

Aufgabe 1:

In dieser Aufgabe sollen Sie den FEV-Datensatz analysieren. Dieser enthält wiederholte Messungen eines Lungenleistungsparamters (FEV= Forciertes Expirations-Volumen) an 300 Mädchen im Alter zwischen 6 bis 19 Jahren. Benutzen Sie den folgenden Code um den FEV-Datensatz von der Homepage zu laden und die Daten zu visualisieren:

```
R>library(nlme)
R>url <- "http://www.statistik.lmu.de/institut/lehrstuhl/semwiso/longitudinal_ss08/"
R>fev <- groupedData(log.fev1 ~ age | subject, read.table(paste(url,
+ "download/fev1.txt", sep = ""), header = T))
R>groups <- unique(fev$subject)
R>x1 <- range(fev$age)
R>y1 <- range(fev$log.fev1)
R>for (i in 1:30) {
+   samp <- groups[(1 + (i - 1) * 10):(i * 10)]
+   d <- subset(fev, subject %in% samp)
+   print(xyplot(log.fev1 ~ age, data = d, groups = subject,
+ type = "b", xlim = x1, ylim = y1))
+   Sys.sleep(0.5)
+ }
R>x1 <- range(fev$height)
R>for (i in 1:30) {
+   samp <- groups[(1 + (i - 1) * 10):(i * 10)]
+   d <- subset(fev, subject %in% samp)
+   print(xyplot(log.fev1 ~ height, data = d, groups = subject,
+ type = "b", xlim = x1, ylim = y1))
+   Sys.sleep(0.5)
+ }
```

- (a) Benutzen Sie die beiden festen Effekte Alter (`age`) und Körpergröße (`height`) in einem gemischten linearen Modell für die Lungenleistung. Verwenden Sie für die zufälligen Effekte subjektspezifische Intercepts und subjektspezifische Trends über die Zeit. Können die Einflüsse von `age` und `height` linear modelliert werden? Vergleichen Sie die Modellgüte verschiedener (Kombinationen von) Transformationen dieser beiden Größen.

Lösung:

Im folgenden werden die Modellierungen linear, log-linear und quadratisch für die beiden Größen und Ihre Kombinationen geschätzt und verglichen:

```
R>m00 <- lme(log.fev1 ~ age + height, random = ~age | subject,
+ data = fev, method = "ML")
R>m10 <- lme(log.fev1 ~ poly(age, 2) + height, random = ~poly(age,
+ 2) | subject, data = fev, method = "ML")
R>m20 <- lme(log.fev1 ~ log(age) + height, random = ~log(age) |
+ subject, data = fev, method = "ML")
R>m01 <- lme(log.fev1 ~ age + log(height), random = ~age | subject,
+ data = fev, method = "ML")
R>m11 <- lme(log.fev1 ~ poly(age, 2) + log(height), random = ~poly(age,
```

```

+      2) | subject, data = fev, method = "ML")
R>m21 <- lme(log.fev1 ~ log(age) + log(height), random = ~log(age) |
+      subject, data = fev, method = "ML")
R>m02 <- lme(log.fev1 ~ age + poly(height, 2), random = ~age |
+      subject, data = fev, method = "ML")
R>m12 <- lme(log.fev1 ~ poly(age, 2) + poly(height, 2), random = ~poly(age,
+      2) | subject, data = fev, method = "ML")
R>m22 <- lme(log.fev1 ~ log(age) + poly(height, 2), random = ~log(age) |
+      subject, data = fev, method = "ML")
R>anova(m22, m12, m02, m21, m11, m01, m20, m10, m00, test = F)

```

	Model	df	AIC	BIC	logLik
m22	1	8	-4601.466	-4556.683	2308.733
m12	2	12	-4619.850	-4552.676	2321.925
m02	3	8	-4584.448	-4539.665	2300.224
m21	4	7	-4471.693	-4432.508	2242.847
m11	5	11	-4549.273	-4487.696	2285.637
m01	6	7	-4515.843	-4476.658	2264.921
m20	7	7	-4571.748	-4532.562	2292.874
m10	8	11	-4607.296	-4545.719	2314.648
m00	9	7	-4580.711	-4541.525	2297.355

Auf Basis von AIC und BIC wählen wir Modell m12.

```
R>summary(m12)
```

Linear mixed-effects model fit by maximum likelihood

Data: fev

	AIC	BIC	logLik
	-4619.85	-4552.676	2321.925

Random effects:

Formula: ~poly(age, 2) | subject

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	0.1071060	(Intr) p(,2)1
poly(age, 2)1	1.2327604	0.199
poly(age, 2)2	0.6904261	-0.267 -0.566
Residual	0.0562798	

Fixed effects: log.fev1 ~ poly(age, 2) + poly(height, 2)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.813264	0.0065748	1690	123.69468	0.0000
poly(age, 2)1	3.590897	0.3637842	1690	9.87095	0.0000
poly(age, 2)2	-0.508445	0.1601138	1690	-3.17552	0.0015
poly(height, 2)1	10.412348	0.3922554	1690	26.54481	0.0000
poly(height, 2)2	0.456211	0.1150166	1690	3.96647	0.0001

Correlation:

	(Intr)	ply(g,2)1	ply(g,2)2	ply(h,2)1
poly(age, 2)1	0.038			
poly(age, 2)2	-0.059	-0.797		
poly(height, 2)1	0.026	-0.956	0.797	
poly(height, 2)2	-0.013	0.343	-0.606	-0.377

Standardized Within-Group Residuals:

	Min	Q1	Med	Q3	Max
	-6.41588588	-0.47132682	0.07359084	0.53976196	2.98484703

Number of Observations: 1994

Number of Groups: 300

Nachtrag: Durch die Interaktion von `age` und `height` kann man das Modell weiter verbessern:

```
R>m12.inter <- update(m12, . ~ . + age:height)
```

```
R>anova(m12, m12.inter)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
m12	1	12	-4619.850	-4552.676	2321.925			
m12.inter	2	13	-4631.247	-4558.474	2328.623	1 vs 2	13.39641	3e-04

(b) Benutzen Sie im Folgenden quadratische Trends für Alter und Körpergröße. Überprüfen Sie (grafisch), ob die Modell-Annahmen über die Residuen $\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{Z}_i \mathbf{b}_i$; $i = 1, \dots, 300$ erfüllt sind.

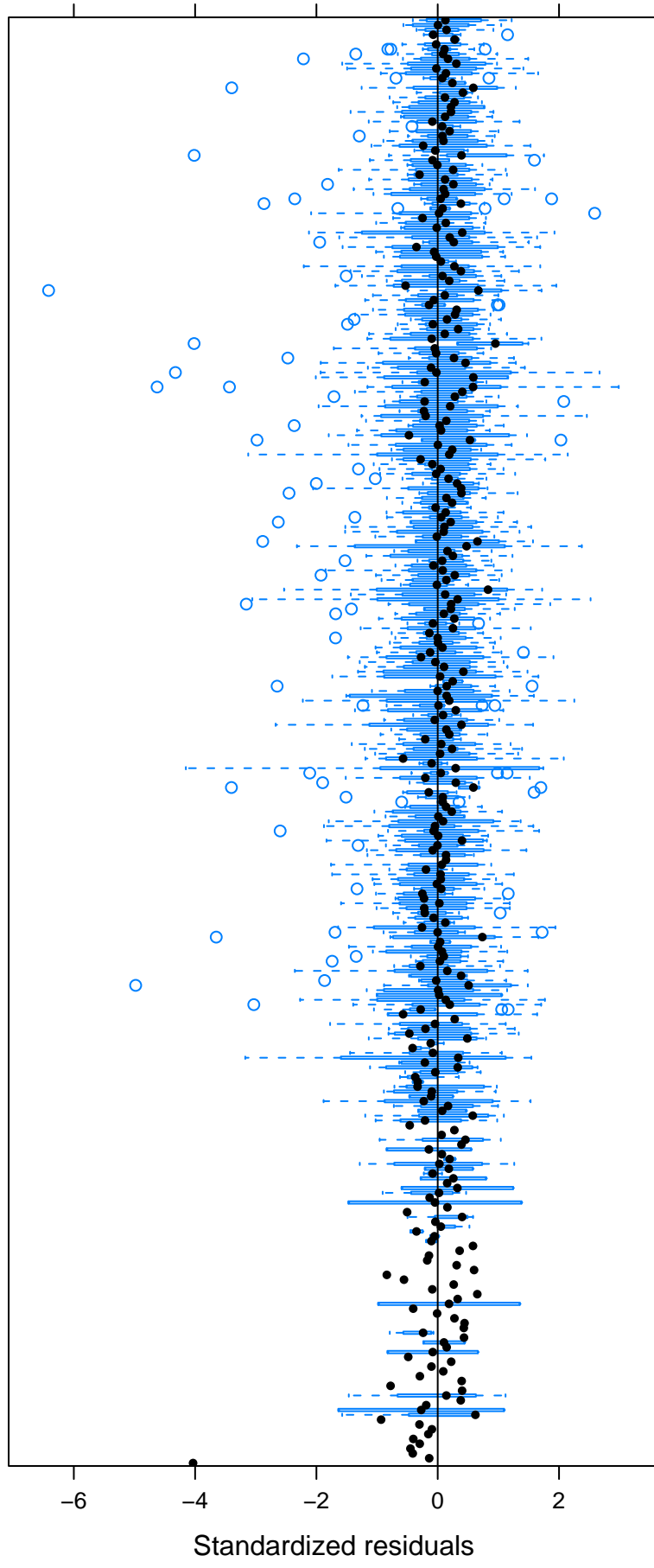
(i) Sind die Residuen für die einzelnen Subjekte in etwa symmetrisch und mit gleicher Varianz verteilt?

Lösung:

Nicht symmetrisch, viele große negative Ausreißer:

```
R>print(plot(m12, subject ~ resid(., type = "n"), abline = 0, cex = 0.5))
```

subject

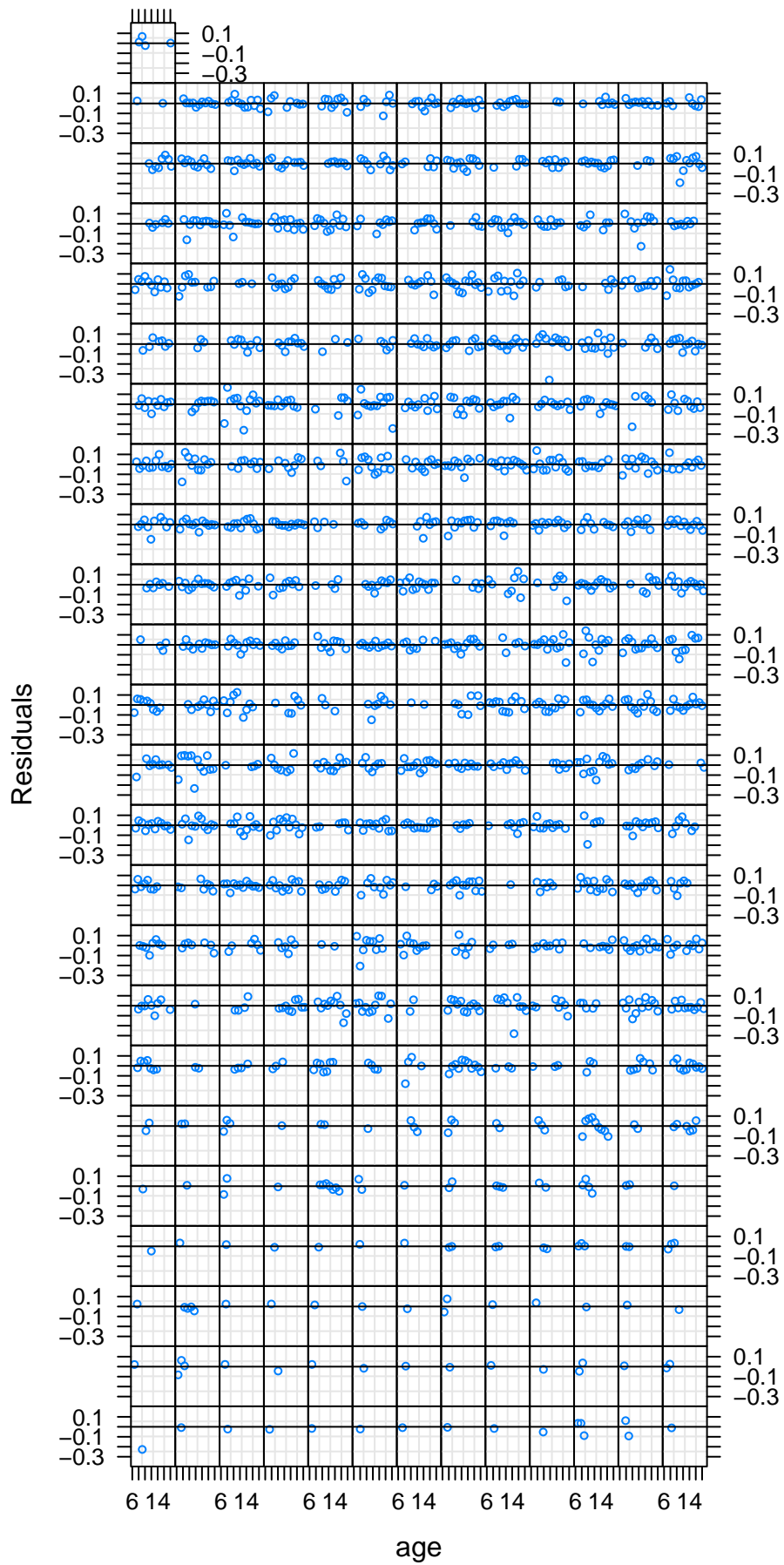


- (ii) Spiegelt das Modell die Form des Trends in Alter adäquat wieder oder gibt es systematische Unter-/Überschätzungen?

Lösung:

S. auch voriger Plot: teilweise deutliche Über-/Unterschätzung, oft noch viel Struktur in subjektspezifischen Residuenverläufen!

