}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 \right]$$

- By maximizing  $\log L_{REML}$ , we obtain the REML estimator  $\hat{\sigma}^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ .
- Hence we can see that the REML estimator adjusts for the degrees of freedom.
- The two estimators will be similar for large  $n$ , however, when the number of mean parameters (i.e. the number of parameters included in the mean part of the model) grows with sample size, the two estimators can be very different.

## Extended Likelihood Principle

- Birnbaum (1962) proved that the classical likelihood function contains all the information in the observed data about the fixed parameter.
- Bjørnstad (1996) extended this concept and showed that all the information in the data  $\mathbf{y}$  for parameters  $\theta$  and unobservables  $\mathbf{v}$  is in the extended likelihood.
- This means that inference about fixed parameters and unobservables, using the information only in the data, requires the extended likelihood function and nothing else. However, these likelihood principles do not show how the information in the data can be retrieved from the likelihood.



- In the absence of  $\mathbf{v}$ , the extended likelihood becomes the marginal likelihood. Fisher advocated the use of ML estimation and established the underlying theory.
- In the absence of  $\theta$ , we see that the extended likelihood gives Bayesian posterior and its use has been advocated by Bayesian statisticians.
- This gives an insight on how to make inferences in at least these two extreme cases, so that we may develop a procedure which gives identical inferences to that using the marginal likelihood for  $\theta$  and that exploiting the property of the predictive probability (posterior) for  $\mathbf{v}$  in these two extreme cases.
- In the context of HGLMs, Lee and Nelder (1996,2005) advocated the use of the h-likelihood and presented how information in the data for unobservables and parameters can be retrieved from it under the extended likelihood framework for all three types of objects ( $\theta$ ,  $\mathbf{v}$  and  $\mathbf{y}$ ).

- Similarly as for classical likelihood inference, we have a model for the data generation process and a corresponding likelihood.
- **Stochastic Model:**
  - Generate an instance of the random quantities  $\mathbf{v}$  from a probability function  $f_{\theta}(\mathbf{v})$ .
  - With  $\mathbf{v}$  fixed, generate an instance of the data  $\mathbf{y}$  from a probability function  $f_{\theta}(\mathbf{y}|\mathbf{v})$ .
  - The combined stochastic model is given by the product of  $f_{\theta}(\mathbf{v})f_{\theta}(\mathbf{y}|\mathbf{v})$ .
- **Statistical Inference:**
  - Given  $\mathbf{y}$ , we make inferences about  $\theta$  by using the marginal likelihood  $L(\theta; \mathbf{y}) \equiv f_{\theta}(\mathbf{y})$ .
  - Given  $\theta$ , we make inferences about  $\mathbf{v}$  by using the conditional likelihood

$$L(\theta, \mathbf{v}; \mathbf{v}|\mathbf{y}) \equiv f_{\theta}(\mathbf{v}|\mathbf{y}). \quad (8)$$

- The extended likelihood for unknowns  $(\mathbf{v}, \theta)$  is given by

$$L(\theta, \mathbf{v}; \mathbf{v}, \mathbf{y}) = L(\theta; \mathbf{y})L(\theta, \mathbf{v}; \mathbf{v}|\mathbf{y}), \quad (9)$$

where

$$\begin{aligned} L(\theta, \mathbf{v}; \mathbf{v}, \mathbf{y}) &\equiv f_{\theta}(\mathbf{v}, \mathbf{y}), \\ L(\theta; \mathbf{y})L(\theta, \mathbf{v}; \mathbf{v}|\mathbf{y}) &\equiv f_{\theta}(\mathbf{y})f_{\theta}(\mathbf{v}|\mathbf{y}). \end{aligned}$$

- The connection between these two processes is given by

$$f_{\theta}(\mathbf{y})f_{\theta}(\mathbf{v}|\mathbf{y}) \equiv L(\theta, \mathbf{v}; \mathbf{v}, \mathbf{y}) \equiv f_{\theta}(\mathbf{v}, \mathbf{y}) = f_{\theta}(\mathbf{v})f_{\theta}(\mathbf{y}|\mathbf{v}). \quad (10)$$

- In the extended likelihood framework,  $\mathbf{v}$  appears in stochastic model as random instances, but it appears in statistical inference as unknowns.
- From (9), we see that the extended likelihood is the product of two likelihoods, the Fisher likelihood  $f_{\theta}(\mathbf{y})$  and the conditional likelihood  $f_{\theta}(\mathbf{v}|\mathbf{y})$ .
- In likelihood theory the product of two likelihoods is a way of gathering information from the two independent source of data (Chapter 1).
- This is straightforward to note the close connection between the Fisher likelihood and the h-likelihood, because it uses the Fisher likelihood for inferences about  $\theta$ .

## Definition of the h-likelihood

- For continuous  $\mathbf{v}$ , Lee and Nelder (1996) proposed the use of the hierarchical likelihood, an extended likelihood limited to a pre-defined scale of  $\mathbf{v}$ .
- Suppose we have two models for the random effects in a linear predictor as

$$\eta_1 = \mathbf{X}\beta + \mathbf{v} \quad \text{and} \quad \eta_2 = \mathbf{X}\beta + \exp(\mathbf{v}).$$

- Then we have two alternative extended likelihoods based on two different scales of random effects:

$$L_1(\theta, \mathbf{v}; \mathbf{y}, \mathbf{v}) = f_\theta(\mathbf{y}|\eta_1)f_\theta(\mathbf{v}) \quad \text{and} \quad L_2(\theta, \mathbf{v}; \mathbf{y}, \mathbf{v}) = f_\theta(\mathbf{y}|\eta_2)f_\theta(\mathbf{v}). \quad (11)$$

- The modes of these two likelihoods differ and the question is which scale of random effects to use for statistical inferences.

## Definition of the h-likelihood

- We define the strong canonical scale of  $\mathbf{v}$  such that the random effects  $\mathbf{v}$  carry no information about the fixed effects  $\theta$  as

$$\frac{\exp\{\ell(\theta_1, \hat{\mathbf{v}}(\theta_1); y, \mathbf{v})\}}{\exp\{\ell(\theta_2, \hat{\mathbf{v}}(\theta_2); y, \mathbf{v})\}} = \frac{f_{\theta_1}(y)}{f_{\theta_2}(y)}$$

where  $\theta_1$  and  $\theta_2$  are two sets of  $\theta$  values and  $\hat{\mathbf{v}}(\theta_1)$  and  $\hat{\mathbf{v}}(\theta_2)$  are the modes of  $\ell(\theta_1, \mathbf{v}; y, \mathbf{v})$  and  $\ell(\theta_2, \mathbf{v}; y, \mathbf{v})$ , respectively. (Lee, Nelder and Pawitan, 2017).

- The h-likelihood is defined as the extended likelihood having  $\mathbf{v}$  on a canonical scale. This means that the marginal likelihood gives the same mode estimators about fixed effects as the h-likelihood, so that there is no conflict between classical likelihood inference and h-likelihood inference.
- For example, in linear mixed models,  $\mathbf{v}$  is on a canonical scale to  $\beta$  which implies that joint maximization of  $h$  with respect to  $\beta$  and  $\mathbf{v}$  gives the MLE for  $\beta$ .
- In HGLMs, the h-likelihood is defined under a weak canonical scale where the random effects combine additively with the fixed effects in a linear predictor.

## Definition of the h-likelihood

- From the linear predictor  $\eta_1$  above, we see that  $L_1(\theta, \mathbf{v}|\mathbf{y}, \mathbf{v})$  is the h-likelihood, which gives a consistent inference framework (Lee, Nelder and Pawitan, 2017)
- The likelihood  $L_1(\theta, \mathbf{v}; \mathbf{y}, \mathbf{v}) = f_\theta(\mathbf{y}|\eta_1)f_\theta(\mathbf{v})$  is called a hierarchical likelihood, as the random effects enter linearly in the linear predictor.
- An important difference between transforming fixed versus random effects is that a transformation of random effects requires the need to multiply the density function for the random effects with a Jacobian.
- In that sense, when  $\mathbf{v}$  is discrete, there is no Jacobian involved so that all extended likelihoods are the h-likelihood (Lee and Bjørnstad, 2013).

## Laplace approximation for the integrals

- For the linear mixed model both the marginal likelihood and REML likelihood are straight forward to derive, but for most other distributions the integral for the marginal likelihood has no analytical form.
- Numerical integration is infeasible if the number of integrands is large and MCMC algorithms are often too slow. As an alternative, we use Laplace approximation in h-likelihood approach.
- The (1st-order) Laplace approximation for some integral  $\int \exp[f(x)]dx$  is

$$\int \exp[f(x)]dx \approx \left\{ \left| -\frac{1}{2\pi} \frac{\partial^2 f(x)}{\partial x^2} \right|^{-\frac{1}{2}} \exp[f(x)] \right\} \bigg|_{x=x_0}$$

where  $x_0$  is a global maximum of  $f(x)$ .

## Laplace approximation for the integrals

- For the marginal likelihood, the Laplace approximation around the fitted random effects is

$$\int f_{\theta}(\mathbf{y}, \mathbf{v}) d\mathbf{v} = \int \exp(\log(f_{\theta}(\mathbf{y}, \mathbf{v}))) d\mathbf{v} \approx \left\{ \left| \frac{-\frac{\partial^2 \log(f_{\theta}(\mathbf{y}, \mathbf{v}))}{\partial \mathbf{v}^2}}{2\pi} \right|^{-\frac{1}{2}} f_{\theta}(\mathbf{y}, \mathbf{v}) \right\} \Big|_{\mathbf{v}=\hat{\mathbf{v}}}$$

where  $\hat{\mathbf{v}}$  is obtained from the mode of  $f_{\theta}(\mathbf{y}, \mathbf{v})$ .

- Applying a Laplace approximation to eliminate random effects together with a quadratic approximation around  $\hat{\beta}$  on the REML likelihood  $f_{\theta}(\mathbf{y}|\hat{\beta})$  to eliminate fixed effects, we get

$$f_{\theta}(\mathbf{y}|\hat{\beta}) \approx \dots \approx \left\{ \left| \frac{\mathbf{I}(\beta, \mathbf{v})}{2\pi} \right|^{-\frac{1}{2}} f_{\theta}(\mathbf{y}, \mathbf{v}) \right\} \Big|_{\beta=\hat{\beta}, \mathbf{v}=\hat{\mathbf{v}}}$$

where  $\mathbf{I}(\beta, \mathbf{v}) = - \begin{pmatrix} \frac{\partial^2 h}{\partial \beta^2} & \frac{\partial^2 h}{\partial \beta \partial \mathbf{v}} \\ \frac{\partial^2 h}{\partial \mathbf{v} \partial \beta} & \frac{\partial^2 h}{\partial \mathbf{v}^2} \end{pmatrix}$  with  $h \equiv \log(f_{\theta}(\mathbf{y}, \mathbf{v}))$ .



- To this end, Laplace approximation for the log-marginal likelihood is specified as an adjusted profile h-likelihood (APHL)

$$p_{\mathbf{v}}(h) = [h - \frac{1}{2} \log(|\mathbf{I}(\mathbf{v})|/2\pi)]|_{\mathbf{v}=\hat{\mathbf{v}}} \quad (12)$$

where  $\mathbf{I}(\mathbf{v})$  is the information matrix for the random effects, and  $\hat{\mathbf{v}}$  is the maximum h-likelihood estimator of the random effects using  $h$  as objective function.

- The approximation for the log-REML likelihood  $\log f_{\theta}(\mathbf{y}|\hat{\beta})$  can also be expressed as an APHL:

$$p_{\beta, \mathbf{v}}(h) = [h - \frac{1}{2} \log(|\mathbf{I}(\beta, \mathbf{v})|/2\pi)]|_{\beta=\hat{\beta}, \mathbf{v}=\hat{\mathbf{v}}} \quad (13)$$

where  $\mathbf{I}(\beta, \mathbf{v})$  is the information matrix for the fixed and random effects.

- The estimates of fixed effects and dispersion parameters are computed by maximizing these two likelihoods.

Here an example is presented to illustrate the fundamental idea of likelihood inference and how it may differ from Bayesian inference.

- A street magician has a small bag with a number of dice. There are two types of dice in the bag; white and blue. The white are numbered 1 to 6, while the blue have three sides with 1 and three sides with 2.
- The magician draws a dice at random from the bag without showing it to you and rolls the dice, then he claims that the number is 2.
  - (a) Which type of dice would you guess he has rolled, a white or a blue?
  - (b) The magician lets you bet on the color of the dice. Which odds would you accept?
  - (c) Now the magician informed that there are 20 white dices and 10 blue dices in the bag. What is your guess on the color of dice, which he rolled?

Solution of (a).

- The likelihood for a white dice is  $1/6$  and for a blue dice is  $1/2$ . Therefore, as a likelihoodist, the maximum likelihood guess is that the dice is blue.
- Let  $Y$  be the number of dice and let  $C$  be a colour of dice and  $c$  be a realized value of the colour of dice. Then, the likelihood ratio is

$$\frac{P(Y = 2|C = \textit{blue})}{P(Y = 2|C = \textit{white})} = \frac{1/2}{1/6} = 3.$$

Solution of (b).

- To be able to make a probability statement we need to know the distribution of the two types of dice in the bag. This is unknown however, which means that for a likelihoodist the odds cannot be computed.
- A Bayesian would guess the distribution and thereby compute the odds

$$\frac{P(Y = 2|C = \textit{blue})\pi(C = \textit{blue})}{P(Y = 2|C = \textit{white})\pi(C = \textit{white})} = 3 \frac{\pi(C = \textit{blue})}{\pi(C = \textit{white})} = \frac{P(C = \textit{blue}|Y = 2)}{P(C = \textit{white}|Y = 2)}.$$

Controversy is how to determine  $\pi(C = \textit{blue})$  and  $\pi(C = \textit{white})$ .

Solution of (c).

- The problem can be solved using a probabilistic argument, but here we also show that both a classical likelihood ratio and the ratio of extended likelihoods can be used to draw the same conclusion.
- Let  $c$  be a realized value of the colour of dice such that

$$L(c = \text{blue}) = P(C = \text{blue}) = 1/3 \quad \text{and} \quad L(c = \text{white}) = P(C = \text{white}) = 2/3.$$

Then, the ratio of extended likelihood is

$$\begin{aligned} \frac{L(c = \text{blue}, Y = 2)}{L(c = \text{white}, Y = 2)} &= \frac{P(Y = 2 | c = \text{blue})L(c = \text{blue})}{P(Y = 2 | c = \text{white})L(c = \text{white})} \\ &= \frac{1/2 \times 1/3}{1/6 \times 2/3} = \frac{3}{2}. \end{aligned}$$

Thus, the maximum extended likelihood guess is that the dice is blue.

- Furthermore, we can compute the conditional likelihood

$$\begin{aligned} L(c = \textit{blue} | Y = 2) &= \frac{L(c = \textit{blue}, Y = 2)}{L(Y = 2)} \\ &= \frac{P(Y = 2 | c = \textit{blue})L(c = \textit{blue})}{P(Y = 2 | c = \textit{blue})L(c = \textit{blue}) + P(Y = 2 | c = \textit{white})L(c = \textit{white})} \\ &= \frac{P(Y = 2 | c = \textit{blue})}{P(Y = 2 | c = \textit{blue}) + P(Y = 2 | c = \textit{white})L(c = \textit{white})/L(c = \textit{blue})} \\ &= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1/6 \times 2/3} = \frac{3}{5} \end{aligned}$$

and

$$L(c = \textit{white} | Y = 2) = \frac{2}{5}.$$

- We call  $L(c = \textit{white} | Y = 2)$  the predictive probability.

- Note that the conditional likelihood  $L(c = \text{blue} | Y = 2)$  depends upon the likelihood ratio  $L(c = \text{white})/L(c = \text{blue})$ , so that it is invariant with respect to the transformation of data and parametrization.
- Furthermore,

$$\frac{L(c = \text{blue}, Y = 2)}{L(c = \text{white}, Y = 2)} = \frac{L(c = \text{blue} | Y = 2)}{L(c = \text{white} | Y = 2)} = \frac{3}{2},$$

i.e. the mode of the conditional likelihood  $L(c | Y = 2)$  is the same as the mode of the extended likelihood  $L(c, Y = 2)$ .

- In (c) we have an information on  $P(C)$  (part of the model), while in (b) no information is available on  $P(C)$ , so that we need a guess  $\pi(C)$ .

- Inference on random effects have important practical use in predictions. A typical example is for instance if there are repeated observations on patients' hospital visits and the life time of these patients are to be predicted. This would require a survival analysis including random effects for patients and the uncertainty in the predictions will include the uncertainty of the fitted random effects.
- When  $\theta$  is known, we can make inferences about  $\mathbf{v}$  using  $f_{\theta}(\mathbf{v}|\mathbf{y})$ . However,  $\theta$  is unknown, so that we may make inferences using  $f_{\hat{\theta}}(\mathbf{v}|\mathbf{y})$  with  $\hat{\theta}$  being the ML estimator. This is the so-called EB approach, which gives consistent estimation for predictive probability because  $\hat{\theta}$  is consistent.



- However, in finite samples this approach often has a poor inferential performance because it cannot account for uncertainty, caused by estimating  $\theta$ ; especially when the number of observations is low and the number of parameters in  $\theta$  is large.
- Such an uncertainty about  $\hat{\theta}$  is included in  $f_{\theta}(\mathbf{y})$ , and can be used for inference on random effects (Lee and Nelder, 1996, 2001).
- Thus, an important question is how to eliminate the nuisance parameter  $\theta$  from the predictive probability  $f_{\theta}(\mathbf{v}|\mathbf{y})$ , using the information on  $\theta$  in the likelihood  $f_{\theta}(\mathbf{y})$ .
- Next slide illustrates the difference using a linear mixed model as an example.

- For a linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

$$\mathbf{v} \sim N(0, \lambda \mathbf{I})$$

$$\mathbf{e} \sim N(0, \phi \mathbf{I})$$

the h-likelihood is

$$\begin{aligned}\log(f_{\theta}(\mathbf{y}, \mathbf{v})) &= \log(f_{\theta}(\mathbf{y}|\mathbf{v})) + \log(f_{\theta}(\mathbf{v})) \\ &= -\frac{n}{2} \log(2\pi\phi) - \frac{1}{2\phi} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{v}) \\ &\quad - \frac{m}{2} \log(2\pi\lambda) - \frac{\mathbf{v}^T \mathbf{v}}{2\lambda}\end{aligned}$$

where  $n$  is the number of observations and  $m$  is the length of  $\mathbf{v}$ .

## H-likelihood and empirical Bayes

- In a linear mixed model, estimates of both  $\beta$  and  $\mathbf{v}$  can be computed by maximizing the h-likelihood.
- The score equations  $\frac{\partial h}{\partial \beta} = 0$  and  $\frac{\partial h}{\partial \mathbf{v}} = 0$  give Henderson's mixed model equations:

$$\begin{pmatrix} \frac{1}{\phi} \mathbf{X}^T \mathbf{X} & \frac{1}{\phi} \mathbf{X}^T \mathbf{Z} \\ \frac{1}{\phi} \mathbf{Z}^T \mathbf{X} & \frac{1}{\phi} \mathbf{Z}^T \mathbf{Z} + \mathbf{I} \frac{1}{\lambda} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} \frac{1}{\phi} \mathbf{X}^T \mathbf{y} \\ \frac{1}{\phi} \mathbf{Z}^T \mathbf{y} \end{pmatrix}$$

and the information matrix (computed from the second derivatives) is

$$\begin{pmatrix} \frac{1}{\phi} \mathbf{X}^T \mathbf{X} & \frac{1}{\phi} \mathbf{X}^T \mathbf{Z} \\ \frac{1}{\phi} \mathbf{Z}^T \mathbf{X} & \frac{1}{\phi} \mathbf{Z}^T \mathbf{Z} + \mathbf{I} \frac{1}{\lambda} \end{pmatrix}.$$

- The above equation is fitting algorithm of regression model

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \beta \\ \mathbf{v} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix}$$

where  $\mathbf{e}_1 \sim N(\mathbf{0}, \phi \mathbf{I})$  and  $\mathbf{e}_2 \sim N(\mathbf{0}, \lambda \mathbf{I})$ . This is called data augmentation method.

- This is different from an EB approach  $f_{\hat{\theta}}(\mathbf{v}|\mathbf{y})$  where the information matrix

$$\left( \frac{1}{\phi} \mathbf{Z}^T \mathbf{Z} + \mathbf{I} \frac{1}{\lambda} \right)$$

would typically be used for inference on the random effects ignoring the uncertainty in the estimates of  $\hat{\beta}$ .

- A more thorough exposition is found in Section 5.4 of Lee, Nelder and Pawitan (2017) showing that the h-likelihood gives correct inference.

- Because the Fisher likelihood  $f_{\theta}(y)$  does not involve  $v$ , the other component, the predictive probability,  $f_{\theta}(v|y)$  carries all the information in the data about the unobservables.
- Thus, the prediction of random effects can be made via the EB method using the estimated predictive probability (or posterior)

$$p(v|y) = f_{\hat{\theta}}(v|y) = \pi(v|y, \hat{\theta}),$$

where  $\hat{\theta}$  is the usual ML estimator (Carlin and Louis, 2000).

- However, using  $f_{\hat{\theta}}(v|y)$  to make inferences about  $v$  is naive and Bjørnstad (1990) has shown how badly it performs in measuring the true uncertainty in estimating  $v$ .

- Note that maximization of the h-likelihood

$$h = \log f_{\theta}(y|v) + \log f_{\theta}(v) = \log f_{\theta}(v|y) + \log f_{\theta}(y)$$

yields EB-mode estimators for  $v$ , without computing  $f_{\theta}(v|y) = f_{\theta}(y, v)/f_{\theta}(y)$ .

- However, the Hessian matrix (i.e. matrix of second derivatives) based upon  $f_{\theta}(v|y)$  gives a naive variance estimate for the prediction  $\hat{v}$  because it does not properly account for the uncertainty caused by estimating  $\theta$ , that is in  $f_{\theta}(y)$ .
- The h-likelihood considers both components and give proper estimators for random effects and their variance estimators. However, the estimation of the first two moments are not enough for accurate inferences of random effects if it is not normal.

- The previous example shows that  $\hat{v}$  is neither a consistent estimator of  $v$  nor follows the asymptotic normal distribution. Thus, interval estimations of random effects differ from those of fixed effects.
- Note that the predictive probability  $f_{\hat{\theta}}(v|y)$  gives an asymptotically correct inference. Thus, it is necessary to have a finite sample adjustment to account for information loss caused by estimating  $\theta$ . This can be generally done.
- Lee and Kim (2016) showed that

$$p(v|y) = E_{\hat{\theta}}(f_{\hat{\theta}}(v|y)) \equiv \int f_t(v|y)f(\hat{\theta} = t)dt = \int f_t(v|y)c(\theta = t)dt,$$

where  $c(\theta)$  is the confidence density in Chapter 1 of Lee, Nelder and Pawitan (2017).

- Because the bootstrap distribution gives an estimate of confidence density, we can have the bootstrap method to get the predictive probability

$$p(v|y) \equiv \frac{1}{B} \sum_{j=1}^B f_{\theta_j^*}(v|y),$$

where  $\theta_1^*, \dots, \theta_B^*$  are the bootstrap replicates of  $\hat{\theta}$ .

- In complex models it may not be easy to design the bootstrap scheme, so that it is convenient to generate the bootstrap replicates of  $\hat{\theta}$  from the asymptotic normal distribution of  $\hat{\theta}$  or the normalized likelihood.
- Via a simulation studies, Lee and Kim (2016) demonstrate that bootstrap methods provide excellent prediction intervals for future random effects, including the prediction of future outcomes in the front.



- Longitudinal data from a clinical trial of 59 epileptics (Thall and Vail, 1990)

**y** : seizure counts during 2-week periods before each of four visits to the clinic

**T** : 1(new drug), 0(placebo)

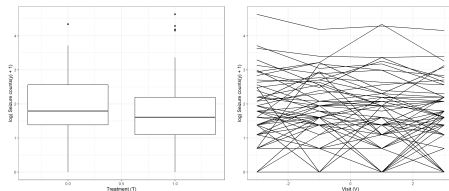
**B** : logarithm of the average number of epileptic seizures recorded in the 8-week period preceding the trial

**A** : logarithm of age

**V** : number of clinic visit(a linear trend, coded -3,-1,1,3)

**patient** : 59 patients

**id** : 236 data (= 59 patients  $\times$  4 clinic visits)



Model 1 : Poisson GLM

$$\log \mu_{ij} = \beta_0 + \beta_{B_i} x_B + \beta_{T_i} x_T + \beta_{A_i} x_A + \beta_{V_j} x_V + \beta_{B_i T_i} x_{BT}$$

Model 2 : Poisson - normal HGLM (GLMM)

Model 3 : Negative binomial - normal HGLM

Model 4 : Negative binomial - gamma HGLM

Model 5 : Over-dispersed Poisson GLM

Model 6 : Over-dispersed Poisson - normal HGLM

Model	cAIC	rAIC
Poisson GLM	1647.9	1664.7
Poisson - normal HGLM	1272.7	1350.5
NB - normal HGLM	1201.1	1310.5
NB - gamma HGLM	1163.9	1274.8
Over-dispersed Poisson GLM	1321.9	1332.8
Over-dispersed Poisson - normal HGLM	1219.4	1320.9

# Chapter 4. HGLMs: algorithm

## Introduction

- HGLMs extend GLMs by allowing random effects in the linear predictor.
- HGLMs also allow regression models for the residual variance and the variance for random effects.
- Lee, Nelder and Pawitan (2017) and Ha, Jeong and Lee (2017) described both the h-likelihood method and IWLS algorithm with related theories.
- In this chapter, we show how HGLMs can be fitted using interconnected and augmented GLMs.

- Consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

with  $\mathbf{e} \sim N(0, \boldsymbol{\Phi})$  where  $\boldsymbol{\Phi} = \text{diag}(\phi_i)$ .

