

Overview of the package LMMstar

Brice Ozenne

August 11, 2021

This vignette describes the main functionalities of the **LMMstar** package. This package implements specific types of linear mixed models mainly useful when having repeated observations over a discrete variable (e.g. time, brain region, ...). Key assumptions are that at the cluster level, observation are independent and identically distributed and that the mean and variance are driven by independent factors. In particular, in large samples the residuals do not have to be normally distributed.

The **LMMstar** package contains four main functions:

- the function `lmm` is the main function of the package which fits linear mixed models. The user can interact with *lmm* objects using:
 - `anova` to test combinations of coefficients (Wald test or Likelihood ratio tests)
 - `coef` to extract the estimates.
 - `confint` to extract estimates, confidence intervals, and p.values.
 - `getVarCov` to extract the modeled residual variance covariance matrix.
 - `logLik` to output the log-likelihood of the estimated model.
 - `predict` to compute the conditional mean for new observations.
 - `residuals` to extract the observed residuals of the fitted model.
 - `summary` to obtain a summary of the results
- the `summarize` function to compute summary statistics stratified on a categorical variable (typically time).
- the `sampleRem` function to simulate longitudinal data.
- the `LMMstar.options` function enables the user to display the default values used in the **LMMstar** package. function. The function can also change the default values to better match the user needs.

Before going further we need to load the **LMMstar** package in the R session:

```
library(LMMstar)
```

To illustrate the functionalities of the package, we will use the **veteran** dataset:

```
data(gastricbypassL, package = "LMMstar")
head(gastricbypassL)
```

	id	visit	time	weight	glucagon
1	1	1	3 months before surgery	127.2	5032.50
2	2	1	3 months before surgery	165.2	12142.50
3	3	1	3 months before surgery	109.7	10321.35
4	4	1	3 months before surgery	146.2	6693.00
5	5	1	3 months before surgery	113.1	7090.50
6	6	1	3 months before surgery	158.8	10386.00

See `?gastricbypassL` for a presentation of the database. We will use a shorter version of the time variable:

```
gastricbypassL$time <- factor(gastricbypassL$time,
  levels = c("3 months before surgery", "1 week before surgery",
    "1 week after surgery", "3 months after surgery" ),
  labels = c("B3_months", "B1_week", "A1_week", "A3_months"))
```

and rescale the glucagon values

```
gastricbypassL$glucagon <- as.double(scale(gastricbypassL$glucagon))
```

Note: the **LMMstar** package is under active development. Newer package versions may include additional functionalities and fix previous bugs. The version of the package that is being is:

```
utils::packageVersion("LMMstar")
```

```
[1] '0.2.4'
```

1 Descriptive statistics

Mean, standard deviation, and other summary statistic can be computed with respect to a categorical variable (typically time) using the `summarize` function:

```
sss <- summarize(weight+glucagon ~ time, data = gastricbypassL, na.rm = TRUE)
print(sss, digits = 3)
```

	outcome	time	observed	missing	mean	sd	min	median	max
1	weight	B3_months	20	0	128.9700	20.269	100.900	123.1000	173.000
2	weight	B1_week	20	0	121.2400	18.910	95.700	114.5000	162.200
3	weight	A1_week	20	0	115.7000	18.275	89.900	110.6000	155.000
4	weight	A3_months	20	0	102.3650	17.054	78.800	98.5000	148.000
5	glucagon	B3_months	20	0	-0.4856	0.641	-1.395	-0.6679	1.030
6	glucagon	B1_week	19	1	-0.6064	0.558	-1.416	-0.7669	0.946
7	glucagon	A1_week	19	1	1.0569	1.044	-0.478	0.9408	3.267
8	glucagon	A3_months	20	0	0.0576	0.760	-1.047	0.0319	2.124

2 Linear mixed model

2.1 Modeling tools

Fit a linear mixed model with **compound symmetry** structure:

```
eCS.lmm <- lmm(weight ~ time + glucagon,
               structure = CS(~time|id),
               data = gastricbypassL)
eCS.lmm
```

Linear Mixed Model with a compound symmetry covariance matrix

```
data          : 78 observations and distributed in 20 clusters
log-likelihood : -243.6005
parameters    : 5 mean ((Intercept) timeB1_week timeA1_week timeA3_months glucagon)
                1 variance (sigma)
                1 correlation (Rho)
```

Fit a linear mixed model with **unstructured** covariance matrix:

```
eUN.lmm <- lmm(weight ~ time + glucagon,
               structure = UN(~time|id),
               data = gastricbypassL)
eUN.lmm
```

Linear Mixed Model with an unstructured covariance matrix

```
data          : 78 observations and distributed in 20 clusters
log-likelihood : -216.3189
parameters    : 5 mean ((Intercept) timeB1_week timeA1_week timeA3_months glucagon)
                4 variance (sigma k.B1_week k.A1_week k.A3_months)
                6 correlation (cor(B1_week,B3_months) cor(A1_week,B3_months) cor(A3_months,B3_months))
```

Note: the calculation of the degrees of freedom, especially when using the observed information can be quite slow. Setting the arguments `df` to `FALSE` and `type.information` to `"expected"` when calling `lmm` should lead to a more reasonable computation time.

2.2 Model output

The `summary` method can be used to display the main information relative to the model fit:

```
summary(eCS.lmm, ci = TRUE)
```

Linear Mixed Model

Dataset: `gastricbypassL`

- 20 clusters
- 78 observations were analyzed, 2 were excluded because of missing values
- between 3 and 4 observations per cluster

Summary of the outcome and covariates:

```
$ weight : num  127 165 110 146 113 ...
$ time    : Factor w/ 4 levels "B3_months","B1_week",...: 1 1 1 1 1 1 1 1 1 1 ...
$ glucagon: num  -0.9654 0.2408 -0.0682 -0.6837 -0.6163 ...
reference level: time=B3_months
```

Estimation procedure

- Restricted Maximum Likelihood (REML)
- log-likelihood : -243.6005
- parameters: mean = 5, variance = 1, correlation = 1

