Clustered data models - Exercises 3

1. When studying characteristics that affect student performance on a battery of exams, we may use a model that takes into account variation between schools, students and tests. The model for the score of student i in school s on test t is

$$y_{ist} = \mathbf{x}'_{ist}\boldsymbol{\beta} + u_s + v_t + w_{is} + \epsilon_{ist},$$

where $\{u_s\} \sim N(0, \sigma_s^2)$ is the school effect, $\{v_t\} \sim N(0, \sigma_t^2)$ is the test effect, $\{w_{is}\} \sim N(0, \sigma_w^2)$ is the student effect and $\{\epsilon_{ist}\} \sim N(0, \sigma_\epsilon^2)$ is the random error. All the random effects are assumed to be independent. This is a slight extension to the model introduced in Section 9.2.2 of Agresti.

Determine a) the intraclass correlation between scores on different exams for a student, b) the intraclass correlation between scores on a particular exam for a pair of students in the same school c) the intraclass correlation between scores on different exams for a pair of students in the same school.

Solution a) We have that

$$\begin{aligned}
\mathrm{Cov}(y_{ist}, y_{ist'}) &= \mathrm{Cov}(u_s + v_t + w_{is} + \epsilon_{ist}, u_s + v_{t'} + w_{is} + \epsilon_{ist'}) \\
&= \mathrm{Var}(u_s) + \mathrm{Cov}(v_t, v_{t'}) + \mathrm{Var}(w_{is}) + \mathrm{Cov}(\epsilon_{ist}, \epsilon_{ist'}) \\
&= \sigma_u^2 + 0 + \sigma_w^2 + 0 = \sigma_u^2 + \sigma_w^2
\end{aligned}$$

and

$$\mathrm{Var}(y_{ist}) = \mathrm{Var}(y_{ist'}) = \sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_\epsilon^2,$$

so that

$$\mathrm{Cor}(y_{ist}, y_{ist'}) = \frac{\sigma_u^2 + \sigma_w^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_\epsilon^2}.$$

b) We have

$$\begin{aligned}
\mathrm{Cov}(y_{ist}, y_{i'st}) &= \mathrm{Cov}(u_s + v_t + w_{is} + \epsilon_{ist}, u_s + v_t + w_{i's} + \epsilon_{i'st}) \\
&= \sigma_u^2 + \sigma_v^2
\end{aligned}$$

so that

$$\mathrm{Cor}(y_{ist}, y_{i'st}) = \frac{\sigma_u^2 + \sigma_v^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_\epsilon^2}.$$

c) We have

$$\begin{aligned}
\mathrm{Cov}(y_{ist}, y_{i'st'}) &= \mathrm{Cov}(u_s + v_t + w_{is} + \epsilon_{ist}, u_s + v_{t'} + w_{i's} + \epsilon_{i'st'}) \\
&= \sigma_u^2
\end{aligned}$$

so that

$$\mathrm{Cor}(y_{ist}, y_{i'st'}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_v^2 + \sigma_w^2 + \sigma_\epsilon^2}.$$

2. (9.6 in Agresti) A crossover study comparing two drugs observes a continuous response (y_{i1}, y_{i2}) for each subject for each drug. Let $\mu_1 = \mathrm{E}(y_{i1})$ and $\mu_2 = \mathrm{E}(y_{i2})$ and consider $H_0 : \mu_1 = \mu_2$.

a) Construct the normal linear mixed model that generates a paired-difference t test (with test statistic $t = \sqrt{n}\bar{d}/s$, using mean and standard deviation of the differences $\{d_i = y_{i2} - y_{i1}\}$ and the corresponding confidence interval for $\mu_2 - \mu_1$.

b) Show the effect of the relative sizes of the variances of the random error and random effect on $\mathrm{Cor}(y_{i1}, y_{i2})$. Based on this, to compare two means, explain why it can be more efficient to use a design with dependent samples than with independent samples.

