GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

CURS 2015-2016 REAVALUCACIÓ – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: 5 de Juliol a les 17:00h **Aula -002-FME)**

Nom de l´alumne:	**DNI:**
Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	Edifici C5 D217 o H6-67
Normativa de l'examen:	ÉS PERMÉS DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	3h 00 min
Sortida de notes:	Abans del 8 de Juliol al Web Docent de MLGz
Revisió de l'examen:	8 de Juliol a 10h a C5-217-C Nord

Problema 1 (5.0 punts): Resposta Binària

En 1846, el grup de Donner (famílies Donner i Reed) va deixar Springfield, Illinois als Califòrnia en vagons tancats. Després d'arribar a Fort Bridger, Wyoming, els líders van decidir buscar una nova ruta a Sacramento, però es van quedar atrapats a les muntanyes orientals de Sierra Nevada quan la regió es va veure afectada per les fortes nevades a finals d'octubre, en un lloc que ara s'anomena el pas de Donner. Els supervivents van ser rescatats el 21 d'abril de 1847, 40 dels 87 havien mort. Les dades s'han obtingut del web https://onlinecourses.science.psu.edu/stat504. Les variables contingudes a l'arxiu de dades subministrat són:

C1. Edat de la persona.
C2. Gènere: 0 Dona 1 Home.
C3. F.survive: Resposta binària target, codificada com 1 si la persona va sobreviure i 0 altrament.

1. Es vol estudiar la relació entre el target (**f.survive**) i el gènere **(sex)**. Formuleu i calculeu el model logit que modela una probabilitat idèntica de sobreviure en els dos sexes. Useu les dades de la taula mostrada a continuació.

```
      age              sex        f.survive        f.age
 Min.    :15.0    female:15    y.no :25     25-29   :16
 1st Qu.:24.0    male   :30    y.yes:20     30-34   : 8
 Median :28.0                               35-39   : 4
 Mean   :31.8                               20-24   : 3
 3rd Qu.:40.0                               40-44   : 3
 Max.   :65.0                               50-54   : 3
                                            (Other): 8

> table(df$f.survive)
y.no y.yes
  25    20
> table(df$sex,df$f.survive)
        y.no y.yes
 female    5    10
 male     20    10
```

Heu d'estimar el model nul:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad i=1:2\equiv \mathit{female}, \mathit{male}$$

La probabilitat marginal de sobreviure (resposta positiva) és: 20/45 = 0.444 i els odds 20 a 25 (o 0.80 a 1), per tant el logodd és -0.86, l'estimador dde la constant en el model nul. Si ho fessim amb R les comandes podrien estar:

```
> table(df$f.survive)

 y.no y.yes
   25    20
> tt0<-prop.table(table(df$f.survive));tt0
```

```
      y.no     y.yes
0.5555556 0.4444444
> tt0[2]
    y.yes
0.4444444
> eta<-log(20/25);eta
[1] -0.2231436
> m0<-glm(df$f.survive~1, family=binomial, data=df)
> summary(m0)

Call: glm(formula = df$f.survive ~ 1, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.2231     0.3000  -0.744    0.457

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 61.827  on 44  degrees of freedom
AIC: 63.827
```

2. Formuleu i calculeu el model logit que modela una probabilitat específica de sobreviure per cadascun dels dos sexes. Useu les dades de la taula mostrada anteriorment.

Heu d'estimar a partir de la taula el model Y-A on Y és el target f.survive i el factor A és sex. És pel nivell d'agregació mostrat per la taula un model saturat i per això po deu estimar-lo directament doncs serà un model que reprodueix exactament les obser vacions. Sigui female (i=1) el nivell de referència, per tant el logodd d'aquest grup co nstitueix l'estimador de la constant en el model Y-A, log(10/5)=0.69:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta+\alpha_i \quad i=1:2 \quad \alpha_{1\equiv female}=0$$

Les diferències dels logodds les homes (resta de grups) respecte el nivell de referència c onstituiran els estimadors dels efectes aditius del nivel male sobre la referencia (fema le):

log(10/20)- log(10/5)=-1.39
Per tant, ser home redueix els logodds de la probabilitat de sobreviure en 1.39 untats o equivalentment els odds de sobreviure en home es redueixen en (1-exp(-1.3863))*1 00 = 75% respecte les dones (grup de referència)
En R podria calcular-se:

```
> ptt<-prop.table(table(df$sex,df$f.survive),1);ptt

             y.no     y.yes
  female 0.3333333 0.6666667
  male   0.6666667 0.3333333
> tt<-table(df$sex,df$f.survive);tt

         y.no y.yes
  female    5    10
  male     20    10
> lodd<-log(tt[,2]/tt[,1]);lodd
    female        male
```

```
 0.6931472 -0.6931472
> lodd[2]-lodd[1]
     male
-1.386294
> m1<-glm(df$f.survive~sex, family=binomial, data=df)
> summary(m1)

Call: glm(formula = df$f.survive ~ sex, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6931     0.5477   1.266   0.2057
sexmale      -1.3863     0.6708  -2.067   0.0388 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 57.286  on 43  degrees of freedom
AIC: 61.286
```

