Labbrapport i Statistik

Laboration 7

732G46

Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-10-18

Innehåll

Introduktion

I denna laboration kommer två dataset användas.

I det första datasetet kommer slumpad data över insekter som getts doser av giftiga substanser att behandlas, i datasetet tittar vi på huruvida insekterna avlider av doserna eller inte.

I Datasetet som består av ett OSU av 113 sjukhus från 338 sjukhus som har undersökts, kommer variabler såsom; medellängden på besöken, åldern, sannolikheten att få en infektion och medelsumman av antal sjukhussängar m.fl. att analyseras och användas i modeller. Det som skiljer denna laboration från den tidigare laborationen är att samtliga 113 observationer inte kommer att användas samtidigt.

Målet med denna laboration är att se användbarheten i logistisk regression samt oddsratio.

Databehandling

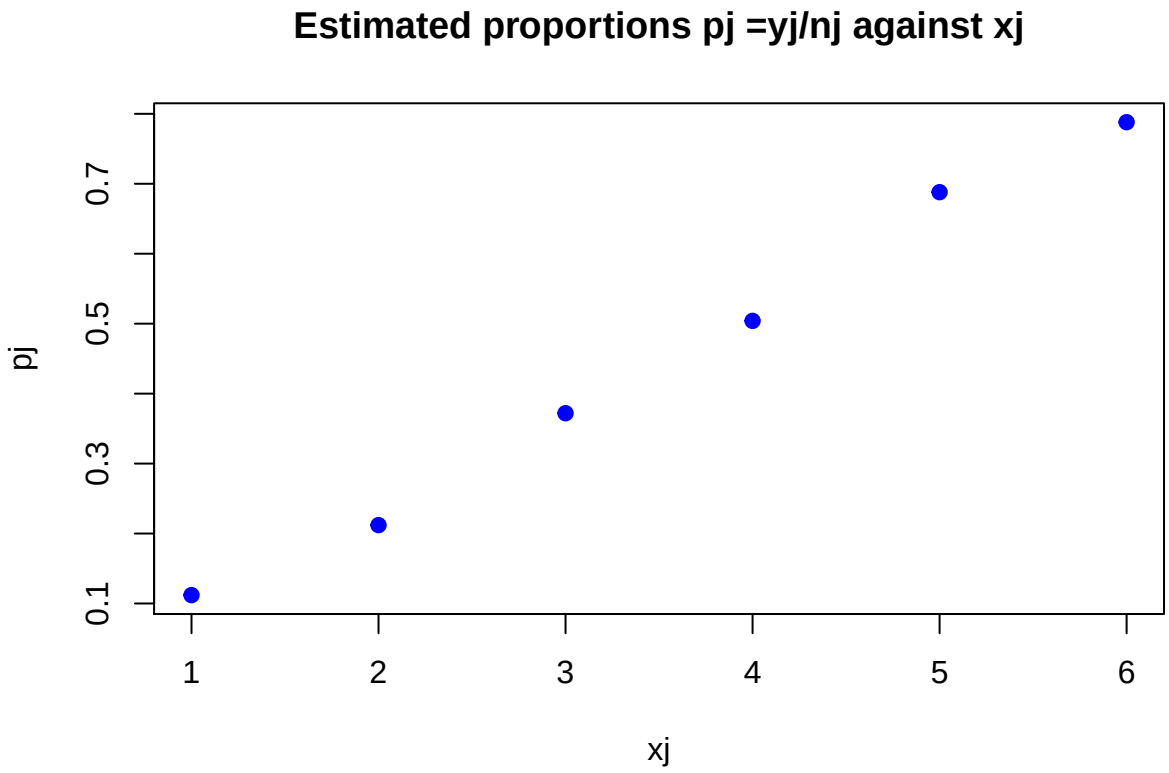
```r
# datahantering
xj <- 1:6
yj <- c(28,53,93,126,172,197)
nj <- rep(250,6)

SENIC <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/Data%20set%20C.1")
SENIC <-SENIC[,2:12]
cols2 <-c("length","Age","infectionRisk","RoutineCR","RoutineChXrR","Nmbrofbeds",
          "Medschal","Reg","ADC","NmbrofNurs","AvaFAS")
colnames(SENIC) <-cols2
```

