Socioeconomic Status of Families [nels]

Assignment

Problem
Find out which characteristics are associated with socio-economic status of the family and describe the association.

Population
The data come from the "*National Education Longitudinal Study*" conducted by the U.S. Department of Education. Each record represents a family having a child in the 8th grade of the elementary school in 1988.

Specifications

(i) Investigate the association of socio-economic status of the family (SES quartile) with the achieved level of education of the father. Use standard methods for the analysis of the two-way contingency table first, then fit a loglinear model. Compare the results. In the loglinear model fit, choose one of the main effects and one of the interaction terms and interpret the estimated parameters.

(ii) Investigate the association of socio-economic status (SES quartile) with father's education and region of residence by a loglinear model. Is the association of socio-economic status with father's education the same in every region? If not, describe how it varies in different regions. In which region is the dependence strongest?

(iii) Investigate the association of socio-economic status (SES below/above median) with father's education, father's work status, and region of residence by a logistic regression model fit on grouped data.

(iv) Investigate the association of socio-economic status (SES below/above median) with father's education, father's work status, and region of residence by the loglinear model that is equivalent to the logistic model fitted in (iii). Show that the relevant parameter estimates and test results are the same.

(v) Build a general loglinear model investigating the mutual associations between all the variables in the data set. For socioeconomic status, use the two-level variant (below/above median), not the SES quartiles. Take care not to fit models with too many parameters (some of the models have thousands of parameters and take too long to fit). Interpret the final model – what does it imply about conditional independence of the variables? Choose one of the main effects and one of the two-way interaction terms in the final model and interpret the estimated parameters.

Requirements
Write a report (prepared by LaTeX, LibreOffice, MS Word, ...) summarizing the most important steps of your solution. Formulate a specific answers to the questions of interest.

Mail the report in the pdf format (file named as `Surname_Firstname_4.pdf`) and related R script (file named as `Surname_Firstname_4.R`) to komarek@karlin.mff.cuni.cz.

Deadline: ***Monday April 18, 2022 [06:59 CEST]***.

Dataset

The dataset can be downloaded from
http://www2.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmst432/Problem_4/AdvRegr_4_nels.RData

The dataframe is called `nels`. It contains 13 580 rows (families) and 9 variables.

Variable list: See Table 1.

Table 1: Variable coding table

Variable Name	Variable Label	Variable Coding
`ses`	Socio-economic status (quartile)	1, 2, 3, 4
`sesmed`	Socio-economic status (median)	below, over
`parents`	Number of adults in family	1, 2
`foreign`	Foreign language spoken at home	yes, no
`fa.educ`	Father's education	factor
`mo.educ`	Mother's education	factor
`region`	Region within US	factor
`fa.wrk`	Father working	yes, no
`mo.wrk`	Mother working	yes, no

Instructions, hints

This document was prepared using Sweave (Leisch, 2002) in R (R Core Team, 2022), version 4.1.3 (2022-03-10). Additionally, the contributed packages colorspace (Zeileis et al., 2009, 2019) and xtable (Dahl et al., 2019) were used.

The rest of the document provides commented R code that *partially* solves the assignment. In your report, it is not necessary to provide answers to all points of the assignment (as many answers are already contained in this document). You should certainly provide some statements to points labeled as **TASK FOR YOU:**. Note that not all output is shown in the document below. It is assumed that you run the code by yourself, supplement it by additional commands if needed and use this document only as a guidance through the code and output (that you create). Even if it is not required to provide an answer explicitely in your document, try to answer the question for yourself.

Before going through this, make sure that you at least partly understand materials included in Section 3.3 of the extended course notes. It may also be useful to go, even quickly, through the R tutorial "*Manipulating multi-way contingency tables*" which is available here: `https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2021_22/nmst432/glm_multitables.html`.

Initial operations

```
> setwd("/home/komarek/teach/mff_2021/nmst432_AdvRegr/Problem_4/")
> #
> library("colorspace")
> library("xtable")
> #
> print(load("AdvRegr_4_nels.RData"))
```

Basic exploration of data (marginal frequencies)

```
>   ### Marginal frequencies
> with(nels, table(ses,     useNA = "ifany"))
> with(nels, table(sesmed,  useNA = "ifany"))
> with(nels, table(parents, useNA = "ifany"))
> with(nels, table(foreign, useNA = "ifany"))
> with(nels, table(fa.educ, useNA = "ifany"))
> with(nels, table(mo.educ, useNA = "ifany"))
> with(nels, table(region,  useNA = "ifany"))
> with(nels, table(fa.wrk,  useNA = "ifany"))
> with(nels, table(mo.wrk,  useNA = "ifany"))
```

Problem (i): `ses` vs. `fa.educ`

The main purpose of this section is to complement Sec. 3.3.2 of the (extended) course notes. The code below exemplifies, on real data, interpretation of parameters related to models built above a two-way contingency table. Some issues are explained a bit differently as compared to the course notes in a hope that two (a bit different) views will lead to full understanding of this part which is crucial for use of loglinear models based on multi-way contingency tables. The idea is to convince you that if you understand interpretation of parameters in the linear model being behind the two-way ANOVA, you should easily understand also interpretation of parameters of the loglinear model behind the two-way contingency table. In the rest, (almost) the same notation as in the course notes will be used, variable $X \in \{1, 2, 3, 4 = I\}$ will now refer to `ses`, variable $Z \in \{1, 2, 3 = J\}$ will refer to `fa.educ`.

```
>   ### Contingency table
> (xtab1 <- with(nels, table(ses, fa.educ)))
   fa.educ
ses Elementary High College
  1       1415 1294      97
  2        574 2354     249
  3        200 2343     688
  4         40  851    3475
```

```
>   ### Exploration: column proportions
> prop.table(xtab1, margin = 2)
```

```
>   ### Formatted numbers
> (ptab1 <- round(prop.table(xtab1, margin = 2) * 100, 1))
```

```
>    ### Table in LaTeX
> print(xtable(ptab1, digits = c(0, 1, 1, 1)), floating = FALSE)
```

	Elementary	High	College
1	63.5	18.9	2.2
2	25.8	34.4	5.5
3	9.0	34.2	15.3
4	1.8	12.4	77.1

Is it possible to conclude anything about the association between `ses` and `fa.educ` from above numbers? Does your finding coincide with your expectations (given some "standard" knowledge of the problem)?

```
>     ### Plot (two slightly different versions)
> par(mfrow = c(1, 2), mar = c(3, 3, 3, 1) + 0.1)
> plot(t(xtab1), col = rainbow_hcl(4), main = "SES by father's education")
> plot(ses ~ fa.educ, data = nels, col = rainbow_hcl(4),
  +      main = "SES by father's education")
> par(mfrow = c(1, 1))
```

See Figure 1.

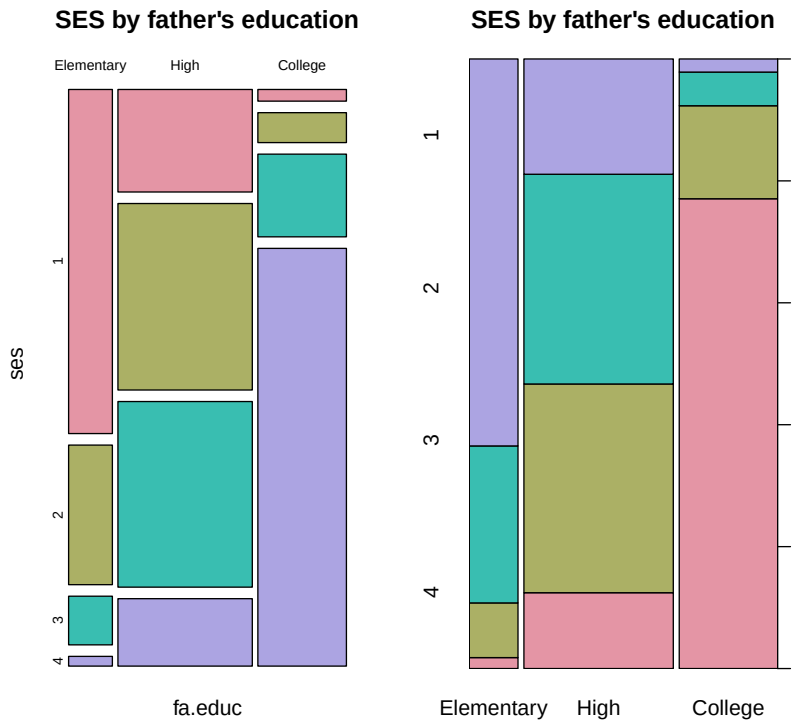

Figure 1: Empirical conditional probabilities of the SES classes given education of father.

