Lecture 2 | 04.03.2024

Multivariate regression model

(likelihood estimation & statistical properties)

Standard notation

- longitudinal observations $\mathcal{D}_L \equiv \{(Y_{ij}, \boldsymbol{X}_{ij}^\top)^\top;\ i = 1, \dots, N; j = 1, \dots, n_i\}$
  - for $N \in \mathbb{N}$ independent subjects observed repeatedly $n_i \in \mathbb{N}$ times
  - for $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, for some $p \in \mathbb{N}$
  - however, the data (random vectors) is $\mathcal{D}_L$ does not form a random sample!

Standard notation

- longitudinal observations $\mathcal{D}_L \equiv \{(Y_{ij}, \boldsymbol{X}_{ij}^\top)^\top;\ i = 1, \dots, N; j = 1, \dots, n_i\}$
  - for $N \in \mathbb{N}$ independent subjects observed repeatedly $n_i \in \mathbb{N}$ times
  - for $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, for some $p \in \mathbb{N}$
  - however, the data (random vectors) is $\mathcal{D}_L$ does not form a random sample!
- independent observations $\mathcal{D}_I \equiv \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top;\ i = 1, \dots, N\}$
  - alternative notation for $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ and $(\boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top \in \mathbb{R}^{p \times n_i}$
  - random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top$ are independent with variable lengths
  - random vectors in $\mathcal{D}_I$ are independent but still not identically distributed!

Standard notation

- longitudinal observations $\mathcal{D}_L \equiv \{(Y_{ij}, \boldsymbol{X}_{ij}^\top)^\top;\ i = 1, \dots, N; j = 1, \dots, n_i\}$
  - for $N \in \mathbb{N}$ independent subjects observed repeatedly $n_i \in \mathbb{N}$ times
  - for $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, for some $p \in \mathbb{N}$
  - however, the data (random vectors) is $\mathcal{D}_L$ does not form a random sample!
- independent observations $\mathcal{D}_I \equiv \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top;\ i = 1, \dots, N\}$
  - alternative notation for $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ and $(\boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top \in \mathbb{R}^{p \times n_i}$
  - random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top$ are independent with variable lengths
  - random vectors in $\mathcal{D}_I$ are independent but still not identically distributed!
- balanced longitudinal profiles $\mathcal{D}_B \equiv \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top;\ i = 1, \dots, N\}$
  - with the same number $n_i = n \in \mathbb{N}$ of repeated observations for $i = 1, \dots, N$
  - random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top$ are independent with the same length
  - however, the vectors in $\mathcal{D}_B$ generally still do not form a random sample!

Standard notation

- longitudinal observations $\mathcal{D}_L \equiv \{(Y_{ij}, \boldsymbol{X}_{ij}^\top)^\top;\ i = 1, \dots, N; j = 1, \dots, n_i\}$
  - for $N \in \mathbb{N}$ independent subjects observed repeatedly $n_i \in \mathbb{N}$ times
  - for $Y_{ij} \in \mathbb{R}$ and $\boldsymbol{X}_{ij} \in \mathbb{R}^p$, for some $p \in \mathbb{N}$
  - however, the data (random vectors) is $\mathcal{D}_L$ does not form a random sample!
- independent observations $\mathcal{D}_I \equiv \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top;\ i = 1, \dots, N\}$
  - alternative notation for $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}$ and $(\boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top \in \mathbb{R}^{p \times n_i}$
  - random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in_i}^\top)^\top$ are independent with variable lengths
  - random vectors in $\mathcal{D}_I$ are independent but still not identically distributed!
- balanced longitudinal profiles $\mathcal{D}_B \equiv \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top;\ i = 1, \dots, N\}$
  - with the same number $n_i = n \in \mathbb{N}$ of repeated observations for $i = 1, \dots, N$
  - random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top$ are independent with the same length
  - however, the vectors in $\mathcal{D}_B$ generally still do not form a random sample!

