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

CURS 2012-2013 Q1 – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: 29/01/2013 a les 15:00h Aula 001-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 5 de Febrer al Web Docent de MLGz

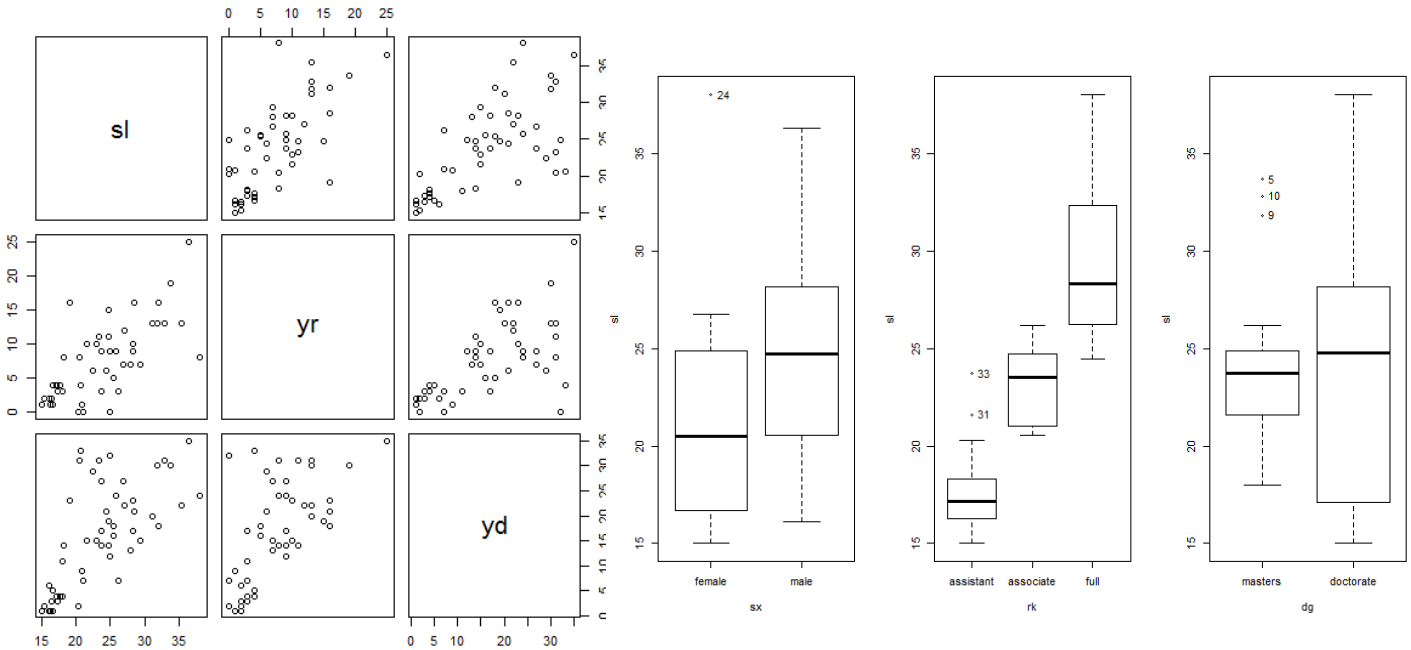
Revisió de l'examen: 5 de Febrer a 10:00 h a C5-217-C Nord o H- P6-67

Problema 1 (3 punts): Resposta Normal

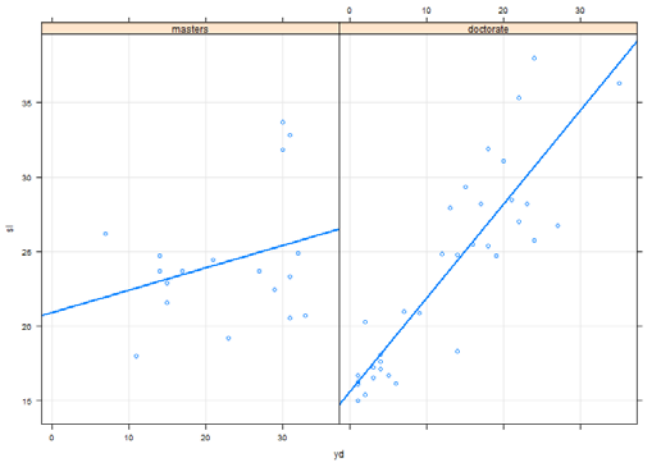
Les dades pertanyen a S. Weisberg (1985). *Applied Linear Regression.* Aquestes són dades de salaris que consisteix en observacions de sis variables per a 52 professors titulars en una universitat petita. Les variables són:

sx = Sexe, codificat 1 per dones i 0 per als homes
rk = càrrec (assistent, associat, full professor)
yr = Nombre d'anys en el càrrec actual
dg = Titulació superior (doctorat, master)
yd = nombre d'anys transcorreguts des que es va obtenir la titulació
sl = salari actual, en dòlars.

```
> summary(salary)
     sx              rk          yr              dg          yd              sl       
 female:14   assistant:18   Min.   : 0.000   masters  :18   Min.   : 1.00   Min.   :15.00  
 male  :38   associate:14   1st Qu.: 3.000   doctorate:34   1st Qu.: 6.75   1st Qu.:18.24  
             full     :20   Median : 7.000                  Median :15.50   Median :23.72  
                            Mean   : 7.481                  Mean   :16.12   Mean   :23.80  
                            3rd Qu.:11.000                  3rd Qu.:23.25   3rd Qu.:27.26  
                            Max.   :25.000                  Max.   :35.00   Max.   :38.05  
```



Plantejem un ANCOVA per veure la relació entre el salari com a resposta i la titulació superior ajustada pel temps que fa que es va assolir:




```
> summary(ma<-lm(sl~yd*dg))

Call:
lm(formula = sl ~ yd * dg)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1067 -1.9284 -0.2985  1.6898  8.2637

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.9403     2.4645   8.497 3.95e-11 ***
yd             0.1497     0.1014   1.477 0.146225
dgdoctorate   -5.3733     2.6775  -2.007 0.050417 .
yd:dgdoctorate  0.4820     0.1217   3.959 0.000248 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.588 on 48 degrees of freedom
Multiple R-squared: 0.654, Adjusted R-squared: 0.6323
F-statistic: 30.24 on 3 and 48 DF,  p-value: 3.987e-11
```

1. Indica quin són els models que s’obtenen pels professors que tenen un títol de màster i pels que ho tenen de doctorat, indicant el que és significatiu i fent una interpretació d’ambdós models en relació a la representació gràfica.