Do you have now some idea about association between `ses` and `fa.educ`? What a P-value do you expect from the χ^2 test of independence?

```
> chisq.test(xtab1)
        Pearson's Chi-squared test

data:  xtab1
X-squared = 8664.6, df = 6, p-value < 2.2e-16
```

For loglinear modelling, a `data.frame` is needed with three columns: (1) observed counts $n_{i,j}$ (for each combination of X and Z variable), (2) values of the X variable (`factor`) and (3) values of the Z variable (`factor`). Function `as.data.frame` applied to the contingency table can be used to do this efficiently as it is shown below. Observed counts will be stored in a column named `N`.

```
>   ### Data frame for loglinear modelling
> qq1 <- as.data.frame(xtab1, responseName = "N")
> print(qq1)
   ses    fa.educ    N
1    1 Elementary 1415
2    2 Elementary  574
3    3 Elementary  200
4    4 Elementary   40
5    1       High 1294
6    2       High 2354
7    3       High 2343
```

```
8    4       High  851
9    1    College   97
10   2    College  249
11   3    College  688
12   4    College 3475
```

The above `data.frame` now contains observed counts $n_{i,j}$ in a column labeled `N` (random variables that underlie them will be denoted as $N_{i,j}$) and also (all) possible values of variables X and Z. It is now useful to realize that for loglinear GLM, $N_{i,j}$'s are *response* variables and both X's and Z's *explanatory* variables that will be used to define a linear predictor. Nevertheless, if data sampling mechanism is multinomial or raw/column multinomial, either X or Z (or both) are in fact *response*s.

It is then useful to view a loglinear model for contingency table as a method which *models expected counts* (does not matter whether they are result of multinomial or Poisson sampling) in a structured way (unless saturated model is taken) where each structure (used form of a linear predictor) imposes some *association structure* on the variables that define the contingency table. With a two-way table, basic association structures are (i) "no structure" (interaction model) and (ii) independence.

Nevertheless, for purposes of interpretation of parameters of a loglinear model, it is (almost always) more useful to think about the cell probabilities (now $\pi_{i,j} = \mathsf{P}(X = i,\, Z = j)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$) rather than about the expected counts $m_{i,j}$. We know that $m_{i,j} = m_{++}\, \pi_{i,j}$ where $m_{++} = \mathbb{E}n$ (which is directly equal to n in case of a multinomial sampling). That is, the loglinear model specifies (with $\eta_{i,j}$ being a linear predictor)

$$\begin{aligned} \log(m_{i,j}) &= \eta_{i,j}. \\ \log(\pi_{i,j}) &= -log(m_{++}) + \eta_{i,j}, \\ \pi_{i,j} &= \frac{\mathsf{e}^{\eta_{i,j}}}{m_{++}}. \end{aligned} \tag{1}$$

It is now directly seen that quantities that are given as *a ratio of two cell probabilities* are always expressed as an exponential of difference of two values of a linear predictor, the (expected) total sample size m_{++} will play no role in such a ratio. So as with a linear model, to interprete parameters of the loglinear model, it is sufficient to be able to interprete "difference of two linear predictors" (which is often equal to one specific regression coefficient). And because now also an exponential appears in the expression (1), parameters must be exponentiated (in the same way as with logistic regression) to get above mentioned interpretation of *a ratio of two cell probabilities*.

There is one more crucial fact to realize which is the following:

$$\frac{\mathsf{P}(X = i1,\, Z = j)}{\mathsf{P}(X = i2,\, Z = j)} = \frac{\mathsf{P}(X = i1 \,|\, Z = j)}{\mathsf{P}(X = i2 \,|\, Z = j)} =: \mathsf{odds}_X(i1, i2 \,|\, Z = j) \tag{2}$$

That is, ratio of two *joint* probabilities is also a ratio of two *conditional* probabilities (if we keep a level of one of the margins the same). Hence we can interprete the expression (2) as the (conditional) *odds* on having X equal to $i1$ rather than $i2$ (e.g., reaching the SES status 2 rather than 1) if $Z = j$ (e.g., if education of father is *High*).

And we can continue in making more complex proportions:

$$\frac{\frac{\mathsf{P}(X=i1,\,Z=j1)}{\mathsf{P}(X=i2,\,Z=j1)}}{\frac{\mathsf{P}(X=i1,\,Z=j2)}{\mathsf{P}(X=i2,\,Z=j2)}} = \frac{\frac{\mathsf{P}(X=i1\,|\,Z=j1)}{\mathsf{P}(X=i2\,|\,Z=j1)}}{\frac{\mathsf{P}(X=i1\,|\,Z=j2)}{\mathsf{P}(X=i2\,|\,Z=j2)}} = \frac{\mathsf{odds}_X(i1,i2\,|\,Z=j1)}{\mathsf{odds}_X(i1,i2\,|\,Z=j2)}$$

$$=: \mathsf{OR}_X\big(i1,i2\,\big|\,Z:j1\leftrightarrow j2\big). \quad (3)$$

That is, if we go back to the loglinear model, exponential of a difference of two differences of two values of a linear predictor provides a ratio of two odds related to the *effect* of the X variable if we compare two groups given by the value of the Z variable, that is, it is certain *odds ratio*. From symmetry, expression (3) is also equal to the "reversed" odds ratio, i.e.,

$$\mathsf{OR}_X\big(i1,i2\,\big|\,Z:j1\leftrightarrow j2\big) = \mathsf{OR}_Z\big(j1,j2\,\big|\,X:i1\leftrightarrow i2\big).$$

As will be shown below, reasonable "difference of two differences of two values of a linear predictor" is often equal to the regression coefficient related to some interaction term. In that case, exponential of the respective regression coefficient will provide certain odds ratio.

Saturated model that is the model that purely somehow parameterizes the expected counts (or cell probabilities) takes the form[1]

$$\log(m_{i,j}) = \beta_0 + \beta_i^X + \beta_j^Z + \beta_{i,j}^{XZ}, \qquad i=1,\ldots,I,\ j=1,\ldots,J.$$

Clearly, the model has "too many" parameters and some identifiability constraints must be imposed on them. One possibility is the following set of constraints:

$$\beta_1^X = 0, \qquad \beta_1^Z = 0, \qquad \beta_{1,j}^{XZ} = 0 \text{ for each } j, \qquad \beta_{i,1}^{XZ} = 0 \text{ for each } i. \quad (4)$$

As with linear model, constraints can be incorporated in the estimation procedure by using properly specified model matrix based on parameterization of categorical variables X and Z by suitable (pseudo)contrasts. Not surprisingly, constraints (4) are achieved by using the "*reference group*" pseudocontrasts, i.e., by using dummy variables for both X and Z while leaving one of the dummies for each variable out of the model. In R, as with the linear model, the reference group pseudocontrasts (`contr.treatment`) are used automatically as soon as both involved variables are `factors`:

```
>   ### Saturated model
> fit1 <- glm(N ~ (ses + fa.educ)^2, family = poisson, data = qq1)
> fit1 <- glm(N ~ ses + fa.educ + ses:fa.educ, family = poisson, data = qq1)
>                       ## the same as above
> summary(fit1)         ### Surprised by (numerically) zero residual deviance?
Call:
glm(formula = N ~ ses + fa.educ + ses:fa.educ, family = poisson,
    data = qq1)

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

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         7.25488    0.02658 272.903   <2e-16 ***
```

[1] In contrast to the course notes, β_0 will be used instead of α to denote the intercept term.

```
ses2                -0.90226    0.04949 -18.233   <2e-16 ***
ses3                -1.95657    0.07554 -25.900   <2e-16 ***
ses4                -3.56601    0.16033 -22.241   <2e-16 ***
fa.educHigh         -0.08939    0.03846  -2.324   0.0201 *
fa.educCollege      -2.68017    0.10496 -25.536   <2e-16 ***
ses2:fa.educHigh     1.50063    0.06039  24.851   <2e-16 ***
ses3:fa.educHigh     2.55026    0.08310  30.688   <2e-16 ***
ses4:fa.educHigh     3.14692    0.16630  18.924   <2e-16 ***
ses2:fa.educCollege  1.84500    0.12952  14.245   <2e-16 ***
ses3:fa.educCollege  3.91565    0.13217  29.625   <2e-16 ***
ses4:fa.educCollege  7.14464    0.19054  37.498   <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: 1.1494e+04  on 11  degrees of freedom
Residual deviance: 4.7251e-13  on  0  degrees of freedom
AIC: 122.87

Number of Fisher Scoring iterations: 2
```

Independence model is obtained by skipping the interaction terms:

```
>   ### Independence model
> fit0 <- glm(N ~ ses + fa.educ, family = poisson, data = qq1)
> summary(fit0)
Call:
glm(formula = N ~ ses + fa.educ, family = poisson, data = qq1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-35.078  -30.629   -7.907   16.860   45.005

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     6.13247    0.02704 226.765  < 2e-16 ***
ses2            0.12418    0.02591   4.793 1.64e-06 ***
ses3            0.14103    0.02580   5.465 4.62e-08 ***
ses4            0.44209    0.02420  18.272  < 2e-16 ***
fa.educHigh     1.12153    0.02439  45.987  < 2e-16 ***
fa.educCollege  0.70452    0.02589  27.210  < 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: 11494.5  on 11  degrees of freedom
Residual deviance:  8649.7  on  6  degrees of freedom
```

```
AIC: 8760.5

Number of Fisher Scoring iterations: 5
```

And now three **tests of indepence** (of X and Z):

```
>    ### Deviance (likelihood-ratio) test of independence
> anova(fit0, fit1, test = "LRT")
Analysis of Deviance Table

Model 1: N ~ ses + fa.educ
Model 2: N ~ ses + fa.educ + ses:fa.educ
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6     8649.7
2         0        0.0  6   8649.7 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
>    ### Score test of independence
> chisq.test(xtab1)
        Pearson's Chi-squared test

data:  xtab1
X-squared = 8664.6, df = 6, p-value < 2.2e-16
```

```
>    ### Yes, the classical chi-sq test of independence
>    ### is the score test in the corresponding loglinear model
> anova(fit0, fit1, test = "Rao")
Analysis of Deviance Table

Model 1: N ~ ses + fa.educ
Model 2: N ~ ses + fa.educ + ses:fa.educ
  Resid. Df Resid. Dev Df Deviance    Rao  Pr(>Chi)
1         6     8649.7
2         0        0.0  6   8649.7 8664.9 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A (small) difference between the value of the test statistic reported by `chisq.test` and by `anova(, test = "Rao")` is caused by the fact that `chisq.test` uses a (well-known) closed form expression of the test statistic, whereas `anova(, test = "Rao")` calculates the test statistic numerically as it does not know that in this situation a closed form expression exists.