Solution. a) We may use a random intercept model

$$y_{ij} = \mu_j + u_i + \epsilon_{ij}, \quad \text{for } i = 1, ..., n, \quad j = 1, 2.$$

(This may also be written in the form $y_{ij} = \beta_0 + \beta_1 x_j + u_i + \epsilon_{ij}$ where x_j is the indicator of using the second drug.) Now

$$d_i = \mu_2 - \mu_1 + \epsilon_{i2} - \epsilon_{i1},$$

so that $d_i,\ i = 1, ..., n$, are independent, and $d_i \sim N(\mu_2 - \mu_1, 2\sigma_\epsilon^2)$. This implies that

$$\frac{\bar{d} - (\mu_2 - \mu_1)}{s_{\bar{d}}/\sqrt{n}} \sim t_{n-1}.$$

The null hypothesis $H_0 :\ \mu_1 = \mu_2$ is rejected with risk level α if $\sqrt{n}|\bar{d}|/s_{\bar{d}} > t_{\alpha/2;n-1}$, and the confidence interval for $\mu_2 - \mu_1$ is $(\bar{d} - t_{\alpha/2;n-1}s_{\bar{d}}/\sqrt{n},\ \bar{d} + t_{\alpha/2;n-1}s_{\bar{d}}/\sqrt{n})$.

b) Since

$$\text{Cov}(y_{i1}, y_{i2}) = \text{Cov}(u_i + \epsilon_{i1}, u_i + \epsilon_{i2}) = \sigma_u^2,$$

$$\text{Cor}(y_{i1}, y_{i2}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2} = \frac{\sigma_u^2/\sigma_\epsilon^2}{\sigma_u^2/\sigma_\epsilon^2 + 1}.$$

Thus, the correlation depends on the ratio $\sigma_u^2/\sigma_\epsilon^2$. It is more effective to use a paired test because it removes the effect of between-subject variation. We can see the difference by comparing the variances of the difference estimates in both cases:

For paired differences:

$$\text{Var}(\bar{d}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(y_{i2} - y_{i1})\right) = \frac{1}{n^2}\sum_{i=1}^{n}(2\sigma_\epsilon^2) = \frac{2\sigma_\epsilon^2}{n}.$$

For independent samples:

$$\begin{aligned}
\text{Var}(\bar{y}_2 - \bar{y}_1) &= \text{Var}(\bar{y}_1) + \text{Var}(\bar{y}_2) \\
&= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(u_i + \epsilon_{i1})\right) + \text{Var}\left(\frac{1}{n}\sum_{i=n+1}^{2n}(u_i + \epsilon_{i2})\right) \\
&= \frac{2}{n}(\sigma_u^2 + \sigma_\epsilon^2).
\end{aligned}$$

Thus, in the latter case, the estimate of $\mu_2 - \mu_1$ has larger variance and is less accurate.

3. (9.8 in Agresti) For the extension of the random-intercept linear mixed model (9.8 in Agresti; page 51 in lecture slides) that assumes $\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \sigma_\epsilon^2\rho^{|j-k|}$, show that

$$\text{Cor}(y_{ij}, y_{ik}) = (\sigma_u^2 + \rho^{|j-k|}\sigma_\epsilon^2)/(\sigma_u^2 + \sigma_\epsilon^2).$$

Solution. The model is

$$y_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + u_i + \epsilon_{ij}.$$

Thus, we have that

$$\text{Cov}(y_{ij}, y_{ik}) = \text{Cov}(u_i + \epsilon_{ij}, u_i + \epsilon_{ik}) = \sigma_u^2 + \rho^{|j-k|}\sigma_\epsilon^2$$

and

$$\text{Var}(y_{ij}) = \text{Var}(y_{ik}) = \sigma_u^2 + \sigma_\epsilon^2.$$

This implies that $\text{Cor}(y_{ij}, y_{ik}) = (\sigma_u^2 + \rho^{|j-k|}\sigma_\epsilon^2)/(\sigma_u^2 + \sigma_\epsilon^2)$.

4. (9.32 in Agresti) For the smoking prevention and cessation study (Section 9.2.3 in Agresti; page 54 in lecture slides), fit multilevel models to analyze whether it helps to add any interaction terms. Interpret fixed and random effects for the model that has a SC × TV interaction.

