CLUSTERED DATA MODELS

Arto Luoma
Tampere University, Finland

Autumn 2024

Contents of the course

Examples of clustered data

Likelihood inference

Quasi-likelihood methods

Linear mixed effect models

Mixed effect model for nonnormal responses

Generalized estimation equations

Missing data

Examples of clustered data

Some concepts

In *repeated measure designs*, measurements are taken repeatedly on several individuals or units. When these repeated measurements are taken over time, we call it a *a longitudinal study* or sometimes *a panel study*.

Measurements taken on one individual compose a *cluster*. Observations within a cluster are often marginally correlated while those from different clusters are uncorrelated. We call the correlation that appears between the observations of each cluster *intraclass or intracluster correlation*. This correlation is due to differences between individuals.

These individual differences are often modeled using *random effects*. Models that include both random and fixed effects are called *mixed-effect models*. Depending on the response variable, we may use a linear mixed-effect model (LME), or a *generalized linear mixed-effect model (GLMM)*.

Example 1: Longitudinal stress in railway rails.

The data shown in Figure 1 are from an experiment testing for longitudinal stress in railway rails (Devore 2000, Example 10.10, p. 427). Six rails were chosen at random and tested three times each by measuring the time it took for a certain type of ultrasonic wave to travel the length of the rail.

We may be interested in the average time for a typical rail (*expected travel time*), the variation in average travel times among rails (*between-rail variability*), and the variation in travel times for a single rail (*within-rail variability*).

Example 1 (cont)

```
> library(nlme)
> plot(Rail)
> Rail
Grouped Data: travel ~ 1 | Rail
     Rail travel
1      1     55
2      1     53
3      1     54
...
18     6     83
```
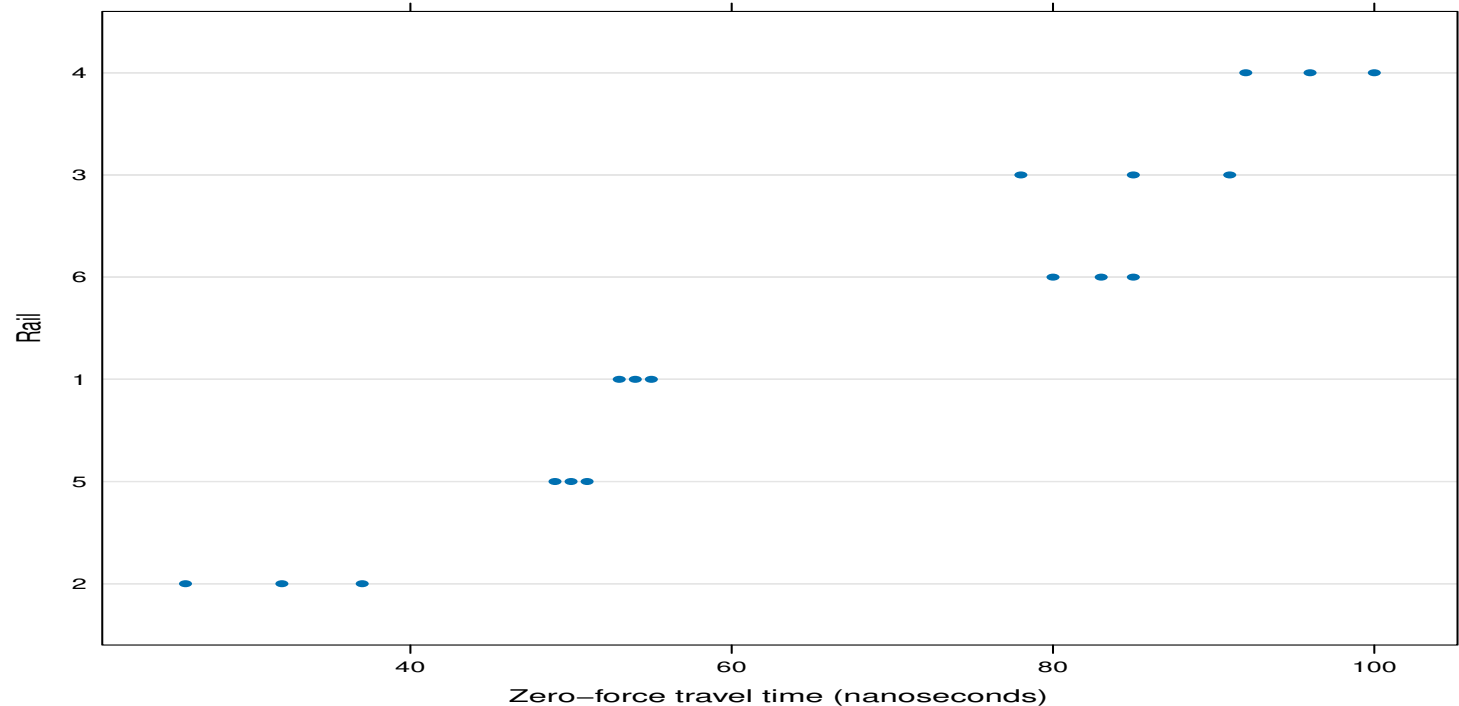
Example 1 (cont)

Figure 1: Travel time for ultrasonic head-waves in the rail (nanoseconds). The value given is the original travel time minus 36,100 nanoseconds

Example 2: Linear growth curve

Growth curves are one type of longitudinal data. A classic example (Potthoff and Roy, 1964) is a set of measurements of the distance from the pituitary gland to the pterygomaxillary fissure taken every two years from 8 years of age until 14 years of age on a sample of 27 children (16 boys and 11 girls).

Figure 2 shows that for most of the subjects, the orthodontic measurement increases with age and the growth is approximately linear. The slopes and intercepts may differ between the children. Further, there may be differences between boys and girls.

Example 2 (cont)

```
> plot(Orthodont)
> Orthodont
Grouped Data: distance ~ age | Subject
    distance age Subject     Sex
1       26.0   8      M01    Male
2       25.0  10      M01    Male
3       29.0  12      M01    Male
4       31.0  14      M01    Male
...
105     24.5   8      F11  Female
106     25.0  10      F11  Female
107     28.0  12      F11  Female
108     28.0  14      F11  Female
```
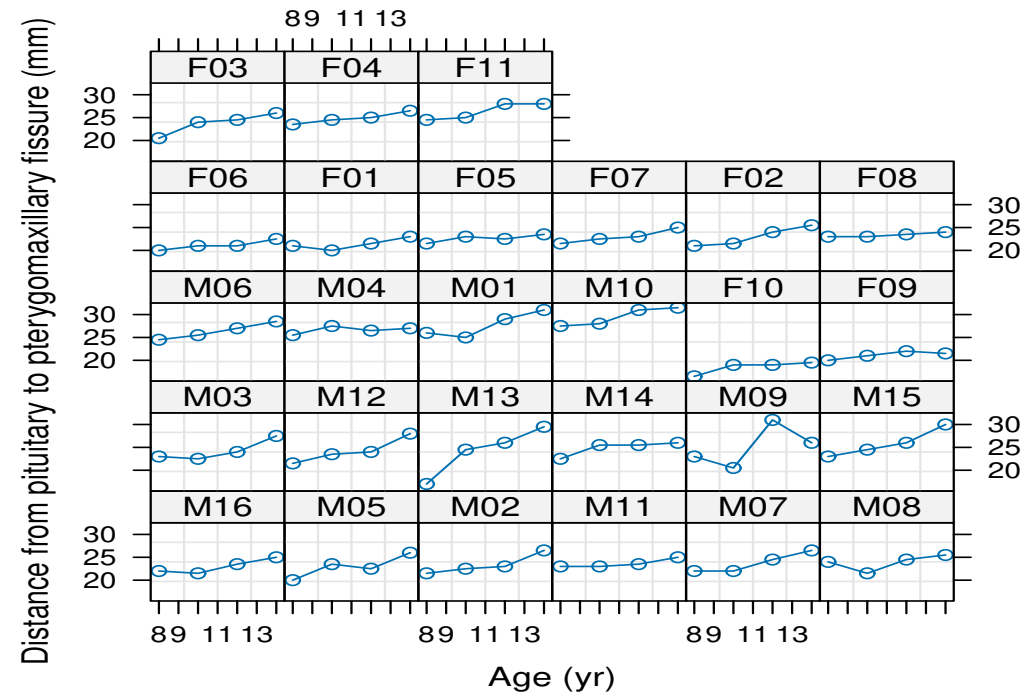
Example 2 (cont)

Figure 2: The distances from the pituitary to the pterygomaxillary fissure (mm) for a sample of 16 boys (subjects M01 to M16) and 11 girls (subjects F01 to F11).

Likelihood inference

Likelihood inference

Most statistical inference is probably performed using so-called likelihood inference. In this approach, all information from the data comes through the likelihood function. Denoting the data as $\mathbf{y} = (y_1, ..., y_n)$ and the (vector of) unknown parameter(s) as $\boldsymbol{\theta}$, the likelihood function is given by the joint density of the observations

$$p(\mathbf{y}; \boldsymbol{\theta}) = p(y_1, ..., y_n; \boldsymbol{\theta})$$

when considered as a function of $\boldsymbol{\theta}$.

The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is obtained by maximising the log-likelihood:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{y}; \boldsymbol{\theta}).$$

The statistical inference on $\boldsymbol{\theta}$ is based on the asymptotic (large sample) normality of the MLE. Under certain regularity conditions, the distribution of $\hat{\boldsymbol{\theta}}$ is approximately multivariate normal with covariance matrix $\mathbf{I}(\boldsymbol{\theta})^{-1}$:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \mathbf{I}(\boldsymbol{\theta})^{-1}),$$

where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher's information matrix, defined as

$$\mathbf{I}(\boldsymbol{\theta}) = -\mathsf{E}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log p(\mathbf{y}; \boldsymbol{\theta}).$$

This result can be used to compute approximate standard errors and confidence intervals for the components of $\boldsymbol{\theta}$.

Likelihood inference (cont)

The most relevant regularity conditions are that a) the log-likelihood function is sufficiently smooth b) the true value of $\boldsymbol{\theta}$ is interior to the parameter space $\boldsymbol{\Theta}$ and c) the sample space of $\mathbf{y}$ is not restricted by $\boldsymbol{\theta}$.

In practice, we often use the observed information matrix

$$\mathbf{J}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log p(\mathbf{y};\boldsymbol{\theta})$$

instead of $\mathbf{I}(\boldsymbol{\theta})$. Further, $\mathbf{J}(\boldsymbol{\theta})$ is usually evaluated at $\hat{\boldsymbol{\theta}}$ and the Hessian matrix (the matrix of the second derivatives) is determined numerically.