Model pels professors amb títol de master: $sl = 20.94 + 0.15yd + e_i$
Model pels professors amb títol de doctorat: $sl = (20.94 - 5.37) + (0.15 + 0.48)yd + e_i$
En el primer model, la pendent no és significativa i podríem admetre que el sou dels professors que tenen un màster no canvia al llarg dels anys. En canvi, la pendent en el model dels professors amb doctorat sembla clarament significativa, indicant una revalorització dels sous al llarg dels anys per aquest col·lectiu. La representació gràfica posa de manifest aquestes diferencies en l’evolució dels sous. En el primer gràfic, tot i que el pendent és positiu, estadísticament no s’estableix la seva significació. La diferència de sous quan s’acaba de rebre el títol és no significativa per poc (p-valor=0.0504) indicant que podem admetre que el sou en aquest moment està al voltant dels 20.000$ per ambdós col·lectius.

```
# Model 1 amb totes les interaccions d’ordre 2 menys la de les variables numèriques
> summary(m<-lm(sl~(yr+yd+sx+rk+dg)^2-yd:yr))

Call:
lm(formula = sl ~ (yr + yd + sx + rk + dg)^2 - yd:yr)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-4.0509 -1.0516 -0.0188  0.6511  6.3098

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              14.0972     5.7249   2.462   0.0194 *
yr                        0.5021     0.5494   0.914   0.3676
yd                        0.1640     0.4805   0.341   0.7352
sxmale                    6.4137     5.1016   1.257   0.2178
rkassociate               2.9465     9.3345   0.316   0.7543
rkfull                   -5.2763    10.6246  -0.497   0.6229
dgdoctorate               2.0714     5.3996   0.384   0.7038
yr:sxmale                -0.4850     0.4410  -1.100   0.2796
yr:rkassociate            0.2936     0.6543   0.449   0.6566
yr:rkfull                 0.3363     0.5941   0.566   0.5752
yr:dgdoctorate            0.2950     0.3350   0.881   0.3851
yd:sxmale                -0.1869     0.3103  -0.602   0.5512
yd:rkassociate           -0.1463     0.5659  -0.259   0.7977
yd:rkfull                 0.3392     0.4976   0.682   0.5003
yd:dgdoctorate           -0.4284     0.4111  -1.042   0.3052
sxmale:rkassociate        1.1189     6.1682   0.181   0.8572
sxmale:rkfull             2.1095     5.6068   0.376   0.7092
sxmale:dgdoctorate       -5.1576     4.6368  -1.112   0.2743
rkassociate:dgdoctorate   4.6862     4.1207   1.137   0.2639
rkfull:dgdoctorate        9.9812     6.8018   1.467   0.1520
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.436 on 32 degrees of freedom
Multiple R-squared: 0.8937,      Adjusted R-squared: 0.8305
F-statistic: 14.16 on 19 and 32 DF,  p-value: 1.349e-10
```

2. A la vista del model 1 amb totes les variables i interaccions d'ordre 2 amb variables categòriques, podem concloure que no té sentit plantejar cap model amb aquests predictors degut a que en el model global tot surt no significatiu? Raoneu la resposta

Tot i que un model lineal tingui tots els coeficients significatius, no es pot afirmar que aquests predictors s'hagin d'eliminar del model. En ocasions, per motius de colinealitat poden haver predictors amb p-valors superiors a 0.05, però si s'elimina un d'ells, l'altre pot passar a ser significatiu. El test que s'inclou en aquesta sortida que permet contrastar si tots els coeficients del model són no significatius simultàniament, vs algún diferent és la taula de l'ANOVA de la regressió que apareix resumida a la darrera línia, i amb un p-valor de $1.3*10^{-10}$ confirma que existeix algun predictor significatiu. Sempre que s'han d'eliminar coeficients del model, s'ha de fer sequencialment, ja que l'eliminació d'un d'ells pot modificar la significació de la resta.

```
> anova(m2,m3)
Analysis of Variance Table

Model 1: sl ~ yr + rk
Model 2: sl ~ yr * rk
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     48 276.99
2     46 261.78  2    15.216 1.3368 0.2727
```

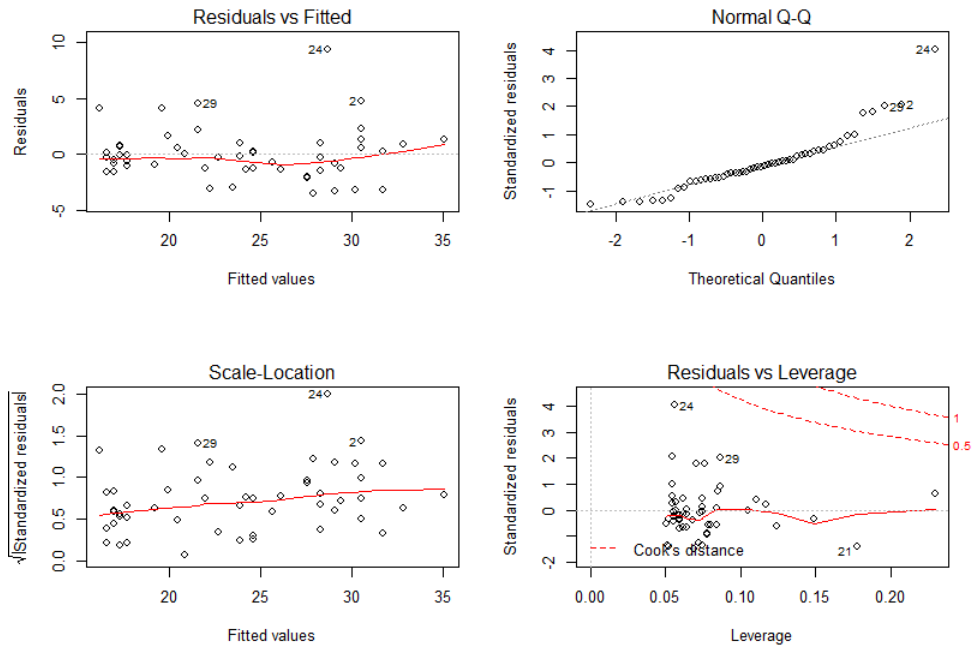
3. Fem la comparació de dos models amb la instrucció anova. Indica el test que s'està aplicant, descriu els models que es comparen, l'estadístic del contrast, la seva distribució de referència sota la hipòtesi nul·la i la decisió que es pren. Amb quin model dels dos et quedaríes?

