Labbrapport i Statistik

Laboration 8

732G46

Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-10-22

Innehåll

Introduktion

I denna laboration kommer två dataset användas.

I det första datasetet kommer data som behandlar ampuler som går sönder i samband med transfsers av paket att behandlas, antalet består av 10 observationer.

I Datasetet som kommer från en prospektiv studie som undersöker effekten av två typer av fall. 100 objekts släpptes och variabler såsom kön, balansindex, hållbarhetsindex, uppfinning togs i beaktan.

Målet med denna laboration är att se användbarheten i Poisson regression, tolka Poisson regressions koefficenter och estimera sannolikheter genom Poissons regressions funktion,

Databehandling

```
# datahantering
Flyg_data <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH01PR21.txt") #Laddar
colnames(Flyg_data) <-c("Ampuler","Transfers") #Namnger kolumnerna

Geriatric <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/Geriatric%20Study")
colnames(Geriatric) <-c("y","x1","x2","x3","x4")
```

Uppgifter

Assigment 1: Simple Poisson regression

a) Fit the Poisson regression model(14.113) with the response function.

```
model <-glm(Ampuler~Transfers,data = Flyg_data,family = "poisson")
summary.glm(model)
```

```
##
## Call:
## glm(formula = Ampuler ~ Transfers, family = "poisson", data = Flyg_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8105  -0.2389  -0.0203   0.3299   0.6074
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3529     0.1317   17.86  < 2e-16 ***
## Transfers     0.2638     0.0792    3.33  0.00087 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12.5687  on 9  degrees of freedom
## Residual deviance:  1.8132  on 8  degrees of freedom
## AIC: 50.39
##
## Number of Fisher Scoring iterations: 4
```

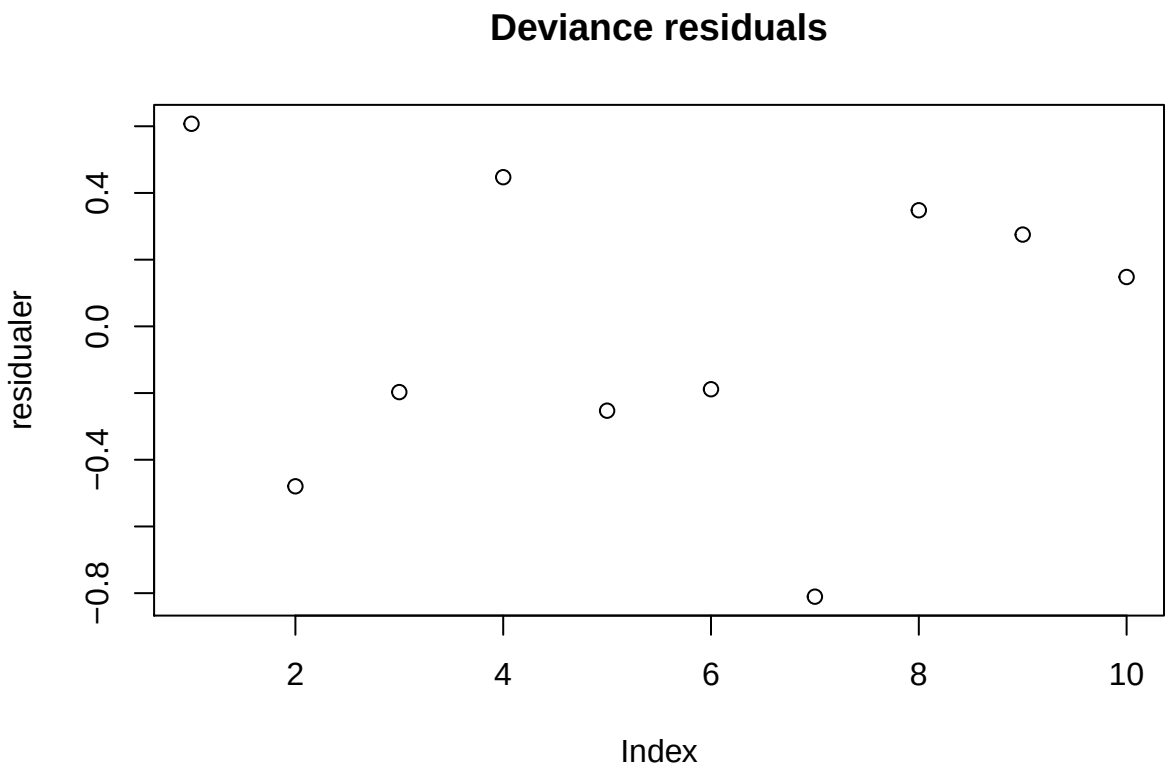
Ovan visas betaparametrarnas skattning och nedan deras standardfel, det som kan utläsas från skattningen är att utan ett transfer kommer ändå 2.353 ampuler att gå sönder under transport, därefter ökar antalet ampuler som går sönder med 0.264 enheter per transfer som sker, eftersom att poisson regression använder sig av länkfunktioner, exponentieras givetvis parametrarna med logarithm som visas nedan. $s\{\beta_0\} = .1317$ $s\{\beta_1\} = .0792$

