Examples of Tukey's trend test in general parametric models

Ludwig A. Hothorn
Leibniz University Hannover,
Christian Ritz
Copenhagen University,
Frank Schaarschmidt
Leibniz University Hannover

November 11, 2024

Contents

1 Introduction **1**

2 The package tukeytrend **2**

3 Case studies **3**
3.1 One primary endpoint assuming normal distributed residuals . . . 3
3.1.1 Small sample size modification . . . 4
3.1.2 Adjustment against multiple covariates . . . 4
3.1.3 Consideration of different covariance-adjusting models . . . 4
3.1.4 Allowing variance heterogeneity . . . 5
3.1.5 Finding an appropriate log-transformation for the zero dose control . . . 6
3.2 Considering dose as a quantitative covariate and a qualitative factor jointly . . . 6
3.2.1 Inclusion of the control vs. high dose comparison . . . 6
3.2.2 Tukey trend test and Williams contrasts . . . 6
3.2.3 Tukey and Dunnett test . . . 7
3.2.4 Taking specific non-monotonic dose-response relationships into account . . . 7
3.2.5 Testing vs. pooled doses . . . 7
3.2.6 Designs without a control group . . . 8
3.3 Multiple endpoints . . . 8
3.3.1 Multiple normal endpoints . . . 8
3.3.2 Multiple normal endpoints with missing data . . . 9
3.3.3 Multiple binary endpoints . . . 9
3.3.4 Multiple, differently scaled endpoints . . . 10
3.4 Using generalized linear models . . . 13
3.4.1 Proportions . . . 13
3.4.2 Multinomial endpoint . . . 13
3.4.3 Overdispersed counts . . . 14
3.4.4 Ordered categorical data . . . 14
3.5 Regression models treating dose as continuous data . . . 15
3.6 Mixed effect model . . . 15
3.6.1 Block design . . . 15
3.6.2 Multi-centre study . . . 16
3.6.3 Technical replicates . . . 16
3.6.4 Cross-over design . . . 17

1 Introduction

Already three decades ago, Tukey and co-workers published a trend test which represents a maximum test on three regression models for the arithmetic, ordinal, and logarithmic-linear dose metameters for the single covariate *dose* in a randomized one-way layout (Tukey et al., 1985). Trend tests in general can be classified into: i) tests which use *dose* either as a qualitative factor or a quantitative covariate, ii) tests which are constructed for a particular shape of the dose-response relation (such as the linear shape, e.g. assumed by Jonckheere (1954), non-parametric test (*factor*), or Cochran-Armitage test on proportions Armitage (1955) (*covariate*)), and iii) tests considering an a-priori defined set of shape-specific alternatives (such as multiple contrast tests (Hothorn, 2006) (*factor*) or multiple models Tukey

et al. (1985) (*covariate*) or a hybrid approach of both (MCPMod (Bornkamp et al., 2009). In the latter, the basic principle a maximum test over either multiple contrasts (of a dose *factor*s levels) or multiple models (Tukey et al., 1985) (*covariate*).

Given the large number of different available methods, the question arises why Tukey's test trend deserves attention today? First, it uses the three common basic dose-response models of biomedical research as an argument of simplicity and interpretability - in contrast to various particular nonlinear models.
Second, it is widely used up to now (365 citations in WebSci), e.g. for dose-response analysis in a placebo-controlled dose-ranging clinical trial on type 2 diabetes Sykes et al. (2015), in toxicogenomics Yager et al. (2013), in developmental toxicology Wise et al. (2011), and in a psychiatric genetic association study Bertelsen et al. (2014), among others.
Third, it is a powerful test Aras et al. (2011) whereas an uniformly powerful test for any alternative can not exists.
In this paper the CRAN package `tukeytrend` is described which includes the Tukey trend test and various modifications using the concept of multiple marginal models (Pipper et al., 2012), particularly for:

1. generalized linear models (Pipper et al., 2012), in analogy to the recent generalization of package `MCPMod` to general parametric models (Pinheiro et al., 2014),
2. linear mixed-effects model according to Ritz et al. (2017),
3. different interpolations between the zero-dose control and the lowest dose in the logarithmic-linear transformation,
4. the inclusion of the "high dose vs. control comparison" into the maximum test instead as an alternative (Aras et al., 2011),
5. for possible downturn effects at high(er) dose(s) according to downturn-protected trend tests (Bretz and Hothorn, 2003),
6. for models with additional covariates and/or secondary factors, i.e. the separation into approaches for *factor* or *covariate* can be overcome,
7. for multiple endpoints (where multiple endpoints were proposed in the original procedure (Tukey et al., 1985) with empirical choice of Sidak inequality),
8. providing both a global decision (trend/no trend) and local decisions by means of either adjusted p-values or simultaneous confidence intervals

The joint distribution of parameter estimates from multiple generalized linear (mixed) models were recently proposed (Pipper et al., 2012). This allows maximum tests on multiple linear models (instead of just multiple contrast tests) and the estimation of adjusted p-values and simultaneous confidence intervals without the explicit formulation of the correlation matrix in cases where it is difficult or even impossible. The variance-covariance matrix of parameter estimates can be obtained using derivatives of the log-likelihood function in a single model, and hence for multiple models a vector of log-likelihood functions is used. This vector-valued asymptotic representation based on standardized score vectors as sum of i.i.d. normally distributed random variables. By plugging in parameter estimates, a consistent sandwich estimator of the variance-covariance matrix is obtained, i.e. the empirical covariance based on functions of the data, not on the data itself. The functions of the `tukeytrend` package make use of the function `mmm` which is available with the R package `multcomp` (Hothorn et al., 2008), e.g.

```
library("multcomp")
mmmTukey <- glht(mmm(arithmetic=zetaA, ordinal=zetaO, linlog=zetaLL),
            mlf(aritmetic="DoseA=0", ordinal="DoseO=0", linlog="DoseLL=0"))
```

where `mlf(ordinal="DoseO=0")` means test on zero slope parameter for the ordinal dose metameter covariate.

2 The package tukeytrend

`tukeytrend` provides wrapper functions, which refit a given initial model with transformed dose variable, using the `update` function of the respective model class.

- By default, 3 linear regression models (arithmetic, ordinal, and logarithmic dose metameters) are included,
- several logarithmic-linear interpolations are available, if one dose level is 0. By default, the version proposed by (Tukey et al., 1985) is computed.
- Optionally, dose levels can be coerced to a factor, and multiple contrast tests can be performed between dose levels (see options in `?contrMat` in package `multcomp`), including options that are appropriate for downturn effects.
- Alternatively, the comparison of highest dose group to control can be added (without making use of multiple contrast tests).

- Currently, model classes `lm`, `glm`, `lmerMod` and `lme` are supported, the right hand side of the (fixed effects) formula may contain secondary factors or covariates, but no interactions with the dose variable.

The refitted models are combined into an list (resembling the structure use in `mmm` of package `multcomp`). Suitable linear functions defining the hypotheses of interest are combined into an `mlf`-like object. by Default, suitable degrees of freedom are extracted from the model objects.

These lists may be passed manually to function `glht` of package `multcomp`, or by using function `asglht`. Tukey trend objects based on different initial models can be combined into one by function `combtt`. In this way, joint inference is possible

- for models with different endpoints, including endpoints of different scale (continuous, binomial, counts)
- models differing w.r.t secondary covariates,
- or transformations of the endpoints.

When passing the objects to `glht`, one may make further use of the built-in functionality of `glht`, e.g.,

- passing a suitable degree of freedom
- require sandwich estimators to account for heterogeneous variance

Finally, the functions `summary` and `confint` of the `multcomp` package can be used to compute adjusted p-values of tests and simultaneous confidence intervals for slope parameter and/or multiple contrasts.

3 Case studies

An advantage of the modified Tukey trend test is mainly the use of the generalized linear (mixed effect) model. This is demonstrated below by several case studies. Next to package `multcomp`, further packages must be installed to successfully run the example code of the case studies: We will make use of the packages `xtable`, `reshape2`, `plyr` for generation of graphics, tables, and for data operations (Dahl, 2016; Wickham, 2007, 2011). In single examples we illustrate additional options for analysis that use functions of the packages `lme4, pbkrtest, sandwich` (Bates et al., 2015; Halekoh and Højsgaard, 2014; Zeileis, 2006). Further, data sets, plot functions and data generating functions are obtained from CRAN packages `DoseFinding`, `SASmixed`, `HSAUR3`, `Corrbin`, `aods3`, `stepp`, `SimComp`.

```
library("xtable")
```

Finally, a number of example data sets are obtained from the R package `SiTuR`, which is available from a `gitHub` repository. The package may be installed via

```
library("devtools")
install_github("lahothorn/SiTuR")
```

However, these data sets are also included in the `.Rnw` source of this vignette, directly preceding the R code first using the example.

