Prof. Dr. M. Kateri,
Dr. W. Herff,
L. Kaufmann, M.Sc.

RWTH Aachen, SS 2022
Release date: June 27th, 2022
Solution: Week starting from July 4th, 2022

Applied Data Analysis

R-Laboratory 10

Log-Linear Models

Useful packages and functions:

- `addmargins()`
- `chisq.test()`
- `vcd::mosaic()`
- `rstandard()`
- `vcd`

Task 32

(a) Load the dataset `GSS` as a contingency table into your workspace.

(b) Test whether party affiliation is independent of gender by testing the goodness of fit of the associated log-linear model. What is the corresponding p-value?

(c) For the fitted model in (b), compute the Pearsonian residuals and deviance residuals as well as the corresponding standardized residuals. Create four contingency tables consisting of these residuals.

(d) Create two mosaic plots, one for the standardized Pearsonian and one for the standardized deviance residuals.

Task 33

(a) Load the dataset *accident.csv* into your workspace. First, transform it to a contingency table, then to a data frame, which has a column `Freq` with the counts of all factor combinations in *accident.csv*.

(b) Fit a log-linear model that predicts `Freq` using `speedlimit`, `street year` and their interaction terms and select an appropriate nested model based on AIC.

(c) Compute the Pearsonian residuals and deviance residuals as well as the corresponding standardized residuals for the selected model. Create four contingency tables consisting of these residuals.

(d) Create two mosaic plots, one for the standardized Pearsonian and one for the standardized deviance residuals of the selected model.

Task 34

Consider the following table consisting of data of a study with variables age of mother (A), length of gestation (G) in days, infant survival (I) and number of cigarettes smoked per day during the prenatal period (S). Treat the counts as response variable and A, G, I and S as explanatory variables.

			Infant Survival	
Age	Smoking	Gestation	No	Yes
< 30	< 5	≤ 260	50	315
		> 260	24	4012
	5 +	≤ 260	9	40
		> 260	6	459
30 +	< 5	≤ 260	41	147
		> 260	14	1594
	5 +	≤ 260	4	11
		> 260	1	124

Source: A. Agresti (2002), Categorical Data Analysis, 2nd ed., Exercise 9.2.

(a) Transfer the data into your R workspace.

(b) Explain why a loglinear model should include the AS interaction term.

(c) Fit the models (AGIS),(AGI, AIS, AGS, GIS),(AG,AI,AS,GI,GS,IS) and (AS,G,I).

