Stockholm University
Mathematical statistics November 2024

ANALYSIS OF CATEGORICAL DATA

Computer assignments, Fall 2024

The purpose of the computer exercises is to improve the understanding of different models for categorical data and to gain knowledge of how to perform and interpret statistical analysis of categorical data.

The work should be conducted in groups of two students. Supervision is offered during five two hour sessions according to the schedule on the course website. The statistical software R (RStudio / RMarkdown) is recommended (but not required) to use and there are some useful R commands added to each assignment.

A written documentation of each assignment should be uploaded on the course website as a pdf-file by the date indicated in the schedule (one submission per group). Make sure that all questions are answered and that all requested interpretations have been made. Relevant R-outputs must be included and referred to, but R-code need not be included.

An approved report submitted in time gives 2 bonus credits on the written exam. A report, submitted in time but with minor errors can after revision give 1 bonus credit. In order to pass the whole course, all computer assignments must be approved before the second exam. The bonus credits are only valid for this academic year.

Computer assignment 1
Analysis of 2×2 tables

Exercise 1:1

In a survey, 500 women och 600 men where asked about their opinion on legal abortion. Of the women, 309 were in favor of legal abortion and of the men the corresponding number was 319. The result of the survey is summarized in the two-way table below.

	In favor	Against	Σ
Women	309	191	500
Men	319	281	600
Σ	628	472	1100

In exercises 1 to 4 below, the calculations should be carried out using basic operations in R ($+ - * /$ etc.) with each step well documented. For exercise 5, the calculations should be redone using certain R functions .

1. Calculate the percentage in favor and against legal abortion for men and women separately.
2. Do men and women have different opinions on legal abortion? Define a null and an alternative hypothesis and report all steps to calculate Pearsons X^2 statistic and the likelihood ratio statistic, G^2. What conclusion can be made from this survey.
3. Report all steps to estimate an odds ratio together with a 95 % confidence interval. Note that it is possible to formulate several odds ratios for a twoway table, depending on which event the odds refer to and what you conditioning on (an odds ratio is an ratio of two conditional odds). Interpret the estimated odds ratio and the corresponding confidence interval in the context of the survey.
4. Report all steps to estimate a risk ratio (relative risk) together with a 95 % confidence interval. The risk ratio should correspond to the odds ratio you reported in question 3. Interpret the risk ratio and the confidence interval in the same way as in question 3.
5. Below is shown how availible R-functions can be used to perform the calculations in question 1 to 4 (other R-functions might work as well). Run the code and check that your previous calculations are correct. Regarding the odds ratio and risk ratio, you might need to change the **rev** option in R (see below) to get agreement between them.

R INSTRUCTION 1

Generate a frequency table and calculate row percentage

```
tab1<- as.table(rbind(c(309, 191), c(319, 281)))
dimnames(tab1) <- list(gender = c("women", "men"),opinion = c("favor","against"))
addmargins(tab1)
addmargins(prop.table(tab1,1),2)
```

Calculate X^2, G^2 and p-values

```
chisq.test(tab1,correct=FALSE)

library(MASS)                    # the MASS package must have been installed
loglm(~gender+opinion,tab1)

# The function loglm fits a loglinear model, but we leave the details about that.
# Loglinear models will be handled in assignment 3 but with the function glm instead of loglm.
```

Calculate risk/odds ratio and confidence interval

```
library(epitools)              # the epitools package must have been installed

oddsratio(tab1, method = "wald", rev="neither")
riskratio(tab1,rev="neither")

# To obtain the intended odds/risk ratio, you may need to reverse the rows or columns:

oddsratio(tab1, method = "wald", rev="row")
oddsratio(tab1, method = "wald", rev="col")
oddsratio(tab1, method = "wald", rev="both")
```

Exercise 1:2

In this exercise you will conduct the same kind of analyses as in exercise 1:1 for several two-way tables. You can use the same R-functions as in 1:1.5 and no step by step calculations is needed. You should write down the result with your own words and refer to the R-outputs that you include.

