NMST432 Advanced Regression Models

# Extended Course Notes

Michal Kulich

Last modified on March 15, 2021.


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

*These course notes contain the whole contents of the course "NMST432 Advanced regression models", which is a part of the curriculum of the Master's program "Probability, Mathematical Statistics and Econometrics".*

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

Michal Kulich
kulich@karlin.mff.cuni.cz

In Karlín on March 15, 2021

# 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 $\varepsilon_i$ are called *error terms (disturbances)*[¶], $\sigma^2$ 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 $\mathsf{E}$, $\mathsf{var}$ for the conditional expectation and variance, respectively, given $\boldsymbol{X}_i$. The symbol $\mathsf{E}_X$ 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 $\beta_{01}$ (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.

*The end of lecture 1 (Mar. 1)*

# 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 $\sigma$-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

- $\theta$ 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 $\chi^2$ 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 $\sigma$-finite measure $\nu$ 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 $\mu$ 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**. $\nabla$

**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, $\theta_0$ is the true canonical parameter, $\varphi_0$ 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 $\theta$ 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. 35.

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 $\varphi_0$ is known. Then the score function for $\theta$ is

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

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

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

The maximum likelihood estimator [MLE] $\widehat{\theta}_n$ solves the equation $U_n(\widehat{\theta}_n \mid \boldsymbol{Y}) = 0$, that is $\sum_{i=1}^{n} Y_i = nb'(\widehat{\theta}_n)$. 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 $\varphi_0$. It can be calculated even if $\varphi_0$ 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 $\varphi_0$ unknown. The MLE of $\theta_0$ is still the same, $\widehat{\theta}_n = (b')^{-1}(\overline{Y}_n)$. However, what is the asymptotic distribution of $\widehat{\theta}_n$ when $\varphi_0$ is unknown? In general, the asymptotic variance may change.

Calculate the joint information matrix for $(\theta,\varphi)$:

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

Thus, the information matrix is diagonal. It follows that the asymptotic distribution of $\widehat{\theta}_n$ is given by (2.2) even if $\varphi_0$ is unknown.

We do not need $\varphi_0$ to estimate $\theta_0$ but we need an estimate of $\varphi_0$ to estimate the asymptotic variance of $\theta_0$. Of course, we could use the MLE of $\varphi_0$ 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).

*The end of lecture 2 (Mar. 4)*

## 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, $\theta_i$ depends on $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$, $\varphi > 0$ is a known or an unknown constant.
3. $\theta_i$ depends on $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$ through the *linear predictor*† $\eta_i \overset{\mathsf{df}}{=} \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.
4. There exists a known strictly monotone, twice continuously differentiable *link function*‡ $g$ such that $g(\mu_i) = \eta_i$. ∇

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

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

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

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

**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 $\varphi$ 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 $\mu_i$ and the linear predictor $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$. 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 **??**). 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 $\eta_i$ with the canonical parameter $\theta_i$. ∇

***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 $\eta_i$: $g(\mu_i) = \eta_i$. 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 $g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta}$.

We know that

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

[*] Česky *kanonický link*

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 $\mu_i$ 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 $\lambda_i$ 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 $g(\mu_i) = \boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}$.

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 $\mathsf{E}\, Y_i = g^{-1}(\eta_i) = \frac{\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}}$. 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 $\mu_i = \exp\{\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\beta}\}$, 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 $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$. However, we are also interested in parametrizing the distributions of the individual $Y_i$'s that

depend on both the primary parameters $\boldsymbol{\beta}$ and the covariates $\boldsymbol{X}_i$. This can be done in three ways:

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

The parametrizations are related to each other as follows:

- $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$;
- $\eta_i = g(\mu_i)$, $\mu_i = g^{-1}(\eta_i)$;
- $\mu_i = b'(\theta_i)$, $\theta_i = (b')^{-1}(\mu_i)$;
- $\eta_i = g\big(b'(\theta_i)\big)$, $\theta_i = (b')^{-1}\big(g^{-1}(\eta_i)\big)$; if the link $g$ is canonical then $\eta_i = \theta_i$.