- The ML estimator for  $\boldsymbol{\beta}$  is  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Phi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Phi}^{-1} \mathbf{y}$  and  $\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \boldsymbol{\Phi}^{-1} \mathbf{X})^{-1}$ .
- Now suppose that we have a regression model for the dispersion  $\phi_i$

$$g(\phi_i) = G_i \boldsymbol{\gamma}$$

where  $g(\cdot)$  is a link function and  $G_i$  is the  $i$ th row in a design matrix  $\mathbf{G}$ .

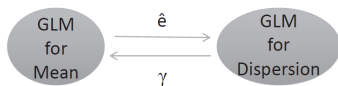
- The ML estimate of the regression coefficient of the dispersion  $\boldsymbol{\gamma}$  can be computed by using  $\hat{e}_i^2$  as response in a gamma GLM with mean  $\phi_i$ .
- The REML estimate can also be computed by using  $\hat{e}_i^2 / (1 - q_i)$  as response in a gamma GLM having a prior weight  $(1 - q_i)/2$  where  $q_i$  is the  $i$ th diagonal element in the hat matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \boldsymbol{\Phi}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Phi}^{-1}$ .

## Joint GLMs for mean and dispersion

- Suppose  $\mathbf{y}$  follows the GLM class of model  $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$  with  $E(y_i) = \mu_i$ ,  $\text{var}(y_i) = \phi_i V(\mu_i)$  and the dispersion  $\phi_i$  follows the regression model  $g(\phi_i) = G_i\gamma$  where  $g(\cdot)$  is a link function and  $G_i$  is the  $i$ th row in a design matrix.
- Given  $\phi_i$ , the ML estimator  $\hat{\boldsymbol{\beta}}$  can be obtained by using an IWLS algorithm for GLM model with prior weight  $1/\phi_i$ . (full algorithm is described in chapter 2)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{s}, \quad \text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1}$$

- Given  $\boldsymbol{\beta}$ , the ML estimate for the regression coefficient of the dispersion model  $\gamma$  can be computed by using the deviance  $d_i$  as response in a Gamma GLM with mean  $\phi_i$ .
- The REML estimate can be computed by using  $d_i/(1 - q_i)$  as response in a gamma GLM having a prior weight  $(1 - q_i)/2$  where  $q_i$  is the  $i$ th diagonal element in the hat matrix  $\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}$ .



- Consider the linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

- The model can be re-written as an augmented linear model

$$\mathbf{y}_a = \mathbf{X}_a\boldsymbol{\delta} + \mathbf{e}_a$$

$$\text{where } \mathbf{y}_a = \begin{pmatrix} \mathbf{y} \\ \mathbf{0} \end{pmatrix}, \mathbf{X}_a = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_m \end{pmatrix}, \boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{v} \end{pmatrix}, \mathbf{e}_a = \begin{pmatrix} \mathbf{e} \\ -\mathbf{v} \end{pmatrix}.$$

- The variance-covariance matrix of the augmented residual vector is given by

$$\text{var}(\mathbf{e}_a) \equiv \mathbf{W}^{-1} = \begin{pmatrix} \phi \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \lambda \mathbf{I}_m \end{pmatrix}$$

- Data augmentation method can be used to fit random effects  $\mathbf{v}$ .

- The estimates from weighted least squares are given by

$$\mathbf{X}_a^T \mathbf{W} \mathbf{X}_a \hat{\delta} = \mathbf{X}_a^T \mathbf{W} \mathbf{y}_a$$

which is identical to Henderson's mixed model equations.

- So we can extend the estimation method for joint GLMs to joint GLMs including random effects by augmenting the response vector.
- The weight matrix  $\mathbf{W}$  may then be updated using the estimated variance components and the algorithm iterates until convergence.
- Lee and Nelder (2001) showed that the augmented linear model can be extended to fit the HGLM class of models.

- HGLM has two random components: a response  $\mathbf{y}$  and unobserved random effect  $\mathbf{v}$ , such that  $\mathbf{y}|\mathbf{v}$  follows a GLM distribution, namely normal, binomial, Poisson, or gamma.
- The expectation of the conditional model  $\mathbf{y}|\mathbf{v}$  is

$$E(\mathbf{y}|\mathbf{u}) = \boldsymbol{\mu}$$

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v}$$

$$\mathbf{v} = r(\mathbf{u})$$

where  $g(\cdot)$  is a link function,  $\mathbf{X}$  and  $\mathbf{Z}$  are design matrices and  $\boldsymbol{\beta}$  is a fixed effect.

- The distribution of  $\mathbf{u}$  is one of the conjugate distributions of GLM family: normal, beta, gamma, or inverse-gamma.
- The random effect  $\mathbf{v}$  is given on an appropriate (weak canonical) scale through the link function  $r(\cdot)$  transforming  $\mathbf{u}$  to guarantee correct model estimator.



- Consider the heteroscedastic linear mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}$$

with independent and heteroscedastic random effects  $v_i \sim N(0, \lambda_i)$  and residuals  $e_i \sim N(0, \phi_i)$ .

- Now we can allow GLMs for the dispersion (residual variance) and random effect variance

$$g_1(\phi_i) = G_{1i}\gamma_1$$

$$g_2(\lambda_i) = G_{2i}\gamma_2$$

- By taking log link for these variance components, we avoid negative estimates for variance components.
- Data augmentation method is used to fit the model.

- The REML estimates for  $\gamma_1$  can be obtained by applying a gamma GLM to the response  $\hat{e}_i^2/(1 - q_i)$  with weights  $(1 - q_i)/2$  for  $i = 1, 2, \dots, n$
- Those for  $\gamma_2$  are computed by applying a gamma GLM to the response  $\hat{v}_i^2/(1 - q_i)$  for  $i = n + 1, n + 2, \dots, n + m$
- The hat value  $q_i$  are obtained from the hat matrix of the augmented model.

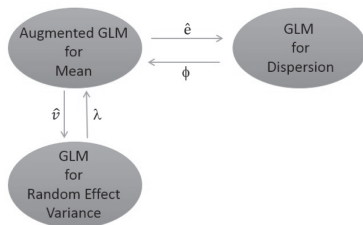


Figure 4.5 Interconnected GLMs for HGLMs with structured dispersions.

- Now we consider a HGLM such that
  - $y|v$  follows a GLM distribution with  $\eta = \mathbf{X}\beta + \mathbf{Z}v$ .
  - $u$  follows any conjugate distribution of GLM family with  $g_1(\phi_i) = G_{1i}\gamma_1$  and  $g_2(\lambda_i) = G_{2i}\gamma_2$ .
- Then the REML estimates for  $\gamma_1$  can be obtained by applying a gamma GLM to the response  $d_i/(1 - q_i)$  with weights  $(1 - q_i)/2$  where  $d_i$  is the deviance from  $y|v$  GLM for  $i = 1, \dots, n$ .
- Those for  $\gamma_2$  are computed by applying a gamma GLM to the response  $d_i/(1 - q_i)$  where  $d_i$  is the deviance from  $v$  GLM for  $i = n + 1, \dots, n + m$ .
- We allow various GLMs to  $y$  and  $\psi$  in the augmented response  $\begin{pmatrix} y \\ \psi \end{pmatrix}$  to fit random effect model.
- We use inter-connected JGLM fit for mean and dispersion of  $\phi$  and  $\lambda$ .

### Review : Poisson GLM

**Model** :  $\log \mu_{ij} = \beta_0 + \beta_{B_i} x_B + \beta_{T_i} x_T + \beta_{A_i} x_A + \beta_{V_j} x_V + \beta_{B_i T_i} x_{BT}$

### Over-dispersed Poisson GLM

- Poisson GLM gives a deviance of 869.9 with degrees of freedom 230, clearly indicating over-dispersion. To accommodate this, we may fit the over-dispersed Poisson model with  $\text{var}(y) = \phi\mu$ .
- For the parameter estimation of  $\phi$ , we may use the deviance or Pearson chi-squared statistic.
- From the deviance we have  $\hat{\phi} = 3.8 = \exp(1.33) = 869.9/230$
- From the Pearson chi-squared statistic we have  $\hat{\phi} = 4.5 = \exp(1.505) = 1036.3/230$

- Because the deviance residuals are the best normalizing transformation under the exponential family, it gives an estimator with small variance, but it gives an inconsistent estimate.
- Hilbe (2014) recommended to use the Pearson chi-squared statistics because it gives a consistent estimator.
- In finite sample, the deviance often gives more efficient estimators (Nelder and Lee, 1992). Thus, it is recommended to use the deviance in small samples.
- Correlation among repeated measures should be considered, so HGLM should be used for further analysis.

- An industrial Taguchi experiment was performed to study the influence of several controllable factors on the mean value and the variation in the percentage of shrinkage of products made by injection molding (Engel, 1992).

y : percentage of shrinkage of products made by injection molding

Controllable factors	Noise factors
A : cycle time	M : percentage regrind
B : mould temperature	N : moisture content
C : cavity thickness	O : ambient temperature
D : holding pressure	
E : injection speed	
F : holding time	
G : gate size	

- This dataset has been attended by many researchers because the model checking plots were not satisfactory.
- Lee and Nelder (1997) gave extensive discussion on how to choose a good model and presented the heteroscedastic log-linear model.

### Heteroscedastic log-linear model

- Model with log-normal distribution and the identity link  $\eta = \mu$

#### Mean Model

$$\eta = \beta_0 + \beta_A A + \beta_C C + \beta_D D + \beta_E E + \beta_G G + \beta_N N + \beta_{C \cdot N} C \cdot N + \beta_{E \cdot N} E \cdot N$$

#### Dispersion Model

$$\log \phi = \gamma_0 + \gamma_A A + \gamma_F F$$

## Review : Gamma GLM

**Model** :  $\eta = \log \mu = \alpha + \beta \text{ crack0}$

## Gamma GLM with structured dispersion

- We may estimate  $\phi$  either based on deviance or Pearson chi-squared statistic.
- In this example, degrees of freedom is large (239). We may prefer the Pearson chi-squared statistic in estimating  $\phi$ .

**Mean Model** :  $\eta = \log \mu = \beta_0 + \beta_1 \text{ crack0}$

**Dispersion Model** :  $\log \phi = \gamma_0 + \gamma_1 \text{ cycle}$



- Tests of the presence of the bacteria *H. influenzae* in children with otitis media in the Northern Territory of Australia (MSHR 1999–2000 Annual Report).

**y** : 1(presence), 0(absence)

**ap** : a(active), p(placebo)

**hilo** : hi(high compliance), lo(low compliance)

**week** : number of week at test (0,2,4,6,11)

**ID** : subject ID

**trt** : placebo, drug(a & lo), drug+(a & hi)

### Binomial GLMM

- $p_{ij} = P(y_{ij} = 1 | v_i)$
- $v_i \sim N(0, \lambda)$

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_1 I(i = \text{drug}) + \beta_2 I(i = \text{drug}+) + v_i$$

## Chapter 5. HGLMs: Modeling

- In this chapter, a number of dataset are modeled using HGLMs.
- In the first few example we show analyses using normal, log-normal, gamma, Poisson, and binomial HGLMs.
- Thereafter, examples using HGLMs including structured dispersion are given.
- We also fit models with correlated random effects, including spatial models.

- Experiment on the preparation of chocolate cakes, conducted at Iowa State College (Cochran and Cox, 1957).

Replicate : 15 replications

Batch : 3 batters

Recipe : R1(Recipe I), R2(Recipe II), R3(Recipe III)

Temperature : 6 different baking temperatures ( $175^{\circ}\text{C} \sim 225^{\circ}\text{C}$ )

Angle : breaking angle

inter :  $\text{Batch}^2$

logAngle : logarithm of Angle

## Normal linear mixed model

- $i = 1, 2, 3$  for recipes,  $j = 1, \dots, 6$  for temperatures and  $k = 1, \dots, 15$  for replicates.
- $y_{ijk} | v_i, v_{ik} \sim N(\mu_{ijk}, \sigma^2)$

$$\mu_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + v_k + v_{ik}$$

## Log-normal linear mixed model

- The same model but with responses  $\log y_{ijk}$  gives a better fit.

$$\log \mu_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + v_k + v_{ik}$$

## Gamma GLMM

- $y_{ijk} | v_i, v_{ik} \sim \text{Gamma}\left(\frac{1}{\phi}, \frac{1}{\mu_{ijk}\phi}\right)$

$$\log \mu_{ijk} = \mu + \gamma_i + \tau_j + (\gamma\tau)_{ij} + v_k + v_{ik}$$

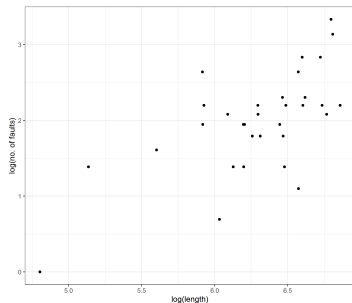
- Fabric data (Bissell, 1972).

**l** : fabric length

**y** : number of faults in a bolt of fabric

**rf** : 32 observations

**x** : logarithm of fabric length



## Poisson GLM

- $y \sim \text{Poi}(\mu)$  and  $x = \log l$

$$\log \mu = \alpha + \beta x$$

- Deviance = 64.5 with 30 df: over-dispersion
- It may be caused by the assumed Poisson regression model being incorrect (Azzalini et al., 1989 and Firth et al., 1991).

## Poisson-gamma HGLM

- Bissell(1972) proposed the use of the negative binomial model, which can be fitted via a Poisson HGLM.
- $y|u \sim \text{Poi}(\mu)$
- When  $u$  follows the gamma distribution with  $E(u) = 1$  and  $\text{var}(u) = \lambda$ ,

$$\log \mu = \alpha + \beta x + \log u$$

### Review : Poisson GLM

- $y \sim Poi(\mu)$

$$\log \mu = \log t + \alpha + \beta x$$

### Poisson-gamma HGLM

- Fitting the data assuming a Poisson GLM, there exist two outliers which give marginally significant lack of fit.
- we fit a negative binomial model via a Poisson-gamma HGLM with saturated random effects for full response, number of train accidents.
- $y|u \sim Poi(\mu)$
- When  $u$  follows the gamma distribution with  $E(u) = 1$  and  $var(u) = \lambda$ ,

$$\log \mu = \log t + \alpha + \beta x + \log u$$

- Three experiments were conducted : two were done with the same salamanders in the summer and autumn and another on in the autumn of the same year using different salamanders (McCullagh and Nelder, 1989).
- In each experiment, 20 females and 20 males were paired six times for mating with individuals from their own and the other population, resulting in 120 observations in each experiment.

Season : Summer, Autumn

Experiment : 3 experiments

TypeM : type of male. 1(whiteside), 0(rough butt)

TypeF : type of female. 1(whiteside), 0(rough butt)

Cross : TypeM  $\times$  TypeF

Male : 60 males (20 males for each experiment)

Female : 60 females (20 females for each experiment)

Mate : success of mating. 1(success), 0(failure)



## Binomial GLMM

- $i, j = 1, \dots, 20$  and  $k = 1, 2, 3$
- $y_{ijk}$  : The outcome (Mate) for the mating of the  $i$ -th female with the  $j$ -th male in the  $k$ -th experiment.
- $p_{ijk} = P(y_{ijk} = 1 | v_{ik}^f, v_{jk}^m)$
- $v_{ik}^f \sim N(0, \sigma_f^2)$ ,  $v_{jk}^m \sim N(0, \sigma_m^2)$

$$\log \left( \frac{p_{ijk}}{1 - p_{ijk}} \right) = \beta_0 + F_i + M_j + (FM)_{ij} + v_{ik}^f + v_{jk}^m$$

- There have been many methods developed to obtain approximate ML estimators.
- Noh and Lee (2007) showed that HL(1,2) has the smallest bias while HL(1,1) is fast with results as follows.

- The width of lines made by a photoresist-nanoline tool were measured in five different locations on silicon wafers, measurement being taken before and after and etching process being treated separately (Phadke et al, 1983).
- 9 experimental factors (A-I) arranged in an  $L_{18}$  orthogonal array and produced 33 measurements at each of 5 locations, giving a total of 165 observations.

Width : width of line

Wafer : 33 silicon wafers

Experimental factors	
A : mask dimension	F : aperture
B : photoresist viscosity	G : exposure time
C : spin speed	H : developing time
D : bake temperature	I : etch time
E : bake time	

## Linear Mixed Models with structured dispersion

- $q$ : index for wafers(1~33),  $r$ : index for observations within wafers
- $i, j, k, l, m, n, o, p$ : index for A-H
- $v_q \sim N(0, \lambda)$  and  $e_{qr} \sim N(0, \phi)$

$$y_{ijkop,qr} = \beta_0 + a_i + b_j + c_k + g_o + h_p + v_q + e_{qr}$$

- $\lambda$  and  $\phi$  represent the between-wafer and within-wafer variances respectively, which can be affected by the experimental factors.
- The dispersion and random effect variance can be modeled as

$$\log \phi_{imno} = \gamma_0^\omega + a_i^\omega + e_m^\omega + f_n^\omega + g_0^\omega$$

$$\log \lambda_m = \gamma_0^b + e_m^b$$

- Designed experiment in a semiconductor plant, which is of interest to study the curvature or camber of the substrate devices produced in the plant (Myers et al., 2002).
- There is a lamination process, and the camber measurement is made four times on each device produced.

Device : 16 devices

x1-x6 : 6 employed factors (each design variable is taken at 2 levels)

y : camber taken in  $10^{-4}$  in./in.

Gamma HGLM with structured dispersion

- When  $y|v \sim \text{Gamma}$  with  $E(y|v) = \mu$  and  $\text{Var}(y|v) = \phi\mu^2$ ,

$$\log \mu = \beta_0 + x_1\beta_1 + x_3\beta_3 + x_5\beta_5 + x_6\beta_6 + v$$

$$\log \phi = \gamma_0 + x_2\gamma_2 + x_3\gamma_3$$

- Data from a clinical trial comparing two treatments for a respiratory illness (Strokes et al., 1995)
- In each of two medical centers, 111 patients were randomly assigned to active treatment (54) or placebo (57). During treatment, respiratory status was determined at 4 visits.

`y` : respiratory status during treatment. 1(good), 0(poor)

`patient` : 111 patients

`treatment, trt` : 1(active treatment), 0(placebo)

`sex, msex` : 1(male), 0(female)

`age` : age of patients

`center` : 2 medical centers

`baseline, base` : baseline respiratory status. 1(good), 0(poor)

`past` : respiratory status for last visit. 1(good), 0(poor)

## Binomial HGLM with structured dispersion

- $i = 1, \dots, 111$  and  $j = 1, 2, 3, 4$
- $p_{ij} = P(y_{ij} = 1 | v_i, y_{i(j-1)})$
- $v_i \sim N(0, \lambda_i)$

$$\begin{aligned} \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = & \beta_0^{(\mu)} + \beta_1^{(\mu)} \text{trt}_i + \beta_2^{(\mu)} \text{mse}_i + \beta_3^{(\mu)} \text{age}_i \\ & + \beta_4^{(\mu)} \text{center}_i + \beta_5^{(\mu)} \text{base}_i + \beta_6^{(\mu)} y_{i(j-1)} + v_i \end{aligned}$$

- The random effects have a structured dispersion.

$$\log \lambda_i = \beta_0^{(\lambda)} + \beta_1^{(\lambda)} \text{age}_i$$

- Data contain the growth measurement of 27 childrens from age 8 until age 14 (Pinheiro and Bates, 2000).
- Every two years, the distance between the pituitary and the pterygomaxillary fissure was recorded using x-ray images of the skull.

**distance** : distance of the subject (mm)

**age** : age (8, 10, 12, 14)

**Subject** : 16 male(boys) and 11 female(girls)

**Sex** : Male, Female

**M** : 1(Male), 0(Female)

**Mage** :  $M \times \text{age}$

**F** : 1(Female), 0(Male)

**Fage** :  $F \times \text{age}$

## Correlated random intercept and slope model

- $y_{ij}$ : distance of the  $i$ -th subject at the  $j$ -th age  $A_{ij}$
- $e_{ij} \sim N(0, \phi_{ij})$
- The random intercept  $v_{1i}$  and random slope  $v_{2i}$  are assumed to be bivariate normal distribution.  $(v_{1i}, v_{2i})^\top \sim \text{BVN} \left( 0, \begin{pmatrix} \lambda_1 & \rho \sqrt{\lambda_1 \lambda_2} \\ \rho \sqrt{\lambda_1 \lambda_2} & \lambda_2 \end{pmatrix} \right)$

$$y_{ij} = \beta_1 F_i + \beta_2 F_i A_{ij} + \beta_3 M_i + \beta_4 M_i A_{ij} + v_{1i} + A_{ij} v_{2i} + \epsilon_{ij}$$



- Clayton and Kaldor (1987) analyzed observed and expected numbers of lip cancer cases in the 56 administrative areas of Scotland with a view to produce a map that would display regional variation in cancer incidence and yet avoid the presentation of unstable rates for the smaller areas.
- Presumably the spatial aggregation is due in large part to the effects of environmental risk factors.

$\log E_n$  : Logarithm of expected numbers of lip cancer cases

$O_y$  : Observed numbers of lip cancer cases

$P_{aff}$  : The percentage of the work force in each area employed in agriculture, fishing, or forestry.

$county$  : 56 administrative areas

$x$  :  $P_{aff}/10$

## Poisson HGLM

- $y_i | v_i \sim \text{Poi}(\mu_i)$

$$\log \mu_i = \log n_i + \beta_0 + \beta_1 x_i / 10 + v_i$$

- The random effect  $v_i$  represented unobserved area-specific log-relative risks. They tried 3 models.

M1  $v_i \sim N(0, \lambda)$

M2  $v_i \sim$  intrinsic autoregressive model (IAR)

M3  $v_i \sim$  MRF in which  $\text{Var}(v)^{-1} = (I - \rho M) / \lambda$ , where  $M$  is the incidence matrix for neighbours.

- Lee and Nelder (2001) chose the model M3 as best.
- The MRF model with  $\rho = 0$  is the M1 model.
- MRF with  $\hat{\rho} = 0.174$  provides a suitable model.
- We found that the main difference between M1 and M3 is the prediction for county 49, which has the highest predicted value because it has the largest  $n_i$ . This gives the very large leverage value (or hat value) of 0.92.
- Though model checking plots are useful, our eyes could be misled, so that objective criteria based upon the likelihood are also required in the model selection.

- Dataset describes prevalence of infection by the nematode *Loa loa* in North Cameroon, 1991-2001 (Rousset et al., 2016).
- The study investigated the relationship between altitude, vegetation indices, and prevalence of the parasite.

id, LOC : 197 locations

longitude : longitude of locations

latitude : latitude of locations

y : number of infected individuals at location

n : number of individuals at location

x1 : altitude (m)

x2-x4 :  $x2 = \max(x1 - 650, 0)$ ,  $x3 = \max(x1 - 1000, 0)$ ,  $x4 = \max(x1 - 1300, 0)$

x5 : maximum normalized-difference vegetation index (NDVI) from repeated satellite scans

x6 : standard error of NDVI

## Binomial HGLM with the logit link

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

$$\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + v_i$$

- The random effect  $v_i$  is for the  $i$ -th location. Rousset et al. (2016) fitted HGLMs

**M1**  $v_i \sim \text{independent } N(0, \lambda)$

**M2**  $v_i \sim \text{normal distribution with variance } \lambda \text{ and Matern correlation for two locations which is represented by}$

$$(1 - \text{Nugget}) \frac{(\rho d)^\nu K_\nu(\rho d)}{2^{\nu-1} \Gamma(\nu)}$$

- Nugget: parameter describing a discontinuous decrease in correlation at zero distance
- $\rho$ : scaling parameter,  $\nu$ : smoothness parameter
- $K_\nu$ : bessel K function of order  $\nu$  and  $d$  is distance computed by longitudes and latitudes for two locations

- Durbin and Koopman (2000) analyzed the lagged quarterly demand for gas in the UK from 1960 to 1986.

`y` : Lagged quarterly demand for gas

`year` : 1960-1986

`quarter` : q1-q4

`time` : 108 times = 27 years  $\times$  4 quarter

`t43, t44` : 1 if time=43 or time=44

`cos1, sin1` :  $\cos(2\pi t/104)$  and  $\sin(2\pi t/104)$  ( $t$  : time)

## Model 1 for gas data

- Durbin and Koopman (2000) considered a local linear-trend model with quarterly seasonals which can be represented as a normal HGLM.
- $f_t = \sum_{j=1}^t r_j$  and  $s_t = \sum_{j=1}^t (t-j+1)p_j$  are random effects for the local linear trend, the quarterly seasonal  $q_t$  with  $w_t = \sum_{j=0}^3 q_{t-j}$ .
- $r_t \sim N(0, \lambda_r)$ ,  $p_t \sim N(0, \lambda_p)$ ,  $w_t \sim N(0, \lambda_w)$ ,  $e_t \sim N(0, \phi_t)$

$$y_t = \alpha + f_t + s_t + q_t + e_t$$

- Lee, Nelder, and Pawitan (2017) add a linear trend  $\beta t$  and found that the random walk  $f_t$  is not necessary. Thus, they considered a model

$$y_t = \alpha + \beta t + s_t + q_t + e_t$$

- The residual plot displays apparent outliers, caused by a disruption in the gas supply in the 3rd and 4th quarters of 1970.