Likelihood inference (cont)

Two models M_1 and M_2 can be compared using the likelihood ratio test if one of the models (say M_1) is nested in the other (M_2), that is, is its special case. Denote the likelihood of M_k as $p_k(\mathbf{y}; \boldsymbol{\theta}_k)$, its MLE as $\hat{\boldsymbol{\theta}}_k$, and the number of estimated parameters as d_k, for $k = 1, 2$. Then the statistic

$$\chi^2 = -2 \log \frac{p_1(\mathbf{y}; \hat{\boldsymbol{\theta}}_1)}{p_2(\mathbf{y}; \hat{\boldsymbol{\theta}}_2)},$$

will approximately be χ^2-distributed with $d = d_2 - d_1$ degrees of freedom if M_1 is true.

The test statistic can also written as $\chi^2 = \hat{D}_1 - \hat{D}_2$, where $\hat{D}_k = -2 \log p_k(\mathbf{y}; \hat{\boldsymbol{\theta}}_k)$, $k = 1, 2$, is the estimated (unstandardized) deviance of M_k.

About 'deviance'

There is a considerable confusion related to the concept of 'deviance' in the statistical literature. Usually, the deviance of a model M_k is defined as

$$-2\left[\log p_k(\mathbf{y}; \hat{\boldsymbol{\theta}}_k) - \log p_S(\mathbf{y}; \hat{\boldsymbol{\theta}}_S)\right]$$

where $p_S(\mathbf{y}; \hat{\boldsymbol{\theta}}_S)$ is the maximized likelihood of the saturated model. The saturated model is the maximally fitting model, having as many parameters as there are observations.

Sometimes this is also called 'standardized deviance', 'residual deviance' (in R) or 'scaled deviance'. In the last-named case, the 'deviance' is obtained from the 'scaled deviance' by multiplying it by a variance parameter.

About 'deviance' (cont)

When comparing two models (M_1 and M_2) it is the same if one uses the standardized or unstandardized deviance, since the term $\log p_S(\mathbf{y}; \hat{\boldsymbol{\theta}}_S)$ cancels out.

The standardized deviance as such can also be used as a measure of model fit. If the model M_k is well-fitting (or true) then the standardized deviance is approximately χ^2-distributed with $d_S - d_k$ degrees of freedom where d_S is the number of parameters in the saturated model. However, this method is not valid for binary data ($y_i = 0$ or 1) of for grouped binary (binomial) data ($Y_i \sim \text{Bin}(n_i, p_i)$) when n_i is small (<5) for some i.

'Null deviance' (in R) is the deviance of the model including only constant terms, no real predictors.

Information criteria

Another method for model comparison is the use of information criteria. The two most commonly used criteria are Akaike's information criterion

$$\text{AIC} = -2\log p(\mathbf{y}; \hat{\boldsymbol{\theta}}) + 2d,$$

and the Bayesian (or Schwarz-Rissanen) information criterion

$$\text{BIC} = -2\log p(\mathbf{y}; \hat{\boldsymbol{\theta}}) + d\ln(n),$$

where d is the number of parameters in the model and n the number of observations.

In general, these criteria consist of the unstandardized deviance and a term penalizing for model complexity. The model obtaining the smallest information criterion value can be considered the best among the models compared.

Software for likelihood estimation

In this course we will employ R software which is freely available for various operating systems. In addition to the basic distribution we will need to install contributed packages.

✔   Package faraway contains the data sets and functions used in Faraway (2006).
✔   Packages nlme and lme4 contain useful functions for the analysis linear and nonlinear mixed effect models
✔   Package gee is useful for the generalized estimation method
✔   Package glmmML is useful for the ML (maximum likelihood) estimation of generalized linear mixed effect models

Example 1: Male satellites for female horseshoe crabs

These data are from a study of female horseshoe crabs (Example 1.5 in Agresti; complete data (n=173) are available at https://users.stat.ufl.edu/~aa/glm/data/Crabs.dat).

During spawning season, a female migrates to the shore to breed. With a male attached to her posterior spine, she burrows into the sand and lays clusters of eggs. The eggs are fertilized externally, in the sand beneath the pair. During spawning, other male crabs may cluster around the pair and may also fertilize the eggs. These male crabs are called *satellites*.

The response variable is y, number of satellites. Explanatory variables: color (1, medium light; 2, medium, 3, medium dark; 4, dark); spine condition (1, both good; 2, one worn or broken; 3, both worn or broken); carapace width (cm); weight (kg.)

Example 1 (cont)

```
># Fit a Poisson regression model and plot the fitted curve
> plot(y ~ weight, data = Crabs, cex=0.5)
> model1 <- glm(y ~ weight, family = poisson, data = Crabs)
> wt <- seq(1, 5, by=0.1)
> fit1 <- predict(model1, newdata = data.frame(weight = wt),
    type = "response")
> lines(wt,fit1)
```

Example 1 (cont)
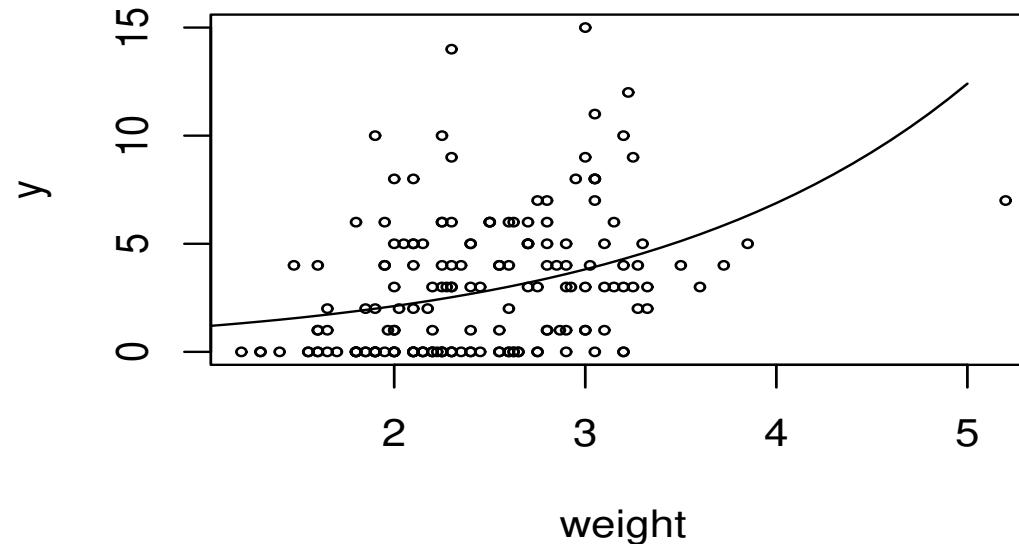
Figure 3: Scatterplot of number of crab satellites against crab weight and a fitted Poisson regression curve.

Example 1 (cont)

```
> # The fit is not very good because the residual deviance
> # is much larger than the degrees of freedom

> model1
Coefficients:
(Intercept)        weight
-0.4284        0.5893

Degrees of Freedom: 172 Total (i.e. Null);  171 Residual
Null Deviance:      632.8
Residual Deviance: 560.9  AIC: 920.2

> 1-pchisq(560, df = 171)
[1] 0
```

Example 1 (cont)

```
> # Fit another model where color is included as a categorical
> # predictor and compare with the first model
> model2 <- glm(y ~ weight + factor(color), family = poisson,
    data = Crabs)
> model2

Coefficients:
(Intercept)     weight  factor(color)2  factor(color)3  factor(color)
-0.04978        0.54618        -0.20511        -0.44980        -0.4520

Degrees of Freedom: 172 Total (i.e. Null);  168 Residual
Null Deviance:      632.8
Residual Deviance: 551.8  AIC: 917.1
```

Example 1 (cont)

```
> anova(model1, model2)
Analysis of Deviance Table

Model 1: y ~ weight
Model 2: y ~ weight + factor(color)
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171       560.87
2       168       551.80  3   9.0615  0.02848 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

According to the likelihood ratio test, this model fits better. Also, AIC is smaller (= better!). However, the residual deviance (551.8) is still too large compared to the degrees of freedom (168).

Generalized linear model

In generalized linear models (GLM) one assumes that the observations y_i come from a distribution that belongs to the exponential family of distributions, so that their density can be written in the form

$$f(y_i; \theta_i, \phi) = \exp\left\{ \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}.$$

It can be shown that $\mu_i = \mathsf{E}(y_i) = b'(\theta)$ and $\mathsf{Var}(y_i) = b''(\theta)a_i(\phi)$.

Further, one assumes that the mean μ_i is related to the linear part η_i of the model and the explanatory variables x_{ij} through a link function $g(.)$ as

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}.$$

Generalized linear model (cont.)

Further, if we have n independent observations $(y_1, ..., y_n)$ from this distribution, we can write the log-likelihood function as

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi) = \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi).$$

By differentiating with respect to β_j we obtain the likelihood equations:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{\mathsf{Var}(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, ..., p.$$

(For, intermediate steps, see Agresti, 2015, p. 124.)

Generalized linear model (cont.)

The likelihood equations can also be written in matrix form

$$\mathbf{X}'\mathbf{D}\mathbf{V}^{-1}(\mathbf{y} - \boldsymbol{\mu}) = \mathbf{0},$$

where $\mathbf{V} = \mathsf{diag}(\mathsf{Var}(y_1), ..., \mathsf{Var}(y_n))$ and $\mathbf{D} = \mathsf{diag}(\partial\mu_1/\partial\eta_1, ..., \partial\mu_n/\partial\eta_n)$.

The Fisher information matrix for $\boldsymbol{\beta}$ is

$$\mathbf{I}(\boldsymbol{\beta}) = \mathsf{E}\left[\left(\frac{\partial l}{\partial\boldsymbol{\beta}}\right)\left(\frac{\partial l}{\partial\boldsymbol{\beta}}\right)'\right] = \mathbf{X}'\mathbf{W}\mathbf{X},$$

where $\mathbf{W} = \mathsf{diag}(w_1, .., w_n),\ w_i = (\partial\mu_i/\partial\eta_i)^2/\mathsf{Var}(y_i)$.

Therefore, approximately,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}).$$

Quasi-likelihood methods

Variance function

In GLMs, the variance is a function of the mean; denote $\mathsf{Var}(y_i) = \phi V(\mu_i)$ where $V(\mu_i) = b''(\theta_i)$ is so-called *variance function*, and we consider the special case $a_i(\phi) = \phi$. When the distribution of y_i belongs to the exponential family, the relation between the mean and the variance characterizes the distribution. For example, $V(\mu_i) = \mu_i$, $\phi = 1$, for the Poisson, $V(\mu_i) = \mu_i(1 - \mu_i)/n_i$, $\phi = 1$, for the binomial proportion and $V(\mu_i) = 1$, $\phi = \sigma^2$, for the normal.

The likelihood equations can be written as

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, ..., p.$$

In *quasi-likelihood estimation* one specifies the variance function $V(\mu_i)$ freely, without correspondence to any probability distribution. The link function $g(\mu_i) = \sum \beta_j x_{ij}$ can be specified as in GLMs.

Variance inflation

Suppose that the standard model implies that $\mathsf{Var}(y_i) = V(\mu_i)$ for the variance, but in reality,

$$\mathsf{Var}(y_i) = \phi V(\mu_i).$$

If this variance function is substituted in the likelihood equations, ϕ will be reduced, so that the solution for $\boldsymbol{\beta}$ will not change.

However, the asymptotic covariance matrix $\mathsf{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X'WX})^{-1}$, where the diagonal elements of $\mathbf{W}$ are

$$w_i = (\partial\mu_i/\partial\eta_i)^2/\mathsf{Var}(y_i) = (\partial\mu_i/\partial\eta_i)^2/[\phi V(\mu_i)],$$

will be ϕ times that for the ordinary GLM.

Estimating the variance inflation parameter ϕ

Let

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

be the Pearson statistic for the standard model. Then, approximately, $\chi^2/\phi \sim \chi^2(n - p)$, where p is the number of predictors, so that $\mathsf{E}(\chi^2/\phi) \approx n - p$, or $\mathsf{E}[\chi^2/(n-p)] \approx \phi$. Therefore, ϕ can be estimated as $\hat{\phi} = \chi^2/(n - p)$. The estimate $\hat{\boldsymbol{\beta}}$ is the same as in the standard case, but $\widehat{\mathsf{Var}}(\hat{\boldsymbol{\beta}})$ should be multiplied by $\hat{\phi}$.

In the case of overdispersion,

$$\mathsf{Var}(y_i) = \phi\mu_i \quad \text{or} \quad \mathsf{Var}(y_i) = \phi\pi_i(1 - \pi_i)/n_i$$

when the response is a Poisson variable, or binomial proportion, respectively.

Example 1: Horseshoe crabs