3.1 One primary endpoint assuming normal distributed residuals

The data set `litter` is used for the standard case where the mean litter weight represent the primary endpoint, `dose` the quantitative variable and `gesttime` (gestation time) and `number` (number of animals in litter) as adjusting covariates (or alternatively as secondary endpoint) (Westfall (1997)).

The first code snippet shows the standard Tukey trend test just for the primary covariate `dosen`.

```
data(litter, package="multcomp")
dl <- litter
dl$dosen <- as.numeric(as.character(dl$dose))
fitw <- lm(weight ~ dosen, data=dl)
ttw <- tukeytrendfit(fitw, dose="dosen", scaling=c("ari", "ord", "arilog"))
exa1<-summary(glht(model=ttw$mmm, linfct=ttw$mlf))
```

None of the p-values is < 0.05. The possible problem in this example is a possible plateau effect for the $5, 50, 500$ doses - a more appropriate model for such shapes is recommended, e.g. in Section 3.2.2.

Table 1: Adjusted p-values for standard Tukey Trend Test

Dose metameter	Test statistics	p-value
dosenari: dosenari	-0.8180	0.5373
dosenord: dosenord	-1.7027	0.1302
dosenarilog: dosenarilog	-1.1283	0.3516

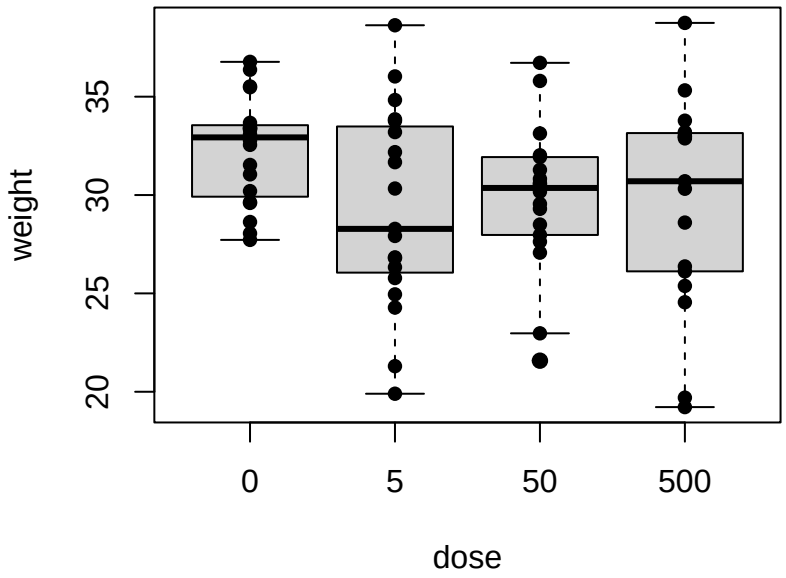

Figure 1: Mean litter weights

3.1.1 Small sample size modification

The underlying `mmm` approach is asymptotic, i.e., too liberal for small sample sizes. The use of a multivariate t-distribution with an appropriate degree of freedom is straightforward. Options are residual degree of freedom (`ddf="residual"`) for linear, generalized and mixed effect models and the Kenward-Roger degree of freedom `ddf="KR"` for mixed effect model, relying on packages `lme4` (Bates et al., 2015) and `pbkrtest` (Halekoh and Højsgaard, 2014).

```
ttdf <- tukeytrendfit(fitw, dose="dosen", scaling=c("ari", "ord", "arilog"), ddf="residual")
exa1a <- summary(glht(model=ttdf$mmm, linfct=ttdf$mlf, df=min(ttdf$df)))
```

The use of this modification is highly recommended for small to medium sample sizes to avoid to liberal results (see the use of Kenward-Rogers df in the mixed effects model in Section 3.6.

3.1.2 Adjustment against multiple covariates

The adjustment against subject-characterizing covariates is a basic principle and easy to achieve in the underlying glm-framework (where both covariates matters in the example, see the object `covars`).

```
fitc <- lm(weight ~ dosen + gesttime+number, data=dl)
covars<-anova(fitc)
ttc<- tukeytrendfit(fitc, dose="dosen", scaling=c("ari", "ord", "arilog"), ddf="residual")
exa3 <- summary(glht(model=ttc$mmm, linfct=ttc$mlf))
```

3.1.3 Consideration of different covariance-adjusting models

An interesting aspect is the simultaneous consideration of different covariance-adjusting models. A maximum test on several `tukeytrendfit` objects can be performed by means of the function `combtt`, for multiple covariance-adjusting models, or multiple models with different endpoints. As an example the evaluation of organ and body weights in a

Table 2: Adjusted p-values for finite Tukey Trend Test- adjusted against covariates

Dose metameter	Test statistics	p-value
dosenari: dosenari	-0.7766	0.5704
dosenord: dosenord	-1.7171	0.1280
dosenarilog: dosenarilog	-1.0356	0.4062

toxicological bioassay is used. When commonly using relative organ weights, it is a priori not clear how to consider the body weight: the relative organ weight, body weight as covariate or ignoring body weight Hot. The liver and body weights from a 13-week study on female F344 rats administered with sodium dichromate dihydrate were used here (NTP). Three different models are formulated: liver weight without considering body weight (`mod1`), relative liver weight (`mod1`), and liver weight with body weight as adjusting covariate (`mod3`)

```
#data("liv", package="SiTuR")

liv$relLiv <-liv$LiverWt/liv$BodyWt
mod1<-lm(LiverWt~Dose, data=liv)
mod2<-lm(relLiv~Dose, data=liv)
mod3<-lm(LiverWt~Dose+BodyWt, data=liv)
tt1<- tukeytrendfit(mod1, dose="Dose", scaling=c("ari", "ord", "arilog"), ddf="residual")
tt2<- tukeytrendfit(mod2, dose="Dose", scaling=c("ari", "ord", "arilog"), ddf="residual")
tt3<- tukeytrendfit(mod3, dose="Dose", scaling=c("ari", "ord", "arilog"), ddf="residual")
cttC <- combtt(tt1,tt2, tt3)
Exa4 <- summary(glht(model=cttC$mmm, linfct=cttC$mlf))
```

In the object `Exa4` the consideration of body weight alone just for the arithmetic dose metameters reveals the smallest p-value (largest test value). Notice, the bivariate consideration of body and organ weight can be a further alternative, see Section 3.3.1.

3.1.4 Allowing variance heterogeneity

In biomedical dose-response problems the assumption of a constant variance is commonly violated. An assumption of constant coefficient of variation seems to be more plausible. More general, heterogeneous variances can be expected. Common linear regression model assumes homogeneous variance. Several modifications for heterogeneous variances are available, among them the sandwich covariance estimator (Zeileis, 2006). Notice that this modification may be liberal for small sample sizes. Creatin kinase data from a toxicological bioassay (Hot) are used as an example.

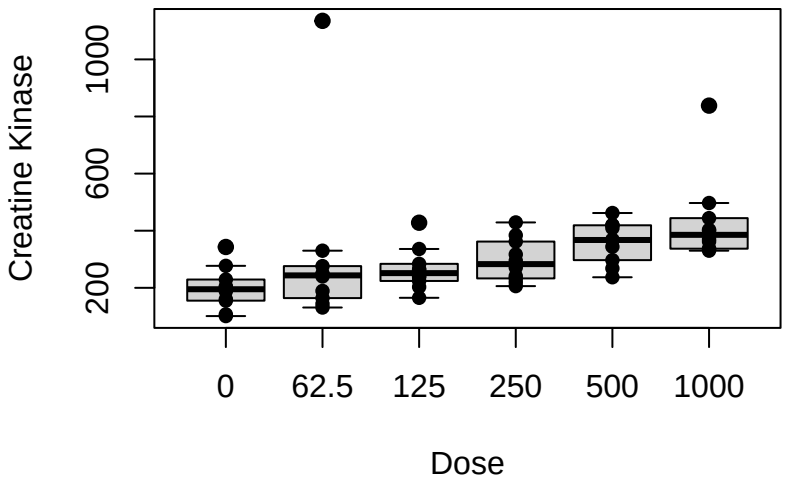

Figure 2: Boxplots of creatin kinase

```
fitG<-lm(value~Dose, data=mclinA)
library("sandwich")
ttO <- tukeytrendfit(fitG, dose="Dose", scaling=c("ari", "ord", "arilog"), ddf="residual")
exa5a<-summary(glht(model=ttO$mmm, linfct=ttO$mlf))
exa5b<-summary(glht(model=ttO$mmm, linfct=ttO$mlf, vcov = sandwich))
```

Notice, for this balanced design the difference between standard and sandwich estimator of the variance-covariance matrix in objects `exa5a,exa5b` is small. Substantial differences may occur when in the dose-response relationship larger variances occur in groups with smaller sample sizes, vice versa.