S'està realitzant un test de la F per hipòtesis compostes per comparar el model que conté només com a efectes principals les variables yr (anys d'antiguitat) i rk (càrrec) de forma aditiva amb el model que conté aquests 2 predictors i la seva interacció. Equival a un test simultani pels coeficients que representen la interacció (H_0:$\beta_{yr*(rk=2)} = \beta_{yr*(rk=3)} = 0$) el model sota la hipòtesi nul·la és el model aditiu.
L'estadístic de contrast és el quocient entre la diferència de suma de quadrats residuals dividit pels seus graus de llibertat i l'estimació de la variància residual.

$$F = \frac{(RSS_{H0} - RSS)/2}{RSS/46} = \frac{15.216/2}{261.78/46} = 1.3368$$

La distribució de referència sota la hipòtesi nul·la és una F-Snedecor amb 2 i 46 graus de llibertat i el p-valor és 0.2727 corresponent a l'àrea de la cua superior. Aquest p-valor indica que no hi ha diferència entre ambdós models i que per tant és adequat prescindir de la interacció en el model, degut a que és una component no significativa. El model seleccionat només inclou els dos efectes principals de forma aditiva (model 1)

A continuació s'ajusta un model i s'obtenen els plots per fer la validació.

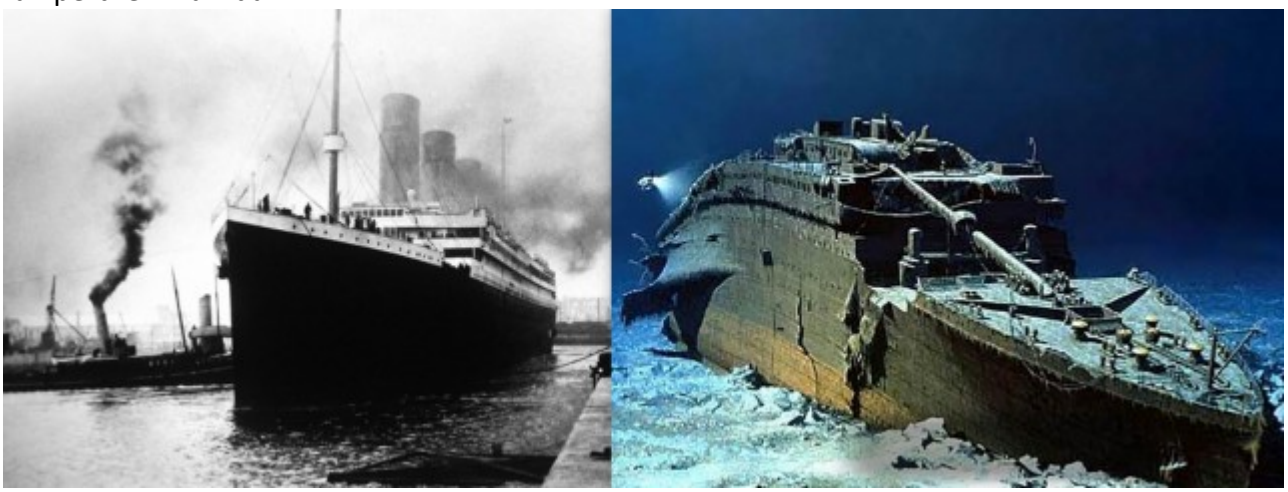


4. Es compleixen les premises del model lineal? Per tal de millorar el model en base a aquests plots, per on continuaries?

En el primer gràfic s'observa que la disposició dels residus al voltant del 0 és aleatòria confirmant la hipòtesi de linealitat, en el plot de normalitat s'observa clarament un valor molt atípic (residu estandarditzat al voltant de 4) i la resta seria coherent amb la hipòtesi de normalitat. El tercer gràfic (i en el primer) s'observa que la variància es pot considerar constant, ja que l'ajust lineal no paràmetric per a la mesura de dispersió és pràcticament una línia horitzontal. En l'últim gràfic podem observar les observacions influents, i si bé existeix alguna observació amb un factor d'apalancament gran (situada a la dreta del gràfic, sembla que la dada més influent és l'observació 24, en base al seu alt residu, i que es situa més a prop de les corbes de nivell que representen la distància de Cook de les observacions. Si bé el model compleix les premises per validar el model, seria convenient veure l'efecte de suprimir la observació 24 sobre els coeficients del model per intentar obtenir un model més precís.

Problema 2 (3 punts): Resposta Binària

El 14 d'abril de 1912 a les 23:40, **fa 100 anys, l'emblemàtic vaixell de luxe Titanic** xocava contra un iceberg **al sud de les costes de l'illa americana Terranova. Dues hores i mitja després,** a les 2:20 del 15 d'abril, el vaixell s'enfonsava. Era el seu primer viatge i portava 2.227 persones a bord. Després de la desgràcia, només 706 se'n van salvar gràcies als bots de salvament. La resta, 1.517, van perdre-hi la vida.

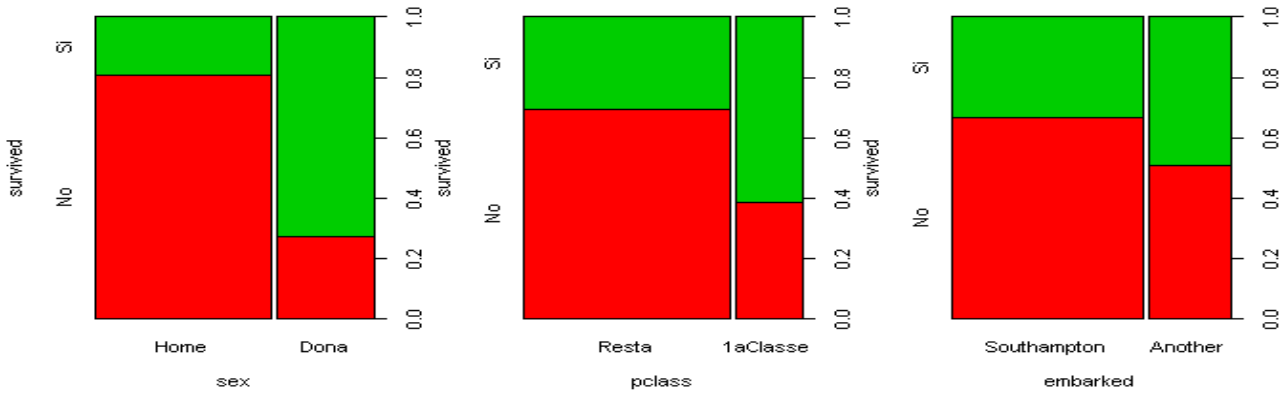


Disposem de 1307 observacions individuals de passatger d'aquell darrer viatge amb 4 característiques:

- pclass: La clase del passatger al vaixell, codificada com 1 -1ª clase i 0 – la resta (2ª o 3ª classes)
- survived: Indicador de si va sobreviure o no (1-Si , 0-No). **Resposta**.
- sex: Codificat com 1-Dona 0-Home.
- embarked: El port on va embarcar el passatger pel viatge final (codificat com to 1=Queenstown or Cherbourg and 0=Southampton).

