UMC Utrecht

Epidemiology and Big Data
Mixed Models part 3: Technical issues in multilevel/longitudinal modelling

Rebecca Stellato

Overview Part 3: technical issues in mixed models

- Choosing a model
  - LRT/AIC
  - REML vs ML estimation in mixed models
  - A model building strategy for MM
  - Testing random effects (variances)
  - Testing fixed effects
- Checking assumptions of the model
- Effect of centering explanatory variables
- Polynomials in linear mixed models
- More than 2 levels

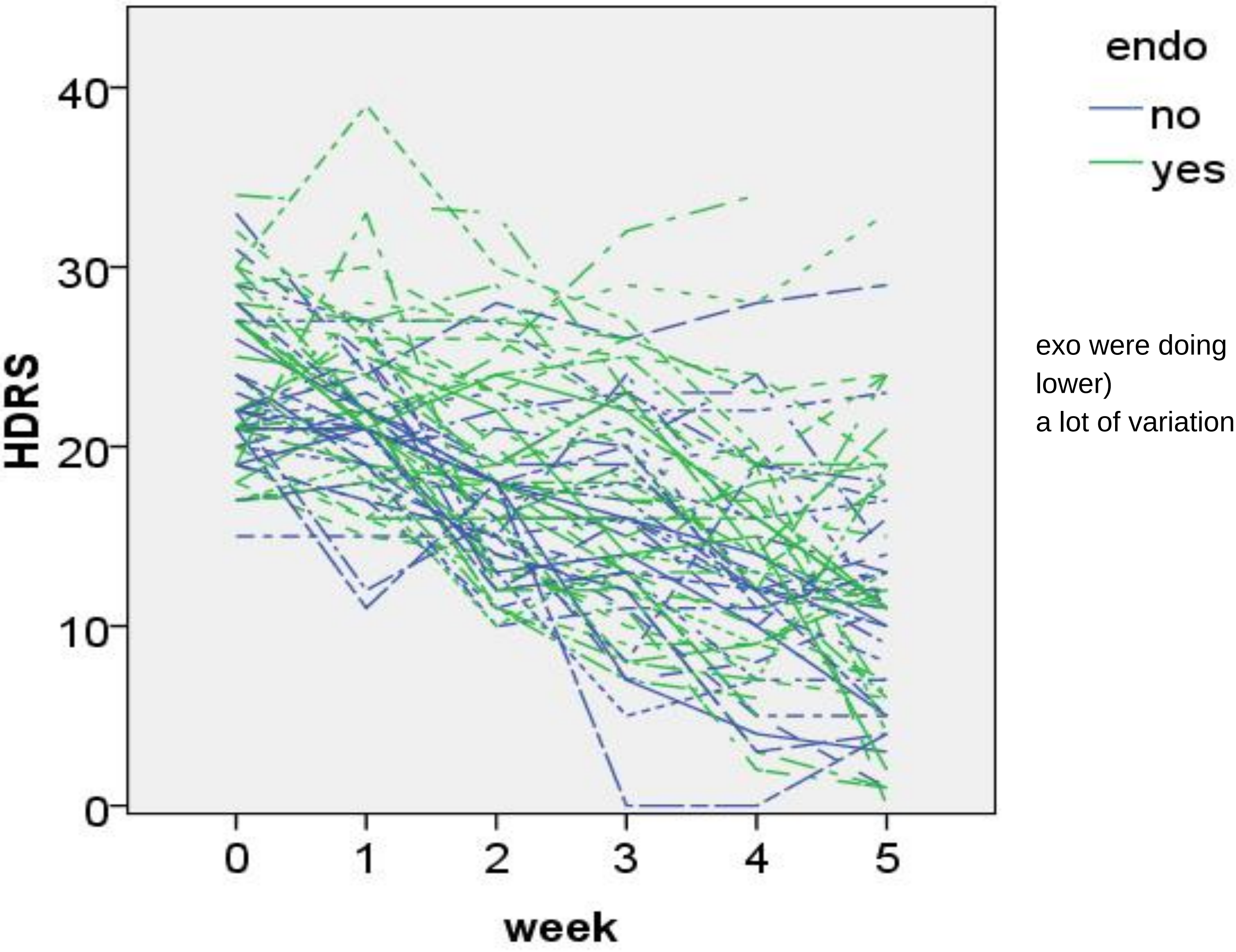
Reisby example, revisited

Hamilton Depression Rating Score on 66 patients measured at 6 time points

how to check the assumptions of the analysis

```
        hdrs.0 hdrs.1 hdrs.2 hdrs.3 hdrs.4 hdrs.5
hdrs.0  1.000  0.493  0.410  0.333  0.227  0.184
hdrs.1  0.493  1.000  0.494  0.412  0.308  0.218
hdrs.2  0.410  0.494  1.000  0.738  0.669  0.461
hdrs.3  0.333  0.412  0.738  1.000  0.817  0.568
hdrs.4  0.227  0.308  0.669  0.817  1.000  0.654
hdrs.5  0.184  0.218  0.461  0.568  0.654  1.000
```

Reisby example, revisited



exo were doing a little bit better (scores lower)
a lot of variation within and between

Reisby example, revisited

- We talked about several logical models to fit the variance-covariance matrix of the HDRS scores over time…
  - CPM with heteregeneous AR(1) correlation structure
  - CPM with unstructured correlation structure
  - LME with random intercept + slope (for time)

all 3 are fairly reasonable given what we know so far but now how to decide which one is the best model

- …and several less logical models:
  - CPM with identity correlation structure
  - CPM with compound symmetry/LME with random intercept
  - CPM with homogeneous AR(1) correlation structure

Testing in Linear Mixed Models

linear mixed effects model

- To decide which LMM fits the data best we can use likelihood-based methods:
  - Likelihood Ratio Test (LRT)

    smaller model nested in bigger model - if not statistical significant then better suited to use the smaller model

    - LRT can be used to test nested models (one is a special case of the other)
    - based on the χ^2-distribution
  - Akaikes Information Criterium (AIC)
    - combination of likelihood and # parameters used in the model (d.f.)
    - model with the lowest AIC (high likelihood with few parameters) is deemed best

lowest AIC= highest likelihood = best model

can be used for nested and non-nested models

(Restricted) Maximum Likelihood Estimation

so far always ML

- Mixed models: maximum likelihood used to estimate fixed regression coefficients and variances of random effects

restricted variation better

  - likelihood quite complex, solved by iteration until convergence

when they don't change much anymore from iteration to iteration

- (Empirical Bayes methods used to estimate individual random effects)
- Problem with ML estimation:
  - variance parameters (residual variance, variance(s) of random effect(s)) biased downwards

like in regular linear regression when we use method of least squares - difference between - is your variance divided by n or n-1 ->n-1 = unbiased estimate

- Solution: REstricted (or: REsidual) Maximum Likelihood (REML)
  - gives unbiased estimates of variance parameters
  - BUT: adjusts likelihood for number of covariates in model, so cannot be used to compare models that differ w.r.t. fixed parts of model

we can't use REML if comparing models that differ on the fixed part of the model. Doesn't work under restricted maximum likelihood assumption

When to use ML, REML?