Residual variance-covariance: compound symmetry

- correlation structure: `~1 | id`

	B3_months	B1_week	A1_week	A3_months
B3_months	1.00	0.97	0.97	0.97
B1_week	0.97	1.00	0.97	0.97
A1_week	0.97	0.97	1.00	0.97
A3_months	0.97	0.97	0.97	1.00
- variance structure: `~1`

	standard.deviation
sigma	18.84957

Fixed effects: `weight ~ time + glucagon`

	estimate	se	df	lower	upper	p.value
(Intercept)	129.369	4.226	20.034	120.556	120.556	<0.001 ***
timeB1_week	-7.619	1.054	53.968	-9.732	-9.732	<0.001 ***
timeA1_week	-14.495	1.428	53.879	-17.358	-17.358	<0.001 ***
timeA3_months	-27.051	1.087	53.943	-29.231	-29.231	<0.001 ***
glucagon	0.822	0.62	53.81	-0.421	-0.421	0.191

The columns lower and upper correspond to the 95% confidence interval
 Degrees of freedom were computed using a Satterthwaite approximation

2.3 Extract estimated coefficients

The value of the estimated coefficients can be output using `coef`:

```
coef(eCS.lmm)
```

(Intercept)	timeB1_week	timeA1_week	timeA3_months	glucagon	sigma	Rho
129.3690995	-7.6194918	-14.4951323	-27.0514694	0.8217879	18.8495684	0.9699341

It is possible to apply specific transformation on the variance coefficients, for instance to obtain the residual variance relative to each outcome:

```
coef(eUN.lmm, effects = "variance", transform.k = "sd")
```

sigma:B3_months	sigma:B1_week	sigma:A1_week	sigma:A3_months
20.28080	19.04553	17.65479	16.76104

2.4 Extract estimated residual variance-covariance structure

The method `getVarCov` can be used to output the covariance structure of the residuals:

```
nlme::getVarCov(eCS.lmm)
```

	B3_months	B1_week	A1_week	A3_months
B3_months	355.3062	344.6236	344.6236	344.6236
B1_week	344.6236	355.3062	344.6236	344.6236
A1_week	344.6236	344.6236	355.3062	344.6236
A3_months	344.6236	344.6236	344.6236	355.3062

It can also be specific to an individual:

```
nlme::getVarCov(eCS.lmm, individual = 5)
```

	B3_months	A1_week	A3_months
B3_months	355.3062	344.6236	344.6236
A1_week	344.6236	355.3062	344.6236
A3_months	344.6236	344.6236	355.3062

2.5 Model diagnostic

The method `residuals` can also be used to extract the residuals in the wide format:

```
eCS.diagW <- residuals(eCS.lmm, type = "normalized", format = "wide")
head(eCS.diagW)
```

	cluster	B3_months	B1_week	A1_week	A3_months
1	1	-0.8042448	-0.709908611	-1.4242831	0.3176640
2	2	1.0863177	-0.133256804	1.1083627	1.5977042
3	3	-0.4597852	-0.612727870	-0.6060136	-0.8589524
4	4	-1.0103075	0.007471088	0.1309862	1.1428822
5	5	-0.1258773	NA	-0.3819184	-0.7874832
6	6	3.5646225	2.333205076	2.8387204	0.3586263

or in the long format:

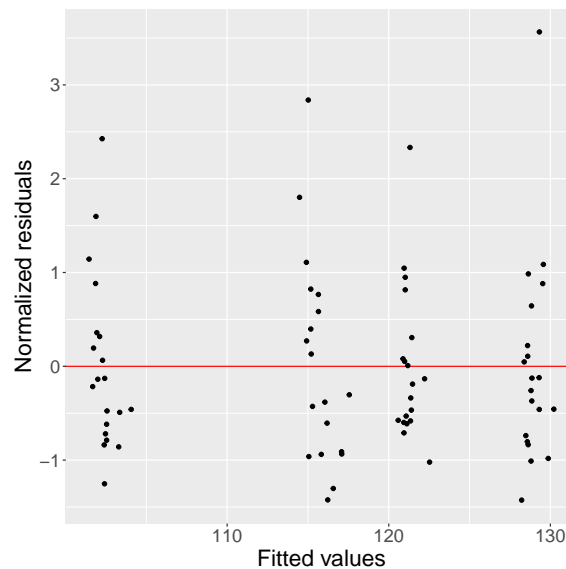
```
eCS.diagL <- residuals(eCS.lmm, type = "normalized", format = "long")
head(eCS.diagL)
```

```
[1] -0.8042448  1.0863177 -0.4597852 -1.0103075 -0.1258773  3.5646225
```

Various type of residuals can be extract but the normalized one are recommended when doing model checking. The method `residuals` can also be used to display diagnostic plots, e.g. about:

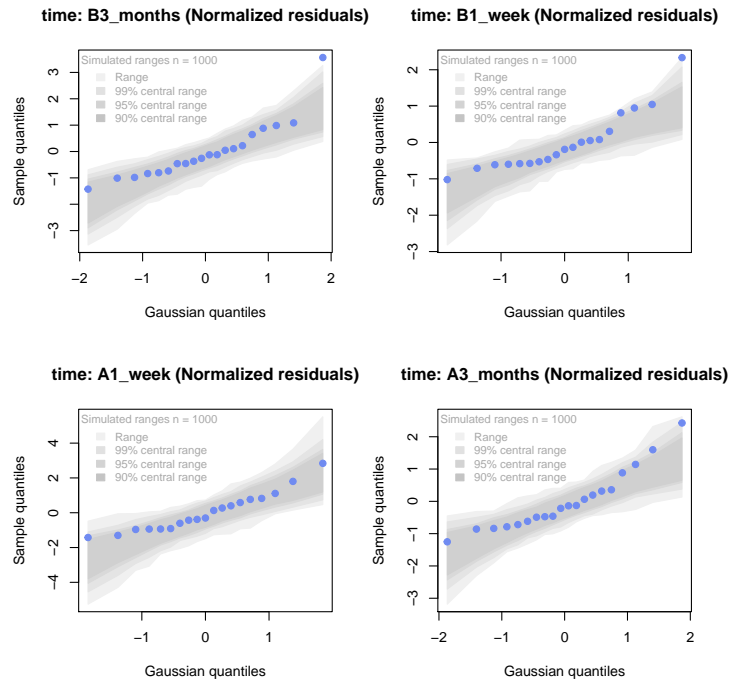
- the distribution of the residuals across fitted values using a scatterplot

```
residuals(eCS.lmm, type = "normalized", plot = "scatterplot", size.text = 20)
```