3. Formuleu i calculeu el model probit que modela una probabilitat específica de sobreviure per cadascun dels dos sexes. Useu les dades de la taula mostrada anteriorment.

Sigui female (i=1) el nivell de referència, per tant el probit de la probabilitat de sobreviure d'aquest grup constitueix l'estimador de la constant en el model Y-A, probit(10/15)=probit(0.667)=qnorm(0.667)=0.431:

$$probit(\pi_i) = \eta + \alpha_i \quad i = 1:2 \quad \alpha_{1\equiv female} = 0$$

Les diferències dels pròbits dels homes (resta de grups) respecte el nivell de referència constituiran els estimadors dels efectes aditius del nivell male sobre la referencia (female):

probit(10/30)- probit(10/15)=-0.862
Per tant, ser home redueix els probit de la probabilitat de sobreviure en 0.862 untats respecte les dones (grup de referència)
En R podria calcular-se:

```
> ptt<-prop.table(table(df$sex,df$f.survive),1);ptt

             y.no     y.yes
  female 0.3333333 0.6666667
  male   0.6666667 0.3333333
> tt<-table(df$sex,df$f.survive);tt

         y.no y.yes
  female    5    10
  male     20    10
> qodd<-qnorm(tt[,2]/(tt[,1]+tt[,2]));qodd
    female       male
 0.4307273 -0.4307273
> qodd[2]-qodd[1]
      male
-0.8614546
> m1<-glm(df$f.survive~sex, family=binomial(probit), data=df)
> summary(m1)

Call:  glm(formula = df$f.survive ~ sex, family = binomial(probit),
```

```
    data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.4307     0.3348   1.287   0.1982
sexmale      -0.8615     0.4100  -2.101   0.0356 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 57.286  on 43  degrees of freedom
AIC: 61.286
```

4. Calculeu el nombre predit d'observacions que sobreviurien i no sobreviurien sota la hipòtesi del model nul descrit al **Punt 1**.

Heu de fer la predicció sobre el model nul que assigna la mateixa probabilitat de sobreviure als homens que a les dones, és a dir, la marginal 20/45=0.444, per tant el nombre predit seria:

Sex	f.survive=NO	f.survive=YES		Under m0	Predict =NO	Predict = YES
Female	5	10	**15**		15*0.666=8.33	15*0.444=6.67
Male	20	10	**30**		30*0.666=16.67	30*0.444=13.33
All	25	20	**45**		25	20

En R podria fer-se :

```
> tt<-table(df$sex,df$f.survive);tt

         y.no y.yes
  female    5    10
  male     20    10
> trow<-apply(tt,1,sum);trow
female   male
    15     30
> tt0<-prop.table(table(df$f.survive));tt0

     y.no     y.yes
0.5555556 0.4444444
> tt0[2]
    y.yes
0.4444444
> fitm0<-round(cbind(trow*(tt0[1]),trow*tt0[2]),dig=2);fitm0
        [,1]  [,2]
female  8.33  6.67
male   16.67 13.33
```

5. Calculeu la deviança del model nul descrit en el **Punt 1** usant les dades calculades per les prediccions en el **Punt 4**.

$$D = 2\sum_{i=1,2}\left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\} =$$

$$= 2\left(10\log\left(\frac{10}{6.67}\right) + 5\log\left(\frac{5}{8.33}\right) + 10\log\left(\frac{10}{13.33}\right) + 20\log\left(\frac{20}{16.67}\right)\right) = 4.53 \qquad \approx \chi^2_{n-p=2-1\ \ =1}$$

