NMST432 Advanced Regression Models

Course notes

Michal Kulich

Last modified on May 22, 2020.

Department of Probability and Mathematical Statistics
Faculty of Mathematics and Physics, Charles University

These course notes contain an overview of notation, definitions, theorems and comments covered by the course "NMST432 Advanced regression models", which is a part of the curriculum of the Master's program "Probability, Mathematical Statistics and Econometrics".

This document undergoes continuing development. The author will appreciate notifications by the reader of potential typos or misprints.

Michal Kulich
kulich@karlin.mff.cuni.cz

In Karlín on May 22, 2020

Contents

1 Review of Linear Regression

1.1 Definition and Assumptions

Consider n independent copies of random vectors $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. Each $\boldsymbol{X}_i$ has $p < n$ components $(X_{i1}, \ldots, X_{ip})$.

Note.

- Y_i is called *the response*[*]. The components of $\boldsymbol{X}_i$ are called *covariates* (explanatory variables, predictors, regressors)[†].
- The covariate X_{i1} is usually taken as 1.
- In certain applications, the covariates can be fixed quantities rather than random variables. Throughout this course, we will consider covariates random. Extensions to fixed covariates usually hold with some additional conditions but the proofs require more effort.

Notation. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$ and

$$\mathbb{X} = \begin{pmatrix} \boldsymbol{X}_1^\mathsf{T} \\ \boldsymbol{X}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_n^\mathsf{T} \end{pmatrix}.$$

The n by p matrix $\mathbb{X}$ is called *the regression matrix*[‡]. We assume $r(\mathbb{X}) = p$ (full rank).

Definition 1.1. The data $(Y_i, \boldsymbol{X}_i)$ satisfy the linear regression model if

$$Y_i = \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta}_0 + \varepsilon_i,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent, $\mathsf{E}\big[\varepsilon_i \,\big|\, \boldsymbol{X}_i\big] = 0$, and $\mathsf{var}\big[\varepsilon_i \,\big|\, \boldsymbol{X}_i\big] = \sigma^2$.

Note. $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots \beta_{0p})^\mathsf{T}$ is a vector of unknown parameters (*regression coefficients*[§]), random variables ε_i are called *error terms (disturbances)*[¶], σ^2 is called *residual variance*[‖].

Note. Another convenient formulation of the model is based on conditional moments (avoids the introduction of error terms). The linear regression model holds if and only if

[*] Česky *odezva* [†] Česky *regresory, nezávisle proměnné, vysvětlující veličiny, prediktory, kovariáty* [‡] Česky *regresní matice* [§] Česky *regresní koeficienty* [¶] Česky *chybové členy* [‖] Česky *residuální rozptyl*

- $Y_1, \dots, Y_n$ are independent
- $\mathsf{E}\big[Y_i \,\big|\boldsymbol{X}_i\big] = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$
- $\mathsf{var}\big[Y_i \,\big|\boldsymbol{X}_i\big] = \sigma^2$

The model specifies the first two conditional moments of Y_i given $\boldsymbol{X}_i$.

Note. We will use the notation E, var for conditional moments given $\boldsymbol{X}_i$. The symbol E_X will be used for unconditional expectation over the distribution of $\boldsymbol{X}_i$.

Note. The regression parameters express the influence of $\boldsymbol{X}_i$ on $\mathsf{E}\,Y_i$. Assuming that $X_{i1} = 1$,

$$\beta_{01} = \mathsf{E}\big[Y_i \,\big|X_{i2} = 0, X_{i3} = 0, \dots, X_{ip} = 0\big]$$

and

$$\beta_{02} = \mathsf{E}\big[Y_i \,\big|X_{i2} = x+1, X_{i3} = x_3, \dots, X_{ip} = x_p\big] - \mathsf{E}\big[Y_i \,\big|X_{i2} = x, X_{i3} = x_3, \dots, X_{ip} = x_p\big]$$

1.2 Estimation

The regression coefficients $\boldsymbol{\beta}_0$ are estimated by *the least squares estimator* (LSE) $\widehat{\boldsymbol{\beta}}$ that minimizes the sum of squares

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^n (Y_i - \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta})^2 = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} (\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta}),$$

i.e., solves the system of *normal equations*

$$\sum_{i=1}^n \boldsymbol{X}_i(Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}) = \mathbf{0}.$$

Because $\mathbb{X}$ is of full rank, the single solution to the system is

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y}.$$

Note.

- $\mathsf{E}\,\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0$ (unbiased), $\mathsf{var}\,\widehat{\boldsymbol{\beta}} = \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$.
- The vector
$$\widehat{\boldsymbol{Y}} = \mathbb{X}\widehat{\boldsymbol{\beta}} = \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\boldsymbol{Y} = \mathbb{H}\boldsymbol{Y}$$
is called *the vector of fitted values*[*].
- The projection matrix $\mathbb{H} \overset{\mathsf{df}}{=} \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ is idempotent, with rank p. It satisfies $\mathbb{H}\mathbb{X} = \mathbb{X}$. The matrix $\mathbb{I}_n - \mathbb{H}$ is also idempotent with rank $n - p$, and satisfies $(\mathbb{I}_n - \mathbb{H})\mathbb{X} = \mathbf{0}$.

[*] Česky *vektor odhadnutých (vyrovnaných) hodnot*

- $\mathsf{E}\,\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{\beta}_0$, $\mathsf{var}\,\widehat{\boldsymbol{Y}} = \sigma^2\mathbb{H}$.
- The random vector $\boldsymbol{u} \overset{\mathsf{df}}{=} \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\mathbb{I}_n - \mathbb{H})\boldsymbol{Y}$ is called *the vector of residuals**. It satisfies $\mathsf{E}\,\boldsymbol{u} = \boldsymbol{0}$, $\mathsf{var}\,\boldsymbol{u} = \sigma^2(\mathbb{I}_n - \mathbb{H})$.
- The random variable

$$SS_e \overset{\mathsf{df}}{=} \boldsymbol{u}^\mathsf{T}\boldsymbol{u} = \sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}})^2 = \boldsymbol{Y}^\mathsf{T}(\mathbb{I}_n - \mathbb{H})\boldsymbol{Y}$$

  is called the *residual sum of squares*†. Because $\mathsf{E}\,SS_e = (n-p)\sigma^2$, we obtain an unbiased estimator of residual variance as $\widehat{\sigma}^2 = SS_e/(n-p)$.

1.3 Normal Linear Regression

For normally distributed errors, additional useful properties can be derived. Assume now that $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2\mathbb{I}_n)$.

Proposition 1.1. *Under normality,*

(i) $\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}_0, \sigma^2\mathbb{I}_n)$

(ii) $\widehat{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}_0, \sigma^2(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1})$

(iii) $\widehat{\boldsymbol{Y}} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}_0, \sigma^2\mathbb{H})$

(iv) $\boldsymbol{u} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma^2(\mathbb{I}_n - \mathbb{H}))$

(v) $SS_e/\sigma^2 \sim \chi^2_{n-p}$

(vi) $\widehat{\boldsymbol{\beta}}$ *and* SS_e *are independent*

(vii) Let $\boldsymbol{c}$ *be any non-zero* p*-vector of real constants. Then*

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta_0}}{\sqrt{\widehat{\sigma}^2\boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

(viii) Assume the model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$*, where* $\mathbb{X} = (\mathbb{X}_A|\mathbb{X}_B)$ *and* $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_A^\mathsf{T}, \boldsymbol{\beta}_B^\mathsf{T})^\mathsf{T}$*,* $\boldsymbol{\beta}_B \in \mathbb{R}_m$*,* $\boldsymbol{\beta}_A \in \mathbb{R}^{p-m}$*, and introduce the submodel* $\boldsymbol{Y} = \mathbb{X}_A\boldsymbol{\beta}_A + \boldsymbol{\varepsilon}'$*. Let* SS_e *and* SS_h *be the residual sums of squares in the model and submodel, respectively. If the submodel is true (*$H_0 : \boldsymbol{\beta}_B = 0$ *holds) then*

$$F = \frac{n-p}{m}\,\frac{SS_h - SS_e}{SS_e} \sim F_{m,n-p}. \tag{1.1}$$

It can be also shown that, under normality, $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator and the maximum likelihood estimator, so it possesses optimality properties.

* Česky *vektor residuí* † Česky *residuální součet čtverců*

1.4 Asymptotic Properties of the LSE

Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, be iid. Assume Definition 1.1 (without normality). Denote $\mathbb{D}_X = \mathsf{E}_X \boldsymbol{X_i} \boldsymbol{X}_i^\mathsf{T}$.

Proposition 1.2. *Let $\mathbb{D}_X$ be a finite regular matrix. Then*

(i) $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$ *as* $n \to \infty$,

(ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \sigma^2 \mathbb{D}_X^{-1})$ *as* $n \to \infty$.

Proposition 1.2(ii) is an asymptotic restatement of Proposition 1.1(ii). Other parts of Proposition 1.1 also hold asymptotically even if the data are not normal.

Now relax the assumption of equal variance: assume only $\mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \boldsymbol{X}_i^\mathsf{T} \boldsymbol{\beta}_0$. Let $\mathsf{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma^2(\boldsymbol{X}_i)$ be stochastically bounded (finite expectation follows). Denote $\mathbb{V}_X = \mathsf{E}_X \sigma^2(\boldsymbol{X}_i) \boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T}$.

Proposition 1.3. *Let $\mathbb{V}_X$ be finite and $\mathbb{D}_X$ be finite and regular. Then*

(i) $\widehat{\boldsymbol{\beta}} \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$ *as* $n \to \infty$,

(ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \mathbb{D}_X^{-1} \mathbb{V}_X \mathbb{D}_X^{-1})$ *as* $n \to \infty$.

When equal variances hold, $\mathbb{V}_X = \sigma^2 \mathbb{D}_X$ and the result in Proposition 1.3(ii) transforms into the result in Proposition 1.2(ii).

Consistent estimates of $\mathbb{D}_X$ and $\mathbb{V}_X$ are

$$\widehat{\mathbb{D}}_n = \frac{1}{n} \mathbb{X}^\mathsf{T} \mathbb{X}$$

and

$$\widehat{\mathbb{V}}_n = \frac{1}{n} \mathbb{X}^\mathsf{T} \mathrm{diag}\,(u_i^2) \mathbb{X}.$$

So, if both normality and homoskedasticity are in doubt, one can use the OLS estimator $\widehat{\boldsymbol{\beta}}$ with variance

$$(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1} \mathbb{X}^\mathsf{T} \mathrm{diag}\,(u_i^2) \mathbb{X}\, (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$$

in place of the usual $\widehat{\sigma}^2 (\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$. This is called *the sandwich estimator**, or, in the econometric context, *White estimator*† (White 1980).

Many variants and improvements of this estimator have been proposed in the literature.

* Česky *sendvičový odhad* † Česky *Whiteův odhad*

1.5 Transformed Response

Recall the linear model $\mathsf{E}\big[Y_i \,\big|\boldsymbol{X}_i\big] = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$ with $\mathsf{var}\big[Y_i \,\big|\boldsymbol{X}_i\big] = \sigma^2$. The regression parameters can be interpreted as

$$\beta_{01} = \mathsf{E}\big[Y_i \,\big|X_{i2} = 0, X_{i3} = 0, \ldots, X_{ip} = 0\big]$$

and

$$\beta_{02} = \mathsf{E}\big[Y_i \,\big|X_{i2} = x+1, X_{i3} = x_3, \ldots, X_{ip} = x_p\big] - \mathsf{E}\big[Y_i \,\big|X_{i2} = x, X_{i3} = x_3, \ldots, X_{ip} = x_p\big].$$

When the response is non-normal, the common practice is to specify a linear model on a transformed response. Let g be some monotone function. The transformed model is

$$g(Y_i) = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0 + \varepsilon_i$$

or $\mathsf{E}\big[g(Y_i) \,\big|\boldsymbol{X}_i\big] = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$ with $\mathsf{var}\big[g(Y_i) \,\big|\boldsymbol{X}_i\big] = \sigma^2$. The induced model for Y_i is

$$Y_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0 + \varepsilon_i).$$

In general, the effect of the covariates on EY_i in this model cannot be expressed.

The only special case (apart from linear g) when the transformed model says anything useful about $\mathsf{E}\big[Y_i \,\big|\boldsymbol{X}_i\big]$ is the log transform. From

$$\log Y_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0 + \varepsilon_i$$

we get a multiplicative model

$$Y_i = \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0}\varepsilon_i^*,$$

where $\varepsilon_i^* = \mathrm{e}^{\varepsilon}_i$, $\mathsf{E}\,\varepsilon_i^* = \mu_\varepsilon > 1$, $\mathsf{var}\,\varepsilon_i^* = \sigma_\varepsilon^2$. Then

$$\begin{aligned}\mathsf{E}\big[Y_i \,\big|\boldsymbol{X}_i\big] &= \exp\{\log\mu_\varepsilon + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0\},\\ \mathsf{var}\big[Y_i \,\big|\boldsymbol{X}_i\big] &= \sigma_\varepsilon^2(\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0\})^2.\end{aligned}$$

While β_{01} (the intercept) does not have useful interpretation, the other parameters express multiplicative effects of $X_{i2}, \ldots, X_{ip}$ on $\mathsf{E}\,Y_i$. E.g.,

$$\mathrm{e}^{\beta_{02}} = \frac{\mathsf{E}\big[Y_i \,\big|X_{i2} = x+1, X_{i3} = x_3, \ldots, X_{ip} = x_p\big]}{\mathsf{E}\big[Y_i \,\big|X_{i2} = x, X_{i3} = x_3, \ldots, X_{ip} = x_p\big]}.$$

So, $\mathrm{e}^{\beta_{0k}}$ is the proportional increase (relative change) in $\mathsf{E}\,Y_i$ after a unit change in X_{ik}.

The problem with the interpretation of the transformed linear model is serious when the primary task is to estimate the effect of $\boldsymbol{X}_i$ on $\mathsf{E}\,Y_i$. If the goal is to predict Y_i from $\boldsymbol{X}_i$, transformations can still be useful even if the interpretation of the parameters is lost.

2 Generalized Linear Model – Theory

The generalized linear model extends the normal linear model in two ways:

- allows a wider choice of distributions for Y_i (distributions from exponential family);
- allows a more general relationship between $\mathsf{E}\, Y_i$ and $\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$.

2.1 Exponential Family

2.1.1 Parametrization, moments

Definition 2.1. If the density of a real random variable (w.r.t. some σ-finite measure μ) can be written in the form

$$f(x;\theta,\varphi) = \exp\left\{\frac{x\theta - b(\theta)}{\varphi} + c(x,\varphi)\right\}, \tag{2.1}$$

where

- θ is called *the canonical parameter*[*];
- $\varphi \in (0,\infty)$ is called *the dispersion parameter*[†];
- b and c are some real functions;

then the distribution belongs to *the exponential family* of distributions[‡]. The expression (2.1) is called *the canonical form of the density*[§].

Examples

Normal distribution $Y \sim \mathsf{N}(\mu,\sigma^2)$

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \qquad \mu\in\mathbb{R},\ \sigma^2>0,\ x\in\mathbb{R} \\ &= \exp\left\{\frac{x\mu-\mu^2/2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\ \theta=\mu,\quad \varphi=\sigma^2,\quad b(\theta)=\frac{\theta^2}{2},\quad c(x,\varphi) &= -\frac{x^2}{2\varphi} - \frac{1}{2}\log(2\pi\varphi) \end{aligned}$$

[*] Česky *kanonický parametr* [†] Česky *disperzní parametr* [‡] Česky *rozdělení exponenciálního typu*
[§] Česky *kanonický tvar hustoty*

Gamma distribution $Y \sim \Gamma(a, p)$

$$\begin{aligned} f(x; a, p) &= \frac{a^p}{\Gamma(p)} x^{p-1} \exp\{-ax\}, \qquad a, p > 0,\ x > 0 \\ &= \exp\Big\{ \frac{-(a/p)x + \log(a/p)}{1/p} + (p-1)\log x + p \log p - \log\Gamma(p) \Big\} \\ \theta = -\frac{a}{p}, &\quad \varphi = 1/p, \quad b(\theta) = -\log(-\theta) \\ c(x, \varphi) &= (1/\varphi - 1)\log x - (\log\varphi)/\varphi - \log\Gamma(1/\varphi) \end{aligned}$$

Inverse Gaussian distribution $Y \sim \mathsf{IG}(\mu, \lambda)$

$$\begin{aligned} f(x; \mu, \lambda) &= \sqrt{\frac{\lambda}{2\pi x^3}} \exp\Big\{ -\frac{\lambda(x-\mu)^2}{2\mu^2 x} \Big\}, \qquad \mu, \lambda > 0,\ x > 0 \\ &= \exp\Big\{ \frac{-x/(2\mu^2) + 1/\mu}{1/\lambda} + \frac{1}{2}\log\frac{\lambda}{2\pi x^3} - \frac{\lambda}{2x} \Big\} \\ \theta = -\frac{1}{2\mu^2}, &\quad \varphi = 1/\lambda, \quad b(\theta) = -\sqrt{-2\theta}, \quad c(x, \varphi) = -\frac{1}{2}\log(2\pi x^3 \varphi) - (2x\varphi)^{-1} \end{aligned}$$

Poisson distribution $Y \sim \mathsf{Po}(\lambda)$

$$\begin{aligned} f(x; \lambda) &= \frac{\lambda^x}{x!} \exp\{-\lambda\}, \qquad \lambda > 0,\ x = 0, 1, 2, \ldots \\ &= \exp\big\{ x \log\lambda - \lambda + \log x! \big\} \\ \theta = \log\lambda, &\quad \varphi = 1, \quad b(\theta) = \exp(\theta), \quad c(x, \varphi) = \log x! \end{aligned}$$

Alternative distribution $Y \sim \mathsf{Alt}(p)$

$$\begin{aligned} f(x; p) &= p^x (1-p)^{1-x}, \qquad p \in (0, 1),\ x = 0, 1 \\ &= \exp\Big\{ x \log\frac{p}{1-p} + \log(1-p) \Big\} \\ \theta = \log\frac{p}{1-p}, &\quad \varphi = 1, \quad b(\theta) = \log(1 + \exp\{\theta\}), \quad c(x, \varphi) = 0 \end{aligned}$$

Lemma 2.1. *Let the random variable Y follow a distribution from the exponential family. Then the moment generation function $m_Y(t) \equiv \mathsf{E}\, e^{tY}$ of Y exists, is finite, and is equal to*

$$m_Y(t) = \exp\Big\{ \frac{b(t\varphi + \theta) - b(\theta)}{\varphi} \Big\}.$$

If $b(\theta)$ is twice continuously differentiable, $m_Y(t)$ is twice differentiable at $t = 0$.

Corollary. If $b(\theta)$ is twice continuously differentiable then Y has finite first two moments, $\mathsf{E}\, Y = b'(\theta)$ and $\mathsf{var}\, Y = \varphi\, b''(\theta)$.

We will always assume that $b(\theta)$ is twice continuously differentiable so that $\mathsf{var}\, Y$ is finite. Let $\mu \stackrel{\mathsf{df}}{=} \mathsf{E}\, Y$.

Note. Since $\mathsf{var}\, Y = \varphi\, b''(\theta) > 0$, b is a strictly convex function and b' is strictly increasing. Hence b' has a well-defined inverse. There exists a function $V(\mu)$ of the mean μ such that $\mathsf{var}\, Y = \varphi V(\mu)$. It satisfies the equation $b''(\theta) = V(b'(\theta))$.

Definition 2.2. The function $V(\mu)$ is called *the variance function**.

Note.

- Each distribution belonging to the exponential family has a different variance function.
- Within the exponential family, the variance function determines the distribution of Y. Not every function V is a variance function of some distribution from the exponential family.

Examples

Normal distribution $Y \sim \mathsf{N}(\mu, \sigma^2)$

$$\theta = \mu, \quad \varphi = \sigma^2, \quad b(\theta) = \frac{\theta^2}{2}$$
$$\mathsf{E}\, Y = b'(\theta) = \mu, \quad \mathsf{var}\, Y = \varphi\, b''(\theta) = \varphi = \sigma^2, \quad V(\mu) = 1$$

This is the only distribution in exponential family with constant variance function, i.e., the variance is unrelated to the mean.

Gamma distribution $Y \sim \Gamma(a, p)$

$$\theta = -\frac{a}{p}, \quad \varphi = 1/p, \quad b(\theta) = -\log(-\theta)$$
$$\mu = \mathsf{E}\, Y = b'(\theta) = -1/\theta = p/a, \quad \mathsf{var}\, Y = \varphi\, b''(\theta) = \varphi/\theta^2 = p/a^2, \quad V(\mu) = \mu^2$$

Inverse Gaussian distribution $Y \sim \mathsf{IG}(\mu, \lambda)$

$$\theta = -\frac{1}{2\mu^2}, \quad \varphi = 1/\lambda, \quad b(\theta) = -\sqrt{-2\theta}$$
$$\mathsf{E}\, Y = b'(\theta) = 1/\sqrt{-2\theta} = \mu, \quad \mathsf{var}\, Y = \varphi\, b''(\theta) = \varphi(-2\theta)^{-3/2} = \mu^3/\lambda, \quad V(\mu) = \mu^3$$

* Česky *rozptylová funkce*

Poisson distribution $Y \sim \mathsf{Po}(\lambda)$

$\theta = \log \lambda, \quad \varphi = 1, \quad b(\theta) = \exp(\theta)$
$\mu = \mathsf{E}\, Y = b'(\theta) = \exp(\theta) = \lambda, \quad \mathsf{var}\, Y = \varphi\, b''(\theta) = \exp(\theta) = \lambda, \quad V(\mu) = \mu$

Alternative distribution $Y \sim \mathsf{Alt}(p)$

$\theta = \log \dfrac{p}{1-p}, \quad \varphi = 1, \quad b(\theta) = \log(1 + \exp\{\theta\}) = \log(1-p)$
$\mu = \mathsf{E}\, Y = b'(\theta) = \dfrac{\mathrm{e}^\theta}{1+\mathrm{e}^\theta} = p, \quad \mathsf{var}\, Y = \varphi\, b''(\theta) = \dfrac{\mathrm{e}^\theta}{(1+\mathrm{e}^\theta)^2} = p(1-p),$
$V(\mu) = \mu(1-\mu)$

2.1.2 Maximum likelihood estimator of the canonical parameter

Let $Y_1, \ldots, Y_n$ be a random sample from the density $f(x; \theta_0, \varphi_0)$ belonging to the exponential family, θ_0 is the true canonical parameter, φ_0 is the true dispersion parameter. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$.

The likelihood is

$$L(\theta, \varphi) = \prod_{i=1}^{n} \exp\left\{ \frac{Y_i\theta - b(\theta)}{\varphi} + c(Y_i, \varphi) \right\},$$

The log-likelihood is

$$\ell(\theta, \varphi) = \log L(\theta, \varphi) = \sum_{i=1}^{n} \left[\frac{Y_i\theta - b(\theta)}{\varphi} + c(Y_i, \varphi) \right].$$

Suppose that the true dispersion parameter φ_0 is known. Then the score function for θ is

$$U(\theta \mid Y_i) = \frac{\partial}{\partial\theta} \log f(x; \theta, \varphi_0) = \frac{1}{\varphi_0} \big[Y_i - b'(\theta) \big].$$

Obviously, $\mathsf{E}\, U(\theta_0 \mid Y_i) = 0$. The score statistic is

$$U_n(\theta \mid \boldsymbol{Y}) = \sum_{i=1}^{n} U(\theta \mid Y_i) = \frac{1}{\varphi_0} \sum_{i=1}^{n} \big[Y_i - b'(\theta) \big].$$

The maximum likelihood estimator [MLE] $\widehat{\theta}_n$ solves the equation $U_n(\widehat{\theta}_n \mid \boldsymbol{Y}) = 0$, that is $\sum_{i=1}^n Y_i = n b'(\widehat{\theta}_n)$. The solution is $\widehat{\theta}_n = (b')^{-1}(\overline{Y}_n)$, where $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$.

The MLE is unique because b is convex, and it does not depend on the dispersion parameter φ_0. It can be calculated even if φ_0 is unknown.

The observed information is

$$I_n(\theta \mid \boldsymbol{Y}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial U(\theta \mid Y_i)}{\partial \theta} = \frac{1}{\varphi_0}b''(\theta) > 0,$$

so the likelihood is strictly concave. The expected (Fisher) information is the same as the observed information,

$$I(\theta) = -\mathsf{E}\,\frac{\partial U(\theta \mid Y_i)}{\partial \theta} = \frac{1}{\varphi_0}b''(\theta).$$

It is easy to check that

$$\mathsf{var}\, U(\theta_0 \mid Y_i) = I(\theta_0).$$

It follows from the theory of MLE that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \varphi_0[b''(\theta_0)]^{-1}). \tag{2.2}$$

Now consider the true dispersion parameter φ_0 unknown. The MLE of θ_0 is still the same, $\widehat{\theta}_n = (b')^{-1}(\overline{Y}_n)$. However, what is the asymptotic distribution of $\widehat{\theta}_n$ when φ_0 is unknown? In general, the asymptotic variance may change.

Calculate the joint information matrix for (θ, φ):

$$I(\theta_0, \varphi_0) = -\mathsf{E}\,\frac{\partial^2 \log f(x; \theta_0, \varphi_0)}{\partial(\theta,\varphi)\partial(\theta,\varphi)^{\mathsf{T}}} = \begin{pmatrix} I_{\theta\theta} & I_{\theta\varphi} \\ I_{\theta\varphi} & I_{\varphi\varphi} \end{pmatrix} = \begin{pmatrix} b''(\theta_0)/\varphi_0 & 0 \\ 0 & I_{\varphi\varphi} \end{pmatrix}.$$

Thus, the information matrix is diagonal. It follows that the asymptotic distribution of $\widehat{\theta}_n$ is given by (2.2) even if φ_0 is unknown.

We do not need φ_0 to estimate θ_0 but we need an estimate of φ_0 to estimate the asymptotic variance of θ_0. Of course, we could use the MLE of φ_0 but it often cannot be calculated explicitely. Instead, we can use the moment estimator

$$\widehat{\varphi} = \frac{S_n^2}{b''(\widehat{\theta}_n)} = \frac{S_n^2}{V(\overline{Y}_n)},$$

where S_n^2 is the sample variance. Since $S_n^2 \xrightarrow{\mathsf{P}} \mathsf{var}\, Y_i = \varphi_0 b''(\theta_0)$, $\widehat{\theta}_n$ is consistent and b'' is continuous, $\widehat{\varphi}$ is consistent (though less efficient than the MLE).

2.2 Definition of the Generalized Linear Model

Consider n independent copies of random vectors $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^{\mathsf{T}}$. We want to express the dependence of $\mu_i \overset{\mathsf{df}}{=} \mathsf{E}\big[Y_i \,\big|\, \boldsymbol{X}_i\big]$ on $\boldsymbol{X}_i$ by a model that is more general than the linear model.

Definition 2.3. (Nelder and Wedderburn 1972) The data $(Y_i, \boldsymbol{X}_i)$ satisfy the *generalized linear model*[*] [GLM] if

1. $Y_1, \ldots, Y_n$ are independent and the distribution of Y_i depends on $\boldsymbol{X}_i$ through regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$.
2. the conditional density of Y_i given $\boldsymbol{X}_i$ has the form
$$f(y; \theta_i, \varphi) = \exp\left\{\frac{y\theta_i - b(\theta_i)}{\varphi} + c(y, \varphi)\right\},$$
(is of exponential type), where $b(\cdot)$ is a known twice continuously differentiable function, θ_i depends on $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$, $\varphi > 0$ is a known or an unknown constant.
3. θ_i depends on $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ through the *linear predictor*[†] $\eta_i \overset{\mathsf{df}}{=} \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.
4. There exists a known strictly monotone, twice continuously differentiable *link function*[‡] g such that $g(\mu_i) = \eta_i$.

Notation. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$ and define the regression matrix
$$\mathbb{X}_{n\times p} = \begin{pmatrix} \boldsymbol{X}_1^\mathsf{T} \\ \boldsymbol{X}_2^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_n^\mathsf{T} \end{pmatrix}.$$

We assume $r(\mathbb{X}) = p$. We sometimes use the notation $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})^\mathsf{T}$ to denote the true regression parameter (but the notation $\boldsymbol{\beta}$ can also mean the true parameter).

Note. The (conditional) means of $Y_1, \ldots, Y_n$ vary because the canonical parameters $\theta_1, \ldots, \theta_n$ depend on $\boldsymbol{X}_i$. The dispersion parameter φ is the same for all observations, it must not depend on $\boldsymbol{X}_i$ (recall homoscedasticity in linear regression). However, the variances of $Y_1, \ldots, Y_n$ depend on the mean through the variance function $V(\mu_i)$, and hence vary with $\boldsymbol{X}_i$.

Note. The link function postulates a possibly non-linear relationship between the expectation of the response μ_i and the linear predictor $\eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$. It has to be specified in advance. There are methods to verify the choice of the link function for a specific data set (see Chapter XX). It is enough to specify the link function up to a non-zero proportionality constant (if $c \neq 0$, g and cg lead to the same model).

Definition 2.4. The link function g is called *the canonical link*[§] for the distribution f if it equates the linear predictor η_i with the canonical parameter θ_i.

[*] Česky *zobecněný lineární model* [†] Česky *lineární prediktor* [‡] Česky *linková funkce* [§] Česky *kanonický link*

Lemma 2.2. *(Properties of canonical link)*

(i) The canonical link is equal to the inverse of b', that is $g(\mu_i) = (b')^{-1}(\mu_i)$.
(ii) The canonical link satisfies the equation $g'(\mu_i) = 1/V(\mu_i)$.

Note. For each distribution f from the exponential family, there is a unique (up to a non-zero proportionality constant) and specific canonical link function. Canonical link functions have certain nice properties that will become apparent later on.

Examples

Normal distribution $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$

$$\theta_i = \mu_i, \quad \varphi = \sigma^2, \quad b(\theta_i) = \frac{\theta_i^2}{2}, \; \mu_i = b'(\theta_i) = \theta_i, \quad \mathsf{var}\, Y_i = \sigma^2, \quad V(\mu) = 1.$$

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) = \mu_i$ (identity link).
This implies $\mathsf{E}\, Y_i = \eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$. This is the normal linear model.

Gamma distribution $Y_i \sim \Gamma(a_i, p)$

$$\theta_i = -\frac{a_i}{p}, \quad \varphi = 1/p, \quad b(\theta_i) = -\log(-\theta_i), \quad \mu_i = b'(\theta_i) = -1/\theta_i, \quad \mathsf{var}\, Y_i = \varphi\mu_i^2.$$

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) \propto 1/\mu_i$ (inverse link*).
This implies $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = 1/\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.

Inverse Gaussian distribution $Y_i \sim \mathsf{IG}(\mu_i, \lambda)$

$$\theta_i = -\frac{1}{2\mu_i^2}, \quad \varphi = 1/\lambda, \quad b(\theta_i) = -\sqrt{-2\theta_i}, \quad \mu_i = b'(\theta_i) = 1/\sqrt{-2\theta_i}, \quad \mathsf{var}\, Y_i = \varphi\mu_i^3.$$

The canonical link is† $g(\mu_i) = (b')^{-1}(\mu_i) \propto 1/\mu_i^2$.
This implies $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = 1/\sqrt{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}}$.

Poisson distribution $Y_i \sim \mathsf{Po}(\lambda_i)$

$$\theta_i = \log\lambda_i, \quad b(\theta_i) = \exp(\theta_i), \quad \mu_i = b'(\theta_i) = \exp(\theta_i), \quad \mathsf{var}\, Y_i = \mu_i.$$

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) = \log\mu_i$.
This implies $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = \exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}$. This is the loglinear model.

---

* we drop the minus sign † we drop the constant -2

Alternative distribution $Y_i \sim \mathsf{Alt}(p_i)$

$$\theta_i = \log\frac{p_i}{1-p_i}, \quad b(\theta_i) = \log(1+\exp\{\theta_i\}), \quad \mu_i = b'(\theta_i) = \frac{\mathrm{e}^{\theta_i}}{1+\mathrm{e}^{\theta_i}}, \quad \mathsf{var}\, Y_i = \mu_i(1-\mu_i).$$

The canonical link is $g(\mu_i) = \log\dfrac{\mu_i}{1-\mu_i}$ (the logistic link).

This implies $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = \dfrac{\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}$. This is the logistic regression model.

Parametrizations of the GLM

The primary parameteres in the GLM are the regression coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$. However, we are also interested in parametrizing the distributions of the individual Y_i's that depend on both the primary parameters $\boldsymbol{\beta}$ and the covariates $\boldsymbol{X}_i$. This can be done in three ways:

- by the *linear predictors* $\eta_1, \ldots, \eta_n$;
- by the *means* $\mu_1 \equiv \mathsf{E}\, Y_1, \ldots, \mu_n \equiv \mathsf{E}\, Y_n$;
- by the *canonical parameters* $\theta_1, \ldots, \theta_n$.

The parametrizations are related to each other as follows:

- $\eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$;
- $\eta_i = g(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$;
- $\mu_i = b'(\theta_i)$, $\theta_i = (b')^{-1}(\mu_i)$;
- $\eta_i = g\big(b'(\theta_i)\big)$, $\theta_i = (b')^{-1}\big(g^{-1}(\eta_i)\big)$; if the link g is canonical then $\eta_i = \theta_i$.

The likelihood function

Let the true dispersion parameter φ_0 be known. The likelihood function for $\boldsymbol{\beta}$ has the form

$$L(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \exp\left\{\frac{Y_i\theta_i - b(\theta_i)}{\varphi_0} + c(Y_i, \varphi_0)\right\},$$

where $\theta_i = (b')^{-1}\big(g^{-1}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta})\big)$.

The log-likelihood is

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n}\left[\frac{Y_i\theta_i - b(\theta_i)}{\varphi_0} + c(Y_i, \varphi_0)\right]. \tag{2.3}$$

The saturated model

Suppose at least one covariate is continuous and consider a model which has the largest possible number of parameters* $p = n$. This is called *the saturated model*†. In the saturated model, each Y_i gets its own canonical parameter θ_i, which is unrelated to the canonical parameters of the other observations. Maximizing $L(\boldsymbol{\beta} \mid \boldsymbol{Y})$ w.r.t all $\boldsymbol{\beta} \in \mathbb{R}^n$ is the same as maximizing $L(\boldsymbol{\theta} \mid \boldsymbol{Y})$ w.r.t all $\boldsymbol{\theta} \in \mathbb{R}^n$. To obtain the MLE in the saturated model, we differentiate (2.3) w.r.t. each θ_i separately and we get n equations

$$\varphi_0^{-1}[Y_i - b'(\theta_i)] = 0, \quad i = 1, \ldots, n.$$

The MLE of μ_i under the saturated model is

$$\widehat{\mu}_i = Y_i.$$

The fitted values $\widehat{\mu}_i \equiv \widehat{Y}_i$ are equal to the observed values Y_i. This model provides a "perfect fit". However, a "perfect fit" of this kind is rarely useful.