a) Response function.

$\hat{\mu} = \exp(2.353 + .264x)$

b) Obtain the deviance residuals and present them in a index plot. Do there appear to be outlying cases?

```
residualer <-residuals.glm(model)
plot(x=1:10,y=residualer,xlab="Index",main="Deviance residuals")
```



I punktdiagrammet som visas ovan kan vi se resdiualerna som anpassats för regressions modellen av Poisson form. Några större avvikelsser syns inte förutom den första och sjunde observationen.

c) Estimate the mean number of ampules broken when X = 0,1,2,3 Compare these estimates with those obtained by menas of the fitted linear regression funcition in 1.21a.

```
new.data <-data.frame(Transfers=c(0,1,2,3))
pred <-predict.glm(model,
            newdata = new.data,
            type = "response"
)

model_0 <-lm(Ampuler~Transfers, data = Flyg_data)

pred_0 <-predict.lm(model_0,
                    newdata =new.data,
                    type = "response")

df <-as.data.frame(rbind(pred,pred_0))
rownames(df) <-c("Poission","Lm")
colnames(df) <-c(0,1,2,3)
df
```

```
##             0    1    2    3
## Poission 10.5 13.7 17.8 23.2
## Lm       10.2 14.2 18.2 22.2
```

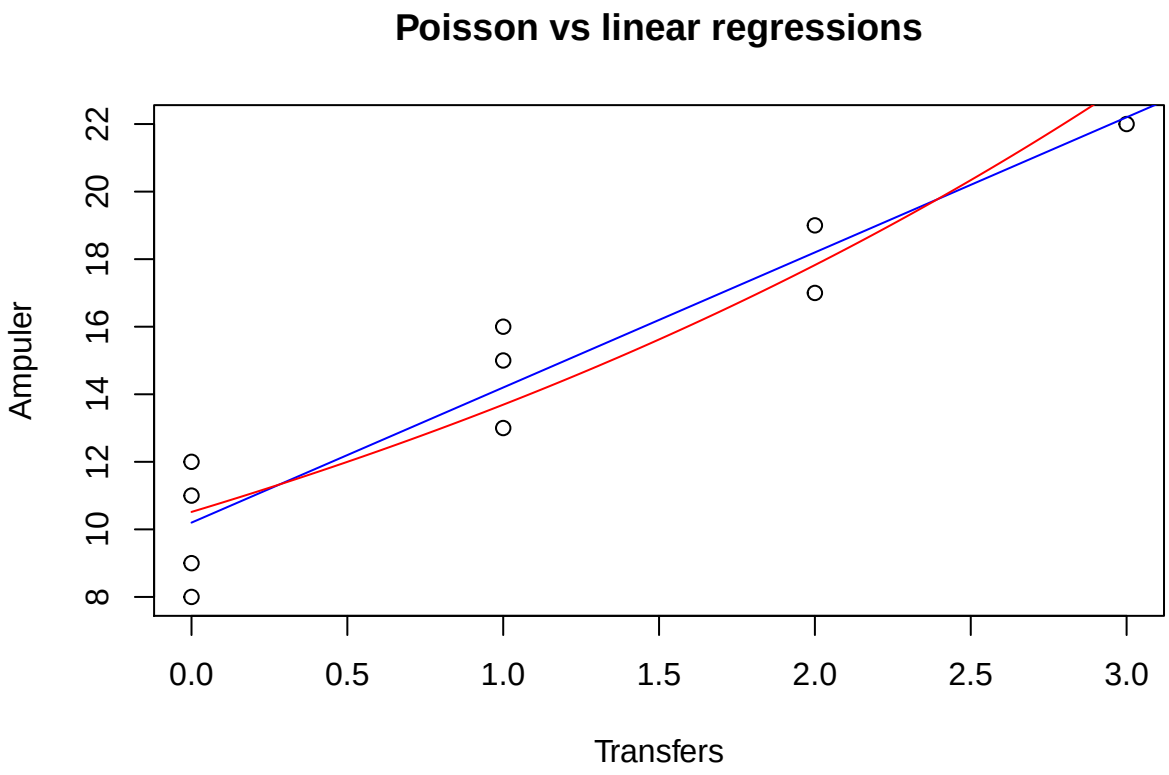
Det som tolkas från tabellen är hur interceptet för Poisson skattning är högre men att b0 är större vilket innebär att prediktioner med poisson modellen retuernar större värden efter 3 eller fler transfers. Den enkla regressions modellen returneras med ett högre intercept som för mellan 0 och 2 transfers i prediktionsmodellen returnerar högre värden än Poissonmodellen.

d. Plot the Poisson and linear regression functions together with the data, which regression function appears to be a better fit here? Discuss

```
Pseq <- seq(0, 10, 0.001)
pweight <-predict(model_0,list(Transfers=Pseq),
                  type = "response")

sweight <-predict(model,list(Transfers=Pseq),