Dades originals: Dawson, Robert J. MacG. (1995), The Unusual Episode Data Revisited. Journal of Statistics Education, 3. http://www.amstat.org/publications/jse/v3n3/datasets.dawson.html

```
> summary(titanic)
      pclass    survived   sex           embarked
 Resta   :986   No:809   Home:843   Southampton:914
 1aClasse:321   Si:498   Dona:464   Another    :393
>
```




```
> table(pclass,survived)
          survived
pclass      No  Si
```

```
  Resta    686 300
  1aClasse 123 198
> table(survived)
survived
 No  Si
809 498
> prop.table(table(pclass,survived),1)
          survived
pclass            No        Si
  Resta    0.6957404 0.3042596
  1aClasse 0.3831776 0.6168224
> prop.table(table(survived))
survived
       No        Si
0.6189748 0.3810252
>
> lm4<-glm(survived~pclass*sex*embarked,family=binomial,data=titanic)
> Anova(lm4)
Analysis of Deviance Table (Type II tests)

Response: survived
                  LR Chisq Df Pr(>Chisq)
pclass               78.29  1  < 2.2e-16 ***
sex                 350.34  1  < 2.2e-16 ***
embarked              5.19  1     0.0227 *
pclass:sex           15.68  1  7.496e-05 ***
pclass:embarked       0.07  1     0.7873
sex:embarked          0.06  1     0.8124
pclass:sex:embarked   0.00  1     0.9936
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> lmf<-step(lm4)
> lmf<-step(lm4)
Start:  AIC=1273.72
survived ~ pclass * sex * embarked

                      Df Deviance    AIC
- pclass:sex:embarked  1   1257.7 1271.7
<none>                     1257.7 1273.7

Step:  AIC=1271.72
survived ~ pclass + sex + embarked + pclass:sex + pclass:embarked +
    sex:embarked

                  Df Deviance    AIC
- sex:embarked     1   1257.8 1269.8
- pclass:embarked  1   1257.8 1269.8
<none>                 1257.7 1271.7
- pclass:sex       1   1273.4 1285.4

Step:  AIC=1269.78
survived ~ pclass + sex + embarked + pclass:sex + pclass:embarked

                  Df Deviance    AIC
- pclass:embarked  1   1257.8 1267.8
<none>                 1257.8 1269.8
- pclass:sex       1   1273.9 1283.9

Step:  AIC=1267.82
survived ~ pclass + sex + embarked + pclass:sex
```

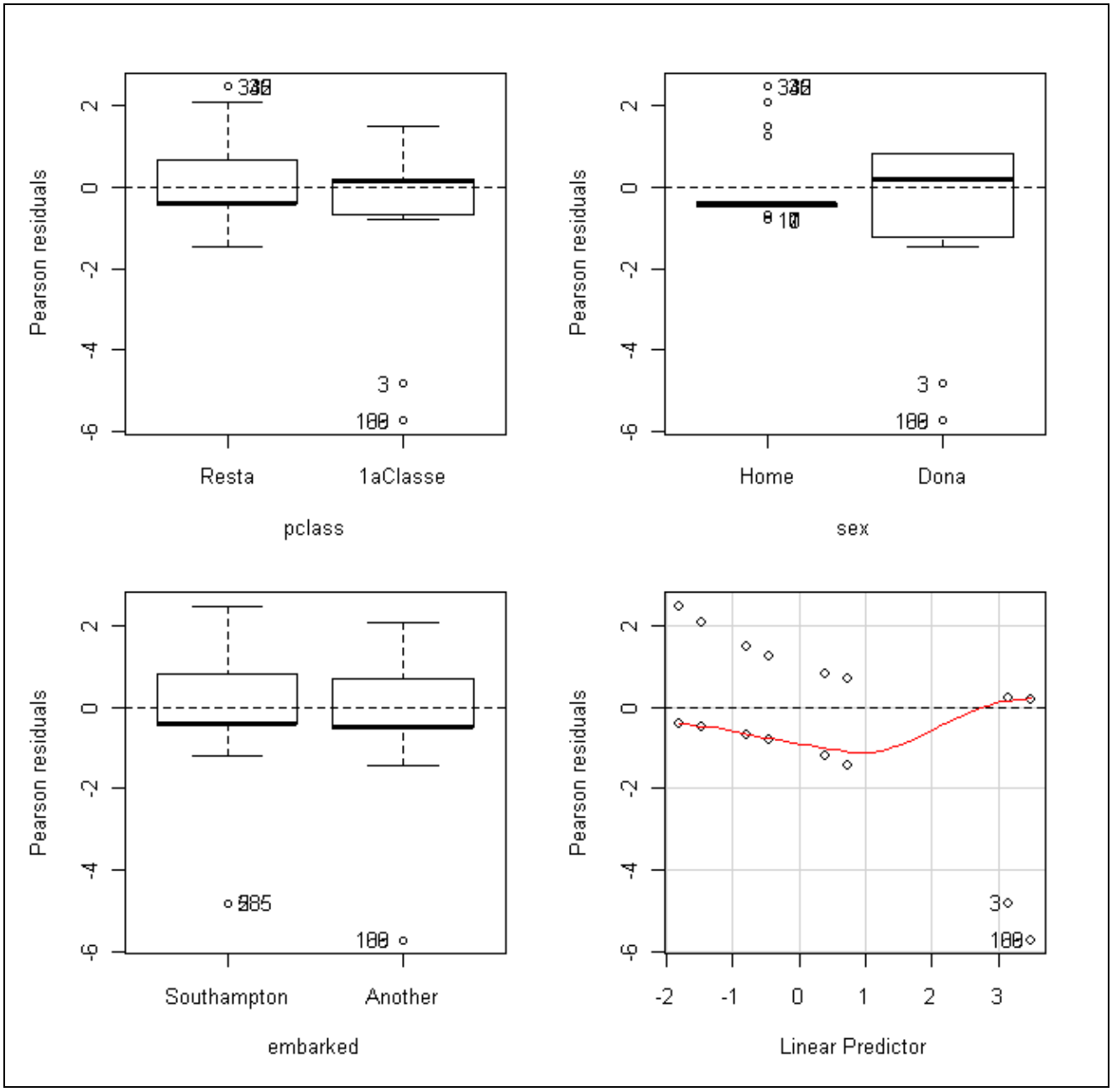
```
             Df Deviance    AIC
<none>            1257.8 1267.8
- embarked    1   1263.0 1271.0
- pclass:sex  1   1274.0 1282.0
>
> summary(lmf)

Call:
glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse         1.0131     0.1934   5.239 1.62e-07 ***
sexDona                2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother        0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona 1.7430     0.5073   3.436 0.000591 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 1267.8

> Model triat pels professors:
> model$dev
[1] 1257.817
> llista<-influencePlot(model,id.n=5,col=3,pch=19)
> titanic[row.names(llista),]
      pclass survived  sex    embarked
1   1aClasse       Si Dona Southampton
3   1aClasse       No Dona Southampton
5   1aClasse       No Dona Southampton
7   1aClasse       Si Dona Southampton
9   1aClasse       Si Dona Southampton
14  1aClasse       Si Dona Southampton
106 1aClasse       No Dona     Another
169 1aClasse       No Dona     Another
285 1aClasse       No Dona Southampton
> residualPlots(model,id.n=3)
```