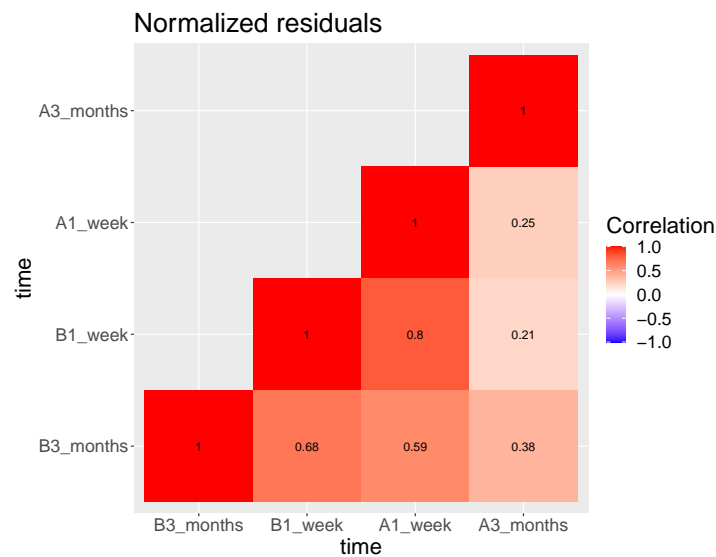
- the "normality" of the residuals at each repetition using a quantile-quantile plot ¹:

```
residuals(eCS.lmm, type = "normalized", format = "wide",
  plot = "qqplot", engine.qqplot = "qqtest")
## Note: the qqtest package to be installed to use the argument engine.plot = "qqtest"
```



- the residual correlation within cluster between the residuals:

```
residuals(eCS.lmm, type = "normalized", plot = "correlation", format = "wide",
  size.text = 20)
```



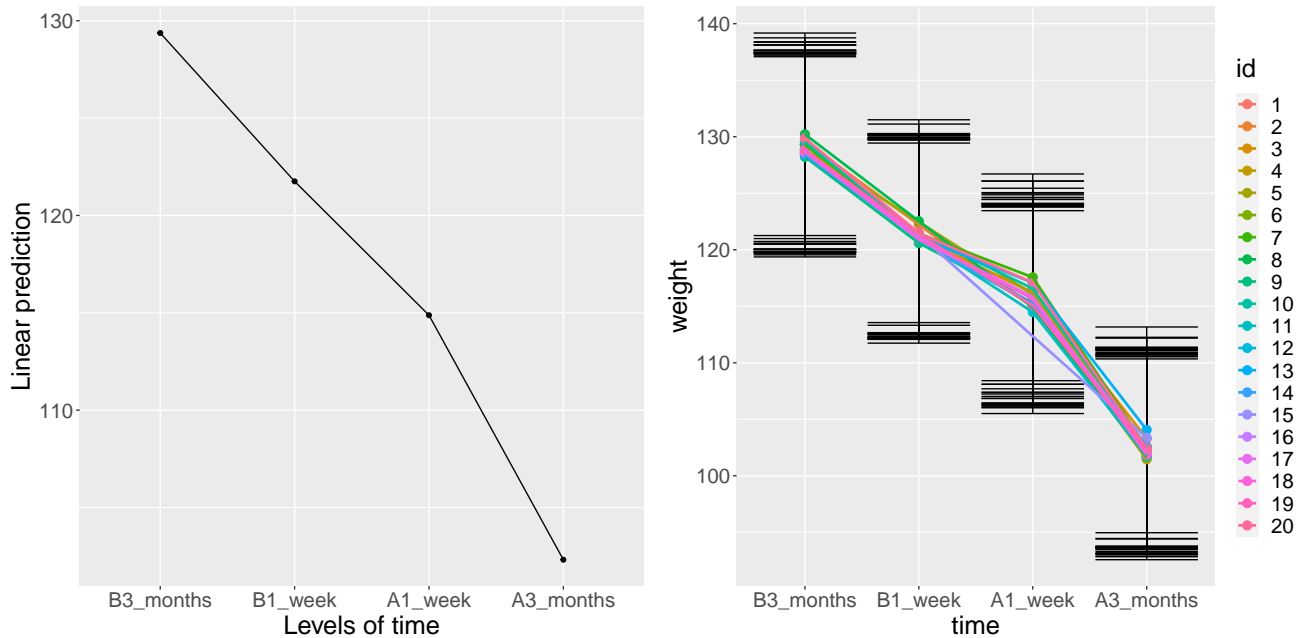
¹see Oldford (2016) for guidance about how to read quantile-quantile plots.

2.6 Model fit

The fitted values can be displayed via the `emmeans` package or using the `autoplot` method:

```
library(emmeans) ## left panel
emmip(eCS.lmm, ~time) + theme(text = element_text(size=20))
```

```
library(ggplot2) ## right panel
autoplot(eCS.lmm, color = "id", size.text = 20)
```



In the first case the average curve (over glucagon values) is displayed while in the latter each possible curve is displayed. With the `autoplot` method, it is possible to display a curve specific to a glucagon value via the argument `at`:

```
autoplot(eCS.lmm, at = data.frame(glucagon = 10), color = "glucagon")
```

2.7 Statistical inference

2.7.1 Model coefficients

The estimated coefficients with their confidence intervals can be accessed via the `confint` method:

```
confint(eCS.lmm)
```

	estimate	lower	upper
(Intercept)	129.369	120.556	138.18
timeB1_week	-7.619	-9.732	-5.51
timeA1_week	-14.495	-17.358	-11.63
timeA3_months	-27.051	-29.231	-24.87
glucagon	0.822	-0.421	2.06

Confidence intervals for the variance and correlation parameters can be displayed too specifying `effect"all"=`:

```
confint(eCS.lmm, effect = "all", backtransform = TRUE,
        columns = c("estimate", "se", "lower", "upper"))
```

	estimate	se	lower	upper
(Intercept)	129.369	4.226	120.556	138.183
timeB1_week	-7.619	1.054	-9.732	-5.507
timeA1_week	-14.495	1.428	-17.358	-11.632
timeA3_months	-27.051	1.087	-29.231	-24.872
glucagon	0.822	0.620	-0.421	2.065
sigma	18.850	0.159	13.479	26.359
Rho	0.970	0.187	0.936	0.986

Note: estimates and confidence intervals for sigma, rho have been back-transformed.
standard errors are not back-transformed.

