# Aufgabenblatt 5

# (Tabellenanalyse)

*UK Erweiterungen des linearen Modells*

**Cordula Eggerth**

Matrikelnummer: 00750881

Kursleiter:

Prof. Dr. Marcus Hudec &

Prof. Dr. Wilfried Grossmann

Sommersemester 2019

## Aufgabe 1 (Zweidimensionale Tabellen):

**Ein Hersteller von Büroartikel fertigt Aktenordner in den Farben gelb, rot und blau an. Zur Analyse des Kaufverhaltens werden für die drei wichtigsten Absatzmärkte A, B und C die Anzahlen der in einem bestimmten Zeitintervall georderten Bestelleinheiten ermittelt. Die Werte sind folgender Tabelle zu entnehmen.**

| Markt | Farbe der Ordner | | | Gesamt |
|---|---|---|---|---|
| | gelb | rot | blau | |
| A | 564 | 672 | 611 | 1847 |
| B | 309 | 198 | 307 | 814 |
| C | 448 | 299 | 425 | 1172 |
| Gesamt | 1321 | 1169 | 1343 | 3833 |

**a) Beantworte mit einem loglinearen Modell die Frage ob in den drei Märkten die Ordnerfarben unterschiedlich beliebt sind.**

**b) Stelle die Daten mit einem Mosaicplot dar.**

**c) Stelle den Zusammenhang mittels Korrespondenzanalyse dar und interpretiere den Plot.**

**Untersuche diese Frage mit einem loglinearen Modell.**

***Überblick über den Datensatz:***

```
> tabelle_ordner
     farbe
markt gelb rot blau
    A  564 672  611
    B  309 198  307
    C  448 299  425
```

***Deskriptive Statistiken:***

```
> summary(df_ordner$y_ordner)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  198.0   307.0   425.0   425.9   564.0   672.0
```

Die Randhäufigkeiten wurden absolut und relativ mittels der Funktion *addmargins()* hinzugefügt.

```
> addmargins(tabelle_ordner)
     farbe
markt gelb  rot blau  Sum
  A    564  672  611 1847
  B    309  198  307  814
  C    448  299  425 1172
  Sum 1321 1169 1343 3833
> addmargins(round(prop.table(tabelle_ordner)*100,
+                  digits=1))
     farbe
markt  gelb   rot  blau   Sum
  A    14.7  17.5  15.9  48.1
  B     8.1   5.2   8.0  21.3
  C    11.7   7.8  11.1  30.6
  Sum  34.5  30.5  35.0 100.0
```

Nach der Farbe sind die Bestellzahlen annähernd gleich aufgeteilt. Nach dem Markt ist die Aufteilung weniger gleichmäßig. In der Folge soll mit einem log-linearen Modell untersucht werden, ob ein Zusammenhang zwischen dem Markt und der Farbe besteht.

***Log-lineares Modell:***

Das hier verwendete log-lineare Modell legt die Betrachtung einer Poissonverteilung zugrunde. Im untenstehenden Output ist erkennbar, dass die Märkte und die Farbe Rot signifikant sind.

```
> summary(mod.ordner)

Call:
glm(formula = y_ordner ~ markt + farbe, family = poisson, data = df_ordner)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9
-2.9329  4.4432  -1.4345  1.6718  -3.3075  1.2744  2.1553  -3.1816  0.7044

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.45606    0.03221 200.432  < 2e-16 ***
marktB      -0.81936    0.04207 -19.476  < 2e-16 ***
marktC      -0.45485    0.03735 -12.180  < 2e-16 ***
farberot    -0.12224    0.04016  -3.044  0.00233 **
farbeblau    0.01652    0.03875   0.426  0.66994
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 500.324  on 8  degrees of freedom
Residual deviance:  61.024  on 4  degrees of freedom
AIC: 141.44

Number of Fisher Scoring iterations: 4
```

Mittels Chi-Quadrat-Test auf Markt und Farbe kann festgestellt werden, dass Markt und Farbe signifikant sind, sodass also ein Zusammenhang besteht.

```
> drop1(mod.ordner, test="Chi")
Single term deletions

Model:
y_ordner ~ markt + farbe
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>       61.02 141.44
markt   2   486.07 562.48 425.04 < 2.2e-16 ***
farbe   2    75.28 151.69  14.25 0.0008029 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Der p-value ist signifikant und der kritische Wert der Chi-Quadrat-Verteilung liegt bei ca. 5.99.

```
> 1 - pchisq(mod.ordner$deviance,2)
[1] 5.606626e-14

> qchisq(0.95,2)
[1] 5.991465
```

Zusätzlich kann man noch das saturierte Modell betrachten, in dem abgesehen von den bereits genannten signifikanten Bestandteilen auch die Interaktion von Markt B und Farbe Rot sowie von Markt C und Farbe Rot veranschaulicht werden.

```
> summary(mod.ordner.sat)

Call:
glm(formula = y_ordner ~ markt * farbe, family = poisson, data = df_ordner)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0  0

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       6.33505    0.04211 150.449  < 2e-16 ***
marktB           -0.60171    0.07078  -8.502  < 2e-16 ***
marktC           -0.23026    0.06329  -3.638 0.000274 ***
farberot          0.17520    0.05711   3.068 0.002155 **
farbeblau         0.08004    0.05839   1.371 0.170448
marktB:farberot  -0.62028    0.10746  -5.772 7.83e-09 ***
marktC:farberot  -0.57955    0.09401  -6.165 7.06e-10 ***
marktB:farbeblau -0.08654    0.09952  -0.870 0.384531
marktC:farbeblau -0.13275    0.08941  -1.485 0.137640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  5.0032e+02  on 8  degrees of freedom
Residual deviance: -1.0925e-13  on 0  degrees of freedom
AIC: 88.411

Number of Fisher Scoring iterations: 2
```

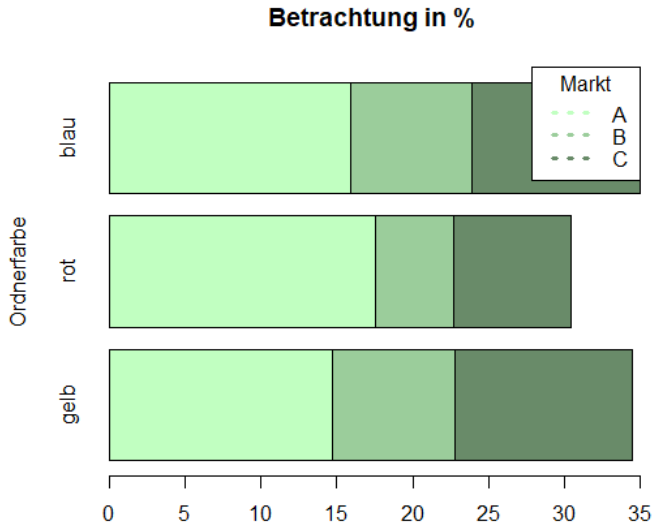
Der Chi-Quadrat-Test auf die Ordneranzahlen hat ein signifikantes Ergebnis.

```
> chisq.test(tab.mod.ordner) # ergebnis: chi-qu.-test signif.

        Pearson's Chi-squared test

data:  tab.mod.ordner
X-squared = 60.857, df = 4, p-value = 1.916e-12
```

***Visualisierung:***

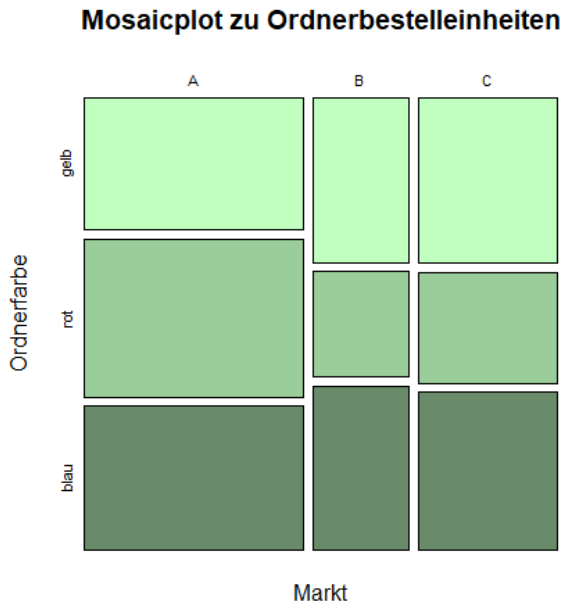
***Balkendiagramm:***

Die Visualisierung kann zunächst mit einem Balkendiagramm stattfinden. Die Ordnerfarben sind insgesamt in etwa ebenmäßig verteilt. Ungefähr 15% der Ordner sind z.B. blau und werden im Markt A bestellt. Hinsichtlich der Farbe-Markt-Kombination sind sich Markt A und Markt C laut Diagramm relativ ähnlich, wohingegen Markt B sich deutlicher unterscheidet.




***Mosaicplot:***

Dieselben Beobachtungen wie für das Balkendiagramm können auch im Mosaicplot in leicht variierter Darstellungsweise gemacht werden.



***Korrespondenzanalyse:***

Für die Korrespondenzanalyse müssen zunächst die Residuen in Matrixform berechnet werden.

```
> residual_matr
      markt
farbe           A          B          C
  gelb -2.8754623  1.6994315  2.1934620
  rot   4.5797688 -3.1896246 -3.0910746
  blau -1.4209891  1.2903801  0.7084683
```

Darauf basierend kann eine Singulärwertzerlegung vorgenommen werden, wobei die Singulärwerte (SV) insbesondere für die ersten zwei Komponenten betrachtet werden.

```
> svd_res
$`d`
[1] 7.781989e+00 5.451758e-01 1.508528e-15

$u
            [,1]       [,2]
[1,] -0.5116026 -0.6273947
[2,]  0.8197561 -0.1517134
[3,] -0.2574157  0.7637793

$v
            [,1]        [,2]
[1,]  0.7184759  0.04386751
[2,] -0.4904031  0.73968812
[3,] -0.4932516 -0.67151853
```

Danach werden die SV in u- und v-Komponenten zerlegt

```
> sv1
            [,1]       [,2]
[1,] -1.4271776 -0.4632433
[2,]  2.2868093 -0.1120191
[3,] -0.7180923  0.5639443
> sv2
           [,1]        [,2]
[1,]  2.004276  0.03239002
[2,] -1.368039  0.54615628
[3,] -1.375985 -0.49582257
```

In der Folge werden die Inertia berechnet. Danach wird ein Chi-Quadrat-Test durchgeführt, wobei dafür zuerst die Voraussetzung der erwarteten Häufigkeiten größer als fünf überprüft wird. Wie untenstehend ersichtlich, ist der Chi-Quadrat-Test für die Tabelle „Ordner“ signifikant. Es besteht demgemäß ein Zusammenhang zwischen Markt und Farbe der Ordner.

```
# inertia
```

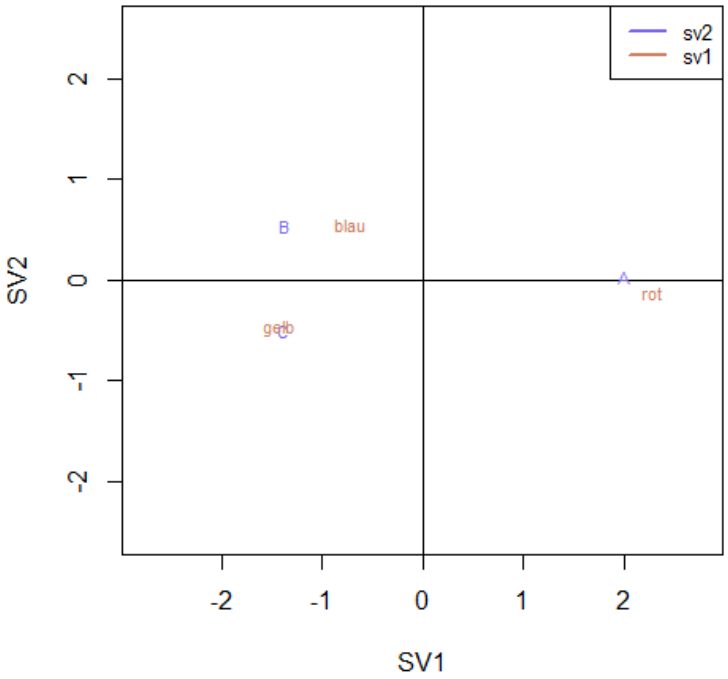
```
> inertia
[1] 60.87
```

```
> chisq.test(tabelle_ordner)

        Pearson's Chi-squared test

data:  tabelle_ordner
X-squared = 60.857, df = 4, p-value = 1.916e-12
```

Der gesamt Wert ergibt dasselbe wie die Chi-Quadrat-Teststatistik bzw. die Inertia, was eine korrekte Skalierung folgern lässt:

```
> gesamt # ergibt selbiges wie chisq teststat. bzw. inertia
          [,1]
[1,] 60.85657
```

Der Korrespondenzanalyseplot stellt nun die ersten beiden Singulärwerte (hier: SV1, SV2) in einem Koordinatensystem dar und belegt die Komponenten mit den entsprechenden Zeilen- bzw. Spaltenlabels. In der untenstehenden Grafik ist zu erkennen, dass Markt A und die Farbe Rot in Kombination häufiger vorkommen als bei Unabhängigkeit, da sie nahe zusammen liegen und weit weg vom Ursprung. Außerdem ist es ein Anzeichen, dass die Verteilung der Kombination nicht mit der Randverteilung übereinstimmt, wenn diese vergleichsweise weit weg vom Ursprung liegt. Der Markt B und die Farbe Blau liegen eigentlich nahe am Ursprung, wofür man interpretieren kann, dass die Randhäufigkeit in etwa dieser Häufigkeit entspricht. Die Kombination von Markt C und Farbe Rot kommt allerdings weniger häufig als bei Unabhängigkeit vor, da dieses Spalten- und Zeilenlabel diametral entgegengesetzt liegen.




```
> tt
            [,1]         [,2]
[1,]  2.0042760  0.03239002
[2,] -1.3680391  0.54615628
[3,] -1.3759854 -0.49582257
[4,] -1.4271776 -0.46324332
[5,]  2.2868093 -0.11201914
[6,] -0.7180923  0.56394429
```

## R-Code zu Aufgabe 1 (Zweidimensionale Tabellen):

```
#*******************************************************************************************
# AUFGABE 1 (Zweidimensionale Tabellen)
#*******************************************************************************************
# Ein Hersteller von Büroartikel fertigt Aktenordner in den Farben gelb, rot und
# blau an. Zur Analyse des Kaufverhaltens werden für die drei wichtigsten Absatzmärkte
# A, B und C die Anzahlen der in einem bestimmten Zeitintervall georderten
# Bestelleinheiten ermittelt.
# Die Werte sind folgender Tabelle zu entnehmen: siehe Angabe.

# a) Beantworte mit einem loglinearen Modell die Frage ob in den drei Märkten die
#    Ordnerfarben unterschiedlich beliebt sind.
# b) Stelle die Daten mit einem Mosaicplot dar.
# c) Stelle den Zusammenhang mittels Korrespondenzanalyse dar und interpretiere
#    den Plot.

# tabelle erstellen
y_ordner <- c(564,672,611, 309,198,307, 448,299,425)
markt <- gl(3,3,labels=c("A","B","C"))
farbe <- gl(3,1,9,labels=c("gelb","rot","blau"))
df_ordner <- data.frame(y_ordner, markt, farbe)
tabelle_ordner <- xtabs(y_ordner ~ markt + farbe)

# DESKRIPTIVE STATISTIKEN
summary(df_ordner$y_ordner)

# randhaeufigkeiten (absolut, relativ)
# absolut
addmargins(tabelle_ordner)
# relativ
addmargins(round(prop.table(tabelle_ordner)*100,
                 digits=1))

# LOG-LINEARES MODELL
# (annahme: basierend auf poissonverteilungsmodell)
# (haupteffekte)
mod.ordner <- glm(y_ordner ~ markt + farbe,
                  family=poisson, df_ordner)
summary(mod.ordner)

drop1(mod.ordner, test="Chi")

# p-value & chi-quadrat krit. wert
1 - pchisq(mod.ordner$deviance,2)
qchisq(0.95,2)

# saturiertes modell
mod.ordner.sat <- glm(y_ordner ~ markt * farbe,
                      family=poisson, df_ordner)
summary(mod.ordner.sat)

# chi-quadrat-test auf die ordneranzahlen
tab.mod.ordner <- matrix(y_ordner,nrow=3,byrow=TRUE)
chisq.test(tab.mod.ordner) # ergebnis: chi-qu.-test signif.

# BARCHART
colors <- c("darkseagreen1","darkseagreen3","darkseagreen4")
percentage <- prop.table(tabelle_ordner)*100
```

```
barplot(percentage, main="Betrachtung in %", ylab="Ordnerfarbe",
        col=colors, horiz=TRUE)
legend("topright", legend=rownames(percentage),
       title="Markt", col=colors, lwd=3, lty=3)

# MOSAICPLOT
mosaicplot(tabelle_ordner, color=colors,
           main="Mosaicplot zu Ordnerbestelleinheiten",
           xlab="Markt", ylab="Ordnerfarbe")

# KORRESPONDENZANALYSE
# residuenmatr.
residual_matr <- xtabs(residuals(mod.ordner, type="pearson") ~
                         farbe + markt, df_ordner)
residual_matr

# SVD (singular value decomposition)
# fuer erste 2 komponenten
svd_res <- svd(residual_matr,2,2)
svd_res

# zerlege in u- und v-komponenten
sv1 <- svd_res$u %*% diag(sqrt(svd_res$d[1:2]))
sv2 <- svd_res$v %*% diag(sqrt(svd_res$d[1:2]))

# inertia
inertia <- svd_res$d[1]^2+svd_res$d[2]^2
inertia

# vgl. der inertia mit chi-quadrat
# (vorauss.: erwartetete haeuf. groesser 5) >> hier erfuellt
res <- matrix(numeric(ncol(tabelle_ordner)*nrow(tabelle_ordner)),
              ncol=ncol(tabelle_ordner))
for(i in 1:nrow(tabelle_ordner)){
  for(j in 1:ncol(tabelle_ordner)){
    res[i,j] <- rowSums(tabelle_ordner)[i] *
      colSums(tabelle_ordner)[j]/sum(tabelle_ordner)
  }
}
all(res>5)

chisq.test(tabelle_ordner)
# ergebnis: chi-quadr.-test signifikant
# >> zshg. zwischen markt und ordnerfarbe

gesamt <- t(svd_res$d) %*% svd_res$d
gesamt # ergibt selbiges wie chisq teststat. bzw. inertia
       # anzeichen fuer korrekte skalierung

# korrespondenzanalyse plot
aa <-1.1 * max(abs(sv1),abs(sv2))
plot(rbind(sv1,sv2), asp=1,
     xlim=c(-aa,aa),ylim=c(-aa,aa),
     xlab="SV1",ylab="SV2",type="n")
abline(h=0,v=0)
text(sv2,c("A","B","C"), cex=0.7, col="mediumslateblue")
text(sv1,c("gelb","rot","blau"), cex=0.7, col="lightsalmon3")
legend("topright", cex=0.8, legend=c("sv2","sv1"),
       col=c("mediumslateblue","lightsalmon3"),
       lwd=2, lty=1)

tt <- rbind(sv2,sv1)
tt
```

## Aufgabe 2 (Zweidimensionale Tabellen):

**Die folgende Tabelle gibt die Beziehungen zwischen den Leistungen von Studenten in Mathematik und Statistik an.**

| | | Mathematik | | |
|---|---|---|---|---|
| | | Sehr gut | Durchschnitt | Schlecht |
| Statistik | Sehr gut | 56 | 71 | 12 |
| | Durchschnitt | 37 | 163 | 38 |
| | Schlecht | 24 | 42 | 85 |

**a) Beantworte mit einem loglinearen Modell die Frage ob es einen Zusammenhang zwischen den Noten in Mathematik und Statistik gibt.**

**b) Stelle die Daten mit einem Mosaicplot dar und mit geeigneten Prozentwerte in einem Barplot dar.**

**c) Stelle den Zusammenhang mittels Korrespondenzanalyse dar und interpretiere den Plot.**

**Überblick** über die aus den Daten (mittels `xtabs()`) erstellte Tabelle:

```
> tabelle_leistungen
             mathematik
statistik     Sehr gut Durchschnitt Schlecht
  Sehr gut          56          71       12
  Durchschnitt      37         163       38
  Schlecht          24          42       85
```

**Deskriptive Statistiken:**

```
> summary(df_leistungen$y_leistungen)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   37.00   42.00   58.67   71.00  163.00
```

Zur Tabelle wurden die **Randhäufigkeiten** hinzugefügt (`mittels addmargins()`):

Absolut:

```
> addmargins(tabelle_leistungen)
             mathematik
statistik     Sehr gut Durchschnitt Schlecht Sum
  Sehr gut          56          71       12 139
  Durchschnitt      37         163       38 238
  Schlecht          24          42       85 151
  Sum              117         276      135 528
```

Relativ (in Prozent):

```
> addmargins(round(prop.table(tabelle_leistungen)*100,
+                  digits=1))
             mathematik
statistik     Sehr gut Durchschnitt Schlecht   Sum
  Sehr gut        10.6         13.4      2.3  26.3
  Durchschnitt     7.0         30.9      7.2  45.1
  Schlecht         4.5          8.0     16.1  28.6
  Sum             22.1         52.3     25.6 100.0
```

**Log-lineares Modell:**

Hierfür wurde ein log-lineares Modell basierend auf dem Poissonverteilungsmodell erstellt. Die Haupteffekte wurden untersucht. Davon waren `statistikDurchschnitt` und `mathematikDurchschnitt` signifikant.