## Model 2 for gas data

- Lee, Nelder, and Pawitan (2017) proposed to delete the random quarterly seasonals and add further fixed effects to model the 1970 disruption and seasonal effects.

$$y_t = \alpha + t\beta + \alpha_i + t\beta_i + \delta_1 I(t = 43) + \delta_2 I(t = 44) \\ + \gamma_1 \sin(2\pi t/104) + \gamma_2 \cos(2\pi t/104) + s_t + e_t$$

- Lee, Nelder, and Pawitan (2017) further found extra dispersion in the 3rd and 4th quarters, which led to a structured dispersion model.

$$\log \phi_t = \varphi + \psi_i$$



- Prestige data from R package "car" (Fox et al., 2016)

id : jobs

education : average education of occupational incumbents (year)

income : average income of incumbents (\$)

women : percentage of incumbents who are women

prestige : Pineo-Porter prestige score for occupation

census : Canadian Census occupational code

type : type of occupation. bc(Blue Collar), prof(Professional, Managerial, and Technical), wc(White Collar), NA

### Additive non-parametric regression model

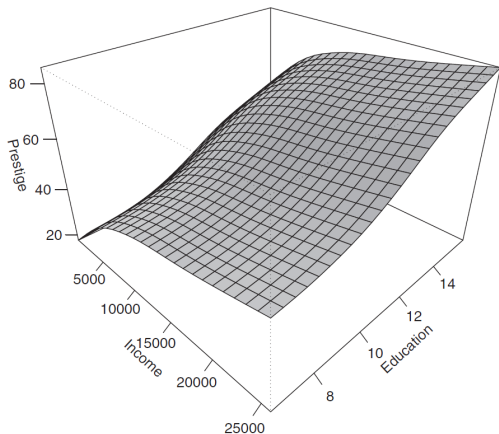
- $f_1(\cdot)$  and  $f_2(\cdot)$  are unknown functions.
- $e_i \sim N(0, \sigma^2)$

$$y_i = f_1(x_{1i}) + f_2(x_{2i}) + e_i$$

- Suppose that cubic smoothing splines are used to fit these unknown functions  $f_1(\cdot)$  and  $f_2(\cdot)$ , which are characterized by singular precision matrices,  $P_1$  and  $P_2$ , respectively (Lee, Nelder, and Pawitan, 2017).
- This additive model can be fitted by using an HGLM.

$$y_i = x_i^T \beta + v_{1i} + v_{2i} + e_i$$

- $x_i^T = (1, x_{1i}, x_{2i})$ ,  $v_1 \sim N(0, P_1^+)$  and  $v_2 \sim N(0, P_2^+)$  are random effects with  $P^+$  being the Moore-Penrose inverse of  $P$ .



- The regression surface  $\hat{f}_1(x_{1i}) + \hat{f}_2(x_{2i})$  from the additive model shows that prestige increases with income and education.

## Chapter 6. DHGLMs

- We represent a DHGLM as  $\{\text{model}(\mu), \text{model}(\phi)\}$
- The original GLM:  $\{\text{GLM}(\mu), \text{constant}\}$

$$\eta^{(\mu)} = g^{(\mu)}(\mu) = \mathbf{X}^{(\mu)}\beta^{(\mu)}$$

- The joint GLM:  $\{\text{GLM}(\mu), \text{GLM}(\phi)\}$

$$\eta^{(\mu)} = g^{(\mu)}(\mu) = \mathbf{X}^{(\mu)}\beta^{(\mu)}$$

$$\eta^{(\phi)} = g^{(\phi)}(\phi) = \mathbf{X}^{(\phi)}\beta^{(\phi)}$$

- The HGLM:  $\{\text{HGLM}(\mu), \text{constant}\}$

$$\eta^{(\mu)} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}$$

- The HGLM with structured dispersion:  $\{\text{HGLM}((\mu)), \text{GLM}(\phi)\}$

$$\eta^{(\mu)} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}$$

$$\eta^{(\phi)} = \mathbf{X}^{(\phi)}\beta^{(\phi)}$$

$$\eta^{(\lambda)} = \mathbf{X}^{(\lambda)}\beta^{(\lambda)}$$

- The DHGLM:  $\{\text{HGLM}(\mu), \text{HGLM}(\phi)\}$

$$\eta^{(\mu)} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}$$

$$\eta^{(\lambda)} = \mathbf{X}^{(\lambda)}\beta^{(\lambda)}$$

$$\eta^{(\phi)} = \mathbf{X}^{(\phi)}\beta^{(\phi)} + \mathbf{Z}^{(\phi)}\mathbf{v}^{(\phi)}$$

$$\eta^{(\alpha)} = \mathbf{X}^{(\alpha)}\beta^{(\alpha)}$$

- The DHGLM:  $\{\text{DHGLM}(\mu), \text{GLM}(\phi)\}$

$$\eta^{(\mu)} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}$$

$$\eta^{(\lambda)} = \mathbf{X}^{(\lambda)}\beta^{(\lambda)} + \mathbf{Z}^{(\lambda)}\mathbf{v}^{(\lambda)}$$

$$\eta^{(\tau)} = \mathbf{X}^{(\tau)}\beta^{(\tau)}$$

$$\eta^{(\phi)} = \mathbf{X}^{(\phi)}\beta^{(\phi)}$$

- The DHGLM:  $\{\text{DHGLM}(\mu), \text{HGLM}(\phi)\}$

$$\eta^{(\mu)} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}$$

$$\eta^{(\lambda)} = \mathbf{X}^{(\lambda)}\beta^{(\lambda)} + \mathbf{Z}^{(\lambda)}\mathbf{v}^{(\lambda)}$$

$$\eta^{(\tau)} = \mathbf{X}^{(\tau)}\beta^{(\tau)}$$

$$\eta^{(\phi)} = \mathbf{X}^{(\phi)}\beta^{(\phi)} + \mathbf{Z}^{(\phi)}\mathbf{v}^{(\phi)}$$

$$\eta^{(\alpha)} = \mathbf{X}^{(\alpha)}\beta^{(\alpha)}$$

### Review : joint GLM

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1}$$

$$\eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)} + \beta_1^{(\phi)} t_j$$

### DHGLM

- when  $v_i^{(\mu)} \sim N(0, \lambda)$  and  $v_i^{(\phi)} \sim N(0, \alpha)$ ,

$$\eta_{ij}^{(\mu)} = \log \mu_{ij} = \beta_0^{(\mu)} + \beta_1^{(\mu)} l_{ij-1} + v_i^{(\mu)}$$

$$\eta_{ij}^{(\phi)} = \log \phi_{ij} = \beta_0^{(\phi)} + \beta_1^{(\phi)} t_j + v_i^{(\phi)}$$

- cAIC selects this DHGLM as the best-fitting model.
- We can conclude that heteroscedasticity between metallic specimens exists significantly in the mean as well as in the dispersion.

## ex. Crack growth continued

- By using the studentized deviance residuals, we can obtain model-checking plots of the model objects.

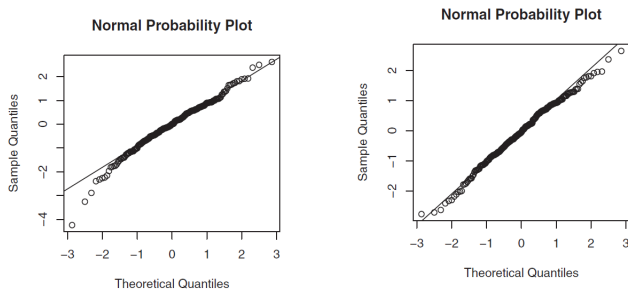


Figure: Normal probability plots for HGLM and DHGLM

- Most of the outliers in HGLM, caused by abrupt changes among repeated measures, disappear when random effects are allowed in the model for the residual variance.

## Review : HGLM

$$\begin{aligned}y_t &= \alpha + t\beta + \alpha_i + t\beta_i + \delta_1 I(t = 43) + \delta_2 I(t = 44) \\&\quad + \gamma_1 \sin(2\pi t/104) + \gamma_2 \cos(2\pi t/104) + s_t + e_t \\ \log \phi_t &= \varphi + \psi_i\end{aligned}$$

## DHGLM

- Consider the follow DHGLM, allowing heavy-tailed distribution for  $e_t$
- When  $v_t^{(\phi)} \sim N(0, \alpha)$ ,

$$\begin{aligned}y_t &= \alpha + t\beta + \alpha_i + t\beta_i + \delta_1 I(t = 43) + \delta_2 I(t = 44) \\&\quad + \gamma_1 \sin(2\pi t/104) + \gamma_2 \cos(2\pi t/104) + s_t + e_t \\ \log \phi_t &= \varphi + \psi_i + v_t^{(\phi)}\end{aligned}$$

- cAIC selects DHGLM as the best-fitting model.



## ex. Gas consumption continued

- The likelihood-ratio test for  $H_0 : \alpha = 0$ , based on the restricted likelihood, rejects the null hypothesis (deviance difference :  $18.8 > \chi^2_{2\delta}(1) = 2.71$  with significant level  $\delta = 0.05$ )

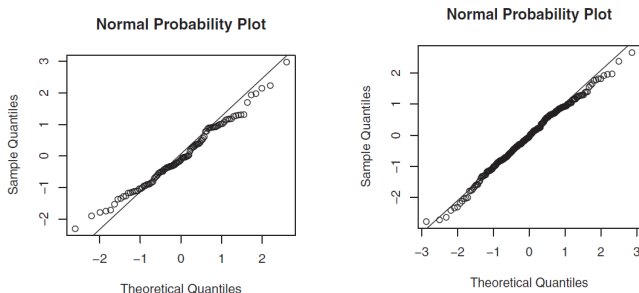


Figure: Normal probability plots for HGLM and DHGLM

- We see that a big outlier in HGLM disappeared under the DHGLM.

- Daily observations for the weekday closing exchange rates for the U.K. Sterling/U.S. Dollar from 1/10/81 to 28/6/85 (Harvey et al., 1994).

**rt** : Exchange rate at time  $t$

**yt** : Mean-corrected returns.  $y_t = 100 \left( \log(r_t/r_{t-1}) - \frac{1}{n} \sum_{i=1}^n \log(r_i/r_{i-1}) \right)$

**yt1** :  $y_{t-1}$

**yt12** :  $y_{t-1}^2$

**date** : 936 observations

- Consider the model

$$y_t = \sqrt{\phi_t} z_t$$

where  $z_t$  is the standard normal random variable and  $\phi_t$  is a volatility at time  $t$ .

## ARCH(1) model

- Engle (1982) introduced the ARCH model of order 1.

$$\phi_t = \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2$$

- This is a joint GLM  $\text{GLM}(\mu = 0)$ ,  $\text{GLM}(\phi)$ , which can be fitted by specifying the identity link function for  $\text{GLM}(\phi)$  and fixing the mean null.

## GARCH(1,1) model

- The ARCH(1) model was extended to the GARCH(1,1) model by Bollerslev (1986).

$$\phi_t = \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2 + \gamma \phi_{t-1}$$

- By letting  $\beta_0^{*(\phi)} = \beta_0^{(\phi)} / (1 - \rho)$  with  $\rho = \beta_1^{(\phi)} + \gamma$ ,

$$\begin{aligned}
 v_t^{(\phi)} &= \phi_t - \beta_0^{*(\phi)} \\
 &= \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2 + \gamma \phi_{t-1} - \beta_0^{*(\phi)} \\
 &= \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2 + \rho(\phi_{t-1} - \beta_0^{*(\phi)}) - \beta_1^{(\phi)} \phi_{t-1} - (1 - \rho)\beta_0^{*(\phi)} \\
 &= \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2 + \rho(\phi_{t-1} - \beta_0^{*(\phi)}) - \beta_1^{(\phi)} \phi_{t-1} - \beta_0^{(\phi)} \\
 &= \rho(\phi_{t-1} - \beta_0^{*(\phi)}) + \beta_1^{(\phi)}(y_{t-1}^2 - \phi_{t-1}) \\
 &= \rho v_{t-1}^{(\phi)} + r_t^{(\phi)}
 \end{aligned}$$

where  $r_t^{(\phi)} = \beta_1^{(\phi)}(y_{t-1}^2 - \phi_{t-1})$ .

- Thus, the GARCH(1,1) can be written as a dispersion model with correlated random effects

$$\phi_t = \beta_0^{*(\phi)} + v_t^{(\phi)}$$

where  $v_t^{(\phi)} = \rho v_{t-1}^{(\phi)} + r_t^{(\phi)}$ .

- To avoid negative volatility, we can consider the exponential GARCH (EGARCH), with a log link  $\eta_t^{(\phi)} = \log \phi_t$

$$\eta_t^{(\phi)} = \beta_0^{(\phi)} + \beta_1^{(\phi)} y_{t-1}^2 + \gamma \eta_{t-1}^{(\phi)}$$

which is equivalent to

$$\eta_t^{(\phi)} = \beta_0^{*(\phi)} + v_t^{(\phi)}$$

### Stochastic volatility (SV) model

- If we take  $r_t^{(\phi)} \sim N(0, \alpha)$ , i.e.,  $v_t^{(\phi)} = \rho v_{t-1}^{(\phi)} + r_t^{(\phi)} \sim AR(1)$ , we have the stochastic volatility (SV) model originating from Harvey et al. (1994).
- For the data, SV model has  $cAIC = 1807$  which has less than  $cAIC = 2006$  for ARCH and  $cAIC = 1863$  for GARCH models, so that SV model is the best one among alternative models.

## Review : HGLM with random slope model

$$y_{ij} = \beta_1 F_i + \beta_2 F_i A_{ij} + \beta_3 M_i + \beta_4 M_i A_{ij} + v_{1i} + A_{ij} v_{2i} + \epsilon_{ij}$$

## DHGLM

- Noh and Lee (2007) showed that a robust analysis against such outliers can be obtained by adding random effects to the residual variance  $\phi_{ij}$ .
- Thus, we consider the following DHGLM

$$y_{ij} = \beta_1^{(\mu)} F_i + \beta_2^{(\mu)} F_i A_{ij} + \beta_3^{(\mu)} M_i + \beta_4^{(\mu)} M_i A_{ij} + v_{1i}^{(\mu)} + A_{ij} v_{2i}^{(\mu)} + \epsilon_{ij}$$
$$\log(\phi_{ij}) = \beta_0^{(\phi)} + v_i^{(\phi)}$$

where  $v_i^{(\phi)} \sim N(0, \alpha)$ .

## ex. Orthodontic growth continued

- Among models we considered, cAIC selects DHGLM as the best fitting model.
- The likelihood-ratio test for  $H_0 : \alpha = 0$  based on the RL, rejects the null hypothesis (deviance difference :  $37.9 > \chi^2_{2\delta}(1) = 2.71$  with significant level  $\delta = 0.05$ )

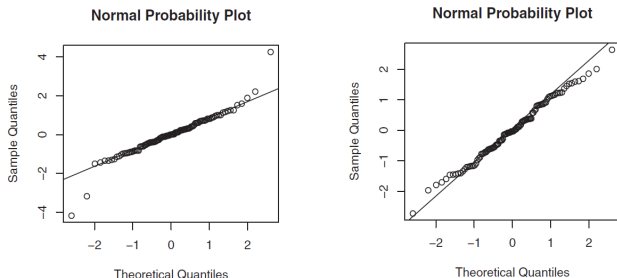


Figure: Normal probability plots for HGLM and DHGLM

- Model checking plots for the DHGLM show that all large outliers (whose sizes are bigger than 4) disappear.

- Schizophrenic behavior data from an eye-tracking experiment with a visual target moving back and forth along a horizontal line on a screen (Rubin and Wu, 1997).
- The outcome measurement is called the gain ratio, and it is recorded repeatedly at the peak velocity of the target during eye-tracking under three conditions (PS:plain sine, CS:color sine, TR:triangular).
- In the experiment, each subject is exposed to 5 trials, usually 3 PS, 1 CS, and 1 TR.
- During each trial, there are 11 cycles. However, for some cycles the gain ratios are missing because of eye blinks.
- On average, there are 34 observations out of 55 cycles for each subject (2906 observations from 4730 cycles).
- We assume (for simplicity) that the missing data are missing at random (MAR). Under MAR assumption, we can perform the analysis using only observed data.



**y** : gain ratio = (eye velocity)/(target velocity)

**x1** : 1/2(PS), -1/2(CS), 0(TR)

**x2** : -1/3(PS or CS), 2/3(TR)

**sex** : -1/2(female), 1/2(male)

**time** : measurement time

**schiz** : 1(schizophrenic), 0(non-schizophrenic)

**subject** : 43 non-schizophrenic subject (22 females and 21 males) and 43 schizophrenic subject (13 females and 30 males)

## HGLM

- $y_{ij}$  : gain ratios for the  $j$ -th measurement of the  $i$ -th subject.

$$y_{ij} = \beta_0^{(\mu)} + x_{1ij}\beta_1^{(\mu)} + x_{2ij}\beta_2^{(\mu)} + t_j\beta_3^{(\mu)} + sch_i\beta_4^{(\mu)} + sch_i \cdot x_{1ij}\beta_5^{(\mu)} \\ + sch_i \cdot x_{2ij}\beta_6^{(\mu)} + v_i^{(\mu)} + e_{ij}$$

where  $v_i^{(\mu)} \sim N(0, \lambda)$  is the subject random effect,  $e_{ij} \sim N(0, \phi)$  is a white noise.

- We find that schizophrenic patients have a larger variance.

$$\log(\phi_i) = \beta_0^{(\phi)} + sch_i\beta_1^{(\phi)}$$

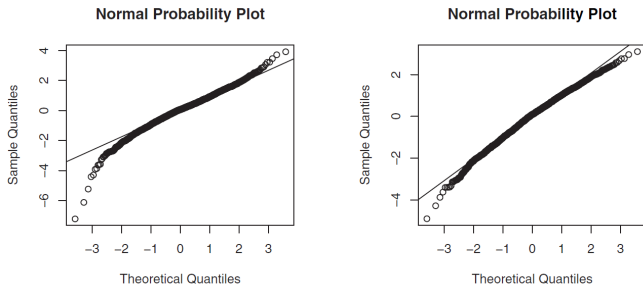
## DHGLM

- Psychologists have known for a long time about large variations in within-schizophrenic performance on almost any task (Silverman, 1967). Thus, abrupt changes among repeated response may be peculiar to schizophrenics and such volatility may differ for each patients.
- Such heteroscedasticity among schizophrenics cannot be modeled by the fixed effect model, but can be modeled by a DHGLM, introducing a random effect in the dispersion.
- When  $v_i^{(\phi)} \sim N(0, \alpha)$ ,

$$\log(\phi_i) = \beta_0^{(\phi)} + sch_i \beta_1^{(\phi)} + sch_i v_i^{(\phi)}$$

- $v_i^{(\mu)}$  and  $v_i^{(\phi)}$  are independent.

- cAIC shows that DHGLM has a better fit than HGLM.
- By using the studentized deviance residuals, we can obtain model-checking plots.



**Figure:** Normal probability plots for HGLM and DHGLM

- We see that most of the outliers in HGLM, caused by abrupt changes among repeated measures, disappear when random effects are allowed in the model for the residual variance.

## Review : HGLM

$$\begin{aligned}\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) &= \beta_0^{(\mu)} + \beta_1^{(\mu)} \text{trt}_i + \beta_2^{(\mu)} \text{mse}_i + \beta_3^{(\mu)} \text{age}_i \\ &\quad + \beta_4^{(\mu)} \text{center}_i + \beta_5^{(\mu)} \text{base}_i + \beta_6^{(\mu)} y_{i(j-1)} + v_i \\ \log \lambda_i &= \beta_0^{(\lambda)} + \beta_1^{(\lambda)} \text{age}_i\end{aligned}$$

## DHGLM

- With binary data, it is difficult to identify the distribution of random effects.
- The use of a heavy-tailed distribution for random effects by allowing random effects for  $\lambda$ , removes sensitivity of the parameter estimation to the choice of random effect distribution (Noh et al, 2005).
- For binary data, they showed that GLMM estimators can give serious biases if the true distribution is not normal.

- Consider the following DHGLM,

$$\begin{aligned} \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) &= \beta_0^{(\mu)} + \beta_1^{(\mu)} \text{trt}_i + \beta_2^{(\mu)} \text{mse}_i + \beta_3^{(\mu)} \text{age}_i \\ &\quad + \beta_4^{(\mu)} \text{center}_i + \beta_5^{(\mu)} \text{base}_i + \beta_6^{(\mu)} y_{i(j-1)} + v_i \\ \log \lambda_i &= \beta_0^{(\lambda)} + \beta_1^{(\lambda)} \text{age}_i + v_i^{(\lambda)} \end{aligned}$$

where  $v_i^{(\mu)} \sim N(0, \lambda_i)$  and  $v_i^{(\lambda)} \sim N(0, \tau)$ .

- The likelihood-ratio test for  $H_0 : \tau = 0$ , based on the restricted likelihood, rejects the null hypothesis (deviance difference :  $3.3 > \chi_{2\delta}^2(1) = 2.71$  with significant level  $\delta = 0.05$ )

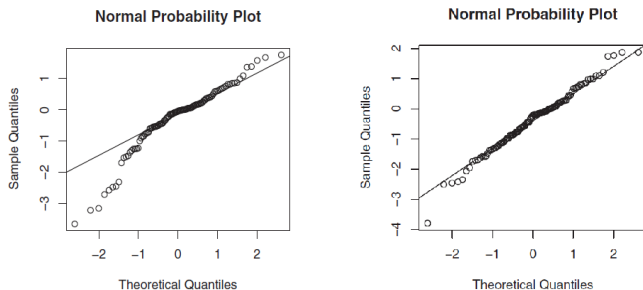


Figure: Normal probability plots of  $\lambda$  for HGLM and DHGLM

- We see that large outliers, and an unpleasant pattern in the normal probability plot under the HGLM, disappear under the DHGLM. Thus, the DHGLM is preferred.
- Furthermore, there are apparent differences between parameter estimates.
- In this case, we should report the results from the DHGLM because a distributional assumption of random effects is hard to identify with the binary data.

## Review : HGLM

$$\log \left( \frac{p_{ijk}}{1 - p_{ijk}} \right) = \beta_0 + F_i + M_j + (FM)_{ij} + v_{ik}^f + v_{jk}^m$$

## DHGLM

- For this binary set, we fit a DHGLM model

$$\log \left( \frac{p_{ijk}}{1 - p_{ijk}} \right) = x_{ijk}^t \beta^{(\mu)} + v_{fik}^{(\mu)} + v_{mjk}^{(\mu)}$$

$$\log(\lambda_{fik}) = \beta_{f0}^{(\lambda)} + b_{fik}^{(\lambda)}$$

$$\log(\lambda_{mik}) = \beta_{m0}^{(\lambda)} + b_{mik}^{(\lambda)}$$

where  $v_{fik}^{(\mu)} \sim N(0, \lambda_{fik})$ ,  $v_{mjk}^{(\mu)} \sim N(0, \lambda_{mjk})$ ,  $b_{fik}^{(\lambda)} \sim N(0, \tau_f)$ , and  $b_{mik}^{(\lambda)} \sim N(0, \tau_m)$ .



- The cAIC difference between HGLM and DHGLM is less than 1, so that there would be no advantage to use the heavy-tailed distribution, compared with the normal distribution.
- The likelihood-ratio test for  $H_0 : \tau_f = 0, \tau_m = 0$  based on RL does not reject the null hypothesis (deviance difference : 2.4 which has p-value of  $0.243 = 0.5 \times P(\chi^2(1) > 2.4) + 0.25 \times P(\chi^2) > 2.4)$  (Self and Liang, 1987).
- Estimates between HGLM and DHGLM are slightly different, which also strongly indicating the adequacy of normality for the distribution of random effects.

### Reviewd : Binomial GLMM

- $p_{ij} = P(y_{ij} = 1 | v_i)$

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + \beta_1 I(i = \text{drug}) + \beta_2 I(i = \text{drug}+) + v_i$$

where  $v_i \sim N(0, \lambda)$

### DHGLM

- We fit a follow DHGLM model.

$$\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0^{(\mu)} + \beta_1^{(\mu)} I(i = \text{drug}) + \beta_2^{(\mu)} I(i = \text{drug}+) + v_i^{(\mu)}$$
$$\log(\lambda_i) = \beta_0^{(\lambda)} + v_i^{(\lambda)}$$

where  $v_i^{(\mu)} \sim N(0, \lambda_i)$  and  $v_i^{(\lambda)} \sim N(0, \tau)$ .

- cAIC from the HGLM is 205.0, while that from the DHGLM is 204.5.
- The cAIC difference is less than 1, so that there would be no advantage to use the heavy-tailed distribution.
- The likelihood=ratio test for  $H_0 : \tau = 0$  based on the RL, does not reject the null hypothesis (deviance difference :  $0.8 < \chi^2_{2\delta}(1) = 2.71$  with significant level  $\delta = 0.05$ ).
- Estimates between HGLM and DHGLM are only slightly different, which also strongly indicates the adequacy of normality for the distribution of random effects.

$$\log(\mu_{ij}) = \beta_0^{(\mu)} + x_{B_i}\beta_B^{(\mu)} + x_{T_i}\beta_T^{(\mu)} + x_{A_i}\beta_A^{(\mu)} + x_{V_j}\beta_V^{(\mu)} + x_{B_iT_i}\beta_{BT}^{(\mu)} + v_i^{(\mu)} + v_{ij}^{(\mu)}$$

$$\log(\lambda_{1i}) = \beta_0^{(\lambda_1)} + v_i^{(\lambda_1)} \quad \text{and} \quad \log(\lambda_{2i}) = \beta_0^{(\lambda_2)} + v_{ij}^{(\lambda_2)}$$

### NB-gamma HGLM

$$v_i^{(\mu)} \sim G(\lambda_{1i}), \quad v_i^{(\lambda_1)} = 0,$$

$$v_{ij}^{(\mu)} \sim G(\lambda_{2ij}), \quad \text{and} \quad v_{ij}^{(\lambda_2)} = 0$$

### Poisson-normal DHGLM

$$v_i^{(\mu)} \sim N(0, \lambda_{1i}), \quad v_i^{(\lambda_1)} \sim N(0, \tau_1),$$

$$v_{ij}^{(\mu)} = 0, \quad \text{and} \quad v_{ij}^{(\lambda_2)} = 0$$

### Poisson-normal-gamma DHGLM

$$v_i^{(\mu)} \sim N(0, \lambda_{1i}), \quad v_i^{(\lambda_1)} = 0,$$

$$v_{ij}^{(\mu)} \sim G(\lambda_{2ij}), \quad \text{and} \quad v_{ij}^{(\lambda_2)} \sim N(0, \tau_2)$$

### Poisson-gamma-gamma DHGLM1

$$v_i^{(\mu)} \sim G(\lambda_{1i}), \quad v_i^{(\lambda_1)} = 0,$$

$$v_{ij}^{(\mu)} \sim G(\lambda_{2ij}), \quad \text{and} \quad v_{ij}^{(\lambda_2)} \sim N(0, \tau_2)$$

### Poisson-gamma-gamma DHGLM2

$$v_i^{(\mu)} \sim G(\lambda_{1i}), \quad v_i^{(\lambda_1)} \sim N(0, \tau_1),$$

$$v_{ij}^{(\mu)} \sim G(\lambda_{2ij}), \quad \text{and} \quad v_{ij}^{(\lambda_2)} \sim N(0, \tau_2)$$

### quasi Poisson-normal DHGLM

$$v_i^{(\mu)} \sim N(0, \lambda_{1i}), \quad v_i^{(\lambda_1)} \sim N(0, \tau_1), \quad \text{and} \quad \text{var}(y_{ij} | v_i^{(\mu)}, v_{ij}^{(\mu)}) = \phi \mu_{ij}$$

- The likelihood-ratio test for  $H_0 : \tau_2 = 0$  based on the RL rejects the null hypothesis (deviance difference :  $32.4 > \chi^2_{2\delta}(1) = 2.71$  with significant level  $\delta = 0.05$  between NB-gamma HGLM and Poisson-gamma-gamma DHGLM1).
- The likelihood-ratio test for  $H_0 : \tau_1 = 0$  based on the RL doesn't reject the null hypothesis (deviance difference : 0 between Poisson-gamma-gamma DHGLM1 and Poisson-gamma-gamma DHGLM2).
- Thus, the likelihood-ratio test selects the Poisson-gamma-gamma DHGLM1.