Because these parameters are constrained (e.g. strictly positive), they uncertainty is by default computed after transformation (e.g. log) and then backtransformed.

2.7.2 Linear combination of the model coefficients

The `anova` method can be use to test one or several linear combinations of the model coefficients using Wald tests. For instance whether there is a change in average weight just after taking the treatment:

```
anova(eUN.lmm, effects = c("timeA1_week-timeB1_week=0"), ci = TRUE)
```

```

** User-specified hypotheses **

- F-test
statistic df.num df.denom      p.value
  43.14145      1 17.87461 3.723244e-06

- P-values and confidence interval (adjusted for multiplicity within each global test)
              estimate      lower      upper
timeA1_week - timeB1_week -3.905721 -5.155641 -2.655801
```

When testing transformed variance or correlation parameters, parentheses (as in `log(k).B1_week`) cause problem for recognizing parameters:

```
try(
  anova(eUN.lmm,
    effects = c("log(k).B1_week=0", "log(k).A1_week=0", "log(k).A3_months=0"))
)
```

```
Error in .anova_Wald(object, effects = effects, rhs = rhs, df = df, ci = ci, :
```

```
Possible misspecification of the argument 'effects' as running mulcomp::glht lead to the following
```

```
Error in parse(text = ex[i]) : <text>:1:7: uventet symbol
```

```
1: log(k).B1_week
```

```
^
```

It is then advised to specify the null hypothesis via a contrast matrix, e.g.:

```
name.coef <- rownames(confint(eUN.lmm, effects = "all", backtransform = FALSE))
name.varcoef <- grep("log(k)", name.coef, value = TRUE, fixed = TRUE)
C <- matrix(0, nrow = 3, ncol = length(name.coef), dimnames = list(name.varcoef, name.coef))
diag(C[name.varcoef, name.varcoef]) <- 1

anova(eUN.lmm, effects = C)
```

**** User-specified hypotheses ****

- F-test

statistic	df.num	df.denom	p.value
6.203176	3	17.99457	0.004417067

2.8 Baseline adjustment

The `lmm` contains an "experimental" feature to drop non-identifiable effects from the model. For instance, let us define two (artificial) groups of patients:

```
gastricbypassL$group <- c("1","2")[as.numeric(gastricbypassL$id) %in% 15:20 + 1]
```

We would like to model group differences only after baseline (i.e. only at 1 week and 3 months after). For this we will define a treatment variable being the group variable except before baseline where it is "none":

```
gastricbypassL$treat <- baselineAdjustment(gastricbypassL, variable = "group",
  repetition = ~time|id, constrain = c("B3_months","B1_week"),
  new.level = "none")
table(treat = gastricbypassL$treat, time = gastricbypassL$time, group = gastricbypassL$group)
```

```
, , group = 1
```

	time			
treat	B3_months	B1_week	A1_week	A3_months
none	14	14	0	0
1	0	0	14	14
2	0	0	0	0

```
, , group = 2
```

	time			
treat	B3_months	B1_week	A1_week	A3_months
none	6	6	0	0
1	0	0	0	0
2	0	0	6	6

Here we will be able to estimate a total of 6 means and therefore can at most identify 6 effects. However the design matrix for the interaction model:

```
colnames(model.matrix(weight ~ treat*time, data = gastricbypassL))
```

```
[1] "(Intercept)"      "treat1"           "treat2"           "timeB1_week"
[5] "timeA1_week"      "timeA3_months"    "treat1:timeB1_week" "treat2:timeB1_week"
[9] "treat1:timeA1_week" "treat2:timeA1_week" "treat1:timeA3_months" "treat2:timeA3_months"
```

contains 12 parameters (i.e. 6 too many). The `lmm` function will internally remove the one that cannot be identified and fit a simplified model:

```
eC.lmm <- lmm(weight ~ treat*time, data = gastricbypassL, structure = UN(~time|id))
```

Advarselsbesked:

```
I model.matrix_regularize(formula.mean, data) :
```

```
Constant values in the design matrix in interactions "treat:time"
```

```
Coefficients "treat1" "treat2" "timeA1_week" "timeA3_months" "treat1:timeB1_week" "treat2:timeB1_w
```

Consider defining manually the interaction, e.g. via `droplevels(interaction(.,.))` to avoid this war

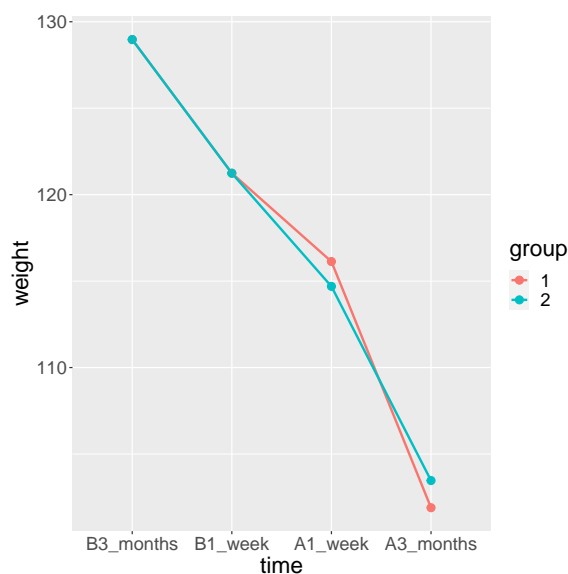
with the following coefficients:

```
coef(eC.lmm, effects = "mean")
```

(Intercept)	timeB1_week	treat1:timeA1_week	treat2:timeA1_week
128.97000	-7.73000	-12.83949	-14.27452
treat1:timeA3_months	treat2:timeA3_months		
-27.07620	-25.50553		

One can visualize the baseline adjustment via the autoplot function:

```
autoplot(eC.lmm, color = "group", ci = FALSE, size.text = 20)
```