```
> summary(mod.leistungen)

Call:
glm(formula = y_leistungen ~ statistik + mathematik, family = poisson,
    data = df_leistungen)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9
 4.0689  -0.1954  -4.5849  -2.2912   3.3008  -3.1494  -1.7233  -4.5680   6.4326

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.42755    0.11767  29.127  < 2e-16 ***
statistikDurchschnitt  0.53780    0.10675   5.038 4.71e-07 ***
statistikSchlecht      0.08281    0.11754   0.704    0.481
mathematikDurchschnitt  0.85823    0.11032   7.780 7.28e-15 ***
mathematikSchlecht     0.14310    0.12631   1.133    0.257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 241.87  on 8  degrees of freedom
Residual deviance: 128.89  on 4  degrees of freedom
AIC: 189.95

Number of Fisher Scoring iterations: 5
```

Bei der Anwendung von `drop1()` stellte sich heraus, dass beide Variablen, `statistik` und `mathematik` signifikant sind.

```
> drop1(mod.leistungen, test="Chi")
Single term deletions

Model:
y_leistungen ~ statistik + mathematik
           Df Deviance    AIC    LRT  Pr(>Chi)
<none>          128.90 189.95
statistik   2   160.67 217.72 31.772 1.261e-07 ***
mathematik  2   210.10 267.15 81.203 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> # p-value
> 1 - pchisq(mod.leistungen$deviance,2)
[1] 0
```

```
> # krit. wert
> qchisq(0.95,2)
[1] 5.991465
```

Im Vergleich dazu kann man zusätzlich das saturierte Modell, das alle Prädiktoren enthält, betrachten:

```
> summary(mod.leistungen.sat)

Call:
glm(formula = y_leistungen ~ statistik * mathematik, family = poisson,
    data = df_leistungen)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0  0

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                4.0254     0.1336  30.123  < 2e-16 ***
statistikDurchschnitt                     -0.4144     0.2119  -1.956 0.050444 .
statistikSchlecht                         -0.8473     0.2440  -3.473 0.000515 ***
mathematikDurchschnitt                     0.2373     0.1787   1.328 0.184206
mathematikSchlecht                        -1.5404     0.3181  -4.843 1.28e-06 ***
statistikDurchschnitt:mathematikDurchschnitt  1.2455     0.2552   4.881 1.05e-06 ***
statistikSchlecht:mathematikDurchschnitt   0.3223     0.3121   1.033 0.301800
statistikDurchschnitt:mathematikSchlecht   1.5671     0.3931   3.986 6.71e-05 ***
statistikSchlecht:mathematikSchlecht       2.8050     0.3932   7.134 9.78e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.4187e+02  on 8  degrees of freedom
Residual deviance: 7.1054e-15  on 0  degrees of freedom
AIC: 69.056

Number of Fisher Scoring iterations: 3
```

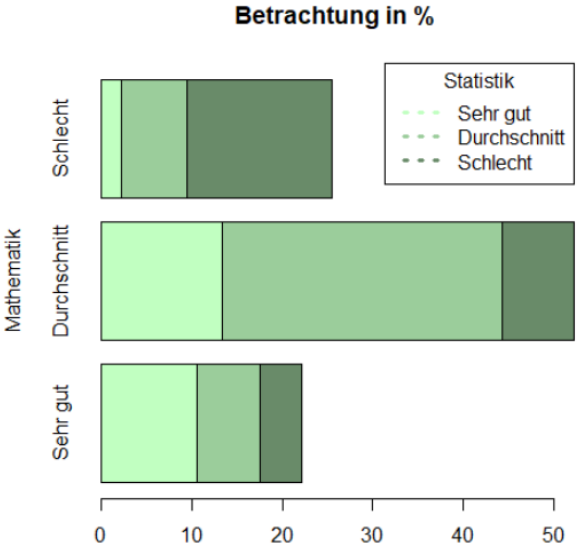
Mit Hilfe des Chi-Quadrat-Tests zeigt sich, dass das Ergebnis signifikant ist:

```
> chisq.test(tab.mod)

        Pearson's Chi-squared test

data:  tab.mod
X-squared = 137.19, df = 4, p-value < 2.2e-16
```

Grafisch kann man die gegebenen Daten beispielsweise mit einem Balkendiagramm oder einem Mosaicplot modellieren, wie es untenstehend gezeigt wird.

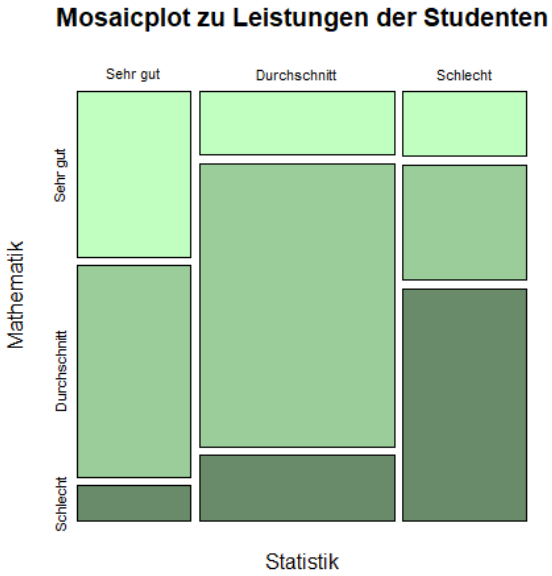
**Balkendiagramm:**

Es zeigte sich, dass von den in Mathematik sehr guten Studenten auch die Mehrheit sehr gut bis durchschnittlich in Statistik waren. Die meisten in Mathematik durchschnittlichen Studenten waren durchschnittlich in Statistik. Die meisten in Mathematik schlechten Studenten waren auch schlecht in Statistik. Nur sehr wenige von den in Mathematik schlechten Studenten waren sehr gut in Statistik.




**Mosaicplot:**

Die Darstellung im Mosaicplot zeigt ebenfalls, dass viele von den in Mathematik durchschnittlichen Studenten auch durchschnittlich in Statistik waren. Die meisten in Statistik schlechten Studenten waren auch in Mathematik schlecht. Die meisten sehr guten Studenten in Mathematik waren auch sehr gut in Statistik.



**Korrespondenzanalyse:**

*Residuenmatrix:*

```
> residual_matr
             statistik
mathematik       Sehr gut Durchschnitt   Schlecht
  Sehr gut      4.5404363   -2.1672183 -1.6354495
  Durchschnitt -0.1946369    3.4598634 -4.1569484
  Schlecht     -3.9486164   -2.9294831  7.4662946
```

*SVD (Singular Value Decomposition) für die ersten zwei Komponenten:*

Diese Zerlegung weist die u- und v-Komponenten aus, die dann in ihren jeweiligen Teilen in einem neuen Koordinatensystem dargestellt werden können.

```
> svd_res
$`d`
[1] 1.036576e+01 5.453972e+00 2.531153e-14

$u
           [,1]        [,2]
[1,] -0.2493875  0.84629484
[2,] -0.4407094 -0.53202248
[3,]  0.8623115 -0.02715022

$v
           [,1]       [,2]
[1,] -0.4294413  0.7431841
[2,] -0.3386576 -0.6592067
[3,]  0.8371925  0.1145598
```

*Trennung der beiden ersten Singulärwerte (SV):*

```
> sv1
           [,1]        [,2]
[1,] -0.8029257  1.97641493
[2,] -1.4189039 -1.24247144
[3,]  2.7762898 -0.06340592
> sv2
          [,1]      [,2]
[1,] -1.382625  1.735613
[2,] -1.090339 -1.539494
[3,]  2.695417  0.267540
```

*Inertia (i.e. quadrierter Abstand der beiden ersten SV vom Koordinatenursprung):*

```
> inertia
[1] 137.1948
```

*Vergleich der Inertia mit Chi-Quadrat-Test:*

(wobei die Voraussetzung, dass die erwarteten Häufigkeiten größer als fünf sein sollten für jede Zelle, hier erfüllt ist)

```
> chisq.test(tabelle_leistungen)

        Pearson's Chi-squared test

data:  tabelle_leistungen
X-squared = 137.19, df = 4, p-value < 2.2e-16
```

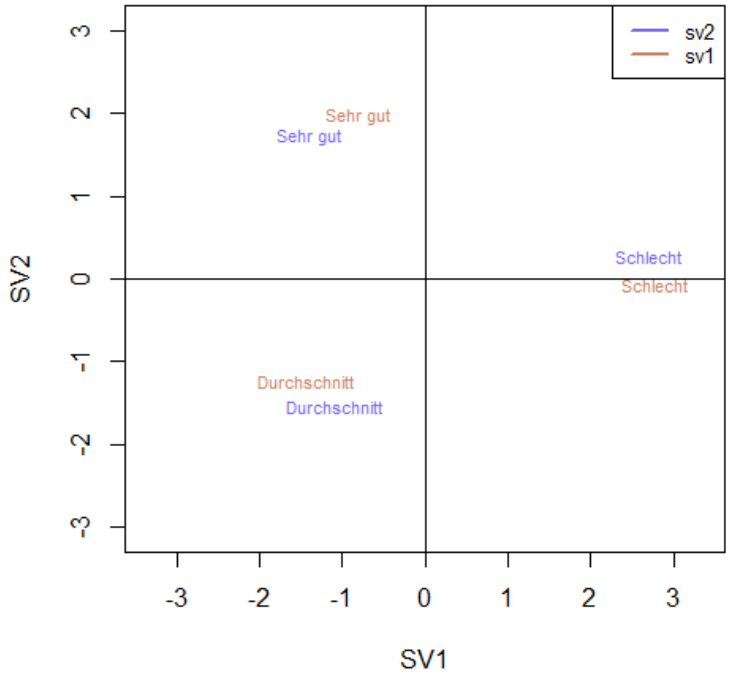
Das Ergebnis des Chi-Quadrat-Tests ist signifikant. Demnach gibt es einen Zusammenhang zwischen den Leistungen der Studenten in Mathematik und Statistik.

Die Zahl gesamt ergibt außerdem dasselbe wie die Chi-Quadrat-Teststatistik und die Inertia, und lässt somit auf eine korrekte Skalierung schließen.

```
> gesamt # ergibt selbiges wie chisq teststat. bzw. inertia
>        # d.h. korrekte skalierung
          [,1]
[1,] 137.1948
```

***Korrespondenzanalyseplot:***

Die jeweiligen u- und v-Komponenten der Singulärwerte SV1 und SV2 wurden im Koordinatensystem eingetragen. Im Plot ist ersichtlich, dass die z.B. die Kombination „schlecht“-„schlecht“ in Statistik und Mathematik häufiger vorkommt als bei Unabhängigkeit. Ebenso kommen die Kombination „sehr gut“-„sehr gut“ und „durschnitt“-„durchschnitt“ häufiger vor als bei Unabhängigkeit. Die Verteilungen der jeweiligen drei Kombinationen liegen nicht so nahe beim Ursprung, aber auch nicht sehr weit weg. Deshalb kann vermutet werden, dass die Verteilungen der Kombinationen nicht genau übereinstimmen mit den entsprechenden Randverteilungen. Die Kombinationen „sehr gut in Statistik“-„schlecht in Mathematik“ oder „sehr gut in Mathematik“-„schlecht in Statistik“ liegen z.B. diametral entgegengesetzt und kommen somit weniger häufig vor als bei Vorliegen von Unabhängigkeit.




```
> tt <- rbind(sv2,sv1)
> tt
           [,1]        [,2]
[1,] -1.3826252  1.73561274
[2,] -1.0903387 -1.53949417
[3,]  2.6954169  0.26754002
[4,] -0.8029257  1.97641493
[5,] -1.4189039 -1.24247144
[6,]  2.7762898 -0.06340592
```

## R-Code zu Aufgabe 2 (Zweidimensionale Tabellen):

```
#*******************************************************************************
# AUFGABE 2 (Zweidimensionale Tabellen)
#*******************************************************************************
# Die folgende Tabelle gibt die Beziehungen zwischen den Leistungen von Studenten in
# Mathematik und Statistik an.
# Tabelle: siehe Angabe.

# a) Beantworte mit einem loglinearen Modell die Frage ob es einen Zusammenhang
#    zwischen den Noten in Mathematik und Statistik gibt.
# b) Stelle die Daten mit einem Mosaicplot dar und mit geeigneten Prozentwerte in
#    einem Barplot dar.
# c) Stelle den Zusammenhang mittels Korrespondenzanalyse dar und interpretiere
#    den Plot.

# tabelle erstellen
y_leistungen <- c(56,71,12, 37,163,38, 24,42,85)
labels_leistungen <- c("Sehr gut", "Durchschnitt", "Schlecht")
statistik <- gl(3,3,labels=labels_leistungen)
mathematik <- gl(3,1,9,labels=labels_leistungen)
df_leistungen <- data.frame(y_leistungen, mathematik, statistik)
tabelle_leistungen <- xtabs(y_leistungen ~ statistik + mathematik)

# DESKRIPTIVE STATISTIKEN
summary(df_leistungen$y_leistungen)


# randhaeufigkeiten (absolut, relativ)
  # absolut
addmargins(tabelle_leistungen)
  # relativ
addmargins(round(prop.table(tabelle_leistungen)*100,
                 digits=1))

# LOG-LINEARES MODELL
# (annahme: basierend auf poissonverteilungsmodell)
# (haupteffekte)
mod.leistungen <- glm(y_leistungen ~ statistik + mathematik,
                      family=poisson, df_leistungen)
summary(mod.leistungen)

drop1(mod.leistungen, test="Chi")

  # p-value
1 - pchisq(mod.leistungen$deviance,2)

  # krit. wert
qchisq(0.95,2)

# ergebnis: zshg. zwischen leistungen in math. und stat.,
#           weil p value signifikant

# saturiertes modell
mod.leistungen.sat <- glm(y_leistungen ~ statistik * mathematik,
                      family=poisson, df_leistungen)
summary(mod.leistungen.sat)

# analyse mit hilfe von chi-quadrat-test
tab.mod <- matrix(y_leistungen,nrow=3,byrow=TRUE)
chisq.test(tab.mod)

# ergebnis: chi-qu.-test ist signifikant
```

```
# BARCHART
colors <- c("darkseagreen1","darkseagreen3","darkseagreen4")
percentage <- prop.table(tabelle_leistungen)*100

barplot(percentage, main="Betrachtung in %", ylab="Mathematik",
        col=colors, horiz=TRUE)
legend("topright", legend=rownames(percentage),
       title="Statistik", col=colors, lwd=3, lty=3)


# MOSAICPLOT
mosaicplot(tabelle_leistungen, color=colors,
           main="Mosaicplot zu Leistungen der Studenten",
           xlab="Statistik", ylab="Mathematik")

# ergebnis: wenn laenge und breite des plots in vierecken
#           gleich lang, dann sind die anteile ca. gleich


# KORRESPONDENZANALYSE
# residuenmatr.
residual_matr <- xtabs(residuals(mod.leistungen, type="pearson") ~
                 mathematik + statistik, df_leistungen)
residual_matr

# SVD (singular value decomposition)
# fuer erste 2 komponenten
svd_res <- svd(residual_matr,2,2)
svd_res

# zerlege in u- und v-komponenten
sv1 <- svd_res$u %*% diag(sqrt(svd_res$d[1:2]))
sv2 <- svd_res$v %*% diag(sqrt(svd_res$d[1:2]))


# inertia
inertia <- svd_res$d[1]^2+svd_res$d[2]^2
inertia

# vgl. der inertia mit chi-quadrat
# (vorauss.: erwartetete haeuf. groesser 5) >> hier erfuellt
res <- matrix(numeric(ncol(tabelle_leistungen)*nrow(tabelle_leistungen)),
                      ncol=ncol(tabelle_leistungen))
for(i in 1:nrow(tabelle_leistungen)){
  for(j in 1:ncol(tabelle_leistungen)){
    res[i,j] <- rowSums(tabelle_leistungen)[i] *
                colSums(tabelle_leistungen)[j]/sum(tabelle_leistungen)
  }
}
all(res>5)

chisq.test(tabelle_leistungen)
# ergebnis: chi-quadr.-test signifikant
# >> zshg. zwischen math.- und stat.-leistungen

gesamt <- t(svd_res$d) %*% svd_res$d
gesamt # ergibt selbiges wie chisq teststat. bzw. inertia
       # d.h. korrekte skalierung
```

```
# korrespondenzanalyse plot
aa <-1.1 * max(abs(sv1),abs(sv2))
plot(rbind(sv1,sv2), asp=1,
     xlim=c(-aa,aa),ylim=c(-aa,aa),
     xlab="SV1",ylab="SV2",type="n")
abline(h=0,v=0)
text(sv2,labels_leistungen, cex=0.7, col="mediumslateblue")
text(sv1,labels_leistungen, cex=0.7, col="lightsalmon3")
legend("topright", cex=0.8, legend=c("sv2","sv1"),
       col=c("mediumslateblue","lightsalmon3"),
       lwd=2, lty=1)

tt <- rbind(sv2,sv1)
tt
```

## Aufgabe 3 (Zweidimensionale Tabellen):

**In einem Experiment sollen zwei Testmethoden verglichen werden. Jeder Test soll dabei von je 10 Personen durchgeführt werden. Die Ergebnisse des Tests sind dabei nur bestanden und nicht bestanden. Man untersuche die folgenden vier Testergebnisse:**

| Ergebnis 1 | | |
|---|---|---|
| | bestanden | nicht bestanden |
| Test 1 | 2 | 8 |
| Test 2 | 8 | 2 |

| Ergebnis 2 | | |
|---|---|---|
| | bestanden | nicht bestanden |
| Test 1 | 2 | 8 |
| Test 2 | 3 | 7 |

| Ergebnis 3 | | |
|---|---|---|
| | bestanden | nicht bestanden |
| Test 1 | 2 | 8 |
| Test 2 | 4 | 6 |

| Ergebnis 4 | | |
|---|---|---|
| | bestanden | nicht bestanden |
| Test 1 | 3 | 7 |
| Test 2 | 7 | 3 |

**Bei welchen Ergebnissen liefert der $\chi^2$-Test signifikante Ergebnisse?**
**Wie ändern sich die Ergebnisse, wenn man anstelle von 10 Personen je Test jeweils 20 oder 40 Personen die Tests durchführen lässt, also die Werte in den Tabellen mit dem Faktor 2 bzw. 4 multipliziert? Begründe die Änderungen.**

***Chi-Quadrat-Tests für 10-Personen-pro-Test-Setting:***

```
> tabelle_ergebnisse1
       bestehen
tests   bestanden nicht bestanden
  Test1         2               8
  Test2         8               2


> chisq.test(tab.ergebnisse1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse1
X-squared = 5, df = 1, p-value = 0.02535


> tabelle_ergebnisse2
       bestehen
tests   bestanden nicht bestanden
  Test1         2               8
  Test2         3               7


> chisq.test(tab.ergebnisse2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse2
X-squared = 0, df = 1, p-value = 1

Warning message:
In chisq.test(tab.ergebnisse2) :
  Chi-Quadrat-Approximation kann inkorrekt sein
```

```
> tabelle_ergebnisse3
       bestehen
tests   bestanden nicht bestanden
  Test1         2               8
  Test2         4               6

> chisq.test(tab.ergebnisse3)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse3
X-squared = 0.2381, df = 1, p-value = 0.6256

Warning message:
In chisq.test(tab.ergebnisse3) :
  Chi-Quadrat-Approximation kann inkorrekt sein

> tabelle_ergebnisse4
       bestehen
tests   bestanden nicht bestanden
  Test1         3               7
  Test2         7               3


> chisq.test(tab.ergebnisse4)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse4
X-squared = 1.8, df = 1, p-value = 0.1797
```

***Ergebnisinterpretation*:**

Von den Chi-Quadrat-Tests auf die Ergebnistabellen ist nur jener signifikant auf dem 0.05 Level, der für `tabelle_ergebnisse1` durchgeführt wurde. Außerdem ist die Voraussetzung, dass mindestens 5 Beobachtungen pro Zelle in jeder Tabelle sein sollten, in diesen Betrachtungen nicht erfüllt. Aufgrund der kleinen Stichproben ist in diesem Setting die Chi-Quadrat-Verteilungsapproximation sehr ungenau, und es sollte eher darauf verzichtet werden, und es sollte daher die exakte Verteilung berücksichtigt werden (z.B. mit Fisher's Exact Test).

***Chi-Quadrat-Tests für 20-Personen-pro-Test-Setting:***

Annahme: gleichbleibende relative Aufteilung der Gruppen, i.e. nur Multiplikation mit Faktor 2

Der Test auf **Ergebnis 1** ist signifikant auf dem 0.05 Level, aber die Voraussetzung der Mindesthäufigkeit pro Zelle ist nicht erfüllt:

```
> tabelle_ergebnisse1.pers20
       bestehen
tests   bestanden nicht bestanden
  Test1         4              16
  Test2        16               4

> tab.ergebnisse1.pers20 <- matrix(y_ergebnisse1.pers20,nrow=2,byrow=TRUE)
> chisq.test(tab.ergebnisse1.pers20)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse1.pers20
```

```
X-squared = 12.1, df = 1, p-value = 0.0005042
```

Der Test auf **Ergebnis 2** ist nicht signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist nicht erfüllt:

```
> tabelle_ergebnisse2.pers20
       bestehen
tests   bestanden nicht bestanden
  Test1         4              16
  Test2         6              14

> chisq.test(tab.ergebnisse2.pers20)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse2.pers20
X-squared = 0.13333, df = 1, p-value = 0.715
```

Der Test auf **Ergebnis 3** ist nicht signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist nicht erfüllt:

```
> tabelle_ergebnisse3.pers20
       bestehen
tests   bestanden nicht bestanden
  Test1         4              16
  Test2         8              12

> chisq.test(tab.ergebnisse3.pers20)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse3.pers20
X-squared = 1.0714, df = 1, p-value = 0.3006
```

Der Test auf **Ergebnis 4** ist signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist erfüllt:

```
> tabelle_ergebnisse4.pers20
       bestehen
tests   bestanden nicht bestanden
  Test1         6              14
  Test2        14               6

> chisq.test(tab.ergebnisse4.pers20)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse4.pers20
X-squared = 4.9, df = 1, p-value = 0.02686
```

***Chi-Quadrat-Tests für 40-Personen-pro-Test-Setting:***

Annahme: gleichbleibende relative Aufteilung der Gruppen, i.e. nur Multiplikation mit Faktor 4

Der Test auf **Ergebnis 1** ist signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist erfüllt:

```
> tabelle_ergebnisse1.pers40
       bestehen
tests   bestanden nicht bestanden
  Test1         8              32
  Test2        32               8
```

```
> chisq.test(tab.ergebnisse1.pers40)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse1.pers40
X-squared = 26.45, df = 1, p-value = 2.704e-07
```

Der Test auf **Ergebnis 2** ist nicht signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist erfüllt:

```
> tabelle_ergebnisse2.pers40
       bestehen
tests   bestanden nicht bestanden
  Test1         8              32
  Test2        12              28
> chisq.test(tab.ergebnisse2.pers40)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse2.pers40
X-squared = 0.6, df = 1, p-value = 0.4386
```

Der Test auf **Ergebnis 3** ist nicht signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist erfüllt:

```
> tabelle_ergebnisse3.pers40
       bestehen
tests   bestanden nicht bestanden
  Test1         8              32
  Test2        16              24
> chisq.test(tab.ergebnisse3.pers40)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse3.pers40
X-squared = 2.9167, df = 1, p-value = 0.08767
```

Der Test auf **Ergebnis 4** ist signifikant auf dem 0.05 Level und die Voraussetzung der Mindesthäufigkeit pro Zelle ist erfüllt:

```
> tabelle_ergebnisse4.pers40
       bestehen
tests   bestanden nicht bestanden
  Test1        12              28
  Test2        28              12
> chisq.test(tab.ergebnisse4.pers40)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.ergebnisse4.pers40
X-squared = 11.25, df = 1, p-value = 0.0007962
```

***Begründung:***

Die Ergebnisse ändern sich, da in den verschiedenen Settings unterschiedliche Stichprobengrößen verwendet werden. Da der Chi-Quadrat-Test eine Approximation an die tatsächliche Verteilung macht, kann es daher sein, dass bei kleinen Stichproben eine schlechte (bzw. sehr ungenaue) Approximation stattfindet, und das Resultat des Chi-Quadrat-Tests

daher nicht mehr sehr aussagekräftig ist. Außerdem ist bei kleinen Stichprobenzahlen, wie oben angesprochen, oftmals in diesem Beispiel die Voraussetzung[1] hinsichtlich der absoluten Häufigkeiten pro Zelle nicht erfüllt. Man sollte deshalb in solchen Situationen lieber exakte Tests verwenden, die keine Approximation machen, sondern die exakte Verteilung und die exakte Abweichung von der H0 berücksichtigen, wie zum Beispiel Fisher's Exact Test.

[1] Anmerkung: Voraussetzung ist eine absolute Häufigkeit pro Zelle von über 5 für die gesamte betrachtete Tabelle.

## R-Code zu Aufgabe 3 (Zweidimensionale Tabellen):

```
#*****************************************************************************************
# AUFGABE 3 (Zweidimensionale Tabellen)
#*****************************************************************************************
# In einem Experiment sollen zwei Testmethoden verglichen werden. Jeder Test soll dabei
# von je 10 Personen durchgeführt werden. Die Ergebnisse des Tests sind dabei nur bestanden
# und nicht bestanden. Man untersuche die folgenden vier Testergebnisse:
# Tabellen: siehe Angabe.

# Bei welchen Ergebnissen liefert der Chi-Quadrat-Test signifikante Ergebnisse?
# Wie ändern sich die Ergebnisse, wenn man anstelle von 10 Personen je Test jeweils 20
# oder 40 Personen die Tests durchführen lässt, also die Werte in den Tabellen mit dem
# Faktor 2 bzw. 4 multipliziert? Begründe die Änderungen.


# daten >> tabellen erstellen
  # factor levels
tests <- gl(2,2,labels=c("Test1","Test2"))
bestehen <- gl(2,1,4,labels=c("bestanden","nicht bestanden"))

  # ERGEBNIS 1
y_ergebnisse1 <- c(2,8, 8,2)
df_ergebnisse1 <- data.frame(y_ergebnisse1, tests, bestehen)
tabelle_ergebnisse1 <- xtabs(y_ergebnisse1 ~ tests + bestehen)

 # ERGEBNIS 2
y_ergebnisse2 <- c(2,8, 3,7)
df_ergebnisse2 <- data.frame(y_ergebnisse2, tests, bestehen)
tabelle_ergebnisse2 <- xtabs(y_ergebnisse2 ~ tests + bestehen)


 # ERGEBNIS 3
y_ergebnisse3 <- c(2,8, 4,6)
df_ergebnisse3 <- data.frame(y_ergebnisse3, tests, bestehen)
tabelle_ergebnisse3 <- xtabs(y_ergebnisse3 ~ tests + bestehen)

 # ERGEBNIS 4
y_ergebnisse4 <- c(3,7, 7,3)
df_ergebnisse4 <- data.frame(y_ergebnisse4, tests, bestehen)
tabelle_ergebnisse4 <- xtabs(y_ergebnisse4 ~ tests + bestehen)

# chi-quadrat-tests fuer 10-personen-pro-test-setting
tab.ergebnisse1 <- matrix(y_ergebnisse1,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse1)

tab.ergebnisse2 <- matrix(y_ergebnisse2,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse2)

tab.ergebnisse3 <- matrix(y_ergebnisse3,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse3)

tab.ergebnisse4 <- matrix(y_ergebnisse4,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse4)
  # ergebnisinterpretation:
  # - keiner der tests ist signifikant auf 0.05 level
  # - ausserdem ist die voraussetzung, dass mind. 5
  #   beobachtungen pro zelle in jeder tabelle sind,
  #   nicht erfuellt
```

```
# chi-quadrat-tests fuer 20-personen-pro-test-setting
# annahme: gleichbleibende relative aufteilung der gruppen

 # ERGEBNIS 1
y_ergebnisse1.pers20 <- c(2,8, 8,2)*2
df_ergebnisse1.pers20 <- data.frame(y_ergebnisse1.pers20, tests, bestehen)
tabelle_ergebnisse1.pers20 <- xtabs(y_ergebnisse1.pers20 ~ tests + bestehen)

tab.ergebnisse1.pers20 <- matrix(y_ergebnisse1.pers20,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse1.pers20) # signifikant
                                   # vorauss. nicht erfuellt

 # ERGEBNIS 2
y_ergebnisse2.pers20 <- c(2,8, 3,7)*2
df_ergebnisse2.pers20 <- data.frame(y_ergebnisse2.pers20, tests, bestehen)
tabelle_ergebnisse2.pers20 <- xtabs(y_ergebnisse2.pers20 ~ tests + bestehen)

tab.ergebnisse2.pers20 <- matrix(y_ergebnisse2.pers20,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse2.pers20) # nicht signifikant
                                   # vorauss. nicht erfuellt


 # ERGEBNIS 3
y_ergebnisse3.pers20 <- c(2,8, 4,6)*2
df_ergebnisse3.pers20 <- data.frame(y_ergebnisse3.pers20, tests, bestehen)
tabelle_ergebnisse3.pers20 <- xtabs(y_ergebnisse3.pers20 ~ tests + bestehen)

tab.ergebnisse3.pers20 <- matrix(y_ergebnisse3.pers20,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse3.pers20) # nicht signifikant
                                   # vorauss. nicht erfuellt

 # ERGEBNIS 4
y_ergebnisse4.pers20 <- c(3,7, 7,3)*2
df_ergebnisse4.pers20 <- data.frame(y_ergebnisse4.pers20, tests, bestehen)
tabelle_ergebnisse4.pers20 <- xtabs(y_ergebnisse4.pers20 ~ tests + bestehen)

tab.ergebnisse4.pers20 <- matrix(y_ergebnisse4.pers20,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse4.pers20) # signifikant
                                   # vorauss. erfuellt


# chi-quadrat-tests fuer 40-personen-pro-test-setting
# annahme: gleichbleibende relative aufteilung der gruppen

# ERGEBNIS 1
y_ergebnisse1.pers40 <- c(2,8, 8,2)*4
df_ergebnisse1.pers40 <- data.frame(y_ergebnisse1.pers40, tests, bestehen)
tabelle_ergebnisse1.pers40 <- xtabs(y_ergebnisse1.pers40 ~ tests + bestehen)

tab.ergebnisse1.pers40 <- matrix(y_ergebnisse1.pers40,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse1.pers40) # signifikant
                                   # vorauss. erfuellt

# ERGEBNIS 2
y_ergebnisse2.pers40 <- c(2,8, 3,7)*4
df_ergebnisse2.pers40 <- data.frame(y_ergebnisse2.pers40, tests, bestehen)
tabelle_ergebnisse2.pers40 <- xtabs(y_ergebnisse2.pers40 ~ tests + bestehen)

tab.ergebnisse2.pers40 <- matrix(y_ergebnisse2.pers40,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse2.pers40) # nicht signifikant
                                   # vorauss. erfuellt
```

```
# ERGEBNIS 3
y_ergebnisse3.pers40 <- c(2,8, 4,6)*4
df_ergebnisse3.pers40 <- data.frame(y_ergebnisse3.pers40, tests, bestehen)
tabelle_ergebnisse3.pers40 <- xtabs(y_ergebnisse3.pers40 ~ tests + bestehen)

tab.ergebnisse3.pers40 <- matrix(y_ergebnisse3.pers40,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse3.pers40) # nicht signifikant
                                   # vorauss. erfuellt

# ERGEBNIS 4
y_ergebnisse4.pers40 <- c(3,7, 7,3)*4
df_ergebnisse4.pers40 <- data.frame(y_ergebnisse4.pers40, tests, bestehen)
tabelle_ergebnisse4.pers40 <- xtabs(y_ergebnisse4.pers40 ~ tests + bestehen)

tab.ergebnisse4.pers40 <- matrix(y_ergebnisse4.pers40,nrow=2,byrow=TRUE)
chisq.test(tab.ergebnisse4.pers40) # signifikant
                                   # vorauss. erfuellt
```

## Aufgabe 4 (Dreidimensionale Tabellen):

**Die Daten femsmoke in der Library faraway zeigen die Ergebnisse einer Studie über das Rauchen bei Frauen in den Jahren 1972 – 1974. Die Variable y gibt die Anzahl der Fälle in den Gruppen an, die durch Raucher, Tot und Altersgruppe gebildet werden. Die kleinen Fallzahlen in manchen Altersgruppen ergeben sich dadurch, dass Personen im Laufe der Untersuchung ausgeschieden werden mussten. Beachte bei der Modellierung, dass dieser Datensatz ein Beispiel für Simpsons Paradoxon ist. In den einzelnen Altersgruppen sind die Ergebnisse anders als das Gesamtergebnis über alle Altersgruppen.**

***Daten und Deskriptives:***

Überblick über den Datensatz `femsmoke`:

```
> head(femsmoke, n=5)
   y smoker dead   age
1  2    yes  yes 18-24
2  1     no  yes 18-24
3  3    yes  yes 25-34
4  5     no  yes 25-34
5 14    yes  yes 35-44
```

Deskriptives zum Datensatz:

```
> summary(femsmoke)
smoker    dead       age
yes:14   yes:14   18-24:4
no :14   no :14   25-34:4
                 35-44:4
                 45-54:4
                 55-64:4
                 65-74:4
                 75+  :4
```

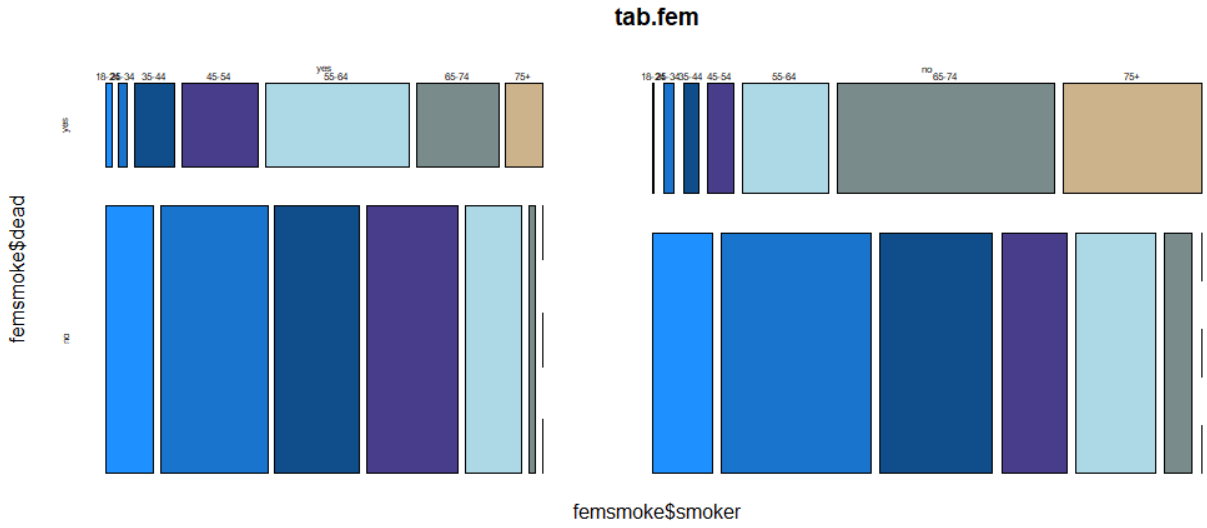
Erstellte Tabelle:

```
> tab.fem
, , femsmoke$age = 18-24

               femsmoke$dead
femsmoke$smoker yes  no
            yes   2  53
            no    1  61

, , femsmoke$age = 25-34

               femsmoke$dead
femsmoke$smoker yes  no
            yes   3 121
            no    5 152

, , femsmoke$age = 35-44

               femsmoke$dead
femsmoke$smoker yes  no
            yes  14  95
            no    7 114
```

```
, , femsmoke$age = 45-54

                femsmoke$dead
femsmoke$smoker yes  no
            yes  27 103
            no   12  66

, , femsmoke$age = 55-64

                femsmoke$dead
femsmoke$smoker yes  no
            yes  51  64
            no   40  81

, , femsmoke$age = 65-74

                femsmoke$dead
femsmoke$smoker yes  no
            yes  29   7
            no  101  28

, , femsmoke$age = 75+

                femsmoke$dead
femsmoke$smoker yes  no
            yes  13   0
            no   64   0
```

***Mosaicplot:***

Der Mosaicplot zeigt die Verhältnisse zwischen den Variablen `dead` und `smoker`, aufgeschlüsselt nach Altersgruppen. Die Anteile von Todesfällen sind bei jüngeren Altersgruppen unter Rauchern höher als bei Nichtrauchern. Die Anteile von Todesfällen in den jüngsten zwei Altersgruppen sind eher ähnlich.



***Log-lineares Modell / Modelle prüfen:***

Zusammenhang `dead` – `smoker`:

Der Zusammenhang zwischen den Variablen `dead` und `smoker` ist gemäß p-Wert signifikant.

```
> summary(mod1.fem)
Call: xtabs(formula = y ~ smoker + dead, data = femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
        Chisq = 9.121, df = 1, p-value = 0.002527

> round(prop.table(mod1.fem,1), digits=3)
      dead
smoker   yes    no
   yes 0.239 0.761
   no  0.314 0.686
```

Zusammenhang `age` – `smoker`:

Der Zusammenhang zwischen den Variablen `age` und `smoker` ist gemäß p-Wert signifikant.

```
> summary(mod2.fem) # signifikanter zshg.
Call: xtabs(formula = y ~ age + smoker, data = femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
        Chisq = 88.3, df = 6, p-value = 6.837e-17

> round(prop.table(mod2.fem,1), digits=3)
       smoker
age       yes    no
  18-24 0.470 0.530
  25-34 0.441 0.559
  35-44 0.474 0.526
  45-54 0.625 0.375
  55-64 0.487 0.513
  65-74 0.218 0.782
  75+   0.169 0.831
```

Zusammenhang `age` – `dead`:

Der Zusammenhang zwischen den Variablen `age` und `dead` ist gemäß p-Wert signifikant.

```
> summary(mod3.fem)
Call: xtabs(formula = y ~ age + dead, data = femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
        Chisq = 596.3, df = 6, p-value = 1.483e-125

> round(prop.table(mod3.fem,1), digits=3)
       dead
age       yes    no
  18-24 0.026 0.974
  25-34 0.028 0.972
  35-44 0.091 0.909
  45-54 0.188 0.812
```

```
  55-64 0.386 0.614
  65-74 0.788 0.212
  75+   1.000 0.000
```

Zusammenhang `age - dead – smoker`:

Der Zusammenhang zwischen den Variablen `age`, `dead` und `smoker` ist gemäß p-Wert signifikant.

```
> summary(mod4.fem)
Call: xtabs(formula = y ~ age + dead + smoker, data = femsmoke)
Number of cases in table: 1314
Number of factors: 3
Test for independence of all factors:
        Chisq = 790.6, df = 19, p-value = 2.14e-155
```

***Modell "Totale Unabhängigkeit":***

Die überprüften Variablen sind signifikant.

```
Call:
glm(formula = y ~ age + dead + smoker, family = poisson, data = femsmoke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.9306  -5.3175  -0.5514   2.4229  11.1895

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.67778    0.10702  25.021  < 2e-16 ***
age25-34     0.87618    0.11003   7.963 1.67e-15 ***
age35-44     0.67591    0.11356   5.952 2.65e-09 ***
age45-54     0.57536    0.11556   4.979 6.40e-07 ***
age55-64     0.70166    0.11307   6.206 5.45e-10 ***
age65-74     0.34377    0.12086   2.844  0.00445 **
age75+      -0.41837    0.14674  -2.851  0.00436 **
deadno       0.94039    0.06139  15.319  < 2e-16 ***
smokerno     0.22931    0.05554   4.129 3.64e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.9  on 27  degrees of freedom
Residual deviance:  735.0  on 19  degrees of freedom
AIC: 887.2

Number of Fisher Scoring iterations: 6

> c(deviance(m1.fem), df.residual(m1.fem))
[1] 735.0028  19.0000
> qchisq(0.95,df.residual(m1.fem))
[1] 30.14353
```

***Modell der 2-fach-Interaktionen:***

Die untenstehend mit Stern markierten Variablen im Modelloutput sind signifikant und es gibt daher einen Zusammenhang. Wenn man das Modell vereinfacht, stellt sich heraus, dass die die Interaktion `age:smoker` hoch signifikant ist.