```
> fit.pois <- glm(y ~ weight, family = poisson)
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.42841    0.17893  -2.394   0.0167 *
weight       0.58930    0.06502   9.064   <2e-16 ***

> (X2 <- sum(residuals(fit.pois, type = "pearson")^2))
[1] 535.8957 # Pearson statistic is the sum of squared
              # Pearson residuals
> (phi <- X2/(173-2))
[1] 3.133893
#  quasi family can use QL inflated Poisson variance directly
> summary(glm(y ~ weight, family = quasi(link = "log",
      variance = "mu")))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4284     0.3168  -1.352    0.178
weight        0.5893     0.1151   5.120 8.17e-07 ***
(Dispersion parameter for quasi family taken to be 3.134159)
```

Example 1: Horseshoe crabs (cont)

```
> summary(glm(y ~ weight, family = quasi(link = "log",
      variance = "mu^2")))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0122     0.3863  -2.621  0.00957 **
weight        0.8184     0.1542   5.306 3.44e-07 ***
(Dispersion parameter for quasi family taken to be 1.362496
# When the variance function is quadratic in mu, the
# parameter estimates are also different. Standard errors
# are similar to those in negative binomial model:

> library(MASS)
> summary(glm.nb(y ~ weight))
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8647     0.4048  -2.136   0.0327 *
weight        0.7603     0.1578   4.817 1.45e-06 ***
---
Theta:  0.931   2 x log-likelihood:  -748.644
```

Overdispersion caused by correlated Bernoulli trials

Denote the n_i Bernoulli trials for observation i by $y_{i1}, y_{i2}, ..., y_{in_i}$. Then $P(y_{ij} = 1) = \pi_i$ and $P(y_{ij} = 0) = 1 - \pi_i$. Denote the sample proportion as $y_i = \sum_j y_{ij}/n_i$. For independent trials, $n_i y_i \sim \mathrm{Bin}(n_i, \pi_i)$, so that $\mathsf{Var}(y_i) = \pi_i(1 - \pi_i)/n_i$.

Let us assume that the Bernoulli trials are clustered so that $\mathsf{Cor}(y_{i_s}, y_{i_t}) = \rho$ for $s \neq t$. Then it can be shown (see Agresti, p. 272, for intermediate steps) that

$$\mathsf{Var}(y_i) = [1 + \rho(n_i - 1)]\frac{\pi_i(1 - \pi_i)}{n_i}.$$

We see that there is overdispersion if $\rho > 0$ and $n_i > 1$. This is not a special case of the inflated binomial variance model unless all n_i are identical.

Correlated Bernoulli trials (cont.)

The Bernoulli trials will be correlated this way if $y_{ij}|\pi_i \sim \text{Ber}(\pi_i)$ for $j = 1, ..., n_i$, and π_i is a random variable with mean μ_i and variance $\rho\mu_i(1 - \mu_i)$. This provides a parametric alternative for the previous QL method.

Let $\pi \sim \text{Beta}(\alpha_1, \alpha_2)$. Then

$$\mathsf{E}(\pi) = \mu \quad \text{and} \quad \mathsf{Var}(\pi) = \rho\mu(1 - \mu),$$

where

$$\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2} \quad \text{and} \quad \rho = \frac{1}{\alpha_1 + \alpha_2 + 1}.$$

Now, if $s|\pi \sim \text{Bin}(n, \pi)$, the marginal distribution of s is so-called beta-binomial distribution. For the sample mean $y = s/n$,

$$\mathsf{E}(y) = \mu \quad \text{and} \quad \mathsf{Var}(y) = [1 + \rho(n - 1)]\mu(1 - \mu)/n.$$

Example 2: Overdispersion in a teratology study (Example 8.2.4 in Agresti)

Teratology is the study of abnormalities of physiological development. Some teratology experiments investigate effects of dietary regimens or chemical agents on the fetal development of rats in a laboratory setting. Female rats on iron-deficient diets were assigned to four groups. Rats in group 1 were given placebo injections, and rats in other groups were given injections of an iron supplement. This was done on days 7 and 10 in group 2, on days 0 and 7 in group 3, and weekly in group 4. The 58 rats were made pregnant, sacrificed after 3 weeks, and then the total number of dead fetuses was counted in each litter, as was the mother's hemoglobin level. The overall sample proportions of deaths for the four groups were $= 0.758$ (placebo), 0.102, 0.034, and 0.048. Because of unmeasured covariates and genetic variability, the probability of death may vary among litters within a particular treatment group and hemoglobin level.

```
> Rats <- read.table("../data/Rats.dat", header=TRUE)
> attach(Rats)
> placebo <- ifelse(group == 1, 1, 0)

# Basic binomial model
> fit.ML <- glm(s/n ~ placebo + h, weights = n, data = Rats,
                family = binomial)
> summary(fit.ML)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.62391    0.78996   -0.790   0.4296
placebo      2.65088    0.48238    5.495  3.9e-08 ***
h           -0.18713    0.07428   -2.519   0.0118 *
        ---
> logLik(fit.ML)
'log Lik.' -121.0219 (df=3)
```

Example 2: Overdispersion in a teratology study (cont)

```
# Variance-inflated model
> summary(glm(s/n ~ placebo + h, weights = n, data = Rats,
      family = quasi(link = "logit", variance = "mu(1-mu)")))

               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -0.6239      1.3466  -0.463  0.64495
placebo         2.6509      0.8223   3.224  0.00213 **
h              -0.1871      0.1266  -1.478  0.14514
---

(Dispersion parameter for quasi family taken to be 2.905728)

> plot(residuals(fit.ML, type = "pearson") ~ n, Rats,
      cex = 0.5, ylab = "Pearson residuals", xlab = "Litter siz
> abline(h = 0, lty = 3)
# The increasing trend in the spread of residuals as the
# cluster size n_i increases suggests that
# the beta-binomial type of variance might be appropriate
```

Example 2 (cont)
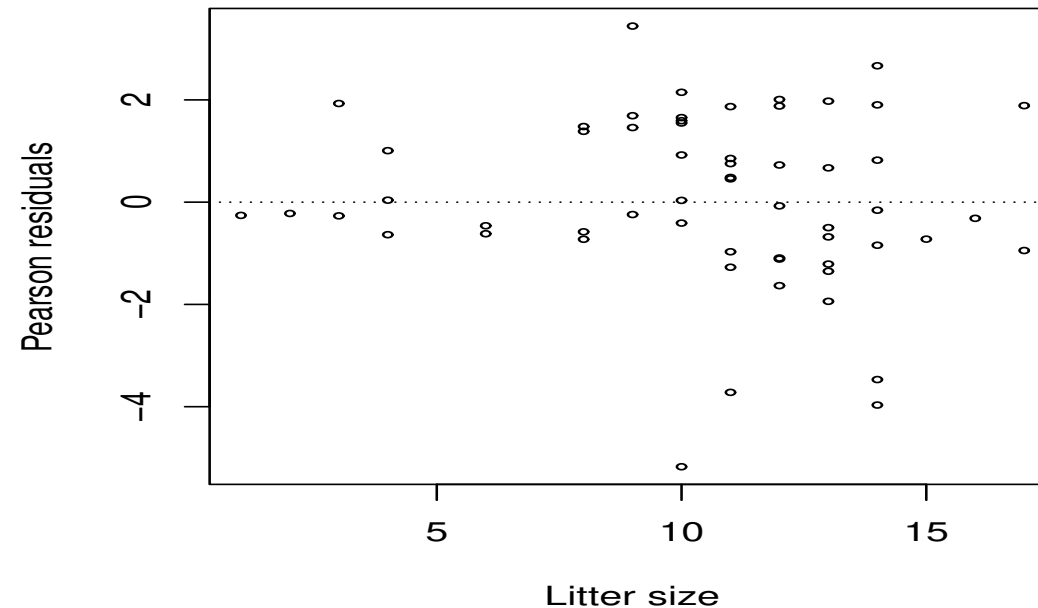
Figure 4: Pearson residuals vs. litter size for binomial logistic model fitted to Rats data.