En R podríeu haver-ho fet:

```
> # Goodness of fit
> devm0a
```

```
[1] 4.531271
> 1-pchisq(devm0a,1)
[1] 0.03328088
> # Test Deviança
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: df$f.survive ~ 1
Model 2: df$f.survive ~ sex
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        44     61.827
2        43     57.286  1   4.5403  0.03311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. Valoreu si sobreviure està estadísticament associat al sexe (**sex**). Empreu, justifiqueu i interpreteu un test estadístic adient.

Amb els resultats dels apartats anteriors amb les dades agregades cal considerar que el model saturat Y-A té una deviança de 0 i el model nul 4.53. Per tant, es pot fer un test de bondat del model nul: H0- El model nul s'ajusta bé a les dades.
La distribució de l'estadístic deviança és assimptòticament una shi quadrat amb n-p =2-1=1 graus de llibertat i el p valor del constrast és per tant P(Shi_df1>4.53)=0.033, per tant, hi ha evidència per rebutjar la H0 i d'aquí que el model nul no s'ajusti bé a les dades, per tant, l'efecte del factor sex és significatiu.

7. Es calcula amb el conjunt de dades individuals el model logístic pel target (**f.survive**) en funció del sexe (**sex**). Indiqueu quins elements seran iguals o diferents en la sortida R del mètode `summary()` aplicat a un objecte de classe glm quan s'empran dades individualitzades i quan s'empran dades agrupades segons la definició del factor **sex**.

Model m1: $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \quad i = 1:2 \quad \alpha_{1\equiv female} = 0$

	Estimador dades individuals	**Estimador dades agrupades per sex**	**Iguals o diferents?**
Terme independent (intercept)	0.693	0.693	Iguals
Estimador dummy per sex	-1.386	-1.386	Iguals
Null deviance	61.827	4.53	Diferents
Graus llibertat Null Deviance	44	1	Diferents
Deviance	57.286	0	Diferents
Graus llibertat Deviance m1	43	0	Diferents
AIC	61.286	¿	Diferents
Dev(m0)-Dev(m1)	4.53	4.53	Idèntics

Empreu els resultats inclosos al final de l'enunciat del Problema 1.

8. S'estudia l'efecte de la covariable edat (**age**) sobre el target **f.survive**. Valoreu i interpreteu l'equació del millor model disponible. Creieu que calen termes quadràtics o cúbics?

Segons les sortides disponibles els termes cúbics i quadràtics no són significatius i poden interpretar-se els pvalors de la sortida directament.
El model que conté el terme lineal de l'edat directament com a covariant mostra un p-valor del coeficient del 0.04, per tant tècnicament per sota del llindar habitual del 5%. L'AIC del model polinòmic fins l'ordre 2 és 62.037 superior al model m8 amb terme lineal original és 60.29, menor per tant millor des de punt de vista d'aquest criteri.

9. Considerar els models pel target **f.survive** amb l'agrupació de l'edat en grups de 5 anys. Quin tractament considereu més adequat per la variable edat (**age**)? Justifiqueu estadísticament les respostes.

El tractament de l'edat com a factor emprant grups d'edat per 5 anys requereix de molts paràmetres, la qual cosa amb un joc de dades tant limitat ja fa sospitar que no donarà bon resultat.
El model (m9) amb tractament factor no pot comparar-se per test de la deviança amb cap dels models amb tractament numèric (m7 o m8), doncs no són encaixats. Per tant, només ens queda el criteri d'Akaike, el menor AIC prové del model amb covariant i terme lineal (m8) amb 60.29, mentre el tractament com a factor dona un AIC de 67.43 major, per tant, és pitjor. El tractament com a covariant de l'edat amb terme lineal és el més adequat.

10. Es necessita controlar per l'edat quan el sexe ja s'ha incorporat al model? Justifiqueu estadísticament les respostes.

Les dades disponibles corresponent al model aditiu age+sex i al model amb sex. El test de la deviança formularia la H0: 'els dos models són equivalent', D(sex)-D(age+sex)=57.286-51.256= 6.03
amb un pvalor obtingut a partir de la distribució asimptòtica de referència una shi quadrat a 1 grau de llibertat p valor = P(Shi quadrat 1>6.03)= 0.01406 < 0.05 (nivell de significació habitual) per tant hi ha evidència per rebutjar la hipòtesi nul.la i un cop entrat el gènere en el model, l'efecte net de l'edat és estadísticament significatiu.

11. Estudieu els models que usen l'edat i el sexe en el predictor lineal. Determineu quin model és el més adient, justificant estadísticament la resposta.

El model (m11) conté el model amb interaccions entre age i sex. Segons la sortida disponible el pvalor de l'efecte net de la interacció entre age i sex és estadísticament significativa, tècnicament, p valor = 0.048 < 0.05, per tant, els efectes principals i les interaccions han de considerar-se i aquest serà el millor model a falta de sortides que ens permetin raonar amb altres criteris (per exemple Akaike).

12. Interpreta el millor model obtingut al punt 11 en termes dels logodds, odds i aproximadament de probabilitats.

El model és m11 que correspon a una formula en R: Y~A*X amb la família binomial i link canònic lògit:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=(\eta+\alpha_i)+(\gamma+\theta_i)age \quad i=1:2 \quad \alpha_{1\equiv female}=0$$

Per les dones:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=(\eta+0)+(\gamma+0)age=7.25-0.19age$$

Pels homes:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=(\eta+\alpha_2)+(\gamma+\theta_2)age=(7.25-6.93)-(0.19+0.16)age=0.32-0.03age$$

La interpretació del model seria que per cada any d'edat el logodds de sobreviure es decrement en en 0.19 unitats per les dones i només en 0.03 unitats pels homes; in terms of odds, 100*(1-exp(0.19))=17.30 and 100*(1-exp(0.03))=2.96, és a dir, els odds de sobreviure es decrementen en un 17.3% a les dones i només en un 2.96% en els homes per cada any de més. The median is 28 for the overall sample, 25.0 for women and 28.0 for men. At age 28, la probabilitat de sobreviure per una dona és de 0.86 i per un home de 0.35, però el pas dels anys no castiga de la mateixa manera als dos gèneres.