- Testing models that differ in variance components:
  - REML will give interpretable LRT, AIC
  - so will ML maximum likelihood will also do
- Testing models that differ in fixed effects:
  - only ML will give interpretable LRT, AIC
- Reporting results (esp if you include the random components):
  - use REML!

When to use ML, REML?

- Leading me to suggest the following model-building strategy:
  1. Start with full fixed model and (using ML estimation), select appropriate random part of model comparing intercepts and random slope or random parts of the model
  2. With the random part chosen, (using ML or REML estimation) try to reduce fixed part of model
  3. Once you have your final model: run that model once more using REML; this is the model you present to your audience
- Testing random effect(s):
  - variance parameters are never <0 variance always 0 or greater 0
  - LRT (REML/ML) for random effects: chi-square test, but divide p-value by 2 because it's a onesided test
  - AIC also okay ~~(but please: no AIC + 2)~~
- Testing fixed effect(s):
  - LRT (ML only!) for fixed effects: chi-square test, usual p-value
  - AIC okay (only under ML)

Reisby example, comparing correlation structures

- Four models with the same fixed structure (endo*week, 12 degrees of freedom) and different random parts.
- Compare these using ML-based methods (LRT and / or AIC):

endo vs exo
5 time points
5 interactions of endo and time points =12

Model	# cov par	-2*logLike	AIC
Identity	1	2388.027	2414.027
Comp Symm	2	2277.381	2305.381
Unstructured	21	2183.227	2249.227
AR(1) homogen	2	2221.847	2249.847
AR(1) heterogen	7	2207.462	2245.462

Reisby example, comparing correlation structures

- From the comparison of the AIC's
  - Taking dependence into account greatly improves the model fit
  - Assuming equal variances and equal correlations is not a good option for these data
  - The parsimonious homog. AR(1) not worse than unstructured
  - The AR(1) with heterogeneous variances is best

- We could also have used LRTs (for nested models only), with corrected p-values:
  - Homogeneous vs heterogeneous AR(1) (for instance):
    - LRT = 2221.847 -2207.462 = 14.385 with 7-2= 5 df; p = 0.0133/2 = 0.0067
    - heterogeneous is significantly better than homogeneous AR(1)

larger better is always better than the reduced model.

Reisby example, comparing correlation structures

- We can also compare LMEs with *linear* time trend with one another:
  - fixed: time, endo, time*endo
  - random: intercept vs. int+slope

```
> anova(lme.ris, lme.ril)
        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lme.ris     1  8 2230.929 2262.345 -1107.465                        
lme.ril     2  6 2294.137 2317.699 -1141.069 1 vs 2 67.20798  <.0001
```

dif. in parameters =2

Model with rand int+slope is better than rand int only (LRT or AIC)

Reisby example, comparing correlation structures

- No LRT to compare LMEs with *linear* time trend to models with CPM
- But we can use AIC to compare these non-nested models

Model	**AIC**
CPM Unstructured	2249.227
CPM AR(1) heterogen	2245.462
LME rand int + slope	2230.929

Model with rand int+slope is best according to AIC

Reisby example, fixed part of the model

- Now we have the random structure, we’ll look at the fixed part of the model
- Three possibilities:
  - only time
  - endo + time
  - endo*time (both main effects + interaction)
- We use ML estimation for testing the fixed part of the model

Reisby example, fixed part of the model

```
> lme2.ris<-update(lme.ris, fixed=hdrs ~ time+endo)
> lme3.ris<-update(lme.ris, fixed=hdrs ~ time)
> anova(lme.ris, lme2.ris, lme3.ris)

         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lme.ris      1  8 2230.929 2262.345 -1107.465
lme2.ris     2  7 2228.933 2256.422 -1107.467 1 vs 2 0.004160  0.9486
lme3.ris     3  6 2231.037 2254.599 -1109.519 2 vs 3 4.104108  0.0428
```

model with random intercept and slope fits way better than just with random intercept

- interaction not significant: no evidence that time effect differs for the groups
- effect of endo (just) significant: evidence for (small) difference between depression scores of people with and without endogenous depression

Reisby example, final model with REML

taking out interaction reducing fixed part of the model

```
> lme3.ris.reml <- update(lme2.ris.CAR1, method="REML")
> summary(lme2.ris.reml)
Linear mixed-effects model fit by REML
 Data: reisby.long
       AIC      BIC    logLik
  2228.116 2255.548 -1107.058

Random effects:
 Formula: ~time | id
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 3.490342 (Intr)
time        1.457808 -0.287
Residual    3.494719
```

Reisby example, final model (cont.)

```
Fixed effects: hdrs ~ time + endo
                Value Std.Error  DF    t-value p-value
(Intercept) 22.492881 0.7598098 308  29.603306  0.0000
time        -2.380472 0.2103154 308 -11.318581  0.0000
endo         1.956867 0.9658720  64   2.026011  0.0469
 Correlation:
     (Intr) time
time -0.318
endo -0.704 -0.008

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.73520482 -0.49503123  0.03559898  0.49317021  3.62063687

Number of Observations: 375
Number of Groups: 66
```

Reisby example, final model (cont.)

```
> intervals(lme2.ris.reml)
Approximate 95% confidence intervals

 Fixed effects:
                  lower      est.     upper
(Intercept) 20.99780674 22.492881 23.987956
time        -2.79430883 -2.380472 -1.966635
endo         0.02731607  1.956867  3.886418
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: id
                           lower       est.      upper
sd((Intercept))        2.6340279  3.4903416 4.62503996
sd(time)               1.1419605  1.4578084 1.86101482
cor((Intercept),time) -0.5695496 -0.2870567 0.05608577

 Within-group standard error:
   lower     est.    upper
3.194723 3.494719 3.822884
```

“Clinical” conclusions imipramine example

- There was no significant interaction between time and group
  - time trends same for patients with endogenous and exogenous depression: the lines run parallel
- There was a significant main effect for group
  - at any given point in time, patients with endogenous depression have HDRS scores on average 1.96 (95% CI: 0.03 – 3.89) points higher than those without
- The effect of time is statistically significant
  - For patients with both endogenous and exogenous depression, HDRS scores decrease, on average, by 2.4 (95% CI: 2.0 - 2.8) points per week
  - On average 5*2.33 = 11.9 points in the course of the study
  - Am & Eur guidelines suggest that a 3-point change is clinically relevant

