Week 7: Mixed Models

MATH-516 Applied Statistics

Tomas Masak

Feb 20th 2023

Section 1

Course Organization (Update)

Content

- **Week 1**: Intro
  - Project 1: Snow Data
- Week 2: Linear Models - Practical Recap
- **Week 3**: Logistic Regression
  - Project 2: Online Shopping Data
- Week 4: Generalized Linear Models
- **Week 5**: Poisson Regression
  - Project 3: Premier League Data
- Week 6: more on GLMs
- **Week 7**: Linear Mixed Models
  - Project 4: U.S. Presidential Elections
- Free Week: Easter Holidays
- Week 8: Linear Mixed and Multilevel Models

Content (cont.)

- **Week 9**: Time Series
  - Project 5: Global Warming
- Week 10: Time Series Regression
- **Week 11**: Functional Data Analysis
  - Project 7: First Wave of Covid in the US
- Week 12: Functional PCA
- **Week 13**: Statistical Consulting
- Week 14: **Oral Exam**
  - discussing your submitted projects

Evaluation: remains the same as announced on Week 1, with Statistical consulting replacing Project 7 (active participation + writing up a suggested solution to a presented problem). Work on Projects 6 and 7 can be combined (to be considered a single submission) and will not be examined.

Section 2

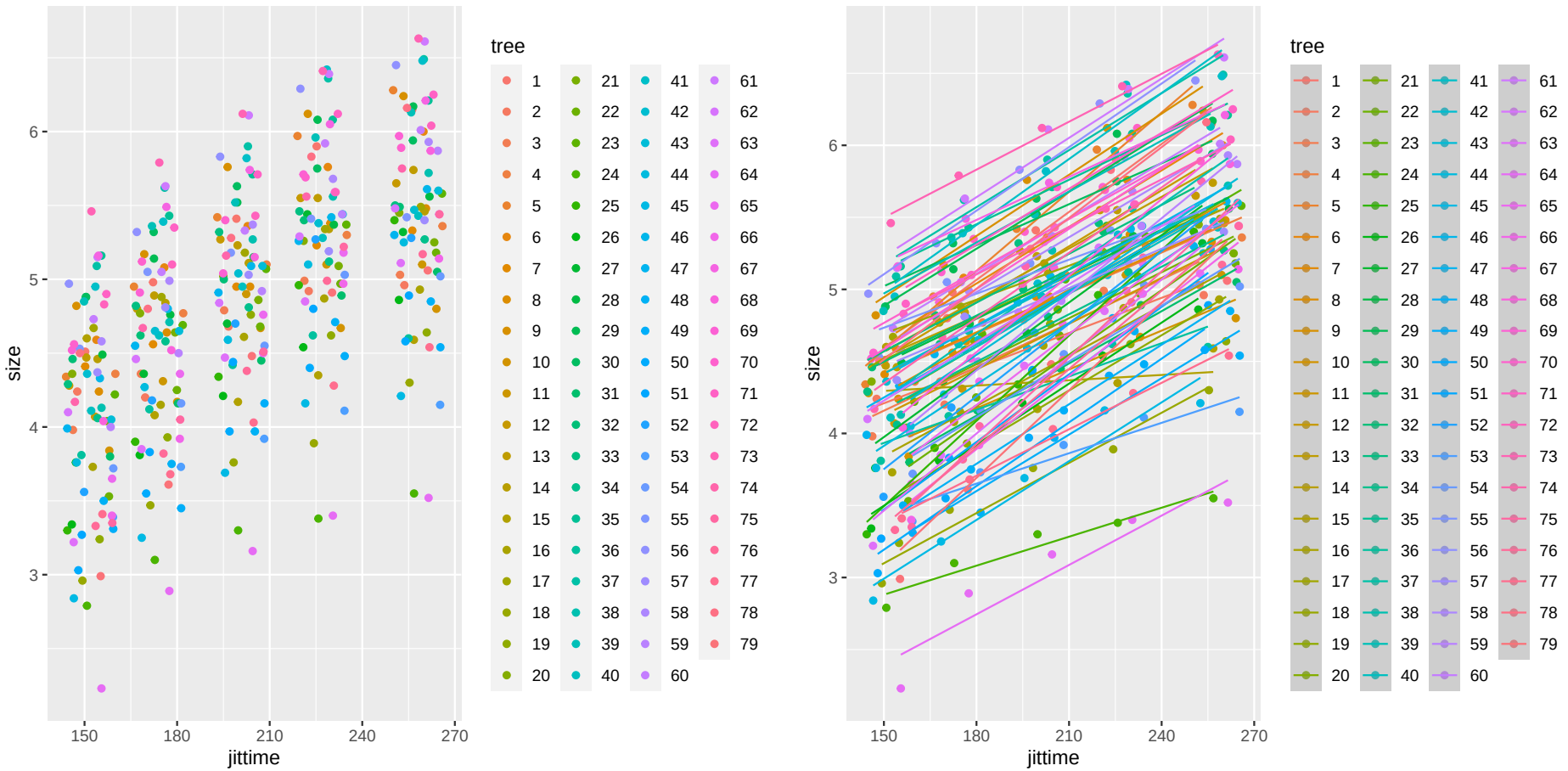
Example: Tree Growth

Data

- log-size (log-height+2log-diameter) of 79 (Sitka spruce) trees measured repeatedly in about 1-month intervals
  - each tree measured 5-times
  - 54 trees grown in ozone-enriched environment (`treat=1`) and 25 were control

```
##   size time tree treat
## 1 4.51  152    1 ozone
## 2 4.98  174    1 ozone
## 3 5.41  201    1 ozone
## 4 5.90  227    1 ozone
## 5 6.15  258    1 ozone
## 6 4.24  152    2 ozone
```

Data Displayed



Right: individual line for every tree corresponding to model

```
y ~ tree*time
```

Models

```
m1 <- lm(size~(time+I(time^2))*tree,data=Sitka)
m0 <- lm(size~(time+I(time^2)), data=Sitka)
anova(m0,m1)
```

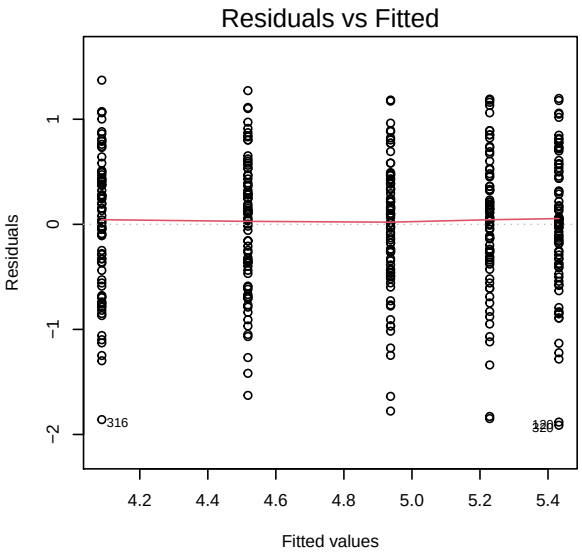
```
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    391 157.107
2    237   6.267 154    150.84 37.043 < 2.2e-16 ***
```

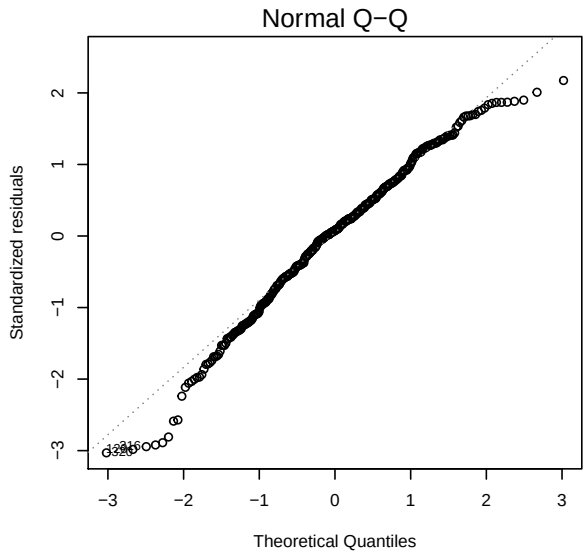
- `m0` allows for a separate curve for the two treatment groups
- `m1` allows for a separate curve for every tree (so `m0` is a submodel of `m1`)

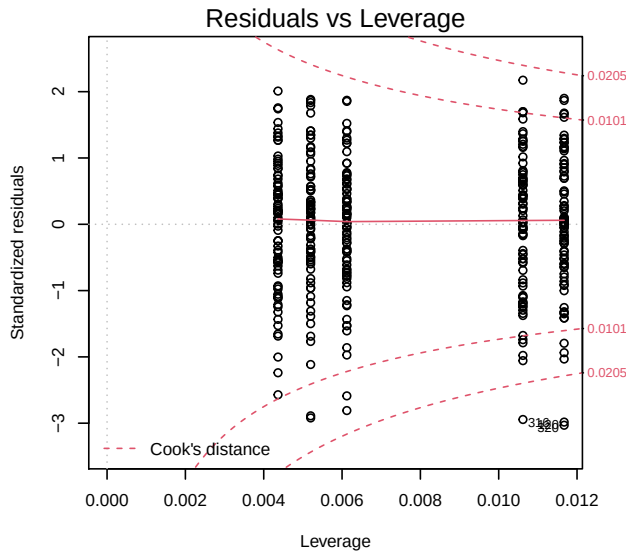
Problems:

- `m1` cannot be simplified to `m0`, but the effect of interest (`treat`) cannot be fitted without this simplification, because every single tree is either treatment or control
  - also what if we had low number of observations for some trees and couldn't afford to fit `m1`?
- but assumptions of `m0` are clearly violated

Diagnostics







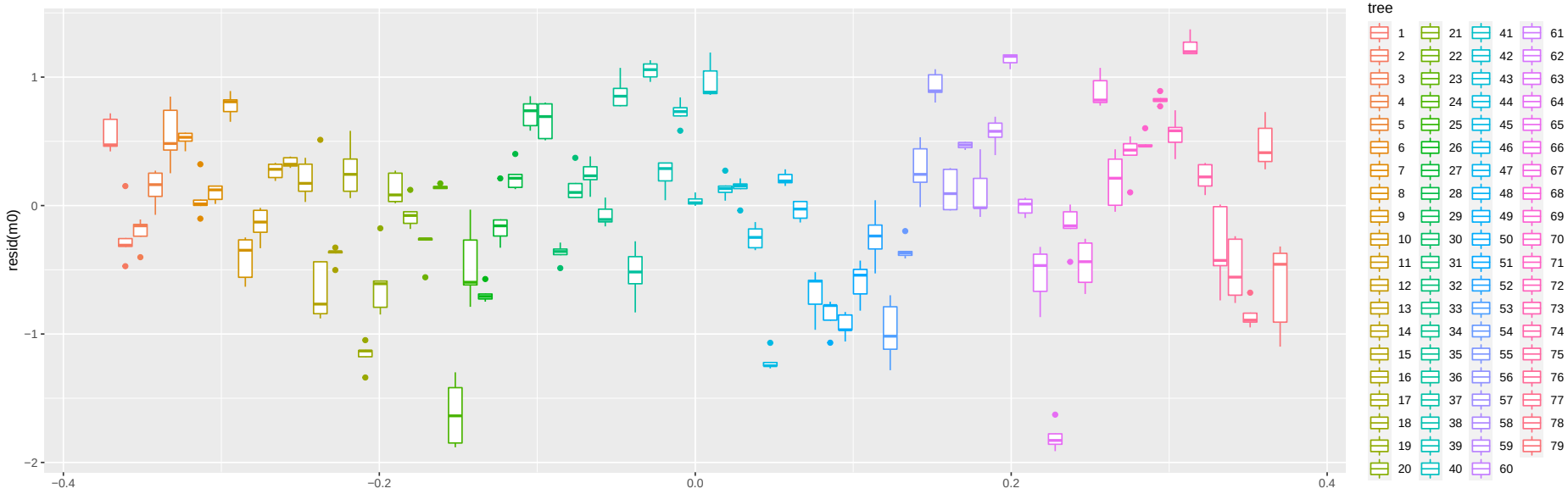


Figure: Residuals grouped by tree.

Section 3

Linear Mixed Models

Definition

- regressions with a large no. of coefficients some of which are themselves being modelled
- extend the linear model

$$Y = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim \mathcal{N}_N(0, \sigma^2 \mathbf{I}_N)$$

to the **linear mixed model**

$$Y = \mathbf{X}\beta + \mathbf{Z}b + \epsilon, \quad b \sim \mathcal{N}_q(0, \mathbf{C}), \quad \epsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$$

  - $\mathbf{X}$ and $\mathbf{Z}$ are known design matrices
  - $\beta \in \mathbb{R}^p$ are fixed (non-random) parameters (effects)
  - $b \in \mathbb{R}^q$ are random effects with mean 0 and covariance matrix $\mathbf{C}$
    - independent of ϵ
- parameters: β, $\mathbf{C}$ and σ^2

Fitting the Model (ML method)

- the linear mixed model has its log-likelihood $\ell(\beta, \mathbf{C}, \sigma^2)$
- if we knew $\mathbf{C}$, we could rewrite the model to

$$Y = \mathbf{X}\beta + e, \quad e \sim \mathcal{N}_n(0, \sigma^2\mathbf{W}), \quad \mathbf{W} = \mathbf{I}_n + \frac{1}{\sigma^2}\mathbf{Z}\mathbf{C}\mathbf{Z}^\top$$

  then the solution $\widehat{\beta}_{\mathbf{C}}$ and $\widehat{\sigma}^2_{\mathbf{C}}$ would be given explicitly by weighted least squares
  - imagine a reparametrization "if we knew $\mathbf{C}/\sigma^2$" instead
- consider the profile log-likelihood for $\mathbf{C}$: $\ell_p(\mathbf{C}) = \ell(\widehat{\beta}_{\mathbf{C}}, \mathbf{C}, \widehat{\sigma}^2_{\mathbf{C}})$