Solution. The likelihood ratio test indicates that the interaction is not significant. (Note that one cannot use the REML estimation method when comparing two models with different fixed parts!) However, if the model with interaction were correct, we could interpret the model as follows: 1) The THK score increases by 0.64 if the student takes part in the school-based curriculum but not in the television-based program. 2) The THK score increases by 0.18 if the student takes part in the television-based program but not in the school-based curriculum. 3) The THK score increases by 0.50 (=0.639+0.178-0.320) if the student takes part in both. (It appears to be illogical that the effect of both methods is smaller than SC alone, but we have to remember that the values are estimates and not exact figures.)

```
library(lme4) # Doug Bates's linear mixed models package
## Loading required package: Matrix
```

```
Smoking <- read.table("../data/Smoking.dat", header=TRUE)
attach(Smoking)

# Model without interaction
fit <- lmer(y ~ PTHK + SC + TV + (1|school) + (1|class), Smoking, REML=FALSE)
summary(fit) # school and classroom random intercepts
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
##    Data: Smoking
##
##      AIC      BIC   logLik deviance df.resid
##   5373.7   5411.4  -2679.9   5359.7     1593
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5483 -0.7011 -0.0152  0.6930  3.1781
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06825  0.2612
##  school   (Intercept) 0.02893  0.1701
##  Residual             1.60045  1.2651
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.77894    0.10724  16.588
## PTHK         0.30653    0.02586  11.854
## SC           0.47416    0.10593   4.476
## TV           0.02073    0.10592   0.196
##
## Correlation of Fixed Effects:
##      (Intr) PTHK   SC
## PTHK -0.517
## SC   -0.496  0.027
## TV   -0.513  0.016 -0.002
```

```
# Model with interaction
fit2 <- lmer(y ~ PTHK + SC*TV + (1|school) + (1|class), Smoking, REML=FALSE)
```

```
summary(fit2)
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: y ~ PTHK + SC * TV + (1 | school) + (1 | class)
##    Data: Smoking
##
##      AIC      BIC   logLik deviance df.resid
##   5373.4   5416.4  -2678.7   5357.4     1592
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5282 -0.7012 -0.0205  0.6840  3.1632
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  class    (Intercept) 0.06358  0.2522
##  school   (Intercept) 0.02575  0.1605
##  Residual             1.60201  1.2657
## Number of obs: 1600, groups:  class, 135; school, 28
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.69700    0.11666  14.547
## PTHK         0.30720    0.02584  11.888
## SC           0.63919    0.14721   4.342
## TV           0.17811    0.14365   1.240
## SC:TV       -0.32042    0.20551  -1.559
##
## Correlation of Fixed Effects:
##       (Intr) PTHK   SC     TV
## PTHK  -0.474
## SC    -0.623  0.017
## TV    -0.634  0.010  0.499
## SC:TV  0.439  0.003 -0.716 -0.699
```

```
anova(fit, fit2)
## Data: Smoking
## Models:
## fit: y ~ PTHK + SC + TV + (1 | school) + (1 | class)
## fit2: y ~ PTHK + SC * TV + (1 | school) + (1 | class)
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## fit     7 5373.7 5411.4 -2679.9   5359.7
## fit2    8 5373.4 5416.4 -2678.7   5357.4 2.3619  1     0.1243
```

5. (9.33 in Agresti) Using the R output shown for the simple analyses of the FEV data in Section 9.2.5, show that the estimated values of $\text{Cor}(y_{i1}, y_{i2})$ and $\text{Cor}(y_{i1}, y_{i8})$ are 0.74 for the random intercept model and 0.86 and 0.62 for the model that also permits autoregressive within-patient errors.