Uppgifter

1 Simple logistic regression

a) Plot the estimated proprtions

```
pj <-yj/nj #Dividerar för att få fram proportionerna
dff <-data.frame(cbind(pj,nj,yj)) #Binder ihop vektorerna på kolumner
plot(y=pj,x=xj,col='blue', pch=19,main="Estimated proportions pj =yj/nj against xj")
```



Ovan syns ett diagram över den logaritmerade nivån av doser adminstrerade till insekter i grupp j, det som syns tydligt är hur andelen som avlider till följd av dosen som de injiceras med ökar i samband med att dosen ökar.

b. Find the maximum likelihood estimates of Bo and Bi. State the fitted response function.

```
MLE <-glm(y~dose,data = df,family = "binomial")#Skapar ett MLE
summary(MLE)
```

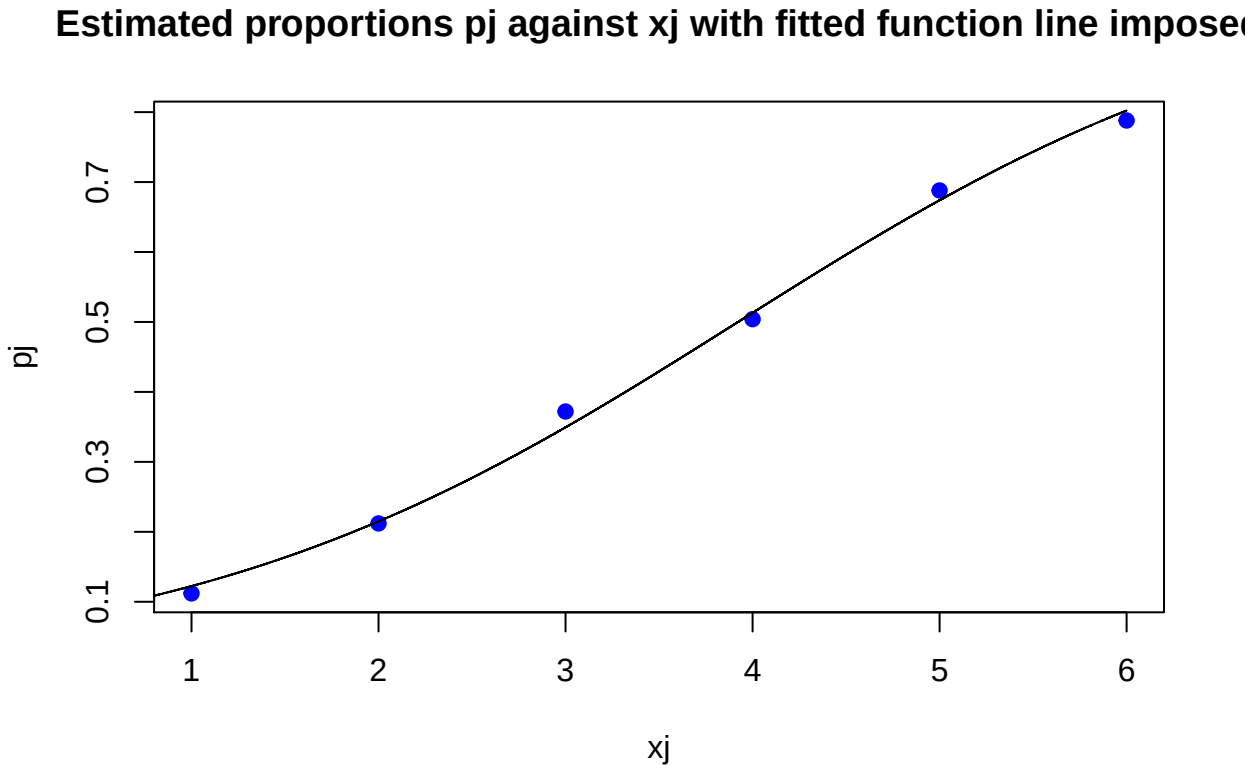
```
##
## Call:
## glm(formula = y ~ dose, family = "binomial", data = df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.800  -0.927  -0.511   0.888   2.050
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6437     0.1561   -16.9   <2e-16 ***
## dose          0.6740     0.0391    17.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2061.9  on 1499  degrees of freedom
## Residual deviance: 1680.3  on 1498  degrees of freedom
## AIC: 1684
##
## Number of Fisher Scoring iterations: 4
```

$$\pi(x) = \frac{\exp(2.644 - .674x_1)}{1 + \exp(2.644 - .674x_1)}$$

Den anpassade funktionen som skapats visar ett negativt intercept, där man behöver nästintill 4 dosenheter för att nå 0. Båda variablerna är signifikanta och har väldigt små p-värden.