Algorithm: starting from an initial $\mathbf{C}^{(0)}$ alternate until convergence for $l = 1, 2, \ldots$ between

1. calculation of $\widehat{\beta}_{\mathbf{C}^{(l-1)}}$ and $\widehat{\sigma}^2_{\mathbf{C}^{(l-1)}}$ by weighted least squares
2. updating $\mathbf{C}^{(l)} = \arg\min_{\mathbf{C}} \ell_p(\mathbf{C})$ by Newton's method (itself iterative)
   - since Newton works well given good starting values, one rather runs EM algorithm for a while (treating b as unobserved data) before switching to this scheme

Fitting the Model (REML method)

- start by integrating out β from the log-likelihood
  - it actually has a closed form and is equivalent to working with likelihood for $\mathbf{A}Y$ such that $\mathbb{E}\mathbf{A}Y = 0$, i.e. it is just a simple transformation of the problem
- use iterative solver
  - closed form for σ^2
  - inner iteration for $\mathbf{C}$
- finally obtain β as with the ML method

ML vs. REML:

- REML is often preferred (and set as default) since it can lead to unbiased variance estimators (which ML never does)
- but REML depends on parametrization of the fixed effects
  - if one wants to compare models with different $\mathbf{X}$ using likelihood criteria, ML needs to be used!
- ML and REML are asymptotically equivalent

Example: Tree Growth

Rewrite the model in a form that shows the grouping:

$$Y_k = \mathbf{X}_k\beta + \mathbf{Z}_k b_k + \epsilon, \quad b_k \sim \mathcal{N}_q(0, \mathbf{C}), \quad \epsilon_k \sim \mathcal{N}_{N_k}(0, \sigma^2 \mathbf{I}_{N_k})$$

- b_i and ϵ_i are i.i.d. for $k = 1, \ldots, K$
- $\{b_i\}_{k=1,\ldots,K} \perp\!\!\!\perp \{\epsilon_k\}_{k=1,\ldots,K}$

Specifically, in the tree growth example `m1` can be replaced by:

- K is the no. of trees
- $N_k = 5$ (for all k) is the number of measurements per tree
- $$Y_k = \begin{pmatrix} Y_{k1} \\ \vdots \\ Y_{k5} \end{pmatrix} \qquad \mathbf{X}_k = \begin{pmatrix} 1 & t_{k1} \\ \vdots & \vdots \\ 1 & t_{k5} \end{pmatrix} = \mathbf{Z}_k$$
- $$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad b_k = \begin{pmatrix} b_{k0} \\ b_{k1} \\ b_{k2} \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Example & Predictors of Random Effects

Fixed effect model: $\mathbb{E}[Y_{ki} \mid X_{ki} = t_{ki}] = \beta_{k0} + \beta_{k1} t_{k1} + \beta_{k2} t_{ki}^2$

- or something similar depending on the parametrization of `tree`, e.g. with `contr.sum` it would be for all but the last tree:

$$\mathbb{E}[Y_{ki} \mid X_{ki} = t_{ki}] = (\beta_0 + \beta_{k0}) + (\beta_1 + \beta_{k1}) t_{ki} + (\beta_2 + \beta_{k2}) t_{ki}^2$$

Mixed effect model:

$$\mathbb{E}[Y_{ki} \mid X_{ki} = t_{ki}, b_k] = (\beta_0 + b_{k0}) + (\beta_1 + b_{k1}) t_{ki} + (\beta_2 + b_{k2}) t_{ki}^2$$

- $\operatorname{cov}(Y_{ki}, Y_{kj}) = \operatorname{cov}(b_{k0} + b_{k1} t_{ki} + b_{k2} t_{ki}^2, b_{k0} + b_{k1} t_{kj} + b_{k2} t_{kj}^2) \neq 0$

In general (no need to read too much into the formula):

$$\widehat{b}_k = \mathbf{C}\mathbf{Z}_k^\top (\mathbf{Z}_k \mathbf{C} \mathbf{Z}_k^\top + \sigma^2 \mathbf{I}_{N_k})^{-1} (Y_k - \mathbf{X}_k^\top \widehat{\beta})$$

- called *predictors* since these are not parameters but random variables, but they are also sort of *shrinkage estimators*, because
- vaguely: $\widehat{b}_{kl}$ is somewhere between 0 and the $\widehat{\beta}_{kl}$ in the `contr.sum` parametrization above

Example: Tree Growth

Consider a simpler model, where only the intercept is random:

$$\mathbb{E}[Y_{ki} \mid X_{ki} = t_{ki}, b_k] = (\beta_0 + b_k) + \beta_{k1} t_{ki}$$

and the corresponding fixed-effect-only model `y ~ tree+time`.