                  type = "response")

plot(x=Flyg_data$Transfers,y=Flyg_data$Ampuler,
     main = "Poisson vs linear regressions",
     xlab = "Transfers",
     ylab="Ampuler")
lines(Pseq,pweight,col="blue")
lines(Pseq,sweight,col="red")
```



För enkelhetens skull kommer jag att benämna Poissons modellen som röd linje och normala regressions modellen blå linje.

Till en början kan vi se att blå linje har en bättre spridningen mot samtliga residualer då den står mellan alla 4 istället för att gå som röd linje som går närmare de 2 högre observationerna. Vid den första observationen

så har röd linje större spridning från observationerna till skillnad från blå linje som allmänt ligger närmare. I detta fall har vi tyvärr bara observationer som sträcker sig från 0 till 3 observationer, vilket tyvärr utesluter möjligheten att se hur linjerna skulle följa datat fortsatt, däremot skulle valet i detta fall varit att använda sig den blå modellen.

e. Management wishes to estimate the probability that 10 or fewer ampules are broken when there is no transfer of the shipment. Use the fitted Poisson regression function obtain this estimate

```
Y <-1:10
W <-exp(2.3529)
sum(W^Y*exp(-W)/factorial(Y))
```

```
## [1] 0.519
```

$$P\left(Y \leq 10 \mid Transfers = 0\right) = \sum_{Y=0}^{10} \frac{(10.516)^{Y} \exp(-10.516)}{Y!} = 0.519$$

Sannolikheten att 10 eller färre ampuler går sönder under transporter där inte en enda transfer sker är 0.519 .

f. Obtain an approximate 95 % confidence interval for B1, interpret your interval estimate

$$\beta_1 \pm t_{\alpha/2}^{(n-1)} \cdot s_{\beta_1} \Rightarrow 0.264 \pm 2.26 \cdot 0.0792 = 0.085 : 0.443$$

För lutningsparametern β_1 kommer ett 95 procentigt konfidensintervall att ligga mellan 0.085 och 0.443 vilket är ett relativt stort intervall. Däremot täcker inte intervallet 0 och kan ses som tillförlitligt.