```
R>print(plot(m12, resid(.) ~ age | subject, abline = 0, cex = 0.5,  
+         strip = F))
```

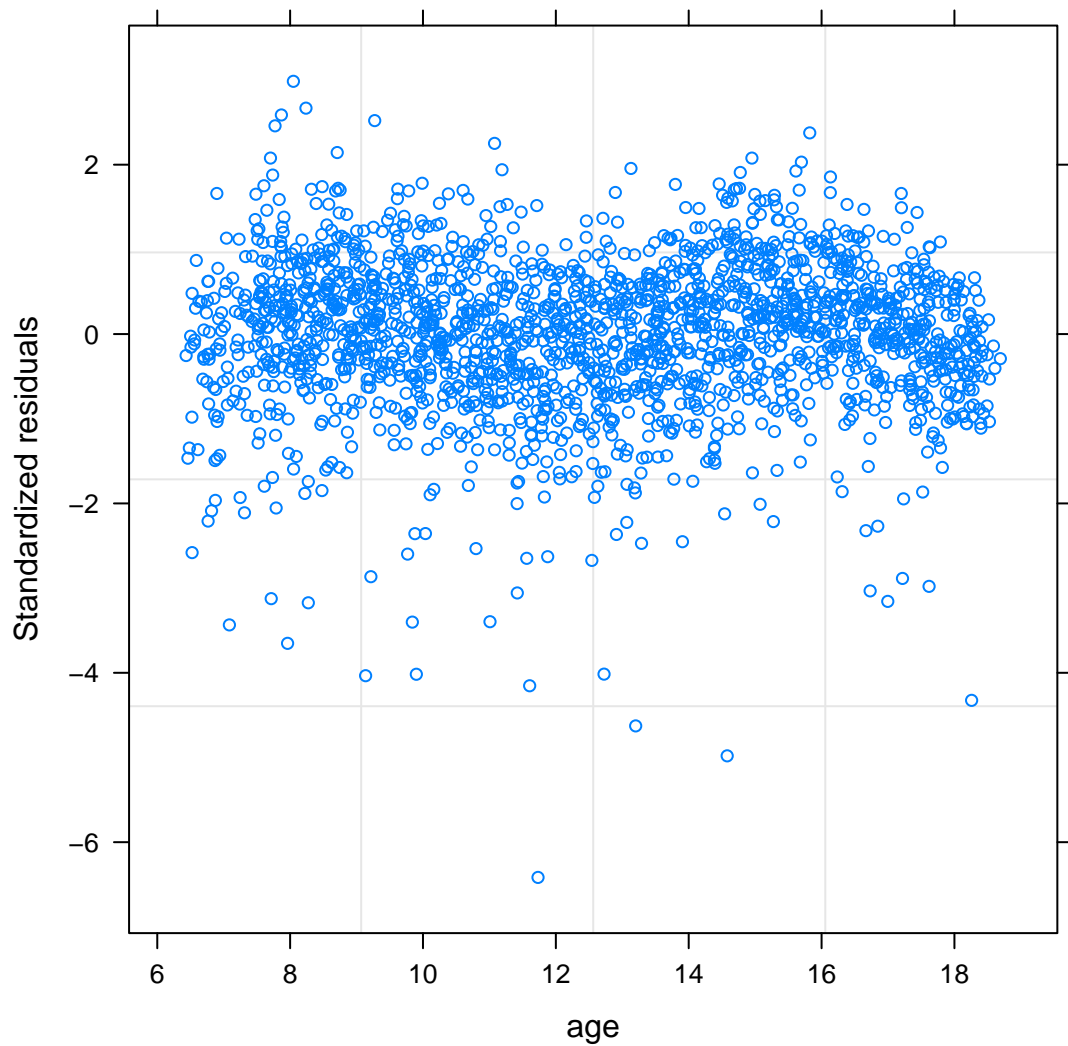


⇒ evtl. Autokorrelation in den Residuen?

(iii) Ist die Varianz der Residuen unabhängig vom Alter?

Lösung:

```
R>print(plot(m12, resid(., type = "p") ~ age, cex = 0.7))
```

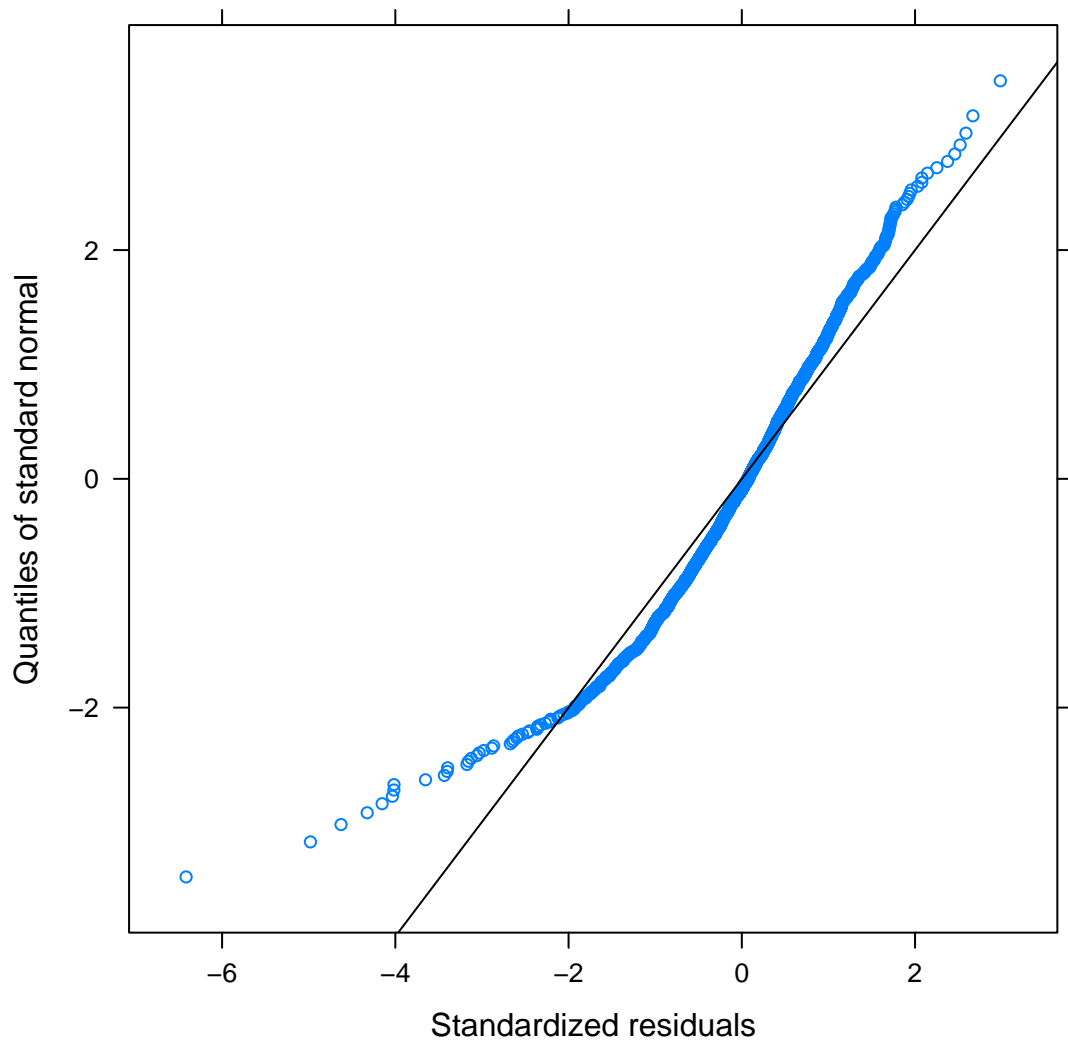


keine Anzeichen von Heteroskedastie.

(iv) Sind die (normalisierten) Residuen in etwa normalverteilt?

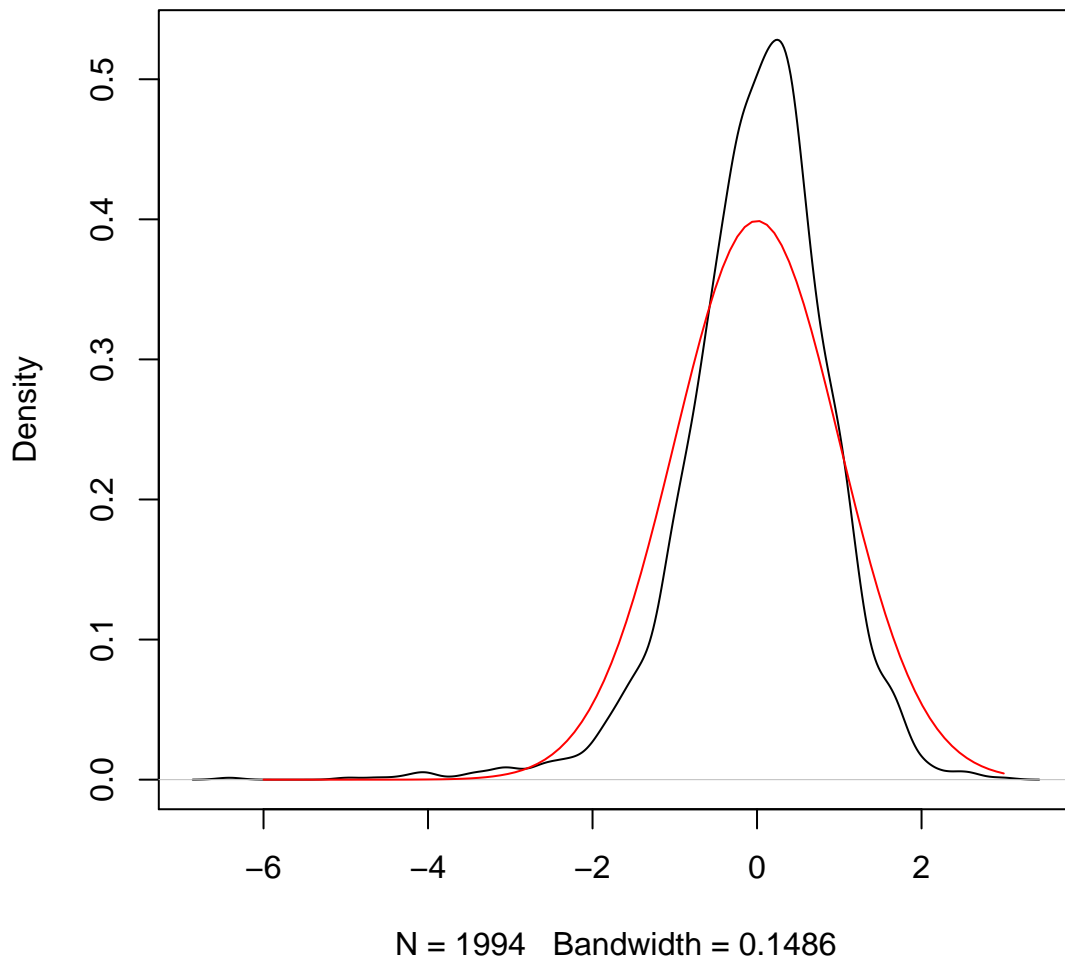
Lösung:

```
R>print(qqnorm(m12, ~resid(., type = "n"), abline = c(0, 1), cex = 0.7))
```



```
R>plot(density(resid(m12, type = "n")), main = "Dichteschätzung der normalisierten Res", col = 2)  
R>curve(dnorm(x), from = -6, to = 3, add = T, col = 2)
```

Dichteschätzung der normalisierten Residuen und Dichte der $N(0,1)$



⇒ Residuen sind stärker um 0 konzentriert und linksschief als die Standard-Normalverteilung.

- (v) Ist die Annahme unkorrelierter Residuen gerechtfertigt oder gibt es Hinweise auf das Vorliegen einer seriellen Korrelationsstruktur in `age`?

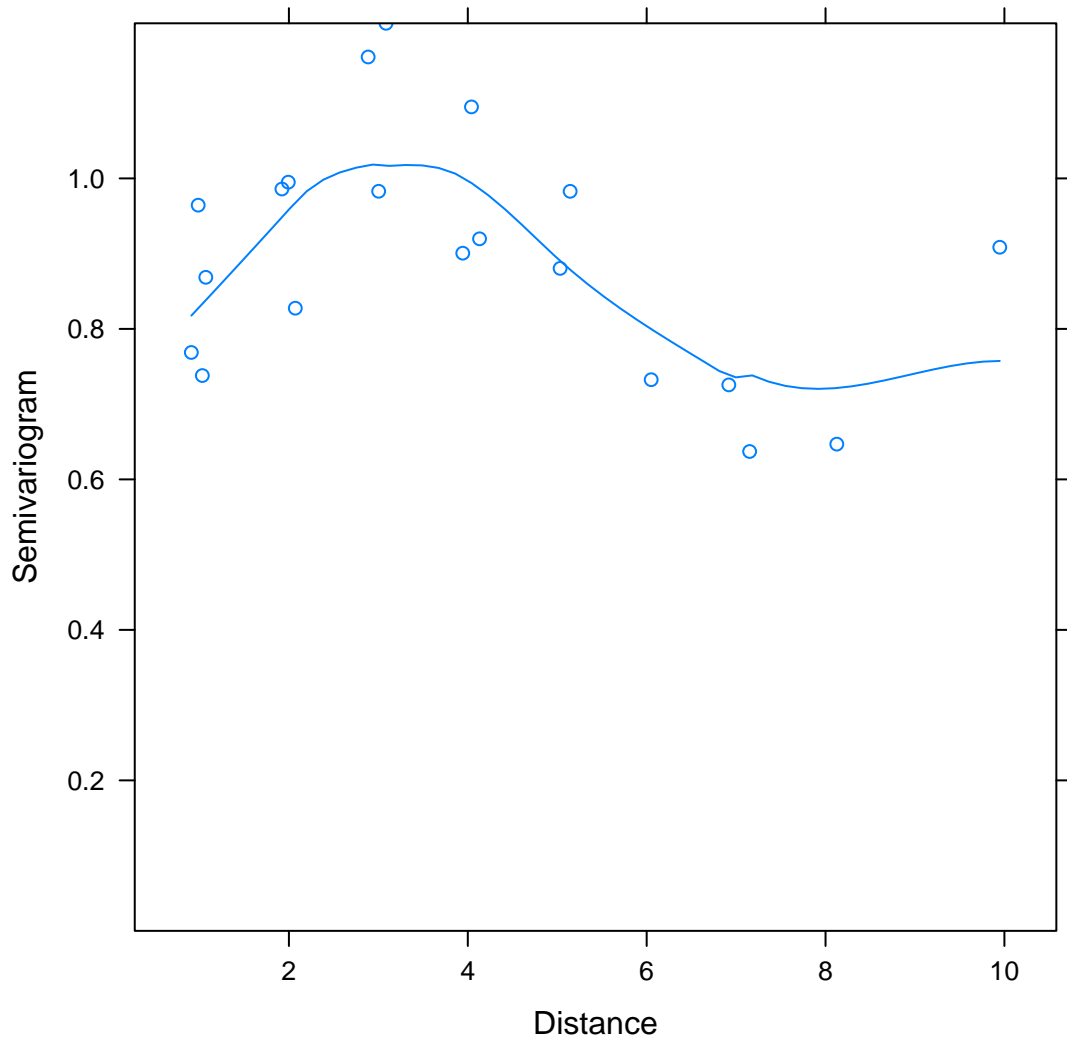
Lösung:

```
R>(Vg.12 <- Variogram(m12, form = ~age | subject))
```

	variog	dist	n.pairs
1	0.7687204	0.90900	395
2	0.9644668	0.98560	396
3	0.7380025	1.03220	385
4	0.8685872	1.07050	392
5	0.9859268	1.92195	392
6	0.9950266	1.99320	393
7	0.8275457	2.07120	396
8	1.1614177	2.88570	393
9	0.9829307	3.00340	394
10	1.2061373	3.08560	390
11	0.9006255	3.94385	392

12	1.0950454	4.04110	389
13	0.9197531	4.13140	393
14	0.8804346	5.03210	393
15	0.9828640	5.14440	393
16	0.7325101	6.05060	391
17	0.7255813	6.91860	392
18	0.6371051	7.15125	396
19	0.6468854	8.12590	389
20	0.9085414	9.94795	392

```
R>print(plot(Vg.12))
```

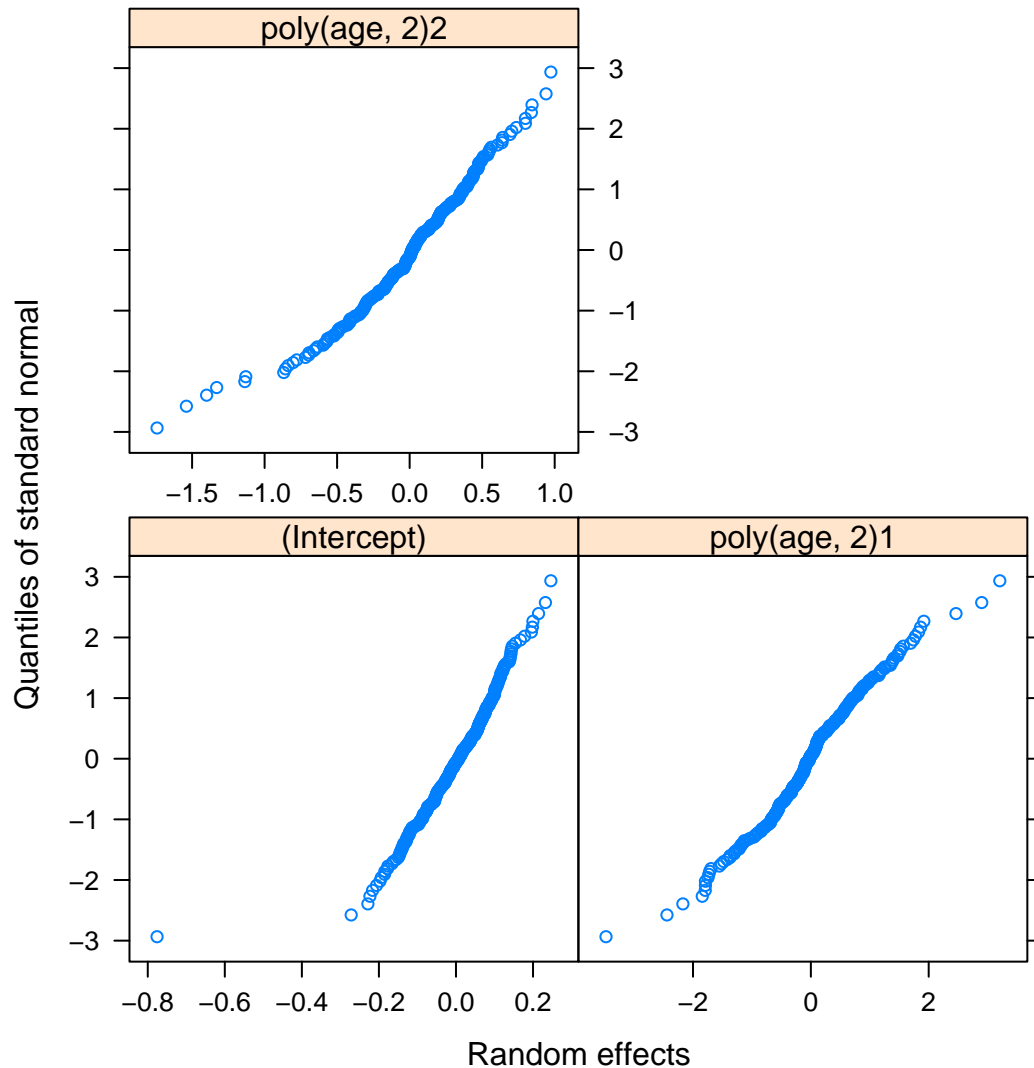


⇒ Variogramm der normalisierten Residuen zeigt keine deutliche serielle Korrelation.

(c) Überprüfen Sie die Verteilungsannahme für die zufälligen Effekte.

Lösung:

```
R>print(qqnorm(m12, ~ranef(.), cex = 0.7))
```



- (d) Erweitern Sie das Modell um eine serielle Korrelation und vergleichen Sie das erweiterte Modell mit dem ohne serielle Korrelationsstruktur. Probieren Sie verschiedene Typen von serieller Korrelation aus.

Hinweis: Falls der Schätzalgorithmus nicht konvergieren sollte, vereinfachen Sie die Kovarianzstruktur der zufälligen Effekte (pdDiag!) bevor Sie die serielle Korrelation mit ins Modell aufnehmen.

Lösung:

```
R>try(m12.Exp <- update(m12, corr = corExp(form = ~age | subject)))
R>try(m12.Gauss <- update(m12, corr = corGaus(form = ~age | subject)))
R>m12.diag <- update(m12, random = pdDiag(~poly(age, 2)))
R>m12.Exp <- update(m12.diag, corr = corExp(form = ~age | subject))
R>m12.Gauss <- update(m12.diag, corr = corGaus(form = ~age | subject))
R>anova(m12, m12.Exp, m12.Gauss, test = F)
```

	Model	df	AIC	BIC	logLik
m12	1	12	-4619.850	-4552.676	2321.925
m12.Exp	2	10	-4665.647	-4609.668	2342.823
m12.Gauss	3	10	-4650.000	-4594.021	2335.000

⇒ Aufnahme der seriellen Korrelation für den Fit offensichtlich wichtig (wichtiger als Korrelationen zwischen zufälligen Effekten). Hätte man aus dem Variogramm (s.o.) so nicht vermutet.

```
R>getVarCov(m12)
```

```
Random effects variance covariance matrix
      (Intercept) poly(age, 2)1 poly(age, 2)2
(Intercept)      0.011472      0.026298      -0.019735
poly(age, 2)1     0.026298      1.519700      -0.481450
poly(age, 2)2    -0.019735     -0.481450      0.476690
Standard Deviations: 0.10711 1.2328 0.69043
```

```
R>getVarCov(m12.diag)
```

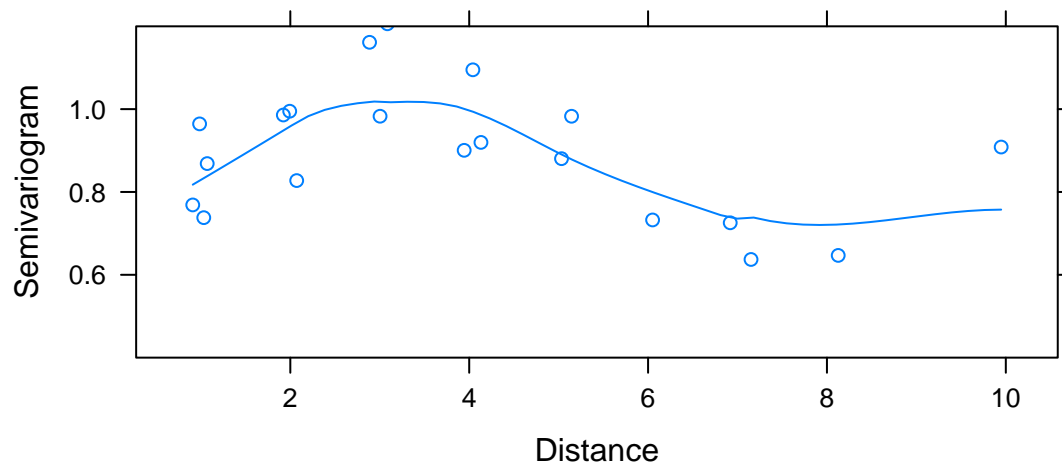
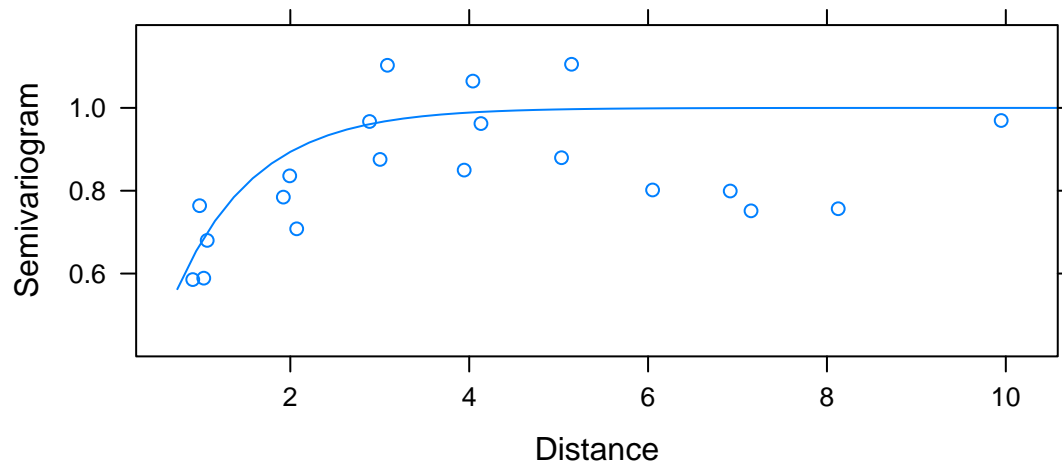
```
Random effects variance covariance matrix
      (Intercept) poly(age, 2)1 poly(age, 2)2
(Intercept)      0.011341      0.0000      0.00000
poly(age, 2)1     0.000000      1.4364      0.00000
poly(age, 2)2     0.000000      0.0000      0.48523
Standard Deviations: 0.1065 1.1985 0.69658
```

