Statistik III

Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

Parameter Estimation

ML and REML estimation

- Parameters may be estimated by *maximum likelihood* methods (ML)
- Alternatively, methods based on the *restricted maximum likelihood* (REML) principle are available
  - They operate on certain contrasts and take into account the number of predictors p, which leads to unbiased estimators for variances and covariances
  - For example, $\hat{\sigma}^2_{ML} = \frac{1}{n}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is a biased estimator, whereas $\hat{\sigma}^2_{REML} = \frac{1}{n-p}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ is unbiased
- Neither approach is superior to the other, both have advantages and limitations
- Likelihood ratio tests of fixed effects, however, must be based on ML estimation

Parameter Estimation

Prediction of random effects

- In some applications (e. g., for predicting individual trajectories), one is interested in "estimating" the random effects $\boldsymbol{v}_i$
- However, random effects are not parameters in the sense of classical statistics, but random variables, and therefore cannot be estimated in the classical way
- Bayesian methods can be used for deriving an optimal prediction of the random effects given the data by determining empirical Bayes estimators (EAP, expected a posteriori) for the posterior distribution of $\boldsymbol{v}_i$ under the condition $\mathbf{Y}_i = \mathbf{y}_i$

Partial Pooling

Prediction

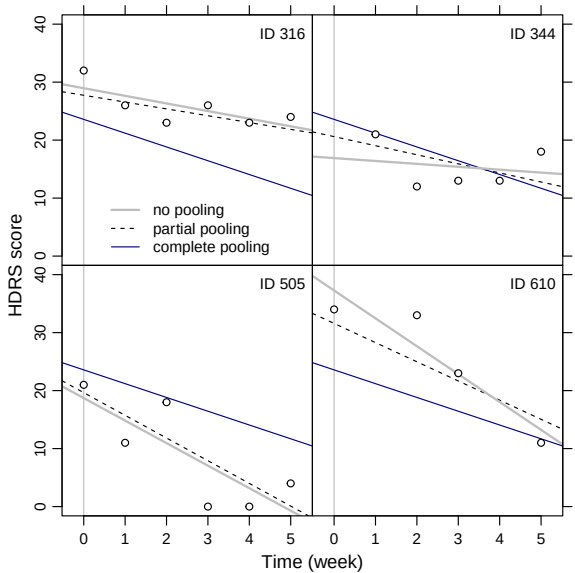
- Given estimates for the fixed and random effects, the prediction for the trajectory of person i is

$$\hat{\mathbf{y}}_i = \mathbf{R}_i \mathbf{\Sigma}_i^{-1} \mathbf{X}_i \hat{\boldsymbol{\beta}} + (\mathbf{I}_{n_i} - \mathbf{R}_i \mathbf{\Sigma}_i^{-1}) \mathbf{y}_i$$

  where $\mathbf{R}_i = \sigma^2 \mathbf{I}_{n_i}$
- Thus, the prediction is computed via a weighted average of the group trend $\mathbf{X}_i \hat{\boldsymbol{\beta}}$ and individual observations $\mathbf{y}_i$ with weights $\mathbf{R}_i \mathbf{\Sigma}_i^{-1}$ and $\mathbf{I}_{n_i} - \mathbf{R}_i \mathbf{\Sigma}_i^{-1}$
- The larger the within-person variance $\mathbf{R}_i$ compared to the total variance $\mathbf{\Sigma}_i$, the more weight is placed on the group trend (sometimes referred to as *partial pooling*)
- This phenomenon is also called *shrinkage* because the prediction of the individual trend "shrinks" to the group trend

Example: Depression and Imipramine

(Reisby et al., 1977)



Polynomial Regression

- Often the assumption of a linear trajectory is too simple, for example, because the change is not constant, but flattens out with time
- In the *quadratic regression model*

