Lecture 4 | 17.03.2025

Statistical inference in a multivariate model for Y

Overview: Two step estimation I

- Motivation for a simple model of the form model $Y_{ij} = a + bX_{ij} + \varepsilon_{ij}$ with no distributional assumption for correlated errors $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{Nn})^\top$

- **Stage 1:** OLS for each subject's specific profile individually (i.e, fixed i)

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

  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:** Assumption about the subject's specific (true) parameters

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

  for errors $\delta_{ai} \sim (0, \sigma_a^2)$ and $\delta_{bi} \sim (0, \sigma_b^2)$ (ie., subject's specific variability) where the primary interest is to estimate the unknown parameters $a, b \in \mathbb{R}$

- **Thus, we obtain** $\widehat{A}_i = a + (\delta_{ai} + Z_{ai})$ and $\widehat{B}_i = b + (\delta_{bi} + Z_{bi})$ with the error term decomposed into 2 parts (within/between variability)

Overview: Two step estimation II

- both stages can be straightforwardly combined together as

$$\begin{aligned}
Y_{ij} &= A_i + B_1 + W_{ij} \\
&= (a + \delta_i) + (b + \delta_{bi})X_{ij} + W_{ij} \\
&= a + bX_{ij} + \delta_{ai} + \delta_{bi}X_{ij} + W_{ij}\underbrace{(\delta_{ai} + \delta_{bi}X_{ij} + W_{ij})}_{\varepsilon_{ij}} \\
&= a + bX_{ij} + \varepsilon_{ij}
\end{aligned}$$

- What is the variance of the of Y_{ij}?
- What is the covariance of two observations Y_{ij} and Y_{ik}, for $j \neq k$?
- What is the covariance of Y_{ij} and Y_{lk}, for $i \neq l$ and $j \neq k$?

Weighted least-squares estimation (WLS)

- Note, that in $\widehat{A}_i = a + (\delta_{ai} + Z_{ai})$ the errors δ_{ai} for $i = 1, \dots, N$ have all the same variance σ_a^2 but Z_{ai} have different variances $v_{ai}^2 > 0$
  Similarly, the same argument also holds for $\widehat{B}_i = B + (\delta_{bi} + Z_{bi})$
- Therefore, proper estimates for $a, b \in \mathbb{R}$ should be the weighted averages of the subject's specific parameter estimates $\widehat{A}_i$ and $\widehat{B}_i$
- Consider again the multivariate model $\boldsymbol{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, for $\boldsymbol{\varepsilon} \sim N_{Nn}(\boldsymbol{0}, \sigma^2\mathbb{V})$ and take an arbitrary symmetric (regular) weighted matrix $\mathbb{W} \in \mathbb{R}^{Nn \times Nn}$

  $\Longrightarrow$ the **weighted LS estimate of β is defined** as

$$\widehat{\boldsymbol{\beta}}_w = \left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbb{W} \boldsymbol{Y}$$

  $\hookrightarrow$ which is an **unbiased (linear) estimate** whatever the choice of $\mathbb{W}$

- However, for the variance of $\widehat{\beta}_w$ it holds that

$$Var(\widehat{\boldsymbol{\beta}}_w) = \sigma^2 \left[\left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1} \mathbb{X}^\top \mathbb{W} \mathbb{V} \mathbb{W} \mathbb{X} \left(\mathbb{X}^\top \mathbb{W} \mathbb{X}\right)^{-1}\right]$$

$$Var(\widehat{\boldsymbol{\beta}}_w) = \sigma^2 \left(\mathbb{X}^\top \mathbb{V}^{-1} \mathbb{X}\right)^{-1} \qquad \text{only for } \mathbb{W} = \mathbb{V}^{-1}$$

  $\hookrightarrow$ can we choose $\mathbb{W}$ such that $\mathbb{W} = \mathbb{V}^{-1}$? How important this choice is?