1. Calculeu manualment el model nul amb l'enllaç ***logit*** per la resposta *survived*.

(LM0) $log(\pi_i) = \eta$

Per tant, l'estimador de $\eta = log\left(\frac{498}{809}\right) = -0.4851988$.

```
> prop.table(table(survived))
survived
       No        Si
0.6189748 0.3810252
> log(prop.table(table(survived))[2]/prop.table(table(survived))[1])
        Si
-0.4851988
>
> lm0<-glm(survived~1,family=binomial,data=titanic)
> summary(lm0)

Call: glm(formula = survived ~ 1, family = binomial, data = titanic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept) -0.48520    0.05696  -8.519   <2e-16 ***
---
 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1737.2  on 1306  degrees of freedom
Residual deviance: 1737.2  on 1306  degrees of freedom
AIC: 1739.2
>
```

2. Calculeu quin el nombre esperat de supervivents segons el model nul per cadascuna de les categories del factor *pclass*.

Per tant, cal calcular-ho manualment, quines són les prediccions a les cel.les de la matriu 2x2 de table (pclass,survived) sota el model nul, que serà com aplicar-li la probabilitat marginal de sobreviure/morir (0.3810/0.6190)a cadascuna de les files.

Observacions pclass	Survived=NO	Survived=SI	**Total**
Resta (Ref)	686	300	986
1aClasse	123	198	321
Total	809	498	1307

Prediccions (LM0) pclass	Survived=NO	Survived=SI	**Total**
Resta (Ref)	610.334	375.666	986
1aClasse	198.699	122.301	321
Total	809	498	1307

3. Quina és la deviança residual del model nul.

Ara ja es pot calcular la deviança residual del model nul

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

$$2\left(300\log\frac{300}{375.666} + 686\log\frac{686}{610.334} + 198\log\frac{198}{122.301} + 123\log\frac{123}{198.699}\right) = 98.2$$

4. Calculeu manualment el model amb el factor *pclass* per la resposta *survived* emprant l'enllaç ***logit***.

(LM1) $\boldsymbol{log(\pi_i) = \eta + \alpha_i}$ **on** $\boldsymbol{\alpha_1 = 0\ i = 1 \equiv Resta}$

Per tant, l'estimador de $\hat{\eta} = log\left(\frac{300}{686}\right) = -0.8270952$ i

l'estimador per $\hat{\alpha}_{2\equiv 1aClasse} = log\left(\frac{198}{123}\right) - log\left(\frac{300}{686}\right) = 1.303178$

```
> # Apartat 2
> table(pclass,survived)
          survived
pclass      No  Si
  Resta    686 300
  1aClasse 123 198
> prop.table(table(pclass,survived),1)
          survived
pclass            No        Si
  Resta    0.6957404 0.3042596
```

```
  1aClasse 0.3831776 0.6168224
> # cnt
> eta=log(table(pclass,survived)[1,2]/table(pclass,survived)[1,1])
>
alpha1aClasse=log(table(pclass,survived)[2,2]/table(pclass,survived)[2
,1])-eta
> eta;alpha1aClasse
[1] -0.8270952
[1] 1.303178
> lm1<-glm(survived~pclass,family=binomial,data=titanic)
> summary(lm1)

Call: glm(formula = survived ~ pclass, family = binomial, data =
titanic)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.82710    0.06922 -11.949   <2e-16 ***
pclass1aClasse  1.30318    0.13406   9.721   <2e-16 ***
---
 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1737.2  on 1306  degrees of freedom
Residual deviance: 1639.0  on 1305  degrees of freedom
AIC: 1643
>
```

5. Calculeu manualment el model nul amb l'enllaç ***probit*** per la resposta *survived*.

(PM0) $probit(\pi_i) = \eta$

Per tant, l'estimador de $\eta = probit(0.3810252) = qnorm(0.3810252) = -0.3027894$

```
> prop.table(table(survived))
survived
       No        Si
0.6189748 0.3810252
> qnorm(0.3810252)
[1] -0.3027894
> pm0<-glm(survived~1,family=binomial(link=probit),data=titanic)
> summary(pm0)

Call:
glm(formula = survived ~ 1, family = binomial(link = probit),
    data = titanic)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.30279    0.03525  -8.589   <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: 1737.2  on 1306  degrees of freedom
Residual deviance: 1737.2  on 1306  degrees of freedom
AIC: 1739.2
>
```

6. Calculeu manualment el model amb el factor *pclass* per la resposta *survived* emprant l'enllaç ***probit***.

```
> # Apartat 6
> table(pclass,survived)
          survived
pclass      No  Si
  Resta    686 300
  1aClasse 123 198
> prop.table(table(pclass,survived),1)
          survived
pclass            No        Si
  Resta    0.6957404 0.3042596=π1
  1aClasse 0.3831776 0.6168224=π2
```

