Practical on logistic regression

Daniela Zugna

These are data from a study on hospital discharge status after burn injury. Data contain 1000 rows and 9 variables. The aim of this practical is to evaluate how the risk of death changes on overage according to other variables.

Variables:
id: Identification code (1 - 1000)
facility: Burn facility (1 - 40)
death: Hospital discharge status (1: Alive, 2: Dead)
age: Age at admission (Years)
gender: Gender (1: Female, 2: Male)
race: Race (1: Non-White, 2: White)
tbsa: Total burn surface area (0 - 100%)
inh_inj: Burn involved inhalation injury (1: No, 2: Yes)
flame: Flame involved in burn injury (1: No, 2: Yes)

1. Open the file burn1000.csv dataset.

```
rm(list=ls())
setwd("")
data<-read.table("burn1000.csv",delimiter(;))
dim(data)
names(data)
head(data)
```

2. Get basic descriptives for the entire data set.

```
summary(data)
```

   - Comment the nature and the distribution of each variable.

3. Declare the categorical variables as factor variables.

```
data$gender<-as.factor(data$gender)
data$flame<-as.factor(data$flame)
data$inh_inj<-as.factor(data$inh_inj)
data$race<-as.factor(data$race)
```

4. Describe the distribution of age and test whether there is a difference in its distribution according to death by t-Student test.

```
summary(data$age[data$death==1])
summary(data$age[data$death==2])
t.test(data$age[data$death==1],data$age[data$death==2])
```

   - What is the mean age in dead patients compared to alive patients?
   - Is there evidence of a difference in the mean age between dead and alive patients? Why?
   - Answer to previous questions for tbsa variable.

```
summary(data$tbsa[data$death==1])
summary(data$tbsa[data$death==2])
t.test(data$tbsa[data$death==1],data$tbsa[data$death==2])
```

5. Cross-tabulate gender and death and test their association by χ^2 test.

```
table(data$death,data$gender)
prop.table(table(data$death,data$gender))
chisq.test(table(data$death,data$gender))
```

   - What is the percentage of females and males among dead patients respectively?
   - What is the percentage of females and males among alive patients respectively?
   - Is there evidence of an association between sex and death? Why?
   - Answer to previous questions for each categorical variable.

```
table(data$death,data$flame)
prop.table(table(data$death,data$flame))
chisq.test(table(data$death,data$flame))
table(data$death,data$inh_inj)
prop.table(table(data$death,data$inh_inj))
chisq.test(table(data$death,data$inh_inj))
table(data$death,data$race)
prop.table(table(data$death,data$race))
chisq.test(table(data$death,data$race))
```

6. Fit a logistic regression model for the risk of death vs. age (continuous variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
data$death[data$death==1]<-0
data$death[data$death==2]<-1
mylogit.age<-glm(death~age,family = binomial, data = data)
summary(mylogit.age)
confint(mylogit.age)
exp(coef(mylogit.age))
exp(cbind(OR = coef(mylogit.age), confint(mylogit.age)))
```

- On the basis of the estimated OR, how does risk of death vary for a unit change of age?
- On the basis of the 95% confidence interval for OR of age, can you conclude that risk of death depends on age? Why?

7. Fit a logistic regression model for the risk of death vs. tbsa (continuous variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
mylogit.tbsa<-glm(death~tbsa,family = binomial, data = data)
summary(mylogit.tbsa)
confint(mylogit.tbsa)
exp(coef(mylogit.tbsa))
exp(cbind(OR = coef(mylogit.tbsa), confint(mylogit.tbsa)))
```

- On the basis of the estimated OR, how does risk of death vary for a unit change of total burn surface area (in percentage)?
- On the basis of the 95% confidence interval for OR of tbsa, can you conclude that risk of death depends on tbsa? Why?

8. Fit a logistic regression model for the risk of death vs. inh_inj (binary variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
mylogit.inj<-glm(death~inh_inj,family = binomial, data = data)
summary(mylogit.inj)
confint(mylogit.inj)
exp(coef(mylogit.inj))
exp(cbind(OR = coef(mylogit.inj), confint(mylogit.inj)))
```

- On the basis of the estimated OR, how does risk of death vary if burning involved inhalation injury?

- On the basis of the 95% confidence interval for OR of inh_inj, can you conclude that risk of death depends on inh_inj? Why?

9. Fit a logistic regression model for the risk of death vs. gender (binary variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
mylogit.gender<-glm(death~gender,family = binomial, data = data)
summary(mylogit.gender)
confint(mylogit.gender)
exp(coef(mylogit.gender))
exp(cbind(OR = coef(mylogit.gender), confint(mylogit.gender)))
```

- On the basis of the estimated OR, how does risk of death vary in males compared to females?
- On the basis of the 95% confidence interval for OR of gender, can you conclude that risk of death depends on gender? Why?

10. Fit a logistic regression model for the risk of death vs. flame (binary variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
mylogit.flame<-glm(death~flame,family = binomial, data = data)
summary(mylogit.flame)
confint(mylogit.flame)
exp(coef(mylogit.flame))
exp(cbind(OR = coef(mylogit.flame), confint(mylogit.flame)))
```

- On the basis of the estimated OR, how does risk of death vary according to flames involved in burn injury?
- On the basis of the 95% confidence interval for OR of flame, can you conclude that risk of death depends on flame? Why?

11. Fit a logistic regression model for the risk of death vs. race (binary variable) and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Calculate odds ratio (OR) and its corresponding 95% confidence interval.

```
mylogit.race<-glm(death~race,family = binomial, data = data)
summary(mylogit.race)
confint(mylogit.race)
exp(coef(mylogit.race))
exp(cbind(OR = coef(mylogit.race), confint(mylogit.race)))
```

- On the basis of the estimated OR, how does risk of death vary according to race?
- On the basis of the 95% confidence interval for OR of race, can you conclude that risk of death depends on race? Why?

12. Fit a logistic regression model for the risk of death with all the variables and examine the regression coefficients. Compute the 95% confidence intervals for the regression coefficients. Describe which conclusions you would draw from this study (which variables affect the risk of death, and which ones might be the most relevant).

```
mylogit<-glm(death~age+tbsa+inh_inj+factor(gender)+factor(flame)+factor(race),
family = binomial,data = data)
data = data)
summary(mylogit)
confint(mylogit)
exp(coef(mylogit))
exp(cbind(OR = coef(mylogit), confint(mylogit)))
```

13. Predict the risk of death with its confidence interval for two patients with the following characteristics:
1. age=mean of age, tbsa=mean of tbsa, inh_inj=2, flame=2 and race=1
2. age=mean of age, tbsa=mean of tbsa, inh_inj=2, flame=2 and race=2

```
newdata<-with(data, data.frame(age=mean(age),tbsa=mean(tbsa),inh_inj=2,
                               gender=2, flame=2,race=1:2))
newdata$inh_inj<-as.factor(newdata$inh_inj)
newdata$gender<-as.factor(newdata$gender)
newdata$flame<-as.factor(newdata$flame)
newdata
prediction<-predict(mylogit,newdata,type="response")
prediction
```