

Aufgabe 1:

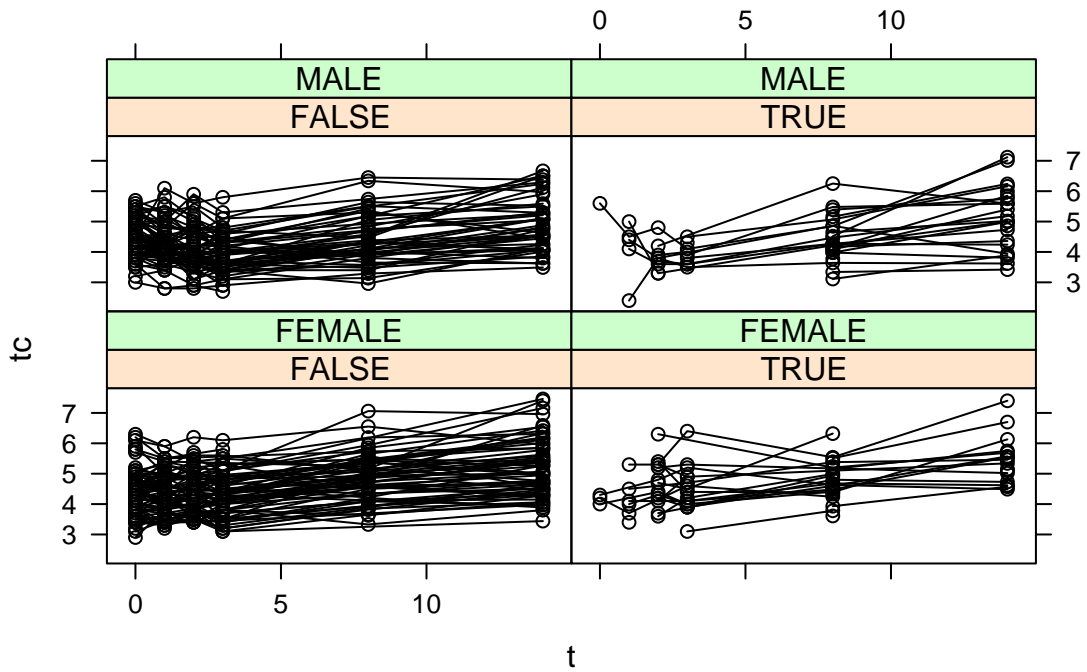
In der Amsterdamer Wachstums- und Gesundheitsstudie wurde das Gesamt-Serumcholesterol (TC) von 147 Probanden insgesamt 6 Mal pro Proband gemessen. Zu Beginn der Studie 1977 ($t = 0$) waren alle Probanden 13 Jahre alt. Die ersten vier Messungen wurden im Jahresabstand, die letzten beiden Messungen 1985 und 1991 durchgeführt. Die Daten sind in der Datei `amsterdam.txt` von der Webseite der Veranstaltung erhältlich – die Datei `amsterdam-description.txt` enthält eine genauere Beschreibung der Daten und Kovariablen.

- (a) Erstellen Sie mit der Funktion `xyplot` einen Spaghettiplot der TC-Messungen für die vier möglichen Kombinationen von `smoker` und `gender`. Benutzen Sie `bwplot`, um für die beiden Geschlechter Boxplots der TC-Messungen über die Zeit zu erstellen. Kommentieren Sie die Plots.

Lösung:

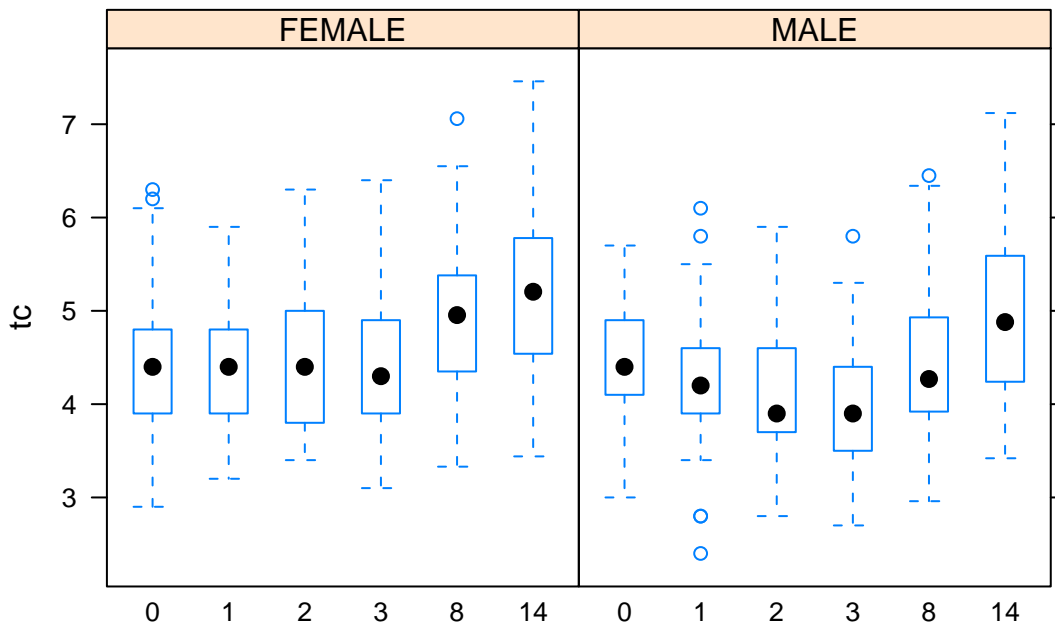
(Aufgabe adaptiert von Michael Hoehle, SoSe07.)

```
R>library(MASS)
R>library(lattice)
R>library(gee)
R>library(nlme)
R>set.seed(123)
R>url <- "http://www.statistik.lmu.de/institut/lehrstuhl/semwiso/longitudinal_ss08/"
R>ex <- read.table(paste(url, "download/amsterdam.txt", sep = ""),
+   header = T)
R>ex$smoker <- factor(ex$smoker)
R>print(xyplot(tc ~ t | smoker * gender, groups = id, data = ex,
+   type = "b", col = 1))
```