(PM1) $\boldsymbol{probit(\pi_i) = qnorm(\pi_i) = \eta + \alpha_i}$ **on** $\boldsymbol{\alpha_1 = 0\ i = 1 \equiv Resta}$

Per tant, l'estimador de $\hat{\eta} = \boldsymbol{qnorm}(\pi_1) = \boldsymbol{qnorm}(0.3042596) = -0.5121882$i
l'estimador per $\hat{\alpha}_{2\equiv 1aClasse} = \boldsymbol{qnorm}(\pi_2) - \boldsymbol{qnorm}(\pi_1) = \boldsymbol{qnorm}(0.6168224) - \boldsymbol{qnorm}(0.3042596) = 0.8093341$

```
> qnorm(prop.table(table(pclass,survived),1)[1,2])
[1] -0.5121882
> qnorm(prop.table(table(pclass,survived),1)[2,2])-
qnorm(prop.table(table(pclass,survived),1)[1,2])
[1] 0.8093341

> pm1<-glm(survived~pclass,family=binomial(link=probit),data=titanic)
> summary(pm1)

Call: glm(formula = survived ~ pclass, family = binomial(link =
probit),     data = titanic)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.51219    0.04188  -12.23   <2e-16 ***
pclass1aClasse  0.80933    0.08250    9.81   <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: 1737.2  on 1306  degrees of freedom
Residual deviance: 1639.0  on 1305  degrees of freedom
AIC: 1643
>
```

7. Valoreu si l'efecte brut del factor *pclass* és significatiu per explicar la probabilitat de sobreviure a la tragèdia del Titanic.

Resulta necessari haver estimat el model nul per coneixer la seva deviança residual i coneixer que el model saturat quan només es considera el factor pclass té deviança residual 0: $\Delta D = 98.2$
Per tant, anova(LM0,LM1,test="Chis") calcularia el test de la deviança per l'efecte brut de pclass i el pvalor de la H0 (H0: Efecte brut pclass no significatiu), és
pvalor= $P\left(\chi_1^2 > 98.2\right)$=0<<0.05, per tant hi ha evidència per rebutjar la H0 i afirmar que l'efecte brut de pclass és significatiu estadísticament al nivell de confiança del 95% habitual.
Amb les dades en R i la disponibilitat d'executar-lo:

```
> anova(lm0,lm1,test="Chis")
Analysis of Deviance Table

Model 1: survived ~ 1
Model 2: survived ~ pclass
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1306     1737.2
2      1305     1639.0  1   98.199 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

8. Amb les variables explicatives disponibles es construeix el model més complex, valoreu quins efectes principals i interaccions són estadísticament significatives amb els resultats disponibles.

Resultaria necessari disposar de més contrastos, però amb la informació disponible es pot afirmar que la interacció d'ordre 3 (com no afecta gens ni mica, podem acceptar que els resultats són efectes nets de les interaccions dobles) i 2 interaccions d'ordre 2 no són significatives (classe amb lloc d'embarcament i sexe amb lloc d'embarcament). Per tant, efectes principals són significatius i la interacció classe amb gènere també.

```
> lm4<-glm(survived~pclass*sex*embarked,family=binomial,data=titanic)
> Anova(lm4)
Analysis of Deviance Table (Type II tests)

Response: survived
                  LR Chisq Df Pr(>Chisq)
pclass               78.29  1  < 2.2e-16 ***
sex                 350.34  1  < 2.2e-16 ***
embarked              5.19  1     0.0227 *
pclass:sex           15.68  1  7.496e-05 ***
pclass:embarked       0.07  1     0.7873
sex:embarked          0.06  1     0.8124
pclass:sex:embarked   0.00  1     0.9936
```

9. S'empra la selecció del millor model mitjançant el mètode step() segons el criteri d'AIC, penseu que la tria coincideix amb la resultant de suprimir els paràmetres no significatius per la via inferencial (contrast de deviances o test de Wald).

A la vista de la discussió del punt anterior (perspectiva inferencial) i dels resultats de la tria del model segons el mètode heurístic del AIC més baix, efectivament els models coincideixen. No surt al llistat, però el detall del millor model segons aquests criteris indica que les dones i la primera classe van tenir una supervivència superior als altres grups (per criteri BIC, el port d'embarcament no és significatiu). El model lmf és:

```
> summary(lmf)

Call: glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse         1.0131     0.1934   5.239 1.62e-07 ***
sexDona                2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother        0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona 1.7430     0.5073   3.436 0.000591 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1737.2  on 1306  degrees of freedom
Residual deviance: 1257.8  on 1302  degrees of freedom
AIC: 1267.8
>
```

10. Intepreteu en totes les escales possibles l’efecte del port d’embarcament en el millor model resultant segons els vostres arguments als punts 8 i 9.

Els arguments no hi compten massa doncs cal fer la interpretació en l’únic model disponible i que a més és el millor segons criteris inferencials i d’AIC.

Embarcar a un port diferent a Southampton (Another) incrementa els logodds de la probabilitat de sobreviure en 0.3485 unitats respecte embarcar a Southampton (ref) dins de la mateixa categoria de la resta de variables.

Embarcar a un port diferent a Southampton (Another) incrementa els odds de sobreviure en (exp(0.3485)-1)100%=41.69% unitats respecte els odds de sobreviure si s’embarca a Southampton (ref) *all else being equal*.

En termes de probabilitat, ***aproximadament*** l’efecte d’embarcar a un lloc diferent de Southampton (Another) suposa un increment de la probabilitat de 0.6189748*0.3810252*0.3485 = 0.08219198 , *ceteris paribus*. O en termes de màxima indeterminació 0.25*0.3485= 0.087125 (evidentment, *ceteris paribus*).

