NMST412/NMFM402 Generalized Linear Models

Course Notes

Michal Kulich

Last modified on July 30, 2026.

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

These course notes contain the whole contents of the course "NMST412/NMFM402 Generalized Linear Models", which is a part of the curricula of the Master's programs "Probability, Mathematical Statistics and Econometrics" and "Financial and Insurance Mathematics".

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 July 30, 2026

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 the response Y_i can be written as

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

where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots \beta_{0p})^\mathsf{T}$ is a vector of unknown *regression parameters (coefficients)*[§] and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables such that $\mathsf{E}\left[\varepsilon_i \,\middle|\, \boldsymbol{X}_i\right] = 0$, and $\mathsf{var}\left[\varepsilon_i \,\middle|\, \boldsymbol{X}_i\right] = \sigma^2$.$\nabla$

Note. The unobserved random variables ε_i are called *error terms (disturbances)*[¶], σ^2 is called *residual variance*[‖].

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

Note. Another convenient formulation of the model is based on conditional moments and it avoids the expression of the error terms:

The linear regression model holds if and only if

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

Thus, the linear regression model specifies the first two conditional moments of Y_i given $\boldsymbol{X}_i$.

Note. We will use the notation E, var for the conditional expectation and variance, respectively, 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 $X_{i1} = 1$, we have

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

and, with $\boldsymbol{e}_j = (0, \dots, 0, 1, 0, \dots, 0)^\mathsf{T}$ being a p-vector of zeros with 1 at the j-th position,

$$\beta_{0j} = \mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i = \boldsymbol{x} + \boldsymbol{e}_j\right] - \mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i = \boldsymbol{x}\right], \quad j = 2, \dots, p.$$

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*

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

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*[*].

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

- The projection matrix $\mathbb{H} \overset{\text{df}}{=} \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ 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 $(\mathbb{I}_n - \mathbb{H})\mathbb{X} = \mathbf{0}$.
- $\mathsf{E}\,\widehat{\boldsymbol{Y}} = \mathbb{X}\boldsymbol{\beta}_0$, $\mathsf{var}\,\widehat{\boldsymbol{Y}} = \sigma^2\mathbb{H}$.
- The random vector $\boldsymbol{u} \overset{\text{df}}{=} \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\mathbb{I}_n - \mathbb{H})\boldsymbol{Y}$ is called *the vector of residuals*[*]. It satisfies $\mathsf{E}\,\boldsymbol{u} = \mathbf{0}$, $\mathsf{var}\,\boldsymbol{u} = \sigma^2(\mathbb{I}_n - \mathbb{H})$.
- The random variable

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

  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 $\boldsymbol{\varepsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2\mathbb{I}_n)$.

Proposition 1.1. *Under normality,*

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

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

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

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

(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*

$$\frac{\boldsymbol{c}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}_0}{\sqrt{\widehat{\sigma}^2\boldsymbol{c}^T(\mathbb{X}^T\mathbb{X})^{-1}\boldsymbol{c}}} \sim t_{n-p}$$

(viii) Assume the model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$*, where* $\mathbb{X} = (\mathbb{X}_A|\mathbb{X}_B)$ *and* $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_A^T, \boldsymbol{\beta}_B^T)^T$*,* $\boldsymbol{\beta}_B \in \mathbb{R}_m$*,* $\boldsymbol{\beta}_A \in \mathbb{R}^{p-m}$*, 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 $(Y_i, \boldsymbol{X}_i)$, $i = 1, \dots, n$, 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) $\widehat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$ *as* $n \to \infty$,

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

(ii) $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, \mathbb{D}_X^{-1} \mathbb{V}_X \mathbb{D}_X^{-1})$ *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. Implications for Data Analysis

The asymptotic results we have just summarized indicate that linear regression model with ordinary least squares estimation of regression parameters can be used to obtain asymptotically correct statistical inference even if the response is not normal and the error terms do not have equal variance. We only need to have enough observations available for analysis so that the asymptotic results provide a reasonable approximation of the true distribution of the parameter estimator and other quantities of interest.

In this aspect, linear regression is actually a robust nonparametric statistical procedure.

(a) If the responses are *normal* and possess *equal variances* we can perform exact statistical inference based on Proposition 1.1 regardless of the size of the dataset (for any $n > p$).

(b) If the responses are *not normal* but have *equal variances* we can perform asymptotic inference based on Proposition 1.2 for large enough number of observations.

(c) If the responses are *not normal* and have *unequal variances* we can perform asymptotic inference based on Proposition 1.3 with sandwich variance estimator for large enough number of observations.

What number of observations is large enough to trust the asymptotic approaches (b) and (c) depends on the complexity of the linear model.

Furthermore, if the error variances are unequal but are known up to a proportionality constant, i.e., $\mathsf{var}\, Y_i = \sigma^2 w_i$ with known w_i, weighted least squares estimation can be used instead of the sandwich.

1.6. Interpretation with Transformed Response

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

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

and, $\boldsymbol{e}_j$ being the j-th unit vector of the length p,

$$\beta_{0j} = \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x} + \boldsymbol{e}_j\right] - \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i = \boldsymbol{x}\right].$$

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 $\mathsf{E}\left[g(Y_i) \,\middle|\, \boldsymbol{X}_i\right] = \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}_0$ with $\mathsf{var}\left[g(Y_i) \,\middle|\, \boldsymbol{X}_i\right] = \sigma^2$. 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 $\mathsf{E}\, Y_i$ in this model cannot be expressed.

The only special case (apart from linear g) when the transformed model says anything useful about $\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i\right]$ 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

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

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

$$\mathrm{e}^{\beta_{0j}} = \frac{\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i = \boldsymbol{x} + \boldsymbol{e}_j\right]}{\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i = \boldsymbol{x}\right]}, \quad j = 2,\ldots,p.$$

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

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 aspects: it admits a wider choice of distributions for Y_i (distributions from the exponential family) and it allows some flexibility in the relationship between $\mathsf{E}\, Y_i$ and $\boldsymbol{X}_i^\top \boldsymbol{\beta}_0$.

2.1. Exponential Family

2.1.1. Parametrization, moments

Definition 2.1. A distribution of a real-valued random variable belongs to *the exponential family* of distributions[*] if its density (w.r.t. some σ-finite measure) 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;

The expression (2.1) is called *the canonical form of the density*[§]. ∇

Example: Normal distribution

$Y \sim \mathsf{N}(\mu,\sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$.

$$\begin{aligned} f(x;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{-\frac{(x-\mu)^2}{2\sigma^2}\Big\} \\ &= \frac{1}{\sqrt{2\pi}\sigma}\exp\Big\{\frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{x^2}{2\sigma^2}\Big\} \\ &= \exp\Big\{\frac{x\mu - \mu^2/2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\Big\}. \end{aligned}$$

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

$$\theta = \mu, \quad \varphi = \sigma^2, \quad b(\theta) = \frac{\theta^2}{2}, \quad c(x,\varphi) = -\frac{x^2}{2\varphi} - \frac{1}{2}\log(2\pi\varphi).$$

Example: Gamma distribution

$Y \sim \Gamma(a,p)$, $a > 0$, $p > 0$, $Y > 0$ a.s.

$$\begin{aligned} f(x;a,p) &= \frac{a^p}{\Gamma(p)} x^{p-1} \exp\{-ax\} \\ &= \exp\Big\{-ax + p\log a + (p-1)\log x - \log\Gamma(p)\Big\} \\ &= \exp\Big\{\frac{-(a/p)x + \log(a/p)}{1/p} + (p-1)\log x + p\log p - \log\Gamma(p)\Big\} \end{aligned}$$

$$\theta = -\frac{a}{p}, \quad \varphi = 1/p, \quad b(\theta) = -\log(-\theta)$$
$$c(x,\varphi) = (1/\varphi - 1)\log x - \log\varphi/\varphi - \log\Gamma(1/\varphi).$$

Example: Inverse Gaussian distribution

$Y \sim \mathsf{IG}(\mu,\lambda)$, $\mu > 0$, $\lambda > 0$, $Y > 0$ a.s.

$$\begin{aligned} f(x;\mu,\lambda) &= \sqrt{\frac{\lambda}{2\pi x^3}} \exp\Big\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\Big\} \\ &= \sqrt{\frac{\lambda}{2\pi x^3}} \exp\Big\{-\frac{\lambda x^2}{2\mu^2 x} + \frac{\lambda x\mu}{\mu^2 x} - \frac{\lambda\mu^2}{2\mu^2 x}\Big\} \\ &= \exp\Big\{\frac{-x/(2\mu^2) + 1/\mu}{1/\lambda} + \frac{1}{2}\log\frac{\lambda}{2\pi x^3} - \frac{\lambda}{2x}\Big\}. \end{aligned}$$

$$\theta = -\frac{1}{2\mu^2}, \quad \varphi = 1/\lambda, \quad b(\theta) = -\sqrt{-2\theta}, \quad c(x,\varphi) = -\frac{1}{2}\log(2\pi x^3\varphi) - (2x\varphi)^{-1}.$$

This is a continuous distribution on the positive halfline. It is related to χ^2 distribution through the transformation

$$\frac{\lambda(X-\mu)^2}{\mu^2 X} \sim \chi^2_1.$$

Example: Poisson distribution

$Y \sim \mathsf{Po}(\lambda)$, $\lambda > 0$, values $0, 1, 2, \ldots$

$$f(x;\lambda) = \frac{\lambda^x}{x!}\exp\{-\lambda\} = \exp\bigl\{x\log\lambda - \lambda - \log x!\bigr\}.$$

$$\theta = \log\lambda, \quad \varphi = 1, \quad b(\theta) = \exp(\theta), \quad c(x,\varphi) = -\log x!$$

Example: Alternative distribution

$Y \sim \mathsf{Alt}(p)$, $p \in (0,1)$, values $0, 1$.

$$f(x;p) = p^x(1-p)^{1-x} = \exp\bigl\{x\log p + (1-x)\log(1-p)\bigr\} = \exp\Bigl\{x\log\frac{p}{1-p} + \log(1-p)\Bigr\}.$$

$$\theta = \log\frac{p}{1-p}, \quad \varphi = 1, \quad b(\theta) = \log(1+\mathrm{e}^\theta), \quad c(x,\varphi) = 0.$$

The next lemma shows that for distributions of exponential family, the first two moments can be obtained from the canonical form of the density by a simple calculation.

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

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

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

$$\begin{aligned} \mathsf{E}\,Y &= b'(\theta), \\ \mathsf{var}\,Y &= \varphi\, b''(\theta). \end{aligned}$$ ♢

Proof. Suppose the density $f(x;\theta,\varphi)$ exists with respect to a σ-finite measure ν and denote the support $A = \{x : f(x;\theta,\varphi) > 0\}$. We have

$$\begin{aligned} m_Y(t) = \mathsf{E}\,\mathrm{e}^{tY} &= \int_A \exp\left\{\frac{x\theta + xt\varphi - b(\theta)}{\varphi} + c(x,\varphi)\right\} d\nu(x) \\ &= \int_A \exp\left\{\frac{x(\theta+t\varphi) - b(\theta+t\varphi)}{\varphi} + c(x,\varphi)\right\} d\nu(x) \cdot \exp\left\{\frac{b(\theta+t\varphi) - b(\theta)}{\varphi}\right\} \\ &= \exp\left\{\frac{b(\theta+t\varphi) - b(\theta)}{\varphi}\right\}. \end{aligned}$$

The moments can be calculated by differentiation of $m_Y(t)$ at $t = 0$. We have $\mathsf{E}\, Y = m'_Y(0)$ and

$$m'_Y(t) = m_Y(t)\frac{b'(\theta + t\varphi)}{\varphi}\varphi,$$

so $\mathsf{E}\, Y = m'_Y(0) = b'(\theta)m_Y(0) = b'(\theta)$. Next, $\mathsf{E}\, Y^2 = m''_Y(0)$ and

$$m''_Y(t) = m_Y(t)\big[\, b'(\theta + t\varphi)\big]^2 + m_Y(t)b''(\theta + t\varphi)\varphi$$

so $\mathsf{E}\, Y^2 = m''_Y(0) = \varphi\, b''(\theta) + \big[\, b'(\theta)\big]^2$. Hence,

$$\mathsf{var}\, Y = \mathsf{E}\, Y^2 - \big(\mathsf{E}\, Y\big)^2 = \varphi\; b''(\theta). \qquad \square$$

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

Note. Since $\mathsf{var}\, Y = \varphi\; b''(\theta) > 0$, b must be a strictly convex function and b' is strictly increasing. Hence b' has a well-defined inverse and there exists a function $V(\mu)$ of the mean μ such that $\mathsf{var}\, Y = \varphi V(\mu)$. It satisfies the equation $b''(\theta) = V(b'(\theta))$ or $V(\mu) = b''\big((b')^{-1}(\mu)\big)$.

Definition 2.2. The function $V(\mu)$ such that $\mathsf{var}\, Y = \varphi\, V(\mu)$ is called *the variance function**. ∇

Note.

- Different distributions that belong to the exponential family must have different variance functions.
- Within the exponential family, the variance function determines the distribution of Y. However, not every function V is a variance function of some distribution from the exponential family.

Example: Normal distribution

For $Y \sim \mathsf{N}(\mu, \sigma^2)$, we have $\theta = \mu$, $\varphi = \sigma^2$, and $b(\theta) = \frac{\theta^2}{2}$. Hence

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

The normal distribution is the only distribution in exponential family with constant variance function, i.e., the variance is unrelated to the mean. (Recall the assumption of homoskedasticity in linear regression!).

* Česky *rozptylová funkce*

Example: Gamma distribution

For $Y \sim \Gamma(a,p)$, we have $\theta = -\frac{a}{p}$, $\varphi = 1/p$, and $b(\theta) = -\log(-\theta)$. Hence

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

Example: Inverse Gaussian distribution

For $Y \sim \mathsf{IG}(\mu,\lambda)$, we have $\theta = -\frac{1}{2\mu^2}$, $\varphi = 1/\lambda$, and $b(\theta) = -\sqrt{-2\theta}$. Hence

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

Example: Poisson distribution

For $Y \sim \mathsf{Po}(\lambda)$, we have $\theta = \log\lambda$, $\varphi = 1$, and $b(\theta) = \exp(\theta)$. Hence

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

Example: Alternative distribution

For $Y \sim \mathsf{Alt}(p)$, we have $\theta = \log\frac{p}{1-p}$, $\varphi = 1$, and $b(\theta) = \log(1+\mathrm{e}^\theta) = \log(1-p)$. Hence

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

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 $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\mathsf{T}$. We will discuss maximum likelihood estimation of the canonical parameter θ with iid data. Summary of the maximum likelihood theory together with notation we use throughout this text is provided in the Appendix starting on p. 99.

The likelihood for exponential family 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

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

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

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

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 $\sum_{i=1}^{n}Y_i=nb'(\widehat{\theta}_n)$. 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 Theorem A.4 in the Appendix 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 $\mu_i \overset{\mathsf{df}}{=} \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X}_i\right]$ 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

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

* Česky *zobecněný lineární model* † Česky *lineární prediktor* ‡ Česky *linková funkce*

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 homoskedasticity 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 Section 2.8.8). 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. ▽

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 $g'(\mu_i) = 1/V(\mu_i)$. ♢

Proof. The link function g maps the mean $\mu_i = b'(\theta_i)$ to the linear predictor η_i: $g(\mu_i) = \eta_i$. The canonical link satisfies $\eta_i = \theta_i$.

For canonical link, $g(b'(\theta_i)) = \theta_i$, hence $g = (b')^{-1}$. This proves (i).

Differentiating the equality $g(b'(\theta_i)) = \theta_i$, we get $g'(b'(\theta_i))b''(\theta_i) = 1$. Because $b'(\theta_i) = \mu_i$ and $b''(\theta_i) = V(\mu_i)$, we get $g'(\mu_i)V(\mu_i) = 1$. This proves (ii). □

Note. For each distribution f from the exponential family, there is a unique (up to a non-zero proportionality constant) canonical link function. Two distributions cannot share the same canonical link. Canonical link functions have certain numerical advantages that will become apparent later on. However, some canonical link functions violate the conditions we require and are difficult to interpret (see examples below).

Example: Normal distribution

For normal distribution, the canonical parameter is $\theta_i = \mu_i$, and the dispersion parameter is $\varphi = \sigma^2$. Let $Y_i \sim \mathsf{N}(\mu_i, \sigma^2)$ with $g(\mu_i) = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.

* Česky *kanonický link*

We know that

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

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) = \mu_i$ (identity link).

So the canonical GLM for the normal distribution is $\mathsf{E}\, Y_i = \eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$, the normal linear model.

Example: Gamma distribution

For gamma distribution, the canonical parameter is $\theta_i = -\frac{a_i}{p}$, and the dispersion parameter is $\varphi = 1/p$. So, we take $Y_i \sim \Gamma(a_i, p)$ with the mean $\mu_i = p/a_i$ and link $g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$.

We know that

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

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) \propto 1/\mu_i$ (inverse link — after dropping the minus sign). It is a function which is discontinuous at 0 and not strictly monotone.

The canonical GLM for the gamma distribution is $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = 1/\boldsymbol{X}_i^\top \boldsymbol{\beta}$. The model can be interpreted only when the linear predictors have all either positive or negative signs.

Example: Inverse Gaussian distribution

For inverse Gaussian distribution, the canonical parameter is $\theta_i = -\frac{1}{2\mu_i^2}$, and the dispersion parameter is $\varphi = 1/\lambda$. So, we take $Y_i \sim \mathsf{IG}(\mu_i, \lambda)$ with the mean μ_i and link $g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$.

We know that

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

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) \propto 1/\mu_i^2$ (squared inverse link — after dropping the constant -2). It is a function which is discontinuous at 0 and not strictly monotone.

The canonical GLM for the inverse Gaussian distribution is $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = 1/\sqrt{\boldsymbol{X}_i^\top \boldsymbol{\beta}}$. The model can be interpreted only when the linear predictors all have positive signs.

Example: Poisson distribution

For Poisson distribution, the canonical parameter is $\theta_i = \log \lambda_i$, and the dispersion parameter is $\varphi = 1$. So, we take $Y_i \sim \mathsf{Po}(\lambda_i)$ with the mean λ_i and link $g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$.

We know that

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

The canonical link is $g(\mu_i) = (b')^{-1}(\mu_i) = \log \mu_i$ (log link).

The canonical GLM for Poisson distribution is $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}$. This is called *the loglinear model*.

Example: Alternative distribution

For alternative distribution, the canonical parameter is $\theta_i = \log \frac{p_i}{1-p_i}$, and the dispersion parameter is $\varphi = 1$. So, we take $Y_i \sim \mathsf{Alt}(p_i)$ with the mean p_i and link $g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$.

We know that

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

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

The canonical GLM for alternative distribution is $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = \frac{\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}$. This is called *the logistic regression model*.

Choice of the link function

Canonical links provide very attractive options for the selection of the link function for normal, Poisson and alternative distributions. For these distributions, we always prefer the canonical link unless there is a very strong reason (given by the nature of the application) to select a different link function. For gamma and inverse Gaussian distributions, the canonical links are problematic because they do not even satisfy the assumptions we put on link functions. Also, they are hard to interpret.

Denote by $\mathscr{M}$ the parametric space for the mean of the response (the set of all possible values of the mean). Then g maps $\mathscr{M}$ to $\mathbb{R}$, which is the space of all possible values of the linear predictor. The inverse link g^{-1} should map $\mathbb{R}$ to $\mathscr{M}$.

For non-negative random variables, such as from gamma or inverse Gaussian distributions, $\mathscr{M} = (0, \infty)$. A reasonable inverse link g^{-1} should map $\mathbb{R}$ to $(0, \infty)$, but this is not the case for the canonical links of these two distributions. On the other hand, a reasonable link that maps the two sets correctly is the log-link. For this link, we get $\mu_i = \exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}$, which is in $\mathscr{M}$ for any value of the parameter vector $\boldsymbol{\beta}$.

For the alternative distribution, $\mathscr{M} = (0, 1)$. A reasonable inverse link g^{-1} should map $\mathbb{R}$ to $(0, 1)$ and be strictly monotone. We can choose such links from distribution functions of continuous random variables with positive densities over $\mathbb{R}$. On the other hand, the link

functions are quantile functions of such distributions. The logistic link is the quantile function of the standard logistic distribution.

Parametrizations of the GLM

The primary parameters in the GLM are the regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\mathsf{T}$. 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* $\eta_1, \dots, \eta_n$;
- by the *means* $\mu_1 \equiv \mathsf{E}\, Y_1, \dots, \mu_n \equiv \mathsf{E}\, Y_n$;
- by the *canonical parameters* $\theta_1, \dots, \theta_n$.

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

[*] Česky *saturovaný model*

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.