```
> library(VGAM)
> fit.bb <- vglm(cbind(s,n-s) ~ placebo + h,
                 betabinomial(zero = 2, irho = 0.2),
                 data = Rats)
# two parameters, mu and rho; zero = 2 specifies 0 covariates
# for 2nd
# irho is initial guess for rho in beta-bin variance
summary(fit.bb)
                Estimate Std. Error z value Pr(>|z|)
(Intercept):1   -0.5009     1.1907   -0.421 0.674002
(Intercept):2   -1.1676     0.3251   -3.592 0.000328 ***
placebo          2.5601     0.7642    3.350 0.000808 ***
h               -0.1546     0.1085   -1.424 0.154362
Names of linear predictors: logitlink(mu), logitlink(rho)
Log-likelihood: -93.1849
> plogis(-1.1676) # logistic function in R
[1] 0.2372891 # Estimate of rho in beta-binomial variance
```

Example 2: Overdispersion in a teratology study (cont)

```
# QL approach using beta-binomial-type variance;
# rho (=phi) = 0.1985
> library(aod)
> quasibin(cbind(s,n-s) ~ placebo + h, data = Rats)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.7237     1.3785 -0.5250   0.5996
placebo       2.7573     0.8522  3.2355   0.0012
h            -0.1758     0.1284 -1.3692   0.1709

Overdispersion parameter:
phi
0.1985
```

Clustered data models, autumn 2024   Tampere University – 42 / 101

Quasi-likelihood and model misspecification

For the model $\eta_i = g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$, the quasi-likelihood estimates $\hat{\boldsymbol{\beta}}$ are the solutions of quasi-score equations

$$\mathbf{u}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\boldsymbol{\beta}} \right) \frac{(y_i - \mu_i)}{\phi V(\mu_i)} = \mathbf{0}.$$

We see that these equations are equivalent to the previous formula if we substitute

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}.$$

It can be shown that $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$ if the mean model $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$ is correctly specified. However, if the variance model is misspecified, $\mathsf{Var}(\hat{\boldsymbol{\beta}})$ is not correct.

Sandwich covariance adjustment

A robust estimator for $\mathsf{Var}(\hat{\boldsymbol{\beta}})$ can be obtained as follows: The first-order Taylor-series approximation for the quasi-score function is

$$\mathbf{u}(\hat{\boldsymbol{\beta}}) \approx \mathbf{u}(\boldsymbol{\beta}) + \frac{\partial \mathbf{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Since $\mathbf{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx -\left(\frac{\partial \mathbf{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)^{-1}\mathbf{u}(\boldsymbol{\beta}),$$

so that

$$\mathsf{Var}(\hat{\boldsymbol{\beta}}) \approx \left(\frac{\partial \mathbf{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)^{-1}\mathsf{Var}(\mathbf{u}(\boldsymbol{\beta}))\left(\frac{\partial \mathbf{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)^{-1}.$$

But

$$-\left(\frac{\partial \mathbf{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)^{-1} = -\left(\frac{\partial l_{quasi}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}\right)^{-1}$$

is an analog of the inverse observed information matrix and approximates the model-based covariance matrix

$$\mathbf{V} = \left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) [\phi V(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)'\right]^{-1},$$

which is based on the assumption that $\mathsf{Var}(y_i) = \phi V(\mu_i)$.

Also,

$$\mathsf{Var}(\mathbf{u}(\boldsymbol{\beta})) = \mathsf{Var}\left[\sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) \frac{(y_i - \mu_i)}{\phi V(\mu_i)}\right] = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right) \frac{\mathsf{Var}(y_i)}{[\phi V(\mu_i)]^2} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)'$$

Sandwich covariance adjustment (cont)

Therefore,

$$\mathsf{Var}(\hat{\boldsymbol{\beta}}) \approx \mathbf{V}\left[\sum_{i=1}^{n}\left(\frac{\partial\mu_i}{\boldsymbol{\beta}}\right)\frac{\mathsf{Var}(y_i)}{[\phi V(\mu_i)]^2}\left(\frac{\partial\mu_i}{\boldsymbol{\beta}}\right)'\right]\mathbf{V}.$$

This matrix simplifies to $\mathbf{V}$ if $\mathsf{Var}(y_i) = \phi V(\mu_i)$.

In practice, $\mathsf{Var}(y_i)$ is unknown. For large n we may replace μ_i with $\hat{\mu}_i$ and $\mathsf{Var}(y_i)$ with $(y_i - \hat{\mu}_i)^2$. This estimator of $\mathsf{Var}(\hat{\boldsymbol{\beta}})$ is called a *sandwich estimator*.

It can be shown that n times this estimator is consistent for the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, regardless of whether the variance specification $\mathsf{Var}(y_i) = \phi V(\mu_i)$ is correct.

Example 3: Horseshoe crabs (Example 8.3.4 in Agresti)

In an earlier example, when predicting male satellite counts by female crab weights, we used the variance-inflation approach to account for overdispersion. Another approach is to use the sandwich estimator to adjust for the standard errors.

```
library(gee)  # sandwich adjustment for generalized estimating
              # equations
obs <- 1:173 # labelling of observations for GEE method
summary(gee(y ~ weight, id = obs, family = poisson,
          scale.fix = TRUE))
              Estimate Naive S.E.    Naive z Robust S.E.  Robust z
(Intercept) -0.4284053 0.17893528 -2.394192   0.3082884 -1.389625
weight       0.5893041 0.06501714  9.063827   0.1103203  5.341755
```

Linear mixed effect models

Linear mixed effect model

In the following, we will consider clustered data. The 'cluster' might consist of consecutive measurements of some variable taken on an individual. Or, it may consist of measurements taken on several individuals (or units) in some group.

For cluster i, denote the multivariate response vector as $\mathbf{y}_i = (y_{i1}, ..., y_{id})'$, $i = 1, ..., n$. Let $\mathbf{x}_{ij}$ denote the p-vector of explanatory variable values for observation y_{ij}, with $\mu_{ij} = \mathsf{E}(y_{ij})$.

The *linear mixed model* for y_{ij} is

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{u}_i + \epsilon_{ij},$$

where $\boldsymbol{\beta}$ is a p-vector of fixed effects and $\mathbf{u}_i$ is a q-vector of random effects. In a *normal linear mixed model,* we assume that $\mathbf{u}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ and $\epsilon \sim N(0, \sigma_\epsilon^2)$.

Linear mixed effect model

Here, $\mathbf{x}'_{ij}\boldsymbol{\beta}$ represents the mean, $\mathbf{z}'_{ij}\mathbf{u}_i$ between-cluster variability, and ϵ_{ij} within-cluster variability. The basic model assumes that $\{\mathbf{u}_i\}$ and $\{\epsilon_{ij}\}$ are independent between clusters (i.e. over i) and of each other. In the simplest model ϵ_{ij} are also independent within clusters.

For $\mathbf{y}_i = (y_{i1}, ..., y_{id})'$, the model has the form

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\epsilon}_i,$$

where $\mathbf{X}_i$ is a $d \times p$ model matrix with $\mathbf{x}'_{ij}$ in row j and $\mathbf{Z}_i$ is a $d \times q$ model matrix with $\mathbf{z}'_{ij}$ in row j for the random effects.

If we assume that $\boldsymbol{\epsilon}_i \sim N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I})$, we have that

$$\mathsf{Var}(\mathbf{y}_i) = \mathbf{Z}_i\boldsymbol{\Sigma}_u\mathbf{Z}'_i + \sigma_\epsilon^2\mathbf{I}.$$

The random-intercept linear mixed model

An important special case is that only the intercept term is random. Then

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + u_i \mathbf{1} + \boldsymbol{\epsilon}_i,$$

so that

$$\mathsf{Var}(\mathbf{y}_i) = \sigma_u^2 \mathbf{1}\mathbf{1}' + \sigma_\epsilon^2 \mathbf{I}.$$

For $k \neq j$,

$$\mathsf{Cor}(y_{ij}, y_{ik}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}.$$

This correlation structure is *exchangeable*, and is called *compound symmetry*.

Multilevel models (Hierarchical models)

In some studies, the data structure is hierarchical, with sampled units nested in clusters that are themselves nested in other clusters. In this case, we may apply *multilevel models*. These models help take into account the effect of explanatory variables at each level of variation.

For example, we may use a battery of exams to study which factors affect student performance. There may be differences between the exams and schools, which may cause extra variation in the results. Let y_{ist} denote the score for student i in school s on test t. Then we may use the model

$$y_{ist} = \mathbf{x}_{ist}\boldsymbol{\beta} + u_s + v_{is} + \epsilon_{ist},$$

where $\{u_s\}$ are random effects for schools, and $\{v_{is}\}$ are random effects for students in school s.

Multilevel models (cont)

The total error variability decomposes into variance components attributable to each level. In the previous model, the total variance is $\sigma_u^2 + \sigma_v^2 + \sigma_\epsilon^2$. The intraclass correlation between scores on different exams for a student is

$$\mathsf{Cor}(y_{ist}, y_{ist'}) = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\epsilon^2},$$

and the intraclass correlation between scores on a particular exam for pairs of students in the same school are

$$\mathsf{Cor}(y_{ist}, y_{i'st}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_\epsilon^2}.$$

Example 1: Smoking prevention and cessation study (Example 9.2.3 in Agresti)

Hedeker and Gibbons (2006, p. 9) analyzed data from a study of the efficacy of two programs for discouraging young people from starting or continuing to smoke. The study compared four groups, defined by a 2×2 factorial design according to whether a student was exposed to a school-based curriculum (SC; $1 =$ yes, $0 =$ no) and a television-based prevention program (TV; $1 =$ yes, $0 =$ no). The subjects were 1600 seventh-grade students from 135 classrooms in 28 Los Angeles schools. The schools were randomly assigned to the four intervention conditions.

The response variable was a tobacco and health knowledge (THK) scale, measured at the end of the study. This variable was also observed at the beginning of the study, and that measure (PTHK = Pre-THK) was used as a covariate. THK took values between 0 and 7. (The data are available in the file Smoking.dat at the text website.)

Example 1: Smoking prevention and cessation study (cont)

We fit the the multilevel model

$$y_{ics} = \beta_0 + \beta_1 \mathrm{PTHK}_{ics} + \beta_2 \mathrm{SC}_{ics} + \beta_3 \mathrm{TV}_{ics} + u_s + v_{cs} + \epsilon_{ics},$$

for student i within classroom c in school s. The random effects are $u_s \sim N(0, \sigma_u^2)$, $v_{cs} \sim N(0, \sigma_v^2)$ and $\epsilon_{ics} \sim N(0, \sigma_\epsilon^2)$.