c. Obtain a scatter plot of the data with the estimated proportions from part (a), and superimpose the fitted logistic response function from part (b). Does the fitted logistic response function appear to fit well?

```
xweight <- seq(0, 6, 0.0001) #Skapar en sekvens som går mellan 0 och 6 med intervall
# på 0.0001
yweight <-predict(MLE,list(dose=xweight),type = "response")
plot(y=pj,x=xj,col='blue', pch=19,main="Estimated proportions pj against xj with fitted function line im
lines(xweight,yweight)
```



Den apassade linjen har en fin anpassning till linjen med små avvikelser, däremot är det alla förutom den tredje punkten som inte direkt berör linjen, sammanfattningsvis skulle jag säga att det är en modell som definitivt kan användas.

d. Obtain exp(bi) and interpret this number.

$$e^{0.674} = 1.962$$

Det är en ganska hög summa som lutningen får, och det blir väldigt tydligt hur Genom att ta e ^ i den summan vi vill logarithemera får vi ut ett värde som motsvarar den parametern i den logarithmerade modellen.

e. What is the estimated probability that an insect dies when the dose level is X = 3.5?

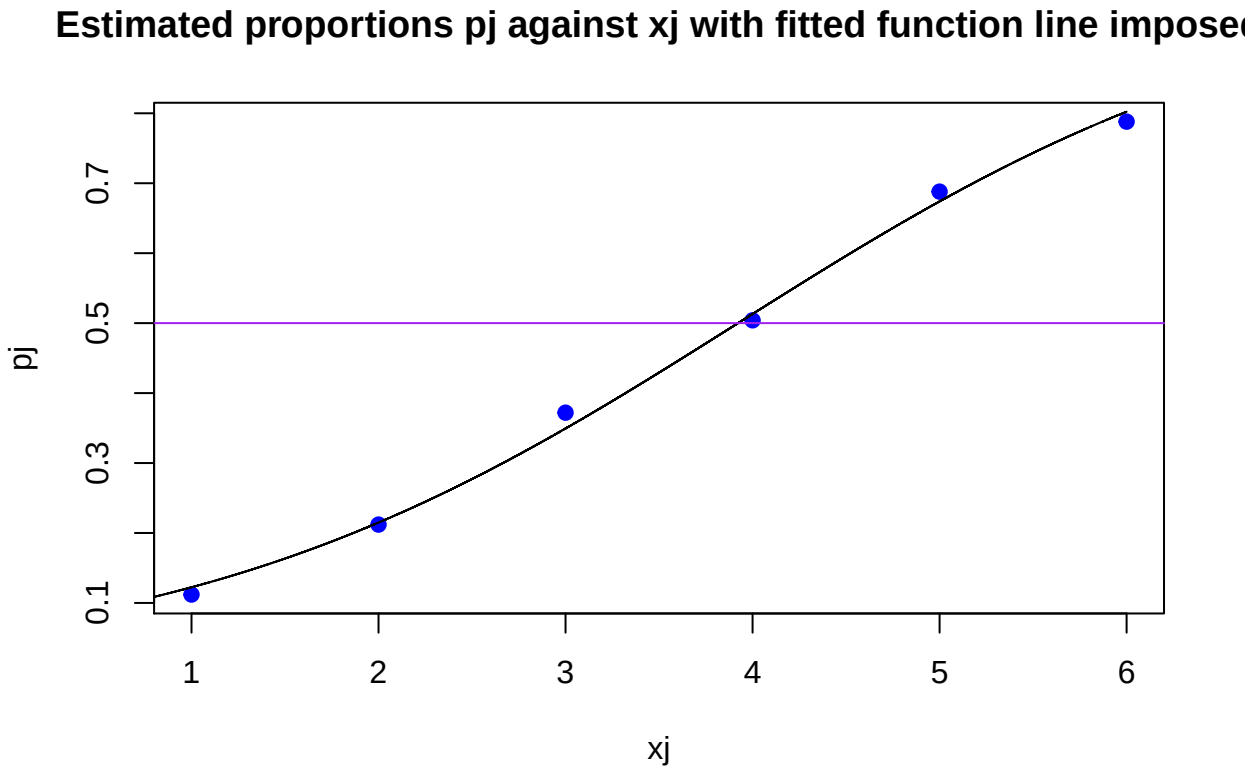
Formel

$$\pi(x) = \frac{\exp\left(\beta_0 + \beta_1 X_1\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1\right)} \Rightarrow \frac{\exp\left(\ 2.644 - .674 \cdot 3.5\right)}{1 + \exp\left(\ 2.644 - .674 \cdot 3.5\right)} = 0.429$$