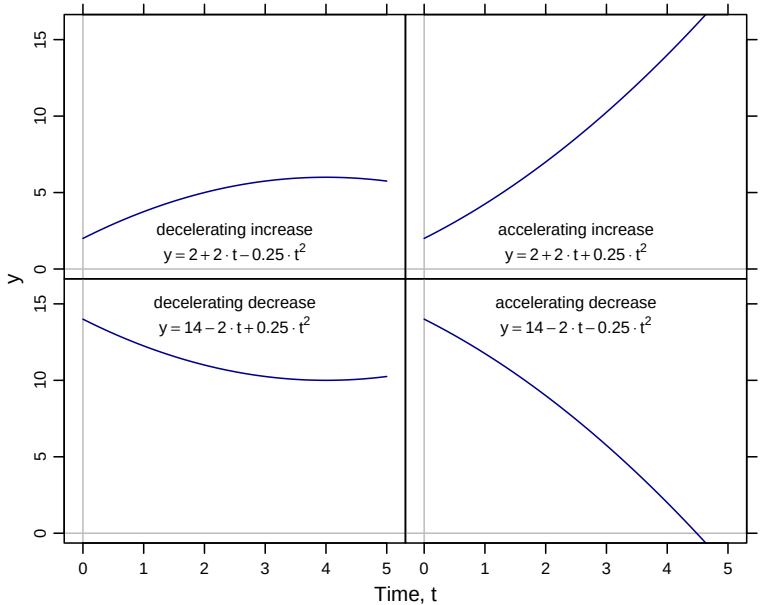
$$\begin{aligned} y_{ij} &= b_{0i} + b_{1i}\, t_{ij} + b_{2i}\, t_{ij}^2 + \varepsilon_{ij} \\ &= b_{0i} + (b_{1i} + b_{2i}\, t_{ij}) t_{ij} + \varepsilon_{ij} \end{aligned}$$

  the linear change depends on time t
- The zero point $t_0 = -b_{1i}/(2b_{2i})$ determined via the partial derivative

$$\frac{\partial y}{\partial t} = b_{1i} + 2b_{2i}\, t_0 = 0$$

  indicates the time at which the positive (negative) trend reverses to a negative (positive) trend

Quadratic Time Trends



Models with Quadratic Trends

- Model with quadratic individual and group trends

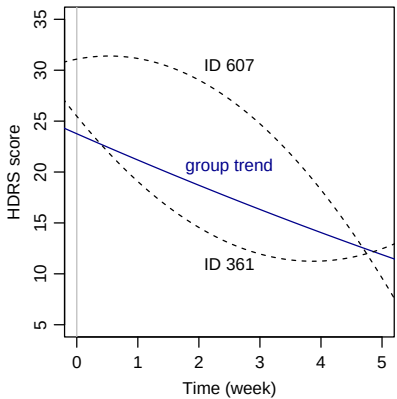
$$
\begin{aligned}
\text{(Level 1)} \quad & y_{ij} = b_{0i} + b_{1i}\, t_{ij} + b_{2i}\, t_{ij}^2 + \varepsilon_{ij} \\
\text{(Level 2)} \quad & b_{0i} = \beta_0 + \upsilon_{0i} \\
& b_{1i} = \beta_1 + \upsilon_{1i} \\
& b_{2i} = \beta_2 + \upsilon_{2i} \\
\text{(2) in (1)} \quad & y_{ij} = \beta_0 + \beta_1\, t_{ij} + \beta_2\, t_{ij}^2 + \upsilon_{0i} + \upsilon_{1i}\, t_{ij} + \upsilon_{2i}\, t_{ij}^2 + \varepsilon_{ij}
\end{aligned}
$$

with

$$
\begin{pmatrix} \upsilon_{0i} \\ \upsilon_{1i} \\ \upsilon_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{\upsilon_0} & \sigma_{\upsilon_0\upsilon_1} & \sigma_{\upsilon_0\upsilon_2} \\ \sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_1} & \sigma_{\upsilon_1\upsilon_2} \\ \sigma_{\upsilon_0\upsilon_2} & \sigma_{\upsilon_1\upsilon_2} & \sigma^2_{\upsilon_2} \end{pmatrix} \right) \text{ i.i.d.}
$$

$$
\boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\, \sigma^2 \mathbf{I}_{n_i}) \text{ i.i.d.}
$$