Note. When all covariates are discrete (with a finite number of values), the largest possible number of parameters in the model is equal to the number of possible distinct values of the covariate vector $\boldsymbol{X}_i$, which is usually smaller than n and does not change as the number of observations increases. In this setting, the saturated model behaves differently.

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,

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

where $\mu_0 = \mathrm{E}\, Y_i$ (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 $h(\boldsymbol{x})$ be the marginal density of $\boldsymbol{X}_i$ (with no assumptions about it except finite second moments). Let $(Y_i, \boldsymbol{X}_i)$,

[*] Česky *nulový model*

$i = 1, \ldots, n$, 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 $f(y \mid \boldsymbol{x}, \boldsymbol{\beta}_0, \varphi_0)h(\boldsymbol{x})$, the full likelihood is

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

and the full log-likelihood is

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

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

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

(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 estimated expectation under the model (conditional on the covariates).

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

Proof (of Theorem 2.3).

(i) The score function is calculated by the chain rule:

$$\boldsymbol{U}(\boldsymbol{\beta} \mid Y_i) = \frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{\varphi_0}[Y_i\theta_i - b(\theta_i)] = \frac{\partial}{\partial \theta} \frac{1}{\varphi_0}[Y_i\theta_i - b(\theta_i)] \cdot \frac{\partial \theta_i}{\partial \mu} \cdot \frac{\partial \mu_i}{\partial \eta} \cdot \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}$$

This is a product of four terms. The first term is $\frac{1}{\varphi_0}(Y_i - \mu_i)$. The next two terms can be calculated by the formula for the derivative of the inverse function. We have

$$\frac{\partial \theta_i}{\partial \mu} = \frac{\partial (b')^{-1}(\mu_i)}{\partial \mu} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)},$$

* Česky *vyrovnané hodnoty*

and

$$\frac{\partial \mu_i}{\partial \eta} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta} = \frac{1}{g'(\mu_i)}.$$

Finally, $\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{X}_i^\top \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_i$. So we have

$$\boldsymbol{U}(\boldsymbol{\beta} \mid Y_i) = \frac{Y_i - \mu_i}{\varphi_0 V(\mu_i) g'(\mu_i)} \boldsymbol{X}_i = \frac{1}{\varphi_0} \underbrace{\frac{1}{V(\mu_i)[g'(\mu_i)]^2}}_{\stackrel{\text{df}}{=} w(\mu_i) > 0} g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i.$$

Because the conditional expectation (given $\boldsymbol{X}_i$) of $Y_i - \mu_i$ is 0 when μ_i is evaluated at the true parameter $\boldsymbol{\beta}_0$, the conditional expectation of $\boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i)$ is zero and the unconditional expectation is zero as well. This proves that $\mathsf{E}\, \boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i) = \boldsymbol{0}$.

The next two points (ii) and (iii) are obvious.

(iv) For the canonical link, we know by Lemma 2.2 that $g'(\mu_i) = 1/V(\mu_i)$. Hence $w(\mu_i) = V(\mu_i)$ and $w(\mu_i)g'(\mu_i) = 1$. The rest is easy. □

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

Theorem 2.4. *(on information matrices in the GLM) Let the definition of the GLM hold. Let $\mathsf{E}_X w(\mu_i)\boldsymbol{X}_i^{\otimes 2}$ 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 $I_n(\boldsymbol{\beta} \mid \boldsymbol{Y}) = n^{-1}\sum_{i=1}^n I(\boldsymbol{\beta} \mid Y_i)$.

(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

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

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

(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

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

where $\hat{\mathbb{W}}$ is the $n \times n$ diagonal matrix $\operatorname{diag}(w(\widehat{\mu}_1),\dots,w(\widehat{\mu}_n))$. When φ_0 is unknown it is replaced by a consistent method-of-moments estimator $\widehat{\varphi}_n$ to be introduced in Section 2.5.

Proof (of Theorem 2.4).

(i) The contribution to the observed information matrix can be calculated as follows.

$$I(\boldsymbol{\beta} \mid Y_i) = -\frac{\partial}{\partial \boldsymbol{\beta}^\top}\boldsymbol{U}(\boldsymbol{\beta} \mid Y_i) = -\frac{1}{\varphi_0}\frac{\partial w(\mu_i)g'(\mu_i)(Y_i-\mu_i)}{\partial \mu}\boldsymbol{X}_i \cdot \frac{\partial \mu_i}{\partial \eta}\cdot\frac{\partial \eta_i}{\partial \boldsymbol{\beta}^\top}.$$

We already know from the proof of Theorem 2.3 that

$$\frac{\partial \mu_i}{\partial \eta} = \frac{1}{g'(\mu_i)} \quad \text{and} \quad \frac{\partial \eta_i}{\partial \boldsymbol{\beta}^\top} = \boldsymbol{X}_i^\top.$$

It remains to calculate the derivative of the product of three functions of μ_i. We get

$$\frac{\partial w(\mu_i)g'(\mu_i)(Y_i-\mu_i)}{\partial \mu} = w'(\mu_i)g'(\mu_i)(Y_i-\mu_i) + w(\mu_i)g''(\mu_i)(Y_i-\mu_i) - w(\mu_i)g'(\mu_i).$$

Putting all the terms together and separating out the part that does not depend on $(Y_i-\mu_i)$, we get

$$I(\boldsymbol{\beta} \mid Y_i) = \frac{1}{\varphi_0}w(\mu_i)\boldsymbol{X}_i\boldsymbol{X}_i^\top - \frac{1}{\varphi_0}\underbrace{\left[w'(\mu_i) + w(\mu_i)\frac{g''(\mu_i)}{g'(\mu_i)}\right](Y_i-\mu_i)\boldsymbol{X}_i\boldsymbol{X}_i^\top}_{\stackrel{\text{df}}{=}\mathbb{J}_i}$$

and the result follows. Notice that the first part is a positive semi-definite matrix while the second part may be anything.

(ii) Because $\mathbb{J}_i$ is a product of $Y_i-\mu_i$ (which has zero conditional expectation given $\boldsymbol{X}_i$ at the true $\boldsymbol{\beta}_0$) and terms that depend on $\boldsymbol{X}_i$ but not on Y_i, its expectation at the true $\boldsymbol{\beta}_0$ is a zero matrix. It follows that

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

Next,

$$\operatorname{var} \boldsymbol{U}(\boldsymbol{\beta}_0 \mid Y_i) = \operatorname{var} \frac{1}{\varphi_0} w(\mu_i) g'(\mu_i)(Y_i - \mu_i)\boldsymbol{X}_i = \mathsf{E}_X \frac{1}{\varphi_0^2} [w(\mu_i) g'(\mu_i)]^2 \operatorname{var}\left[Y_i \,\middle|\, \boldsymbol{X}_i \right] \boldsymbol{X}_i^{\otimes 2}$$
$$= \mathsf{E}_X \frac{[w(\mu_i) g'(\mu_i)]^2 \varphi_0 V(\mu_i)}{\varphi_0^2} \boldsymbol{X}_i^{\otimes 2} = \frac{1}{\varphi_0} \mathsf{E}_X w(\mu_i) \boldsymbol{X}_i^{\otimes 2} = I(\boldsymbol{\beta}_0 \mid Y_i).$$

(iii) We have $w(\mu_i) = \frac{1}{V(\mu_i)[g'(\mu_i)]^2}$. For the canonical link, $g'(\mu_i) = 1/V(\mu_i)$ by Lemma 2.2, hence $g'(\mu_i) = 1/w(\mu_i)$. Next,

$$g''(\mu_i) = -\frac{w'(\mu_i)}{w^2(\mu_i)}.$$

Hence

$$\frac{g''(\mu_i)}{g'(\mu_i)} w(\mu_i) = -w'(\mu_i) \quad \text{and} \quad \left[w'(\mu_i) + w(\mu_i) \frac{g''(\mu_i)}{g'(\mu_i)} \right] = 0.$$

□

2.4. The Iterative Weighted Least Squares Algorithm for Fitting the GLM

The parameters of the GLM can be estimated by a numerical algorithm called *iterative weighted least squares*[*] [IWLS]. It is based on the following result.

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

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

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

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

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

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 $\widehat{\mu}_i$:

$$g(Y_i) \approx g(\widehat{\mu}_i) + g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i).$$

Unlike $g(Y_i)$, the adjusted dependent variable can be calculated even if Y_i is outside of the domain of g, for example when $g \equiv \log$ and $Y_i \sim \mathsf{Po}(\mu_i)$ attains the value of zero.

[*] 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

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

Proof (of Theorem 2.5). Take the obvious equality

$$\Big(\sum_{i=1}^{n} w(\widehat{\mu}_i) \boldsymbol{X}_i \boldsymbol{X}_i^{\mathsf{T}}\Big)\widehat{\boldsymbol{\beta}}_n = \Big(\sum_{i=1}^{n} w(\widehat{\mu}_i) \boldsymbol{X}_i \boldsymbol{X}_i^{\mathsf{T}}\Big)\widehat{\boldsymbol{\beta}}_n$$

and add zero to the right-hand side in the form of the likelihood equations

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

Rearrange the right-hand side to get

$$\Big(\sum_{i=1}^{n} w(\widehat{\mu}_i) \boldsymbol{X}_i \boldsymbol{X}_i^{\mathsf{T}}\Big)\widehat{\boldsymbol{\beta}}_n = \sum_{i=1}^{n} w(\widehat{\mu}_i) \boldsymbol{X}_i \big[\boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}_n + g'(\widehat{\mu}_i)(Y_i - \widehat{\mu}_i)\big],$$

where the bracket contains the value $\widehat{Z}_i$ of the adjusted dependent variable. Rewrite the result in a matrix form as

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

This completes the proof. □

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 (see Appendix A.1.2, bottom of page 101).
- According to (2.9), the matrix $(\mathbb{X}^\mathsf{T}\hat{\mathbb{W}}^{(k)}\mathbb{X})^{-1}$ estimates (up to the proportionality constant φ_0) 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

$$\widehat{\varphi}_n = \frac{X^2}{n-p}. \tag{2.11}$$

▽

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

[*] Česky *Pearsonovo chí kvadrát*

Note. The notation $\|\cdot\|$ means the Euclidean norm. 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 $\boldsymbol{\beta}_0$.

Note. The moment estimator $\widehat{\varphi}_n$ is used instead of φ_0 in all statistics that need to be evaluated. The asymptotic distributions of these statistics are not affected (Cramér-Slutski Theorem).

Proof. We have

$$\widehat{\varphi}_n = \frac{1}{n-p}\sum_{i=1}^{n} h(Y_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}_n).$$

Decompose this as follows:

$$\widehat{\varphi}_n = \frac{1}{n-p}\sum_{i=1}^{n} h(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}_0) + \frac{1}{n-p}\sum_{i=1}^{n} [h(Y_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}_n) - h(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}_0)].$$

The first summand is an average of iid terms that converges in probability by the weak law of large numbers to

$$\mathsf{E}\, h(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}_0) = \mathsf{E}\,\mathsf{E}\Big[\frac{(Y_i-\mu_i)^2}{V(\mu_i)}\,\Big|\boldsymbol{X}_i\Big] = \mathsf{E}\,\frac{\varphi_0 V(\mu_i)}{V(\mu_i)} = \varphi_0.$$

We need to prove that the second summand converges in probability to 0. Take its Euclidean norm, ignore the subtraction of p from n in the denominator, and bound it from above using a one-step Taylor expansion

$$\left\| \frac{1}{n}\sum_{i=1}^{n} [h(Y_i, \boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}_n) - h(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}_0)] \right\| \le \left\| \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right\| \frac{1}{n}\sum_{i=1}^{n} \left\| h'(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}^*) \right\|,$$

where $\boldsymbol{\beta}^*$ lies on the line segment between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_0$, and $h'(y, \boldsymbol{x}, \beta) = \partial h / \partial \boldsymbol{\beta}$. The estimator $\widehat{\boldsymbol{\beta}}_n$ is consistent, so $\left\| \widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right\| \xrightarrow{\mathsf{P}} 0$ and $\|\boldsymbol{\beta}^* - \boldsymbol{\beta}_0\| \xrightarrow{\mathsf{P}} 0$.

It remains to show that $\frac{1}{n}\sum_{i=1}^{n} \left\| h'(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}^*) \right\|$ is bounded from above in probability by a constant. Since $\boldsymbol{\beta}^*$ is consistent, for n large enough $\boldsymbol{\beta}^* \in \mathscr{B}_0$. For such n,

$$\frac{1}{n}\sum_{i=1}^{n} \left\| h'(Y_i, \boldsymbol{X}_i, \boldsymbol{\beta}^*) \right\| \le \frac{1}{n}\sum_{i=1}^{n} C(Y_i, \boldsymbol{X}_i) \xrightarrow{\mathsf{P}} \mathsf{E}\, C(Y_i, \boldsymbol{X}_i) < \infty.$$

This completes the proof. □

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. 23) 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, $\widetilde{\ell}_n(\boldsymbol{Y}) \geq \ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y})$.

The unscaled deviance can be expressed as

$$D(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}}_n) = 2 \sum_{i=1}^{n} \big[Y_i(\widetilde{\theta}_i - \widehat{\theta}_i) - b(\widetilde{\theta}_i) + b(\widehat{\theta}_i) \big]. \tag{2.12}$$

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 $\widehat{\varphi}_n$ defined by (2.11).

2.7. Asymptotic Results

Asymptotic results for the GLM follow from the general theory of maximum likelihood estimation. The theory is reviewed in the Appendix starting on p. 99.

The following theorem transcribes the results of Theorems A.2 – A.5 from the Appendix 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 Pearson X^2 is another generalization.

The Fisher information matrix

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

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)

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

(iii)

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

(iv)

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

♢

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

$$\widehat{I}_n^{-1} / n = \widehat{\varphi}_n (\mathbb{X}^{\mathsf{T}} \hat{\mathbb{W}} \mathbb{X})^{-1}. \tag{2.13}$$

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

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 A.5 and Theorem A.7 from the Appendix.

Theorem 2.8.

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

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

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

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

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

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

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 components 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 $\boldsymbol{\beta}_M = (\beta_{p-m+1},\dots,\beta_p)^\mathsf{T}$ and $\boldsymbol{X}_i^M = (X_{i,p-m+1},\dots,X_{ip})^\mathsf{T}$. Let $\widehat{\boldsymbol{\beta}}_M = (\widehat{\beta}_{p-m+1},\dots,\widehat{\beta}_p)^\mathsf{T}$ 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 $\widehat{\Sigma} = \widehat{I}_n^{-1}/(n\widehat{\varphi}_n) = (\mathbb{X}^\mathsf{T}\widehat{\mathbb{W}}\mathbb{X})^{-1}$ (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* $\widetilde{\Sigma} = (\mathbb{X}^T \widetilde{\mathbb{W}} \mathbb{X})^{-1}$. *Denote by* $\widetilde{\varphi}_n$ *the estimator of the dispersion parameter calculated under the submodel (under* H_0^**). If* H_0^* *holds then*

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

(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*

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

(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*

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

Note.

- Theorem 2.9 follows from Definition A.6 and Theorem A.9 in the Appendix. 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

$$\left(\frac{\widehat{\beta}_j}{\sqrt{\widehat{\varphi}_n \widehat{\sigma}^2_{jj}}}\right)^2, \tag{2.14}$$

  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 A.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 (2.14). 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 $\boldsymbol{c}^\top \boldsymbol{\beta}_0$ where $\boldsymbol{0} \neq \boldsymbol{c} \in \mathbb{R}^p$ can be obtained easily from Theorem 2.7 part (iii). An asymptotic confidence interval with coverage probability converging to $1-\alpha$ is

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

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 $\boldsymbol{Y}^* = \hat{\mathbb{W}}^{1/2}\widehat{\boldsymbol{Z}}$. Recall that $\hat{\mathbb{W}} = \operatorname{diag}(w(\widehat{\mu}_1), \ldots, w(\widehat{\mu}_n))$, $w(\mu_i) = \frac{1}{V(\mu_i)[g'(\mu_i)]^2}$, and

$$\widehat{Z}_i = \boldsymbol{X}_i^\top \widehat{\boldsymbol{\beta}}_n + (Y_i - \widehat{\mu}_i) g'(\widehat{\mu}_i).$$

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

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

be the projection matrix and write the fitted values as

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

2.8.1. Pearson residuals

Pearson residuals are defined by

$$\boldsymbol{r}^P = \boldsymbol{Y}^* - \hat{\boldsymbol{Y}}^* = \hat{\mathbb{W}}^{1/2}\widehat{\boldsymbol{Z}} - \hat{\mathbb{W}}^{1/2}\mathbb{X}\widehat{\boldsymbol{\beta}} = \hat{\mathbb{W}}^{1/2}(\widehat{\boldsymbol{Z}} - \mathbb{X}\widehat{\boldsymbol{\beta}}).$$

The Pearson residual for an individual observation is

$$r_i^P = \sqrt{w(\widehat{\mu}_i)}(\widehat{Z}_i - \boldsymbol{X}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}) = \frac{(Y_i - \widehat{\mu}_i)g'(\widehat{\mu}_i)}{\sqrt{[g'(\widehat{\mu}_i)]^2 V(\widehat{\mu}_i)}} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}.$$

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

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

We have $\boldsymbol{r}^P = (\mathbb{I}_n - \mathbb{H}^*)\boldsymbol{Y}^*$. We can approximate $\mathsf{var}\,\boldsymbol{Y}^* = \mathsf{var}\,\hat{\mathbb{W}}^{1/2}\widehat{\boldsymbol{Z}}$ by $\mathbb{W}^{1/2}\,\mathsf{var}\,\boldsymbol{Z}\,\mathbb{W}^{1/2}$, where $\boldsymbol{Z}$ has elements

$$Z_i = \boldsymbol{X}_i^{\mathsf{T}}\boldsymbol{\beta}_0 + (Y_i - \mu_i)g'(\mu_i)$$

and $\mathbb{W} = \mathrm{diag}\,(w(\mu_1), \ldots, w(\mu_n))$. Here, $\boldsymbol{\beta}_0$ and μ_i are the true values to which the estimates converge in probability and the variance is conditional given the covariates. Further, $\mathsf{var}\,Z_i = \mathsf{var}\,[Y_i g'(\mu_i)] = \varphi_0 V(\mu_i)[g'(\mu_i)]^2$ and

$$\mathsf{var}\,\boldsymbol{Y}^* \approx \mathbb{W}^{1/2}\mathrm{diag}\,\{\varphi_0 V(\mu_i)[g'(\mu_i)]^2\}\mathbb{W}^{1/2} = \varphi_0\mathrm{diag}\left\{\frac{V(\mu_i)[g'(\mu_i)]^2}{V(\mu_i)[g'(\mu_i)]^2}\right\} = \varphi_0\mathbb{I}_n.$$

Because $\mathbb{I}_n - \mathbb{H}^*$ is idempotent,

$$\mathsf{var}\,\boldsymbol{r}^P \approx \varphi_0(\mathbb{I}_n - \mathbb{H}^*).$$

2.8.2. Leverages

It has been shown above 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

The standardized Pearson residual normalizes r_i^P by division by the square root of its approximate variance:

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

These residuals 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

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

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. Regression Models for Binary Data

3.1. Alternative vs. binomial data

Let $Y_{ij}^* \sim \mathsf{Alt}(\pi_i)$, $\pi \in (0,1)$, be independent variables for $i = 1,\dots,K$, $j = 1,\dots,m_i$. For a fixed i, $Y_{i1}^*,\dots,Y_{im_i}^*$ are identically distributed. The total number of observations is $N = \sum_{i=1}^K m_i$. Let π_i depend on $\boldsymbol{X}_i = (X_{i1},\dots,X_{ip})^\mathsf{T}$ 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 $Y_{i1}^*,\dots,Y_{im_i}^*$ 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 $Y_i \sim \mathsf{Bi}(m_i,\pi_i)$. Because a binomial response can be always written as a sum of independent responses with an alternative distribution, the GLM developed for the alternative distribution can be also used to fit regression models to binomial responses even though the binomial distribution does not strictly belong to the exponential family as we defined it.

The dataset with alternative or binomial responses can be arranged in two different formats (see Figure 3.1) that can be transformed one to the other.

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

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

Note. It is a bad idea to mix the two data formats in a single dataset.

Note. If the covariate vector has at least one continuous component (with no rounding) then $m_i = 1$ for all i, $N = K$ and the two data formats are the same.

The presence or absence of at least one continuous covariate leads to one of two different kinds of asymptotics when $N \to \infty$.

Figure 3.1.: Binary data written in the alternative format A (left panel) vs. the binomial format B (right panel).