Smoker variiert zeitlich.

```
R>print(bwplot(tc ~ as.factor(t) | gender, groups = id, data = ex,
+ type = "b", horizontal = FALSE))
```



(b) Berechnen Sie für jeden Zeitpunkt Mittelwerte und Standardabweichungen der TC-Messungen für die beiden Geschlechter.

Lösung:

```
R>t <- unique(ex$t)
R>E <- with(ex, tapply(tc, list(t, gender), mean))
R>SD <- with(ex, tapply(tc, list(t, gender), sd))
R>res <- cbind(E[, 1], SD[, 1], E[, 2], SD[, 2])
R>colnames(res) <- c("E-FEMALE", "SD-FEMALE", "E-MALE", "SD-MALE")
R>res
```

```
      E-FEMALE SD-FEMALE  E-MALE  SD-MALE
0  4.408974 0.7263475 4.463768 0.6128456
1  4.379487 0.6665201 4.259420 0.6640603
2  4.434615 0.6982031 4.075362 0.6706901
3  4.370513 0.7222096 3.950725 0.6098754
8  4.880769 0.7230936 4.431739 0.7832098
14 5.252949 0.9068056 4.973043 0.9262405
```

- (c) Berechnen Sie die empirischen Korrelationen zwischen den Messzeitpunkten. Kommentieren Sie im Hinblick auf die empirischen Korrelation Vor- und Nachteile der Korrelationsstrukturen `independence`, `unstructured`, `exchangeable`/compound symmetry und `stat_M_dep`/banded Toeplitz als Arbeitskorrelationsmatrizen bei einer GEE-Modellierung der Daten. Geben Sie für jede Korrelationsstruktur auch die Anzahl zu schätzenden Parameter an.

Lösung:

```
R>ex.wide <- reshape(ex, timevar = "t", idvar = "id", direction = "wide",
+   v.names = c("tc", "bfatness", "smoker"))
R>cor <- cor(ex.wide[, paste("tc.", t, sep = "")])
R>print(cor, digits = 2)
```

```
      tc.0 tc.1 tc.2 tc.3 tc.8 tc.14
tc.0  1.00 0.76 0.70 0.67 0.64  0.59
tc.1  0.76 1.00 0.77 0.78 0.67  0.59
tc.2  0.70 0.77 1.00 0.85 0.71  0.63
tc.3  0.67 0.78 0.85 1.00 0.74  0.65
tc.8  0.64 0.67 0.71 0.74 1.00  0.69
tc.14 0.59 0.59 0.63 0.65 0.69  1.00
```

```
R>round(range(cor[row(cor) != col(cor)]), 2)
```

```
[1] 0.59 0.85
```

Die Korrelationen zwischen den unterschiedlichen Messzeitpunkten liegen im Bereich 0.59 – 0.85. Das ist definitiv mehr als bei Unabhängigkeit (`independence`: 1 Varianzparameter, 0 Korrelationsparameter), und eine zu große Variation, um Compound-Symmetry (`exchangeable`: 1 Varianzparameter, 1 Korrelationsparameter) zu rechtfertigen – außerdem ist diese Form problematisch, weil die Zeitintervalle nicht äquidistant sind. `Unstructured` hat 1 Varianzparameter und $(5 \cdot 4) / 2 = 10$ Korrelationsparameter. Das Problem bei banded-Toeplitz (in `gee`: `stat_M_dep`) (bei Ordnung $M = 5$: 1 Varianzparameter, 5 Korrelationsparameter) ist, dass die letzten zwei Messungen in anderen Zeitabständen erhoben wurden als die ersten vier - d.h. es macht z.B. keinen Sinn anzunehmen, dass die serielle Korrelation zwischen Y_{i1} und Y_{i3} (zeitlicher Abstand: 2 Jahre) gleich der seriellen Korrelation zwischen

Y_{i4} und Y_{i6} (zeitlicher Abstand: 10 Jahre) ist.

- (d) Fitten Sie GEE-Modelle, bei denen nur Intercept sowie die Haupteffekte für t , $fitness$, $bfatness$, $gender$ und $smoker$ eingehen. Erstellen Sie eine Tabelle, in der Sie Regressionskoeffizienten und robuste Standardfehler für jede der Korrelationsstrukturen aus Teil (c) angeben. Kommentieren Sie die Unterschiede in $\hat{\beta}$ bzw. $\hat{se}(\hat{\beta})$.