```
> summary(lmf)

Call:
glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse         1.0131     0.1934   5.239 1.62e-07 ***
sexDona                2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother        0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona 1.7430     0.5073   3.436 0.000591 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 1267.8
```

11. Intepreteu en totes les escales possibles l’efecte del pclass en el millor model resultant segons els vostres arguments als punts 8 i 9.

Els arguments no hi compten massa doncs cal fer la interpretació en l’únic model disponible i que a més és el millor segons criteris inferencials i d’AIC.

Estar a 1a Classe incrementa els logodds de la probabilitat de sobreviure en **1.0131** unitats en els homes i en **1.0131+1.7430=2.7561** unitats en les dones respecte No Estar a 1aClasse (ref) dins de la mateixa categoria del factor embarked (port d’embarcament).

Estar a 1a Classe incrementa els odds de sobreviure un (exp(**1.0131**)-1)100%=175% en els homes i un (exp(**2.7561**)-1)100%=1474% en les dones respecte No Estar a 1aClasse (ref) dins de la mateixa categoria del factor embarked (port d'embarcament). Escandalós, si.

En termes de probabilitat, ***aproximadament*** l'efecte de pclass en 1a Classe suposa un increment de la probabilitat en 0.6189748*0.3810252***1.0131**= 0.2398 unitats en els homes i 0.6189748*0.3810252***2.7561**=0.65 unitats en les dones, dins de la mateixa categoria del factor embarked (port d'embarcament).

```
> summary(lmf)

Call:
glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse         1.0131     0.1934   5.239 1.62e-07 ***
sexDona                2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother        0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona 1.7430     0.5073   3.436 0.000591 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

AIC: 1267.8
```

12. Quina és la probabilitat predita de sobreviure en el millor model resultant segons els vostres arguments als punts 8 i 9 per un home que va embarcar a Southampton en primera classe?

```
> summary(lmf)

Call:
glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse         1.0131     0.1934   5.239 1.62e-07 ***
sexDona                2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother        0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona 1.7430     0.5073   3.436 0.000591 ***


> eta= 0 -1.8156207        +      1.0130522   = -0.8025685
> exp(-0.8025685)/(1+exp(-0.8025685))
[1] 0.3094764
```

La probabilitat de sobreviure un home en 1a Classe embarcat a Southampton és de casi el 31% segons el model lmf.

13. Valoreu els residus del model triat (dit model).

Els residus de Pearson mostren outliers, valors molt negatius que amb la informació subministrada no se sap a qui poden pertanyer, per altre banda amb els residus d'Student dels

glm() es tindria una idea més aproximada per valorar si són atípics o no realment. És a dir, atípics dels residus n'hi ha.

14. A quí penseu que pertanyen les observacions influents o amb residus elevats?

Amb la informació disponible només es pot endevinar que són dones en 1a Classe (no queda clar on embarcaren) i que no van sobreviure.

El detall concret és:
Les observacions 3, 5 i 285 tenen un residus elevats i un anclatge elevat, per tant són sospitosos de ser *influent data segons criteris estàndar (són dones en 1a Classe que embarquen a Southampton i que no sobrevieun, res de particular)*, per altra banda les observacions 106 i 169 tenen un residu elevat: són dones de 1a Classe que NO van pujar a Southampton i que van morir, ara bé a mi no em sembla res remarcable, potser caldria disposar de més variables, el model queda molt limitat.

```
> llista
       StudRes         Hat       CookD
1    0.2909592 0.008383435 0.008567535
3   -2.5635120 0.008383435 0.199025441
5   -2.5635120 0.008383435 0.199025441
7    0.2909592 0.008383435 0.008567535
9    0.2909592 0.008383435 0.008567535
14   0.2909592 0.008383435 0.008567535
22   0.2909592 0.008383435 0.008567535
106 -2.6928739 0.006151455 0.202485295
169 -2.6928739 0.006151455 0.202485295
285 -2.5635120 0.008383435 0.199025441
> titanic[row.names(llista),]
      pclass survived  sex    embarked
1   1aClasse       Si Dona Southampton
3   1aClasse       No Dona Southampton
5   1aClasse       No Dona Southampton
7   1aClasse       Si Dona Southampton
9   1aClasse       Si Dona Southampton
14  1aClasse       Si Dona Southampton
22  1aClasse       Si Dona Southampton
106 1aClasse       No Dona     Another
169 1aClasse       No Dona     Another
285 1aClasse       No Dona Southampton
>
```

15. Valoreu la qualitat del model triat (dit model) en funció de la informació subministrada sobre la deviança residual.

El millor model disponible té una deviança residual de 1257.8 unitats amb 1302 gll, per tant, són del mateix de magnitud i podem valorar satisfactòriament l'ajust del model a les dades. El test de Bondat de l'Ajust, no convenient donada la desagregació de les dades, mostra un pvalor >0.80 i per tant, no hi ha evidència per rebutjar la H0, s'accepta que el model ajusta bé les dades.

```
> 1-pchisq(lmf$dev,lmf$df.residual)
[1] 0.8058561
>

> summary(lmf)

Call:  glm(formula = survived ~ pclass + sex + embarked + pclass:sex,
    family = binomial, data = titanic)
```

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.8156     0.1161 -15.640  < 2e-16 ***
pclass1aClasse           1.0131     0.1934   5.239 1.62e-07 ***
sexDona                  2.2051     0.1584  13.919  < 2e-16 ***
embarkedAnother          0.3485     0.1524   2.287 0.022182 *
pclass1aClasse:sexDona   1.7430     0.5073   3.436 0.000591 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1737.2  on 1306  degrees of freedom
Residual deviance: 1257.8  on 1302  degrees of freedom
AIC: 1267.8
>
```

Problema 3 (3 Punts)

Les dades de Woolridge (2002) disponibles a la pàgina web d'Stata mostren 807 individus amb descripció del nb total de cigarretes que fumen en un dia (la resposta) i descripció de les variables explicatives:

1. cigpric (preu en cèntims de la caixa de cigarretes)
2. lcigpric (log del preu en cèntims de la caixa de cigarretes)
3. income, lincome (ingressos anuals en $ i el seu log)
4. restaurn (dummy amb 1 si hi ha restriccions per fumar en els restaurants habituals de l'individu)
5. white (dummy amb 1 si raça blanca
6. educ (anys d'estudis)
7. age (edat)
8. fumador (factor amb 1 si cigs>0)

La descriptiva del joc de dades es mostra a continuació.

```
> summary(tabac)
      educ          cigpric      white        age           income
 Min.   : 6.00   Min.   :44.00   0: 98   Min.   :17.00   Min.   :  500
 1st Qu.:10.00   1st Qu.:58.14   1:709   1st Qu.:28.00   1st Qu.:12500
 Median :12.00   Median :61.05           Median :38.00   Median :20000
 Mean   :12.47   Mean   :60.30           Mean   :41.24   Mean   :19305
 3rd Qu.:13.50   3rd Qu.:63.18           3rd Qu.:54.00   3rd Qu.:30000
 Max.   :18.00   Max.   :70.13           Max.   :88.00   Max.   :30000
      cigs        restaurn    lincome          lcigpric     fumador      uns
 Min.   : 0.000   0:608    Min.   : 6.215   Min.   :3.784   0:497   Min.   :1
 1st Qu.: 0.000   1:199    1st Qu.: 9.433   1st Qu.:4.063   1:310   1st Qu.:1
 Median : 0.000            Median : 9.903   Median :4.112           Median :1
 Mean   : 8.686            Mean   : 9.687   Mean   :4.096           Mean   :1
 3rd Qu.:20.000            3rd Qu.:10.309   3rd Qu.:4.146           3rd Qu.:1
 Max.   :80.000            Max.   :10.309   Max.   :4.250           Max.   :1
>
```