```
>    ### Also the Wald test can be considered
> (be1 <- coef(fit1))
       (Intercept)               ses2               ses3               ses4
         7.2548848         -0.9022554         -1.9565674         -3.5660054
       fa.educHigh      fa.educCollege    ses2:fa.educHigh    ses3:fa.educHigh
```

```
       -0.08939134         -2.68017383          1.50063323          2.55026141
  ses4:fa.educHigh ses2:fa.educCollege ses3:fa.educCollege ses4:fa.educCollege
        3.14692401          1.84499733          3.91564530          7.14464414
```

```
> V1 <- vcov(fit1)
> (interIndex <- grep(":fa.educ", names(be1)))
[1]  7  8  9 10 11 12
```

```
> W <- as.numeric(be1[interIndex] %*% solve(V1[interIndex, interIndex], be1[interIndex]))
> pW <- pchisq(W, df = length(interIndex), lower.tail = FALSE)
> Wald  <- data.frame(W = W, df = length(interIndex), Pvalue = pW)
> print(Wald)
         W df Pvalue
1 5514.305  6      0
```

Interpretation of coefficients, saturated model

Here are (exponentiated) values of the fitted coefficients of the saturated models (intercept excluded):

```
> #exp(coef(fit1)[-1])
> round(exp(coef(fit1)[-1]), 2)    ### Meaning of each coefficient?
              ses2               ses3               ses4        fa.educHigh
              0.41               0.14               0.03               0.91
    fa.educCollege   ses2:fa.educHigh   ses3:fa.educHigh   ses4:fa.educHigh
              0.07               4.48              12.81              23.26
ses2:fa.educCollege ses3:fa.educCollege ses4:fa.educCollege
              6.33              50.18            1267.30
```

Before we proceed, let us check ordering of the levels of both "explanatory" variables and also let us check, which level is taken as a reference one by the fitting function.

```
>     ### Check levels of the two factors
> with(nels, table(ses,     useNA = "ifany"))
ses
   1    2    3    4
2806 3177 3231 4366
```

```
> with(nels, table(fa.educ, useNA = "ifany"))
fa.educ
Elementary       High    College
      2229       6842       4509
```

```
> contr.treatment(3)
  2 3
1 0 0
2 1 0
3 0 1
```

Yes, the first level of each factor is a reference (only zeros in the first row of above matrix).

Effects of X, coefficients β_i^X and $\beta_{i,j}^{XZ}$

As we all (hopefully) know from linear regression, with the *reference group* pseudocontrasts parameterization, β_i^X provides effect of changing X from reference to group i under the condition that **the second variable (Z) takes a reference value**. With our model, "effect of changing X" is quantified by the corresponding odds (2) as soon as the coefficient is exponentiated. That is,

$$\exp(\beta_i^X) = \mathsf{odds}_X(i, 1 \,|\, Z = 1) = \mathsf{odds}_{\mathtt{ses}}(i, 1 \,|\, \mathtt{fa.educ} = \mathtt{Elementary}), \qquad i = 2, 3, 4. \tag{5}$$

Similarly,

$$\exp(\beta_{i1}^X - \beta_{i2}^X) = \mathsf{odds}_X(i1, i2 \,|\, Z = 1) = \mathsf{odds}_{\mathtt{ses}}(i1, i2 \,|\, \mathtt{fa.educ} = \mathtt{Elementary}), \qquad i1,\, i2 = 2, 3, 4.$$

Further, as we also (hopefully) know, interaction terms *modify* the effect of one variable depending on the value of the other variable. With the reference group pseudocontrasts, $\beta_{i,j}^{XZ}$ gives "correction" of the "slope" for change of X from reference to class i if $Z = j$. Hence (see also detailed derivation in the course notes):

$$\mathsf{odds}_X(i, 1 \,|\, Z = j) = \exp(\beta_i^X + \beta_{i,j}^{ZX}), \qquad i = 2, \ldots, 4,\ j = 2, 3. \tag{6}$$

An arbitary odds related to the effect of X is then

$$\mathsf{odds}_X(i2, i2 \,|\, Z = j) = \exp(\beta_{i1}^X - \beta_{i2}^X + \beta_{i1,j}^{ZX} - \beta_{i2,j}^{ZX}), \qquad i1, i2 = 1, 2, 3, 4,\ j = 1, 2, 3,$$

where zeros from the identifiability constraints (4) must be taken into account.

Combination of expressions (5) and (6) now gives interpretation of the interaction regression coefficients:

$$\exp(\beta_{i,j}^{XZ}) = \frac{\mathsf{odds}_X(i, 1 \,|\, Z = j)}{\mathsf{odds}_X(i, 1 \,|\, Z = 1)} = \mathsf{OR}_X\big(i, 1 \,\big|\, Z : j \leftrightarrow 1\big).$$

Odds ratio related to arbitrary combination of the X and Z values is then clearly given as (again while taking zeros from the identifiability constraints into account)

$$\mathsf{OR}_X\big(i1, i2 \,\big|\, Z : j1 \leftrightarrow j2\big) = \frac{\mathsf{odds}_X(i1, i2 \,|\, Z = j1)}{\mathsf{odds}_X(i1, i2 \,|\, Z = j2)} = \exp(\beta_{i1,j1}^{XZ} - \beta_{i2,j1}^{XZ} - \beta_{i1,j2}^{XZ} + \beta_{i2,j2}^{XZ}),$$
$$i1, i2 = 1, 2, 3, 4,\ j1, j2 = 1, 2, 3.$$

In the same way, with obvious changes of indeces, **effects of Z** which is based on coefficients β_j^Z and $\beta_{i,j}^{XZ}$, can be quantified.

TO REMEMBER:	• Main effect $\longrightarrow$ certain (conditional) ODDS; • Interaction term $\longrightarrow$ certain ODDS RATIO.

NOTE 1: Any other parameterization of the two categorical variables can be used (sum contrasts, ...). The only thing which changes will be interpretation of the regression coefficients.

NOTE 2: If the **independence** model holds, the interaction coefficients are all equal to zero and all odds ratios are equal to one. The main effects then provide not only odds in the reference group but odds in all groups. That is, under **independence**

$$\exp(\beta_i^X) = \mathsf{odds}_X(i, 1 \,|\, Z = j), \qquad i = 2, 3, 4,\ j = 1, 2, 3,$$
$$\exp(\beta_{i1}^X - \beta_{i2}^X) = \mathsf{odds}_X(i1, i2 \,|\, Z = j), \qquad i1, i2 = 2, 3, 4,\ j = 1, 2, 3.$$

And here, estimated **odds on a better SES** (as compared to the lowest SES equal to 1) in the groups according to father's education (based on the saturated model):

```
>   ### ODDS on better ses (compared to ses = 1)
>   ### given father's education
>   ### ------------------------------------
> #
>     ### fa.educ = Elementary
> be1[paste("ses", 2:4, sep = "")]
> (oddsElem <- exp(be1[paste("ses", 2:4, sep = "")]))
> #
>     ### fa.educ = High
> be1[paste("ses", 2:4, sep = "")]
> be1[paste("ses", 2:4, ":fa.educHigh", sep = "")]
> (oddsHigh <- exp(be1[paste("ses", 2:4, sep = "")] +
  +                  be1[paste("ses", 2:4, ":fa.educHigh", sep = "")]))
> #
>     ### fa.educ = College
> be1[paste("ses", 2:4, sep = "")]
> be1[paste("ses", 2:4, ":fa.educCollege", sep = "")]
> (oddsColl <- exp(be1[paste("ses", 2:4, sep = "")] +
  +                  be1[paste("ses", 2:4, ":fa.educCollege", sep = "")]))
```