The saturated model with $p = n$ does not satisfy the regularity assumptions of the MLE theory (the number of parameters must be constant for the theory to apply; here $p \to \infty$ as $n \to \infty$). The estimates obtained from this model are not even consistent.

The null model

The null model‡ is the opposite extreme. It assumes $p = 1$ and $\boldsymbol{X}_i = 1$ so that the model includes only the intercept and all Y_i are equally distributed.

The MLE of the common canonical parameter θ of the null model is derived in Section 2.1.2. Using $\beta_0 = \eta = g\big(b'(\theta)\big)$, we get the MLE of β_0 as $\widehat{\beta}_n = g(b'(\widehat{\theta}_n)) = g(\overline{Y}_n)$. From the central limit theorem for iid random variables and the delta method,

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) \xrightarrow{\mathsf{D}} \mathsf{N}(0, \varphi_0 V(\mu_0)[g'(\mu_0)]^2),$$

where $\mu_0 = \mathsf{E}\, Y_i$ (compare this with (2.2)).

Neither the null model nor the saturated model are particularly interesting. We aim to build a model which has more structure than the null model, fewer parameters than the saturated model, and fits the observed data well.

2.3 Maximum Likelihood Estimation in the GLM

Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ be iid random vectors of dimension $p + 1$. Let $h_i(\boldsymbol{x})$ be the marginal density of $\boldsymbol{X}_i$ (with no assumptions about it except finite second moments).

* When all covariates are discrete, the largest possible number of parameters is p^* — the number of possible distinct values of the covariate vector $\boldsymbol{X}_i$. † Česky *saturovaný model* ‡ Česky *nulový model*

Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, satisfy the generalized linear model (Definition 2.3) with true parameters $\boldsymbol{\beta}_0$ and φ_0. Consider φ_0 known. Write the conditional density of Y given $\boldsymbol{X} = \boldsymbol{x}$ as $f(y \mid \boldsymbol{x}, \boldsymbol{\beta}_0, \varphi_0)$. Then the joint density of $(Y_i, \boldsymbol{X}_i)$ is $f(y \mid \boldsymbol{x}, \boldsymbol{\beta}_0, \varphi_0)h_i(\boldsymbol{x})$, the full likelihood is

$$L^*(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}, \varphi_0)\, h_i(\boldsymbol{X}_i)$$

and the full log-likelihood is

$$\ell^*(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}, \varphi_0) + \sum_{i=1}^{n} \log h_i(\boldsymbol{X}_i).$$

Since the rightmost sum does not depend on $\boldsymbol{\beta}$, it suffices to maximize

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(Y_i \mid \boldsymbol{X}_i, \boldsymbol{\beta}, \varphi_0). \tag{2.4}$$

This is the log-likelihood shown previously in (2.3) (without the detailed derivation and justification needed for the validity of asymptotic results).

When the covariates are random, it is not necessary to consider, know or estimate their distribution. If the covariates were constants, the log-likelihood and the score statistic would be sums of nonidentically distributed terms. Feller-Lindeberg or Lyapunov central limit theorems would have to be applied to validate the asymptotic results, and additional assumptions would have to be imposed on the covariates. The asymptotic results for constant covariates would then turn out to be the same as the results for iid data.

The core term in the log-likelihood (2.4) that we are going to maximize can be written as

$$\sum_{i=1}^{n} \frac{Y_i\theta_i - b(\theta_i)}{\varphi_0}, \tag{2.5}$$

where $g(\mu_i) = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$ and $\mu_i = b'(\theta_i)$. The following theorem summarizes the main results for maximum likelihood estimation of $\boldsymbol{\beta}$.

Theorem 2.3. *(likelihood equations in the GLM; Nelder and Wedderburn 1972) Let the definition of the GLM hold. Denote by $\boldsymbol{\beta}_0$ the true parameter. Let*

$$w(\mu_i) = \frac{1}{V(\mu_i)[g'(\mu_i)]^2} > 0. \tag{2.6}$$

(i) The score function for $\boldsymbol{\beta}$ is

$$\boldsymbol{U}(\boldsymbol{\beta} \mid Y_i) = \varphi_0^{-1} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i,$$

where $\mu_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta})$. It satisfies $\mathsf{E}\,\boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i) = \boldsymbol{0}$.

(ii) The score statistic for $\boldsymbol{\beta}$ is

$$\boldsymbol{U}_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \frac{1}{\varphi_0} \sum_{i=1}^{n} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i.$$

(iii) The maximum likelihood estimator $\widehat{\boldsymbol{\beta}}_n$ solves the system of equations

$$\sum_{i=1}^{n} w(\widehat{\mu}_i) g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i)\boldsymbol{X}_i = \boldsymbol{0}, \tag{2.7}$$

where $\widehat{\mu}_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)$.

(iv) When the link g is canonical then

$$w(\mu_i) = V(\mu_i) = \frac{1}{g'(\mu_i)},$$

the score statistic can be written as

$$\boldsymbol{U}_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \frac{1}{\varphi_0} \sum_{i=1}^{n} (Y_i - \mu_i)\boldsymbol{X}_i,$$

and the likelihood equations are

$$\sum_{i=1}^{n} Y_i \boldsymbol{X}_i = \sum_{i=1}^{n} \widehat{\mu}_i \boldsymbol{X}_i.$$

Note. When the link g is canonical then $\boldsymbol{S} = \sum_{i=1}^{n} Y_i \boldsymbol{X}_i$ is the sufficient statistic and the MLE equates the observed value of $\boldsymbol{S}$ to its expectation under the model (conditional on the covariates).

Definition 2.5. $\widehat{\mu}_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)$ are called *the fitted values**.

The next step is to investigate the observed and expected information matrices for $\boldsymbol{\beta}$. Let $\boldsymbol{a}^{\otimes 2} \stackrel{\mathsf{df}}{=} \boldsymbol{a}\boldsymbol{a}^\mathsf{T}$.

Theorem 2.4. *(on information matrices in the GLM) Let the definition of the GLM hold. Let $\mathsf{E}_{\boldsymbol{X}} w(\mu_i)\boldsymbol{X}_i^{\otimes 2}$ be finite and of full rank.*

(i) The contribution of the i-th observation to the observed information matrix is

$$I(\boldsymbol{\beta} \mid Y_i) = \frac{1}{\varphi_0}\big[w(\mu_i)\boldsymbol{X}_i^{\otimes 2} - \mathbb{J}_i\big],$$

where

$$\mathbb{J}_i = \left[w'(\mu_i) + w(\mu_i)\frac{g''(\mu_i)}{g'(\mu_i)}\right](Y_i - \mu_i)\boldsymbol{X}_i^{\otimes 2}.$$

The observed information matrix is $I_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = n^{-1}\sum_{i=1}^{n} I(\boldsymbol{\beta} \mid Y_i)$.

* Česky *vyrovnané hodnoty*

(ii) *When evaluated at the true* $\boldsymbol{\beta}_0$, $\mathsf{E}\,\mathbb{J}_i = 0$. *The Fisher (expected) information matrix at the true* $\boldsymbol{\beta}_0$ *is*

$$I(\boldsymbol{\beta}_0) = \mathsf{E}\, I(\boldsymbol{\beta}_0 \mid Y_i) = \frac{1}{\varphi_0}\mathsf{E}_{\boldsymbol{X}}\, w(\mu_i)\boldsymbol{X}_i^{\otimes 2}. \tag{2.8}$$

By assumptions, it is finite and of full rank. It holds that $\mathsf{var}\,\boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i) = I(\boldsymbol{\beta}_0)$.

(iii) *When the link* g *is canonical then* $\mathbb{J}_i = 0$ *at any* $\boldsymbol{\beta}$ *for all* i, *the observed information matrix is positive definite at all* $\boldsymbol{\beta}$, *the log-likelihood is concave, the likelihood equations have just one solution and it is the MLE.*

Note. If the link g is not canonical, there is no guarantee that a solution to the likelihood equations is the MLE. The likelihood is not concave, the equations may have multiple solutions. Numerical algorithms for solving the likelihood equations may iterate slowly and converge to the wrong solution.

The Fisher information matrix $I(\boldsymbol{\beta}_0)$ can be consistently estimated by the empirical estimator

$$\widehat{I}_n = \frac{1}{n\varphi_0}\sum_{i=1}^{n} w(\widehat{\mu}_i)\boldsymbol{X}_i^{\otimes 2} = \frac{1}{n\varphi_0}\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X}, \tag{2.9}$$

where $\hat{\mathbb{W}}$ is the $n \times n$ diagonal matrix $\mathrm{diag}\,(w(\widehat{\mu}_1), \ldots, w(\widehat{\mu}_n))$. When φ_0 is unknown it is replaced by a consistent estimator $\widehat{\varphi}_n$, which will be introduced in Section 2.5.

2.4 Algorithm for Fitting the GLM

The parameters of the GLM can be estimated by a numerical algorithm called *iterative weighted least squares** [IWLS].

Theorem 2.5. *(Nelder and Wedderburn 1972) The MLE* $\widehat{\boldsymbol{\beta}}_n$ *in the GLM solves the system of equations*

$$\widehat{\boldsymbol{\beta}}_n = (\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X})^{-1}(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\widehat{\boldsymbol{Z}}),$$

where $\hat{\mathbb{W}} = \mathrm{diag}\,(w(\widehat{\mu}_1), \ldots, w(\widehat{\mu}_n))$, $\widehat{\boldsymbol{Z}}$ *is an* n*-vector with components*

$$\widehat{Z}_i = \widehat{\eta}_i + (Y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i),$$

$\widehat{\mu}_i = g^{-1}(\widehat{\eta}_i)$, *and* $\widehat{\eta}_i = \boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_n$.

Note. $\widehat{\boldsymbol{Z}}$ is called *the adjusted dependent variable*†. Notice that $\widehat{Z}_i$ is the linear approximation to $g(Y_i)$ by Taylor expansion around $\widehat{\mu}_i$. Unlike $g(Y_i)$, it can be calculated even if Y_i is outside of the domain of g.

* Česky *iterativní vážené nejmenší čtverce* † Česky *upravená odezva*

Note. When the link g is canonical then $\hat{\mathbb{W}} = \operatorname{diag}(V(\widehat{\mu}_1), \ldots, V(\widehat{\mu}_n))$ and

$$\widehat{Z}_i = \widehat{\eta}_i + \frac{Y_i - \widehat{\mu}_i}{V(\widehat{\mu}_i)}.$$

One cannot calculate $\widehat{\boldsymbol{\beta}}_n$ directly from Theorem 2.5 because it appears on both the left-hand side as well as the right-hand side. However, the result motivates the following iterative algorithm.

Iterative weighted least squares algorithm

Step 1. Take initial values $\widehat{\mu}_i^{(0)} = Y_i$ (or $Y_i \pm \varepsilon$ if Y_i is not within the domain of g). Set $k := 0$.

Step 2. Calculate $\hat{\mathbb{W}}^{(k)} = \operatorname{diag}(w(\widehat{\mu}_1^{(k)}), \ldots, w(\widehat{\mu}_n^{(k)}))$ and $\widehat{Z}_i^{(k)} = g(\widehat{\mu}_i^{(k)}) + (Y_i - \widehat{\mu}_i^{(k)}) g'(\widehat{\mu}_i^{(k)})$.

Step 3. Take

$$\widehat{\boldsymbol{\beta}}_n^{(k+1)} = (\mathbb{X}^\mathsf{T} \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1} (\mathbb{X}^\mathsf{T} \hat{\mathbb{W}}^{(k)} \widehat{\boldsymbol{Z}}^{(k)}).$$

Step 4. Calculate $\widehat{\mu}_i^{(k+1)} = g^{-1}(\boldsymbol{X}_i^\mathsf{T} \widehat{\boldsymbol{\beta}}_n^{(k+1)})$.

Step 5. Set $k := k + 1$.

Iterate steps 2–5 until convergence, for example until $\left\| \widehat{\boldsymbol{\beta}}_n^{(k)} - \widehat{\boldsymbol{\beta}}_n^{(k-1)} \right\| < \delta$, where δ is a pre-specified tolerance parameter. If the model is well formulated, the algorithm usually converges in 5–7 steps.

Note.

- The IWLS algorithm is a special case of the Fisher scoring algorithm.
- According to (2.9), the matrix $(\mathbb{X}^\mathsf{T} \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1}$ estimates (up to a proportionality constant) the inverse information matrix. Thus, an estimate of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ is obtained by the IWLS as well (just make sure to update it after the last iteration of $\widehat{\boldsymbol{\beta}}_n^{(k)}$).
- Let $\mathbb{X}^* = \hat{\mathbb{W}}^{1/2} \mathbb{X}$ and $\boldsymbol{Y}^* = \hat{\mathbb{W}}^{1/2} \widehat{\boldsymbol{Z}}$. Then $\widehat{\boldsymbol{\beta}}_n$ can be written as an ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_n = (\mathbb{X}^{*\mathsf{T}} \mathbb{X}^*)^{-1} \mathbb{X}^{*\mathsf{T}} \boldsymbol{Y}^*$. This is useful for extending the diagnostic methods available for the linear model to the GLM.

2.5 Estimation of the Dispersion Parameter

The dispersion parameter φ_0 is usually unknown (unless we work with Poisson or alternative distributions). This fact does not alter the estimation of $\boldsymbol{\beta}_0$ or the asymptotic properties of $\widehat{\boldsymbol{\beta}}_n$ but we occassionally need an estimator for φ_0. Instead of using the method of maximum likelihood, φ_0 is estimated by a modified method of moments.

Definition 2.6. The statistic

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)} \tag{2.10}$$

is called *the Pearson chi-square statistic**. An estimator for φ_0 is given by

$$\widehat{\varphi}_n = \frac{X^2}{n-p}.$$

Note. When the distribution of Y_i is normal, X^2 is the residual sum of squares SS_e and $\widehat{\varphi}_n$ is the usual estimator of residual variance.

The next theorem provides conditions for consistency of $\widehat{\varphi}_n$.

Theorem 2.6. *Let* $h(y, \boldsymbol{x}, \boldsymbol{\beta}) = \frac{[y - g^{-1}(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta})]^2}{V(g^{-1}(\boldsymbol{x}^\mathsf{T}\boldsymbol{\beta}))}$. *Suppose there exists a function* $C(y, \boldsymbol{x})$ *such that* $|\partial h/\partial \boldsymbol{\beta}| \leq C(y, \boldsymbol{x})$ *in a neighborhood* $\mathcal{B}_0$ *of* $\boldsymbol{\beta}_0$ *and* $\mathsf{E}\, C(Y_i, \boldsymbol{X}_i)$ *exists and is finite. Then* $\widehat{\varphi}_n \xrightarrow{\mathsf{P}} \varphi_0$.

Note. The condition of Theorem 2.6 is fulfilled when V and g' are bounded away from zero and V has a bounded derivative in a neighborhood of $\boldsymbol{\beta}_0$.

2.6 Deviance

Definition 2.7. The statistic

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_n) = 2\varphi_0 \big[\widetilde{\ell}_n(\boldsymbol{Y}) - \ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y})\big],$$

where $\widetilde{\ell}_n(\boldsymbol{Y})$ is the maximized log-likelihood of the saturated model, is called *the (unscaled) deviance* of the model with parameters $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ and observations $\boldsymbol{Y}$.

Note. In the saturated model, the MLE of μ_i is Y_i (see p. 19) and the MLE of θ_i is $\widetilde{\theta}_i = (b')^{-1}(Y_i)$. The maximized log likelihood (2.5) of the saturated model is

$$\widetilde{\ell}_n(\boldsymbol{Y}) = \frac{1}{\varphi_0} \sum_{i=1}^{n} \big[Y_i \widetilde{\theta}_i - b(\widetilde{\theta}_i)\big].$$

In the model with parameters $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, the maximized log likelihood (2.5) is

$$\ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y}) = \frac{1}{\varphi_0} \sum_{i=1}^{n} \big[Y_i \widehat{\theta}_i - b(\widehat{\theta}_i)\big],$$

where $\widehat{\theta}_i = (b')^{-1}(\widehat{\mu}_i)$. Obviously, $\widetilde{\ell}_n(\boldsymbol{Y}) \geq \ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y})$.

* Česky *Pearsonovo chí kvadrát*

The unscaled deviance can be expressed as

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_n) = 2 \sum_{i=1}^{n} \big[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i) \big]. \tag{2.11}$$

The deviance is always non-negative, does not depend on φ_0, and is zero if and only if the model provides a "perfect fit".

Note.

- The deviance is a goodness-of-fit measure. When the data are normal, the deviance is equal to the residual sums of squares. It generalizes the term residual sums of squares to the GLM*.
- $D^*(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_n, \varphi_0) = \varphi_0^{-1} D(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_n)$ is called *the scaled deviance*. If φ_0 is unknown, use the moment estimator $\widehat{\varphi}_n$.

2.7 Asymptotic Results

Asymptotic results for the GLM follow from the general theory of maximum likelihood estimation. The theory is reviewed in *"Summary of maximum likelihood estimation theory"* (here referred to as *MLE Summary*) available from the course website.

The following theorem transcribes the results of Theorems 2 – 5 from *MLE Summary* in the context of the GLM. The regularity conditions R1 – R4 are assured by the specification of the model. Condition R6 has been verified in Theorem 2.3, part (i) and Theorem 2.4, part (ii).

The Fisher information matrix

$$I(\boldsymbol{\beta}_0) = \mathsf{E}\, I(\boldsymbol{\beta}_0 \mid Y_i) = \mathsf{var}\, \boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i) = \frac{1}{\varphi_0} \mathsf{E}_{\boldsymbol{X}} w(\mu_i) \boldsymbol{X}_i^{\otimes 2}$$

is finite and of full rank by assumptions imposed on the covariates (finiteness of all necessary moments and linear independence of covariates).

Theorem 2.7.

(i) The MLE $\widehat{\boldsymbol{\beta}}_n$ is consistent (as long as the likelihood equations (2.7) have a unique solution).

(ii)

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, I(\boldsymbol{\beta}_0)).$$

(iii)

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\boldsymbol{0}, I^{-1}(\boldsymbol{\beta}_0)).$$

* The Pearson X^2 is another generalization.

(iv)

$$2\log\frac{L_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y})}{L_n(\boldsymbol{\beta}_0 \mid \boldsymbol{Y})} \xrightarrow{\mathsf{D}} \chi^2_p.$$

The information matrix $I(\boldsymbol{\beta}_0)$ can be consistently estimated by

$$\widehat{I}_n = \frac{1}{n\widehat{\varphi}_n}\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X}.$$

According to part (iii) of Theorem 2.7, the estimated asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ is

$$\widehat{I}_n^{-1}/n = \widehat{\varphi}_n(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X})^{-1}. \tag{2.12}$$

Denote $\widehat{\Sigma} \equiv (\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X})^{-1}$ so that $\widehat{\varphi}_n\widehat{\Sigma}$ estimates $\mathsf{var}\,\widehat{\boldsymbol{\beta}}_n$.

Let us consider the problem of testing the simple hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad \text{against} \qquad H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0.$$

The test statistics and their null distributions are established by the following theorem, which is based on Definition 5 and Theorem 7 from *MLE Summary*.

Theorem 2.8.

(i) ***Score (Rao) test.*** *Let* $\mu_i^0 = g^{-1}(\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}_0)$, $\mathbb{W}^0 = \operatorname{diag}(w(\mu_1^0),\dots,w(\mu_n^0))$, *denote* $\Sigma^0 = (\mathbb{X}^{\mathsf{T}}\mathbb{W}^0\mathbb{X})^{-1}$. *If* H_0 *holds then*

$$\begin{aligned} R_n &= \frac{1}{n}\boldsymbol{U}_n(\boldsymbol{\beta}_0)^{\mathsf{T}}\widehat{I}_n^{-1}\boldsymbol{U}_n(\boldsymbol{\beta}_0) \\ &= \frac{1}{\widehat{\varphi}_n}\Big(\sum_{i=1}^n w(\mu_i^0)g'(\mu_i^0)(Y_i-\mu_i^0)\boldsymbol{X}_i\Big)^{\mathsf{T}}\Sigma^0\Big(\sum_{i=1}^n w(\mu_i^0)g'(\mu_i^0)(Y_i-\mu_i^0)\boldsymbol{X}_i\Big) \\ &\xrightarrow{\mathsf{D}} \chi^2_p \end{aligned}$$

(ii) ***Wald test.*** *If* H_0 *holds then*

$$W_n = n\big(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big)^{\mathsf{T}}\widehat{I}_n\big(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big) = \frac{1}{\widehat{\varphi}_n}\big(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big)^{\mathsf{T}}\widehat{\Sigma}^{-1}\big(\widehat{\boldsymbol{\beta}}_n-\boldsymbol{\beta}_0\big) \xrightarrow{\mathsf{D}} \chi^2_p$$

(iii) ***Likelihood ratio test.*** *Let* $\theta_i^0 = (b')^{-1}(\mu_i^0)$. *If* H_0 *holds then*

$$\lambda_n = 2[\ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y}) - \ell_n(\boldsymbol{\beta}_0 \mid \boldsymbol{Y})] = \frac{2}{\widehat{\varphi}_n}\sum_{i=1}^n\big[Y_i(\widehat{\theta}_i-\theta_i^0) - b(\widehat{\theta}_i) + b(\theta_i^0)\big] \xrightarrow{\mathsf{D}} \chi^2_p$$

The simple hypothesis is rarely of interest for applications. We are more interested in composite hypotheses, for example, in testing that the last m components of the

regression parameter vector are all zero (without loss of generality: the components of $\boldsymbol{\beta}$ can be always rearranged in this way). Take

$$H_0^*: \begin{pmatrix} \beta_{p-m+1} \\ \beta_{p-m+2} \\ \vdots \\ \beta_p \end{pmatrix} = \mathbf{0} \qquad \text{against} \qquad H_1^*: \begin{pmatrix} \beta_{p-m+1} \\ \beta_{p-m+2} \\ \vdots \\ \beta_p \end{pmatrix} \neq \mathbf{0}$$

for some $m < p$. If H_0^* is true then the last m parameters attain zero value and the last m colums of the covariate vector can be excluded from the model. The null hypothesis specifies a submodel (with $p - m$ parameters) of the full model with (p parameters).

Denote $\boldsymbol{\beta}_M = (\beta_{p-m+1}, \ldots, \beta_p)^\mathsf{T}$ and $\boldsymbol{X}_i^M = (X_{p-m+1}, \ldots, X_p)^\mathsf{T}$. Let $\widehat{\boldsymbol{\beta}}_M = (\widehat{\beta}_{p-m+1}, \ldots, \widehat{\beta}_p)^\mathsf{T}$ be the MLE of $\boldsymbol{\beta}_M$ under the larger model. Let $\widetilde{\boldsymbol{\beta}}_n$ be the MLE of $\boldsymbol{\beta}$ under the submodel (subject to the constraint $\boldsymbol{\beta}_M = \mathbf{0}$), let $\widetilde{\mu}_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\widetilde{\boldsymbol{\beta}}_n)$ be the fitted values under the submodel.

Partition the $p \times p$ matrix $\widehat{\Sigma} = \widehat{I}_n^{-1}/(n\widehat{\varphi}_n) = (\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}\mathbb{X})^{-1}$ (the estimated asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ without $\widehat{\varphi}_n$) into four blocks

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\Sigma}_A & \widehat{\Sigma}_B \\ \widehat{\Sigma}_B^\mathsf{T} & \widehat{\Sigma}_M \end{pmatrix},$$

where the lower right block $\widehat{\Sigma}_M$ is of size $m \times m$.

Theorem 2.9.

(i) ***Score (Rao) test.*** *Let* $\widetilde{\mathbb{W}} = \operatorname{diag}(w(\widetilde{\mu}_1), \ldots, w(\widetilde{\mu}_n))$. *Let* $\widetilde{\Sigma}_M$ *be the* $m \times m$ *lower right block of the matrix* $\widetilde{\Sigma} = (\mathbb{X}^\mathsf{T}\widetilde{\mathbb{W}}\mathbb{X})^{-1}$. *Denote by* $\widetilde{\varphi}_n$ *the estimator of the dispersion parameter calculated under the submodel (under* H_0^**). If* H_0^* *holds then*

$$R_n^* = \frac{1}{\widetilde{\varphi}_n}\left(\sum_{i=1}^n w(\widetilde{\mu}_i)g'(\widetilde{\mu}_i)(Y_i - \widetilde{\mu}_i)\boldsymbol{X}_i^M\right)^\mathsf{T} \widetilde{\Sigma}_M \left(\sum_{i=1}^n w(\widetilde{\mu}_i)g'(\widetilde{\mu}_i)(Y_i - \widetilde{\mu}_i)\boldsymbol{X}_i^M\right) \xrightarrow{\mathsf{D}} \chi^2_m.$$

(ii) ***Wald test.*** *Denote by* $\widehat{\varphi}_n$ *the estimator of the dispersion parameter calculated under the larger model (not assuming that* H_0^* *is true). If* H_0^* *holds then*

$$W_n^* = \frac{1}{\widehat{\varphi}_n}\big(\widehat{\boldsymbol{\beta}}^M\big)^\mathsf{T}\widehat{\Sigma}_M^{-1}\big(\widehat{\boldsymbol{\beta}}^M\big) \xrightarrow{\mathsf{D}} \chi^2_m.$$

(iii) ***Likelihood ratio (deviance) test.*** *Let* $D(\boldsymbol{Y} \mid \widetilde{\boldsymbol{\beta}})$ *be the (unscaled) deviance of the submodel, let* $D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}})$ *be the (unscaled) deviance of the larger model. Let the estimate* $\widehat{\varphi}_n$ *be calculated under the larger model (not assuming that* H_0^* *is true). If* H_0^* *holds then*

$$\lambda_n^* = \frac{1}{\widehat{\varphi}_n}\big[D(\boldsymbol{Y} \mid \widetilde{\boldsymbol{\beta}}) - D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}})\big] \xrightarrow{\mathsf{D}} \chi^2_m.$$

Note.

- Theorem 2.9 follows from Definition 6 and Theorem 9 in *MLE Summary*. The hypothesis H_0^* is rejected at the asymptotic level of α if the chosen test statistic (it must be selected in advance) exceeds the $1-\alpha$ quantile of the χ^2_m distribution.
- Under the standard linear regression model with normal distribution, these three test statistics are all equal to the F test statistic (1.1) for submodel testing. In that case, the exact distribution of the test statistics under the null hypothesis is $F_{m,n-p}$. When normality does not hold or the link is not identity, the three test statistics are not the same and we only know that their asymptotic distribution is χ^2_m.
- Generally, the likelihood ratio test statistic is twice the difference in the log likelihoods between the model and the submodel. However, it can be also expressed as a properly scaled difference in deviances between the submodel and the model. *The deviance test is the preferred tool for testing submodels in generalized linear models.*
- The Wald and Rao statistics are asymptotically equivalent to the likelihood ratio test statistic. However, in finite samples they may be different. Unlike the likelihood ratio test statistic, the Wald test statistic depends on the parametrization of the model and tends to have the slowest convergence to the asymptotic distribution. For these reasons, the Wald statistic is the least desirable of the three.
- An important special case is $m = 1$ (testing of a single parameter). Then the Wald statistic for testing zero value of the j-th parameter is

$$\left(\frac{\widehat{\beta}_j}{\sqrt{\widehat{\varphi}_n \widehat{\sigma}^2_{jj}}}\right)^2, \tag{2.13}$$

  where $\widehat{\sigma}^2_{jj}$ is the j-th diagonal element of $\widehat{\Sigma}$. Before applying the square, these statistics are asymptotically standard normal; in this form they are automatically provided in the output of almost any statistical software for fitting the GLM.
- The deviance of the current model $D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}})$ is twice the difference in log likelihoods between the saturated model and the current model. However, the deviance cannot be in general used as a test statistic to compare the goodness-of-fit of the current model to the saturated model unless all covariates are discrete (otherwise the number of parameters of the saturated model grows to infinity and Theorem 9 from *MLE Summary* does not hold). Differences in deviances between a submodel and a larger model do not have this problem.

Confidence intervals

The simplest confidence intervals for the individual parameters are based on Wald test statistics (2.13). The interval with end points

$$\widehat{\beta}_j \pm u_{1-\alpha/2}\sqrt{\widehat{\varphi}_n \widehat{\sigma}^2_{jj}},$$

covers β_j with probability converging to $1-\alpha$.

Better confidence intervals would be obtained from inverting acceptance regions of the Rao or likelihood ratio test statistics or using profile likelihood methods.

Wald-type confidence intervals for linear combinations of parameters $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}_0$ where $\mathbf{0} \neq \boldsymbol{c} \in \mathbb{R}^p$ can be obtained easily from Theorem 2.7 part (iii). An asymptotic confidence interval with coverage probability converging to $1-\alpha$ has end points

$$\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n \pm u_{1-\alpha/2}\sqrt{\widehat{\varphi}_n \boldsymbol{c}^\mathsf{T}\widehat{\Sigma}\boldsymbol{c}}.$$

2.8 Diagnostic Methods for the GLM

Diagnostic methods can be derived from the linear model using Theorem 2.5. Let $\mathbb{X}^* = \hat{\mathbb{W}}^{1/2}\mathbb{X}$ and $\boldsymbol{Y}^* = \hat{\mathbb{W}}^{1/2}\widehat{\boldsymbol{Z}}$. Write $\widehat{\boldsymbol{\beta}}_n$ as an ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_n = (\mathbb{X}^{*\mathsf{T}}\mathbb{X}^*)^{-1}\mathbb{X}^{*\mathsf{T}}\boldsymbol{Y}^*$. Let

$$\mathbb{H}^* = \mathbb{X}^*(\mathbb{X}^{*\mathsf{T}}\mathbb{X}^*)^{-1}\mathbb{X}^{*\mathsf{T}} = \hat{\mathbb{W}}^{1/2}\mathbb{X}(\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}^{1/2},$$

and $\hat{\boldsymbol{Y}}^* = \mathbb{X}^*\widehat{\boldsymbol{\beta}} = \mathbb{H}^*\boldsymbol{Y}^* = \hat{\mathbb{W}}^{1/2}\mathbb{X}(\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}\widehat{\boldsymbol{Z}}$.

2.8.1 Pearson residuals

Pearson residuals are defined by the identity $\boldsymbol{Y}^* - \hat{\boldsymbol{Y}}^* = \hat{\mathbb{W}}^{1/2}\widehat{\boldsymbol{Z}} - \hat{\mathbb{W}}^{1/2}\mathbb{X}\widehat{\boldsymbol{\beta}}$, which gives the following residuals for the individual observations

$$r_i^P = \frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}.$$

Sum of squares of Pearson residuals is equal to the Pearson X^2 statistic:

$$\sum_{i=1}^{n}\left(r_i^P\right)^2 = X^2.$$

2.8.2 Leverages

It can be shown that

$$\mathsf{var}\, r_i^P \approx \varphi_0(1 - h_{ii}^*),$$

where h_{ii}^*, the i-th diagonal element of $\mathbb{H}^*$, is called *the leverage*. Potentially influential observations can be identified by the rule of thumb $h_{ii}^* > 2p/(n-2p)$. These observations are sort of atypical in their covariates and thus may have unduly strong influence on the results of the model fit.

2.8.3 Standardized Pearson residuals

Standardized Pearson residuals divide r_i^P by the square root of the estimated approximate variance of r_i^P:

$$r_i^{PS} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{\varphi}_n V(\widehat{\mu}_i)(1 - h_{ii}^*)}}.$$

They have approximately unit variance.

2.8.4 Deviance residuals

Deviance residuals are signed square roots of the contributions of the observations to the deviance. Let $\widetilde{\theta}_i = (b')^{-1}(Y_i)$, $d_i = 2\big\{Y_i[\widetilde{\theta}_i - \widehat{\theta}_i] - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i)\big\}$, and define the deviance residual as

$$r_i^D = \operatorname{sgn}(Y_i - \widehat{\mu}_i)\sqrt{d_i}.$$

Sum of squares of deviance residuals is equal to the deviance:

$$\sum_{i=1}^{n} \big(r_i^D\big)^2 = D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}}).$$

2.8.5 Standardized deviance residuals

Standardized deviance residuals use the same normalization as standardized Pearson residuals.

$$r_i^{DS} = \frac{\operatorname{sgn}(Y_i - \widehat{\mu}_i)\sqrt{d_i}}{\sqrt{\widehat{\varphi}_n(1 - h_{ii}^*)}}.$$

These are the default residuals in R.

2.8.6 Cook's distance

Cook's distance measures the influence of the i-th observation on the estimates of regression parameters $\widehat{\boldsymbol{\beta}}$. Let $\widehat{\boldsymbol{\beta}}_{(i)}$ denote the estimates calculated after deletion of the i-th observation from the data set. Cook's distance is defined as

$$CD_i = \frac{1}{p\widehat{\varphi}_n}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})^{\mathsf{T}}\mathbb{X}^{*\mathsf{T}}\mathbb{X}^*(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}).$$

In linear regression, it can be shown that

$$CD_i = \frac{1}{p\widehat{\varphi}_n}\left(\frac{Y_i^* - \hat{Y}_i^*}{\sqrt{1-h_{ii}^*}}\right)^2 \frac{h_{ii}^*}{1-h_{ii}^*} = \frac{1}{p}\left(r_i^{PS}\right)^2 \frac{h_{ii}^*}{1-h_{ii}^*}.$$

This is how Cook's distance is calculated in the GLM. An observation is considered influential if $CD_i > \frac{8}{n-2p}$.

2.8.7 Residual plots

Residual plots are created and used in a direct analogy with the linear model. However, for some data types (e.g. binary data) the residual plots are much less informative and require smoothing to yield any useful information. In general, residual plots are somewhat less useful in the GLM than they are in the linear model.

2.8.8 Diagnostics of the link function

We only mention two simple methods for checking that the correct link function was selected. Plotting the adjusted dependent variable $\widehat{Z}_i$ against the linear predictor $\widehat{\eta}_i$ provides a graphical check. If the link is correct the plot should reveal a linear pattern. A formal test can be obtained by adding $(\widehat{\eta}_i)^2$ to the model as an additional covariate and testing that its parameter is zero. If the hypothesis is rejected the link may be incorrect.

Both methods are sensitive to inappropriate transformations of the regressors. If the transformations are not chosen well, both methods may indicate a problem even if the link is correct.

Incorrect link functions do not have a serious effect on deciding which regressors affect the response or on the results of submodel testing. The choice of the link function is important if the primary goal of the analysis is prediction.