Solution. (Compare with exercise 3!) In the first case, the estimate of intraclass correlation is $\hat{\sigma}_u^2/(\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2) = 0.4526834^2/(0.4526834^2 + 0.2716699^2) \approx 0.74$, and in the second case $(\hat{\sigma}_u^2 + \sigma_\epsilon^2\hat{\rho}^{2-1})/(\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2) = (0.4075485^2 + 0.3354964^2 \cdot 0.6480841)/(0.4075485^2 + 0.3354964^2) \approx 0.86$ and $(\hat{\sigma}_u^2 + \sigma_\epsilon^2\hat{\rho}^{8-1})/(\hat{\sigma}_u^2 + \hat{\sigma}_\epsilon^2) = (0.4075485^2 + 0.3354964^2 \cdot 0.6480841^7)/(0.4075485^2 + 0.3354964^2) \approx 0.62$.

```
library(nlme)
##
```

```
## Attaching package: 'nlme'
## The following object is masked from 'package:lme4':
##
##     lmList
```

```
FEV2 <- read.table("../data/FEV2.dat", header=TRUE)
attach(FEV2)
summary(lme(fev ~ base + factor(drug) + hour, random = ~ 1|patient))
## Linear mixed-effects model fit by REML
##   Data: NULL
##        AIC      BIC    logLik
##   388.9149 419.3466 -187.4575
##
## Random effects:
##  Formula: ~1 | patient
##         (Intercept)  Residual
## StdDev:   0.4526834 0.2716699
##
## Fixed effects:  fev ~ base + factor(drug) + hour
##                    Value  Std.Error  DF    t-value p-value
## (Intercept)    1.0492317 0.29217514 503   3.591105  0.0004
## base           0.9028516 0.10328130  68   8.741675  0.0000
## factor(drug)c  0.2258930 0.13361174  68   1.690667  0.0955
## factor(drug)p -0.2814907 0.13362978  68  -2.106496  0.0389
## hour          -0.0745734 0.00494027 503 -15.095011  0.0000
##  Correlation:
##                (Intr) base   fctr(drg)c fctr(drg)p
## base           -0.943
## factor(drug)c -0.246  0.019
## factor(drug)p -0.252  0.025  0.500
## hour          -0.076  0.000  0.000      0.000
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.24541534 -0.53211567  0.02615837  0.55222515  2.57172148
##
## Number of Observations: 576
## Number of Groups: 72
```

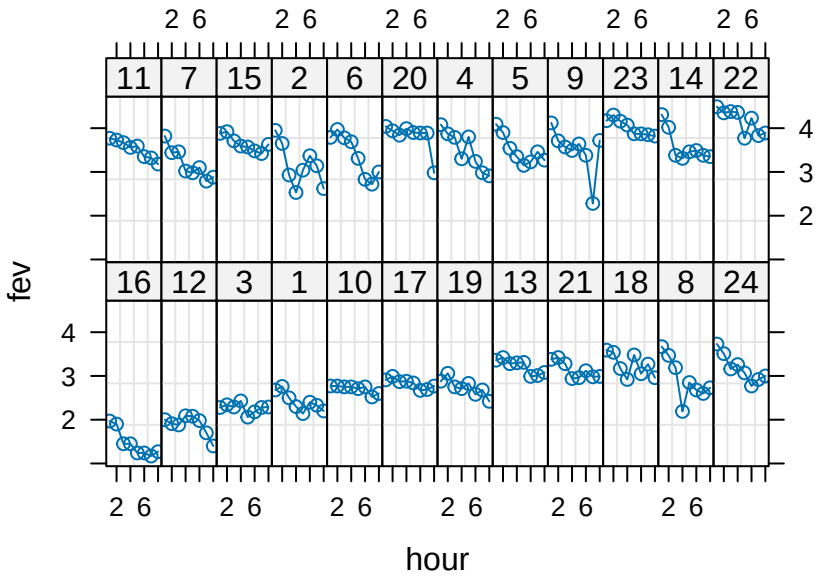
```
summary(lme(fev ~ base + factor(drug) + hour,
            random = ~ 1|patient, correlation = corAR1(form = ~ hour|patient)))
## Linear mixed-effects model fit by REML
##   Data: NULL
##        AIC      BIC   logLik
##   243.3501 278.1292 -113.675
##
## Random effects:
##  Formula: ~1 | patient
##         (Intercept)  Residual
## StdDev:   0.4075485 0.3354964
##
## Correlation Structure: AR(1)
##  Formula: ~hour | patient
##  Parameter estimate(s):
```

```
##       Phi
## 0.6480841
## Fixed effects:  fev ~ base + factor(drug) + hour
##                    Value  Std.Error  DF   t-value p-value
## (Intercept)    1.0723482 0.29138890 503  3.680127  0.0003
## base           0.8917791 0.10257147  68  8.694222  0.0000
## factor(drug)c  0.2129614 0.13269345  68  1.604912  0.1131
## factor(drug)p -0.3141641 0.13271136  68 -2.367274  0.0208
## hour          -0.0690618 0.00769164 503 -8.978814  0.0000
##  Correlation:
##               (Intr) base   fctr(drg)c fctr(drg)p
## base          -0.939
## factor(drug)c -0.245  0.019
## factor(drug)p -0.251  0.025  0.500
## hour          -0.119  0.000  0.000      0.000
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -3.430597097 -0.497835839 -0.005822408  0.507114029  2.392245519
##
## Number of Observations: 576
## Number of Groups: 72
```