```
>     ### All in one table
> ODDSbetterSES <- data.frame(Elementary = oddsElem, High = oddsHigh,
  +                           College = oddsColl)
> print(ODDSbetterSES)
     Elementary      High   College
ses2 0.40565371 1.8191654  2.567010
ses3 0.14134276 1.8106646  7.092784
ses4 0.02826855 0.6576507 35.824742
```

In agreement with our findings from the descriptive part of the analysis (and probably also with our expectations from the subject matter knowledge), with the *College* education (multiplicative) difference between probabilities (i.e., the odds) in being in higher SES classes rather than in low SES classes is much more profound than with *High* or even only *Elementary* education. You can also notice that if X and Z were independent, the theoretical (population) counterparts of the odds in the above table should show three identical columns.

Reversely, we can calculate **odds on better education** (as compared to the *Elementary* one) in the four SES classes:

```
>   ### ODDS on better education (compared to fa.educ = Elementary)
>   ### given family SES
>   ### ------------------------------------------------------------------------
> #
>     ### ses = 1
> (odds1 <- exp(be1[paste("fa.educ", c("High", "College"), sep = "")]))
> #
```

```
>     ### ses = 2
> (odds2 <- exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +               be1[paste("ses2:fa.educ", c("High", "College"), sep = "")]))
> #
>     ### ses = 3
> (odds3 <- exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +               be1[paste("ses3:fa.educ", c("High", "College"), sep = "")]))
> #
>     ### ses = 4
> (odds4 <- exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +               be1[paste("ses4:fa.educ", c("High", "College"), sep = "")]))
```

```
>     ### All in one table
> ODDShigherEduc <- data.frame(ses1 = odds1, ses2 = odds2, ses3 = odds3, ses4 = odds4)
> print(ODDShigherEduc)
                     ses1      ses2   ses3   ses4
fa.educHigh    0.91448763 4.1010453 11.715 21.275
fa.educCollege 0.06855124 0.4337979  3.440 86.875
```

In agreement with previous findings, in a "low level society", father's education is rather *Elementary* whereas in a "high level society", *High* or even *College* education is more likely than *Elementary* education.

And finally, we can also provide estimated **odds ratios**.

```
>   ### ODDS RATIOS (ratios of odds on higher ses compared to ses = 1
>   ###   when comparing higher levels of education with Elementary one)
>   ###
>   ### = ODDS RATIOS (ratios of odds on higher level of education
>   ###   compared to Elementary one)
>   ###   when comparing higher ses with ses = 1)
>   ###
>   ### ---------------------------------------------------------------------------
> exp(be1[grep(":fa.educ", names(be1))])
   ses2:fa.educHigh    ses3:fa.educHigh    ses4:fa.educHigh ses2:fa.educCollege
           4.484528           12.810452           23.264393            6.328083
ses3:fa.educCollege ses4:fa.educCollege
          50.181443         1267.300258
```

```
>   ### ODDS RATIOS (ratios of odds on higher ses compared to ses = 1
>   ###   when comparing College education with High education)
>   ###
>   ### = ODDS RATIOS (ratios of odds on College education compared to High education
>   ###   when comparing higher ses with ses = 1
>   ### ---------------------------------------------------------------------------
> exp(be1[grep(":fa.educCollege", names(be1))] - be1[grep(":fa.educHigh", names(be1))])
ses2:fa.educCollege ses3:fa.educCollege ses4:fa.educCollege
           1.411092            3.917227           54.473815
```

Looking for trends

If either of the two variables X and Z is ordinal, it may make sense to look for trends in evolution of the respective log-odds by replacing the (pseudo)contrast parameterization of either of the two variables by some parameterization of numeric variables based on a suitable set of scores. For instance, the SES categories could reasonably be represented by scores 1, 2, 3, 4 and the following model could then be considered:

$$\log(m_{i,j}) = \beta_0 + \beta^X\, i + \beta_j^Z + \beta_j^{XZ}\, i, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J. \tag{7}$$

In R, such a model is simply fitted by replacing the `factor` version of `ses` by its numeric counterpart having values 1, 2, 3, 4:

```
>   ### Are fit0 and fit1 the only reasonable models in this situation?
>   ###
>   ### How about the model below?
>   ### Does it have reasonable interpretation?
>   ### ---------------------------------------------------------------------
> qq1 <- transform(qq1, nses = as.numeric(ses))
> print(qq1)
> summary(qq1)
```

```
> fit1n <- glm(N ~ (nses + fa.educ)^2, family = poisson, data = qq1)
> summary(fit1n)
Call:
glm(formula = N ~ (nses + fa.educ)^2, family = poisson, data = qq1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-18.604   -4.204    1.396    2.928   16.342

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         8.31000    0.04603  180.52   <2e-16 ***
nses               -1.02727    0.02742  -37.47   <2e-16 ***
fa.educHigh        -0.67311    0.05425  -12.41   <2e-16 ***
fa.educCollege     -5.63103    0.09965  -56.51   <2e-16 ***
nses:fa.educHigh    0.94879    0.02948   32.18   <2e-16 ***
nses:fa.educCollege 2.38798    0.03625   65.87   <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: 11494.5  on 11  degrees of freedom
Residual deviance:  1103.3  on  6  degrees of freedom
AIC: 1214.2

Number of Fisher Scoring iterations: 4
```

```
> be1n <- coef(fit1n)
```

And how about the interpretation of coefficients? It is simple, the coefficient β_X now expresses a change

in odds related to a unity change of X when Z is from the reference category, i.e.,

$$\exp(\beta^X) = \mathsf{odds}_X(i+1, i \,|\, Z = 1) = \mathsf{odds}_{\mathtt{ses}}(i+1, i \,|\, \mathtt{fa.educ} = \mathtt{Elementary}), \qquad i = 1, 2, 3.$$

The interaction coefficient β_j^{XZ} then modifies the "slope" β_X if $Z = j$, that is,

$$\exp(\beta^X + \beta_j^{XZ}) = \mathsf{odds}_X(i+1, i \,|\, Z = j). \qquad i = 1, 2, 3,\ j = 2, 3.$$

And again, $\exp(\beta_j^{XZ})$ is a certain odds ratio, namely,

$$\exp(\beta_j^{XZ}) = \frac{\mathsf{odds}_X(i+1, i \,|\, Z = j)}{\mathsf{odds}_X(i+1, i \,|\, Z = 1)} = \mathsf{OR}_X\big(i+1, i \,\big|\, Z : j \leftrightarrow 1\big), \qquad i = 1, 2, 3,\ j = 2, 3.$$

If the trend model (7) holds, the estimated odds on higher SES in the three categories by father's education would be as follows (in the code, we also compare it to estimates based on the saturated model)

```
>   ### Odds on better ses (by 1)
>   ### ------------------------------------
> #
>     ### fa.educ = Elementary
> exp(be1n["nses"])
> #
> (oddsnElem <- exp(be1n["nses"] * 1:3))  ## trend
> exp(be1[paste("ses", 2:4, sep = "")])   ## saturated
> #
>   ### fa.educ = High
> exp(be1n["nses"] + be1n["nses:fa.educHigh"])
> #
> (oddsnHigh <- exp((be1n["nses"] + be1n["nses:fa.educHigh"]) * 1:3))  ## trend
> exp(be1[paste("ses", 2:4, sep = "")] +                               ## saturated
  +     be1[paste("ses", 2:4, ":fa.educHigh", sep = "")])
> #
>     ### fa.educ = College
> exp(be1n["nses"] + be1n["nses:fa.educCollege"])
> #
> (oddsnColl <- exp((be1n["nses"] + be1n["nses:fa.educCollege"]) * 1:3)) ## trend
> exp(be1[paste("ses", 2:4, sep = "")] +                                 ## saturated
  +     be1[paste("ses", 2:4, ":fa.educCollege", sep = "")])
```

All odds in a table:

```
>     ### All in one table
> ODDSnbetterSES <- data.frame(Elementary = oddsnElem, High = oddsnHigh,
  +                              College = oddsnColl)
> #
>     ### Compare
> print(ODDSnbetterSES)          ## SES ordinal (numeric)
  Elementary      High   College
1 0.35798319 0.9245240  3.898963
2 0.12815197 0.8547446 15.201909
3 0.04587625 0.7902319 59.271677
```

```
> print(ODDSbetterSES)           ## SES nominal
     Elementary      High   College
ses2 0.40565371 1.8191654  2.567010
ses3 0.14134276 1.8106646  7.092784
ses4 0.02826855 0.6576507 35.824742
```

When comparing the odds estimated using the two models, it seems that the linear trend does not really fit well for *High* and *College* education.

But we can also test whether the linear trend is appropriate, e.g., by the deviance test:

```
>   ### Is (saturated) fit1 significantly better than fit1n?
>   ### ---------------------------------------------------
> anova(fit1n, fit1, test = "LRT")
Analysis of Deviance Table

Model 1: N ~ (nses + fa.educ)^2
Model 2: N ~ ses + fa.educ + ses:fa.educ
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1         6     1103.3
2         0        0.0  6   1103.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Apparently, the trend model is significantly worse than the saturated model. Given the comparison of the two odds sets above, we also have idea why it is so.