Format A.

$$\left.\begin{matrix} Y_{11}^* & \boldsymbol{X}_1^\top \\ \vdots & \vdots \\ Y_{1m_1}^* & \boldsymbol{X}_1^\top \end{matrix}\right\} m_1 \times$$

$$\vdots$$

$$\left.\begin{matrix} Y_{K1}^* & \boldsymbol{X}_K^\top \\ \vdots & \vdots \\ Y_{Km_K}^* & \boldsymbol{X}_K^\top \end{matrix}\right\} m_K \times$$

Format B.

$$\left.\begin{matrix} Y_1 & m_1 & \boldsymbol{X}_1^\top \\ Y_2 & m_2 & \boldsymbol{X}_2^\top \\ \vdots & \vdots & \vdots \\ Y_K & m_K & \boldsymbol{X}_K^\top \end{matrix}\right\} K \times$$

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

Most of the results are the same for both data formats and both kinds of asymptotics but there are certain important differences that will be pointed out later.

3.2. 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:

3.2.1. Logistic link

Take the quantile function of the standard logistic distribution.

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

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*[*].

[*] Česky *logistická regrese*

3.2.2. Probit link

Take the quantile function of the standard normal distribution.

$$g(\mu_i) = \Phi^{-1}(\mu_i), \qquad \mu_i = \Phi(X_i^\top \boldsymbol{\beta}).$$

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

3.2.3. Cauchit link

Take the quantile function of the standard Cauchy distribution.

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

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 $\eta_i \to \pm\infty$.

3.2.4. Complementary log-log link

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

$$g(\mu_i) = \log(-\log(1-\mu_i)), \qquad \mu_i = 1 - \mathrm{e}^{-\exp\{X_i^\top \boldsymbol{\beta}\}}.$$

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

$$g(\mu_i) = -\log(-\log(\mu_i)), \qquad \mu_i = \mathrm{e}^{-\exp\{-X_i^\top \boldsymbol{\beta}\}}.$$

The inverse link functions are plotted in Figure 3.2. 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.

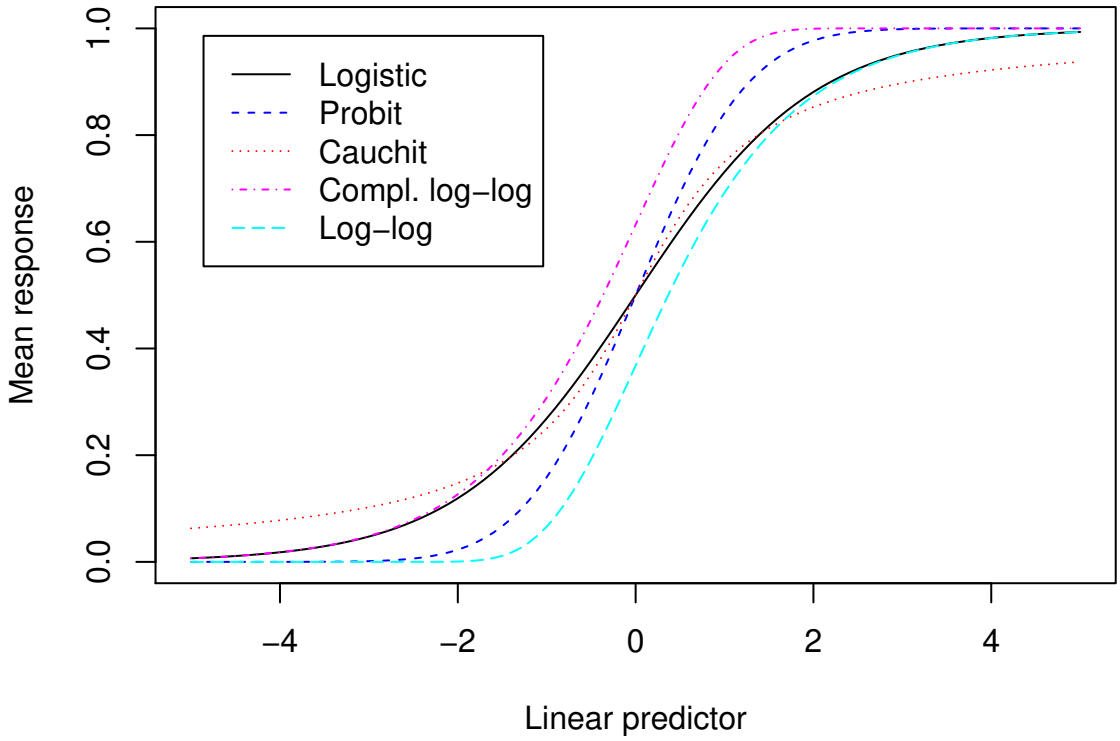
3.3. Binary data likelihood

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.

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

Figure 3.2.: 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.



(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 the

framework of Chapter 2 strictly but the core of the likelihood is the same and all the results have exactly the same form and properties. Therefore we do not have to distinguish the two sampling schemes.

3.4. 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

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

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 $Y_i = \mathbb{1}(U_i < C_i)$, 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

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

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

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

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 $\boldsymbol{X}_i^\top \boldsymbol{\beta}$. 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.

3.5. 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 $\pi_i = \frac{\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\top \boldsymbol{\beta}\}}$.

Interpretation of regression parameters

Let $\boldsymbol{X}_i^\top \boldsymbol{\beta} = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. 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 $\boldsymbol{x}^0 = (1, x_2, \ldots, x_p)^\top$, the other with observed covariates increased at the k-th component by one: $\boldsymbol{x}^k = \boldsymbol{x}^0 + \boldsymbol{e}_k$. Denote $\pi_{X0} = \mathrm{P}\big(Y_{ij}^* = 1 \,\big|\, \boldsymbol{X} = \boldsymbol{x}^0\big)$ and $\pi_{Xk} = \mathrm{P}\big(Y_{ij}^* = 1 \,\big|\, \boldsymbol{X} = \boldsymbol{x}^k\big)$. Then

$$\boldsymbol{\beta}^\top \boldsymbol{x}^0 = \log \frac{\pi_{X0}}{1-\pi_{X0}} \qquad \text{and} \qquad \boldsymbol{\beta}^\top \boldsymbol{x}^k = \boldsymbol{\beta}^\top \boldsymbol{x}^0 + \beta_k = \log \frac{\pi_{Xk}}{1-\pi_{Xk}}.$$

It follows that

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

Thus e^{β_k} is the odds ratio for success comparing two individuals differing by one unit in the covariate X_k. E.g., if $\beta_k = 0.431$ one can say that a unit increase in the covariate X_k increases the odds of success $\mathrm{e}^{0.431} = 1.539$ times (or by 53.9%). When $\beta_k = 0$ 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	$Y = 1$	$Y = 0$
$\boldsymbol{X} = \boldsymbol{x}^k$	π_{Xk}	$1-\pi_{Xk}$
$\boldsymbol{X} = \boldsymbol{x}^0$	π_{X0}	$1-\pi_{X0}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\widehat{\boldsymbol{\beta}}_n$ (see (2.13) on p. 33) is

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

It can be easily obtained from the IWLS for either data format.

Deviance

The next thing we need to do is to evaluate the deviance of a logistic model

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

(see Def. 2.7 on p. 31, with dispersion $\varphi = 1$), where $\ell(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{Y})$ is the maximized log-likelihood of our model and $\widetilde{\ell}(\boldsymbol{Y})$ is the maximized log-likelihood of the saturated model.

Let us consider the binomial formulation first. We have K observations with K different values of the covariate vector $\boldsymbol{X}_i$, each observed m_i times, $i = 1, \ldots, K$. The saturated model has K parameters that generate K distinct fitted values $\widetilde{\pi}_i = \frac{Y_i}{m_i}$. The canonical parameter θ_i is $\log\frac{\pi_i}{1-\pi_i}$ and $b(\theta_i) = \log\frac{1}{1-\pi_i}$. Hence we can write the deviance as

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

where the first part of the summand for the i-th group summarizes the successes (the number of successes times the log of the ratio of the observed number of successes divided by the fitted number of successes) and the second part summarizes the failures (the number of failures times the log of the ratio of the observed number of failures divided by the fitted number of failures).

It is important to realize that the deviance can be calculated even if $Y_i = 0$ or $Y_i = m_i$ (then the associated log term that becomes zero is simply omitted).

With the Bernoulli data format, the saturated model has fitted values $\widetilde{\pi}_{ij} = Y^*_{ij}$ and the deviance becomes (consider the special case of the above with $m_i = 1$)

$$2\sum_{i=1}^{K}\sum_{j=1}^{m_i}\left\{Y^*_{ij}\log\frac{Y^*_{ij}}{\widehat{\pi}_i} + (1-Y^*_{ij})\log\frac{1-Y^*_{ij}}{1-\widehat{\pi}_i}\right\},$$

which is *different* from the deviance calculated from the binomial data format unless $m_i = 1$ for all i.

Which deviance is the right one? The difference between them stems from the selection of the saturated model. If the data have K distinct values of the covariate vector, the most general model that can be fitted has K distinct fitted values and hence at most K parameters. So, the saturated model that was used to develop the deviance for the Bernoulli data format does not in fact exist. It also follows that the deviance tests that subtract deviances of larger models from deviances of submodels (Theorem 2.9(iii)) are not affected by the form of the deviance (the log-likelihood of the saturated model is canceled) as long as the same saturated model is used in both.

Statistical software will calculate the deviance blindly according to the format the data are entered in (binomial deviance for binomial data format, alternative deviance for Bernoulli data format). Thus, the deviance the software reports for the Bernoulli data format will be wrong but deviance tests of submodels will still be correct.

When all covariates in the dataset are discrete with finite support and data are entered in the binomial data format, K stays constant. Then the saturated model with K parameters satisfies the assumptions of maximum likelihood theory and the binomial deviance $D(\boldsymbol{Y}, \widehat{\boldsymbol{\beta}})$ converges in distribution to χ^2_{K-p} as $m_i \to \infty$ for all i if the current model is valid. Thus, when all m_i are sufficiently large, the deviance can be used as a goodness of fit statistic for deciding whether the current model describes the data sufficiently well (deviance larger than the quantile $\chi^2_{K-p}(1-\alpha)$ indicates that the model does not fit well). However, we must remember that such tests can be done only when all their assumptions are fulfilled:

1. All covariates are discrete
2. There are enough observations in each group
3. The deviance was calculated from the binomial data format

Deviances calculated from the Bernoulli data format cannot ever be used for goodness-of-fit testing but work for submodel testing.

Pearson X^2

The situation with Pearson X^2 statistic is similar. Consider first the Pearson residuals calculated as $\hat{\mathbb{W}}^{1/2}(\widehat{\boldsymbol{Z}} - \mathbb{X}\widehat{\boldsymbol{\beta}})$ (see Sec. 2.8.1).

Bernoulli data format

With *Bernoulli data format*, the Pearson residuals are

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

When $Y_{ij}^* = 1$, the Pearson residual is $\sqrt{\frac{1-\widehat{\pi}_i}{\widehat{\pi}_i}}$. When $Y_{ij}^* = 0$, the Pearson residual is $-\sqrt{\frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}}$. [Think what happens when you plot these residuals against linear predictor or against one of the covariates.]

The Pearson X^2 statistic is obtained by summing squares of Pearson residuals. We get

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

For a well-fitting model, $Y_i \approx m_i\widehat{\pi}_i$ and $m_i - Y_i \approx m_i(1-\widehat{\pi}_i)$. Hence,

$$X^2 \approx \sum_{i=1}^{K}\left[m_i(1-\widehat{\pi}_i) + m_i\widehat{\pi}_i\right] = \sum_{i=1}^{K} m_i = N.$$

The Pearson X^2 statistic calculated from the Bernoulli format is about equal to the sample size, if the model fits well. This is not a desirable behavior of a goodness-of-fit statistic.

Binomial data format

Pearson residuals for the *binomial data format* are calculated from the reduced data as $\hat{\mathbb{W}}_R^{1/2}(\widehat{\boldsymbol{Z}}_R - \mathbb{X}_R\widehat{\boldsymbol{\beta}})$. There is one residual for each group $i = 1,\ldots,K$ and

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

which is the binomial variable Y_i standardized by subtracting the estimated mean and dividing by its estimated standard deviation. For large m_i, these residuals are approximately standard normal.

The Pearson X^2 statistic for the binomial data format is

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

where we used the equality $\frac{1}{\widehat{\pi}_i(1-\widehat{\pi}_i)} = \frac{1}{\widehat{\pi}_i} + \frac{1}{1-\widehat{\pi}_i}$. This is the chi-square statistic for testing goodness of fit in a $2 \times K$ contingency table with p estimated parameters. In a saturated model with $p = K$, the Pearson X^2 statistic is zero. If the fitted model holds and $m_i \to \infty$ for all i, then $X^2 \xrightarrow{\mathsf{D}} \chi^2_{K-p}$. Thus, when all covariates are discrete and the number of successes and failures is large enough in each group, we can test the validity of the fitted model by

Pearson X^2 statistic calculated from the binomial format (but not by the Pearson X^2 statistic calculated from the Bernoulli format of the same dataset).

Statistical software computes Pearson residuals and Pearson X^2 statistics (and the deviance) from the format in which the data are entered. So, it is the responsibility of the analyst to reshape the data *into the binomial format* if these statistics are important for the analysis. Of course, if there is at least one continuous covariate in the model, the two data formats do not differ from each other (or only negligibly).

Hints on logistic regression practice

For submodel testing, use deviance tests. Do not trust Wald tests of individual coefficients reported in model output, especially not for factor covariates. These tests depend on the parametrization of the factor and may give a misleading impression about the significance of the factor. The recommended way to test model terms in R by deviance tests is

```
drop1(...,test="Chisq")
```

These tests work for both Bernoulli and binomial data formats.

Appropriate transformations of continuous covariates can be deduced, e.g., (i) by factorization of the covariate into subintervals (`cut(x,c(-Inf,x1,x2,...,xm,Inf))`) and evaluating trends in the estimated parameters or (ii) by adding a few different transformations to the linear term and testing their significance. Residual plots can be used, too, but they must be smoothed properly.

A common problem in logistic regression is caused by fitted values converging to zero or one in some subgroup. E.g., if gender is included in the model and all men have response $Y_{ij} = 1$, the MLE of π_i is 1 for all men. This sets the diagonal terms of $\hat{\mathbb{W}}$ to zero and estimated coefficients and their standard errors blow up. So if some of the estimated coefficients are incredibly large in absolute value, and have incredibly large standard errors, or if the IWLS algorithm fails to converge, this is the likely reason. Data subgroups for which no success or no failure are observed must be removed from the data set.

4. Loglinear Models for Poisson Data

Responses with Poisson distribution are typically counts recording the number of occurrences of some event. When a large number of independent Bernoulli trials are performed, with a small success probability at each trial, the observed total number of successes will approximately follow a Poisson distribution. Poisson responses can also arise by observations of independent Poisson processes evaluated at a fixed time. The primary tool for the analysis of Poisson responses is the loglinear regression model.

4.1. Poisson loglinear model

Let $Y_1, \dots, Y_n$ 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

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

Interpretation of regression parameters

Let $\boldsymbol{X}^\mathsf{T} \boldsymbol{\beta} = \beta_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. Denote $\lambda_0 = \mathsf{E}\left[Y_i \,\middle|\, X_2 = \cdots = X_p = 0\right]$. Then $\log \lambda_0 = \beta_1$ so e^{β_1} is the expected value of Y_i for an individual with zero values in all covariates.

Now consider two individuals with observed covariates $\boldsymbol{x}^0 = (1, x_2, \dots, x_p)^\mathsf{T}$ and $\boldsymbol{x}^j = \boldsymbol{x}^0 + \boldsymbol{e}_j$ (the j-th covariate is increased by 1, the others are the same). Denote $\lambda_{X0} = \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X} = \boldsymbol{x}^0\right]$ and $\lambda_{Xj} = \mathsf{E}\left[Y_i \,\middle|\, \boldsymbol{X} = \boldsymbol{x}^j\right]$. 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 of expectations is 1 and the covariate has no effect on the expectation given the other covariates.

Estimation of parameters

The likelihood is

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

and the log-likelihood is

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

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

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

The information matrix is $I(\boldsymbol{\beta}) = \mathsf{E}_X \lambda_i \boldsymbol{X}_i^{\otimes 2}$, 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

$$D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}}) = 2\sum_{i=1}^{n}\Big[Y_i \log\frac{Y_i}{\widehat{\lambda}_i} - (Y_i - \widehat{\lambda}_i)\Big]. \tag{4.1}$$

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

$$X^2 = \sum_{i=1}^{n}\frac{(Y_i - \widehat{\lambda}_i)^2}{\widehat{\lambda}_i}. \tag{4.2}$$

Aggregated Poisson responses

What if multiple observations share the same covariate vector? Then they have the same mean and the same fitted value. We can apply an analogue to the binomial data format we considered in logistic regression context and come to very similar conclusions.

Change our notation as follows: let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ be K distinct values of the covariate vector recorded among the $n \geq K$ observations. Let m_i be the number of observations that share the same covariate vector, so that $\sum_{i=1}^{K} m_i = n$. Change the meaning of Y_i: now, let Y_i be **the sum** of the m_i independent Poisson responses sharing the covariate vector $\boldsymbol{X}_i$. Because the sum of independent Poisson variables is Poisson, we have $Y_i \sim \mathsf{Po}(m_i\lambda_i)$. This distribution is in the exponential family, so all results apply without a change (except writing the mean as $m_i\lambda_i$ at each occurrence).

With the aggregated data format, the responses are $Y_1, \ldots, Y_K$, the regression matrix is $K \times p$, and the weight matrix $\hat{\mathbb{W}}$ is $K \times K$. The data set is written more compactly, especially when all covariates are categorical and $K \ll n$. The IWLS uses

$$\hat{\mathbb{W}} = \mathrm{diag}\,(m_1\widehat{\lambda}_1, \ldots, m_K\widehat{\lambda}_K)$$

and

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

The MLE of λ_i in the saturated model are $\widetilde{\lambda}_i = \frac{Y_i}{m_i}$. Hence, the aggregated deviance is

$$D(\boldsymbol{Y} \mid \widehat{\boldsymbol{\beta}}) = 2\sum_{i=1}^{K}\Big[Y_i \log \frac{Y_i}{m_i\widehat{\lambda}_i} - (Y_i/m_i - \widehat{\lambda}_i)\Big]. \tag{4.3}$$

This is different from the deviance defined by (4.1). When all covariates are categorical, K is finite and $m_i \to \infty$ for all $i = 1,\ldots,K$, then our aggregated deviance (4.3) converges in distribution to χ^2_{K-p} as long as the current model is valid. This convergence does not hold for the deviance defined by (4.1). On the other hand, differences in deviances between a submodel and a larger model are always the same no matter if the deviances are calculated from aggregated data or not.

The Pearson residual for aggregated Poisson data is

$$r_i^P = \frac{Y_i - m_i\widehat{\lambda}_i}{\sqrt{m_i\widehat{\lambda}_i}}, \quad i = 1,\ldots,K,$$

and the Pearson X^2 statistic is

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

This is the χ^2 goodness-of-fit test statistic for multinomial distribution with K outcome levels and p estimated parameters. If all covariates are categorical, K is finite and $m_i \to \infty$ for all $i = 1,\ldots,K$, then the aggregated Pearson X^2 statistic converges in distribution to χ^2_{K-p} as long as the current model is valid.

Pearson X^2 ***statistic*** *can be used for goodness-of-fit testing of models with categorical covariates (in any model, not only logistic or loglinear). The deviance can be used for this purpose as well, and has the same asymptotic distribution. However, only deviance can be used to compare a model with a submodel. Differences in Pearson* X^2 *statistics do not have asymptotic* χ^2 *distribution. This is why we prefer deviance tests, while Pearson* X^2 *tests are considered secondary and of limited importance.*

4.2. Modelling Poisson process intensity

So far we considered independent responses Y_i with distribution $\mathsf{Po}(\lambda_i)$ and expressed the effect of covariates on λ_i. However, this assumes that the responses are counts observed over some standardized time interval, e.g., month, week, year — otherwise the means would be affected by the different duration.

Imagine that we want to compare the expected number of traffic accidents caused by men vs. women across different ages. Each observation is one driver and the response is the number of accidents caused by that driver. The covariates are gender and current age. But – for how long are those accidents recorded? Over the whole lifetime? That would induce an artificial age effect, older drivers having more accidents than younger ones. So we would use the number of accidents over the past five years. That would be OK if we remove drivers who have had their driver's licences for less than five years. But they are a particularly interesting subgroup, suspected of causing more accidents than the more experienced drivers. We do not want to exclude them. Also, it is quite possible that men are driving longer trips than women, and younger drivers longer and more frequent trips than elderly drivers: we should take that into account, too. So the best approach would be to standardize the number of accidents for the number of km driven during the last five years and compare the mean number of accidents per 100,000 km driven between men and women, and between different ages. Suppose that we observe the number of km driven over the past 5 years for each participant. How do we perform the analysis?