To more easily compare the two groups, one could set the baseline treatment to the treatment in the control arm by omitting the argument `new.level`:

```
gastricbypassL$treat2 <- baselineAdjustment(gastricbypassL, variable = "group",
      repetition = ~time|id, constrain = c("B3_months", "B1_week"))
table(treat = gastricbypassL$treat2, time = gastricbypassL$time, group = gastricbypassL$group)
```

```
, , group = 1
```

	time				
treat	B3_months	B1_week	A1_week	A3_months	
1	14	14	14	14	
2	0	0	0	0	

```
, , group = 2
```

	time				
treat	B3_months	B1_week	A1_week	A3_months	
1	6	6	0	0	
2	0	0	6	6	

Fitting the model

```
eC2.lmm <- suppressWarnings(lmm(weight ~ treat2*time, data = gastricbypassL, structure = UN(~time|id)))
```

will directly output group differences:

```
confint(eC2.lmm, effects = "mean", columns = c("estimate", "lower", "upper", "p.value"))[5:6,]
```

	estimate	lower	upper	p.value
treat22:timeA1_week	-1.44	-2.75	-0.12	0.0343
treat22:timeA3_months	1.57	-3.64	6.78	0.5325

2.9 Marginal means

The `lmm` function can be used in conjunction with the `emmeans` package to compute marginal means. Consider the following model:

```
e.group <- lmm(weight ~ time*group, data = gastricbypassL, structure = UN(~time|id))
```

We can for instance compute the average value over time *assuming balanced groups*:

```
library(emmeans)
emmeans(e.group, specs=~time)
```

NOTE: Results may be misleading due to involvement in interactions

time	emmean	SE	df	lower.CL	upper.CL
B3_months	130	5.05	18.0	119.3	141
B1_week	122	4.69	18.0	112.5	132
A1_week	117	4.55	18.0	107.0	126
A3_months	104	4.20	18.1	94.9	113

Results are averaged over the levels of: group

Confidence level used: 0.95

This differs from the average value over time over the whole sample:

```
df.pred <- cbind(gastricbypassL, predict(e.group, newdata = gastricbypassL))
summarize(formula = estimate~time, data = df.pred)
```

	outcome	time	observed	missing	mean	sd	min	median	max
1	estimate	B3_months	20	0	128.970	2.270212	127.5214	127.5214	132.35
2	estimate	B1_week	20	0	121.240	2.726942	119.5000	119.5000	125.30
3	estimate	A1_week	20	0	115.700	2.014981	114.4143	114.4143	118.70
4	estimate	A3_months	20	0	102.365	3.146729	100.3571	100.3571	107.05

as the groups are not balanced and with this approach more "weight" is given to the expected value group 1 as it contains more individuals.

```
table(group = gastricbypassL$group, time = gastricbypassL$time)
```

```

time
group B3_months B1_week A1_week A3_months
1      14      14      14      14
2       6       6       6       6

```

By hand:

```

mu.group1 <- as.double(coef(e.group)["(Intercept)"])
mu.group2 <- as.double(coef(e.group)["(Intercept)"] + coef(e.group)["group2"])
p.group1 <- 14/20
p.group2 <- 6/20
c(emmeans = (mu.group1+mu.group2)/2,
  predict = mu.group1 * p.group1 + mu.group2 * p.group2)

```

```

emmeans predict
129.9357 128.9700

```

which one is relevant depends on the application. The `emmeans` function can also be used to display expected value in each group over time:

```

emmeans.group <- emmeans(e.group, specs = ~group|time)
emmeans.group

```

```

time = B3_months:
group emmean SE df lower.CL upper.CL
1      128 5.53 18.0    115.9    139
2      132 8.45 18.0    114.6    150

```

```

time = B1_week:
group emmean SE df lower.CL upper.CL
1      120 5.14 18.0    108.7    130
2      125 7.85 18.0    108.8    142

```

```

time = A1_week:
group emmean SE df lower.CL upper.CL
1      114 4.99 18.0    103.9    125
2      119 7.62 18.0    102.7    135

```

```

time = A3_months:
group emmean SE df lower.CL upper.CL
1      100 4.60 18.1     90.7    110
2      107 7.03 18.1     92.3    122

```

Confidence level used: 0.95

Using the `pair` function displays the differences:

```
epairs.group <- pairs(emmeans.group, reverse = TRUE)
epairs.group
```

```
time = B3_months:
  contrast estimate    SE   df t.ratio p.value
2 - 1          4.83 10.10 18.0 0.478   0.6383
```

```
time = B1_week:
  contrast estimate    SE   df t.ratio p.value
2 - 1          5.80  9.38 18.0 0.618   0.5441
```

```
time = A1_week:
  contrast estimate    SE   df t.ratio p.value
2 - 1          4.29  9.11 18.0 0.471   0.6435
```

```
time = A3_months:
  contrast estimate    SE   df t.ratio p.value
2 - 1          6.69  8.40 18.1 0.797   0.4361
```

One can adjust for multiple comparison via the `adjust` argument and display confidence intervals setting the argument `infer` to `TRUE`:

```
summary(epairs.group, by = NULL, adjust = "mvt", infer = TRUE)
```

contrast	time	estimate	SE	df	lower.CL	upper.CL	t.ratio	p.value
2 - 1	B3_months	4.83	10.10	18.0	-18.0	27.6	0.478	0.7496
2 - 1	B1_week	5.80	9.38	18.0	-15.4	27.0	0.618	0.6482
2 - 1	A1_week	4.29	9.11	18.0	-16.3	24.9	0.471	0.7556
2 - 1	A3_months	6.69	8.40	18.1	-12.3	25.7	0.797	0.5287

Confidence level used: 0.95

Conf-level adjustment: mvt method for 4 estimates

P value adjustment: mvt method for 4 tests

This should also work when doing baseline adjustment (because of baseline adjustment no difference is expected at the first two timepoints):

```
summary(pairs(emmeans(eC2.lmm , specs = ~treat2|time), reverse = TRUE), by = NULL)
```