Of course, also odds related to father's education can be compared when estimated using the two models.

```
>   ### Odds on better education (compared to fa.educ = Elementary)
>   ### ------------------------------------------------------------------
> #
>     ### ses = 1
> (oddsn1 <- exp(be1n[paste("fa.educ", c("High", "College"), sep = "")] +
  +                be1n[paste("nses:fa.educ", c("High", "College"), sep = "")]))
> #
> exp(be1[paste("fa.educ", c("High", "College"), sep = "")])   ## Compare (saturated)
> #
>     ### ses = 2
> (oddsn2 <- exp(be1n[paste("fa.educ", c("High", "College"), sep = "")] +
  +                2 * be1n[paste("nses:fa.educ", c("High", "College"), sep = "")]))
> #
> exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +     be1[paste("ses2:fa.educ", c("High", "College"), sep = "")])
> #
>     ### ses = 3
> (oddsn3 <- exp(be1n[paste("fa.educ", c("High", "College"), sep = "")] +
  +                3 * be1n[paste("nses:fa.educ", c("High", "College"), sep = "")]))
> #
```

```
> exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +     be1[paste("ses3:fa.educ", c("High", "College"), sep = "")])
> #
>     ### ses = 4
> (oddsn4 <- exp(be1n[paste("fa.educ", c("High", "College"), sep = "")] +
  +                4 * be1n[paste("nses:fa.educ", c("High", "College"), sep = "")]))
> #
> exp(be1[paste("fa.educ", c("High", "College"), sep = "")] +
  +     be1[paste("ses4:fa.educ", c("High", "College"), sep = "")])
```

```
>     ### All in one table
> ODDSnhigherEduc <- data.frame(ses1 = oddsn1, ses2 = oddsn2, ses3 = oddsn3,
  +                               ses4 = oddsn4)
> #
>     ### Compare
> print(ODDSnhigherEduc)          ## SES ordinal (numeric)
                     ses1      ses2     ses3     ses4
fa.educHigh    1.31743002 3.4023822 8.786960 22.69312
fa.educCollege 0.03904468 0.4252539 4.631639 50.44535
```

```
> print(ODDShigherEduc)           ## SES nominal
                     ses1      ses2   ses3   ses4
fa.educHigh    0.91448763 4.1010453 11.715 21.275
fa.educCollege 0.06855124 0.4337979  3.440 86.875
```

Goodness-of-fit test

With a loglinear model based on a contingency table, the saturated model has a reasonable interpretation (no structure imposed on the cell probabilities). Comparison of the model M to the saturated model by the deviance test can then be interpreted also as a goodness-of-fit test of model M. Nevertheless, especially with multi-dimensional tables, we should be aware of problems with asymptotic properties of the test for which the sample size n is not the critical value. Remember that for validity of asymptotic arguments behind the classical χ^2 test of independence (which is a score test for certain loglinear model), it is requested to have sufficiently high (magical number 5 often appears here) *expected* counts under the independence model. Similarly, for asymptotics in comparison of two loglinear models, sufficiently high *expected* counts (= fitted values) are needed in all cells of the underlying table under the null hypothesis model.

Look at the code below and later use it with understanding that asymptotics may not always work.

```
> ##### (Useful?) function to calculate
> ##### a goodness-of-fit test
> ##### - perhaps useful for multi-dimensional tables
> ##### ============================================
> gof <- function(m){
  +   DD <- deviance(m)
  +   df <- m$df.residual
  +   pval <- pchisq(DD, df, lower.tail = FALSE)
  +   nparm <- length(coef(m))
```

```
  +
  +   LowCount <- sum(fitted(m) <= 5)
  +
  +   cat("Goodness-of-fit test, model with ", nparm, " parameters\n", sep = "")
  +   cat("Deviance = ", DD, ", df = ", df, "\n", sep = "")
  +   cat("P-value: ", ifelse(pval < 0.001, "<0.001", format(round(pval, 3), nsmall = 3)), "\n
  +
  +   if (LowCount){
  +     cat("Number of cells with low fitted counts: ", LowCount, "\n\n")
  +     print(summary(fitted(m)))
  +   }
  + }
```

```
> gof(fit1n)  ## Test of a linear trend for X again
Goodness-of-fit test, model with 6 parameters
Deviance = 1103.284, df = 6
P-value: <0.001
```

```
> #
> gof(fit1)      ### Hmmmm...
Goodness-of-fit test, model with 12 parameters
Deviance = 4.725109e-13, df = 0
P-value: <0.001
```

```
>      ### Do you have a better name for chi^2 distribution
>      ### with 0 degrees of freedom?
>      ### Correct p-value:
>      pchisq(0, df = 0, lower.tail = FALSE)
[1] 1
```

```
> #
> gof(fit0)    ## Test of independence again
Goodness-of-fit test, model with 6 parameters
Deviance = 8649.664, df = 6
P-value: <0.001
```

Problem (ii): `ses` **vs.** `fa.educ` **vs.** `region`

Purpose of the code below is again to exemplify, on real data, interpretation of parameters related to models built above a three-way contingency table. That is, this section complements Section 3.3.3 of the course notes. Variable $X \in \{1, 2, 3, 4 = I\}$ again refers to `ses`, variable $Z \in \{1, 2, 3 = J\}$ again refers to `fa.educ` and finally, variable $V \in \{1, 2, 3, 4 = K\}$ refers to `region`.

First, we create the corresponding (three-way) contingency table and also related `data.frame`.

```
> (xtab2 <- with(nels, table(ses, fa.educ, region)))
, , region = Northeast

   fa.educ
ses Elementary High College
  1        160  206      17
  2         84  385      37
  3         31  404     137
  4         10  155     831

, , region = Midwest

   fa.educ
ses Elementary High College
  1        299  421      28
  2        134  862      69
  3         39  658     189
  4         10  215     866

, , region = South

   fa.educ
ses Elementary High College
  1        687  480      39
  2        245  748      83
  3         78  807     214
  4         17  313    1140

, , region = West

   fa.educ
ses Elementary High College
  1        269  187      13
  2        111  359      60
  3         52  474     148
  4          3  168     638
```

```
> qq2 <- as.data.frame(xtab2, responseName = "N")
```

Saturated model

We start by considering a saturated loglinear model which basically just some parameterization of joint probabilities $\pi_{i,j,k} = \mathsf{P}(X = i, Z = j, V = k)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. The saturated model is specified (for related expected counts $m_{i,j,k}$) as

$$\log(m_{i,j,k}) = \beta_0 + \beta_i^X + \beta_j^Z + \beta_k^V + \beta_{i,j}^{XZ} + \beta_{i,k}^{XV} + \beta_{j,k}^{ZV} + \beta_{i,j,k}^{XZV}, \qquad i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K.$$

For identifiability reasons, the following set of constraints can be considered

$$\begin{aligned}
&\beta_1^X = 0, \quad \beta_1^Z = 0, \quad \beta_1^V = 0,\\
&\beta_{1,j}^{XZ} = 0 \text{ for each } j, \quad \beta_{i,1}^{XZ} = 0 \text{ for each } i,\\
&\beta_{1,k}^{XV} = 0 \text{ for each } k, \quad \beta_{i,1}^{XV} = 0 \text{ for each } i,\\
&\beta_{1,k}^{ZV} = 0 \text{ for each } k, \quad \beta_{j,1}^{ZV} = 0 \text{ for each } j,\\
&\beta_{1,j,k}^{XZV} = 0 \text{ for each } j \text{ and } k, \quad \beta_{i,1,k}^{XZV} = 0 \text{ for each } i \text{ and } k, \quad \beta_{i,j,1}^{XZV} = 0 \text{ for each } i \text{ and } j.
\end{aligned}$$

The model under the above constraints is again easily fitted if the *reference group* pseudocontrast parameterization (`contr.treatment`) is used for each of the three "explanatory" variables:

```
> fit2 <- glm(N ~ (ses + fa.educ + region)^3, family = poisson, data = qq2)
> summary(fit2)
```