Model	cAIC	rAIC
NB - gamma HGLM	1163.9	1274.8
Poisson - normal DHGLM	1270.5	1349.1
Poisson - normal - gamma DHGLM	1183.0	1282.7
Poisson - gamma - gamma DHGLM1	1144.2	1244.4
Poisson - gamma - gamma DHGLM2	1146.1	1246.4
quasi Poisson - normal DHGLM	1217.2	1319.6

- Approximately 30% of hospitalized patients due to acute ischemic stroke are placed in the risk of early neurologic deterioration (END) at their hospital stay.
- The patient's risk to END can be monitored by following their blood pressure (BP).
- Data has systolic BP (SBP) with time in hours after arriving at the emergency room for two stroke patients (one is END; the other is non-END).

time, time1 : Times after arriving at the emergency room (hrs)

y1 : SBP of END stroke patient

y2 : SBP of non-END stroke patient

## Joint spline model

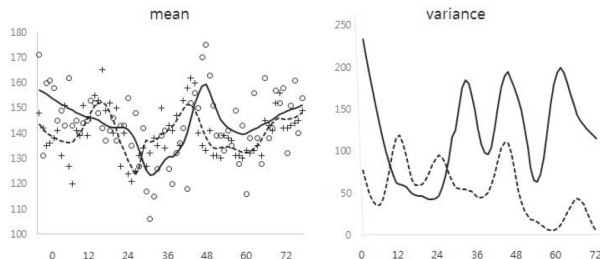
- For detection of changes in SBP with respect to time, we use cubic splines (Silverman, 1967; Green and Silverman, 1994) not only for the mean changes but also for variance changes, using the joint cubic splines model (Lee and Nelder, 2006).
- $y_t$  : SBP measurement at time  $t$ ,  $e_t \sim N(0, \phi_t)$
- $f_m(t)$  and  $f_d(t)$  : unknown functions of the mean and variance.

$$y_t = f_m(t) + e_t \text{ and } \log \phi_t = f_d(t)$$

- For joint fitting of the mean  $\mu_t$  and variance  $\phi_t$ , we use the DHGLM.

$$\begin{aligned}\mu_t &= \beta_0^{(\mu)} + \beta_1^{(\mu)} t + v_t^{(\mu)} \\ \log \phi_t &= \beta_0^{(\phi)} + \beta_1^{(\phi)} t + v_t^{(\phi)}\end{aligned}$$

where  $v_t^{(\mu)} \begin{bmatrix} v_t^{(\phi)} \end{bmatrix}$  is the random component with mean 0 and a singular precision matrix  $P/\lambda^{(\mu)} \begin{bmatrix} P/\lambda^{(\phi)} \end{bmatrix}$  (Lee, Nelder, and Pawitan, 2017).

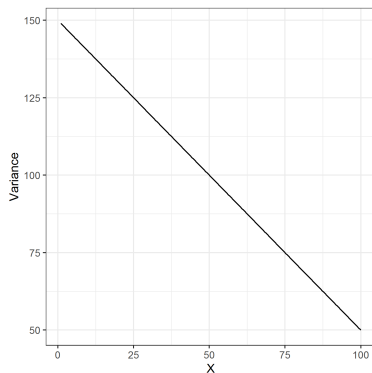
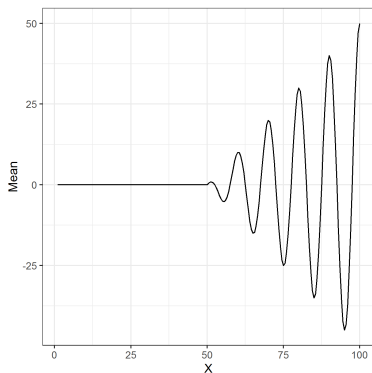


**Figure:** Joint cubic splines for SBP (END: solid line, non-END: dashed line)

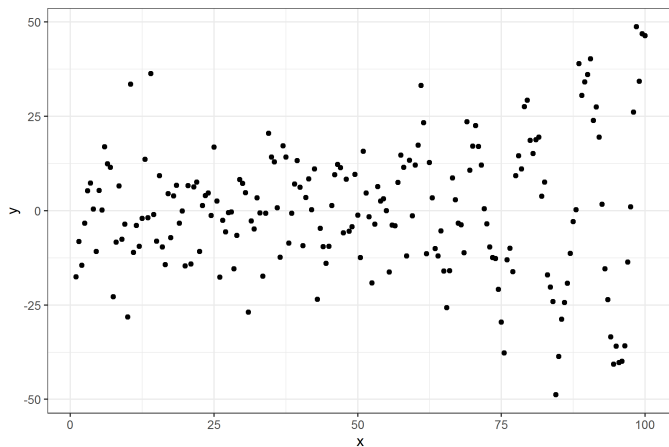
- Mean patterns for END and non-END patients are similar, so that it has been very difficult to predict the potential END patients. However, it can be noticed from the plot that the END patient has higher variance in SBP than non-END patient.
- Thus, the variance of the SBP is used as a covariate for predicting an END event, which greatly prevents the occurrence of END patients in the emergency room in Korea.



- The raw data are generated from normal distribution with the true mean and variance, described in plot.



## Raw data



# An extension of linear mixed models via DHGLM

- The IWLS algorithm gives fast computations using GLM estimations.

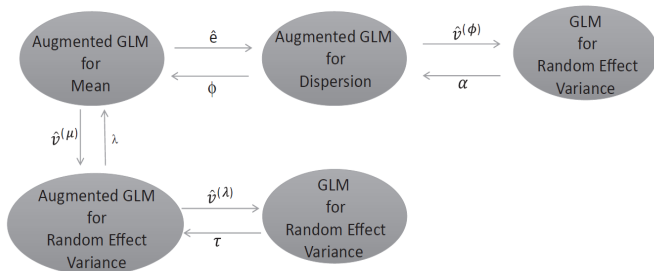


Figure 6.11 *Interconnected GLMs for fitting DHGLMs.*

# An extension of linear mixed models via DHGLM

- Consider the DHGLM introduced in Chapter 1(1.4)

$$\mathbf{y} = \mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)} + \mathbf{e}$$

$$\mathbf{e} \sim N(0, \exp(\mathbf{X}^{(\phi)}\beta^{(\phi)} + \mathbf{Z}^{(\phi)}\mathbf{v}^{(\phi)}))$$

with  $\mathbf{v}^{(\mu)} \sim N(0, \lambda\mathbf{I})$ ,  $\mathbf{v}^{(\phi)} \sim N(0, \alpha\mathbf{I})$ ,  $\text{cor}(\mathbf{v}^{(\mu)}, \mathbf{v}^{(\phi)}) = 0$

- We specify the h-likelihood to show that it is rather easy to specify even though the model is rather advanced.

$$\begin{aligned} h &= \log(f(\mathbf{y}|\mathbf{v}^{(\mu)}, \mathbf{v}^{(\phi)})) + \log(f(\mathbf{v}^{(\mu)})) + \log(f(\mathbf{v}^{(\phi)})) \\ &= \frac{1}{2} \log(|\mathbf{V}|) - \frac{1}{2} (\mathbf{y} - (\mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)}))^{\top} \mathbf{V}^{-1} \\ &\quad \times (\mathbf{y} - (\mathbf{X}^{(\mu)}\beta^{(\mu)} + \mathbf{Z}^{(\mu)}\mathbf{v}^{(\mu)})) \\ &\quad - \frac{m}{2} \log(\lambda) - \frac{1}{2\lambda^2} (\mathbf{v}^{(\mu)})^{\top} (\mathbf{v}^{(\mu)}) \\ &\quad - \frac{m}{2} \log(\alpha) - \frac{1}{2\alpha^2} (\mathbf{v}^{(\phi)})^{\top} (\mathbf{v}^{(\phi)}) \end{aligned}$$

where  $\mathbf{V} = \text{diag}(\exp(\mathbf{X}^{(\phi)}\beta^{(\phi)} + \mathbf{Z}^{(\phi)}\mathbf{v}^{(\phi)}))$ .

- We could allow correlations among all random components  $\mathbf{e}$ ,  $\mathbf{v}^{(\mu)}$ ,  $\mathbf{v}^{(\phi)}$ , which leads to many other interesting models to explore.

## Chapter 7. MDHGLMs

- As a most general model for a single response, we presented a DHGLM, which has a great room for further generalization by including more general correlation patterns among random effects.
- In this chapter, we introduce multivariate models for various types of responses including continuous, proportion, counts, events, etc.
- We show that general multivariate models can be generated by connecting DHGLMs for various responses with correlated random effects.
- Correlation between random components is essential in the definition of joint models, where correlations among multivariate responses are modeled via correlated random effects.

- Data from a study on the developmental toxicity of ethylene glycol (EG) in mice (Price et al., 1985).
- Times-pregnant CD-1 mice were dosed by gavage with EG in distilled water on gestational days 6 through 15.

litter, id : 94 dams

dose : dose (g/kg)

y1 : fetal weight (g)

y2 : 1(malformation), 0(not)

dose2 : dose<sup>2</sup>

Dose (g/kg)	Dams	Live	Malformations		Weight (g)	
			No.	%	Mean	(S.D)
0.00	25	297	1	(0.34)	0.972	(0.0976)
0.75	24	276	26	(9.42)	0.877	(0.1041)
1.50	22	229	89	(38.86)	0.764	(0.1066)
3.00	23	226	129	(57.08)	0.704	(0.1238)

## Bivariate HGLM

- $y_{ij} = (y_{1ij}, y_{2ij})^T$  : bivariate responses from  $j$ -th mouse, born from  $i$ -th dam
- $v_{ij} = (w_i, u_i)^T$  : unobserved random effects for the  $i$ -th dam
- It is assumed that  $y_{1ij}$  and  $y_{2ij}$  are conditionally independent given  $v_i$ .

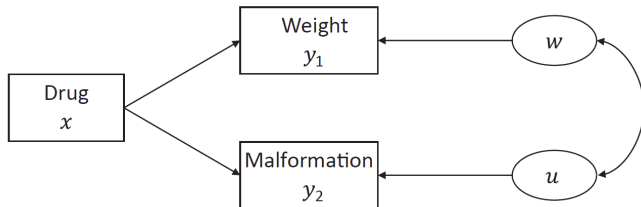


Figure: Path diagram for the MDHGLM fitted to the EG data

- Hence, the following bivariate HGLM is proposed

$$y_{1ij}|w_i \sim N(\mu_{ij}, \phi)$$

where  $\mu_{ij} = x_{1ij}\beta_1 + w_i$ ,

$$y_{2ij}|u_i \sim \text{Bernoulli}(p_{ij})$$

where  $\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = x_{2ij}\beta_2 + u_i$ , and

$$v_i \sim N\left(0, \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\right)$$

- We consider three models

**M1** Independent random-effects model where  $\rho = 0$

**M2** Random-effects model with a saturated variance-covariance matrix

**M3** Shared random-effects models where  $u_i = \delta w_i$  for some constant  $\delta$



- The Rheumatoid Arthritis Patients rePort Onset Re-activation sTudy (RAPPORT study) : longitudinal study that aims to identify an increase in disease activity by self-reported questionnaires.
- Self-reported questionnaires are provided for patients every 3 months together with clinical evaluations of patients' disease status.
- HAQ and RADAI were used for patients to self-report their functional status.
- A clinical examination was recorded using the DAS28, which is a composite score that includes for example the swollen joints counts. The DAS28 score varies between 0 and 10.
- There are 159 patients and 5 visits for each patients.
- Not all patients gave information for each  $k$ -th response and not all patients were measured at each of the 5 visits.
- HAQ : Health assessment questionnaires (20 questions from 8 categories)
- RADAI : Rheumatoid arthritis disease activity index (5 items)
- DAS28 : Disease activity score with 28 joint counts

y1 : DAS28

y2 : 1(HAQ > 0.5), 0(HAQ < 0.5)

y3 : 1(RADAI > 2.2), 0(RADAI < 2.2)

time : month of measurement (0,3,6,9,12)

age : age at the baseline

sex : 1(female), 0(male)

subject : 159 patients

### Multivariate model with 3 responses

- $y_{ij} = (y_{1ij}, y_{2ij}, y_{3ij})^T$  : response from  $j$ -th visits of the  $i$ -th (patients  $i = 1, 2, \dots, 159$  and  $j = 1, 2, \dots, 5$ )
- $X_k$  : designed matrix for  $k$ -th response
- As covariates we use the intercept, time, age, and sex.

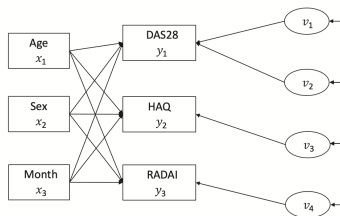


Figure: Path diagram for the MDHGLM fitted to the rheumatoid arthritis data

- We consider the following multivariate model with three responses.

$$y_{1ij} | v_{11i}, v_{12i} \sim N(X_{1ij}\beta_1 + v_{11i} + v_{12i} \cdot \text{time}, \phi)$$

$$y_{2ij} | v_{21i} \sim \text{Bernoulli} \left( \frac{\exp(X_{2ij}\beta_2 + v_{21i})}{1 + \exp(X_{2ij}\beta_2 + v_{21i})} \right)$$

$$y_{3ij} | v_{31i} \sim \text{Bernoulli} \left( \frac{\exp(X_{3ij}\beta_3 + v_{31i})}{1 + \exp(X_{3ij}\beta_3 + v_{31i})} \right)$$

- The model for DAS28 includes a random intercept and slope, while HAQ and RADAI have only random intercepts.
- We assume a 4-dimensional latent structure :

$$\begin{pmatrix} v_{11i} \\ v_{12i} \\ v_{21i} \\ v_{31i} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_{11} & \rho_1 \lambda_{11,12}^* & \rho_2 \lambda_{11,21}^* & \rho_3 \lambda_{11,31}^* \\ \rho_1 \lambda_{12,11}^* & \lambda_{12} & \rho_4 \lambda_{12,21}^* & \rho_5 \lambda_{12,31}^* \\ \rho_2 \lambda_{21,11}^* & \rho_4 \lambda_{21,12}^* & \lambda_{21} & \rho_6 \lambda_{21,31}^* \\ \rho_3 \lambda_{31,11}^* & \rho_5 \lambda_{31,12}^* & \rho_6 \lambda_{31,21}^* & \lambda_{31} \end{pmatrix} \right)$$

where  $\lambda_{ij,kl}^* = \sqrt{\lambda_{ij}\lambda_{kl}}$ .

- National merit twins data including extensive questionnaires from 839 adolescent twins, who took the national merit scholarship qualifying test (NMSQT) in 1962 among the roughly 600,000 US high school juniors (Loehlin and Nichols, 1976).
- They were diagnosed as identical (509 pairs) or same-sex fraternal (330 pairs) by a brief mail questionnaire.
- Later, they completed a 1082-item questionnaire covering a variety of behaviors, attitudes, personality, life experiences, health, vocational preferences, etc., plus the 480-item California psychological inventory.
- Twins' scores on the NMSQT and their five subscales are also included.
- The 285-item questionnaire filled out by the parent was mainly focused on the life histories and experiences of the twins.

pairnum : 768 pairs

y1, y2, y3, y4 : NMSQT scores recorded within 0-100. English( $y_1$ ), mathematics( $y_2$ ), social science( $y_3$ ) and natural science( $y_4$ )

variables	code	definition
Gender	$x_1$	1(male), 2(female)
Mother's educational level	$x_2$	1( $\leq$ 8th grade), 2(part high school), 3(high school grad), 4(part college), 5(college grad), 6(graduate degree)
Father's educational level	$x_3$	1( $\leq$ 8th grade), 2(part high school), 3(high school grad), 4(part college), 5(college grad), 6(graduate degree)
Family income level	$x_4$	1( $\leq$ \$5000), 2(\$5000 to \$7500) 3(\$7500 to \$10000), 4(\$10000 to \$15000) 5(\$15000 to \$20000), 6(\$20000 to \$25000) 7( $\geq$ \$25000)
Zygotosity	$x_5$	0(identical), 1(fraternal)

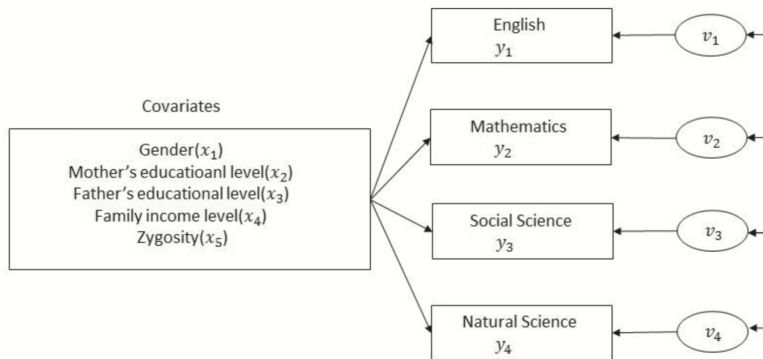


Figure: Path diagram for the MDHGLM fitted to the NMSQT for twins data

## Multivariate HGLM

- We consider a multivariate HGLM with 4 response variables for the  $j$ -th person of the  $i$ -th twin. For  $k = 1, 2, 3, 4$ ,

$$y_{kij}|v_{ki} \sim N(X_{ij}\beta_k^{(\mu)} + v_{ki}, \phi_{kij})$$

where random effects follow multivariate normal distribution

$$\begin{pmatrix} v_{1i} \\ v_{2i} \\ v_{3i} \\ v_{4i} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_{1i} & \rho_1 \lambda_{1i,2i}^* & \rho_2 \lambda_{1i,3i}^* & \rho_3 \lambda_{1i,4i}^* \\ \rho_1 \lambda_{2i,1i}^* & \lambda_{2i} & \rho_4 \lambda_{2i,3i}^* & \rho_5 \lambda_{2i,4i}^* \\ \rho_2 \lambda_{3i,1i}^* & \rho_4 \lambda_{3i,2i}^* & \lambda_{3i} & \rho_6 \lambda_{3i,4i}^* \\ \rho_3 \lambda_{4i,1i}^* & \rho_5 \lambda_{4i,2i}^* & \rho_6 \lambda_{4i,3i}^* & \lambda_{4i} \end{pmatrix} \right),$$

$\lambda_{ji,k}^* = \sqrt{\lambda_{ji} \lambda_{ki}}$  and  $\lambda_{ki} = \exp(\beta_{k0}^{(\lambda)})$  is the variance of random effects.

- To allow heterogeneity between type of zygosity, we consider the model for residual variance

$$\log \phi_{kij} = \beta_{k0}^{(\phi)} + \beta_{k5}^{(\phi)} x_{5i}$$



- Random effects of social science and natural science scores show the strongest correlation, 0.738. Correlation between English and mathematics scores has the lowest value, 0.622.
- For gender effect, men have higher significant scores on mathematics, social science and natural science, but women have higher significant scores on English.
- Mother's educational level is not significant at almost all subject's scores. But father's educational level 4 and 6 are significant.
- Family's income level 5 has a significant positive effect and it has the highest estimate.
- In dispersion models for residual variances, we see that fraternal twins have greater heterogeneity than identical twins.

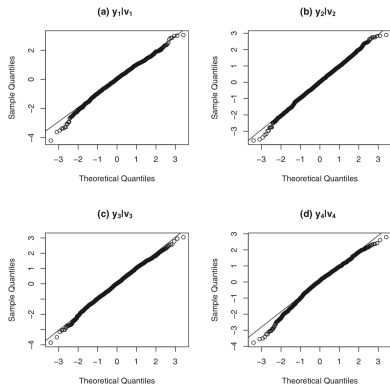


Figure 7.4 Normal probability plots for (a)  $y_1|v_1$ , (b)  $y_2|v_2$ , (c)  $y_3|v_3$ , and (d)  $y_4|v_4$  under the multivariate HGLM on the national merit scholarship qualifying test for twins data.

- We see that the normal probability plots are approximately linear in the absence of outliers. Thus, the fitted model is satisfactory.

- Lee, Nelder and Pawitan (2017) considered the Vascular Cognitive Impairment (VCI) data.
- The VCI measurements are increased among stroke patients, because cognitive function is declined due to stroke. However, through an early intervention based on the VCI, the cognitive function can be improved.
- The purpose of the study is to examine the effects of 10 demographic and 10 acute neuroimaging variables on the cognitive function in the ischemic stroke patients.

$y_1, y_2, y_3, y_4$  : the standardized VCI scores. Executive( $y_1$ ), memory( $y_2$ ), visuoapatial( $y_3$ ), and language( $y_4$ )

$id$  : 372 patients

Variable	Code	Definition
Demographic variables		
Age	x <sub>1</sub>	integer of age/10
Gender	x <sub>2</sub>	1(male), 0(female)
Edu	x <sub>3</sub>	0(none), 1(elementary), 2(middle), 3(high), 4(over college)
HTN	x <sub>4</sub>	1(hypertension), 0(none)
DM	x <sub>5</sub>	1(diabetes mellitus), 0(none)
Af	x <sub>6</sub>	1(atrial fibrillation), 0(none)
HxStroke	x <sub>7</sub>	1(history of stroke), 0(none)
NIHSS	x <sub>8</sub>	national institute of health stroke scale score at admission
VCINP	x <sub>9</sub>	time interval from stroke onset to first K-VCiHS-NP
PCI	x <sub>10</sub>	1(IQCODE $\geq$ 3.6), 0(otherwise)
Neuroimaging variables		
AcuteLeft	x <sub>11</sub>	Left or bilateral involvement
AcuteMulti	x <sub>12</sub>	lesion multiplicity in acute DWI imaging
AcuteCS	x <sub>13</sub>	cortical involvement of acute lesions
ChrCS	x <sub>14</sub>	cortical involvement of chronic territorial infarction
PVWM	x <sub>15</sub>	Periventricular white matter lesions (PVWM). 0(PVWM 0,1), 1(PVWM 2,3)
SCWM	x <sub>16</sub>	Subcortical white matter lesions (SCWM). 0(SCWM 0,1), 1(SCWM 2,3)
LAC	x <sub>17</sub>	The presence of lacunes
CMB	x <sub>18</sub>	The presence of cerebral microbleeds
		Medial temporal lobe atrophy (MTA)
MTA1	x <sub>19</sub>	1(MTA 2), 0(not 2)
MTA2	x <sub>20</sub>	1(MTA 3,4), 0(not 3,4)

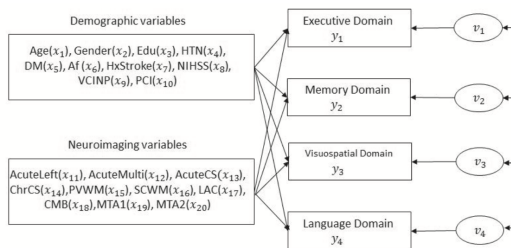


Figure: Path diagram for the MDHGLM fitted to the VCI data

## Multivariate HGLM

- Consider a multivariate HGLM for four response variables for the  $t$ -th visit of the  $i$ -th patient. For  $k = 1, 2, 3, 4$ ,

$$y_{kit} | v_{ki} \sim N(X_{it}\beta_k^{(\mu)} + v_{ki}, \phi_{kit})$$

where  $X_{it}$  are covariates,  $\phi_{kit} = \exp(\beta_{k0}^{(\phi)})$  is the residual variance.

- The random effects follow a multivariate normal distribution :

$$\begin{pmatrix} v_{1i} \\ v_{2i} \\ v_{3i} \\ v_{4i} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_{1i} & \rho_1 \lambda_{1i,2i}^* & \rho_2 \lambda_{1i,3i}^* & \rho_3 \lambda_{1i,4i}^* \\ \rho_1 \lambda_{2i,1i}^* & \lambda_{2i} & \rho_4 \lambda_{2i,3i}^* & \rho_5 \lambda_{2i,4i}^* \\ \rho_2 \lambda_{3i,1i}^* & \rho_4 \lambda_{3i,2i}^* & \lambda_{3i} & \rho_6 \lambda_{3i,4i}^* \\ \rho_3 \lambda_{4i,1i}^* & \rho_5 \lambda_{4i,2i}^* & \rho_6 \lambda_{4i,3i}^* & \lambda_{4i} \end{pmatrix} \right),$$

where  $\lambda_{ji}^* = \sqrt{\lambda_{ji} \lambda_{ki}}$  and  $\lambda_{ki} = \exp(\beta_{k0}^{(\lambda)})$  is the variance of random effects.