Note: `adjust = "tukey"` was changed to `"sidak"`

because `"tukey"` is only appropriate for one set of pairwise comparisons

contrast	time	estimate	SE	df	t.ratio	p.value
2 - 1	B3_months	0.00	0.000	NaN	NaN	NaN
2 - 1	B1_week	0.00	0.000	NaN	NaN	NaN
2 - 1	A1_week	-1.44	0.621	16.2	-2.311	0.1303
2 - 1	A3_months	1.57	2.463	16.3	0.638	0.9522

P value adjustment: sidak method for 4 tests

3 Data generation

Simulate some data in the wide format:

```
set.seed(10) ## ensure reproductibility
n.obs <- 100
n.times <- 4
mu <- rep(0,4)
gamma <- matrix(0, nrow = n.times, ncol = 10) ## add interaction
gamma[,6] <- c(0,1,1.5,1.5)
dW <- sampleRem(n.obs, n.times = n.times, mu = mu, gamma = gamma, format = "wide")
head(round(dW,3))
```

	id	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	Y1	Y2	Y3	Y4
1	1	1	0	1	1	0	-0.367	1.534	-1.894	1.729	0.959	1.791	2.429	3.958	2.991
2	2	1	0	1	2	0	-0.410	2.065	1.766	0.761	-0.563	2.500	4.272	3.002	2.019
3	3	0	0	2	1	0	-1.720	-0.178	2.357	1.966	1.215	-3.208	-5.908	-4.277	-5.154
4	4	0	0	0	1	0	0.923	-2.089	0.233	1.307	-0.906	-2.062	0.397	1.757	-1.380
5	5	0	0	2	1	0	0.987	5.880	0.385	0.028	0.820	7.963	7.870	7.388	8.609
6	6	0	0	1	1	2	-1.075	0.479	2.202	0.900	-0.739	0.109	-1.602	-1.496	-1.841

Simulate some data in the long format:

```
set.seed(10) ## ensure reproductibility
dL <- sampleRem(n.obs, n.times = n.times, mu = mu, gamma = gamma, format = "long")
head(dL)
```

	id	visit	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
1	1	1	1.791444	1	0	1	1	0	-0.3665251	1.533815	-1.894425	1.7288665	0.9592499
2	1	2	2.428570	1	0	1	1	0	-0.3665251	1.533815	-1.894425	1.7288665	0.9592499
3	1	3	3.958350	1	0	1	1	0	-0.3665251	1.533815	-1.894425	1.7288665	0.9592499
4	1	4	2.991198	1	0	1	1	0	-0.3665251	1.533815	-1.894425	1.7288665	0.9592499
5	2	1	2.500179	1	0	1	2	0	-0.4097541	2.065413	1.765841	0.7613348	-0.5630173
6	2	2	4.272357	1	0	1	2	0	-0.4097541	2.065413	1.765841	0.7613348	-0.5630173

4 Modifying default options

The `LMMstar.options` method enable to get and set the default options used by the package. For instance, the default option for the information matrix is:

```
LMMstar.options("type.information")
```

```
$type.information  
[1] "observed"
```

To change the default option to "expected" (faster to compute but less accurate p-values and confidence intervals in small samples) use:

```
LMMstar.options(type.information = "expected")
```

To restore the original default options do:

```
LMMstar.options(reinitialise = TRUE)
```

5 R session

Details of the R session used to generate this document:

```
sessionInfo()
```

```
R version 4.0.3 (2020-10-10)
```

```
Platform: x86_64-w64-mingw32/x64 (64-bit)
```

```
Running under: Windows 10 x64 (build 19042)
```

```
Matrix products: default
```

```
locale:
```

```
[1] LC_COLLATE=Danish_Denmark.1252 LC_CTYPE=Danish_Denmark.1252 LC_MONETARY=Danish_Denmark.1252
```

```
[4] LC_NUMERIC=C LC_TIME=Danish_Denmark.1252
```

```
attached base packages:
```

```
[1] stats      graphics  grDevices  utils      datasets  methods    base
```

```
other attached packages:
```

```
[1] ggplot2_3.3.5 emmeans_1.6.1 LMMstar_0.2.4 nlme_3.1-149
```

```
loaded via a namespace (and not attached):
```

```
[1] Rcpp_1.0.7          plyr_1.8.6          compiler_4.0.3      pillar_1.6.2
[5] tools_4.0.3         digest_0.6.27       lifecycle_1.0.0     tibble_3.1.3
[9] gtable_0.3.0        lattice_0.20-41     pkgconfig_2.0.3     rlang_0.4.11
[13] Matrix_1.2-18       mvtnorm_1.1-1       coda_0.19-4         withr_2.4.2
[17] stringr_1.4.0       dplyr_1.0.5         generics_0.1.0      vctrs_0.3.8
[21] qqtest_1.2.0        grid_4.0.3          tidyselect_1.1.0    glue_1.4.2
[25] R6_2.5.0            fansi_0.5.0         survival_3.2-10     multcomp_1.4-16
[29] lava_1.6.9          TH.data_1.0-10      farver_2.1.0        reshape2_1.4.4
[33] purrr_0.3.4         magrittr_2.0.1      scales_1.1.1        codetools_0.2-16
[37] ellipsis_0.3.2      MASS_7.3-53         splines_4.0.3       xtable_1.8-4
[41] colorspace_2.0-2    numDeriv_2016.8-1.1 labeling_0.4.2       sandwich_3.0-0
[45] utf8_1.2.2          stringi_1.5.3       estimability_1.3     munsell_0.5.0
[49] crayon_1.4.1        zoo_1.8-9
```

References

Oldford, R. W. (2016). Self-calibrating quantile–quantile plots. *The American Statistician*, 70(1):74–90.