2.9 Model-building strategies

Model-building strategies for generalized linear models do not differ from the strategies applied to other regression models, including linear regression. The primary tool for

model building are deviance tests comparing a larger model with a submodel. If the deviance test is significant it means that the terms in the larger model cannot be removed without a significant decrease in the quality of model fit.

Since the development of the final model usually involves repeated applications of deviance tests, each performed on a selected level α (usually $\alpha = 0.05$), it is clear that the overall procedure does not preserve the desired level. If many tests are done then the final model is likely to include terms that in fact do not affect the response at all (*overfitting*). There is no universal and reliable method for adjusting the levels of the individual tests so that the overall probability of including irrelevant terms is under control. Nevertheless the analyst should be aware of this problem and should not interpret the p-values of submodel tests too dogmatically.

Approaches for developing reasonable models vary with the nature of the problem, structure of the data and questions to be addressed by the analysis. There is no universal solution to be recommended. Each problem requires careful consideration by the analyst taking into account the nature of the problem, the data-collection methods and tools, the meaning of the variables included in the dataset, their mutual relationships, and the goals of the analysis.

If *prediction* is the primary goal, it is useful to consider rich and flexible models. Omission of an important term from the model or its inclusion with an inappropriate transformation may have detrimental biasing effects on the predictions. If unnecessary covariates are left in, the variability in the predicted response is increased but the predictions are not biased. Interpretation of regression parameters is usually not that important. In prediction analyses, validation of the prediction model should be performed either by dividing the data set into disjoint training (used for model building) and validation (used for evaluation of the predictions) subsets or at least by cross-validation (predictions of each observation by a model fitted on data excluding that observation). Validation is a very useful tool for selection of the best prediction model out of several candidates.

If the goal is to *evaluate covariate effects* ("how does covariate X affect the mean of the response Y?"), one must be really careful about several things. First, the covariate of interest must be kept in the model even if it is not significant – otherwise its effect cannot be evaluated. Second, the regression parameters expressing the influence of the covariate of interest should have a straightforward interpretation. Thus, we cannot afford to model the effect of X by a complicated function that cannot be easily summarized (splines of order > 1, polynomials), or to use complex transformations of the response or link functions that are difficult to interpret. Third, there might be covariates that should be kept in the model regardless of their significance (suspected confounders) and/or covariates that should not be included in the model no matter how significant they are (variables on the causal pathway between X and Y, variables that are influenced by the value of Y). Thus, making reasonable decisions about which covariates should be included in the model and which should be dropped is not based solely on significance tests but also

on external expert knowledge of the problem to be analyzed. It is precisely this issue that makes automated computer-based algorithms (unsupervised stepwise regression, regression trees, neural networks, deep learning, etc.) unable to solve certain problems acceptably.

Another common problem in model-building strategies is the inclusion of *interactions*, especially when the number of covariates that can be considered for interactions is quite large. The strategy that starts with a model that includes a lot of main effects as well as all possible two-way interactions between them, and tries to gradually eliminate the superfluous terms usually does not lead to a good model. With this approach, we are likely to end up with a model that suffers from overfitting, keeps a lot of unnecessary interactions and is hard to interpret. It is better to fit only the main effects first, eliminate those that are not contributing to the model, and then try to add two-way interactions of the remaining terms one by one. This strategy is much more likely to end up only with interactions that really matter. Considering higher order interactions (three-way, four-way,...) is usually a hopeless task. It is better not to consider them at all, except in analyses where, for some reason, such interactions are among the terms of interest.

There is one principle about building models with interactions, which is almost universally valid and the analyst should take care not to violate it. The models should be built *hierarchically*, meaning that if a covariate is present in a higher-order interaction, then all its corresponding lower-order interactions as well as the main effects should be included in the model as well, no matter if they are significant or not. This principle should be ignored only in analyses where there is a sound justification for its violation.

This brief exposition of model-building strategies cannot be complete and should be understood in the whole context of the particular task to be done. As noted earlier, each problem should be carefully considered in order to choose a tailor-made strategy that works well for it. This requires practical experience. The analyst should be aware that there is no such thing as the true model and that his task is not to discover it. All models are wrong – we are only looking for an acceptable model that provides satisfactory answers to the questions of interest.

3 Generalized Linear Model for Discrete Responses

3.1 Analysis of Binary Data

3.1.1 Alternative vs. binomial data

Let $Y_{ij}^* \sim \mathsf{Alt}(\pi_i)$, $\pi \in (0,1)$, be independent variables for $i = 1, \ldots, K$, $j = 1, \ldots, m_i$. For a fixed i, $Y_{i1}^*, \ldots, Y_{im_i}^*$ are identically distributed. The total number of observations is $N = \sum_{i=1}^{K} m_i$. Let π_i depend on $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ through the linear predictor $\eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$, $\boldsymbol{\beta}$ is the vector of unknown regression coefficients to be estimated. Therefore $Y_{i1}^*, \ldots, Y_{im_i}^*$ share the same covariate vector $\boldsymbol{X}_i$.

The response $Y_{ij}^* \sim \mathsf{Alt}(\pi_i)$ has a distribution of exponential family with $\mu_i \equiv \mathsf{E}\, Y_{ij}^* = \pi_i$ and $\mathsf{var}\, Y_{ij}^* = \pi_i(1-\pi_i)$. The variance function is $V(\mu) = \mu(1-\mu)$, the dispersion parameter is $\varphi = 1$, the canonical parameter is $\theta_i = \log \frac{\pi_i}{1-\pi_i}$. Finally, $b(\theta_i) = \log(1 + \mathrm{e}^{\theta_i}) = \log \frac{1}{1-\pi_i}$.

Denote $Y_i = \sum_{j=1}^{m_i} Y_{ij}^*$. Then $Y_i \sim \mathsf{Bi}(m_i, \pi_i)$. Thus, binomial responses can be treated together with alternative responses even though the binomial distribution does not strictly belong to the exponential family.

The dataset for alternative/binomial regression can be arranged in two different ways. It is recommended not to mix the two data formats in a single dataset.

Format A. The dataset is arranged so that there are N rows and each value of $\boldsymbol{X}_i$ appears in m_i different rows. The row corresponding to the ij-th observation includes Y_{ij}^* and $\boldsymbol{X}_i$. This is the Bernoulli format of the data.

Format B. The dataset is arranged so that there are K rows and each value of $\boldsymbol{X}_i$ appears only once. The i-th row includes Y_i, m_i, and $\boldsymbol{X}_i$. This is the binomial format of the data.

There are two different kinds of asymptotics when $N \to \infty$.

1. K is constant, $m_i \to \infty$ at the same rate for all i. This happens when all covariates are discrete with a finite support.
2. $K \to \infty$, m_i are small (typically $m_i = 1$). This happens when at least one covariate is continuous.

Most of the results are the same for both kinds of asymptotics but there are certain

important differences that will be pointed out later.

3.1.2 Link functions for binary data

Because $\mu_i \equiv \pi_i \in (0,1)$, suitable link functions are maps $(0,1) \to \mathbb{R}$. Any quantile function of a continuous distribution on $\mathbb{R}$ could be used as a link function for binary responses. Here are some examples:

Logistic link

Take the quantile function of the standard logistic distribution.

$$g(\mu_i) = \log \frac{\mu_i}{1-\mu_i}, \qquad \mu_i = \frac{\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}.$$

This is the logistic link, the canonical link function, the most commonly used link for binary data. The model is called *the logistic regression model*[*].

Probit link

Take the quantile function of the standard normal distribution.

$$g(\mu_i) = \Phi^{-1}(\mu_i), \qquad \mu_i = \Phi(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}).$$

This is the probit link, the model is called *the probit regression model*[†]. It is used in threshold analysis, toxicology and pharmacokinetics.

Cauchit link

Take the quantile function of the standard Cauchy distribution.

$$g(\mu_i) = \tan[\pi(\mu_i - 0.5)], \qquad \mu_i = \frac{1}{\pi}\arctan(\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}) + \frac{1}{2}.$$

This is the cauchit link, the model is called *the cauchit regression model*[‡]. It is suitable when π_i converges to 0 (1) extremely slowly for $\eta_i \to \pm\infty$.

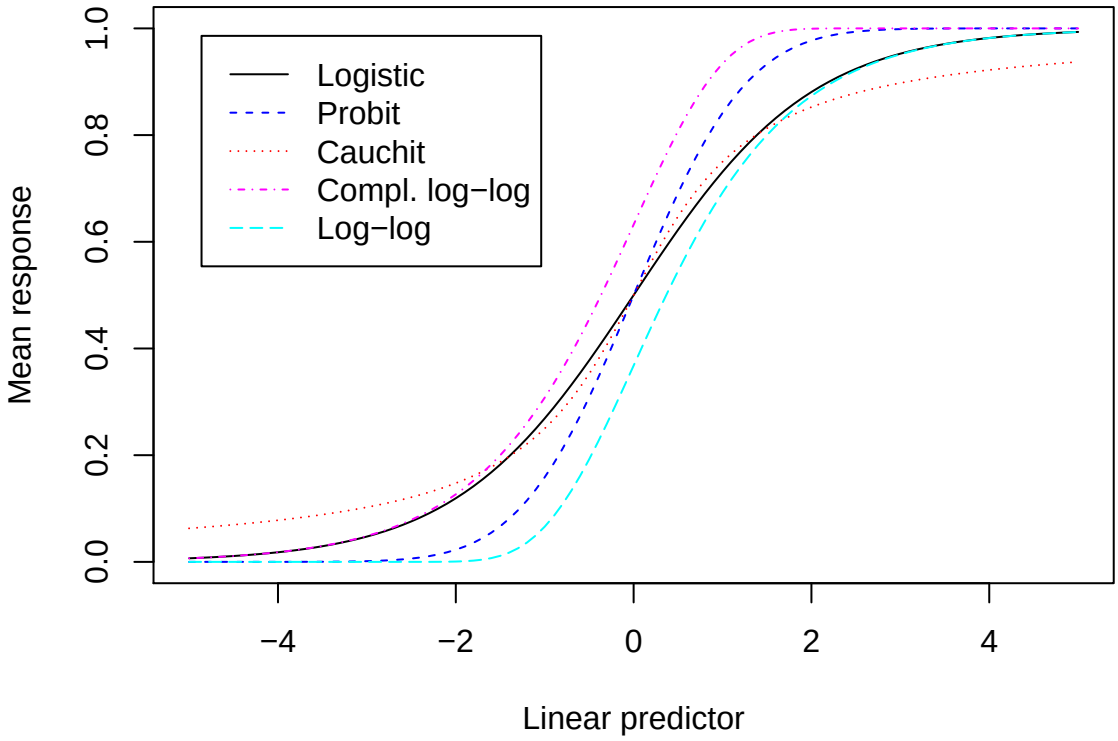
Complementary log-log link

Take the quantile function of the negative Gumbel (extreme value) random variable.

$$g(\mu_i) = \log(-\log(1-\mu_i)), \qquad \mu_i = 1 - \mathrm{e}^{-\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}.$$

[*] Česky *logistická regrese* [†] Česky *probitová regrese* [‡] Česky *cauchitová regrese*

Figure 3.1: Inverse link functions for binary data. The linear predictor η_i is on the horizontal axis, the success probability π_i is on the vertical axis.



This link function does not possess symetry properties. It is used in the analysis of discrete survival data. Its counterpart is the log-log link

$$g(\mu_i) = -\log(-\log(\mu_i)), \qquad \mu_i = \mathrm{e}^{-\exp\{-\boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}\}}.$$

The inverse link functions are plotted in Figure 3.1. The choice of the link function should be governed by the desired interpretation of the fitted model rather than by the data. The canonical logistic link should be the first choice unless a different interpretation is needed or there is a strong prior reason to choose a different link.

3.1.3 Binary data likelihood

There are two different sampling schemes to be considered.

(i) Alternative responses are observed independently of each other together with the

covariates. Then m_i are random variables.

(ii) m_i is fixed in advance, then m_i independent observations are obtained for each combination of the covariates.

The likelihoods for these two schemes only differ by a constant that does not affect the analysis. If m_i is random then the likelihood is a product of independent alternative distributions

$$\prod_{i=1}^{K}\prod_{j=1}^{m_i}\pi_i^{Y_{ij}^*}(1-\pi_i)^{1-Y_{ij}^*} = \prod_{i=1}^{K}\pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i}.$$

If m_i is fixed then the likelihood is a product of independent binomial distributions

$$\prod_{i=1}^{K}\binom{m_i}{Y_i}\pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i} = \prod_{i=1}^{K}\pi_i^{Y_i}(1-\pi_i)^{m_i-Y_i}\prod_{i=1}^{K}\binom{m_i}{Y_i}.$$

The product of the binomial numbers does not include the parameters, so it is not relevant. The first scheme follows the framework of independent observations from a distribution of exponential type so the theory of Chapter 2 applies. The second scheme does not follow this framework but all the results have exactly the same form and properties. Therefore we do not have to distinguish the sampling schemes.

3.1.4 Threshold analysis by probit regression

The probit link has an interesting application in threshold analysis of normally distributed data.

Consider random variables U_i that follow the normal linear regression model

$$U_i = \boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\alpha} + \varepsilon_i, \tag{3.1}$$

where $\boldsymbol{Z}_i$ are p-dimensional covariate vectors, $\boldsymbol{\alpha}$ are regression coefficients and $\varepsilon_i \sim \mathsf{N}(0,\sigma^2)$ are error terms. Now suppose that the responses U_i cannot be observed directly. Instead, a threshold C_i is provided and we learn whether the unobserved response U_i exceeds the threshold or not.

Assume that C_i is independent of U_i. The observed response is $Y_i = I(U_i < C_i)$, together with the values of the covariates $\boldsymbol{Z}_i$ and the threshold C_i. The goal is to estimate the regression coefficients α and the residual variance σ^2 of the underlying linear regression model (3.1).

The observations come in the form of iid triplets $(Y_i, C_i, \boldsymbol{Z}_i)$. The response Y_i follows an alternative distribution with

$$\mathrm{P}\left[Y_i = 1\right] \equiv p_i = \mathrm{P}\left[U_i < C_i\right].$$

Conditionally on the value of the observed threshold C_i, we get

$$p_i = \mathrm{P}\left[\frac{U_i - \boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\alpha}}{\sigma} < \frac{C_i - \boldsymbol{Z}_i^\mathsf{T}\boldsymbol{\alpha}}{\sigma}\right] = \Phi\left(\frac{C_i}{\sigma} - \boldsymbol{Z}_i^\mathsf{T}\frac{\boldsymbol{\alpha}}{\sigma}\right).$$

Define

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{Z}_i \\ C_i \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} -\boldsymbol{\alpha}/\sigma \\ 1/\sigma \end{pmatrix}.$$

This translates the problem into binary probit regression model with the linear predictor $\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$. The parameters $\boldsymbol{\beta}$ can be estimated by the usual procedures for the analysis of the GLM. The parameters of interest can be obtained from $\widehat{\boldsymbol{\beta}}$ as $\widehat{\sigma}^2 = \frac{1}{\widehat{\beta}_{p+1}^2}$ and $\widehat{\alpha}_j = -\frac{\widehat{\beta}_j}{\widehat{\beta}_{p+1}}$.

Of course, this can only be done if the threshold values C_i are linearly independent of the covariates $\boldsymbol{Z}_i$. For example, if C_i are all set to the same value, the intercept term cannot be distinguished from the residual variance and the parameters of the original linear regression model cannot be determined.

3.1.5 Logistic regression

The logistic regression model is the most commonly used model for the analysis of binary and binomial responses.

The logistic link has the form $g(\pi_i) = \log \frac{\pi_i}{1-\pi_i}$, where $\pi_i/(1-\pi_i)$ is the odds of success. The success probabilities can be expressed as $\pi_i = \frac{\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}$.

Interpretation of regression parameters

Let $\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta} = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. Denote $\pi_0 = \mathrm{P}\big(Y_{ij}^* = 1 \,\big|\, X_2 = \cdots = X_p = 0\big)$. Then

$$\log \frac{\pi_0}{1-\pi_0} = \beta_1$$

so $\mathrm{e}^{\boldsymbol{\beta}_1}$ is the odds of success for an individual with zero values in all covariates.

Now consider two individuals: one with observed covariates $\boldsymbol{x}^0 = (1, x_2, \ldots, x_p)^\mathsf{T}$, the other with observed covariates $\boldsymbol{x}^j = (1, x_2, \ldots, x_j + 1, \ldots, x_p)^\mathsf{T}$ (the j-th covariate is increased by 1, the others are the same). Denote $\pi_{X0} = \mathrm{P}\big(Y_{ij}^* = 1 \,\big|\, \boldsymbol{X} = \boldsymbol{x}^0\big)$ and $\pi_{Xj} = \mathrm{P}\big(Y_{ij}^* = 1 \,\big|\, \boldsymbol{X} = \boldsymbol{x}^j\big)$. Then

$$\boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}^0 = \log \frac{\pi_{X0}}{1-\pi_{X0}} \qquad \text{and} \qquad \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}^j = \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}^0 + \beta_j = \log \frac{\pi_{Xj}}{1-\pi_{Xj}}.$$

It follows that

$$\beta_j = \log\left(\frac{\pi_{Xj}}{1-\pi_{Xj}} \cdot \frac{1-\pi_{X0}}{\pi_{X0}}\right) \qquad \text{and} \qquad \mathrm{e}^{\beta_j} = \frac{\pi_{Xj}(1-\pi_{X0})}{\pi_{X0}(1-\pi_{Xj})}.$$

Thus e^{β_j} is the odds ratio for success comparing two individuals differing by one unit in the covariate X_j. E.g., if $\beta_j = 0.431$ one can say that a unit increase in the covariate

X_j increases the odds of success $\mathrm{e}^{0.431} = 1.539$ times (or by 53.9%). When $\beta_j = 0$ the odds ratio is 1 and the covariate has no effect on the odds of success (or the probability of success) given the other covariates.

Consider a two-by-two contingency table of conditional probabilities

Covariates	$Y = 1$	$Y = 0$
$\boldsymbol{X} = \boldsymbol{x}^j$	π_{Xj}	$1 - \pi_{Xj}$
$\boldsymbol{X} = \boldsymbol{x}^0$	π_{X0}	$1 - \pi_{X0}$

The odds ratio e^{β_j} describes the association between $\boldsymbol{X}$ and Y in this contingency table. The odds ratio is one if and only if there is independence.

Estimation of parameters

By Theorem 2.3, the score statistic with the canonical link is

$$\boldsymbol{U}_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{K} \sum_{j=1}^{m_i} (Y_{ij}^* - \pi_i)\boldsymbol{X}_i = \sum_{i=1}^{K} (Y_i - m_i\pi_i)\boldsymbol{X}_i$$

and $\widehat{\boldsymbol{\beta}}_n$ solves the equations

$$\sum_{i=1}^{K} Y_i \boldsymbol{X}_i = \sum_{i=1}^{K} m_i \widehat{\pi}_i \boldsymbol{X}_i,$$

where

$$\widehat{\pi}_i = \frac{\exp\{\boldsymbol{X}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_n\}}{1 + \exp\{\boldsymbol{X}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_n\}}.$$

The IWLS algorithm can be implemented in two different ways depending on the data format. With the Bernoulli format A, the regression matrix $\mathbb{X}$ includes each observed value of $\boldsymbol{X}_i$ in m_i different rows, and its dimension is $N \times p$. Suppose the observations ij are ordered by the two indices $11, \ldots, 1m_1, 21, \ldots, 2m_2, \ldots, Km_K$. Let

$$\hat{\mathbb{W}}^{(k)} = \operatorname{diag}\left(\widehat{\pi}_1^{(k)}(1 - \widehat{\pi}_1^{(k)}), \ldots, \widehat{\pi}_K^{(k)}(1 - \widehat{\pi}_K^{(k)})\right)$$

be an $N \times N$ matrix, where the i-th element is repeated m_i times, define

$$\widehat{Z}_{ij}^{(k)} = \boldsymbol{X}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}^{(k)} + \frac{Y_{ij}^* - \widehat{\pi}_i^{(k)}}{\widehat{\pi}_i^{(k)}(1 - \widehat{\pi}_i^{(k)})},$$

and create an N-vector $\widehat{\boldsymbol{Z}}^{(k)} = (\widehat{Z}_{11}^{(k)}, \ldots, \widehat{Z}_{Km_K}^{(k)})^{\mathsf{T}}$. The IWLS algorithm iterates

$$\widehat{\boldsymbol{\beta}}_n^{(k+1)} = (\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1} (\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}}^{(k)} \widehat{\boldsymbol{Z}}^{(k)})$$

until convergence.

With the binomial format B, the regression matrix $\mathbb{X}_R$ includes each observed value of $\boldsymbol{X}_i$ only once, and its dimension is $K \times p$. Let

$$\hat{\mathbb{W}}_R^{(k)} = \operatorname{diag}\left(m_1\widehat{\pi}_1^{(k)}(1-\widehat{\pi}_1^{(k)}), \ldots, m_K\widehat{\pi}_K^{(k)}(1-\widehat{\pi}_K^{(k)})\right)$$

be an $K \times K$ matrix, where each element appears just once, define

$$\widehat{Z}_{Ri}^{(k)} = \boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}^{(k)} + \frac{Y_i - m_i\widehat{\pi}_i^{(k)}}{m_i\widehat{\pi}_i^{(k)}(1-\widehat{\pi}_i^{(k)})},$$

and create a K-vector $\widehat{\boldsymbol{Z}}_R^{(k)} = (\widehat{Z}_{R1}^{(k)}, \ldots, \widehat{Z}_{RK}^{(k)})^{\mathsf{T}}$. The IWLS algorithm iterates

$$\widehat{\boldsymbol{\beta}}_n^{(k+1)} = (\mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R^{(k)}\mathbb{X}_R)^{-1}(\mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R^{(k)}\widehat{\boldsymbol{Z}}_R^{(k)})$$

until convergence.

Obviously, $\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}^{(k)}\mathbb{X} = \mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R^{(k)}\mathbb{X}_R$ and $\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}^{(k)}\widehat{\boldsymbol{Z}}^{(k)} = \mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R^{(k)}\widehat{\boldsymbol{Z}}_R^{(k)}$, so the two implementations of the IWLS algorithm for the two data formats are equivalent.

The information matrix is

$$I(\boldsymbol{\beta}) = \mathsf{E}_{\boldsymbol{X}}\widehat{\pi}_i(1-\widehat{\pi}_i)\boldsymbol{X}_i^{\otimes 2},$$

and it can be estimated by

$$\widehat{I}_n = \frac{1}{N}\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X} = \frac{1}{N}\mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R\mathbb{X}_R.$$

The estimated variance of $\widehat{\boldsymbol{\beta}}_n$ is

$$(\mathbb{X}^{\mathsf{T}}\hat{\mathbb{W}}\mathbb{X})^{-1} = (\mathbb{X}_R^{\mathsf{T}}\hat{\mathbb{W}}_R\mathbb{X}_R)^{-1}.$$

Deviance

The saturated model has K parameters $\pi_1, \ldots, \pi_K$ and their MLEs are $\widetilde{\pi}_i = Y_i/m_i$. Under the first kind of asymptotics (K constant, $m_i \to \infty\ \forall i$), the MLE theory holds for the saturated model and the MLEs are consistent and asymptotically normal. For the second kind of asymptotics ($K \to \infty$, m_i small), the MLEs are inconsistent and the theory fails.

The deviance for logistic regression is

$$\begin{aligned} D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}}) &= 2\sum_{i=1}^{K}\sum_{j=1}^{m_i}\left[Y_{ij}^*\left(\log\frac{Y_i/m_i}{1-Y_i/m_i} - \log\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}\right) - \log\frac{1}{1-Y_i/m_i} + \log\frac{1}{1-\widehat{\pi}_i}\right] \\ &= 2\sum_{i=1}^{K}\left[Y_i\log\frac{Y_i}{m_i\widehat{\pi}_i} + (m_i - Y_i)\log\frac{m_i - Y_i}{m_i - m_i\widehat{\pi}_i}\right]. \end{aligned}$$

According to Theorem 2.9(iii), the difference in deviances between a submodel and a wider model has a limiting χ^2 distribution if the submodel holds. This is used for model building in logistic regression.

Under the first kind of asymptotics (K constant, $m_i \to \infty \ \forall i$), the deviance alone has asymptotically χ^2 distribution with $K - p$ degrees of freedom if the current model is valid. Thus, in these circumstances a formal goodness-of-fit test can be obtained by comparing the fit of the current model to the saturated model. If the deviance exceeds $1-\alpha$ quantile of χ^2_{K-p} distribution, the current model should be rejected. However, this test is only valid if all covariates are discrete and there are enough observations at each value of $\boldsymbol{X}$ (e.g., $m_i\widehat{\pi}_i > 5$, $m_i(1-\widehat{\pi}_i) > 5$).

Pearson X^2

The Pearson residuals and the Pearson X^2 statistic have different versions and different interpretations depending on the data format. The two versions will be the same if and only if $m_i = 1$ for all i. It is important to keep in mind that most statistical software will calculate the version that corresponds to the input data format for the particular dataset.

Bernoulli data format

With the Bernoulli data format A, the Pearson residual is

$$r_{ij}^P = \frac{Y_{ij}^* - \widehat{\pi}_i}{\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}} = \begin{cases} \sqrt{\frac{1-\widehat{\pi}_i}{\widehat{\pi}_i}} & Y_{ij}^* = 1 \\ -\sqrt{\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}} & Y_{ij}^* = 0. \end{cases}$$

When plotted, these residuals form two separated clouds, one for successes and one for failures. Their usefulness for various model checking procedures is limited. The Pearson X^2 statistic is

$$X^2 = \sum_{i=1}^{K}\sum_{j=1}^{m_i} \frac{(Y_{ij}^* - \widehat{\pi}_i)^2}{\widehat{\pi}_i(1-\widehat{\pi}_i)} = \sum_{i=1}^{K}\left[\frac{Y_i}{m_i\widehat{\pi}_i} m_i(1-\widehat{\pi}_i) + \frac{m_i - Y_i}{m_i(1-\widehat{\pi}_i)} m_i\widehat{\pi}_i\right].$$

If the model fits well then $X^2 \approx N$. This version of the Pearson X^2 statistic does not possess a useful interpretation.

Binomial data format

With the binomial data format B, the Pearson residuals are calculated as $\hat{\mathbb{W}}_R^{1/2}(\widehat{\boldsymbol{Z}}_R - \mathbb{X}_R\widehat{\boldsymbol{\beta}})$. The i-th residual is

$$r_i^P = \frac{Y_i - m_i\widehat{\pi}_i}{\sqrt{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)}}$$

As long as $m_i \gg 1$, these residuals have much more useful distribution. They are

approximately normal as $m_i \to \infty$. The Pearson X^2 statistic is

$$X^2 = \sum_{i=1}^{K} \frac{(Y_i - m_i\widehat{\pi}_i)^2}{m_i\widehat{\pi}_i(1-\widehat{\pi}_i)} = \sum_{i=1}^{K} \frac{(Y_i - m_i\widehat{\pi}_i)^2}{m_i\widehat{\pi}_i} + \sum_{i=1}^{K} \frac{[m_i - Y_i - m_i(1-\widehat{\pi}_i)]^2}{m_i(1-\widehat{\pi}_i)},$$

which is the χ^2 test statistic for testing goodness of fit in a $2 \times K$ table with p estimated parameters. When K is constant and $m_i \to \infty$ for all i (at the same rate) then $X^2 \xrightarrow{\mathsf{D}} \chi^2_{K-p}$ if the model is valid. For the saturated model ($p = K$), $X^2 = 0$. Thus, the binomial version of the Pearson X^2 statistic can be used for testing the goodness of fit if all covariates are discrete and there are enough observations at each value of $\boldsymbol{X}$. However, differences in Pearson X^2 statistics between the model and a submodel do not converge to a χ^2 distribution. This is why we always prefer deviance tests to tests based on the Pearson X^2 statistic.

3.2 Analysis of Poisson Count Data

3.2.1 Poisson loglinear model

Let $Y_1, \ldots, Y_n$ be independent random variables, $Y_i \sim \mathsf{Po}(\lambda_i)$. Let λ_i depend on covariates $\boldsymbol{X}_i$ through the identity $\log \lambda_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.

Poisson distribution belongs to the exponential family with $\mu_i = \lambda_i$ and $\mathsf{var}\, Y_i = \lambda_i$. The variance function is $V(\mu) = \mu$, the dispersion parameter is $\varphi = 1$, the canonical parameter is $\theta_i = \log \lambda_i$, $b(\theta_i) = \mathrm{e}^{\theta_i}$. The log link is canonical for Poisson distribution.

We have

$$\mathsf{E}\big[Y_i \,\big|\, \boldsymbol{X}_i\big] = \mathsf{var}\big[Y_i \,\big|\, \boldsymbol{X}_i\big] = \lambda_i = \exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}.$$

Interpretation of regression parameters

Let $\boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. Denote $\lambda_0 = \mathsf{E}\big[Y_i \,\big|\, X_2 = \cdots = X_p = 0\big]$. Then $\log \lambda_0 = \beta_1$ so $\mathrm{e}^{\boldsymbol{\beta}_1}$ is the expected value of Y_i for an individual with zero values in all covariates.

Now consider two individuals with observed covariates $\boldsymbol{x}^0 = (1, x_2, \ldots, x_p)^\mathsf{T}$ and $\boldsymbol{x}^j = (1, x_2, \ldots, x_j + 1, \ldots, x_p)^\mathsf{T}$ (the j-th covariate is increased by 1, the others are the same). Denote $\lambda_{X0} = \mathsf{E}\big[Y_i \,\big|\, \boldsymbol{X} = \boldsymbol{x}^0\big]$ and $\lambda_{Xj} = \mathsf{E}\big[Y_i \,\big|\, \boldsymbol{X} = \boldsymbol{x}^j\big]$. Then

$$\beta_j = \log \frac{\lambda_{Xj}}{\lambda_{X0}} \qquad \text{and} \qquad \mathrm{e}^{\beta_j} = \frac{\lambda_{Xj}}{\lambda_{X0}}.$$

Thus e^{β_j} is the proportional increase in $\mathsf{E}\, Y_i$ per unit difference in the covariate X_j. When $\beta_j = 0$ the ratio is 1 and the covariate has no effect on the expectation given the other covariates.

Estimation of parameters

The likelihood is

$$L_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \prod_{i=1}^{n} \exp\{Y_i \log \lambda_i - \lambda_i - \log Y_i!\}$$

log-likelihood

$$\ell_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} (Y_i \log \lambda_i - \lambda_i - \log Y_i!).$$

By Theorem 2.3, the score statistic with the canonical link is

$$\boldsymbol{U}_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = \sum_{i=1}^{n} (Y_i - \lambda_i)\boldsymbol{X}_i$$

and $\widehat{\boldsymbol{\beta}}_n$ solves the equations

$$\sum_{i=1}^{n} Y_i \boldsymbol{X}_i = \sum_{i=1}^{n} \widehat{\lambda}_i \boldsymbol{X}_i,$$

where

$$\widehat{\lambda}_i = \exp\{\boldsymbol{X}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}_n\}.$$

The MLE of $\boldsymbol{\beta}$ is calculated by the IWLS algorithm

$$\widehat{\boldsymbol{\beta}}_n^{(k+1)} = (\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1} (\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}}^{(k)} \widehat{\boldsymbol{Z}}^{(k)})$$

with

$$\hat{\mathbb{W}}^{(k)} = \operatorname{diag}(\widehat{\lambda}_1^{(k)}, \ldots, \widehat{\lambda}_n^{(k)})$$

and $\widehat{\boldsymbol{Z}}^{(k)} = (\widehat{Z}_1^{(k)}, \ldots, \widehat{Z}_n^{(k)})^{\mathsf{T}}$, where

$$\widehat{Z}_i^{(k)} = \boldsymbol{X}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}^{(k)} + \frac{Y_i - \widehat{\lambda}_i^{(k)}}{\widehat{\lambda}_i^{(k)}}.$$

The information matrix is $I(\boldsymbol{\beta}) = \mathsf{E}_{\boldsymbol{X}} \lambda_i \boldsymbol{X}_i^{\otimes 2}$, which can be estimated by

$$\widehat{I}_n = \frac{1}{n} \mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}} \mathbb{X} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\lambda}_i \boldsymbol{X}_i^{\otimes 2}.$$

The estimated variance of $\widehat{\boldsymbol{\beta}}_n$ is $(\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}} \mathbb{X})^{-1}$.

Deviance

The MLEs of the saturated model parameters are $\widetilde{\lambda}_i = Y_i$ The deviance for loglinear model is

$$D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}}) = 2\sum_{i=1}^{n}\Big[Y_i \log\frac{Y_i}{\widehat{\lambda}_i} - (Y_i - \widehat{\lambda}_i)\Big]. \tag{3.2}$$

According to Theorem 2.9(iii), the difference in deviances between a submodel and a wider model has a limiting χ^2 distribution if the submodel holds. This is used for loglinear model building.

Pearson X^2

The Pearson residual is

$$r_i^P = \frac{Y_i - \widehat{\lambda}_i}{\sqrt{\widehat{\lambda}_i}},$$

the Pearson X^2 statistic is

$$X^2 = \sum_{i=1}^{n}\frac{(Y_i - \widehat{\lambda}_i)^2}{\widehat{\lambda}_i}. \tag{3.3}$$

3.2.2 Modelling Poisson process intensity

Let Y_i be a number of events observed during t_i units of time on a certain individual (experimental unit), let $\boldsymbol{X}_i$ be covariates. We have independent realizations of $(Y_i, t_i, \boldsymbol{X}_i)$ for $i = 1, \ldots, n$.

Suppose $Y_i \sim \mathsf{Po}(\lambda_i t_i)$, that is, the outcomes are generated by a homogeneous Poisson process with intensity λ_i. Let λ_i depend on the covariates $\boldsymbol{X}_i$ through the identity $\log\lambda_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$. Then

$$\mathsf{E}\big[Y_i \,\big|\, \boldsymbol{X}_i\big] = \mathsf{var}\big[Y_i \,\big|\, \boldsymbol{X}_i\big] = t_i\lambda_i = t_i\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\} = \exp\{\log t_i + \beta_1 + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}\}.$$

The intensity $\lambda_i = \mathsf{E}\, Y_i/t_i$ describes the expected number of events observed during a unit time interval.

This model can be fitted as a Poisson loglinear model with the term $\log t_i$ added to the linear predictor. It is a term similar to a covariate, except that its parameter is equal to 1 by default and is not estimated. Such a term is called *an offset*[*] in the GLM terminology. Estimation and testing proceeds in exactly the same way as if offset was not present.