At the University of California, Berkeley, in total 2691 men and 1835 women were applied to graduate studies, in year 1975. Among those, 1198 men and 557 women were admitted.

1. Enter these data into R and make an analysis as in Excercise 1:1 (Pearsons χ^2 and G^2, odds ratio and risk ratio with confidence interval). Interpret the result! Be careful to specify which odds and risk ratios you have calculated and how they should be interpreted.

2. Investigate what will happen if you replace all numbers in the table to one tenth of the original values. Compare the output with that in question 1. Which values have changed and how? Repeat and replace the numbers to one hundredth of the original and comment on what you observe.
   Hint: The R command **round('table name'/10,0)** may be useful.

3. Based on the result above you should now create a completely new two-way table with cell counts that you chose by yourself, such that the sample odds ratio ($\hat{\theta}$) is within the interval $(0.99, 1.01)$ and the corresponding population odds ratio (θ) differ from 1 significantly. Make a comment about the relevance to declare 'statistical significance' in a situation like this. How would you prefer to report the result? Make sure that you understand this excercise and ask for help if needed.

Computer assignment 2
Logistic regression

Excercise 2:1

Data on periodontitis (severe inflammation of the gums) were collected from a group of adult patients at a large dental clinic. The patients were compared with a group of adult people with normal gums who visited the same clinic. The individuals in both groups were interviewed about the use of dental floss in the last five years, with the following result:

Regularly use of dental floss	Periodontitis Yes	No	Σ
Yes	22	75	97
No	148	265	413
Σ	170	340	510

1. Suppose that you fit the following logistic regression model of the probability for periodontitis in the population (which we assume is defined) as a function of dental floss use:

$$\begin{aligned} \text{logit } (p_x) &= \beta_0 + \beta_1 x \\ \text{where } x &= 1 \text{ if dental floss is used, and 0 else} \\ \text{and } p_x &= P(\text{periodontitis} \mid x) \end{aligned}$$

   Which parameters in the model can be estimated considering how the data is collected? What estimates do you get and how do you interpret them?

2. Suppose that you instead fit the following logistic regression model of the probability for using dental floss as a function of periodontitis status:

$$\begin{aligned} \text{logit } (p_y) &= \gamma_0 + \gamma_1 y \\ \text{where } y &= 1 \text{ if periodontitis is present and 0 else} \\ \text{and } p_y &= P(\text{using dental floss} \mid y) \end{aligned}$$

   Which parameters in this model can be estimated from the study? What estimates do you get and how do you interpret them?

3. Show how the parameter estimates in both models are expressed by the use of the actual numbers in the table. Show that β_1 and γ_1 are equal.

R INSTRUCTION 2:1

In order to fit a logistic regression model it is convenient to first organize the data in a 'data frame'.

Input data and create a data frame

```
use<-c("no","no", "yes","yes")
per<-c("no","yes","no","yes")
n<-c(265,148,75,22)
data21 <- data.frame(use,per,n)
data21
```

Fit a logistic regression model

```
# First the characters "per" and "use" must be converted to dummy variables.
data21$use <- as.factor(data$use)
data21$per <- as.factor(data$per)

model21<-glm(per~use, weights =n, family = binomial(link=logit), data=data21)
summary(model21)
```

Note:

The R-function "as.factor()" create dummy variables in alphabetical order and zero is assigned to the level which is first alphabetically. In this case it means that R is modelling the probability, P(per=yes) and the coefficient of "use" refer to the difference in logodds(per=yes) between use=yes and use=no. The level use=no is the reference or base level (to change the reference level the R function "relevel" can be used). To have a better control over the analysis you can manually create dummy variables during the data input. Then, zero always represents the reference level.

```
x<-c(0,0,1,1)  # x=0 if dental fluss is not used, x=1 if dental fluss is used
y<-c(0,1,0,1)  # y=0 if periodontitis is No, y=1 if periodontitis is Yes
n<-c(265,148,75,22)
data21a <- data.frame(x,y,n)
model21a<-glm(y~x, weights=n, family=binomial(link=logit), data=data21a)
summary(model21a)
```

R is modelling the probability, P(y=1) and the coefficient of x refer to the difference in logodds(y=1) between x=1 and x=0 (which is in agreement with the first model above).