Model Predictions



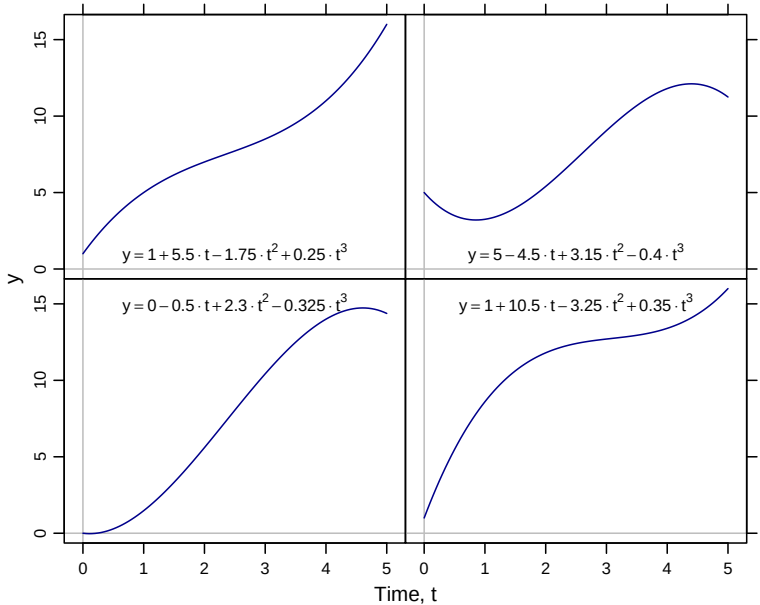
- Averaged across individuals, the group trend turns out to be approximately linear with $\hat{\beta}_1 = -2.63$ and $\hat{\beta}_2 = 0.05$
- Some of the predicted individual trajectories are clearly nonlinear
- Test against a model without individual quadratic trends

$$H_0 \colon \sigma^2_{\upsilon_2} = \sigma_{\upsilon_0\upsilon_2} = \sigma_{\upsilon_1\upsilon_2} = 0$$
$$G^2(3) = 10.98, p = .012$$

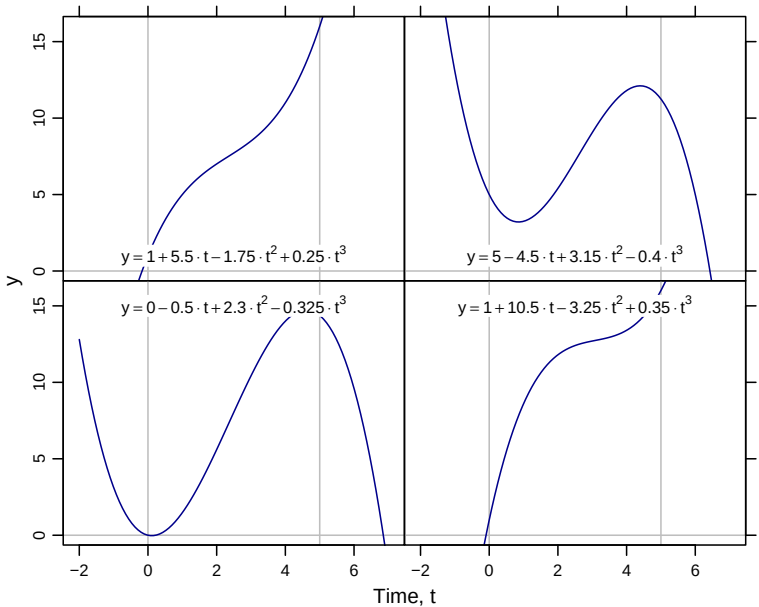
Higher Order Polynomials

- With higher order polynomials, nonlinear time trends can be modeled flexibly and parsimoniously (unlike contrasts in ANOVA models), which is another advantage of mixed models compared to ANOVA models
- For example, they can describe saddle points or reversal points in the time course
- Advantage: The regression model remains linear with respect to the parameters
- Disadvantage: Extrapolations can quickly produce values outside a reasonable range

Cubic Time Trends



Cubic Time Trends: Extrapolation



Example: Depression and Imipramine

Model with quadratic time trends

```
m4 <- lmer(hamd ~ week + I(week^2) + (week + I(week^2) | id),
           dat, REML = FALSE, control = lmerControl(optimizer = "bobyqa"))
```