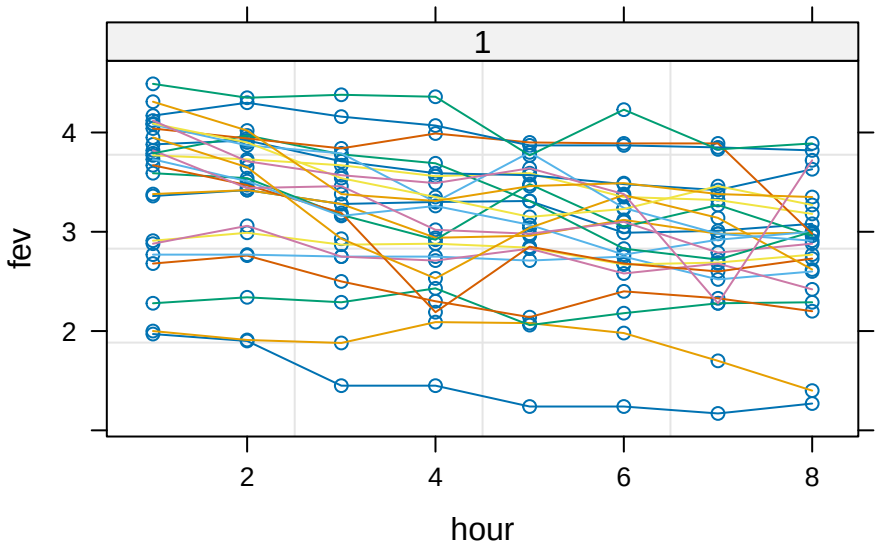
6. (9.34 in Agresti) Refer to Exercise 1.21 (in Agresti) and the longitudinal analysis in Section 9.2.5. Analyze the data in file FEV2.dat at www.stat.ufl.edu/~aa/glm/data, investigating the correlation structure for the eight FEV responses and modeling how FEV depends on the hour and the drug, adjusting for the baseline observation. Take into account whether to treat hour as qualitative or quantitative, whether you need interaction terms, whether to have random slopes or only random intercepts, and whether to treat within-patient errors as correlated. Interpret results for your final chosen model. (You may want to read Littell et al. (2000). The book SAS for Mixed Models, 2nd ed., by Littell et al. (2006, SAS Institute), uses SAS to fit various models to these data.)

Solution. The figures suggest that there may be differences in slopes both between patients and between treatments.

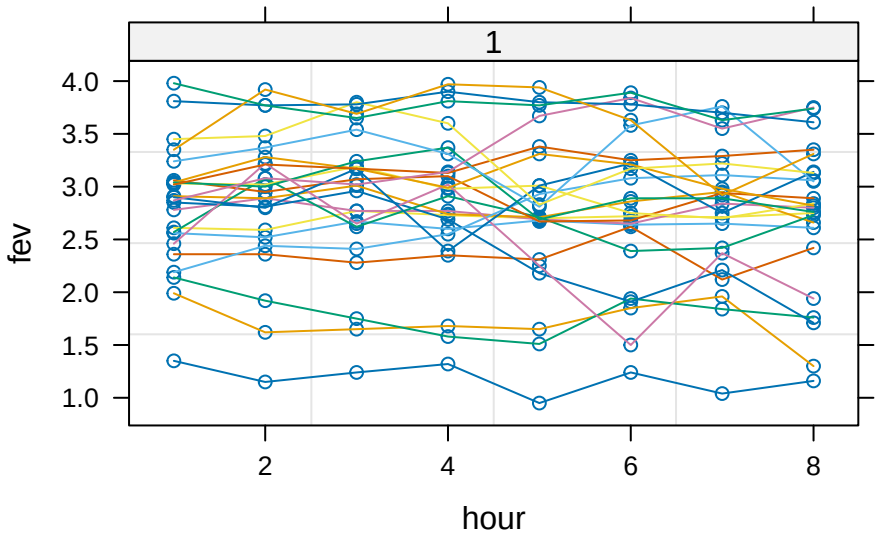
```
library(nlme)
FEV2 <- groupedData(fev ~ hour | patient, data = FEV2)
plot(subset(FEV2, drug == "a"), asp = 3)
```




```
plot(subset(FEV2, drug == "a"), outer = 1, key = FALSE, asp = 0.5)
```




```
plot(subset(FEV2, drug == "p"), outer = 1, key = FALSE, asp = 0.5)
```