```
> tapply(df$age,df$sex,summary)
$female
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   22.50   25.00   31.07   42.50   50.00

$male
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   25.00   28.00   32.17   34.25   65.00

> 100*(1-exp(-0.19)) ;100*(1-exp(-0.03))
[1] 17.30409
[1] 2.955447
> predict(m11, newdata=data.frame(age=28,sex="female"),type="response")
        1
0.8596407
> predict(m11, newdata=data.frame(age=28,sex="male"),type="response")
       1
0.356399
```

13. En el model del Punt 11 es vol fer una diagnosi per determinar la presència d'observacions influents i residuals atípics. Amb els resultats disponibles indiqueu les observacions potencialment influents, les observacions que constitueixen definitivament dades influents i aquelles que són *outliers* dels residus.

Són molt poques observacions, el més remarcable és una dada influent a jutjar per la seva distància de Cook i residus studentitzat: d'una dona de 25 anys que va morir (obs 11). La següent observació remarcable, però molt menys és la 2, pertanyent a una dona de 40 anys que va sobreviure, no és un outlier dels residus. L'home de 65 anys (obs 9) té un factor d'apalancament generalitzat de 0.205 > 0.18 (2*p/n=2*4/45=0.18, llindar de referencia), però no afecta al càlcul dels coeficients al tenir una distancia de Cook moderada

RESULTATS PEL PROBLEMA 1

```
> summary(m8a)
Call: glm(formula = f.survive ~ poly(age, 3), family = binomial, data = df)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.146      1.083  -1.058    0.290
poly(age, 3)1  -19.718     17.138  -1.151    0.250
poly(age, 3)2  -12.797     13.591  -0.942    0.346
poly(age, 3)3   -7.771      7.602  -1.022    0.307

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 54.037  on 41  degrees of freedom
AIC: 62.037
> summary(m8)
```

```
Call:glm(formula = f.survive ~ age, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.81852    0.99937   1.820   0.0688 .
age         -0.06647    0.03222  -2.063   0.0391 *
---
    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 56.291  on 43  degrees of freedom
AIC: 60.291

> summary(m9)
Call:   glm(formula = f.survive ~ f.age, family = binomial, data = df)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.493e-16  1.414e+00   0.000    1.000
f.age20-24   1.857e+01  3.766e+03   0.005    0.996
f.age25-29  -2.513e-01  1.501e+00  -0.167    0.867
f.age30-34  -5.108e-01  1.592e+00  -0.321    0.748
f.age35-39   1.099e+00  1.826e+00   0.602    0.547
f.age40-44   6.931e-01  1.871e+00   0.371    0.711
f.age45-49  -1.857e+01  4.612e+03  -0.004    0.997
f.age50-54  -6.931e-01  1.871e+00  -0.371    0.711
f.age60-64  -1.857e+01  4.612e+03  -0.004    0.997
f.age65-70  -1.857e+01  4.612e+03  -0.004    0.997

    Null deviance: 61.827  on 44  degrees of freedom
Residual deviance: 47.425  on 35  degrees of freedom
AIC: 67.425

> Anova(m9)
Analysis of Deviance Table (Type II tests)
Response: f.survive
      LR Chisq Df Pr(>Chisq)
f.age   14.402  9     0.1087
```