3.1.5 Finding an appropriate log-transformation for the zero dose control

An approximation must be used to log-transform the zero dose control. Tukey et al. (1985) proposed $log(d_1) - f * (d_1 - d_0)/(d_2 - d_1) * (log(d_2/d_1))$ with $f = 1$. The function `tukeytrendfit` allows an arbitrary vector of positive f-values by the option `d0shift` (again adjusted for these multiple models).

```
fitG<-lm(value~Dose, data=mclinA)
library("sandwich")
ttG <- tukeytrendfit(fitG, dose="Dose", scaling=c("ari", "ord", "arilog"),
                     ddf="residual", d0shift=c(0.2,0.5,1, 2) )
exa5c<-summary(glht(model=ttG$mmm, linfct=ttG$mlf, vcov = sandwich))
```

The object `exa5c` shows the clear effect of an *optimal* transformation: for $f = 10$ the smallest p-value is achieved-much smaller than for the ordinal model selected in object `exa5a`.

3.2 Considering dose as a quantitative covariate and a qualitative factor jointly

A main property to be emphasized is the joint modeling of dose as quantitative covariate and a qualitative factor, shown for different setups.

3.2.1 Inclusion of the control vs. high dose comparison

A power comparison by Aras et al. (2011) of the test control vs. highdose (CvsH) and the Tukey trend test revealed that the CvsH-test is superior to Tukey's trend test only when all the middle doses have identical or near identical response. Because such pattern occur not too frequently, they recommend Tukey's trend test alone. Instead focusing on the decision between CvsH-test and Tukey-test, the CvsH-test can be included as a fourth model in Tukey's trend test. A two-sample CvsH test is identical to a linear regression model for only two groups. Ignoring all other dose groups is technically achieved in the `tukeytrend` package by replac9ing the data of these groups by NA's). The multiplicity penalty for including a further model will be small, because a CvsH-model is highly correlated with the three other models. Notice, that the comparison of the CvsH-test alone (i.e. ignoring all other doses) has a special meaning in trend testing. It is a single contrast within the Williams (1971) or Dunnett (1955) multiple contrast test and its combination (Jaki and Hothorn (2013)) and is the first test in the closed testing procedure under order restriction (Hothorn et al. (1997)).

```
ttw6 <- tukeytrendfit(fitw, dose="dosen", scaling=c("ari", "ord", "log", "arilog", "highvslow"))
Exa6 <- summary(glht(model=ttw6$mmm, linfct=ttw6$mlf, alternative="less"))
```

3.2.2 Tukey trend test and Williams contrasts

The two main representatives of trend test for quantitative and qualitative dose, namely Tukey trend tests and Williams-type contrasts (Bretz and Hothorn, 2002), can be considered jointly.

```
ttw8 <- tukeytrendfit(fitc, dose="dosen", ddf="residual",
                      scaling=c("ari", "ord", "log", "arilog", "treat"),
                      ctype="Williams")
exa8<-summary(glht(ttw8$mmm, ttw8$mlf, alternative="less"))
```

The p-value for comparing all pooled dose groups against control (C3) is < 0.05, i.e., a plateau trend shows the clearest significance (but all quasilinear models reveal rather large p-values).

Table 3: Tukey and Williams Trend Test

Dose metameter	Test statistics	p-value
dosenari: dosenari	-0.7766	0.4435
dosenord: dosenord	-1.7171	0.1171
dosenlog: dosenlog	0.5757	0.9275
dosenarilog: dosenarilog	-1.0356	0.3324
dosentreat: C 1	-2.0050	0.0668
dosentreat: C 2	-2.2706	0.0375
dosentreat: C 3	-2.8467	0.0082

3.2.3 Tukey and Dunnett test

In designs including a zero-dose control, alternatives according to a monotonic trend or differences to control may be of interest Jaki and Hothorn (2013). Therefore Tukey trend test and Dunnett test can be considered jointly.

```
ttw9 <- tukeytrendfit(fitw, dose="dosen",
                      scaling=c("ari", "ord", "log", "arilog", "treat"),
                      ctype="Dunnett")
Exa9<-summary(glht(ttw9$mmm, ttw9$mlf, alternative="less"))
```

3.2.4 Taking specific non-monotonic dose-response relationships into account

In some dose-response assay, particularly in toxicology, a downturn effect at high doses is likely, i.e. a non-monotonic dose-response relationship may occur Hot. Trend tests may be seriously biased and tests without any order restriction, such as Dunnett test, allows only group wise inference. A double maximum test can be used, a maximum over all possible peak point doses (from $k, (k-1), (k-2), ..., 1$) and a maximum on the multiple contrasts, e.g. Williams-type contrasts (Bretz and Hothorn, 2003). This idea can be easily transferred to the Tukey trend test approach by i) defining an so-called `UmbrellaWilliams` contrast (Bretz and Hothorn, 2003), i.e. taking dose qualitatively or, ii) using a double maximum test on the possible peak points and the three regression models, where these dose metameters up to certain dose are replaced with `NA` (see below object `comptt`). Although $3k$ instead of 3 models are subject to a maximum test, the power will be not sacrificed because many of these models are highly correlated.

i) dose qualitatively

```
tt10 <- tukeytrendfit(fitw, dose="dosen", scaling=c("ari", "ord", "arilog", "treat"),
                      ctype="UmbrellaWilliams", ddf="residual")
Exa10<-summary(glht(model=tt10$mmm, linfct=tt10$mlf, alternative="less"))
```

ii) dose quantitatively

```
dl$dos500<-dl$dosen; dl$dos50<-dl$dosen
dl$dos500[dl$dosen==500] <-NA
dl$dos50[dl$dosen==500] <-NA
dl$dos50[dl$dos50==50] <-NA
fitall<-lm(weight ~ dosen, data=dl)
fit500 <- lm(weight ~ dos500, data=dl)
fit50 <- lm(weight ~ dos50, data=dl)
tt50<- tukeytrendfit(fit50, dose="dos50", scaling=c("ari"), ddf="residual")
tt500<- tukeytrendfit(fit500, dose="dos500", scaling=c("ari", "ord", "arilog"), ddf="residual")
ttall<- tukeytrendfit(fitall, dose="dosen", scaling=c("ari", "ord", "arilog"), ddf="residual")
combi10 <- combtt(tt50,tt500,ttall)
comptt<- summary(glht(model=combi10$mmm, linfct=combi10$mlf, alternative="less"))
```

3.2.5 Testing vs. pooled doses

For plateau-shaped dose-response relationships testing vs. the pooled doses may be of interest. For example, in the randomized placebo-controlled dose finding trial with golimumab to treat ulcerative colitis (Rutgeerts et al., 2015), unadjusted p-values have been used for comparisons of all individual doses vs. placebo and as well against the pooled doses. As an example the health-related quality according to IBD-Questionnaire (change from baseline after 6 weeks) was used (see Table 2 in Rutgeerts et al. (2015)). Only summary data are available and therefor simulated normal distributed raw data were generated for the analysis here.

```
library("SimComp")
set.seed(170549)
d1 <- ermvnorm(n=73,mean=c(12.9, 22.0, 23.0, 24.4),sd=c(29.07,32.82, 30.99, 31.77))
d2 <- ermvnorm(n=61,mean=c(12.9, 22.0, 23.0, 24.4),sd=c(29.07,32.82, 30.99, 31.77))
d3 <- ermvnorm(n=75,mean=c(12.9, 22.0, 23.0, 24.4),sd=c(29.07,32.82, 30.99, 31.77))
d4 <- ermvnorm(n=77,mean=c(12.9, 22.0, 23.0, 24.4),sd=c(29.07,32.82, 30.99, 31.77))
dose <- as.numeric(rep(c(0, 1, 2, 4),c(73,61, 75, 77)))
IBDQ<-c(d1[,1], d2[,2], d3[,3], d4[,4])
rut<-data.frame(IBDQ,dose)
lmRU <-glm(IBDQ~dose, data=rut)
Cmat3 <-  c(-3, 1, 1, 1)
EXRU <- tukeytrendfit(lmRU, dose="dose", scaling=c("ari", "ord", "arilog", "treat"), ctype=Cmat3)
EXRRU <- summary(glht(model=EXRU$mmm, linfct=EXRU$mlf))
```

Table 4: Tukey Trend Test and pooled dose vs. placebo test

Dose metameter	Test statistics	p-value
doseari: doseari	2.0435	0.0751
doseord: doseord	2.2206	0.0499
dosearilog: dosearilog	2.2206	0.0496
dosetreat: 1	-2.4785	0.0252

The p-values for a trend (log-transformed) is about 0.05, but for the pooled doses vs. placebo it is about 0.025, not surprising for such a plateau shape.