Next, we fit different models and test if they differ significantly.

```
model1 <- lme(fev ~ base + factor(drug) + hour, random = ~ 1|patient, method= "ML", data = FEV2)
model2 <- lme(fev ~ base + factor(drug) + factor(hour), random = ~ 1|patient, method= "ML", data = FEV2)
anova(model1, model2)
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model1     1  7 368.6692 399.1620 -177.3346
## model2     2 13 375.8491 432.4785 -174.9245 1 vs 2 4.820128  0.5671
```

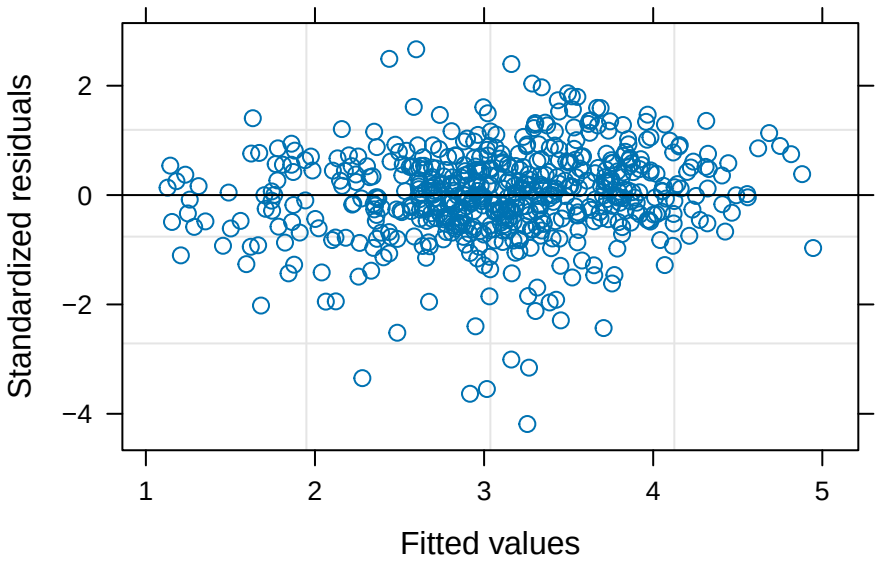
```
# Treating hour as qualitative does not improve the fit
model3 <- lme(fev ~ base + factor(drug)*hour, random = ~ 1|patient, method= "ML", data = FEV2)
anova(model1, model3)
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model1     1  7 368.6692 399.1620 -177.3346
## model3     2  9 298.4815 337.6864 -140.2407 1 vs 2 74.18776  <.0001
```

```
# Interaction of drug and hour is significant

model3b <- lme(fev - base ~ factor(drug)*hour, random = ~ 1|patient, method= "ML", data = FEV2)
# We could as well consider fev - base as the response variable and remove base
# from the explanatory variables


model3c <- lme(fev ~ base + factor(drug)*hour, random = ~ 1|patient, data = FEV2)
model4 <- lme(fev ~ base + factor(drug)*hour, random = ~ hour|patient, data = FEV2)
anova(model3c, model4)
##         Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model3c     1  9 333.4345 372.5295 -157.7173
## model4      2 11 241.4482 289.2309 -109.7241 1 vs 2 95.98636  <.0001
```