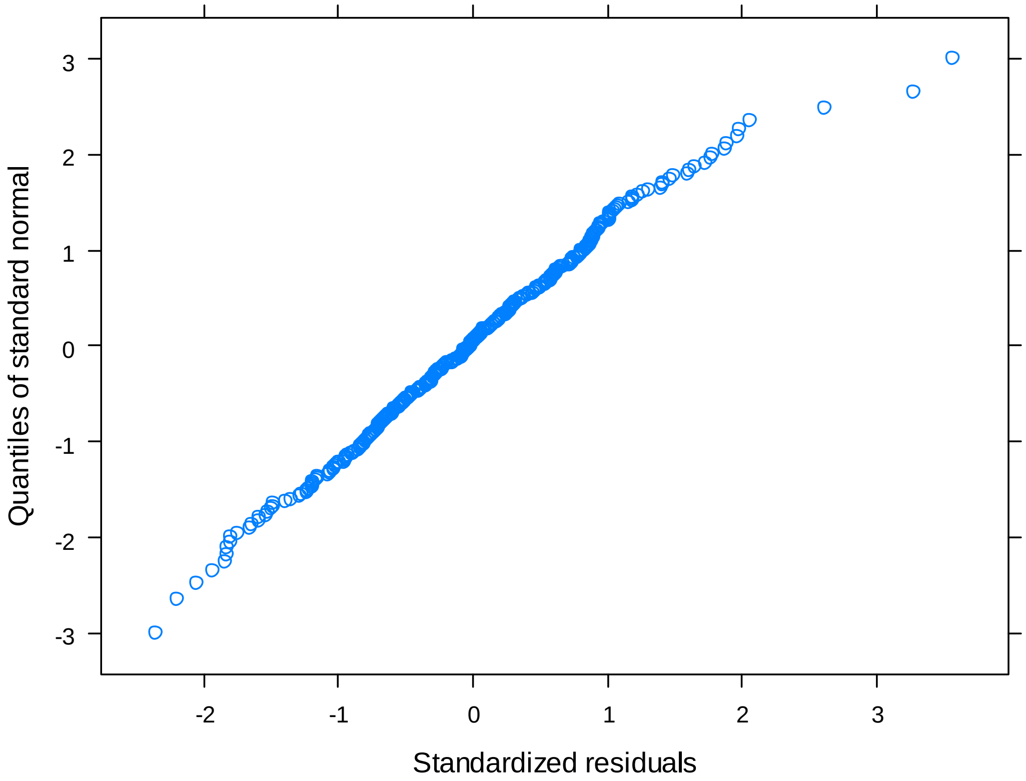
- Before presenting these results, we need to check our model assumptions!

Checking assumptions of the model

- Model assumptions:
  - linearity (if we use time – or other covariates – as linear)
    - check with individual plots, spaghetti plots, residual plots
  - normality of residuals
  - normality of random intercepts (& slopes, if used)
    - these three can be saved and checked using Q-Q plots, boxplots, histograms
    - but: generally not helpful
      1. because deviations from normality probably not a big problem for inference on fixed effects (if your interest is in inference on random effects, there could be a problem)
      2. model 'inflicts' normality on the random effects, so normality of the estimated random effects may partly reflect model assumptions
  - independence of residuals (once fixed and random effects are taken into account)
    - as in linear models: keep your fingers crossed!

Checking assumptions of the model in R

Diagnostic plots for (level-1) residuals: Q-Q plot of residuals

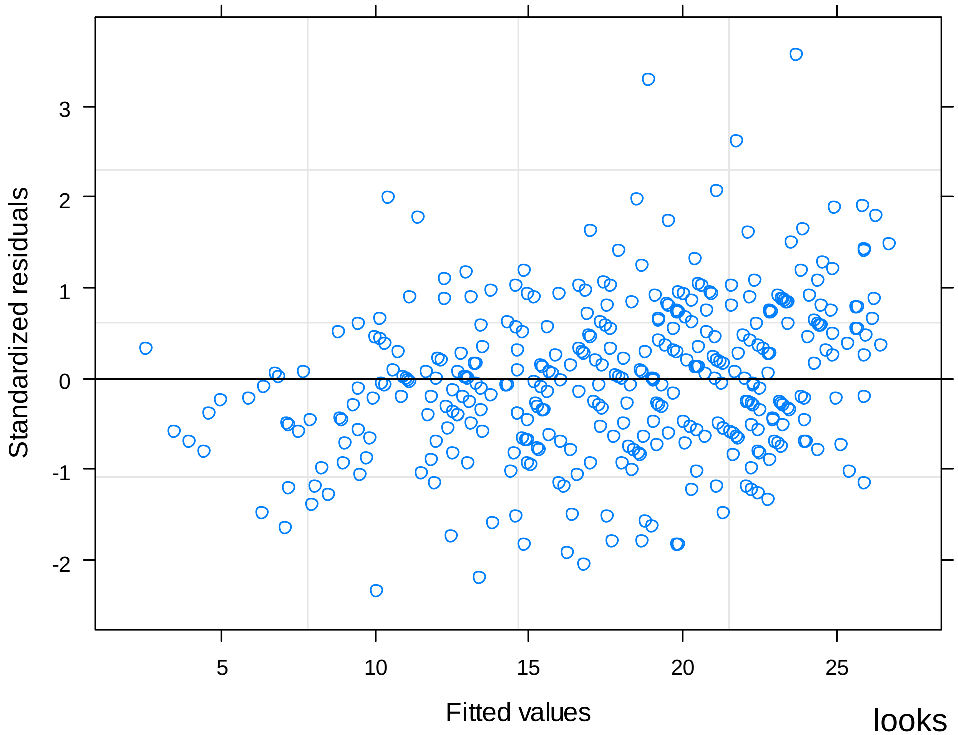


- Aside from three outliers, no departures from normality

very clean plot - looks good

Checking assumptions of the model in R

Diagnostic plots for (level-1) residuals: residuals vs. fitted

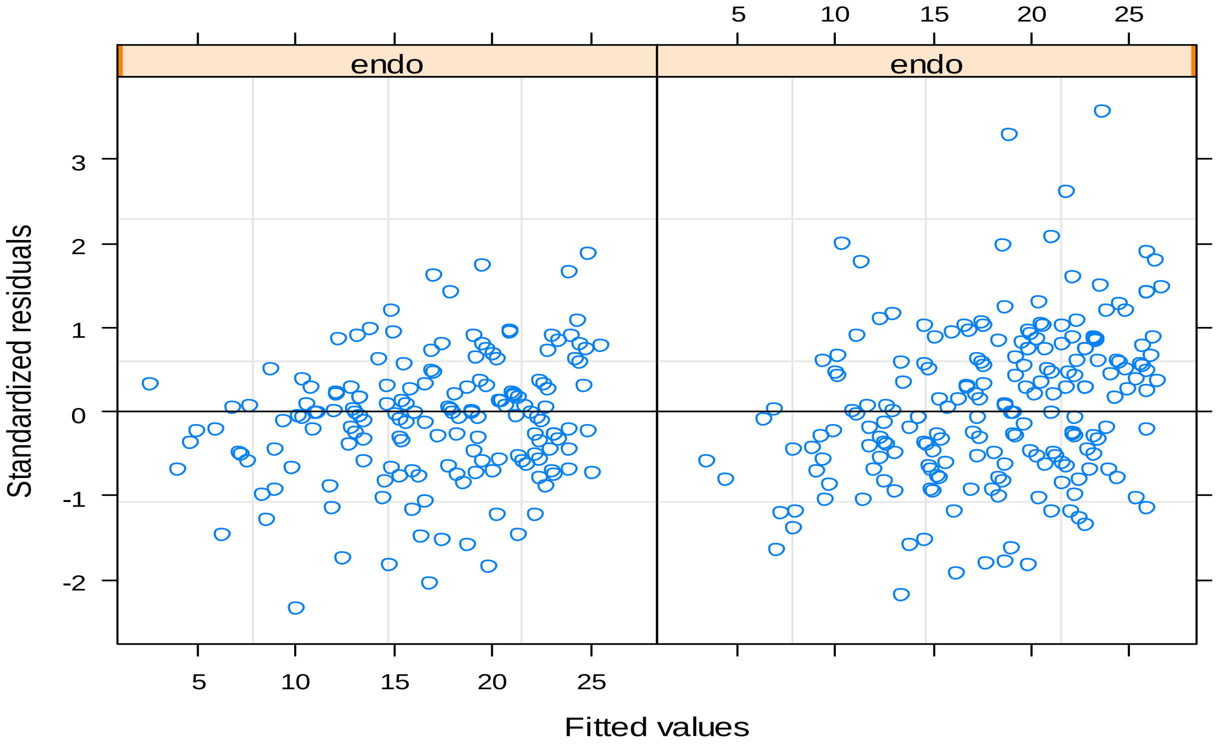


- (We're hoping for a graph with no patterns)
- This looks a bit problematic: a slight trend towards higher residuals with higher fitted values
- Problems with linearity and/or missing covariates?