In general, observations of Poisson counts can be understood as realizations of Poisson processes recorded at some time. A homogeneous Poisson process with intensity λ is a random process $N(t)$, $t \geq 0$, with the following properties:

- $N(0) = 0$;
- N has independent increments, that is for any $t_1 < t_2 < \cdots < t_j$, the random variables $N(t_2) - N(t_1)$, $N(t_3) - N(t_2)$, …, $N(t_j) - N(t_{j-1})$ are independent;
- $N(t_2) - N(t_1) \sim \mathsf{Po}(\lambda(t_2 - t_1))$ for any $t_1 < t_2$.

It is a non-decreasing piecewise constant process with values in $\{0, 1, 2, \ldots\}$. Marginally, $N(t) \sim \mathsf{Po}(\lambda t)$. It can be shown that the times between successive jumps of $N(t)$ are independent variables with distribution $\mathsf{Exp}(\lambda)$.

Now suppose that we observe iid vectors $(Y_i, t_i, \boldsymbol{X}_i)$, where t_i is the observation time and Y_i is a realization of a Poisson process with intensity λ_i observed at the time t_i. We are interested in estimating the effects of $\boldsymbol{X}_i$ on the intensity λ_i taking into account the observation time t_i.

Note: *The times t_i can be measured on some real time scale (days, months, years,…) but they can also be non-decreasing transformations of time, such as the number of km driven (see the example with traffic accidents), amount of money spent, etc. The times need not even be continuous, discrete time scales are OK, too.*

We have $Y_i \sim \mathsf{Po}(\lambda_i t_i)$, hence $\mathsf{E}\, Y_i = \mathsf{var}\, Y_i = \lambda_i t_i$. The intensity $\lambda_i = \mathsf{E}\, Y_i / t_i$ describes the expected number of events observed during a unit time interval. We use the log link for λ_i: $\log \lambda_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$, hence $\lambda_i = \mathrm{e}^{\boldsymbol{X}_i^\top \boldsymbol{\beta}}$. It follows that

$$\mathsf{E}\, Y_i = \lambda_i t_i = t_i \mathrm{e}^{\boldsymbol{X}_i^\top \boldsymbol{\beta}} = \mathrm{e}^{\log t_i + \boldsymbol{X}_i^\top \boldsymbol{\beta}}.$$

So, the observation time can be simply taken into account by adding $\log t_i$ to the linear predictor. We can consider it another covariate, with a regression parameter that is *a priori* known to be 1 and hence need not be estimated. Such a term is called *an offset* in GLM terminology. Adding an offset to the linear predictor preserves the structure of the GLM, all formulae and results remain valid without change. In R function `glm()`, an offset term is specified by adding `+offset(var)` to the model formula.

The extension of the Poisson loglinear model to homogeneous Poisson processes is important because it provides a very simple solution to a frequently encountered practical problem. It is very useful to remember this.

4.3. Two-Way Contingency Tables

4.3.1. Notation

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 $n_{ij} = \sum_{l=1}^{n} \mathbb{1}(X_l = i, Z_l = j)$. 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

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

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

The cell probabilities can be also arranged into a table:

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

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

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

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

The conditional probabilities can be expressed as follows:

$$\mathrm{P}\big(X = i \,\big|\, Z = j\big) = \frac{\pi_{ij}}{\pi_{+j}} = \frac{m_{ij}}{m_{+j}} \quad \text{and} \quad \mathrm{P}\big(Z = j \,\big|\, X = i\big) = \frac{\pi_{ij}}{\pi_{i+}} = \frac{m_{ij}}{m_{i+}}.$$

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.

4.3.2. Distributions of observed counts

In order to analyze a contingency table by maximum likelihood methods, we have to specify the joint distribution of the observed data, i.e., the counts in the contingency table. There are three reasonable models that arise by different ways of collecting data summarized in the table:

Poisson distribution

Poisson distribution of observed counts arises when the observations (X_i, Z_i) from which the table is built arrive randomly over a given period of time.

For example, we investigate associations between tooth decay ($X = 1$ yes, $X = 2$ no) and soft drink consumption ($Z = 1$ never or rarely, $Z = 2$ sometimes, $Z = 3$ frequently). We include young children (age 6–12) who come to a dentist's office for prevention check-up between January and June of a certain year (if their parents are willing to provide the required information). With this data collection method, the total sample size n is random and the observed cell counts in the resulting table can be assumed to be realizations of independent Poisson processes with different intensity. Then, the whole table can be modelled as independent Poisson variables.

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

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

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

Multinomial distribution of observed counts arises when the total number of observations n is fixed in advance.

If, in the previous example, we plan to enroll $n = 100$ children and collect data until the planned sample size is reached, we obviously end up with multinomial distribution for the observed contingency table.

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

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

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 $\mathsf{Mult}_{IJ}(1, \pi)$. The asymptotics works through letting $n \to \infty$. The asymptotic MLE theory for iid data applies to this case.

Row multinomial distribution

Row multinomial distribution of observed counts arises when all the row totals n_{i+} are fixed in advance.

We obtain row multinomial distribution if, in the previous example, we plan to enroll $n_{1+} = 50$ children with tooth decay and $n_{2+} = 50$ children without tooth decay and collect data until the planned sample size is reached in both subgroups.

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

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

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.

Of course, we could consider column multinomial distribution as well, but it is just a transposition of the row multinomial case.

4.3.3. 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 4.1. *Let $X_i\sim Po(\lambda_i)$ be independent random variables, $i=1,\dots,n$. Then the conditional joint distribution of the random vector $(X_1,\dots,X_n)^T$ given $\sum_{i=1}^n X_i=s$ is $Mult_n(s,\boldsymbol{p})$, where $\boldsymbol{p}=(p_1,\dots,p_n)^T$ and $p_i=\lambda_i/\sum_{j=1}^n\lambda_j$.* ♢

Proof. Calculate the conditional probability

$$\mathrm{P}\big(X_1=k_1,\dots,X_n=k_n\,\big|\sum_{i=1}^n X_i=s\big)=\frac{\mathrm{P}\big[X_1=k_1,\dots,X_n=k_n,\sum_{i=1}^n X_i=s\big]}{\mathrm{P}\big[\sum_{i=1}^n X_i=s\big]}.\qquad(*)$$

The probability in the numerator is zero unless $\sum_{i=1}^n k_i=s$; in that case the event $\sum_{i=1}^n X_i=s$ can be dropped and we have an intersection of independent events. The probability in the denominator is determined from the known result about summing independent Poisson variables $\sum_{i=1}^n X_i\sim\mathsf{Po}(\sum_{i=1}^n\lambda_i)$. Hence

$$(*)=\frac{\prod_{i=1}^n\lambda_i^{k_i}\mathrm{e}^{-\lambda_i}\frac{1}{k_i!}}{\big(\sum_i\lambda_i\big)^s\mathrm{e}^{-\sum_i\lambda_i}\frac{1}{s!}}=\frac{s!}{k_1!\cdots k_n!}\,\frac{\lambda_1^{k_1}\cdots\lambda_n^{k_n}}{\big(\sum_i\lambda_i\big)^{k_1}\cdots\big(\sum_i\lambda_i\big)^{k_n}}=\frac{s!}{k_1!\cdots k_n!}\,\pi_1^{k_1}\cdots\pi_n^{k_n}$$

with $\pi_i=\frac{\lambda_i}{\sum_i\lambda_i}$. The proof is completed. □

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

- The conditional joint distribution of $(n_{11},\dots,n_{IJ})^\mathsf{T}$ given $n_{++}=n$ is $\mathsf{Mult}_{IJ}(n,\pi)$, where the components of π are $\pi_{ij}=m_{ij}/m_{++}$ ($\Rightarrow$ *multinomial distribution*).

- The conditional joint distribution of $(n_{i1}, \ldots, n_{iJ})^\top$ given n_{i+} is $\mathsf{Mult}_J(n_{i+}, \pi_i)$, where the components of π_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,

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

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:

$$\pi_{ij} = \frac{m_{ij}}{m_{++}} = \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}}{\sum_{k,l} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{kl}}}. \tag{4.4}$$

Theorem 4.2. *The likelihood functions for estimation of parameters $\boldsymbol{\beta}$ in the loglinear model $\log m_{ij} = \alpha + \boldsymbol{\beta}^T \boldsymbol{X}_{ij}$ 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 4.2 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 (4.4).

Proof. First, write the Poisson likelihood for the loglinear model $\log m_{ij} = \alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}$ with data $\boldsymbol{n} = (n_{11}, \ldots, n_{IJ})^\top$.

$$L_P(\alpha, \boldsymbol{\beta} \mid \boldsymbol{n}) = \prod_{i,j} \frac{1}{n_{ij}!} m_{ij}^{n_{ij}} \mathrm{e}^{-mij}$$

and express the log-likelihood as

$$\ell_p(\alpha, \boldsymbol{\beta}) = \sum_{i,j} n_{ij} \log m_{ij} - m_{++} - \underbrace{\sum_{i,j} \log n_{ij}!}_{\equiv c_P},$$

where the constant c_P can be ignored.

Next, write the multinomial likelihood. When we express cell probabilities in terms of the expected frequencies using the loglinear model, we get

$$\pi_{ij} = \frac{m_{ij}}{m_{++}} = \frac{\mathrm{e}^{\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}}{\sum_{i,j} \mathrm{e}^{\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}} = \frac{\mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}}{\sum_{i,j} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}}. \tag{4.5}$$

The parameter α dropped out, the multinomial likelihood is only a function of $\boldsymbol{\beta}$. Hence

$$L_M(\boldsymbol{\beta} \mid \boldsymbol{n}) = n! \prod_{i,j} \frac{1}{n_{ij}!} \left(\frac{m_{ij}}{m_{++}} \right)^{n_{ij}}$$

and the log-likelihood is

$$\ell_M(\boldsymbol{\beta}) = \sum_{i,j} n_{ij} \log m_{ij} - n \log m_{++} + \underbrace{\log \frac{n!}{n_{11}! \cdots n_{IJ}!}}_{\equiv c_M}.$$

Now go back to the Poisson log-likelihood ℓ_p and reparametrize it as a function of parameters $(m_{++}, \boldsymbol{\beta})$ instead of $(\alpha, \boldsymbol{\beta})$ – this is a one-to-one transformation of parameters. The Poisson log-likelihood can be written as

$$\ell_p(m_{++}, \boldsymbol{\beta}) = \underbrace{\sum_{i,j} n_{ij} \log m_{ij} - n \log m_{++}}_{= \ell_M(\boldsymbol{\beta}) - c_M} + \underbrace{n \log m_{++} - m_{++}}_{\text{denote } \ell^*(m_{++})} - c_p.$$

The first part is the multinomial log-likelihood (the conditional log-likelihood given the total size n) without its irrelevant constant, the second part is the marginal log-likelihood for the sample size $n \sim \mathsf{Po}(m_{++})$. The first part depends only on $\boldsymbol{\beta}$ but not on α or m_{++}, the second part depends on m_{++} but not on $\boldsymbol{\beta}$.

Thus, in both Poisson and multinomial models, the MLE of the parameters $\boldsymbol{\beta}$ are obtained by maximizing the same function

$$\begin{aligned}
\sum_{i,j} n_{ij} \log m_{ij} - n \log m_{++} &= \sum_{i,j} n_{ij} (\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}) - n \log \sum_{i,j} \mathrm{e}^{\alpha + \boldsymbol{\beta}^\top \boldsymbol{X}_{ij}} \\
&= \sum_{i,j} n_{ij} \boldsymbol{\beta}^\top \boldsymbol{X}_{ij} + n\alpha - n \log \mathrm{e}^{\alpha} - n \log \sum_{i,j} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}} \\
&= \sum_{i,j} n_{ij} \boldsymbol{\beta}^\top \boldsymbol{X}_{ij} - n \log \sum_{i,j} \mathrm{e}^{\boldsymbol{\beta}^\top \boldsymbol{X}_{ij}}
\end{aligned}$$

over $\boldsymbol{\beta}$. The likelihoods of both models are equivalent as far as estimation of $\boldsymbol{\beta}$ is concerned. □

Theorem 4.2 can be extended to row multinomial distribution as follows:

Theorem 4.3. *(Palmgren 1981) The likelihood functions for estimation of parameters $\boldsymbol{\beta}$ in the loglinear model $\log m_{ij} = \alpha_i + \boldsymbol{\beta}^T \boldsymbol{X}_{ij}$ 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.

4.4. Loglinear models for two-way tables

In this section we will show examples of loglinear models for the analysis of two-way tables. We will pay attention to the interpretation of the models and their parameters and the form of the relevant test statistics.

4.4.1. The independence model (X, Z)

In this model, factor variables X and Z are entered into the model as main effects, without any interaction. The response is the observed cell count.

The expected cell counts are expressed as

$$\log m_{ij} = \alpha + \beta_i^X + \beta_j^Z. \tag{4.6}$$

The variable X has I levels, hence the main effects β_i^X include $I-1$ linearly independent parameters. The variable Z has J levels, its main effects β_j^Z include $J-1$ parameters. We will use the constraints $\beta_1^X = \beta_1^Z = 0$. With this parametrization, the top-left cell of the table serves as *the reference cell* to which all the other cells are compared. The vector of regression parameters to be estimated is

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

and its dimension is $p = 1 + I - 1 + J - 1 = I + J - 1$.

This model is obtained by setting the dummy covariate vector $\boldsymbol{X}_{ij}$ for the cell (i, j) as follows:

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

Recall that $\pi_{ij} = m_{ij}/m_{++}$. Because $\beta_1^X = \beta_1^Z = 0$, we can express the intercept α as

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

Next,

$$\log m_{ij} = \log m_{++} + \log \pi_{ij}.$$

It follows that the loglinear model (4.6) can be also stated in terms of cell probabilities:

$$\log \pi_{ij} = \log \pi_{11} + \beta_i^X + \beta_j^Z. \tag{4.7}$$

From (4.7) , we can deduce the meaning of the regression parameters β_i^X and β_j^Z. For any $j = 1, \ldots, J$, we have

$$\log \pi_{ij} - \log \pi_{1j} = \log \pi_{11} + \beta_i^X + \beta_j^Z - \log \pi_{11} - \beta_j^Z = \beta_i^X$$

and

$$e^{\beta_i^X} = \pi_{ij}/\pi_{1j}$$

the ratio of probabilities in the j column. Dividing the numerator and the denominator by π_{+j}, we get

$$e^{\beta_i^X} = \frac{\pi_{ij}}{\pi_{1j}} = \frac{P\big(X = i \,\big|\, Z = j\big)}{P\big(X = 1 \,\big|\, Z = j\big)} \tag{$*$}$$

for any $j = 1, \ldots, J$, that is, the conditional distribution of X given Z is the same in all columns of the table.

Similar interpretation holds for β_j^Z:

$$e^{\beta_j^Z} = \frac{\pi_{ij}}{\pi_{i1}} = \frac{P\big(Z = j \,\big|\, X = i\big)}{P\big(Z = 1 \,\big|\, X = i\big)}.$$

for any $i = 1, \ldots, I$. The conditional distribution of Z given X is the same in all rows of the table.

But we are not done yet. From $(*)$, it follows

$$\pi_{ij} = \pi_{1j} e^{\beta_i^X}$$

for any $j = 1, \ldots, J$, and summing these equalities over j, we get

$$\pi_{i+} = \pi_{1+} e^{\beta_i^X}.$$

Hence, in this model, $e^{\beta_i^X}$ also has the interpretation of ratio of marginal probabilities (and similarly $e^{\beta_j^Z}$):

$$e^{\beta_i^X} = \frac{\pi_{i+}}{\pi_{1+}} = \frac{P[X = i]}{P[X = 1]} \quad \text{and} \quad e^{\beta_j^Z} = \frac{\pi_{+j}}{\pi_{+1}} = \frac{P[Z = j]}{P[Z = 1]}. \tag{$**$}$$

If we ignore all other outcomes but 1 and i, we can say that $e^{\beta_i^X}$ is the odds of observing $X = i$ rather than observing $X = 1$.

Now take $(**)$ and plug it into (4.7) to get

$$\log \pi_{ij} = \log \pi_{11} + \log \frac{\pi_{i+}}{\pi_{1+}} + \log \frac{\pi_{+j}}{\pi_{+1}}.$$

From this,

$$\pi_{ij} = \frac{\pi_{11}}{\pi_{1+}\pi_{+1}} \pi_{i+}\pi_{+j}.$$

Sum these equations over all i and j to get

$$1 = \frac{\pi_{11}}{\pi_{1+}\pi_{+1}}.$$

We have shown that the cell probabilities satisfy the equation

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad \Leftrightarrow \quad \mathrm{P}[X = i, Z = j] = \mathrm{P}[X = i]\,\mathrm{P}[Z = j] \tag{†}$$

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

The MLE's of π_{i+} and π_{+j} are the empirical relative frequencies $\widehat{\pi}_{i+} = \frac{n_{i+}}{n}$ and $\widehat{\pi}_{+j} = \frac{n_{+j}}{n}$. The fitted values (expected cell counts under independence) can be expressed explicitly from (†):

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

4.4.2. The interaction model (XZ)

Now we add interaction between the two factor variables into the previous model. The model (XZ) for expected cell counts is defined by the equation

$$\log m_{ij} = \alpha + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ} \tag{4.8}$$

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 added interaction terms is $(I-1)(J-1)$, so the total number of parameters in this model is $p = I + J - 1 + (I-1)(J-1) = IJ$. This is the saturated model; 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

$$\log \pi_{ij} = \log \pi_{11} + \beta_i^X + \beta_j^Z + \beta_{ij}^{XZ}. \tag{4.9}$$

The interpretation of the main effects is as in $(*)$, but only at the first level of the other factor.

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

that is, $e^{\beta_i^X}$ is the odds of observing $X = i$ compared to observing $X = 1$ when $Z = 1$. For other levels of Z, the conditional distribution of X is different.

From (4.9),

$$\log \pi_{ij} - \log \pi_{i1} - \log \pi_{1j} + \log \pi_{11} = \beta_{ij}^{XZ}.$$

Hence

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

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.

	$Z=1$	$\cdots$	$Z=j$	$\cdots$	$Z=J$
$X=1$	π_{11}	$\cdots$	π_{1j}	$\cdots$	π_{1J}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X=i$	π_{i1}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$X=I$	π_{I1}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}

The interaction term β_{ij}^{XZ} is related to the odds ratio in the sub-table composed of the four red probabilities.

The cell probabilities can be expressed as

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

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.

4.4.3. 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 (it is a submodel of (XZ) but its deviance is zero). The fitted values in the independence model are

given by (‡): $\widehat{m}_{ij} = n_{i+}n_{+j}/n$. From (4.1), the deviance of (X, Z) is

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

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 (4.2), 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 $\chi^2_{(I-1)(J-1)}$.

We will prefer the deviance test to the classical χ^2 test of independence because it can be generalized to tables of higher dimension, while the χ^2 test cannot.

Miscellaneous comments

Usually, we do not have to fit the saturated model. We know that its deviance is zero and the fitted values are equal to the observed cell counts. However, we may want to fit it to get estimates of the interaction terms. If independence is rejected, the interaction terms will tell us which levels of X and Z most strongly contribute to the dependence.

In general we have only two candidate models for two-way tables: full independence and full dependence (which is saturated). Sometimes we may want to get a model, that allows dependence but is not saturated. We can get it when the levels of X or Z are ordered in some way (so called *ordinal data*). Then we can assign some increasing numerical scores to the levels of X and Z; for example $x_1 < \cdots < x_I$ and $z_1 < \ldots < z_J$ and include them in the interaction as numeric terms instead of as factors. If we do it for both variables, we get the model

$$\log m_{ij} = \alpha + \beta_i^X + \beta_j^Z + \gamma x_i z_j.$$

This model has a single interaction parameter γ, which describes the dependence between X and Z, and it is not saturated. If we set $x_1 = z_1 = 0$, we do not change the interpretation of the other parameters and the log odds ratio $\log\frac{\pi_{ij}\pi_{11}}{\pi_{i1}\pi_{1j}}$ gets expressed as $\gamma x_i z_j$. Thus, this is a model for which the dependence of X on Z gets stronger at higher levels of X and Z (for $\gamma > 0$). The test of $\gamma = 0$ is an independence test that is particularly sensitive against alternatives of this kind.

4.5. Loglinear models for three-way tables

4.5.1. Three-way contingency table

Consider three categorical random variables $X \in \{1, \ldots, I\}$, $Z \in \{1, \ldots, J\}$, and $V \in \{1, \ldots, K\}$. Sample n independent subjects and observe n independent realizations of the triplets $(X_1, Z_1, V_1), \ldots, (X_n, Z_n, V_n)$. Denote the observed count of the outcome $(X = i, Z = j, V = k)$ by

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

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

Let

$$n_{ij+} = \sum_{k=1}^{K} n_{ijk}, \quad n_{i++} = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{ijk}, \quad \text{etc.}$$

Denote

$$m_{ijk} = \mathsf{E}\, n_{ijk} \quad \text{and} \quad \pi_{ijk} = \mathrm{P}[X = i, Z = j, V = k] = \frac{m_{ijk}}{m_{+++}}.$$

The symbols m_{ij+}, m_{++k}, π_{i+k}, π_{+j+} etc. all have the obvious meaning (summation over the indices replaced by +). Obviously, $\pi_{+++} = \sum_{i,j,k} \pi_{ijk} = 1$ and $n_{+++} = n$.

