Lecture 3 | 11.03.2024

Statistical inference in a multivariate regression model

Notation overview

- balanced longitudinal profiles $\mathcal{D}_B \equiv \{(\boldsymbol{Y}_i, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top;\ i = 1, \dots, N\}$
  - for $n_i = n \in \mathbb{N}$ for all $i = 1, \dots, N$
  - random vectors $(\boldsymbol{Y}_i, \boldsymbol{X}_{i1}^\top, \dots, \boldsymbol{X}_{in}^\top)^\top$ are independent with the same length
  - for longitudinal data we do not assume that subject specific measurements are taken at the same time $\Rightarrow$ $\mathcal{D}_B$ generally not a random sample!
  - for multivariate regression model we already assume that the observations in $\mathcal{D}_B$ form a random sample $\Rightarrow$ notation $\mathcal{D}_S$
- population and data model formulation (theoretical vs. empirical)

$$\boldsymbol{Y} = \boldsymbol{X}^\top \mathbb{B} + \boldsymbol{\varepsilon} \qquad\qquad \mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U}$$

for generic random vectors $\boldsymbol{Y} \in \mathbb{R}^n$ and $\boldsymbol{X} \in \mathbb{R}^p$ and some matrix of the unknown parameters $\mathbb{B} \in \mathbb{R}^{p \times n}$

The corresponding data: $\mathbb{Y} = (\boldsymbol{Y}_1^\top, \dots, \boldsymbol{Y}_N^\top)^\top$, $\mathbb{X} = (\boldsymbol{X}_1^\top, \dots, \boldsymbol{X}_N^\top)^\top$, and $\mathbb{U} = (\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_N^\top)^\top \equiv (\varepsilon_1^\top, \dots, \varepsilon_N^\top)^\top$

Statistical inference: Likelihood ratio test

- Inference in terms of confidence intervals/regions and hypothesis tests
- General form of the null hypothesis:

$$H_0 : \mathbb{C}_1\mathbb{B}\mathbb{M}_1 = \mathbb{D}$$

where $\mathbb{C}_1, \mathbb{M}_1$, and $\mathbb{D}$ are some (suitable) matrices

- The rows of $\mathbb{C}_1$ do inference on the effects of independent variables while the columns of $\mathbb{M}_1$ do inference on particular linear combinations of dependent variables
- In practical applications it is common that $\mathbb{D}$ is a zero matrix (all elements are zeros) and $\mathbb{M}_1 = \mathbb{I}$ (i.e. a unit matrix with ones on the main diagonal) $\hookrightarrow$ alternatively, the model of the form $\mathbb{Y}\mathbb{M}_1 = \mathbb{X}\mathbb{B}\mathbb{M}_1 + \mathbb{U}\mathbb{M}_1$
- Thus, the null hypothesis reduces to

$$H_0 : \mathbb{C}_1\mathbb{B} = \mathbf{0}$$

against a general alternative hypothesis of the form $H_A : \mathbb{C}_1\mathbb{B} \neq \mathbf{0} \in \mathbb{R}^{q\times n}$ (with the rank of the matrix $\mathbb{C}_1$ being equal to $q \in \mathbb{N}$)

Inference: Likelihood ratio test

- consider the null hypothesis of the form $H_0: \ \mathbb{C}_1\mathbb{B} = \mathbb{D}$
- the model $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U}$ can be equivalently expressed as

$$\widetilde{\mathbb{Y}} = \mathbb{Z}\widetilde{\mathbb{B}} + \mathbb{U},$$

  for $\widetilde{\mathbb{Y}} = \mathbb{Y} - \mathbb{X}\mathbb{B}_0$, where $\mathbb{C}_1\mathbb{B}_0 = \mathbb{D}$ (satisfies the null hypothesis), $\mathbb{Z} = \mathbb{X}\mathbb{C}^{-1}$ where $\mathbb{C}^\top = (\mathbb{C}_1^\top, \mathbb{C}_2^\top)$ and $\widetilde{\mathbb{B}} = (\widetilde{\mathbb{B}}_1^\top, \widetilde{\mathbb{B}}_2^\top)^\top = \mathbb{C}(\mathbb{B} - \mathbb{B}_0)$
- the null hypothesis $\mathbb{C}_1\mathbb{B} = \mathbb{D}$ gives that $\widetilde{\mathbb{B}}_1 = \mathbf{0}$ and for the matrix partition $\mathbb{C}^{-1} = (\mathbb{C}^{(1)}, \mathbb{C}^{(2)})$ the projection matrix