```
> m10<-glm(f.survive~sex, family=binomial, data=df)
> m10a<-glm(f.survive~age+sex, family=binomial, data=df)
> anova(m10a)
Analysis of Deviance Table
Model: binomial, link: logit
Response: f.survive
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                    44     61.827
age   1   5.5358        43     56.291
sex   1   5.0344        42     51.256
> anova(m10)
Analysis of Deviance Table
Model: binomial, link: logit
Response: f.survive
Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                    44     61.827
sex   1   4.5403        43     57.286
>
```

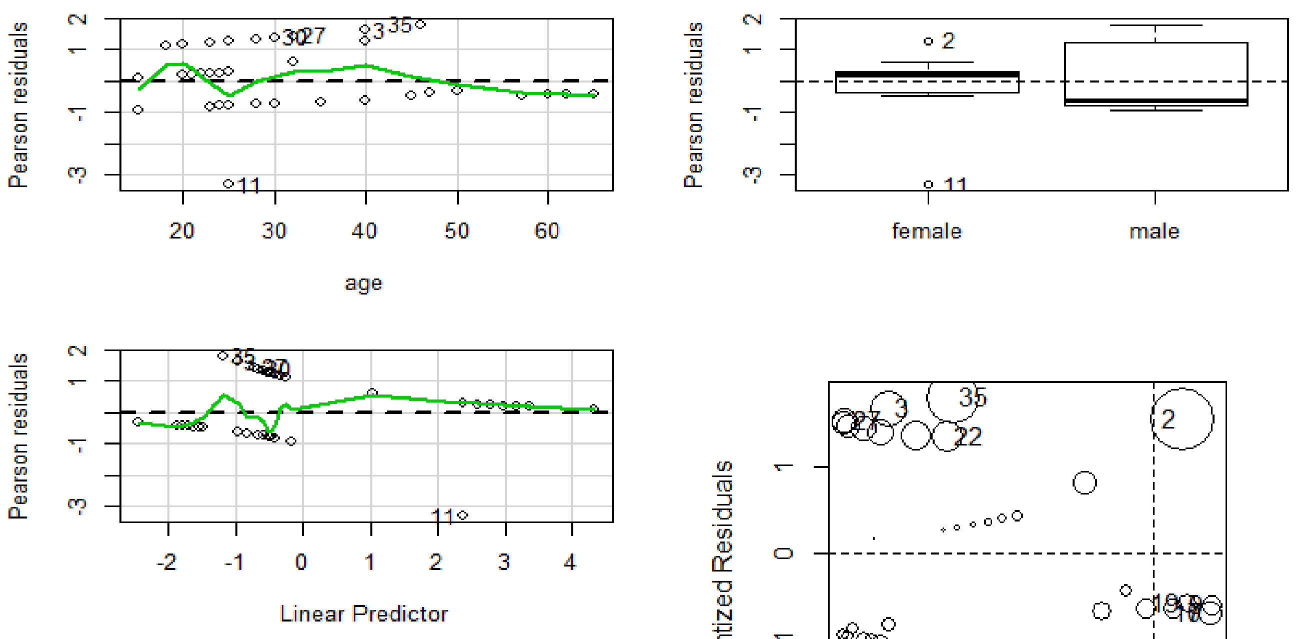
```
> m11<-glm(f.survive~age*sex, family=binomial, data=df)
> Anova(m11)
Analysis of Deviance Table (Type II tests)

Response: f.survive
        LR Chisq Df Pr(>Chisq)
```

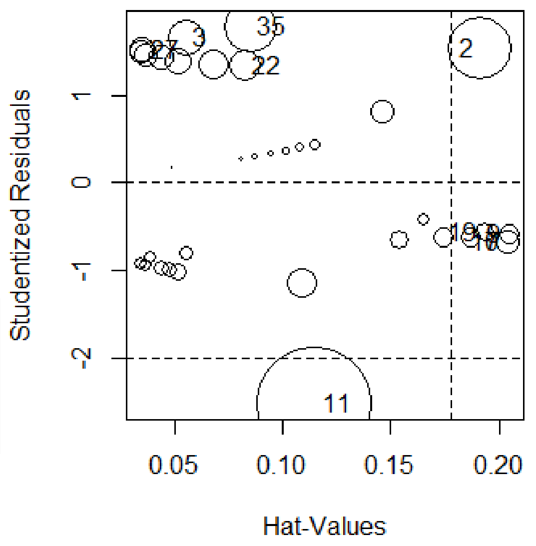
```
age       6.0300  1    0.01406 *
sex       5.0344  1    0.02485 *
age:sex   3.9099  1    0.04800 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> summary(m11)
Call: glm(formula = f.survive ~ age * sex, family = binomial, data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  7.24638    3.20517   2.261   0.0238 *
age         -0.19407    0.08742  -2.220   0.0264 *
sexmale     -6.92805    3.39887  -2.038   0.0415 *
age:sexmale  0.16160    0.09426   1.714   0.0865 .
---
```