The observed cell counts have a joint multinomial distribution in this framework. However, the equivalence of the likelihood for independent Poisson counts $n_{ijk} \sim \mathsf{Po}(m_{ijk})$ with the multinomial likelihood is still true. This result can be extended to multi-way tables, though we do not try to show this.

4.5.2. Marginal and conditional associations in a three-way table

Suppose we are particularly interested in investigating the associations between the variables X and Z. The third variable V is a sort of "nuisance covariate," not of main interest. Its existence, however, changes the way we look at the associations of the other two variables.

Consider the following idea: we want to investigate the association between X and Z and we have methods available for the analysis of two-way tables. So we will try to transform the problem into the context of two-way tables. We can do this in two different ways:

1. Ignore the existence of the variable V and analyze the associations between X and Z as if V did not exist. This approach leads us to *marginal associations* between X and Z.
2. Analyze the associations between X and Z at each level of V — perform separate analyses of K two-way tables. This approach investigates *conditional associations* between X and Z.

The problem is, as we will see, that the marginal and conditional associations need not agree.

Notation for marginal associations

Ignoring the existence of V, we arrive at a two-way contingency table for X and Z. It is created by collapsing the original three-way table n_{ijk} across the levels of V. The collapsed two-way table has observed cell counts n_{ij+}. The marginal associations between X and Z can be described and investigated by the methods for a two-way table introduced in Section 4.4.

Definition 4.1. Discrete variables X and Z are called *marginally independent* if and only if

$$\mathrm{P}[X=i, Z=j] = \mathrm{P}[X=i]\,\mathrm{P}[Z=j] \quad \text{for all } i, j,$$

that is,

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

Marginal associations between X and Z are described by marginal odds ratios, which correspond to exponentiated interaction terms in the model (4.9) applied to the collapsed two-way table.

Definition 4.2. The marginal odds ratio for the i-th level of X and the j-th level of Z are defined as

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

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

Notice that the marginal odds ratios exactly correspond to the alternative expressions for $\mathrm{e}^{\beta_{ij}^{XZ}}$ *shown at the bottom of p. 67, in particular,*

$$\theta_{ij}^{XZ} = \frac{\mathrm{P}[X=i, Z=j]\,\mathrm{P}[X=1, Z=1]}{\mathrm{P}[X=i, Z=1]\,\mathrm{P}[X=1, Z=j]} = \frac{\mathrm{P}\big(X=i \,\big|\, Z=j\big)\mathrm{P}\big(X=1 \,\big|\, Z=1\big)}{\mathrm{P}\big(X=1 \,\big|\, Z=j\big)\mathrm{P}\big(X=i \,\big|\, Z=1\big)}.$$

Notation for 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 π_{ijk}/π_{++k}. The conditional associations between X and Z can be described and investigated by the methods described in Section 4.4 applied to n_{ijk} for each fixed k.

Definition 4.3. Variables X and Z are called *conditionally independent given* V if and only if

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

The next theorem shows how conditional independence can be characterized in terms of cell probabilities.

Theorem 4.4. *Variables X and Z are conditionally independent given V if and only if*

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

Proof. Write

$$\mathrm{P}\big(X=i, Z=j \,\big|\, V=k\big) = \frac{\pi_{ijk}}{\pi_{++k}}, \quad \mathrm{P}\big(X=i \,\big|\, V=k\big) = \frac{\pi_{i+k}}{\pi_{++k}}, \quad \text{and} \quad \mathrm{P}\big(Z=j \,\big|\, V=k\big) = \frac{\pi_{+jk}}{\pi_{++k}}.$$

Plug this into Definition 4.3. □

The odds ratios that describe conditional associations between X and Z are called *conditional odds ratios*.

Definition 4.4. 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

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

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

Notice that

$$\begin{aligned}\theta^{XZ}_{ij(k)} &= \frac{\mathrm{P}\big(X=i, Z=j \,\big|\, V=k\big)\mathrm{P}\big(X=1, Z=1 \,\big|\, V=k\big)}{\mathrm{P}\big(X=i, Z=1 \,\big|\, V=k\big)\mathrm{P}\big(X=1, Z=j \,\big|\, V=k\big)} \\ &= \frac{\mathrm{P}\big(X=i \,\big|\, Z=j, V=k\big)\mathrm{P}\big(X=1 \,\big|\, Z=1, V=k\big)}{\mathrm{P}\big(X=1 \,\big|\, Z=j, V=k\big)\mathrm{P}\big(X=i \,\big|\, Z=1, V=k\big)}.\end{aligned}$$

4.5.3. Example of Simpson's Paradox

We explain the concept of marginal and conditional associations on a hypothetical example: a study of salaries among university graduates. The example is artificial but it illustrates several extremely important aspects not only about the analysis of three-way tables but about statistical reasoning in general. Consider three categorical variables:

Gender denoted by X, with levels 1 = female and 2 = male,
Salary denoted by Z, with levels 1 = low and 2 = high,
Field of education denoted by V, with levels 1 = Humanities and 2 = Natural Sciences & Technology.

These variables generate a three way contingency table with $I = J = K = 2$, containing a total of 8 cells. We want to study the association between gender X and salary Z. The table of cell probabilities π_{ijk} — the true joint distribution of (X, Z, V) — is given by Table 4.1.

Table 4.1.: Example about salaries of university graduates: True cell probabilities.

Field (V)	Gender (X)	Salary (Z)	
		low ($j = 1$)	*high* ($j = 2$)
Humanities ($k = 1$)	*Female* ($i = 1$)	0.18	0.12
	Male ($i = 2$)	0.12	0.08
Nat. Sci & Tech. ($k = 2$)	*Female* ($i = 1$)	0.02	0.08
	Male ($i = 2$)	0.08	0.32

First, let's look at the marginal association between gender and salary. We must collapse the 3-way table into a 2×2 table, ignoring field of education. We get Table 4.2.

Table 4.2.: Example about salaries of university graduates: Marginal probabilities.

Gender (X)	Salary (Z)	
	low ($j = 1$)	*high* ($j = 2$)
Female ($i = 1$)	0.20	0.20
Male ($i = 2$)	0.20	0.40

The marginal odds ratio for having high salary comparing men to women is

$$\theta_{22}^{XZ} = \frac{0.4 \cdot 0.2}{0.2 \cdot 0.2} = 2.$$

Hence, men have twice the odds for a high salary than women. *"Discrimination!! But... wait a minute."* Let's look at the conditional associations now. They can be determined from Table 4.1 by looking at gender-salary sub-table at each level of the field of study.

For *humanities* ($k = 1$), we get

$$\theta_{22(1)}^{XZ} = \frac{0.08 \cdot 0.18}{0.12 \cdot 0.12} = 1.$$

Hence, men who studied humanities have the same odds for a high salary as women who studied humanities.

For *natural sciences & technology* ($k = 2$), we get

$$\theta^{XZ}_{22(2)} = \frac{0.32 \cdot 0.02}{0.08 \cdot 0.08} = 1.$$

Hence, men who studied natural sciences & technology have the same odds for a high salary as women who studied natural sciences & technology.

So, there is no difference in salaries between men and women in any field of study, men and women are perfectly equal. However, overall, men have twice the odds for a high salary than women. How is this possible and what does it mean?

The explanation is that the field of study is strongly associated with both gender and salary. The marginal odds ratio between field of study and gender is 6 (see Table 4.3) meaning that men have 6-times the odds of studying natural sciences & technology than women. Similarly, the marginal odds ratio between field of study and salary is also 6 (see Table 4.4) meaning that graduates of natural sciences & technology have 6-times higher odds for high salary than graduates of humanities. So, men are much more likely to study the field that is much more likely to provide a high salary. Otherwise, there is no difference in the salaries between men and women.

Table 4.3.: Example about salaries of university graduates: Association of field of study with gender.

Field of study (V)	Gender (X)	
	Female ($i = 1$)	*Male* ($i = 2$)
Humanities ($k = 1$)	0.30	0.20
Nat. Sci & Tech. ($k = 2$)	0.10	0.40

$$\theta^{XV}_{22} = \frac{0.3 \cdot 0.4}{0.1 \cdot 0.2} = 6.$$

4.5.4. Simpson's paradox and confounding: statistics and causality

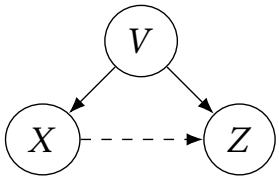
The apparent contradiction we have seen in the previous example is called *Simpson's paradox*. It is an example of the fact that marginal and conditional associations of two variables in the presence of a third variable need not be the same. They can even reverse themselves: marginal associations may show a positive relationship while conditional associations are negative (or vice versa). This feature is not limited to categorical variables or contingency tables. It can be observed between any three variables and is frequently encountered in all kinds of regression problems.

Table 4.4.: Example about salaries of university graduates: Association of field of study with salary.

Field of study (V)	Salary (Z)	
	Low ($j = 1$)	*High* ($j = 2$)
Humanities ($k = 1$)	0.30	0.20
Nat. Sci & Tech. ($k = 2$)	0.10	0.40

$$\theta_{22}^{ZV} = \frac{0.3 \cdot 0.4}{0.1 \cdot 0.2} = 6.$$

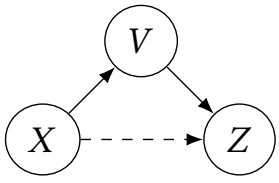
Figure 4.1.: A causal diagram: variable V confounds the effect of X on Z.



In epidemiology, this phenomenon is called "confounding". Confounding occurs when there exists a third variable that distorts the association between the two variables that we want to investigate.

A situation where the variable V acts as a confounder for the effect of X on Z is demonstrated in Figure 4.1. This a causal diagram: the variables are displayed as the nodes of an oriented graph, the edges are associations between the variables and arrows show causality. We want to know whether X is a cause of Z (in other words, whether X has any effect on Z). This is the dashed arrow in the graph. The third variable V acts as a confounder if it is a direct cause of both X and Z. In this situation, the marginal association between X and Z will give a false result; it will indicate an effect even if X does not actually have any influence on Z at all. The correct conclusion could be determined from the conditional association of X and Z given V – that is, by inclusion of V in the regression model as an additional covariate.

Figure 4.2.: A causal diagram: variable V acts as a mediator of the effect of X on Z.



An intuitive illustration of confounding is provided by the following example: let X be "carrying matches or lighters in the pocket", Z is lung cancer, and V is smoking. So, we want to know whether carrying matches or lighters in the pocket causes lung cancer. In a marginal analysis of these two variables, we are sure to find a positive association and would be incorrectly tempted to conclude that matches and lighters indeed cause lung cancer. However, this is a false result brought about by the confounding effect of smoking, which is the real cause of both X (smokers carry matches and lighters) and Z (smokers are at increased risk of lung cancer). From the marginal analysis, we would arrive at the erroneous decision that matches and lighters are detrimental to the population's health and should be banned immediately.

A different situation is illustrated by Figure 4.2. Here, the variable V is not the cause but a consequence of X. Thus, a part of the effect of X on Z is mediated through V. In this case, the full effect of X on Z can be only determined from the marginal association that ignores V. The conditional association would only reveal the part of the effect of X on Z that is not mediated through V (and there may be none).

The difference between the two figures is only in the direction of the arrow between X and V. The direction of causality between these two variables decides whether (1) V is a confounder, must be measured, taken into account, and included in the model as a covariate or (2) V is a mediating variable and should be ignored. The distinction between these two possibilities cannot be made from the data alone. If we analyze the data to evaluate the relationship of X and V, we only find out that these variables are correlated but we cannot infer the direction of causality between them. We need a deep external knowledge of the problem we are solving, we need to understand what the variables really mean in the context of the problem and what mechanisms govern their relationships. If we view the variables contained in the dataset just as some abstract objects we cannot construct a valid model and reach the correct conclusions.

We cannot make this problem disappear by ignoring it. The distorting effect of a confounder persists even if the confounder is not measured, and even if we do not have any idea about the existence of the confounder. This is the reason why statistical analyses of observational studies ("routinely collected data") are fraught with problems. If such studies do not have a thoughtfully chosen design (= data collection mechanism) and if they are not analyzed with sufficient expertise in both statistical methodology and the field of application, their results are likely to be false. The only reliable way to prevent confounding is to conduct active experiments which assign levels of the variable X in such a way that associations between X and any external variables cannot arise. These topics are covered by courses in *experimental design*. However, many practical problems do not allow the conduct of active experiments.

The problem of making false causal conclusions from observational data affected by confounding is aggravated by the availability of large databases containing routinely collected data. Such databases often contain haphazard sets of data items whose collection mechanism is not under control and therefore do not allow appropriate measures to remove

confounding.* As noted above, this leads to invalid regression models which do not produce correct conclusions. Building such regression models on larger and larger data sets ("big data") only leads to producing larger and larger errors.

The confounding issue also represents a substantial weakness of the recently popular data analysis methods called "artificial intelligence", "neural networks", "deep learning" etc. Even though these methods focus on making predictions, there is a danger that differences in the predictions between various subjects will be interpreted causally. Also, predictions made from models that suffer from confounding will become wrong if there is any change in the associations of the unaccounted confounder with the predictors or the response.

Finally, let's provide an interpretation for the gender-salary example. We want to know whether gender X has an effect on the salary Z. The third variable is V = field of study. Because gender cannot be changed by the field of study, we know that we are in the framework of Fig. 4.2: the field of study acts as a mediator for the effect of gender on the salary. In the conditional analysis (after controlling V), gender has no effect on salary. In the marginal analysis (ignoring V), men have twice the odds for high salary than women. We can conclude that men indeed are paid better, but only because they study the fields that provide better salaries more frequently than women. If the choice of the field of study is the result of the free will rather than the result of coercion or discrimination in admission procedures, we can conclude that there is no salary discrimination (in this artificial example, at least).

4.5.5. The independence model (X, Z, V)

This model includes only main effects and no interactions. Its equation is

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V. \tag{4.10}$$

Again, we adopt the constraints that set any parameter with subscript 1 to zero. Here, we have $\beta_1^X = \beta_1^Z = \beta_1^V = 0$.

Because $m_{ijk} = \pi_{ijk} m_{+++}$, the model can be also written as

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V. \tag{4.11}$$

The dimension of the parameter vector is $p = 1 + I - 1 + J - 1 + K - 1 = I + J + K - 2$.

Interpretation of the main effects:

The interpretation of the main effects can be determined in the same way as with the model (X, Z) for a two way table. In particular, start with

$$\log \frac{\pi_{ijk}}{\pi_{1jk}} = \beta_i^X \quad \text{for any } j, k.$$

* The presence or absence of a certain data item in the database represents a potential very strong confounding factor that cannot be adjusted for.

Transform it into

$$\pi_{ijk} = \pi_{1jk} e^{\beta_i^X} \quad \text{for any } j,k,$$

sum it over all j and k to get

$$\pi_{i++} = \pi_{1++} e^{\beta_i^X}$$

and thus

$$e^{\beta_i^X} = \frac{\pi_{i++}}{\pi_{1++}} = \frac{P[X=i]}{P[X=1]}.$$

This can be repeated for the other main effects, so

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

Expressing the cell probabilities:

To express the cell probabilities, plug the main effects into (4.11) to get

$$\pi_{ijk} = \pi_{111} \frac{\pi_{i++}}{\pi_{1++}} \frac{\pi_{+j+}}{\pi_{+1+}} \frac{\pi_{++k}}{\pi_{++1}}$$

for all i,j,k. Summing these equations over i,j,k gives

$$1 = \pi_{+++} = \frac{\pi_{111}}{\pi_{1++}\pi_{+1+}\pi_{++1}}$$

and hence

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

for all i,j,k.

Interpretation of the model:

The model (X,Z,V) holds if and only if

$$P[X=i, Z=j, V=k] = P[X=i]P[Z=j]P[V=k] \quad \text{for all } i,j,k,$$

that is, if the variables X, Z and V are all marginally independent. Conditional independence follows from marginal independence – see Theorem 4.4 and use fact that $\pi_{i+k} = \pi_{i++}\pi_{++k}$ and $\pi_{+jk} = \pi_{+j+}\pi_{++k}$, $\pi_{ij+} = \pi_{i++}\pi_{+j+}$.

4.5.6. Model (XV, Z)

Now, add a single two-way interaction, for example between variables X and V.

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

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} \tag{4.12}$$

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 $p = I + J + K - 2 + (I-1)(K-1) = IK + J - 1$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV}. \tag{4.13}$$

Interpretation of the main effects:

Because Z does not enter any interaction term, the interpretation of the main effect β_j^Z is the same as in the previous model, i.e.,

$$e^{\beta_j^Z} = \frac{\pi_{+j+}}{\pi_{+1+}} = \frac{P[Z=j]}{P[Z=1]}. \tag{$*$}$$

For the main effect of X, we get from (4.13)

$$\beta_i^X = \log \pi_{ij1} - \log \pi_{1j1} \quad \text{for any } j \quad \text{and} \quad \pi_{ij1} = \pi_{1j1} e^{\beta_i^X}.$$

Sum this over all j to get

$$e^{\beta_i^X} = \frac{\pi_{i+1}}{\pi_{1+1}} = \frac{P[X=i, V=1]}{P[X=1, V=1]} = \frac{P\big(X=i \,\big|\, V=1\big)}{P\big(X=1 \,\big|\, V=1\big)}.$$

Similarly,

$$e^{\beta_k^V} = \frac{\pi_{1+k}}{\pi_{1+1}} = \frac{P[V=k, X=1]}{P[V=1, X=1]} = \frac{P\big(V=k \,\big|\, X=1\big)}{P\big(V=1 \,\big|\, X=1\big)}.$$

Interpretation of the interaction:

The interaction parameter β_{ik}^{XV} can be separated directly from (4.13) as follows:

$$\beta_{ik}^{XV} = \log \pi_{ijk} - \log \pi_{1jk} - \log \pi_{ij1} + \log \pi_{1j1}. \tag{$\dagger$}$$

Thus,

$$e^{\beta_{ik}^{XV}} = \frac{\pi_{ijk}\pi_{1j1}}{\pi_{1jk}\pi_{ij1}} = \theta_{ik(j)}^{XV},$$

the conditional odds ratio between X and V given $Z = j$. But the same interaction parameter also expresses the marginal odds ratio between X and V, which can be shown as follows. Start from the expression for π_{i+k} based on (4.13):

$$\pi_{i+k} = \sum_{j=1}^{J} \pi_{111} e^{\beta_i^X} e^{\beta_j^Z} e^{\beta_k^V} e^{\beta_{ik}^{XV}}$$

and see from $(*)$ that $\sum_{j=1}^{J} e^{\beta_j^Z} = 1/\pi_{+1+}$. This gives us the equation

$$\log \pi_{i+k} = \log \pi_{111} - \log \pi_{+1+} + \beta_i^X + \beta_k^V + \beta_{ik}^{XV}$$

and, from this,

$$\beta_{ik}^{XV} = \log \pi_{i+k} - \log \pi_{i+1} - \log \pi_{1+k} + \log \pi_{1+1}. \qquad (**)$$

We confirmed that

$$e^{\beta_{ik}^{XV}} = \frac{\pi_{i+k}\pi_{1+1}}{\pi_{i+1}\pi_{1+k}} = \theta_{ik}^{XV},$$

the marginal odds ratio between X and V.

Expressing the cell probabilities:

Plug the expression $(**)$ for the interaction term and the expressions for the main effects into the model equation (4.13). We get

$$\log \pi_{ijk} = \log \pi_{111} + \log \frac{\pi_{i+1}}{\pi_{1+1}} + \log \frac{\pi_{+j+}}{\pi_{+1+}} + \log \frac{\pi_{1+k}}{\pi_{1+1}} + \log \frac{\pi_{i+k}\pi_{1+1}}{\pi_{i+1}\pi_{1+k}}.$$

Delete the terms that cancel each other out and rearrange the rest to get

$$\pi_{ijk} = \frac{\pi_{111}}{\pi_{1+1}\pi_{+1+}} \pi_{i+k}\pi_{+j+}.$$

It is easy to find out by summation that $\frac{\pi_{111}}{\pi_{1+1}\pi_{+1+}} = 1$ and hence the cell probabilities satisfy the equation

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

for all i, j, k.

Interpretation of the model:

In the model (XV, Z), we have

$$\mathrm{P}[X = i, Z = j, V = k] = \mathrm{P}[X = i, V = k]\,\mathrm{P}[Z = j] \quad \text{for all } i, j, k.$$

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 are the same.