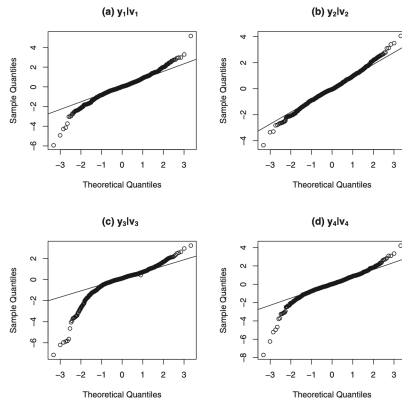


Figure 7.6 Normal probability plots for (a)  $y_1|v_1$ , (b)  $y_2|v_2$ , (c)  $y_3|v_3$ , and (d)  $y_4|v_4$  under the multivariate HGLM on the vascular cognitive impairment data.

- We see many large outliers.

## MDHGLM1

- We consider a multivariate DHGLM (called MDHGLM1) that allows a heavy-tailed distribution for  $y_{kit}|v_{ki}$  ( $k = 1, 2, 3, 4$ ) as follows. For  $k = 1, 2, 3, 4$ ,

$$\log \phi_{kit} = \beta_{k0}^{(\phi)} + v_{ki}^{(\phi)},$$

where  $v_{ki}^{(\phi)} \sim N(0, \alpha_k)$  and  $\lambda_{ki} = \exp(\beta_{k0}^{(\lambda)})$ .

## MDHGLM2

- We further consider a MDHGLM (called MDHGLM2) also allowing heavy-tailed distribution for  $v_{ki}$  as follows. For  $k = 1, 2, 3, 4$ ,

$$\begin{aligned} \log \phi_{kit} &= \beta_{k0}^{(\phi)} + v_{ki}^{(\phi)} \\ \log \lambda_{ki} &= \beta_{k0}^{(\lambda)} + v_{ki}^{(\lambda)}, \end{aligned}$$

where  $v_{ki}^{(\phi)} \sim N(0, \alpha_k)$  and  $v_{ki}^{(\lambda)} \sim N(0, \tau_k)$ .



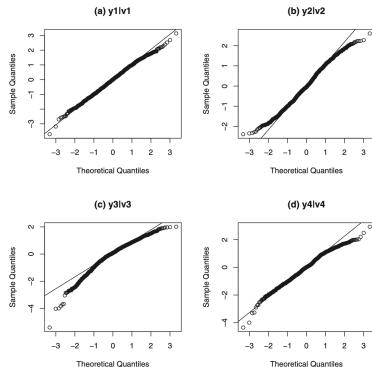


Figure 7.7 Normal probability plots for (a)  $y_1|v_1$ , (b)  $y_2|v_2$ , (c)  $y_3|v_3$ , and (d)  $y_4|v_4$  under the MDHGLM1 on the vascular cognitive impairment data.

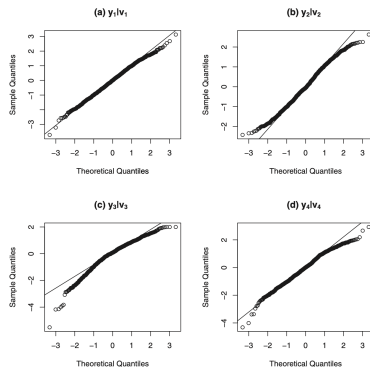


Figure 7.8 Normal probability plots for (a)  $y_1|v_1$ , (b)  $y_2|v_2$ , (c)  $y_3|v_3$ , and (d)  $y_4|v_4$  under the MDHGLM2 on the vascular cognitive impairment data.

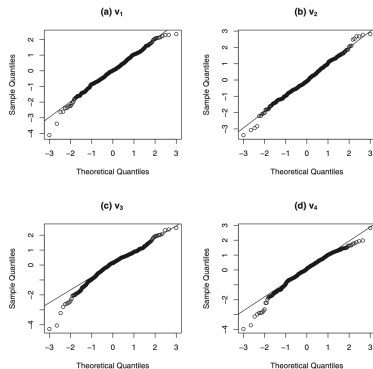


Figure 7.9 Normal probability plots for (a)  $v_1$ , (b)  $v_2$ , (c)  $v_3$ , and (d)  $v_4$  under the MDHGLM1 on the vascular cognitive impairment data.

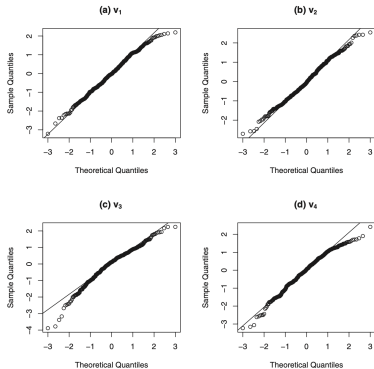


Figure 7.10 Normal probability plots for (a)  $v_1$ , (b)  $v_2$ , (c)  $v_3$ , and (d)  $v_4$  under the MDHGLM2 on the vascular cognitive impairment data.

- cAIC selects MDHGLM2 (cAIC=10437.4) as the best-fitting model among 3 models, because cAIC for the multivariate HGLM (cAIC=13260.0) and MDHGLM1 (cAIC=10548.1) are larger.
- We see that most outliers in the multivariate HGLMs disappear by using MDHGLM1 or MDHGLM2.
- From the normal probability plots for  $v_{ki}^{(\lambda)}$ , MDHGLM2 is preferred to the MDHGLM1 because  $\hat{v}_{ki}^{(\lambda)}$  leans more toward the line.
- Thus, we select the MDHGLM2 as the final model, which gives robust estimators against outliers as well as robustness against misspecification of distributional assumptions on random effects.

- Longitudinal data set from mother's stress and children's morbidity study (MSCM) (Asar and Ilk, 2014).
- In this MSCM study, 167 mothers and their preschool children were enrolled for 28 days.
- Investigation of the serial dependence structures of the 2 longitudinal responses suggested a weak correlation structure for the period of days 1~16. Therefore, only the period of days 17 ~ 28 is considered in this dataset.
- $167 \times 12 = 2004$  observations are in dataset.

**stress** =  $y_1$  : mother's stress. 1(presence), 0(absence)

**illness** =  $y_2$  : children's illness. 1(presence), 0(absence)

**married** : marriage status. 1(married), 0(other)

**education** : highest education level. 1( $\geq$  high school), 0( $<$  high school)

**employed** : employment status. 1(employed), 0(unemployed)

**race** : race. 1(non-white), 0(white)

**csex** : gender of children. 1(female), 0(male)

**chlth** : health statuses of children at baseline. 3(very good), 2(good), 1(fair),  
0(poor/very poor)

**mlhth** : health statuses of mothers at baseline. 3(very good), 2(good), 1(fair),  
0(poor/very poor)

**houseize** : household size. 1(more than 3 people), 0(2-3 people)

**bstress** : rhe average stress values of the 1~16 days

**billness** : rhe average illness values of the 1~16 days

**week** : study time.  $\text{week} = (\text{day}-22)/7$

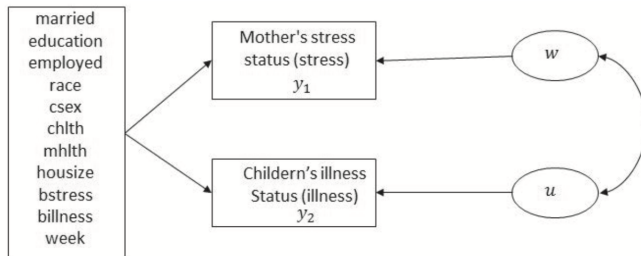


Figure: Path diagram for the MDHGLM fitted to the mother's stress and children's morbidity data

## Bivariate Bernoulli HGLM

- $y_{ij} = (y_{1ij}, y_{2ij})^\top$  : bivariate binary responses for the  $j$ -th visit of the  $i$ -th family
- $v_i^{(\mu)} = (w_i^{(\mu)}, u_i^{(\mu)})^\top$  : unobserved random effects for the  $i$ -th family
- $y_{1ij}|v_i^{(\mu)} \sim \text{Bernoulli}(p_{1ij})$ ,  $y_{2ij}|v_i^{(\mu)} \sim \text{Bernoulli}(p_{2ij})$

$$\log\left(\frac{p_{1ij}}{1 - p_{1ij}}\right) = X_{ij}\beta_1^{(\mu)} + w_i^{(\mu)}, \quad \log\left(\frac{p_{2ij}}{1 - p_{2ij}}\right) = X_{ij}\beta_2^{(\mu)} + u_i^{(\mu)}$$

where  $v_i^{(\mu)} \sim N(0, \Sigma_i)$  with  $\Sigma_i = \begin{pmatrix} \lambda_{1i} & \rho \sqrt{\lambda_{1i}\lambda_{2i}} \\ \rho \sqrt{\lambda_{1i}\lambda_{2i}} & \lambda_{2i} \end{pmatrix}$  and  $-1 < \rho < 1$ .

- Thus, given  $v_i^{(\mu)}$ ,  $y_{1ij}$  and  $y_{2ij}$  are independent.
- We first consider three models with  $\log \lambda_{1i} = \beta_{10}^{(\lambda)}$  and  $\log \lambda_{2i} = \beta_{20}^{(\lambda)}$ .

**M1** Independent model, having  $\rho = 0$

**M2** Random-effect model with a saturated variance-covariance matrix

**M3** Shared random-effects model, having  $u_i^{(\mu)} = \delta w_i^{(\mu)}$  for some constant  $\delta$

- The cAIC has values of 2653.7 (M1), 2428.9 (M2), and 2517.3 (M3).
- Thus, cAIC selects the full model M2 among 3 models.

### Robust bivariate DHGLM

- In binary data, GLMMs are sensitive to a distributional assumption of random effects, which is difficult to identify.
- Thus, we consider the robust bivariate DHGLM by allowing random effects in the variance for random effects.

**M4** the same as M2, but having  $\log \lambda_{1i} = \beta_{10}^{(\lambda)} + w_i^{(\lambda)}$  and  $\log \lambda_{2i} = \beta_{20}^{(\lambda)} + u_i^{(\lambda)}$  where  $w_i^{(\lambda)} \sim N(0, \tau_1)$  and  $u_i^{(\lambda)} \sim N(0, \tau_2)$ .

- The cAIC has value of 2103.9 for M4. Thus, cAIC selects M4 as the best-fitting model.



- Longitudinal data set in the R package JM (Komarek, 2015) from a Mayo Clinic trial on 312 patients with primary biliary cirrhosis (PBC) conducted in 1974-1984.
- There are 1 to 5 visits per subject performed at time of months. At each visit, measurements of 3 response variables are observed.
- Komarek (2015) used 260 subjects known to be alive at 910 days of follow-up, and only the longitudinal measurements by this point will be considered.

subject : 260 subjects

day : time of day = month  $\times$  30.4375

month =  $x$  : time of month

lbili =  $y_1$  : continuous logarithmic bilirubin

platelet =  $y_2$  : discrete platelet count

spiders =  $y_3$  : dichotomous indication of blood vessel malformations

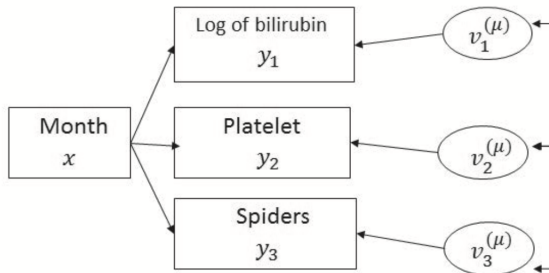


Figure: Path diagram for the MDHGLM fitted to the PBC data

## Multivariate model for 3 responses

- We consider a multivariate model for three response variables with a covariate  $x_{it}$  for the  $t$ th visit of the  $i$ th patient.

$$y_{1it}|v_{1i} \sim N(\mu_{1it}, \phi_{1i})$$

$$\text{with } \mu_{1it} = \beta_{10}^{(\mu)} + \beta_{11}^{(\mu)} x_{it} + v_{1i}^{(\mu)} \text{ and } \log \phi_{1i} = \beta_{10}^{(\phi)} + \beta_{11}^{(\phi)} x_{it}$$

$$y_{2it}|v_{2i} \sim N(\mu_{2it}, \phi_{2i})$$

$$\text{with } \mu_{2it} = \beta_{20}^{(\mu)} + \beta_{21}^{(\mu)} x_{it} + v_{2i}^{(\mu)} \text{ and } \log \phi_{2i} = \beta_{20}^{(\phi)} + \beta_{21}^{(\phi)} x_{it}$$

$$y_{3it}|v_{3i} \sim \text{Bernoulli}(p_{3it})$$

$$\text{with } \log \left( \frac{p_{3it}}{1 - p_{3it}} \right) = \beta_{30}^{(\mu)} + \beta_{31}^{(\mu)} x_{it} + v_{3i}^{(\mu)}$$

where the random effects follow multivariate normal distribution :

$$\begin{pmatrix} v_{1i}^{(\mu)} \\ v_{2i}^{(\mu)} \\ v_{3i}^{(\mu)} \end{pmatrix} \sim \text{MVN} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_1 & \rho_1 \lambda_{1,2}^* & \rho_2 \lambda_{1,3}^* \\ \rho_1 \lambda_{2,1}^* & \lambda_2 & \rho_3 \lambda_{2,3}^* \\ \rho_2 \lambda_{3,1}^* & \rho_3 \lambda_{3,2}^* & \lambda_3 \end{pmatrix} \right) \quad \text{with } \lambda_{j,k}^* = \sqrt{\lambda_j \lambda_k}.$$

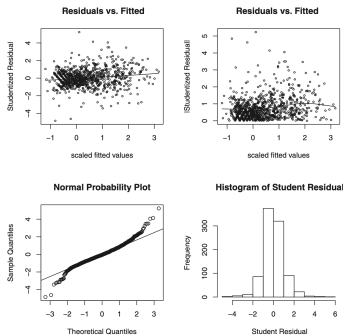


Figure 7.13 Model checking plots for multivariate HGLM of  $y_1$  on the primary biliary cirrhosis data.

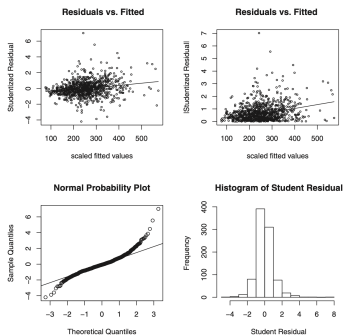


Figure 7.14 Model checking plots for multivariate DHGLM of  $y_2$  on the primary biliary cirrhosis data.

- Under the multivariate HGLM, we see that many large outliers exist.

Multivariate DHGLM allowing heavy-tailed distributions for  $y_1$  and  $y_2$

$$\log \phi_{1i} = \beta_{10}^{(\phi)} + \beta_{11}^{(\phi)} x_{it} + v_{1i}^{(\phi)} \text{ with } v_{1i}^{(\phi)} \sim N(0, \alpha_1)$$

$$\log \phi_{2i} = \beta_{20}^{(\phi)} + \beta_{21}^{(\phi)} x_{it} + v_{2i}^{(\phi)} \text{ with } v_{2i}^{(\phi)} \sim N(0, \alpha_2)$$

- cAIC shows that DHGLM (cAIC=13068.1) is better fit than HGLM (cAIC=19776.5).
- We can see that most outliers in multivariate HGLM disappear allowing heavy-tailed distribution for  $y_1$  and  $y_2$ .
- Thus, we select DHGLM which gives robust estimators against outliers.

## Review : DHGLM with ignorable missingness

- In Chapter 6, we analyzed the schizophrenic behavior data from an eye-tracking experiment with a visual target moving back and forth along a horizontal line on a screen (Rubin and Wu, 1997).
- We assume that the missing data are missing at random (MAR).
- We proposed using a DHGLM with

$$y_{ij} = \beta_0^{(\mu)} + x_{1ij}\beta_1^{(\mu)} + x_{2ij}\beta_2^{(\mu)} + t_j\beta_3^{(\mu)} + sch_i\beta_4^{(\mu)} + sch_i \cdot x_{1ij}\beta_5^{(\mu)} \\ + sch_i \cdot x_{2ij}\beta_6^{(\mu)} + v_i^{(\mu)} + e_{ij}$$

where  $v_i^{(\mu)} \sim N(0, \lambda)$  is the subject random effect, and  $e_{ij} \sim N(0, \phi)$ .

$$\log(\phi_i) = \beta_0^{(\phi)} + sch_i\beta_1^{(\phi)} + sch_i v_i^{(\phi)}$$

where  $v_i^{(\phi)} \sim N(0, \tau)$  are the dispersion random effects.

- We call this model DI (DHGLM with ignorable missingness).

## DN : DHGLM with non-ignorable missingness

- According to the physicians, missingness could be caused by eye blinks which are related to eye movements (responses) (Goossens and Opstal, 2000).
- This leads to the following model for missing data.
- $\delta_{ij} = y_{2ij}$  : indicator variables. 1(missing), 0(otherwise)

$$\eta = \Phi^{-1}(p_{ij}) = \delta_0 + x_{1ij}\delta_1 + x_{2ij}\delta_2 + sex_i\delta_3 + schi_i\delta_4 + sex_i \cdot x_{1ij}\delta_5 \\ + sex_i \cdot x_{2ij}\delta_6 + sex_i \cdot schi_i\delta_7 + \rho y_{ij}^*$$

where  $p_{ij} = P(\delta_{ij} = 1)$ .

- We can consider the model DI as well as DN with the probit model having two responses :  $y_1$  for a continuous response and  $y_2$  for a missing indicator.

DI DHGLM with ignorable missingness where  $\rho = 0$

DN DHGLM with non-ignorable missingness where  $\rho \neq 0$

- The negative value of  $\hat{\rho}$  supports the physicians' opinions that lower values of the response are more likely to be missing at each cycle.
- However, the conclusions concerning non-ignorable missingness depend crucially on untestable distributional assumptions. Thus, sensitivity analysis has been recommended.
- Fortunately, the analysis of the responses in there data indicates that they are not sensitive to the assumptions about the heavy tails or the missing mechanism (Yun and Lee, 2006).



- Lee, Nelder, and Pawitan (2017) considered law school admission data of Bock and Lieberman (1970), consisting of 6 items for law school admission test with 350 subjects.

$y_1 \sim y_6$  : items for law school admission test. 1(correct), 0(not correct)

subject,  $x$  : 350 subjects

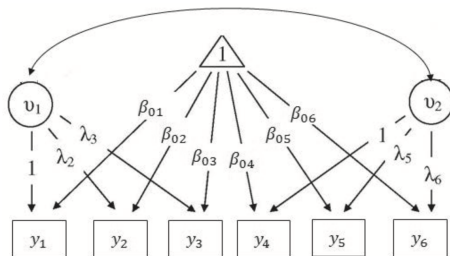


Figure: Path diagram for the binary 2-factor model

## Binary 2-factor model

- $\pi_{ij} = P(y_{ij} = 1 | \mathbf{v}_i)$
- Consider a binary 2-factor model.

$$\text{logit}(\boldsymbol{\pi}_i) = \boldsymbol{\beta}_0 + \boldsymbol{\Lambda} \mathbf{v}_i$$

where  $\boldsymbol{\pi}_i = (\pi_{i1}, \dots, \pi_{i6})^\top$  and  $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{06})^\top$ . Respectively,

$$\boldsymbol{\Lambda}^\top = \begin{pmatrix} 1 & \lambda_2 & \lambda_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \lambda_5 & \lambda_6 \end{pmatrix}$$

and  $\mathbf{v}_i = (v_{i1}, v_{i2})^\top \sim \text{BVN} \left( \mathbf{0}, \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \right)$ .

## 1-factor model

- We also consider 1 factor model, which is equivalent to assume the correlation between  $v_{1i}$  and  $v_{2i}$  being  $\pm 1$ .

$$\text{logit}(\pi_i) = \beta_0 + \mathbf{\Lambda} w_{1i}$$

where

$$\mathbf{\Lambda}^\top = (1 \quad \lambda_2 \quad \lambda_3 \quad \lambda_4 \quad \lambda_5 \quad \lambda_6)$$

and  $w_{1i} \sim N(0, \gamma_{11})$ .

- 1-factor model has  $\text{cAIC} = 2371.7$  which is less than 2-factor model ( $\text{cAIC} = 2548.6$ ). Thus, cAICs clearly prefers the 1-factor model.

## Chapter 8. Survival Analysis

- In this chapter we study the analysis of incomplete data, caused by censoring in event-time survival data.
- Cox's proportional hazards model is widely used for the analysis of survival data.
- Frailty models with a non-parametric baseline hazard extend proportional hazards model by allowing random effects in hazards and have been widely adopted for the analysis of survival data (Hougaard, 2000; Duchateau and Janssen, 2008).
- Using h-likelihood theory we can show that Poisson HGLM algorithms can be used to fit these models.
- Ha, Lee, and Song (2001) showed that with the h-likelihood it is easy to eliminate nuisance parameters by using a plug-in method and a fast estimation algorithm can thereby be used.
- Either a log-normal or gamma distribution can be used as the frailty distribution. Therefore, normal and log-gamma distribution can be adopted for the log frailties.

- Data consists of right censored observations from  $q$  subjects, with  $n_i$  observations each ( $i = 1, \dots, q$ ).
- $n = \sum_i n_i$  : total sample size
- $T_{ij}$  : survival time for the  $j$ -th observation of the  $i$ -th subject ( $j = 1, \dots, n_i$ ).
- $C_{ij}$  : corresponding censoring time
- $y_{ij} = \min\{T_{ij}, C_{ij}\}, \quad \delta_{ij} = I(T_{ij} \leq C_{ij})$
- $u_i$  : unobserved frailty for the  $i$ -th subject
- The conditional hazard function of  $T_{ij}$  is of the form

$$\lambda_{ij}(t|u_i) = \lambda_0(t) \exp(x_{ij}^\top \beta) u_i$$

where  $\lambda_0(\cdot)$  is an unspecified baseline hazard function and  $\beta = (\beta_1, \dots, \beta_p)^\top$  is a vector of regression parameters for the fixed covariates  $x_{ij}$ .

- Here, the term  $x_{it}^\top \beta$  doesn't include an intercept term because of identifiability.

## Frailty models

- We assume that the frailties  $u_i$  are i.i.d. random variables with a frailty parameter  $\alpha$ .
- We can assume gamma and log-normal distributions for  $u_i$ .
  - (i) gamma frailty with  $E(u_i) = 1$  and  $\text{var}(u_i) = \alpha$
  - (ii) log-normal frailty having  $v_i = \log u_i \sim N(0, \alpha)$

### Multi-component frailty models

- $\mathbf{X} : n \times p$  model matrix
- $\mathbf{Z}^{(r)} : n \times q_r$  model matrices correspond to the frailties  $\mathbf{v}^{(r)}$
- $\mathbf{v}^{(r)}, \mathbf{v}^{(l)}$  are independent for  $r \neq l$

$$\mathbf{X}\beta + \mathbf{Z}^{(1)}\mathbf{v}^{(1)} + \mathbf{Z}^{(k)}\mathbf{v}^{(k)} + \dots + \mathbf{Z}^{(k)}\mathbf{v}^{(k)}$$

- $\mathbf{Z}^{(r)}$  has indicator values such that  $Z_{st}^{(r)} = 1$  if observation  $s$  is a member of subject  $t$  in the  $r$ -th frailty component, and 0 otherwise.

- Data from study on the recurrence of infections in kidney patients who are using a portable dialysis machine (McGilchrist and Aisbett, 1991).
- Times until the 1st and 2nd recurrences of kidney infection in 38 patients are recorded.
- The catheter is later removed if infection occurs and can be removed for other reasons, which we regard as censoring (about 24%).

**id** : 38 patients

**time** : time until infection since the insertion of the catheter

**status** : censoring indicator. 1(infection), 0(censoring)

**age** : age of patient

**sex** : 1(male), 2(female)

**disease** : disease types. GN, AN, PKD, other

**frail** : estimated frailty (McGilchrist and Aisbett, 1991)

### Frailty model with 2 covariates

- We fit frailty models with 2 covariates, the *sex* and *age*.
- The survival times for the same patient are likely to be correlated because of a shared frailty describing the common patient's effect. So we consider *patient* as the frailty.
- The standard shared frailty model assumes that censoring times are independent of event times within clusters.
- For further discussions in survival analysis, see Ha, Jeong, and Lee (2017).



- Dataset is based on a tumorigenesis study of 50 litters of female rats (Mantel et al., 1977).
- For each litter, 1 rat was selected to receive the drug and the other 2 rats were placebo-treated controls.
- Death before occurrence of tumor yields a right-censored observation. 40 rats developed a tumor, leading to censoring of about 73%.
- The survival times for rats in a given litter may be correlated due to a random effect representing shared genetic or environmental effects.

**litter** : 50 litters

**rx** : 1(drug), 0(placebo)

**time** : time to development of tumor or death (weeks)

**status** : censoring indicator. 1(occurrence), 0(death, censored)

## Log-normal frailty model

- We fit models with 1 covariate, the *rx*. Also, we consider *litter* as the frailty.
- From the results, the *rx* group has significantly higher risk than the control group.
- The variance estimate of the frailty is  $\hat{\alpha} = 0.4272$  (SE=0.4232).
- Although we report the SE of the  $\alpha$ , one should not use it for testing the absence of frailty  $\alpha = 0$  (Vaida and Xu, 2000).
- A null hypothesis is on the boundary of the parameter space, so that the critical value of an asymptotic  $(\chi^2(0) + \chi^2(1))/2$  distribution is 2.71 at 5% significant level (Lee, Nelder, and Pawitan, 2017; Ha, Su, vester. Legrand, and MacKenzie, 2011).
- The difference in deviance  $-2p_{\beta,v}(h_p)$  between Cox's PHM without frailty (364.15) and log-normal frailty model (362.56) is 1.59(< 2.71), indicating that the frailty effect is non-significant.

- For the selection of a model between non-nested models, we may use 3 AIC criteria (Lee, Nelder, and Pawitan, 2017; Ha, Lee, and MacKenzie, 2007; Donohue, Overholser, Xu, and Vaida, 2011).

$$cAIC = -2h_0 + 2df_c$$

$$mAIC = -2p_v(h_p) + 2df_m$$

$$rAIC = -2p_{\beta,v}(h_p) + 2df_r$$

where  $h_0 = \ell_0^*$ .