Exercise 2:2

In an experiment, 165 mice were exposed to various doses of bensoapyren (BaP) over a ten month period. It was then examined how many of the mice that has developed a lung tumor. The result as a function of logarithmic dose is:

log(dose)	-7.60	-6.22	-4.60	-3.00	-1.39	0.92	Σ
Tumor	1	2	4	9	12	32	60
No tumor	17	17	24	23	16	8	105
Σ	18	19	28	32	28	40	165

1. Estimate for each dose separately the risk (probability) to develop a lung tumor. Make a plot of the risk against log(dos). Calculate the logodds for tumor and plot it against log(dose). Seems the logistic regression model appropriate for these data? Motivate your answer!

2. Fit a logistic regression model to this data and interpret the parameter estimates, particularly the slope parameter.

3. Find the covariance matrix for the estimates and assess if they are correlated. Find a 95 % confidence interval for the parameters. Find a 95 % confidence interval for the tumor risk at dose 0.25 (log(dose)= -1.39).

4. Perform a Wald-test of $H_0 : \beta_1 = 0$. Explain and show with an own calculation, based on R-output, how the test statistic is constructed. Interpret the result from the hypothesis test.

5. The variance of a parameter estimate is generally inversely proportional to the sample size. Is this true also here? Since we have no expression explicitly for the variances, we cannot confirm it straightforward. In order to investigate it, you should multiply all counts in the table by 10 repeatedly (10, 100, 1000) and report what you observe.

R INSTRUCTION 2:2

Input data and create a data frame

```
logdos<-c(-7.60,-6.22,-4.60,-3.00,-1.39,0.92)
n<-c(18,19,28,32,28,40)
x<-c(1,2,4,9,12,32)
data22 <- data.frame(logdos,x,n)
data22
```

Fit a logistic regression model.

```
# Note the difference in model specification compared to 2.1

model22<-glm(x/n~logdos, family = binomial(link=logit), weights=n, data=data22)
summary(model22)
```

Extract covariance matrix and confidence intervals

```
vcov(model22)

confint.default(model22,level=0.95)
```

Computer assignment 3
Log-linear models

Exercise 3:1 (Higher dimension table)

Wermuth (1976, Biometrics) reports data collected from a birth clinic, which includes information on the mother's age, her smoking habits (number of cigarettes per day), gestational age (in days) and the survival status of the child. Data were collected during a given period and neither the total sample size nor any margins were held fixed. The aim of the assignment is to find a log-linear model that can be used to gain an understanding of the association between the variables. The model should fit the data well but not be too complex to interpret.

Mother's age	Smoking habits	Gestational age	Child survival	
			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

1. In order to find a 'good' model several models have to be fitted. You should start with the saturated model (which has a perfect fit) and remove interaction terms in a systematic way, where higher order interactions are removed before lower order interactions. No main effect should be removed, since the interest here is the association between the variables. The goodness-of-fit of the different models should be evaluated with deviance (compared to the saturated model) and AIC, and the table below should be completed with the calculated values. The variable names to be used are: X = Mother's age, Y = Smoking habits, Z = Gestational age and V = Child survival.

   The R-instruction below may be useful for this excercise. It also include code for a stepwise approach, that can be used.

Model	*Deviance*	*df*	$p - value$	*AIC*
XYZV	?	?	?	?
XYZ, XYV, XZV, YZV	?	?	?	?
.	?	?	?	?
.	?	?	?	?
.	?	?	?	?
XY, XZ, XV, YZ, YV, ZV	?	?	?	?
.	?	?	?	?
.	?	?	?	?
.	?	?	?	?
X, Y, Z, V	?	?	?	?

2. Choose from your table a model with few parameters and a good fit. Describe the procedure to compare different models.

3. Interpret the model you chose. Which associations are significant? Quantify the associations with odds ratios together with confidence intervals.