```
R>getVarCov(m12.Exp)
```

```
Random effects variance covariance matrix
      (Intercept) poly(age, 2)1 poly(age, 2)2
(Intercept)      0.010818      0.00000      0.000000
poly(age, 2)1     0.000000      0.67118      0.000000
poly(age, 2)2     0.000000      0.00000      0.024072
Standard Deviations: 0.10401 0.81925 0.15515
```

⇒ Offensichtlich wird (Ko-)Varianzstruktur der Daten durch serielle Korrelation gut beschrieben- Varianzen der zufälligen Effekte werden bei Berücksichtigung der seriellen Korrelation deutlich kleiner.

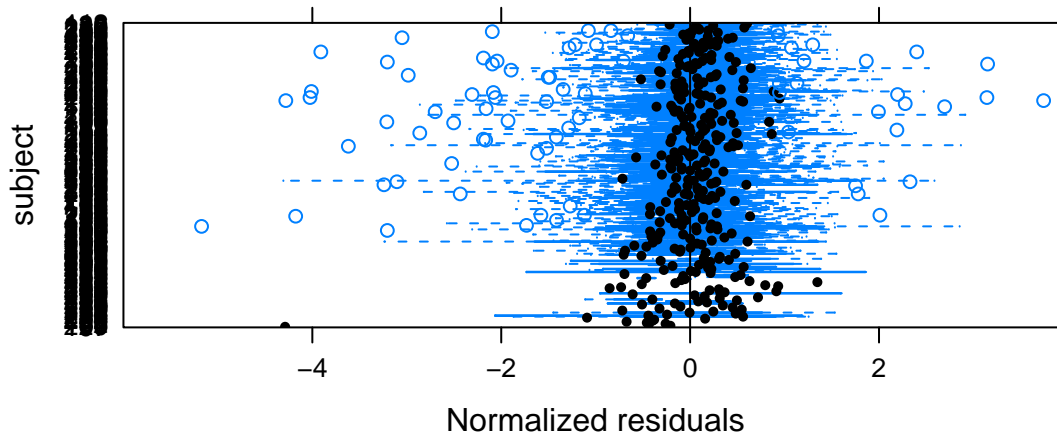
```
R>Vg.Exp <- plot(Variogram(m12.Exp, form = ~age | subject, resType = "p"),
+               ylim = c(0.4, 1.2))
R>Vg.12 <- plot(Variogram(m12, form = ~age | subject, resType = "p"),
+               ylim = c(0.4, 1.2))
R>print(Vg.Exp, split = c(1, 1, 1, 2), more = TRUE)
R>print(Vg.12, split = c(1, 2, 1, 2))
```



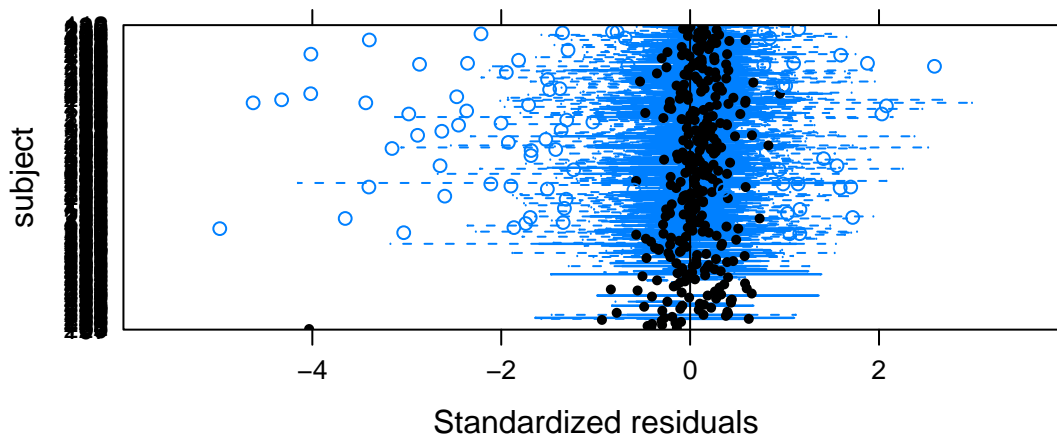
Grafische Überprüfung zeigt einigermaßen gute Anpassung an das Semi-Variogramm bis Distance= 5 und leichte Verbesserung durch Hereinnahme der Exponential-Korrelationsfunktion.

```
R>res.serial <- plot(m12.Exp, subject ~ resid(., type = "n"), xlim = c(-6,
+ 4), cex = 0.5, abline = 0, main = "Normalisierte Residuen, mit Exponential-Korrelati
R>res.noserial <- plot(m12, subject ~ resid(., type = "n"), , xlim = c(-6,
+ 4), cex = 0.5, abline = 0, main = "ohne Exponential-Korrelation")
R>print(res.serial, split = c(1, 1, 1, 2), more = TRUE)
R>print(res.noserial, split = c(1, 2, 1, 2))
```

Normalisierte Residuen, mit Exponential-Korrelation



ohne Exponential-Korrelation



Teilweise kleine Verbesserung sichtbar - Residuen symmetrischer verteilt.

Aufgabe 2:

- (a) Programmieren Sie eine R-Funktion, die die auf Folie 8.29 präsentierte Methode zur Identifikation von extremen Subjekten implementiert.

Lösung:

```
R>mahalanobis.residcheck <- function(m, p = 0.01) {  
+   grps <- getGroups(m)  
+   ng <- table(grps)  
+   grplevels <- levels(grps)  
+   r <- resid(m, level = 0, type = "response")  
+   r <- split(r, grps)  
+   for (i in grplevels) {  
+     if (ng[i] > 1) {  
+       V <- getVarCov(m, individual = i, type = "marginal")
```

```

+           r[[i]] <- backsolve(chol(V[[1]]), diag(ng[i])) %*%
+             r[[i]]
+       }
+       else r[[i]] <- NA
+     }
+     d <- sapply(r, crossprod)
+     pvals <- 1 - pchisq(q = d, df = ng)
+     res <- data.frame(N = as.numeric(ng), P = pvals, out = pvals <
+       p)
+     return(res[complete.cases(res), ])
+ }

```

(m ist ein lme-Objekt)

- (b) Benutzen Sie die Funktion um auf Basis des End-Modells aus Aufgabe 1 Ausreißer-Subjekte zu identifizieren. Benutzen Sie $p < 0.01$ als Kriterium für Ausreißer.

Lösung:

```

R>mah <- mahalanobis.residcheck(m12.Exp, 0.01)
R>(out <- rownames(mah)[mah$out])

```

```

 [1] "147" "223" "81"  "207" "293" "237" "69"  "79"  "85"  "246" "80"  "9"
[13] "7"   "241" "263" "120" "73"  "163" "2"   "101" "10"  "208" "126" "286"
[25] "118" "131" "252" "259" "30"  "37"  "265" "95"  "150" "66"  "42"  "270"
[37] "140" "139" "112"

```

- (c) Überprüfen Sie den Einfluss dieser extremen Beobachtungen auf die Modellschätzung.

Lösung:

```

R>m12.Exp.noout <- update(m12.Exp, data = subset(fev, !(subject %in%
+   out)))
R>summary(m12.Exp.noout)

```

```

Linear mixed-effects model fit by maximum likelihood
Data: subset(fev, !(subject %in% out))
      AIC      BIC    logLik
-4318.182 -4263.958 2169.091

```

Random effects:

```

Formula: ~poly(age, 2) | subject
Structure: Diagonal
      (Intercept) poly(age, 2)1 poly(age, 2)2 Residual
StdDev:  0.09221563      0.7291606  0.0003207516 0.05699173

```

Correlation Structure: Exponential spatial correlation