Estimation under the normal model (MLE)

- Using, in addition, the assumption of the normal multivariate model i.e., $\boldsymbol{Y} \sim N_{Nn}(\mathbb{X}\beta, \sigma^2\mathbb{V})$ (or $\varepsilon \sim N_{Nn}(\mathbf{0}, \sigma^2\mathbb{V})$ alternatively) we can use the maximum likelihood estimation approach to construct the estimates...
- The **log-likelihood** for the observed data in $\mathcal{D}_S$ takes the form

$$\ell(\beta, \sigma^2, \mathbb{V}_0, \mathcal{D}_S) = -\frac{1}{2}\left[Nn\log(\pi\sigma^2) + N\log|\mathbb{V}_0| + \frac{(\boldsymbol{Y} - \mathbb{X}\beta)^\top\mathbb{V}^{-1}(\boldsymbol{Y} - \mathbb{X}\beta)}{\sigma^2}\right]$$

- For a particular choice of $\mathbb{V}_0 \in \mathbb{R}^{n\times n}$ the MLE of β is given by the expression

$$\widehat{\beta}(\mathbb{V}_0) = \left(\mathbb{X}^\top\mathbb{V}^{-1}\mathbb{X}\right)^{-1}\mathbb{X}^\top\mathbb{V}^{-1}\boldsymbol{Y}$$

- By substituting the estimate $\widehat{\beta}(\mathbb{V}_0)$ into the likelihood form we obtain

$$\ell(\widehat{\beta}(\mathbb{V}_0), \sigma^2, \mathbb{V}_0, \mathcal{D}_S) = -\frac{1}{2}\left[Nn\log(\pi\sigma^2) + N\log|\mathbb{V}_0| + \frac{(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))^\top\mathbb{V}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))}{\sigma^2}\right]$$

- Consequently, the partial derivative with respect to σ^2 gives the MLE of σ^2 as

$$\widehat{\sigma^2}(\mathbb{V}_0) = \frac{(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))^\top\mathbb{V}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))}{Nn}$$

Estimation of the covariance structure

- The covariance structure in $\mathbb{V}_0$ must be still estimated – it can be done using the **reduced log-likelihood** by using the estimated quantities $\widehat{\beta}(\mathbb{V}_0)$ and $\widehat{\sigma^2}(\mathbb{V}_0)$

- The **reduced log-likelihood** for $\mathbb{V}_0$ can be expressed (except some constant) as

$$\ell(\mathbb{V}_0) \equiv \ell(\widehat{\beta}(\mathbb{V}_0), \widehat{\sigma^2}(\mathbb{V}_0), \mathbb{V}_0, \mathcal{D}_S) =$$
$$= -\frac{1}{2}\left[Nn\log(\pi\widehat{\sigma^2}(\mathbb{V}_0)) + N\log|\mathbb{V}_0| + 0\right] =$$
$$= -\frac{N}{2}\left[n\log\left((\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))^\top\mathbb{V}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))\right) + \log|\mathbb{V}_0|\right] + const$$

- Finally, the ML estimate $\widehat{\mathbb{V}_0}$ is used to obtain the estimates for the mean and variance, i.e.,

$$\widehat{\beta} = \widehat{\beta}(\widehat{\mathbb{V}_0}) \qquad \text{and} \qquad \widehat{\sigma^2} = \widehat{\sigma^2}(\widehat{\mathbb{V}_0})$$

(however, the maximization of $\ell(\mathbb{V}_0)$ with respect to the parameters in $\mathbb{V}_0$ requires nontrivial optimization techniques and algorithms – generally, the dimensionality of the optimization problem for $\mathbb{V}_0$ is $\frac{n(n-1)}{2}$ – calculation of the determinant and inverse of a $n \times n$ matrix)

