# Part II: Generalized Linear Models

1. Preliminaries
2. Exponential Dispersion Family of Distributions
3. Generalized Linear Models (GLMs)
4. Logistic Regression
5. Poisson Regression
6. Log-linear Models
7. Regularized GLMs

# GLMs...

**In the second part, we consider...**

**generalized linear models**,

$$g[E(\mathbf{Y})] = \mathbf{X}\beta$$

with

- *random component*: $\mathbf{Y} = (Y_1, \ldots, Y_n)'$
- *linear predictor*: $\mathbf{X}\beta$,
  $\mathbf{X}$ $n \times p$ model matrix
  $\beta = (\beta_1, \ldots, \beta_p)'$ parameter vector
- *link function*: $g$
  (relates $E(\mathbf{Y})$ to the linear predictor).

- GLMs extend LMs to embrace *non-normal response distributions* and possibly *nonlinear functions* for the mean response.

# Part II: Generalized Linear Models

## Chapter II.1

## Preliminaries

Notation, Linear Algebra & Probability, Likelihood, Linear Models

# Topics

**To be discussed/refreshed...**

- properties of vectors & matrices (see Chapter I.1: Linear Algebra)
- random vectors, expectations, covariance matrix (see Chapter I.1: Probability)
- selected probability distributions (see also Chapter I.1: Probability)
- likelihood & basic results useful for statistical inference
- linear models (Part I)

# Notation & basic definitions

## II.1.1 Notation (vectors and matrices)

- $\mathbb{R}^p$: $p$-dimensional Euclidean space
- $\mathbb{R}^{p \times q}$: set of all $(p \times q)$-matrices
- vectors are written in bold italics: $\boldsymbol{x} = (x_i)_{1 \leqslant i \leqslant p} = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix}$
- random vectors are written in capital bold italics: $\boldsymbol{X} = (X_i)_{1 \leqslant i \leqslant p} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$
- matrices are written in bold capitals: $\mathbf{A} = (a_{ij})_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant q} = \begin{pmatrix} a_{11} & \cdots & a_{1q} \\ \vdots & \cdots & \vdots \\ a_{p1} & \cdots & a_{pq} \end{pmatrix}$

# Notation & basic definitions

- matrices of higher dimension are written analogously: $\mathbf{B} = (b_{ijk})_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant q, 1 \leqslant k \leqslant r}$, etc.
- sums of entries of a matrix over one (or more) dimensions are denoted by replacing the corresponding indicator(s) through '$+$':

$$a_{i+} = \sum_{j=1}^{q} a_{ij} \ , \quad b_{++k} = \sum_{i=1}^{p} \sum_{j=1}^{q} b_{ijk}$$

# Probability distributions for continuous random variables

## II.1.2 Remark (probability density functions (pdf) of distributions on $\mathbb{R}$)

- Normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu \in \mathbb{R}$, $\sigma > 0$:

$$f(y;\mu,\sigma^2) = \varphi_{\mu,\sigma^2}(y) = \frac{1}{\sqrt{2\pi}\,\sigma}\ \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad y \in \mathbb{R}; \quad \varphi_{0,1} = \varphi$$

- $\chi^2$-distribution $\chi^2(p)$ with $p \in \mathbb{N}$ degrees of freedom:

$$f(y;p) = \frac{1}{2^{p/2}\Gamma(p/2)}\, y^{p/2-1}\, e^{-y/2}, \quad y > 0$$

- Exponential distribution $\operatorname{Exp}(\lambda)$ with parameter $\lambda > 0$:

$$f(y;\lambda) = \lambda e^{-\lambda y}\ , \quad y > 0$$

- Gamma distribution $\mathcal{G}(\alpha,\beta)$ with parameters $\alpha > 0$, $\beta > 0$:

$$f(y;\alpha,\beta) = \frac{\alpha^\beta}{\Gamma(\beta)} y^{\beta-1} e^{-\alpha y}\ , \quad y > 0$$

  - $\beta = 1$: Exponential distribution $(\alpha = \lambda)$

# Probability distributions for discrete random variables

### II.1.3 Remark (probability mass functions (pmf) of distributions on $\mathbb{R}$)

- Bernoulli distribution $\mathcal{B}(1,\pi)$ with parameter $\pi \in [0,1]$:

$$p_y = f(y;\pi) = \pi^y (1-\pi)^{1-y}, \quad y \in \{0,1\}$$

- Binomial distribution $\mathcal{B}(n,\pi)$, with $n \in \mathbb{N}$ and parameter $\pi \in [0,1]$.