[*] Česky *offset*

3.3 Loglinear Models for Contingency Tables

3.3.1 Two-way contingency table

Consider discrete random variables $X \in \{1, \dots, I\}$ and $Z \in \{1, \dots, J\}$. Observe n independent realizations $(X_1, Z_1), \dots, (X_n, Z_n)$ of this pair. Denote the observed count of the pair $(X = i, Z = j)$ by $n_{ij} = \sum_{l=1}^{n} I(X_l = i, Z_l = j)$. The observed counts (also called *frequencies*) can be arranged into a two-way contingency table

	$Z = 1$	$\cdots$	$Z = J$	Total
$X = 1$	n_{11}	$\cdots$	n_{1J}	n_{1+}
$\vdots$	$\vdots$		$\vdots$	$\vdots$
$X = I$	n_{I1}	$\cdots$	n_{IJ}	n_{I+}
Total	n_{+1}	$\cdots$	n_{+J}	$n_{++} = n$

where $n_{i+} = \sum_{j=1}^{J} n_{ij}$ and $n_{+j} = \sum_{i=1}^{I} n_{ij}$.

Denote the expected cell frequencies

$$m_{ij} = \mathsf{E}\, n_{ij}, \qquad m_{i+} = \sum_{j=1}^{J} m_{ij}, \qquad m_{+j} = \sum_{i=1}^{I} m_{ij},$$

and $m_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} m_{ij}$. The cell probabilities are be

$$\pi_{ij} = \mathrm{P}\left[X = i, Z = j\right], \qquad \pi_{i+} = \sum_{j=1}^{J} \pi_{ij} = \mathrm{P}\left[X = i\right], \qquad \pi_{+j} = \sum_{i=1}^{I} \pi_{ij} = \mathrm{P}\left[Z = j\right].$$

Obviously, $\pi_{++} = \sum_{i=1}^{I} \sum_{j=1}^{J} \pi_{ij} = 1$.

The expected frequencies are related to the cell probabilities as follows (we allow n to be random):

$$m_{ij} = \mathsf{E}\, n_{ij} = \mathsf{E} \sum_{l=1}^{n} I(X_l = i, Z_l = j) = \mathsf{E}\,\mathsf{E}\big[\sum_{l=1}^{n} I(X_l = i, Z_l = j) \,\big|n\big] = \mathsf{E}\, n\pi_{ij} = m_{++}\pi_{ij},$$

so $\pi_{ij} = m_{ij}/m_{++}$.

The goal is to use the observed counts n_{ij} to model the cell probabilities π_{ij}, investigate the marginal distributions of X and Z and the associations between X and Z.

3.3.2 Distributions of observed counts

There are several distributions we can use for observed counts in the contingency table.

Poisson distribution

Let $n_{11}, \ldots, n_{IJ}$ be independent random variables with Poisson distributions $n_{ij} \sim \mathsf{Po}(m_{ij})$. It follows $\mathsf{E}\, n_{ij} = \mathsf{var}\, n_{ij} = m_{ij}$. The joint density of the whole table is

$$\mathrm{P}\left[n_{11} = k_{11}, \ldots, n_{IJ} = k_{IJ}\right] = \prod_{i,j} \frac{1}{k_{ij}!} m_{ij}^{k_{ij}} \mathrm{e}^{-m_{ij}}, \qquad k_{ij} = 0, 1, 2, \ldots.$$

The total number of observations $n = \sum_{i,j} n_{ij}$ is a random variable with the distribution $\mathsf{Po}(m_{++})$.

The asymptotics does not work by observing an increasing number of independent Poisson variables (the total IJ is fixed) but by letting $m_{++} \to \infty$. The asymptotic MLE theory for iid data does not apply to this case.

Multinomial distribution

Let the vector $(n_{11}, \ldots, n_{IJ})$ follow the multinomial distribution $\mathsf{Mult}_{IJ}(n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_{11}, \ldots, \pi_{IJ})^{\mathsf{T}}$. The joint density of the whole table is

$$\mathrm{P}\left[n_{11} = k_{11}, \ldots, n_{IJ} = k_{IJ}\right] = n! \prod_{i,j} \frac{1}{k_{ij}!} \pi_{ij}^{k_{ij}}, \qquad k_{ij} = 0, 1, \ldots, n, \quad \sum_{i,j} k_{ij} = n.$$

The total number of observations n is fixed, $n_{ij} \sim \mathsf{Bi}(n, \pi_{ij})$, $\mathsf{E}\, n_{ij} = n\pi_{ij}$, $\mathsf{var}\, n_{ij} = n\pi_{ij}(1 - \pi_{ij})$, the counts are not independent.

The contingency table can be expressed by summing n iid random vectors, each with distribution $\mathsf{Mult}_{IJ}(1, \boldsymbol{\pi})$. The asymptotics works through letting $n \to \infty$. The asymptotic MLE theory for iid data applies to this case.

Row multinomial distribution

Let the vectors $(n_{i1}, \ldots, n_{iJ})$, $i = 1, \ldots, I$, be iid with the multinomial distribution $\mathsf{Mult}_J(n_{i+}, \boldsymbol{\pi}_i)$, where $\boldsymbol{\pi}_i = (\pi_{i1}/\pi_{i+}, \ldots, \pi_{iJ}/\pi_{i+})^{\mathsf{T}}$. The joint density of the whole table is

$$\mathrm{P}\left[n_{11} = k_{11}, \ldots, n_{IJ} = k_{IJ}\right] = \prod_i n_{i+}! \prod_j \frac{1}{k_{ij}!} \left(\frac{\pi_{ij}}{\pi_{i+}}\right)^{k_{ij}}, \quad k_{ij} = 0, 1, \ldots, n, \; \sum_j k_{ij} = n_{i+}.$$

The numbers of observations n_{i+} in the I rows of the table are fixed, $n_{ij} \sim \mathsf{Bi}(n_{i+}, \frac{\pi_{ij}}{\pi_{i+}})$, $\mathsf{E}\, n_{ij} = n_{i+}\frac{\pi_{ij}}{\pi_{i+}}$, $\mathsf{var}\, n_{ij} = n_{i+}\frac{\pi_{ij}}{\pi_{i+}}(1 - \frac{\pi_{ij}}{\pi_{i+}})$, the counts are independent between rows but dependent within rows.

The asymptotics works through letting $n_{i+} \to \infty$ for all i at the same rate. The asymptotic MLE theory for iid data applies to this case.

Equivalence of Poisson and multinomial models

We start with a result stating that Poisson and multinomial distributions are related through conditioning on the total count.

Lemma 3.1. *Let $X_i \sim \mathsf{Po}(\lambda_i)$ be independent random variables, $i = 1, \ldots, n$. Then the conditional joint distribution of the random vector $(X_1, \ldots, X_n)^\mathsf{T}$ given $\sum_{i=1}^n X_i = s$ is $\mathsf{Mult}_n(s, \boldsymbol{p})$, where $\boldsymbol{p} = (p_1, \ldots, p_n)^\mathsf{T}$ and $p_i = \lambda_i / \sum_{j=1}^n \lambda_j$.*

Corollary. Let $n_{ij} \sim \mathsf{Po}(m_{ij})$ be independent, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Then:

- The conditional joint distribution of $(n_{11}, \ldots, n_{IJ})^\mathsf{T}$ given $n_{++} = n$ is $\mathsf{Mult}_{IJ}(n, \boldsymbol{\pi})$, where the components of $\boldsymbol{\pi}$ are $\pi_{ij} = m_{ij}/m_{++}$ ($\Rightarrow$ *multinomial distribution*).
- The conditional joint distribution of $(n_{i1}, \ldots, n_{iJ})^\mathsf{T}$ given n_{i+} is $\mathsf{Mult}_J(n_{i+}, \boldsymbol{\pi}_i)$, where the components of $\boldsymbol{\pi}_i$ are $\pi_{ij} = m_{ij}/m_{i+} = \pi_{ij}/\pi_{i+}$ ($\Rightarrow$ *row multinomial distribution*).

The corollary states that both multinomial and row multinomial distributions of a contingency table can be obtained from Poisson distribution by conditioning on some observed totals.

Assume that the loglinear model holds for the expected frequencies m_{ij}, in particular,

$$\log \mathsf{E}\, n_{ij} = \alpha + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{ij} \quad \text{or} \quad m_{ij} = \mathrm{e}^{\alpha + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{ij}},$$

where α is the intercept and $\boldsymbol{X}_{ij}$ is a vector of covariates characterizing the (i, j)-th cell. The maximum dimension of $\boldsymbol{X}_{ij}$ is $IJ - 1$. The cell probabilities π_{ij} can be expressed as follows:

$$\pi_{ij} = \frac{m_{ij}}{m_{++}} = \frac{\mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{ij}}}{\sum_{k,l} \mathrm{e}^{\boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{kl}}}. \tag{3.4}$$

Theorem 3.2. *The likelihood functions for estimation of parameters $\boldsymbol{\beta}$ in the loglinear model $\log m_{ij} = \alpha + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{ij}$ arising from Poisson or multinomial sampling distributions are equivalent (they differ only by a multiplicative constant that does not depend on $\boldsymbol{\beta}$).*

Note. Theorem 3.2 does not deal with estimation of the intercept α – in fact, the intercept is not even identifiable in the multinomial model. This is obvious from expression (3.4).

Theorem 3.2 can be extended to row multinomial distribution as follows:

Theorem 3.3. *(Palmgren 1981) The likelihood functions for estimation of parameters $\boldsymbol{\beta}$ in the loglinear model $\log m_{ij} = \alpha_i + \boldsymbol{\beta}^\mathsf{T} \boldsymbol{X}_{ij}$ arising from Poisson or row-multinomial sampling distributions are equivalent (they differ only by a multiplicative constant that does not depend on $\boldsymbol{\beta}$).*

Note. Row multinomial sampling requires row-specific intercept in the loglinear model.

Corollary. Expressions for any quantity derived from the likelihood function for $\boldsymbol{\beta}$ (score function, information matrix, the MLE) and their properties (asymptotic distributions, test statistics, confidence intervals) are the same no matter which of the three distributions generated the contingency table.

When the data are generated by the Poisson model, they can be transformed to the multinomial model by conditioning on the observed cell count total $n = n_{++}$. The asymptotic results hold in the multinomial model (the data are equivalent to n independent observations from Mult_1). The formulae can be derived from the theory of GLM for the loglinear model with Poisson distribution.

In the rest of this section we assume the loglinear model with Poisson distribution but the results apply to multinomial models without change.

3.3.3 Loglinear models for two-way tables

The independence model (X, Z)

The model (X, Z) for expected cell counts is defined by the equation

$$\log m_{ij} = \alpha + \beta_i^X + \beta_j^Z \tag{3.5}$$

with the constraints $\beta_1^X = \beta_1^Z = 0$. The regressors that generate this model are

$$\begin{aligned}
\boldsymbol{X}_{11} &= (1, 0, \ldots, 0, \ldots, 0, \ldots, 0, \ldots, 0)^\mathsf{T} && \text{for the cell } (1,1),\\
\boldsymbol{X}_{i1} &= (1, 0, \ldots, 1, \ldots, 0, \ldots, 0, \ldots, 0)^\mathsf{T} && \text{for the cell } (i,1),\ i \neq 1,\\
\boldsymbol{X}_{1j} &= (1, 0, \ldots, 0, \ldots, 0, \ldots, 1, \ldots, 0)^\mathsf{T} && \text{for the cell } (1,j),\ j \neq 1,\\
\boldsymbol{X}_{ij} &= (1, 0, \ldots, 1, \ldots, 0, \ldots, 1, \ldots, 0)^\mathsf{T} && \text{for the cell } (i,j),\ i \neq 1,\ j \neq 1.
\end{aligned}$$

The parameters are

$$\boldsymbol{\beta} = (\alpha, \beta_2^X, \ldots, \beta_I^X, \beta_2^Z, \ldots, \beta_J^Z)^\mathsf{T}.$$

the dimension of the parameter vector is $p = 1 + I - 1 + J - 1 = I + J - 1$.

The parameter α can be expressed as

$$\alpha = \log m_{11} = \log m_{++} + \log \pi_{11},$$

hence the model can be also stated in terms of cell probabilities

$$\log \pi_{ij} = \log \frac{m_{ij}}{m_{++}} = \log \pi_{11} + \beta_i^X + \beta_j^Z. \tag{3.6}$$

The interpretation of the regression parameters is as follows:

$$\mathrm{e}^{\beta_i^X} = \frac{\pi_{i+}}{\pi_{1+}} = \frac{\mathrm{P}\left[X = i\right]}{\mathrm{P}\left[X = 1\right]} \quad \text{and} \quad \mathrm{e}^{\beta_j^Z} = \frac{\pi_{+j}}{\pi_{+1}} = \frac{\mathrm{P}\left[Z = j\right]}{\mathrm{P}\left[Z = 1\right]},$$

that is, $e^{\beta_i^X}$ is the odds of observing the i-th level of the variable X compared to observing the first level.

The cell probabilities satisfy the equation

$$\pi_{ij} = \pi_{i+}\pi_{+j}$$

for all i and j, which means that the model (X, Z) holds if and only if the variables X and Z are independent.

The fitted cell frequencies in this model can be expressed explicitely

$$\widehat{m}_{ij} = n\widehat{\pi}_{ij} = n\widehat{\pi}_{i+}\widehat{\pi}_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

because $\frac{n_{i+}}{n}$ and $\frac{n_{+j}}{n}$ are the MLE's of π_{i+} and π_{+j}.

The interaction model (XZ)

The model (XZ) for expected cell counts is defined by the equation

$$\log m_{ij} = \alpha + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ} \tag{3.7}$$

with the constraints $\beta_1^X = \beta_1^Z = 0$, $\beta_{i1}^{XZ} = 0$ for all $i = 1, \ldots, I$, and $\beta_{1j}^{XZ} = 0$ for all $j = 1, \ldots, J$. The number of parameters in this model is $p = I+J-1+(I-1)(J-1) = IJ$. This is the saturated model, hence the estimated expected cell counts $\widehat{m}_{ij}$ (fitted values) are equal to the observed cell counts n_{ij} for all i, j.

The equivalent model for cell probabilities is

$$\log \pi_{ij} = \log \pi_{11} + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}. \tag{3.8}$$

The interpretation of the main effects is as follows:

$$e^{\beta_i^X} = \frac{\pi_{i1}}{\pi_{11}} = \frac{\mathrm{P}\big(X = i \,\big|\, Z = 1\big)}{\mathrm{P}\big(X = 1 \,\big|\, Z = 1\big)} \quad \text{and} \quad e^{\beta_j^Z} = \frac{\pi_{1j}}{\pi_{11}} = \frac{\mathrm{P}\big(X = 1 \,\big|\, Z = j\big)}{\mathrm{P}\big(X = 1 \,\big|\, Z = 1\big)},$$

that is, $e^{\beta_i^X}$ is the odds of observing the i-th level of the variable X compared to observing the first level of X when $Z = 1$.

The interaction parameters can be written as

$$e^{\beta_{ij}^{XZ}} = \frac{\pi_{ij}\pi_{11}}{\pi_{i1}\pi_{1j}} = \frac{\mathrm{P}\left[X = i, Z = j\right]\mathrm{P}\left[X = 1, Z = 1\right]}{\mathrm{P}\left[X = i, Z = 1\right]\mathrm{P}\left[X = 1, Z = j\right]} = \frac{\mathrm{P}\big(X = i \,\big|\, Z = j\big)\mathrm{P}\big(X = 1 \,\big|\, Z = 1\big)}{\mathrm{P}\big(X = 1 \,\big|\, Z = j\big)\mathrm{P}\big(X = i \,\big|\, Z = 1\big)},$$

which is the odds ratio in the 2×2 sub-table that includes the first and the i-th rows and the first and j-th columns from the original table. The odds ratio expresses the

proportional change in the odds of the event $X = i$ (relative to $X = 1$) when Z changes from 1 to j.

The cell probabilities can be expressed as

$$\pi_{ij} = \mathrm{P}\left[X = i, Z = j\right] = \frac{\mathrm{e}^{\beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}}}{\sum_{k,l} \mathrm{e}^{\beta_k^X + \beta_l^Z + \beta_{kl}^{XZ}}}.$$

Since this is the saturated model, its deviance $D(XZ)$ is zero. The MLE theory holds for this model because all covariates are discrete and the number of parameters $p = IJ$ is constant.

3.3.4 Testing independence in a two-way table

The variables X and Z are independent if and only if the model (X, Z) holds, that is, all the interaction parameters are zero. So, a test of independence should test all $(I-1)(J-1)$ interaction parameters simultaneously.

Such a test can be based on the deviance $D(X, Z)$ of the independence model. From (3.2), the deviance is

$$D(X, Z) = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \Big[n_{ij} \log \frac{n_{ij}}{\widehat{m}_{ij}} - (n_{ij} - \widehat{m}_{ij}) \Big],$$

where $\widehat{m}_{ij} = n_{i+} n_{+j} / n$. If independence holds then $D(X, Z) \xrightarrow{\mathsf{D}} \chi^2_{(I-1)(J-1)}$, so the hypothesis is rejected at the level of α when $D(X, Z) \geq \chi^2_{(I-1)(J-1)}(1 - \alpha)$.

The independence hypothesis can be also tested by the Pearson X^2 statistic. By (3.3), the Pearson X^2 statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \widehat{m}_{ij})^2}{\widehat{m}_{ij}}.$$

This is the classical χ^2 statistic for testing independence in a two-way table. Under the null hypothesis, it also converges in distribution to $\chi^2_{(I-1)(J-1)}$.

3.3.5 Loglinear models for three-way tables

Three-way contingency table

Consider categorical random variables $X \in \{1, \ldots, I\}$, $Z \in \{1, \ldots, J\}$, and $V \in \{1, \ldots, K\}$. Observe n independent realizations $(X_1, Z_1, V_1), \ldots, (X_n, Z_n, V_n)$ of this triplet. Denote

the observed count of the outcome $(X = i, Z = j, V = k)$ by

$$n_{ijk} = \sum_{l=1}^{n} I(X_l = i, Z_l = j, V_l = k).$$

The observed counts n_{ijk} form a three-way contingency table.

Let $n_{ij+} = \sum_{k=1}^{K} n_{ijk}$, $n_{i++} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk}$, etc. Denote $m_{ijk} = \mathsf{E}\, n_{ijk}$ and $\pi_{ijk} = \mathrm{P}\,[X = i, Z = j, V = k]$. The symbols m_{ij+}, m_{++k}, π_{i+k}, π_{+j+} etc. all have the obvious meaning (summing over the indices replaced by +). Obviously, $\pi_{+++} = \sum_{i,j,k} \pi_{ijk} = 1$.

Marginal and conditional associations in a three-way table

Suppose we would like to describe the associations between the variables X and Z in the presence of a third classifier V. We have two ways to do this:

Marginal associations

Let us ignore the existence of the third variable. We can form a two-way contingency table for X and Z by collapsing the original three-way table n_{ijk} across the levels of V, getting a two-way table n_{ij+}. Now the associations between X and Z can be described and investigated by the methods for a two-way table described in Section 3.3.3 applied to n_{ij+}.

Definition 3.1. Discrete variables X and Z are called *marginally independent* (in the presence of a third discrete variable V) if and only if

$$\pi_{ij+} = \pi_{i++}\pi_{+j+} \quad \text{for all } i, j.$$

Marginal associations between X and Z are described by marginal odds ratios, which correspond to exponentiated interaction terms in the model (3.8) applied to the collapsed two-way table.

Definition 3.2. The marginal odds ratio for the i-th level of X and the j-th level of Z are defined as

$$\theta_{ij}^{XZ} = \frac{\pi_{ij+}\pi_{11+}}{\pi_{i1+}\pi_{1j+}}.$$

X and Z are marginally independent if and only if $\theta_{ij}^{XZ} = 1$ for all i and j.

Conditional associations

Let us fix the value of variable V at some value $k \in \{1, \ldots, K\}$. We get K separate two-way contingency tables for X and Z by considering each of the layers of the three-way table n_{ijk} formed by fixing $V = k$. The k-th two-way table has observed counts

n_{ijk} $(i = 1, \ldots, I,\ j = 1, \ldots, J)$ and cell probabilities proportional to π_{ijk}. The associations between X and Z can be described and investigated by the methods described in Section 3.3.3 applied to n_{ijk} for each fixed k.

Definition 3.3. Variables X and Z are called *conditionally independent given* V if and only if

$$\mathrm{P}\big(X = i, Z = j \,\big|V = k\big) = \mathrm{P}\big(X = i \,\big|V = k\big)\mathrm{P}\big(Z = j \,\big|V = k\big) \quad \text{for all } i, j, k.$$

Theorem 3.4. *Variables X and Z are conditionally independent given V if and only if*

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \quad \textit{for all } i, j, k.$$

Conditional associations between X and Z are described by conditional odds ratios.

Definition 3.4. The conditional odds ratios for the i-th level of X and the j-th level of Z given the k-th level of V are defined as

$$\theta^{XZ}_{ij(k)} = \frac{\pi_{ijk}\pi_{11k}}{\pi_{i1k}\pi_{1jk}}.$$

X and Z are conditionally independent if and only if $\theta^{XZ}_{ij(k)} = 1$ for all i, j, and k.

Comments

Marginal and conditional associations can be quite different. In each application, the analyst must decide whether marginal or conditional associations are more relevant for the particular case. The decision is based on the context of the problem and the desired interpretation of the results.

The variable V usually acts as *a confounder*. A confounder is a variable that affects both X and Z and thus masks the true relationship between X and Z. In the conditional analysis, we adjust for the confounding effect of V and reveal the true association of X and Z. Therefore, the conditional associations are the correct approach in this case.

However, it may happen that V is a part of the causal pathway between X and Z (X affects V and V affects Z). Then V cannot be considered a confounder. The conditional analysis of the association between X and Z removes the part of the effect that is mediated through V and thus gives misleading results about the true association. In this case, the marginal associations provide the correct approach to the problem.

The independence model (X, Z, V)

The model (X, Z, V) is defined by the equation

$$\log m_{ijk} = \alpha + \beta^X_i + \beta^Z_j + \beta^V_k \tag{3.9}$$

with the constraints $\beta_1^X = \beta_1^Z = \beta_1^V = 0$.

In terms of cell probabilities,

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V . \tag{3.10}$$

The dimension of the parameter vector is $p = I + J + K - 2$.

The interpretation of the regression parameters is as follows:

$$\mathrm{e}^{\beta_i^X} = \frac{\pi_{i++}}{\pi_{1++}} = \frac{\mathrm{P}\left[X = i\right]}{\mathrm{P}\left[X = 1\right]},\ \mathrm{e}^{\beta_j^Z} = \frac{\pi_{+j+}}{\pi_{+1+}} = \frac{\mathrm{P}\left[Z = j\right]}{\mathrm{P}\left[Z = 1\right]},\ \mathrm{e}^{\beta_k^V} = \frac{\pi_{++k}}{\pi_{++1}} = \frac{\mathrm{P}\left[V = k\right]}{\mathrm{P}\left[V = 1\right]}.$$

The cell probabilities satisfy the equation

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$$

for all i, j, k. The model (X, Z, V) holds if and only if the variables X, Z and V are both marginally and conditionally independent.

Model (XV, Z)

The model (XV, Z) for expected cell counts is defined by the equation

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} \tag{3.11}$$

with the constraints $\beta_1^X = \beta_1^Z = \beta_1^V = 0$, $\beta_{i1}^{XV} = 0$ for all i, and $\beta_{1k}^{XV} = 0$ for all k. The number of parameters in this model is $p = IK + J - 1$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} . \tag{3.12}$$

The interpretation of the main effects:

$$\mathrm{e}^{\beta_i^X} = \frac{\pi_{i+1}}{\pi_{1+1}} = \frac{\mathrm{P}\big(X = i \,\big|V = 1\big)}{\mathrm{P}\big(X = 1 \,\big|V = 1\big)},\ \mathrm{e}^{\beta_k^V} = \frac{\pi_{1+k}}{\pi_{1+1}} = \frac{\mathrm{P}\big(V = k \,\big|X = 1\big)}{\mathrm{P}\big(V = 1 \,\big|X = 1\big)},$$
$$\mathrm{e}^{\beta_j^Z} = \frac{\pi_{+j+}}{\pi_{+1+}} = \frac{\mathrm{P}\left[Z = j\right]}{\mathrm{P}\left[Z = 1\right]}.$$

It can be easily shown that

$$\mathrm{e}^{\beta_{ik}^{XV}} = \theta_{ik}^{XV} = \theta_{ik(j)}^{XV},$$

which is both the marginal and the conditional odds ratio between X and V.

The cell probabilities satisfy the equation

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}$$

for all i, j, k. Hence in this model,

- the pair (X, V) is independent of Z,
- X is both marginally and conditionally independent of Z,
- X is associated with V, and
- the marginal and conditional associations between X and V agree.

Model (XV, ZV)

The model (XV, ZV) for expected cell counts is defined by the equation

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} \tag{3.13}$$

with the constraints $\beta_1^X = \beta_1^Z = \beta_1^V = 0$, $\beta_{i1}^{XV} = 0$ for all i, $\beta_{1k}^{XV} = 0$ for all k, $\beta_{j1}^{ZV} = 0$ for all j, and $\beta_{1k}^{ZV} = 0$ for all k. The number of parameters in this model is $p = K(I + J - 1)$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV}. \tag{3.14}$$

The interpretation of the main effects:

$$\mathrm{e}^{\beta_i^X} = \frac{\pi_{i+1}}{\pi_{1+1}} = \frac{\mathrm{P}\big(X = i \,\big|\, V = 1\big)}{\mathrm{P}\big(X = 1 \,\big|\, V = 1\big)},\ \mathrm{e}^{\beta_j^Z} = \frac{\pi_{+j1}}{\pi_{+11}} = \frac{\mathrm{P}\big(Z = j \,\big|\, V = 1\big)}{\mathrm{P}\big(Z = 1 \,\big|\, V = 1\big)},$$
$$\mathrm{e}^{\beta_k^V} = \frac{\pi_{11k}}{\pi_{111}} = \frac{\mathrm{P}\big(V = k \,\big|\, X = 1, Z = 1\big)}{\mathrm{P}\big(V = 1 \,\big|\, X = 1, Z = 1\big)}.$$

The cell probabilities satisfy the equation

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}$$

for all i, j, k.

The association between X and Z:

Conditional OR for $X \leftrightarrow Z$: $\quad \theta_{ij(k)}^{XZ} = \dfrac{\pi_{ijk}\pi_{11k}}{\pi_{i1k}\pi_{1jk}} = 1$

Marginal OR for $X \leftrightarrow Z$: $\quad \theta_{ij}^{XZ} = \dfrac{\pi_{ij+}\pi_{11+}}{\pi_{i1+}\pi_{1j+}} \neq 1$

The association between X and V:

Conditional OR for $X \leftrightarrow V$: $\quad \theta_{ik(j)}^{XV} = \dfrac{\pi_{ijk}\pi_{1j1}}{\pi_{ij1}\pi_{1jk}} = \mathrm{e}^{\beta_{ik}^{XV}}$

Marginal OR for $X \leftrightarrow V$: $\quad \theta_{ik}^{XV} = \dfrac{\pi_{i+k}\pi_{1+1}}{\pi_{i+1}\pi_{1+k}} = \mathrm{e}^{\beta_{ik}^{XV}}$

The association between Z and V is similar to that between X and V (rotate indices). In particular,

$$\theta^{ZV}_{jk(i)} = \theta^{ZV}_{jk} = \mathrm{e}^{\beta^{ZV}_{jk}}.$$

Hence in this model,

- all three variables are marginally dependent,
- X is conditionally independent of Z given V (but marginally dependent),
- X is associated with V and the marginal and conditional associations between X and V agree,
- Z is associated with V and the marginal and conditional associations between Z and V agree.

Model (XV, ZV, XZ)

The model (XV, ZV, XZ) for expected cell counts is defined by the equation

$$\log m_{ijk} = \alpha + \beta^X_i + \beta^Z_j + \beta^V_k + \beta^{XV}_{ik} + \beta^{ZV}_{jk} + \beta^{XZ}_{ij} \tag{3.15}$$

with the constraints $\beta^X_1 = \beta^Z_1 = \beta^V_1 = 0$, $\beta^{XV}_{i1} = 0$ for all i, $\beta^{XV}_{1k} = 0$ for all k, $\beta^{ZV}_{j1} = 0$ for all j, $\beta^{ZV}_{1k} = 0$ for all k, $\beta^{XZ}_{i1} = 0$ for all i, and $\beta^{XZ}_{1j} = 0$ for all j. The number of parameters in this model is $p = KI + KJ + IJ - (I + J + K) + 1$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta^X_i + \beta^Z_j + \beta^V_k + \beta^{XV}_{ik} + \beta^{ZV}_{jk} + \beta^{XZ}_{ij} \tag{3.16}$$

The interpretation of the main effects:

$$\mathrm{e}^{\beta^X_i} = \frac{\pi_{i11}}{\pi_{111}},\ \mathrm{e}^{\beta^Z_j} = \frac{\pi_{1j1}}{\pi_{111}},\ \mathrm{e}^{\beta^V_k} = \frac{\pi_{11k}}{\pi_{111}}.$$

The interaction parameters determine the conditional odds ratios:

Conditional OR for $X \leftrightarrow Z$:	$\theta^{XZ}_{ij(k)} = \mathrm{e}^{\beta^{XZ}_{ij}}$
Conditional OR for $X \leftrightarrow V$:	$\theta^{XV}_{ik(j)} = \mathrm{e}^{\beta^{XV}_{ik}}$
Conditional OR for $Z \leftrightarrow V$:	$\theta^{ZV}_{jk(i)} = \mathrm{e}^{\beta^{ZV}_{jk}}$

Marginal odds ratios are all different from conditional odds ratios and do not have nice expressions. There is no decomposition of the cell probabilities π_{ijk}.

In this model,

- all three variables are marginally and conditionally dependent,
- marginal associations are different from conditional associations,
- conditional associations do not depend on the value of the conditioning variable.

Model (XZV)

The model (XZV) for expected cell counts is defined by the equation

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} + \beta_{ijk}^{XZV} \tag{3.17}$$

with the usual constraints (any parameter with 1 anywhere among the indices is set to 0). The number of parameters in this model is $p = IJK$. It is the saturated model.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} + \beta_{ijk}^{XZV} \tag{3.18}$$

The interpretation of the main effects is the same as in the model (XV, ZV, XZ). The second-order interaction parameters determine the conditional odds ratios at the first level of the conditioning variable:

$$\begin{aligned}
&\text{Conditional OR for } X \leftrightarrow Z \text{ given } V = 1: && \theta_{ij(1)}^{XZ} = \mathrm{e}^{\beta_{ij}^{XZ}} \\
&\text{Conditional OR for } X \leftrightarrow V \text{ given } Z = 1: && \theta_{ik(1)}^{XV} = \mathrm{e}^{\beta_{ik}^{XV}} \\
&\text{Conditional OR for } Z \leftrightarrow V \text{ given } X = 1: && \theta_{jk(1)}^{ZV} = \mathrm{e}^{\beta_{jk}^{ZV}}.
\end{aligned}$$

The third-order interaction parameters determine how the conditional odds ratios change if the conditioning is not on the first level:

$$\mathrm{e}^{\beta_{ijk}^{XZV}} = \frac{\theta_{ij(k)}^{XZ}}{\theta_{ij(1)}^{XZ}} = \frac{\theta_{ik(j)}^{XV}}{\theta_{ik(1)}^{XV}} = \frac{\theta_{jk(i)}^{ZV}}{\theta_{jk(1)}^{ZV}}.$$

Marginal odds ratios are all different from conditional odds ratios and do not have nice expressions. There is no decomposition of the cell probabilities π_{ijk}.

This model is similar to (XV, ZV, XZ) except that conditional associations do depend on the value of the conditioning variable.

Model selection

The model selection strategy that usually works best is to fit the saturated model (XZV) first and then test whether the higher order interactions can be removed, starting with the three-way interaction. Model selection should be hierarchical (do not remove lower order terms if a higher order term involving the same variable is still in the model). Deviance tests are used to decide whether a term can be removed. Because the saturated model satisfies the MLE theory assumptions, the deviance of each fitted model can be compared with a quantile of a suitable χ^2 distribution to check whether the current model fits well.

The final model that cannot be further reduced reveals the structure of the associations between the three variables, as explained above.

3.3.6 Loglinear models for multi-way tables

Theoretically, a loglinear model can fit contingency tables of arbitrarily high dimension. The interpretation of the main effects, two-way and three-way interaction remains the same as in a three-way table, unless there are interactions of even higher order.

The dependence structure among multiple categorical variables can be deduced from an undirected graph where each variable plays the role of a node and two-way interactions are the edges between the nodes. If there is at least one path connecting two nodes, then the two variables corresponding to the nodes are marginally dependent. If all the paths connecting the two nodes can be interrupted by removing a certain set of nodes from the graph, then the two variables are conditionally independent given the set of variables corresponding to the removed nodes.

Notes on model selection for multi-way tables

- Model building proceeds by performing deviance tests comparing a model vs. a submodel.
- The number of possible models in a multi-way table can be very large. There is no way to fit all of them.
- The starting model cannot be too complex. In most cases, we cannot start by the saturated model. Interactions of the fourth or higher orders are difficult to interpret and we try to avoid them even if they seem to be statistically significant. A reasonable strategy could be to start with a model containing all three-way interactions, test its goodness-of-fit using deviance, and remove the insignificant three-way (and then two-way) interactions in a backward step-wise procedure. It is better to do this interactively rather than to use some automated model-building procedure.
- Multi-way tables require a lot of data. If there are four or five factors with a moderate number of levels, the number of cells in the table is huge and the observed counts can be quite low even if the total number of observations is in the thousands. If too many of the fitted cell counts are below 5, the asymptotic approximations tend to be unreliable. The analyst must make sure that there are enough observations in the cells, otherwise some of the variables must be removed from the analysis or their levels must be merged to reduce the number of cells in the table.

3.3.7 Equivalence of loglinear and logistic models

Consider three categorical variables: an outcome Y with two possible values $\{1, 2\}$ (where 2 codes a success), and covariates $X \in \{1, \ldots, I\}$ and $Z \in \{1, \ldots, J\}$. The data from n independent observations of (X, Z, Y) can be summarized as a 3-way contingency table of the size $I \times J \times 2$ with observed counts n_{ijk} of the combinations $X = i$, $Z = j$ and

$Y = k$, expected counts m_{ijk}, and cell probabilities π_{ijk}.

We are interested in estimating the conditional probabilities p_{ij} of success given the covariates, that is $p_{ij} = \mathrm{P}\big(Y = 2 \,\big|\, X = i, Z = j\big) = m_{ij2}/m_{ij+} = \pi_{ij2}/\pi_{ij+}$. The problem can be addressed either by logistic regression or by a loglinear model. We will show that, with an appropriately selected loglinear model, the results from the two approaches are equivalent.