$$\mathbb{P}_1 = \mathbb{I} - \mathbb{X}\mathbb{C}^{(2)}(\mathbb{C}^{(2)\top}\mathbb{X}^\top\mathbb{X}\mathbb{C}^{(2)})^{-1}\mathbb{C}^{(2)\top}\mathbb{X}^\top$$

  defines the projection onto the linear subspace orthogonal to the columns of the matrix $\mathbb{X}\mathbb{C}^{(2)}$ (i.e., residuals for the regression onto $\mathbb{C}^{(2)}$ – under the null hypothesis, thus $\widetilde{\mathbb{B}}_1 = \mathbf{0}$)

LRT: Likelihood under the null and alternative

- maximized likelihood under the null hypothesis

$$\ell_0 = |2\pi N^{-1}\widetilde{\mathbb{Y}}^\top \mathbb{P}_1 \widetilde{\mathbb{Y}}|^{-N/2} \cdot \exp\{-\frac{1}{2}Nn\}$$

- maximized likelihood under the alternative hypothesis

$$\ell_1 = |2\pi N^{-1}\widetilde{\mathbb{Y}}^\top \widetilde{\mathbb{P}} \widetilde{\mathbb{Y}}|^{-N/2} \cdot \exp\{-\frac{1}{2}Nn\}$$

- the likelihood ratio test statistic is given as

$$\lambda^{2/N} = |\widetilde{\mathbb{Y}}^\top \widetilde{\mathbb{P}} \widetilde{\mathbb{Y}}| / |\widetilde{\mathbb{Y}}^\top \mathbb{P}_1 \widetilde{\mathbb{Y}}| = |\widetilde{\mathbb{Y}}^\top \widetilde{\mathbb{P}} \widetilde{\mathbb{Y}}| / |\widetilde{\mathbb{Y}}^\top \widetilde{\mathbb{P}} \widetilde{\mathbb{Y}} + \widetilde{\mathbb{Y}}^\top \mathbb{P}_2 \widetilde{\mathbb{Y}}|$$

  and it follows the $\Lambda(n, N-p, q)$ distribution, where $q \in \mathbb{N}$ is the number of rows in $\mathbb{C}_1$ (for $\mathbb{P}_2 = \mathbb{P}_1 - \widetilde{\mathbb{P}}$ – what does it mean geometrically?)

Examples

- Repeated measurements for two groups (two-sample problems):

$$\boldsymbol{Y}_i^{(1)} \sim N_n(\boldsymbol{\mu}_1, \Sigma), \qquad i = 1, \ldots, N_1$$
$$\boldsymbol{Y}_i^{(2)} \sim N_n(\boldsymbol{\mu}_2, \Sigma), \qquad i = 1, \ldots, N_2$$

- Typical testing problems:
  - parallel profiles of two groups $H_0 : \mathbb{C}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \boldsymbol{0}$
  - identical profiles for both groups $H_0 : \boldsymbol{1}^\top(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = 0$
  - treatment effect $H_0 : \mathbb{C}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) = \boldsymbol{0}$

- Multiple testing problem: testing for identical profiles only makes sense if the profiles are parallel; Similarly, if the profiles are parallel, is there any treatment effect at all?

Two sample problems:

Parallel profiles

$$T = \frac{N_1 N_2}{(N_1+N_2)^2}(N_1+N_2-2)\left[\mathbb{C}(\overline{Y}^{(1)} - \overline{Y}^{(2)})\right]^\top \left(\mathbb{C}\mathcal{S}\mathbb{C}^\top\right)^{-1}\left[\mathbb{C}(\overline{Y}^{(1)} - \overline{Y}^{(2)})\right]$$

and (under the null hypothesis) $T \sim T^2(n-1, N_1+N_2-2)$

Equality of two levels

$$T = \frac{N_1 N_2}{(N_1+N_2)^2}(N_1+N_2-2)\frac{\left[\mathbf{1}^\top(\overline{Y}^{(1)} - \overline{Y}^{(2)})\right]^2}{\mathbf{1}^\top \mathcal{S}\mathbf{1}}$$