$$p_y = f(y;n,\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad 0 \leqslant y \leqslant n,\ y \in \mathbb{N}_0$$

- Poisson distribution $\mathcal{P}(\mu)$ with parameter $\lambda > 0$:

lambda, not mu

$$p_y = f(y;\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y \in \mathbb{N}_0$$

- Negative Binomial distribution $\mathcal{NB}(\mu,k)$ with parameters $\mu > 0$ and $k > 0$:

$$p_y = f(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)} \left(\frac{\mu}{\mu+k}\right)^y \left(\frac{k}{\mu+k}\right)^k, \quad y \in \mathbb{N}_0$$

# Probability distributions for discrete random vectors

Seems somehow similar to binomial; however, multinomial distribution can have multiple outcomes (binomial has only „Success“ or „Failure“)

## II.1.4 Remark (probability mass function (pmf) of a distribution on $\mathbb{R}^m$, $m > 1$)

- Multinomial distribution $\mathcal{M}(n, \boldsymbol{\pi})$ with $n \in \mathbb{N}$ and parameters $\pi_1, \ldots, \pi_{m+1} \in [0, 1]$ such that $\sum_{j=1}^{m+1} \pi_j = 1$ (i.e. $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{m+1})'$ is a probability vector):

$$p_{\mathbf{y}} = f(y_1, \ldots, y_{m+1}) = \binom{n}{y_1, \ldots, y_{m+1}} \prod_{j=1}^{m+1} \pi_j^{y_j},$$

$$\mathbf{y} = (y_1, \ldots, y_{m+1})' \in \{(i_1, \ldots, i_{m+1})' \in \mathbb{N}_0^m | \sum_{j=1}^{m+1} i_j = n\}$$

  - $\binom{n}{y_1, \ldots, y_{m+1}} = \frac{n!}{y_1! \ldots y_{m+1}!}$ (multinomial coefficient)
  - $m = 1$: Binomial distribution

# Connections of probability distributions

**II.1.5 Proposition**

1. Let $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} \mathcal{B}(1, \pi)$. Then, $\sum_{j=1}^{n} Y_j \overset{\text{iid}}{\sim} \mathcal{B}(n, \pi)$.

2. Let $\mathbf{Y} \sim \mathcal{M}(n, \boldsymbol{\pi})$. Then,
   - $Y_j \sim \mathcal{B}(n, \pi_j)$, for $j \in \{1, \dots, m+1\}$
   - $\sum_{j=1}^{k} Y_j \sim \mathcal{B}(n, \sum_{j=1}^{k} \pi_j)$, for $k \in \{1, \dots, m\}$
   - $\mathbf{Y}_J = (Y_{J_1}, \dots, Y_{J_k}, n - \sum_{j \in J} Y_{J_j})' \sim \mathcal{M}(n, \boldsymbol{\pi}_J)$,
     with $\boldsymbol{\pi}_J = (\pi_{J_1}, \dots, \pi_{J_k}, 1 - \sum_{j \in J} \pi_{J_j})'$ for $J = \{J_1, \dots, J_k\} \subset \{1, \dots, m+1\}$

3. Let $Y_1, \dots, Y_k$ be independent Poisson random variables with $Y_j \sim \mathcal{P}(\lambda_j)$, $j \in \{1, \dots, k\}$ and consider the random vector $\mathbf{Y} = (Y_1, \dots, Y_k)'$. Then, $\mathbf{Y} \Big| \sum_{j=1}^{k} Y_j = n \sim \mathcal{M}(n, \boldsymbol{\pi})$, where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)'$ with $\pi_j = \frac{\lambda_j}{\sum_{j=1}^{k} \lambda_j}$, i.e the conditional distribution of $\mathbf{Y}$ given $\sum_{j=1}^{k} Y_j = n$ is $\mathcal{M}(n, \boldsymbol{\pi})$.

# Likelihood

**II.1.6 Definition (likelihood function)**

Given an observed sample $\mathbf{y} = (y_1, \ldots, y_n)'$, $n \in \mathbb{N}$, and assuming a statistical model $f_{\mathbf{Y}}(\mathbf{y}; \vartheta)$ depending on an unknown parameter $\vartheta \in \Theta \subseteq \mathbb{R}^p$, the likelihood $L(\vartheta|\mathbf{y})$ is defined as
$L(\vartheta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \vartheta)$.