Lösung:

```
R>lcorstr <- list("independence", "stat_M_dep", "exchangeable",
+ "unstructured")
R>gees <- lapply(lcorstr, function(corstr) {
+ print(corstr)
+ gee(tc ~ 1 + t + fitness + bfatness + gender + smoker, corstr = corstr,
+ Mv = ifelse(corstr == "stat_M_dep", 5, 0), data = ex)
+ })
```

```
[1] "independence"
(Intercept)      t      fitness      bfatness  genderMALE  smokerTRUE
 3.55132356 0.04891857 0.07634050 0.16077213 -0.05419639 -0.04969867
[1] "stat_M_dep"
(Intercept)      t      fitness      bfatness  genderMALE  smokerTRUE
 3.55132356 0.04891857 0.07634050 0.16077213 -0.05419639 -0.04969867
[1] "exchangeable"
(Intercept)      t      fitness      bfatness  genderMALE  smokerTRUE
 3.55132356 0.04891857 0.07634050 0.16077213 -0.05419639 -0.04969867
[1] "unstructured"
(Intercept)      t      fitness      bfatness  genderMALE  smokerTRUE
 3.55132356 0.04891857 0.07634050 0.16077213 -0.05419639 -0.04969867
```

```
R>estimate <- sapply(1:length(gees), function(i) coef(gees[[i]]))
R>sd <- sapply(1:length(gees), function(i) sqrt(diag(gees[[i]]$robust.variance)))
R>colnames(estimate) <- lcorstr
R>colnames(sd) <- lcorstr
R>print(estimate, digits = 2)
```

	independence	stat_M_dep	exchangeable	unstructured
(Intercept)	3.551	4.227	4.227	4.310
t	0.049	0.055	0.057	0.062
fitness	0.076	-0.096	-0.101	-0.154
bfatness	0.161	0.077	0.078	0.074
genderMALE	-0.054	-0.130	-0.128	-0.093
smokerTRUE	-0.050	-0.136	-0.160	-0.134

```
R>print(sd, digits = 2)
```

	independence	stat_M_dep	exchangeable	unstructured
(Intercept)	0.5175	0.5206	0.5207	0.4920
t	0.0048	0.0048	0.0047	0.0051
fitness	0.2712	0.2789	0.2788	0.2592
bfatness	0.0259	0.0193	0.0198	0.0234
genderMALE	0.1313	0.1322	0.1317	0.1302
smokerTRUE	0.0892	0.0529	0.0540	0.0579

Die Standardfehler sind alle ähnlich, was auf die robuste Schätzung zurückzuführen ist: Eine Arbeitskorrelationsmatrix wird angenommen, mittels Sandwichschätzer wird jedoch die eigentliche Korrelationsstruktur der Daten geschätzt. Die Variation in den Parameterschätzern ist relativ groß, z.B. ist für `fitness` sogar das Vorzeichen bei `independence` unterschiedlich von dem Vorzeichen bei den anderen Arbeitskorrelationen. Dieser Parameter hat aber auch einen entsprechend großen Standardfehler.

- (e) Passen Sie an die Daten ein LME-Modell mit Random-Intercept und Random-Slope an. Geben Sie die Modellformel und die entsprechenden Verteilungsannahmen für die vorkommenden Zufallsvariablen an. Identifizieren Sie aus dem Output die Schätzer aller auftretenden Parameter.

Lösung:

Die Modellformel ist:

$$y_{ij} | \mathbf{b}_i = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + \varepsilon_{ij},$$

wobei $\mathbf{b}_i = (b_{0i}, b_{1i})' \stackrel{i.i.d.}{\sim} \mathcal{N}_2(0, \mathbf{D})$ und $\varepsilon_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$.

```
R>m.slope.lme <- lme(tc ~ 1 + t, random = ~1 + t | id, data = ex)
R>summary(m.slope.lme)
```

Linear mixed-effects model fit by REML

```
Data: ex
      AIC      BIC    logLik
1422.894 1451.574 -705.4472
```

Random effects:

```
Formula: ~1 + t | id
Structure: General positive-definite, Log-Cholesky parametrization
      StdDev   Corr
(Intercept) 0.57919877 (Intr)
t            0.03780896 0.077
Residual    0.40019466
```

Fixed effects: tc ~ 1 + t

```
      Value Std.Error DF t-value p-value
(Intercept) 4.221739 0.05127607 734 82.33351 0
t           0.059229 0.00416242 734 14.22945 0
```

Correlation:

```
(Intr)
t -0.113
```

Standardized Within-Group Residuals:

```
      Min      Q1      Med      Q3      Max
-2.73505106 -0.59986461 -0.01514336 0.56513568 2.99102873
```

Number of Observations: 882

Number of Groups: 147

Das heißt also: $\hat{\beta} =$

```
(Intercept)      t
4.22174      0.05923
```

und $\sqrt{\hat{\sigma}_\varepsilon^2}=0.4002$, und \hat{D} ist

```
R>getVarCov(m.slope.lme)
```

```
Random effects variance covariance matrix
      (Intercept)      t
(Intercept)  0.3354700 0.0016764
t            0.0016764 0.0014295
Standard Deviations: 0.5792 0.037809
```

- (f) Testen Sie, ob der Random-Slope benötigt wird. Was müssen Sie bei diesem Test beachten? Wie wäre eine Ablehnung der Nullhypothese inhaltlich zu interpretieren?