3.2.6 Designs without a control group

Many designs in preclinical or clinical trials include a negative control or placebo. The comparison of the doses vs. control is mainly of interest. However, designs without a control exists, e.g., Verrier et al. (2014). Whereas the Williams trend test can not be used and other multiple contrast tests such as Changepoint contrasts must be used, the Tukey trend test can easily used for such data.

```
verrier<-data.frame(
   dose = c(5,10, 20, 40, 60),
   events = c(28, 26,14,12,4),
   n = c(70, 59, 90,88, 77))
lmV <-glm(cbind(events,n-events)~dose, data=verrier, family= binomial(link="logit"))
EXV <- tukeytrendfit(lmV, dose="dose", scaling=c("ari", "ord", "log"))
EXVV <- summary(glht(model=EXV$mmm, linfct=EXV$mlf))
```

3.3 Multiple endpoints

Although the independent dose-response analysis is quite common, the joint analysis of p correlated endpoints is needed and already proposed in the original procedure (Tukey et al., 1985) with empirical choice of the Sidak inequality. The extension of the multiple marginal model approach is straightforward by using $p * 3$ models. Here, the advantage of this new method is clearly shown: the variance matrix needs not be explicitly formulated, but it is estimated from the data in each specific design. Moreover, it allows also differently scaled endpoints in the generalized linear model-alternative evaluations are rather complicated. This allows the analysis of equally scaled multiple endpoints, such as multiple normals (as an alternative the a related max(max)-test proposed by Hasler and Hothorn (2013)) or multiple binary endpoints (as an alternative the a related max-test in Klingenberg and Satopaeae (2013)). Moreover, it allows also the consideration differently scaled multiple endpoints, such a normal and binary, as an alternative to the joint modeling approach (Faes et al., 2006).

3.3.1 Multiple normal endpoints

In the above liver weight example, liver and body weights can be considered also as two primary endpoints Hot.

```
fitLl <-lm(LiverWt~Dose, data=liv)
fitLb <-lm(BodyWt~Dose, data=liv)
ttLl <- tukeytrendfit(fitLl, dose="Dose", scaling=c("ari", "ord", "arilog"))
ttLb <- tukeytrendfit(fitLb, dose="Dose", scaling=c("ari", "ord", "arilog"))
cttL <- combtt(ttLl, ttLb)
EXA11<-summary(glht(model=cttL$mmm, linfct=cttL$mlf))
```

For both organ and body weight a decreasing trend exist for arithmetic dose scores, clearer for organ weight.

Table 5: Bivariate Tukey Trend Test

Dose metameter	Test statistics	p-value
ttLl.lm.LiverWt.Doseari: Doseari	-6.0359	0.0000
ttLl.lm.LiverWt.Doseord: Doseord	-5.1679	0.0000
ttLl.lm.LiverWt.Dosearilog: Dosearilog	-5.1679	0.0000
ttLb.lm.BodyWt.Doseari: Doseari	-5.3800	0.0000
ttLb.lm.BodyWt.Doseord: Doseord	-3.9784	0.0002
ttLb.lm.BodyWt.Dosearilog: Dosearilog	-3.9784	0.0001

3.3.2 Multiple normal endpoints with missing data

The mmm-approach can deal with missing values under the MCAR assumption (Pipper et al., 2012). Below, an artificial pattern of missing values is introduced, just to illustrate the ability of the `tukeytrend` package to deal with missing values.

```
# data("liv", package="SiTuR")
livv<-liv
livv[3,2] <- livv[14,2] <- livv[19,2] <- livv[46,2] <- livv[53,2] <- livv[54,2]<- NA
livv[1,3] <- livv[11,3] <- livv[18,3] <- livv[42,3] <- livv[55,3]<- NA #missing MCAR
fitLlv <-lm(LiverWt~Dose, data=livv)
fitLbv <-lm(BodyWt~Dose, data=livv)
ttLlv <- tukeytrendfit(fitLlv, dose="Dose", scaling=c("ari", "ord", "arilog"))
ttLbv <- tukeytrendfit(fitLbv, dose="Dose", scaling=c("ari", "ord", "arilog"))
cttLv <- combtt(ttLlv, ttLbv)
Exa15<-summary(glht(model=cttLv$mmm, linfct=cttLv$mlf))
```

3.3.3 Multiple binary endpoints

The analysis of multiple binary endpoints is commonly a challenge because the correlation is rather small, zero cells may occur and the sample sizes are often too small to assume reasonable performance of asymptotic Wald-type intervals or tests. In the following multiple tumors example (Hot), 4 treatment groups (doses 0, 37, 75, 150), each containing 50 mice, have been investigated for presence or absence of 89 different tumor classifications (t01,...,t89). We will restrict our analysis on those 10 tumor classifications, that show an overall abundance more than 5. Most of the remaining 89 classifications occur just once or twice across all 200 mice.

```
# data("miceF", package="SiTuR")

miceF$dose <- miceF$group
      miceF$dose[miceF$dose==0] <-0
      miceF$dose[miceF$dose==1] <-37
      miceF$dose[miceF$dose==2] <-75
      miceF$dose[miceF$dose==3] <-150
cst <- colSums(miceF[, 4:92])
tt5 <-as.matrix(miceF[, names(cst[cst>5])])
```

The Figure shows the resulting 10 x 200 data matrix with presence/absence of the 10 most abundant tumors in black/white, along with the dose groups as a color scale. Some of the tumor classes are mutually exclusive subclasses of the same tumor (t26,t27; t28, t29), other tumor classes may be weakly correlated.

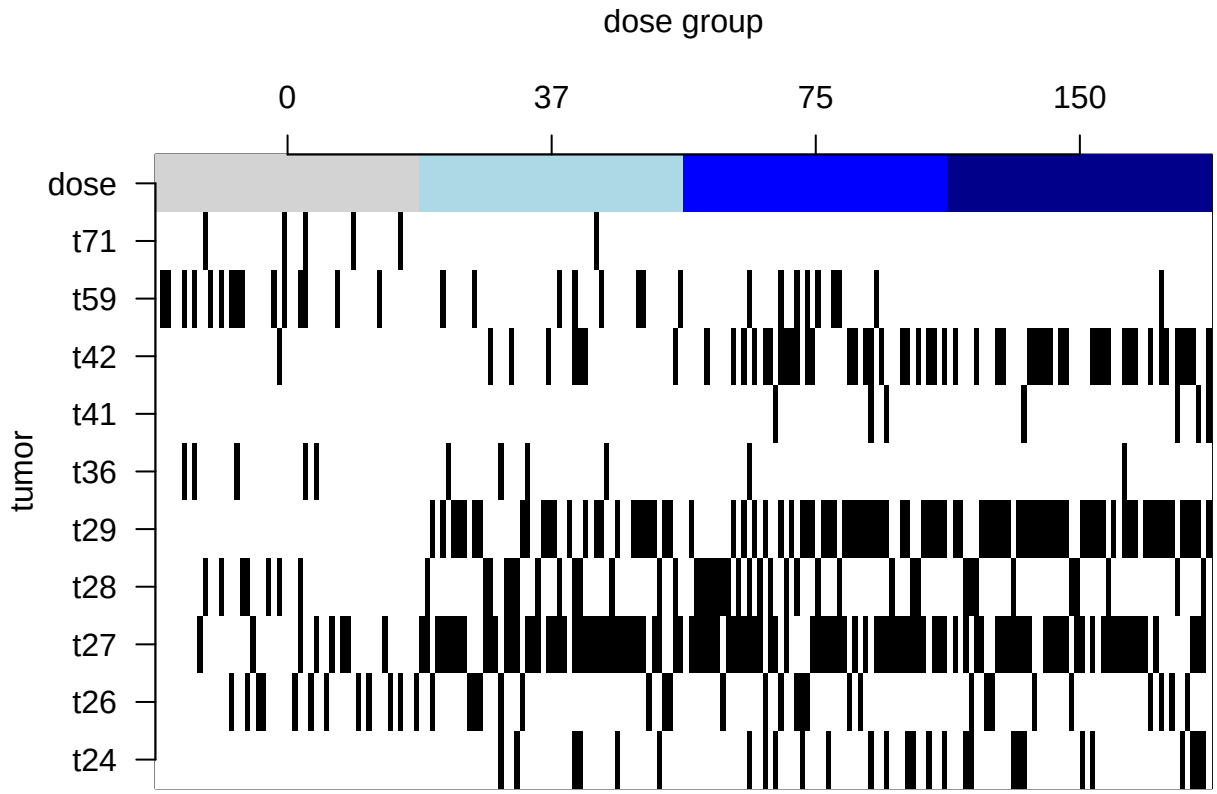

Figure 3: Presence/absence of 10 tumor classes in 200 mice

Below, Tukey trend test (arithmetic, ordinal and interpolated logarithmic dose metameters) is applied on 10 generalized linear models with logit link and binomial assumption.

```
N24i <- glm(t24 ~ dose, data=miceF, family=binomial())
N26i <- glm(t26 ~ dose, data=miceF, family=binomial())
N27i <- glm(t27 ~ dose, data=miceF, family=binomial())
N28i <- glm(t28 ~ dose, data=miceF, family=binomial())
N29i <- glm(t29 ~ dose, data=miceF, family=binomial())
N36i <- glm(t36 ~ dose, data=miceF, family=binomial())
N41i <- glm(t41 ~ dose, data=miceF, family=binomial())
N42i <- glm(t42 ~ dose, data=miceF, family=binomial())
N59i <- glm(t59 ~ dose, data=miceF, family=binomial())
N71i <- glm(t71 ~ dose, data=miceF, family=binomial())

tu24i <- tukeytrendfit(N24i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu26i <- tukeytrendfit(N26i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu27i <- tukeytrendfit(N27i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu28i <- tukeytrendfit(N28i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu29i <- tukeytrendfit(N29i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu36i <- tukeytrendfit(N36i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu41i <- tukeytrendfit(N41i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu42i <- tukeytrendfit(N42i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu59i <- tukeytrendfit(N59i, dose="dose", scaling=c("ari", "ord", "arilog"))
tu71i <- tukeytrendfit(N71i, dose="dose", scaling=c("ari", "ord", "arilog"))

tt10 <- combtt(tu24i, tu26i, tu27i, tu28i, tu29i, tu36i, tu41i, tu42i, tu59i, tu71i)
stt10 <- summary(asglht(tt10))
```

For tumor classes t24, t27, t29 and t42 we find significantly increasing proportions with increasing dose.