Appendix A Likelihood in a linear mixed model

A.1 Log-likelihood

Denote by \mathbf{Y} a vector of m outcomes, \mathbf{X} a vector of p covariates, $\mu(\boldsymbol{\Theta}, \mathbf{X})$ the modeled mean, and $\Omega(\boldsymbol{\Theta}, \mathbf{X})$ the modeled residual variance-covariance. The restricted log-likelihood in a linear mixed model can then be written:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{X}) = & \frac{p}{2} \log(2\pi) - \frac{1}{2} \log \left(\left| \sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \mathbf{X}_i^\top \right| \right) \\ & + \sum_{i=1}^n \left(-\frac{m}{2} \log(2\pi) - \frac{1}{2} \log |\Omega_i(\boldsymbol{\Theta})| - \frac{1}{2} (\mathbf{Y}_i - \mu(\boldsymbol{\Theta}, \mathbf{X}_i)) \Omega_i(\boldsymbol{\Theta})^{-1} (\mathbf{Y}_i - \mu(\boldsymbol{\Theta}, \mathbf{X}_i))^\top \right) \quad (\text{A}) \end{aligned}$$

This is what the `logLik` method is computing for the REML criteria. The red term is specific to the REML criteria and prevents from computing individual contributions to the likelihood². The blue term is what `logLik` outputs for the ML criteria when setting the argument `indiv` to `TRUE`.

A.2 Score

Using that $\partial \log(\det(X)) = \text{tr}(X^{-1} \partial(X))$, the score is obtained by derivating once the log-likelihood, i.e., for $\theta \in \boldsymbol{\Theta}$:

$$\begin{aligned} \mathcal{S}(\theta) = & \frac{\partial \mathcal{L}(\boldsymbol{\Theta}|\mathbf{Y}, \mathbf{X})}{\partial \theta} = \frac{1}{2} \text{tr} \left(\left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \mathbf{X}_i^\top \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \frac{\partial \Omega_i(\boldsymbol{\Theta})}{\partial \theta} \Omega_i(\boldsymbol{\Theta})^{-1} \mathbf{X}_i^\top \right) \right) \\ & + \sum_{i=1}^n \left(-\frac{1}{2} \text{tr} \left(\Omega_i(\boldsymbol{\Theta})^{-1} \frac{\partial \Omega_i(\boldsymbol{\Theta})}{\partial \theta} \right) + \frac{\partial \mu(\boldsymbol{\Theta}, \mathbf{X}_i)}{\partial \theta} \Omega_i(\boldsymbol{\Theta})^{-1} (\mathbf{Y}_i - \mu(\boldsymbol{\Theta}, \mathbf{X}_i))^\top \right. \\ & \quad \left. + \frac{1}{2} (\mathbf{Y}_i - \mu(\boldsymbol{\Theta}, \mathbf{X}_i)) \Omega_i(\boldsymbol{\Theta})^{-1} \frac{\partial \Omega_i(\boldsymbol{\Theta})}{\partial \theta} \Omega_i(\boldsymbol{\Theta})^{-1} (\mathbf{Y}_i - \mu(\boldsymbol{\Theta}, \mathbf{X}_i))^\top \right). \end{aligned}$$

This is what the `score` method is computing for the REML criteria. The red term is specific to the REML criteria and prevents from computing the score relative to each cluster. The blue term is what `score` outputs for the ML criteria when setting the argument `indiv` to `TRUE`.

²The REML is the likelihood of the observations divided by the prior on the estimated mean parameters $\hat{\boldsymbol{\Theta}}_\mu \sim \mathcal{N}(\mu, (\mathbf{X} \Omega^{-1}(\boldsymbol{\Theta}) \mathbf{X}^\top)^{-1})$. This corresponds to $\frac{1}{\sqrt{2\pi^p} |(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \mathbf{X}_i^\top)^{-1}|} \exp \left(-(\hat{\boldsymbol{\Theta}}_\mu - \mu) (2 \sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \mathbf{X}_i^\top)^{-1} (\hat{\boldsymbol{\Theta}}_\mu - \mu)^\top \right)$. Since μ will be estimated to be $\boldsymbol{\Theta}_\mu$, the exponential term equals 1 and thus does not contribute to the log-likelihood. One divided by the other term gives $\sqrt{2\pi^p} |(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\boldsymbol{\Theta}) \mathbf{X}_i^\top)|^{-1}$. The log of this term equals the red term

A.3 Hessian

Derivating a second time the log-likelihood gives the hessian, $\mathcal{H}(\Theta)$, with element³:

$$\begin{aligned}\mathcal{H}(\theta, \theta') &= \frac{\partial^2 \mathcal{L}(\Theta | \mathbf{Y}, \mathbf{X})}{\partial \theta \partial \theta'} = \frac{\partial \mathcal{S}(\theta)}{\partial \theta'} \\ &= \frac{1}{2} tr \left(\left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \mathbf{X}_i^\top \right)^{-1} \left\{ \sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \left(\frac{\partial^2 \Omega_i(\Theta)}{\partial \theta \partial \theta'} - 2 \frac{\partial \Omega_i(\Theta)}{\partial \theta} \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \right) \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right. \right. \\ &\quad \left. \left. + \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta} \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right) \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \mathbf{X}_i^\top \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right) \right\} \right) \\ &\quad + \sum_{i=1}^n \left(\frac{1}{2} tr \left(\Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta} - \Omega_i(\Theta)^{-1} \frac{\partial^2 \Omega_i(\Theta)}{\partial \theta \partial \theta'} \right) \right. \\ &\quad \left. - \frac{\partial \mu(\Theta, \mathbf{X}_i)}{\partial \theta} \Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \varepsilon_i(\Theta)^\top - \frac{\partial \mu(\Theta, \mathbf{X}_i)}{\partial \theta} \Omega_i(\Theta)^{-1} \frac{\partial \mu(\Theta, \mathbf{X}_i)}{\partial \theta'} \right. \\ &\quad \left. + \frac{1}{2} \varepsilon_i(\Theta) \Omega_i(\Theta)^{-1} \left(\frac{\partial^2 \Omega_i(\Theta)}{\partial \theta \partial \theta'} - \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta} - \frac{\partial \Omega_i(\Theta)}{\partial \theta} \Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \right) \Omega_i(\Theta)^{-1} \varepsilon_i(\Theta)^\top \right).\end{aligned}$$

where $\varepsilon_i(\Theta) = \mathbf{Y}_i - \mu(\Theta, \mathbf{X}_i)$.

The `information` method will (by default) return the (observed) information which is the opposite of the hessian. So multiplying the previous formula by -1 gives what `information` output for the REML criteria. The red term is specific to the REML criteria and prevents from computing the information relative to each cluster. The blue term is what `information` outputs for the ML criteria (up to a factor -1) when setting the argument `indiv` to `TRUE`.