and (under the null hypothesis) $T \sim T^2(1, N_1+N_2-2)$

Same treatment effect

$$T = (N_1+N_2-2)(\mathbb{C}\overline{Y})^\top \left(\mathbb{C}\mathcal{S}\mathbb{C}^\top\right)^{-1}\mathbb{C}\overline{Y}, \quad \text{for } \overline{Y} = \frac{N_1\overline{Y}^{(1)} + N_2\overline{Y}^{(2)}}{N_1+N_2}$$

and (under the null hypothesis) $T \sim T^2(n-1, N_1+N_2-2)$

Overview

- statistical test about some (multivariate) mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ can be often expressed in terms of the null hypothesis $H_0: \ \mathbb{A}\boldsymbol{\mu} = \boldsymbol{a}$ vs. $H_A: \ \mathbb{A}\boldsymbol{\mu} \neq \boldsymbol{a}$. where $\mathbb{A} \in \mathbb{R}^{q \times n}$ and $\boldsymbol{a} \in \mathbb{R}^q$
- for $\boldsymbol{X}_i \sim N_n(\boldsymbol{\mu}, \Sigma)$ for $i = 1, \dots, N$, with Σ known, the log-likelihood based test statistic $-2\log\lambda = N(\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a})^\top(\mathbb{A}\Sigma\mathbb{A}^\top)^{-1}(\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a})$ follows (exactly) the χ^2 distribution with $q \in \mathbb{N}$ degrees of freedom
- for $\boldsymbol{X}_i \sim N_n(\boldsymbol{\mu}, \Sigma)$ for $i = 1, \dots, N$, with Σ unknown, the log-likelihood test statistic $-2\log\lambda = N\log\left\{1 + (\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a})^\top(\mathbb{A}\widehat{\Sigma}_N\mathbb{A}^\top)^{-1}(\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a})\right\}$ follows asymptotically the χ^2 distribution with $q \in \mathbb{N}$ degrees of freedom and the exact Hotelling test is based on the test statistic

$$(N-1)(\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a})^\top(\mathbb{A}\widehat{\Sigma}_N\mathbb{A}^\top)^{-1}(\mathbb{A}\overline{\boldsymbol{X}}_N - \boldsymbol{a}) \sim T^2(q, N-1)$$

- for $Y_i \sim N(\boldsymbol{X}_i^\top\beta, \sigma^2)$, for $i = 1, \dots, N$, with $\sigma^2 > 0$ unknown, the test of the null hypothesis $H_0: \ \mathbb{A}\beta = \boldsymbol{a}$, for $\boldsymbol{a} \in \mathbb{R}^q$, leads to the test statistic

$$\frac{N-n}{q} \cdot \frac{(\mathbb{A}\widehat{\beta} - \boldsymbol{a})^\top\left[\mathbb{A}(\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{A}\right]^{-1}(\mathbb{A}\widehat{\beta} - \boldsymbol{a})}{(\boldsymbol{Y} - \mathbb{X}\widehat{\beta})^\top(\boldsymbol{Y} - \mathbb{X}\widehat{\beta})} \sim F_{q,N-n}$$

Multivariate model vs. general linear model

- Multivariate regression model $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U}$
  - $n \in \mathbb{N}$ repeated measurements within $N \in \mathbb{N}$ subjects (random sample)
  - repeated measurements taken at the same time-points across subjects
  - time evolution modeled by the set of $\beta_j \in \mathbb{R}^p$ parameters ($j = 1, \dots, n$)
  - the vector of subject's specific covariates $\boldsymbol{X}_i \in \mathbb{R}^p$ fixed over time
  - covariance structure modeled by the matrix Σ, where $\boldsymbol{u}_i \sim N_n(\boldsymbol{0}, \Sigma)$
  - the data usually form a random sample from the joint distribution $F_{\boldsymbol{Y},\boldsymbol{X}}$

- General linear model for correlated errors $\boldsymbol{Y} = \mathbb{X}\beta + \varepsilon$
  - $n \in \mathbb{N}$ repeated measurements within $N \in \mathbb{N}$ subjects (balanced data)
  - the vector of unknown parameters $\beta \in \mathbb{R}^p$ is fixed over time
  - subject's specific covariates $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ may vary with $j \in \{1, \dots, n\}$
  - subjects' independence and within subject's covariance modeled by the variance covariance Σ, where $\varepsilon \sim N(\boldsymbol{0}, \Sigma)$ (overall dimensionality: Nn)
  - the model can be further generalized for unbalanced data ($n_i \in \mathbb{N}$)