Let the correct logistic model be

$$\log \frac{p_{ij}}{1 - p_{ij}} = \gamma^0 + \gamma_i^X + \gamma_j^Z, \tag{3.19}$$

where $\gamma_1^X = \gamma_1^Z = 0$. The left-hand side of the model can be rewritten as

$$\log \frac{p_{ij}}{1 - p_{ij}} = \log \frac{m_{ij2}}{m_{ij1}} = \log \frac{\pi_{ij2}}{\pi_{ij1}}.$$

The regression parameters of the logistic model have the following interpretation:

$$\begin{aligned} \mathrm{e}^{\gamma_i^X} &= \frac{\mathrm{P}\big(Y = 2 \,\big|\, X = i, Z = j\big)}{\mathrm{P}\big(Y = 1 \,\big|\, X = i, Z = j\big)} \frac{\mathrm{P}\big(Y = 1 \,\big|\, X = 1, Z = j\big)}{\mathrm{P}\big(Y = 2 \,\big|\, X = 1, Z = j\big)} \\ &= \frac{p_{ij}}{1 - p_{ij}} \frac{1 - p_{1j}}{p_{1j}} = \frac{\pi_{ij2}\pi_{1j1}}{\pi_{ij1}\pi_{1j2}} = \theta_{i2(j)}^{XY}, \end{aligned}$$

which is the conditional odds ratio for the association between X and Y given $Z = j$ (and it does not depend on j). Next,

$$\begin{aligned} \mathrm{e}^{\gamma_j^Z} &= \frac{\mathrm{P}\big(Y = 2 \,\big|\, X = i, Z = j\big)}{\mathrm{P}\big(Y = 1 \,\big|\, X = i, Z = j\big)} \frac{\mathrm{P}\big(Y = 1 \,\big|\, X = i, Z = 1\big)}{\mathrm{P}\big(Y = 2 \,\big|\, X = i, Z = 1\big)} \\ &= \frac{p_{ij}}{1 - p_{ij}} \frac{1 - p_{i1}}{p_{i1}} = \frac{\pi_{ij2}\pi_{i11}}{\pi_{ij1}\pi_{i12}} = \theta_{j2(i)}^{ZY}, \end{aligned}$$

which is the conditional odds ratio for the association between Z and Y given $X = i$ (and it does not depend on i).

Now consider the loglinear model with all two-way interactions, that is (XY, ZY, XZ):

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^Y + \beta_{ik}^{XY} + \beta_{jk}^{ZY} + \beta_{ij}^{XZ} \tag{3.20}$$

with the usual constraints on the parameters. In this model,

$$\log \frac{p_{ij}}{1 - p_{ij}} = \log \frac{m_{ij2}}{m_{ij1}} = \beta_2^Y + \beta_{i2}^{XY} + \beta_{j2}^{ZY}.$$

Thus, the loglinear model (3.20) induces the same structure on the conditional log odds of success as the logistic model (3.19). Clearly, $\beta_{i2}^{XY} = \gamma_i^X$, $\beta_{j2}^{ZY} = \gamma_j^Z$, and $\beta_2^Y = \gamma^0$. Thus, the logistic model estimates a subset of the parameters of the loglinear model that

determine the associations between Y and (X, Z). The results about these associations (parameter estimates, hypotheses tests) are the same no matter if they were obtained from the logistic model (3.19) or the loglinear model (3.20).

The only difference between the logistic model (3.19) and the loglinear model (3.20) is that the loglinear model also estimates the associations between X and Z. These associations are not estimated by the logistic model (they are not of interest).

One can easily generalize this observation about equivalence between loglinear and logistic models to arbitrary categorical covariates.

Note. The logistic model $Y \sim M$ is equivalent to the loglinear model $(MY, \mathfrak{I}(M))$, where MY includes all the interactions between the terms in M with Y and $\mathfrak{I}(M)$ is the most general interaction between all the terms in M.

4 Extensions of Generalized Linear Models

4.1 Quasi-likelihood and Overdispersion

4.1.1 Overdispersion in binomial data

Let $Y_1, \dots, Y_n \sim \mathsf{Bi}(m, \pi_0)$ be independent observations with moments $\mathsf{E}\, Y_i = m\pi_0$ and $\mathsf{var}\, Y_i = m\pi_0(1-\pi_0)$. Since each of them can be decomposed into iid alternative variables, methods for exponential family distributions apply to this case.

Now suppose that π_0 is not a constant but a random variable with mean π_0: the ith binomial variable has the distribution $\mathsf{Bi}(m, \pi_i)$ (given π_i), where $\pi_1, \dots, \pi_n$ is a random sample from a distribution with a density $g(\pi)$, mean $\mathsf{E}\, \pi_i = \pi_0$ and variance $\mathsf{var}\, \pi_i = \sigma_\pi^2$. Then $\mathsf{E}\, Y_i = m\pi_0$ but $\mathsf{var}\, Y_i = m\pi_0(1-\pi_0) + m(m-1)\sigma_\pi^2$. The variability in Y_i is larger than for the binomial distribution unless $m = 1$. This effect is called *overdispersion*[*].

Let $m > 1$ and consider a specific distribution for π_i, for example $\pi_i \sim \mathsf{B}(\alpha, \beta)$. Then $\pi_0 = \alpha/(\alpha+\beta)$ and $\sigma_\pi^2 = \pi_0(1-\pi_0)/(\alpha+\beta+1)$. The distribution of Y_i is

$$\mathrm{P}\left[Y_i = j\right] = \binom{m}{j} \frac{B(\alpha+j, \beta+m-j)}{B(\alpha,\beta)}, \quad j = 0, \dots, m.$$

This is called *a beta-binomial distribution*. Its variance is

$$\varphi\, m\pi_0(1-\pi_0), \quad \text{where} \quad \varphi = 1 + \frac{m-1}{\alpha+\beta+1} > 1$$

plays the role of a dispersion parameter.

This is not a distribution from the exponential family but we would like to extend the theory of GLM to responses following distributions of this type.

4.1.2 Overdispersion in Poisson data

Let $Y_1, \dots, Y_n \sim \mathsf{Po}(\lambda_0)$ be independent observations with moments $\mathsf{E}\, Y_i = \lambda_0$ and $\mathsf{var}\, Y_i = \lambda_0$. This is a distribution from the exponential family.

Let the Poisson parameters for the observations be random variables with mean λ_0 rather than a constant: the ith Poisson variable has the distribution $\mathsf{Po}(\lambda_i)$ (given

[*] Česky *nadměrná disperse*

λ_i), where $\lambda_1, \ldots, \lambda_n$ is a random sample from a distribution with a density $g(\lambda)$, mean $\mathsf{E}\,\lambda_i = \lambda_0$ and variance $\mathsf{var}\,\lambda_i = \sigma_\lambda^2$. Then $\mathsf{E}\,Y_i = \lambda_0$ but $\mathsf{var}\,Y_i = \lambda_0 + \sigma_\lambda^2 > \lambda_0$. The variability in Y_i is larger than for the Poisson distribution, another example of overdispersion.

Now suppose $\lambda_i \sim \Gamma(a, a\lambda_0)$. Then λ_0 is the mean of λ_i and $\sigma_\lambda^2 = \lambda_0/a$. The distribution of Y_i is

$$\mathrm{P}\,[Y_i = j] = \frac{1}{jB(j, a\lambda_0)} \left(\frac{a}{a+1}\right)^{a\lambda_0} \frac{1}{(a+1)^j}, \quad j = 0, 1, 2, \ldots .$$

This is called *a Poisson-gamma distribution*. Its variance is $\varphi\lambda_0$, where $\varphi = 1 + 1/a > 1$ plays the role of a dispersion parameter. Special cases of this distribution are negative binomial $\mathsf{NB}(m, p)$ (for $a\lambda_0 = m \in \mathsf{N}$, $a/(a+1) = p$) and geometric $\mathsf{Geo}(p)$ (for $a\lambda_0 = 1$, $a/(a+1) = p$).

Again, this distribution does not belong to the exponential family. The results for the GLM need to be extended to responses with distributions of this type.

4.1.3 Quasi-likelihood

Consider n independent copies of random vectors $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ are the covariates and Y_i is the response.

Assumptions.

1. $Y_1, \ldots, Y_n$ are independent
2. The mean $\mu_i = \mathsf{E}\,Y_i$ satisfies the identity $g(\mu_i) = \eta_i$, where $\eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$ is the linear predictor and g is a known strictly monotone, twice continuously differentiable link function.
3. The variance $\mathsf{var}\,Y_i$ satisfies the identity $\mathsf{var}\,Y_i = \varphi V(\mu_i)$, where $\varphi > 0$ is a dispersion parameter and V is a known positive continuously differentiable variance function.

Note. The form of the distribution of Y_i is not specified, it does not have to belong to the exponential family. Instead, a variance function describing the relationship of the mean to the variance is supplied.

Note. The original GLM was a parametric model. This is a semi-parametric model: the form of the distribution of Y_i is not specified, we only specify conditions on the first two moments of Y_i.

Because this is not a parametric model maximum likelihood estimator cannot be used. The regression parameters are estimated by the *maximum quasi-likelihood estimator*.

Definition 4.1. (Wedderburn 1974) The quasi-(log)likelihood $Q(\boldsymbol{\beta})$ is defined as $Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} Q_i(\boldsymbol{\beta})$, where

$$Q_i(\boldsymbol{\beta}) = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\varphi V(t)}\, dt.$$

The maximum quasi-likelihood estimator $\widehat{\boldsymbol{\beta}}_n$ is the point that maximizes the quasi-likelihood*.

The maximum quasi-likelihood estimator solves the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}_n) = \mathbf{0}$, where

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\beta})$$

is the quasi-score† with the terms

$$\boldsymbol{U}_i(\boldsymbol{\beta}) = \frac{\partial Q_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{Y_i - \mu_i}{\varphi V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}.$$

The quasi-score has exactly the same form as the score in the GLM, that is, it can be written as a sum of the terms

$$\frac{1}{\varphi} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i.$$

This fact was the primary motivation to introduce the quasi-likelihood in the form specified in Definition 4.1. The following lemma says that the quasi-score has exactly the same properties as the GLM score.

Lemma 4.1. *(Wedderburn 1974)*

(i) $\boldsymbol{U}_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$, *are iid random vectors.*
(ii) If $\boldsymbol{\beta}_0$ *is the true parameter then* $\mathsf{E}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) = \mathbf{0}$.
(iii) If $\boldsymbol{\beta}_0$ *is the true parameter then* $\mathsf{var}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) = -\mathsf{E}\, \frac{\partial}{\partial \boldsymbol{\beta}^\intercal} \boldsymbol{U}_i(\boldsymbol{\beta}_0) = I(\boldsymbol{\beta}_0)$, *where* $I(\boldsymbol{\beta}_0)$ *is defined by* (2.8).

Proposition 4.2. *There exists a sequence* $\widehat{\boldsymbol{\beta}}_n$ *of solutions to the quasi-likelihood equations such that* $\widehat{\boldsymbol{\beta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$.

The proposition claims consistency as long as the solutions exist and are unique. The key for the validity of the proposition is Lemma 4.1 (ii). Proposition 4.2 together with Lemma 4.1 justify the validity of two of the key asymptotic results that hold in the GLM. The results that we are getting in this section generalize Proposition 1.2.

* Česky *kvazivěrohodnost* † Česky *kvaziskóre*

Theorem 4.3.

(i) $\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I(\boldsymbol{\beta}_0))$,

(ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I^{-1}(\boldsymbol{\beta}_0))$.

Thus, Wald tests and score tests (and confidence intervals) derived for the GLM also hold when quasi-likelihood is used instead of full likelihood. On the other hand, likelihood ratio (and hence deviance) tests do not work.

The dispersion parameter φ can be estimated by the method of moments based on Pearson X^2 statistic as described in Section 2.5.

The theory of GLM (except deviance tests) holds even for distributions that are not of exponential type as long as the data are independent and the variance function is correctly specified. Thus, we can fit GLM to distributions such as beta-binomial, negative binomial or geometric (and many more).

4.2 Sandwich Variance Estimation in the GLM

4.2.1 Behavior of the MLE under a misspecified model

Let $X_1, \ldots, X_n$ be iid random variables (vectors) on the space $(\mathcal{X}, \mathcal{A})$ with distribution P and density p with respect to a σ-finite masure μ.

Consider the model $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ on the space $(\mathcal{X}, \mathcal{A})$ with densities $p_{\boldsymbol{\theta}}$ with respect to μ. Let $\Theta \subseteq \mathbb{R}^d$. Suppose the model $\mathcal{P}$ is regular. We do not assume that $P \in \mathcal{P}$. There need not exist any $\boldsymbol{\theta} \in \Theta$ such that $P = P_{\boldsymbol{\theta}}$.

We use the data $X_1, \ldots, X_n \overset{\mathsf{iid}}{\sim} P$ and the model $\mathcal{P}$ to estimate $\boldsymbol{\theta}$ by the method of maximum likelihood. This section summarizes the results obtained by White (1982).

Define

$$\begin{aligned}
\ell_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(X_i), \\
\widehat{\boldsymbol{\theta}}_n &= \arg\max_{\boldsymbol{\theta} \in \Theta} \ell_n(\boldsymbol{\theta}), \\
\boldsymbol{U}_i(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(X_i), \\
\boldsymbol{U}_n(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \boldsymbol{U}_i(\boldsymbol{\theta}).
\end{aligned}$$

The estimator $\widehat{\boldsymbol{\theta}}_n$ solves the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. White (1982) calls it the *quasi-maximum likelihood estimator* but we will prefer the term *pseudo-maximum likelihood estimator*. What does this estimator estimate when the model does not hold?

Theorem 4.4. *(White 1982)* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$ *as* $n \to \infty$, *where*

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} K(P, P_{\boldsymbol{\theta}}) = \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{E}_P \log p_{\boldsymbol{\theta}}(X_i)$$

and

$$K(P, P_{\boldsymbol{\theta}}) = \mathsf{E}_P \log \frac{p(X_i)}{p_{\boldsymbol{\theta}}(X_i)} \geq 0.$$

The pseudo-MLE converges to the point $\boldsymbol{\theta}_0$ that minimizes the Kullback-Leibler distance between all the distributions belonging to the model and the true distribution of the data.

The pseudo-score statistic is asymptotically normal when evaluated at $\boldsymbol{\theta}_0$ and the estimator is also asymptotically normal.

Theorem 4.5. *(White 1982)*

(i) The probability limit $\boldsymbol{\theta}_0$ *of* $\widehat{\boldsymbol{\theta}}_n$ *satisfies* $\mathsf{E}_P \boldsymbol{U}_n(\boldsymbol{\theta}_0) = \mathbf{0}$.

(ii)

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \mathsf{var}_P \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{\boldsymbol{\theta}_0}(X_i).$$

(iii)

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, \mathbb{D}^{-1}\Sigma\mathbb{D}^{-1}),$$

where

$$\mathbb{D} = -\mathsf{E}_P \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \log p_{\boldsymbol{\theta}_0}(X_i).$$

The asymptotic variance of $\widehat{\boldsymbol{\theta}}_n$ has the sandwich form. The matrix $\mathbb{D}$ plays the role of an information matrix. However, $\mathbb{D}$ is not equal to the asymptotic variance Σ of the pseudo-score statistic.

Theorem 4.6. *The asymptotic variance matrix* $\mathbb{D}^{-1}\Sigma\mathbb{D}^{-1}$ *of* $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ *can be consistently estimated by* $\widehat{\mathbb{D}}^{-1}\widehat{\Sigma}\widehat{\mathbb{D}}^{-1}$, *where*

$$\widehat{\mathbb{D}} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} \log p_{\widehat{\boldsymbol{\theta}}_n}(X_i)$$

and

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i(\widehat{\theta}_n)^{\otimes 2}.$$

This is the *sandwich estimator** of the asymptotic variance. Compare this to Proposition 1.3 and its use in linear regression.

* Česky *sendvičový odhad*

Note.

- If the model holds, that is $\exists \boldsymbol{\theta}_0 \in \Theta$ such that $P = P_{\boldsymbol{\theta}_0}$, then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_0$ and $\mathbb{D} \equiv \Sigma$.
- The sandwich estimator $\widehat{\mathbb{D}}^{-1}\widehat{\Sigma}\widehat{\mathbb{D}}^{-1}$ of the asymptotic variance tends to underestimate the true variance $\mathbb{D}^{-1}\Sigma\mathbb{D}^{-1}$ unless n is very large. Various modifications have been proposed to reduce the small sample bias of the sandwich estimator.
- This theory is especially useful when at least some components of $\boldsymbol{\theta}_0$ are equal to the true parameters that we wish to estimate.

4.2.2 Applications to the GLM

Consider n independent copies of random vectors $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^\mathsf{T}$ are the covariates and Y_i is the response.

Assumptions.

1. $Y_1, \ldots, Y_n$ are independent
2. The mean $\mu_i = \mathsf{E}\, Y_i$ satisfies the identity $g(\mu_i) = \eta_i$, where $\eta_i = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}_0$ is the linear predictor and g is a known strictly monotone, twice continuously differentiable link function.

Note.

- We will choose a *working variance function** $V(\mu_i)$ and use it in the estimation of $\boldsymbol{\beta}_0$ but we will not asume that this variance function is correct.
- No assumptions are made about the form of the density of Y_i.

The estimation of $\boldsymbol{\beta}_0$ proceeds as if Y_i had a distribution from the exponential family with mean $\mu_i = \mathsf{E}\, Y_i = g^{-1}(\eta_i)$ and variance $\mathsf{var}\, Y_i = \varphi V(\mu_i)$. The *pseudo-score function*† is the same as in the GLM,

$$\boldsymbol{U}_i(\boldsymbol{\beta}) = \frac{Y_i - \mu_i}{\varphi V(\mu_i)} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{1}{\varphi} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i.$$

The estimator $\widehat{\boldsymbol{\beta}}_n$ is the solution to the system of pseudo-score equations

$$\boldsymbol{U}_n(\widehat{\boldsymbol{\beta}}_n) = \sum_{i=1}^{n} w(\widehat{\mu}_i) g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i)\boldsymbol{X}_i = \mathbf{0},$$

where $\widehat{\mu}_i = g^{-1}(\boldsymbol{X}_i^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)$. This system can be solved by the IWLS algorithm.

Lemma 4.7.

(i) $\boldsymbol{U}_i(\boldsymbol{\beta})$, $i = 1, \ldots, n$, *are iid random vectors.*

* Česky *pracovní rozptylová funkce* † Česky *pseudoskórová funkce*

(ii) If $\boldsymbol{\beta}_0$ is the true parameter then $\mathsf{E}\,\boldsymbol{U}_i(\boldsymbol{\beta}_0) = \mathbf{0}$.

We use the theory from the previous section to derive the probability limit and the asymptotic distribution of $\widehat{\boldsymbol{\beta}}_n$. Let

$$I(\boldsymbol{\beta}_0) = -\mathsf{E}\,\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\boldsymbol{U}_i(\boldsymbol{\beta}_0) = \frac{1}{\varphi}\mathsf{E}_{\boldsymbol{X}}\, w(\mu_i)\boldsymbol{X}_i^{\otimes 2}$$

and

$$\Sigma = \mathsf{var}\,\boldsymbol{U}_i(\boldsymbol{\beta}_0) = \frac{1}{\varphi^2}\mathsf{E}_{\boldsymbol{X}}\, w^2(\mu_i)[g'(\mu_i)]^2 \mathsf{var}\, Y_i\, \boldsymbol{X}_i^{\otimes 2} \neq I(\boldsymbol{\beta}_0).$$

It follows from Theorem 4.5 (note that $I(\boldsymbol{\beta}_0)$ plays the role of the matrix $\mathbb{D}$) that

Theorem 4.8.

(i) $\widehat{\boldsymbol{\beta}}_n$ *converges in probability to* $\boldsymbol{\beta}_0$.

(ii) $\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \Sigma)$,

(iii) $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I^{-1}(\boldsymbol{\beta}_0)\Sigma I^{-1}(\boldsymbol{\beta}_0))$.

Corollary. The asymptotic variance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ can be consistently estimated by $\widehat{I}^{-1}\widehat{\Sigma}\widehat{I}^{-1}$, where $\widehat{I}$ is defined by (2.9) and

$$\widehat{\Sigma} = \frac{1}{n\widehat{\varphi}^2}\sum_{i=1}^{n}\left[w(\widehat{\mu}_i)g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i)\right]^2 \boldsymbol{X}_i^{\otimes 2}.$$

The corollary follows from Theorem 4.6. The dispersion parameter φ can be estimated by the Pearson X^2 statistic (see Section 2.5) but it is not needed to calculate either $\widehat{\boldsymbol{\beta}}_n$ or its estimated asymptotic variance $\widehat{I}^{-1}\widehat{\Sigma}\widehat{I}^{-1}$.

Likelihood ratio (deviance) tests cannot be used but Wald tests and score tests based on Theorem 4.8 are available. Thus, even if the distribution of the responses is not of exponential family and the variance function is unknown, the theory of the GLM can be used for parameter estimation and asymptotic variance can be estimated by the sandwich estimator. If the working variance function $V(\mu)$ is guessed correctly then $\Sigma \approx I(\boldsymbol{\beta}_0)$ and the results will be close to those obtained by the quasilikelihood approach. If the working variance function $V(\mu)$ is far from the truth the asymptotic variance will increase and estimates will be less efficient. The most serious danger of this approach is the potential underestimation of the true variance by the sandwich.

5 Generalized Estimating Equations

5.1 Group-Dependent Data

In this chapter, we extend the generalized linear model even further: we will develop a regression model that can be used for certain kinds of correlated data.

Suppose we observe K independent random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$, where

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}, \quad i = 1, \ldots, K.$$

So the data consists of K independent groups (subjects), which include different numbers of correlated observations. Within each group, the observations are dependent but between groups, they are independent.

We will call such data structures *group-dependent data*[*] but they may by called by several different terms depending on the context or application field.

Clustered data are measurements collected on groups of objects such as teeth, animal litters, or siblings. The measurements in the same group do not have a well defined ordering.[†]

Repeated measures are measurements made repeatedly on the same subject so that there is ordering among them (first, second, third,...).[‡]

Longitudinal data are repeated measures with captured time information. This means, each measurement is made at a different time, which is recorded in the data. Longitudinal data arise by observing independent short pieces of time series.[§]

Panel data is the term used for group-dependent data in econometrics.[¶]

Each observation Y_{ij} is accompanied by a vector of covariates $\boldsymbol{X}_{ij}$ of the size p. We would like to describe the dependence of $\mu_{ij} = \mathsf{E}\, Y_{ij}$ on the covariates $\boldsymbol{X}_{ij}$ by a regression model. As in the GLM, we will assume that $g(\mu_{ij}) = \boldsymbol{X}_{ij}^\mathsf{T}\boldsymbol{\beta}_0$, where g is a known strictly monotone, twice continuously differentiable link function and $\boldsymbol{\beta}_0$ is an unknown true parameter vector.

Thus, we assume that

$$\mathsf{E}\, \boldsymbol{Y}_i = \boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{in_i})^\mathsf{T},$$

where $\mu_{ij} = g^{-1}(\boldsymbol{X}_{ij}^\mathsf{T}\boldsymbol{\beta}_0)$. Like in Section 4.2.2, We leave $\mathsf{var}\, \boldsymbol{Y}_i$ unspecified – we do not impose any assumptions on variances and covariances of the measurements.

[*] Česky *skupinově závislá data* [†] Česky *shluková data* [‡] Česky *opakovaná měření* [§] Česky *longitudinální data* [¶] Česky *panelová data*

5.2 Estimation of Regression Parameters by Generalized Estimating Equations

A method for estimating regression parameters under this general extension of the GLM to group-dependent data was first proposed by Liang and Zeger (1986). They called it *generalized estimating equations*[*] [GEE]. Let

$$\mathbb{X}_i = \begin{pmatrix} \boldsymbol{X}_{i1}^\mathsf{T} \\ \vdots \\ \boldsymbol{X}_{in_i}^\mathsf{T} \end{pmatrix}_{n_i \times p} \quad \text{and} \quad \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)_{p \times n_i} = \mathbb{X}_i^\mathsf{T} \begin{pmatrix} g'(\mu_{i1}) & \dots & 0 \\ 0 & \ddots & 0 \\ 0 & \dots & g'(\mu_{in_i}) \end{pmatrix}^{-1}.$$

Recall that the score function in the GLM (with $n_i = 1$ for all i) can be written as

$$\boldsymbol{U}_i(\boldsymbol{\beta}) = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{1}{\varphi V(\mu_i)} (Y_i - \mu_i).$$

When written in this way, the score can be easily generalized to multivariate $\boldsymbol{Y}_i$. We define a pseudo-score function

$$\boldsymbol{U}_i(\boldsymbol{\beta}) = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) \mathbb{Q}_i^{-1}(\boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

where

$$\mathbb{Q}_i(\boldsymbol{\mu}_i) = \varphi \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i) \mathbb{R}_i \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$$

represents our guess about $\mathsf{var}\, \boldsymbol{Y}_i$. The diagonal $n_i \times n_i$ matrix $\mathbb{V}_i$ provides working variances of the observations, $V(\mu_{i1}), \dots, V(\mu_{in_i})$, on its diagonal. The $n_i \times n_i$ matrix $\mathbb{R}_i$ is a *working correlation matrix*[†], our guess about $\mathsf{cor}\,(Y_{ij}, Y_{ik})$.

Neither the variance function $V(\cdot)$ nor the correlation matrix $\mathbb{R}_i$ is assumed to be be correct. If we knew $\mathsf{var}\, Y_{ij}$ and $\mathsf{cor}\, \boldsymbol{Y}_i$ we would use them in place of $\varphi V(\mu_{ij})$ and $\mathbb{R}_i$. Since we do not know them, we just use our best guess.

Because $\mathsf{E}\, \boldsymbol{Y}_i = \boldsymbol{\mu}_i$, we easily get the moments of $\boldsymbol{U}_i(\boldsymbol{\beta}_0)$:

$$\mathsf{E}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) = \boldsymbol{0} \quad \text{and} \quad \mathsf{var}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) \equiv \Sigma = \mathsf{E} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) \mathbb{Q}_i^{-1}(\boldsymbol{\mu}_i) \mathsf{var}\, \boldsymbol{Y}_i \mathbb{Q}_i^{-1}(\boldsymbol{\mu}_i) \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^\mathsf{T}.$$

The estimator $\widehat{\boldsymbol{\beta}}_K$ is defined as the solution to the system of equations

$$\boldsymbol{U}_K(\widehat{\boldsymbol{\beta}}_K) \equiv \sum_{i=1}^{K} \boldsymbol{U}_i(\widehat{\boldsymbol{\beta}}_K) = \boldsymbol{0}. \tag{5.1}$$

This is called *the GEE estimator*. It is another special case of the methods of Section 4.2.1. Heuristically, the properties of $\widehat{\boldsymbol{\beta}}_K$ follow from Theorem 4.5, though a rigorous proof of consistency would require more work.

[*] Česky *zobecněné odhadovací rovnice* [†] Česky *pracovní korelační matice*

Let

$$\mathbb{D} = -\mathsf{E}\,\frac{\partial}{\partial \boldsymbol{\beta}^{\mathsf{T}}}\boldsymbol{U}_i(\boldsymbol{\beta}_0) = \mathsf{E}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)\mathbb{Q}_i^{-1}(\boldsymbol{\mu}_i)\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}}.$$

Proposition 5.1. *(Liang and Zeger 1986) As $K \to \infty$,*

(i) $\widehat{\boldsymbol{\beta}}_K \xrightarrow{\mathsf{P}} \boldsymbol{\beta}_0$,
(ii) $\frac{1}{\sqrt{K}}\boldsymbol{U}_K(\boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \Sigma)$,
(iii) $\sqrt{K}(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_0) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, \mathbb{D}^{-1}\Sigma\mathbb{D}^{-1})$.

Notice that the asymptotics requires that the number K of independent groups tends to infinity. The number of observations within the groups is irrelevant. By Theorem 4.6, the asymptotic variance of $\mathbb{D}^{-1}\Sigma\mathbb{D}^{-1}$ can be consistently estimated by the sandwich $\widehat{\mathbb{D}}^{-1}\widehat{\Sigma}\widehat{\mathbb{D}}^{-1}$, where

$$\widehat{\mathbb{D}} = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)\mathbb{Q}_i^{-1}(\widehat{\boldsymbol{\mu}}_i)\left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}}$$

and

$$\widehat{\Sigma} = \frac{1}{K}\sum_{i=1}^{K}\boldsymbol{U}_i(\widehat{\boldsymbol{\beta}}_K)^{\otimes 2}.$$

The system of equations (5.1) can be solved by a modified IWLS method. The solver iterates

$$\widehat{\boldsymbol{\beta}} = \left[\sum_{i=1}^{K}\left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)\mathbb{Q}_i^{-1}(\widehat{\boldsymbol{\mu}}_i)\left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)^{\mathsf{T}}\right]^{-1}\left[\sum_{i=1}^{K}\left(\frac{\partial \widehat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}}\right)\mathbb{Q}_i^{-1}(\widehat{\boldsymbol{\mu}}_i)\widehat{\boldsymbol{Z}}_i\right],$$

where

$$\widehat{\boldsymbol{Z}}_i = (\widehat{Z}_{i1}, \ldots, \widehat{Z}_{in_i})^{\mathsf{T}} \quad \text{and} \quad \widehat{Z}_{ij} = \frac{\widehat{\eta}_{ij} + (Y_{ij} - \widehat{\mu}_{ij})g'(\widehat{\mu}_{ij})}{g'(\widehat{\mu}_{ij})}.$$

The dispersion parameter φ can be estimated by

$$\widehat{\varphi} = \frac{1}{n-p}\sum_{i=1}^{K}\sum_{j=1}^{n_i}\frac{(Y_{ij} - \widehat{\mu}_{ij})^2}{V(\widehat{\mu}_{ij})},$$

which is based on a modified Pearson X^2 statistic (ignoring the correlations). Here, $n = \sum_{i=1}^{K} n_i$ is the total sample size. This estimator is consistent if the variance function V is correctly specified, but it is not efficient. Estimated φ is not needed to calculate either $\widehat{\boldsymbol{\beta}}_n$ or its estimated asymptotic variance $\widehat{\mathbb{D}}^{-1}\widehat{\Sigma}\widehat{\mathbb{D}}^{-1}$.

5.3 Correlation Structures

How should we choose the working covariance matrix $\mathbb{Q}_i(\boldsymbol{\mu}_i) = \varphi\mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)\mathbb{R}_i\mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$? The variance function $V(\cdot)$ expresses our belief about the dependence of the variance on

the mean. It is more complicated to choose a suitable working correlation matrix $\mathbb{R}_i$ for each of the independent groups. The current section is devoted to this problem.

5.3.1 Working independence

We take $\mathbb{R}_i = \mathbb{I}_{n_i}$ (identity matrix). Then $\mathbb{Q}_i(\boldsymbol{\mu}_i) = \varphi \mathbb{V}_i(\boldsymbol{\mu}_i)$ is diagonal and $\boldsymbol{\beta}_0$ is estimated as if the data were independent, by fitting the standard IWLS. After the estimates are obtained under the independence assumption (they are still consistent), their variance is adjusted by the sandwich to take into account correlations and also perhaps the wrong choice of $V(\cdot)$. This is the easiest way to fit GEE. The estimates will be consistent but inefficient if the correlations are strong.

5.3.2 Parametrized correlations

The other option is to introduce some non-independent correlation structure and parametrize it by an m-dimensional parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^m$. We take $\mathbb{R}_i = \mathbb{R}_i(\boldsymbol{\alpha})$,

$$\mathbb{Q}_i(\boldsymbol{\mu}_i) \equiv Q_i(\boldsymbol{\mu}_i, \boldsymbol{\alpha}) = \varphi \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i) \mathbb{R}_i(\boldsymbol{\alpha}) \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i)$$

and

$$\widehat{\mathbb{Q}}_i(\boldsymbol{\mu}_i) \equiv Q_i(\boldsymbol{\mu}_i, \widehat{\boldsymbol{\alpha}}) = \widehat{\varphi} \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i) \mathbb{R}_i(\widehat{\boldsymbol{\alpha}}) \mathbb{V}_i^{1/2}(\boldsymbol{\mu}_i),$$

where $\widehat{\boldsymbol{\alpha}}$ is a $\sqrt{K}$-consistent estimator of $\boldsymbol{\alpha}$, for example some moment estimator.

The score is modified as follows

$$\widehat{\boldsymbol{U}}_i(\boldsymbol{\beta}) = \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right) \widehat{\mathbb{Q}}_i^{-1}(\boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

but these vectors are no longer independent for $i = 1, \ldots, n$. One needs to show that

$$\frac{1}{\sqrt{K}} \sum_{i=1}^{K} \widehat{\boldsymbol{U}}_i(\boldsymbol{\beta}_0) = \frac{1}{\sqrt{K}} \sum_{i=1}^{K} \boldsymbol{U}_i(\boldsymbol{\beta}_0) + o_P(1).$$

(see Liang and Zeger 1986). The estimator $\widehat{\boldsymbol{\beta}}$ solves $\sum_{i=1}^{K} \widehat{\boldsymbol{U}}_i(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$ and Proposition 5.1 still holds.

Here is a general strategy how to estimate parametrized correlations by the method of moments:

1. Estimate $\boldsymbol{\beta}$ under working independence (take $\mathbb{R}_i = \mathbb{I}_{n_i}$).
2. Calculate Pearson residuals
$$r_{ij}^P = \frac{Y_{ij} - \widehat{\mu}_{ij}}{\sqrt{V(\widehat{\mu}_{ij})}}.$$

3. If the mean structure and variance function are correct we have

$$\mathsf{E}\, r_{ij}^P \approx 0, \quad \mathsf{var}\, r_{ij}^P \approx \varphi, \quad \mathsf{E}\, r_{ij}^P r_{ik}^P \approx \varphi R_{ijk}(\boldsymbol{\alpha}).$$

We can use moment estimators of $\boldsymbol{\alpha}$ based on the products $r_{ij}^P r_{ik}^P$ of Pearson residuals from the same group. The next section shows examples of such estimators for selected correlation structures.

1-band correlation

Let

$$\mathsf{cor}\,(Y_{ij}, Y_{ik}) = \begin{cases} 1 & \text{if } j = k, \\ \alpha & \text{if } |j-k| = 1, \\ 0 & \text{if } |j-k| > 1. \end{cases}$$

The parameter α is consistently estimated by

$$\widehat{\alpha} = \frac{1}{\widehat{\varphi}} \frac{1}{N-K-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i-1} r_{ij}^P r_{i,j+1}^P.$$

m-band correlation

Take some positive whole number m. Let

$$\mathsf{cor}\,(Y_{ij}, Y_{ik}) = \begin{cases} 1 & \text{if } j = k, \\ \alpha_l & \text{if } |j-k| = l,\ l = 1, \ldots, m, \\ 0 & \text{if } |j-k| > m. \end{cases}$$

For $l \in \{1, \ldots, m\}$, the parameter α_l is consistently estimated by

$$\widehat{\alpha}_l = \frac{1}{\widehat{\varphi}} \frac{1}{N-Kl-p} \sum_{i=1}^{K} \sum_{j=1}^{n_i-l} r_{ij}^P r_{i,j+l}^P.$$

Note that m should not be too large; especially it should not exceed too many n_i's.