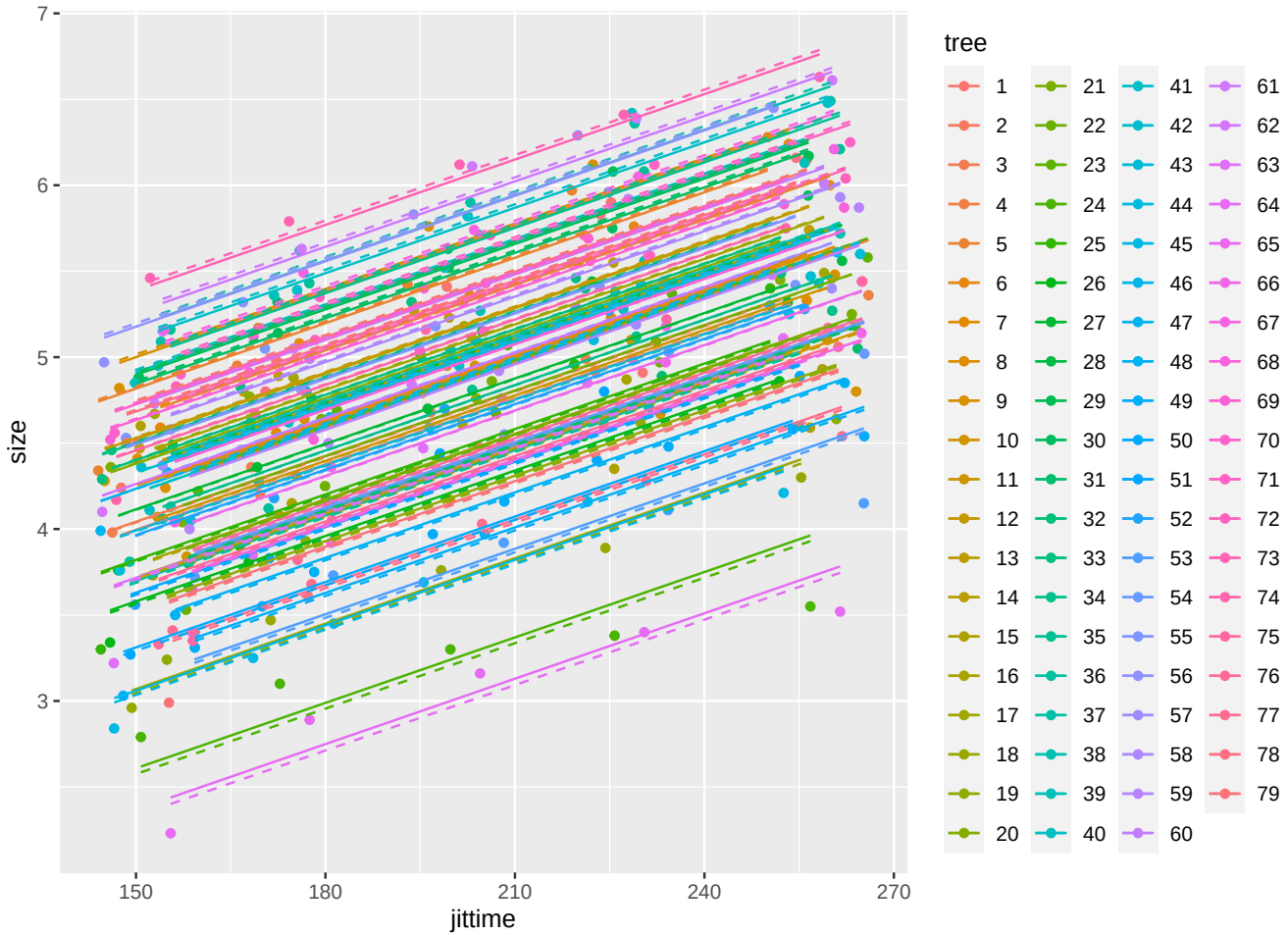


Figure: Tree Growth data and lines by fixed−effect−only model (dashed) and random intercept model (solid).