3.3.4 Multiple, differently scaled endpoints

A first example considers a normal distributed and count endpoint, a second example a normal distributed and binary endpoint. Rather common are multiple, but differently scaled endpoints, such as normal and binomial, or binomial

Model	Test stats	p-value
tu24i.glm.t24.doseari: doseari	3.05	0.03022
tu24i.glm.t24.doseord: doseord	3.29	0.01444
tu24i.glm.t24.dosearilog: dosearilog	3.29	0.01419
tu26i.glm.t26.doseari: doseari	-0.67	0.99938
tu26i.glm.t26.doseord: doseord	-0.80	0.99668
tu26i.glm.t26.dosearilog: dosearilog	-0.80	0.99672
tu27i.glm.t27.doseari: doseari	3.60	0.00490
tu27i.glm.t27.doseord: doseord	4.41	0.00022
tu27i.glm.t27.dosearilog: dosearilog	4.40	0.00025
tu28i.glm.t28.doseari: doseari	0.31	1.00000
tu28i.glm.t28.doseord: doseord	0.82	0.99613
tu28i.glm.t28.dosearilog: dosearilog	0.82	0.99612
tu29i.glm.t29.doseari: doseari	6.44	0.00000
tu29i.glm.t29.doseord: doseord	6.79	0.00000
tu29i.glm.t29.dosearilog: dosearilog	6.79	0.00000
tu36i.glm.t36.doseari: doseari	-1.84	0.52293
tu36i.glm.t36.doseord: doseord	-1.98	0.41804
tu36i.glm.t36.dosearilog: dosearilog	-1.98	0.41703
tu41i.glm.t41.doseari: doseari	2.26	0.24287
tu41i.glm.t41.doseord: doseord	2.21	0.26880
tu41i.glm.t41.dosearilog: dosearilog	2.21	0.26888
tu42i.glm.t42.doseari: doseari	5.65	0.00000
tu42i.glm.t42.doseord: doseord	5.78	0.00000
tu42i.glm.t42.dosearilog: dosearilog	5.79	0.00000
tu59i.glm.t59.doseari: doseari	-3.41	0.00950
tu59i.glm.t59.doseord: doseord	-3.45	0.00801
tu59i.glm.t59.dosearilog: dosearilog	-3.44	0.00857
tu71i.glm.t71.doseari: doseari	-2.02	0.39005
tu71i.glm.t71.doseord: doseord	-2.11	0.33266
tu71i.glm.t71.dosearilog: dosearilog	-2.10	0.33908

Table 6: Tukey trend test for bivariate binary

and time-to-event. In the above litter weight example, the number of litter mates can be considered as a second endpoint,that cna be analysed ibn a generalized linear model with quasipoisson assumption (McCullagh and Nelder, 1989).

```
fitw <- lm(weight ~ dosen + gesttime, data=dl) #lm-model
ttw <- tukeytrendfit(fitw, dose="dosen", scaling=c("ari", "ord", "arilog"))
fitqp <- glm(number~dosen + gesttime, data=dl, family=quasipoisson) #  glm-model
ttqp <- tukeytrendfit(fitqp, dose="dosen", scaling=c("ari", "ord", "arilog"))
cttqw <- combtt(ttqp, ttw) # combine both models
Exa12<-summary(glht(model=cttqw$mmm, linfct=cttqw$mlf))
```

The dependencies between differently scaled multiple endpoints pup weight, fetal death and malformation and the covariate dose can be analyzed by specific regression models Catalano and Ryan (1992), weighted potential outcomes using principal strata Elliott et al. (2006),threshold models Faes et al. (2004) or bivariate random effects models Najita et al. (2009). As an example the ethylene glycol data from a developmental toxicity study conducted by the National Toxicology Program are available Wu and de Leon (2014).

94 pregnant mice were randomly exposed to ethylen glycol at four different dose levels, 0, 0.75, 1.5, and 3 g/kg/day (with dose 0 as control) and 1028 live fetuses (litter sizes ranging from 1 to 16), were examined for various defects, including fetal weight (continuous) and the presence/absence of fetal malformations (binary), both believed to be sensitive indicators of toxicity. Figure 4 contains pup weights and number of malformations on a per-foetus level, see the raw data in Table 7.

Litter	Dose	Weight	Malformation
60	0	0.90	0
60	0	0.83	0
60	0	0.95	0
60	0	0.95	0
60	0	1.07	0
60	0	1.06	0
60	0	0.96	0
61	0	1.02	0
61	0	1.01	0
61	0	1.07	0
61	0	0.82	0
61	0	1.01	0
61	0	0.99	0
61	0	1.06	0
61	0	1.04	0
61	0	1.02	0
61	0	1.04	0
61	0	0.91	0
...	...	...	...
156	3000	0.79	0
156	3000	0.90	0
156	3000	0.86	0
156	3000	0.86	0
156	3000	0.80	0
156	3000	0.84	0
156	3000	0.87	0
156	3000	0.72	0
156	3000	0.83	0

Table 7: Raw data of a fetal weight and malformation

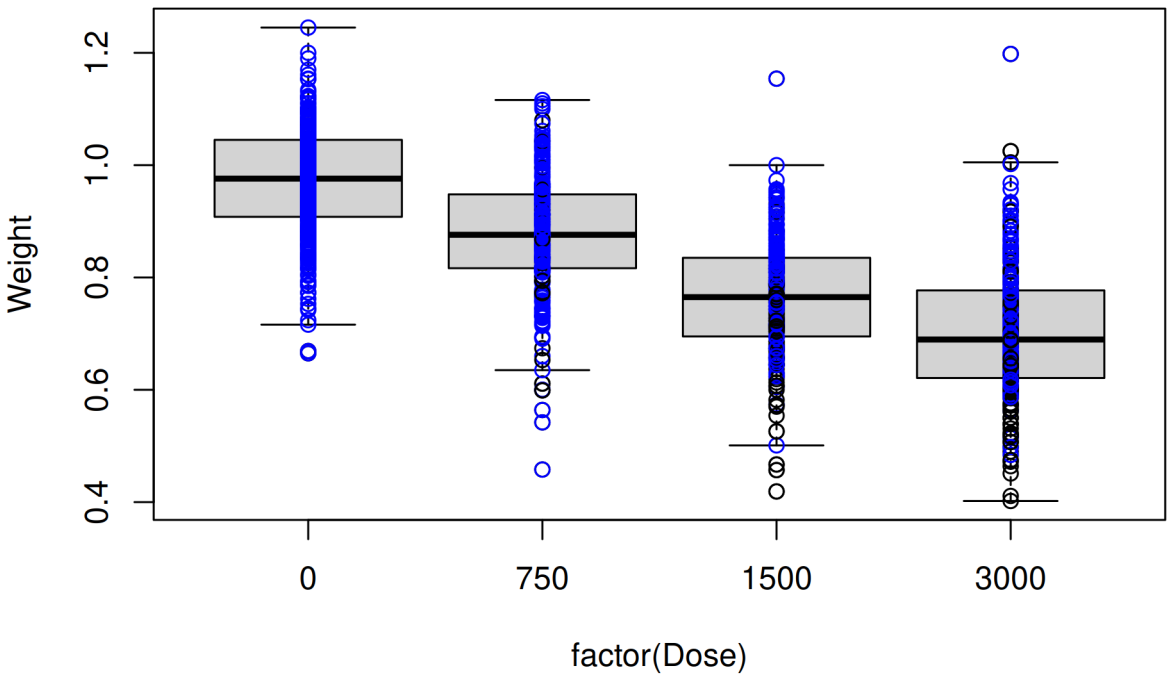

Figure 4: Ethylen glycol example

```
wuleon$dose<-as.factor(wuleon$Dose)
fitwu <- lm(Weight~Dose, data=wuleon)
ttwu <- tukeytrendfit(fitwu, dose="Dose", scaling=c("ari", "ord", "arilog"))
fitwm <- glm(Malformation~Dose, family=binomial(),data=wuleon)
ttwm <- tukeytrendfit(fitwm, dose="Dose", scaling=c("ari", "ord", "arilog"))
cttwum <- combtt(ttwu, ttwm) # combine both models
Exawu<-summary(glht(model=cttwum$mmm, linfct=cttwum$mlf))
```

Model	Test stats	p-value
ttwu.lm.Weight.Doseari: Doseari	-28.69	0.0000000
ttwu.lm.Weight.Doseord: Doseord	-30.63	0.0000000
ttwu.lm.Weight.Dosearilog: Dosearilog	-30.63	0.0000000
ttwm.glm.Malformation.Doseari: Doseari	14.64	0.0000000
ttwm.glm.Malformation.Doseord: Doseord	14.24	0.0000000
ttwm.glm.Malformation.Dosearilog: Dosearilog	14.24	0.0000000

Table 8: Tukey trend test for joint modeling of normal and binomial endpoint

3.4 Using generalized linear models

One advantage of the above Tukey trend test approach is its generalization in the generalized linear model (GLM), demonstrated??below for proportions, proportions with overdispersion, a multinomial endpoint decomposed into multiple binary endpoints, overdispersed counts and ordered categorical data.