```
> residualPlots(m11,id.n=5)
    Test stat Pr(>|t|)
age     1.516    0.218
sex        NA       NA
> llista<-influencePlot(m11,id.n=5);llista
      StudRes        Hat     CookD
2   1.5381571 0.19150978 0.3504005
3   1.6594231 0.05501176 0.2026468
7  -0.6822812 0.20388668 0.1348349
9  -0.5923527 0.20465145 0.1160494
10 -0.6822812 0.20388668 0.1348349
11 -2.5268811 0.11477866 0.6335992
19 -0.5675311 0.19340541 0.1067521
22  1.3369906 0.08248291 0.1787946
27  1.5195979 0.03506587 0.1391439
35  1.7858609 0.08476873 0.2863119
```




```
> df[row.names(llista),]
  age    sex f.survive f.age
2   40 female     y.yes 40-44
3   40   male     y.yes 40-44
7   45 female      y.no 45-49
9   65   male      y.no 65-70
10  45 female      y.no 45-49
11  25 female      y.no 25-29
```

Problema 2 (5 Punts): Targetes de Crèdit de Viatge

El quadre següent es refereix a una mostra d'individus seleccionats a l'atzar per a un estudi italià sobre la relació entre els ingressos (income) i si un posseeix una targeta de crèdit de viatge (com American Express o Diner Club). A cada nivell d'ingressos anuals en milions de lires (la moneda a Itàlia abans de l'euro), la taula indica el nombre d'individus inclosos en la mostra i el nombre d'aquests individus que posseeixen com a mínim una targeta de crèdit de viatge.

En aquest exemple es té informació sobre els individus agrupats pel seu ingrés, el nombre d'individus (casos) dins d'aquest grup d'ingressos i el nombre de targetes de crèdit. Dades procedents de https://onlinecourses.science.psu.edu/stat504.

```
> df$logsize<-log(df$size)
> dim(df)
[1] 31  4
> summary(df)
     income            size             ntcc       logsize
 Min.   : 24.00   Min.   : 1.000   Min.   :0   Min.   :0.0000
 1st Qu.: 33.50   1st Qu.: 1.000   1st Qu.:0   1st Qu.:0.0000
 Median : 45.00   Median : 2.000   Median :0   Median :0.6931
 Mean   : 53.42   Mean   : 3.226   Mean   :1   Mean   :0.8236
 3rd Qu.: 66.50   3rd Qu.: 5.000   3rd Qu.:1   3rd Qu.:1.6094
 Max.   :130.00   Max.   :10.000   Max.   :6   Max.   :2.3026
> sum(df$ntcc);sum(df$size)
[1] 31
[1] 100
```

1. Considereu el model nul: calculeu-lo amb les dades disponibles. Quin és el nombre esperat de targetes de crèdit de viatge que té un individu amb ingressos al voltant dels 120 milions de lires?.

El model nul s'escriu

$$\log(\mathrm{E}[Y_i]) = \log(n_i\mu_{ij}) = \log(n_i\mu_i) = \log(n_i) + \log(\mu_i) = \log(n_i) + \eta$$

El nombre esperat de targetes en el grup i-èssim pel tamany del grup i-èssim serà el nombre total de targetes de viatge en el grup i-èssim. En el cas del model nul, cal modelar el nombre esperat de targetes per persona de qualsevol grup de la mateixa manera. Amb les dades disponibles el nombre mig de targetes per persona a la mostra és 31/100, en aplicar el log(0.31)=-1.1712, precisament el nombre esperat de targetes de viatge per persona en qualsevol dels grups i en concret en el grup sol.licitat. En R dona:

```
> summary(m0)
Call:
glm(formula = ntcc ~ offset(logsize), family = poisson, data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1712     0.1796  -6.521 6.99e-11 ***
---
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 42.078  on 30  degrees of freedom
AIC: 79.218
```

2. Els ingressos són una variable estadísticament significativa per explicar el nombre de targetes de viatge que disposa un individu?

El model que conté els ingressos (income) linealment té un efecte brut estadísticament significatiu en ser el pvalor de la variable en el model m1 de 5.85e-05 (test de Wald).

Un altre argument seria emprar el test de deviances i comparar la deviança entre el model m1 i model nul, ambdós disponibles: D(m0)-D(m1)=42.078-28.465=13.613 i P(X21>13.613)=0.000225<<0.05 per tant es rebutja la hipòtesi d'equivalència i per tant els dos models no són equivalents i l'efecte de l'income és estadísticament significatiu. En R:

```
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: ntcc ~ offset(logsize)
Model 2: ntcc ~ offset(logsize) + income
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        30     42.078
2        29     28.465  1   13.613 0.0002246 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> waldtest(m0,m1,test="Chisq")
Wald test

Model 1: ntcc ~ offset(logsize)
Model 2: ntcc ~ offset(logsize) + income
  Res.Df Df  Chisq Pr(>Chisq)
1     30
2     29  1 16.154  5.839e-05 ***
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Penseu que el model loglineal que conté els ingressos amb els termes lineal, quadràtic i cúbic és significativament millor que el model lineal amb ingressos? Justifiqueu l'argument amb un contrast d'hipòtesi adequat.

Només en veure el summary del model m2 calculat amb els polinomis ortogonals per els ingressos, els pvalors dels termes quadràtics i cúbics són superiors a 0.05, i per tant, no són significatius.

```
> summary(m2)
Call: glm(ntcc ~ offset(logsize) + poly(income, 3), family = poisson,data = df)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.3300     0.2118  -6.281 3.36e-10 ***
poly(income, 3)1    3.8118     1.0165   3.750 0.000177 ***
poly(income, 3)2   -1.3903     0.9728  -1.429 0.152952
poly(income, 3)3    0.5350     1.2060   0.444 0.657321
---
```

Adicionalment es disposa les dades del contrast de la deviança entre el model lineal i el cúbilc dels ingressos, que amb un pvalor de 0.33 >>0.05 no hi ha evidència per rebutjar la hipòtesi nulla i per tant són equivalents, el que indica que els termes quadràtics i cúbics no són significatius, no cal complicar el model amb income lineal, no es millora substancialment.

```
> anova(m1,m2,test="Chisq")
Analysis of Deviance Table

Model 1: ntcc ~ offset(logsize) + income
Model 2: ntcc ~ offset(logsize) + poly(income, 3)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        29     28.465
2        27     26.256  2   2.2083   0.3315
```

4. Interpreteu el model m1, en l'escala del predictor lineal i de la resposta. Escriviu les equacions del model indicat.

L'equació del model amb la covariant ingressos (income) és:

$$log(\mathrm{E}[Y_i]) = \log(n_i\mu_{ij}) = \log(n_i\mu_i) = \log(n_i) + log(\mu_i) =$$
$$= \log(n_i) + \eta + \theta \cdot income = \log(n_i) - 2.387 + 0.0208 income$$

Per cada unitat d'increment d'income (és a dir per cada milió de lires addicionals), el logaritme del nombre mig de targetes de viatge per persona s'incrementa en 0.0208 unitats. En l'escala de la resposta, el nombre esperat de targetes de viatge per persona per income+1 multiplica per exp(0.0208) =1.021 respecte el nb obtingut per income.

5. Considereu el model m1. Quin és el nombre esperat de targetes de crèdit de viatge que té un individu amb ingressos al voltant dels 120 milions de lires?.

$$log(\mathrm{E}[Y_i]) = \log(n_i\mu_{ij}) = \log(n_i\mu_i) = \log(n_i) + log(\mu_i) \rightarrow$$
$$\rightarrow log(\mu_i) = \eta + \theta \cdot income = -2.387 + 0.0208 income = -2.387 + 0.0208 \cdot 120 = 0.109$$

$$\mu_i = \exp(0.109) = 1.1152$$

El nombre predit de targetes per individu es pot obtenir a partir del model estimat amb les dades agrupades gràcies a l´ús de l'offset com el logaritme del tamany de cada grup (nivell d'ingressos). Aplicant el model a un income =120, el logaritme del nombre de targetes de viatge per persona és 0.109 i exponenciant s'obté un nombre esperat de targetes pel valor 120 milions de lliures d'ingressos és de 1.1152.

6. Quanta gent esperaríeu que tinguessin com a mínim una targeta de crèdit de viatge en un grup de 10 persones que guanyen al voltant de 120 milions de lires?