4. Let us now assume that the main purpose is to investigate the relationship between Child survival and the variables Mother's age, Gestational age and Smoking habits. Let the response variable be Child survival and fit a logistic regression model for the probability that a child survive as a function of the explanatory variables Mother's age, Gestational age and Smoking habits (and interactions between them if necessary). Interpret the model you chose, especially regarding the relationship between survival and smoking habits.

5. To illustrate the relationship between a log-linear model and a logistic regression model you should now fit a log-linear model that corresponds to the logistic regression model you chose in question 3. Confirm that the two models gives identical estimates and standard errors of the corresponding parameters.

R INSTRUCTION 3

Data is available as a csv-file "data_ca3" on the course website.

Set the working directory and read in data

```
setwd("---")                      # Input the path to your desired working directory
data3<-read.csv("data_ca3.csv")
```

Fitting a loglinear model

```
# Saturated model:
msat<-glm(n~x*y*z*v, family=poisson(link=log), data=data3)
summary(msat)

# All three-way interactions:
m3<-glm(n~(x*y*z+x*y*v+x*z*v+y*z*v), family=poisson(link=log), data=data3)
summary(m3)

# Alternative coding for a model with all three-way interactions:
m3<-glm(n~(x+y+z+v)^3, family=poisson(link=log), data=data3)
summary(m3)
```

LR-test between two nested models:

```
anova(m3,msat,test="LRT")
```

Stepwise model selection, using AIC

```
# In this example, m3 is the most complex model that can be considered and m2 is the simplest.

mstep_AIC<-step(m3, direction="both", trace=TRUE, scope = list(upper = m3, lower = m2))
summary(mstep_AIC)
```

Fitting a logistic regression model for this data

```
# See R-instruction 2.1
```

Computer assignment 4
Multiple logistic regression
Decision Tree

In this exercise you will carry out several analyses of *The ICU Study* data (Hosmer & Lemeshow (1989): *Applied Logistic Regression*). The data consists of records on 200 patients admitted to an Intensive Care Unit (ICU) during a certain period and the overall task is to use multiple logistic regression and decision tree models to identify factors that affect the survival of such patients.

The data can be downloaded from the course website (data_ca4.csv) and the variables are explained in the table below. Note! For two of the variables, it is recommended to combine categories due to low cell counts. See the R-instruction.

Column	Variable	Coding
1	Patient	ID-code
2	Survival	0=yes 1=no
3	Age	Year
4	Sex	0=male 1=female
5	Ethnicity	1=white 2=black 3=other
6	Treatment at admission	0=medical 1=surgical
7	Cancer	0=no 1=yes
8	Previous kidney failure	0=no 1=yes
9	Infection	0=no 1=yes
10	Heart/lung treatment before admission	0=no 1=yes
11	Blood pressure	mm Hg

Column	Variable	Coding
12	Heart rate	beats/min.
13	Admitted to ICU within the last 6 months	0=no 1=yes
14	Type of Admission	0=not acute 1=acute
15	Fracture of neck or spine	0=no 1=yes
16	Blood oxygen	0=above 60 1=below 60
17	Blood pH	0=above 7.25 1=below 7.25
18	Blood carbon dioxide	0=below 45 1=above 45
19	Blood bikarbonate	0=above 18 1=below 18
20	Blood creatine	0=below 2.0 1=above 2.0
21	Consciousness level	0=awake 1=unconscious 2=coma

Exercise 4:1 (Multiple logistic regression)

Which variables affect the survival of a patient admitted to ICU? And how do they affect the survival? To answer that, you should create a multiple logistic regression model for the probability to not survive (i.e. the probability to die). For variable selection, stepwise methods can be used. We will assume that there are no interactions between the variables (you are free to investigate whether this assumption is reasonable).

1. Report the model selection process in short. Based on your chosen model, which factors affect the probability to not survive? Report odds ratio with confidence in-

terval for the most important variables/factors and interpret them. Use the variable names in the table above when you report this (not V3, V4, ...).