3.4.1 Proportions

As an example, 2-by-5 table data with non-zero event in the control from a dose finding study to assess the dose dependent suppression of serum prolactin by cabergoline in hyperprolactinaemia certain adverse events (Bretz and Hothorn (2002)).

```
adverse <-c(rep("absent",11), rep("present",9), rep("absent",24), rep("present",19),
         rep("absent",21),rep("present",21),rep("absent",21),rep("present",21),
         rep("absent",17),rep("present",24))
#dose <-c(rep("0", 20), rep("0.125", 43),rep("0.5", 42),rep("0.75", 42), rep(1, 41))
dose <- c(rep(0, 20), rep(0.125, 43),rep(0.5, 42),rep(0.75, 42), rep(1, 41))
webster <- data.frame(adverse=factor(adverse),dose=dose)
t(table(webster))
```

adverse dose absent present 0 11 9 0.125 24 19 0.5 21 21 0.75 21 21 1 17 24

```
lmW <-glm(adverse~dose, data=webster, family= binomial(link="logit"))
EX16 <- tukeytrendfit(lmW, dose="dose", scaling=c("ari", "ord", "arilog"))
EXA16 <- summary(glht(model=EX16$mmm, linfct=EX16$mlf))
```

3.4.2 Multinomial endpoint

The comparison of multinomial vectors represents a rather specific problem, such as differential blood count Schaarschmidt et al. (2017). The problem is even more complicated when overdispersion may occur. For example, in a reproductive toxicity experiment n_{ij} females are treated within the dose groups (and zero-dose control) and the health status of each pup within a single female is classified into unaffected, malformed or death. Up to now, approaches for simultaneous inference for overdispersed multinomial vectors seems to be not available. Therefore, this problem is first split into two proportions $x_{death}/(x_{unaffected} + x_{malformed} + x_{dead})$ and $x_{malformed}/(x_{unaffected} + x_{malformed} + x_{dead})$, followed by a trend test for two correlated overdispersed binomial endpoints.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Linear Hypotheses:
##                                                  Estimate Std. Error
## tMmN.glm.cbind(NResp.1,.Doseari: Doseari == 0   0.0012496  0.0002479
## tMmN.glm.cbind(NResp.1,.Doseord: Doseord == 0   0.5522732  0.1035582
## tMmN.glm.cbind(NResp.1,.Dosearilog: Dosearilog == 0  0.9493020  0.1746926
```

```
## tDmN.glm.cbind(NResp.2,.Doseari: Doseari == 0        0.0023947  0.0002691
## tDmN.glm.cbind(NResp.2,.Doseord: Doseord == 0        0.8683509  0.1139120
## tDmN.glm.cbind(NResp.2,.Dosearilog: Dosearilog == 0  1.3182634  0.1825614
## tNmN.glm.cbind(NResp.3,.Doseari: Doseari == 0       -0.0031460  0.0003042
## tNmN.glm.cbind(NResp.3,.Doseord: Doseord == 0       -1.0901653  0.1196304
## tNmN.glm.cbind(NResp.3,.Dosearilog: Dosearilog == 0 -1.6306381  0.1853978
##                                                     z value Pr(>|z|)
## tMmN.glm.cbind(NResp.1,.Doseari: Doseari == 0         5.042 9.41e-07 ***
## tMmN.glm.cbind(NResp.1,.Doseord: Doseord == 0         5.333 1.94e-07 ***
## tMmN.glm.cbind(NResp.1,.Dosearilog: Dosearilog == 0   5.434 1.10e-07 ***
## tDmN.glm.cbind(NResp.2,.Doseari: Doseari == 0         8.898  < 1e-07 ***
## tDmN.glm.cbind(NResp.2,.Doseord: Doseord == 0         7.623  < 1e-07 ***
## tDmN.glm.cbind(NResp.2,.Dosearilog: Dosearilog == 0   7.221  < 1e-07 ***
## tNmN.glm.cbind(NResp.3,.Doseari: Doseari == 0       -10.342  < 1e-07 ***
## tNmN.glm.cbind(NResp.3,.Doseord: Doseord == 0        -9.113  < 1e-07 ***
## tNmN.glm.cbind(NResp.3,.Dosearilog: Dosearilog == 0  -8.795  < 1e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

A difference to existing methods to fit models for overdispersed multinomial data (Yee), is the allowance for category-specific overdispersion.

3.4.3 Overdispersed counts

As an example for overdispersed count data a micronucleus assay on three doses of phenylethanol, together with a negative control and a positive control Engelhardt (2006) was selected

```
#data("mn", package="SiTuR")

Mn<-droplevels(mn[mn$group != "Positive", ])
Mn$dose<-c(rep(0,5), rep(188, 5), rep(375,5), rep(750,5))
fitMN <- glm(MN ~ dose, data=Mn, family=quasipoisson(link="log"))
Exa20 <- tukeytrendfit(fitMN, dose="dose", scaling=c("ari", "ord", "arilog", "treat"),
                       ctype="UmbrellaWilliams")
EXA20<-summary(glht(model=Exa20$mmm, linfct=Exa20$mlf))
```

3.4.4 Ordered categorical data

Rather different approaches can be used for the evaluation ordered categorical data, among a glm-approach with quasipoisson link function. As an example the hyaline droplets findings in male rats treated with hexachloro-butadiene, four severity scores were used: 1 - no abnormality detected; 2 - minimal, occasional small hyaline droplets in occasional tubules; 3 - mild, scattered small hyaline droplets, 4 - moderate, high number of variable sized droplets (Swain et al., 2012).

```
#data("hyalin", package="SiTuR")

exP<-glm(droplets~dose, data=hyalin, family=quasipoisson(link = "log"))

ntab <- table(hyalin$dose)  # sample sizes needed for unbalanced designs
cmd <- contrMat(n=ntab, type="Dunnett")
cmuw <- contrMat(n=ntab, type="UmbrellaWilliams")
cmduw <- rbind(cmd, cmuw) # combine both matrices
aa<-summary(glht(exP, linfct=mcp(dose=cmduw),alternative="greater"))

Dlevel0<-as.numeric(levels(hyalin$dose))
S00<-log(Dlevel0[2])-log(Dlevel0[3]/Dlevel0[2])*(Dlevel0[2]-Dlevel0[1])/(Dlevel0[3]-Dlevel0[2])
hyalin$DoseN<-as.numeric(as.character(hyalin$dose))
hyalin$Dose0<-as.numeric(hyalin$dose)
hyalin$DoseL<-log(hyalin$DoseN)
hyalin$DoseLL<-hyalin$DoseL
hyalin$DoseLL[hyalin$DoseN==Dlevel0[1]] <-S00
```

```
# umbrella needed?
lmN <-glm(droplets~DoseN, data=hyalin,family=quasipoisson(link = "log"))
lmO <-glm(droplets~DoseO, data=hyalin, family=quasipoisson(link = "log"))
lmLL <-glm(droplets~DoseLL, data=hyalin, family=quasipoisson(link = "log"))
ordC <- summary(glht(mmm(covariate=lmN, ordinal=lmO,  linlog=lmLL),
             mlf(covariate="DoseN=0", ordinal="DoseO=0", linlog="DoseLL=0")))
```

Model	Test stats	p-value
covariate: DoseN	4.2641077	0.0001172
ordinal: DoseO	7.0549311	0.0000000
linlog: DoseLL	7.7194447	0.0000000