Lösung:

```
R>m.noslope.lme <- update(m.slope.lme, random = ~1 | id)
R>anova(m.noslope.lme, m.slope.lme)
```

	Model	df	AIC	BIC	logLik	Test	L.Ratio	p-value
m.noslope.lme	1	4	1475.129	1494.249	-733.5645			
m.slope.lme	2	6	1422.894	1451.574	-705.4472	1 vs 2	56.23466	<.0001

Die χ_2^2 -Verteilung ist hier eine konservative Approximation der Null-Verteilung für die LQ-Statistik, denn beim Testen eines Varianzparameter auf Null ist die Nullhypothese am Rand des Parameterraums. Dadurch ist die gewöhnliche Asymptotik ungültig. Alternativen sind die χ^2 -Mischung von Stram & Lee oder Simulation der LQ-Statistik mittels Bootstrap (z.B. über `simulate.lme()`). Da hier jedoch schon der einfache LQ-Test H_0 ablehnt sind diese Erweiterungen nicht nötig und ein Random-Slope ist notwendig. Inhaltlich wäre eine Ablehnung so zu interpretieren dass keine signifikante Variabilität in der Veränderung des TC-Wertes zwischen den Probanden besteht.

- (g) Skizzieren Sie mit Hilfe eines Plots von t gegen die TC-Werte, was positive bzw. negative Korrelation zwischen den beiden Random-Effects bedeutet.

Lösung:

Positive Korrelation: Individuen mit großen Intercept haben tendenziell auch eine größere Slope, Individuen mit kleinem Intercept haben tendenziell auch eine kleinere Slope; d.h. inhaltlich: für Jugendliche mit zu Studienbeginn überdurchschnittlichem TC steigen tendenziell auch im Verlauf der Studiendauer die TC-Werte überdurchschnittlich stark an, für Jugendliche mit zu Studienbeginn unterdurchschnittlichem TC steigen tendenziell auch im Verlauf der Studiendauer die TC-Werte weniger stark an.

Negative Korrelation: Individuen mit großen Intercept haben tendenziell eine kleinere Slope als Individuen mit kleinem Intercept. D.h. inhaltlich: für Jugendliche mit zu Studienbeginn überdurchschnittlichem TC steigen tendenziell im Verlauf der Studiendauer die TC-Werte unterdurchschnittlich stark an, für Jugendliche mit zu Studienbeginn unterdurchschnittlichem TC dagegen steigen tendenziell im Verlauf der Studiendauer die TC-Werte überdurchschnittlich stark an.

```

R>cor.pos <- matrix(c(1, 0.9, 0.9, 1), 2, 2)
R>cor.neg <- matrix(c(1, -0.9, -0.9, 1), 2, 2)
R>library(mvtnorm)
R>do.plot <- function(corM, main) {
+   res <- mvrnorm(n = 10, mu = c(5, 2), Sigma = corM)
+   plot(NA, xlim = c(0, 5), ylim = c(3, 10), type = "n", xlab = "t",
+        ylab = "", main = main)
+   abline(coef = c(5, 2), lty = 1, lwd = 4, col = "darkgrey")
+   for (i in 1:nrow(res)) abline(coef = res[i, ])
+ }
R>set.seed(123)
R>par(mfcol = c(1, 2))
R>do.plot(cor.pos, "Positive Korrelation")
R>do.plot(cor.neg, "Negative Korrelation")

```

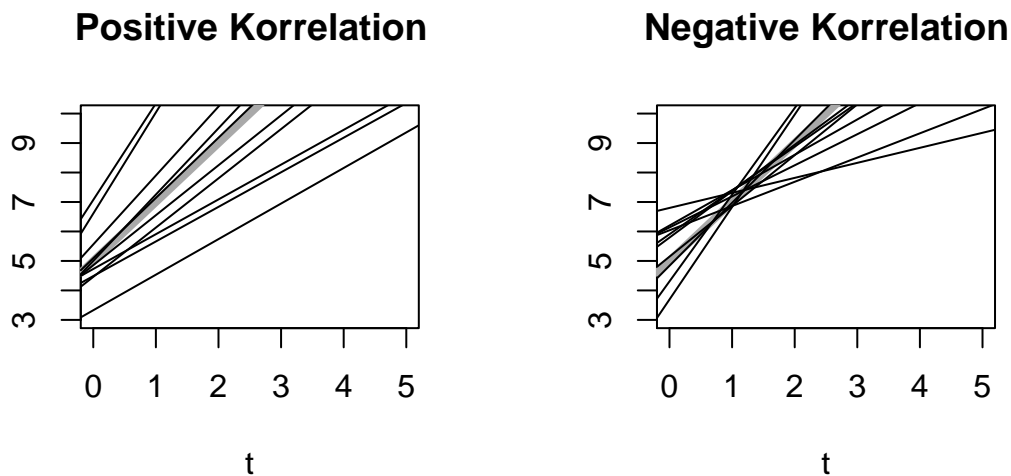


Abbildung 1: Positiv und negativ korrelierte Random Slopes und Intercepts. Dicke graue Linie entspricht Populationsmittel mit $\beta_0 = 5$ und $\beta_1 = 2$.