looks like a trend in the data slightly towards higher values

Checking assumptions of the model in R

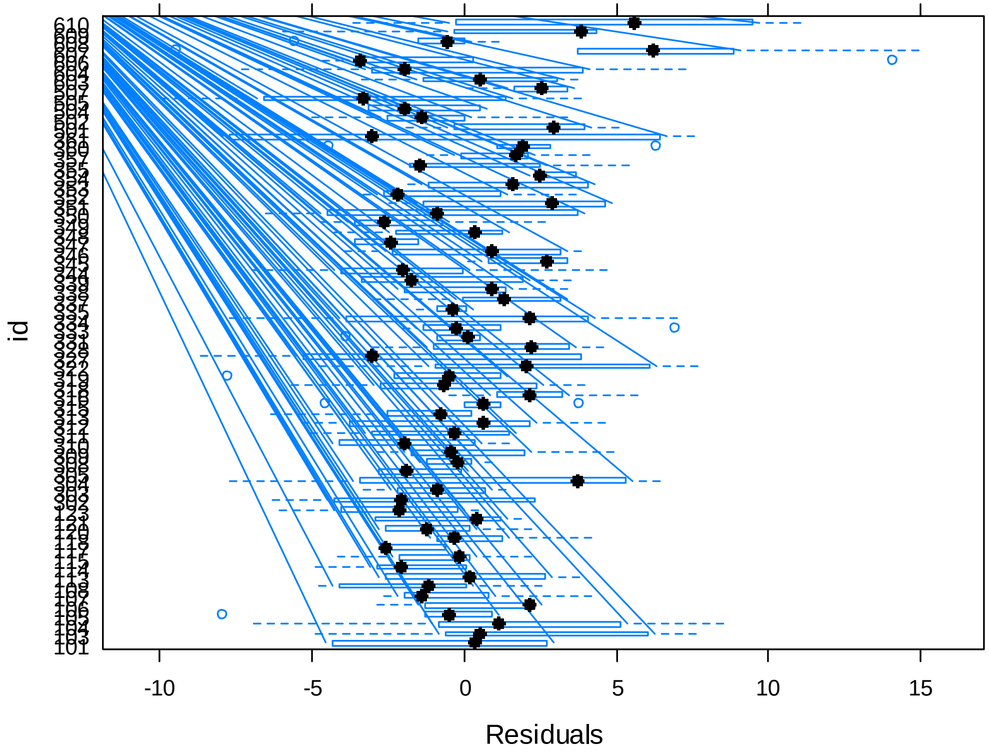
Diagnostic plots for (level-1) residuals: residuals vs. fitted by endo