General linear model with correlated errors

- instead of time-varying β_j and fixed $\boldsymbol{X}_j \in \mathbb{R}^p$ the time evolution can be modeled in terms of time-varying covariates $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ and fixed $\boldsymbol{\beta} \in \mathbb{R}^p$
- Simplification in terms of the vectors of unknown parameters $\boldsymbol{\beta}_j \in \mathbb{R}^p$ for $j = 1, \dots, n$ (in the matrix $\mathbb{B} \in \mathbb{R}^{p \times n}$): $\Rightarrow \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n$
- Relaxation in terms of the subject's specific covariates $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ that are now allowed to change with $j \in \{1, \dots, n\}$: $\Rightarrow \boldsymbol{X}_{ij} = (X_{ij1}, \dots, X_{ijp})^\top \in \mathbb{R}^p$
- this allows for an alternative formulation of the multivariate (data) model (where $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U}$ follows as a special case) in a form

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{Nn} \end{pmatrix} = \begin{pmatrix} X_{111} & \dots & X_{11p} \\ \vdots & \ddots & \vdots \\ X_{1n1} & \dots & X_{1np} \\ X_{211} & \dots & X_{21p} \\ \vdots & \ddots & \vdots \\ X_{Nn1} & \dots & X_{Nnp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{Nn} \end{pmatrix}$$

General linear model with correlated errors

- instead of time-varying β_j and fixed $\boldsymbol{X}_j \in \mathbb{R}^p$ the time evolution can be modeled in terms of time-varying covariates $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ and fixed $\boldsymbol{\beta} \in \mathbb{R}^p$
- Simplification in terms of the vectors of unknown parameters $\boldsymbol{\beta}_j \in \mathbb{R}^p$ for $j = 1, \dots, n$ (in the matrix $\mathbb{B} \in \mathbb{R}^{p \times n}$): $\Rightarrow \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_n$
- Relaxation in terms of the subject's specific covariates $\boldsymbol{X}_{ij} \in \mathbb{R}^p$ that are now allowed to change with $j \in \{1, \dots, n\}$: $\Rightarrow \boldsymbol{X}_{ij} = (X_{ij1}, \dots, X_{ijp})^\top \in \mathbb{R}^p$
- this allows for an alternative formulation of the multivariate (data) model (where $\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U}$ follows as a special case) in a form

$$\begin{pmatrix} Y_{11} \\ \vdots \\ Y_{1n} \\ Y_{21} \\ \vdots \\ Y_{Nn} \end{pmatrix} = \begin{pmatrix} X_{111} & \dots & X_{11p} \\ \vdots & \ddots & \vdots \\ X_{1n1} & \dots & X_{1np} \\ X_{211} & \dots & X_{21p} \\ \vdots & \ddots & \vdots \\ X_{Nn1} & \dots & X_{Nnp} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{Nn} \end{pmatrix}$$

- What are the advantages and disadvantages of both model formulations?

Matrix formulation

- typically we use the notation (under multivariate normal assumption)

$$\boldsymbol{Y} \sim N_{Nn}(\mathbb{X}\beta, \sigma^2\mathbb{V}),$$

  where $\mathbb{V}$ is a block-diagonal matrix with non-zero blocks of size $n \times n$ (each block $\sigma^2\mathbb{V}_0$ represents the variance-covariance of repeated measurements within a single subject)
- the variance covariance matrix $\sigma^2\mathbb{V}$ is estimated by borrowing power across subject (i.e., replication of $\sigma^2\mathbb{V}_0$ across the units)
- there can be no specific (parametric) structure assumed for $\mathbb{V}_0$ but it is common to postulate some parametric form of $\mathbb{V}_0$
- the correlation structure within $\sigma^2\mathbb{V}$ is crucial for a proper inference