(d) Use backward selection based on AIC for the models fitted in (c). Compare the results of the strategies and interpret the chosen models.

```
# Task 32 (a)
tab.party=xtabs(count~sex+party, data=GSS)
# Add margins
party.margins = addmargins(tab.party)
party.margins

# (b)
# model of independence
glm.ind = glm(count ~ sex + party, family=poisson, data=GSS)
glm.ind
summary(glm.ind)

# asymptotic p-value
p.val = 1 - pchisq(glm.ind$deviance,df=glm.ind$df.residual)

# Chi-squared Test
chisq.test(tab.party)

# saturated model
glm.sat = glm(count ~ sex * party, family=poisson, data=GSS)
summary(glm.sat)

# (c)
# Pearsonian residuals and corresponding standardized residuals
res.p = residuals(glm.ind, type="pearson")
stdres.p = rstandard(glm.ind, type="pearson")
xtabs(res.p ~ sex + party, data=GSS)
stdres.p.tab = xtabs(stdres.p ~ sex + party, data=GSS)

# Comparison with the statistic of the Chi-squared Test in (b)
sum(res.p^2)
chisq.test(tab.party)$statistic

# deviance residuals and corresponding standardized residuals
res.d = residuals(glm.ind, type="deviance")
stdres.d = rstandard(glm.ind, type="deviance")
xtabs(res.d ~ sex + party, data=GSS)
stdres.d.tab = xtabs(stdres.d ~ sex + party, data=GSS)

# Comparison with the deviance of the log linear model
summary(glm.ind)
sum(res.d^2)

# own implementation
2 * (sum(GSS$count * log(GSS$count/glm.ind$fitted.values)))

# (d) mosaic-plot
# standardized Pearsonian residuals
mosaic(tab.party,gp=shading_Friendly,residuals=stdres.p.tab,
      residuals_type="Std\nresiduals",labeling=labeling_residuals)
# standardized deviance residuals
mosaic(tab.party,gp=shading_Friendly,residuals=stdres.d.tab,
      residuals_type="Std\nresiduals",labeling=labeling_residuals)
```

```
#############
#
# Task 33
#
#############

# (a)
accident = read.csv2("accident.csv",header=TRUE)
accident

# transform to a contingency table
accident.tab = table(accident)
accident.tab
# add margins
addmargins(accident.tab)

# transform to a dataframe
accident.dat = data.frame(accident.tab)
accident.dat

# (b)
# saturated model
glm.sat = glm(Freq ~ speedlimit * street * year, family=poisson, data=accident.dat)
glm.sat
# model selection
glm.select  = step(glm.sat, direction="backward")

# (c)
# Pearsonian residuals and corresponding standardized residuals
res.p = residuals(glm.select,type="pearson")
stdres.p = rstandard(glm.select,type="pearson")
xtabs(res.p ~ speedlimit + street + year, data=accident.dat)
stdres.p.tab = xtabs(stdres.p ~ speedlimit + street + year, data=accident.dat)
stdres.p.tab

# deviance and corresponding standardized residuals
res.d = residuals(glm.select,type="deviance")
stdres.d = rstandard(glm.select,type="deviance")
xtabs(res.d ~ speedlimit + street + year, data=accident.dat)
stdres.d.tab = xtabs(stdres.d ~ speedlimit + street + year, data=accident.dat)
stdres.d.tab

# (d)
library(vcd)
# mosaic plot of standardized Pearsonian residuals
mosaic(accident.tab,gp=shading_Friendly,residuals=stdres.p.tab,
      residuals_type="Std\nresiduals",labeling=labeling_residuals)

# mosaic plot of standardized deviance residuals
mosaic(accident.tab,gp=shading_Friendly,residuals=stdres.d.tab,
      residuals_type="Std\nresiduals",labeling=labeling_residuals)
```

```
################
### Task 34
################

#(a)
freq<- c(50,24,9,6,41,14,4,1,315,4021,40,459,147,1594,11,124)
row<-c(rep(1,4),rep(2,4))
lay1<-rep(1:2,2,each=2)
lay2<-rep(1:2,4)
col<-c(rep(1,8),rep(0,8))

row.lb<-c("<30","30+")
lay1.lb<-c("<5","5+")
lay2.lb<-c("<=260",">260")
col.lb<-c("yes","no")

A<-factor(row,labels=row.lb)
G<-factor(lay2,labels=lay2.lb)
S<-factor(lay1,labels=lay1.lb)
I<-factor(col,labels=col.lb)

data<-data.frame(freq,A,G,S,I)

#(b)
#to observe interaction between A (age of mother) and S (number of cigarettes smoked per day during the prenatal period) we include the interaction term
#as we will see, the selected models with backward selection will include this term

#(c)
sat.model<-glm(freq~A*G*S*I,poisson, data=data) #model (AGIS)

model.three<-glm(freq~A*G*I+A*I*S+A*G*S+G*I*S,poisson, data=data) #model (AGI,AIS,AGS,GIS)
model.two<-glm(freq~A*G+A*I+A*S+G*I+G*S+I*S,poisson, data=data) #model(AG,AI,AS,GI,GS,IS)
model<-glm(freq~A*S+G+I,poisson, data=data) #model(AS,G,I)

#(d)
model.AIC.sat=step(sat.model,direction="backward")
model.AIC.three=step(model.three,direction="backward")
model.AIC.two=step(model.two,direction="backward")
model.AIC=step(model,direction="backward")

#chosen model (except for model.AIC, because we use backward) is model with all two way interactions contained
#this is the model of homogeneous association
```