- When parameter estimation does not converge, recommended steps include
  - Adjust stopping (convergence) tolerances for the optimizer
  - Re-fit the model with another available optimizer
  - See `?convergence`, `?lmerControl` for help

Graphical illustration of model predictions

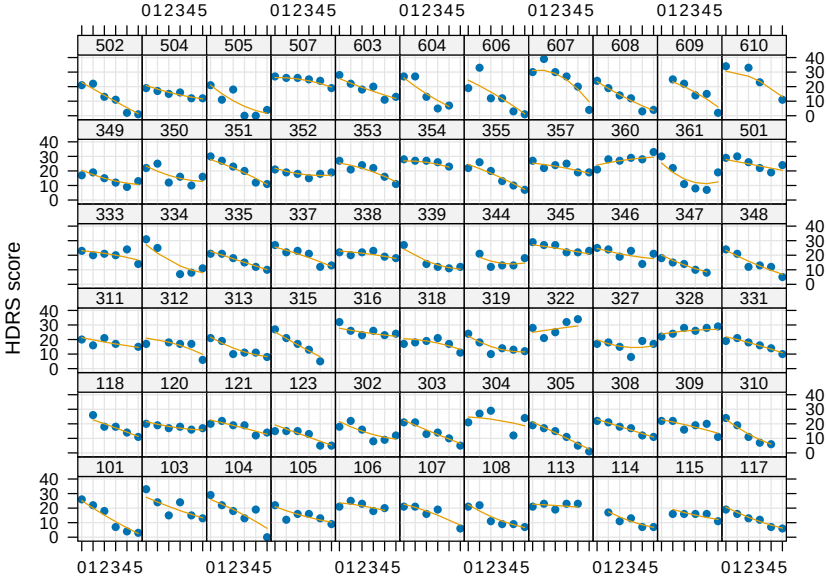
```
xyplot(hamd + predict(m4) ~ week | id, dat, type = c("p", "l"), pch = 16,
  grid = TRUE, distribute.type = TRUE, par.strip.text = list(cex = .8),
  layout = c(11, 6), ylab = "HDRS score", xlab = "Time (week)")
```

Example: Depression and Imipramine

ML parameter estimation (`m4`)

```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: hamd ~ week + I(week^2) + (week + I(week^2) | id)
   Data: dat
      AIC       BIC    logLik -2*log(L)  df.resid
 2227.648  2266.917 -1103.824  2207.648       365
Random effects:
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 3.23
          week        2.58     -0.11
          I(week^2)   0.44     -0.08 -0.83
 Residual             3.24
Number of obs: 375, groups:  id, 66
Fixed Effects:
(Intercept)         week    I(week^2)
     23.760       -2.633        0.051
```

Example: Observed and Predicted Time Trends



Marginal and Conditional Models

Marginal models

- The constraints imposed on the distribution of variable $\mathbf{y}$ in the mixed model by assuming random effects implies the *marginal model*

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\,\boldsymbol{\beta} + \boldsymbol{\delta}_i \\ \boldsymbol{\delta}_i &= \mathbf{Z}_i\,\boldsymbol{\upsilon}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{y}_i &\sim N(\mathbf{X}_i\,\boldsymbol{\beta},\, \boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\Sigma}_{\upsilon}\mathbf{Z}_i' + \sigma^2\mathbf{I}_{n_i})\end{aligned}$$

- The marginal model, however, can also be formulated directly, without assuming random effects, by specifying a general covariance matrix $\boldsymbol{\Sigma}_i$ in a covariance structure model

$$\begin{aligned}\boldsymbol{\delta}_i &\sim N(\mathbf{0},\, \boldsymbol{\Sigma}_i) \\ \mathbf{y}_i &\sim N(\mathbf{X}_i\,\boldsymbol{\beta},\, \boldsymbol{\Sigma}_i)\end{aligned}$$

Marginal and Conditional Models

Conditional models

- Linear mixed models also allow for deriving a distribution of the observed values $\mathbf{y}$ given the random effects
- For this *conditional model* we then have

$$\begin{aligned} \mathbf{y}_i \mid \boldsymbol{\upsilon}_i &\sim N(\mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \boldsymbol{\upsilon}_i, \sigma^2 \mathbf{I}_{n_i}) \\ \boldsymbol{\upsilon}_i &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_\upsilon) \end{aligned}$$