Exchangeable correlation

Let

$$\mathsf{cor}\,(Y_{ij}, Y_{ik}) = \begin{cases} 1 & \text{if } j = k, \\ \alpha & \text{if } j \neq k. \end{cases}$$

The parameter α is consistently estimated by

$$\widehat{\alpha} = \frac{1}{\widehat{\varphi}} \frac{1}{\sum_{i=1}^{K} \binom{N_i}{2} - p} \sum_{i=1}^{K} \sum_{j=1}^{n_i} \sum_{k=j+1}^{n_i} r_{ij}^P r_{ik}^P.$$

AR(1) correlation

Let $Y_{i1}, \ldots, Y_{in_i}$ form an AR(1) series. Then

$$\mathsf{cor}\,(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}.$$

Since

$$\begin{aligned} \mathsf{E}\, r_{ij}^P r_{ik}^P &\approx \varphi\alpha^{|j-k|} \\ \text{and} \quad \log \mathsf{E}\, r_{ij}^P r_{ik}^P &\approx \log\varphi + |j-k| \log\alpha, \end{aligned}$$

we can estimate $\log\alpha$ as the slope in the linear regression problem with $\log r_{ij}^P r_{ik}^P$ as the response and $|j-k|$ as the covariate.

Other correlation structures

See methods for estimating correlations from time series data.

5.3.3 Joint estimation of mean and correlation structures

Prentice (1988) and Prentice and Zhao (1991) proposed to estimate the correlation parameters by introducing another set of estimating equations for $\boldsymbol{\alpha}$ and solving them jointly with the generalized estimating equations (5.1) for $\boldsymbol{\beta}$. Yan and Fine (2004) extended this idea even further: they proposed a third set of estimating equations for φ and allow φ to depend on the covariates and thus vary between subjects.

For details, see the references.

5.3.4 Summary of GEE methods

1. GEE works for regression analysis of data with K independent groups, which are correlated within each group. The number K of independent groups must be large enough for the asymptotics to work.
2. It is not necessary to correctly specify the distribution of the response, the variance of the response, or the correlations within each group. If the variance and the correlations are seriously misspecified, the variance of the estimators increases but the estimators are still consistent and asymptotically normal.
3. The parameters $\boldsymbol{\beta}$ have population-averaged interpretation, not subject-specific interpretation (see the discussion in Chapter 7).

6 Linear Mixed Effects Models

6.1 Introduction

6.1.1 One-way ANOVA

Let $Y_{ij} \sim \mathsf{N}(\mu_i, \sigma_e^2)$ be independent. The index $i = 1, \ldots, I$ denotes a subject. Subjects represent the levels of factor A that may potentially affect the mean of the outcome. The index $j = 1, \ldots, n_i$ denotes observations within subject. The total number of observations is $n = \sum_{i=1}^{I} n_i$. The null hypothesis of interest is

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I,$$

that is, the means $\mathsf{E}\, Y_{ij}$ are the same for all subjects (levels of factor A).

This is the setup of the *one-way ANOVA problem*[*] familiar from an introductory statistics class. The model can be written in the form

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij},$$

where $\beta_1, \ldots, \beta_I$ are constants subject to the linear constraint $\sum_{i=1}^{I} \beta_i = 0$, μ is the overall mean, and ε_{ij} are random errors (expressing the unexplained variability in Y_{ij}) satisfying $\mathsf{E}\, \varepsilon_{ij} = 0$, $\mathsf{var}\, \varepsilon_{ij} = \sigma_e^2$.

There are $I + 1$ unknown parameters in this model, in particular $\mu, \beta_1, \ldots, \beta_{I-1}, \sigma_e^2$.

The hypothesis test in one-way ANOVA can be presented in the form of ANOVA table (Table 6.1). The second column (SS = sums of quares) contains sums of squares

[*] Česky *analýza rozptylu – jednoduché třídění*

Table 6.1: One way ANOVA table with fixed effects

Source	SS	df	MS	EMS	F
Subject (A)	SS_A	$I-1$	$MS_A = \dfrac{SS_A}{I-1}$	$\sigma_e^2 + Q_\beta$	$F_A = \dfrac{MS_A}{MS_e}$
Residual	SS_e	$n-I$	$MS_e = \dfrac{SS_e}{n-I}$	σ_e^2	
Total	SS_T	$n-1$			

for subjects

$$SS_A = \sum_{i=1}^{I} n_i(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot})^2,$$

where $\overline{Y}_{i\cdot} = n_i^{-1} \sum_{j=1}^{n_i} Y_{ij}$ and $\overline{Y}_{\cdot\cdot} = n^{-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$, and residual sums of squares

$$SS_e = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \overline{Y}_{i\cdot})^2.$$

It can be shown that

$$SS_A = \boldsymbol{Z}^\mathsf{T} \mathbb{A} \boldsymbol{Z},$$

where

$$\boldsymbol{Z} = \begin{pmatrix} \overline{Y}_{1\cdot} \\ \vdots \\ \overline{Y}_{I\cdot} \end{pmatrix} - \mu \mathbf{1}_I \quad \text{and} \quad \mathbb{A} = \operatorname{diag}(n_1, \ldots, n_I) - \frac{1}{n} \boldsymbol{n}^{\otimes 2},$$

$\boldsymbol{n} = (n_1, \ldots, n_I)^\mathsf{T}$ and $\mathbf{1}_I$ is a column vector of 1's of length I.

Also,

$$SS_e = \sum_{i=1}^{I} \boldsymbol{Z}_i^\mathsf{T} \mathbb{A}_i \boldsymbol{Z}_i,$$

where

$$\boldsymbol{Z}_i = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix} - \mu_i \mathbf{1}_{n_i} \quad \text{and} \quad \mathbb{A}_i = \mathbb{I}_{n_i} - \frac{1}{n_i} \mathbf{1}_{n_i}^{\otimes 2}.$$

Proposition 6.1. *Let $\boldsymbol{X}$ be a random vector of length n with mean $\mathsf{E}\,\boldsymbol{X} = \boldsymbol{\mu}$ and variance matrix $\mathsf{var}\,\boldsymbol{X} = \mathbb{V}$. Let $\mathbb{A}$ be any $n \times n$ matrix. Then*

$$\mathsf{E}\,\boldsymbol{X}^\mathsf{T} \mathbb{A} \boldsymbol{X} = \boldsymbol{\mu}^\mathsf{T} \mathbb{A} \boldsymbol{\mu} + \operatorname{tr} \mathbb{A} \mathbb{V}.$$

According to Proposition 6.1,

$$\mathsf{E}\, SS_A = \mathsf{E}\, \boldsymbol{Z}^\mathsf{T} \mathbb{A} \boldsymbol{Z} = (I-1)\sigma_e^2 + \sum_{i=1}^{I} n_i(\mu_i - \overline{\mu})^2,$$

where $\overline{\mu} = n^{-1} \sum n_i \mu_i = \mu + n^{-1} \sum n_i \beta_i = \mu + \overline{\beta}$, and

$$\mathsf{E}\, SS_e = \sum_{i=1}^{I} \mathsf{E}\, \boldsymbol{Z}_i^\mathsf{T} \mathbb{A}_i \boldsymbol{Z}_i = (n-I)\sigma_e^2.$$

Hence the expectations of mean squares (Table 6.1, column EMS) are

$$\mathsf{E}\, MS_A = \sigma_e^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(\mu_i - \overline{\mu})^2 = \sigma_e^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i(\beta_i - \overline{\beta})^2 = \sigma_e^2 + Q_\beta,$$
$$\mathsf{E}\, MS_e = \sigma_e^2,$$

not assuming that H_0 holds and that the data are normal. $\mathsf{E}\, MS_A$ is the sum of the residual variance σ_e^2 and weighted between-subject variability (sample variance of β_i).

When H_0 is true, $\mathsf{E}\, MS_A = \mathsf{E}\, MS_e$ and, under normality, $F = MS_A/MS_e$ has an F distribution with $I-1$ and $n-I$ degrees of freedom.

6.1.2 One-way ANOVA with random effects

Suppose that the subjects (levels of factor A) represent a random sample from some general population of subjects. Then we may not want to test differences in the means of the I particular subjects participating in the experiment (as classical one-way ANOVA does) but to test whether EY_{ij} depends on the subject in the context of all subjects belonging to the whole general population. To do this, we need to acknowledge in the model that we may have taken different subjects who would have had different means than the I subjects we were observing. A natural way to do this is to consider the subject effects random rather than fixed.

Assume

$$Y_{ij} = \mu + b_i + \varepsilon_{ij},$$

where μ is the overall mean, ε_{ij} are iid error terms satisfying $\mathsf{E}\,\varepsilon_{ij} = 0$, $\mathsf{var}\,\varepsilon_{ij} = \sigma_e^2$, and $b_1, \ldots, b_I$ are iid random variables with $\mathsf{E}\, b_i = 0$, $\mathsf{var}\, b_i = \sigma_b^2$. We also assume independence between $(b_1, \ldots, b_I)$ and $(\varepsilon_{11}, \ldots, \varepsilon_{In_I})$. The random variables b_i are called *random effects*[*].

There are 3 unknown parameters in this model, in particular $\mu, \sigma_b^2, \sigma_e^2$. The parameter σ_b^2 expresses how different are the subjects in the general populations in terms of the subject-specific mean response.

It is easy to calculate the conditional (given subject) and unconditional (in the general population) moments of Y_{ij}:

$$\mathsf{E}\big[Y_{ij}\,\big|b_i\big] = \mu + b_i, \qquad \mathsf{E}\, Y_{ij} = \mu,$$
$$\mathsf{var}\big[Y_{ij}\,\big|b_i\big] = \sigma_e^2, \qquad \mathsf{var}\, Y_{ij} = \sigma_b^2 + \sigma_e^2.$$

The parameters σ_b^2 and σ_e^2 are called *variance components*[†] because the total variability in Y_{ij} is decomposed into variability between subjects σ_b^2 (how different are the subjects from each other) and variability within subjects σ_e^2 (how different are the observations made on the same subject from each other).

[*] Česky *náhodné efekty* [†] Česky *komponenty rozptylu*

Unlike in classical one-way ANOVA, the observations are not independent:

$$\begin{aligned}
\mathsf{cov}\,(Y_{ij}, Y_{i'j'}) &= 0 && \text{for} \quad i \neq i', \\
\mathsf{cov}\,(Y_{ij}, Y_{ij'}) &= \mathsf{var}\, b_i = \sigma_b^2 && \text{for} \quad j \neq j', \\
\mathsf{cor}\,(Y_{ij}, Y_{ij'}) &= \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \geq 0 && \text{for} \quad j \neq j'.
\end{aligned}$$

The observations are correlated within each subject unless $\sigma_b^2 = 0$. The data has group-dependent structure. The correlations arise because the differences between subjects are not expressed in the mean structure (μ) but added to the random component of the model ($b_i + \varepsilon_{ij}$). In classical ANOVA, the mean structure ($\mu + \beta_i$) captures the differences between subjects and the random component of the model (ε_{ij}) consists of independent variables.*

We would like to test the hypothesis that all subjects in the general population have the same mean response. Within our model, this hypothesis can be expressed as

$$H_0 : \sigma_b^2 = 0 \quad \text{against} \quad H_1 : \sigma_b^2 > 0.$$

The analysis proceeds exactly as in the classical one-way ANOVA. Define SS_A, SS_e, MS_A and MS_e as in the previous section. It can be shown using Proposition 6.1 that

$$\begin{aligned}
\mathsf{E}\, MS_A &= \sigma_e^2 + \frac{1}{I-1}\left(n - \frac{\sum n_i^2}{n}\right)\sigma_b^2, \\
\mathsf{E}\, MS_e &= \sigma_e^2.
\end{aligned}$$

If the null hypothesis is true then $\mathsf{E}\, MS_A = \mathsf{E}\, MS_e = \sigma_e^2$. If the null hypothesis is true and normality holds for both ε_{ij} and b_i then $F = MS_A/MS_e$ has an F distribution with $I - 1$ and $n - I$ degrees of freedom. Thus, the analysis proceeds according to Table 6.1 but the null hypothesis and the interpretation of the results are different.

The expected mean squares allow us to find an unbiased estimator of σ_b^2 quite easily. In particular,

$$\widehat{\sigma}_b^2 = (MS_A - MS_e)\frac{I-1}{n - \frac{\sum n_i^2}{n}}$$

is unbiased. However, the probability of $MS_A < MS_e$ is positive so this estimator can sometimes attain negative values. We can take $\max(0, \widehat{\sigma}_b^2)$ as an estimator but it is no longer unbiased.

* This an illustration of the fact that difference between dependence and independence can be a matter of point of view. The same observations can be considered independent by one analyst and dependent by another analyst without any of them being wrong.

6.1.3 Two-way ANOVA with random effects

Now consider two classification factors A and B, both with randomly selected levels from some general populations. Let

$$Y_{ijk} = \mu + b_i^A + b_j^B + b_{ij}^{AB} + \varepsilon_{ijk},$$

where $i = 1, \ldots, I$ are levels of factor A, $j = 1, \ldots, J$ are levels of factor B, and $k = 1, \ldots, m$ indexes observations made for each combination of the levels of A and B. The design is balanced (equal number of observations for each combination of i and j). The total number of observations is $n = mIJ$.

The parameter μ is the overall mean (a constant), b_i^A, b_j^B, and b_{ij}^{AB} are mutually independent random effects* satisfying $\mathsf{E}\, b_i^A = 0$, $\mathsf{var}\, b_i^A = \sigma_A^2$, $\mathsf{E}\, b_j^B = 0$, $\mathsf{var}\, b_j^B = \sigma_B^2$, and $\mathsf{E}\, b_{ij}^{AB} = 0$, $\mathsf{var}\, b_{ij}^{AB} = \sigma_{AB}^2$. Finally, ε_{ijk} are iid error terms satisfying $\mathsf{E}\, \varepsilon_{ijk} = 0$, $\mathsf{var}\, \varepsilon_{ijk} = \sigma_e^2$, independent from all the random effects.

There are 5 unknown parameters in this model, in particular $\mu, \sigma_A^2, \sigma_B^2, \sigma_{AB}^2, \sigma_e^2$.

We get

$$\begin{aligned}
\mathsf{E}\big[Y_{ijk} \,\big|\, b_i^A, b_j^B, b_{ij}^{AB}\big] &= \mu + b_i^A + b_j^B + b_{ij}^{AB} \\
\mathsf{var}\big[Y_{ijk} \,\big|\, b_i^A, b_j^B, b_{ij}^{AB}\big] &= \sigma_e^2 \\
\mathsf{E}\, Y_{ijk} &= \mu \\
\mathsf{var}\, Y_{ijk} &= \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 + \sigma_e^2.
\end{aligned}$$

The total variability in Y_{ijk} is decomposed into four variance components.

The covariances depend on how many levels of A and B are shared by the observations:

$$\begin{aligned}
\mathsf{cov}\,(Y_{ijk}, Y_{ijl}) &= \sigma_A^2 + \sigma_B^2 + \sigma_{AB}^2 && \text{for} \quad k \neq l, \\
\mathsf{cov}\,(Y_{ijk}, Y_{ij'l}) &= \sigma_A^2 < \mathsf{cov}\,(Y_{ijk}, Y_{ijl}) && \text{for} \quad j \neq j', \\
\mathsf{cov}\,(Y_{ijk}, Y_{i'jl}) &= \sigma_B^2 < \mathsf{cov}\,(Y_{ijk}, Y_{ijl}) && \text{for} \quad i \neq i', \\
\mathsf{cov}\,(Y_{ijk}, Y_{i'j'l}) &= 0 && \text{for} \quad i \neq i',\ j \neq j'.
\end{aligned}$$

The covariance (and correlation) is largest between two observations made on the same levels of both A and B. It is smaller between two observations made on the same level of either A or B (not both). Two observations made on different levels of both A and B are independent. This data does not have group-dependent structure, there are no mutually independent groups.

* The random effects are independent from each other as well as from random effects pertaining to other levels of A and B

Define the sums of squares and mean squares exactly as in the balanced two-way ANOVA problem with fixed effects:

$$
\begin{aligned}
SS_A &= mJ\sum_{i=1}^{I}(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2, & MS_A &= \frac{SS_A}{I-1},\\
SS_B &= mI\sum_{j=1}^{J}(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdot\cdot\cdot})^2, & MS_B &= \frac{SS_B}{J-1},\\
SS_{AB} &= m\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdot\cdot\cdot})^2, & MS_{AB} &= \frac{SS_{AB}}{(I-1)(J-1)},\\
SS_e &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{m}(Y_{ijk} - \overline{Y}_{ij\cdot})^2, & MS_e &= \frac{SS_e}{(m-1)IJ}.
\end{aligned}
$$

Rewriting the sums of squares as quadratic forms and using Proposition 6.1, we get expected mean squares as follows:

$$
\begin{aligned}
\mathsf{E}\, MS_A &= mJ\sigma_A^2 + m\sigma_{AB}^2 + \sigma_e^2,\\
\mathsf{E}\, MS_B &= mI\sigma_B^2 + m\sigma_{AB}^2 + \sigma_e^2,\\
\mathsf{E}\, MS_{AB} &= m\sigma_{AB}^2 + \sigma_e^2,\\
\mathsf{E}\, MS_e &= \sigma_e^2.
\end{aligned}
\tag{6.1}
$$

Consider testing the null hypothesis $H_0 : \sigma_A^2 = 0$. The only mean squares term that involves the tested parameter is MS_A. Under the null hypothesis, $\mathsf{E}\, MS_A = \mathsf{E}\, MS_{AB}$ and the test statistic is $F_A = MS_A/MS_{AB}$, distributed as $F_{I-1,(I-1)(J-1)}$ under normality and validity of the null hypothesis. This is different from two-way ANOVA with fixed effects, where the test statistic for testing factor A main effects is MS_A/MS_e, distributed as $F_{I-1,(m-1)IJ}$ under normality and validity of the null hypothesis. So here, the test for random effects is different from the test for fixed effects. Because $\mathsf{E}\, MS_{AB} \geq \mathsf{E}\, MS_e$, the test usually results in smaller test statistic than the test for fixed effects and rejects less frequently. Thus, in more complex models than simple one-way ANOVA, it is more difficult to demonstrate violation of the null hypothesis when the subjects represent a random sample than if they are considered fixed.

The null hypothesis $H_0 : \sigma_B^2 = 0$ is tested by the statistic $F_B = MS_B/MS_{AB}$, distributed as $F_{J-1,(I-1)(J-1)}$ under normality and validity of the null hypothesis. The null hypothesis $H_0 : \sigma_{AB}^2 = 0$ is tested by the statistic $F_{AB} = MS_{AB}/MS_e$, distributed as $F_{(I-1)(J-1),(m-1)IJ}$ under normality and validity of the null hypothesis. This is the only test in this model that is the same with random effects as with fixed effects.

The system of equations (6.1) is linear in the parameters. It can be used to derive

these unbiased moment estimators of the variance components:

$$\begin{aligned}\widehat{\sigma}_e^2 &= MS_e,\\ \widehat{\sigma}_{AB}^2 &= \frac{1}{m}(MS_{AB} - MS_e),\\ \widehat{\sigma}_A^2 &= \frac{1}{mJ}(MS_A - MS_{AB}),\\ \widehat{\sigma}_B^2 &= \frac{1}{mI}(MS_B - MS_{AB}).\end{aligned}$$

All these estimators except $\widehat{\sigma}_e^2$ may be negative.

The idea of generalizing the results known from classical ANOVA theory to random effects can be followed for more complex models than this one. However, the results quickly become too difficult to derive and apply. Even the two-way ANOVA model with unbalanced data (number of observations n_{ij} vary with i and j) results in very complicated expressions for $\mathsf{E}\,MS$. The F-test statistics get more complex and do not have exact F distribution even under normality. Various approximations to the denominator degrees of freedom are used to find approximate F distributions of these statistics (one of them is called *Satterthwaite's approximation*).

6.1.4 Random intercept and slope

Let Y_{ij} be measured on a subject i at the time t_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$. The number of observations on a subject and the timing of the observations may vary. Suppose that $\mathsf{E}\,Y_{ij}$ depends linearly on time.

Consider the model

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + \varepsilon_{ij}, \tag{6.2}$$

where ε_{ij} are independent, $\mathsf{E}\,\varepsilon_{ij} = 0$, $\mathsf{var}\,\varepsilon_{ij} = \sigma_e^2$. This is the classical linear model. It assures that $\mathsf{E}\,Y_{ij}$ is a linear function of time but implies that $Y_{i1}, \ldots, Y_{in_i}$ are mutually independent. This is not a realistic assumption for observations that are measured sequentially on the same subject.

Now let us extend the model so that different subjects follow different lines. Let b_i^0 be the deviation of the intercept of the i-th subject from the population intercept β_0 and let b_i^1 be the deviation of the slope of the i-th subject from the population slope β_1. We could include these parameters as fixed effects into model (6.2). However, if the number of subjects grew to infinity the number of subject-specific parameters would grow also and we would not be able to find their consistent estimators. It is better to view the subjects as a random sample coming from some general population of subjects and to consider their parameters random.

This consideration leads us to the model

$$Y_{ij} = \beta_0 + b_i^0 + (\beta_1 + b_i^1)t_{ij} + \varepsilon_{ij},$$

where $\mathsf{E}\,\varepsilon_{ij} = 0$, $\mathsf{var}\,\varepsilon_{ij} = \sigma_e^2$, $\mathsf{E}\,b_i^0 = 0$, $\mathsf{var}\,b_i^0 = \sigma_0^2$, $\mathsf{E}\,b_i^1 = 0$, $\mathsf{var}\,b_i^1 = \sigma_1^2$, ε_{ij} are independent, (b_i^0, b_i^1) are independent, ε_{ij} and (b_i^0, b_i^1) are independent, and $\mathsf{cov}\,(b_i^0, b_i^1) = \sigma_{01}$.

Separating the fixed and random part, we get

$$Y_{ij} = \underbrace{\beta_0 + \beta_1 t_{ij}}_{\text{fixed}} + \underbrace{b_i^0 + b_i^1 t_{ij} + \varepsilon_{ij}}_{\text{random}}. \tag{6.3}$$

The fixed part of (6.3) is the same as in the linear model (6.2). However, the random part is richer, it involves time and induces correlations between measurements taken on the same subject.

We have $\mathsf{E}\big[Y_{ij} \,\big|\, b_i^0, b_i^1\big] = \beta_0 + b_i^0 + (\beta_1 + b_i^1)t_{ij}$, hence $\mathsf{E}\,Y_{ij} = \beta_0 + \beta_1 t_{ij}$. The linear dependence on time is preserved, with the same parameters as in model (6.2). Also, $\mathsf{var}\big[Y_{ij} \,\big|\, b_i^0, b_i^1\big] = \sigma_e^2$ and hence

$$\mathsf{var}\,Y_{ij} = \sigma_e^2 + \mathsf{var}\,(b_i^0 + b_i^1 t_{ij}) = \sigma_e^2 + \sigma_0^2 + 2t_{ij}\sigma_{01} + t_{ij}^2\sigma_1^2.$$

So, under model (6.3), the population variance $\mathsf{var}\,Y_{ij}$ is a quadratic function of time. Also,

$$\mathsf{cov}\,(Y_{ij}, Y_{ik}) = \mathsf{cov}\,(b_i^0 + b_i^1 t_{ij},\, b_i^0 + b_i^1 t_{ik}) = \sigma_0^2 + (t_{ij} + t_{ik})\sigma_{01} + t_{ij}t_{ik}\sigma_1^2.$$

It can be shown that $\mathsf{cov}\,(Y_{ij}, Y_{ik}) \geq (\sigma_0 - t_{ij}\sigma_1)(\sigma_0 - t_{ik}\sigma_1)$.

This model can be written as

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, K$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\mathsf{T}$,

$$\mathbb{X}_i = \mathbb{Z}_i = \begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{in_i} \end{pmatrix},$$

$\boldsymbol{\beta} = (\beta_0, \beta_1)^\mathsf{T}$, $\boldsymbol{b}_i = (\boldsymbol{b}_i^0, \boldsymbol{b}_i^1)^\mathsf{T}$, $\mathsf{E}\,\boldsymbol{b}_i = \boldsymbol{0}$, $\mathsf{var}\,\boldsymbol{b}_i = \boldsymbol{\Psi} = \left(\begin{smallmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{smallmatrix}\right)$, $\mathsf{E}\,\boldsymbol{\varepsilon}_i = 0$, and $\mathsf{var}\,\boldsymbol{\varepsilon}_i = \sigma_e^2\mathbb{I}_{n_i}$.

This is an example of a linear mixed effects model to be studied in the next sections.

6.2 Definition of Linear Mixed Effects Model

6.2.1 Single-level LME model

Consider independent subjects $i = 1, \ldots, K$ and denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^\mathsf{T}$. Let $\mathbb{X}_i$ be an $n_i \times p$ regression matrix for fixed effects, $\boldsymbol{\beta}$ a p-vector of fixed effects, let $\mathbb{Z}_i$ be an $n_i \times q$ regression matrix for random effects.

Definition 6.1. The responses $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ satisfy a (single-level) *linear mixed effects model*[*] [LME model] if

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, K, \tag{6.4}$$

where $\boldsymbol{b}_i$ (*the random effects*[†]) are independent vectors,

$$\boldsymbol{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathbb{D}),$$

$\boldsymbol{\varepsilon}_i$ are independent vectors,

$$\boldsymbol{\varepsilon}_i \sim \mathsf{N}_{n_i}(\mathbf{0}, \sigma_e^2\mathbb{I}_{n_i}),$$

and $(\boldsymbol{\varepsilon}_1, \ldots \boldsymbol{\varepsilon}_K)^\mathsf{T}$ is independent of $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)^\mathsf{T}$.

This model was proposed by Laird and Ware (1982). Its mean structure is described by the term $\mathbb{X}_i\boldsymbol{\beta}$, the random structure is given by $\mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i$. Obviously, the random part has expectation $\mathbf{0}$ and variance $\mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\mathsf{T} + \sigma_e^2\mathbb{I}_{n_i}$. The assumption $\mathsf{var}\,\boldsymbol{\varepsilon}_i = \sigma_e^2\mathbb{I}_{n_i}$ will be relaxed in Section 6.5.

The parameters of the LME model are the fixed effects $\boldsymbol{\beta}$, the variance of the residual term σ_e^2, and the covariance matrix of random effects $\mathbb{D}$, a symmetric positive definite matrix. For computational reasons, it is convenient to replace $\mathbb{D}$ by another matrix Δ such that $\Delta^\mathsf{T}\Delta = \sigma_e^2\mathbb{D}^{-1}$. The matrix Δ (which is not unique) is called *the relative precision factor*. It can be obtained, for example, by Choleski decomposition of a positive definite matrix[‡].

Model (6.4) can be also written as

$$\boldsymbol{Y}_i \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \sigma_e^2\Sigma_i), \tag{6.5}$$

where $\Sigma_i = \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\mathsf{T}/\sigma_e^2 + \mathbb{I}_{n_i}$. This is the marginal form of the LME model, which hides the random effects structure. The marginal model (6.5) cannot be always written as a mixed effects model (6.4) – for example, if the matrix $\mathbb{D}$ is not positive definite. The marginal form can be extended to include all the observations

$$\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2\Sigma),$$

where $n = \sum_{i=1}^K n_i$,

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_K \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_K \end{pmatrix},$$

and Σ is a block-diagonal matrix with diagonal blocks $\Sigma_1, \ldots, \Sigma_K$.

[*] Česky *lineární smíšený model* [†] Česky *náhodné efekty* [‡] $\mathbb{A} > 0 \Rightarrow$ there exists a lower triangular $\mathbb{L}$ s.t. $\mathbb{A} = \mathbb{L}\mathbb{L}^\mathsf{T}$.

6.2.2 Multi-level LME model

Another grouping level can be added to the LME as follows. Consider primary groups $i = 1, \ldots, K$, and secondary groups $j = 1, \ldots, m_i$ nested within the primary groups (such as family j within town i). Observe a multivariate vector $\boldsymbol{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijn_{ij}})^\mathsf{T}$ for each combination of grouping levels i and j.

Definition 6.2. The independent vectors $\boldsymbol{Y}_{11}, \ldots, \boldsymbol{Y}_{Km_K}$ satisfy *a two-level linear mixed effects model* if

$$\boldsymbol{Y}_{ij} = \mathbb{X}_{ij}\boldsymbol{\beta} + \mathbb{Z}_{i,j}\boldsymbol{b}_i + \mathbb{Z}_{ij}\boldsymbol{b}_{ij} + \boldsymbol{\varepsilon}_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, m_i, \tag{6.6}$$

where $\boldsymbol{b}_i$ are independent first-level random effects

$$\boldsymbol{b}_i \sim \mathsf{N}_{q_1}(\boldsymbol{0}, \mathbb{D}_1),$$

$\boldsymbol{b}_{ij}$ are independent second-level random effects

$$\boldsymbol{b}_{ij} \sim \mathsf{N}_{q_2}(\boldsymbol{0}, \mathbb{D}_2),$$

$\boldsymbol{\varepsilon}_{ij}$ are independent residual terms

$$\boldsymbol{\varepsilon}_{ij} \sim \mathsf{N}_{n_{ij}}(\boldsymbol{0}, \sigma_e^2 \mathbb{I}_{n_{ij}})$$

(all the random effects and residual terms are mutually independent), $\mathbb{X}_{ij}$ is an $n_{ij} \times p$ regression matrix for fixed effects, $\boldsymbol{\beta}$ is a p-vector of fixed effects, $\mathbb{Z}_{i,j}$ is an $n_{ij} \times q_1$ regression matrix for first-level random effects, and $\mathbb{Z}_{ij}$ is an $n_{ij} \times q_2$ regression matrix for second-level random effects.

This model can be further extended to three or more levels (*multilevel modelling*). In the next sections, however, we will only present results for single-level LME. Extensions to multi-level models are possible but the notation gets a bit more complex.

6.3 Parameter Estimation

Consider a single-level LME model satisfying Definition 6.1. Suppose there exists a matrix Δ such that $\Delta^\mathsf{T}\Delta = \sigma_e^2 \mathbb{D}^{-1}$. Let $\boldsymbol{\theta}$ be an unconstrained parameter that uniquely determines the matrix Δ (and, together with σ_e^2, also $\mathbb{D}$).

6.3.1 Marginal likelihood

Consider the marginal form of the model

$$\boldsymbol{Y}_i \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \sigma_e^2 \Sigma_i),$$

where

$$\Sigma_i = \mathbb{Z}_i \mathbb{D} \mathbb{Z}_i^\mathsf{T} / \sigma_e^2 + \mathbb{I}_{n_i} = \mathbb{Z}_i (\Delta^\mathsf{T} \Delta)^{-1} \mathbb{Z}_i^\mathsf{T} + \mathbb{I}_{n_i}$$

is a function of parameters $\boldsymbol{\theta}$ only. The log-likelihood can be written as

$$\ell_n(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2) = c - \frac{n}{2} \log \sigma_e^2 - \frac{1}{2} \sum_{i=1}^{K} \log |\Sigma_i| - \frac{1}{2\sigma_e^2} \sum_{i=1}^{K} (\boldsymbol{Y}_i - \mathbb{X}_i \boldsymbol{\beta})^\mathsf{T} \Sigma_i^{-1} (\boldsymbol{Y}_i - \mathbb{X}_i \boldsymbol{\beta}). \quad (6.7)$$

When $\boldsymbol{\theta}$ and σ_e^2 are fixed, maximizing this with respect to $\boldsymbol{\beta}$ is equivalent to minimizing the weighted least squares criterion

$$\widehat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{K} (\boldsymbol{Y}_i - \mathbb{X}_i \boldsymbol{\beta})^\mathsf{T} \Sigma_i^{-1} (\boldsymbol{Y}_i - \mathbb{X}_i \boldsymbol{\beta}).$$

The solution is the weighted least squares estimator

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{K} \mathbb{X}_i^\mathsf{T} \Sigma_i^{-1} \mathbb{X}_i \right)^{-1} \sum_{i=1}^{K} \mathbb{X}_i^\mathsf{T} \Sigma_i^{-1} \boldsymbol{Y}_i.$$

This estimator depends on $\boldsymbol{\theta}$ but not on σ_e^2. Next, maximize the profile likelihood $\ell_n(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \sigma_e^2)$ with respect to σ_e^2 given $\boldsymbol{\theta}$. The estimator is obtained by the normalized residual weighted least squares

$$\widehat{\sigma}_e^2(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{K} \big(\boldsymbol{Y}_i - \mathbb{X}_i \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\big)^\mathsf{T} \Sigma_i^{-1} \big(\boldsymbol{Y}_i - \mathbb{X}_i \widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\big).$$

Finally, take $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ and $\widehat{\sigma}_e^2(\boldsymbol{\theta})$, plug them into ℓ_n to get profile likelihood for $\boldsymbol{\theta}$

$$\ell_n^*(\boldsymbol{\theta}) = \ell_n(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma}_e^2(\boldsymbol{\theta})),$$

and maximize it over $\boldsymbol{\theta}$. The form of this profile likelihood is very complicated; analytical calculation of the score statistic and information matrix is possible only in simple special cases. The problem can be handled by numerical optimization methods that do not rely on analytical derivatives.

In the next sections, we will develop a different approach to parameter estimation in LME models that takes advantage of the random effects structure combined with a clever decomposition of the log-likelihood.

6.3.2 Henderson's mixed model equations

Write the model in the more general form

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \mathbb{Z}\boldsymbol{b} + \boldsymbol{\varepsilon},$$

where $\mathbb{X}$ has n rows and p columns, $\mathbb{Z}$ has n rows and q^* columns, $\boldsymbol{b} \sim \mathsf{N}_{q^*}(\mathbf{0}, \mathbb{D}_*)$, $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \Lambda_*)$, $\mathsf{cov}\,(\boldsymbol{b}, \boldsymbol{\varepsilon}) = 0$, and $\Sigma = \mathsf{var}\,\boldsymbol{Y} = \mathbb{Z}\mathbb{D}_*\mathbb{Z}^\mathsf{T} + \Lambda_*$.