```
> summary(m2.fem)

Call:
glm(formula = y ~ (age + dead + smoker)^2, family = poisson,
    data = femsmoke)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.70006  -0.11004  -0.00002   0.12254   0.67272

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.54284    0.58736   0.924 0.355384
age25-34             0.92902    0.68381   1.359 0.174273
age35-44             1.94048    0.62486   3.105 0.001900 **
age45-54             2.76845    0.60657   4.564 5.02e-06 ***
age55-64             3.37507    0.59550   5.668 1.45e-08 ***
age65-74             2.86586    0.60894   4.706 2.52e-06 ***
age75+               2.02211    0.64955   3.113 0.001851 **
deadno               3.43271    0.59014   5.817 6.00e-09 ***
smokerno            -0.29666    0.25324  -1.171 0.241401
age25-34:deadno     -0.12006    0.68655  -0.175 0.861178
age35-44:deadno     -1.34112    0.62857  -2.134 0.032874 *
age45-54:deadno     -2.11336    0.61210  -3.453 0.000555 ***
age55-64:deadno     -3.18077    0.60057  -5.296 1.18e-07 ***
age65-74:deadno     -5.08798    0.61951  -8.213  < 2e-16 ***
age75+:deadno      -27.31727 8839.01146  -0.003 0.997534
age25-34:smokerno    0.11752    0.22091   0.532 0.594749
age35-44:smokerno    0.01268    0.22800   0.056 0.955654
age45-54:smokerno   -0.56538    0.23585  -2.397 0.016522 *
age55-64:smokerno    0.08512    0.23573   0.361 0.718030
age65-74:smokerno    1.49088    0.30039   4.963 6.93e-07 ***
age75+:smokerno      1.89060    0.39582   4.776 1.78e-06 ***
deadno:smokerno      0.42741    0.17703   2.414 0.015762 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.9378  on 27  degrees of freedom
Residual deviance:    2.3809  on  6  degrees of freedom
AIC: 180.58

Number of Fisher Scoring iterations: 18

> c(deviance(m2.fem), df.residual(m2.fem))
[1] 2.380927 6.000000
> qchisq(0.95,df.residual(m2.fem))
[1] 12.59159
> drop1(m2.fem, test="Chisq")
Single term deletions

Model:
y ~ (age + dead + smoker)^2
```

```
            Df Deviance    AIC    LRT Pr(>Chi)
<none>              2.38 180.58
age:dead     6   632.30 798.49 629.92  < 2e-16 ***
age:smoker   6    92.63 258.83  90.25  < 2e-16 ***
dead:smoker  1     8.33 184.52   5.95  0.01475 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

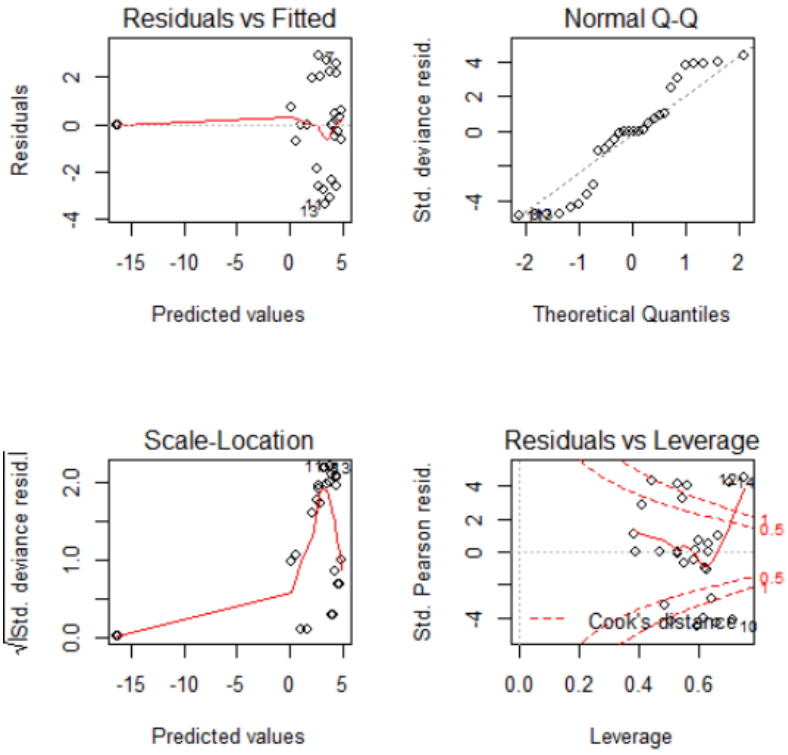
***Modell der bedingten Unabhängigkeit*** *(von* `smoker` *und* `age` *gegeben* `dead`*)*:

```
> summary(m3.fem)

Call:
glm(formula = y ~ dead * smoker + dead * age, family = poisson,
    data = femsmoke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3384  -0.9860  -0.0004   1.0406   2.8717

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.1223     0.5812   0.210 0.833355
deadno            3.8563     0.5897   6.539 6.19e-11 ***
smokerno          0.5036     0.1074   4.688 2.76e-06 ***
age25-34          0.9808     0.6770   1.449 0.147399
age35-44          1.9459     0.6172   3.153 0.001617 **
age45-54          2.5649     0.5991   4.281 1.86e-05 ***
age55-64          3.4122     0.5868   5.815 6.06e-09 ***
age65-74          3.7689     0.5840   6.454 1.09e-10 ***
age75+            3.2452     0.5885   5.514 3.50e-08 ***
deadno:smokerno  -0.3786     0.1257  -3.013 0.002590 **
deadno:age25-34  -0.1076     0.6861  -0.157 0.875435
deadno:age35-44  -1.3398     0.6281  -2.133 0.032920 *
deadno:age45-54  -2.1712     0.6113  -3.552 0.000382 ***
deadno:age55-64  -3.1717     0.6000  -5.286 1.25e-07 ***
deadno:age65-74  -4.9498     0.6151  -8.047 8.49e-16 ***
deadno:age75+   -23.5891  1485.9944  -0.016 0.987335
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.938  on 27  degrees of freedom
Residual deviance:   92.633  on 12  degrees of freedom
AIC: 258.83

Number of Fisher Scoring iterations: 14

> c(deviance(m3.fem), df.residual(m3.fem))
[1] 92.6332 12.0000
> qchisq(0.95,df.residual(m3.fem))
[1] 21.02607
> drop1(m3.fem, test="Chisq")
Single term deletions

Model:
y ~ dead * smoker + dead * age
            Df Deviance    AIC    LRT Pr(>Chi)
```

```
<none>              92.63 258.83
dead:smoker  1    101.83 266.03   9.20  0.00242 **
dead:age     6    725.80 880.00 633.17  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Diagnostische Plots:***

Der Plot Residuals vs. Fitted zeigt hier, dass die Residuen nicht rein zufällig um die Nulllinie verstreut sind. Im Normal Q-Q Plot weichen die standardisierten Residuen stärker von den theoretischen Quantilen ab. Wie der Residuals vs. Leverage Plot zeigt, dass vergleichsweise viele Punkte außerhalb der Cook's Distance Linie liegen.




***Suche nach Möglichkeiten einfacherer Modelle:***

Hier sind die Variable `age` und die Interaktion `dead` mit `smoker` signifikant.

```
> summary(m4.fem)

Call:
glm(formula = y ~ age + dead * smoker, family = poisson, data = femsmoke)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2606  -5.1564  -0.5933   2.5373  10.4236

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.51582    0.12239  20.555  < 2e-16 ***
age25-34         0.87618    0.11003   7.963 1.67e-15 ***
age35-44         0.67591    0.11356   5.952 2.65e-09 ***
age45-54         0.57536    0.11556   4.979 6.40e-07 ***
age55-64         0.70166    0.11307   6.206 5.45e-10 ***
```

```
age65-74          0.34377    0.12086   2.844  0.00445 **
age75+           -0.41837    0.14674  -2.851  0.00436 **
deadno            1.15910    0.09722  11.922  < 2e-16 ***
smokerno          0.50361    0.10743   4.688 2.76e-06 ***
deadno:smokerno  -0.37858    0.12566  -3.013  0.00259 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1193.9  on 27  degrees of freedom
Residual deviance:  725.8  on 18  degrees of freedom
AIC: 880

Number of Fisher Scoring iterations: 6


> c(deviance(m4.fem), df.residual(m4.fem))
[1] 725.8025  18.0000
> qchisq(0.95,df.residual(m4.fem))
[1] 28.8693
> drop1(m4.fem, test="Chisq")
Single term deletions

Model:
y ~ age + dead * smoker
            Df Deviance    AIC   LRT Pr(>Chi)
<none>           725.8  880.0
age          6    906.3 1048.5 180.5  < 2e-16 ***
dead:smoker  1    735.0  887.2   9.2  0.00242 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Modellvergleich:***

Die ANOVA ergibt, dass sich die Modelle 3 und 4 unterscheiden.

```
> anova(m3.fem, m4.fem, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ dead * smoker + dead * age
Model 2: y ~ age + dead * smoker
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        12      92.63
2        18     725.80 -6  -633.17 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Zusammenfassung für die jeweiligen Variablen[2]:***

***Zusammenfassung über Variable*** `smoker`:

Variable `smoker` gesamt:

Der Chi-Quadrat-Test ist signifikant, wenn man die gesamten Daten zur Variable betrachtet.

```
> tab.smoke
```

[2] Anmerkung: Wenn die Voraussetzungen erfüllt sind, wird der Chi-Quadrat-Test verwendet. Sind die Voraussetzungen nicht erfüllt, wird versucht den exakten Fisher-Test anzuwenden.

```
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74 75+
  yes     3     8    21    39    91   130  77
  no    114   273   209   169   145    35   0
> round(prop.table(tab.smoke), digits=2)
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74  75+
  yes  0.00  0.01  0.02  0.03  0.07  0.10 0.06
  no   0.09  0.21  0.16  0.13  0.11  0.03 0.00
> chisq.test(tab.smoke)

        Pearson's Chi-squared test

data:  tab.smoke
X-squared = 596.28, df = 6, p-value < 2.2e-16
```

Variable `smoker=="yes"`:

Der Chi-Quadrat-Test ist signifikant, wenn man die Gruppe Raucher betrachtet.

```
> tab.smoke.yes
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74 75+
  yes     2     3    14    27    51    29  13
  no     53   121    95   103    64     7   0
> round(prop.table(tab.smoke.yes), digits=2)
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74  75+
  yes  0.00  0.01  0.02  0.05  0.09  0.05 0.02
  no   0.09  0.21  0.16  0.18  0.11  0.01 0.00


> summary(tab.smoke.yes)
Call: xtabs(formula = y ~ dead + age, data = femsmoke, subset = (smoker ==
    "yes"))
Number of cases in table: 582
Number of factors: 2
Test for independence of all factors:
        Chisq = 183.35, df = 6, p-value = 6.58e-37
        Chi-squared approximation may be incorrect
```

Variable `smoker=="no"`:

Der Chi-Quadrat-Test ist signifikant, wenn man die Gruppe Nichtraucher betrachtet.

```
> tab.smoke.no
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74 75+
  yes     1     5     7    12    40   101  64
  no     61   152   114    66    81    28   0
> round(prop.table(tab.smoke.no), digits=2)
     age
dead  18-24 25-34 35-44 45-54 55-64 65-74  75+
  yes  0.00  0.01  0.01  0.02  0.05  0.14 0.09
  no   0.08  0.21  0.16  0.09  0.11  0.04 0.00


> summary(tab.smoke.no)
Call: xtabs(formula = y ~ dead + age, data = femsmoke, subset = (smoker ==
    "no"))
```

```
Number of cases in table: 732
Number of factors: 2
Test for independence of all factors:
        Chisq = 401.2, df = 6, p-value = 1.514e-83
```

***Zusammenfassung über Variable*** `dead`:

Variable `dead`:

Der Chi-Quadrat-Test ist signifikant, wenn man die gesamte Variable betrachtet.

```
> tab.dead
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74 75+
   yes    55   124   109   130   115    36  13
   no     62   157   121    78   121   129  64
> round(prop.table(tab.dead,1), digits=2)
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74  75+
   yes  0.09  0.21  0.19  0.22  0.20  0.06 0.02
   no   0.08  0.21  0.17  0.11  0.17  0.18 0.09
> chisq.test(tab.dead)

        Pearson's Chi-squared test

data:  tab.dead
X-squared = 88.298, df = 6, p-value < 2.2e-16
```

Variable `dead=="yes"`:

Der Chi-Quadrat-Test ist möglicherweise keine gute Approximation, wenn man die Gruppe `"yes"` der Variable betrachtet. Der Fisher-Test erzielte kein Ergebnis.

```
> tab.dead.yes
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74 75+
   yes     2     3    14    27    51    29  13
   no      1     5     7    12    40   101  64
> round(prop.table(tab.dead.yes,1), digits=2)
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74  75+
   yes  0.01  0.02  0.10  0.19  0.37  0.21 0.09
   no   0.00  0.02  0.03  0.05  0.17  0.44 0.28

> chisq.test(tab.dead.yes) # voraussetzung nicht erfuellt

        Pearson's Chi-squared test

data:  tab.dead.yes
X-squared = 65.461, df = 6, p-value = 3.473e-12

Warning message:
In chisq.test(tab.dead.yes) : Chi-Quadrat-Approximation kann inkorrekt sein
```

Variable `dead=="no"`:

Der Chi-Quadrat-Test kann nicht berechnet werden, wenn man die Gruppe `"no"` der Variable betrachtet.

```
> tab.dead.no
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74 75+
   yes    53   121    95   103    64     7   0
   no     61   152   114    66    81    28   0
> round(prop.table(tab.dead.no,1), digits=2)
      age
smoker 18-24 25-34 35-44 45-54 55-64 65-74  75+
   yes  0.12  0.27  0.21  0.23  0.14  0.02 0.00
   no   0.12  0.30  0.23  0.13  0.16  0.06 0.00
> summary(tab.dead.no)
Call: xtabs(formula = y ~ smoker + age, data = femsmoke, subset = (dead ==
    "no"))
Number of cases in table: 945
Number of factors: 2
Test for independence of all factors:
        Chisq = NaN, df = 6, p-value = NA
        Chi-squared approximation may be incorrect
```

***Zusammenfassung über Variable*** `age`:

Variable `age`:

Der Chi-Quadrat-Test ist signifikant, wenn man die gesamte Variable betrachtet.

```
> tab.age
      dead
smoker yes  no
   yes 139 443
   no  230 502
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.24 0.76
   no  0.31 0.69
> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 8.7515, df = 1, p-value = 0.003093
```

Variable `age` Gruppe `"18-25"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 18-25 der Variable betrachtet.

```
> tab.age
      dead
smoker yes no
   yes   2 53
   no    1 61
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.04 0.96
   no  0.02 0.98
> fisher.test(tab.age)
```

```
        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 0.6002
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.1159566 137.8768839
sample estimates:
odds ratio
  2.285972

> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 0.011061, df = 1, p-value = 0.9162
```

Variable `age` Gruppe `"25-34"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 25-34 der Variable betrachtet.

```
> tab.age
      dead
smoker yes  no
   yes   3 121
   no    5 152
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.02 0.98
   no  0.03 0.97
> fisher.test(tab.age) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1149111 3.9658960
sample estimates:
odds ratio
 0.7544629

> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 0.0004775, df = 1, p-value = 0.9826

Warning message:
In chisq.test(tab.age) : Chi-Quadrat-Approximation kann inkorrekt sein
```

Variable `age` Gruppe `"35-44"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 35-44 der Variable betrachtet.

```
> tab.age
      dead
smoker yes  no
   yes  14  95
   no    7 114
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.13 0.87
   no  0.06 0.94
> fisher.test(tab.age)

        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 0.07056
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.860952 7.300904
sample estimates:
odds ratio
   2.39102

> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 2.6456, df = 1, p-value = 0.1038
```

Variable `age` Gruppe `"45-54"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 45-54 der Variable betrachtet. Da die Häufigkeiten Null sind, lässt sich aus den Tests eigentlich keine sinnvolle Aussage ableiten zu dieser Altersgruppe.

```
> tab.age
      dead
smoker yes no
   yes   0  0
   no    0  0
> round(prop.table(tab.age,1), digits=2)
      dead
smoker yes no
   yes
   no
> fisher.test(tab.age) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0 Inf
sample estimates:
```

```
odds ratio
         0

> chisq.test(tab.age)
Error in chisq.test(tab.age) :
  mindestens ein Eintrag von 'x' muss positiv sein
```

Variable `age` Gruppe `"55-64"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 55-64 der Variable betrachtet.

```
> tab.age
      dead
smoker yes no
   yes  51 64
   no   40 81
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.44 0.56
   no  0.33 0.67
> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 2.7136, df = 1, p-value = 0.09949

> summary(tab.age)
Call: xtabs(formula = y ~ smoker + dead, data = femsmoke, subset = (age ==
    "55-64"))
Number of cases in table: 236
Number of factors: 2
Test for independence of all factors:
        Chisq = 3.172, df = 1, p-value = 0.0749
```

Variable `age` Gruppe `"65-74"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 65-74 der Variable betrachtet.

```
> tab.age
      dead
smoker yes  no
   yes  29   7
   no  101  28
> round(prop.table(tab.age,1), digits=2)
      dead
smoker  yes   no
   yes 0.81 0.19
   no  0.78 0.22
> fisher.test(tab.age)

        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
```

```
95 percent confidence interval:
 0.4307119 3.4358242
sample estimates:
odds ratio
  1.147591

> chisq.test(tab.age)

        Pearson's Chi-squared test with Yates' continuity correction

data:  tab.age
X-squared = 0.0039532, df = 1, p-value = 0.9499
```

Variable `age` Gruppe `"75+"`:

Der Chi-Quadrat-Test bzw. exakte Fisher-Test ist nicht signifikant, wenn man die Gruppe 75+ der Variable betrachtet. Da einige Zellenhäufigkeiten Null sind, ist das Ergebnis eigentlich nicht aufschlussreich.

```
> tab.age
      dead
smoker yes no
   yes  13  0
   no   64  0
> round(prop.table(tab.age,1), digits=2)
      dead
smoker yes no
   yes   1  0
   no    1  0
> fisher.test(tab.age) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0 Inf
sample estimates:
odds ratio
         0

> chisq.test(tab.age)

        Pearson's Chi-squared test

data:  tab.age
X-squared = NaN, df = 1, p-value = NA

Warning message:
In chisq.test(tab.age) : Chi-Quadrat-Approximation kann inkorrekt sein
```

## R-Code zu Aufgabe 4 (Dreidimensionale Tabellen):

```
#*****************************************************************************************
# AUFGABE 4 (Dreidimensionale Tabellen)
#*****************************************************************************************
# Die Daten femsmoke in der Library faraway zeigen die Ergebnisse einer Studie
# über das Rauchen bei Frauen in den Jahren 1972 - 1974. Die Variable y gibt die
# Anzahl der Fälle in den Gruppen an, die durch Raucher, Tot und Altersgruppe
# gebildet werden.
# Die kleinen Fallzahlen in manchen Altersgruppen ergeben sich dadurch, dass Personen
# im Laufe der Untersuchung ausgeschieden werden mussten. Beachte bei der Modellierung,
# dass dieser Datensatz ein Beispiel für Simpsons Paradoxon ist. In den einzelnen
# Altersgruppen sind die Ergebnisse anders als das Gesamtergebnis über alle Altersgruppen.

# DATEN UND DESKRIPTIVES:
# daten laden und tabelle erstellen
data(femsmoke)
head(femsmoke, n=5)
summary(femsmoke)

is.data.frame(femsmoke)     # ist schon data.frame
is.factor(femsmoke$smoker) # ist schon factor
is.factor(femsmoke$dead)    # ist schon factor
is.factor(femsmoke$age)     # ist schon factor

tab.fem <- xtabs(femsmoke$y ~ femsmoke$smoker + femsmoke$dead + femsmoke$age)
tab.fem
summary(tab.fem)


# MOSAICPLOT:
par(mfrow=c(1,1))
mosaicplot(tab.fem, color=c("dodgerblue","dodgerblue3",
                            "dodgerblue4","darkslateblue"),
           cex=0.5)

# LOG-LINEARES MODELL / MODELLE PRUEFEN:
# zshg. dead - smoker
mod1.fem <- xtabs(y ~ smoker + dead, femsmoke)
summary(mod1.fem) # signifikanter zshg.
round(prop.table(mod1.fem,1), digits=3)

# zshg. age - smoker
mod2.fem <- xtabs(y ~ age + smoker, femsmoke)
summary(mod2.fem) # signifikanter zshg.
round(prop.table(mod2.fem,1), digits=3)

# zshg. age - dead
mod3.fem <- xtabs(y ~ age + dead, femsmoke)
summary(mod3.fem) # signifikanter zshg.
round(prop.table(mod3.fem,1), digits=3)

# zshg. age - dead - smoker
mod4.fem <- xtabs(y ~ age + dead + smoker, femsmoke)
summary(mod4.fem) # signifikanter zshg.
round(prop.table(mod4.fem,1),digits=3)


# MODELL "TOTALE UNABHAENGIGKEIT":
m1.fem <- glm(y ~ age + dead + smoker, data = femsmoke, family = poisson)
summary(m1.fem)
c(deviance(m1.fem), df.residual(m1.fem))
qchisq(0.95,df.residual(m1.fem))
```

```
# MODELL "TOTALE UNABHAENGIGKEIT":
m1.fem <- glm(y ~ age + dead + smoker, data = femsmoke, family = poisson)
summary(m1.fem)
c(deviance(m1.fem), df.residual(m1.fem))
qchisq(0.95,df.residual(m1.fem))

# MODELL DER 2-FACH-INTERAKTIONEN:
m2.fem <- glm(y ~ (age + dead + smoker)^2, data=femsmoke, family=poisson)
summary(m2.fem)
c(deviance(m2.fem), df.residual(m2.fem))
qchisq(0.95,df.residual(m2.fem))
drop1(m2.fem, test="Chisq")

# MODELL D. BEDINGTEN UNABH. (von smoker und age gegeben dead)
m3.fem <- glm(y ~ dead*smoker + dead*age, data=femsmoke, family=poisson)
summary(m3.fem)
c(deviance(m3.fem), df.residual(m3.fem))
qchisq(0.95,df.residual(m3.fem))
drop1(m3.fem, test="Chisq")
par(mfrow=c(2,2))
plot(m3.fem)

# SUCHE NACH MOEGLICHKEITEN EINFACHERER MODELLE
m4.fem <- glm(y ~ age + dead*smoker, data=femsmoke, family=poisson)
c(deviance(m4.fem), df.residual(m4.fem))
qchisq(0.95,df.residual(m4.fem))
drop1(m4.fem, test="Chisq")


# MODELLVERGLEICH
anova(m3.fem, m4.fem, test="Chisq")

# ZSFG. UEBER VARIABLE smoker
  # smoker variable gesamt
tab.smoke <- xtabs(y ~ dead + age, data=femsmoke)
tab.smoke
chisq.test(tab.smoke)
round(prop.table(tab.smoke), digits=2)

  # smoker=="yes"
tab.smoke.yes <- xtabs(y ~ dead + age, data=femsmoke,
                       subset=(smoker=="yes"))
tab.smoke.yes
fisher.test(tab.smoke.yes) # voraussetzung nicht erfuellt
chisq.test(tab.smoke.yes) # voraussetzung nicht erfuellt
summary(tab.smoke.yes)
addmargins(tab.smoke.yes)
round(prop.table(tab.smoke.yes), digits=2)

  # smoker=="no"
tab.smoke.no <- xtabs(y ~ dead + age, data=femsmoke,
                       subset=(smoker=="no"))
tab.smoke.no
fisher.test(tab.smoke.no) # voraussetzung nicht erfuellt
chisq.test(tab.smoke.no)
summary(tab.smoke.no)
addmargins(tab.smoke.no)
round(prop.table(tab.smoke.no), digits=2)
```

```
# ZSFG. UEBER VARIABLE dead
  # dead variable gesamt
tab.dead <- xtabs(y ~ smoker + age, data=femsmoke)
tab.dead
chisq.test(tab.dead)
round(prop.table(tab.dead,1), digits=2)

  # dead=="yes"
tab.dead.yes <- xtabs(y ~ smoker + age, data=femsmoke,
                      subset=(dead=="yes"))
tab.dead.yes
fisher.test(tab.dead.yes) # voraussetzung nicht erfuellt
chisq.test(tab.dead.yes) # voraussetzung nicht erfuellt
summary(tab.dead.yes)
addmargins(tab.dead.yes)
round(prop.table(tab.dead.yes,1), digits=2)