- $df_c = \text{trace}\{D^{-1}(h_p, (\beta, v))D(h_0, (\beta, v))\}$  is an effective degrees of freedom adjustment for estimating the fixed and random effects. It is computed by using the Hessian matrices  $D(h_p, (\beta, v)) = -\partial^2 h_p / \partial(\beta, v)^2$ ,  $D(h_0, (\beta, v)) = -\partial^2 h_0 / \partial(\beta, v)^2$ .
- $df_m$  is the number of fixed parameters.
- $df_r$  is the number of dispersion parameters (Ha et al., 2007).

- Dataset consists of a placebo-controlled randomized trial of gamma interferon (rIFN-g) in the treatment of chronic granulomatous disease (CGD) (Fleming and Harrington, 1991).
- 128 patients from 13 centers were tracked for around 1 year.
- The survival times are the recurrent infection times of each patient.
- Censoring occurred at the last observation for all patients, except one, who experienced a serious infection on the date he left the study.
- About 63% of the data were censored.
- The recurrent infection times for a given patient are likely to be correlated. Also, each patient belongs to the 1 of the 13 centers.
- The correlation may be attributed to patient effect and center effect.

ex. CGD infection - cgd(page230).csv

**tstart - tstop** : recurrent infection times of each patient or censoring time

**id** : 128 patients

**center** : 13 centers

**treat** : rIFN-g or placebo

**status** : censoring indicator. 1(infection observed), 0(censored)

**random** : data of randomization

**sex, age, height, weight** : information about patients at study entry

**inherit** : pattern of inheritance

**steroids** : use of steroids at study entry. 1(yes), 0(no)

**propylac** : use of propylac antibiotics at study entry. 1(yes), 0(no)

**hos.cat** : categorization of the centers into 4 groups

**enum** : observation number within subject

## Multilevel log-normal frailty model

- We fit a multilevel log-normal frailty with 2 frailties and a single covariate, *treatment*. Here, the 2 frailties are random *center* and *patient* effects.

$$X\beta + Z^{(1)}v^{(1)} + Z^{(2)}v^{(2)}$$

$$v^{(1)} \sim N(0, \alpha_1 I_{q_1})$$

$$v^{(2)} \sim N(0, \alpha_2 I_{q_2})$$

where  $v^{(1)}$  is center frailty, and  $v^{(2)}$  is patient frailty.

- For testing the need for a random component ( $\alpha_1 = 0$  or  $\alpha_2 = 0$ ), we use the deviance  $-2p_{\beta,v}(h_p)$ , and fit the following 4 models.

**M1** Cox's model without frailty ( $\alpha_1 = 0, \alpha_2 = 0$ ) :  $-2p_{\beta,v}(h_p) = 707.48$

**M2** model without patient effect ( $\alpha_1 > 0, \alpha_2 = 0$ ) :  $-2p_{\beta,v}(h_p) = 703.66$

**M3** model without center effect ( $\alpha_1 = 0, \alpha_2 > 0$ ) :  $-2p_{\beta,v}(h_p) = 692.99$

**M4** multilevel model ( $\alpha_1 > 0, \alpha_2 > 0$ ) :  $-2p_{\beta,v}(h_p) = 692.95$

- The deviance difference between M3 and M4 ( $0.04 < 2.71 = \chi^2_{0.10}(1)$ ) indicates the absence of the random center effects.
- The deviance difference between M2 and M4 (10.71) indicates the necessity of random patient effects.
- The deviance difference between M1 and M3 (14.49) indicates the necessity of random patient effect even without random center effects.
- cAIC, mAIC and rAIC also choose M3 among the M1 - M4.

- Therneau and Lumley (2015) reported data on recurrences of bladder cancer, which were used to demonstrate methodology for recurrent event modeling (Wei et al., 1989).
- 85 patients were assigned to either thiotepa or placebo, and reports up to 4 recurrences for any patients.

**start** : start of interval (0 or previous recurrence time) (month)

**stop** : tumor recurrence or censoring time (month)

**event** : censoring indicator. 1(recurrence), 0(otherwise)

**id** : 85 patients

**rx** : treatment. 1(placebo), 2(thiotepa)

**number** : initial number of tumours. (8=8 or more)

**size** : size of largest initial tumor (cm)

**enum** : observation number within subject



## Log-normal frailty models

- We fit log-normal frailty models with 3 covariates, the *rx*, the *number*, and the *size* using HL(1,1).
- The thiotepa treatment has a marginally significant lower recurrent risk than in the placebo group controlling initial number of tumors.
- The deviance difference between Cox's PHM (1029.4) and log-normal frailty model (1024.1) is 5.3 ( $> 2.71$ ), indicating that the frailty effect is significant ( $p=0.011$ ).

## Gamma frailty model

- The results from gamma frailty model using HL(1,2) are slightly different to those of log-normal frailty, particularly for estimation of  $\beta$ .
- AIC indicates that log-normal and gamma frailty models are better than Cox's PHM.
- Between log-normal and gamma frailty models, AICs indicate that the log-normal frailty model is better than the gamma frailty model.

## Grouped duration model

- $T_i$  : duration time until occurrence of event for the  $i$ -th individual
- $T_i$  is not observed exactly, but we have information that the event happened in a specific interval.
- The durations are observed at the  $t$ -th time point  $a_t$  ( $t = 1, \dots, r$ ) with the  $a_0 = 0$ .

$$d_{it} = \begin{cases} 1 & i\text{-th individual experienced event during the } t\text{-th time interval} \\ 0 & \text{o.w} \end{cases}$$

- We considered the binary variable  $d_{it}$  as the response variable with the corresponding  $x_{it}$  observed at the  $(t - 1)$ -th time point  $a_{t-1}$ .

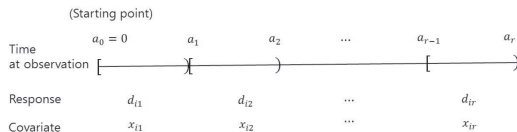


Figure: Structure of grouped duration data

- Given the random effect  $v_i$ , the conditional hazard rate at time  $T_i = u$  for  $a_{t-1} \leq u < a_t$  with  $t = 1, \dots, r$  of the form

$$\lambda(u|v_i) = \lambda_0(u) \exp(x_{ij}^\top \beta + v_i)$$

- $\lambda_0(\cdot)$  : baseline hazard function
- $\beta$  : regression coefficients of covariates of interests
- $x_{it}$  : risk factors observed over multiple time points ( $t = 1, \dots, r$ )
- $v_i$  : frailties of individuals
- Ha, Jeong, and Lee (2017) showed that the responses  $d_{it}$  follow the Bernoulli HGLM with the complementary log-log link

$$\log(-\log(1 - p_{it})) = \gamma_t + x_{it}^\top \beta + v_i$$

where  $p_{it} = Pr(d_{it} = 1|v_i)$  and  $\gamma_t = \log \int_{a_{t-1}}^{a_t} \lambda_0(u) du$ .

- For 1556 students in the Los Angeles area, onset of smoking is observed at each of 3 timepoints  $a_1$ ,  $a_2$ , and  $a_3$ .
- $a_1$  : starting time for investigation
- $a_2$  : 1-year follow-up and  $a_3$  : 2-year follow-up
- These event times are grouped at the 3 intervals  $[0, a_1)$ ,  $[a_1, a_2)$ ,  $[a_2, a_3)$ .
- For each student, we generate the following 4 responses.
  - (i)  $d_{i1} = 1$  if he/she started smoking at intervals at  $[0, a_1)$   
(smkonset = 1)
  - (ii)  $(d_{i1}, d_{i2}) = (0, 1)$  if he/she started smoking at intervals at  $[a_1, a_2)$   
(smkonset = 2)
  - (iii)  $(d_{i1}, d_{i2}, d_{i3}) = (0, 0, 1)$  if he/she started smoking at intervals at  $[a_2, a_3)$   
(smkonset = 3)
  - (iv)  $(d_{i1}, d_{i2}, d_{i3}) = (0, 0, 0)$  if he/she had not smoked until  $a_3$  (censored)  
(smkonset = 3)

`school`, `class`, `student` : 28 school, 134 class, 1556 students

`smkonset` :  $i$ -th time interval when the event occur

`event` : censoring indicator. 1(smoked), 0(otherwise)

`int` : constant value 1

`SexMale` : gender of student. 1(male), 0(female)

`cc` : indicating whether the school was randomized to a social-resistance classroom curriculum. 1(yes), 0(no)

`tv` : indicating whether the school was randomized to a media (television) intervention. 1(yes), 0(no)

`cctv` :  $cc \times tv$

### Grouped duration model

- 3 covariates are considered *SexMale*, *cc*, *tv*.
- Deviance difference between Cox's PHM (40189.8) and log-normal frailty model (40123.6) is 66.2 ( $> 2.71$ ), indicating the necessity of frailty.
- From the output, male has higher risk for smoking than female.
- Schools with *cc* or *tv* give lower risk for smoking to their students.

- For  $i = 1, \dots, q$ ,  $j = 1, \dots, n_i$ , and  $k = 1, \dots, K$ ,
- $T_{ijk}$  : time to type  $k$  for the  $j$ -th observation in the  $i$ -th cluster
- $C_{ij}$  : independent censoring time
- Observed event  $y_{ij} = \min(T_{ij1}, T_{ij2}, \dots, T_{ijK}, C_{ij})$
- Event indicator  $\delta_{ijk} = I(y_{ij} = T_{ijk})$
- The cause-specific hazard function conditional on the log-frailty  $\mathbf{v}_i = (v_{i1}, \dots, v_{iK})$  is

$$\lambda_{ijk}(t|\mathbf{v}_i) = \lambda_{0k}(t) \exp(\mathbf{x}_{ij}^\top \boldsymbol{\beta}_k + v_{ik})$$

where  $\lambda_{0k}(t)$  is the unspecified baseline hazard function for event type  $k$ .

- $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kp})^\top$  : fixed parameters for event type  $k$
- $\mathbf{x}_{ij}$  : fixed covariates
- $v_{ik}$  : random effect for type  $k$  event in cluster  $i$

## Competing risk models

- Consider  $K = 2$ .
- Event times from cause 1 and 2 would follow a cause-specific proportional hazards model

$$\lambda_{ij1}(t|v_i) = \lambda_{01}(t) \exp(x_{ij}^T \beta_1 + v_{i1})$$

$$\lambda_{ij2}(t|v_i) = \lambda_{02}(t) \exp(x_{ij}^T \beta_2 + v_{i2})$$

where  $v_{i1}$  and  $v_{i2}$  might be correlated.

- In the traditional cause-specific analysis, patients who failed from cause 2 are treated as censored for the analysis of type 1 events, which ignores a potential correlation between  $v_{i1}$  and  $v_{i2}$ .
- Competing risks data usually arise when an occurrence of a competing event prevents the occurrence of the event of interest.
- Treating the competing event as a censoring can lead to biased results (Pepe and Mori, 1993).



- We used a simulated data set generated in the R package `crrSC` (Zhou et al., 2012, 2015).
- The data consists of a data frame with 200 observations.

`ftime = time` : event time

`fstatus = status` : event type. 1(event of interest, 112 observations), 2(competing event, 47 observations), 0(censoring, 41 observations)

`x = z` : binary covariate generated with probability of 0.5

`ID` : 100 cluster with each cluster size 2

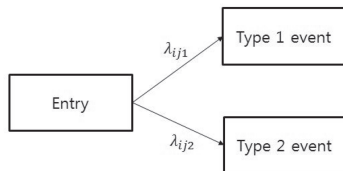


Figure: Path diagram for the competing risk frailty model

## Cause-specific hazard frailty model

- Consider the cause-specific hazard frailty model (Ha, Jeong, and Lee, 2017).
- $\lambda_{ijk}$  : conditional hazard function for the  $j$ -th observation in the  $i$ -th cluster that failed from cause  $k$  (given a shared log-frailty  $v_i$ )

$$\lambda_{ij1}(t|v_i) = \lambda_{01}(t) \exp(x_{ij}^\top \beta_1 + v_i)$$

$$\lambda_{ij2}(t|v_i) = \lambda_{02}(t) \exp(x_{ij}^\top \beta_2 + \gamma v_i)$$

where  $v_i \sim N(0, \sigma^2)$

- If  $\gamma > 0$  [ $\gamma < 0$ ], a cluster with higher frailty in type 1 event will experience an earlier [delayed] type 2 events (Huang and Wolfe, 2002).
- $\gamma = 1$  : the effect of the frailty is identical for both events.
- $\gamma = 0$  : two event rates are not associated.
- The estimate of shared parameter  $\hat{\gamma} = -1.218$  shows a negative association between 2 events.

# H-likelihood theory for the frailty model

- The h-likelihood gives a straightforward way of handling non-parametric baseline hazards.
- The h-likelihood is defined by

$$h = h(\beta, \lambda_0, \alpha) = \ell_0 + \ell_1$$

- $\ell_0 = \sum_{ij} \log f(y_{ij}, \delta_{ij} | u_i; \beta, \lambda_0) = \sum_{ij} \delta_{ij} \{\log \lambda_0(y_{ij}) + \eta_{ij}\} - \sum_{ij} \Lambda_0(y_{ij}) \exp(\eta_{ij})$
- $\ell_1 = \sum_i \log f(v_i; \alpha).$

$$\begin{aligned}\ell_0 &= \sum_{ij} \log \left( \{S(y_{ij})\}^{1-\delta_{ij}} \{f(y_{ij})\}^{\delta_{ij}} \right) \\ &= \sum_{ij} \log \left( \exp(-\Lambda(y_{ij})) \{\lambda(y_{ij})\}^{\delta_{ij}} \right) \\ &= \sum_{ij} \{-\Lambda(y_{ij}) + \delta_{ij} \log \lambda(y_{ij})\} \\ &= \sum_{ij} -\Lambda_0(y_{ij}) \exp(\eta_{ij}) + \sum_{ij} \delta_{ij} \{\log \lambda_0(y_{ij}) + \eta_{ij}\}\end{aligned}$$

## H-likelihood theory for the frailty model

- The functional form of  $\lambda_0(t)$  is unknown. Hence, we consider  $\Lambda_0(t)$  to be a step function with jumps at the observed event time (Breslow, 1972).

$$\Lambda_0(t) = \sum_{k: y_{(k)} \leq t} \lambda_{0k}$$

where  $y_{(k)}$  is the  $k$ -th smallest distinct event time among the  $y_{ij}$ 's, and  $\lambda_{0k} = \lambda_0(y_{(k)})$ .

- Ha, Lee and Song(2001) proposed the use of the profile h-likelihood with  $\lambda_0$  eliminated,  $r^* := h|_{\lambda_0 = \hat{\lambda}_0}$ , given by

$$r^* = r^*(\beta, \alpha) = \ell_0^* + \ell_1$$

where  $\ell_0^* = \sum_{ij} \log f^*(y_{ij}, \delta_{ij} | u_i; \beta, \hat{\lambda}_0)$  does not depend on  $\lambda_0$ . And

$$\hat{\lambda}_{0k}(\beta, v) = \frac{d_{(k)}}{\sum_{(i,j) \in R_{(k)}} \exp(\eta_{ij})}$$

are solutions of the estimating equations,  $\partial h / \partial \lambda_{0k} = 0$ .  $d_{(k)}$  is the number of events at  $y_{(k)}$  and  $R_{(k)} = \{(i, j) : y_{ij} \geq y_{(k)}\}$  is the risk set at  $y_{(k)}$ .

# H-likelihood theory for the frailty model

- Therneau and Grambsch (2000) and Ripatti and Palmgren (2000) proposed h-likelihood, called penalized partial likelihood (PPL)  $h_p$ .

$$h_p(\beta, \nu, \alpha) = \sum_{ij} \delta_{ij} \eta_{ij} - \sum_k d_{(k)} \log \left\{ \sum_{ij \in R_{(k)}} \exp(\eta_{ij}) \right\} + \ell_1$$

- Ha, Lee, and Song (2001) and Ha et al. (2010) have shown that  $r^*$  is proportional to the PPL  $h_p$ .

$$\begin{aligned} r^* &= \sum_k d_{(k)} \log \hat{\lambda}_{0k} + \sum_{ij} \delta_{ij} \eta_{ij} - \sum_k d_{(k)} + \ell_1 \\ &= h_p + \sum_k d_{(k)} \{\log d_{(k)} - 1\} \end{aligned}$$

where  $\sum_k d_{(k)} \{\log d_{(k)} - 1\}$  is a constant which does not depend upon unknown parameters.

- Thus, the h-likelihood procedure for HGLMS of Lee and Nelder (1996, 2001) can be extended to frailty models based on  $h_p$  (Ha et al., 2010).

## Estimator of baseline hazard function, $\lambda_0(t)$

- When there is no such random effects,

$$L_p(h_0(t)) = \left[ \prod_{i=1}^D \lambda_0(y_{(i)}) \exp(\beta^T X_{(i)}) \right] \exp \left[ - \sum_{j=1}^n \Lambda_0(y_j) \exp(\beta^T X_j) \right]$$

- Let  $\lambda_{0i} = \lambda(y_{(i)})$  ( $i = 1, \dots, D$ ) and  $\Lambda_0(y_j) = \sum_{y_{(i)} \leq y_j} \lambda_{0i} = \sum_{i=1}^D R_j(y_{(i)}) \lambda_{0i}$ . Then,

$$L_p(\lambda_{01}, \lambda_{02}, \dots, \lambda_{0D}) = \left[ \prod_{i=1}^D \lambda_{0i} \exp(\beta^T X_{(i)}) \exp \left[ - \lambda_{0i} \sum_{j=1}^n R_j(y_{(i)}) \exp(\beta^T X_j) \right] \right]$$

- The maximum likelihood estimator of  $\lambda_{0i}$  is given by

$$\hat{\lambda}_{0i} = \frac{1}{\sum_{j=1}^n R_j(y_{(i)}) \exp(\beta^T X_j)}$$
$$\hat{\Lambda}_0(t) = \int_0^t \frac{\sum_{i=1}^n dN_i(u)}{\sum_{j=1}^n R_j(u) \exp(\beta^T X_j)}$$

where  $N_i(t)$  counts the number of events in  $[0, t]$  for unit  $i$  and  $\sum N_i(t) = N(t)$ .

## Estimator of baseline hazard function, $\lambda_0(t)$

- Note that

$$\begin{aligned}\sum_{j=1}^n \hat{\lambda}_0(y_j) \exp(\beta^\top X_j) &= \sum_{j=1}^n \int_0^\infty I(y_j \geq t) \exp(\beta^\top X_j) \frac{\sum_{i=1}^n dN_i(t)}{\sum_{i=1}^n R_i(t) \exp(\beta^\top X_j)} \\ &= \int_0^\infty dN(t)\end{aligned}$$

- Then,

$$\begin{aligned}L_p(\hat{\lambda}_0(t)) &= \left[ \prod_{i=1}^D \lambda_0(y_{(i)}) \exp(\beta^\top X_{(i)}) \right] \exp \left[ - \sum_{j=1}^n \lambda_0(y_j) \exp(\beta^\top X_j) \right] \\ &= \left[ \prod_{i=1}^D \hat{\lambda}_0(y_{(i)}) \exp(\beta^\top X_{(i)}) \right] \exp \left[ - \int_0^\infty dN(t) \right]\end{aligned}$$

## Estimator of baseline hazard function, $\lambda_0(t)$

- Note that

$$\begin{aligned}\ell_0 &= \sum_{ij} \{\delta_{ij} \{\log \lambda_0(y_{ij}) + \eta_{ij}\} - \Lambda_0(y_{ij}) \exp(\eta_{ij})\} \\ &= \sum_k d_{(k)} \log \lambda_{0k} + \sum_{i,j} \delta_{ij} \eta_{ij} - \sum_k \lambda_{(0k)} \left\{ \sum_{(i,j) \in R(y_{(k)})} \exp(\eta_{ij}) \right\}\end{aligned}$$

- Plugging in  $\hat{\lambda}_{0k}(\beta, \nu) = \frac{d_{(k)}}{\sum_{(i,j) \in R(y_{(k)})} \exp(\eta_{ij})}$ ,

$$\ell_0^* = \sum_k d_{(k)} \log \hat{\lambda}_{0k} + \sum_{i,j} \delta_{ij} \eta_{ij} - \sum_k d_{(k)}$$



## Chapter 9. Joint Models

- In this chapter, we consider data analysis for multivariate responses where at least one response is time-to-event.
- Separated analysis ignoring the inherent association between the outcomes from the subject can lead to a biased result (Guo and Carlin, 2004).
- Thus, joint modeling has been widely studied (Henderson et al. 2000; Ha et al., 2003; Rizopoulos, 2012).
- An unobserved random effect can be used to account for the association among multivariate outcomes.
- For the analysis of such dataset, the h-likelihood approach is very effective.

- Dataset from the clinical study to investigate the chronic renal allograft dysfunction in renal transplants (Sung et al., 1998).
- The renal function is evaluated from the serum creatinine (sCr) values. Since the time interval between the consecutive measurements differs from patient to patient, we focus on the mean creatinine levels over 6 months.
- A Graft-loss time is observed from each patient.
- During the study period, there were 13 graft losses due to the kidney dysfunction. For other remaining patients, we assumed that the censoring occurred at the last follow-up time (about 88%).

id : 112 patients

month : visiting time (month)

cr : serum creatinine value (mg/dL)

sex : gender. 1(male), 0(female)

age : age of patients

icr : reciprocal of serum creatinine value =  $1/\text{cr}$

sur\_time : graft-loss time (month)

status : censoring indicator. 1(occurrence), 0(censoring)

- We are interested in investigating the effects of covariates over 2 response (sCr values and a graft-loss time).
- Ha et al. (2003) considered *month*, *sex* and *age* as covariates for sCr. Also they considered *sex* and *age* as covariates for the loss time.
- We consider the standard mixed linear model we use values  $1/\text{sCr}$  as responses  $y_{ij}$ .

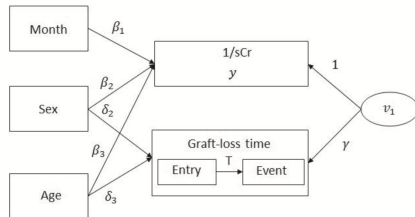


Figure: Path diagram for the joint model for repeated measures and survival time

## Joint model

- For the 1/sCr values, consider a linear mixed model

$$y_{ij} = x_{1ij}^T \beta + v_{1i} + e_{ij}$$

where  $x_{1ij}$  are covariates,  $v_{1i} \sim N(0, \sigma_{v1}^2)$ , and  $e_{ij} \sim N(0, \sigma_e^2)$ .

- For graft-loss time  $t_i$ , consider a frailty model with the conditional hazard function

$$\lambda(t_i | v_{1i}) = \lambda_0(t_i) \exp(x_{2i}^T \delta + \gamma v_{1i})$$

where  $\lambda_0(t)$  is the baseline hazard function,  $x_{2i}$  are between-subject covariates, and  $\gamma$  is the shared parameter.

- Ha et al. (2003) considered a Weibull model for the baseline hazard function where  $\lambda_0(t_i) = \tau t^{\tau-1}$  with a shape parameter  $\tau$ .
- Also, we can fit the non-parametric baseline hazard model.
- The values of cAIC show that non-parametric baseline hazard model is preferred to the Weibull baseline hazard model.

## Separate model

- We can fit 2 random effect models separately with LMM and following frailty model.

$$y_{ij} = \mathbf{x}_{1ij}^T \beta + v_{1i} + e_{ij}$$

where  $\mathbf{x}_{1ij}$  are covariates,  $v_{1i} \sim N(0, \sigma_{v1}^2)$ , and  $e_{ij} \sim N(0, \sigma_e^2)$ .

$$\lambda(t_i | v_{2i}) = \lambda_0(t_i) \exp(\mathbf{x}_{2i}^T \delta + v_{2i})$$

where  $v_{2i} \sim N(0, \sigma_{v2}^2)$ .

- The cAIC can be computed by adding cAIC from two models.
- We can see that joint models are preferred to corresponding separate models.

- Data were collected in a recent clinical trial to compare the efficacy and safety of 2 antiretroviral drugs in treating patients who had failed or were intolerant of zidovudine (AZT) therapy (Rizopoulos, 2015).
- 467 HIV-infected patients were enrolled and randomly assigned to receive either didanosine (ddI) or zalcitabine (ddC).
- The number of CD4 cells per  $\text{mm}^3$  of blood were recorded at study entry, and again at the 2, 6, 12, 18 month visits.
- Times to death were also recorded with a 40% censoring rate.

patient : 467 patients

time : the time to death or censoring

death : censoring indicator. 1(death), 0(censoring)

CD4 : the CD4 cells count

month : recorded time points

drug : ddC(zalcitabine), ddI(didanosine)

gender : male, female

prevOI : AIDS diagnosis at study entry

AZT : intolerance(AZT intolerance), failure(AZT failure)

start : start of time in the first interval

stop : end of time in the first interval

event : 1(death in the first interval), 0(censoring)

y :  $CD4^{1/2}$



- Rizopoulos (2015) considered a joint model for the square root of CD4 value  $y_{ij}$  for the  $j$ -th visit and the time to death  $t_i$  of the  $i$ -th patient.
- We consider *month* and *drug* as covariates for  $y_{ij}$ , and drug for  $t_i$ .

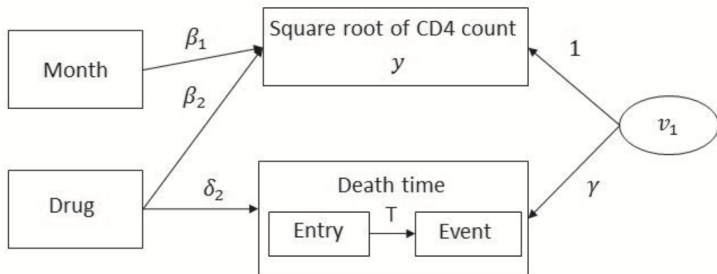


Figure: Path diagram for the joint model for repeated measures and survival time on AIDS data

## Joint model

- For the response  $y_{ij}$ , consider a linear mixed model.

$$y_{ij} = \mathbf{x}_{1ij}^T \beta + v_{1i} + e_{ij}$$

where  $\mathbf{x}_{1ij}^T$  are covariates,  $v_{1i} \sim N(0, \sigma_{v1}^2)$  and  $e_{ij} \sim N(0, \sigma_e^2)$ .