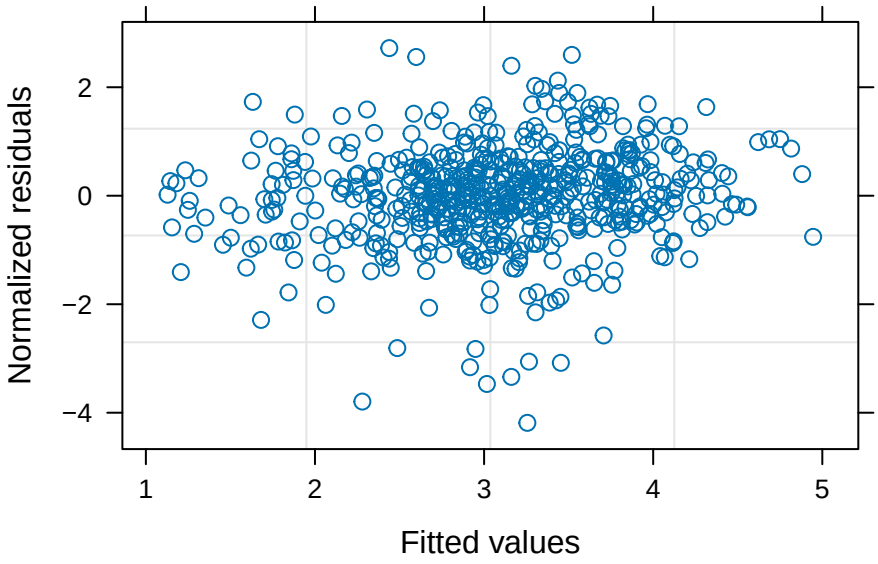
```
# Modeling slope as random improves the fit

model5 <- lme(fev ~ base + factor(drug)*hour, random = ~ hour|patient,
              correlation = corAR1(form = ~ hour|patient), data = FEV2)
anova(model4, model5)
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model4     1 11 241.4482 289.2308 -109.7241
## model5     2 12 225.4022 277.5287 -100.7011 1 vs 2 18.04602  <.0001
```

```r
# AR(1) sturcture improves the model

model6 <- lme(fev ~ base + factor(drug)*hour, random = ~ hour|patient,
              correlation = corARMA(form = ~ hour|patient, p = 2), data = FEV2)
anova(model5, model6)
##        Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## model5     1 12 225.4022 277.5287 -100.7011
## model6     2 13 226.3975 282.8680 -100.1988 1 vs 2 1.004597  0.3162
```

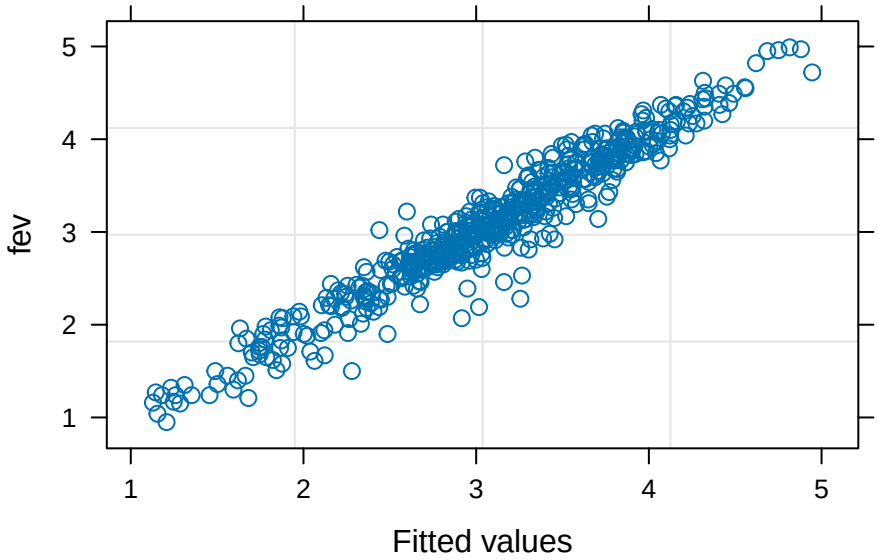
```r
# AR(2) does not provide improvement over AR(1)


# Plotting some diagnostic figures
plot(model5)
```