  # dead=="no"
tab.dead.no <- xtabs(y ~ smoker + age, data=femsmoke,
                     subset=(dead=="no"))
tab.dead.no
fisher.test(tab.dead.no) # voraussetzung nicht erfuellt
chisq.test(tab.dead.no)
summary(tab.dead.no)
addmargins(tab.dead.no)
round(prop.table(tab.dead.no,1), digits=2)


# ZSFG. UEBER VARIABLE age
# age variable gesamt
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke)
tab.age
chisq.test(tab.age)
round(prop.table(tab.age,1), digits=2)
femsmoke$age

# age=="18-24"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                     subset=(age=="18-24"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)

# age=="25-34"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                      subset=(age=="25-34"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)
```

```
# age=="35-44"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                 subset=(age=="35-44"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)

# age=="45-55"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                 subset=(age=="45-55"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)

# age=="55-64"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                 subset=(age=="55-64"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)


# age=="75+"
tab.age <- xtabs(y ~ smoker + dead, data=femsmoke,
                 subset=(age=="75+"))
tab.age
fisher.test(tab.age) # voraussetzung nicht erfuellt
chisq.test(tab.age)
summary(tab.age)
addmargins(tab.age)
round(prop.table(tab.age,1), digits=2)
```

## Aufgabe 5 (Dreidimensionale Tabellen):

**Im Datensatz suicide in der Library faraway findet man die Ergebnisse von Selbstmorden in Großbritannien. Die Kreuzklassifizierungsvariablen sind:**

**• Cause = Methode des Selbstmordes**

**• Age = Altersgruppe (y = young, m = middle, o = old)**

**• Sex = Geschlecht (m = male, f = female)**

**Lassen sich bestimmte "Vorlieben" für Methoden in bestimmten Kombinationen von Alter und Geschlecht erkennen?**

**In diesem Beispiel könnte man auch eine Korrespondenzanalyse für die zweidimensionale Tabelle durchführen, die sich aus Methode und der Kombination von Altersgruppe und Geschlecht ergibt.**

***Überblick über den Datensatz `suicide`:***

```
> head(suicide, n=5)
    y cause age sex
1 398  drug   y   m
2 121   gas   y   m
3 455  hang   y   m
4 155   gun   y   m
5  55  jump   y   m
```

***Deskriptive Zusammenfassung:***

```
> summary(suicide)
       y           cause   age     sex
 Min.   :  5.0   drug :6   m:12   f:18
 1st Qu.: 26.0   gas  :6   o:12   m:18
 Median : 76.5   gun  :6   y:12
 Mean   :147.4   hang :6
 3rd Qu.:172.2   jump :6
 Max.   :797.0   other:6
```

***Aus dem Datensatz erstellte Tabelle:***

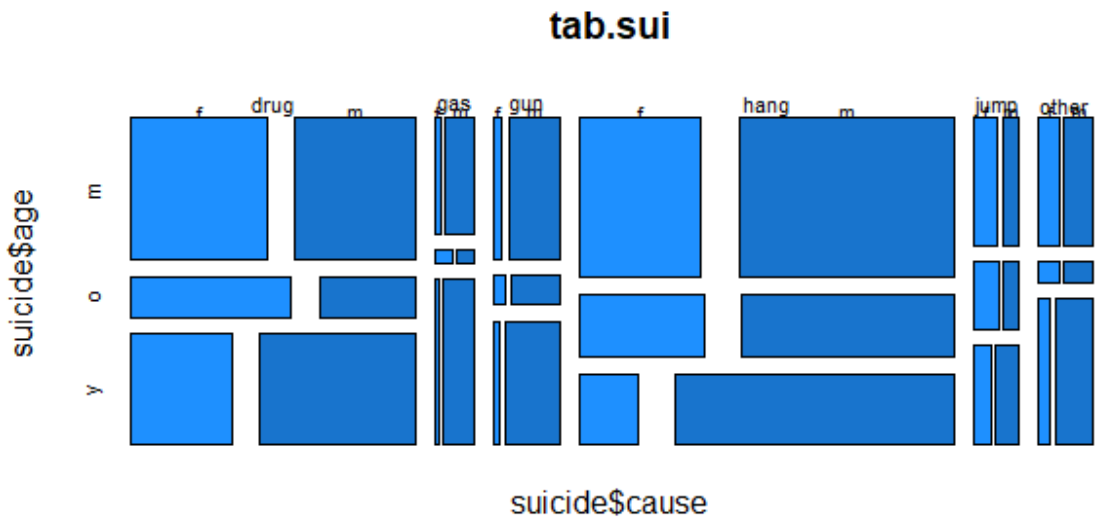
```
              suicide$age
suicide$cause   m   o   y
        drug  399  93 398
        gas    82   6 121
        gun   168  33 155
        hang  797 316 455
        jump   51  26  55
        other  82  14 124
```

Der Test auf die Unabhängigkeit aller Variablen ist signifikant, d.h. demgemäß gibt es einen Zusammenhang zwischen ihnen.

```
> summary(tab.sui)
Call: xtabs(formula = suicide$y ~ suicide$cause + suicide$age + suicide$sex
)
Number of cases in table: 5305
Number of factors: 3
Test for independence of all factors:
```

```
        Chisq = 747.4, df = 27, p-value = 4.449e-140
```

***Mosaicplot:***




Aus dem Mosaicplot ist ersichtlich, dass Männer mittleren Alters zu der Methode „hang“ tendieren. Bei den Frauen mittleren Alters ist die Methode „hang“ auch häufiger vorgekommen als bei den anderen weiblichen Altersgruppen, aber weniger im Vergleich zu den Männern. Bei den Frauen scheint im mittleren Alter die Methode „drug“ zu überwiegen. Die Methode „drug“ kommt auch bei den Männern häufig vor, i.e. die zweithäufigste Methode. Bei jungen Männern ist „hang“ und „drug“ ebenfalls die häufigste Methode, während die Methoden „jump“, „gas“ und „other“ eher selten vorkommen. Ihr Vorkommen ist allerdings bei Männern häufiger als bei Frauen. Bei Frauen überwiegen insgesamt die Methoden „drug“, gefolgt von „hang“. Die Methoden „gas“, „gun“, und „other“ kommen bei Frauen fast nicht vor.

***Log-lineares Modell / Modelle prüfen:***

- Der Zusammenhang `cause – age` ist signifikant:

```
> summary(mod1.sui)
Call: xtabs(formula = y ~ cause + age, data = suicide)
Number of cases in table: 5305
Number of factors: 2
Test for independence of all factors:
        Chisq = 261.51, df = 10, p-value = 2.057e-50
```

- Der Zusammenhang `cause – sex` ist signifikant:

```
> summary(mod2.sui)
Call: xtabs(formula = y ~ cause + sex, data = suicide)
Number of cases in table: 5305
Number of factors: 2
Test for independence of all factors:
        Chisq = 342, df = 5, p-value = 9.324e-72
```

- Der Zusammenhang `age – sex` ist signifikant:

```
> summary(mod3.sui)
```

```
Call: xtabs(formula = y ~ age + sex, data = suicide)
Number of cases in table: 5305
Number of factors: 2
Test for independence of all factors:
        Chisq = 128.19, df = 2, p-value = 1.456e-28
```

- Der Zusammenhang `age – sex - cause` ist signifikant:

```
> summary(mod4.sui)
Call: xtabs(formula = y ~ age + sex + cause, data = suicide)
Number of cases in table: 5305
Number of factors: 3
Test for independence of all factors:
        Chisq = 747.4, df = 27, p-value = 4.449e-140
```

***Modell der totalen Unabhängigkeit:***

In diesem Modell sind alle enthaltenen Variablen signifikant auf dem 0.001 Level.

```
> summary(m1.sui)

Call:
glm(formula = y ~ age + sex + cause, family = poisson, data = suicide)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12.7679   -4.0651   -0.4458    3.1459    8.2556

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.76349    0.03300 174.644   <2e-16 ***
ageo        -1.09409    0.03879 -28.203   <2e-16 ***
agey        -0.40377    0.03070 -13.150   <2e-16 ***
sexm         0.55888    0.02854  19.583   <2e-16 ***
causegas    -1.98015    0.06858 -28.875   <2e-16 ***
causegun    -1.47015    0.05524 -26.612   <2e-16 ***
causehang    0.27071    0.03171   8.537   <2e-16 ***
causejump   -1.83073    0.06426 -28.490   <2e-16 ***
causeother  -1.67607    0.06016 -27.860   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6341.8  on 35  degrees of freedom
Residual deviance:  790.3  on 27  degrees of freedom
AIC: 1026.4

Number of Fisher Scoring iterations: 5

> c(deviance(m1.sui), df.residual(m1.sui))
[1] 790.2961  27.0000
> qchisq(0.95,df.residual(m1.sui))
[1] 40.11327
```

***Modell der Zweifachinteraktionen:***

In diesem Modell sind `causehang, ageo:causegun, agey:causegun, agey:causejump` und `sexm:causejump` nicht signifikant. Die anderen Variablen sind alle signifikant. Wenn man `drop1()` anwendet, stellt sich heraus, dass `age:sex, age:cause` und `sex:cause` auf dem 0.001 Level signifikant sind.

```
> summary(m2.sui)

Call:
glm(formula = y ~ (age + sex + cause)^2, family = poisson, data = suicide)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.53676  -0.31894  -0.00096   0.38909   1.52264

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.15179    0.04290 143.404  < 2e-16 ***
ageo            -1.15935    0.07976 -14.536  < 2e-16 ***
agey            -0.64587    0.06557  -9.851  < 2e-16 ***
sexm            -0.21309    0.05759  -3.700 0.000215 ***
causegas        -3.29650    0.19016 -17.335  < 2e-16 ***
causegun        -2.82923    0.15745 -17.969  < 2e-16 ***
causehang       -0.09473    0.05827  -1.626 0.104031
causejump       -1.89039    0.11157 -16.943  < 2e-16 ***
causeother      -2.10341    0.11747 -17.906  < 2e-16 ***
ageo:sexm       -0.17715    0.08125  -2.180 0.029241 *
agey:sexm        0.72540    0.07058  10.278  < 2e-16 ***
ageo:causegas   -0.85433    0.32809  -2.604 0.009217 **
agey:causegas    0.37837    0.14628   2.587 0.009694 **
ageo:causegun   -0.26981    0.19129  -1.410 0.158403
agey:causegun   -0.14089    0.12072  -1.167 0.243179
ageo:causehang   0.36033    0.09131   3.946 7.93e-05 ***
agey:causehang  -0.70369    0.07520  -9.357  < 2e-16 ***
ageo:causejump   0.58462    0.17054   3.428 0.000608 ***
agey:causejump   0.02676    0.14825   0.180 0.856763
ageo:causeother -0.51658    0.23262  -2.221 0.026369 *
agey:causeother  0.28566    0.12816   2.229 0.025819 *
sexm:causegas    1.70963    0.19514   8.761  < 2e-16 ***
sexm:causegun    2.00413    0.16368  12.244  < 2e-16 ***
sexm:causehang   0.86519    0.06773  12.774  < 2e-16 ***
sexm:causejump  -0.11470    0.13117  -0.874 0.381887
sexm:causeother  0.60377    0.12897   4.681 2.85e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6341.809  on 35  degrees of freedom
Residual deviance:   14.901  on 10  degrees of freedom
AIC: 285

Number of Fisher Scoring iterations: 4

> c(deviance(m2.sui), df.residual(m2.sui))
[1] 14.90066 10.00000
> qchisq(0.95,df.residual(m2.sui))
[1] 18.30704
```

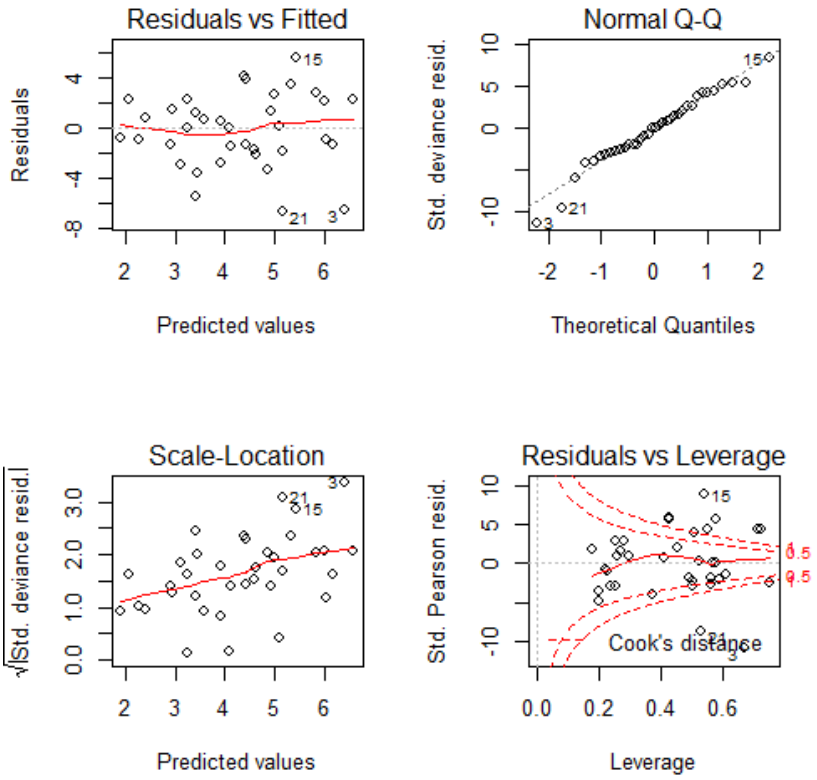
```
> drop1(m2.sui, test="Chisq")
Single term deletions

Model:
y ~ (age + sex + cause)^2
          Df Deviance    AIC    LRT  Pr(>Chi)
<none>          14.90 285.00
age:sex    2   154.70 420.80 139.80 < 2.2e-16 ***
age:cause 10   293.18 543.27 278.28 < 2.2e-16 ***
sex:cause  5   389.00 649.09 374.10 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Modell der bedingten Unabhängigkeit (von*** `cause` ***und*** `age` ***gegeben*** `sex`***):***

Das Modell ist signifikant.

```
> summary(m3.sui)

Call:
glm(formula = y ~ sex * age + sex * cause, family = poisson,
    data = suicide)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.5841  -1.6995   0.0024   2.2189   5.5928

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.17055    0.03969 155.474  < 2e-16 ***
sexm            -0.13894    0.05510  -2.522 0.011682 *
ageo            -0.98645    0.05866 -16.817  < 2e-16 ***
agey            -0.84202    0.05571 -15.114  < 2e-16 ***
causegas        -3.26391    0.17737 -18.401  < 2e-16 ***
causegun        -2.91027    0.14978 -19.430  < 2e-16 ***
causehang       -0.16737    0.05029  -3.328 0.000873 ***
causejump       -1.75647    0.08871 -19.799  < 2e-16 ***
causeother      -2.07828    0.10207 -20.362  < 2e-16 ***
sexm:ageo       -0.18778    0.07825  -2.400 0.016409 *
sexm:agey        0.65372    0.06709   9.743  < 2e-16 ***
sexm:causegas    1.81502    0.19331   9.389  < 2e-16 ***
sexm:causegun    1.99398    0.16238  12.280  < 2e-16 ***
sexm:causehang   0.73370    0.06550  11.202  < 2e-16 ***
sexm:causejump  -0.15195    0.12872  -1.180 0.237819
sexm:causeother  0.68069    0.12683   5.367 8.02e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6341.81  on 35  degrees of freedom
Residual deviance:  293.18  on 20  degrees of freedom
AIC: 543.27

Number of Fisher Scoring iterations: 4

> c(deviance(m3.sui), df.residual(m3.sui))
[1] 293.1775  20.0000
```

```
> qchisq(0.95,df.residual(m3.sui))
[1] 31.41043
> drop1(m3.sui, test="Chisq")
Single term deletions

Model:
y ~ sex * age + sex * cause
          Df Deviance    AIC    LRT  Pr(>Chi)
<none>         293.18 543.27
sex:age    2   424.59 670.68 131.41 < 2.2e-16 ***
sex:cause  5   658.88 898.98 365.71 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```




Die diagnostischen Plots zeigen hinsichtlich Residuals vs. Fitted Values, dass die Daten zufällig um die Nulllinie verteilt zu sein scheinen. Im Normal Q-Q Plot liegen die Datenpunkte nahe an den theoretischen Quantilen, d.h. die Residuen scheinen annähernd normalverteilt zu sein. Es gibt nur links unten und rechts oben einige wenige Punkte, für die noch entschieden werden sollte, ob sie überhaupt in der Modellanalyse berücksichtigt werden sollten. Im Plot Residuals vs. Leverage liegen die eben genannten Punkte zum Teil außerhalb der Cook's Distance Linien, d.h. sie sollten nochmals hinsichtlich Einfluss auf die Ergebnisse der Modellierung bedacht werden.

***Suche nach einfacheren Modellen:***

```
> summary(m4.sui)

Call:
```

```
glm(formula = y ~ age + sex * cause, family = poisson, data = suicide)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-10.866   -2.103   -0.268    1.254    5.582

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      6.06595    0.03671 165.232  < 2e-16 ***
ageo            -1.09409    0.03879 -28.204  < 2e-16 ***
agey            -0.40377    0.03070 -13.150  < 2e-16 ***
sexm             0.03081    0.04777   0.645 0.519027
causegas        -3.26391    0.17737 -18.401  < 2e-16 ***
causegun        -2.91027    0.14978 -19.430  < 2e-16 ***
causehang       -0.16737    0.05028  -3.328 0.000873 ***
causejump       -1.75647    0.08871 -19.799  < 2e-16 ***
causeother      -2.07828    0.10207 -20.362  < 2e-16 ***
sexm:causegas    1.81502    0.19331   9.389  < 2e-16 ***
sexm:causegun    1.99398    0.16238  12.280  < 2e-16 ***
sexm:causehang   0.73370    0.06550  11.202  < 2e-16 ***
sexm:causejump  -0.15195    0.12872  -1.180 0.237819
sexm:causeother  0.68069    0.12683   5.367 8.02e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6341.81  on 35  degrees of freedom
Residual deviance:  424.59  on 22  degrees of freedom
AIC: 670.68

Number of Fisher Scoring iterations: 4

> c(deviance(m4.sui), df.residual(m4.sui))
[1] 424.5889  22.0000
> qchisq(0.95,df.residual(m4.sui))
[1] 33.92444
> drop1(m4.sui, test="Chisq")
Single term deletions

Model:
y ~ age + sex * cause
          Df Deviance     AIC    LRT  Pr(>Chi)
<none>         424.59  670.68
age        2  1343.11 1585.21 918.52 < 2.2e-16 ***
sex:cause  5   790.30 1026.39 365.71 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Modellvergleich zwischen*** `m3.sui` ***und*** `m4.sui`***:***

Der p-Wert gemäß ANOVA ist signifikant.

```
> anova(m3.sui, m4.sui, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ sex * age + sex * cause
Model 2: y ~ age + sex * cause
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        20     293.18
```

```
2        22     424.59 -2  -131.41 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

***Zusammenfassung über Variable*** `sex`***:***

- Für die Variable `sex` insgesamt ist der Chi-Quadrat-Test signifikant:

```
> tab.sex
    cause
age drug  gas  gun hang jump other
  m  849   95  194 1247  122   142
  o  247   11   40  501   64    24
  y  657  136  169  550   95   162
> round(prop.table(tab.sex), digits=2)
    cause
age drug  gas  gun hang jump other
  m 0.16 0.02 0.04 0.24 0.02  0.03
  o 0.05 0.00 0.01 0.09 0.01  0.00
  y 0.12 0.03 0.03 0.10 0.02  0.03
> chisq.test(tab.sex) # signifikant

        Pearson's Chi-squared test

data:  tab.sex
X-squared = 261.51, df = 10, p-value < 2.2e-16
```

- Für die Gruppe `m` der Variable `sex` ist der Chi-Quadrat-Test signifikant und der exakte Fisher-Test produziert kein Ergebnis:

```
> tab.sex.m
    cause
age drug gas gun hang jump other
  m  399  82 168  797   51    82
  o   93   6  33  316   26    14
  y  398 121 155  455   55   124
> round(prop.table(tab.sex.m), digits=2)
    cause
age drug  gas  gun hang jump other
  m 0.12 0.02 0.05 0.24 0.02  0.02
  o 0.03 0.00 0.01 0.09 0.01  0.00
  y 0.12 0.04 0.05 0.13 0.02  0.04
> fisher.test(tab.sex.m) # voraussetzung nicht erfuellt
Error in fisher.test(tab.sex.m) :
  FEXACT error 6 (f5xact).  LDKEY=601 is too small for this problem: kval=9
48616471.
Try increasing the size of the workspace.
> chisq.test(tab.sex.m)  # signifikant

        Pearson's Chi-squared test

data:  tab.sex.m
X-squared = 184.74, df = 10, p-value < 2.2e-16
```

- Für die Gruppe `f` der Variable `sex` insgesamt ist der Chi-Quadrat-Test signifikant und der exakte Fisher-Test produziert kein Ergebnis:

```
> tab.sex.f
```

```
   cause
age drug gas gun hang jump other
  m  450  13  26  450   71    60
  o  154   5   7  185   38    10
  y  259  15  14   95   40    38
> round(prop.table(tab.sex.f), digits=2)
   cause
age drug  gas  gun hang jump other
  m 0.23 0.01 0.01 0.23 0.04  0.03
  o 0.08 0.00 0.00 0.10 0.02  0.01
  y 0.13 0.01 0.01 0.05 0.02  0.02
> fisher.test(tab.sex.f) # voraussetzung nicht erfuellt
Error in fisher.test(tab.sex.f) :
  FEXACT error 6.  LDKEY=610 is too small for this problem,
  (ii := key2[itp=23] = 194392381, ldstp=18300)
Try increasing the size of the workspace and possibly 'mult'
> chisq.test(tab.sex.f)  # signifikant