Consistency of the estimates $\widehat{\sigma^2}$ and $\widehat{\mathbb{V}_0}$

- Note, that in the simultaneous estimation of the mean, variance, and the covariance parameters ($\boldsymbol{\beta}, \sigma^2$, and $\mathbb{V}_0$) the design/model matrix $\mathbb{X}$ is explicitly involved in the estimate for σ^2 as well as $\mathbb{V}_0$
- If the matrix $\mathbb{X}$ is specified incorrectly, the estimates for σ^2 and $\mathbb{V}_0$ are not even consistent $\Longrightarrow$ using a full saturated model for the mean structure can offer a possible solution (large number of the estimated parameters)
- Saturated model for the conditional mean structure guarantees consistent estimates of the variance-covariance structure which can be further used to do inference about the mean structure (to reduce its complexity)
- **Good strategy but very often not feasible!**
- **The maximum likelihood estimation works relatively well if the model matrix $\mathbb{X}$ is well specified... otherwise, it can be more appropriate to use the restricted maximum likelihood (REML) approach**

Restricted maximum likelihood (REML)

The main idea is to somehow restrict the dependency of the estimates $\widehat{\sigma^2}$ and $\widehat{\mathbb{V}_0}$ on the mean structure postulated by the design/model matrix $\mathbb{X}$...

(Patterrson and Thompson, 1971)

- standard maximum likelihood typically gives biased variance estimate (even in classical regression, compare RSS/n versus $RSS/(n-p)$)
- the principal idea is to perform standard MLE for transformed data $\boldsymbol{Y}^*$ such that the distribution of $\boldsymbol{Y}^* = \mathbb{A}\boldsymbol{Y}$ does not depend on $\boldsymbol{\beta} \in \mathbb{R}^p$
- one possible option for $\mathbb{A}$ is a transformation of $\boldsymbol{Y}$ into OLS residuals which means that the matrix $\mathbb{A}$ takes the form $\mathbb{A} = \mathbb{I} - \mathbb{X}(\mathbb{X}^{-1}\mathbb{X})^{-1}\mathbb{X}$
- however, any (full-rank) matrix which satisfies $E\boldsymbol{Y}^* = \boldsymbol{0}$, $\forall \boldsymbol{\beta} \in \mathbb{R}^p$ will give unbiased estimates for the variance-covariance parameters
- nevertheless, both methods (maximim likelihood and REML) are asymptotically equivalent whenever the sample size tends to infinity and $p \in \mathbb{N}$ is fixed (for $p \to \infty$ the problem is more complex, REML)

REML – some calculation details

- let's assume that $\boldsymbol{Y} \sim N_{Nn}(\mathbb{X}\boldsymbol{\beta}, \mathbb{H}(\boldsymbol{\alpha}))$ for $\boldsymbol{\alpha} \in \mathbb{R}^q$ where $\mathbb{H}(\boldsymbol{\alpha})$ fully captures the variance-covariance structure (i.e., including the variance σ^2)
- for the projection matrix $\mathbb{A} = \mathbb{I} - \mathbb{X}(\mathbb{X}^{-1}\mathbb{X})^{-1}\mathbb{X}$, let $\mathbb{B} \in \mathbb{R}^{Nn \times (Nn-p)}$ is a matrix which satisfies $\mathbb{B}\mathbb{B}^\top = \mathbb{A}$ and $\mathbb{B}^\top\mathbb{B} = \mathbb{I}_{(Nn-p)\times(Nn-p)}$
- let $\boldsymbol{Z} = \mathbb{B}^\top \boldsymbol{Y}$ be the vector of transformed response vector $\boldsymbol{Y}$ where, from the normality property, we have $\boldsymbol{Z} \sim N_{(Nn-p)}(\mathbb{B}^\top\mathbb{X}\boldsymbol{\beta}, \mathbb{B}^\top\mathbb{H}(\boldsymbol{\alpha})\mathbb{B})$
  (Why not to use the random $\boldsymbol{Z} = \mathbb{A}\boldsymbol{Y} \sim N_n(\mathbb{A}^\top\mathbb{X}\beta, \mathbb{A}^\top\mathbb{H}(\alpha)\mathbb{A})$ instead?)
- the corresponding maximum likelihood estimate of β based on $\boldsymbol{Y}$ (fixed $\boldsymbol{\alpha}$) is the generalized least-squares estimator $\widehat{\beta} = (\mathbb{X}^\top\mathbb{H}^{-1}\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{H}^{-1}\boldsymbol{Y}$
- random vector $\boldsymbol{Z}$ and $\widehat{\beta}$ are independent – whatever the true value of $\boldsymbol{\beta} \in \mathbb{R}^p$ and, moreover, it also holds that $E\boldsymbol{Z} = \boldsymbol{0}$
- thus, we have $\boldsymbol{Z} \sim N_{Nn-p}(\mathbb{B}^\top\mathbb{X}\boldsymbol{\beta}, \mathbb{B}^\top\mathbb{H}(\boldsymbol{\alpha})\mathbb{B}) \equiv N_{Nn-p}(\boldsymbol{0}, \mathbb{B}^\top\mathbb{H}(\boldsymbol{\alpha})\mathbb{B})$, where $\boldsymbol{Z}$ is independent of $\widehat{\beta}$ therefore, the inference for $\boldsymbol{\alpha} \in \mathbb{R}^q$ can be performed independently of β based on $\boldsymbol{Z}$
- the multivariate normal density of $\boldsymbol{Z}$ (expressed in terms of $\boldsymbol{Y}$) is proportional to the ratio of the density of $\boldsymbol{Y}$ and the density of $\widehat{\beta}$