Den skattade sannolikheten att en insekt dör när dosen uppgår till 3.5 enheter är 0.429, vilket kan tolkas som moderat, alltså inte högt eller lågt.

f. What is the estimated median lethal dose—that is, the dose for which 50 percent of the experimental insects are expected to die?

```
plot(y=pj,x=xj,col='blue', pch=19,main="Estimated proportions pj against xj with fitted function line im
lines(xweight,yweight)
abline(h=0.5,col="purple")
```



Genom att införa en linje precis vid 0.5 i dos visas hur den skattade median dosen skär precis innan 4 på x-axeln, alltså måste medianen vara lite mindre än 4, exakt 3.922.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1)}{1 + \exp(\beta_0 + \beta_1 X_1)} \Rightarrow \frac{\exp(-2.644 + .674 \cdot 3.922)}{1 + \exp(-2.644 + .674 \cdot 3.922)} = 0.4998$$

Refer to Toxicity experiment Problem 14.12. Assume that the fitted model is appropriate and

that large-sample inferences are applicable.

a. Obtain an approximate 99 percent confidence interval for Bl. Convert this confidence

interval into one for the odds ratio. Interpret this latter interval.

```
z <-qnorm(1-(0.01/2)) #Z-värde
S_b1 <-0.03911 #Standardfel
B1 <-0.67399 #Skattningen som tagits fram.
LCL <-B1-z*S_b1 #Lower confidence limit
UCL <-B1+z*S_b1 #Upper confidence limit
cbind(LCL,B1,UCL) #Binder ihop på kolumn
```

```
##         LCL    B1    UCL
## [1,] 0.5732 0.674 0.7747
```

I ett 99 procentigt konfidensinterval kommer B1 att ligga mellan 0.57 och 0.775. Intervallet är relativt litet sett till storleken på B1.

```
z <-qnorm(1-(0.01/2))
S_b1 <-0.03911
B1 <-0.67399
LCL_1 <-exp(B1-z*S_b1)
UCL_1 <-exp(B1+z*S_b1)
B1_1 <-exp(B1)
cbind(LCL_1,B1_1,UCL_1)
```

```
##      LCL_1  B1_1 UCL_1
## [1,] 1.774 1.962  2.17
```

För de exponentiella värdet kommer konfidensintervallet att ligga mellan 1.774 och 2.17 med 1 % signifikans.

b. Conduct a Wald test to determine whether dose level (X) is related to the probability that an insect dies; use a = .01. State the alternatives, decision rule, and conclusion. What is the approximate P-value of the test?

H0: B1=0

H1: B1≠0

```
z <-qnorm(1-(0.01/2))
S_b1 <-0.03911
B1 <-0.67399
teststatistiska <-B1/S_b1
```

I fall av att T $>$ 2.575 förkastas H_0, om inte, antas H_0.

Då T överstiger det kritiska värdet (17.23 $>$ 2.575), kan vi förkasta H_0 om att $B1 = 0$. Alltså kan vi inte droppa B1 då det är en statistiskt signifikant variabel på 1 % signifikansnivå.