        Pearson's Chi-squared test

data:  tab.sex.f
X-squared = 93.809, df = 10, p-value = 9.379e-16
```

***Zusammenfassung über Variable*** `age`***:***

- Für die Variable `age` insgesamt ist der Chi-Quadrat-Test signifikant:

```
> tab.age
   cause
sex drug  gas  gun hang jump other
  f  863   33   47  730  149   108
  m  890  209  356 1568  132   220
> round(prop.table(tab.age), digits=2)
   cause
sex drug  gas  gun hang jump other
  f 0.16 0.01 0.01 0.14 0.03  0.02
  m 0.17 0.04 0.07 0.30 0.02  0.04
> chisq.test(tab.age) # signifikant

        Pearson's Chi-squared test

data:  tab.age
X-squared = 341.98, df = 5, p-value < 2.2e-16
```

- Für die Gruppe `y` der Variable `age` produzieren die Tests kein adäquates Ergebnis:

```
> tab.age.y
   cause
age drug gas gun hang jump other
  m    0   0   0    0    0     0
  o    0   0   0    0    0     0
  y  657 136 169  550   95   162
> round(prop.table(tab.age.y), digits=2)
   cause
age drug  gas  gun hang jump other
  m 0.00 0.00 0.00 0.00 0.00  0.00
  o 0.00 0.00 0.00 0.00 0.00  0.00
  y 0.37 0.08 0.10 0.31 0.05  0.09
```

```
> fisher.test(tab.age.y) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age.y
p-value = 1
alternative hypothesis: two.sided

> chisq.test(tab.age.y)  # signifikant

        Pearson's Chi-squared test

data:  tab.age.y
X-squared = NaN, df = 10, p-value = NA

Warning message:
In chisq.test(tab.age.y) : Chi-Quadrat-Approximation kann inkorrekt sein
```

- Für die Gruppe `o` der Variable `age` produzieren die Tests kein adäquates Ergebnis:

```
> tab.age.o
   cause
age drug gas gun hang jump other
  m    0   0   0    0    0     0
  o  247  11  40  501   64    24
  y    0   0   0    0    0     0
> round(prop.table(tab.age.o), digits=2)
   cause
age drug  gas  gun hang jump other
  m 0.00 0.00 0.00 0.00 0.00  0.00
  o 0.28 0.01 0.05 0.56 0.07  0.03
  y 0.00 0.00 0.00 0.00 0.00  0.00
> fisher.test(tab.age.o) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age.o
p-value = 1
alternative hypothesis: two.sided

> chisq.test(tab.age.o)  # signifikant

        Pearson's Chi-squared test

data:  tab.age.o
X-squared = NaN, df = 10, p-value = NA

Warning message:
In chisq.test(tab.age.o) : Chi-Quadrat-Approximation kann inkorrekt sein
```

- Für die Gruppe `m` der Variable `age` produzieren die Tests kein adäquates Ergebnis:

```
> tab.age.m
   cause
age drug  gas  gun hang jump other
  m  849   95  194 1247  122   142
  o    0    0    0    0    0     0
  y    0    0    0    0    0     0
> round(prop.table(tab.age.m), digits=2)
```

```
   cause
age drug  gas  gun hang jump other
  m 0.32 0.04 0.07 0.47 0.05  0.05
  o 0.00 0.00 0.00 0.00 0.00  0.00
  y 0.00 0.00 0.00 0.00 0.00  0.00
> fisher.test(tab.age.m) # voraussetzung nicht erfuellt

        Fisher's Exact Test for Count Data

data:  tab.age.m
p-value = 1
alternative hypothesis: two.sided

> chisq.test(tab.age.m)  # signifikant

        Pearson's Chi-squared test

data:  tab.age.m
X-squared = NaN, df = 10, p-value = NA

Warning message:
In chisq.test(tab.age.m) : Chi-Quadrat-Approximation kann inkorrekt sein
```

***Zusammenfassung über Variable*** `cause`***:***

- Für die Variable `cause` insgesamt ist der Chi-Quadrat-Test signifikant:

```
> tab.cause
   age
sex    m    o    y
  f 1070  399  461
  m 1579  488 1308
> round(prop.table(tab.cause), digits=2)
   age
sex    m    o    y
  f 0.20 0.08 0.09
  m 0.30 0.09 0.25
> chisq.test(tab.cause)

        Pearson's Chi-squared test

data:  tab.cause
X-squared = 128.19, df = 2, p-value < 2.2e-16
```

- Für die Gruppe `drug` der Variable `cause` sind die Tests signifikant.:

```
> tab.cause.drug
   age
sex   m   o   y
  f 450 154 259
  m 399  93 398
> round(prop.table(tab.cause.drug), digits=2)
   age
sex    m    o    y
  f 0.26 0.09 0.15
  m 0.23 0.05 0.23
> fisher.test(tab.cause.drug)

        Fisher's Exact Test for Count Data
```

```
data:  tab.cause.drug
p-value = 5.092e-11
alternative hypothesis: two.sided

> chisq.test(tab.cause.drug)

        Pearson's Chi-squared test

data:  tab.cause.drug
X-squared = 47.132, df = 2, p-value = 5.828e-11
```

- Für die Gruppe `gas` der Variable `cause` sind die Tests signifikant:

```
> round(prop.table(tab.cause.gas), digits=2)
   age
sex    m    o    y
  f 0.05 0.02 0.06
  m 0.34 0.02 0.50
> tab.cause.gas
   age
sex   m   o   y
  f  13   5  15
  m  82   6 121
> round(prop.table(tab.cause.gas), digits=2)
   age
sex    m    o    y
  f 0.05 0.02 0.06
  m 0.34 0.02 0.50
> fisher.test(tab.cause.gas)

        Fisher's Exact Test for Count Data

data:  tab.cause.gas
p-value = 0.01518
alternative hypothesis: two.sided

> chisq.test(tab.cause.gas)

        Pearson's Chi-squared test

data:  tab.cause.gas
X-squared = 10.241, df = 2, p-value = 0.005973

Warning message:
In chisq.test(tab.cause.gas) :
  Chi-Quadrat-Approximation kann inkorrekt sein
```

- Für die Gruppe `drug` der Variable `cause` sind die Tests nicht signifikant:

```
> tab.cause.gun
   age
sex   m   o   y
  f  26   7  14
  m 168  33 155
> round(prop.table(tab.cause.gun), digits=2)
```

```
   age
sex    m    o    y
  f 0.06 0.02 0.03
  m 0.42 0.08 0.38
> fisher.test(tab.cause.gun)

        Fisher's Exact Test for Count Data

data:  tab.cause.gun
p-value = 0.1335
alternative hypothesis: two.sided

> chisq.test(tab.cause.gun)

        Pearson's Chi-squared test

data:  tab.cause.gun
X-squared = 3.7652, df = 2, p-value = 0.1522

Warning message:
In chisq.test(tab.cause.gun) :
  Chi-Quadrat-Approximation kann inkorrekt sein
```

- Für die Gruppe `hang` der Variable `cause` sind die Tests signifikant.:

```
> tab.cause.hang
   age
sex   m   o   y
  f 450 185  95
  m 797 316 455
> round(prop.table(tab.cause.hang), digits=2)
   age
sex    m    o    y
  f 0.20 0.08 0.04
  m 0.35 0.14 0.20
> fisher.test(tab.cause.hang)

        Fisher's Exact Test for Count Data

data:  tab.cause.hang
p-value < 2.2e-16
alternative hypothesis: two.sided

> chisq.test(tab.cause.hang)

        Pearson's Chi-squared test

data:  tab.cause.hang
X-squared = 70.194, df = 2, p-value = 5.722e-16
```

- Für die Gruppe `jump` der Variable `cause` sind die Tests signifikant.:

```
> tab.cause.jump
   age
sex  m  o  y
  f 71 38 40
```

```
  m 51 26 55
> round(prop.table(tab.cause.jump), digits=2)
   age
sex    m    o    y
  f 0.25 0.14 0.14
  m 0.18 0.09 0.20
> fisher.test(tab.cause.jump)

        Fisher's Exact Test for Count Data

data:  tab.cause.jump
p-value = 0.03325
alternative hypothesis: two.sided

> chisq.test(tab.cause.jump)  # signifikant

        Pearson's Chi-squared test

data:  tab.cause.jump
X-squared = 6.8939, df = 2, p-value = 0.03184
```

- Für die Gruppe `other` der Variable `cause` sind die Tests signifikant.:

```
> tab.cause.other
   age
sex   m   o   y
  f  60  10  38
  m  82  14 124
> round(prop.table(tab.cause.other), digits=2)
   age
sex    m    o    y
  f 0.18 0.03 0.12
  m 0.25 0.04 0.38
> fisher.test(tab.cause.other)

        Fisher's Exact Test for Count Data

data:  tab.cause.other
p-value = 0.001252
alternative hypothesis: two.sided

> chisq.test(tab.cause.other)  # signifikant

        Pearson's Chi-squared test

data:  tab.cause.other
X-squared = 13.001, df = 2, p-value = 0.001502
```

***Analyse als zweidimensionale Tabelle und Korrespondenzanalyse:***

Altersgruppe und Geschlecht werden in einen Factor zusammengelegt, um eine zweidimensionale Tabellenanalyse und Korrespondenzanalyse durchzuführen. Dadurch ergeben sich sechs Kombinationen nach Zusammenlegung von Geschlecht und Altersgruppe:

- m-m … male – middle age
- m-o … male – old age
- m-y … male – young age
- f-m … female – middle age
- f-o … female – old age
- f-y … female – young age

Das `data.frame` sieht vor Verbindung zu einem Datensatz pro Gruppe wie (anhand des Beispiels der Gruppe `m-m` wie folgt aus:

```
> m_m
     y cause age sex
7  399  drug   m   m
8   82   gas   m   m
9  797  hang   m   m
10 168   gun   m   m
11  51  jump   m   m
12  82 other   m   m
```

Das `data.frame` sieht in der Version, in der `sex` und `age` zum Factor `sex_age` kombiniert wurden folgendermaßen aus:

```
> head(combined.df, n=10)
     y cause sex_age
1  399  drug      mm
2   82   gas      mm
3  797  hang      mm
4  168   gun      mm
5   51  jump      mm
6   82 other      mm
7   93  drug      mo
8    6   gas      mo
9  316  hang      mo
10  33   gun      mo
```

Basierend auf diesem data.frame wird ein Modell für die Korrespondenzanalyse vorbereitet:

```
> summary(mod.sui)

Call:
glm(formula = y ~ cause + sex_age, family = poisson, data = combined.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2605  -3.8986  -0.7253   2.9870   7.8484

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.25723    0.03186 196.378  < 2e-16 ***
causegas    -1.98015    0.06858 -28.875  < 2e-16 ***
causegun    -1.47015    0.05524 -26.612  < 2e-16 ***
```

```
causehang    0.27071    0.03171   8.537  < 2e-16 ***
causejump   -1.83073    0.06426 -28.490  < 2e-16 ***
causeother  -1.67607    0.06016 -27.860  < 2e-16 ***
sex_agemo   -1.17423    0.05179 -22.672  < 2e-16 ***
sex_agemy   -0.18829    0.03739  -5.036 4.75e-07 ***
sex_agefm   -0.38913    0.03960  -9.827  < 2e-16 ***
sex_agefo   -1.37559    0.05603 -24.550  < 2e-16 ***
sex_agefy   -1.23115    0.05294 -23.256  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6341.81  on 35  degrees of freedom
Residual deviance:  658.88  on 25  degrees of freedom
AIC: 898.98

Number of Fisher Scoring iterations: 5
```

Die (zweidimensionale) Tabelle dazu sieht folgendermaßen aus:

```
> tabelle_sui
                 combined.df$sex_age
combined.df$cause  mm  mo  my  fm  fo  fy
            drug  399  93 398 450 154 259
            gas    82   6 121  13   5  15
            gun   168  33 155  26   7  14
            hang  797 316 455 450 185  95
            jump   51  26  55  71  38  40
            other  82  14 124  60  10  38
```

Dann wird aus der Tabelle die Residuenmatrix berechnen:

```
> residual_matr
       sex_age
cause            mm          mo          my          fm          fo          fy
  drug  -5.37466038 -5.37506552 -1.64596446  5.12807134  1.92931335  8.64224184
  gas    1.17475837 -3.44650865  7.94002598 -5.12571184 -3.09432485 -1.31483928
  gun    4.38721244 -0.66869568  5.58142598 -6.13190392 -4.23403653 -3.55205613
  hang   4.32127162  7.19500225 -4.68820648 -0.62699500  0.92514626 -7.40866662
  jump  -3.56878617  0.02973505 -1.71599001  1.90256772  3.66859686  3.15314418
  other -1.58159152 -2.94420055  4.79584280 -0.75691075 -2.95349557  1.77887456
```

Weiters erfolgt eine Singulärwertzerlegung, wobei insbesondere die ersten beiden Singulärwerte betrachtet werden:

```
> svd_res
$`d`
[1] 1.901170e+01 1.614435e+01 3.020823e+00 1.983648e+00 9.008452e-01 1.4745
47e-14

$u
           [,1]        [,2]
[1,] -0.6541315 -0.18830515
[2,]  0.2983911 -0.55347160
[3,]  0.4942363 -0.33565471
[4,]  0.3793917  0.60816417
[5,] -0.3078007  0.08921827
[6,] -0.0112234 -0.40953965
```

```
$v
           [,1]       [,2]
[1,]  0.4623612  0.1143842
[2,]  0.2582996  0.5406417
[3,]  0.2577436 -0.6767971
[4,] -0.4591648  0.2494933
[5,] -0.2642065  0.3016544
[6,] -0.6102740 -0.2886629
```

Die Singulärwerte 1 und 2 werden in u- und v-Komponenten zerlegt:

```
> sv1
            [,1]       [,2]
[1,] -2.85217067 -0.7566108
[2,]  1.30105689 -2.2238509
[3,]  2.15498918 -1.3486619
[4,]  1.65423895  2.4436059
[5,] -1.34208512  0.3584793
[6,] -0.04893672 -1.6455319
> sv2
          [,1]      [,2]
[1,]  2.016006  0.459596
[2,]  1.126248  2.172300
[3,]  1.123824 -2.719373
[4,] -2.002069  1.002465
[5,] -1.152004  1.212049
[6,] -2.660941 -1.159848
```

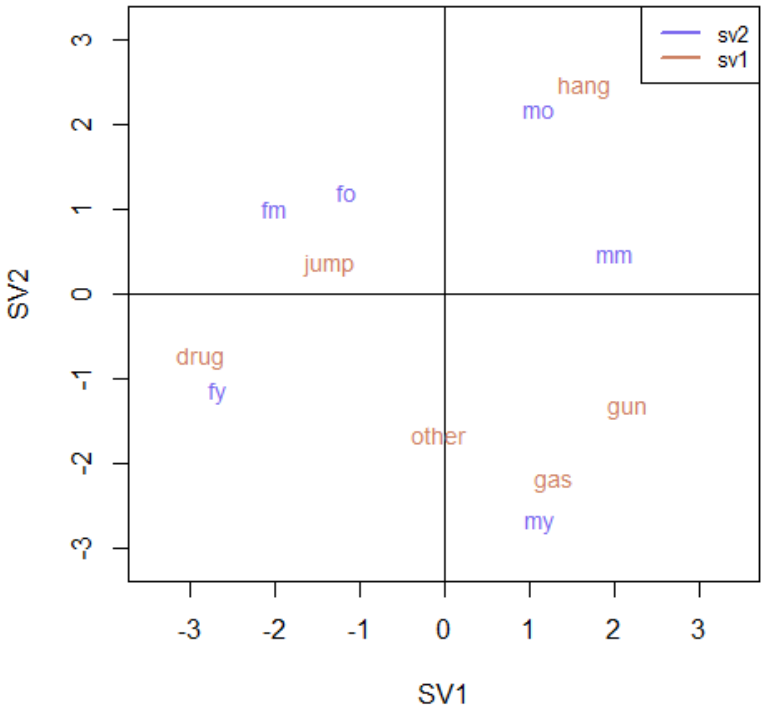
Ergebnis aus der Berechnung der Inertia:

```
> inertia
[1] 622.0849
```

Die Voraussetzung für den Chi-Quadrat-Test (i.e. mindestes erwartete Häufigkeit größer fünf pro Zelle) ist erfüllt. Der Chi-Quadrat-Test für die zweidimensionale Tabelle hat ein signifikantes Ergebnis. Dies lässt auf eine Zusammenhang zwischen den Variablen `cause` und `sex_age` schließen.

```
> chisq.test(tabelle_sui)

        Pearson's Chi-squared test

data:  tabelle_sui
X-squared = 635.96, df = 25, p-value < 2.2e-16
```

Die Ergebnisse der Korrespondenzanalyse können in einem Korrespondenzanalyseplot dargestellt werden:




Im obigen Plot ist erkennbar, dass die Kombination männlich und ältere Gruppe (`mo`) und Methode hang verhältnismäßig öfter als bei Unabhängigkeit vorkommt. Dahingegen kommt die Kombination `mo` und `gas`, `gun` oder `other` nicht sehr oft vor. Die Verteilung der Kombination der Gruppe `female-oldage` und `female-middleage` mit der Methode `jump` ist eher nahe am Koordinatenursprung, d.h. ihre Verteilung ist ähnlich wie die Randverteilung. In der Gruppe `female-youngage` kommt die Kombination mit der Methode `drug` öfter vor als bei Unabhängigkeit. Bei den Gruppen `female-youngage`, `female-oldage` und `female-middleage` kommen die Methoden `other`, `gas` und `gun` generell weniger häufig vor als bei Unabhängigkeit und sie sind nicht nahe der Randverteilung.

## R-Code zu Aufgabe 5 (Dreidimensionale Tabellen):

```
#*********************************************************************************************
# AUFGABE 5 (Dreidimensionale Tabellen)
#*********************************************************************************************
# Im Datensatz suicide in der Library faraway findet man die Ergebnisse von Selbstmorden
# in Großbritannien. Die Kreuzklassifizierungsvariablen sind:
#    . Cause = Methode des Selbstmordes
#    . Age = Altersgruppe (y = young, m = middle, o = old)
#    . Sex = Geschlecht (m = male, f = female)
# Lassen sich bestimmte "Vorlieben" für Methoden in bestimmten Kombinationen von Alter
# und Geschlecht erkennen?
# In diesem Beispiel könnte man auch eine Korrespondenzanalyse für die zweidimensionale
# Tabelle durchführen, die sich aus Methode und der Kombination von Altersgruppe und
# Geschlecht ergibt.