El nombre esperat de targetes de crèdit per viatge per uns ingressos anuals de 120 mil.lions de lires és de 1.1152 tal com s'ha calculat en l'apartat anterior. Aquesta és l'esperança matemàtica d'una variable de Poisson de paràmetre 1.1152. La probabilitat que aquesta variable prengui un valor més gran que 0 és 0.6721, que és la probabilitat que un individu tingui alguna targeta de viatge en aquest grup. Tècnicament, el nombre de persones d'ingressos iguals a 120 milions de lires entre n=10 persones és una variable binomial B(n=10, p=0.6722) i per tant, l'esperança d'aquesta variable dona el nombre esperat de persones en un grup de 10 amb ingressos de 120 Mlires que tenen almenys una targeta és n x p x (1-p) = 10*0.6722*(1-0.6722)= 2.2035 persones.

$$\mu_i = 1.1152 \rightarrow Poisson(\mu_i = 1.1152) \rightarrow \mathrm{P}([Y > 0]) = 1 - \frac{\mu_i^0}{0!}\exp(-\mu_i) = 1 - \frac{1.1152^0}{0!}\exp(-1.1152) = 0.6722$$

7. Valoreu els gràfics de residus facilitats pel model m1. Hi ha *outliers* i/o (possibles) valors influents? Indiqueu les observacions en cadascuna de les condicions anteriors, tot justificant el llindar emprat per l'estadístic implicat en la determinació de la tipologia de l'observació.

La sortida de influencePlot(m1,labels=df$income), treu un buble plot on les etiquetes són els ingressos en millions de lires, són els individus 23 i 30 que tenen una distància de Cook elevada, concretament la observació 30 té uns ingressos de 120, però en canvi el grup de 6 persones només té un total de 6 targetes. El grup 23 té 6 persones amb ingressos de 65 milions de lires i un total de 6 targetes, és un outlier dels residus (rstudent superior a 2). El grup 30 de 120 d'ingressos és atípic, per tenir uns ingressos molt elevats.

RESULTATS PEL PROBLEMA 2

```
> m0<-glm(ntcc~offset(logsize),family=poisson,data=df)
> m1<-glm(ntcc~offset(logsize)+income,family=poisson,data=df)
> m2<-glm(ntcc~offset(logsize)+poly(income,3),family=poisson,data=df)
> summary(m0)
Call:
glm(formula = ntcc ~ offset(logsize), family = poisson, data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept[redacted]     0.1796  -6.521 6.99e-11 ***
---
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 42.078  on 30  degrees of freedom
AIC: 79.218
> summary(m1)

Call:
glm(formula = ntcc ~ offset(logsize) + income, family = poisson,
    data = df)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.386586   0.399655  -5.972 2.35e-09 ***
income       0.020758   0.005165   4.019 5.84e-05 ***
---
 (Dispersion parameter for poisson family taken to be 1)
    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 28.465  on 29  degrees of freedom
AIC: 67.604

> summary(m2)
Call:
glm(formula = ntcc ~ offset(logsize) + poly(income, 3), family = poisson,
    data = df)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.3300     0.2118  -6.281 3.36e-10 ***
poly(income, 3)1   3.8118     1.0165   3.750 0.000177 ***
poly(income, 3)2  -1.3903     0.9728  -1.429 0.152952
poly(income, 3)3   0.5350     1.2060   0.444 0.657321
---
 (Dispersion parameter for poisson family taken to be 1)

    Null deviance: 42.078  on 30  degrees of freedom
Residual deviance: 26.257  on 27  degrees of freedom
AIC: 69.396

> anova(m1,m2,test="Chisq")
Analysis of Deviance Table
Model 1: ntcc ~ offset(logsize) + income
Model 2: ntcc ~ offset(logsize) + poly(income, 3)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        29     28.465
2        27     26.256  2   2.2083   0.3315
>
```

```
> round(predict(m1,type="response"),dig=2)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17
0.15 0.16 0.82 0.50 1.54 0.87 1.43 0.18 1.30 0.19 0.61 0.41 1.05 0.43 0.44 0.23 0.25
  18   19   20   21   22   23   24   25   26   27   28   29   30   31
0.25 2.60 0.27 0.31 1.60 2.13 1.13 1.97 0.47 0.48 0.53 0.65 6.66 1.37
```

```
> residualPlot(m1)
> influencePlot(m1,labels=df$income)
       StudRes        Hat     CookD
65   2.2865417 0.06956068 0.5324392
120 -0.4508049 0.67455275 0.4563497
> llista<-influencePlot(m1,id.n=3);llista
      StudRes        Hat     CookD
5  -0.4923412 0.11274890 0.1168827
7  -1.7363193 0.09865545 0.2945773
23  2.2865417 0.06956068 0.5324392
24  1.4934413 0.03653902 0.2464268
30 -0.4508049 0.67455275 0.4563497
31 -0.3600818 0.17907709 0.1141626
> df[row.names(llista),]
   income size ntcc  logsize
5      30    9    1 2.197225
7      32    8    0 2.079442
23     65    6    6 1.791759
24     68    3    3 1.098612
30    120    6    6 1.791759
31    130    1    1 0.000000
```