Index of Extra Lecture 2

Introduction to Session

Today's session:

- presentations and discussions of data sets for multilevel course,
- material on repeated measures data as a special case of *multivariate data*, which essentially views the series of observations per subject,

$$Y = (Y_1, \ldots, Y_n),$$

  as a multivariate observation ($\sim$ distinct variables, or the wide data format).[1]

What's different with the multivariate approach?

- no modelling of within-subject correlation structure, instead all correlations are estimated without restrictions,
- time points are considered as categories instead of a quantitative dimension.

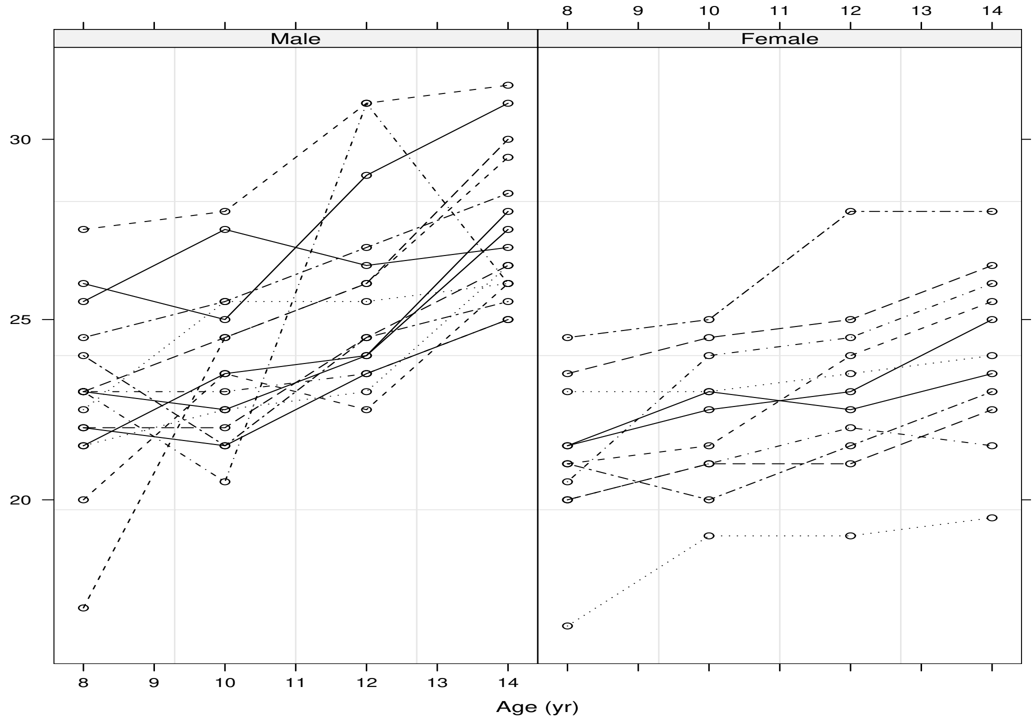
Supplementary reading:

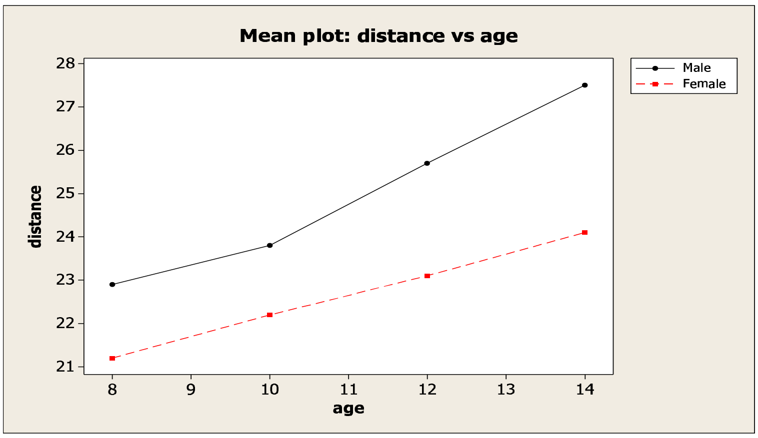
- material not part of standard VHM 802 curriculum,
- textbook: Chapters 3-4 of Davis (2002): *Statistical Methods for the Analysis of Repeated Measurements.*

[1] Here multivariate is absolutely not the same as multivariable!