$\hookrightarrow$ what should be postulated in addition to be able to say that the data in $\mathcal{D}_B$ already form a random sample (independent and identically distributed random vectors)?

The main aims of longitudinal analysis

- Estimation of the cross-sectional dependence structure (between subjects)
- Estimation of the time/spatial dependence structure (within subjects)

- Inference on marginal vs. hierarchical means
- Inference on subject-specific profiles and their developments

The main aims of longitudinal analysis

- ❑ Estimation of the cross-sectional dependence structure (between subjects)
- ❑ Estimation of the time/spatial dependence structure (within subjects)

- ❑ Inference on marginal vs. hierarchical means
- ❑ Inference on subject-specific profiles and their developments

$\hookrightarrow$ the estimation and the following inference can be performed in terms of various characteristics and different inference tools

$\hookrightarrow$ the main interest will be given to the conditional distribution characterized by the conditional expectation in particular

Common approaches to longitudinal data

- **Naive methods**
  The longitudinal structure within a subject is firstly summarized into one (or more) characteristics and independent characteristics are regressed over independent subjects (e.g., separate time points analysis, area under the curve, analysis of endpoints, increments, covariance)

- **Simple methods**
  Marginal models similar to a standard cross-sectional study, however, with an additional assumption on the variance – generally $E\boldsymbol{Y}_i = \boldsymbol{X}_i^\top \beta$ and $Var\,\boldsymbol{Y}_i = \mathbb{V}_i(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} \in \mathbb{R}^q$ and $\beta \in \mathbb{R}^p$ must be estimated

- **Radom effects models**
  Allow for marginal as well as hierarchical interpretation – the regression coefficients may vary across subjects (modifications due to random effects) and the models apriori assume a specific correlation structure among repeated observations within the subjects

- **Transition models**
  Modelling the conditional expectation of Y_{ij} given past observations within the same subject and the explanatory variables X_{ij}

More general: multivariate regression

- multivariate linear regression as an extension of ordinary linear regression *(multivariate linear regression vs. multiple (multi-variable) regression)*
- general model formulation for $\mathbb{Y} \in \mathbb{R}^{N \times n}$ and $\mathbb{X} \in \mathbb{R}^{N \times p}$ *($N \in \mathbb{N}$ and $n \in \mathbb{N}$ repeated measurements within each subject)*

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U} \qquad \Big| \qquad Y_{ij} = \boldsymbol{X}_i^\top \boldsymbol{\beta}_j + \varepsilon_{ij}$$

where $\mathbb{Y} = (Y_{ij})_{i,j=1}^{N,n}$, $\mathbb{X} = (X_{ij})_{i,j=1}^{N,p}$, $\mathbb{B} = (\beta_{ij})_{i,j=1}^{p,n}$, and $\mathbb{U} = (\varepsilon_{ij})_{i,j=1}^{N,n}$

- What are the corresponding data (let's denote the data as $\mathcal{D}_S$)? *(recall, that the vector of the explanatory covariates is subject specific)*
- What is the meaning of the formulae above? *(note, that the time dependence is only reflected within $\boldsymbol{\beta}_j \in \mathbb{R}^p$)*
- What are the objects appearing in the expression?
- What are typical assumptions for such linear model?

More general: multivariate regression

- multivariate linear regression as an extension of ordinary linear regression
  (multivariate linear regression vs. multiple (multi-variable) regression)
- general model formulation for $\mathbb{Y} \in \mathbb{R}^{N \times n}$ and $\mathbb{X} \in \mathbb{R}^{N \times p}$
  ($N \in \mathbb{N}$ and $n \in \mathbb{N}$ repeated measurements within each subject)

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U} \qquad \Big| \qquad Y_{ij} = \boldsymbol{X}_i^\top \beta_j + \varepsilon_{ij}$$

where $\mathbb{Y} = (Y_{ij})_{i,j=1}^{N,n}$, $\mathbb{X} = (X_{ij})_{i,j=1}^{N,p}$, $\mathbb{B} = (\beta_{ij})_{i,j=1}^{p,n}$, and $\mathbb{U} = (\varepsilon_{ij})_{i,j=1}^{N,n}$

- What are the corresponding data (let's denote the data as $\mathcal{D}_S$)?
  (recall, that the vector of the explanatory covariates is subject specific)
- What is the meaning of the formulae above?
  (note, that the time dependence is only reflected within $\beta_j \in \mathbb{R}^p$)
- What are the objects appearing in the expression?
- What are typical assumptions for such linear model?