Uniform correlation model

- **Assumption:** positive correlation $\rho \in (0,1)$ between any two repeated observations within a given subject
- **Matrix notation:** $\mathbb{V}_0 = (1-\rho)\mathbb{I}_{n\times n} + \rho \mathbf{1}_{n\times n}$
- **Motivation:** the response (random) variable Y_{ij} can be decomposed as

$$Y_{ij} = \mu_{ij} + Z_i + V_{ij},$$

where $\mu_{ij} = EY_{ij}$ and $Z_i \sim N(0,\nu^2)$ independent of $V_{ij} \sim N(0,\tau^2)$ and it holds that $\rho = \nu^2/(\nu^2+\tau^2)$ and $\sigma^2 = \nu^2 + \tau^2$ (for $\varepsilon_{ij} = Z_i + V_{ij}$)

- **Interpretation:** linear model for the mean of the response with a random intercept (with the variance between subjects $\nu^2 > 0$)

Exponential correlation model

- **Assumption:** covariance between Y_{ij} and Y_{ik} for $i \neq k$ is of the form

$$v_{jk} = \sigma^2 \exp\{-\phi|t_j - t_k|\}$$

and it decays towards zero as the time separation between repeated observations increases (with the rate of decay given by $\phi > 0$)

- **Matrix notation:** $\mathbb{V}_0 = (v_{jk})_{j,k=1}^{n}$
- **Motivation:** the response (random) variable Y_{ij} can be decomposed as

$$Y_{ij} = \mu_{ij} + W_{ij},$$

where $W_{ij} = \rho W_{i(j-1)} + Z_{ij}$ for $Z_{ij} \sim N(0, \sigma^2(1-\rho^2))$ independent (verify, that it holds that $VarY_{ij} = VarW_{ij} = \sigma^2$)

- **Interpretation:** linear model for the mean of the response with with the first order autoregressive correlation structure
- **Generalization:** $Y_{ij} = \mu_{ij} + W_i(t_j)$ for continuous time Gaussian processes $\{X_i(t);\ t \in \mathbb{R}\}$ independent for $i = 1, \dots, N$ and general time points $t_1 < \cdots < t_{ni}$

Towards least squares – two step estimation

- For simplification assume the model $Y_{ij} = a + bX_{ij} + \varepsilon_{ij}$ and no distributional assumption for the error vector $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{Nn})^\top$
- Two-stage least squares as a simple estimation method for models where it is needed to distinguish the variance sources (within/between subjects)

Towards least squares – two step estimation

- For simplification assume the model $Y_{ij} = a + bX_{ij} + \varepsilon_{ij}$ and no distributional assumption for the error vector $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{Nn})^\top$
- Two-stage least squares as a simple estimation method for models where it is needed to distinguish the variance sources (within/between subjects)
- **Stage 1:** longitudinal profiles for each subject $i \in \{1, \dots, N\}$ individually

$$Y_{ij} = A_i + B_i X_{ij} + W_{ij}, \quad j = 1, \dots, n, \quad \text{and } W_{ij} \sim (0, \tau^2), \ i.i.d.$$

  to obtain $\widehat{A}_i = A_i + Z_{ai}$ and $\widehat{B}_i = B_i + Z_{bi}$, for $Z_{ai} \sim (0, v_{ai}^2)$, $Z_{bi} \sim (0, v_{bi}^2)$
- **Stage 2:** OLS analysis of the subject's specific parameter estimates

$$A_i = a + \delta_{ai} \qquad \text{and} \qquad B_i = b + \delta_{bi}$$

  for independent errors $\delta_{ai} \sim (0, \sigma_a^2)$ and $\delta_{bi} \sim (0, \sigma_b^2)$
- **Therefore:** $\widehat{A}_i = a + (\delta_{ai} + Z_{ai})$ and $\widehat{B}_i = b + (\delta_{bi} + Z_{bi})$

Summary

- Two alternative but not equivalent multivariate model formulations

$$\mathbb{Y} = \mathbb{X}\mathbb{B} + \mathbb{U} \qquad \text{versus} \qquad \boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Estimation of the unknown parameters in $\mathbb{B} \in \mathbb{R}^{p \times n}$ or $\boldsymbol{\beta} \in \mathbb{R}^{p}$
  (either in terms of the least squares or the maximum likelihood estimation)
- Decomposition of the overall data variability into two different sources
  (the within subject's variability and the between subjects' variability)
- Marginal or hierarchical inference (in terms of the confidence intervals/regions or the statistical tests)
- Two stage estimation approach in the model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\beta}$
  (towards the mixed effect model with fixed and random effects)