Orthodont Data

- a study conducted in two groups of children (16 boys and 11 girls); Univ. North Carolina, Potthoff & Roy (1964),
- at ages 8, 10, 12, and 14, the distance (mm) from the center of the pituitary to the pterygomaxillary fissure was measured (on x-rays),
- objective: compare growth profiles of boys and girls.





ONE-SAMPLE MULTIVARIATE NORMAL MODEL

Notation and assumptions:

- for each "subject" i, the n-dimensional observation vector $Y_i = (Y_{i1}, \ldots, Y_{in})$ follows a MVN (multivariate normal) distribution[2] (μ, Σ); i.e., with mean $\mu = (\mu_1, \ldots, \mu_n)$ and $n \times n$ variance/covariance matrix Σ,

$$\Sigma = \begin{pmatrix} \sigma_{11} & \ldots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \ldots & \sigma_{nn} \end{pmatrix},$$

- the m subjects ($i = 1, \ldots, m$) are independent and have the same mean and variance (i.i.d.),
- combining the equations for all subjects, we may write

$$\begin{aligned} Y &\equiv \begin{pmatrix} Y_{11} & \ldots & Y_{1n} \\ \vdots & \ddots & \vdots \\ Y_{m1} & \ldots & Y_{mn} \end{pmatrix} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_m \end{pmatrix} \\ &= \begin{pmatrix} \mu_1 & \ldots & \mu_n \\ \vdots & \ddots & \vdots \\ \mu_1 & \ldots & \mu_n \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \ldots & \varepsilon_{1n} \\ \vdots & \ddots & \vdots \\ \varepsilon_{m1} & \ldots & \varepsilon_{mn} \end{pmatrix} \equiv X\beta + \varepsilon, \end{aligned}$$

  with a $(1 \times n)$ parameter matrix $\beta = \mu$, a $(m \times 1)$ design matrix $X = (1)$, and a $(m \times n)$ error matrix ε,
- the errors $\varepsilon \sim \text{MVN}(0, I_m \otimes \Sigma)$, where $I_m \otimes \Sigma$ is short for a block-diagonal covariance matrix.

[2] This assumes a normal distribution for each component, *and* the conditional distributions of subsets of components given the others to be normal, but this extra condition is often ignored in practice.

Multivariate Analysis for Orthodont Data

Model: (denoting as before by Y_{ij} the distance for child i at time point j)

$$Y_{ij} = \mu_{\text{sex}(i),j} + \varepsilon_{ij} \quad \text{or} \quad Y = X\mu + \varepsilon,$$

where X is a 27×2 design matrix giving the children's sex, μ is a 2×4 matrix of means,

$$\mu = \begin{pmatrix} \mu_{b,8} & \mu_{b,10} & \mu_{b,12} & \mu_{b,14} \\ \mu_{g,8} & \mu_{g,10} & \mu_{g,12} & \mu_{g,14} \end{pmatrix},$$

and ε is a 27×4 multivariate normal error matrix.

Hypotheses of (potential) interest

- $\mu_b = \mu_g$ — same profiles for boys and girls,
- $\mu_{sj} = \alpha_s + \beta_j$ — parallel curves,
- $\mu_{s1} = \mu_{s2} = \ldots = \mu_{s4}$ — no time effects,
- same covariance matrices for boys and girls (assumed equal in our model formulation).

These hypotheses can be tested by multivariate analysis, in which no specific assumptions are made about the correlation within subjects.