# DATEN LADEN UND DESKRIPTIVES:
data(suicide)
head(suicide, n=5)

summary(suicide)

is.data.frame(suicide)    # ist schon data.frame
is.factor(suicide$cause) # ist schon factor
is.factor(suicide$age)    # ist schon factor
is.factor(suicide$sex)    # ist schon factor

tab.sui <- xtabs(suicide$y ~ suicide$cause + suicide$age + suicide$sex)
tab.sui
summary(tab.sui)

 # MOSAICPLOT:
 par(mfrow=c(1,1))
 mosaicplot(tab.sui, color=c("dodgerblue","dodgerblue3",
                             "dodgerblue4","darkslateblue",
                             "lightblue","lightcyan4","navajowhite3"),
            cex=0.7)

 # LOG-LINEARES MODELL / MODELLE PRUEFEN:
 # zshg. cause - age
 mod1.sui <- xtabs(y ~ cause + age, suicide)
 summary(mod1.sui) # signifikanter zshg.
 round(prop.table(mod1.sui,1), digits=3)

 # zshg. cause - sex
 mod2.sui <- xtabs(y ~ cause + sex, suicide)
 summary(mod2.sui) # signifikanter zshg.
 round(prop.table(mod2.sui,1), digits=3)

 # zshg. age - sex
 mod3.sui <- xtabs(y ~ age + sex, suicide)
 summary(mod3.sui) # signifikanter zshg.
 round(prop.table(mod3.sui,1), digits=3)

 # zshg. age - sex - cause
 mod4.sui <- xtabs(y ~ age + sex + cause, suicide)
 summary(mod4.sui) # signifikanter zshg.
 round(prop.table(mod4.sui,1),digits=3)


# MODELL DER TOTALEN UNABHAENGIGKEIT:
m1.sui <- glm(y ~ age + sex + cause, data = suicide, family = poisson)
summary(m1.sui)
c(deviance(m1.sui), df.residual(m1.sui))
qchisq(0.95,df.residual(m1.sui))
```

```
# MODELL DER 2-FACH-INTERAKTIONEN:
m2.sui <- glm(y ~ (age + sex + cause)^2, data=suicide, family=poisson)
summary(m2.sui)
c(deviance(m2.sui), df.residual(m2.sui))
qchisq(0.95,df.residual(m2.sui))
drop1(m2.sui, test="Chisq")

# MODELL D. BEDINGTEN UNABH. (von cause und age gegeben sex):
m3.sui <- glm(y ~ sex*age + sex*cause, data=suicide, family=poisson)
summary(m3.sui)
c(deviance(m3.sui), df.residual(m3.sui))
qchisq(0.95,df.residual(m3.sui))
drop1(m3.sui, test="Chisq")
par(mfrow=c(2,2))
plot(m3.sui)

# SUCHE NACH EINFACHEREN MODELLEN:
m4.sui <- glm(y ~ age + sex*cause, data=suicide, family=poisson)
summary(m4.sui)
c(deviance(m4.sui), df.residual(m4.sui))
qchisq(0.95,df.residual(m4.sui))
drop1(m4.sui, test="Chisq")

# MODELLVERGLEICH
anova(m3.sui, m4.sui, test="Chisq")


# ZSFG. UEBER VARIABLE sex
# sex variable gesamt
tab.sex <- xtabs(y ~ age + cause, data=suicide)
tab.sex
chisq.test(tab.sex) # signifikant
round(prop.table(tab.sex), digits=2)

# sex=="m"
tab.sex.m <- xtabs(y ~ age + cause, data=suicide,
                   subset=(sex=="m"))
tab.sex.m
fisher.test(tab.sex.m) # voraussetzung nicht erfuellt
chisq.test(tab.sex.m)  # signifikant
summary(tab.sex.m)
addmargins(tab.sex.m)
round(prop.table(tab.sex.m), digits=2)

# sex=="f"
tab.sex.f <- xtabs(y ~ age + cause, data=suicide,
                   subset=(sex=="f"))
tab.sex.f
fisher.test(tab.sex.f) # voraussetzung nicht erfuellt
chisq.test(tab.sex.f)  # signifikant
summary(tab.sex.f)
addmargins(tab.sex.f)
round(prop.table(tab.sex.f), digits=2)


# ZSFG. UEBER VARIABLE age
# age variable gesamt
tab.age <- xtabs(y ~ sex + cause, data=suicide)
tab.age
chisq.test(tab.age) # signifikant
round(prop.table(tab.age), digits=2)
```

```
# age=="y"
tab.age.y <- xtabs(y ~ age + cause, data=suicide,
                   subset=(age=="y"))
tab.age.y
fisher.test(tab.age.y) # voraussetzung nicht erfuellt
chisq.test(tab.age.y)  # signifikant
summary(tab.age.y)
addmargins(tab.age.y)
round(prop.table(tab.age.y), digits=2)

# age=="o"
tab.age.o <- xtabs(y ~ age + cause, data=suicide,
                   subset=(age=="o"))
tab.age.o
fisher.test(tab.age.o) # voraussetzung nicht erfuellt
chisq.test(tab.age.o)  # signifikant
summary(tab.age.o)
addmargins(tab.age.o)
round(prop.table(tab.age.o), digits=2)

# age=="m"
tab.age.m <- xtabs(y ~ age + cause, data=suicide,
                   subset=(age=="m"))
tab.age.m
fisher.test(tab.age.m) # voraussetzung nicht erfuellt
chisq.test(tab.age.m)  # signifikant
summary(tab.age.m)
addmargins(tab.age.m)
round(prop.table(tab.age.m), digits=2)


# ZSFG. UEBER VARIABLE cause
# cause variable gesamt
tab.cause <- xtabs(y ~ sex + age, data=suicide)
tab.cause
chisq.test(tab.cause) # signifikant
round(prop.table(tab.cause), digits=2)

# cause=="drug"
tab.cause.drug <- xtabs(y ~ sex + age, data=suicide,
                   subset=(cause=="drug"))
tab.cause.drug
fisher.test(tab.cause.drug)
chisq.test(tab.cause.drug)  # signifikant
summary(tab.cause.drug)
addmargins(tab.cause.drug)
round(prop.table(tab.cause.drug), digits=2)

# cause=="gas"
tab.cause.gas <- xtabs(y ~ sex + age, data=suicide,
                           subset=(cause=="gas"))
tab.cause.gas
fisher.test(tab.cause.gas)
chisq.test(tab.cause.gas)  # signifikant
summary(tab.cause.gas)
addmargins(tab.cause.gas)
round(prop.table(tab.cause.gas), digits=2)


# cause=="gun"
tab.cause.gun <- xtabs(y ~ sex + age, data=suicide,
                       subset=(cause=="gun"))
tab.cause.gun
fisher.test(tab.cause.gun)
chisq.test(tab.cause.gun)  # signifikant
summary(tab.cause.gun)
addmargins(tab.cause.gun)
round(prop.table(tab.cause.gun), digits=2)
```

```
# cause=="hang"
tab.cause.hang <- xtabs(y ~ sex + age, data=suicide,
                        subset=(cause=="hang"))
tab.cause.hang
fisher.test(tab.cause.hang)
chisq.test(tab.cause.hang)  # signifikant
summary(tab.cause.hang)
addmargins(tab.cause.hang)
round(prop.table(tab.cause.hang), digits=2)

# cause=="jump"
tab.cause.jump <- xtabs(y ~ sex + age, data=suicide,
                        subset=(cause=="jump"))
tab.cause.jump
fisher.test(tab.cause.jump)
chisq.test(tab.cause.jump)  # signifikant
summary(tab.cause.jump)
addmargins(tab.cause.jump)
round(prop.table(tab.cause.jump), digits=2)

# cause=="other"
tab.cause.other <- xtabs(y ~ sex + age, data=suicide,
                         subset=(cause=="other"))
tab.cause.other
fisher.test(tab.cause.other)
chisq.test(tab.cause.other)  # signifikant
summary(tab.cause.other)
addmargins(tab.cause.other)
round(prop.table(tab.cause.other), digits=2)


# ALTERSGRUPPE UND GESCHLECHT ZUSAMMENLEGEN IN 1 FACTOR:
  # kombinationen geschlecht-altersgruppe:
  # m-m, m-o, m-y, f-m, f-o, f-y
m_m <- suicide[suicide$sex=="m" & suicide$age=="m", ]
m_o <- suicide[suicide$sex=="m" & suicide$age=="o", ]
m_y <- suicide[suicide$sex=="m" & suicide$age=="y", ]
f_m <- suicide[suicide$sex=="f" & suicide$age=="m", ]
f_o <- suicide[suicide$sex=="f" & suicide$age=="o", ]
f_y <- suicide[suicide$sex=="f" & suicide$age=="y", ]

combined.df <- data.frame(y=m_m$y, cause=m_m$cause, sex_age=rep("mm",6))
df_mo <- data.frame(y=m_o$y, cause=m_o$cause, sex_age=rep("mo",6))
df_my <- data.frame(y=m_y$y, cause=m_y$cause, sex_age=rep("my",6))
df_fm <- data.frame(y=f_m$y, cause=f_m$cause, sex_age=rep("fm",6))
df_fo <- data.frame(y=f_o$y, cause=f_o$cause, sex_age=rep("fo",6))
df_fy <- data.frame(y=f_y$y, cause=f_y$cause, sex_age=rep("fy",6))
combined.df <- rbind(combined.df, df_mo, df_my, df_fm, df_fo, df_fy)

head(combined.df, n=10)


# KORRESPONDENZANALYSE
# basis
mod.sui <- glm(y ~ cause + sex_age,
               family=poisson, combined.df)
summary(mod.sui)

tabelle_sui <- xtabs(combined.df$y ~ combined.df$cause + combined.df$sex_age)
tabelle_sui

# residuenmatrix
residual_matr <- xtabs(residuals(mod.sui, type="pearson") ~
                         cause + sex_age, combined.df)
residual_matr

# SVD (singular value decomposition)
# fuer erste 2 komponenten
svd_res <- svd(residual_matr,2,2)
svd_res
```

```
# zerlege in u- und v-komponenten
sv1 <- svd_res$u %*% diag(sqrt(svd_res$d[1:2]))
sv2 <- svd_res$v %*% diag(sqrt(svd_res$d[1:2]))

# inertia
inertia <- svd_res$d[1]^2+svd_res$d[2]^2
inertia

# vgl. der inertia mit chi-quadrat
# (vorauss.: erwartetete haeuf. groesser 5) >> hier erfuellt
res <- matrix(numeric(ncol(tabelle_sui)*nrow(tabelle_sui)),
              ncol=ncol(tabelle_sui))
for(i in 1:nrow(tabelle_sui)){
  for(j in 1:ncol(tabelle_sui)){
    res[i,j] <- rowSums(tabelle_sui)[i] *
      colSums(tabelle_sui)[j]/sum(tabelle_sui)
  }
}
all(res>5)

chisq.test(tabelle_sui)
# ergebnis: chi-quadr.-test signifikant
# >> zshg. zwischen cause und sex_age

gesamt <- t(svd_res$d) %*% svd_res$d
gesamt # ergibt ungefaehr selbiges wie chisq teststat.
       # bzw. inertia
# anzeichen fuer korrekte skalierung

# korrespondenzanalyse plot
aa <-1.1 * max(abs(sv1),abs(sv2))
par(mfrow=c(1,1))
plot(rbind(sv1,sv2), asp=1,
     xlim=c(-aa,aa),ylim=c(-aa,aa),
     xlab="SV1",ylab="SV2",type="n")
abline(h=0,v=0)
text(sv2,c("mm","mo","my","fm","fo","fy"), cex=0.9, col="mediumslateblue")
text(sv1,c("drug","gas","gun","hang","jump","other"), cex=0.9, col="lightsalmon3")
legend("topright", cex=0.8, legend=c("sv2","sv1"),
       col=c("mediumslateblue","lightsalmon3"),
       lwd=2, lty=1)

tt <- rbind(sv2,sv1)
tt
```

## Aufgabe 6 (Dreidimensionale Tabellen):

**Die folgenden beiden Tabellen stellen die Überlebenden und Toten beim Untergang der Titanic gegliedert nach Geschlecht, und Passagierklasse dar.**

| Geschlecht | Passagier-klasse | Überleben | |
|---|---|---|---|
| | | nein | ja |
| maennlich | 1. Klasse | 119 | 62 |
| | 2. Klasse | 155 | 25 |
| | 3. Klasse | 422 | 88 |
| | Mannschaft | 670 | 192 |
| weiblich | 1. Klasse | 5 | 141 |
| | 2. Klasse | 14 | 93 |
| | 3. Klasse | 106 | 90 |
| | Mannschaft | 3 | 20 |

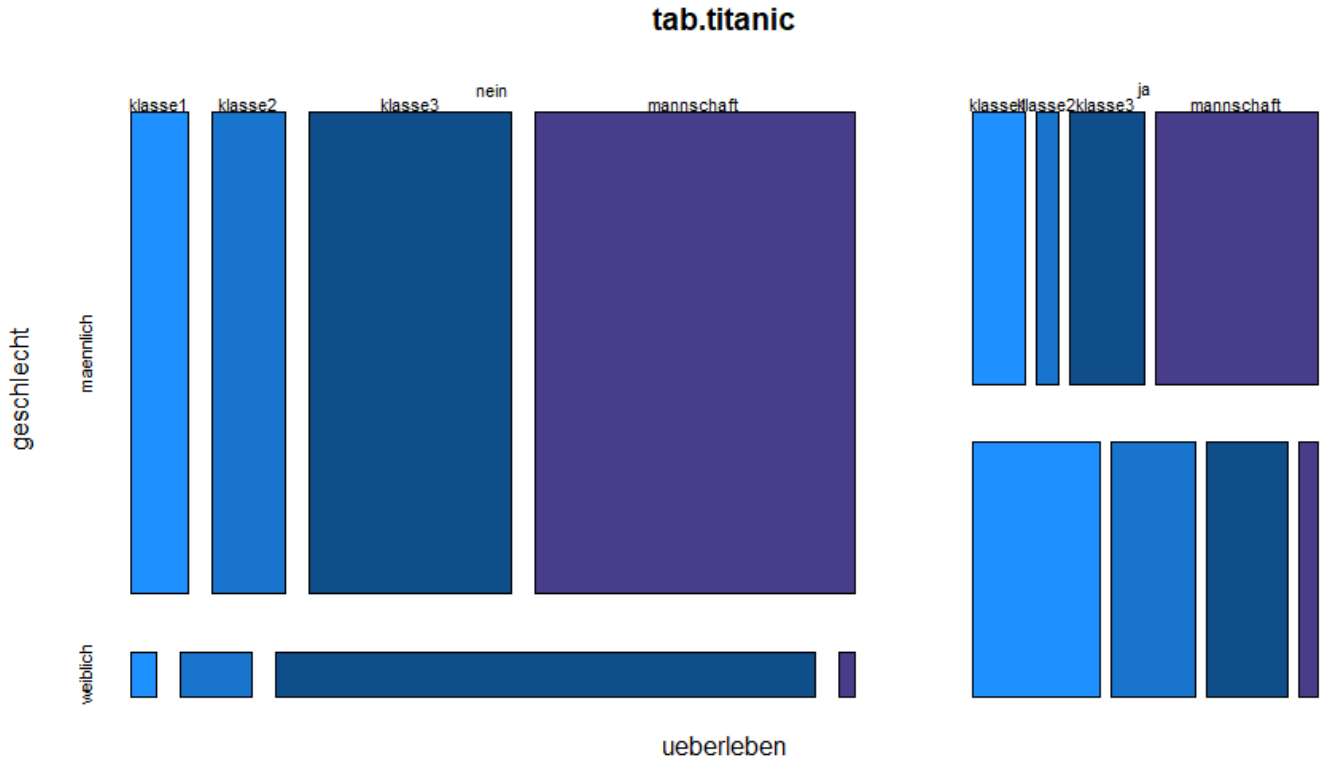
**Man untersuche den Zusammenhang mit einem loglinearen Modell und vergleiche die Ergebnisse mit einer Analyse mit einer logistischen Regression.**

***Tabelle, die aus den Daten der Angabe erstellt wurde:***

```
> titanic
     y ueberleben geschlecht passagierklasse
1  119       nein  maennlich         klasse1
2  155       nein  maennlich         klasse2
3  422       nein  maennlich         klasse3
4  670       nein  maennlich      mannschaft
5   62         ja  maennlich         klasse1
6   25         ja  maennlich         klasse2
7   88         ja  maennlich         klasse3
8  192         ja  maennlich      mannschaft
9    5       nein   weiblich         klasse1
10  14       nein   weiblich         klasse2
11 106       nein   weiblich         klasse3
12   3       nein   weiblich      mannschaft
13 141         ja   weiblich         klasse1
14  93         ja   weiblich         klasse2
15  90         ja   weiblich         klasse3
16  20         ja   weiblich      mannschaft
```

```
> tab.titanic
, , passagierklasse = klasse1

          geschlecht
ueberleben maennlich weiblich
      nein       119        5
      ja          62      141

, , passagierklasse = klasse2

          geschlecht
ueberleben maennlich weiblich
      nein       155       14
      ja          25       93

, , passagierklasse = klasse3

          geschlecht
ueberleben maennlich weiblich
      nein       422      106
      ja          88       90

, , passagierklasse = mannschaft

          geschlecht
ueberleben maennlich weiblich
      nein       670        3
      ja         192       20
```

***Mosaicplot:***




***Log-lineares Modell:***

- Zusammenhang zwischen Überleben und Geschlecht:

```
> summary(mod1.titanic) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht, data = titanic)
Number of cases in table: 2205
Number of factors: 2
Test for independence of all factors:
        Chisq = 453.9, df = 1, p-value = 1.011e-100


> prop.table(mod1.titanic,1)
           geschlecht
ueberleben  maennlich   weiblich
      nein 0.91432396 0.08567604
      ja   0.51617440 0.48382560
```

- Zusammenhang zwischen Passagierklasse und Geschlecht:

```
> summary(mod2.titanic) # signifikant
Call: xtabs(formula = y ~ passagierklasse + geschlecht, data = titanic)
Number of cases in table: 2205
Number of factors: 2
Test for independence of all factors:
        Chisq = 351, df = 3, p-value = 8.991e-76


> prop.table(mod2.titanic,1)
                 geschlecht
passagierklasse maennlich  weiblich
     klasse1    0.5535168 0.4464832
     klasse2    0.6271777 0.3728223
     klasse3    0.7223796 0.2776204
     mannschaft 0.9740113 0.0259887
```

- Zusammenhang zwischen Überleben und Passagierklasse:

```
> summary(mod3.titanic) # signifikant
Call: xtabs(formula = y ~ ueberleben + passagierklasse, data = titanic)
Number of cases in table: 2205
Number of factors: 2
Test for independence of all factors:
        Chisq = 187.38, df = 3, p-value = 2.245e-40
```

- Zusammenhang zwischen Überleben, Geschlecht und Passagierklasse:

```
> summary(mod4.titanic) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht + passagierklasse,
    data = titanic)
Number of cases in table: 2205
Number of factors: 3
Test for independence of all factors:
        Chisq = 1327.6, df = 10, p-value = 4.221e-279


> prop.table(mod4.titanic,1)
, , passagierklasse = klasse1

           geschlecht
ueberleben   maennlich    weiblich
      nein 0.079651941 0.003346720
      ja   0.087201125 0.198312236

, , passagierklasse = klasse2

           geschlecht
ueberleben   maennlich    weiblich
      nein 0.103748327 0.009370817
      ja   0.035161744 0.130801688

, , passagierklasse = klasse3

           geschlecht
ueberleben   maennlich    weiblich
      nein 0.282463186 0.070950469
      ja   0.123769339 0.126582278

, , passagierklasse = mannschaft

           geschlecht
ueberleben   maennlich    weiblich
      nein 0.448460509 0.002008032
      ja   0.270042194 0.028129395
```

***Analyse in den Passagierklassen:***

- Klasse 1:

```
> mod.pk1 <- xtabs(y ~ ueberleben + geschlecht, titanic,
+                  subset=(passagierklasse=="klasse1"))
> mod.pk1
           geschlecht
ueberleben maennlich weiblich
      nein       119        5
      ja          62      141


> summary(mod.pk1) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht, data = titanic,
    subset = (passagierklasse == "klasse1"))
```

```
Number of cases in table: 327
Number of factors: 2
Test for independence of all factors:
        Chisq = 133.33, df = 1, p-value = 7.65e-31
```

- Klasse 2:

```
> mod.pk2 <- xtabs(y ~ ueberleben + geschlecht, titanic,
+                  subset=(passagierklasse=="klasse2"))
> mod.pk2
          geschlecht
ueberleben maennlich weiblich
      nein       155       14
      ja          25       93

> summary(mod.pk2) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht, data = titanic,
    subset = (passagierklasse == "klasse2"))
Number of cases in table: 287
Number of factors: 2
Test for independence of all factors:
        Chisq = 147.82, df = 1, p-value = 5.191e-34
```

- Klasse 3:

```
> mod.pk3 <- xtabs(y ~ ueberleben + geschlecht, titanic,
+                  subset=(passagierklasse=="klasse3"))
> mod.pk3
          geschlecht
ueberleben maennlich weiblich
      nein       422      106
      ja          88       90


> summary(mod.pk3) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht, data = titanic,
    subset = (passagierklasse == "klasse3"))
Number of cases in table: 706
Number of factors: 2
Test for independence of all factors:
        Chisq = 61.69, df = 1, p-value = 4.014e-15
```

- Mannschaft:

```
> mod.pk4 <- xtabs(y ~ ueberleben + geschlecht, titanic,
+                  subset=(passagierklasse=="klasse1"))
> mod.pk4
          geschlecht
ueberleben maennlich weiblich
      nein       119        5
      ja          62      141

> summary(mod.pk4) # signifikant
Call: xtabs(formula = y ~ ueberleben + geschlecht, data = titanic,
    subset = (passagierklasse == "klasse1"))
Number of cases in table: 327
Number of factors: 2
Test for independence of all factors:
        Chisq = 133.33, df = 1, p-value = 7.65e-31
```

***Modell der totalen Unabhängigkeit:***

In diesem Modell sind Überleben „ja“, Geschlecht „weiblich“, Passagierklasse „klasse3“ und „mannschaft“ signifikant.

```
> summary(mod.totaleunabh)

Call:
glm(formula = y ~ ueberleben + geschlecht + passagierklasse,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-15.105   -6.215   -2.306    3.067   16.727

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                5.15982    0.05829  88.523   <2e-16 ***
ueberlebenja              -0.74254    0.04556 -16.298   <2e-16 ***
geschlechtweiblich        -1.30063    0.05192 -25.051   <2e-16 ***
passagierklasseklasse2    -0.13048    0.08089  -1.613    0.107
passagierklasseklasse3     0.76966    0.06689  11.506   <2e-16 ***
passagierklassemannschaft  0.99563    0.06472  15.385   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2532.7  on 15  degrees of freedom
Residual deviance: 1010.6  on 10  degrees of freedom
AIC: 1118.6

Number of Fisher Scoring iterations: 6


> c(deviance(mod.totaleunabh), df.residual(mod.totaleunabh))
[1] 1010.643   10.000


> qchisq(0.95, df.residual(mod.totaleunabh))
[1] 18.30704
```

***Modell der 2-fach-Interaktionen:***

In diesem Modell sind Überleben „ja“, Geschlecht „weiblich“, Passagierklasse „klasse2“, „klasse3, „mannschaft“ sowie die Interaktionen zwischen Geschlecht „weiblich“ und Überleben „ja“ signifikant. Außerdem sind die Interaktionen zwischen Überleben „ja“ und „klasse2“/„klasse3“/“mannschaft“ signifikant. Die Interaktion zwischen Geschlecht „weiblich“ und Passagierklasse „mannschaft“ ist signifikant.