- (h) Wieso liefern ein LME mit Random-Intercept und ein GEE mit Sandwich-Schätzung (= robust) basierend auf einer `exchangeable`-Arbeitskovarianz fast die gleichen Ergebnisse für $\hat{\beta}$, nicht aber für deren Standardfehler? Warum sind die Ergebnisse für die modellbasierte (=naive) Schätzung so ähnlich?

Lösung:

```

R>m.lme <- lme(tc ~ 1 + t, random = ~1 | id, data = ex, method = "ML")
R>m.gee <- gee(tc ~ 1 + t, id = id, corstr = "exchangeable", data = ex)

```

```

(Intercept)          t
4.22173897  0.05922892

```

```

R>fixed.effects(m.lme)

```

```

(Intercept)          t
4.22173897  0.05922892

```

```
R>vcov(m.lme)
```

```
              (Intercept)              t
(Intercept)  2.970363e-03 -4.443834e-05
t            -4.443834e-05  9.522501e-06
```

```
R>m.gee$robust.variance
```

```
              (Intercept)              t
(Intercept)  2.611350e-03 -2.390567e-05
t            -2.390567e-05  1.720807e-05
```

```
R>m.gee$naive.variance
```

```
              (Intercept)              t
(Intercept)  2.974755e-03 -4.465201e-05
t            -4.465201e-05  9.568287e-06
```

Ein LME mit Random-Intercept spezifiziert (fast) die gleiche Korrelationsstruktur wie die `exchangeable` (=Compound Symmetry) Korrelationsstruktur im GEE-Modell, nämlich $\text{Corr}(Y_{it_j}, Y_{it_k}) = \rho \forall t_j, t_k$.

Im LME mit Random Intercepts b_{0i} ist $\rho = \text{Var}(b_0)/(\text{Var}(\varepsilon) + \text{Var}(b_0))$, also immer positiv. Im GEE-Modell können auch negative within-subject Korrelationen modelliert/geschätzt werden.

Die robuste Schätzung bei GEEs berücksichtigt die beobachtete Korrelationsstruktur während sie beim LME und bei der naiven Schätzung aus dem GEE fest durch das Modell vorgegeben ist (und in diesem Spezialfall hier übereinstimmt).

Aufgabe 2:

Für die Amsterdam-Daten ist außerdem von Interesse ob die Studienteilnehmer in eine kardiovaskuläre Risikogruppe fallen. Für jeden Zeitpunkt t ist der kardiovaskuläre Risikoindikator Y_{it}^* (0=klein, 1=groß) von Individuum i durch $Y_{it}^* = I(Y_{it} > q_{0.75}(t))$ gegeben, wobei $q_{0.75}(t)$ das 75%-Quantil der beobachteten TC-Messungen zum Zeitpunkt t ist.

Führen Sie diese Dichotomisierung für die Amsterdam-Daten durch und fügen Sie diese Größe als `tc.dich` zu den Daten hinzu.

Hinweis: Zur Kontrolle enthält die Datei `tc.dich.txt` die gewünschten Werte.

Lösung:

```
R>t <- unique(ex$t)
R>q75 <- sapply(t, function(time) {
+   quantile(subset(ex, t == time)$tc, 0.75)
+ })
R>tc.dich <- (ex$tc > rep(q75, length(unique(ex$id)))) * 1
R>ex.dich <- cbind(ex, tc.dich)
```

Im Folgenden sollen Sie die Auswirkung 3 verschiedener Missing-Mechanismen an diesen Daten ausprobieren:

- (a) **Szenario 1:** Das für die Cholesterol-Messungen zuständige Labor arbeitete schlecht. Es zeigt sich dass im Durchschnitt etwa ein Drittel der Proben falsch analysiert wurde und als fehlende Messung eingestuft werden muss.
- (b) **Szenario 2:** Kinder und Jugendliche, die als Risikopatienten eingestuft wurden ($Y_{it_{c,i}}^* = 1$), werden zur Behandlung an entsprechende Spezialisten verwiesen und im Folgenden ($t > t_{c,i}$) nicht mehr im Rahmen der vorliegenden Studie betrachtet, weil durch die entsprechende Medikation das natürliche Entwicklungsmuster gestört wird.
- (c) **Szenario 3:** Mit einer Wahrscheinlichkeit von 0.8 fallen Kinder und Jugendliche, deren TC-Wert zum Zeitpunkt t über dem entsprechenden 75%-Quantil liegt, bereits vor der Messung zum Zeitpunkt t aus der Studie, weil ihre besorgte Eltern sie in hermetisch abgeriegelte Diät-Camps haben einliefern lassen, in denen sie den Rest ihrer traurigen Jugend ohne Frühstückseier und Mayonnaise verbringen müssen.

Überlegen Sie für jedes der 3 Szenarien um was für einen Missing-Mechanismus es sich handelt. Erzeugen Sie nach den hier beschriebenen Mechanismen aus den echten Daten Datensätze mit fehlenden Werten und vergleichen Sie die Punktschätzer der festen Effekte und der Varianzparameter auf Basis dieser Datensätze mit denen aus den echten Daten.