The single-level LME model can be obtained as a special case with

$$q^* = Kq, \quad n = \sum_{i=1}^{K} n_i, \quad \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_K \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_K \end{pmatrix}, \quad \mathbb{Z} = \begin{pmatrix} \mathbb{Z}_1 & 0 & \cdots & 0 \\ 0 & \mathbb{Z}_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{Z}_K \end{pmatrix},$$

$$\boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_1 \\ \vdots \\ \boldsymbol{b}_K \end{pmatrix}, \quad \mathbb{D}_* = \begin{pmatrix} \mathbb{D} & 0 & \cdots & 0 \\ 0 & \mathbb{D} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbb{D} \end{pmatrix}, \quad \Lambda_* = \sigma_e^2 \mathbb{I}_n.$$

Consider $\mathbb{D}_*$ and Λ_* known and write the joint density of $(\boldsymbol{Y}, \boldsymbol{b})$ with parameters $\boldsymbol{\beta}$:

$$f(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\beta}) = f(\boldsymbol{y} \mid \boldsymbol{b}; \boldsymbol{\beta}) f(\boldsymbol{b}) = \frac{1}{(2\pi)^{n/2} \left|\Lambda_*\right|^{1/2}} \exp\Big\{-\frac{1}{2}(\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta} - \mathbb{Z}\boldsymbol{b})^\mathsf{T} \Lambda_*^{-1} (\boldsymbol{y} - \mathbb{X}\boldsymbol{\beta} - \mathbb{Z}\boldsymbol{b})\Big\}$$
$$\times \frac{1}{(2\pi)^{q^*/2} \left|\mathbb{D}_*\right|^{1/2}} \exp\Big\{-\frac{1}{2}\boldsymbol{b}^\mathsf{T} \mathbb{D}_*^{-1} \boldsymbol{b}\Big\}.$$

Treat this density as a likelihood for unknown parameters $(\boldsymbol{\beta}, \boldsymbol{b})$ and maximize it jointly to obtain estimators of both $\boldsymbol{\beta}$ and $\boldsymbol{b}$. We get

$$(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}) = \arg\min_{\boldsymbol{\beta}, \boldsymbol{b}} \begin{pmatrix} \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} - \mathbb{Z}\boldsymbol{b} \\ \boldsymbol{b} \end{pmatrix}^\mathsf{T} \begin{pmatrix} \Lambda_*^{-1} & 0 \\ 0 & \mathbb{D}_*^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{Y} - \mathbb{X}\boldsymbol{\beta} - \mathbb{Z}\boldsymbol{b} \\ \boldsymbol{b} \end{pmatrix}.$$

This can be rewritten as weighted sum of squares

$$\arg\min_{\boldsymbol{\beta}, \boldsymbol{b}} \left[\begin{pmatrix} \boldsymbol{Y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbb{X} & \mathbb{Z} \\ 0 & -\mathbb{I}_{q^*} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{pmatrix} \right]^\mathsf{T} \begin{pmatrix} \Lambda_*^{-1} & 0 \\ 0 & \mathbb{D}_*^{-1} \end{pmatrix} \left[\begin{pmatrix} \boldsymbol{Y} \\ \mathbf{0} \end{pmatrix} - \begin{pmatrix} \mathbb{X} & \mathbb{Z} \\ 0 & -\mathbb{I}_{q^*} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{pmatrix} \right].$$

This is a weighted least squares problem. The estimators satisfy the system of equations

$$\begin{pmatrix} \mathbb{X}^\mathsf{T}\Lambda_*^{-1}\mathbb{X} & \mathbb{X}^\mathsf{T}\Lambda_*^{-1}\mathbb{Z} \\ \mathbb{Z}^\mathsf{T}\Lambda_*^{-1}\mathbb{X} & \mathbb{Z}^\mathsf{T}\Lambda_*^{-1}\mathbb{Z} + \mathbb{D}_*^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{b}} \end{pmatrix} = \begin{pmatrix} \mathbb{X}^\mathsf{T}\Lambda_*^{-1}\boldsymbol{Y} \\ \mathbb{Z}^\mathsf{T}\Lambda_*^{-1}\boldsymbol{Y} \end{pmatrix}. \tag{6.8}$$

Equations (6.8) are called *Henderson's mixed model equations* (Henderson 1984).

If all the inverses exist (which we assume they do), the estimators can be written down explicitely:

$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\boldsymbol{Y}),$$
$$\widehat{\boldsymbol{b}} = \mathbb{D}_*\mathbb{Z}^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}),$$

where $\Sigma = \mathbb{Z}\mathbb{D}_*\mathbb{Z}^\mathsf{T} + \Lambda_*$. In the special case of the single-level LME model, we get

$$\widehat{\boldsymbol{b}}_i = \mathbb{D}\mathbb{Z}_i^\mathsf{T}(\mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\mathsf{T} + \sigma_e^2\mathbb{I}_{n_i})^{-1}(\boldsymbol{Y}_i - \mathbb{X}_i\widehat{\boldsymbol{\beta}}).$$

Note.

1. $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator (BLUE) and consistent estimator of $\boldsymbol{\beta}$ regardless of the distribution of $\boldsymbol{Y}$.
2. $\widehat{\boldsymbol{b}}$ is the best linear unbiased predictor (BLUP) of $\boldsymbol{b}$. This means that $\forall \boldsymbol{c} \in \mathbb{R}^{q^*}$ $\mathsf{E}\,(\boldsymbol{c}^\mathsf{T}\widetilde{\boldsymbol{b}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{b})^2$ is minimized by $\widehat{\boldsymbol{b}}$ among all zero-mean linear functions $\widetilde{\boldsymbol{b}}$ of $\boldsymbol{Y}$.
3. $\widehat{\boldsymbol{b}}$ is the posterior mean of $\boldsymbol{b}$ given $\boldsymbol{Y}$ and $\widehat{\boldsymbol{\beta}}$ (the prior mean is $\mathbf{0}$).
4. $\widehat{\boldsymbol{b}}$ is a weighted average of $\mathbf{0}$ (the prior mean) and $\overline{\boldsymbol{b}}$ estimated by weighted least squares as fixed effects from the same data. $\widehat{\boldsymbol{b}}$ is called *shrinkage estimator* because it shrinks the estimated fixed effects towards zero.

Denote

$$\mathbb{C} \equiv \begin{pmatrix} \mathbb{C}_{11} & \mathbb{C}_{12} \\ \mathbb{C}_{21} & \mathbb{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbb{X}^\mathsf{T}\Lambda_*^{-1}\mathbb{X} & \mathbb{X}^\mathsf{T}\Lambda_*^{-1}\mathbb{Z} \\ \mathbb{Z}^\mathsf{T}\Lambda_*^{-1}\mathbb{X} & \mathbb{Z}^\mathsf{T}\Lambda_*^{-1}\mathbb{Z} + \mathbb{D}_*^{-1} \end{pmatrix}^{-1}.$$

It can be shown that $\mathbb{C}_{11} = \big(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\big)^{-1}$.

Proposition 6.2. *(Henderson 1984) Suppose $\mathbb{D}_*$ and Λ_* are known. Then*

(i) $\widehat{\boldsymbol{b}}$ *maximizes* $\mathsf{cor}\,(\widetilde{\boldsymbol{b}}, \boldsymbol{b})$ *among all linear unbiased predictors* $\widetilde{\boldsymbol{b}}$.

(ii) $\mathsf{E}\big[\boldsymbol{b}\,\big|\widehat{\boldsymbol{b}}\big] = \widehat{\boldsymbol{b}}$.

(iii) $\mathsf{var}\,\mathbb{A}\widehat{\boldsymbol{\beta}} = \mathbb{A}\mathbb{C}_{11}\mathbb{A}^\mathsf{T} = \mathbb{A}\big(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\big)^{-1}\mathbb{A}^\mathsf{T}$.

(iv) $\mathsf{cov}\,(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}) = \mathbf{0}$, $\mathsf{cov}\,(\widehat{\boldsymbol{\beta}}, \boldsymbol{b}) = -\mathbb{C}_{21}$.

(v) $\mathsf{var}\,\widehat{\boldsymbol{b}} = \mathsf{cov}\,(\widehat{\boldsymbol{b}}, \boldsymbol{b}) = \mathbb{D}_* - \mathbb{C}_{22}$.

(vi) $\mathsf{var}\,(\widehat{\boldsymbol{b}} - \boldsymbol{b}) = \mathbb{C}_{22}$.

Corollary. For a single-level LME model with known $\mathbb{D}$ and σ_e^2, the following holds:

(i) $\mathsf{var}\,\widehat{\boldsymbol{\beta}} = \big(\sum_{i=1}^K \mathbb{X}_i^\mathsf{T}\Sigma_i^{-1}\mathbb{X}_i\big)^{-1}$,

(ii) $\mathsf{var}\,\widehat{\boldsymbol{b}}_i = \mathbb{D}\mathbb{Z}_i^\mathsf{T}\Big[\Sigma_i^{-1} - \Sigma_i^{-1}\mathbb{X}_i\big(\sum_{i=1}^K \mathbb{X}_i^\mathsf{T}\Sigma_i^{-1}\mathbb{X}_i\big)^{-1}\mathbb{X}_i^\mathsf{T}\Sigma_i^{-1}\Big]\mathbb{Z}_i\mathbb{D}$,

(iii) $\mathsf{var}\,(\widehat{\boldsymbol{b}}_i - \boldsymbol{b}_i) = \mathbb{D} - \mathsf{var}\,\widehat{\boldsymbol{b}}_i$.

6.3.3 Maximum likelihood estimation of variance parameters

The likelihood function for a single-level LME model can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2 \mid \boldsymbol{Y}) = \prod_{i=1}^K f(\boldsymbol{Y}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2) = \prod_{i=1}^K \int f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma_e^2) f(\boldsymbol{b}_i \mid \boldsymbol{\theta}, \sigma_e^2)\, d\boldsymbol{b}_i,$$

where

$$f(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{\beta}, \sigma_e^2) = \frac{1}{(2\pi\sigma_e^2)^{n_i/2}} \exp\Big\{-\frac{1}{2\sigma_e^2}\,\|\boldsymbol{Y}_i - \mathbb{X}_i\boldsymbol{\beta} - \mathbb{Z}_i\boldsymbol{b}_i\|^2\Big\}$$

and

$$f(\boldsymbol{b}_i \mid \boldsymbol{\theta}, \sigma_e^2) = \frac{1}{(2\pi)^{q/2}\,|\mathbb{D}|^{1/2}} \exp\Big\{-\frac{1}{2}\boldsymbol{b}_i^{\mathsf{T}}\mathbb{D}^{-1}\boldsymbol{b}_i\Big\} = \\ = \frac{1}{(2\pi\sigma_e^2)^{q/2}\,|\Delta|^{-1}} \exp\Big\{-\frac{1}{2\sigma_e^2}\,\|\Delta\boldsymbol{b}_i\|^2\Big\},$$

where Δ is a relative precision factor matrix of size $q \times q$ such that $|\Delta| > 0$ and $\Delta^{\mathsf{T}}\Delta = \sigma_e^2\mathbb{D}^{-1}$.

The whole likelihood can be rearranged as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2 \mid \boldsymbol{Y}) = \prod_{i=1}^{K} \frac{|\Delta|}{(2\pi\sigma_e^2)^{n_i/2}} \int \frac{1}{(2\pi\sigma_e^2)^{q/2}} \exp\Big\{-\frac{1}{2\sigma_e^2}\left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\boldsymbol{b}_i\right\|^2\Big\}\, d\boldsymbol{b}_i,$$

where

$$\widetilde{\boldsymbol{Y}}_i = \begin{pmatrix}\boldsymbol{Y}_i \\ \boldsymbol{0}\end{pmatrix}, \quad \widetilde{\mathbb{X}}_i = \begin{pmatrix}\mathbb{X}_i \\ \boldsymbol{0}\end{pmatrix}, \quad \text{and} \quad \widetilde{\mathbb{Z}}_i = \begin{pmatrix}\mathbb{Z}_i \\ \Delta\end{pmatrix}$$

all have $n_i + q$ rows. For fixed $\boldsymbol{\beta}$, the Euclidean norm $\left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\boldsymbol{b}_i\right\|^2$ is minimized over $\boldsymbol{b}_i$ for

$$\widehat{\boldsymbol{b}}_i = (\widetilde{\mathbb{Z}}_i^{\mathsf{T}}\widetilde{\mathbb{Z}}_i)^{-1}\widetilde{\mathbb{Z}}_i^{\mathsf{T}}(\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta})$$

(see Henderson's equations). Thus

$$\left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\boldsymbol{b}_i\right\|^2 = \left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2 + (\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i)^{\mathsf{T}}\widetilde{\mathbb{Z}}_i^{\mathsf{T}}\widetilde{\mathbb{Z}}_i(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i).$$

The first term does not involve $\boldsymbol{b}_i$ and

$$\int \frac{1}{(2\pi\sigma_e^2)^{q/2}} \exp\Big\{-\frac{1}{2\sigma_e^2}(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i)^{\mathsf{T}}\widetilde{\mathbb{Z}}_i^{\mathsf{T}}\widetilde{\mathbb{Z}}_i(\boldsymbol{b}_i - \widehat{\boldsymbol{b}}_i)\Big\}\, d\boldsymbol{b}_i = \frac{1}{\sqrt{|\widetilde{\mathbb{Z}}_i^{\mathsf{T}}\widetilde{\mathbb{Z}}_i|}}. \tag{6.9}$$

Because $\left|\widetilde{\mathbb{Z}}_i^{\mathsf{T}}\widetilde{\mathbb{Z}}_i\right| = \left|\mathbb{Z}_i^{\mathsf{T}}\mathbb{Z}_i + \Delta^{\mathsf{T}}\Delta\right|$, we get

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2 \mid \boldsymbol{Y}) = \frac{1}{(2\pi\sigma_e^2)^{n/2}} \exp\Big\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{K}\left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\widehat{\boldsymbol{b}}_i\right\|^2\Big\} \prod_{i=1}^{K} \frac{|\Delta|}{\sqrt{|\mathbb{Z}_i^{\mathsf{T}}\mathbb{Z}_i + \Delta^{\mathsf{T}}\Delta|}}.$$

This can be maximized in three steps:

1. For fixed $\boldsymbol{\theta}$ and σ_e^2, minimize $\sum_{i=1}^{K}\left\|\widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i\boldsymbol{\beta} - \widetilde{\mathbb{Z}}_i\boldsymbol{b}_i\right\|^2$ jointly with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}_i$. Henderson's equations provide explicit formulae for $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \sigma_e^2)$ and $\widehat{\boldsymbol{b}}_i(\boldsymbol{\theta}, \sigma_e^2)$.

2. Maximize $L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}, \sigma_e^2), \boldsymbol{\theta}, \sigma_e^2)$ with respect to σ_e^2 for fixed $\boldsymbol{\theta}$. This step yields

$$\widehat{\sigma}_e^2(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{K} \left\| \widetilde{\boldsymbol{Y}}_i - \widetilde{\mathbb{X}}_i \widehat{\boldsymbol{\beta}} - \widetilde{\mathbb{Z}}_i \widehat{\boldsymbol{b}}_i \right\|^2.$$

3. Plug $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{b}}_i$, and $\widehat{\sigma}_e^2$ into $L(\cdot \mid \boldsymbol{Y})$ and get the profiled likelihood

$$L^*(\boldsymbol{\theta}) = L(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \widehat{\sigma}_e^2(\boldsymbol{\theta}) \mid \boldsymbol{Y}) = \frac{\exp\{-n/2\}}{(2\pi \widehat{\sigma}_e^2(\boldsymbol{\theta}))^{n/2}} \prod_{i=1}^{K} \frac{|\Delta|}{\sqrt{\left|\mathbb{Z}_i^\mathsf{T} \mathbb{Z}_i + \Delta^\mathsf{T} \Delta\right|}}.$$

The profile log-likelihood is

$$\ell^*(\boldsymbol{\theta}) = \log L^*(\boldsymbol{\theta}) = -\frac{1}{2}\Big(n + n \log 2\pi + n \log \widehat{\sigma}_e^2(\boldsymbol{\theta}) - 2K \log |\Delta| + \sum_{i=1}^{K} \log \left| \mathbb{Z}_i^\mathsf{T} \mathbb{Z}_i + \Delta^\mathsf{T} \Delta \right| \Big).$$

This can be simplified further by QR and/or Choleski decompositions. Once the MLE $\widehat{\boldsymbol{\theta}}$ is found, it is plugged into $\widehat{\sigma}_e^2(\widehat{\boldsymbol{\theta}})$ and then into $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}_e^2)$ and $\widehat{\boldsymbol{b}}_i(\widehat{\boldsymbol{\theta}}, \widehat{\sigma}_e^2)$. For details, see Jennrich and Schluchter (1986) and Lindstrom and Bates (1988).

6.3.4 Restricted maximum likelihood (REML) estimation of variance parameters

The main idea of REML is to get rid of the fixed effects before the variance parameters are estimated. There are two different ways to derive REML estimators, which lead to the same result. Both derivations are shown for the general case $\boldsymbol{Y} \sim \mathsf{N}_n(\mathbb{X}\boldsymbol{\beta}, \sigma_e^2 \Sigma)$. The single-level LME model is a special case of this model.

Derivation of REML estimators by zero-mean contrasts

The REML method was first proposed by Patterson and Thompson (1971), who took the following strategy to derive it. Take a matrix $\mathbb{A}$ such that $\mathsf{E}\, \mathbb{A}\boldsymbol{Y} = \boldsymbol{0}$ and use the likelihood of $\mathbb{A}\boldsymbol{Y}$ instead of the likelihood of $\boldsymbol{Y}$.

The matrix $\mathbb{I}_n - \mathbb{X}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\Sigma^{-1}$ generates the residuals of weighted least squares, is idempotent and has rank $n - p$. Take $\mathbb{A}$ a $(n-p) \times n$ matrix such that

$$\mathbb{A}^\mathsf{T}\mathbb{A} = \mathbb{I}_n - \mathbb{X}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\Sigma^{-1} \quad \text{and} \quad \mathbb{A}\mathbb{A}^\mathsf{T} = \mathbb{I}_{n-p}.$$

Since $\mathbb{A}^\mathsf{T}\mathbb{A}\mathbb{X} = \boldsymbol{0}$ then also $\mathbb{A}\mathbb{X} = \boldsymbol{0}$ and $\mathsf{E}\, \mathbb{A}\boldsymbol{Y} = \mathbb{A}\mathbb{X}\boldsymbol{\beta} = \boldsymbol{0}$. We have

$$\mathbb{A}\boldsymbol{Y} \equiv \boldsymbol{Y}_A \sim \mathsf{N}_{n-p}(\boldsymbol{0}, \sigma_e^2 \mathbb{A}\Sigma\mathbb{A}^\mathsf{T}).$$

Take $\mathbb{G} = (\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}\Sigma^{-1}$. It is a $p \times n$ matrix such that $\mathbb{G}\mathbb{X} = \mathbb{I}_p$, $\mathbb{G}\boldsymbol{Y} = \widehat{\boldsymbol{\beta}}$ (the WLS estimator of $\boldsymbol{\beta}$) and

$$\mathbb{G}\boldsymbol{Y} \equiv \widehat{\boldsymbol{\beta}} \sim \mathsf{N}_p(\boldsymbol{\beta}, \sigma_e^2(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}).$$

The random vector $(\boldsymbol{Y}_A, \widehat{\boldsymbol{\beta}})$ is a linear one-to-one transformation of $\boldsymbol{Y}$, hence it is multivariate normal. Because $\mathsf{cov}\,(\mathbb{A}\boldsymbol{Y}, \mathbb{G}\boldsymbol{Y}) = \mathbf{0}$, the components $\mathbb{A}\boldsymbol{Y}$ and $\mathbb{G}\boldsymbol{Y}$ are independent.

Because of independence, the joint density of $f_{\boldsymbol{Y}_A,\widehat{\boldsymbol{\beta}}}$ of $(\boldsymbol{Y}_A, \widehat{\boldsymbol{\beta}})$ is the product of the marginal densities $f_{\boldsymbol{Y}_A}$ and $f_{\widehat{\boldsymbol{\beta}}}$, so we can express $f_{\boldsymbol{Y}_A} = f_{\boldsymbol{Y}_A,\widehat{\boldsymbol{\beta}}}/f_{\widehat{\boldsymbol{\beta}}}$. Instead of this ratio, we take $f_{\boldsymbol{Y}}/f_{\widehat{\boldsymbol{\beta}}}$ and use it as a likelihood function for estimating σ_e^2 and $\boldsymbol{\theta}$.

We have

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi\sigma_e^2)^{n/2}}\frac{1}{|\Sigma|^{1/2}}\exp\Big\{-\frac{1}{2\sigma_e^2}(\boldsymbol{y}-\mathbb{X}\boldsymbol{\beta})^\mathsf{T}\Sigma^{-1}(\boldsymbol{y}-\mathbb{X}\boldsymbol{\beta})\Big\}. \tag{6.10}$$

and

$$f_{\widehat{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\beta}}) = \frac{1}{(2\pi\sigma_e^2)^{p/2}}\left|\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\right|^{1/2}\exp\Big\{-\frac{1}{2\sigma_e^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\mathsf{T}\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\Big\}.$$

Realizing that $\mathbb{X}^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}}) = 0$, we get

$$(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\boldsymbol{\beta}) = (\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\mathsf{T}\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}). \tag{6.11}$$

The ratio $f_{\boldsymbol{Y}}(\boldsymbol{Y})/f_{\widehat{\boldsymbol{\beta}}}(\widehat{\boldsymbol{\beta}})$, simplified by the equality (6.11), does not depend on $\boldsymbol{\beta}$. We call it the restricted (residual) likelihood. It is not a proper likelihood because the ratio of densities in not in general a density.

Definition 6.3. The restricted (REML) likelihood is

$$L_R(\boldsymbol{\theta}, \sigma_e^2) = \frac{1}{(2\pi\sigma_e^2)^{(n-p)/2}}\ |\Sigma|^{-1/2}\ \left|\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\right|^{-1/2} \times \\ \times \exp\Big\{-\frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})\Big\}.$$

Note. The REML log-likelihood is

$$\ell_R(\boldsymbol{\theta}, \sigma_e^2) = C - \frac{n-p}{2}\log\sigma_e^2 - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\log\left|\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\right| \\ - \frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}}), \tag{6.12}$$

where Σ and $\widehat{\boldsymbol{\beta}}$ depend on $\boldsymbol{\theta}$. It differs from the marginal log-likelihood (6.7) as follows:

$$\ell_R(\boldsymbol{\theta}, \sigma_e^2) = \ell_n(\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}), \boldsymbol{\theta}, \sigma_e^2) + \frac{p}{2}\log\sigma_e^2 - \frac{1}{2}\log\left|\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X}\right| + C_1.$$

Maximizing (6.12) over σ_e^2 for fixed $\boldsymbol{\theta}$ yields the following REML estimator of σ_e^2

$$\widehat{\sigma}_e^2(\boldsymbol{\theta}) = \frac{1}{n-p}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}).$$

This differs from the MLE by the subtraction of the number of parameters p from the number of observations n in the denominator. Unlike the MLE, the REML estimator is unbiased when $\boldsymbol{\theta}$ is known and agrees with the usual unbiased estimator of residual variance in linear models.

Repeating the steps leading to the derivation of the usual profile likelihood in Section 6.3.3, we end up with REML profile log-likelihood for estimating $\boldsymbol{\theta}$ in the single-level LME model written in the form

$$\ell_R^*(\boldsymbol{\theta}) = \ell_R(\boldsymbol{\theta}, \sigma_e^2(\boldsymbol{\theta})) = -\frac{1}{2}\Bigg[(n-p)\log\widehat{\sigma}_e^2(\boldsymbol{\theta}) - 2K\log|\Delta| + \\ + \sum_{i=1}^{K}\log\left|\boldsymbol{Z}_i^\mathsf{T}\boldsymbol{Z}_i + \Delta^\mathsf{T}\Delta\right| + \log\left|\sum_{i=1}^{K}\mathbb{X}_i^\mathsf{T}\Sigma_i^{-1}\mathbb{X}_i\right|\Bigg] + C_1.$$

This is maximized to obtain the REML estimator $\widehat{\boldsymbol{\theta}}_R$ of of $\boldsymbol{\theta}$.

Next, we calculate $\widehat{\sigma}_e^2(\widehat{\boldsymbol{\theta}}_R)$ and, finally, $\widehat{\boldsymbol{\theta}}_R$ is plugged into the weighted least squares estimator to obtain $\widehat{\boldsymbol{\beta}}$.

Derivation of REML estimators by integration over $\boldsymbol{\beta}$

This is a completely different way to derive REML estimators, which nevertheless leads to the same result. It was proposed by Harville (1974).

We want to get rid of the unknown $\boldsymbol{\beta}$ in the likelihood. We can do this by averaging the density of $\boldsymbol{Y}$ over all possible values of $\boldsymbol{\beta}$ and take the result as a new likelihood for estimation of σ_e^2 and $\boldsymbol{\theta}$.

So the idea is to take

$$L_R(\boldsymbol{\theta}, \sigma_e^2) = \int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2)\, d\boldsymbol{\beta}.$$

This can be also viewed in a Bayesian context. Consider $\boldsymbol{\beta}$ random and assign a non-informative improper prior density $h(\boldsymbol{\beta}) = 1$ on $\mathbb{R}^p$ to the random $\boldsymbol{\beta}$. Calculating the expectation of $L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2)$ over $\boldsymbol{\beta}$ yields $L_R(\boldsymbol{\theta}, \sigma_e^2)$.

We will show that this likelihood is the same as the REML likelihood from Definition 6.3. We need to calculate

$$\int L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2)\, d\boldsymbol{\beta} = \int f_{\boldsymbol{Y}}(\boldsymbol{Y}; \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2)\, d\boldsymbol{\beta},$$

where $f_{\boldsymbol{Y}}(\boldsymbol{y})$ is defined by (6.10). From the decomposition (6.11), we get

$$\int f_{\boldsymbol{Y}}(\boldsymbol{Y};\boldsymbol{\beta},\boldsymbol{\theta},\sigma_e^2)\,d\boldsymbol{\beta} = \frac{1}{(2\pi\sigma_e^2)^{n/2}}\frac{1}{|\Sigma|^{1/2}}\exp\Big\{-\frac{1}{2\sigma_e^2}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})^{\mathsf{T}}\Sigma^{-1}(\boldsymbol{Y}-\mathbb{X}\widehat{\boldsymbol{\beta}})\Big\}\times$$
$$\times\int\exp\left\{-\frac{1}{2\sigma_e^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\Sigma^{-1}\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}d\boldsymbol{\beta}. \quad (6.13)$$

Using the same trick as in (6.9), we can see that

$$\int\exp\left\{-\frac{1}{2\sigma_e^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^{\mathsf{T}}\mathbb{X}^{\mathsf{T}}\Sigma^{-1}\mathbb{X}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right\}d\boldsymbol{\beta} = (2\pi\sigma_e^2)^{q/2}\left|\mathbb{X}^{\mathsf{T}}\Sigma^{-1}\mathbb{X}\right|^{-1/2}.$$

Thus, the expression calculated from (6.13) is the same as the restricted likelihood in Definition 6.3. Both approaches to deriving REML estimators lead to the same result.

6.3.5 Comparison of REML versus ML estimators

To understand REML estimation, one has to realize three important facts:

- REML likelihood only estimates the variance parameters, not the fixed effects;
- REML likelihood is not a proper likelihood;
- the difference between REML likelihood and ordinary likelihood is asymptotically negligible.

We can summarize the strengths and weaknesses of REML (relative to MLE) as follows:

Advantages of REML

- The variance estimators agree with those used in the analysis of variance and linear regression.
- Under certain conditions (balanced design), REML estimators of variance parameters are unbiased.

Disadvantages of REML

- Because of the extra term $\left|\mathbb{X}^{\mathsf{T}}\Sigma^{-1}\mathbb{X}\right|^{-1/2}$ in REML likelihood, the estimates of σ_e^2 and $\boldsymbol{\theta}$ depend on the parametrization of the fixed part of the model.

  Consider $\mathbb{X} \neq \mathbb{X}^*$ such that $\mathcal{M}(\mathbb{X}) = \mathcal{M}(\mathbb{X}^*)$. If MLE is used, then the likelihood, variance parameter estimates, residuals, and fitted values, are the same for both parametrizations. If REML is used, then the likelihood, variance parameter estimates, residuals, and fitted values differ even though the two models are equivalent.

- REML likelihood (or any quantity derived from the likelihood, e.g. AIC or BIC) cannot be used to compare two models with different fixed effects structures or even parametrizations of fixed effects.

6.4 Hypothesis Testing and Confidence Intervals

6.4.1 Asymptotic approach based on MLE theory

It can be shown that in the single-level LME model the terms that express the difference between the regular likelihood and REML likelihood vanish as the number of groups K grows. Thus, the asymptotic theory for ML and REML estimators is the same, with some exceptions explained below.

Theorem 6.3. *If the model holds and regularity assumptions for the MLE theory are satisfied*

(i) The estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_e^2$ and $\widehat{\boldsymbol{\theta}}$ are consistent as $K \to \infty$.

(ii) Let r be the dimension of $\boldsymbol{\theta}$. Then

$$\sqrt{K}\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\sigma}_e^2 - \sigma_e^2 \\ \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \end{pmatrix} \xrightarrow{\mathsf{D}} \mathsf{N}_{p+r+1}(\mathbf{0}, I^{-1}) \quad as \quad K \to \infty,$$

where I is the Fisher information matrix with the following structure

$$I = \begin{pmatrix} I_{\boldsymbol{\beta}} & 0 & 0 \\ 0 & I_{\sigma_e^2} & I_{\sigma_e^2,\boldsymbol{\theta}} \\ 0 & I_{\boldsymbol{\theta},\sigma_e^2} & I_{\boldsymbol{\theta}} \end{pmatrix}.$$

(iii) Let M be a model, ℓ_M its log-likelihood, and $\widehat{\boldsymbol{\beta}}$, $\widehat{\sigma}_e^2$, and $\widehat{\boldsymbol{\theta}}$ parameter estimates. Let S be a submodel, ℓ_S its log-likelihood, and $\widetilde{\boldsymbol{\beta}}$, $\widetilde{\sigma}_e^2$, and $\widetilde{\boldsymbol{\theta}}$ parameter estimates. If the submodel holds and if the submodel does not specify parameters to be at the boundary of the parameter space then

$$LR = 2[\ell_M(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}_e^2, \widehat{\boldsymbol{\theta}}) - \ell_S(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}_e^2, \widetilde{\boldsymbol{\theta}})] \xrightarrow{\mathsf{D}} \chi_m^2,$$

where m is the difference in the number of parameters between the model M and the submodel S.

The meaning of Theorem 6.3 needs to be carefully explained in a series of important notes.

Note.

1. The asymptotic results require that the number of independent groups K grows to infinity. It is not enough to keep K bounded and to increase all n_i.
2. **Estimation of fixed effects:**
   - $\sqrt{K}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathsf{D}} \mathsf{N}_p(\mathbf{0}, I_{\boldsymbol{\beta}}^{-1})$ as $K \to \infty$ if the model holds, where $I_{\boldsymbol{\beta}} = \mathsf{E}\, \mathbb{X}_i^{\mathsf{T}} \Sigma_i^{-1} \mathbb{X}_i / \sigma_e^2$.
   - $\widehat{\boldsymbol{\beta}}$ is consistent even if the random part of the model is incorrectly specified but its asymptotic variance is different from $I_{\boldsymbol{\beta}}^{-1}$.
   - If the model holds then the asymptotic variance $I_{\boldsymbol{\beta}}^{-1}$ can be consistently estimated by
   $$\Big[\frac{1}{K}\sum_{i=1}^{K} \mathbb{X}_i^{\mathsf{T}} (\widehat{\sigma}_e^2 \mathbb{I}_{n_i} + \mathbb{Z}_i D(\widehat{\boldsymbol{\theta}}) \mathbb{Z}_i^{\mathsf{T}})^{-1} \mathbb{X}_i\Big]^{-1}.$$
3. **Likelihood ratio tests:**
   - Theorem 6.3(iii) also holds for REML log-likelihood as long as the fixed part $\mathbb{X}$ is the same in the model as in the submodel.
   - Likelihood ratio tests for reduction in the dimension of $\boldsymbol{\beta}$ (tests of the fixed part of the model) are not recommended because of slow convergence to the limiting χ^2 distribution; these tests are more suitable for testing variance components.
   - Theorem 6.3(iii) does not hold when the submodel involves a removal of a variance component. This issue is discussed in more detail in the next section.

6.4.2 Likelihood ratio tests

Likelihood ratio tests are not recommended for testing hypotheses about $\boldsymbol{\beta}$ but they are useful for testing hypotheses about variance components (functions of $\boldsymbol{\theta}$). However, variance component testing frequently violates the regularity assumptions of MLE theory, which makes the asymptotic distribution of the LR test statistic specified by Theorem 6.3(iii) invalid.

Consider a model with a single random intercept b_i added to the linear predictor describing the fixed part of the model. Let the variance of b_i be denoted $\sigma_b^2 \geq 0$. The parameter space for the variance component σ_b^2 is not an open set because it includes the zero value. Indeed, testing the presence of the random intercept in the model is equivalent to testing the hypothesis $H_0 : \sigma_b^2 = 0$ against the alternative $H_1 : \sigma_b^2 > 0$. The null hypothesis specifies a value of the parameter σ_b^2 that lies at the border of the parameter space. Because of this, the LR test statistic comparing log-likelihoods with and without the random intercept does not converge to the usual χ_1^2 distribution.

This problem persists in all likelihood ratio tests about a removal of a variance

component (which reduces the dimension of $\boldsymbol{\theta}$ from q to $q-m$). Such test statistics violate the conditions of Theorem 6.3(iii) and do not have asymptotic χ^2_m distribution.

We provide several examples of the true limiting distributions of likelihood ratio test statistics in important special cases addressed by Stram and Lee (1994; 1995). Proofs are omitted. Recall that $\mathbb{D} = \mathsf{var}\, \boldsymbol{b}_i$ and consider a few important hypotheses about $\mathbb{D}$.

1. $H_0 : \mathbb{D} = 0$ against $H_1 : \mathbb{D} = d_{11} > 0$
   This is a test for the presence of a single variance component, the number of tested parameters is $m = 1$. If the null hypothesis is true then $LR \xrightarrow{\mathsf{D}} Z$ where the random variable Z is distributed as an equal mixture of the constant 0 and the χ^2_1 distribution.

   If the observed value of the LR test statistic is λ the correct asymptotic P-value for the LR test can be calculated as $0.5\mathrm{P}\left[\chi^2_1 \geq \lambda\right]$.