In case of discrete data, the likelihood $L(\vartheta|\mathbf{y})$ is the probability of the observed data $\mathbf{y}$ under the specific model assumption.

**II.1.7 Definition (likelihood function based on iid random variables)**

Given a realization $\mathbf{y} = (y_1, \ldots, y_n)'$ of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $n \in \mathbb{N}$, if the components of $\mathbf{Y}$ are stochastically independent and identically distributed (iid) having a pdf or pmf $f_{Y_1}$, for continuous or discrete random variables, respectively, i.e. $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\cdot; \vartheta)$, then it holds

$$L(\vartheta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}) = \prod_{i=1}^{n} f_{Y_1}(y_i; \vartheta) \ .$$

# Quantities derived from the likelihood

### II.1.8 Definition (score function)

Given an observed sample $\mathbf{y} \in \mathbb{R}^n$ and the log-likelihood $\ell(\boldsymbol{\vartheta}|\mathbf{y}) = \ln\left(L(\boldsymbol{\vartheta}|\mathbf{y})\right)$, with $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$, the score function is defined as the gradient of the log likelihood

$$S(\boldsymbol{\vartheta}) = S(\boldsymbol{\vartheta}|\mathbf{y}) := \nabla_{\boldsymbol{\vartheta}}\{\ell(\boldsymbol{\vartheta}|\mathbf{y})\} = \left(\frac{\partial \ell(\boldsymbol{\vartheta}|\mathbf{y})}{\partial \vartheta_1}, \dots, \frac{\partial \ell(\boldsymbol{\vartheta}|\mathbf{y})}{\partial \vartheta_p}\right)' .$$

### II.1.9 Remark

In most regular problems (where the likelihood is of quadratic form), the analysis of the likelihood function can focus on the *location* of the maximum and the *curvature* around it.

☞ In such cases, the *maximum likelihood estimate* $\hat{\boldsymbol{\vartheta}}(\mathbf{y})$ is the solution of the *score equation(s)*:

$$S(\boldsymbol{\vartheta}) = \mathbf{0} .$$

The corresponding **maximum likelihood estimator (MLE)** is then $\hat{\boldsymbol{\vartheta}}(\mathbf{Y})$.

## II.1.10 Definition (Fisher Information)

For $\mathbf{Y} \in \mathbb{R}^n$ and under a statistical model $f_{\mathbf{Y}}(\mathbf{Y}; \vartheta)$ with unkown parameter $\vartheta \in \Theta \subset \mathbb{R}$, the (expected) Fisher information $\mathcal{I}_n(\vartheta)$ is defined as

$$\mathcal{I}_n(\vartheta) = \mathsf{E}(I_n(\vartheta)) := \mathsf{E}\left[\left(\frac{\partial \ell(\vartheta)}{\partial \vartheta}\right)^2\right] = \mathsf{E}\left[\left(\frac{\partial \log f_{\mathbf{Y}}(\mathbf{Y}; \vartheta)}{\partial \vartheta}\right)^2\right].$$

Under mild conditions[a] it can equivalently be defined as

$$\mathcal{I}_n(\vartheta) = \mathsf{E}(I_n(\vartheta)) = \mathsf{E}\left[-\frac{\partial^2 \log f^{\mathbf{Y}}(\mathbf{Y}; \vartheta)}{\partial \vartheta^2}\right].$$

[a] see Casela and Berger (2002, Section 7.3)

- The curvature of the loglikelihood at $\hat{\vartheta}$ is $I_n(\hat{\vartheta})$, called the **observed Fisher information** at $\hat{\vartheta}$ $[I_n(\hat{\vartheta}) = I_n(\hat{\vartheta}(\mathbf{y}))]$.

☞ A large curvature is associated with a strong peak, indicating less uncertainty about $\vartheta$.

The Cramer-Rao lower bound gives the minimal possible variance for an estimator and is linked to the Fisher Information.