Table 9: Tukey trend test for ordered categorical data

3.5 Regression models treating dose as continuous data

Not all dose-response problems are based on few pre-defined dose levels, but continuous exposure measures are quite common, e.g. in epidemiology. When assuming dose as quantitative covariate, these type of data can be easily analysed by means of Tukey's trend test. Post-hoc groups definition for using multiple contrast test can be problematic. As an example patient's blood pressure during surgery, using individual doses of a hypotensive drug is used.

```
if(requireNamespace("HSAUR3")){
data(bp, package="HSAUR3")
bp$Dose<-10^(bp$logdose)
fitbp <-lm(bloodp~Dose+recovtime, data=bp)
ttbp <- tukeytrendfit(fitbp, dose="Dose", scaling=c("ari", "ord", "log"))
EXA31 <- summary(glht(model=ttbp$mmm, linfct=ttbp$mlf))
EXA31}
```

Simultaneous Tests for General Linear Hypotheses
Linear Hypotheses: Estimate Std. Error z value Pr(¿—z—) Doseari: Doseari == 0 0.040248 0.008623 4.667 ¡1e-04 *** Doseord: Doseord == 0 0.434789 0.105583 4.118 ¡1e-04 *** Doselog: Doselog == 0 5.915333 1.438424 4.112 ¡1e-04 *** — Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 (Adjusted p values reported – single-step method)

3.6 Mixed effect model

Tukey trend test can be applied in mixed effect models, where `dose` is assumed as a fixed effect, and further effects may be included as random effects to account for repeated measures at the same subject, blocks, multiple centres, technical or biological replicates, etc. Currently, the `tukeytrend` package does not transform the dose variable when it is part of the random effect structure, e.g. to include between-subject variation of reaction to increasing doses by random slopes. However, Sets of models that contain transformed versions of the dose variable in fixed as well as random effects may be fitted, and analyzed jointly by direct input into `mmm`.

3.6.1 Block design

The first example used a block design where some animals are housed in different barns, treated with 0, 10, 20 or 30 g additive in feed, where for each dose of fee, one animal is present in each barn (RCBD). The weight gains of 160 days are measured together with their baseline values (InitWt). A simple covariance-adjusted mixed effect model was used and the p-value for the Tukey trend test (using Kenward-Roger df) are reported in Table **??**, revealing a clear trend according to a rather shifted log-linear dose metameter.

```
if(requireNamespace("SASmixed")){
data(AvgDailyGain, package="SASmixed")
AVG<-AvgDailyGain

library("lme4")
mmNN <-lmer(adg~Trt+InitWt+(1|Block), data=AVG)
tt21 <- tukeytrendfit(mmNN, dose="Trt", scaling=c("ari", "ord", "arilog"),
                      ddf="residual", d0shift=c(0.1, 0.5, 1,  5) )
comptt21 <- glht(model=tt21$mmm, linfct=tt21$mlf)
Exa21a<-summary(comptt21)
Exa21a
}
```

3.6.2 Multi-centre study

The dose-response relationships for liver weight in eight labs were investigated in an immunotoxicological ring study Hothorn (2003). A possible approach for such multicentre studies is a mixed effects model where the centre is considered as random.

```
library("lme4")
mixedA <- lmer(weight ~ Dose+ (1|centre), data=atla)

tt24 <- tukeytrendfit(mixedA, dose="Dose", scaling=c("ari", "ord", "arilog"),
                      ddf="KR")
comptt24 <- glht(model=tt24$mmm, linfct=tt24$mlf)
Exa24<-summary(comptt24)
```

Comparison	Test stats	p-value
Doseari: Doseari	28.64	0.00000000
Doseord: Doseord	28.04	0.00000000
Dosearilog: Dosearilog	31.40	0.00000000

Table 10: Multicentre trial: mixed model

3.6.3 Technical replicates

Technical replicates are quiet common in biomedical research. As a data example the Comet mutagenicity assay is used where tail intensities for liver in each 5 animals, with duplicated samples and each 50 cells are considered, see Hot, p.144ff.

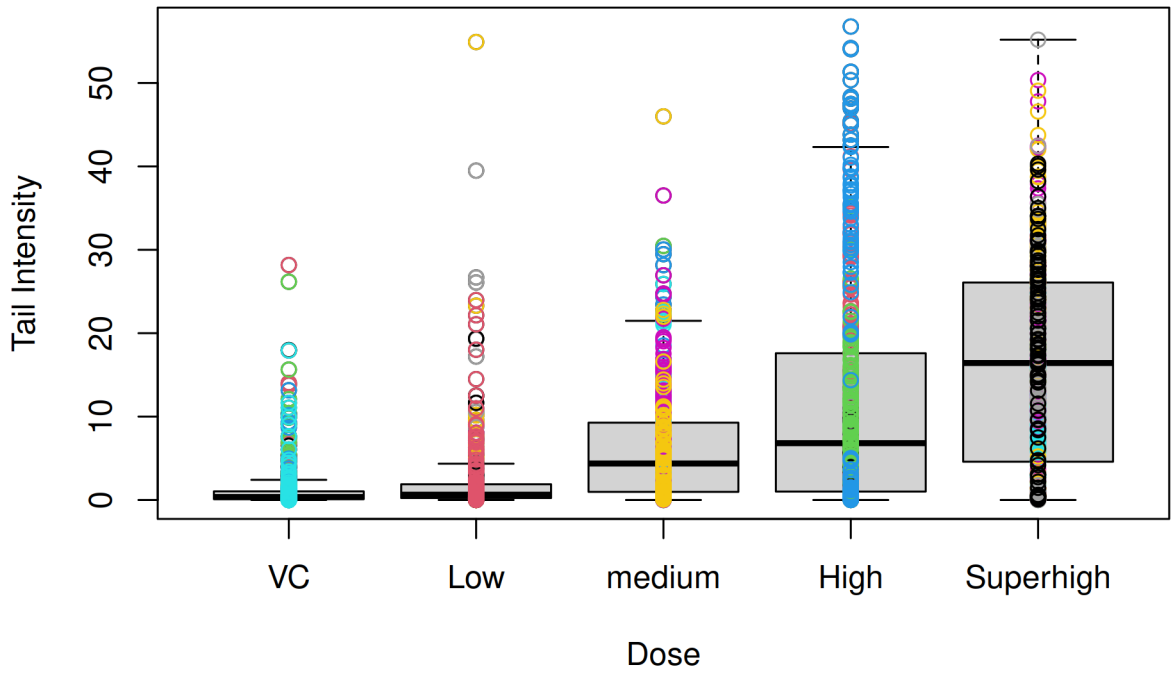

Figure 5: Boxplots for tail intensities in Comet assay.

The boxplots in Figure 5 show dose-dependent skewness and bimodality and the between-animal variability. Here log-transformed tail intensities are considered, where the between-animals and between-samples variability is modeled by means of a mixed effects model.

```
TIComet$Dose[TIComet$Treatment=="VC"] <-0
TIComet$Dose[TIComet$Treatment=="Low"] <-25
TIComet$Dose[TIComet$Treatment=="medium"] <-75
TIComet$Dose[TIComet$Treatment=="High"] <-150
TIComet$Dose[TIComet$Treatment=="Superhigh"] <-500

TIComet$logTI <-log(TIComet$Tail.intensity+0.001)
mixL <- lmer(logTI ~ Dose+(1|Animal_no/Sample), data=TIComet)
EXB <- tukeytrendfit(mixL, dose="Dose", scaling=c("ari", "ord", "arilog"))
EXBB <- summary(glht(model=EXB$mmm, linfct=EXB$mlf))
```

3.6.4 Cross-over design

In the cross-over design within the subjects the period effects should ne modeled for the primary covariate dose. The reduction of appetite is investigated in a 3 by 3 cross-over trial of three dose (0, 4, or 8g) fenugreek (a food supplemented) in 18 adults where 3 participants were randomly assigned to each of the 6 sequences **(personal communication, T. Jaki)**.

Note, that with these three doses the three versions of rescaling, arithmetic (0,4,8), ordinal (0,1,2) and interpolated logarithmic (0.693, 1.386, 2.079), result in exactly the same regressions except that estimates and standard errors are scaled up or down by constants. That is, the three corresponding test statistics test are perfectly correlated. The methods for multiplicity adjustment based on quantiles of the multivariate t distribution (packages `mvtnorm` (Genz et al.; Genz and Bretz, 2009) used by `multcomp` (Hothorn et al., 2008)) deal with that problem, such that no multiplicity adjustment is performed within the group of these three regression models.

```
library("lme4")
modFE <- lmer(Ad_Lib_Lunch ~ trt + period + (1| subject), REML = TRUE, data=fenu)
ttfe <- tukeytrendfit(modFE, dose="trt", scaling=c("ari", "ord", "arilog"), ddf="KR")
compfe <- summary(glht(model=ttfe$mmm, linfct=ttfe$mlf))
compfe

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Linear Hypotheses:
##                            Estimate Std. Error z value Pr(>|z|)
## trtari: trtari == 0         -15.907      9.937  -1.601    0.109
## trtord: trtord == 0         -63.628     39.750  -1.601    0.109
## trtarilog: trtarilog == 0   -91.796     57.347  -1.601    0.109
## (Adjusted p values reported -- single-step method)
```

The correlation matrix used internally for adjusting for multiplicity (rounded to 2 digits) can be inspected by:

```
round(cov2cor(vcov(compfe)), digits=2)

##                      trtari: trtari trtord: trtord trtarilog: trtarilog
## trtari: trtari                    1              1                    1
## trtord: trtord                    1              1                    1
## trtarilog: trtarilog              1              1                    1
```

References

Hothorn, L.A. Statistics in Toxicology-using R (2015).