4.5.7. Model (XV, ZV)

Add another two-way interaction to the preceding model, this time between Z and V.

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

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} \qquad (4.14)$$

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 $p = IK + J - 1 + (J - 1)(K - 1) = K(I + J - 1)$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV}. \tag{4.15}$$

Interpretation of the main effects:

The main effects for X have the same interpretation as in the previous model. The interpretation of the main effects for Z is analogous to X (X and Z can be exchanged and the model still applies). The main effects of V can be only separated when $i = j = 1$ and no summation is possible (because V is simultaneously present in two interaction terms). Hence

$$\begin{aligned}
e^{\beta_i^X} &= \frac{\pi_{i+1}}{\pi_{1+1}} = \frac{P\big(X = i \,\big|\, V = 1\big)}{P\big(X = 1 \,\big|\, V = 1\big)}, \\
e^{\beta_j^Z} &= \frac{\pi_{+j1}}{\pi_{+11}} = \frac{P\big(Z = j \,\big|\, V = 1\big)}{P\big(Z = 1 \,\big|\, V = 1\big)}, \quad \text{and} \\
e^{\beta_k^V} &= \frac{\pi_{11k}}{\pi_{111}} = \frac{P\big(V = k \,\big|\, X = 1, Z = 1\big)}{P\big(V = 1 \,\big|\, X = 1, Z = 1\big)}.
\end{aligned}$$

Interpretation of the interactions:

For β_{ik}^{XV}, the expression (†) is still valid. Hence,

$$e^{\beta_{ik}^{XV}} = \frac{\pi_{ijk}\pi_{1j1}}{\pi_{1jk}\pi_{ij1}} = \theta_{ik(j)}^{XV},$$

which is the conditional odds ratio between X and V given $Z = j$. The interaction β_{jk}^{ZV} can be interpreted by analogy in the same way:

$$e^{\beta_{jk}^{ZV}} = \frac{\pi_{ijk}\pi_{i11}}{\pi_{ij1}\pi_{i1k}} = \theta_{jk(i)}^{ZV},$$

which is the conditional odds ratio between Z and V given $X = i$.

The lack of the XZ interaction term means that X and Z are conditionally independent given V. It is easy to verify that their conditional odds ratios are all one:

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

Expressing the cell probabilities:

In order to express cell probabilities in the model (XV, ZV) and to investigate marginal associations, we need to start with expressions for π_{i+k}, π_{+jk} and π_{++k} obtained from (4.15).

We have

$$\log \pi_{i+k} = \log \pi_{111} + \beta_i^X + \beta_k^V + \beta_{ik}^{XV} + \underbrace{\log \sum_{j=1}^{J} e^{\beta_j^Z + \beta_{jk}^{ZV}}}_{\text{denote } \equiv \gamma_k}, \qquad (\ddagger)$$

$$\log \pi_{+jk} = \log \pi_{111} + \beta_j^Z + \beta_k^V + \beta_{jk}^{ZV} + \underbrace{\log \sum_{i=1}^{I} e^{\beta_i^X + \beta_{ik}^{XV}}}_{\text{denote } \equiv \delta_k},$$

and

$$\log \pi_{++k} = \log \pi_{111} + \beta_k^V + \underbrace{\log \sum_{i=1}^{I} e^{\beta_i^X + \beta_{ik}^{XV}}}_{=\delta_k} + \underbrace{\log \sum_{j=1}^{J} e^{\beta_j^Z + \beta_{jk}^{ZV}}}_{=\gamma_k}.$$

Now we can calculate

$$P\big(X=i, Z=j \,\big|\, V=k\big) = \frac{\pi_{ijk}}{\pi_{++k}} = e^{\beta_i^X + \beta_j^Z + \beta_{ik}^{XV} + \beta_{jk}^{ZV} - \delta_k - \gamma_k} = e^{\beta_i^X + \beta_{ik}^{XV} - \delta_k} e^{\beta_j^Z + \beta_{jk}^{ZV} - \gamma_k} = \frac{\pi_{i+k}}{\pi_{++k}} \frac{\pi_{+jk}}{\pi_{++k}}.$$

Thus, the cell probabilities satisfy the equation

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

for all i, j, k. This means that X and Z are conditionally independent given V. However, X and Z are not marginally independent because their marginal odds ratio is

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

and this cannot be equal to 1 for all i, j.

From (‡), we can express the marginal odds ratios between X and V:

$$\theta_{ik}^{XV} = \frac{\pi_{i+k}\pi_{1+1}}{\pi_{i+1}\pi_{1+k}} = e^{\beta_{ik}^{XV}}.$$

We can see that they are equal to the conditional odds ratios. The same is true for the marginal and conditional associations between Z and V.

Interpretation of the model:

In the model (XV, ZV), X is conditionally independent of Z given V but marginally dependent. This is the setup of Simpson's paradox.

Also, (i) X is associated with V and the marginal and conditional associations between X and V agree, (ii) Z is associated with V and the marginal and conditional associations between Z and V agree.

4.5.8. Model (XV, ZV, XZ)

At this step, we add another two-way interaction. The model (XV, ZV, XZ) includes all possible two-way interactions between the three factors. The model equation is

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} \tag{4.16}$$

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, $\beta_{1k}^{ZV} = 0$ for all k, $\beta_{i1}^{XZ} = 0$ for all i, and $\beta_{1j}^{XZ} = 0$ for all j. The number of parameters in this model is $p = K(I+J-1) + (I-1)(J-1) = KI + KJ + IJ - (I+J+K) + 1$.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} \tag{4.17}$$

Interpretation of the main effects:

The main effects have the same expressions as the main effects for V in the previous model. They compare probabilities of a given level to the first level of the factor, when the other factors are at the first level.

$$e^{\beta_i^X} = \frac{\pi_{i11}}{\pi_{111}}, \quad e^{\beta_j^Z} = \frac{\pi_{1j1}}{\pi_{111}}, \quad e^{\beta_k^V} = \frac{\pi_{11k}}{\pi_{111}}.$$

Interpretation of the interactions:

All the interaction parameters can be interpreted as the conditional odds ratios:

$$\begin{aligned} e^{\beta_{ij}^{XZ}} &= \frac{\pi_{ijk}\pi_{11k}}{\pi_{i1k}\pi_{1jk}} = \theta_{ij(k)}^{XZ} \\ e^{\beta_{ik}^{XV}} &= \frac{\pi_{ijk}\pi_{1j1}}{\pi_{1jk}\pi_{ij1}} = \theta_{ik(j)}^{XV} \\ e^{\beta_{jk}^{ZV}} &= \frac{\pi_{ijk}\pi_{i11}}{\pi_{ij1}\pi_{i1k}} = \theta_{jk(i)}^{ZV} \end{aligned}$$

Expressing the cell probabilities:

The cell probabilities π_{ijk} cannot be expressed in terms of marginal probabilities. Marginal odds ratios are all different from conditional odds ratios and do not have convenient expressions.

Interpretation of the model:

In the model (XV, ZV, XZ),

- 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 (e.g., associations between X and Z are the same at each level k of V).

4.5.9. Model (XZV)

The model (XZV) includes a three-way interaction (in addition to all two-way interactions) and is saturated. Its model equation is

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} + \beta_{ijk}^{XZV} \tag{4.18}$$

with the usual constraints (any parameter with 1 anywhere among the indices is set to 0). The number of parameters in this model is equal to the number of cells in the table, $p = IJK$. It is the saturated model for the three-way table.

The model for cell probabilities is

$$\log \pi_{ijk} = \log \pi_{111} + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{ik}^{XV} + \beta_{jk}^{ZV} + \beta_{ij}^{XZ} + \beta_{ijk}^{XZV} \tag{4.19}$$

Interpretation of the main effects:

The interpretation of the main effects is the same as in the model (XV, ZV, XZ).

Interpretation of the interactions:

The second-order interactions determine the conditional odds ratios at the first level of the conditioning variable:

$$\begin{aligned} e^{\beta_{ij}^{XZ}} &= \frac{\pi_{ij1}\pi_{111}}{\pi_{i11}\pi_{1j1}} = \theta_{ij(1)}^{XZ} \\ e^{\beta_{ik}^{XV}} &= \frac{\pi_{i1k}\pi_{111}}{\pi_{11k}\pi_{i11}} = \theta_{ik(1)}^{XV} \\ e^{\beta_{jk}^{ZV}} &= \frac{\pi_{1jk}\pi_{111}}{\pi_{1j1}\pi_{11k}} = \theta_{jk(1)}^{ZV} \end{aligned}$$

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

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

Expressing the cell probabilities:

The cell probabilities π_{ijk} cannot be expressed in terms of marginal probabilities. Marginal odds ratios are all different from conditional odds ratios and do not have nice expressions.

Interpretation of the model:

In the model (XZV),

- all three variables are marginally and conditionally dependent,
- marginal associations are different from conditional associations,
- conditional associations depend on the value of the conditioning variable (e.g., associations between X and Z vary with the level k of V).

4.5.10. Model selection

The model selection strategy that usually works best is to start from the saturated model (XZV) and test whether the 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, according to the interpretations provided above.

4.6. 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.

Example:

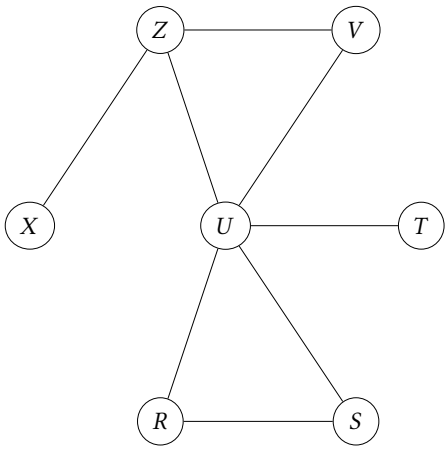
Suppose we have observations of seven categorical variables X, Z, U, V, R, S, and T. We build a seven-way contingency table of observed counts out of this data, fit a sequence of loglinear models and test significance of various interactions. We end up with the model

$$(XZ, ZUV, UT, URS).$$

There are

- two three-way interactions:
  - between Z, U, and V (plus all possible two-way interactions between these variables) and

Figure 4.3.: Graphical representation of the model (XZ, ZUV, UT, URS).



  - between U, R, and S (plus all possible two-way interactions between these variables), and
- additional two two-way interactions:
  - between X and Z and
  - between U and T.

Suppose we have enough data to fit such a complex model. Then, what does it tell us about the associations between variables?

Let us create a graph, put variables in the nodes and turn each two-way interaction that was included in the model into an edge connecting two nodes. The graphical representation of our model is shown in Figure 4.3.

What are the associations between the variables?

- all variables are marginally dependent with each other (there is a path leading from any variable to any other);
- X and V are conditionally independent given Z (the only path that connects X and V leads through Z);
- (X, Z, V), T, and (R, S) are conditionally independent given U (the three vectors are subgraphs that are only connected with each other through U).

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.

4.7. 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 $X \in \{1,\dots,I\}$ and $Z \in \{1,\dots,J\}$. 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

$$\log \frac{p_{ij}}{1-p_{ij}} = \gamma^0 + \gamma_i^X + \gamma_j^Z, \tag{4.20}$$

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,

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

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):

$$\log m_{ijk} = \alpha + \beta_i^X + \beta_j^Z + \beta_k^Y + \beta_{ik}^{XY} + \beta_{jk}^{ZY} + \beta_{ij}^{XZ} \tag{4.21}$$

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 (4.21) induces the same structure on the conditional log odds of success as the logistic model (4.20). 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 (4.20) or the loglinear model (4.21).

The only difference between the logistic model (4.20) and the loglinear model (4.21) 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.

5. Quasi-likelihood and Sandwich Variance

In this chapter, we will show in several steps that the assumptions of the GLM can be substantially relaxed. The conclusions will be similar to linear regression with normally distributed data: we can show that the results obtained for linear regression under the normality assumption hold asymptotically even if the distribution of the data is not normal. Next, we can show that the assumption of equal variances can be also removed, if we replace the ordinary variance estimator by the sandwich. In this chapter, we will see that something similar is true for the GLM as well. So, we can extend the methods for the analysis of the GLM to distributions that do not belong to the exponential family.

5.1. Quasi-likelihood and Overdispersion

This section is motivated by the problem of overdispersion, which we discuss first. We will show that even if we start with variables that have distributions in the exponential family, we can easily get into situations where the responses do not follow those distributions. Such situations violate the assumptions of the GLM. The idea of quasi-likelihood offers an approach to deal with such data.

5.1.1. Overdispersion in binomial data

Consider the following example. We have iid random variables $Y_1, \dots, Y_n \sim \mathsf{Bi}(m, \pi_0)$. Their moments are $\mathsf{E}\, Y_i = m\pi_0$ and $\mathsf{var}\, Y_i = m\pi_0(1-\pi_0)$. Such variables are sums of independent Bernoulli variables and can be analyzed by the methods of the previous chapters.

But let us change the setup slightly. Instead of

$$Y_i \sim \mathsf{Bi}(m, \pi_0)$$

take

$$Y_i \sim \mathsf{Bi}(m, \pi_i), \qquad (*)$$

where $\pi_1, \dots, \pi_n$ are iid random variables with $\mathsf{E}\, \pi_i = \pi_0$. So, the groups do not have the same success probability π_0 any more. Each of them has a different (unobserved) success probability π_i and π_0 is the mean of all those.

We want to estimate π_0 from the observations $Y_1, \dots, Y_n$ that are still iid random variables. However, their distribution is not binomial when $\pi_1, \dots, \pi_n$ are not observed.

Denote $g(\pi)$ the density of π_i. The expectation of π_i is $\mathsf{E}\,\pi_i = \pi_0$, as required above. Denote the variance of π_i by $\mathsf{var}\,\pi_i = \sigma_\pi^2$

Let us calculate the moments of Y_i. Conditionally on π_i, we have from $(*)$ $\mathsf{E}\big[Y_i \,\big|\, \pi_i\big] = m\pi_i$ and $\mathsf{var}\big[Y_i \,\big|\, \pi_i\big] = m\pi_i(1-\pi_i)$. So,

$$\begin{aligned}
\mathsf{E}\,Y_i &= \mathsf{E}\,\mathsf{E}\big[Y_i \,\big|\, \pi_i\big] = \mathsf{E}\,m\pi_i = m\pi_0, \\
\mathsf{var}\,Y_i &= \mathsf{E}\,\mathsf{var}\big[Y_i \,\big|\, \pi_i\big] + \mathsf{var}\,\mathsf{E}\big[Y_i \,\big|\, \pi_i\big] = \mathsf{E}\,m\pi_i(1-\pi_i) + \mathsf{var}\,m\pi_i \\
&= m\pi_0 - m(\sigma_\pi^2 + \pi_0^2) + m^2\sigma_\pi^2 = m\pi_0(1-\pi_0) + m(m-1)\sigma_\pi^2 \geq m\pi_0(1-\pi_0).
\end{aligned}$$

When $m = 1$, the distribution of Y_i is Bernoulli no matter if the success probabilities are random or not. There is no problem. However, when $m > 1$, the distribution of Y_i is not binomial and the variance of Y_i is larger than that of a binomial distribution. This phenomenon is called *overdispersion*. It arises because the Bernoulli variables that sum into Y_i are no longer independent; the random π_i induces a positive correlation between them.

Now let us be more specific about the distribution of π_i. As a special case, consider $\pi_i \sim \mathsf{B}(\alpha, \beta)$ (the beta distribution) for some unknown $\alpha > 0$ and $\beta > 0$. The moments of the beta distribution are

$$\begin{aligned}
\pi_0 &= \mathsf{E}\,\pi_i = \frac{\alpha}{\alpha+\beta}, \\
\sigma_\pi^2 &= \mathsf{var}\,\pi_i = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} = \frac{1}{\alpha+\beta+1}\pi_0(1-\pi_0).
\end{aligned}$$

Plugging the variance σ_π^2 into the expression for $\mathsf{var}\,Y_i$, we get

$$\mathsf{var}\,Y_i = m\pi_0(1-\pi_0) + \frac{m(m-1)}{\alpha+\beta+1}\pi_0(1-\pi_0) = \varphi\, m\pi_0(1-\pi_0),$$

where

$$\varphi = 1 + \frac{m-1}{\alpha+\beta+1} > 1$$

is a dispersion parameter. The distribution of Y_i has the same variance function as binomial (or Bernoulli) distribution but with an additional dispersion parameter that is > 1: *overdispersion*. This distribution cannot belong to the exponential family.

We can calculate what this distribution is. Start with writing down the density of π:

$$g(\pi) = \frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}, \quad \pi \in (0,1),$$

where

$$B(\alpha,\beta)=\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

is the beta function.

Take $j=0,\dots,m$ and calculate the density of Y_i using $f(y)=\int f(y|x)f(x)dx$

$$\begin{aligned}\mathrm{P}[Y_i=j]&=\int_0^1\binom{m}{j}\pi^j(1-\pi)^{m-j}\frac{1}{B(\alpha,\beta)}\pi^{\alpha-1}(1-\pi)^{\beta-1}d\pi\\&=\frac{1}{B(\alpha,\beta)}\binom{m}{j}\int_0^1\pi^{j+\alpha-1}(1-\pi)^{m-j+\beta-1}d\pi=\binom{m}{j}\frac{B(\alpha+j,\beta+m-j)}{B(\alpha,\beta)},\end{aligned}$$

where, at the last step, the integral was recognized as a beta density without the normalizing constant. This distribution is called *a beta-binomial distribution* with parameters $m=1,2,\dots$, $\alpha>0$, and $\beta>0$. Its first and second moments have been calculated above.

When α and β are both natural numbers, we can get a more explicit expression for the density. Using the relationship between beta and gamma functions and the fact that $\Gamma(p)=(p-1)!$ for a natural p, we have

$$B(\alpha,\beta)=\frac{(\alpha-1)!(\beta-1)!}{(\alpha+\beta-1)!},\quad B(\alpha+j,\beta+m-j)=\frac{(\alpha+j-1)!(\beta+m-j-1)!}{(\alpha+\beta+m-1)!},$$

and

$$\begin{aligned}\mathrm{P}[Y_i=j]&=\frac{m!}{(m-j)!j!}\,\frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\,\frac{(\alpha+j-1)!(\beta+m-j-1)!}{(\alpha+\beta+m-1)!}\\&=\frac{\binom{\alpha+j-1}{j}\binom{m-j+\beta-1}{m-j}}{\binom{m+\alpha+\beta-1}{m}}.\end{aligned}$$

This distribution arises in so called ***Pólya urn scheme****.*
Suppose there are α white balls and β black balls in an urn. Conduct a sampling experiment as follows: draw a ball from the urn randomly, note its color, return it into the urn, add another ball of the same color as the last drawn ball, mix the balls. Continue the process in the same way until you have drawn m balls.
Then the number of white balls that were drawn in m steps of this process has the ***beta-binomial distribution*** *given above.*
If the drawn ball is returned into the urn but no additional balls are added (sampling with replacement), the number of white balls drawn has a ***binomial distribution****.*
If the drawn ball is not returned into the urn (sampling without replacement), the number of white balls drawn has a ***hypergeometric distribution****.*

Beta-binomial distribution does not belong to the exponential family and we would like to extend the theory of GLM to responses following distributions of this type.

5.1.2. Overdispersion in Poisson data

By a similar consideration, overdispersion can be induced in Poisson distribution as well. Start with independent random variables $Y_1, \dots, Y_n \sim \mathsf{Po}(\lambda_0)$. We have $\mathsf{E}\, Y_i = \mathsf{var}\, Y_i = \lambda_0$.

Now modify their distribution as follows. Take $\lambda_1, \dots, \lambda_n$ as iid random variables with $\mathsf{E}\,\lambda_i = \lambda_0$ and $\mathsf{var}\,\lambda_i = \sigma_\lambda^2$ and generate Y_i from

$$Y_i \sim \mathsf{Po}(\lambda_i).$$

We want to estimate λ_0 from the iid observations $Y_1, \dots, Y_n$. However, their distribution is not Poisson.

Let us calculate the moments of Y_i. Conditionally on λ_i, we have $\mathsf{E}\big[Y_i \,\big|\lambda_i\big] = \lambda_i$ and $\mathsf{var}\big[Y_i \,\big|\lambda_i\big] = \lambda_i$. So,

$$\begin{aligned}\mathsf{E}\, Y_i &= \mathsf{E}\,\mathsf{E}\big[Y_i \,\big|\lambda_i\big] = \mathsf{E}\,\lambda_i = \lambda_0,\\ \mathsf{var}\, Y_i &= \mathsf{E}\,\mathsf{var}\big[Y_i \,\big|\lambda_i\big] + \mathsf{var}\,\mathsf{E}\big[Y_i \,\big|\lambda_i\big] = \mathsf{E}\,\lambda_i + \mathsf{var}\,\lambda_i = \lambda_0 + \sigma_\lambda^2 > \lambda_0.\end{aligned}$$

Thus, the variance of Y_i is larger than that of a Poisson distribution. Again, we have encountered *overdispersion*.