Question: What are the advantages or disadvantages of the longitudinal model formulation and the multivariate model formulation?

Parameter estimation

The main goal:

Estimation of the unknown parameters in $\mathbb{B} \in \mathbb{R}^{p \times n}$ and the variance-covariance matrix of the random (row) vectors in $\mathbb{U}$ (error terms)

In general:

Under different assumptions $\Rightarrow$ different estimation approaches $\Rightarrow$ different statistical properties of the estimates

- least squares
- maximum likelihood
- generalized method of moments
- likelihood-based estimation
- ...

$\hookrightarrow$ specific set of the postulated assumptions implies certain statistical properties

(in most applications it is assumed that $\mathbb{U}$ has uncorrelated, normally distributed rows with a zero mean vector and some specific variance-covariance matrix Σ)

Multivariate normal distribution

- **Multivariate normal model:** $\boldsymbol{u}_i \sim N_n(\mathbf{0}, \Sigma)$ where $\mathbb{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N)^\top$
  $\hookrightarrow$ *where subject specific error vectors in* $\mathbb{U}$ *are* $\boldsymbol{u}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in})^\top$

- starting with the multivariate normal regression model the unknown parameters can be estimated by the method of the maximum likelihood
- general form of the density of the multivariate normal distribution

$$f(\boldsymbol{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \qquad \boldsymbol{x} \in \mathbb{R}^n$$

- random sample $\mathcal{D}_S = \{(\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)^\top;\ i = 1, \ldots, N\}$, $\boldsymbol{Y}_i \in \mathbb{R}^n$ and $\boldsymbol{X}_i \in \mathbb{R}^p$
- the joint distribution of the random vectors $(\boldsymbol{Y}_i^\top, \boldsymbol{X}_i^\top)^\top$ can be expressed/factorized as

$$F_{(\boldsymbol{Y},\boldsymbol{X})}(\boldsymbol{y}, \boldsymbol{x}) = F_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) \cdot F_{\boldsymbol{X}}(\boldsymbol{x}) \qquad \forall \boldsymbol{y} \in \mathbb{R}^n \quad \forall \boldsymbol{x} \in \mathbb{R}^p$$

Likelihood and log-likelihood functions

- likelihood function for the data in $\mathcal{D}_S$ and unknown means $\mu_{ij} = \boldsymbol{X}_i^\top \beta_j$
  $\hookrightarrow$ *thus, the mean vector in the conditional distribution* $F_{\boldsymbol{Y}|\boldsymbol{X}}$ *is* $\boldsymbol{\mu} = \boldsymbol{X}_i^\top \mathbb{B}$

$$L(\mathbb{B}, \Sigma, \mathcal{D}_S) = \left[|2\pi\Sigma|^{-N/2} \cdot \exp\left\{ -\frac{1}{2} \sum_{i=1}^{N} (\boldsymbol{Y}_i - \boldsymbol{X}_i^\top \mathbb{B})^\top \Sigma^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i^\top \mathbb{B}) \right\} \right]$$

- hence, the log-likelihood function can be expressed as

$$l(\mathbb{B}, \Sigma, \mathcal{D}_S) = -\frac{N}{2} \log |2\pi\Sigma| - \frac{1}{2} trace \Big[(\mathbb{Y} - \mathbb{X}\mathbb{B}) \Sigma^{-1} (\mathbb{Y} - \mathbb{X}\mathbb{B})^\top \Big]$$