Uncertainty Quantification

Let $\theta \in \mathbb{R}^r$ denote the vector of parameters determining $\mathbf{C}$.

Theorem. Under validity of the model above and the MLE regularity conditions, we have for $K \to \infty$:

1. estimators $\widehat{\beta}, \widehat{\theta}, \widehat{\sigma}^2$ are consistent,
2. $$\sqrt{K}\begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\theta} - \theta \\ \widehat{\sigma}^2 - \sigma^2 \end{pmatrix} \to \mathcal{N}_{p+r+1}(0, \mathbf{J}^{-1}), \quad \text{where } \mathbf{J} = \begin{pmatrix} \mathbf{J}_\beta & 0 & 0 \\ 0 & \mathbf{J}_\theta & \mathbf{J}_{\theta,\sigma^2} \\ 0 & \mathbf{J}_{\theta,\sigma^2} & \mathbf{J}_{\sigma^2} \end{pmatrix}$$
   is the Fisher information matrix,
3. when $\widehat{\ell}$ denotes the maximized log-likelihood of the model and $\widehat{\ell}_0$ denotes the maximized log-likelihood of a submodel then
   $$2[\widehat{\ell} - \widehat{\ell}_0] \to \chi^2_m,$$
   where m is the difference in the no. of parameters between the model and the submodel.

Testing for Model Components

- testing fixed effects, i.e. $H_0 : \beta_{p-m+1} = \ldots = \beta_p = 0$ against $H_1 : \neg H_0$
  - can be done via LRT due to point 3. of the previous theorem
  - ML needs to be used instead of REML
  - however, p-values tend to be too small, sometimes overstating importance of some effects
- testing random effects, i.e. $H_0 : \theta_{p-m+1} = \ldots = \theta_p = 0$
  - usually cannot be done using the previous theorem, because MLE regularity assumptions are typically not met
    - one of the components of θ is typically the variance for one of the components of b_i's, which lies on the edge of the parameter space
  - it can still be shown that the LR statistic still follows χ^2-distribution, but with smaller degrees of freedom
    - p-values of the test the LRT from point 3. of the previous theorem are too big, sometimes understating importance of some effects

While solutions based on theory exist, a simpler road for us is the parametric bootstrap.

Example: Tree Growth

```
library(lme4)
# standardize time, otherwise convergence issues
Sitka <- Sitka %>% mutate(time=(time-mean(time))/sd(time))
mm1 <- lmer(size~treat*(time+I(time^2)) + (time+I(time^2)|tree),
            data=Sitka,REML=F)
mm0 <- lmer(size~time+I(time^2) + (time+I(time^2)|tree),
            data=Sitka,REML=F)
anova(mm0,mm1)
```

```
## Data: Sitka
## Models:
## mm0: size ~ time + I(time^2) + (time + I(time^2) | tree)
## mm1: size ~ treat * (time + I(time^2)) + (time + I(time^2) | tree)
##     npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## mm0   10 -111.88 -72.089 65.939  -131.88
## mm1   13 -118.50 -66.772 72.249  -144.50 12.62  3   0.005535 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- REML=F because the fixed-effect structure differs between the two models
- treatment seems significant, but... let's do the bootstrap

Example: Tree Growth

Bootstrap still rejects, although the p-value is doubled:

```
lrstat <- as.numeric(2*(logLik(mm1)-logLik(mm0)))
lrstats <- rep(0,1000)
for(i in 1:1000){
  set.seed(517*i)
  if(i %% 10 ==0) print(i)
  newDat <- Sitka
  newDat$size <- unlist(simulate(mm0))
  bnull <- lmer(size~time+I(time^2) + (time+I(time^2)|tree),
                data=newDat,REML=F)
  balt <- lmer(size~treat*(time+I(time^2)) + (time+I(time^2)|tree),
                data=newDat,REML=F)
  lrstats[i] <- as.numeric(2*(logLik(balt)-logLik(bnull)))
}
mean(lrstats > lrstat)
```