The problems we saw in the whole sample are present in both groups

Checking assumptions of the model in R

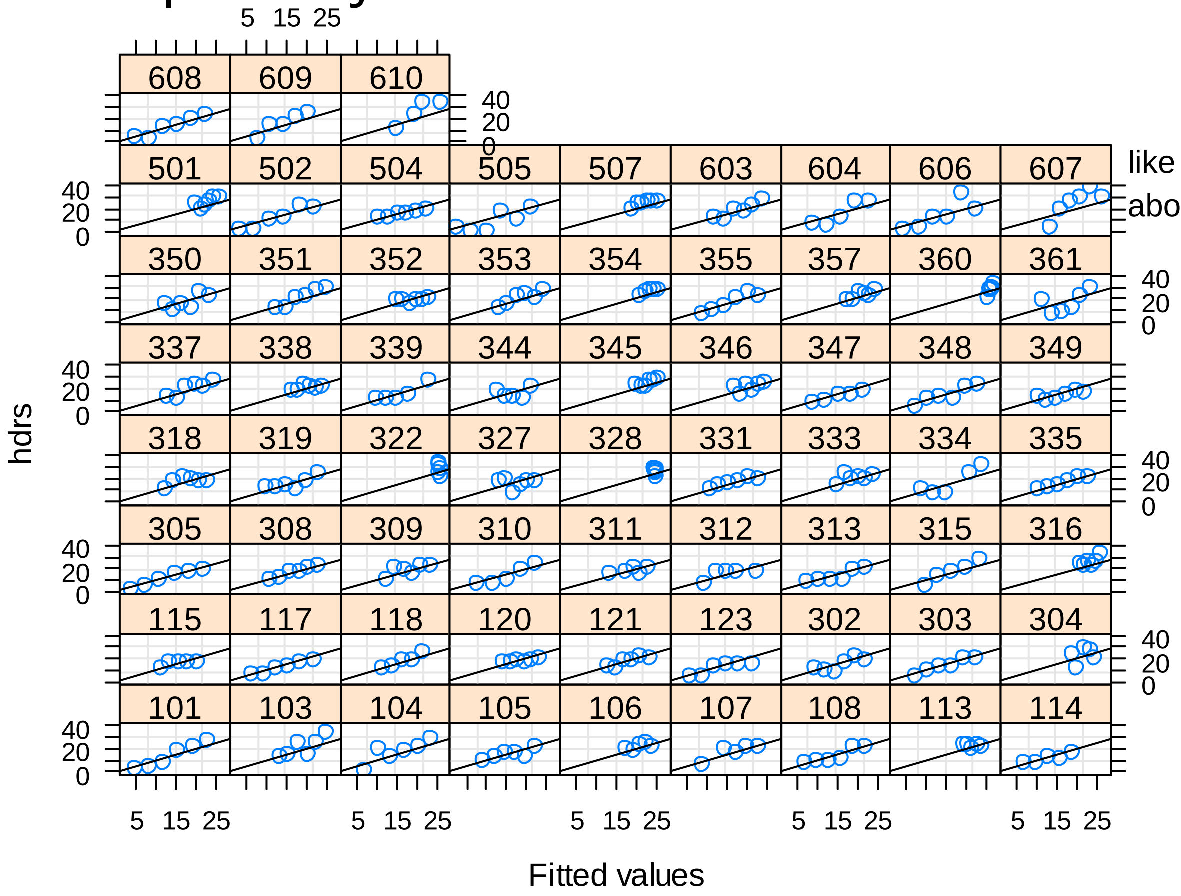
Boxplots residuals per subject



We can use this plot to check for large outliers in residuals per person

Checking assumptions of the model in R

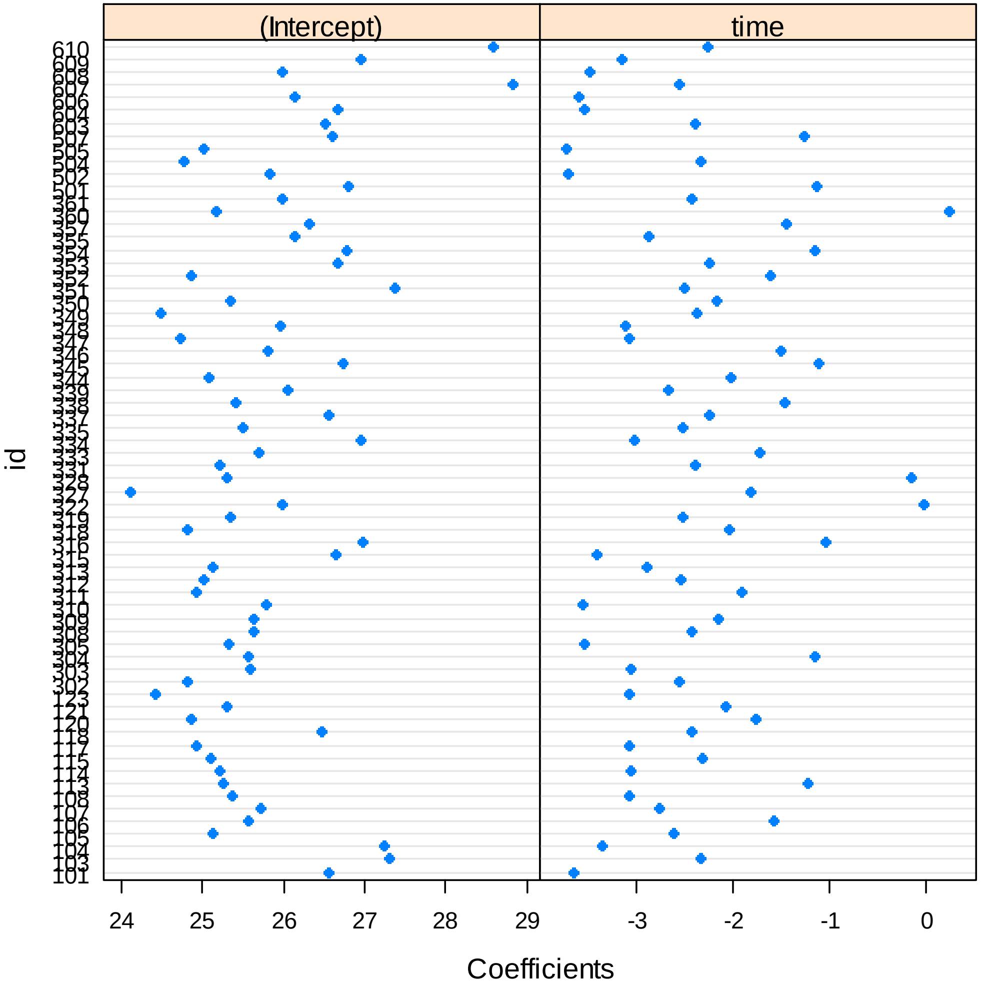
Observed vs. fitted per subject



like this one where the values are above

This plot can be used to check for individuals with poor agreement between observed and fitted HDRS scores.

Checking assumptions of the model in R



- Quick plot of random intercepts and slopes for time
- (We're not looking for patterns here, just for large outliers)
- No obvious outliers, though a few high-ish slopes

Statistical conclusions imipramine example

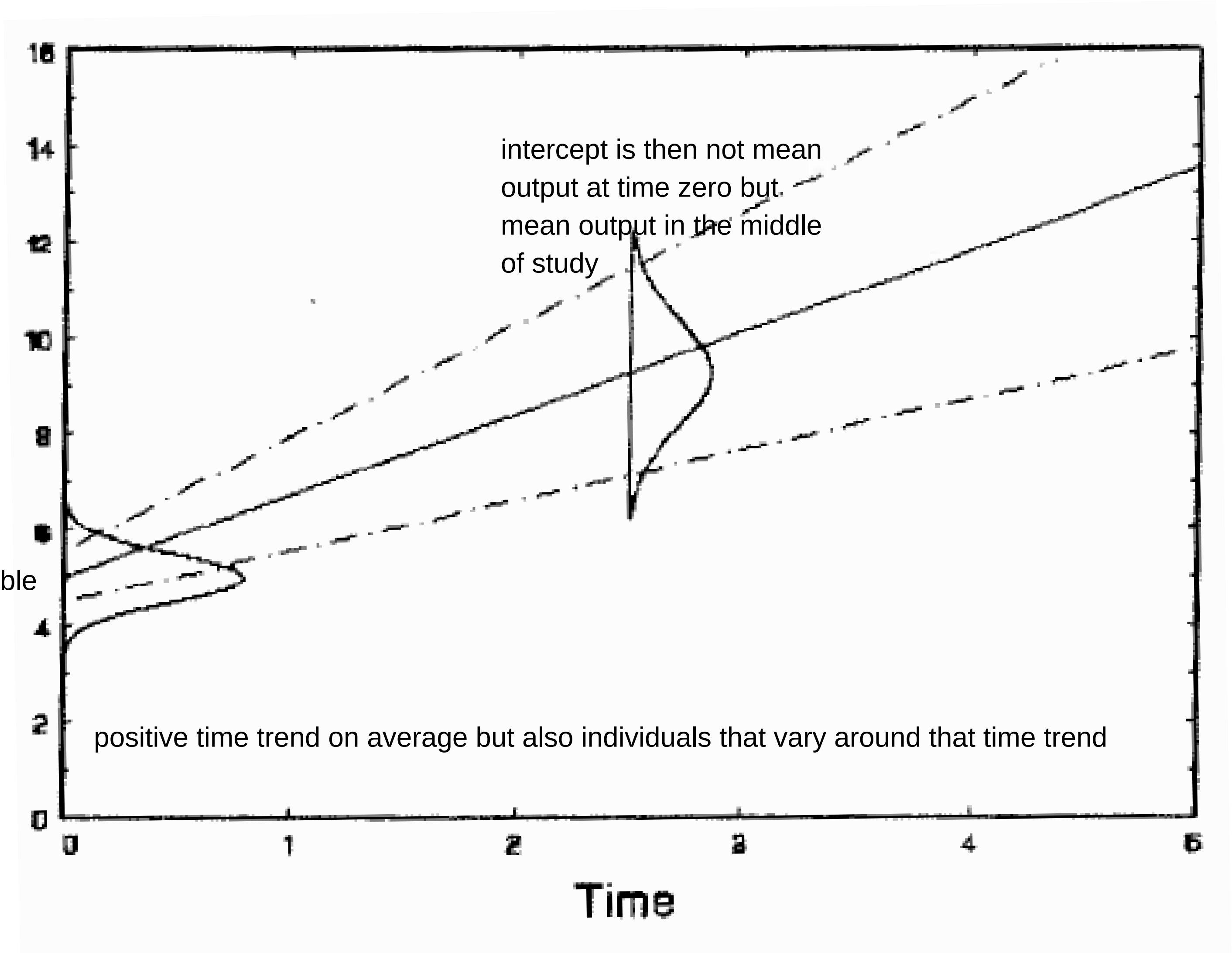
- A model with fixed linear time effect and a fixed effect for group, random intercept and random slope for time for the within subject residuals seems to provide the "best" fit for these data
- The assumptions of normality for the level-1 and level-2 random effects seem reasonable
- The assumption of constant variance of residuals (given the random effects) might be violated