The empirical estimation of $\mathbb{B}$

$\hookrightarrow$ under the assumption that the matrix $\mathbb{X}^\top\mathbb{X}$ has a full rank ($p \in \mathbb{N}$), the maximum likelihood estimates of the mean parameters in $\mathbb{B} \in \mathbb{R}^{p\times n}$ are given by the expression

$$\widehat{\mathbb{B}}_N = \left(\mathbb{X}^\top\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{Y}$$

- denote the fitted values as $\widehat{\mathbb{Y}} = \mathbb{X}\widehat{\mathbb{B}}_N$
- denote the residuals as $\widehat{\mathbb{U}} = \mathbb{Y} - \widehat{\mathbb{Y}} = \mathbb{Y} - \mathbb{X}\widehat{\mathbb{B}}_N$
- denote the corresponding (regression) projection matrix as $\mathbb{H} = \mathbb{X}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$ and the residual projection matrix as $\mathbb{P} = (\mathbb{I} - \mathbb{H})$

Estimation of variance-covariance matrix

$\hookrightarrow$ under the assumption that the matrix $\mathbb{X}^\top\mathbb{X}$ has a full rank ($p \in \mathbb{N}$), the maximum likelihood estimate of the variance-covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$ is given by the expression

$$\widehat{\Sigma}_N = \frac{1}{N}\mathbb{Y}^\top(\mathbb{I} - \mathbb{H})\mathbb{Y} = \frac{1}{N}\widehat{\mathbb{U}}^\top\widehat{\mathbb{U}},$$

$\hookrightarrow$ the projection matrix $\mathbb{H}$ is also called the hat matrix and it projects from the N-dimensional real space $\mathbb{R}^N$ into a p-dimensional linear subspace. Alternatively, the matrix $(\mathbb{I} - \mathbb{H})$ is the projection matrix of the orthogonal projection into the residual subspace (the $(N-p)$-dimensional subspace of $\mathbb{R}^N$)

Useful formulae for derivations

- Linear form for $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{x} \in \mathbb{R}^p$

$$\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$$

- Quadratic form for $\mathbb{A} \in \mathbb{R}^{p \times p}$ (symmetric matrix)

$$\frac{\partial \boldsymbol{x}^\top \mathbb{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\mathbb{A}\boldsymbol{x} \qquad \text{and} \qquad \frac{\partial^2 \boldsymbol{x}^\top \mathbb{A} \boldsymbol{x}}{\partial \boldsymbol{x} \partial \boldsymbol{x}^\top} = 2\mathbb{A}$$

- Trace of a matrix $\mathbb{X}$

$$\frac{\partial \mathit{trace} \mathbb{X}\mathbb{A}}{\partial \mathbb{X}} = \begin{cases} \mathbb{A}^\top & \text{for general } \mathbb{X} \\ \mathbb{A} + \mathbb{A}^\top - \mathit{diag}(\mathbb{A}) & \text{for } \mathbb{X} \text{ symmetric} \end{cases}$$

Statistical properties $\widehat{\mathbb{B}}_N$ and $\widehat{\Sigma}_N$

- the estimates in $\widehat{\mathbb{B}}_N$ are unbiased estimates for $\mathbb{B}$
- for $\widehat{\mathbb{U}} = \mathbb{Y} - \mathbb{X}\widehat{\mathbb{B}}_N$ it holds that $E\widehat{\mathbb{U}} = \mathbf{0}$
- $\widehat{\mathbb{B}}_N$ and $\widehat{\mathbb{U}}$ are multivariate normal
- $\widehat{\mathbb{B}}_N$ and $\widehat{\mathbb{U}}$ are statistically independent
- covariance between $\widehat{\beta}_{ij}$ and $\widehat{\beta}_{kl}$ is equal to $\sigma_{jl} \cdot (\mathbb{X}^\top \mathbb{X})^{-1}_{(ik)}$
- $N\widehat{\Sigma}_N \sim W_n(\Sigma, N-p)$