- For death time  $t_i$ , consider a frailty model with the conditional hazard function

$$\lambda(t_i | v_{1i}) = \lambda_0(t_i) \exp(\mathbf{x}_{2i}^T \delta + \gamma v_{1i})$$

where  $\lambda_0(t)$  is the baseline hazard function,  $\mathbf{x}_{2i}^T$  are between-subject covariates and  $\gamma$  is the shared parameter.

- Rizopoulos (2015) considered a Weibull model for the baseline hazard function where  $\lambda_0(t_i) = \tau t^{\tau-1}$  with a shape parameter  $\tau$ .
- We can also fit a non-parametric baseline hazard model.
- The values of cAIC show that the Weibull baseline hazard model is preferred to the non-parametric baseline hazard model.

## Separate model

- We can fit 2 random effect models separately with following frailty model.

$$y_{ij} = x_{1ij}^T \beta + v_{1i} + e_{ij}$$

where  $x_{1ij}^T$  are covariates,  $v_{1i} \sim N(0, \sigma_{v1}^2)$  and  $e_{ij} \sim N(0, \sigma_e^2)$ .

$$\lambda(t_i | v_{2i}) = \lambda_0(t_i) \exp(x_{2i}^T \delta + v_{2i})$$

where  $v_{2i} \sim N(0, \sigma_{v2}^2)$ .

- We can see that joint models are preferred to corresponding separate models.

- In chapter 7, we analyzed the PBC data available in the R package JM (Rizopoulos, 2015).
- We fit joint model for the logarithm of serum bilirubin (mg/dL)  $y_{ij}$  for the  $j$ -th visit and the time to event  $t_i$  of the  $i$ -th event.
- We consider *year*, *sex*, and *drug* as covariates for  $y_{ij}$ .
- We also consider *sex* and *drug* for  $t_i$ .

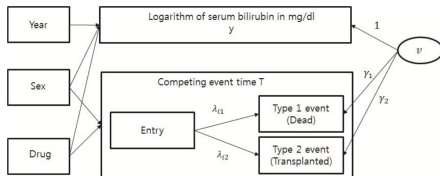


Figure 9.3 Path diagram for the joint model of repeated measure and competing event time on PBC data.

**Figure:** Path diagram for the joint model for repeated measures and competing event time on PBC data

id : 312 patients

serBilir : serum bilirubin (mg/dL)

y : serBilir<sup>1/2</sup>

years : number of years between registration and the earlier of death, transplantation, or study analysis time

status : censoring indicator. 2(transplanted), 1(dead), 0(alive)

year : number of years between enrollment and this visit date

drug : 1(D-penicillamine), 0(placebo)

sex : gender of patients. 1(male), 0(female)

Variable	Description
ascites	Yes or No
hepatomegaly	Yes or No
spiders	Yes or No
edema	No edema, edema no diuretics, edema despite diuretics
serChol	serum cholesterol (mg/dL)
albumin	albumin (mg/dL)
alkaline	alkaline phosphatase in
SGOT	SGOT (U/ml)
platelets	platelets per cubic ml / 1000
prothrombin	prothrombin time (sec)
histologic	histologic stage of disease
status2	1(death), 0(transplanted or alive)

## Joint Model

- For  $y_{ij}$ , consider a linear mixed model.

$$y_{ij} = x_{1ij}^T \beta + v_i + e_{ij}$$

where  $x_{1ij}^T$  are covariates,  $v_i \sim N(0, \sigma_v^2)$  and  $e_{ij} \sim N(0, \sigma_e^2)$ .

- For the time event  $t_i$ , consider the cause-specific hazard frailty model for competing risk.
- Given a shared log-frailty  $v_{1i}$ , the conditional hazard function  $\lambda_{ik}$  for the  $i$ -th patient that failed from cause  $k$  ( $k = 1, 2$ ) can be expressed as

$$\lambda_{i1}(t|v_i) = \lambda_{01}(t_i) \exp(x_{2i}^T \delta_1 + \gamma_1 v_i)$$

$$\lambda_{i2}(t|v_i) = \lambda_{02}(t_i) \exp(x_{2i}^T \delta_2 + \gamma_2 v_i)$$

where  $\lambda_{0k}(t)$  is an unspecified baseline hazard function for cause  $k$ ,  $\delta_k$  is regression parameters for cause  $k$ .

- The estimates of shared parameters  $\hat{\gamma}_1 = 1.271$  and  $\hat{\gamma}_2 = 1.189$  show a positive associations between  $y_{ij}$  and 2 events.
- The visiting year effect for  $y_{ij}$  is positively very significant.
- The effect of drug is not significant for  $y_{ij}$  and for death event, but it is negatively significant for trasplanted event.
- The effect of sex is positively significant for  $y_{ij}$  and for death event, but it is not significant for transplanted event.
- However, when we fit the competing risk model for  $t_i$  removing response  $y_{ij}$ , the effect of drug is not significant for transplanted event.



- $y_{ij}$  : the  $j$ th repeated response of  $i$ -th subject ( $i = 1, \dots, q, j = 1, \dots, n_i$ )
- $T_i$  : a single event time of  $i$ -th subject
- $C_i$  : the corresponding censoring time
- We observe  $t_i^* = \min(T_i, C_i)$  and  $\delta_i = I(T_i \leq C_i)$ .
- Linear Mixed Model for  $y$  :

$$y_{ij} = x_{1ij}^T \beta_1 + v_i + \epsilon_{ij}$$

where  $v_i \sim N(0, \alpha)$  and  $\epsilon \sim N(0, \phi)$  are independent.

- Frailty Model for  $T$  :

$$\lambda_i(t|v_i) = \lambda_0(t) \exp(x_{2i}^T \beta_2 + \gamma v_i)$$

where  $\lambda_0$  is an unspecified baseline hazard function and  $\gamma$  is a real-valued association parameter that allows the magnitude of the association to be different between two outcomes,  $y_{ij}$  and  $T_i$ .

- The h-likelihood becomes

$$h = \sum_{i,j} \ell_{1ij} + \sum_i \ell_{2i} + \sum_i \ell_{3i}$$

where

$$\begin{aligned}\ell_{1ij} &= \ell_{1ij}(\beta_1, \phi; y_{ij} | v_i) \\ &= -\frac{1}{2} \log(2\pi\phi) - \frac{1}{2\phi} (y_{ij} - \eta_{1ij})^2 \\ \ell_{2i} &= \ell_{2i}(\beta_2, \lambda_0; t_i^*, \delta_i | v_i) \\ &= \delta_i (\log \lambda_0(t_i^*) + \eta_{2i}) - \Lambda_0(t_i^*) \exp(\eta_{2i}) \\ \ell_{3i} &= \ell_{3i}(\alpha; v_i) r \\ &= -\frac{1}{2} \log(2\pi\alpha) - \frac{1}{2\alpha} v_i^2\end{aligned}$$

- $\eta_{1ij} = x_{1ij}^T \beta_1 + v_i$  and  $\eta_{2i} = x_{2i}^T \beta_2 + v_i$  are linear predictors.

- Following Breslow (1972), we define the baseline cumulative hazard function  $\Lambda_0$  to be a step function with jumps  $\lambda_{0r} = \lambda_0(t_{(r)})$  at the observed event times  $t_{(r)}$ .

$$\Lambda_0(t) = \sum_{r: t_{(r)} \leq t} \lambda_{0r}.$$

where  $t_{(r)}$  is the  $r$ -th smallest distinct event time ( $r = 1, \dots, D$ ).

- The second term  $\sum_i \ell_{2i}$  of  $h$  becomes

$$\sum_i \ell_{2i} = \sum_r d_r \log \lambda_{0r} + \sum_i \delta_i \eta_{2i} - \sum_r \lambda_{0r} \left\{ \sum_{i \in R_r} \exp(\eta_{2i}) \right\}$$

where  $d_r$  is the number of events at  $t_{(r)}$  and  $R_r = \{i : t_i^* \geq t_{(r)}\}$  is the risk set at  $t_{(r)}$ .

- Following Ha et al. (2001), we use the profile h-likelihood  $h^*$  :

$$h^* = h|_{\lambda_0 = \hat{\lambda}_0} = \sum_{i,j} \ell_{1ij} + \sum_i \ell_{2i}^* + \sum_i \ell_{3i}$$

where

$$\begin{aligned} \sum_i \ell_{2i}^* &= \sum_i \ell_{2i}|_{\lambda_0 = \hat{\lambda}_0} = \sum_r d_r \log \hat{\lambda}_{0r} + \sum_i \delta_i \eta_{2i} - \sum_r d_r \\ \hat{\lambda}_{0r} &= \hat{\lambda}_{0r}(\beta_2, \nu) = \frac{d_r}{\sum_{i \in R_r} \exp(\eta_{2i})} \end{aligned}$$

are the solution of the estimating equations  $\frac{\partial h}{\partial \lambda_{0r}} = 0$  for  $r = 1, \dots, D$ .

- The penalized partial h-likelihood  $h_p$  is given by

$$h_p = \sum_{i,j} \ell_{1ij} + \sum_i \delta_i \eta_{2i} - \sum_r d_r \log \left\{ \sum_{i \in R_r} \exp(\eta_{2i}) \right\} + \sum_i \ell_{3i}$$

- The score equations for fixed and random effects  $(\beta_1, \beta_2, \nu)$  given dispersion parameters  $\psi = (\phi, \alpha, \gamma)^\top$  are

$$\frac{\partial h_p}{\partial \beta_1} = \frac{1}{\phi} X_1^\top (y - \mu_1)$$

$$\frac{\partial h_p}{\partial \beta_2} = X_2^\top (\delta - \hat{\mu}_2)$$

$$\frac{\partial h_p}{\partial \nu} = \frac{1}{\phi} Z_1^\top (y - \mu_1) + \gamma Z_2^\top (\delta - \hat{\mu}_2) - \frac{\nu}{\alpha}$$

where  $\mu_1 = X_1 \beta_1 + Z_1 \nu = \eta_1$ ,  $\hat{\mu}_2 = \exp \left( \log \hat{\Lambda}_0(t^*) + \eta_2 \right)$  with  $\eta_2 = X_2 \beta_2 + \gamma Z_2 \nu$ .

- $Z_1$  is  $n \times q$  group indicator matrix, and  $Z_2 = I_1$  which denotes a  $q \times q$  identity matrix.
- $\hat{\Lambda}_0(t) = \sum_{r: t_{(r)} \leq t} \hat{\lambda}_{0r}$  is the estimator of cumulative baseline hazard.

- This leads to the iterative least squares (ILS; see Ha et al. (2017)) joint equations for  $\theta = (\beta_1^\top, \beta_2^\top, \nu^\top)^\top$ , given by

$$\left( \begin{array}{ccc} X_1^\top W_1 X_1 & 0 & X_1^\top W_1 Z_1 \\ 0 & X_2^\top W_2 X_2 & X_2^\top (\gamma W_2) Z_2 \\ Z_1^\top W_1 X_1 & Z_2^\top (\gamma W_2) X_2 & \mathbf{Z}^\top \mathbf{W} \mathbf{Z} + Q \end{array} \right) \bigg|_{\theta=\theta^{(s)}} \theta^{(s+1)} = \left( \begin{array}{c} X_1^\top W_1 w_1 \\ X_2^\top w_2 \\ \mathbf{Z}^\top \mathbf{w}^* \end{array} \right) \bigg|_{\theta=\theta^{(s)}}$$

where  $W_1 = -\frac{\partial^2 h_p}{\partial \eta_1 \partial \eta_1^\top} = \frac{1}{\phi} I_n$ ,  $W_2 = -\frac{\partial^2 h_p}{\partial \eta_2 \partial \eta_2^\top}$ ,  $Q = -\frac{\partial^2 \ell_3}{\partial \nu \partial \nu^\top} = \frac{1}{\alpha} I_q$ ,  
 $w_1 = y$ ,  $w_2 = W_2 \eta_2 + (\delta - \hat{\mu}_2)$ , and

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ \gamma Z_2 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} W_1 & 0 \\ 0 & W_2 \end{pmatrix}, \quad \text{and} \quad \mathbf{w}^* = \begin{pmatrix} W_1 w_1 \\ w_2 \end{pmatrix}$$

Note here that  $\mathbf{Z}^\top \mathbf{W} \mathbf{Z} = Z_1^\top W_1 Z_1 + Z_2^\top (\gamma^2 W_2) Z_2$  and  $\mathbf{Z}^\top \mathbf{w}^* = Z_1^\top W_1 w_1 + \gamma Z_2^\top w_2$ .

- The fitting procedure consists of the following two steps.

(S1) Estimation of fixed and random effects  $\theta = (\beta_1^T, \beta_2^T, \nu^T)^T$  via the ILS equations.

(S2) Estimation of dispersion parameters  $\psi = (\phi, \alpha, \gamma)^T$  as follows.

### Estimation of $\psi$

- We used the adjusted profile h-likelihood, given by

$$p_{\theta}(h_p) = \left[ h_p - \frac{1}{2} \log \det \left\{ \frac{1}{2\pi} H(h_p, \theta) \right\} \right] \Big|_{\theta=\hat{\theta}}$$

where  $\hat{\theta} = \hat{\theta}(\psi)$  are solutions of  $\frac{\partial h_p}{\partial \theta} = 0$  for given  $\psi$ , and

$$H(h_p, \psi) = -\frac{\partial^2 h_p}{\partial \theta \partial \theta^T}$$

is observed information matrix for  $\theta$ .

- The estimating equations of  $\psi$  are given by

$$\frac{\partial p_{\theta}(h_p)}{\partial \psi} = 0$$

leading to the estimating equations

$$\hat{\phi} = \frac{(y - \hat{\mu}_1)^{\top} (y - \hat{\mu}_1)}{n - \kappa_0} \quad \text{and} \quad \hat{\alpha} = \frac{\hat{v}^{\top} \hat{v}}{q - \kappa_1}$$

where  $\kappa_0 = -\phi \operatorname{tr} \left\{ \hat{H}^{-1} \frac{\partial \hat{H}}{\partial \phi} \right\}$ ,  $\kappa_1 = -\alpha \operatorname{tr} \left\{ \hat{H}^{-1} \frac{\partial \hat{H}}{\partial \alpha} \right\}$ , and  $\hat{H} = H(h_p, \theta)|_{\theta=\hat{\theta}(\psi)}$ .

- The estimate of  $\gamma$  is also easily implemented via the Newton-Raphson method using the first and second derivatives.
- This approach can be extended to a joint model with competing-risk data (Ha et al., 2017).



# Chapter 10. Further Topics: Variable Selection

## Penalized least-square methods

- Many classical subset selection methods, such as forward/backward selection or best-subset selection, cannot be easily adapted to applications where the number of variables is much greater than the sample size.
- PLS methods is another way to perform variable selection. The general version of the PLS is the penalized likelihood criterion:

$$Q_{\lambda}(\beta) = \ell(\beta) - p_{\lambda}(\beta),$$

where  $\ell(\beta) = \sum_{i=1}^n \log f_{\phi}(y_i|\beta)$  is log-likelihood and  $p_{\lambda}(\beta)$  is penalty function.

- We can in general put variable selection of any GLM-based regression model in this framework.

- Consider the regression model

$$y_i = x_i^T \beta + e_i, \quad i = 1, \dots, n \quad (1)$$

where  $\beta$  is a  $p \times 1$  vector of fixed unknown parameters and  $e_i$ 's are i.i.d. with  $(0, \phi)$ .

- Variable selection procedure can be described as PLS estimation that minimizes

$$Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^T \beta)^2 + \sum_{j=1}^d p_\lambda(|\beta_j|)$$

where  $p_\lambda(\cdot)$  is a penalty function controlling model complexity.

- With the  $L_1$ -penalty, the PLS becomes **LASSO**:

$$Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|,$$

which automatically sets to zero those predictors with small estimated OLS coefficients, thus performing simultaneous estimation and variable selection.

- LASSO has been criticized on the ground that it typically selects too many variables to prevent over-shrinkage of the regression coefficients (Radchenko and James, 2008); otherwise, regression coefficients of selected variables are often over-shrunk.
- To improve LASSO, various other penalties have been proposed: SCAD penalty for oracle estimators (Fan and Li, 2001), adaptive LASSO (Zou, 2006), elastic net (Zou and Hastie, 2005).

- With the  $L_2$ -penalty, the PLS becomes **ridge regression**:

$$Q_\lambda(\beta) = \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top \beta)^2 + \lambda \sum_{j=1}^p |\beta_j|^2.$$

- In this case, all variables are kept in the model but the resulting estimates are the shrunk versions of the OLS estimates.
- Ridge regression often achieves good prediction performance, but it cannot produce a parsimonious model.
- The ridge estimator is the same as random-effect estimator where  $\beta_j$  are i.i.d. normal random effects.

- We describe a random effect model that generates a family of penalties, including the normal type, LASSO type and a new unbounded penalty at the origin.
- In regression model (1), suppose  $\beta$  are random effects; conditional on  $u_j$ , we have

$$\beta_j|u_j \sim N(0, u_j\theta), \quad (2)$$

where  $\theta$  is a fixed dispersion parameter and  $u_j$ 's are i.i.d. random variables.

- In this random effect model, sparseness or selection is achieved in a transparent way, since  $u_j \approx 0$  implies  $\beta_j \approx 0$ .
- Since  $\theta u_j = (a\theta)(u_j/a)$  for any  $a > 0$ ,  $\theta$  and  $u_j$  are not separately identifiable. Thus, we constrain  $E(u_j) = 1$  as in HGLMs, which imposes a constraint on random effect estimates such that  $\sum_{j=1}^p \hat{u}_j/p = 1$ .

- Assume that  $u_j$  's are from the gamma distribution with a parameter  $w$  such that

$$f_w(u_j) = (1/w)^{1/w} \frac{1}{\Gamma(1/w)} u_j^{1/w-1} e^{-u_j/w},$$

having  $E(u_j) = 1$  and  $Var(u_j) = w$ .

- Model (2) can be re-written as  $\beta_j = \sqrt{\tau_j} e_j$  with  $e_j \sim N(0, 1)$  and

$$\log \tau_j = \log \theta + v_j$$

where  $v_j \equiv \log u_j$ , which defines a DHGLM together with model (1).

- Then h-loglikelihood  $h = h_1 + h_2$  is given by

$$h_1 = \sum_{i=1}^n \log f_{\phi}(y_i|\beta) = -\frac{n}{2} \log(2\pi\phi) - \frac{1}{2\phi} \sum_{i=1}^n (y_i - \mathbf{x}_i^{\top} \beta)^2,$$

$$h_2 = \sum_{j=1}^p \{\log f_{\theta}(\beta_j|u_j) + \log f_w(v_j)\},$$

$$\log f_{\theta}(\beta_j|u_j) = -\frac{1}{2} \{\log(2\pi\theta) + \log u_j + \beta_j^2/(\theta u_j)\},$$

$$\log f_w(v_j) = -\log(w)/w - \log \Gamma(1/w) + v_j/w - \exp(v_j)/w.$$

The outline of the estimation scheme using IWLS as follows:

- For given  $(\beta, w, \phi, \theta)$  solving  $\partial h / \partial u = 0$  gives the random effect estimator

$$\hat{u}_j \equiv \hat{u}_j(\beta) = \frac{1}{4} \{8w\beta_j^2 / \theta + (2 - w)^2\}^{1/2} + (2 - w). \quad (3)$$

- For given  $\hat{u}$ , Lee and Oh (2014) proposed to update  $\beta$  based on the model (1) with  $\beta$  satisfying (2). This is a purely random effect model

$$Y = X\beta + e$$

where  $e \sim N(0, \Sigma \equiv \text{diag}\{\phi\})$  and  $\beta \sim N(0, D \equiv \text{diag}\{\hat{u}_j\theta\})$ .

- From the mixed model equation, we update  $\beta$  by solving

$$(X^\top X + W_\lambda)\beta = X^\top y \quad (4)$$

where  $W_\lambda \equiv \text{diag}\{\lambda / \hat{u}_j\}$  and  $\lambda = \phi / \theta$ .



- It is clear that  $\hat{\beta}_j = 0$  when  $\hat{u}_j = 0$ . If we allow threshold by setting small  $\hat{u}_j$  to zero, then the corresponding weight  $1/\hat{u}_j$  in  $W_\lambda$  is undefined.
- We could exclude the corresponding predictors from (4), but instead we employ a perturbed random effect estimate  $\hat{u}_{\delta,k} = \lambda(|\beta_k| + \delta)/|p'_\lambda(|\beta_k|)|$  for a small positive  $\delta = 10^{-8}$ . Then the weight is always defined and the solution is nearly identical to the original IWLS as long as  $\delta$  is small.
- In random effect models, we used ML or REML estimates for  $(w, \phi, \theta)$  and computed tuning parameter  $\lambda$  as the ratio  $\phi/\theta$ . On the other hand, in variable selection, it is common to estimate  $\lambda$  by using K-fold cross validation since  $\lambda$  is not a model parameter in PLS procedure.

- Given  $(w, \phi, \theta)$ , the estimator of  $\beta$  is obtained by maximizing the profile h-loglikelihood

$$h_p = (h_1 + h_2)|_{u=\hat{u}},$$

where  $\hat{u}$  solves  $dh/du = 0$ .

- Since  $h_1$  is the classical loglikelihood, the procedure corresponds to a penalized loglikelihood with implied penalty

$$p_\lambda(\beta) = -\phi h_2|_{u=\hat{u}},$$

where  $\hat{u}_j$  is computed in the first step of the IWLS.

- Specifically, for fixed  $w$ , taking only terms that involve  $\beta_j$  and  $\hat{u}_j$ , the  $j$ -th term of the penalty function is

$$p_\lambda(\beta_j) = \frac{\phi}{2\theta} \frac{\beta_j^2}{\hat{u}_j} + \frac{\phi(w-2)}{2w} \log \hat{u}_j + \frac{\phi}{w} \hat{u}_j. \quad (5)$$

- Thus the random effect model leads to a family of potentially unbounded penalty functions  $p_\lambda(\beta)$  indexed by  $w$ :
  - (1)  $w \rightarrow 0$ : ridge penalty ( $\because \hat{u}_j \rightarrow 1$  if  $w \rightarrow 0$ )
  - (2)  $w = 2$ : LASSO penalty ( $\because \hat{u}_j = |\beta_j| / \sqrt{\theta}$ )
  - (3)  $w > 2$ : penalty with infinite value and derivative at 0
- As the concavity near the origin increases, the sparsity of local solutions increases, and as the slope becomes flat, the amount of shrinkage lessens.
- From the next figure, we can see that HL controls the sparsity and shrinkage amount by choosing the values of  $w$  and  $\lambda$  simultaneously.

# Implied penalty functions

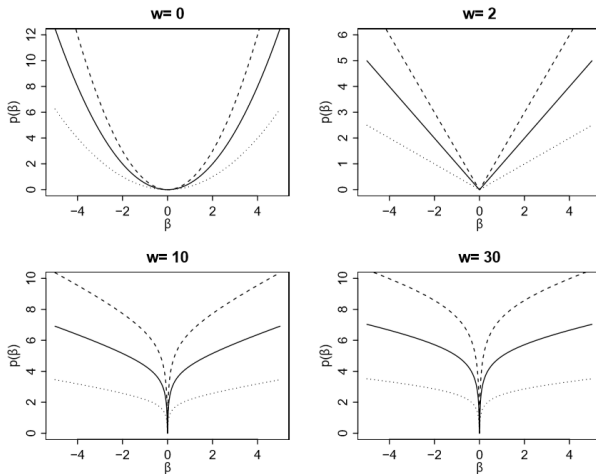


Figure 11.1 Penalty function  $p_\lambda(\beta)$  at different values of  $w$ , for  $\lambda = 1$  (solid),  $\lambda = 1.5$  (dashed) and  $\lambda = 0.5$  (dotted). In general, larger values of  $\lambda$  are associated with larger penalties, hence more shrinkage and more sparseness.

- By controlling the amount of sparsity and shrinkage simultaneously, the HL has much higher chances of selecting the correct models without losing prediction accuracy than the other methods (Kwon et al., 2017).
- Ng et al. (2006) showed the consistency of all local solutions of the HL method, which implies the uniqueness of HL solution under certain conditions
- Ng et al. (2017) showed that HL estimator achieves consistent estimation of number of change points, their locations, and their sizes, while LASSO and SCAD may not.
- Advantage of the HL method is to achieve asymptotic selection consistency without losing prediction accuracy in finite sample.

- Consider the simplest case that  $\beta$  is the population mean and  $z$  is the sample mean. Here we can illustrate various variable selection procedures.
- The IWLS step (4) gives

$$\hat{\beta} = \frac{z}{1 + \lambda/\hat{u}}, \quad (6)$$

and the corresponding PLS criterion is

$$Q_{\lambda}(\beta) = \frac{1}{2}(z - \beta)^2 + p_{\lambda}(\beta). \quad (7)$$

- The next figure shows the penalized likelihood surfaces at different values of  $z$ . Given  $\lambda$  as  $z$  approaches zero (when  $z \leq 2$ ), there is only one maximum at zero, so in this case the estimate is zero and the corresponding predictor is not selected. Otherwise, bimodality occurs.