➢ Now we may present our results (with caution)

Centering explanatory variables

- In the London schools dataset, both the outcome and the intake test had been "centered" (actually, both were standardized)
- What is the effect of centering an explanatory variable?
    - changes the interpretation of the fixed intercept
    - can change the variance of the random intercepts, and the correlation of random intercepts with random slopes

Centering explanatory variables



intercept is then not mean output at time zero but mean output in the middle of study

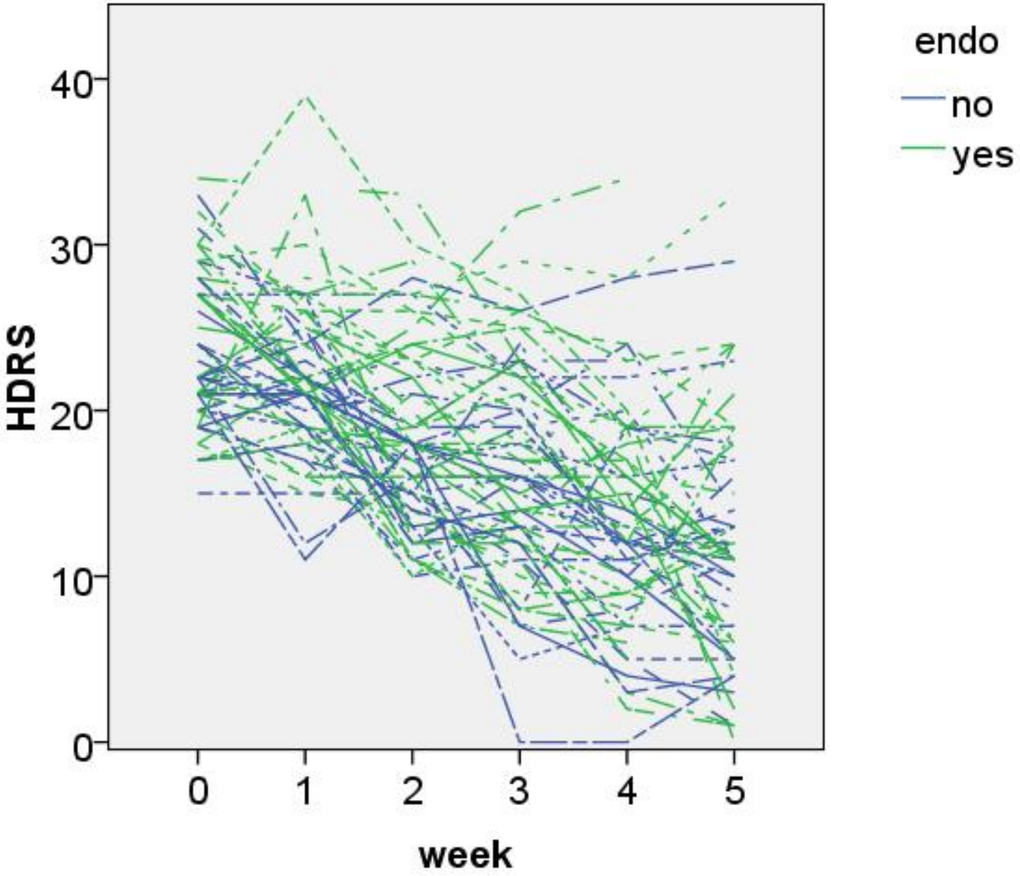
center time variable

positive time trend on average but also individuals that vary around that time trend

Source: Hedeker and Gibbons, Longitudinal Data Analysis. Wiley & Sons 2006, p. 60
https://www.wiley.com/en-nl/Longitudinal+Data+Analysis-p-9780471420279
Inclusion is permitted according to the agreement with the publisher.

Centering explanatory variables

- Take Reisby data, center time (week 2.5 becomes 0 point)
  - for sake of simplicity, using model with just fixed effect of time, random effects for intercept and time



Centering explanatory variables

Parameter estimate	Model 1 (time not centered)	Model 2 (time centered)
Fixed: intercept average dep. score at beginning of study	25.95	20.01
Fixed: time	-2.38	-2.38
Random: intercept (s.d.)	intercept will change too when not fixed 4.25	3.65
Random: slope of time (s.d.)	1.46	1.46
Random: corr (int-slope)	tends to be negative at time point zero -0.582	0.322 half way through study
Residual (s.d.)	3.49	3.49

only thing that really changes is random intercept and more/less variance that will affect the interpretation of the fixed intercept
usually makes it possible to interpret the intercept better - as usually all variables would have to be zero

Estimates of intercept, variation of random intercepts and correlation rand int-slope all changed!