```

Formula: ~age | subject
Parameter estimate(s):
      range
0.9261662

```

```
Fixed effects: log.fev1 ~ poly(age, 2) + poly(height, 2)
              Value Std.Error   DF   t-value p-value
(Intercept)   0.814443 0.0061859 1408 131.66073  0.0000
poly(age, 2)1   3.340506 0.3559605 1408   9.38448  0.0000
poly(age, 2)2  -0.450149 0.1499318 1408  -3.00236  0.0027
poly(height, 2)1 9.513892 0.3868323 1408 24.59436  0.0000
poly(height, 2)2 0.283038 0.1090745 1408   2.59490  0.0096
```

```
Correlation:
              (Intr) ply(g,2)1 ply(g,2)2 ply(h,2)1
poly(age, 2)1  -0.005
poly(age, 2)2   0.010 -0.778
poly(height, 2)1 0.030 -0.965    0.805
poly(height, 2)2 -0.019  0.353   -0.623   -0.364
```

```
Standardized Within-Group Residuals:
              Min           Q1           Med           Q3           Max
-4.79259420 -0.50763153  0.05431366  0.58572296  2.80711173
```

```
Number of Observations: 1673
Number of Groups: 261
```

```
R>summary(m12.Exp)
```

```
Linear mixed-effects model fit by maximum likelihood
Data: fev
              AIC           BIC    logLik
-4665.647 -4609.668 2342.823
```

```
Random effects:
Formula: ~poly(age, 2) | subject
Structure: Diagonal
              (Intercept) poly(age, 2)1 poly(age, 2)2 Residual
StdDev:      0.1040084      0.8192537      0.1551504 0.0643803
```

```
Correlation Structure: Exponential spatial correlation
Formula: ~age | subject
Parameter estimate(s):
range
0.8915407
```

```
Fixed effects: log.fev1 ~ poly(age, 2) + poly(height, 2)
              Value Std.Error   DF   t-value p-value
(Intercept)   0.813080 0.0064637 1690 125.79077  0.0000
poly(age, 2)1   3.807507 0.3946520 1690   9.64776  0.0000
poly(age, 2)2  -0.595127 0.1669998 1690  -3.56364  0.0004
poly(height, 2)1 10.163900 0.4319583 1690 23.52982  0.0000
poly(height, 2)2 0.306701 0.1211550 1690   2.53148  0.0114
```

```
Correlation:
              (Intr) ply(g,2)1 ply(g,2)2 ply(h,2)1
poly(age, 2)1  -0.003
poly(age, 2)2   0.004 -0.773
poly(height, 2)1 0.024 -0.966    0.799
poly(height, 2)2 -0.013  0.346   -0.618   -0.358
```

```
Standardized Within-Group Residuals:
```

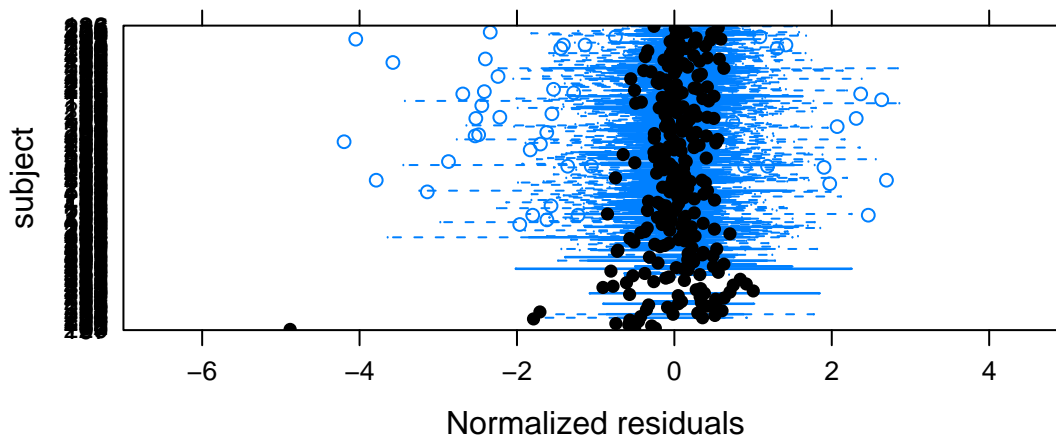
Min	Q1	Med	Q3	Max
-5.91484309	-0.47513143	0.07299316	0.55869144	2.73445757

Number of Observations: 1994

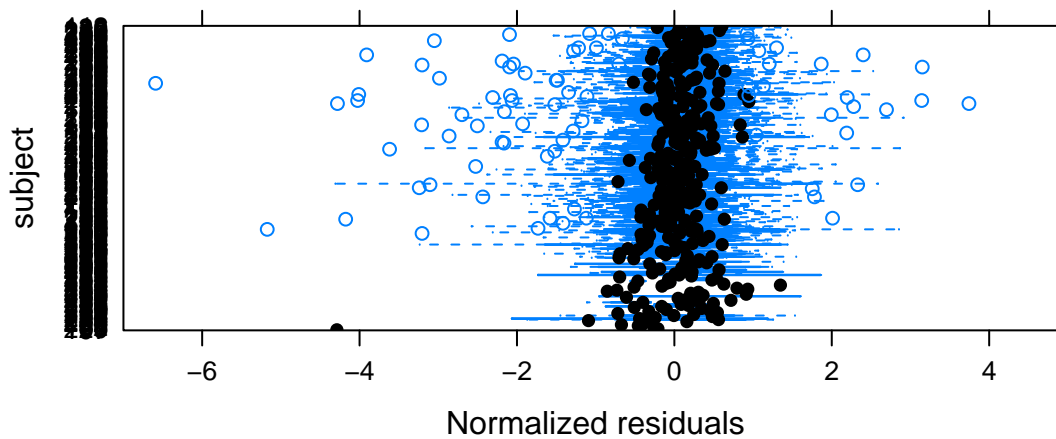
Number of Groups: 300

```
R>r.m12.noout <- plot(m12.Exp.noout, subject ~ resid(., type = "n"),
+   cex = 0.7, main = "Ausreißer Subjekte entfernt", xlim = c(-7,
+   5))
R>r.m12 <- plot(m12.Exp, subject ~ resid(., type = "n"), cex = 0.7,
+   main = "mit Ausreißern", xlim = c(-7, 5))
R>print(r.m12.noout, split = c(1, 1, 1, 2), more = TRUE)
R>print(r.m12, split = c(1, 2, 1, 2))
```

Ausreißer Subjekte entfernt



mit Ausreißern



⇒ geringere Variabilität in den Daten führt zu geringfügig kleineren Schätzungen der Varianzkomponenten und etwas stärkerer serieller Korrelation. Schätzung der festen Effekte ebenfalls nur geringfügig verändert. Standardisierte Residuen symmetrischer (Klar, weil 'problematische' Subjekte nicht mehr im Datensatz).