2. $H_0 : \mathbb{D} = \begin{pmatrix} \mathbb{D}_{11} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & 0 \end{pmatrix}$ against $H_1 : \mathbb{D} = \begin{pmatrix} \mathbb{D}_{11} & \boldsymbol{d}_{12} \\ \boldsymbol{d}_{12}^\mathsf{T} & d_{22} \end{pmatrix}$,
   where the dimension of $\mathbb{D}_{11}$ is $q \times q$ and $d_{22} > 0$ is a scalar. This is a test for adding a single variance component to $q > 0$ existing ones, the number of tested parameters is $m = q + 1$. If the null hypothesis is true then $LR \xrightarrow{\mathsf{D}} Z$ where the random variable Z is distributed as an equal mixture of the χ^2_q and the χ^2_{q+1} distributions.

   The correct asymptotic P-value for the LR test is $0.5\mathrm{P}\left[\chi^2_q \geq \lambda\right] + 0.5\mathrm{P}\left[\chi^2_{q+1} \geq \lambda\right]$.

3. $H_0 : \mathbb{D} = \begin{pmatrix} \mathbb{D}_{11} & \mathbf{0} \\ \mathbf{0}^\mathsf{T} & \mathbf{0} \end{pmatrix}$ against $H_1 : \mathbb{D} = \begin{pmatrix} \mathbb{D}_{11} & \mathbb{D}_{12} \\ \mathbb{D}_{12}^\mathsf{T} & \mathbb{D}_{22} \end{pmatrix}$,

   where the dimension of $\mathbb{D}_{11}$ is $q \times q$ and the dimension of $\mathbb{D}_{22} > 0$ is $k \times k$. This is a test for adding k variance components to $q > 0$ existing ones, the number of tested parameters is $m = k(k+1)/2 + qk$. If the null hypothesis is true then $LR \xrightarrow{\mathsf{D}} Z$ where the random variable Z is distributed as an unequal mixture of χ^2 distributions, all of them with degrees of freedom not larger than $k(k+1)/2 + qk$. The weights in the limiting mixture are rather difficult to calculate.

In all these special cases, the true limiting distribution of the LR statistic is stochastically smaller than the asymptotic distribution χ^2_m implied by Theorem 6.3(iii). Therefore, the critical values and p-values based on the standard MLE theory are larger than the true critical values and p-values. Likelihood ratio tests of variance components based on standard MLE theory are *conservative*; they provide tests with true level much smaller than the desired α and have much smaller power to detect deviations from the null hypothesis.

6.4.3 t tests and F tests for fixed effects

Suppose Σ is known and take $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\boldsymbol{Y})$. Consider the problem of estimation and testing linear combinations $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$ of fixed effects for some vector of constants $\boldsymbol{c} \neq \boldsymbol{0}$.

Suppose that σ_e^2 is known, too. Then it is easy to show that for any $\boldsymbol{c} \neq \boldsymbol{0}$

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\sigma_e^2\boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\boldsymbol{c}}} \sim \mathsf{N}(0,1).$$

This could be used to test $H_0 : \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta} = c_0$ and build exact confidence intervals for $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$ if the true variance parameters were known.

Now let σ_e^2 be unknown and use the REML estimator

$$\widehat{\sigma}_e^2 = \frac{1}{n-p}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\mathsf{T}\Sigma^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}}).$$

Then it is equally easy to show that for any $\boldsymbol{c} \neq \boldsymbol{0}$

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2\boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}.$$

This could be used to test $H_0 : \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta} = c_0$ and build exact confidence intervals for $\boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}$ if Σ were known and σ_e^2 unknown.

With both Σ and σ_e^2 unknown, take ML or REML estimators $\widehat{\Sigma}$ and $\widehat{\sigma}_e^2$ of Σ and σ_e^2, respectively and redefine $\widehat{\boldsymbol{\beta}} = (\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\boldsymbol{Y})$. By Theorem 6.3(ii) and the note below it,

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2\boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}}} \xrightarrow{\mathsf{D}} \mathsf{N}(0,1)$$

for any $\boldsymbol{c} \neq \boldsymbol{0}$. However, the asymptotics requires $K \to \infty$ and the approximation may not work well in practical data analyses. Unfortunately, it is no longer possible to obtain the exact small-sample distribution of the left hand side.

There are several options for better approximations of the small-sample distribution than by standard normal:

- Use t_{n-p} as if Σ were known. This is implemeted in the R library `nlme`.
- Use a better approximation $t_{\widehat{f}}$, where $\widehat{f}$ is calculated from the data.
  - Satterthwaite approximation (Satterthwaite 1946)
  - Kenward-Roger approximation (Kenward and Roger 1997)

  These approximations should be more precise than t_{n-p}. Kenward-Roger approximation is the preferred one. Both are implemeted in the SAS procedure `proc MIXED`.

Now take a fixed $m \times p$ matrix $\mathbb{A}$, where $m < p$, $r(\mathbb{A}) = m$. Consider the problem of testing the hypothesis $H_0 : \mathbb{A}\boldsymbol{\beta} = \boldsymbol{\gamma}_0$ against $H_1 : \mathbb{A}\boldsymbol{\beta} \neq \boldsymbol{\gamma}_0$, where $\boldsymbol{\gamma}_0$ is some vector of constants, usually zeros.

If σ_e^2 and Σ are known then

$$\mathbb{A}\widehat{\boldsymbol{\beta}} \sim \mathsf{N}_m\big(\mathbb{A}\boldsymbol{\beta}, \sigma_e^2 \mathbb{A}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{A}^\mathsf{T}\big)$$

and

$$(\mathbb{A}\widehat{\boldsymbol{\beta}} - \mathbb{A}\boldsymbol{\beta})^\mathsf{T}\big[\sigma_e^2 \mathbb{A}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{A}^\mathsf{T}\big]^{-1}(\mathbb{A}\widehat{\boldsymbol{\beta}} - \mathbb{A}\boldsymbol{\beta}) \sim \chi^2_m.$$

If σ_e^2 is unknown and Σ is known then

$$F = \frac{1}{m}(\mathbb{A}\widehat{\boldsymbol{\beta}} - \mathbb{A}\boldsymbol{\beta})^\mathsf{T}\big[\widehat{\sigma}_e^2 \mathbb{A}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\mathbb{A}^\mathsf{T}\big]^{-1}(\mathbb{A}\widehat{\boldsymbol{\beta}} - \mathbb{A}\boldsymbol{\beta}) \sim F_{m,n-p}.$$

If σ_e^2 and Σ are unknown then the statistic F does not have exact $F_{m,n-p}$ distribution but it can be used as an approximation valid for large K. This is implemented in R library `nlme`. Alternatively, one can use an approximation $F_{m,\widehat{f}}$, where $\widehat{f}$ is calculated from the data (Satterthwaite approximation). However, $\widehat{f}$ is different from and more complicated than $\widehat{f}$ used in the t-test.

Satterthwaite approximation for t-test

Let us briefly describe how Satterthwaite approximation for t-test can be obtained.

Write

$$\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}}} = \frac{\frac{\boldsymbol{c}^\mathsf{T}\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^\mathsf{T}\boldsymbol{\beta}}{\sqrt{\sigma_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\boldsymbol{c}}}}{\sqrt{\frac{f\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}}{\sigma_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\Sigma^{-1}\mathbb{X})^{-1}\boldsymbol{c}} \frac{1}{f}}}.$$

The numerator has standard normal distribution. To show that the distribution of this statistic can be approximated by χ^2_f, we seek constants f and ψ such that

$$\frac{f\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}}{\psi} \dot{\sim} \chi^2_f.$$

We use the method of moments to find f and ψ from the two equations

$$\frac{f}{\psi}\,\mathsf{E}\,\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c} = f,$$

$$\frac{f^2}{\psi^2}\,\mathsf{var}\,\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c} = 2f.$$

We get

$$f = \frac{2\big[\mathsf{E}\,\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}\big]^2}{\mathsf{var}\,\widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c}}.$$

Hence the number of degrees of freedom f is estimated by $\widehat{f} = (2\widehat{E}^2)/\widehat{V}$, where

$$\widehat{E} = \widehat{\sigma}_e^2 \boldsymbol{c}^\mathsf{T}(\mathbb{X}^\mathsf{T}\widehat{\Sigma}^{-1}\mathbb{X})^{-1}\boldsymbol{c} \stackrel{\mathsf{df}}{=} h(\widehat{\sigma}_e^2, \widehat{\boldsymbol{\theta}}),$$
$$\widehat{V} = \widehat{\boldsymbol{g}}^\mathsf{T}\widehat{\mathbb{I}}^{-1}\widehat{\boldsymbol{g}},$$
$$\widehat{\mathbb{I}}^{-1} \text{ is the estimated covariance matrix of } (\widehat{\sigma}_e^2, \widehat{\boldsymbol{\theta}}^\mathsf{T})^\mathsf{T}, \text{ and}$$
$$\widehat{\boldsymbol{g}} = \frac{\partial}{\partial(\sigma_e^2, \boldsymbol{\theta})} h(\widehat{\sigma}_e^2, \widehat{\boldsymbol{\theta}}).$$

6.5 Extended Linear Mixed Effects Model

6.5.1 Introduction

So far we have considered the model

$$\boldsymbol{Y}_i = \mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, K,$$

where $\boldsymbol{b}_i \sim \mathsf{N}_q(\boldsymbol{0}, \mathbb{D})$ and $\boldsymbol{\varepsilon}_i \sim \mathsf{N}_{n_i}(\boldsymbol{0}, \sigma_e^2\mathbb{I}_{n_i})$.

Now we remove the assumption of independence and homoskedasticity on the residual terms and assume instead

$$\boldsymbol{\varepsilon}_i \sim \mathsf{N}_{n_i}(\boldsymbol{0}, \sigma_e^2\Lambda_i),$$

where Λ_i is $n_i \times n_i$ positive definite matrix parametrized by a vector of parameters $\boldsymbol{\lambda}$ of a fixed dimension. Marginally,

$$\boldsymbol{Y}_i \sim \mathsf{N}_{n_i}(\mathbb{X}_i\boldsymbol{\beta}, \mathbb{Z}_i\mathbb{D}\mathbb{Z}_i^\mathsf{T} + \sigma_e^2\Lambda_i).$$

Thus, a part of the variance and covariance structure is ascribed to the random effects $\boldsymbol{b}_i$, another part is ascribed to the residual terms $\boldsymbol{\varepsilon}_i$. For a given covariance structure, there may be multiple ways to partition it between the random effects and residual terms, so one has to be careful with the specification of this model to make sure that all the parameters are identifiable.

6.5.2 Parameter estimation

There exists an invertible square root of Λ_i such that

$$\Lambda_i = (\Lambda_i^{1/2})^\mathsf{T}\Lambda_i^{1/2} \quad \text{and} \quad \Lambda_i^{-1} = \Lambda_i^{-1/2}(\Lambda_i^{-1/2})^\mathsf{T}.$$

Transform the observations, regressors and residual terms

$$\boldsymbol{Y}_i^* = (\Lambda_i^{-1/2})^\mathsf{T}\boldsymbol{Y}_i, \quad \mathbb{X}_i^* = (\Lambda_i^{-1/2})^\mathsf{T}\mathbb{X}_i, \quad \mathbb{Z}_i^* = (\Lambda_i^{-1/2})^\mathsf{T}\mathbb{Z}_i, \quad \boldsymbol{\varepsilon}_i^* = (\Lambda_i^{-1/2})^\mathsf{T}\boldsymbol{\varepsilon}_i.$$

Then $\mathsf{var}\,\boldsymbol{\varepsilon}_i^* = \sigma_e^2 \mathbb{I}_{n_i}$ and $\boldsymbol{Y}_i^*$ satisfy the standard one-level LME model

$$\boldsymbol{Y}_i^* = \mathbb{X}_i^* \boldsymbol{\beta} + \mathbb{Z}_i^* \boldsymbol{b}_i + \boldsymbol{\varepsilon}_i^*.$$

The Jacobian of the transformation is $|\Lambda_i|^{-1/2}$. The likelihood of the extended LME model is

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2, \boldsymbol{\lambda} \mid \boldsymbol{Y}) = L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_e^2, \boldsymbol{\lambda} \mid \boldsymbol{Y}^*) \prod_{i=1}^{K} |\Lambda_i|^{-1/2}.$$

The first part is the likelihood of the standard LME model. The same holds for REML:

$$L_R(\boldsymbol{\theta}, \sigma_e^2, \boldsymbol{\lambda} \mid \boldsymbol{Y}) = L_R(\boldsymbol{\theta}, \sigma_e^2, \boldsymbol{\lambda} \mid \boldsymbol{Y}^*) \prod_{i=1}^{K} |\Lambda_i|^{-1/2}.$$

The parameters are estimated by the same decompositions of the log-likelihood as before, with an additional term $-1/2 \sum_{i=1}^{K} \log |\Lambda_i|$.

6.5.3 Generalized least squares

Consider a special case of the extended LME model with no random effects

$$\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

where $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\boldsymbol{0}, \sigma_e^2 \Lambda)$. Suppose $\boldsymbol{\lambda}$ is known. Then

$$\boldsymbol{Y}^* = \mathbb{X}^* \boldsymbol{\beta} + \boldsymbol{\varepsilon}^*$$

satisfies the classical linear model. The LSE of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \left[(\mathbb{X}^*)^\mathsf{T} \mathbb{X}^*\right]^{-1} (\mathbb{X}^*)^\mathsf{T} \boldsymbol{Y}^*$. The parameter σ_e^2 is estimated by the residual sum of squares

$$\widehat{\sigma}_e^2(\boldsymbol{\lambda}) = \frac{1}{n-p} \left\| \boldsymbol{Y}^* - \mathbb{X}^* \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \right\|^2.$$

(this is REML estimator). Plug this into the likelihood to get

$$\ell_M(\boldsymbol{\lambda}) = C - n \log \left\| \boldsymbol{Y}^* - \mathbb{X}^* \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \right\| - \frac{1}{2} \log |\Lambda|$$

and maximize it over $\boldsymbol{\lambda}$ to get the MLE. Or use the REML log-likelihood

$$\ell_R(\boldsymbol{\lambda}) = C - (n-p) \log \left\| \boldsymbol{Y}^* - \mathbb{X}^* \widehat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) \right\| - \frac{1}{2} \log |\Lambda| - \frac{1}{2} \log \left| (\mathbb{X}^*)^\mathsf{T} \mathbb{X}^* \right|$$

and maximize it over $\boldsymbol{\lambda}$ to get the REML estimator.

This method is called *generalized least squares*.

6.5.4 Decomposing variance structure

The ability to specify arbitrary Λ_i (subject to identifiability assumption) allows to include in the model correlated residual terms with time series or spatial structure and/or to model unequal variances of residual terms.

Decompose

$$\varepsilon_{ij} = W_{ij} + \eta_{ij},$$

where η_{ij} are independent variables with equal variance σ_e^2 and W_{ij} are values of a random Gaussian processes with zero mean, independent of η_{ij}. We can take, for example:

- $W_{ij} = W_i(j)$ random processes in discrete time ($j = 1, \dots, n_i$ are indices of ordered observations within subject), independent between subjects.
- $W_{ij} = W_i(t_{ij})$ random processes in continuous time (t_{ij} is the time of the observation Y_{ij}), independent between subjects.
- $W_{ij} = W_i(u_{ij}, v_{ij})$ random fields on the plane ($[u_{ij}, v_{ij}]$ are the coordinates of the observation Y_{ij}), independent between subjects.

Consider $W_{ij} = W_i(t_{ij})$ as an example. Let $W_i(t)$ be weakly stationary with $\mathsf{E}\, W_i(t) = 0$, $\mathsf{var}\, W_i(t) = \tau^2$, $\mathsf{cor}\,(W_i(t), W_i(s)) = \rho(|t-s|)$. Then the variance is decomposed into three components

$$\mathsf{var}\, Y_{ij} = \boldsymbol{Z}_{ij}^{\mathsf{T}} \mathbb{D} \boldsymbol{Z}_{ij} + \tau^2 + \sigma_e^2,$$

the random effects variance $\boldsymbol{Z}_{ij}^{\mathsf{T}} \mathbb{D} \boldsymbol{Z}_{ij}$ that depends on regressors $\boldsymbol{Z}_{ij}$, the variance of the serial component τ^2 and the white noise variance σ_e^2. The covariance has two components

$$\mathsf{cov}\,(Y_{ij}, Y_{ik}) = \boldsymbol{Z}_{ij}^{\mathsf{T}} \mathbb{D} \boldsymbol{Z}_{ik} + \tau^2 \rho(|t_{ij} - t_{ik}|).$$

Denote by $\mathbb{R}_i$ the correlation matrix of $(W_i(t_{i1}), \dots, W_i(t_{in_i}))$. We get

$$\mathsf{var}\, \boldsymbol{Y}_i = \mathbb{Z}_i \mathbb{D} \mathbb{Z}_i^{\mathsf{T}} + \tau^2 \mathbb{R}_i + \sigma_e^2 \mathbb{I}_{n_i}$$

and

$$\Lambda_i = \frac{\tau^2}{\sigma_e^2} \mathbb{R}_i + \mathbb{I}_{n_i}.$$

The process $W_i(t)$ is chosen to generate the desired autocorrelation function, for example

- exponential correlation $\rho(u) = \exp\{-\lambda u\}$
- Gaussian correlation $\rho(u) = \exp\{-\lambda u^2\}$
- compound symmetry $\rho(u) = \lambda$ (beware nonidentifiability of the random intercept)
- $AR(1)$ process $W_{ij} = \lambda W_{i,j-1} + \nu_{ij}$, with correlations $\rho(u) = \lambda^{|j-k|}$
- $ARMA(p, q)$ process

For spatial correlations, one can take some distance function d (Euclidean, L_1, maximal) and calculate distances between measurements $d_{ijk} = d([u_{ij}, v_{ij}], [u_{ik}, v_{ik}])$. Then specify $\rho(d)$, the correlation between two measurements taken at the distance d of each other, for example

- exponential correlation $\rho(d) = \exp\{-\lambda d\}$
- Gaussian correlation $\rho(d) = \exp\{-\lambda d^2\}$
- linear correlation $\rho(d) = (1 - \lambda d) I(d < 1/\lambda)$

There are graphical methods for the choice of the right correlation structure (sample autocorrelation function, sample variogram). For more details, see Diggle et al. (2002, Chapter 5).

6.6 Comparison of LME and GEE Approaches

Group-dependent data with linear mean structure can be analyzed either by LME or by GEE methods. Both methods provide consistent estimators of mean effects $\boldsymbol{\beta}$ as long as the mean structure is correct and the number of independent groups K grows to infinity.

However, LME and GEE differ in some important aspects:

LME

- specifies a detailed model for $\mathsf{var}\, \boldsymbol{Y}_i$
- assumes normality of $\boldsymbol{b}_i$ and $\boldsymbol{\varepsilon}_i$
- yields important information on the variance structure (decomposition of variance, estimates of variance components, random effects, hypotheses tests about variance structure)
- inference on $\boldsymbol{\beta}$ (tests, confidence intervals) is invalid if the variance structure is not correctly specified
- needs relatively large K, its performance when K is moderate or small is uncertain

GEE

- uses a working model for $\mathsf{var}\, \boldsymbol{Y}_i$, which is not assumed to be correct
- does not make assumptions on the distribution of $\boldsymbol{Y}_i$
- does not provide much information on the variance structure
- inference on $\boldsymbol{\beta}$ is valid even if the working variance structure is incorrect, as long as K is large enough
- needs relatively large K, fails when K is small

7 Generalized Linear Mixed Models

7.1 Model and Assumptions

Again, we observe K independent random vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$, where

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^\mathsf{T}, \quad i = 1, \ldots, K.$$

In this chapter, we develop a model for non-normal group-dependent data by introducing random effects into the generalized linear model.

Definition 7.1. The data $\boldsymbol{Y}_i$ satisfy the *generalized linear mixed model*[*] [GLMM] if

1. $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_K$ are independent.
2. There exist unobserved iid q-dimensional random vectors $\boldsymbol{b}_i$ with density $h(\boldsymbol{b}; \boldsymbol{\psi})$ parametrized by an s-dimensional parameter $\boldsymbol{\psi}$ such that $Y_{i1}, \ldots, Y_{in_i}$ are conditionally independent given $\boldsymbol{b}_i$ and the conditional density of Y_{ij} given $\boldsymbol{b}_i$ has the form of the GLM
$$f(y \mid \boldsymbol{b}_i) = \exp\left\{ \frac{y\,\theta_{ij} - b(\theta_{ij})}{\varphi} + c(y, \varphi) \right\}.$$
3. The canonical parameter θ_{ij} depends on the p-dimensional fixed effects covariates $\boldsymbol{X}_{ij}$, fixed regression parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, random effects covariates $\boldsymbol{Z}_{ij}$ of dimension $q \leq p$, and random effects $\boldsymbol{b}_i$ through the linear predictor
$$\eta_{ij} = \boldsymbol{X}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\mathsf{T}\boldsymbol{b}_i.$$
4. There exists a link function g such that $g(\mu_{ij}) = \eta_{ij}$, where $\mu_{ij} = b'(\theta_{ij}) \equiv \mathsf{E}\big[Y_{ij} \,\big|\, \boldsymbol{b}_i\big]$.

Note.

- Conditionally on $\boldsymbol{b}_i$, the response Y_{ij} satisfies the GLM. The presence of the common $\boldsymbol{b}_i$ in all $Y_{i1}, \ldots, Y_{in_i}$ brings in within-group correlations.
- The random effects covariates $\boldsymbol{Z}_{ij}$ are usually taken as a subset of the fixed effects covariates $\boldsymbol{X}_{ij}$.

[*] Česky *zobecněný lineární smíšený model*

- Commonly, the random effects are assumed to be normally distributed

$$\boldsymbol{b}_i \sim \mathsf{N}_q(\mathbf{0}, \mathbb{D}),$$

  where the covariance matrix $\mathbb{D}$ is parametrized by a vector of parameters $\boldsymbol{\psi}$. Then $h(\boldsymbol{b}; \boldsymbol{\psi})$ is a multivariate normal density.

Note. The conditional moments of Y_{ij} given $\boldsymbol{b}_i$ are

$$\begin{aligned} \mathsf{E}\big[Y_{ij} \,\big|\boldsymbol{b}_i\big] &= g^{-1}(\boldsymbol{X}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\mathsf{T}\boldsymbol{b}_i) \\ \mathsf{var}\big[Y_{ij} \,\big|\boldsymbol{b}_i\big] &= \varphi V(\mu_{ij}), \end{aligned}$$

where $V(\mu) = b''(\theta)$ is the variance function. The expectation of Y_{ij} given the covariates is

$$\mathsf{E}\, Y_{ij} = \mathsf{E}\,\mathsf{E}\big[Y_{ij} \,\big|\boldsymbol{b}_i\big] = \mathsf{E}\, g^{-1}(\boldsymbol{X}_{ij}^\mathsf{T}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^\mathsf{T}\boldsymbol{b}_i),$$

where the expectation on the right-hand side is over the distribution of the random effects (conditionally on all covariates). Unless g is linear, the distribution of Y_{ij} given the covariates does not follow the generalized linear model.

7.2 Parameter Estimation

The likelihood for data that satisfy Definition 7.1 can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\psi} \mid \boldsymbol{Y}) = \prod_{i=1}^{K} \int \prod_{j=1}^{n_i} f(Y_{ij} \mid \boldsymbol{\beta}, \boldsymbol{b}_i) h(\boldsymbol{b}_i \mid \boldsymbol{\psi})\, d\boldsymbol{b}_i.$$

Unlike in the linear mixed effects model, here the integral cannot be calculated explicitely even if h is normal.

If the link g is canonical and h is multivariate normal, the likelihood can be expressed as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\psi} \mid \boldsymbol{Y}) &= \prod_{i=1}^{K} (2\pi)^{-q/2} \,|\mathbb{D}|^{-1/2} \times \\ &\times \int \exp\left\{ \frac{1}{\varphi} \Big[\boldsymbol{Y}_i^\mathsf{T} (\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i) - \mathbf{1}_{n_i}^\mathsf{T} b(\mathbb{X}_i\boldsymbol{\beta} + \mathbb{Z}_i\boldsymbol{b}_i) \Big] + \mathbf{1}_{n_i}^\mathsf{T} c(\boldsymbol{Y}_i, \varphi) - \frac{1}{2}\boldsymbol{b}_i^\mathsf{T}\mathbb{D}^{-1}\boldsymbol{b}_i \right\} d\boldsymbol{b}_i, \end{aligned}$$

where the functions b and c are applied to vectors by element-by-element calculation. The log-likelihood has the form

$$\ell(\boldsymbol{\beta}, \boldsymbol{\psi} \mid \boldsymbol{Y}) = -\frac{K}{2} \log |\mathbb{D}| + \sum_{i=1}^{K} \log \int \exp\Big\{ \cdots \Big\} d\boldsymbol{b}_i + C.$$

Many different approaches have been suggested to maximize $\ell(\boldsymbol{\beta}, \boldsymbol{\psi} \mid \boldsymbol{Y})$, the use of the Laplace approximation is one among them.

The Laplace approximation is

$$\int_{\mathbb{R}^q} \mathrm{e}^{Q(\boldsymbol{b})} d\boldsymbol{b} \approx (2\pi)^{q/2} \left| -Q''(\widetilde{\boldsymbol{b}}) \right|^{-1/2} \exp\{Q(\widetilde{\boldsymbol{b}})\},$$

where $\widetilde{\boldsymbol{b}}$ is the point where the maximum of the function Q is attained. This approximation is obtained by replacing Q in the integrand by the second order Taylor expansion of Q around $\widetilde{\boldsymbol{b}}$ and integrating the exponentiated quadratic function as a Gaussian density.

After plugging the Laplace approximation into the log-likelihood, the integral disappears and the resulting expression can be maximized using a combination of a modified IWLS algorithm for estimating $\boldsymbol{\beta}$, modified Henderson's equations for estimating $\boldsymbol{b}_i$, and moment estimation for $\boldsymbol{\psi}$ and φ. There are many different approaches to implementing these ideas. The calculations that need to be performed and the development of formulas and algorithms are rather tedious.

One of approaches leads to the following estimation procedure. At each iterative step, using the parameter estimates obtained at the previous step, calculate successively:

1. The IWLS step:
$$\widehat{\boldsymbol{\beta}} = (\mathbb{X}^{\mathsf{T}} \widehat{\mathbb{W}}^{-1} \mathbb{X})^{-1} \mathbb{X}^{\mathsf{T}} \widehat{\mathbb{W}}^{-1} \widehat{\boldsymbol{Y}}^*,$$
where
$$\widehat{\mathbb{W}} = \begin{pmatrix} \widehat{\mathbb{W}}_1 & \cdots & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \widehat{\mathbb{W}}_K \end{pmatrix}, \quad \widehat{\mathbb{W}}_i = \operatorname{diag}\{\widehat{\varphi} V(\widehat{\mu}_{ij})[g'(\widehat{\mu}_{ij})]^2\}_{j=1}^{n_i} + \mathbb{Z}_i \widehat{\mathbb{D}} \mathbb{Z}_i^{\mathsf{T}}$$
and $\widehat{Y}_{ij}^* = g(\widehat{\mu}_{ij}) + (Y_{ij} - \widehat{\mu}_{ij}) g'(\widehat{\mu}_{ij})$.
2. The Henderson step:
$$\widehat{\boldsymbol{b}}_i = \widehat{\mathbb{D}} \mathbb{Z}_i^{\mathsf{T}} \widehat{\mathbb{W}}_i^{-1} (\widehat{\boldsymbol{Y}}_i^* - \mathbb{X}_i \widehat{\boldsymbol{\beta}}).$$
3. Moment estimation of an unstructured $\mathbb{D}$:
$$\widehat{\mathbb{D}} = \frac{1}{K} \sum_{i=1}^{K} \widehat{\mathsf{E}}(\boldsymbol{b}_i \boldsymbol{b}_i^{\mathsf{T}}) = \frac{1}{K} \sum_{i=1}^{K} [\widehat{\mathsf{E}}(\boldsymbol{b}_i \mid \boldsymbol{Y}_i)]^{\otimes 2} + \frac{1}{K} \sum_{i=1}^{K} \widehat{\mathsf{var}}(\boldsymbol{b}_i \mid \boldsymbol{Y}_i) =$$
$$= \frac{1}{K} \sum_{i=1}^{K} \widehat{\boldsymbol{b}}_i^{\otimes 2} + \frac{1}{K} \sum_{i=1}^{K} \widehat{\mathsf{var}}(\boldsymbol{b}_i \mid \widehat{\boldsymbol{b}}_i).$$
The last term is calculated according to Proposition 6.2.

7.3 Interpretation

Recall that

$$\mathsf{E}\left[Y_{ij} \,\middle|\, \boldsymbol{b}_i\right] = g^{-1}(\boldsymbol{X}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i) \tag{7.1}$$

and

$$\mathsf{E}\, Y_{ij} = \mathsf{E}\, g^{-1}(\boldsymbol{X}_{ij}^{\mathsf{T}}\boldsymbol{\beta} + \boldsymbol{Z}_{ij}^{\mathsf{T}}\boldsymbol{b}_i). \tag{7.2}$$

Thus, $\boldsymbol{\beta}$ in (7.1) expresses the effect of $\boldsymbol{X}_{ij}$ on $\mathsf{E}\, Y_{ij}$ conditionally on $\boldsymbol{b}_i$, that is, when the latent characteristics of the subject do not change. These effects are called *the subject-specific effects*. They describe the effects on $\mathsf{E}\left[Y_{ij} \,\middle|\, \boldsymbol{b}_i\right]$ when the given subject changes the value of $\boldsymbol{X}_{ij}$. In general, the parameters $\boldsymbol{\beta}$ do not compare two *different* subjects that differ in the value of $\boldsymbol{X}_{ij}$.

The effect of $\boldsymbol{X}_{ij}$ on $\mathsf{E}\, Y_{ij}$ unconditionally on $\boldsymbol{b}_i$, that is, in the general population, is called *the population-averaged effect*. Because $\mathsf{E}\, Y_{ij}$ in (7.2) does not necessarily have the form of a GLM with the link function g, the parameters $\boldsymbol{\beta}$ from (7.1) in general do not possess population-averaged interpretation.

There are several special cases when a relationship between the subject-specific and population-averaged model can be found:

- If the link function g is linear (including identity link as a special case) then the subject-specific and population-averaged model have the same form and the parameters $\boldsymbol{\beta}$ are the same in both. This is the case of the LME model of the previous chapter.
- If the link function g is logarithm and $\boldsymbol{b}_i$ are normal then the subject-specific and population-averaged model have the same loglinear form and the parameters $\boldsymbol{\beta}$ are the same in both as long as the covariate does not appear in the random effects structure.
- If the link function g is probit and $\boldsymbol{b}_i$ are normal then the subject-specific and population-averaged model have the same probit form. The parameters $\boldsymbol{\beta}$ in the subject-specific and population averaged model differ.

When the parameters in the subject-specific and population averaged model differ, their value in the population averaged model is always closer to zero than in the subject-specific model.

As an example, take binary Y_{ij} and logistic regression with a single continuous covariate X_{ij} and a random intercept. Let $\pi(x_{ij})$ be the conditional probability of success given x_{ij} and b_i. The model equation is

$$\log \frac{\pi(x_{ij})}{1-\pi(x_{ij})} = \beta_0 + \beta_1 x_{ij} + b_i,$$

which can be rewritten as

$$\pi(x_{ij}) = \frac{\exp\{\beta_0 + \beta_1 x_{ij} + b_i\}}{1 + \exp\{\beta_0 + \beta_1 x_{ij} + b_i\}}.$$

The random intercept b_i can be interpreted as an innate propensity for success of the i-th subject that affects the odds of success at all trials performed by this subject. The parameter β_1 is the subject-specific effect of the covariate X_{ij}; it explains what happens to the success probability of a given subject when the same subject is tested in different conditions.

The population-averaged model is

$$\pi^*(x_{ij}) = \mathsf{E}\,\frac{\exp\{\beta_0 + \beta_1 x_{ij} + b_i\}}{1 + \exp\{\beta_0 + \beta_1 x_{ij} + b_i\}},$$

where the expectation is taken over the distribution of b_i in the whole population. This is no longer a logistic regression model. However, if b_i is normally distributed with a small variance, $\pi^*(x_{ij})$ can be approximated by

$$\pi^*(x_{ij}) \approx \frac{\exp\{\beta_0^* + \beta_1^* x_{ij}\}}{1 + \exp\{\beta_0^* + \beta_1^* x_{ij}\}},$$

where $|\beta_1^*| < |\beta_1|$. The parameter β_1^* is the population-averaged effect of the covariate X_{ij}; it explains what happens to the mean success probability of the whole population when the conditions change. The population-averaged effects are always weaker than the subject-specific effects.

7.4 Comparison of GLMM vs. GEE models

GLMM is a subject-specific model, its parameters have subject-specific interpretation. GEE is a population-averaged model, its parameters have population-averaged interpretation. In general, GLMM estimates different parameters than GEE. Also, GLMM and GEE models with the same link function cannot be both correct (except in special cases, some of which were mentioned in the previous section).

Otherwise, the main points of the discussion in Section 6.6 are still valid.

The choice between the GLMM and the GEE model should be driven mainly by the desired interpretation of the parameters.

Bibliography

Diggle, P., Heagerty, P., Liang, K. and Zeger, S. (2002). *Analysis of Longitudinal Data*, Oxford Statistical Science Series, 2nd edn, Oxford University Press, Oxford.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**(2): 383–385.

Henderson, C. (1984). *Applications of linear models in animal breeding*, University of Guelph.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**(4): 805–820.

Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood, *Biometrics* **53**(3): 983–997.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data, *Biometrics* **38**(4): 963.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data, *Journal of the American Statistical Association* **83**(404): 1014–1022.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models, *Journal of the Royal Statistical Society. Series A* **135**(3): 370–384.

Palmgren, J. (1981). The Fisher information matrix for log linear-models arguing conditionally on observed explanatory variables, *Biometrika* **68**(2): 563–566.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**(3): 545–554.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44**(4): 1033–1048.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses, *Biometrics* **47**(3): 825–839.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components, *Biometrics Bulletin* **2**(6): 110–114.

Stram, D. O. and Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model, *Biometrics* **50**(4): 1171–1177.

Stram, D. O. and Lee, J. W. (1995). Corrections: Variance component testing in the longitudinal mixed effects model, *Biometrics* **51**(3): 1196–1196.

Wedderburn, R. (1974). Quasi-likelihood functions, generalized linear-models, and Gauss-Newton method, *Biometrika* **61**(3): 439–447.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

White, H. (1982). Maximum-likelihood estimation of mis-specified models, *Econometrica* **50**(1): 1–25.

Yan, J. and Fine, J. (2004). Estimating equations for association structures, *Statistics in Medicine* **23**(6): 859–874.

Index