- As its central feature, the observations of a person are conditionally independent given the random effects
- One can also consider more general models (e. g., AR1, MA1, ARMA) that assume dependencies, usually called *autocorrelations*
- These models can be formulated within the framework offered by the `nmle` package

Example: Marginal Covariance Matrix

Reisby et al. (1977): model with quadratic time trends

- Predicted

$$\hat{\mathbf{\Sigma}}_i = \mathbf{Z}_i \hat{\mathbf{\Sigma}}_v \mathbf{Z}_i' + \hat{\sigma}^2 \mathbf{I}_{n_i} = \begin{pmatrix} 20.96 & 9.41 & 8.16 & 6.68 & 4.98 & 3.06 \\ 9.41 & 23.86 & 15.57 & 16.08 & 14.88 & 11.97 \\ 8.16 & 15.57 & 31.07 & 23.11 & 23.26 & 20.98 \\ 6.68 & 16.08 & 23.11 & 38.31 & 30.12 & 30.09 \\ 4.98 & 14.88 & 23.26 & 30.12 & 45.98 & 39.29 \\ 3.06 & 11.97 & 20.98 & 30.09 & 39.29 & 59.11 \end{pmatrix}$$

- Observed

$$\widehat{Cov}(\mathbf{y}_i) = \begin{pmatrix} 20.55 & 10.11 & 10.14 & 10.09 & 7.19 & 6.28 \\ 10.11 & 22.07 & 12.28 & 12.55 & 10.26 & 7.72 \\ 10.14 & 12.28 & 30.09 & 25.13 & 24.63 & 18.38 \\ 10.09 & 12.55 & 25.13 & 41.15 & 37.34 & 23.99 \\ 7.19 & 10.26 & 24.63 & 37.34 & 48.59 & 30.51 \\ 6.28 & 7.72 & 18.38 & 23.99 & 30.51 & 52.12 \end{pmatrix}$$

Example: Marginal Covariance Matrix

Reisby et al. (1977): model with quadratic time trends

- Covariance matrix of random effects $\hat{\mathbf{\Sigma}}_v$

```
Sigma.u <- VarCorr(m4)$id
```

- Marginal covariance matrix $\hat{\mathbf{\Sigma}}_i = \mathbf{Z}_i \hat{\mathbf{\Sigma}}_v \mathbf{Z}_i' + \hat{\sigma}^2 \mathbf{I}_{n_i}$

```
Z <- cbind(1, 0:5, (0:5)^2)
Z %*% Sigma.u %*% t(Z) + sigma(m4)^2 * diag(6)
```

- Empirical covariance matrix $\widehat{Cov}(\mathbf{y}_i)$

```
cov(matrix(dat.raw$hamd, ncol = 6, byrow = TRUE),
    use = "pairwise.complete.obs")
```

Centering Variables

Centering time variables

- If powers of the time variable (t, t^2, t^3, etc.) enter the regression equation, problems may arise due to multicollinearity
  - For example, $t = 0, 1, 2, 3$ and $t^2 = 0, 1, 4, 9$ are almost perfectly correlated
- Centering can mitigate the problem
  - $(t - \bar{t}) = -1.5, -0.5, 0.5, 1.5$ and $(t - \bar{t})^2 = 2.25, 0.25, 0.25, 2.25$ are uncorrelated
- Centering changes the interpretation of the intercepts
  - If uncentered, then intercepts represent differences at the first time point ($t = 0$)
  - If centered, then they represent differences after half the time

Centering Variables

Options for centering in models with two levels

- Level 1
  - Centering around the group mean
  - Centering around the total mean
- Level 2
  - Centering around the total mean
- The interpretation of the intercepts is tied to the zeroes of the predictor variables here as well
- Caution
  - Notice that centering around the respective group means eliminates any differences in the mean values between the groups that may exist in the data
  - In order to avoid this, an additional variable carrying the group means may be introduced