```
> library(lme4) # Doug Bates's linear mixed models package
> attach(Smoking)
> Smoking # data in file Smoking.dat at
            # www.stat.ufl.edu/~aa/glm/data
   school class SC TV PTHK y
1     403 403101 1   0    2 3
2     403 403101 1   0    4 4
...
1600 515 515113 0   0    3 3
```

Example 1: Smoking prevention and cessation study (cont)

```
> fit <- lmer(y ~ PTHK + SC + TV + (1|school) + (1|class))
> summary(fit) # school and classroom random intercepts
Random effects: # These are "REML" variance estimates
 Groups    Name         Variance Std.Dev.
class    (Intercept) 0.06853  0.2618
school   (Intercept) 0.03925  0.1981
Residual              1.60108  1.2653
Number of obs: 1600, groups:  class, 135; school, 28
Fixed effects:
Estimate Std. Error t value
(Intercept)   1.78493     0.11295  15.803
PTHK          0.30524     0.02590  11.786
SC            0.47147     0.11330   4.161
TV            0.01956     0.11330   0.173
# use ranef(fit) to predict random effects at levels of
# school, class
# use predict(fit) to get predicted values for the observatio
```

Linear models with random intercept and random slope

In addition to random intercepts, slopes may also be modeled as random. For example, in longitudinal studies there may be individual differences in the rate of change of some variable.

A simple example of such models, applied to a clinical trial, is

$$y_{ij} = (\beta_0 + u_{i1}) + (\beta_1 + u_{i2})t_j + \beta_2 x_i + \beta_3 t_j x_i + \epsilon_{ij},$$

where y_{ij} is the response of subject i at time t_j and x_i is a treatment indicator. The mean intercept is β_0 for the placebo group and $\beta_0 + \beta_2$ for the treatment group. Further, the mean slope is β_1 for the placebo group and $\beta_1 + \beta_3$ for the treatment group. In addition, there is between-subject variation in both intercepts and slopes.

Linear models with correlated errors

In a linear mixed model $\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\epsilon}_i$ the within-subject errors ϵ_{ij}, $j = 1, ..., d$, may be allowed to be correlated. This may be a reasonable assumption when subjects are observed spatially or over time.

In this case,

$$\mathsf{Var}(\mathbf{y}_i) = \mathbf{Z}_i \boldsymbol{\Sigma}_u \mathbf{Z}_i' + \mathbf{R},$$

where $\mathbf{R} = \mathsf{Var}(\boldsymbol{\epsilon}_i)$ is not a diagonal matrix.

In an AR(1) correlation structure, $\mathsf{Cor}(\epsilon_{ij}, \epsilon_{ik}) = \rho^{|j-k|}$. In the more general *Toeplitz*-structure, where $\mathsf{Cor}(\epsilon_{ij}, \epsilon_{ik}) = \rho_{|j-k|}$, the correlation also depends on the time separation of the observations. In the most general *unstructured* case, the correlation matrix has $\binom{d}{2}$ free elements.

Example 2: FEV study (Section 9.2.5 in Agresti)

In a pharmaceutical clinical trial analyzed by Little et al. (2000), 24 patients were randomly assigned to three treatment groups (drug A, drug C, placebo) and compared on a measure of respiratory ability, FEV (forced expiratory volume in liters in 1 second). The study observed FEV at baseline and then at intervals of 8 hours.

A simple model for the FEV measurement at hour j for subject i is

$$y_{ij} = \beta_0 + \beta_1 \text{baseline}_i + \beta_2 c_i + \beta_3 p_i + \beta_4 j + u_i + \epsilon_{ij},$$

where $c_i = 1$ for drug C and 0 otherwise and $p_i = 1$ for placebo and 0 otherwise. We fit two models: the first with independent within-patient errors and the second with an AR(1) correlation structure.

Example 2: FEV study (cont)

```
> library(nlme)
> attach(FEV2)
> summary(lme(fev ~ base + factor(drug) + hour,
                random = ~ 1|patient))
AIC       BIC     logLik
388.9149 419.3466 -187.4575
Random effects:
        (Intercept)  Residual
StdDev:   0.4526834 0.2716699
Fixed effects:  fev ~ base + factor(drug) + hour
                  Value  Std.Error  DF    t-value  p-value
(Intercept)    1.0492317 0.29217514 503   3.591105  0.0004
base           0.9028516 0.10328130  68   8.741675  0.0000
factor(drug)c  0.2258930 0.13361174  68   1.690667  0.0955
factor(drug)p -0.2814907 0.13362978  68  -2.106496  0.0389
hour          -0.0745734 0.00494027 503 -15.095011  0.0000
```

Example 2: FEV study (cont)