Linear mixed effects models with polynomial terms

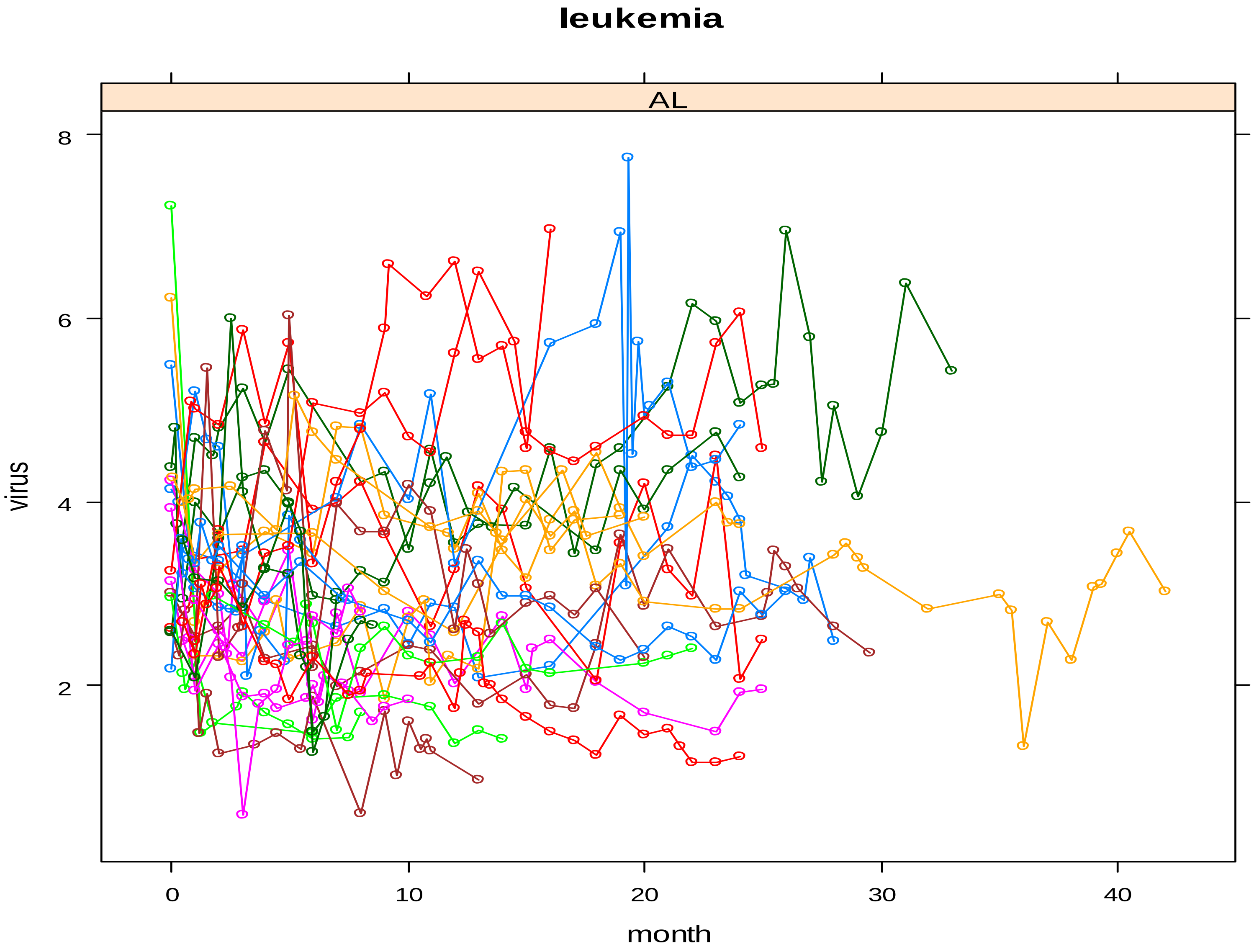
- Instead of linear trends over time, it is quite possible to observe non-linear trends (think of children’s growth, for instance)
- There are many non-linear models that can be used within mixed models (beyond the scope of this course)
- It is possible to fit polynomials as part of a “linear” mixed model

Example: Herpes Antibody Levels

- 45 children suffering from
  - solid lump tumour (N=18)
  - leukemia (N=27)
- Measurements of antibody levels to a herpes virus taken during hospital visits for courses of chemotherapy
- Duration: 1 mo - 3 yrs (median 12 mo)
- Intervals between measurements differed per child
- Questions:
  - are antibody levels affected by chemo?
  - if so, is change related to cancer type?

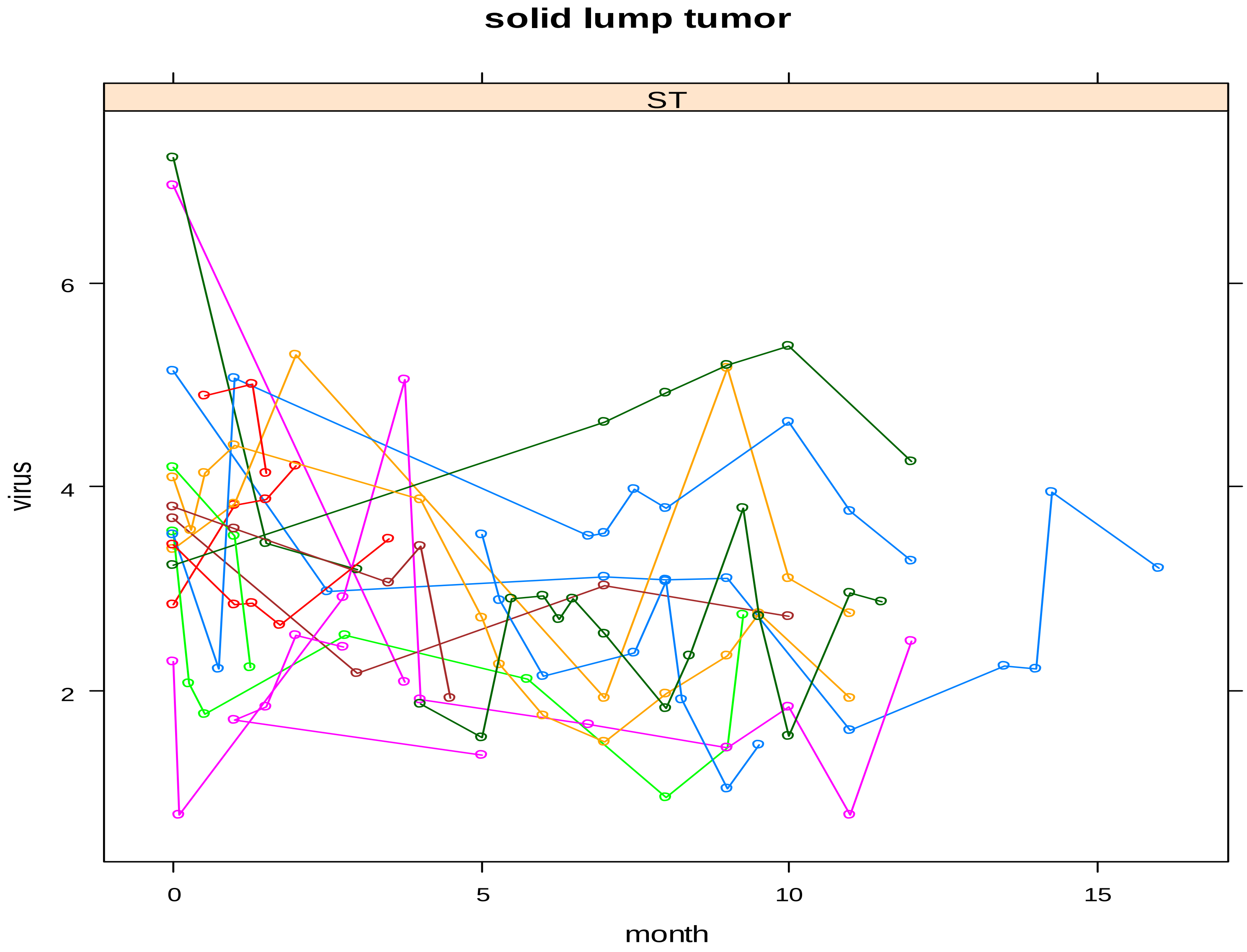
mixed effect model as they all start at different times and for different time periods
what's the change in antibody levels