### II.1.11 Definition (Cramer-Rao Lower Bound)

Under 'certain' regularity conditions[a], the variance of any unbiased estimator $\hat{\vartheta}$ of $\vartheta$ with finite variance satisfies

$$\mathrm{Var}(\hat{\vartheta}) \geqslant \frac{1}{\mathcal{I}_n(\vartheta)} \, .$$

[a] see Casela & Berger (2002, Theorem 7.3.9)

☞ If additional to the assumptions required in II.1.11, it holds $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f_{Y_1}(\cdot;\vartheta)$, then* $\mathrm{Var}(\hat{\vartheta}) \geqslant \frac{1}{n\mathrm{E}(\mathrm{I}(\vartheta))}$, with $\mathrm{E}(\mathrm{I}(\vartheta)) = \mathrm{E}\left(\frac{\partial \log f_{Y_1}(Y;\vartheta)}{\partial \vartheta}\right)^2$ .

### II.1.12 Definition (asymptotically efficient estimator)

A sequence of estimators $(\hat{\vartheta}_n)_n$ is said to be *asymptotically efficient* for a parameter $\vartheta$, if it holds $\sqrt{n}(\hat{\vartheta}_n - \vartheta) \longrightarrow \mathcal{N}(0, \nu(\vartheta))$ in distribution and $\nu(\vartheta) = \frac{1}{\mathcal{I}_n(\vartheta))}$. That is, the asymptotic variance of $\hat{\vartheta}_n$ achieves the Cramer-Rao Lower Bound.

*see Casela & Berger (2002, Corollary 7.3.10)

## II.1.13 Definition (Fisher Information Matrix)

For $\boldsymbol{\vartheta} \in \Theta \subset \mathbb{R}^p$, the (expected) Fisher information matrix $\mathcal{I}_n(\boldsymbol{\vartheta})$ is the $p \times p$ matrix with his elements defined as

$$(\mathcal{I}_n(\boldsymbol{\vartheta}))_{ij} := \mathrm{E}I_n(\boldsymbol{\vartheta}) = \mathrm{E}\left[\left(\frac{\partial \log f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\vartheta})}{\partial \vartheta_i}\right)\left(\frac{\partial \log f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\vartheta})}{\partial \vartheta_j}\right)\right].$$

Under certain regularity conditions, the elements of the Fisher information matrix may also be written as $(\mathcal{I}_n(\boldsymbol{\vartheta}))_{ij} := -\mathrm{E}\left[\left(\frac{\partial^2 \log f_{\mathbf{Y}}(\mathbf{Y};\boldsymbol{\vartheta})}{\partial \vartheta_i \partial \vartheta_j}\right)\right]$.

## II.1.14 Proposition

Under some conditions[a], the score function has the following properties

$$\begin{aligned} \mathrm{E}S(\boldsymbol{\vartheta}) &= 0\,, \\ \mathrm{Cov}S(\boldsymbol{\vartheta}) &= \mathcal{I}_n(\boldsymbol{\vartheta})\,. \end{aligned}$$

[a] see Cassela & Berger (2002, Sections 7.3, 10.3)

## II.1.15 Proposition of MLEs

Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} P_\vartheta$, $\vartheta \in \Theta \subset \mathbb{R}^p$, and $\mathbf{y} = (y_1, \ldots, y_n)'$ be an observed sample, $n \in \mathbb{N}$. Let further $\vartheta_0 \in \Theta$ be the true parameter value and $\widehat{\vartheta}_n = \widehat{\vartheta}(y_1, \ldots, y_n)$ a solution of the likelihood equations (score equations). Then, under 'certain' regularity conditions[a] it holds:

1. $\widehat{\vartheta}_n \xrightarrow{P} \vartheta_0$, $n \to \infty$ (consistent; also strong consistency is possible)
2. $\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) \xrightarrow{P} \mathcal{N}_p(0, \mathcal{J}_0^{-1})$ with $\mathcal{J}_0 = \lim\limits_{n\to\infty} \frac{1}{n}\mathcal{J}_n(\vartheta_0)$.
3. $\lim\limits_{n\to\infty} E_\vartheta \widehat{\vartheta}_n = \vartheta_0$ (asymptotic unbiased)
4. $(\widehat{\vartheta}_n)_n$ asymptotic efficient

[a] Casela & Berger (2002, Section 10.6.2)

## II.1.16 Invariance Principle of MLEs

Let $g : \Theta \longrightarrow \mathbb{R}^k$ and $\widehat{\vartheta}$ the MLE for $\vartheta$. Then $g(\widehat{\vartheta})$ is the MLE for $g(\vartheta)$.