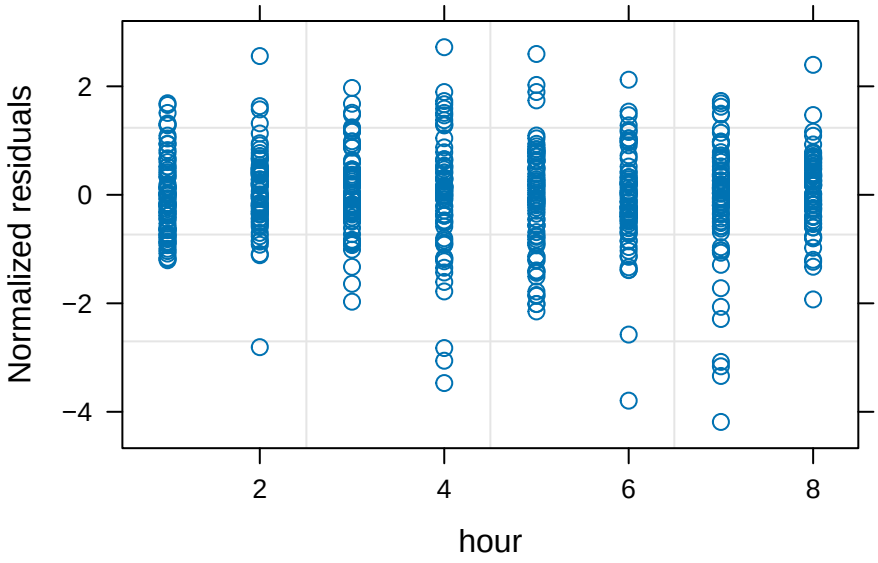

```r
plot(model5, form = resid(., type = "normalized") ~ fitted(.))
```



```
# In "normalized residuals" the effect of autocorrelation is removed.
# The residuals do not show heteroscedasticity. However their distribution is
# skewed to negative values.
plot(model5, form = fev ~ fitted(.))
```




```
plot(model5, form = resid(., type = "normalized") ~ hour)
```



Interpretation of the final model: Both intercepts and slopes vary between the patients. The standard deviations of the intercept and slope are 0.47 and 0.050, respectively, and their correlation is -0.347. The within-patient errors has standard deviation 0.233 and have an AR(1) structure with correlation coefficient 0.31.

The baseline value is a significant explanatory variable: When its value increases by 1, the response variable increases by 0.90. (But this coefficient does not differ significantly from 1. It might have been wiser to consider *fev* - *base* as the response variable and remove *base* from the explanatory variables!) The intercepts differ significantly between the treatments: the intercepts are 1.12, 1.44 (1.1213280+0.3162339) and 0.504 (1.1213280-0.6171436) for drug A, drug C and placebo, respectively. Thus, drug C seems to provide the best improvement for the FEV value. Because the variables drug and hour have interaction, the slopes also differ between the treatments. The slopes are -0.089, -0.110 (-0.0894006-0.0209822) and -0.017

(-0.0894006+0.0722745) for drug A, drug C and placebo, respectively. There is no significant difference between the slopes of drug A and drug C. For the placebo, the slope does not differ significantly from 0. Because the slope is negative for the drugs, we can conclude that their effect decreases with time.

```
summary(model5)
## Linear mixed-effects model fit by REML
##   Data: FEV2
##        AIC      BIC    logLik
##   225.4022 277.5287 -100.7011
##
## Random effects:
##  Formula: ~hour | patient
##  Structure: General positive-definite, Log-Cholesky parametrization
##             StdDev     Corr
## (Intercept) 0.47398988 (Intr)
## hour        0.04997068 -0.347
## Residual    0.23304457
##
## Correlation Structure: AR(1)
##  Formula: ~hour | patient
##  Parameter estimate(s):
##       Phi
## 0.3117025
## Fixed effects:  fev ~ base + factor(drug) * hour
##                         Value  Std.Error  DF   t-value p-value
## (Intercept)         1.1213280 0.29341809 501  3.821605  0.0001
## base                0.9030590 0.10245262  68  8.814407  0.0000
## factor(drug)c       0.3162339 0.15074458  68  2.097813  0.0396
## factor(drug)p      -0.6171436 0.15076031  68 -4.093541  0.0001
## hour               -0.0894006 0.01337098 501 -6.686167  0.0000
## factor(drug)c:hour -0.0209822 0.01890942 501 -1.109614  0.2677
## factor(drug)p:hour  0.0722745 0.01890942 501  3.822143  0.0001
##  Correlation:
##                    (Intr) base   fctr(drg)c fctr(drg)p hour   fctr(drg)c:
## base               -0.932
## factor(drug)c      -0.272  0.017
## factor(drug)p      -0.277  0.022  0.500
## hour               -0.173  0.000  0.337      0.337
## factor(drug)c:hour  0.122  0.000 -0.476     -0.238     -0.707
## factor(drug)p:hour  0.122  0.000 -0.238     -0.476     -0.707  0.500
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -4.186967227 -0.437897596  0.003773089  0.501995700  2.666238024
##
## Number of Observations: 576
## Number of Groups: 72
```