As a special case, consider $\lambda_i \sim \Gamma(a, a\lambda_0)$ (the gamma distribution) for some unknown $a > 0$. The moments of this distribution are

$$\begin{aligned}\mathsf{E}\,\lambda_i &= \frac{a\lambda_0}{a} = \lambda_0,\\ \sigma_\lambda^2 = \mathsf{var}\,\lambda_i &= \frac{a\lambda_0}{a^2} = \frac{1}{a}\lambda_0\end{aligned}$$

Plugging the variance σ_λ^2 into the expression for $\mathsf{var}\, Y_i$, we get

$$\mathsf{var}\, Y_i = \lambda_0 + \frac{1}{a}\lambda_0 = \Big(1 + \frac{1}{a}\Big)\lambda_0 = \varphi\lambda_0,$$

where

$$\varphi = 1 + \frac{1}{a} > 1$$

is a dispersion parameter. The distribution of Y_i has the same variance function as the Poisson distribution but with an additional dispersion parameter: *overdispersion*. This distribution cannot belong to the exponential family.

Let us calculate the density of this distribution. The density of λ_i is

$$g(\lambda) = \frac{a^{a\lambda_0}}{\Gamma(a\lambda_0)}\lambda^{a\lambda_0 - 1}\mathrm{e}^{-a\lambda}, \quad \lambda > 0.$$

So, for $k = 0, 1, 2, \ldots$,

$$\begin{aligned}
\mathrm{P}[Y_i = k] &= \int_0^\infty \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda} \frac{a^{a\lambda_0}}{\Gamma(a\lambda_0)} \lambda^{a\lambda_0 - 1} \mathrm{e}^{-a\lambda} d\lambda \\
&= \frac{a^{a\lambda_0}}{\Gamma(a\lambda_0)} \frac{1}{k!} \int_0^\infty \lambda^{k + a\lambda_0 - 1} \mathrm{e}^{-(a+1)\lambda} d\lambda = \frac{a^{a\lambda_0}}{\Gamma(a\lambda_0)} \frac{1}{k!} \frac{\Gamma(k + a\lambda_0)}{(a+1)^{k+a\lambda_0}} \\
&= \frac{1}{kB(k, a\lambda_0)} \left(\frac{a}{a+1} \right)^{a\lambda_0} \frac{1}{(a+1)^k}.
\end{aligned}$$

This distribution is called *a Poisson-gamma distribution* with parameters $a > 0$ and $\lambda_0 > 0$. Its first and second moments have been calculated above.

The Poisson-gamma distribution has an interesting special case: the *negative binomial distribution* $\mathsf{NB}(m, p)$ is obtained by setting $a\lambda_0 = m \in \mathsf{N}$ and $a/(a+1) = p$. In turn, the *geometric distribution* $\mathsf{Geo}(p)$ is a special case of negative binomial, with $m = 1$.

5.1.3. Quasi-likelihood

Regression analyses of overdispersed responses (and other cases) can be put in the framework of the GLM by so called *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. Compared to the GLM the assumptions have changed. We do not require the distribution of the response to belong to the exponential family. Instead, we assume we know the variance function of the response.

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. Instead, the regression parameters are estimated by the *maximum quasi-likelihood estimator*.

The strategy is to specify a quasi-likelihood function so that its score function is the same as in the GLM. That score only depends on quantities that have been specified in the assumptions above.

Definition 5.1. (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 5.1. The following lemma says that the quasi-score has exactly the same properties as the GLM score.

Lemma 5.1. *(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* $\mathrm{var}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) = -\mathsf{E} \frac{\partial}{\partial \boldsymbol{\beta}^T} \boldsymbol{U}_i(\boldsymbol{\beta}_0) = I(\boldsymbol{\beta}_0)$, *where* $I(\boldsymbol{\beta}_0)$ *is defined by* (2.8). $\diamondsuit$

The proof of lemma 5.1 follows the same steps that were done in Section 2.3.

The asymptotic properties of the maximum quasi-likelihood estimator follow from the general theory on consistency and asymptotic normality of Z-estimators (see Lecture notes for NMST434 Modern Statistical Methods).

* Česky *kvazivěrohodnost* † Česky *kvaziskóre*

Proposition 5.2. *There exists a sequence $\widehat{\boldsymbol{\beta}}_n$ of solutions to the quasi-likelihood equations such that $\widehat{\boldsymbol{\beta}}_n \xrightarrow{P} \boldsymbol{\beta}_0$.* ♢

The proposition claims consistency as long as the solutions exist and are unique. The key for the validity of the proposition is Lemma 5.1 (ii). Proposition 5.2 together with Lemma 5.1 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.

Theorem 5.3.

(i) $\frac{1}{\sqrt{n}} U_n(\boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, I(\boldsymbol{\beta}_0))$,

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

The asymptotic variance of $\widehat{\boldsymbol{\beta}}_n$ does not have the sandwich form because the two components of the sandwich are the same by Lemma 5.1 (iii).

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).

5.1.4. Quasi-likelihood: Advice on practical use

In order to use quasi-likelihood methods appropriately, one has to correctly specify both the mean and the variance function. Let us illustrate the principles on two instructive examples.

This part shows you how to use quasi-likelihood to solve two specific examples. There could be many other similar examples and we do not have time to go through each of them. The important thing is to be able to take advantage of the flexibility of quasi-likelihood by thinking about the problem to be solved.

Beta-binomial distribution

Beta-binomial distribution should be used when we observe numbers of successes and failures from groups of observations that are correlated by sharing their specific success probability.

In Section 3, we had $Y_{ij}^* \sim \mathsf{Alt}(\pi_i)$, $i = 1,\ldots,K$, $j = 1,\ldots,m_i$, where $\pi_i = \mathrm{e}^{\eta_i}/(1+\mathrm{e}^{\eta_i})$ and the linear predictor was $\eta_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_p X_{pi}$. These Bernoulli variables were iid and their sum (the number of successes in the group) $Y_i = \sum_{j=1}^{m_i} Y_{ij}^*$ had binomial distribution $Y_i \sim \mathsf{Bi}(m_i, \pi_i)$.

Now suppose that the covariates $X_{2i},\ldots,X_{pi}$ do not fully describe the differences in π_i between the groups — for example because some important covariate was not recorded among those that we measured. This can be captured by considering π_i a random variable with the mean $\mu_i = \mathrm{e}^{\eta_i}/(1+\mathrm{e}^{\eta_i})$ and some variability expressing the effect of that unobserved covariate (or any other influences on π_i). Now, $Y_{i1}^*,\ldots,Y_{im_i}^*$ are not independent and their sum Y_i does not have a binomial distribution. A reasonable model for such data is the beta-binomial model, through the specification

- $\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i\right] \equiv m_i\mu_i = m_i \dfrac{\mathrm{e}^{\eta_i}}{1+\mathrm{e}^{\eta_i}}$ and
- $\mathsf{var}\left[Y_i \middle| \boldsymbol{X}_i\right] = \varphi m_i \mu_i (1-\mu_i)$.

This model can be fitted by quasi-likelihood using the GLM formulae for aggregated binary data (the reduced "binomial" format) from Section 3.5. The dispersion parameter is estimated as

$$\hat{\varphi} = \frac{1}{K-p}\sum_{i=1}^{K}\frac{(Y_i - m_i\widehat{\mu}_i)^2}{m_i\widehat{\mu}_i(1-\widehat{\mu}_i)}.$$

Here, each group contributes one term to the Pearson X^2 statistic. If $\widehat{\varphi} \approx 1$, we can suspect that there was not a serious overdispersion after all and that the binomial model would have fitted the data well (but remember: have not introduced a formal test of $H_0 : \varphi = 1$ that controls the Type I error).

Negative binomial and geometric distributions

Another application of quasi-likelihood allows fitting negative binomial or geometric distribution. Suppose Y_i have negative binomial distributions with a known common parameter m and mean μ_i. This means that Y_i is the number of failures suffered before the m-th success is observed (in a sequence of independent Bernoulli trials with success probability p_i depending on the subject). We'll choose a loglinear model for the mean, i.e., $\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i\right] = \mu_i = \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}}$ and want to estimate and test the components of $\boldsymbol{\beta}$.

The general Poisson-gamma distribution has the expectation λ_0 and variance $(1+1/a)\lambda_0$. Negative binomial distribution is a special case with $\lambda_0 = m/a$ and $a/(a+1) = p$. Hence, $a = p/(1-p)$. In the regression case, the parameter p (and hence, a) depends on the index i while m is constant. The mean is

$$\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i\right] = \mu_i = \frac{m}{a_i} = \frac{m(1-p_i)}{p_i}.$$

The variance can be expressed in two ways: as a function of the parameters m and p_i like

this

$$\mathsf{var}\left[Y_i \middle| \boldsymbol{X}_i\right] = \left(1 + \frac{1}{a_i}\right)\mu_i = \frac{a_i + 1}{a_i}\mu_i = \frac{m(1-p_i)}{p_i^2}$$

or as a function of the mean μ_i like this

$$\mathsf{var}\left[Y_i \middle| \boldsymbol{X}_i\right] = \left(1 + \frac{1}{a_i}\right)\mu_i = \mu_i + \frac{\mu_i}{a_i} = \mu_i + \frac{\mu_i^2}{m}.$$

This is the variance function of the negative binomial distribution. So, we take

- $\mathsf{E}\left[Y_i \middle| \boldsymbol{X}_i\right] \equiv \mu_i = \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}}$ and
- $\mathsf{var}\left[Y_i \middle| \boldsymbol{X}_i\right] = V(\mu_i) = \mu_i + \frac{\mu_i^2}{m}$.

Note that the dispersion parameter is $\varphi = 1$.

Now, quasi-likelihood can be used to fit this model. If $m = 1$, we have the geometric distribution as a special case. If m varied by observations, we would need to extend the approach to allow variance functions of the form $V(\mu_i, m_i)$ with observed m_i plugged in.

Last, you can notice this: the loglinear model for the expectation of the response gives you

$$\frac{m(1-p_i)}{p_i} = \mathrm{e}^{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}}$$

and hence

$$\log\frac{1-p_i}{p_i} = -\log m + \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}.$$

This means that the model can be also interpreted as a logistic model for the success probabilities p_i! The exponentiated parameters give you the odds ratios for success per unit change in the covariate.

5.2. Sandwich Variance Estimation in the GLM

In this section we will weaken the assumptions we are working with even further. We will only require that the conditional mean is specified correctly by the model but the variance function may be wrong. The results we provide in this chapter are justified by the theory about the behavior of maximum likelihood estimators when the model is not correctly specified. This theory was proposed by White (1982) and is summarized in the Appendix, section A.2.

5.2.1. 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. $Y_1, \dots, Y_n$ 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

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

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 5.4.

(i) $\boldsymbol{U}_i(\boldsymbol{\beta})$, $i = 1, \dots, n$, *are iid random vectors.*
(ii) If $\boldsymbol{\beta}_0$ *is the true parameter then* $\mathsf{E}\, \boldsymbol{U}_i(\boldsymbol{\beta}_0) = \boldsymbol{0}$. ♢

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

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

and

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

It follows from Theorem A.11 (note that $I(\boldsymbol{\beta}_0)$ plays the role of the matrix $\mathbb{D}$) that

Theorem 5.5.

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

[*] Česky *pracovní rozptylová funkce* [†] Česky *pseudoskórová funkce*

(ii) $\frac{1}{\sqrt{n}} U_n(\boldsymbol{\beta}_0) \xrightarrow{D} N_p(\mathbf{0}, \Sigma)$,

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

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 A.12. 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 5.5 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.

5.2.2. Sandwich variance: Advice on practical use

In order to use sandwich methods appropriately, one needs to correctly specify the mean, make a best guess about the variance function and use it as the working variance in the estimating procedures. You should be careful with the sandwich variance estimator: it is known to underestimate the true variability and does not work well with small to moderate sample sizes. You can use, e.g., bootstrap to improve it but remember that bootstrap is also an asymptotic method and cannot do miracles with small sample sizes.

In R, sandwich estimation can be performed with the functions available in libraries `sandwich` and `lmtest`. Both libraries work for model fits obtained from the function `glm()`.

A. Appendix

A.1. Summary of Classical Maximum Likelihood Theory

A.1.1. Definition

Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of independent identically distributed random variables (or vectors), each with density $f(x|\boldsymbol{\theta}_X)$ with respect to a σ-finite measure μ. We assume that $f(x|\boldsymbol{\theta}_X) \in \mathscr{F}$, where

$$\mathscr{F} = \{\text{distributions with density } f(x|\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$$

represents a parametric model for the distribution of the data.

The model $\mathscr{F}$ must satisfy the model identifiability condition: For any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ it holds $f(x|\boldsymbol{\theta}_1) \neq f(x|\boldsymbol{\theta}_2)$. In other words, no distribution can be parametrized by several different parameter vectors.

Because of independence, the joint denstity of the random sample $X_1, \ldots, X_n$ is $\prod_{i=1}^n f(x_i|\boldsymbol{\theta}_X)$. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}_X$ is the point from Θ that maximizes the joint density evaluated at the observed values of $X_1, \ldots, X_n$.

Definition A.1 (likelihood, log-likelihood).

- The random function
$$L_n(\boldsymbol{\theta}) \stackrel{\text{df}}{=} \prod_{i=1}^n f(X_i|\boldsymbol{\theta})$$
is called *the likelihood function* for the parameter $\boldsymbol{\theta}$ in the model $\mathscr{F}$.
- The random function
$$\ell_n(\boldsymbol{\theta}) \stackrel{\text{df}}{=} \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(X_i|\boldsymbol{\theta})$$
is called *the log-likelihood function*. ∇

Definition A.2 (maximum likelihood estimator). *The maximum likelihood estimator* (MLE) of the parameter $\boldsymbol{\theta}_X$ in the model $\mathscr{F}$ is defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta} \in \Theta} L_n(\boldsymbol{\theta}).$$ ∇

Note. Since the logarithm is strictly increasing, $L_n(\boldsymbol{\theta})$ and $\ell_n(\boldsymbol{\theta})$ attain the maximum at the same point.

Definition A.3. Let P and Q be probability measures on the same probability space with densities p and q with respect to the same σ-finite measure μ (for example, $\mu = P + Q$). Define

$$K(P,Q) = \begin{cases} \mathsf{E}_P \log \frac{p(X)}{q(X)} = \int_{\{x:p(x)>0\}} \log \frac{p(x)}{q(x)} p(x)\, d\mu(x) & \text{if } \mathrm{P}[q(X)=0]=0 \\ +\infty & \text{otherwise.} \end{cases}$$

$K(P,Q)$ is called the *Kullback-Leibler distance (divergence)*. ∇

Note. In fact, $K(P,Q)$ is a pseudo-distance: it holds $K(P,Q) \geq 0$, and $K(P,Q) = 0$ if and only if $P = Q$, but it is not symmetric: $K(P,Q) \neq K(Q,P)$.

Theorem A.1. *Suppose the support set $S = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on the parameter $\boldsymbol{\theta}$. Denote P_X the induced probability measure of the random variable X_i and P_θ the probability measure associated with the density $f(x|\boldsymbol{\theta})$. Then for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_X$*

$$\frac{1}{n} \log \frac{L_n(\boldsymbol{\theta}_X)}{L_n(\boldsymbol{\theta})} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i|\boldsymbol{\theta}_X)}{f(X_i|\boldsymbol{\theta})} \to K(P_X, P_\theta) \quad P_X - \textit{ almost surely,}$$

and hence

$$\mathrm{P}[\ell_n(\boldsymbol{\theta}_X) > \ell_n(\boldsymbol{\theta})] \to 1 \quad \textit{as } n \to \infty. \qquad \diamondsuit$$

Note. When the number of observations increases to infinity, the (log-)likelihood function at the true parameter will be with a large probability larger than the (log-)likelihood function at any other parameter. This observation justifies the idea of estimating the parameters by maximizing the log-likelihood over all possible parameter vectors.

A.1.2. The calculation of the maximum likelihood estimator

The maximum likelihood estimator is usually determined by differentiation of the log-likelihood. The first derivative is set to zero and it is verified that the second derivative is negative definite.

Definition A.4 (score, information).

- The random vector
$$\boldsymbol{U}(\boldsymbol{\theta}|X_i) \stackrel{\mathrm{df}}{=} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta})$$
is called *the score function* for the parameter $\boldsymbol{\theta}$ in the model $\mathscr{F}$.

- The random vector
$$\boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) \stackrel{\text{df}}{=} \sum_{i=1}^{n} \boldsymbol{U}(\boldsymbol{\theta}|X_i) = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\boldsymbol{\theta})$$
is called *the score statistic*.
- The random matrix
$$I(\boldsymbol{\theta}|X_i) \stackrel{\text{df}}{=} -\frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{U}(\boldsymbol{\theta}|X_i) = -\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \log f(X_i|\boldsymbol{\theta})$$
is called the contribution of the i-th observation to the information matrix.
- The random matrix
$$I_n(\boldsymbol{\theta}|\boldsymbol{X}) \stackrel{\text{df}}{=} -\frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}^\top} \boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{\theta}|X_i)$$
is called *the observed information matrix*.
- The matrix
$$I(\boldsymbol{\theta}) \stackrel{\text{df}}{=} \mathsf{E}\, I(\boldsymbol{\theta}|X_i) = -\mathsf{E}\, \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top} \log f(X_i|\boldsymbol{\theta})$$
is called *the expected (Fisher) information matrix*. ∇

If the set Θ is open, the MLE $\widehat{\boldsymbol{\theta}}_n$ solves the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$, that is
$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}.$$
This system is called the *likelihood equations*.

The solution to the likelihood equations need not exist. Sometimes there may be multiple solutions, at most one of which is the MLE. If $I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) > 0$ (the observed information is positive definite at $\widehat{\boldsymbol{\theta}}_n$), we know that $\widehat{\boldsymbol{\theta}}_n$ is at least a local maximum. If $I_n(\boldsymbol{\theta}|\boldsymbol{X}) > 0$ for every $\boldsymbol{\theta} \in \Theta$, the log-likelihood function is concave and the solution to the likelihood equations must be the global maximum and hence the MLE.

In most cases no explicit solution can be found and the MLE must be calculated by numerical methods. There are two commonly used numerical methods for solving the likelihood equations. Let $\widehat{\boldsymbol{\theta}}^{(r)}$ be the r-th iteration to the solution.

- **The Newton-Raphson method:** $\widehat{\boldsymbol{\theta}}^{(r+1)} = \widehat{\boldsymbol{\theta}}^{(r)} + [nI_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})]^{-1} \boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})$
- **The Fisher Scoring method:** $\widehat{\boldsymbol{\theta}}^{(r+1)} = \widehat{\boldsymbol{\theta}}^{(r)} + [nI(\widehat{\boldsymbol{\theta}}^{(r)})]^{-1} \boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})$

They are iterated until the change in $\widehat{\boldsymbol{\theta}}$ from one iteration to the next is sufficiently small or until $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}})$ is sufficiently close to $\mathbf{0}$. The only difference between the two methods is in the information matrix: N-R uses the observed information, FS uses the expected information.

Both require setting $\widehat{\boldsymbol{\theta}}^{(1)}$, the starting value for numerical approximation, and are sensitive to its choice.

A.1.3. Properties of the maximum likelihood estimator

Maximum likelihood estimators are consistent and asymptotically normal as long as so called *regularity conditions* are satisfied.

Conditions (Regularity conditions for maximum likelihood estimators).

R1. The number of parameters d in the model $\mathscr{F}$ is constant.

R2. The support set $S = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on the parameter $\boldsymbol{\theta}$.

R3. The parameter space Θ is an open set.

R4. The density $f(x|\boldsymbol{\theta})$ is sufficiently smooth function of $\boldsymbol{\theta}$ (at least twice continuously differentiable).

R5. The Fisher information matrix $I(\boldsymbol{\theta})$ is finite, regular, and positive definite in a neighborhood of $\boldsymbol{\theta}_X$.