```
> summary(lme(fev ~ base + factor(drug) + hour,
                random = ~ 1|patient, correlation =
                    corAR1(form = ~ hour|patient)))
AIC       BIC     logLik
243.3501 278.1292 -113.675
Random effects:
        (Intercept)  Residual
StdDev:   0.4075485 0.3354964
Correlation Structure: AR(1)
0.6480841 # autoregressive correlation estimate
Fixed effects:  fev ~ base + factor(drug) + hour
                  Value  Std.Error  DF   t-value p-value
(Intercept)    1.0723482 0.29138890 503  3.680127  0.0003
base           0.8917791 0.10257147  68  8.694222  0.0000
factor(drug)c  0.2129614 0.13269345  68  1.604912  0.1131
factor(drug)p -0.3141641 0.13271136  68 -2.367274  0.0208
hour          -0.0690618 0.00769164 503 -8.978814  0.0000
```

Maximum likelihood model fitting

We combine the n observation vectors $\mathbf{y}_i$ into an nd-vector $\mathbf{y}$ and express the linear mixed model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{pmatrix}, \ \mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{Z}_n \end{pmatrix}, \ \mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_n \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_n \end{pmatrix}.$$

Marginally, assuming that $\mathbf{y}_i$ are independent and multivariate normal,

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}' + \mathbf{R}_{\boldsymbol{\epsilon}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{u}} = \text{diag}(\boldsymbol{\Sigma}_u, ..., \boldsymbol{\Sigma}_u)$ and $\mathbf{R}_{\boldsymbol{\epsilon}} = \text{diag}(\mathbf{R}, ..., \mathbf{R})$.

Mixed effect model: Estimation

Ignoring the constant terms, the log-likelihood function is

$$L(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2} \log|\mathbf{V}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

where $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma_u}\mathbf{Z}' + \mathbf{R_\epsilon}$.

If $\mathbf{V}$ is known, then maximizing $L(\boldsymbol{\beta}, \mathbf{V})$ with respect to $\boldsymbol{\beta}$ yields the generalized least squares solution

$$
\begin{aligned}
\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\mathbf{V}) &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\
&= \left(\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\sum_{i=1}^{n}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i.
\end{aligned}
$$

Mixed effect model: Estimation

Since $\mathsf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, this estimator is unbiased. Since it is a linear function of $\mathbf{y}$, it is multivariate normal with

$$\mathsf{Var}(\tilde{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1}.$$

In practice, $\mathbf{V}$ is rarely known. Substituting an estimate $\hat{\mathbf{V}}$ in $\tilde{\boldsymbol{\beta}}$ we obtain

$$\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{\mathbf{V}}) = \left(\sum_{i=1}^{n} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \hat{\mathbf{V}}_i^{-1} \mathbf{y}_i.$$

Under regularity conditions, its asymptotic distribution is the same as when $\mathbf{V}$ is known.

Prediction of random components

After estimating $\boldsymbol{\beta}$ and $\mathbf{V}$ we can predict the random effects $\{\mathbf{u}_i\}$. This is useful in many applications. For example, in multilevel performance models they can be used to rank institutions, such as schools and hospitals. In *small area estimation*, characteristics are estimated for many geographical areas which each have relatively few observations. Random effect models, which take clustering into account, can provide improved estimates, because those estimates borrow from the whole, rather than using only cluster-specific data.

A predictor $\tilde{u}_i$ is called *the best linear unbiased predictor* (BLUP) of the random effect $\mathbf{u}_i$ if $\tilde{\mathbf{u}}_i$ is linear in $\mathbf{y}$, $\mathsf{E}(\tilde{\mathbf{u}}_i) = \mathbf{0}$ and $\mathsf{E}\{[\mathbf{a}'(\tilde{\mathbf{u}}_i - \mathbf{u}_i)]^2\}$ is minimized for any linear combination $\mathbf{a}'(\tilde{\mathbf{u}}_i - \mathbf{u}_i)$.

Prediction of random components (cont)

For the normal linear mixed model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}$ with $\mathsf{Var}(\mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{u}}$ and $\mathsf{Var}(\boldsymbol{\epsilon}) = \mathbf{R}_{\boldsymbol{\epsilon}}$, $\mathsf{Cov}(\mathbf{y}, \mathbf{u}) = \mathsf{Cov}(\mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \mathbf{u}) = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{u}}$, so

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{u} \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}' + \mathbf{R}_{\boldsymbol{\epsilon}} & \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{u}} \\ \boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}' & \boldsymbol{\Sigma}_{\mathbf{u}} \end{pmatrix}\right].$$

Using the properties of the conditional distributions of multivariate normal variables, we obtain that

$$\mathsf{E}(\mathbf{u}|\mathbf{y}) = \boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

for $\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}' + \mathbf{R}_{\boldsymbol{\epsilon}}$. With known variances, the BLUP $\tilde{\mathbf{u}}$ of $\mathbf{u}$ is therefore

$$\tilde{\mathbf{u}} = \boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \boldsymbol{\Sigma}_{\mathbf{u}}\mathbf{Z}'\mathbf{V}^{-1}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}]\mathbf{y}.$$

Example: balanced one-way layout

The prediction $\tilde{\mathbf{u}}$ is a weighted combination of $\mathbf{0}$ and the GLS estimate based on treating $\mathbf{u}$ as a fixed effect. As an example, consider one-way layout

$$y_{ij} = \beta_0 + u_i + \epsilon_{ij}, \; i = 1, ..., c, \; j = 1, ..., n,$$

where $\mathsf{Var}(\boldsymbol{\epsilon}_i) = \sigma_\epsilon^2 \mathbf{I}$. In matrix form this is

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_n \\ \vdots \\ \mathbf{1}_n \end{pmatrix}, \; \boldsymbol{\beta} = \beta_0, \; \mathbf{Z} = \begin{pmatrix} \mathbf{1}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_2 & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_n \end{pmatrix}, \; \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_c \end{pmatrix},$$

with $\boldsymbol{\Sigma}_{\mathbf{u}} = \mathsf{Var}(\mathbf{u}) = \sigma_u^2 \mathbf{I}_c$ and $\mathbf{R}_{\boldsymbol{\epsilon}} = \sigma_\epsilon^2 \mathbf{I}_{nc}$.

Example: balanced one-way layout (cont)

For this structure, $\hat{\beta}_0 = \bar{y}$ and

$$
\tilde{u}_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n}(\bar{y}_i - \bar{y}),
$$

$$
\hat{\beta}_0 + \tilde{u}_i = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2/n}\bar{y}_i + \frac{\sigma_\epsilon^2}{\sigma_u^2 + \sigma_\epsilon^2/n}\bar{y}.
$$

Thus, $\tilde{u}_i$ is a weighted average of 0 and the LS estimate $(\bar{y}_i - \bar{y})$ from treating $\mathbf{u}$ as a fixed effect. The estimated mean $\hat{\beta}_0 + \tilde{u}_i$ for group i is a weighted average of $\bar{y}_i$ and $\bar{y}$, with greater weight for $\bar{y}_i$ as n increases.

Residual (or restricted) maximum likelihood method

A problem in ML estimation is that it produces biased estimates for the variance components. It is possible reduce or completely remove the bias using restricted ML (REML) estimation. In this approach, the fixed part is first removed by considering a likelihood based on $\mathbf{z} = \mathbf{L}\mathbf{y}$ where $\mathbf{L}$ is a matrix satisfying $\mathbf{L}\mathbf{X} = \mathbf{0}$. Then $\mathsf{E}(\mathbf{L}\mathbf{y}) = \mathbf{0}$, and the distribution of so-called *error contrasts*, $\mathbf{L}\mathbf{y} = \mathbf{L}(\mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon})$, does not depend on the fixed effects $\boldsymbol{\beta}$.

We may choose $\mathbf{L} = \mathbf{I} - \mathbf{P_X}$, where $\mathbf{P_X}$ is the orthogonal projector matrix on the space $\mathbf{C}(\mathbf{X})$. Then the error contrasts $\mathbf{L}\mathbf{y} = (\mathbf{I} - \mathbf{P_X})\mathbf{y} = \mathbf{y} - \hat{\boldsymbol{\mu}}$ are residuals, which explains the name 'residual ML'.

For example, in the normal linear model, the REML estimator of σ^2 is the unbiased estimator $s^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/(n-r)$, where r is the number of regression parameters.

Mixed effect model: Testing

Hypotheses concerning the fixed part of the model can be tested using the likelihood ratio test (LRT). However, in this case one should use the ML estimation method, since two restricted likelihood functions are not comparable when the fixed parts are different. Even in the case that the ML method is used, the LRT statistic is only approximately χ^2-distributed under H_0, and a more accurate p-value can be obtained by simulating the distribution of the test statistic.

The significance of variance components can also be tested using the LRT. If the fixed part is the same in both hypotheses, REML can also be used in estimation. However, when testing the significance of a variance component, this is on the border of the parameter space under H_0. Therefore, the distribution of the test statistic is not even approximately χ^2. Here, again, the simulation approach can be used.

Mixed effect models: prediction

When predicting new observations it is important to consider on what level of the model the prediction is made. For example, in a block experiment the treatment effect might be modelled as a fixed effect and the block effect as a random effect. The model can be written as

$$y_{ijk} = \mu + \tau_i + \gamma_j + \epsilon_{ijk}$$

where μ is the overall mean, τ_i the treatment effect, γ_j the random effect and ϵ_{ijk} the error term. It is assumed that $\gamma_j \sim N(0, \sigma_\gamma^2)$ and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

When predicting a new observation for treatment i and block j, the best linear unbiased predictor (BLUP) is $\hat{\mu} + \hat{\tau}_i + \hat{\gamma}_j$. When predicting for a new block (outside the original design) the BLUP becomes $\hat{\mu} + \hat{\tau}_i$.

Mixed effect models: Example

We consider data from Box, Hunter, and Hunter (1978), and compare four processes A, B, C and D in the production of penicillin. The raw material, corn steep liquor, can be made in blends sufficient for four runs. In order to control for the blend effect, a randomized complete block design is used. Each process is applied to 5 blends, so we have altogether 20 observations. The lmer function in R package lme4 can be used for estimation.

The blends can be regarded as a sample of a larger population, so it is natural to model the blend effect as random. The treatment effect can be tested with LRT if the estimation method is ML. Under H_0, the test statistic is approximately χ_3^2-distributed. The result is insignificant, with p-value 0.26. A more accurate p-value, 0.336, is obtained by simulation (parametric bootstrap).

Mixed effect models: Example (cont)

The significance of the random effect can also be tested using LRT. Under $H_0 : \sigma_\gamma^2 = 0$ the variance component is on the border of its parameter space, so the test statistic is not distributed as χ_1^2. Thus, simulation is needed. Both ML and REML methods can be used in estimation, but the same method must be used for both hypotheses. The p-value is 0.043, giving a significant result in the 5% risk level. The density of the test statistic is shown in Figure 5 (left) and the density of χ_1^2 (right).

The estimated random effects

$$4.29, -2.14, -0.71, 1.43, -2.86,$$

are a shrinked version of the 2-way ANOVA effects

$$6, -3, -1, 2, -4.$$
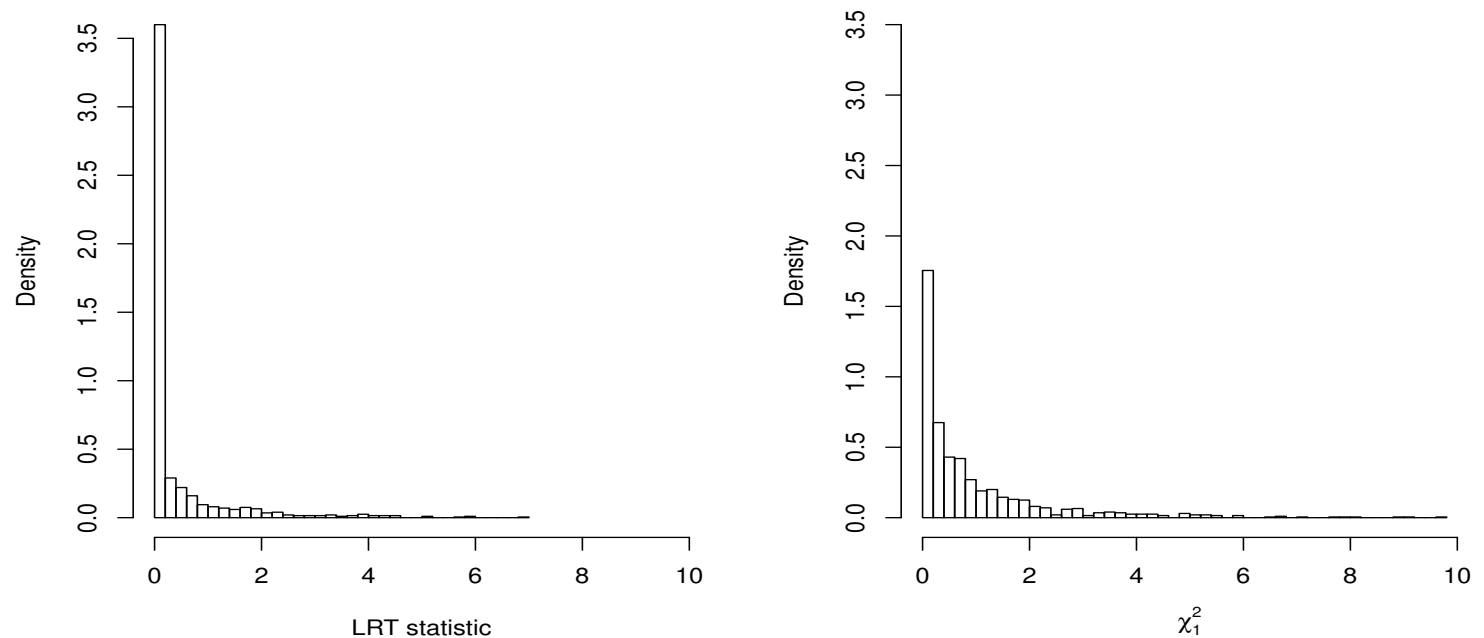
Distribution of the test statistic

Figure 5: The density of the LRT statistic when testing the significance of random effects in the penicilline experiment (left), and the density of the χ_1^2 random variable (right).

Longitudinal data

In a *longitudinal study*, repeated measurements on individuals are taken over time. This type of study is also sometimes called a *panel study*, especially if the measurements are taken at equal time intervals.

It is usually assumed that longitudinal data have variation on more than one level, and it is natural to model such data using mixed effect models. Typically, population level effects are modelled with fixed parameters, and individual level variation with random parameters. The measurements within an individual may be assumed either to be independent or have some correlation structure.

Example: PSID

As an example we consider data set psid from faraway package. It is a small subset of the total data of the Panel Study of Income Dynamics (PSID), which begun in 1968 and is a longitudinal study of a representative sample of U.S. individuals. The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan and is still continuing. Some cases are shown in Figure 6.

We model the income as a function of time and the following predictors: age, sex and educational level. We also assume that there is variation between individuals in the overall level and growth rate of income, not explained by the predictors. This variation can be modeled using random components.

PSID data
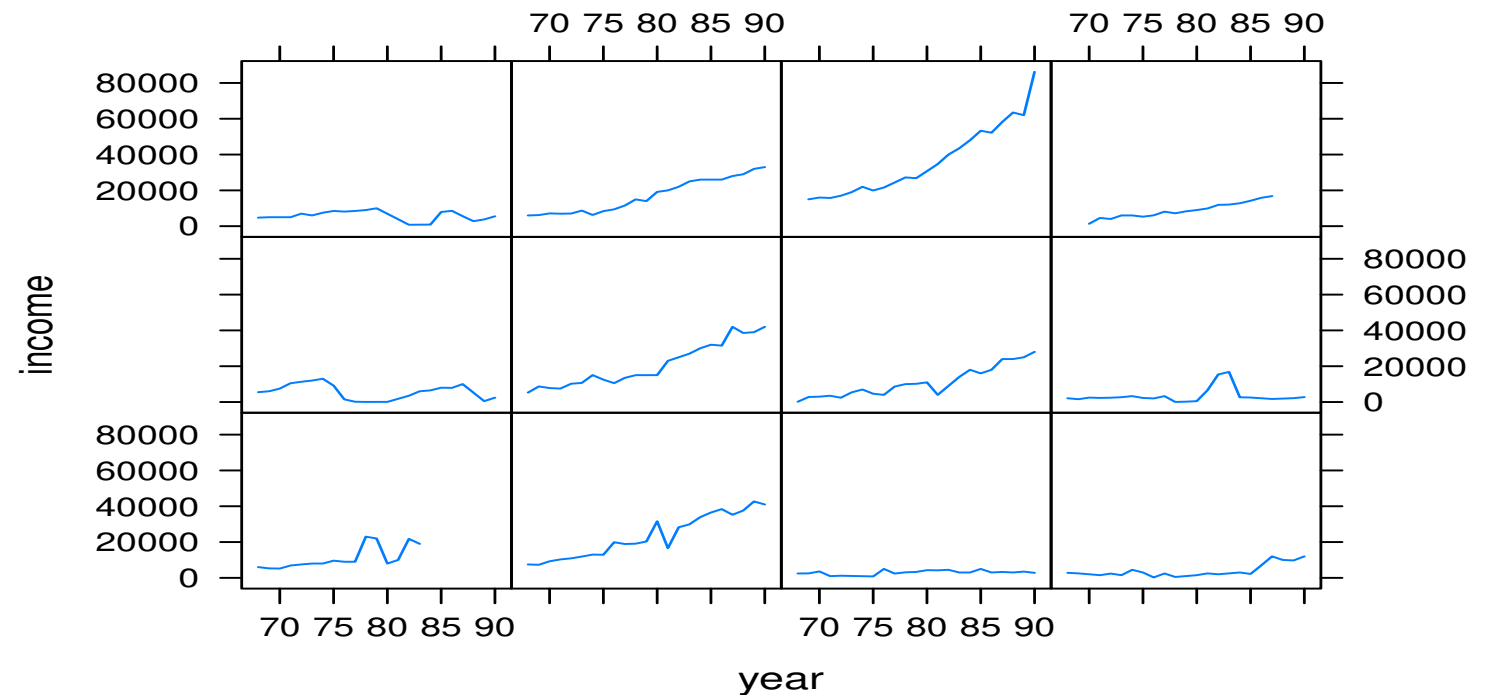
Figure 6: Income as a function of time for 12 individuals in the PSID data.

Example: PSID (cont)

The model for individual i and year j can be written as:

$$
\log(income)_{ij} = \quad \beta_0 + \beta_1 year_j + \beta_2 sex_i + \beta_3 sex_i \times year_j
$$
$$
+ \beta_4 educ_i + \beta_5 age_i + u_{i0} + u_{i1} year_j + \epsilon_{ij},
$$

and

$$
\begin{pmatrix} u_{i0} \\ u_{i1} \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_u).
$$

Here are some results:

```
Random effects:
 Groups    Name          Variance Std.Dev. Corr
 person    (Intercept)   0.2817   0.53071
           cyear         0.0024   0.04899  0.19
 Residual                0.4673   0.68357
```
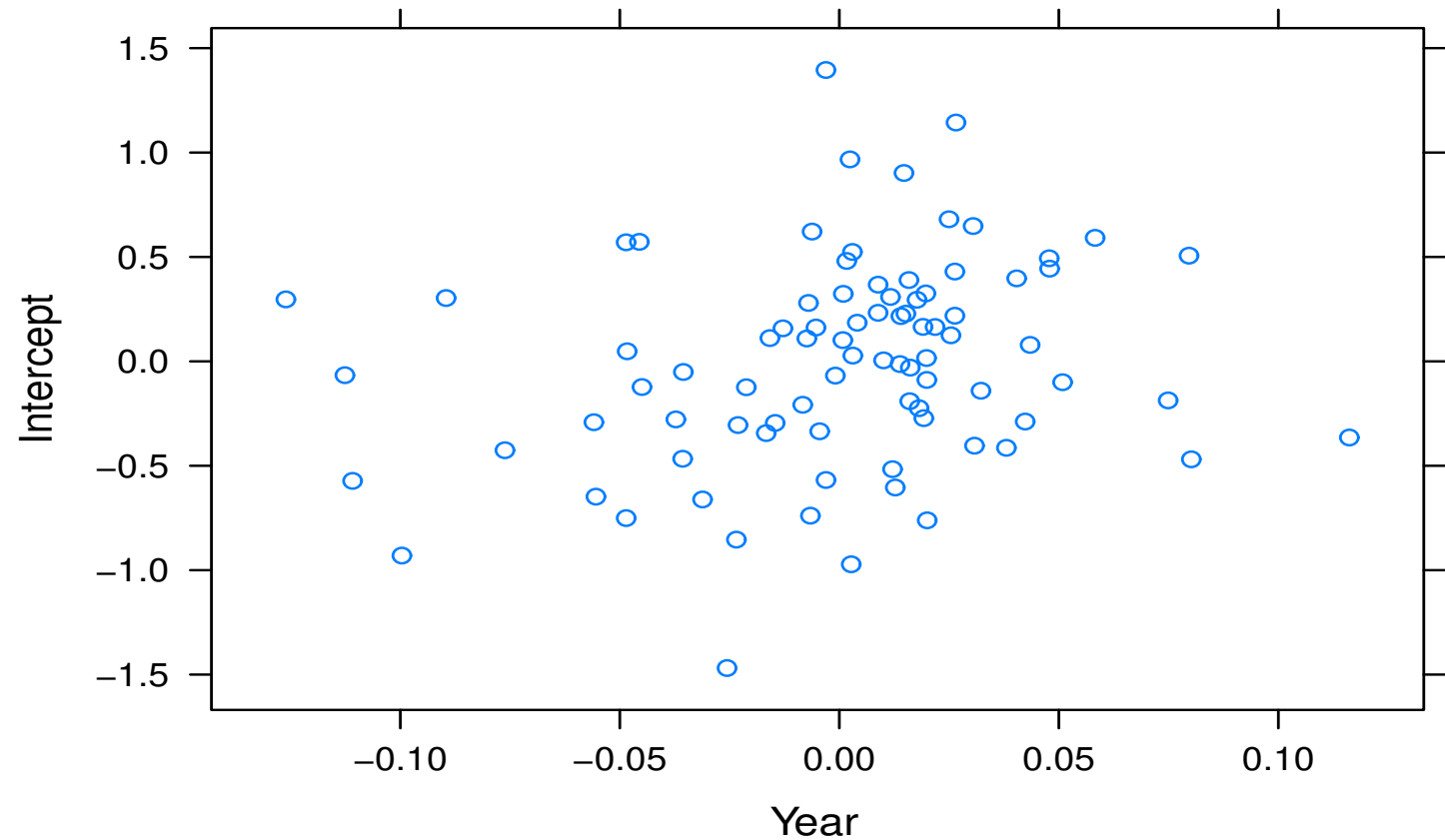
Distribution of the random effects

Figure 7: Bivariate distribution of the random effect estimates.

Mixed effect model for nonnormal responses

Generalized linear mixed model (GLMM)

The GLM can be generalized to allow for clustered data. One possibility is introducing random effects. For example, binary data may be modeled using *logistic-normal models*, which combine the logistic regression model with normally distributed random effects. The model form is

$$\text{logit}[P(y_{ij} = 1|\mathbf{u}_i)] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i,$$

where $\{\mathbf{u}_i\}$ are independent $N(\mathbf{0}, \boldsymbol{\Sigma}_u)$ random variables.

For example, a simple random intercept model for binary matched observations in d clusters is

$$\text{logit}(P(y_{ij} = 1|u_i)) = \beta_0 + \beta_j + u_i, \; j = 2, ..., d.$$

This might describe the outcomes (correct, incorrect) for d questions in an exam.

General binary intercept model: correlation

In binary intercept models, the observations y_{ij} and y_{ik}, $(j \neq k)$, are assumed to be conditionally independent, given u_i. In a general case where the inverse link function is an arbitrary cdf F, the marginal covariance is

$$
\begin{aligned}
\mathsf{Cov}(y_{ij}, y_{ik}) &= \mathsf{E}[\mathsf{Cov}(y_{ij}, y_{ik}|u_i)] + \mathsf{Cov}[\mathsf{E}(y_{ij}|u_i), \mathsf{E}(y_{ik}|u_i)] \\
&= 0 + \mathsf{Cov}[F(\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i), F(\mathbf{x}'_{ik}\boldsymbol{\beta} + u_i)].
\end{aligned}
$$

The functions in the last covariance term are both monotonicallly increasing in u_i, and are hence positively correlated. Thus, $\mathsf{Cor}(y_{ij}, y_{ik}) > 0$.

Binary random effect models vs. marginal models

Effects in binary random-effects models tend to be larger than those in corresponding marginal models. For random-intercept probit regression, the marginal model is also a probit model. The model implies that

$$P(y_{ij} = 1 | u_i) = \Phi(\mathbf{x}_{ij}'\boldsymbol{\beta} + u_i) = P(z \le \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i),$$

where $z \sim N(0, 1)$. Since $z - u_i \sim N(0, 1 + \sigma_u^2)$,

$$P(y_{ij} = 1) = P(z - u_i \le \mathbf{x}_{ij}'\boldsymbol{\beta}) = \Phi\left(\frac{\mathbf{x}_{ij}'\boldsymbol{\beta}}{\sqrt{1 + \sigma_u^2}}\right).$$

We see that the coefficients of the marginal model are obtained by dividing those from the GLMM by $\sqrt{1 + \sigma_u^2}$. (The corresponding transformation for the logistic model is approximately $\boldsymbol{\beta}/\sqrt{1 + (\sigma_u/c)^2}$ for $c \approx 1.7$.)

Poisson GLMM for correlated count data

The Poisson regression model with a normal random effect and log link is

$$\log[\mathsf{E}(y_{ij}|\mathbf{u}_i)] = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i,$$

where $\{\mathbf{u}_i\}$ are independent $N(\mathbf{0}, \boldsymbol{\Sigma}_u)$. Conditional on $\mathbf{u}_i$, y_{ij} has a Poisson distribution.

The random intercept model is of the form

$$\log[\mathsf{E}(y_{ij}|\mathbf{u}_i)] = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i.$$

Note that y_{ij} will have a negative-binomial distribution if instead $\exp(u_i)$ has a gamma distribution. The Poisson GLMM with normal random effects has the advantage that it can be easily generalized to multivariate random effects and multilevel models.

For the random-intercept Poisson GLMM, the mean of the marginal model is

$$\mathsf{E}(y_{ij}) = \mathsf{E}[\mathsf{E}(y_{ij}|u_i)] = \mathsf{E}[e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i}] = e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+\sigma_u^2/2}.$$

Thus, the log mean conditionally equals $\mathbf{x}'_{ij}\boldsymbol{\beta} + u_i$ and marginally equals $\mathbf{x}'_{ij}\boldsymbol{\beta} + \sigma_u^2/2$.

As in the negative-binomial distribution, the marginal variance is a quadratic function of the marginal mean:

$$
\begin{aligned}
\mathsf{Var}(y_{ij}) &= \mathsf{E}[\mathsf{Var}(y_{ij}|u_i)] + \mathsf{Var}[\mathsf{E}(y_{ij}|u_i)] \\
&= \mathsf{E}[e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+u_i}] + e^{2\mathbf{x}'_{ij}\boldsymbol{\beta}}\mathsf{Var}(e^{u_i}) \\
&= e^{\mathbf{x}'_{ij}\boldsymbol{\beta}+\sigma_u^2/2} + e^{2\mathbf{x}'_{ij}\boldsymbol{\beta}}\left(e^{2\sigma_u^2} - e^{\sigma_u^2}\right) \\
&= \mathsf{E}(y_{ij}) + [\mathsf{E}(y_{ij})]^2\left(e^{\sigma_u^2} - 1\right).
\end{aligned}
$$

GLMM fitting

In order to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_u$ we need to compute their marginal likelihood. Let $f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta})$ denote the conditional probability mass function or density function of $\mathbf{y}$, given $\mathbf{u}$, and let $f(\mathbf{u}; \boldsymbol{\Sigma_u})$ denote the normal pdf of $\mathbf{u}$. Then the marginal likelihood is

$$f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma_u}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma_u}) d\mathbf{u}.$$

Many methods can approximate this numerically.

One method with relatively easy implementation is the penalized quasi-likelihood method (PQL). This method is based on the linearization of the response equation. We can compute p-values using likelihood-based methods, but the inference is only asymptotically correct. Biased results may arise especially for binary responses and binomial responses with small groups.

GLMM fitting (cont)

Another approach is numerical integration. When the dimension of $\mathbf{u}_i$ is small, like in random intercept models, standard numerical integration methods, such as *Gauss-Hermite quadrature*, can approximate the likelihood well. The approximation can be improved by increasing quadrature points. Liu and Pierce (1994) suggested an adaptive version which centers the quadrature points with respect to the mode of the integrand and scales them. This approach is usually superior to PQL by providing an approximation to the true likelihood, rather than a quasi-likelihood.

A third approach is to use Monte Carlo -based methods (such as MCMC) which may help overcome the "curse of dimensionality". The parameter estimates converge to the ML estimates as the Monte Carlo sample size increases.

The respondents in a General Social Survey indicated whether they supported legalizing abortion in each of $d = 3$ situations.

Let y_{ij} denote the response for subject i in situation j, with $y_{ij} = 1$ representing support of legalization. We first fit the random-intercept GLMM

$$\text{logit}(P(y_{ij} = 1 | u_i)) = \beta_0 + \beta_j + \gamma x_i + u_i,$$

where $x_i = 1$ for females and 0 for males and $\{u_i\}$ are independent $N(0, \sigma_u^2)$. The situation effects $\{\beta_j\}$ satisfy a constraint, $\beta_3 = 0$ in the R output.

Situations are (1) if the family has a very low income and cannot afford any more children, (2) when the woman is not married and does not want to marry the man, and (3) when the woman wants it for any reason.

Frequency table. Support for legalized abortion in three situations, by gender.