REML – overview

- the maximum likelihood estimate of $\boldsymbol{\alpha} \in \mathbb{R}^q$ maximizes the log-likelihood

$$\ell(\boldsymbol{\alpha}) = -\frac{Nn}{2}\log(2\pi) - \frac{1}{2}\log|\mathbb{H}| - \frac{1}{2}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\top \mathbb{H}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})$$

- the restricted maximum likelihood estimate (REML) of $\boldsymbol{\alpha} \in \mathbb{R}^q$ maximizes

$$\ell^*(\boldsymbol{\alpha}) = -\frac{(Nn-p)}{2}\log(2\pi) - \frac{1}{2}\log|\mathbb{H}| - \frac{1}{2}\log|\mathbb{X}^\top\mathbb{H}^{-1}\mathbb{X}| - \frac{1}{2}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})^\top \mathbb{H}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\boldsymbol{\beta}})$$

ML and REML estimates:

- Thus, for given $\mathbb{V}_0$ the MLE estimate of β is $\widehat{\beta}(\mathbb{V}_0) = (\mathbb{X}^\top\mathbb{V}^{-1}\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{V}^{-1}\boldsymbol{Y}$
- The REML of the variance parameter $\sigma^2 > 0$ is as

$$\widehat{\sigma^2}(\mathbb{V}_0) = \frac{1}{Nn-p}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))^\top \mathbb{V}_0^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))$$

- and the REML estimate of $\mathbb{V}_0$ maximizes the reduced log-likelihood function

$$\ell^*(\mathbb{V}_0) = -\frac{1}{2}N\left[n\log(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0))^\top\mathbb{V}^{-1}(\boldsymbol{Y} - \mathbb{X}\widehat{\beta}(\mathbb{V}_0)) + \log|\mathbb{V}_0|\right] - \frac{1}{2}\log|\mathbb{X}^\top\mathbb{V}^{-1}\mathbb{X}|$$

Robust estimation of the standard errors

- the idea is to allow for a robust inference for $\boldsymbol{\beta} \in \mathbb{R}^p$ by using a generalized least-squares estimator $\widehat{\boldsymbol{\beta}}_W = (\mathbb{X}^\top \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{W} \boldsymbol{Y}$ and the variance-covariance $\widehat{R}_W = \left[(\mathbb{X}^\top \mathbb{W} \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{W}\right] \widehat{\mathbb{V}} \left[\mathbb{W} \mathbb{X} (\mathbb{X}^\top \mathbb{W} \mathbb{X})^{-1}\right]$
- statistical inference for $\boldsymbol{\beta}$ is based on the assumption that

$$\widehat{\boldsymbol{\beta}}_W \sim N_p(\boldsymbol{\beta}, \widehat{R}_W)$$

- **Matrix $\mathbb{W}^{-1}$ is called the working correlation matrix** (qualitative)
- **Matrix $\mathbb{V}$ is the unknown true variance-covariance matrix**