```
> summary(mod.2fach)

Call:
glm(formula = y ~ (ueberleben + geschlecht + passagierklasse)^2,
    family = poisson)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10
11      12
 1.1304  1.0758 -1.3045  0.0947 -1.4252 -2.2521  3.2903 -0.1759 -3.4108 -2.7606  2.9
291 -1.1470
     13      14      15      16
 1.0326  1.4178 -2.6353  0.5713

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)                                4.67368    0.09259  50.476  < 2e-16 ***
ueberlebenja                              -0.37085    0.13522  -2.743  0.00609 **
geschlechtweiblich                        -1.84584    0.16543 -11.158  < 2e-16 ***
passagierklasseklasse2                     0.28207    0.12173   2.317  0.02049 *
passagierklasseklasse3                     1.43416    0.10253  13.988  < 2e-16 ***
passagierklassemannschaft                  1.82994    0.10019  18.264  < 2e-16 ***
ueberlebenja:geschlechtweiblich            2.40352    0.13830  17.379  < 2e-16 ***
ueberlebenja:passagierklasseklasse2       -0.94703    0.19336  -4.898 9.69e-07 ***
ueberlebenja:passagierklasseklasse3       -1.63216    0.16693  -9.778  < 2e-16 ***
ueberlebenja:passagierklassemannschaft    -0.86261    0.15652  -5.511 3.56e-08 ***
geschlechtweiblich:passagierklasseklasse2  0.18633    0.19466   0.957  0.33845
geschlechtweiblich:passagierklasseklasse3  0.10279    0.16876   0.609  0.54248
geschlechtweiblich:passagierklassemannschaft -2.96264    0.25049 -11.828  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2532.672  on 15  degrees of freedom
Residual deviance:   61.603  on  3  degrees of freedom
AIC: 183.58

Number of Fisher Scoring iterations: 5


> qchisq(0.95, df.residual(mod.2fach))
[1] 7.814728
```

Nun wird geschaut, ob das Modell der 2-fach-Interaktionen noch vereinfacht werden kann:

```
> drop1(mod.2fach, test="Chi")
Single term deletions

Model:
y ~ (ueberleben + geschlecht + passagierklasse)^2
                          Df Deviance    AIC    LRT  Pr(>Chi)
<none>                         61.60 183.58
ueberleben:geschlecht      1   418.62 538.60 357.02 < 2.2e-16 ***
ueberleben:passagierklasse 3   165.22 281.20 103.62 < 2.2e-16 ***
geschlecht:passagierklasse 3   401.26 517.24 339.66 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Die Bestandteile sind alle signifikant (auf dem 0.001 Level) und daher wird keine dieser erklärenden Variablen weggelassen.

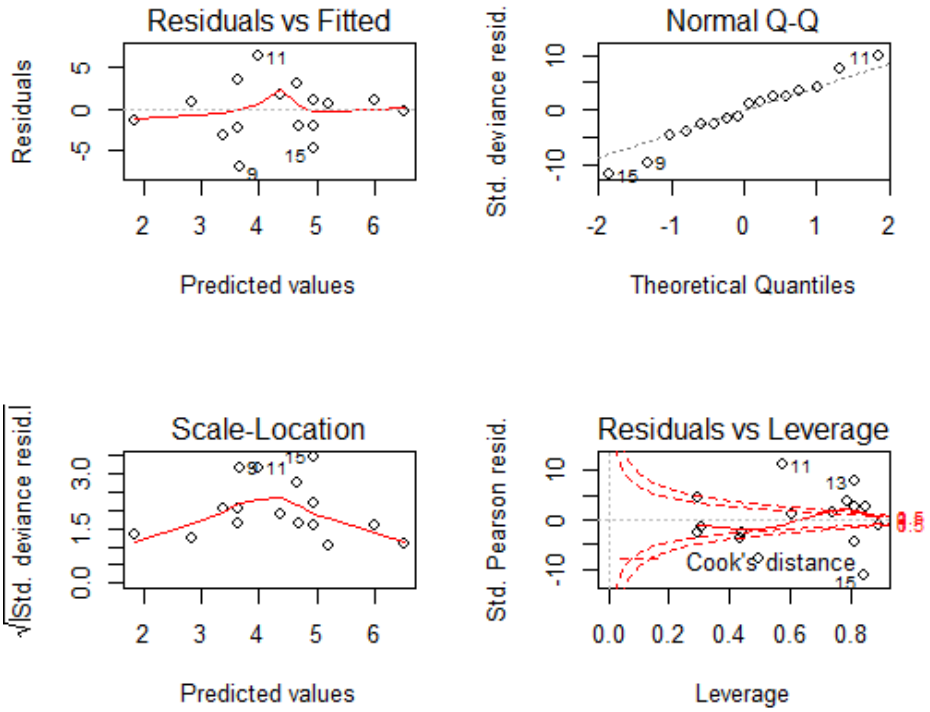
***Modell der bedingten Unabhängigkeit von Überleben und Passagierklasse gegeben Geschlecht:***

```
> summary(mod.bedingt)

Call:
glm(formula = y ~ ueberleben * geschlecht + geschlecht * passagierklasse,
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.9638  -2.0976   0.1651   1.2262   6.3752

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                         4.96053    0.07537  65.820  < 2e-16 ***
ueberlebenja                       -1.31428    0.05880 -22.354  < 2e-16 ***
geschlechtweiblich                 -1.28187    0.13499  -9.496  < 2e-16 ***
passagierklasseklasse2             -0.00554    0.10526  -0.053   0.9580
passagierklasseklasse3              1.03591    0.08652  11.973  < 2e-16 ***
passagierklassemannschaft           1.56076    0.08176  19.089  < 2e-16 ***
```

```
ueberlebenja:geschlechtweiblich               2.30289    0.11906  19.342  < 2e-16 ***
geschlechtweiblich:passagierklasseklasse2    -0.30524    0.16515  -1.848   0.0646 .
geschlechtweiblich:passagierklasseklasse3    -0.74141    0.13942  -5.318 1.05e-07 ***
geschlechtweiblich:passagierklassemannschaft -3.40887    0.23877 -14.277  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2532.67  on 15  degrees of freedom
Residual deviance:  165.22  on  6  degrees of freedom
AIC: 281.2

Number of Fisher Scoring iterations: 5


> qchisq(0.95,df.residual(mod.bedingt))
[1] 12.59159
```

### *Diagnostic Plots für das Modell:*

Der Plot Residuls vs. Fitted zeigt ein unauffälliges Bild, denn die Residuen scheinen weitgehend zufällig um die Nulllinie verstreut zu sein. Im Normal Q-Q Plot ist ersichtlich, dass die theoretischen Quantile nicht ganz den tatsächlichen entsprechen – links unten und rechts oben gibt es mitunter deutlichere Abweichungen von der Gerade. Im Plot Residuals vs. Leverage befinden sich einige Leverage-Punkte, die einerseits mit einer Nummer beziffert sind bzw. außerhalb den roten Linien, die die Cook’s Distance anzeigen, liegen. In Bezug auf diese Punkte sollte noch entschieden werden, ob sie tatsächlich im Modell bleiben sollten.




### *Versuch weiterer Vereinfachungen:*

Hier sind im Ergebnis alle Variablen signifikant. Somit wird keine Variable mehr aus dem Modell genommen.

```
> drop1(mod.bedingt, test="Chi")
Single term deletions

Model:
```

```
y ~ ueberleben * geschlecht + geschlecht * passagierklasse
                           Df Deviance    AIC    LRT  Pr(>Chi)
<none>                          165.22 281.20
ueberleben:geschlecht       1   596.61 710.59 431.39 < 2.2e-16 ***
geschlecht:passagierklasse  3   579.25 689.23 414.03 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(mod.bedingt2)

Call:
glm(formula = y ~ ueberleben + geschlecht * passagierklasse,
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-12.5696   -5.0375    0.0706    3.5581   11.0227

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                                4.80923    0.07577  63.474  < 2e-16 ***
ueberlebenja                              -0.74254    0.04556 -16.298  < 2e-16 ***
geschlechtweiblich                        -0.21489    0.11124  -1.932   0.0534 .
passagierklasseklasse2                    -0.00554    0.10526  -0.053   0.9580
passagierklasseklasse3                     1.03591    0.08652  11.973  < 2e-16 ***
passagierklassemannschaft                  1.56076    0.08176  19.089  < 2e-16 ***
geschlechtweiblich:passagierklasseklasse2 -0.30524    0.16515  -1.848   0.0646 .
geschlechtweiblich:passagierklasseklasse3 -0.74141    0.13942  -5.318 1.05e-07 ***
geschlechtweiblich:passagierklassemannschaft -3.40887    0.23877 -14.277  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2532.67  on 15  degrees of freedom
Residual deviance:  596.61  on  7  degrees of freedom
AIC: 710.59

Number of Fisher Scoring iterations: 6

> qchisq(0.95, df.residual(mod.bedingt2))
[1] 14.06714
```

### *Vergleich von Modellen der bedingten Unabhängigkeit:*

Mittels ANOVA werden die beiden Modelle verglichen:

```
> anova(mod.bedingt,mod.bedingt2)
Analysis of Deviance Table

Model 1: y ~ ueberleben * geschlecht + geschlecht * passagierklasse
Model 2: y ~ ueberleben + geschlecht * passagierklasse
  Resid. Df Resid. Dev Df Deviance
1         6     165.22
2         7     596.61 -1  -431.39
```

### *Untersuchung des Zusammenhangs zwischen Überleben und Passagierklasse:*

Der Zusammenhang ist signifikant.

```
> # zsfg. ueberleben - passagierklasse
> (zsfg_ue_pa<-xtabs(y~ueberleben+passagierklasse, titanic))
           passagierklasse
ueberleben klasse1 klasse2 klasse3 mannschaft
      nein     124     169     528        673
      ja       203     118     178        212

> prop.table(zsfg_ue_pa,1)
```

```
          passagierklasse
ueberleben    klasse1    klasse2    klasse3 mannschaft
      nein 0.08299866 0.11311914 0.35341365 0.45046854
      ja   0.28551336 0.16596343 0.25035162 0.29817159

> summary(zsfg_ue_pa) # signifikant
Call: xtabs(formula = y ~ ueberleben + passagierklasse, data = titanic)
Number of cases in table: 2205
Number of factors: 2
Test for independence of all factors:
        Chisq = 187.38, df = 3, p-value = 2.245e-40
```

***Untersuchung des Zusammenhangs zwischen Überleben und Passagierklasse (für Geschlecht „maennlich“:***

Der Zusammenhang ist signifikant.

```
> (zsfg_ue_pa_m <-xtabs(y~ueberleben+passagierklasse, titanic,
+                       subset=(geschlecht=="maennlich")))
          passagierklasse
ueberleben klasse1 klasse2 klasse3 mannschaft
      nein     119     155     422        670
      ja        62      25      88        192

> prop.table(zsfg_ue_pa_m,1)
          passagierklasse
ueberleben    klasse1    klasse2    klasse3 mannschaft
      nein 0.08711567 0.11346999 0.30893119 0.49048316
      ja   0.16893733 0.06811989 0.23978202 0.52316076

> summary(zsfg_ue_pa_m) # signifikant
Call: xtabs(formula = y ~ ueberleben + passagierklasse, data = titanic,
    subset = (geschlecht == "maennlich"))
Number of cases in table: 1733
Number of factors: 2
Test for independence of all factors:
        Chisq = 29.592, df = 3, p-value = 1.682e-06
```

***Untersuchung des Zusammenhangs zwischen Überleben und Passagierklasse (für Geschlecht „weiblich“:***

Der Zusammenhang ist signifikant.

```
> (zsfg_ue_pa_w <-xtabs(y~ueberleben+passagierklasse, titanic,
+                       subset=(geschlecht=="weiblich")))
          passagierklasse
ueberleben klasse1 klasse2 klasse3 mannschaft
      nein       5      14     106          3
      ja       141      93      90         20

> prop.table(zsfg_ue_pa_w,1)
          passagierklasse
ueberleben    klasse1    klasse2    klasse3 mannschaft
      nein 0.03906250 0.10937500 0.82812500 0.02343750
      ja   0.40988372 0.27034884 0.26162791 0.05813953

> summary(zsfg_ue_pa_w) # signifikant
Call: xtabs(formula = y ~ ueberleben + passagierklasse, data = titanic,
    subset = (geschlecht == "weiblich"))
Number of cases in table: 472
Number of factors: 2
Test for independence of all factors:
        Chisq = 126.54, df = 3, p-value = 3.018e-27
```

***Logistische Regression:***

Es wurden mehrere logistische Regressionen (i.e. volles Modell, additives Modell, Modell mit nur Passagierklasse, Modell mit nur Geschlecht) zur Erklärung der Variable Überleben durchgeführt. Insgesamt kann nach dem AIC-Kriterium argumentiert werden, dass das Modell mit nur dem Geschlecht als erklärende Variable das niedrigste AIC hat, also das „beste“ nach diesem Kriterium wäre.

- Volles Modell:

```
> summary(res.logit.full)

Call:
glm(formula = ueberleben ~ geschlecht * passagierklasse, family = binomial(link = logit),
    data = titanic)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -5.351e-16  1.414e+00       0        1
geschlechtweiblich                       8.157e-16  2.000e+00       0        1
passagierklasseklasse2                  -3.600e-16  2.000e+00       0        1
passagierklasseklasse3                   3.561e-16  2.000e+00       0        1
passagierklassemannschaft                1.256e-15  2.000e+00       0        1
geschlechtweiblich:passagierklasseklasse2  3.975e-16  2.828e+00       0        1
geschlechtweiblich:passagierklasseklasse3 -6.280e-16  2.828e+00       0        1
geschlechtweiblich:passagierklassemannschaft -1.256e-15  2.828e+00       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 22.181  on  8  degrees of freedom
AIC: 38.181

Number of Fisher Scoring iterations: 2
```

- Additives Modell:

```
> summary(res.logit)

Call:
glm(formula = ueberleben ~ geschlecht + passagierklasse, family = binomial(link = logit),
    data = titanic)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             6.625e-16  1.118e+00       0        1
geschlechtweiblich     -1.110e-16  1.000e+00       0        1
passagierklasseklasse2 -6.070e-16  1.414e+00       0        1
passagierklasseklasse3 -8.789e-16  1.414e+00       0        1
passagierklassemannschaft -9.421e-16  1.414e+00       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 22.181  on 11  degrees of freedom
AIC: 32.181

Number of Fisher Scoring iterations: 2
```

- Modell mit nur Passagierklasse:

```
> summary(res.logit2)

Call:
glm(formula = ueberleben ~ passagierklasse, family = binomial(link = logit),
    data = titanic)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -3.914e-16  1.000e+00       0        1
passagierklasseklasse2    4.875e-16  1.414e+00       0        1
passagierklasseklasse3    4.500e-16  1.414e+00       0        1
passagierklassemannschaft 6.280e-16  1.414e+00       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 22.181  on 12  degrees of freedom
AIC: 30.181

Number of Fisher Scoring iterations: 2
```

- Modell mit nur Geschlecht:

```
> summary(res.logit3)

Call:
glm(formula = ueberleben ~ geschlecht, family = binomial(link = logit),
    data = titanic)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.177  -1.177   0.000   1.177   1.177

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.220e-16  7.071e-01       0        1
geschlechtweiblich  4.441e-16  1.000e+00       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 22.181  on 14  degrees of freedom
AIC: 26.181

Number of Fisher Scoring iterations: 2
```

Vergleich des vollen Modells mit res.logit2 und res.logit3:

```
> # vergleiche modelle
> anova(res.logit, res.logit2, test ="Chisq")
Analysis of Deviance Table

Model 1: ueberleben ~ geschlecht + passagierklasse
Model 2: ueberleben ~ passagierklasse
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        11     22.181
2        12     22.181 -1        0        1
> anova(res.logit, res.logit3, test ="Chisq")
Analysis of Deviance Table

Model 1: ueberleben ~ geschlecht + passagierklasse
Model 2: ueberleben ~ geschlecht
  Resid. Df Resid. Dev Df    Deviance Pr(>Chi)
1        11     22.181
2        14     22.181 -3 -3.5527e-15        1
```

## R-Code zu Aufgabe 6 (Dreidimensionale Tabellen):

```
#*******************************************************************************************
# AUFGABE 6 (Dreidimensionale Tabellen)
#*******************************************************************************************
# Die folgenden beiden Tabellen stellen die Überlebenden und Toten beim Untergang
# der Titanic gegliedert nach Geschlecht, und Passagierklasse dar.
# Tabelle: siehe Angabe.

# Man untersuche den Zusammenhang mit einem loglinearen Modell und vergleiche
# die Ergebnisse mit einer Analyse mit einer logistischen Regression.


# daten anlegen und tabelle erstellen
y_titanic <- c(119,155,422,670, 62,25,88,192,
               5,14,106,3, 141,93,90,20)
ueberleben<-gl(2,4,16,labels=c("nein","ja"))
geschlecht<-gl(2,8,16,labels=c("maennlich","weiblich"))
passagierklasse<-gl(4,1,16,labels=c("klasse1","klasse2",
                                    "klasse3","mannschaft"))
titanic<-data.frame(y,ueberleben,geschlecht,passagierklasse)
titanic

tab.titanic <-xtabs(y_titanic ~ ueberleben + geschlecht +
                                passagierklasse)
tab.titanic

# MOSAICPLOT
mosaicplot(tab.titanic, color=c("dodgerblue","dodgerblue3",
                                "dodgerblue4","darkslateblue"))


# LOG-LINEARES MODELL

# zshg. ueberleben - geschlecht
mod1.titanic <- xtabs(y ~ ueberleben + geschlecht, titanic)
summary(mod1.titanic) # signifikant

prop.table(mod1.titanic,1)

# zshg. passagierklasse - geschlecht
mod2.titanic <- xtabs(y ~ passagierklasse + geschlecht, titanic)
summary(mod2.titanic) # signifikant

prop.table(mod2.titanic,1)

# zshg. ueberleben - passagierklasse
mod3.titanic <- xtabs(y ~ ueberleben + passagierklasse, titanic)
summary(mod3.titanic) # signifikant

prop.table(mod3.titanic,1)

# zshg. ueberleben - geschlecht - passagierklasse
mod4.titanic <- xtabs(y ~ ueberleben + geschlecht +
                      passagierklasse, titanic)
summary(mod4.titanic) # signifikant

prop.table(mod4.titanic,1)
```

```
# analyse in den passagierklassen
mod.pk1 <- xtabs(y ~ ueberleben + geschlecht, titanic,
                 subset=(passagierklasse=="klasse1"))
mod.pk1
summary(mod.pk1) # signifikant


mod.pk2 <- xtabs(y ~ ueberleben + geschlecht, titanic,
                 subset=(passagierklasse=="klasse2"))
mod.pk2
summary(mod.pk2) # signifikant


mod.pk3 <- xtabs(y ~ ueberleben + geschlecht, titanic,
                 subset=(passagierklasse=="klasse3"))
mod.pk3
summary(mod.pk3) # signifikant


mod.pk4 <- xtabs(y ~ ueberleben + geschlecht, titanic,
                 subset=(passagierklasse=="klasse1"))
mod.pk4
summary(mod.pk4) # signifikant


# modell der totalen unabhaengigkeit
mod.totaleunabh <- glm(y ~ ueberleben + geschlecht +
                       passagierklasse, family=poisson)

summary(mod.totaleunabh)
c(deviance(mod.totaleunabh), df.residual(mod.totaleunabh))

qchisq(0.95, df.residual(mod.totaleunabh))

# modell der 2fach-interaktionen
mod.2fach <- glm(y ~ (ueberleben + geschlecht +
                      passagierklasse)^2, family=poisson)
summary(mod.2fach)

qchisq(0.95, df.residual(mod.2fach))

# modell vereinfachen
drop1(mod.2fach, test="Chi")

# modell der bedingten unabhaengigkeit
# von ueberleben und passagierklasse
# geg. geschlecht
mod.bedingt <- glm(y ~ ueberleben*geschlecht +
                       geschlecht*passagierklasse,
                   family=poisson)
summary(mod.bedingt)

qchisq(0.95,df.residual(mod.bedingt))

par(mfrow=c(2,2))
plot(mod.bedingt)
```

```
# sonstige vereinfachung beurteilen
drop1(mod.bedingt, test="Chi")

mod.bedingt2 <- glm(y ~ ueberleben +
                      geschlecht*passagierklasse,
                    family=poisson)
summary(mod.bedingt2)

qchisq(0.95, df.residual(mod.bedingt2))

# vgl. modelle der bedingten unabhaengigkeit
anova(mod.bedingt,mod.bedingt2)

# zsfg. hinsichtlich geschlecht

  # zsfg. ueberleben - passagierklasse
(zsfg_ue_pa<-xtabs(y~ueberleben+passagierklasse, titanic))
prop.table(zsfg_ue_pa,1)
summary(zsfg_ue_pa) # signifikant

# fuer geschlecht==maennlich
(zsfg_ue_pa_m <-xtabs(y~ueberleben+passagierklasse, titanic,
                      subset=(geschlecht=="maennlich")))
prop.table(zsfg_ue_pa_m,1)
summary(zsfg_ue_pa_m) # signifikant


# fuer geschlecht==weiblich
(zsfg_ue_pa_w <-xtabs(y~ueberleben+passagierklasse, titanic,
                      subset=(geschlecht=="weiblich")))
prop.table(zsfg_ue_pa_w,1)
summary(zsfg_ue_pa_w) # signifikant


# LOGISTISCHE REGRESSION
res.logit.full <- glm(ueberleben ~ geschlecht*passagierklasse,
                      family=binomial(link=logit), data=titanic)
res.logit.full
summary(res.logit.full)

res.logit <- glm(ueberleben ~ geschlecht + passagierklasse,
                 family=binomial(link=logit), data=titanic)
res.logit
summary(res.logit)

res.logit2 <- glm(ueberleben ~ passagierklasse,
                  family=binomial(link=logit), data=titanic)
res.logit2
summary(res.logit2)

res.logit3 <- glm(ueberleben ~ geschlecht,
                  family=binomial(link=logit), data=titanic)
res.logit3
summary(res.logit3)


 # vergleiche modelle
anova(res.logit, res.logit2, test ="Chisq")
anova(res.logit, res.logit3, test ="Chisq")
```

## Literaturquellen:


- Folien und R-Codes zu den bisher vorgetragenen Kapiteln aus UK Erweiterungen des linearen Modells (Prof. Wilfried Grossmann, 2019).
- Chi-Square-Test Assumptions (2019): http://www.simafore.com/blog/bid/56480/2-key-assumptions-to-be-aware-of-before-applying-the-chi-square-test.
- Chi-Quadrat-Test (2019): https://us.sagepub.com/sites/default/files/upm-binaries/82020_Cha pter_11.pdf.
- John McDonald (2014): Small Numbers in Chi-Square and G-Tests, Handbook of Biological Statistics. http://www.biostathandbook.com/small.html.