A possible simplification is to use the expected hessian. Indeed for any deterministic matrix A :

- $\mathbb{E}[A(\mathbf{Y}_i - \mu(\Theta, \mathbf{X}_i))^\top | \mathbf{X}_i] = 0$
- $\mathbb{E}[(\mathbf{Y}_i - \mu(\Theta, \mathbf{X}_i))A(\mathbf{Y}_i - \mu(\Theta, \mathbf{X}_i))^\top | \mathbf{X}_i] = tr(A \mathbb{V}ar(\mathbf{Y}_i - \mu(\Theta, \mathbf{X}_i)))$

Leading to:

$$\begin{aligned}\mathbb{E}[\mathcal{H}(\theta, \theta') | \mathbf{X}] &= \frac{1}{2} tr \left(\left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \mathbf{X}_i^\top \right)^{-1} \left\{ \sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \left(\frac{\partial^2 \Omega_i(\Theta)}{\partial \theta \partial \theta'} - 2 \frac{\partial \Omega_i(\Theta)}{\partial \theta} \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \right) \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right. \right. \\ &\quad \left. \left. + \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta} \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right) \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \mathbf{X}_i^\top \right)^{-1} \left(\sum_{i=1}^n \mathbf{X}_i \Omega_i^{-1}(\Theta) \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \mathbf{X}_i^\top \right) \right\} \right) \\ &\quad + \sum_{i=1}^n \left(-\frac{1}{2} tr \left(\Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta'} \Omega_i(\Theta)^{-1} \frac{\partial \Omega_i(\Theta)}{\partial \theta} \right) - \frac{\partial \mu(\Theta, \mathbf{X}_i)}{\partial \theta} \Omega_i(\Theta)^{-1} \frac{\partial \mu(\Theta, \mathbf{X}_i)}{\partial \theta'} \right)\end{aligned}$$

This is what `information` output when the argument `type.information` is set to "expected" (up to a factor -1).

³if one is relative to the mean and the other to the variance then they are respectively θ and θ'

A.4 Degrees of freedom

Degrees of freedom are computed using a Satterthwaite approximation, i.e. for an estimate coefficient $\hat{\beta} \in \widehat{\Theta}$ with standard error $\sigma_{\widehat{\beta}}$, the degree of freedom is:

$$df(\sigma_{\widehat{\beta}}) = \frac{2\sigma_{\widehat{\beta}}^2}{\text{Var}[\sigma_{\widehat{\beta}}^2]}$$

Using a first order Taylor expansion we can approximate the variance term as:

$$\begin{aligned} \text{Var}[\sigma_{\widehat{\beta}}^2] &\approx \frac{\partial \sigma_{\widehat{\beta}}^2}{\partial \Theta} \Sigma_{\Theta} \frac{\partial \sigma_{\widehat{\beta}}^2}{\partial \Theta}^{\top} \\ &\approx c_{\beta} (\widehat{\mathcal{I}}_{\Theta})^{-1} \frac{\partial \widehat{\mathcal{I}}_{\Theta}}{\partial \Theta} (\widehat{\mathcal{I}}_{\Theta})^{-1} c_{\beta}^{\top} \Sigma_{\Theta} c_{\beta} (\widehat{\mathcal{I}}_{\Theta})^{-1} \frac{\partial \widehat{\mathcal{I}}_{\Theta}}{\partial \Theta}^{\top} (\widehat{\mathcal{I}}_{\Theta})^{-1} c_{\beta} \end{aligned}$$

where Σ_{Θ} is the variance-covariance matrix of all model coefficients, \mathcal{I}_{Θ} the information matrix for all model coefficients, c_{β} a matrix used to select the element relative to β in the first derivative of the information matrix, and $\frac{\partial}{\partial \Theta}$ denotes the vector of derivatives with respect to all model coefficients.

Appendix B Likelihood ratio test with the REML criterion

The blue term of Equation A in the log-likelihood is invariant to re-parameterisation while the red term is not. This means that a re-parametrisation of X into $\tilde{X} = BX$ with B invertible would not change the likelihood when using ML but would decrease the log-likelihood by $\log(|B|)$ when using REML.

Let's take an example:

```
## data(gastricbypassL, package = "LMMstar")
dfTest <- gastricbypassL
dfTest$glucagon2 <- dfTest$glucagon*2
```

where we multiply one column of the design matrix by 2. As mentioned previously this does not affect the log-likelihood when using ML:

```
logLik(lmm(weight ~ glucagon, data = dfTest, structure = UN(~time|id), method = "ML"))
logLik(lmm(weight ~ glucagon2, data = dfTest, structure = UN(~time|id), method = "ML"))
```

```
[1] -245.7909
```

```
[1] -245.7909
```

but it does when using REML:

```
logLik(lmm(weight ~ glucagon, data = dfTest, structure = UN(~time|id), method = "REML"))
logLik(lmm(weight ~ glucagon2, data = dfTest, structure = UN(~time|id), method = "REML"))
log(2)
```

```
[1] -245.0382
```

```
[1] -245.7313
```

```
[1] 0.6931472
```

Therefore, when comparing models with different mean effects there is a risk that the difference (or part of it) in log-likelihood is due to a new parametrisation and not only to a difference in model fit. This would typically be the case when adding an interaction where we can have a smaller restricted log-likelihood when considering a more complex model:

```
set.seed(10)
dfTest$ff <- rbinom(NROW(dfTest), size = 1, prob = 0.5)
logLik(lmm(weight ~ glucagon, data = dfTest, structure = UN(~time|id), method = "REML"))
logLik(lmm(weight ~ glucagon*ff, data = dfTest, structure = UN(~time|id), method = "REML"))
```

```
[1] -245.0382
```

```
[1] -239.2056
```

This is quite counter-intuitive as more complex model should lead to better fit and would never happen when using ML:

```
logLik(lmm(weight ~ glucagon, data = dfTest, structure = UN(~time|id), method = "ML"))
logLik(lmm(weight ~ glucagon*ff, data = dfTest, structure = UN(~time|id), method = "ML"))
```


[1] -245.7909

[1] -237.3642

This is why, unless one knows what he/she is doing, it is not recommended to use likelihood ratio test to assess relevance of mean parameters in mixed models estimated with REML.