Assignment 2: Multiple Poisson regression

a. Fit the Poisson regression model with response function.

```
Ger_mm <-glm(y~x1+x2+x3+x4,data = Geriatric,family = "poisson")
summary.glm(Ger_mm)
```

```
##
## Call:
## glm(formula = y ~ x1 + x2 + x3 + x4, family = "poisson", data = Geriatric)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -2.185  -0.782  -0.256   0.545   2.362
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.48947    0.33687    1.45   0.1462
## x1          -1.06940    0.13315   -8.03  9.6e-16 ***
## x2          -0.04661    0.11997   -0.39   0.6977
## x3           0.00947    0.00295    3.21   0.0013 **
## x4           0.00857    0.00431    1.99   0.0470 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 199.19  on 99  degrees of freedom
## Residual deviance: 108.79  on 95  degrees of freedom
## AIC: 377.3
##
## Number of Fisher Scoring iterations: 5
```

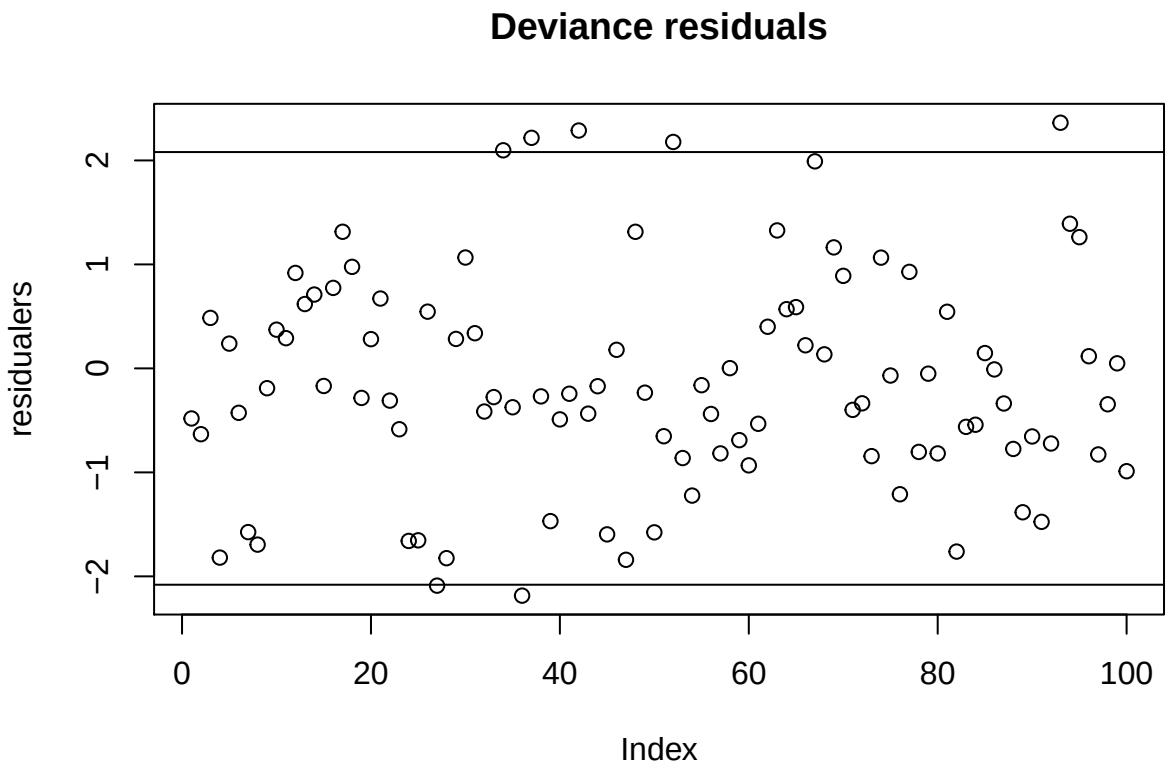
Ovan visas den skattade regressionsmodellen, ett intercept som är positivt medan de två första lutningsparametrarna har en negativ inverkan samt att de två sista har en positiv inverkan.