Before we proceed, we check whether the model could be simplified by dropping the three-way interaction term.

```
> drop1(fit2, test = "LRT")
Single term deletions

Model:
N ~ (ses + fa.educ + region)^3
                    Df Deviance    AIC   LRT Pr(>Chi)
<none>                    0.00 421.66
ses:fa.educ:region 18    32.89 418.55 32.89  0.01721 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we now see, the three-way interaction term cannot be omitted. Consequently, no "simpler" association structure for the three variables exist. This will further be exemplified once we express (estimated) association parameters. Before we do that, let us briefly repeat what you know from Sec. 3.3.5 of the course notes. Perhaps not everything mentioned below is explicitely stated in the course notes, nevertheless, it is assumed that the course notes were read first.

Main effects

It can easily be find (using similar derivations as in case of two-way table) that the main effects (with the *reference group* pseudocontrasts) lead to *conditional odds* that relate conditional probability of reaching level i (or j or k) of variable X (or Z or V) versus the reference level 1 of that variable given the condition that the remaining variables are set to their reference. That is, for example for the variable X:

$$\exp(\beta_i^X) = \frac{\mathsf{P}(X = i \,|,\, Z = 1, V = 1)}{\mathsf{P}(X = 1 \,|,\, Z = 1, V = 1)} =: \mathsf{odds}_X(i, 1 \,|\, Z = 1, V = 1), \qquad i = 2, \ldots, I. \tag{8}$$

Two-way interactions

Again, using similar derivations as in case of two-way table, we find that each two-way interaction "modifies" (multiplicatively after being exponentiated) related conditional odds. For example,

$$\exp(\beta_i^X + \beta_{i,j}^{XZ}) = \frac{\mathsf{P}(X = i\,|,\, Z = j,\, V = 1)}{\mathsf{P}(X = 1\,|,\, Z = j,\, V = 1)} =: \mathsf{odds}_X(i,\, 1 \,|\, Z = j,\, V = 1),$$
$$i = 2, \ldots, I,\ j = 2, \ldots, J.$$

That is, each two-way interaction leads to a ratio of two conditional odds where in the condition, one of the conditioning variables is not set to the reference as it is in (8). Namely,

$$\exp(\beta_{i,j}^{XZ}) = \frac{\mathsf{odds}_X(i,\, 1 \,|\, Z = j,\, V = 1)}{\mathsf{odds}_X(i,\, 1 \,|\, Z = 1,\, V = 1)} =: \mathsf{OR}_X\big(i,\, 1 \,\big|\, Z : j \leftrightarrow 1,\, V = 1\big),$$
$$i = 2, \ldots, I,\ j = 2, \ldots, J.$$

From symmetry, it is also

$$\exp(\beta_{i,j}^{XZ}) = \frac{\mathsf{odds}_Z(j,\, 1 \,|\, X = i,\, V = 1)}{\mathsf{odds}_Z(j,\, 1 \,|\, X = 1,\, V = 1)} =: \mathsf{OR}_Z\big(j,\, 1 \,\big|\, X : i \leftrightarrow 1,\, V = 1\big),$$
$$i = 2, \ldots, I,\ j = 2, \ldots, J.$$

That is, each two-way interaction term leads to certain conditional odds ratio where we (a) compare two conditional probabilities of one variable (odds) when (b) changing the second variable and (c) setting the third variable to reference.

Let us now calculate estimated conditional odds ratios that compare odds on different `ses` levels if `fa.educ` changes (or compare odds on different `fa.educ` if the `ses` level changes) given the reference `region` which is `NorthEast`. Check the estimated coefficients first and related P-values (those are now not really of interest):

```
> (be2 <- coef(fit2))
> pv2 <- summary(fit2)$coefficients[, "Pr(>|z|)"]
> (pv2 <- round(pv2, 3))
```

```
>   ### Odds ratios (odds on higher ses compared to ses = 1) when comparing higher
>   ### educations with elementary ones
>   ### = odds ratios (odds on higher education compared to elementary one) when
>   ###   comparing higher ses with ses = 1
>   ### --> conditional ones (given region)
> #
>   ## Region = Northeast (reference)
>   ## ----------------------------
> (lorNE <- c(be2[grep("^ses[0-9]:fa.educHigh$",    names(be2))],
+             be2[grep("^ses[0-9]:fa.educCollege$", names(be2))]))
> (orNE <- exp(lorNE))
> (orNE <- matrix(round(orNE, 2), nrow = 3))
> rownames(orNE) <- paste("ses", 2:4, sep = "")
> colnames(orNE) <- paste("educ", c("High", "College"))
```

```
> print(orNE)
```

```
     educ High educ College
ses2      3.56         4.15
ses3     10.12        41.59
ses4     12.04       782.12
```

That is, for example, estimated odds on SES $= 2$ as compared to SES $= 1$ is in the *NorthEast* 3.56 times higher for fathers with *High* education when comparing them to fathers with *Elementary education*. Or reversely, estimated odds on *High* education as compared to `Elementary` education is in the *NorthEast* 3.56 times higher for people with SES $= 2$ when comparing them to people with SES $= 1$.

Three-way interactions

As we can easily find and as it is also explained in Sec. 3.3.5 of the course notes, each three-way interaction "modifies" (multiplicatively after being exponentiated) related conditional odds ratio. In particular,

$$\exp(\beta^{XYZ}_{i,j,k}) = \frac{\mathsf{OR}_X\big(i,\,1\,\big|\,Z:j\leftrightarrow 1,\,V=k\big)}{\mathsf{OR}_X\big(i,\,1\,\big|\,Z:j\leftrightarrow 1,\,V=1\big)} = \frac{\mathsf{OR}_Z\big(j,\,1\,\big|\,X:i\leftrightarrow 1,\,V=k\big)}{\mathsf{OR}_Z\big(j,\,1\,\big|\,X:i\leftrightarrow 1,\,V=1\big)} = \cdots,$$
$$i=2,\ldots,I,\ j=2,\ldots,J,\ k=2,\ldots,K. \quad (9)$$

TASK FOR YOU: Calculate and also provide in your report a table with six conditional *odds ratios* that quantify association between `ses` and `fa.educ` in *MidWest* (similar table as that one above calculated for *NorthEast*). For yourself, calculate the same table also for *South* and *West* and then try to comment (in 2–3 sentences) in which region the association between SES and father's education is the strongest and why.

```
>   ## Region = Midwest
>   ## ------------------------------
> ### Try by yourself and report in your document.
> ### Some three-way interactions must be involved in your calculations ;-)
> #
>   ## Region = South
>   ## ------------------------------
> ### Try by yourself, no need to report in your document.
> #
>   ## Region = West
>   ## ------------------------------
> ### Try by yourself, no need to report in your document.
```

Finally, we explicitely calculate estimates of changes in conditional odds ratios, i.e., estimates of quantities given by (9). On top of that, we extract related P-values.

First, we calculate it for $k = 2$. By doing that, we in fact compare association structures `ses`:fa.educ between the second region (*MidWest*) and the reference region (*NorthEast*).

```
> ### Three-way interactions: changes of above conditional odds ratios
>   ### when comparing given region with Northeast (reference)
>   ### ---------------------------------------------------------------
> #
>     ### Midwest - Northeast (reference)
>     ### ++++++++++++++++++++++++++++++
```

```
> (orChangeMW <- exp(be2[grep("^ses[0-9]:fa.educ(College|High):regionMidwest",
  +                           names(be2))]))
   ses2:fa.educHigh:regionMidwest    ses3:fa.educHigh:regionMidwest
                         1.283386                         1.183799
   ses4:fa.educHigh:regionMidwest ses2:fa.educCollege:regionMidwest
                         1.268362                         1.326368
ses3:fa.educCollege:regionMidwest ses4:fa.educCollege:regionMidwest
                         1.244172                         1.182385
```

```
> (pvChangeMW <- pv2[grep("^ses[0-9]:fa.educ(College|High):regionMidwest",
  +                       names(pv2))])
   ses2:fa.educHigh:regionMidwest    ses3:fa.educHigh:regionMidwest
                            0.212                            0.548
   ses4:fa.educHigh:regionMidwest ses2:fa.educCollege:regionMidwest
                            0.619                            0.487
ses3:fa.educCollege:regionMidwest ses4:fa.educCollege:regionMidwest
                            0.601                            0.762
```

None of conditional OR's is significantly different between *MidWest* and *NorthEast* (let alone if we adjust for multiple comparison). That is, associations between `ses` and `fa.educ` in *MidWest* do not differ significantly from associations in *NorthEast*.

Similary, we can compare *West* and *NorthEast*:

```
>     ### West - Northeast (reference)
>     ### ++++++++++++++++++++++++++++++
> (orChangeWE <- exp(be2[grep("^ses[0-9]:fa.educ(College|High):regionWest",
  +                           names(be2))]))
   ses2:fa.educHigh:regionWest    ses3:fa.educHigh:regionWest
                      1.306917                       1.295428
   ses4:fa.educHigh:regionWest ses2:fa.educCollege:regionWest
                      6.691358                       2.698011
ses3:fa.educCollege:regionWest ses4:fa.educCollege:regionWest
                      1.415916                       5.626473
```

```
> (pvChangeWE <- pv2[grep("^ses[0-9]:fa.educ(College|High):regionWest",
  +                       names(pv2))])
   ses2:fa.educHigh:regionWest    ses3:fa.educHigh:regionWest
                         0.214                          0.349
   ses4:fa.educHigh:regionWest ses2:fa.educCollege:regionWest
                         0.005                          0.030
ses3:fa.educCollege:regionWest ses4:fa.educCollege:regionWest
                         0.449                          0.024
```

Some of OR's significantly differ between *West* and *Northeast* (even if MCP is taken into account by Bonferroni procedure). That is, associations between `ses` and `fa.educ` in *West* differ significantly from associations in *NortEast*.

We can also easily compare *South* and *NorthEast*:

```
>     ### South and Northeast (reference)
>     ### ++++++++++++++++++++++++++++++
>     ### Try by yourself, just for yourself.
```

With some effort (differences between two three-way interactions would have to be involved) also any other pair of regions can be compared.