Linear mixed effects models with polynomial terms



seems non linear

Linear mixed effects models with polynomial terms



Linear mixed effects models with polynomial terms

just to show that it's possible to fit non linear time trend int linear models

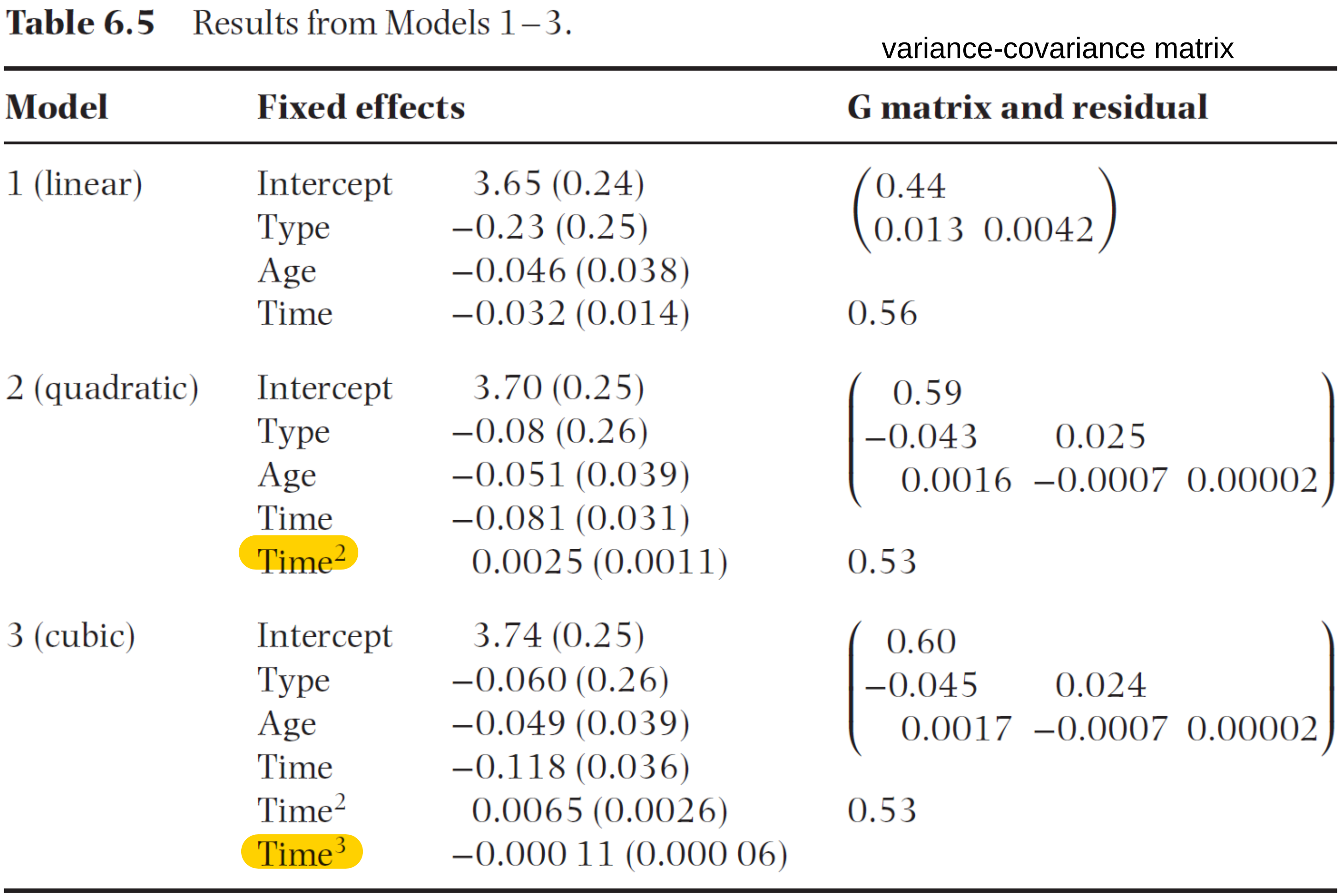


Table 6.5 Results from Models 1–3.

variance-covariance matrix

Model	Fixed effects		G matrix and residual
1 (linear)	Intercept	3.65 (0.24)	$\begin{pmatrix} 0.44 & \\ 0.013 & 0.0042 \end{pmatrix}$
	Type	−0.23 (0.25)	
	Age	−0.046 (0.038)	
	Time	−0.032 (0.014)	0.56
2 (quadratic)	Intercept	3.70 (0.25)	$\begin{pmatrix} 0.59 & & \\ -0.043 & 0.025 & \\ 0.0016 & -0.0007 & 0.00002 \end{pmatrix}$
	Type	−0.08 (0.26)	
	Age	−0.051 (0.039)	
	Time	−0.081 (0.031)	
	Time^2	0.0025 (0.0011)	0.53
3 (cubic)	Intercept	3.74 (0.25)	$\begin{pmatrix} 0.60 & & \\ -0.045 & 0.024 & \\ 0.0017 & -0.0007 & 0.00002 \end{pmatrix}$
	Type	−0.060 (0.26)	
	Age	−0.049 (0.039)	
	Time	−0.118 (0.036)	
	Time^2	0.0065 (0.0026)	0.53
	Time^3	−0.000 11 (0.000 06)	

every child can vary around the parabular (fixed quadratic effect) - variance of time^2 is very small so when time cubed they only added to the fixed part

Source: Brown & Prescott, Applied Mixed Models in Medicine, 3nd Edition. Wiley, 2015, p. 272

Other possibilities for nonlinear trends

just to mention it

- Orthogonal polynomials
- Natural cubic splines
- Nonlinear mixed models

Three-level models

- So far: two levels
  - children within schools, patients within hospitals
  - measurements within individuals over time
- What about three levels?
  - children within classrooms within schools
  - longitudinal measurements within patients within hospitals

Analyzing three-level models

- Variance at 3 levels

probably want to add random effects to ecplai

  - random effects (which??) at 2 levels
- Variables measured at 3 levels?
  - main effects
  - "cross-level" interactions (SES of school * SES of child, gender of teacher * gender of child)
- Think carefully about design

question and design should lead to logical design - possible sources of variation and effects in lower level that could differ in other levels

  - possible sources of variation
  - effects at lower level that could possibly differ at higher level
    - teacher-level variables (gender, experience) could have different effects at different schools
    - child-level variables (gender, entrance exam score) could have different effects in different classrooms or at different schools
- Think about research question: simplicity vs generalizability

always a trade of between generalizing and it being too simple

Example three-level data

- Monday lab: multi-center hypertension trial: 27 centers, 193 patients, 4 post-randomization DBP
- Sources of variation:
  - centers:
    - may serve different populations, with (on average) higher or lower BP
  - patients:
    - patients vary greatly in their blood pressure levels
      - age, gender, baseline BMI, treatment
    - patients may vary (greatly?) in trend over time
  - measurements in time:
    - BP varies considerably from moment to moment, day to day within individuals
      - stress level, tx adherence, BMI at the moment

Example three-level data

- Design:
  - randomized trial, so interest in tx & tx*time
  - hospital-level variables: none provided
  - patient-level variables: treatment
  - measurement (time)-level variables: none provided
- Fixed effects:
  - (intercept,) time (linear or categorical?), tx, tx*time
  - baseline DBP (why?)
- Random effects?
  - differences in avg DBP among centers → random intercept per center
  - difference in tx effect per center? → random tx per center (tricky!!)
  - difference in time trend per center? → random time trend per center
  - differences among patients → random intercept (& time trend) per patient

Summary technical issues MM

- Model building better done in protocol
- Otherwise: use LRT or AIC to build random part of model, then to simplify fixed part of model
- Use ML estimation for likelihood-based tests
- Use REML estimation for presenting results
- Some model assumptions (linearity, normality of res) can be checked
- Centering explanatory variables has effect on interpretation of several parameters
- “Linear” mixed models may also include polynomials
- 3+ levels also possible (complicated, but possible)