ANALYSIS OF MULTIVARIATE LINEAR MODEL

Model: $Y = XB + E$:

- parameters B and Σ estimated by
  * $\hat{B} = (X'X)^{-1}X'Y \sim$ least squares estimate,
  * $\hat{\Sigma} = S = (Y - X\hat{B})'(Y - X\hat{B})/(n-p)$ unbiased estimate,
- usual ANOVA methods for single time point parameters $\beta_{1j}, \ldots, \beta_{pj}$,
- hypothesis of general form $\mathrm{H}_0 : ABC = D$, where
  * A is $(a \times p)$ matrix of coef. for "within time" hypotheses, and $\mathrm{rank}(A) = a \leq p$,
  * C is $(n \times c)$ matrix of coef. for "between time" hypotheses, and $\mathrm{rank}(C) = cp$,
  * D is $(a \times c)$ matrix of constants (often zeros),
- test of H_0:
  * 4 different test statistics: Wilk's lambda (likelihood ratio test), Pillai's trace, Hotelling-Lawley trace, and Roy's root statistic,
  * "one-dimensional" cases: agreement w. Hotelling's T^2,
  * except in simple cases, their distribution is known only approximately (by suitable F-distributions, usually indicated in the software),
  * no test generally "best" (power/robustness), but Wilk's lambda performs reasonably well in all situations.

MANOVA for Orthodont Data

Estimation:

$$\hat{\mu} = \begin{pmatrix} \mu_b \\ \mu_g \end{pmatrix} = \begin{pmatrix} 22.9 & 23.8 & 25.7 & 27.5 \\ 21.2 & 22.2 & 23.1 & 24.1 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} 5.4 & 2.7 & 3.9 & 2.7 \\ 2.7 & 4.2 & 2.9 & 3.3 \\ 3.9 & 2.9 & 6.5 & 4.1 \\ 2.7 & 3.3 & 4.1 & 5.0 \end{pmatrix}$$

Testing

- test of parallelism: MANOVA $F = 2.70$, df $= (3, 23)$, $P = 0.070$,
- test of no group differences — two versions:
  * without assuming parallelism: MANOVA, $F = 3.63$, df $= (4, 22)$, $P = 0.020$,
  * assuming parallelism: univariate t-test between two groups

    $$t = 3.05, \quad \text{df} = 25, \quad P = 0.005,$$

- test of no time differences — two versions:
  * without assuming parallelism: Wilk's lambda converted to

    $$F = 11.46, \quad \text{df} = (6, 46), \quad P < 0.0001,$$

  * assuming parallelism: Wilk's lambda converted to

    $$F = 36.4, \quad \text{df} = (3, 23), \quad P < 0.0001,$$

- separate tests of no time differences:

  $$\begin{aligned} \text{boys}: \quad & F = 31.9, \quad \text{df} = (3, 23), \quad P < 0.0001, \\ \text{girls}: \quad & F = 7.09, \quad \text{df} = (3, 23), \quad P = 0.0015. \end{aligned}$$

- conclusions:
  * some indication of non-parallel curves,
  * clear group and time differences.

Drawbacks and Advantages of MV Approach

Limitations of multivariate approach:

- requires balanced data and complete data (usually, if one time point is missing, the entire subject drops out),
- requires sufficient no. of subjects relative to no. of time points[3],
- difficult to deal with time-varying predictors (unless incorporated into a balanced design).
- no extension to hierarchical data structures,
- sensitivity to assumptions: some robustness to normality assumption has been shown, but procedures can be very sensitive to outliers.

— however, MLwiN implementation (multilevel course) relaxes many of the limitations.

Advantages of multivariate approach:

- valid and exact[4] statistical inference,
- avoids specific assumptions on correlation structure,
- may perform better in small/moderate samples[4] than linear mixed models.[5]

[3] Recommended to have more subjects than outcomes in every "cell" formed by subject predictors in order to ensure estimability and adequate power.

[4] Does not refer to inference in MLwiN.

[5] Simulation studies discussed in Davis (2002), Section 6.5.3.