2. How well does your chosen model fits the data? In assignment 3, the deviance was used to assess model fit but since the data now is on individual level, deviance is not suitable and instead, the Hosmer-Lemeshow goodness-of-fit test (Agresti 5.2.5) should be used. Perform this test in R (see the R-instruction) and interpret the result.

3. Now we shift the focus to prediction. How well does your multiple logistic regression models predict the outcome? There exists several different metrics of prediction performance for binary responses and here we will use the following: 'Accuracy', 'Sensitivity' (True positive rate), 'Specificity' (True negative rate) and 'AUC' (area under the ROC curve). To explain these measures we first notice that the model predictions (or fitted values) are probabilities to not survive. If we let patients be classified as 'Survive' or 'Not survive' when theirs predicted probability is below or above a certain cut-off value (treshold value), a confusion matrix (a 2x2 classification table) can be constructed, as shown below.

	Predicted		
Actual	Survive	Not survive	Total
Survive	n_{11}	n_{12}	$n_{1.}$
Not survive	n_{21}	n_{22}	$n_{2.}$
			$n_{..}$

The definition of accuracy, sensitivity and specificity is then:

- Accuracy $= \frac{n_{11}+n_{22}}{n_{..}}$
- Sensitivity $= \frac{n_{11}}{n_{1.}}$
- Specificity $= \frac{n_{22}}{n_{2.}}$

Construct a confusion matrix based on your chosen model and calculate accuracy, sensitivity and specificity, for a cut-off value of 0.5 and two more of your choice (see the R-instruction). Describe and explain the result.

4. As shown above, the sensitivity and specificity dependens on the cut-off setting. A ROC-curve is a plot of sensitivity (true positive rate) against 1-specificity (false postive rate) at each possible cut-off setting. AUC is the area under the ROC curve and can be used as a summary metsure of a binary classifier's performance. AUC=1 indicates perfect classification and AUC=0.5 indicates that the model's performance is no better than random guessing. Create plots of ROC curves and calculate the AUC for the full model and two more models of your choice (see the R-instruction). Which of the model performs best based on AUC?

5. The AUC-values obtained in the previous exercise may be subject to optimistic bias, as they were computed for the same dataset that was used for model fitting. This

might result in overfitting, where the model become well adapted to the training data but may not perform as good to new data. To get a more reliable AUC-value, a validation method can be used. We will use the Leave-One-Out-Cross-Validation (LOOCV). For LOOCV, a model is fitted (or trained) using all observations except one, which is held out for validation. The model is then used to make a prediction on the held-out observation. This process is repeated for each observation, resulting in LOOCV predicted probabilities for each observation (see the R-instruction). These LOOCV predicted probabilities can then be used to calculate AUC. Do this for the full model and two more models of your choice (the same models as in the previous exercise). Which of the model performs best based on LOOCV-adjusted AUC? Compare with the result in the previous exercise.
Note. The stepwise procedure should not be performed within the cross-validation loop.

Excercise 4:2 Decision Tree

In this exercise you will experiment with decision tree models for binary classification, focusing on its practical application without delving into the underlying theory. A decision tree model can be visualized as a flowchart that shows how objects (patients) are classified into subgroups based on the predictor variables that are most important for predicting the target variable (survival status). A decision tree is constructed by a supervised learning algorithm that starts with a root node, representing the entire dataset. The algorithm iteratively splits the dataset into subsets, based on specific splitting rules and this process continues until a stopping criterion is met (such as a maximum depth, a minimum number of samples per leaf node, or a minimum information gain). Once the decision tree is constructed, it can be used to make predictions on new data.

Fore details about the theory, see
`https://en.wikipedia.org/wiki/Decision_tree_learning`
`https://cran.r-project.org/web/packages/rpart/rpart.pdf`

1. To illustrate the method you should run the first code chunk - Fitting a decision tree model and plot the tree - in the R-instruction 4.2. Examine then how changes of the complexity parameter (cp) and the splitting method affect the decision tree diagram. Choose a 'good' decision tree and briefly describe what it shows. Include the tree plot in your report.