Example: Math Achievements

Data

- The `hsbdataset.txt` file contains data from the National Center for Education Statistics' (NCES; cf. Bryk & Raudenbush, 2002; Singer, 1998, 2003) "High School & Beyond" national survey of U.S. public and Catholic high schools.
- The data set consists of information on 7185 students from 160 schools
- The variables described below can be assigned to two levels
  - Level 1 refers to students
  - Level 2 refers to schools

Example: Math Achievements

Data

Level	Variable	Description
1	`mathach`	Performance in mathematics test
1	`ses`	Socioeconomic status (overall mean 0)
2	`meanses`	Mean socioeconomic status of the school (overall mean 0)
1	`cses`	Centered socioeconomic status of the student (mean for each school 0, difference `ses` − `meanses`)
2	`school`	School ID
2	`sector`	Public (0) or Catholic high school (1)

Example: Math Achievements

Model equation

$$
\begin{aligned}
\text{(Level 1)} \quad y_{ij} &= b_{0i} + b_{1i}\,\texttt{cses}_{ij} + \varepsilon_{ij} \\
\text{(Level 2)} \quad b_{0i} &= \beta_0 + \beta_2\,\texttt{meanses}_i + \beta_4\,\texttt{sector}_i + \upsilon_{0i} \\
b_{1i} &= \beta_1 + \beta_3\,\texttt{meanses}_i + \beta_5\,\texttt{sector}_i + \upsilon_{1i} \\
\text{(2) in (1)} \quad y_{ij} &= \beta_0 + \beta_1\,\texttt{cses}_{ij} + \beta_2\,\texttt{meanses}_i + \beta_4\,\texttt{sector}_i \\
&\quad + \beta_3\,(\texttt{cses}_{ij} \times \texttt{meanses}_i) + \beta_5\,(\texttt{cses}_{ij} \times \texttt{sector}_i) \\
&\quad + \upsilon_{0i} + \upsilon_{1i}\,\texttt{cses}_{ij} + \varepsilon_{ij}
\end{aligned}
$$

with

$$
\begin{pmatrix} \upsilon_{0i} \\ \upsilon_{1i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \boldsymbol{\Sigma}_\upsilon = \begin{pmatrix} \sigma^2_{\upsilon_0} & \sigma_{\upsilon_0\upsilon_1} \\ \sigma_{\upsilon_0\upsilon_1} & \sigma^2_{\upsilon_1} \end{pmatrix} \right) \text{ i.i.d.}
$$

$$
\varepsilon_i \sim N(\mathbf{0}, \boldsymbol{\sigma}^2 \mathbf{I}_{n_i}) \text{ i.i.d.}
$$

Example: Math Achievements

- In this model, by decomposing the socioeconomic status according to the equation

$$\texttt{ses} = \texttt{cses} + \texttt{meanses}$$

  its differential effectiveness is considered at both Level 1 and Level 2
- At the same time, the effect of the type of school is examined via the variable `sector`
- Notice that the formulation of the model assumes dependencies of the slope b_{1i} on both mean socioeconomic status and school type, which is captured by the interactions of `cses` with `meanses` and `sector`, respectively

Example: Math Achievements

Read data

```
dat <- read.table("hsbdataset.txt", header = TRUE)
dat$school <- factor(dat$school)
```

Model

```
library(lme4)
m1 <- lmer(mathach ~ meanses*cses + sector*cses +
                     (1 + cses | school), dat, REML = FALSE)
summary(m1)
```

Interaction between status and school type

```
lattice::xyplot(mathach ~ cses |
  factor(sector, labels = c("Public high school", "Catholic high school")),
  data = dat, groups = school, type = "r",
  xlab = "Socioeconomic status (cses)", ylab = "Math performance (mathach)"
)
```

Example: Math Achievements

ML parameter estimation (`m1`)