Lösung:

Szenario 1 ist MCAR; Szenario 2 ist MAR weil das Fehlen nur von beobachteten Werten der Response abhängt; Szenario 3 ist NMAR weil das Fehlen von unbeobachteten Werten der Response abhängt.

```
R>set.seed(123)
R>MCAR <- MAR <- NMAR <- ex.dich
R>for (i in unique(ex$id)) {
+   for (j in unique(ex$t)) {
+     if (runif(1) < 0.3333)
+       MCAR[(ex$id == i & ex$t == j), "tc"] <- NA
+     if (MAR[(ex$id == i & ex$t == j), "tc.dich"] == 1)
+       MAR[(ex$id == i & ex$t > j), "tc"] <- NA
+     if ((NMAR[(ex$id == i & ex$t == j), "tc.dich"] == 1) &
+         (runif(1) < 0.8))
+       NMAR[(ex$id == i & ex$t >= j), "tc"] <- NA
+   }
+ }
R>MCAR <- groupedData(tc ~ t | id, MCAR)
R>MAR <- groupedData(tc ~ t | id, MAR)
R>NMAR <- groupedData(tc ~ t | id, NMAR)
R>m.noslope.lme
```

Linear mixed-effects model fit by REML

```
Data: ex
Log-restricted-likelihood: -733.5645
Fixed: tc ~ 1 + t
(Intercept)          t
4.22173897  0.05922892
```

Random effects:

```
Formula: ~1 | id
(Intercept) Residual
StdDev:    0.6127439 0.4482326
```

Number of Observations: 882
 Number of Groups: 147

```
R>mMCAR <- update(m.noslope.lme, data = MCAR, na.action = na.exclude)
R>mMAR <- update(m.noslope.lme, data = MAR, na.action = na.exclude)
R>mNMAR <- update(m.noslope.lme, data = NMAR, na.action = na.exclude)
R>compareEstimates <- cbind(NoMiss = c(fixef(m.noslope.lme), as.numeric(VarCorr(m.noslope.lme)
+ 2])), MCAR = c(fixef(mMCAR), as.numeric(VarCorr(mMCAR)[,
+ 2])), MAR = c(fixef(mMAR), as.numeric(VarCorr(mMAR)[, 2])),
+ NMAR = c(fixef(mNMAR), as.numeric(VarCorr(mNMAR)[, 2])))
R>rownames(compareEstimates) <- c("Intercept", "t", "sd(RandomIntercept)",
+ "sd(Error)")
R>compareEstimates
```

	NoMiss	MCAR	MAR	NMAR
Intercept	4.22173897	4.23543190	4.27602332	4.04384144
t	0.05922892	0.05880858	0.05163562	0.04383018
sd(RandomIntercept)	0.61274390	0.61784190	0.64496170	0.46749090
sd(Error)	0.44823260	0.44317010	0.40509540	0.37166170

NMAR führt zu deutlich anderen Ergebnissen, MAR zu leicht veränderten, MCAR ist unproblematisch.

Aufgabe 3:

Benutzen Sie im folgenden wieder den kompletten Datensatz. Von Interesse ist nun die Frage, wovon es abhängt ob Studienteilnehmer in die kardiovaskuläre Risikogruppe fallen.

- (a) Fitten Sie ein gemischtes logistisches Regressionsmodell mit festen Effekten (nur Haupteffekte) für `t`, `gender` und `smoker` sowie einem Random-Intercept.

Hinweis: Benutzen Sie `glmmML()` aus dem gleichnamigen Paket oder `lmer()` aus dem Paket `lme4`.

Lösung:

```
R>library(glmmML)
R>m.glme.1 <- glmmML(tc.dich ~ 1 + t + smoker + gender, cluster = id,
+ family = binomial, data = ex.dich)
R>m.glme.2 <- glmmML(tc.dich ~ 1 + t + smoker + gender, cluster = id,
+ data = ex.dich, method = "ghq", n.points = 25)
R>round((coef(m.glme.2) - coef(m.glme.1))/coef(m.glme.2), 4)
```

(Intercept)	t	smokerTRUE	genderMALE
-0.0427	0.0170	0.0176	-0.0621

```
R>summary(m.glme.1)
```

Call: `glmmML(formula = tc.dich ~ 1 + t + smoker + gender, family = binomial, data = ex.dich)`

	coef	se(coef)	z	Pr(> z)
(Intercept)	-2.01804	0.39096	-5.162	2.45e-07

```
t          0.02827  0.02290  1.235 2.17e-01
smokerTRUE -0.46441  0.36585 -1.269 2.04e-01
genderMALE -1.01906  0.52975 -1.924 5.44e-02
```

```
Standard deviation in mixing distribution: 2.565
Std. Error:                               0.2539
```

```
Residual deviance: 749.3 on 877 degrees of freedom  AIC: 759.3
```

```
R>summary(m.glme.2)
```

```
Call: glmmML(formula = tc.dich ~ 1 + t + smoker + gender, data = ex.dich, cluster = )
```

```
              coef se(coef)      z Pr(>|z|)
(Intercept) -1.93549  0.38840 -4.983 6.25e-07
t            0.02875  0.02308  1.246 2.13e-01
smokerTRUE  -0.47275  0.36786 -1.285 1.99e-01
genderMALE  -0.95947  0.52755 -1.819 6.90e-02
```

```
Standard deviation in mixing distribution: 2.585
Std. Error:                               0.3128
```

```
Residual deviance: 745.6 on 877 degrees of freedom  AIC: 755.6
```

```
R>detach(package:nlme)
```

```
R>library(lme4)
```

```
R>m.glme.3 <- lmer(tc.dich ~ 1 + t + smoker + gender + (1 | id),
+   family = binomial, data = ex.dich, method = "Laplace")
R>m.glme.4 <- lmer(tc.dich ~ 1 + t + smoker + gender + (1 | id),
+   family = binomial, data = ex.dich, method = "PQL")
```

```
R>summary(m.glme.3)
```

```
Generalized linear mixed model fit using Laplace
Formula: tc.dich ~ 1 + t + smoker + gender + (1 | id)
Data: ex.dich
Family: binomial(logit link)
AIC   BIC logLik deviance
759.3 783.2 -374.6   749.3
Random effects:
Groups Name      Variance Std.Dev.
id      (Intercept) 6.5768   2.5645
number of obs: 882, groups: id, 147
```

```
Estimated scale (compare to 1 ) 0.7387302
```

```
Fixed effects:
```