```
[1] 0.011
```

So we can finally conclude that ozone treatment matters :)

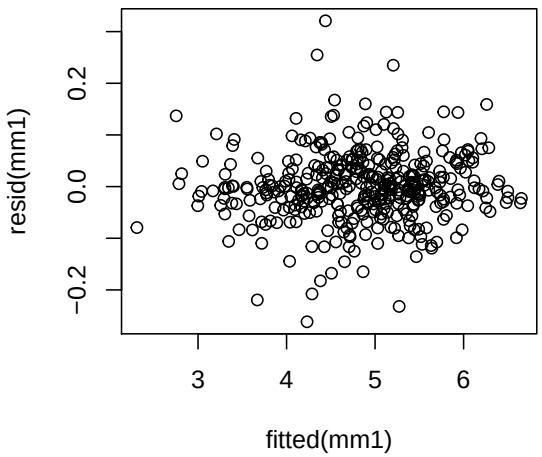
Diagnostics

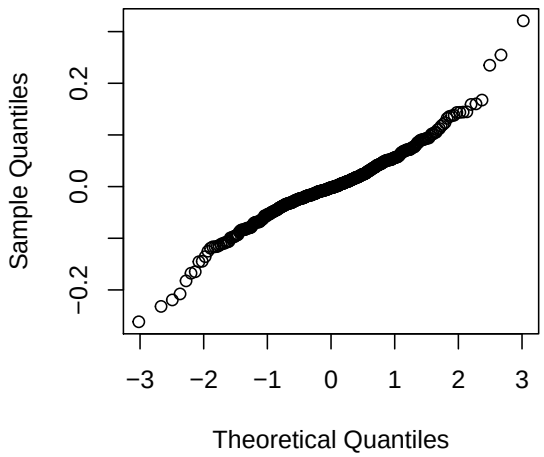
Similar to the standard linear model, with the additional

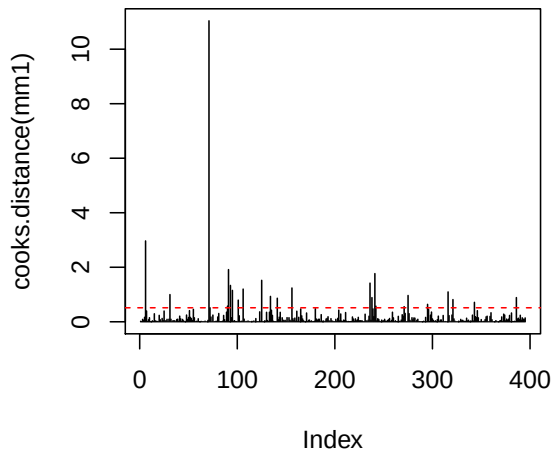
- check of normality for the random effects

```
plot(fitted(mm1),resid(mm1)) # the only thing `plot(mm1)` gives
qqnorm(resid(mm1),main="") # a bit heavy tails
plot(cooks.distance(mm1),type="h") # the old ROTs not useful here
abline(h=3*mean(cooks.distance(mm1)),col="red",lty=2) # another ROT
qqnorm(ranef(mm1)$tree[,1], main="QQ for Intercept")
qqnorm(ranef(mm1)$tree[,2], main="QQ for Linear Term")
qqnorm(ranef(mm1)$tree[,3], main="QQ for Quadratic Term")
# this is only checking marginals; multivariate GoF tests are tricky
```

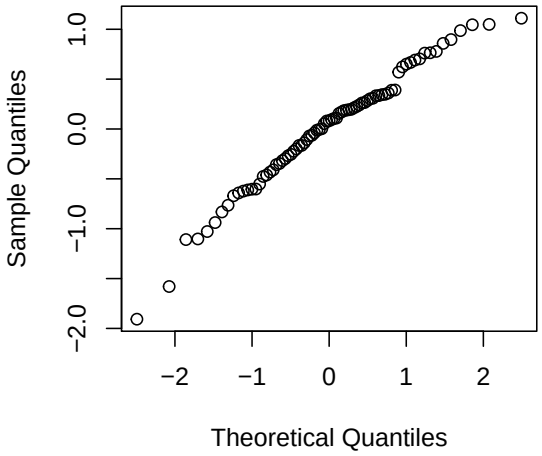
Diagnostics



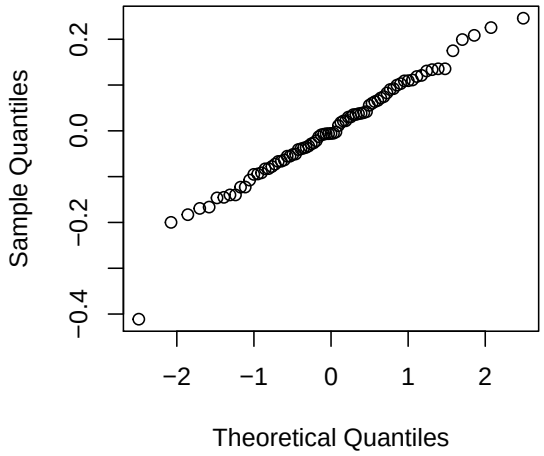




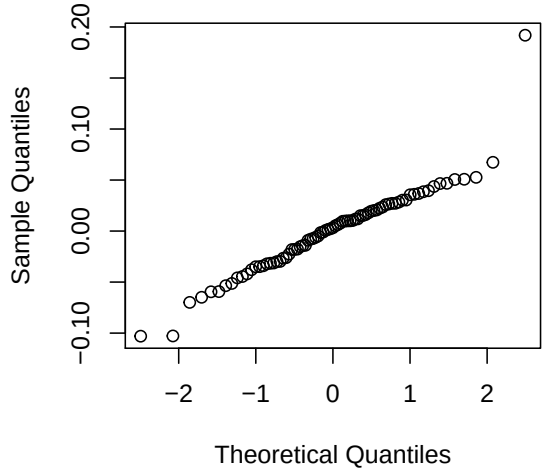
QQ for Intercept



QQ for Linear Term



QQ for Quadratic Term



summary(mm1)