Uppgift 2 Multiple logistic regression

a. Fit logistic regression model (14.4!) containing all predictor variables in the pool in first-order terms and interaction terms for all pairs of predictor variables. State the fitted response function.

```
SENIC$Medschal <-ifelse(SENIC$Medschal == 1,1,0)
model <-glm(Medschal~(Age+RoutineChXrR+ADC+NmbrofNurs)^2,family=binomial,data=SENIC)
summary.glm(model)
```

```
##
## Call:
## glm(formula = Medschal ~ (Age + RoutineChXrR + ADC + NmbrofNurs)^2,
##     family = binomial, data = SENIC)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9339  -0.2729  -0.1499  -0.0916   2.6446
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.84e+00   3.67e+01   -0.24    0.810
## Age                      2.24e-02   6.89e-01    0.03    0.974
## RoutineChXrR             5.64e-03   3.66e-01    0.02    0.988
## ADC                      1.47e-01   8.44e-02    1.74    0.081 .
## NmbrofNurs              -1.05e-01   7.78e-02   -1.35    0.178
## Age:RoutineChXrR         2.53e-04   6.94e-03    0.04    0.971
## Age:ADC                 -1.99e-03   1.64e-03   -1.22    0.224
## Age:NmbrofNurs           1.44e-03   1.52e-03    0.95    0.344
## RoutineChXrR:ADC        -3.35e-04   3.31e-04   -1.01    0.311
## RoutineChXrR:NmbrofNurs  3.91e-04   3.29e-04    1.19    0.234
## ADC:NmbrofNurs          -5.19e-06   2.14e-05   -0.24    0.809
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 95.706  on 112  degrees of freedom
## Residual deviance: 43.984  on 102  degrees of freedom
## AIC: 65.98
##
## Number of Fisher Scoring iterations: 7
```

Utifrån denna utskrift syns det hur majoriteten av variablerna inte är signifikanta sett till p-värde på 5 % signifikansnivå, den variabeln som är mest signifikant sett till p-värde är ADC.

```
SENICs <-SENIC
default <-SENICs
```

```r
#split dataset into training and testing set
data <-default

#split dataset into training and testing set
set.seed(2222)

sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE)
train <- data[sample, ]
test <- data[!sample, ]

#fit logistic regression model
model <- glm(Medschal~(Age+RoutineChXrR+ADC+NmbrofNurs)^2,family=binomial,data=train)

pred_y <- predict(model, test, type="response")

#convert defaults from "Yes" and "No" to 1's and 0's
test$Medschal <- ifelse(test$Medschal >= 0.5, 1, 0)

#find optimal cutoff probability to use to maximize accuracy
optimal <- optimalCutoff(test$Medschal, pred_y)[1]

#create confusion matrix
confusionMatrix(test$Medschal, pred_y,threshold = optimal)
```

```
##    0 1
## 0 53 3
## 1  3 8
```

```r
misClassError(test$Medschal, pred_y, threshold=optimal)
```

```
## [1] 0.0896
```

I detta fall gjorde modellen ett väldigt bra jobb, detta utifrån att modellen hade 91 % exakthet vilket är väldigt mycket, däremot kan säkert modellen göras bättre än detta.

b. Test whether all interaction terms can be dropped from the regression model; use a = .05. State the full and reduced models, clecision rule, and conclusion. What is the approximale P-value of the test?

$$H0: B5 = B6 = B7 = B8 = B9 = B10 = 0$$

$$H1: B5 \neq B6 \neq B7 \neq B8 \neq B9 \neq B10 = 0$$

```
anova(model,
      update(model, ~1),    # update here produces null model for comparison
      test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: Medschal ~ (Age + RoutineChXrR + ADC + NmbrofNurs)^2
## Model 2: Medschal ~ 1
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1        35       13.1
## 2        45       35.6 -10    -22.5    0.013 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(model)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Medschal
##                          LR Chisq Df Pr(>Chisq)
## Age                          4.10  1      0.043 *
## RoutineChXrR                 0.51  1      0.476
## ADC                          1.96  1      0.162
## NmbrofNurs                   0.14  1      0.710
## Age:RoutineChXrR             0.19  1      0.664
## Age:ADC                      0.08  1      0.781
## Age:NmbrofNurs               0.00  1      0.962
## RoutineChXrR:ADC             0.79  1      0.374
## RoutineChXrR:NmbrofNurs      1.33  1      0.248
## ADC:NmbrofNurs               0.94  1      0.333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\chi^2(0.95, 6) = 12.59$$

Det utskriften i detta fall säger är att ingen av interaktionstermerna har en statistiskt signifikant påverkan på 5 % signifikansnivå, vilket innebär att de kan droppas från modellen, ingen av dem överstiger heller de kritiska chitvå-värdet. Alltså kan vi på 5 % signifikansnivå inte förkasta nollhypotesen om att interaktionstermerna summerar till 0.

Lärdomar

Lärdomarna från denna laboration har varit många, det har varit väldigt annorlunda att jobba med logistiska regressions modeller, termer som tidigare varit signifikanta i vanliga regressionsmodeller är inte längre signifikanta när man har en annan variabel som responsvariabel och mäter på ett annat sätt.