*The end of lecture 3 (Mar. 4)*

### The likelihood function

Let the true dispersion parameter $\varphi_0$ 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 $\theta_i = (b')^{-1}\big(g^{-1}(\boldsymbol{X}_i^\top \boldsymbol{\beta})\big)$.

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 $\theta_i$, which is unrelated to the canonical parameters of the other observations. Maximizing $L(\boldsymbol{\beta} \mid \boldsymbol{Y})$ w.r.t all $\boldsymbol{\beta} \in \mathbb{R}^n$ is the same as maximizing $L(\boldsymbol{\theta} \mid \boldsymbol{Y})$ w.r.t all $\boldsymbol{\theta} \in \mathbb{R}^n$. To obtain the MLE in the saturated model, we differentiate (2.3) w.r.t. each $\theta_i$ separately and we get $n$ equations

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

The MLE of $\mu_i$ under the saturated model is

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

[*] Česky *saturovaný model*

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 $\theta$ 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 $\beta_0$ as $\widehat{\beta}_n = g(b'(\widehat{\theta}_n)) = g(\overline{Y}_n)$. From the central limit theorem for iid random variables and the delta method,

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

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

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

## 2.3. Maximum Likelihood Estimation in the GLM

Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$ be iid random vectors of dimension $p+1$. Let $h_i(\boldsymbol{x})$ be the marginal density of $\boldsymbol{X}_i$ (with no assumptions about it except finite second moments). Let $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, satisfy the generalized linear model (Definition 2.3) with true parameters $\boldsymbol{\beta}_0$ and $\varphi_0$. Consider $\varphi_0$ known. Write the conditional density of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$ as $f(y \mid \boldsymbol{x}, \boldsymbol{\beta}_0, \varphi_0)$. Then the joint density of $(Y_i, \boldsymbol{X}_i)$ is $f(y \mid \boldsymbol{x}, \boldsymbol{\beta}_0, \varphi_0) h_i(\boldsymbol{x})$, the full likelihood is

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

and the full log-likelihood is

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

[*] Česky *nulový model*

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

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

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

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

* Česky *vyrovnané hodnoty*

Because the conditional expectation (given $\boldsymbol{X}_i$) of $Y_i - \mu_i$ is 0 when $\mu_i$ 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 end of lecture 4 (Mar. 11)*

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

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

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

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

*where*

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

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

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

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

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

*(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

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

where $\hat{\mathbb{W}}$ is the $n \times n$ diagonal matrix $\mathrm{diag}\,(w(\widehat{\mu}_1), \ldots, w(\widehat{\mu}_n))$. When $\varphi_0$ is unknown it is replaced by a consistent estimator $\widehat{\varphi}_n$, which will 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 $\mu_i$. 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

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

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,

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

(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. 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*

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

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

**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.

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

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

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

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

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. □

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

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 $\hat{\mathbb{W}}^{(k)} = \mathrm{diag}\,(w(\widehat{\mu}_1^{(k)}), \ldots, w(\widehat{\mu}_n^{(k)}))$ 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 $\delta$ 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.2, bottom of page 37).
- According to (2.9), the matrix $(\mathbb{X}^\mathsf{T} \hat{\mathbb{W}}^{(k)} \mathbb{X})^{-1}$ estimates (up to the proportionality constant $\varphi_0$) 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.

*The end of lecture 5 (Mar. 11)*

## 2.5. Estimation of the Dispersion Parameter

The dispersion parameter $\varphi_0$ 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 $\varphi_0$. Instead of using the method of maximum likelihood, $\varphi_0$ 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 $\varphi_0$ 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 $\|\partial h/\partial \boldsymbol{\beta}\| \le C(y, \boldsymbol{x})$ in a neighborhood $\mathcal{B}_0$ of $\boldsymbol{\beta}_0$ and $\mathsf{E}\, C(Y_i, \boldsymbol{X}_i)$ exists and is finite. Then $\widehat{\varphi}_n \xrightarrow{P} \varphi_0$.* ◇

**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 $\varphi_0$ 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

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

[*] Česky *Pearsonovo chí kvadrát*

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 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 $\mu_i$ is $Y_i$ (see p. 21) and the MLE of $\theta_i$ 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, $\widetilde{\ell}_n(\boldsymbol{Y}) \ge \ell_n(\widehat{\boldsymbol{\beta}}_n \mid \boldsymbol{Y})$.