```
        Sequence of responses (1=yes,0=no) in three situations

        1,1,1 1,1,0 1,0,1 1,0,0 0,1,1 0,1,0 0,0,1 0,0,0
Male      342    26    11    32     6    21    19   356
Female    440    25    14    47    14    18    22   457

> Abortion # data file Abortion.dat at
           # www.stat.ufl.edu/~aa/glm/data
    gender response situation case
1        1        1         1    1
2        1        1         2    1
3        1        1         3    1
4        1        1         1    2
```

```
> z1 <- ifelse(Abortion$situation==1,1,0)
> z2 <- ifelse(Abortion$situation==2,1,0)
> library(glmmML)  # ML fitting of GLMMs
> # Alternative: glmer function in lme4 package
> fit.glmm <- glmmML(response ~ gender + z1 + z2,
+                cluster=Abortion$case, family=binomial, data=
+                method = "ghq", n.points=70, start.sigma=9)
                 # uses adaptive GHQ
> summary(fit.glmm)
                coef    se(coef)       z  Pr(>|z|)
(Intercept) -0.61874    0.3777 -1.63839 1.01e-01
gender       0.01259    0.4888  0.02575 9.79e-01
z1           0.83470    0.1601  5.21347 1.85e-07
z2           0.29240    0.1567  1.86621 6.20e-02
Scale parameter in mixing distribution:  8.736 gaussian
Std. Error:                              0.5421
LR p-value for H_0: sigma = 0:  0
```

Generalized estimation equations

Generalized estimation equations (GEE)

Another possibility to model panel-type data with non-normal response is generalized estimation equations (GEE). While in GLMM intra-cluster correlation is modeled using cluster-specific random effects, in GEE a marginal model is specified for population-wide effects.

We only need to specify the mean and variance functions for the response variable, and a correlation structure between the observations in the same cluster. Let y_{ij} denote the observation j in cluster i, $j = 1, ..., d$, and let $\mathbf{x}_{ij}$ denote the vector of explanatory variables. The mean $\mathsf{E}(y_{ij}) = \mu_{ij}$ is connected to the predictors $\mathbf{x}_{ij}$ with a link function g: $g(\mu_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta}$.

Further, we assume that the variance of y_{ij} is given as a function of μ_{ij}: $\mathsf{Var}(y_{ij}) = \phi V(\mu_{ij})$.

Generalized estimation equations (cont)

We also need to define a correlation structure between the observations in the same cluster. Let $\mathbf{R}(\boldsymbol{\alpha})$ denote the correlation matrix and let $\mathbf{B}_i$ be a diagonal matrix with diagonal elements $V(\mu_{ij})$. Then the covariance matrix is

$$\mathbf{V}_i = \phi \mathbf{B}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{B}_i^{1/2}.$$

The unknown $\boldsymbol{\beta}$ can be solved from the *generalized estimation equations*

$$\sum_i \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)' \mathbf{V}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}.$$

These can be understood as a generalization of the likelihood equations. However GEE need not correspond to the likelihood of any standard distribution. Therefore, GEE are a quasi-likelihood method.

Generalized estimation equations (cont)

The correlation parameters $\boldsymbol{\alpha}$ can be solved from similar equations. Finally, the dispersion (or scale) parameter ϕ is usually estimated as

$$\hat{\phi} = \frac{1}{nd - p} \sum_{i=1}^{n} \sum_{j=1}^{d} \hat{r}_{ij}^2,$$

where p is the number of explanatory variables and $\hat{r}_{ij}$ is the Pearson residual, $\hat{r}_{ij} = (y_{ij} - \hat{\mu}_{ij})/\sqrt{V(\hat{\mu}_{ij})}$.

The estimator of $\boldsymbol{\beta}$ is consistent ($\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ when $n \to \infty$) and asymptotically multivariate normal even if the variance function or the correlation structure is misspecified. Moreover, the *sandwich estimator*, introduced in the context of QL methods, is available as a robust covariance matrix of $\hat{\boldsymbol{\beta}}$. These properties make it possible to make inference about $\boldsymbol{\beta}$.

Correlation structures

Two examples of possible correlation structures:

✔ Exchangeable $R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix}$

✔ AR(1): autoregressive of order 1

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{n_i-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{n_i-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{n_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{n_i-1} & \alpha^{n_i-2} & \alpha^{n_i-3} & \cdots & 1 \end{bmatrix}$$

Marginal model analog for the survey response example is

$$\text{logit}[P(y_{ij} = 1)] = \beta_0 + \beta_j + \gamma x_i.$$

```
> library(gee)
> fit.gee <- gee(response ~ gender + z1 + z2, id=case, family=binomi
+               corstr="exchangeable", data=Abortion)
> summary(fit.gee)
                Estimate Naive S.E.      Naive z Robust S.E. Robust z
(Intercept) -0.125325730 0.06782579 -1.84775925  0.06758212 -1.85442
gender       0.003437873 0.08790630  0.03910838  0.08784072  0.03913
z1           0.149347107 0.02814374  5.30658404  0.02973865  5.02198
z2           0.052017986 0.02815145  1.84779075  0.02704703  1.92324
Working Correlation
            [,1]      [,2]      [,3]
[1,] 1.0000000 0.8173308 0.8173308
[2,] 0.8173308 1.0000000 0.8173308
[3,] 0.8173308 0.8173308 1.0000000
```

Example: Modeling correlated survey responses (cont)

```
> fit.gee2 <- gee(response ~ gender + z1 + z2, id=case, family=binom
+                 corstr="independence", data=Abortion)
> summary(fit.gee2) # same estimates as ML, assuming independence
                 Estimate Naive S.E.      Naive z Robust S.E. Robust z
(Intercept) -0.125407576 0.05562131 -2.25466795  0.06758236 -1.85562
gender       0.003582051 0.05415761  0.06614123  0.08784012  0.04077
z1           0.149347113 0.06584875  2.26803253  0.02973865  5.02198
z2           0.052017989 0.06586692  0.78974374  0.02704704  1.92324
Working Correlation
[,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

Missing data

Missing data types

It is very common in statistical data sets that some observations are missing. For instance, in longitudinal studies, some individuals may fail to respond after a certain time point. In cross-sectional studies, there may occur unit non-response (no response at all from some individuals) or item non-response (certain questions in a survey are not answered).

A common technique is to exclude subjects with totally or partially missing data. However, this may lead to biased inference if the probability of missingness varies in different subpopulations. A more valid approach is to model the mechanism which produces missingness. The main techniques here are multiple imputation and full likelihood modelling.

In the following, we introduce three categories of missingness. It depends on the type of missingness which approaches to deal with missing data are valid.

Missing data types (cont)

✔ Missingness completely at random (MCAR). The probability of a missing response is independent of both observed and unobserved data. In this case, complete case analysis is valid.

✔ Missingness at random (MAR). The probability of a missing response is independent of unobserved data but may depend on observed data. In this case, it is valid to use the likelihood function of the observed data, and to ignore the missingness mechanism, provided that the parameters of the likelihood are not related to the parameters of the missingness process.

✔ Missingness not at random (MNAR). Missingness of an item may depend on the unobserved missing values. For example, a subject may exit from a clinical trial due to adverse consequences of the treatment.

Literature

✔ Agresti, Alan, *Foundations of linear and generalized linear models*, Wiley, 2015

✔ Faraway, Julian J., *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Chapman& Hall/CRC, 1st Ed 2006, 2nd Ed 2016

✔ Pinheiro, Bates, Mixed-Effects Models in S and S-Plus, Springer, 2000

✔ Stroup, Walter W., *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, Chapman& Hall/CRC, 2013

✔ Zhou, Xiao-Hua, Zhou, Chuan, Liu, Danping, and Ding, Xiaobo, *Applied Missing Data Analysis in the Health Sciences*, Wiley, 2014