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.01950    0.35594  -5.674 1.40e-08 ***
t            0.02829    0.02265   1.249  0.2116
smokerTRUE  -0.46460    0.35915  -1.294  0.1958
genderMALE  -1.01729    0.51787  -1.964  0.0495 *
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
              (Intr) t      smTRUE
t              -0.257
smokerTRUE    -0.061 -0.332
genderMALE    -0.625  0.000  0.000
```

```
R>summary(m.glme.4)
```

```
Generalized linear mixed model fit using PQL
Formula: tc.dich ~ 1 + t + smoker + gender + (1 | id)
Data: ex.dich
Family: binomial(logit link)
AIC   BIC logLik deviance
768.4 792.3 -379.2   758.4
Random effects:
Groups Name      Variance Std.Dev.
id      (Intercept) 9.0587   3.0098
number of obs: 882, groups: id, 147
```

```
Estimated scale (compare to 1 ) 0.720212
```

```
Fixed effects:
```

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.61919    0.40098  -4.038 5.39e-05 ***
t              0.02881    0.02272   1.268  0.205
smokerTRUE   -0.48874    0.36320  -1.346  0.178
genderMALE   -0.76755    0.57937  -1.325  0.185
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Correlation of Fixed Effects:
              (Intr) t      smTRUE
t              -0.228
smokerTRUE    -0.053 -0.339
genderMALE    -0.643 -0.001  0.001
```

```
R>m.glme.1$deviance
```

```
[1] 749.2843
```

```
R>m.glme.2$deviance
```

```
[1] 745.5649
```

```
R>deviance(m.glme.3)
```

```
[1] 749.2843
```

```
R>deviance(m.glme.4)
```

```
[1] 758.3629
```

⇒ adaptive Gauss-Quadratur mit 25 Stützstellen erreicht deutlich beste Anpassung, PQL am schlechtesten, Laplace in etwa gleich der adaptive Gauss-Quadratur mit 8 Stützstellen. Betrachte die relativen Veränderungen in $\hat{\beta}$:

```
R>round((fixef(m.glme.3) - coef(m.glme.1))/fixef(m.glme.3), 4)
```

```
(Intercept)          t  smokerTRUE  genderMALE
      0.0007      0.0007      0.0004      -0.0017
```

```
R>round((fixef(m.glme.3) - coef(m.glme.2))/coef(m.glme.2), 4)
```

```
(Intercept)          t  smokerTRUE  genderMALE
      0.0434     -0.0163     -0.0172      0.0603
```

```
R>round((fixef(m.glme.4) - coef(m.glme.2))/coef(m.glme.2), 4)
```

```
(Intercept)          t  smokerTRUE  genderMALE
     -0.1634      0.0018      0.0338     -0.2000
```

- (b) Geben Sie die entsprechende Modellformel für dieses Modell an und identifizieren Sie aus dem Output die Schätzer für alle relevanten Größen. Interpretieren Sie die Regressionskoeffizienten für `smoker` und `gender` auf der Odds-Ratio-Skala und geben Sie für diese Kovariablen auch 95%-Konfidenzintervalle an.

Lösung:

$$\begin{aligned} \text{logit}(\mu_{it}) &= \beta_0 + b_{0i} + \beta_1 t + \beta_2 I(\text{smoker}_{it} = \text{TRUE}) + \beta_3 I(\text{gender}_{it} = \text{MALE}), \\ Y_{it}^* | b_{0i} &\sim Be(\mu_{it}), \\ b_{0i} &\overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_{b_0}^2) \end{aligned}$$

Siehe `summary(m.glme.2)`, also $\hat{\beta} =$

```
R>coef(m.glme.2)
```

```
(Intercept)          t  smokerTRUE  genderMALE
-1.93549200  0.02875451 -0.47274794 -0.95946925
```

und $\sqrt{\sigma_{b_0}^2} = 2.5852$.