```
>     ### Also Midwest-South, Midwest-West, South-West comparisons can be done...
>     ### ++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
>     ### Would you know how?
>     ### Try by yourself, just for yourself.
```

TO REMEMBER:
- Main effect $\longrightarrow$ certain (conditional) ODDS;
- Two-way interaction term $\longrightarrow$ certain (conditional) ODDS RATIO;
- Three-way interaction term $\longrightarrow$ certain ratio of two (conditional) ODDS RATIOs.

Association structures implied by different models

In our analyzis, the saturated model could not be simplified, i.e., the three considered variables exhibit a general association structure. Especially for point (v) of the assignment, it may be useful to summarize other possible association structures that we may encounter, see also Section 3.3.5. On this place, however, associations will be discussed from a bit different perspective. We now concentrate on conditional associations implied by different models for two "more important" variables (let's say X and Z) under the presence of the third variable V. In other words, we now concentrate on association aspects of the conditional distribution of (X, Z) given V being implied by different models. In particular, we explain what different models imply for the following two sets of odds ratios:

$$\mathsf{OR}_X(i1, i2 \,|\, Z : j1 \leftrightarrow j2,\, V = k) = \frac{\mathsf{odds}_X(i1,\, i2 \,|\, Z = j1,\, V = k)}{\mathsf{odds}_X(i1,\, i2 \,|\, Z = j2,\, V = k)}, \tag{10}$$

$$\mathsf{OR}_Z(j1, j2 \,|\, X : i1 \leftrightarrow i2,\, V = k) = \frac{\mathsf{odds}_Z(j1,\, j2 \,|\, X = i1,\, V = k)}{\mathsf{odds}_Z(j1,\, j2 \,|\, X = i2,\, V = k)},$$
$$i1,\, i2 = 1, \ldots, I,\; j1,\, j2 = 1, \ldots, J,\; k = 1, \ldots, K.$$

Different models will be described symbolically in a form of "full" R formula.

Saturated model $X + Z + V + X:Z + X:V + Z:V + X:Z:V$

As expressions (9) show, the odds ratios (10) depend on k, a value of the conditioning variable V and in general are not necessarily equal to one. That is, conditionally, given V, X and Z are dependent and the form of their association varies with V. This can also be seen if we rewrite the symbolic description of the model as

$$V + \big(X + X(V)\big) + \big(Z + Z(V)\big) + \big(X:Z + X:Z(V)\big),$$

which, if V (by which we condition) is considered as a "constant" leads to saturated (dependence) model based on a two-way table determined by X and Z where, however, association structure, given by the interaction terms, depends on that "constant".

Model without the three-way interaction $X + Z + V + X:Z + X:V + Z:V$

If the three-way interaction is not present in the model, all odds ratio changes (9) are equal to one. From here, we can easily derive that the odds ratios (10) are not necessarily one but do not depend on k, the value of the conditioning variable V. That is,

$$\begin{aligned} &\mathsf{OR}_X(i1, i2 \,|\, Z : j1 \leftrightarrow j2,\ V = 1) = \cdots = \mathsf{OR}_X(i1, i2 \,|\, Z : j1 \leftrightarrow j2,\ V = K), \\ &\mathsf{OR}_Z(j1, j2 \,|\, X : i1 \leftrightarrow i2,\ V = 1) = \cdots = \mathsf{OR}_Z(j1, j2 \,|\, X : i1 \leftrightarrow i2,\ V = K), \\ &\qquad\qquad i1,\ i2 = 1, \ldots, I,\ j1,\ j2 = 1, \ldots, J. \end{aligned} \tag{11}$$

That is, conditionally, given V, X and Z are dependent but the form of their association is the same for all levels of the third variable V (does not depend on V). This can also be seen if we rewrite the symbolic description of the model as

$$V + \big(X + X(V)\big) + \big(Z + Z(V)\big) + X:Z.$$

Now, if V is considered as a "constant" we again obtain a saturated (dependence) model based on a two-way table determined by X and Z where, however, association structure, given by the interaction terms, does not depend on that "constant". We also say that X and Z exhibit **homogeneous associations** given V.

Model that includes $X:Z$ interaction

Once the model becomes either of the following
$$\begin{aligned} &X + Z + V + X:Z \\ &X + Z + V + X:Z + X:V \\ &X + Z + V + X:Z + Z:V \end{aligned}$$

That is, it includes the $X:Z$ interaction (plus perhaps other two-way interactions), we again obtain **homogeneous associations** for X and Z given V. If the model is just $X + Z + V + X:Z$ then, of course, even stronger statement holds that (X, Z) and V are independent in which case, all conditional odds ratios in one row of (11) are not only equal but also equal to unconditional odds ratios.

Model that does not include $X:Z$ interaction

Once the model becomes either of the following
$$\begin{aligned} &X + Z + V + X:V + Z:V \\ &X + Z + V + X:V \\ &X + Z + V + Z:V \\ &X + Z + V \end{aligned}$$

That is, if it does not include the $X:Z$ interaction but other two-way interactions are perhaps included, the model can symbolically be written as
$$\begin{aligned} &V + \big(X + X(V)\big) + \big(Z + Z(V)\big) \\ &V + \big(X + X(V)\big) + Z \\ &V + X + \big(Z + Z(V)\big) \\ &V + X + Z \end{aligned}$$

which suggests (and can also be derived rigorously, see the course notes) that in this case, given V, variables X and Z are conditionally *independent* and all conditional odds ratios (10) are equal to one. If the model is just $X + Z + V$ then, of course, even stronger statement holds that (X, Z) and V are independent in which case, all conditional odds ratios in one row of (11) are not only equal to one but also equal to unconditional odds ratios (which are also all equal to one).

TO REMEMBER: In the following hierarchically well formulated (HWF) model is always assumed.

- $X:Z:V$ in the model

$\longrightarrow$ conditional association between X and Z depends on V;

- $X:Z:V$ not in the model but $X:Z$ included

$\longrightarrow$ homogeneous conditional association between X and Z given V;

- $X:Z$ interaction not in the model

$\longrightarrow$ conditional independence of X and Z given V.

NOTE: Once it is possible to assume homogeneous associations (no three-way interaction in the model), conditional independence of X and Z given V is tested by testing significance of the $X:Z$ term.

Problem (iii): `sesmed` vs. `fa.educ`, `fa.wrk` and `region` by logistic regression

This problem is just application of logistic regression...

TASK FOR YOU: While using "standard" model building tools, develop a reasonable logistic model with `sesmed` as response and remaining three variables as covariates. In your report, state which model have you developed. There is no need to report all estimated parameters. Only briefly (in 2–3 sentences) explain what the model implies for dependence of `sesmed` on the remaining variables.

Since all explanatory variables are categorical and there are only three of them, we could reasonably start with the saturated model:

```
> fit3 <- glm(sesmed ~ (fa.educ + fa.wrk + region)^3, family = binomial, data = nels)
> drop1(fit3, test = "LRT")
Single term deletions

Model:
sesmed ~ (fa.educ + fa.wrk + region)^3
                     Df Deviance   AIC    LRT Pr(>Chi)
<none>                    13187 13235
fa.educ:fa.wrk:region  6    13202 13238 15.429  0.01717 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It looks like the three-way interaction is significant. But, to let you exercise a bit, forget about model `fit3` and start your model building from a model that includes all two-way interactions but not the three-way interaction term. That is, assume that the following model is "correct" one and perhaps can further be simplified.

```
> fit3alt <- glm(sesmed ~ (fa.educ + fa.wrk + region)^2, family = binomial, data = nels)
> drop1(fit3alt, test = "LRT")
Single term deletions

Model:
sesmed ~ (fa.educ + fa.wrk + region)^2
               Df Deviance   AIC     LRT  Pr(>Chi)
<none>               13202 13238
fa.educ:fa.wrk  2    13207 13239  4.3755 0.1121662
fa.educ:region  6    13225 13249 22.6062 0.0009397 ***
fa.wrk:region   3    13205 13235  2.7833 0.4262546
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now, continue by yourself to derive a model which will not be significantly worse than model `fit3alt`. Given the above output, this does not have to be a difficult task... Report the final model.

Problem (iv): `sesmed` vs. `fa.educ`, `fa.wrk` and `region` by loglinear model

Let us first create a `data.frame` for loglinear modelling:

```
> (xtab4 <- with(nels, table(sesmed, fa.educ, fa.wrk, region)))
> qq4 <- as.data.frame(xtab4, responseName = "N")
```

TASK FOR YOU: Read (again) carefully Section 3.3.7 of the course notes and state (symbolically in a form of the R formula in your report) the loglinear model which is equivalent to the final logistic model developed in the previous part. By fitting the model, check by yourself whether your solution is correct. Estimates (and other quantities) of equivalent regression coefficient should really be the same. Possible rounding error is perhaps present on some distant decimal place, certainly not on the first, second, third or fourth one. Some discrepances observed there are a sign that your solution is not correct...

Problem (v): Analysis of a multi-way table

Background and motivation

Categorical data are quite frequent in social sciences where almost nothing can be "exactly" measured and data usually come from various questionnaires. The first thing which is (should be) of interest for a social scientist is to reveal a nature of the association structure among the available variables (items in a questionnaire). To do that, statistician may help by developing a reasonable (= still fitting the data but being as simple as possible) loglinear model and then by explaining what the model implies for associations to a social scientist. A model that both social scientits and also statistician may understand is a (hierarchically well formulated) model that includes at most the three-way interactions which implies that

1. Variable X is *independent* of all other variables if X appears as a main effect only in the model. Of course, in this case X is also conditionally independent with any other variable given remaining variables;
2. Two variables X and Z are *conditionally* (given all other variables) *independent* if $X:Z$ interaction is not included in the model but both X and Z appear in some other two-way (or three-way) interactions;
3. Two variables X and Z exhibit *homogeneous conditional associations* (given all other variables) if $X:Z$ interaction is present in the model but there is no three-way interaction in which $X:Z$ is nested
4. Conditional association (given other variables) of X and Z changes with (some) of other variables if there is at least one three-way interaction of a type $X:Z:V$ in the model.

This is a simpler description of the whole association structure that concentrates on conditional associations within each pair of variables given the others (without being explicit what "others" means, it is usually less than all remaining variables) and directly follows from the summary provided at the end of our analysis of the three-way table. For a particular problem, everything can be summarized by (1) a list of variables which are (unconditionally) independent of others, (2) a matrix indexed by the remaining variables showing which pair satisfies 2, 3, or 4 from the above list of possibilities. Such a matrix may look as follows:

Table 2: Hypothetical table describing conditional association structure among the analyzed variables implied by the model.

	sesmed	parents	foreign	fa.educ	mo.educ	region	fa.wrk	mo.wrk
sesmed				H				
parents								
foreign				★	H			
fa.educ	H		★		★			
mo.educ			H	★				
region								
fa.wrk								
mo.wrk								

where '★' would stand for conditional independence, 'H' for homogeneous associations and empty cells for general conditional association. More details about the association structure can be provided by an undirected graph as explained in the course notes. Nevertheless, if we deal with a higher number of

variables, such a graph may easily become quite messy and non interpretable by statistician, let alone by a poor social scientist.

This final problem of the assignment is example of a task to help social scientist understand the association structure among several variables.

First, we need the corresponding contingency table:

```
> xtab5 <- with(nels, table(sesmed, parents, foreign, fa.educ, mo.educ, region,
  +                          fa.wrk, mo.wrk))
> dim(xtab5)
[1] 2 2 2 3 3 4 2 2
```

```
> prod(dim(xtab5))       ## 1152
[1] 1152
```

Table is not really small, having 1152 cells...

Our methods highly rely on asymptotics for which the sample size is not that important. We always need sufficiently high *expected* counts. But if already *observed* counts are not really high then we can really expect problems. We will not worry about counts in all 1152 cells (we will try to deal with models which are far from saturated one anyway). But as a minimum, all marginal observed counts should be checked and if really small cells found, some action should be taken. Either some cells can be merged or some items (variables) thrown away (upon mutual agreement to do so with the social scientist).

```
>   ### Some exploration (do we have at least marginally reasonable counts?)
> margin.table(xtab5, margin = 1)
> margin.table(xtab5, margin = 2)
> margin.table(xtab5, margin = 3)
> margin.table(xtab5, margin = 4)
> margin.table(xtab5, margin = 5)
> margin.table(xtab5, margin = 6)
> margin.table(xtab5, margin = 7)
> margin.table(xtab5, margin = 8)
```

There seem to be no problem here.

We can also explore some pairwise dependencies (to have some initial view into the problem).

```
>   ### Some pairwise dependencies
> with(nels, table(fa.wrk, mo.wrk))
> prop.table(with(nels, table(fa.wrk, mo.wrk)), margin = 2)
> chisq.test(with(nels, table(fa.wrk, mo.wrk)))
> #
> with(nels, table(fa.educ, mo.educ))
> prop.table(with(nels, table(fa.educ, mo.educ)), margin = 2)
> chisq.test(with(nels, table(fa.educ, mo.educ)))
> #
> with(nels, table(mo.wrk, mo.educ))
> prop.table(with(nels, table(mo.wrk, mo.educ)), margin = 2)
> chisq.test(with(nels, table(mo.wrk, mo.educ)))
```

Let us start with a model that involves all three-way interactions and no higher order terms, compare it to the saturated model (previously written function `gof` is used here) and hope that it is not significantly worse:

```
> qq5 <- as.data.frame(xtab5, responseName = "N")
> m1 <- glm(N ~ (sesmed + parents + foreign + fa.educ + mo.educ + region +
  +                fa.wrk + mo.wrk)^3, family = poisson, data = qq5)
> gof(m1)
Goodness-of-fit test, model with 246 parameters
Deviance = 843.3029, df = 906
P-value: 0.932

Number of cells with low fitted counts:  878

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.0008   0.2204   0.8997  11.7882   4.5562 732.7289
```

We are lucky, model with at most three-way interactions is safely not significantly worse than the saturated model.

How about if we consider only two-way interactions:

```
> mTwoWay <- glm(N ~ (sesmed + parents + foreign + fa.educ + mo.educ + region +
  +                    fa.wrk + mo.wrk)^2, family = poisson, data = qq5)
> #
> gof(mTwoWay)
Goodness-of-fit test, model with 74 parameters
Deviance = 1351.458, df = 1078
P-value: <0.001

Number of cells with low fitted counts:  879

    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
  0.0023   0.1587   0.8303  11.7882   4.5505 671.9030
```

```
> anova(mTwoWay, m1, test = "LR")        ### Some three-way interactions needed...
Analysis of Deviance Table

Model 1: N ~ (sesmed + parents + foreign + fa.educ + mo.educ + region +
    fa.wrk + mo.wrk)^2
Model 2: N ~ (sesmed + parents + foreign + fa.educ + mo.educ + region +
    fa.wrk + mo.wrk)^3
  Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
1      1078     1351.5
2       906      843.3 172   508.16 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It is worse, both against the saturated and three-way interaction model.

TASK FOR YOU: Apparently, some three-way interactions are needed and once those included, perhaps some two-way interactions become obsolete. Use now standard model building strategy based on deviance tests to arrive at as simple model as possible which, however, still fits the data. Describe (by a table like Table 2) the conditional association structure among the variables. Choose one of the main effects and one of the two-way interaction terms in the final model and interpret the estimated parameter.

As usually, start from the "big" (all three-way interactions) model and try to remove always several (the least significant) terms at once. Especially in later stages of the process, if you remove more terms in one step, try to put each of them back in the model and check whether it does not significantly improve the model.

We now have quite lots of terms included in the model. It is perhaps wise to partly automatize the process as follows:

```
> D1 <- drop1(m1, test = "LR")            ## takes some time, be patient...
> print(D1)
```

You can now check by eyes which terms are the least significant and candidates for removal or which terms are the most significant and candidates for being held in the model. Or you can exploit some of the programming capabilities of R. Either, you can list candidates for removal, then fit the model without them and compare it to the previous model, the all three-way interactions model or even saturated model:

```
> (Drop1 <- attr(D1, "row.names")[-1][D1[["Pr(>Chi)"]][-1] > 0.5])
> m2 <- update(m1, paste(". ~ . - ", paste(Drop1, collapse = "-")))
> gof(m2)
> anova(m2, m1, test = "LRT")
```

If we really want to be sure that nothing "important" has been removed, standard back-check can be performed (rather automatically) by adding each of removed terms back to the model:

```
> (Return1  <- add1(m2, scope = m1, test = "LRT"))
> min(Return1[, "Pr(>Chi)"], na.rm = TRUE)
```

Or, you can list candidates for being kept and then fit some simpler model PLUS kept terms and do the comparison:

```
> (Keep1 <- attr(D1, "row.names")[-1][D1[["Pr(>Chi)"]][-1] < 0.2])
> m3 <- update(mTwoWay, paste(". ~ . + ", paste(Keep1, collapse = "+")))
> gof(m3)
> anova(m3, m1, test = "LRT")
```

And now your turn...

References

Dahl, D. B., Scott, D., Roosen, C., Magnusson, A., and Swinton, J. (2019). *xtable: Export tables to LaTeX or HTML*. URL http://CRAN.R-project.org/package=xtable. R package version 1.8-4.

Leisch, F. (2002). Dynamic generation of statistical reports using literate data analysis. In Härdle, W. and Rönz, B., editors, *COMPSTAT 2002 – Proceedings in Computational Statistics*, pages 575–580, Heidelberg, 2002. Physica-Verlag.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Zeileis, A., Hornik, K., and Murrell, P. (2009). Escaping RGBland: Selecting colors for statistical graphics. *Computational Statistics and Data Analysis*, **53**(9), 3259–3270. doi: 10.1016/j.csda.2008.11.033.

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., and Wilke, C. O. (2019). colorspace: A toolbox for manipulating and assessing colors and palettes. arXiv 1903.06490, arXiv.org E-Print Archive. URL http://arxiv.org/abs/1903.06490.