National Toxicology Program. 13 Weeks gavage study on female F344 rats administered with Sodium dichromate dihydrate (VI) (CASRN: 7789-12-0, Study Number: C20114,TDMS Number:2011402.

G. Aras, A. Xue, and T. Liu. Tukey's contrast test versus two-sample test in a dose-response clinical trial. *Statistics In Biopharmaceutical Research*, 3(1):31–39, Feb. 2011. doi: 10.1198/sbr.2010.09036.

P. Armitage. Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386, 1955.

D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

B. Bertelsen, L. Melchior, Z. Tumer, C. Groth, N. M. Debes, L. Skov, K. K. Hoist, B. Fagerlund, and J. D. Mikkelsen. Association of the chrna7 promoter variant-86t with tourette syndrome and comorbid obsessive-compulsive disorder. *Psychiatry Research*, 219(3):710–711, Nov. 2014. doi: 10.1016/j.psychres.2014.06.032.

B. Bornkamp, J. Pinheiro, and F. Bretz. Mcpmod: An r package for the design and analysis of dose-finding studies. *Journal of Statistical Software*, 29(7):1–23, Feb. 2009.

F. Bretz and L. Hothorn. Detecting dose-response using contrasts: asymptotic power and sample size determination for binomial data. *STATISTICS IN MEDICINE*, 21:3325–3335, 2002.

F. Bretz and L. Hothorn. Statistical analysis of monotone or non-monotone dose-response data from in vitro toxicological assays. *ATLA-Altern Lab Anim*, 31(Suppl. 1):81–96, JUN 2003. ISSN 0261-1929.

P. Catalano and L. Ryan. Bivariate latent variable models for clustered discrete and continuous outcomes. *Journal of the American Statistical Association*, 87:651–658, 1992.

D. B. Dahl. *xtable: Export Tables to LaTeX or HTML*, 2016. URL `https://CRAN.R-project.org/package=xtable`. R package version 1.8-2.

C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc*, 50(272):1096–1121, 1955.

M. Elliott, M. Joffe, and Z. Chen. A potential outcomes approach to developmental toxicity analyses. *Biometrics*, 62: 352–360, 2006.

G. Engelhardt. In vivo micronucleus test in mice with 1-phenylethanol. *Archives of Toxicology*, 80:868–872, 2006.

C. Faes, M. Aerts, H. Geys, G. Molenberghs, and Declerck. Bayesian testing for trend in a power model for clustered binary data. *Environmental and Ecologoical Statistics*, 11:305–322, 2004.

C. Faes, H. Geys, M. Aerts, and G. Molenberghs. A hierarchical modeling approach for risk assessment in developmental toxicity studies. *Computational Statistics & Data Analysis*, 51(3):1848–1861, Dec. 2006. doi: 10.1016/j.csda.2005.12.002.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195.* Springer-Verlag, Heidelberg, 2009.

A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. *mvtnorm: Multivariate Normal and t Distributions. R package version 1.0-5.* URL `http://CRAN.R-project.org/package=mvtnorm`.

U. Halekoh and S. Højsgaard. A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software*, 59(9):1–30, 2014. URL `http://www.jstatsoft.org/v59/i09/`.

M. Hasler and L. A. Hothorn. Simultaneous confidence intervals on multivariate non-inferiority. *Statistics In Medicine*, 32(10):1720–1729, May 2013. doi: 10.1002/sim.5633.

L. Hothorn. Statistics with interlaboratory in vitro toxicological studies. *ATLA- Alternative to Laboratory Animals*, 31:43–63, 2003.

L. A. Hothorn. Multiple comparisons and multiple contrasts in randomized dose-response trials-confidence interval oriented approaches. *Journal Of Biopharmaceutical Statistics*, 16(5):711–731, 2006.

L. A. Hothorn, M. Neuhauser, and H. F. Koch. Analysis of randomized dose-finding-studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal*, 39(4):467–479, 1997.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50 (3):346–363, 2008.

T. Jaki and L. A. Hothorn. Statistical evaluation of toxicological assays: Dunnett or williams test-take both. *Archives of Toxicology*, 87(11):1901–1910, Nov. 2013. doi: 10.1007/s00204-013-1065-x.

A. R. Jonckheere. A distribution-free kappa-sample test against ordered alternatives. *Biometrika*, 41(1-2):133–145, 1954.

B. Klingenberg and V. Satopaeae. Simultaneous confidence intervals for comparing margins of multivariate binary data. *Computational Statistics & Data Analysis*, 64:87–98, Aug. 2013. doi: 10.1016/j.csda.2013.02.016.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. CRC Press, 8 1989.

J. Najita, Y. Li, and P. Catalano. A novel application of a bivariate regression model for binary and continuous outcomes to studies of fetal toxicity. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 58:555–573, 2009.

J. Pinheiro, B. Bornkamp, E. Glimm, and F. Bretz. Model-based dose finding under model uncertainty using general parametric models. *Statistics In Medicine*, 33(10):1646–1661, May 2014. doi: 10.1002/sim.6052.

C. B. Pipper, C. Ritz, and H. Bisgaard. A versatile method for confirmatory evaluation of the effects of a covariate in multiple models. *Journal of the Royal Statistical Society Series C-applied Statistics*, 61:315–326, 2012. doi: 10.1111/j.1467-9876.2011.01005.x.

C. Ritz, R. P. Laursen, and C. T. Damsgaard. Simultaneous inference for multilevel linear mixed models - with an application to a large-scale school meal study. *Journal of the Royal Statistical Society Series C - Applied Statistics*, 66(2):295–311, 2017. doi: 10.1111/rssc.1216.

P. Rutgeerts, B. Feagan, and C. e. l. Marano. Randomised clinical trial: a placebo-controlled study of intravenous golimumab induction therapy for ulcerative colitis. *Alimentary Pharmacology and Therapeutics*, 42:504–514, 2015.

F. Schaarschmidt, D. Gerhard, and C. Vogel. Simultaneous confidence intervals for comparisons of several multinomial samples. *Computational Statistics and Data Analysis*, 106:65–76, 2017.

A. Swain, J. Turton, and C. e. a. Scudamore. Nephrotoxicity of hexachloro-1:3-butadiene in the male hanover wistar rat; correlation of minimal histopathological changes with biomarkers of renal injury. *Journal of Applied Toxicology*, 32:417–428, 2012.

A. P. Sykes, R. O'Connor-Semmes, R. Dobbins, D. J. Dorey, J. D. Lorimer, S. Walker, W. O. Wilkison, and L. Kler. Randomized trial showing efficacy and safety of twice-daily remogliflozin etabonate for the treatment of type 2 diabetes. *Diabetes Obesity & Metabolism*, 17(1):94–97, Jan. 2015. doi: 10.1111/dom.12391.

J. W. Tukey, J. L. Ciminera, and J. F. Heyse. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics*, 41(1):295–301, 1985. doi: 10.2307/2530666.

D. Verrier, S. Sivapregassam, and A.-C. Solente. Dose-finding studies, mcp-mod, model selection, and model averaging: two applications in the real world. *Clinical Trials*, 11:476–484, 2014.

P. H. Westfall. Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92(437):299–306, Mar. 1997. doi: 10.2307/2291474.

H. Wickham. Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20, 2007. URL `http://www.jstatsoft.org/v21/i12/`.

H. Wickham. The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1):1–29, 2011. URL `http://www.jstatsoft.org/v40/i01/`.

D. A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971.

L. D. Wise, D. A. Stoffregen, C.-M. Hoe, and G. R. Lankas. Juvenile toxicity assessment of open-acid lovastatin in rats. *Birth Defects Research Part B-developmental and Reproductive Toxicology*, 92(4):314–322, Aug. 2011. doi: 10.1002/bdrb.20296.

B. Wu and A. de Leon. Gaussion copula mixed models for clusterd mixed outcomes, with application in developmental toxicology. *Journal of Agricultural Biological and Environmental Statistics*, 19:39–56, 2014.

J. W. Yager, P. R. Gentry, R. S. Thomas, L. Pluta, A. Efremenko, M. Black, L. L. Arnold, J. M. McKim, P. Wilga, G. Gill, K.-Y. Choe, and H. J. Clewell. Evaluation of gene expression changes in human primary uroepithelial cells following 24-hr exposures to inorganic arsenic and its methylated metabolites. *Environmental and Molecular Mutagenesis*, 54(2):82–98, Mar. 2013. doi: 10.1002/em.21749.

T. W. Yee. *VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-3.* URL `https://CRAN.R-project.org/package=VGAM`.

A. Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9), Aug. 2006.