```
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 tree     (Intercept) 0.356490 0.59707
          time        0.013509 0.11623   0.04
          I(time^2)   0.002593 0.05092   0.12 -0.71
 Residual             0.008096 0.08998
Number of obs: 395, groups:  tree, 79

Fixed effects:
                     Estimate Std. Error t value
(Intercept)          5.093296   0.120072  42.419
treatozone          -0.201347   0.145231  -1.386
time                 0.546397   0.024638  22.177
I(time^2)           -0.108450   0.014041  -7.724
treatozone:time     -0.079045   0.029800  -2.652
treatozone:I(time^2) -0.009835   0.016984  -0.579
```

- the random quadratic term is very significant (LR test's p-value 10^{-8} even though understate), but of a low variance and high correlation ... problems

Section 4

Multilevel Models

Grouping

There can be more than one type of grouping:

$$Y_{kl} = \mathbf{X}_{kl}\beta + \mathbf{Z}^{(1)}_{k,l} b_k + \mathbf{Z}^{(2)}_{l,k} b_l + \epsilon_{kl}$$

- Y_{kl} $in\mathbb{R}^{N_{kl}}$ is the vector of individual observations belonging to
  - k-th level of grouping (1)
  - l-th level of grouping (2)
- e.g. if in the tree growth example we needed to group not only by tree but also by time
  - all i-th observations per tree could be correlated
  - e.g. if a different person would measure tree size for different times i, but the same person for a single i

Or groupings can be nested:

$$Y_{kl} = \mathbf{X}_{kl}\beta + \mathbf{Z}_{k,l} b_k + \mathbf{Z}_{kl} b_{kl} + \epsilon_{kl}$$

- these are called multilevel models:
  - $k = 1, \ldots, K$ is the first-level grouping (e.g. school)
  - $l = 1, \ldots, L$ is the second-level grouping (e.g. class)
  - $Y_{kl} \in \mathbb{R}^{N_{kl}}$ is the vector of individual observations (e.g. students)

R syntax (lme4 package)

formula	meaning
`(1|group)`	random group intercept
`(x|group)` = `(1+x|group)`	random slope of x within group with correlated intercept
`(0+x|group)` = `(-1+x|group)`	random slope of x within group: no variation in intercept
`(1|group) + (0+x|group)`	uncorrelated random intercept and random slope within group
`(1|site/block)` = `(1|site)+(1|site:block)`	intercept varying among sites and among blocks within sites (nested random effects)
`site+(1|site:block)`	*fixed* effect of sites plus random variation in intercept among blocks within sites
`(x|site/block)` = `(x|site)+(x|site:block)` = `(1 + x|site)+(1+x|site:block)`	slope and intercept varying among sites and among blocks within sites
`(x1|site)+(x2|block)`	two different effects, varying at different levels
`x*site+(x|site:block)`	fixed effect variation of slope and intercept varying among sites and random variation of slope and intercept among blocks within sites
`(1|group1)+(1|group2)`	intercept varying among crossed random effects (e.g. site, year)

source: link

GLMMs

- the linear mixed model generalizes to the GLMM in the same way that the standard linear model generalizes to the GLM
- GLMMs are not as useful, because β's can only be interpreted as the population-average effects if the link g is linear:

$$\mathbb{E}[Y_n \mid X_n] = \mathbb{E}g^{-1}(X_n^\top \beta + Z_n \top b_n)$$

  - cannot go inside g^{-1} with the expectation unless it is linear, i.e. cannot get rid of b_n unless g^{-1} is linear
- only Gaussian models have the linear link as the canonical link
- if one does not use a canonical link, issues may arise
  - numerical convergence issues
  - calculated estimators are not guaranteed to be MLEs $\Rightarrow$ theory does not work, etc.

Section 5

Project 4