Schätzer und 95%-Konfidenzintervalle auf der Skala der Odds:

```
R>(smokerCI.glm <- exp(coef(m.glme.2)[3] + c(est = 0, lo = -1,
+      hi = 1) * qnorm(0.975) * m.glme.2[["coef.sd"]][3]))
```

```
      est      lo      hi
0.6232872 0.3030852 1.2817745
```

```
R>(maleCI.glm <- exp(coef(m.glme.2)[4] + c(est = 0, lo = -1, hi = 1) *
+      qnorm(0.975) * m.glme.2[["coef.sd"]][4]))
```

```
      est      lo      hi
0.3830962 0.1362261 1.0773465
```

Interpretation als multiplikative Veränderung der Odds (Chancen) über $\exp(\hat{\beta})$. Die festen Parameter lassen sich nur subjektspezifisch, d.h. gegeben den zufälligem Intercept, interpretieren. Für eine rauchende Person ist verändert sich das individuelle Risiko ($= \hat{\mu}_{it}/(1-\hat{\mu}_{it}) \neq \hat{P}(Y_{it}^* = 1)$) in der Risikogruppe zu sein um den Faktor $\exp(\hat{\beta}_2) = 0.6233$ gegenüber dem individuellen Risiko, falls diese Person Nichtraucher wäre.

Um $\hat{\beta}_3$ subjektspezifisch interpretieren zu können bedarf es folgenden Tricks: Ein Mann und eine Frau mit gleichen Kovariablen und gleichem Random Effekt werden verglichen. Somit ist die Chance in der Risikogruppe zu sein für einen solchen Mann das $\exp(\hat{\beta}_3) = 0.3831$ -fache des Risikos einer entsprechenden Frau. Da aber beide Konfidenzintervalle die 1 überlappen sind die beiden Effekte hier nicht signifikant zum 5%-Niveau.

- (c) Fitten Sie ein binäres GEE mit Intercept und Haupteffekten für `t`, `gender` und `smoker`. Benutzen Sie eine unstrukturierte Arbeitskorrelationsmatrix. Geben Sie die Regressionskoeffizienten und robuste 95%-Konfidenzintervalle für `smoker` und `gender` auf der Odds-Ratio-Skala an und interpretieren Sie diese.

Lösung:

```
R>library(gee)
R>m.gee <- gee(tc.dich ~ 1 + t + smoker + gender, id = id, data = ex,
+ family = binomial, corstr = "unstructured")
```

```
(Intercept)          t  smokerTRUE  genderMALE
-0.97248825  0.01218794 -0.10964710 -0.54330819
```

```
R>summary(m.gee)
```

```
GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gee S-function, version 4.13 modified 98/01/27 (1998)
```

Model:

```
Link:                               Logit
Variance to Mean Relation: Binomial
Correlation Structure:               Unstructured
```

Call:

```
gee(formula = tc.dich ~ 1 + t + smoker + gender, id = id, data = ex,
    family = binomial, corstr = "unstructured")
```

Summary of Residuals:

```
      Min      1Q      Median      3Q      Max
-0.3053489 -0.2755336 -0.1960555 -0.1575110  0.8452141
```

Coefficients:

```
              Estimate Naive S.E.   Naive z Robust S.E.   Robust z
(Intercept) -0.96672566  0.20136668 -4.8008224  0.20913005 -4.6226052
t            0.01034079  0.01403167  0.7369608  0.01401819  0.7376695
smokerTRUE  -0.25539258  0.20531225 -1.2439227  0.20283705 -1.2591022
genderMALE  -0.47542889  0.28639156 -1.6600660  0.28816640 -1.6498415
```

```
Estimated Scale Parameter:  1.001650
```

Number of Iterations: 4

Working Correlation

```
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000 0.5451807 0.4832505 0.4091587 0.4704457 0.3745552
[2,] 0.5451807 1.0000000 0.5677776 0.4403347 0.3637341 0.2944319
[3,] 0.4832505 0.5677776 1.0000000 0.6109509 0.4573108 0.3245128
[4,] 0.4091587 0.4403347 0.6109509 1.0000000 0.5326561 0.3842712
[5,] 0.4704457 0.3637341 0.4573108 0.5326561 1.0000000 0.3704741
[6,] 0.3745552 0.2944319 0.3245128 0.3842712 0.3704741 1.0000000
```

```
R>(smokerCI.gee <- exp(coef(m.gee)[3] + c(est = 0, lo = -1, hi = 1) *
+   qnorm(0.975) * sqrt(m.gee$robust.variance[3, 3])))
```

```
      est      lo      hi
0.7746123 0.5205102 1.1527619
```

```
R>(maleCI.gee <- exp(coef(m.gee)[4] + c(est = 0, lo = -1, hi = 1) *
+   qnorm(0.975) * sqrt(m.gee$robust.variance[4, 4])))
```

```
      est      lo      hi
0.6216184 0.3533753 1.0934818
```

Beim GEE werden die Effekte nicht subjektspezifisch sondern populationsspezifisch interpretiert. Also gilt die Aussage, dass sich die Chance in der Risikogruppe zu sein bei Männern gegenüber Frauen (*ceteris paribus*) um den Faktor $\exp(\hat{\beta}_3) = 0.6216$ verändert. Also direkt auf Populationsebene interpretierbar im Gegensatz zum LME.

- (d) Betrachten Sie den Unterschied zwischen der Interpretation der Odds-Ratios mit GEEs und GLMMs. In welcher Situation ist welche Methodik von Vorteil?

Lösung:

Falls Interesse in Prädiktion von Individuen sind GLMMs vorzuziehen. Ist man an populationsspezifischen Aussagen über Kovariableneffekten interessiert, sind GEEs der geeignete Ansatz. Bei der Identitätslinkfunktion ist die Interpretation in beiden Modellklassen gleich.