$\hookrightarrow$ however, poor choice of $\mathbb{W}$ will only effect the efficiency of the inference about $\boldsymbol{\beta} \in \mathbb{R}^p$ but not the its validity $\Longrightarrow$ confidence intervals and statistical tests will be asymptotically correct whatever the true form of $\mathbb{V}$

$\hookrightarrow$ typically, it is either common to use $\mathbb{W}^{-1} = \mathbb{I}$ or, for smoothly decaying autocorrelation, a block-diagonal matrix $\mathbb{W}^{-1}$ with elements $\exp\{-c|t_j - t_k|\}$, $c > 0$

Example: Designed experiment

- measurements Y_{ijg}, for $i = 1, \ldots, N_g$, $g = 1, \ldots, G$, and $j = 1, \ldots, n$
- saturated model for the response $EY_{ijg} = \mu_{jg}$
- variance-covariance $Var\,\boldsymbol{Y} = \mathbb{V}$ with diagonal blocks $\mathbb{V}_0 \in \mathbb{R}^{n \times n}$

Example

- REML estimate for $\mathbb{X}$ using a specific form of the model matrix $\mathbb{X}$

$$\mathbb{X} = \begin{pmatrix} \mathbb{I} & \mathbb{O} \\ \mathbb{I} & \mathbb{O} \\ \mathbb{O} & \mathbb{I} \\ \mathbb{O} & \mathbb{I} \\ \mathbb{O} & \mathbb{I} \end{pmatrix}$$

for a particular choice of $G = 2$ (number of groups), $N_1 = 2$ (individuals in the first group), and $N_2 = 3$ (individuals in the second group) and $n \in \mathbb{N}$ (number of repeated observations for each subject)

- The vector of unknown parameters $\boldsymbol{\beta} = (\beta_1^{(g_1)}, \ldots, \beta_n^{(g1)}, \beta_1^{(g2)}, \ldots, \beta_n^{(g2)})^\top$

Example: Designed experiment – estimates

- **Mean estimates**

$$\widehat{\mu}_{jg} = \frac{1}{N_g}\sum_{i=1}^{N_g} Y_{ijg}$$

- **REML estimate for $\mathbb{V}_0$**

$$\widehat{\mathbb{V}}_0 = \Big(\sum_{g=1}^{G} N_g - G\Big)^{-1} \sum_{g=1}^{G}\sum_{i=1}^{N_g}(\boldsymbol{Y}_{ig} - \widehat{\mu}_g)(\boldsymbol{Y}_{ig} - \widehat{\mu}_g)^\top$$

- **REML estimate for $\mathbb{V}$**
  is a block-diagonal matrix with blocks formed by the estimate $\widehat{\mathbb{V}}_0$

$\hookrightarrow$ the saturated model for the mean structure may not be useful in practice – its only purpose is **to provide a consistent estimate of $\mathbb{V}_0$**... for observational studies with continuously varying covariates it is no longer applicable...

However, the principal idea remains the same...

Summary

- weighted least-squares estimation vs. maximum likelihood estimation
  (with or without the assumption of the normal model)
- maximum likelihood vs. restricted maximum likelihood estimation
  (robust estimates for β - limiting the dependence on $\mathbb{X}$)
- inference about the mean structure based on $\widehat{\beta}_W \sim N_p(\beta, \widehat{R}_W)$
  (using the assumption of the multivariate normal model for the response)
- special attention given to a consistent estimation of $\mathbb{V}$
  (saturated or most elaborated model is used to get the estimate $\widehat{\mathbb{V}}$)