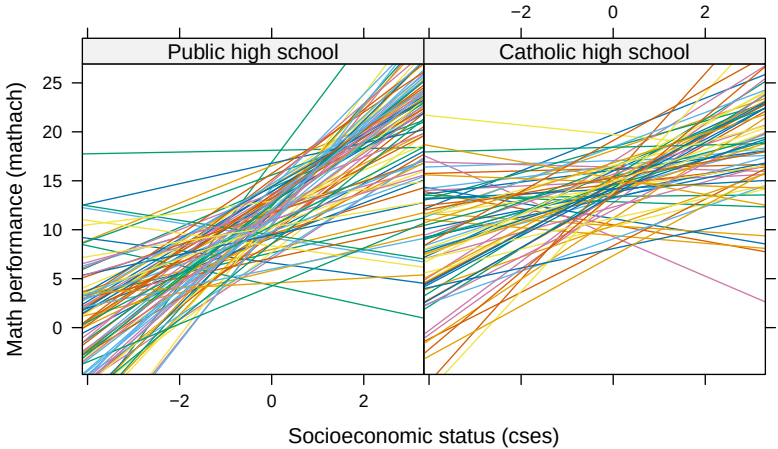
```
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: mathach ~ meanses * cses + sector * cses + (1 + cses | school)
   Data: dat

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 school   (Intercept)  2.319   1.52
          cses         0.065   0.26     0.48
 Residual             36.721   6.06
Number of obs: 7185, groups:  school, 160

Fixed effects:
             Estimate Std. Error t value
(Intercept)     12.11       0.20    61.5
meanses          5.34       0.37    14.6
cses             2.94       0.15    19.1
sector           1.22       0.30     4.0
meanses:cses     1.04       0.30     3.5
cses:sector     -1.64       0.24    -6.9
```

Example: Math Achievements

Interaction between status and school type



Example: Fixed Effects

- Mean math achievement (i. e., for a student with a mean `cses` score in a school with a mean `meanses` score) is 12.11 in Public high schools and $12.11 + 1.22$ in Catholic high schools
- Effects of socioeconomic status at the two levels
  - The effect at the student level depends on the type of school: math achievement increases by 2.94 points in Public high schools and by $2.94 - 1.64 = 1.30$ points in Catholic high schools for a unit increase in `cses`
  - Higher math achievements are obtained in schools with higher mean socioeconomic status
  - In addition, the dependence of math achievement on `cses` scores is more pronounced in schools with higher `meanses` scores (estimated interaction > 0)

Example: Random Effects

- The estimate $\hat{\sigma}_{v_0} = 1.52$ of the standard deviation of mean school performance leaves room for improving prediction by including additional predictors
- However, there is little variation in the dependence of math achievement on cses across schools ($\hat{\sigma}_{v_1} = 0.26$), which should also be noted when interpreting the reported correlation of 0.48
- The corresponding covariance has an estimated value of $\hat{\sigma}_{v_0 v_1} = 0.48 \cdot \hat{\sigma}_{v_0} \cdot \hat{\sigma}_{v_1} = 0.19$
- These results suggest a simplified model of the dependence of math achievement on cses, where the intercept, but not the slope varies across schools

Summary

Regression models with fixed and random effects

- allow for adequately modeling hierarchical data structures
  - longitudinal data
  - individuals organized in groups (e.g., students in classes, or schools)
- allow for adequately modeling the correlations occurring in this context
- offer an optimal trade-off between individual and aggregate data analysis
  - while individual differences are modeled, information aggregated over the sample is exploited, too

In psychology, therefore, linear mixed models allow for integrating differential and general psychological aspects within a common theoretical framework

References

Bryk, A.S., & Raudenbush, S.W. (2002). Hierarchical linear models. Thousand Oaks, CA: Sage Publications.

Hedeker, D. R., & Gibbons, R. D. (2006). Longitudinal data analysis. Hoboken, NJ: Wiley. Chap. 5

Reisby, N., Gram, L. F., Bech, P., Nagy, A., Petersen, G. O., Ortmann, J., Ibsen, I., Dencker, S. J., Jacobsen, O., Krautwald, O., Sondergaard, I., & Christiansen, J. (1977). Imipramine: Clinical effects and pharmacokinetic variability. *Psychopharmacology*, 54, 263–272.

Singer, J.D. (1998). Using SAS Proc Mixed to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 24, 323-355. [Link]

Singer, J.D., & Willett, J.B. (2003). Applied longitudinal data analysis: Modeling change and event occurrence. New York: Oxford University Press.

Verbeke, G. & Molenberghs, G. (2000). Linear mixed models for longitudinal data. New York: Springer.