2. How well can your chosen tree model predict whether or not a patient admitted to the intensive care unit will survive or not? As a measure of predictive power, the AUC

will be used, as for the multiple logistic regression model. How AUC is calulated for a tree model is shown in R Instruction 4.2. Compare the predictive power for you tree model with that for the multiple logistic regression models. Which model performs best based on AUC?

3. Finally, calculate a LOOCV-corrected AUC for your tree model (see R-instruction 4.2) and make a comment on the result.

R INSTRUCTION 4

Data is available as a csv-file "data_ca4" on the course website.

Set the working directory, read in data and combine categories

```
setwd("---")
data4 <- read.csv("data_ca4.csv")

# Combine categories (this should be done for one more variable)
table(data4$v5)
data4$v5[data4$v5>1] <- 0
table(data4$v5)
```

4.1 Multiple logistic regression

Fitting multiple logistic regression models

```
m1 <- glm(v2~1, family=binomial,data=data4)              # Empty model (only intercept)
m_full <- glm(v2~., family=binomial,data=data4)          # Full model (all explanatory variables)
summary(m_full)
```

Stepwise model selection - forward, backward and both - with AIC or BIC as inclusion criteria

```
mstep1_AIC <- step(m_empty, direction="forward", scope = list(upper = m_full), trace = TRUE)
mstep2_AIC <- step(m_full, direction="backward", trace = FALSE)
mstep3_AIC <- step(m_full, direction="both", trace = FALSE)

mstep1_BIC <- step(m_empty, direction="forward", scope = list(upper = m_full),
  k = log(nrow(data4)), trace = TRUE)
mstep2_BIC <- step(m_full, direction="backward", k = log(nrow(data4)), trace = FALSE)
mstep3_BIC <- step(m_full, direction="both", k = log(nrow(data4)), trace = FALSE)

# Set trace = FALSE to suppress the detailed output of each step
```

Perform Hosmer-Lemeshow goodness of fit test

```
predprob_full<- predict(m_full, type = "response")  # Obtain predicted probabilities for the full model
library(ResourceSelection)         # The package "ResourceSelection" must have been installed
hoslem.test(survive, predprob_full), g=10)
```

Create a confusion matrix (classification table)

```
pred01_full <- ifelse(predprob_full > 0.5, 1, 0)    # Binary classification, cut-off = 0.5
confmatrix <- table(1-data4$v2,1-pred01_full)
dimnames(confmatrix) <- list(Actual = c("Survive", "Not-survive"),
  Predicted = c("Survive", "Not-survive"))
confmatrix               # To add margin and obtain proportions, see R-instructin 1
```

Create a ROC-curve and calculate AUC

```
library(pROC)     # Package pROC must have been installed
roc(data4$v2, predprob_full)
plot(roc(data4$v2, predprob_full),legacy.axes = TRUE)
```

Leave-one-out cross-validation, AUC

```
predprob_LOOCV <- numeric(nrow(data4))  # Create a vector for the predicted values

for (i in 1:nrow(data4)) {
  # Create training and validation sets
  data_training <- data4[-i,]
  data_validation <- data4[i, ,drop=FALSE]

  # Fit the model on the training data
  m1<-glm(v2~v3+v4+v5, family=binomial,data=data_training)

  # Predict the value for the held-out observation
  predprob_LOOCV[i] <- predict(m1, newdata = data_validation, type = "response")
}

roc(data4$v2, predprob_LOOCV)
```

4.2 Decision tree

Fitting a decision tree model and plot the tree

```
library(rpart)              # The package "rpart" must have been installed
library(rpart.plot)         # The package "rpart.plot" must have been installed

#Fitting a decision tree model
tm1 <- rpart(v2 ~.,method = "class", data=data4, parms = list(split = "information"), cp = 0.1)
    # use split = "gini" or  split = "information"
    # cp "complexity parameter" can be adjusted

rpart.plot(tm, digits=3)   #Plot the decision tree
```

Get predicted probabilities

```
predprob_tm <- predict(tm_full,type = "prob")[,2]

# predprob_tm can then be used to calculate the AUC and the LOOCV-corrected AUC,
# using the same code as for the multiple logistic regression.
```