a) Response functioon

$$\hat{\mu} = \exp\left(.4895 - 1.0694x_1 - .0466x_2 + .0095x_3 + .0086x_4\right)$$

b) Obtain the deviance residuals and present them in a index plot. Do there appear to be outlying cases?

```
residualers <-residuals.glm(Ger_mm)
s <-2*sd(residualers)
plot(x=1:100,y=residualers,xlab="Index",main="Deviance residuals")
abline(h=s)
abline(h=-s)
```



Genom att införa en linje vid två standardavvikelser för residualerna kan vi se att det är 5 observationer som ligger utanför, en observation under den nedre standardavvikelserna och 4 ovanför, dessa observationer kan vara bra att ha med sig men slutsatsen kring att de inte påverkar de stora hela dras.

c) Assuming the fitted model is appropriate, use the likelihood ratio test to determine whether gender (x2) can be dropped from the model at a = .05.

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

Full modell med variabeln kön

$$\hat{\mu} = \exp\left(.4895 - 1.0694x_1 - .0466x_2 + .0095x_3 + .0086x_4\right)$$

Reducerad modell utan variabeln

$$\hat{\mu} = \exp\left(.4438 - 1.077x_1 + .0095x_3 + .009x_4\right)$$

$$LR_{\text{Gender}} = -2\ln L_R - (-2\ln L_F) = 0.15$$

```
## Likelihood ratio test
##
## Model 1: y ~ x1 + x3 + x4
## Model 2: y ~ x1 + x2 + x3 + x4
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   4   -184
## 2   5   -184  1  0.15        0.7
```

$$\chi^2(.95, 1) = 3.84$$

I fall av att teststatistikan överstiger det kritiska värdet ($\chi^2 > 3.84$) förkastas H0 om att B2 = 0.

I detta fall då ($0.15 < 3.84$) samtidigt som att vi får ett p-värde som summeras till (0.7) kan vi med 95 % konfidens inte förkasta H0 om att B2 (Kön) inte har en statistiskt signifikant påverkan på modellen. Alltså har kön inte en statistiskt signifikant påverkan på modellen på 5 % signifikansgrad, vilket innebär att variabeln kan droppas från modellen.

d) For the fitted model containing only X1,X2 and X4 in first-order terms, obtain an approximate 95 percent confidence interval for B1. Interpret your confidence interval

$$\beta_1 \pm t_{\alpha/2}^{(n-1)} \cdot s_{\beta_1} \Rightarrow -1.078 \pm 1.98 \cdot 0.1314 = -1.34 : -0.0818$$

För en modell där samtliga förklarande variabler förutom x3 är inkluderade kommer β_1 i ett 95 % konfidensintervall att ligga mellan -1.34 och -.0818.

For the fitted model in 14.39d, predict the number of falls during the six months for a person that has been assigned to the intervention "only education", with balance index 52 and strength index 60.

```
Ger_mm_0 <-glm(y~x1+x3+x4,data = Geriatric,family = "poisson")
new_data_2 <-data.frame(
  x1=0,x3=52,x4=60
)
predict.glm(Ger_mm_0,newdata = new_data_2,type = "response")
```

```
##    1
## 4.37
```

För en person som 52 i balansindex, 60 i styrkeindex och utbildning enbart förväntas 4.4 som avrundas uppåt till 5 fall under en 6 månaders period.

Lärdomar