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 $\varphi_0$, 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 $\varphi_0$ is unknown, use the moment estimator $\widehat{\varphi}_n$ defined by (2.11).

*The end of lecture 6 (Mar. 18)*

## 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. 35.

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 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)*

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

*(iii)*

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

*(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 Pearson $X^2$ is another generalization.

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 $\mathsf{var}\,\widehat{\boldsymbol{\beta}}_n$ .

Let us consider the problem of testing the simple hypothesis

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

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

***Theorem 2.8.***

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

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

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

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

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

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

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 $m$ colums of the covariate vector can be excluded from the model. The null hypothesis specifies a submodel (with $p-m$ parameters) of the full model with ($p$ parameters).

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

Partition the $p \times p$ matrix $\widehat{\Sigma} = \widehat{I}_n^{-1}/(n\widehat{\varphi}_n) = (\mathbb{X}^\mathsf{T}\widehat{\mathbb{W}}\mathbb{X})^{-1}$ (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*

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

*(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*

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

*(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*

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

**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 $\alpha$ if the chosen test statistic (it must be selected in advance) exceeds the $1-\alpha$ quantile of the $\chi_m^2$ 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 $\chi^2_m$.
- 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 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 $\beta_j$ 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 $\mathbf{0} \neq \boldsymbol{c} \in \mathbb{R}^p$ can be obtained easily from Theorem 2.7 part (iii). An asymptotic confidence interval with coverage probability converging to $1-\alpha$ is

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

*The end of lecture 7 (Mar. 18)*

# A. Appendix: Maximum Likelihood Theory

## A.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 $\sigma$-finite measure $\mu$. 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 $\Theta$ 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 $\sigma$-finite measure $\mu$ (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 $P_\theta$ 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.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
$$I(\boldsymbol{\theta}|X_i) \stackrel{\text{df}}{=} -\frac{\partial}{\partial \boldsymbol{\theta}^\mathsf{T}} \boldsymbol{U}(\boldsymbol{\theta}|X_i) = -\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} \log f(X_i|\boldsymbol{\theta})$$
is called the contribution of the $i$-th observation to the information matrix.
- The random matrix
$$I_n(\boldsymbol{\theta}|\boldsymbol{X}) \stackrel{\text{df}}{=} -\frac{1}{n} \frac{\partial}{\partial \boldsymbol{\theta}^\mathsf{T}} \boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^{n} I(\boldsymbol{\theta}|X_i)$$
is called *the observed information matrix*.
- The matrix
$$I(\boldsymbol{\theta}) \stackrel{\text{df}}{=} \mathsf{E}\, I(\boldsymbol{\theta}|X_i) = -\mathsf{E}\, \frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\mathsf{T}} \log f(X_i|\boldsymbol{\theta})$$
is called *the expected (Fisher) information matrix*. ∇

If the set $\Theta$ 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.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 $\Theta$ 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.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 $\boldsymbol{\theta}_X$.

### A.4.1. 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 $\ell_n$. 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 $\chi^2_d$ 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 $\alpha$.

**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.

### A.4.2. 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 **??** 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
$$R_n^* = \frac{1}{n}\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n)^\top \widehat{I}_{AA.B}^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n).$$
- Theh 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 $\chi_m^2$ 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 $\alpha$.

**Note.** The number of degrees of freedom in the reference $\chi^2_m$ 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.

# Bibliography

Nelder, J. and Wedderburn, R. (1972). Generalized linear models, *Journal of the Royal Statistical Society. Series A* **135**(3): 370–384.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* **48**(4): 817–838.

# Index