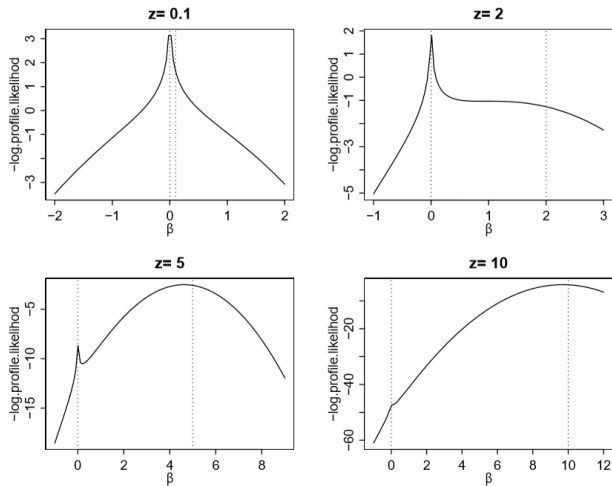


Figure 11.2 Implied penalized log-likelihood functions equal to  $-Q_{\lambda}(\beta)$  in (11.9) at different values of  $z$  and fixed  $\lambda = 1$ .

- Note that the implied penalized likelihood  $Q_\lambda(\beta)$  is not convex but the model can be expressed hierarchically as (a)  $y_i|\beta$  is normal and (b)  $\beta_j|u_j$  is normal with (c) gamma  $u_j$ ; all three models are convex.
- Thus the IWLS algorithm overcomes the difficulties of a non-convex optimization by solving three interlinked convex optimizations.
- Equalizing the score equations for  $\beta$  from (5) and from the PLS (6), we have

$$\beta(1 + \lambda/\hat{u}) - z = \partial Q_\lambda / \partial \beta = -(z - \beta) + p'_\lambda(\beta),$$

and get a useful general formula

$$\hat{u}(\beta) = \lambda\beta / p'_\lambda(\beta), \tag{8}$$

which allows us to obtain results for LASSO, SCAD or the so called adaptive LASSO by using different random effect estimates  $\hat{u}$  in the IWLS of (5).



- Examples of the penalty derivatives for some methods are given in the next table.

Types	$p'_\lambda(\beta)$
LASSO	$\lambda \operatorname{sign}(\beta)$
SCAD	$\lambda \operatorname{sign}(\beta) \left\{ I( \beta  < \lambda) + \frac{(a\lambda -  \beta )_+}{(a-1)\lambda} I( \beta  > \lambda) \right\}$
HL	$\lambda\beta / \{w\{(2/w - 1) + \kappa_j\}/4\}$ where $\kappa_j = \{8\beta^2/(w\theta) + (2/w - 1)^2\}^{1/2}$

Table. Derivative of penalty functions for some methods

- For the LASSO,  $p_\lambda(\beta) = \lambda|\beta|$ , so  $\hat{u} = |\beta|$ .
- For the adaptive LASSO,  $p_\lambda(\beta) = 2\lambda|\beta|/|z|$ , so  $\hat{u} = |\beta||z|/2$ .
- For the SCAD,  $\hat{u} = |\beta|/\{I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda} I(|\beta| > \lambda)\}$  for some  $a > 2$ .

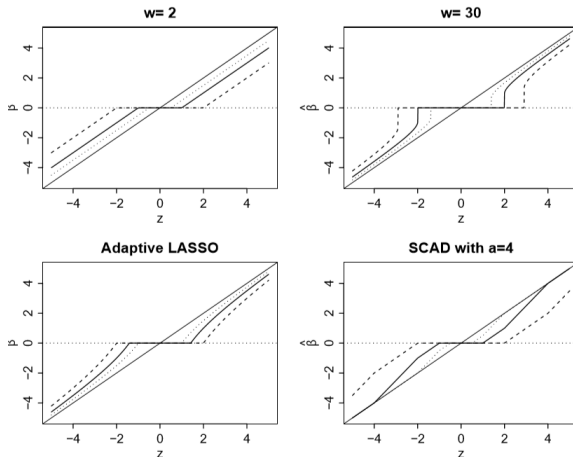


Figure 11.3 Different IWLS solutions (11.8) as a function of  $z$  at fixed  $\lambda = 1$  (solid),  $\lambda = 2$  (dashed) and  $\lambda = 0.5$  (dotted). The formula for  $\hat{u}$  is given by (11.5) for  $w = 2$  and 30, and by (11.11) and (11.12) for the adaptive LASSO and SCAD estimates, respectively.

- In regression problems, explanatory variables often possess a natural group structure.
  - categorical factors are often represented by a group of indicator variables
  - to capture flexible functional shapes, continuous factors can be represented by a linear combination of basis functions such as splines or polynomials.
- In these situations, the problem of selecting relevant variables involves selecting groups rather than selecting individuals.
- Depending on the situation, the individual variables in a group may or may not be meaningful scientifically
  - If they are not, we are typically not interested in selecting individual variables and the interest is limited to group selection.
  - However, if the individual variables are meaningful, then we would be interested in selecting individual variables within each selected group; we refer to this as bi-level selection. (Huang et al., 2012)

- Suppose that the explanatory variables can be divided into  $K$  groups and the outcome  $y = (y_1, \dots, y_n)^\top$  has mean  $\mu = (\mu_1, \dots, \mu_n)^\top$  that follows a GLM with link function  $\eta_i \equiv h(\mu_i)$ , such that we have a linear predictor  $\eta = (\eta_1, \dots, \eta_n)^\top$ ,

$$\eta = X\beta \equiv X_1\beta_1 + \dots + X_K\beta_K \quad (9)$$

where  $X \equiv (X_1, \dots, X_K)$  and  $\beta = (\beta_1, \dots, \beta_K)^\top$  are collection of  $n \times p_k$  design matrices and  $p_k$  regression coefficients, respectively.

- For group selection, Lee et al. (2015) considered a random effect model

$$\beta_{kj} | u_k \sim N(0, u_k \theta), \quad k = 1, \dots, K \text{ and } j = 1, \dots, p_k \quad (10)$$

$$u_k \sim \text{gamma}(w_k), \quad k = 1, \dots, K \quad (11)$$

where  $\theta$  and  $w_k$  are regularization parameters that control the degree of shrinkage and sparseness of the estimates.

- For a given  $\theta$ , the sparsity among the groups increases as  $w_k$ 's get larger, while for fixed  $w_k$ 's the shrinkage becomes smaller as  $\theta$  increases.
- Group selection is achieved as follows.
  - If  $\hat{u}_k = 0$ , then  $\hat{\beta}_{kj} = 0$  for all  $j$ .
  - If  $\hat{u}_k > 0$ , then  $\hat{\beta}_{kj} \neq 0$  for all  $j$ .
- This means that the model is limited to group-only selection, as it does not impose sparsity within the selected groups.

- Bi-level selection can be done by extending the model (10) as follows:

$$\beta_{kj} | u_k, v_{kj} \sim N(0, u_k v_{kj} \theta), \quad k = 1, \dots, K \text{ and } j = 1, \dots, p_k \quad (12)$$

$$u_k \sim \text{gamma}(w_k) \quad (13)$$

$$v_{kj} \sim \text{gamma}(\tau). \quad (14)$$

where  $u_k$  is the random effect corresponding to the  $k$ -th group and  $v_{kj}$  is the random effect corresponding to the  $j$ -th variable in the  $k$ -th group.

- Hence this model selects variables at both the group level and the individual variable level within selected groups.
  - If  $\hat{u}_k = 0$ , then  $\hat{\beta}_{kj} = 0$  for all  $j = 1, \dots, p_k$ .
  - If  $\hat{u}_k > 0$ , then  $\hat{\beta}_{kj} = 0$  when  $\hat{v}_{kj} = 0$ .

- Interaction terms in regression models form a natural hierarchy with the main effects, so their selection requires special consideration.
- It is common practice that the presence of an interaction term requires both of the corresponding main effects in the model. This may be called a strong hierarchy constraint, while the weak version requires only one of the main effects to be present.
- We can use a random effect model to impose sparse selection of interaction terms under the hierarchy constraints.

- Consider a  $p$ -predictor GLM with both main and interaction terms.

$$\eta_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \sum_{j < k} x_{ij}x_{ik}\delta_{jk}, \quad i = 1, \dots, n,$$

which we write in matrix form as

$$\eta = X\beta + Z\delta,$$

where  $\eta = (\eta_1, \dots, \eta_n)$  is the vector of linear predictors,  $\beta = (\beta_1, \dots, \beta_p)$  and  $\delta = (\delta_{12}, \dots, \delta_{p-1,p})$  are the vectors of the corresponding regression coefficients for main and interaction terms, respectively. Similarly,  $X$  is the design matrix of the intercept and linear terms for the main effects, and  $Z$  is that of the cross product terms for the interactions.



- Lee et al. (2015) proposed the use of random effect model.
- Under the strong hierarchy constraint,

$$\begin{aligned}\beta_j | u_j &\sim N(0, u_j \theta), \\ \delta_{kj} | u_k, u_j, v_{kj} &\sim N(0, u_k u_j v_{kj} \theta) \text{ for } k > j \\ u_j &\sim \text{gamma}(w_1) \text{ and } v_{kj} \sim \text{gamma}(w_2).\end{aligned}$$

- Under the weak hierarchy constraint,

$$\begin{aligned}\beta_j | u_j &\sim N(0, u_j \theta), \\ \delta_{kj} | u_k, u_j, v_{kj} &\sim N(0, (u_k + u_j) v_{kj} \theta) \\ u_j &\sim \text{gamma}(w_1) \text{ and } v_{kj} \sim \text{gamma}(w_2).\end{aligned}$$

## Interaction and hierarchy constraints

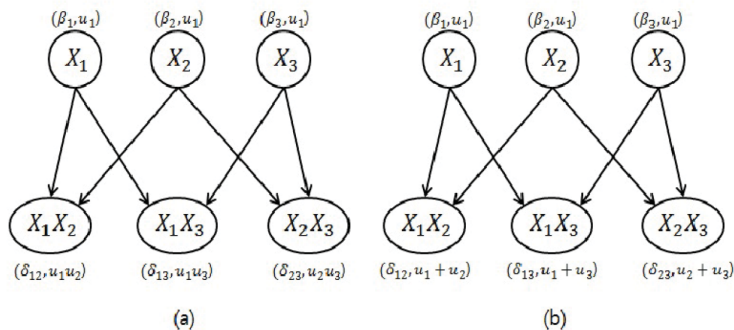


Figure 10.2 A model with main effects and interaction terms under (a) strong hierarchy and (b) weak hierarchy constraints.

- For completeness, we describe here other statistical models in which the notion of hierarchy applies, and show how to model them using the random effects approach.
- Suppose we want to fit the second-order mixed polynomial model

$$\eta = X_1\beta_1 + \cdots + X_p\beta_p + X_1^2\delta_{11} + X_1X_2\delta_{12} \cdots + X_p^2\delta_{pp}, \quad (15)$$

where  $X_kX_j$  denotes the component-wise product between the two column vectors.

- To maintain the functional marginality rule, we consider a random effect model

$$\begin{aligned}\beta_j|u_j &\sim N(0, u_j\theta), \\ \delta_{jj}|u_j, v_{jj} &\sim N(0, u_jv_{jj}\theta), \\ \delta_{kj}|u_k, u_j, v_{kj} &\sim N(0, u_ku_jv_{kj}\theta), \\ u_j &\sim \text{gamma}(w_1) \text{ and } v_{kj} \sim \text{gamma}(w_2).\end{aligned}$$

- This is analogous to the strong hierarchy in previous model, but now we include  $\delta_{jj}$ . It can be easily extended to general higher-order models.

## Functional marginality and general graph structure

- Various hierarchical structures can be represented by a directed graph.

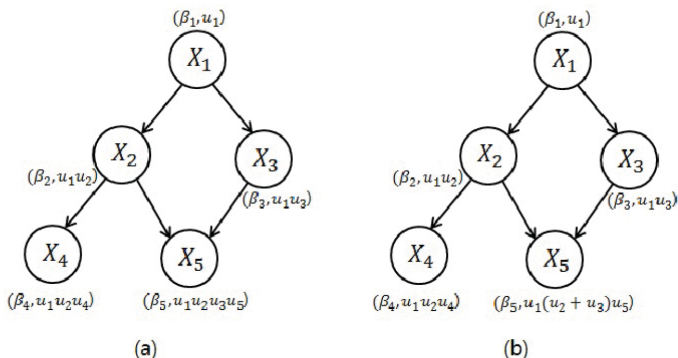


Figure 10.3 The directed graph structure representing hierarchy of variables under (a) strong hierarchy and (b) weak hierarchy constraints.

- In Figure 10.3 (a) for strong hierarchy,  $X_5$  can be included if  $(X_1, X_2, X_3)$  are included in the model. This graph can be modeled by the following random effect model:

$$\beta_1 | u_1 \sim N(0, u_1 \theta),$$

$$\beta_2 | u_1, u_2 \sim N(0, u_1 u_2 \theta),$$

$$\beta_3 | u_1, u_3 \sim N(0, u_1 u_3 \theta),$$

$$\beta_4 | u_1, u_2, u_4 \sim N(0, u_1 u_2 u_4 \theta),$$

$$\beta_5 | u_1, u_2, u_3, u_5 \sim N(0, u_1 u_2 u_3 u_5 \theta),$$

$$u_j \sim \text{gamma}(w) \text{ for } j = 1, \dots, 5.$$

- For weak hierarchy,  $X_5$  can be included if the model includes, besides  $X_1$ , at least one of  $X_2$  and  $X_3$ . This graph can be modeled by

$$\beta_5 | u_1, u_2, u_3, u_5 \sim N(0, u_1 (u_2 + u_3) u_5 \theta).$$

- This illustrates how the random effect model can be adapted to describe various hierarchical structures in the covariates.
- The HL method can easily applied to produce sparse versions of classical multivariate techniques, such as the principle component analysis, canonical covariance analysis, partial-least squares for Gaussian and that for survival outcomes (Lee et al., 2010, 2011a,b, 2013).
- Furthermore, it is straight forwards to apply HL method to various class of HGLM models via penalized h-loglikelihood; general frailty models (Ha et al., 2014a) and competing risks models (Ha et al., 2014b)

- Disease progression of diabetes in Efron(2004)
  - 442 diabetes patients
  - 10 predictive variables: age, sex, bmi, bp and 6 types of serum measurements
  - Response variable: a measure of disease progression
- Consider a quadratic model having  $p = 64$  predictive variables
  - 10 original terms
  - 9 quadratic terms (except for binary variable)
  - ${}_{10}C_9 = 45$  interaction terms
- We compare three methods: LASSO, SCAD and HL ( $w = 30$ ).

Method	LASSO	SCAD	HL
Number of variables	15	12	14
CV error	2988.69	2982.85	2891.76

	LASSO	SCAD	HL
sex	-5.43	-11.07	-10.86
bmi	23.89	25.14	23.63
map	12.04	15.16	15.17
hdl	-9.00	-12.98	-12.52
ltg	22.28	23.49	22.89
glu	0.89		2.93
age			
age <sup>2</sup>	0.35	0.95	2.76
bmi <sup>2</sup>	1.29	0.06	2.13
glu <sup>2</sup>	2.25	2.31	3.51
age:sex	5.26	7.33	7.46
age:map	1.53	0.68	1.69
age:ltg	0.43	0.01	1.55
age:glu	0.58		
sex:map	0.03		2.29
bmi:map	3.87	5.23	5.13



- The numbers of variables selected by the three methods are similar, varying from 10 to 15, though the HL method has the smallest cross-validated error (Kwon et al., 2016).
- If we look at estimates of main effects, the LASSO estimators are shrunk the most and the SCAD estimators the least.
- We see that all methods include the age:sex interaction in their final model, consistent with the known result that diabetes progression behaves differently in women after menopause
- As seen in Table 10.2, with an automatic variable selection method with large  $p$ , the marginality rule will be easily violated. A systematic way of handling such a problem is grouped model selection as we shall show.

## ex. Gene-gene interaction

- As an illustration, we analyse gene-gene interaction in a cohort study called ULSAM (Uppsala Longitudinal Study of Adult Men).
- Ongoing population-based study of all available men born between 1920 to 1924 in Uppsala County, Sweden.
- Analyse a subset of  $n = 1179$  subjects for which we have genetic data.
- The primary outcome is body-mass index (BMI), a major risk factor for many cardiovascular diseases.
- Based on several criteria, we selected 10 single-nucleotide polymorphisms (SNPs) as the predictive variables. (Lee et al., 2015)

## ex. Gene-gene interaction

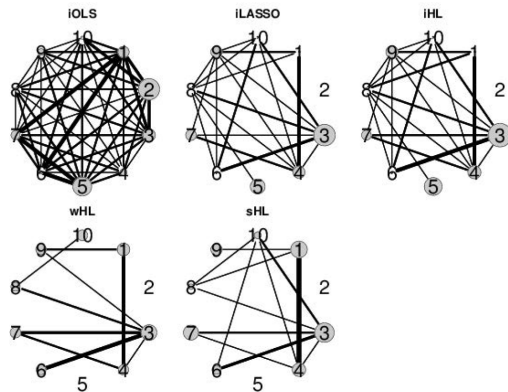


Figure 10.4 Results from various methods applied to ULSAM data. In each graph, each node represents a SNP, the size of the main effect is represented by the circle size and the interaction by the thickness of the line between two nodes.

## ex. Gene-gene interaction

- The ordinary least squares (iOLS) method estimates all the interaction terms and cannot recognize the linkage disequilibrium between SNPs 1, 2 and 5.
- The largest interactions in iOLS are (1,6), (1,7), (5,7). In contrast, all the sparse methods select (1,4) and (3,6) as the most interesting pairs.
- As expected, unconstrained methods (iLASSO and iHL) select interaction term without main effects, which can lead to misleading conclusions.
- The hierarchy constrained method (wHL and sHL) have comparable sparsity and they both select only one of the linked SNPs 1, 2 and 5. If strong hierarchy is desired, sHL method provides a sensibly sparse solution in this case.

# Chapter 11. Further Topics : Multiple Testing

## Single hypothesis testing

- Single hypothesis testing problem on the mean of  $Y \sim N(\mu, 1)$ .

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1$$

- The classical Neyman-Pearson likelihood ratio is

$$L = \frac{f(y|H_1)}{f(y|H_0)}$$

- Let a discrete random effect be  $o = 0$  if  $H_0$  is true and  $o = 1$  if  $H_1$  is true. Then the h-likelihood is

$$f(y, o) = f(y|o)P(o)$$

- Hence, the h-likelihood ratio is

$$R = \frac{f(y, o = 1)}{f(y, o = 0)} = \frac{f(y|H_1)P(o = 1)}{f(y|H_0)P(o = 0)} = \frac{1 - p_0}{p_0} L$$

- For the test depending on the value of  $L$ ,  $p_0$  should be strictly between 0 and 1. However, in single hypothesis testing,  $p_0$  is not estimable.
- Both  $L$  and  $R$  give equivalent optimal tests, but the h-likelihood ratio  $R$  opens up a way for testing multiple hypotheses.
- The h-likelihood ratio can also be interpreted as a ratio of predictive probabilities.

$$R = \frac{f(y, o = 1)}{f(y, o = 0)} = \frac{P(o = 1|y)f(y)}{P(o = 0|y)f(y)} = \frac{P(o = 1|y)}{P(o = 0|y)}$$

- We can show that the optimal test is determined by the ratio of predictive probabilities  $R$ , equivalent to the h-likelihood ratio.

- With the loss function that depends on  $\lambda$ ,

$$o(1 - \delta) + \lambda(1 - o)\delta$$

we have the risk

$$E(o(1 - \delta) + \lambda(1 - o)\delta|y) = P(o = 1|y) + P(o = 0|y)(\lambda - R)\delta$$

- The optimal test  $\delta^\lambda$  is determined by the h-likelihood ratio,

$$\delta^\lambda = I(R > \lambda)$$

- In the single hypothesis testing,  $p_0$  may not be estimable, ( $R = \frac{1-p_0}{p_0} L$  may not be calculated), so that need to define the optimal test without  $p_0$ .
- Define the optimal test as  $\delta^{\lambda^*} = I(L > \lambda^*) (= I(R > \lambda))$  where  $\lambda^* = \frac{\lambda p_0}{1-p_0}$ , for some  $0 < p_0 < 1$ . And choose  $\lambda^*$  to satisfy  $P(\delta^{\lambda^*} = 1|H_0) \leq \alpha$ .

- Most literature on multiple testing has focused on the error control, not the power of the test.
- The h-likelihood gives the optimal test maximizing the power of the test.
- Suppose that we have  $N$  null hypotheses  $H_1, \dots, H_N$  to test simultaneously.

	$\delta = 0$	$\delta = 1$	Total
$o = 0$	$V_{00}$	$V_{01}$ (Type 1 error)	$N_0$
$o = 1$	$V_{10}$ (Type 2 error)	$V_{11}$	$N_1$
Total	$M_0$	$M_1$	$N$

- There are methods choosing the threshold of test.
  - control the family wise error rate(FWER)
  - control false discovery rate(FDR)



- **Family wise error rate**

The probability of at least one false positive.

$$FWER = P(V_{01} \geq 1)$$

- **False discovery rate**

The expected proportion of errors among rejected hypotheses.

$$FDR = E\left(\frac{V_{01}}{M_1}\right)$$

Following Efron (2004), we use the marginal FDR

$$mFDR = \frac{E(V_{01})}{E(M_1)}$$

- Similar to single case, with the loss

$$\sum o_i(1 - \delta_i) + \lambda(1 - o_i)\delta_i$$

the optimal rule  $\delta^\lambda = \{\delta_1^\lambda, \dots, \delta_N^\lambda\}$  becomes

$$\delta_i^\lambda = I(R_i > \lambda)$$

- In multiple testing case,  $p_o = \frac{E(N_0)}{N}$  is estimable, so that  $R_i$  can be directly used.
- With the optimal rule  $\delta^\lambda$ , the marginal false discovery rate is given by

$$mFDR(\lambda) = \frac{E(V_{01})}{E(M_1)} = \frac{\sum P(o_i = 0, \delta_i^\lambda = 1)}{E(\sum \delta_i^\lambda)}$$

- And the estimated mFDR is given by

$$\widehat{mFDR}(\lambda) = \frac{\hat{p}_0 \sum P(R_i > \lambda | H_{0i})}{\sum I(R_i > \lambda)}$$

- $mFDR(\lambda)$  can be controlled by  $\widehat{mFDR}(\lambda)$  at a specific level by varying  $\lambda$ .
- Parameters can be estimated by maximizing marginal likelihood. And if the MLE for  $\theta$  is consistent, the likelihood ratio test is asymptotically optimal.
- **Random effect model for multiple testing** Suppose that  $y_{ij1}$  for the  $i$ th site of the  $j$ th individual in the control group and  $y_{ij2}$  in the treatment group can be modeled for  $i = 1, 2, \dots, N$  as

$$\begin{aligned}y_{ij1} &= \xi_i + \epsilon_{ij1} \\ y_{ij2} &= \xi_i + w_i + \epsilon_{ij2}\end{aligned}$$

where  $\xi_i$  is the site effect,  $w_i$  is the random treatment effect and  $\epsilon_{ijm}$  is the error with  $E(\epsilon_{ijm}) = 0$  and  $Var(\epsilon_{ijm}) = \phi_{im}$ .

Assume that the random treatment effect  $w_i$ s are independent with

$$E(w_i|H_{0i}) = 0 \text{ and } \text{Var}(w_i|H_{0i}) = \sigma^2$$

$$E(w_i|H_{1i}) = \mu \neq 0 \text{ and } \text{Var}(w_i|H_{1i}) = \tau^2$$

- Then, for the difference in means  $d_i = \bar{y}_{i2} - \bar{y}_{i1}$ , we have the following hierarchical model:

$$\begin{aligned} \text{Conditional on } w_i \text{ and } o_i, E(d_i|w_i, o_i) &= w_i \\ &\text{and } \text{Var}(d_i|w_i, o_i) = \psi_i \end{aligned}$$

$$\text{Conditional on } o_i = 0, E(w_i|H_{0i}) = 0 \text{ and } \text{Var}(w_i|H_{0i}) = \sigma^2$$

$$\text{Conditional on } o_i = 1, E(w_i|H_{1i}) = \mu \text{ and } \text{Var}(w_i|H_{1i}) = \tau^2$$

where  $\psi_i = \phi_{i1}/n_1 + \phi_{i2}/n_2$ .

- Let  $v = (w, o)$  be unobservables, and  $y$  be the set of all observations. The h-likelihood is defined to be

$$L(v, \theta; y, v) = f_{\theta}(y, v) = f_{\theta}(y)P_{\theta}(v|y)$$

- Suppose that we are not interested in effect size  $w_i$ , we can integrate them out. It leads the model for  $d = (d_1, \dots, d_N)$

Given  $o_i = 0$ ,  $E(d_i|H_{0i}) = 0$  and  $Var(d_i|H_{0i}) = \psi_i + \sigma^2$

Given  $o_i = 1$ ,  $E(d_i|H_{1i}) = \mu$  and  $Var(d_i|H_{1i}) = \psi_i + \tau^2$

- Then the h-likelihood is given by

$$L(o, \theta; d, o) = f_{\theta}(d, o) = \prod_{i=1}^N L(o_i)$$

where

$$L(o_i = 1) = P(o_i = 1)f_{\theta}(d_i|o_i = 1) = (1 - p_0)f_{\theta}(d_i|H_{1i})$$

$$L(o_i = 0) = P(o_i = 0)f_{\theta}(d_i|o_i = 0) = p_0f_{\theta}(d_i|H_{0i})$$

## ex. Neuroimaging data

- PET data from the study of the Korean standard template.
- The data consists of scans of 28 healthy males and 22 healthy females.
- Each image has  $N = 189,201$  voxels.
- Previous methods have not identified any voxel in the brain to be significant and Lee and Lee (2017) identified some significant voxels. So the method based on h-likelihood ratio test is the most powerful one.

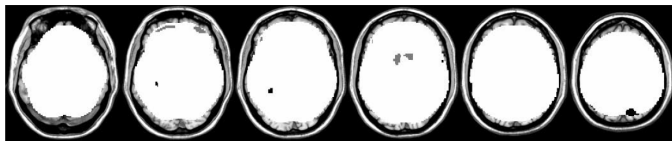


Figure 10.5 *Multiple testing for the neuroimage data by using likelihood-ratio testing. The gray-colored (black-colored) region are negatively (positively) activated.*