R6. The order of differentiation and integration can be interchanged in expressions such as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int h(x, \boldsymbol{\theta})\, d\mu(x) = \int \frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})\, d\mu(x),$$

where $h(x, \boldsymbol{\theta})$ is either $f(x|\boldsymbol{\theta})$ or $\partial f(x|\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

Note. Take the identity

$$\int_{-\infty}^{\infty} f(x|\boldsymbol{\theta})\, d\mu(x) = 1$$

and differentiate both sides of the equation twice with respect to $\boldsymbol{\theta}$. Regularity condition R6 implies

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} f(x|\boldsymbol{\theta})\, d\mu(x) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}} f(x|\boldsymbol{\theta})\, d\mu(x) = \mathbf{0}. \tag{A.1}$$

Theorem A.2 (consistency of the MLE). *Let conditions R1–R6 hold. Then there exists n_0 and a sequence $\widehat{\boldsymbol{\theta}}_n$ ($n \geq n_0$) of solutions to the likelihood equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$ such that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_X$.* ♢

Note. If the log-likelihood is strictly concave, the likelihood equations have a unique solution, which is the MLE. It converges in probability to the true parameter. If the log-likelihood is not strictly concave, the likelihood equations may have multiple solutions representing local maxima and minima of the log-likelihood. There is one solution among them (the closest to $\boldsymbol{\theta}_X$), which provides a consistence sequence of estimators. Other solutions may not be close to $\boldsymbol{\theta}_X$ and may not converge to it.

Note. If there exists a sequence $\widetilde{\boldsymbol{\theta}}_n$ of other estimators that are guaranteed to be consistent (for example, moment estimators of $\boldsymbol{\theta}_X$), a consistent MLE can be obtained by taking the root of the likelihood equations, which is closest to $\widetilde{\boldsymbol{\theta}}_n$. Alternatively, one can perform one step of the Newton-Raphson algorithm with $\widetilde{\boldsymbol{\theta}}_n$ as the starting value.

Theorem A.3 (Score function properties). *Let conditions R1–R6 hold. Then*

(i) $E\,\boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = 0$, $var\,\boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = I(\boldsymbol{\theta}_X)$.

(ii) $\frac{1}{\sqrt{n}}\boldsymbol{U}_n(\boldsymbol{\theta}_X|\boldsymbol{X}) \xrightarrow{D} N_d(\mathbf{0}, I(\boldsymbol{\theta}_X))$. ♢

Note. The Fisher information matrix at $\boldsymbol{\theta}_X$ can be calculated in two different ways: from Definition A.4 (the expectation of minus the second derivative of the log density) or from Theorem A.3 (the score function variance).

Theorem A.4 (asymptotic normality of the MLE). *Suppose conditions R1–R6 hold. Let* $\widehat{\boldsymbol{\theta}}_n$ *be a consistent sequence of solutions to the likelihood equations. Then*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \xrightarrow{D} N_d(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_X)).$$ ♢

Note.

- The asymptotic variance of the MLE is equal to the inverse of the Fisher information. More information means better precision for estimation.
- The asymptotic variance of the MLE is in a certain sense optimal. Other estimators (e.g., moment estimators) cannot have a smaller asymptotic variance.

Theorem A.5 (asymptotic distribution of the likelihood ratio). *Suppose conditions R1–R6 hold. Let* $\widehat{\boldsymbol{\theta}}_n$ *be a consistent sequence of solutions to the likelihood equations. Then*

$$2\log\frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_X)} = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_X)) \xrightarrow{D} \chi^2_d.$$ ♢

Theorem A.6 (the Δ method for the MLE). *Suppose conditions R1–R6 hold. Let* $\widehat{\boldsymbol{\theta}}_n$ *be a consistent sequence of solutions to the likelihood equations. Take* $q : \Theta \to \mathbb{R}^k$ *a continuously differentiable function. Denote* $\boldsymbol{\nu}_X = q(\boldsymbol{\theta}_X)$ *a* $D(\boldsymbol{\theta}) = \partial q(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. *Then* $\widehat{\boldsymbol{\nu}}_n = q(\widehat{\boldsymbol{\theta}}_n)$ *is the MLE of the parameter* $\boldsymbol{\nu}_X$ *and*

$$\sqrt{n}(\widehat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_X) \xrightarrow{D} N_k(\mathbf{0}, D(\boldsymbol{\theta}_X)I^{-1}(\boldsymbol{\theta}_X)D(\boldsymbol{\theta}_X)^T).$$ ♢

A.1.4. Tests based on maximum likelihood theory

The theory of the MLE can be used to derive tests of simple and composite hypotheses about the parameter $\boldsymbol{\theta}_X$.

Testing of simple hypotheses

We want to test the null hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 \in \Theta$. It is a simple hypothesis because there is just a single distribution in the model $\mathscr{F}$ with the density $f(x|\boldsymbol{\theta}_0)$.

We will introduce three different test statistics for testing H_0.

Definition A.5.

(i) The statistic
$$\lambda_n = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_0)}$$
is called *the likelihood ratio*.

(ii) The statistic
$$W_n = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^{\mathsf{T}} \widehat{I}_n(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$
is called *the Wald statistic*.

(iii) The statistic
$$R_n = \frac{1}{n} \boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})^{\mathsf{T}} \widehat{I}_n^{-1}(\boldsymbol{\theta}_0) \boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$$
is called *the Rao (score) statistic*. ∇

Note. The symbol $\widehat{I}_n$ denotes any consistent estimator of the Fisher information matrix. Three different estimators can be used in Wald and Rao statistics:

1. $\widehat{I}_n(\boldsymbol{\theta}) = I_n(\boldsymbol{\theta}|\boldsymbol{X}) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^{\mathsf{T}}} \log f(\boldsymbol{\theta}|X_i)$ (the observed information matrix)
2. $\widehat{I}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}|X_i)^{\otimes 2}$ (the empirical variance of the score function)
3. $\widehat{I}_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ (the Fisher information matrix)

The most common choice for the Wald statistic is $\widehat{I}_n(\widehat{\boldsymbol{\theta}}_n) = I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X})$. The most common choice for the Rao statistic is $\widehat{I}_n(\boldsymbol{\theta}_0) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}_0|X_i)^{\otimes 2}$.

Note.

- The likelihood ratio requires the calculation of $\widehat{\boldsymbol{\theta}}_n$ and L_n or ℓ_n. It does not require the calculation of $\boldsymbol{U}_n$ and $\widehat{I}_n$.
- The Wald statistic requires the calculation of $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{I}_n$. It does not require the calculation of L_n and $\boldsymbol{U}_n$.
- Rao statistic requires the calculation of $\boldsymbol{U}_n$ a $\widehat{I}_n$. It does not require the calculation of $\widehat{\boldsymbol{\theta}}_n$ and L_n.

Note. If $d = 1$ (one parameter) and $\theta_0 = 0$, then the Wald statistic can be written as
$$W_n = \left[\frac{\widehat{\theta}_n}{\sqrt{n^{-1} \widehat{I}_n^{-1}(\widehat{\theta}_n)}} \right]^2,$$

where $n^{-1}\widehat{I}_n^{-1}(\widehat{\theta}_n)$ is the estimator of the asymptotic variance of $\widehat{\theta}_n$.

Theorem A.7. *Suppose conditions R1–R6 are satisfied. Let the hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ hold. Then:*

(i)

$$2\log\lambda_n = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)) \xrightarrow{D} \chi^2_d$$

(ii)

$$W_n \xrightarrow{D} \chi^2_d$$

(iii)

$$R_n \xrightarrow{D} \chi^2_d \qquad \diamondsuit$$

Note. If H_0 holds, $\widehat{\boldsymbol{\theta}}_n$ should be close to $\boldsymbol{\theta}_0$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ should be close to $L_n(\boldsymbol{\theta}_0)$, and $\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$ should be close to $\mathbf{0}$. Under H_0, all three test statistics have values close to 0. Their large values testify against H_0.

Corollary. Denote by $\chi^2_d(1-\alpha)$ the $(1-\alpha)$-quantile of χ^2_d distribution. Consider tests of $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ defined by the rule: reject H_0 in favor of H_1, if

(i) $2\log\lambda_n \geq \chi^2_d(1-\alpha)$ *(likelihood ratio test)*
(ii) $W_n \geq \chi^2_d(1-\alpha)$ *(Wald test)*
(iii) $R_n \geq \chi^2_d(1-\alpha)$ *(score test)*

Each of these tests has asymptotically (for $n \to \infty$) the level α.

Note. It can be shown that these three tests are asymptotically equivalent. For large sample sizes, their results are almost identical. With smaller sample sizes, their results can differ. Investigations of small sample behavior of these test statistics revealed that the likelihood ratio test has the best properties, the Wald test is the worst of the three.

Thus, in practical applications, the likelihood ratio test should be preferred.

Note. Under normality, the three test statistics are identical.

Estimation in the presence of nuisance parameters and testing of composite hypotheses

It is frequently desirable to estimate and test just a small number of parameters in a model that contains a much larger number of parameters. We divide the parameter vector into two subsets: the parameters of interest and the other paramaters – *nuisance parameters*.

Let $\boldsymbol{\theta}$ be divided into $\boldsymbol{\theta}_A$ containing the first m components of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_B$ containing the remaining $d-m$ components of $\boldsymbol{\theta}$. We have

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)^\mathsf{T} = (\theta_1, \ldots, \theta_m, \theta_{m+1}, \ldots, \theta_d)^\mathsf{T}$$

We want to test the hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$. We want to know whether the first m components of $\boldsymbol{\theta}_X$ are equal to the vector of constants $\boldsymbol{\theta}_{A0}$ regardless of the other $d-m$ components of $\boldsymbol{\theta}_X$.

This is not a simple null hypothesis because there are many distributions in the model $\mathscr{F}$ that satisfy H_0^*.

All the vectors and matrices appearing in the notation of maximum likelihood estimation theory are decomposed into the first m components (part A) and the remaining $d-m$ components (part B). For example,

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{An} \\ \widehat{\boldsymbol{\theta}}_{Bn} \end{pmatrix}, \qquad \boldsymbol{U}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_{An}(\boldsymbol{\theta}) \\ \boldsymbol{U}_{Bn}(\boldsymbol{\theta}) \end{pmatrix}, \qquad I(\boldsymbol{\theta}) = \begin{pmatrix} I_{AA}(\boldsymbol{\theta}) & I_{AB}(\boldsymbol{\theta}) \\ I_{BA}(\boldsymbol{\theta}) & I_{BB}(\boldsymbol{\theta}) \end{pmatrix}, \quad \text{etc.}$$

The following lemma is useful for inverting the decomposed information matrix.

Lemma A.8 (Block matrix inversion). *Let the matrix*

$$I = \begin{pmatrix} I_{AA} & I_{AB} \\ I_{BA} & I_{BB} \end{pmatrix}$$

be of full rank. Then there exists an inverse matrix to I and it can be expressed as

$$I^{-1} = \begin{pmatrix} I^{AA} & I^{AB} \\ I^{BA} & I^{BB} \end{pmatrix},$$

where

$$\begin{aligned} I^{AA} &= I_{AA.B}^{-1}, \\ I^{AB} &= -I_{AA.B}^{-1} I_{AB} I_{BB}^{-1}, \\ I^{BA} &= -I_{BB.A}^{-1} I_{BA} I_{AA}^{-1}, \\ I^{BB} &= I_{BB.A}^{-1}, \\ I_{AA.B} &= I_{AA} - I_{AB} I_{BB}^{-1} I_{BA} \\ I_{BB.A} &= I_{BB} - I_{BA} I_{AA}^{-1} I_{AB}. \end{aligned}$$ ♢

If the null hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ holds we know that $\boldsymbol{\theta}_{AX} = \boldsymbol{\theta}_{A0}$, but we do not know the value of $\boldsymbol{\theta}_{BX}$. We can estimate $\boldsymbol{\theta}_{BX}$ by the maximum likelihood method applied to the nested submodel

$$\mathscr{F}_0 = \{\text{distributions with density } f(x|(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)),\ \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0},\ \boldsymbol{\theta}_B \in \Theta_B \subseteq \mathbb{R}^{d-m}\},$$

with $d-m$ unknown parameters.

Denote the maximum likelihood estimator of $\boldsymbol{\theta}_X$ in the submodel $\mathscr{F}_0$ by $\widetilde{\boldsymbol{\theta}}_n = \begin{pmatrix} \widetilde{\boldsymbol{\theta}}_{An} \\ \widetilde{\boldsymbol{\theta}}_{Bn} \end{pmatrix}$, where $\widetilde{\boldsymbol{\theta}}_{An} = \boldsymbol{\theta}_{A0}$ and $\widetilde{\boldsymbol{\theta}}_{Bn}$ solves the system of likelihood equations

$$\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_{A0}, \widetilde{\boldsymbol{\theta}}_{Bn}) = \mathbf{0}.$$

The Fisher information matrix for $\boldsymbol{\theta}_B$ in this model is $I_{BB}(\boldsymbol{\theta}_X)$.

By Theorems A.3 and A.4 applied to the submodel $\mathscr{F}_0$, we get

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

and

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}^{-1}(\boldsymbol{\theta}_X)).$$

On the other hand, Theorems A.3 and A.4 and Lemma A.8 applied to the larger model imply

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB.A}^{-1}(\boldsymbol{\theta}_X)),$$

where (dropping the arguments $\boldsymbol{\theta}_X$)

$$I_{BB.A}^{-1} = (I_{BB} - I_{BA}I_{AA}^{-1}I_{AB})^{-1} \geq I_{BB}^{-1}.$$

Thus, the asymptotic variance of the MLE of the parameter $\boldsymbol{\theta}_{BX}$ depends on whether or not $\boldsymbol{\theta}_{AX}$ is known. If $\boldsymbol{\theta}_{AX}$ is known (which is true if H_0^* holds), the asymptotic variance of the MLE $\widetilde{\boldsymbol{\theta}}_{Bn}$ is generally larger than the asymptotic variance of the MLE $\widehat{\boldsymbol{\theta}}_{Bn}$ that does not assume a known $\boldsymbol{\theta}_{AX}$.

However, when $I_{BA} = 0$ (the estimators of $\boldsymbol{\theta}_{AX}$ and $\boldsymbol{\theta}_{BX}$ are asymptotically independent), then the asymptotic variances of $\widetilde{\boldsymbol{\theta}}_{Bn}$ and $\widehat{\boldsymbol{\theta}}_{Bn}$ are the same. Then it does not matter whether or not $\boldsymbol{\theta}_{AX}$ is known.

Let us generalize the three test statistics introduced in Definition A.5 of the previous section to testing the composite hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$.

Definition A.6.

(i) The statistic

$$\lambda_n^* = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\widetilde{\boldsymbol{\theta}}_n)}$$

is called *the likelihood ratio*.

(ii) The statistic
$$W_n^* = n(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})^\top \widehat{I}_{AA.B}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})$$
is called *the Wald statistic*.

(iii) The statistic
$$R_n^* = \frac{1}{n}\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)^\top \widehat{I}_n^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$$
is called *the Rao (score) statistic*. ∇

Note.

- Obviously, $\lambda_n^* \geq 1$.
- The expression $\widehat{I}_{AA.B}$ in the Wald statistic means the inverse of the upper left block of the the matrix $\widehat{I}_n^{-1}$.
- Since $\boldsymbol{U}_{Bn}(\widetilde{\boldsymbol{\theta}}_n) = \mathbf{0}$, the Rao statistic can be written as
$$R_n^* = \frac{1}{n}\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n)^\top \widehat{I}_{AA.B}^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n).$$
- The Rao statistic does not require the calculation of the MLE $\widehat{\boldsymbol{\theta}}_n$ in the larger model, it only needs the MLE $\widetilde{\boldsymbol{\theta}}_n$ in the submodel. This is often much easier to get.

Theorem A.9. *Let the null hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$, hold. Then*

(i)
$$2\log\lambda_n^* = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\widetilde{\boldsymbol{\theta}}_n)) \xrightarrow{D} \chi_m^2;$$

(ii)
$$W_n^* \xrightarrow{D} \chi_m^2;$$

(iii)
$$R_n^* \xrightarrow{D} \chi_m^2.$$ ♢

Note. Under H_0^*, we expect $\widehat{\boldsymbol{\theta}}_n$ to be close to $\widetilde{\boldsymbol{\theta}}_n$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ to be close to $L_n(\widetilde{\boldsymbol{\theta}}_n)$, and $\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$ to be close to $\mathbf{0}$. The large values of the three test statistics testify against the null hypothesis.

Corollary. Let $\chi_m^2(1-\alpha)$ be $(1-\alpha)$-quantile of the χ_m^2 distribution. Consider tests of $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$, against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$ given by the rule: reject H_0^* in favor of H_1^* if

(i) $2\log\lambda_n^* \geq \chi_m^2(1-\alpha)$ *(the likelihood ratio test)*
(ii) $W_n^* \geq \chi_m^2(1-\alpha)$ *(the Wald test)*

(iii) $R_n^* \geq \chi^2_m(1-\alpha)$ *(the score test)*

Then each of these three tests has asymptotically (for $n \to \infty$) the level α.

Note. The number of degrees of freedom in the reference χ^2_m distribution is equal to the number of tested parameters.

Note. These three tests are asymptotically equivalent under the null hypothesis as well as under local alternatives. With small or moderate sample sizes, the likelihood ratio test has the best properties and the Wald test is the worst of the three. In practical applications, the likelihood ratio test should be preferred.

Note. Let $m = 1$, $\boldsymbol{\theta}_{AX} = \theta_{Xj}$, and $\boldsymbol{\theta}_{A0} = 0$. Consider the test of the hypothesis $H_0^* : \theta_{Xj} = 0$ against $H_1^* : \theta_{Xj} \neq 0$ (zero value of the j-th parameter in the presence of other parameters that are unspecified by the hypothesis). Then the Wald statistic can be written as

$$W_n = \left[\frac{\widehat{\theta}_{jn}}{\sqrt{n^{-1}\widehat{I}^{-1}_{jj}}} \right]^2,$$

where $n^{-1}\widehat{I}^{-1}_{jj}$ is the estimator of the asymptotic variance of $\widehat{\theta}_{jn}$. This is the square of the test statistic that statistical software typicaly evaluates to test zero value of a single model parameter.

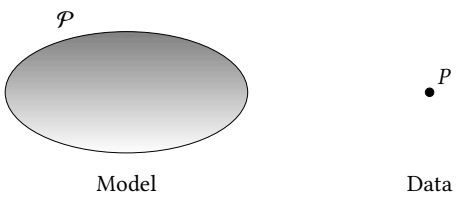
A.2. Behavior of the MLE under a misspecified model

The question is: what happens with the maximum likelihood estimator when the distribution of the data is not correctly determined (the data we work with have a different density). The results we summarize here are taken from White (1982).

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.

So far everything is arranged as if we were doing maximum likelihood theory. However, we do not assume that $P \in \mathcal{P}$. The data we observe need not satisfy the model, there need not exist any $\boldsymbol{\theta} \in \Theta$ such that $P = P_{\boldsymbol{\theta}}$.



We use the data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$ and the model $\mathscr{P}$ to estimate $\boldsymbol{\theta}$ by the method of maximum likelihood. Define the likelihood, MLE, the score function and the score statistic as usual.

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

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 A.10. *(White 1982)* $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $n \to \infty$, *where*

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

and

$$K(P, P_{\boldsymbol{\theta}}) = E_P \log \frac{p(X_i)}{p_{\boldsymbol{\theta}}(X_i)} \geq 0. \qquad \diamondsuit$$

Here, $K(P, P_{\boldsymbol{\theta}})$ is called *the Kullback-Leibler distance* between the distributions P and $P_{\boldsymbol{\theta}}$. It has all the properties of a distance measure except that it is not symmetric, $K(P, P_{\boldsymbol{\theta}}) \neq K(P_{\boldsymbol{\theta}}, P)$.*

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

So, the pseudo-MLE converges somewhere after all but what is it? What is the pseudo-MLE actually estimating? The answer is provided by the first point of this theorem:

Theorem A.11. *(White 1982)*

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

(ii)

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

where

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

* Therefore it is sometimes called *the Kullback-Leibler divergence* instead of *distance*.

(iii)

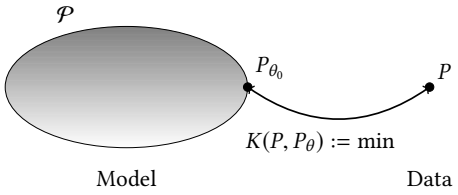
$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{D} N_d(\mathbf{0}, \mathbb{D}^{-1}\Sigma\mathbb{D}^{-1}),$$

where

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

♢

Thus, the incorrectly applied MLE procedure tries to estimate the point $\boldsymbol{\theta}_0$, which solves the set of equations $\mathsf{E}_P \boldsymbol{U}_n(\boldsymbol{\theta}_0) = \mathbf{0}$ (there may be multiple solutions, of course – only one of them is the point of convergence). This identifies the distribution that is "closest" to P within the model, in the sense of the Kullback-Leibler distance. This is illustrated by the following picture.



The other parts of the theorem show that the pseudo-score statistic (when evaluated at $\boldsymbol{\theta}_0$) and the estimator are asymptotically normal.

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 (in general) not equal to the asymptotic variance Σ of the pseudo-score statistic.

Theorem A.12. *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*

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

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.

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 $\mathbb{D} \equiv \Sigma$.

[*] Česky *sendvičový odhad*

- 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.

Bibliography

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.

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.

Index