Inicialment es tanteja un model loglineal amb component aleatòria modelada segons una llei de Poisson. Es mostren alguns dels resultats facilitats per R. Responegueu a les preguntes formulades.

```
> summary(pm1)

Call:
glm(formula = cigs ~ lcigpric + lincome + restaurn + white +
    educ + age + I((age - 38)^2), family = poisson(link = log),
    data = tabac)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.376e+00  6.160e-01   3.857 0.000115 ***
```

```
lcigpric        -1.060e-01  1.434e-01  -0.739 0.459926
lincome          1.037e-01  2.028e-02   5.115 3.14e-07 ***
restaurn1       -3.636e-01  3.122e-02 -11.646  < 2e-16 ***
white1          -5.520e-02  3.742e-02  -1.475 0.140162
educ            -5.942e-02  4.256e-03 -13.961  < 2e-16 ***
age              1.007e-02  1.035e-03   9.738  < 2e-16 ***
I((age - 38)^2) -1.371e-03  5.695e-05 -24.070  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 15821  on 806  degrees of freedom
Residual deviance: 14752  on 799  degrees of freedom
AIC: 16239

Number of Fisher Scoring iterations: 6

> pm1a <-glm(cigs~lcigpric + restaurn+white +educ+ age+ I((age-38)^2),
family=poisson(link=log), data=tabac )
> summary(pm1a)

Call:
glm(formula = cigs ~ lcigpric + restaurn + white + educ + age +
    I((age - 38)^2), family = poisson(link = log), data = tabac)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.267e+00  5.894e-01   5.544 2.96e-08 ***
lcigpric        -1.023e-01  1.432e-01  -0.715    0.475
restaurn1       -3.517e-01  3.114e-02 -11.293  < 2e-16 ***
white1          -6.281e-02  3.740e-02  -1.680    0.093 .
educ            -5.317e-02  4.067e-03 -13.073  < 2e-16 ***
age              1.126e-02  1.017e-03  11.074  < 2e-16 ***
I((age - 38)^2) -1.447e-03  5.534e-05 -26.148  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 15821  on 806  degrees of freedom
Residual deviance: 14780  on 800  degrees of freedom
AIC: 16264

> (sum(residuals(pm1,type="pearson")^2))
[1] 16232.71
```

1. Interpreteu l’efecte dels ingressos.

El coeficient de lincome 1.037e-01 indica que un increment de un 1% en els ingressos (no en el seu logaritme) està associat amb un increment de 0.1% en el nombre esperat de cigarretes fumades per dia, deixant constants la resta de variables explicatives.

2. Contrasteu la significació estadística de l’efecte dels ingressos.

```
> anova(pm1a,pm1,test="Chisq")
Analysis of Deviance Table

Model 1: cigs ~ lcigpric + restaurn + white + educ + age + I((age -
38)^2)
Model 2: cigs ~ lcigpric + lincome + restaurn + white + educ + age +
I((age -
    38)^2)
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       800      14780
2       799      14752   1       27 1.999e-07
```

Els ingressos tenen un efecte estadísticament significatiu.

3. Interpreteu l’efecte de les restriccions de fumar als restaurants habituals.

```
> anova(pm1b,pm1,test="Chisq")
Analysis of Deviance Table

Model 1: cigs ~ lcigpric + lincome + white + educ + age + I((age - 38)^2)
Model 2: cigs ~ lcigpric + lincome + restaurn + white + educ + age +
I((age -
    38)^2)
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       800      14898
2       799      14752   1      145 2.123e-33
>  (1-exp(coef(pm1)[4]))*100
restaurn1
 30.48349
```

Resulta estadísticament significativa. El coeficient de la dummy implica que les persones que freqüenten restaurants amb restriccions per fumar, fumen en promig un 30.5% menys cigarretes diàries que els qui freqüenten restaurants sense restriccions, deixant constant la resta de les explicatives.

4. Trobeu evidència de sobredispersió en el joc de dades.

```
> (sum(residuals(pm1,type="pearson")^2))
[1] 16232.71
Amb 799 g.ll, la distribució de referència és una shi quadrat.
```

El model no ajusta bé les dades tal com indiquen els estadístics de la deviança residual i de Pearson. El promig de l'estadístic de Pearson per grau de llibertat està en 20.32 unitats de deviança suggerint que la variança està al voltant de 20 vegades la mitjana i per tant, es mostra una forta evidència de sobredispersió.

Suposeu que la variança és proporcional a la mitjana, en comptes d'igual a ella en el model :
`cigs ~ lcigpric + lincome + restaurn + white + educ + age + I((age - 38)^2)`

```
> summary(qm1)

Call:
glm(formula = cigs ~ lcigpric + lincome + restaurn + white +
    educ + age + I((age - 38)^2), family = quasi(link = log,
    variance = "mu"), data = tabac)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.3759125  2.7766779   0.856  0.39244  ...
---

(Dispersion parameter for quasi family taken to be 20.31783)

    Null deviance: 15821  on 806  degrees of freedom
Residual deviance: 14752  on 799  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
> summary(qm1a)

Call:
glm(formula = cigs ~ lcigpric + restaurn + white + educ + age +
    I((age - 38)^2), family = quasi(link = log, variance = "mu"),
    data = tabac)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)      3.2674136  ...
---

(Dispersion parameter for quasi family taken to be 20.35514)

    Null deviance: 15821  on 806  degrees of freedom
Residual deviance: 14780  on 800  degrees of freedom

> mu <- predict(qm1,type="response")
> zero<- exp(-mu)
> summary(zero)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
2.004e-09 1.959e-05 1.755e-04 8.705e-03 1.575e-03 5.607e-01
```

5. Determineu si la covariant ingressos és estadísticament significativa segons l'estadístic de Wald.

Les estimacions puntuals són idèntiques, no així els estàndard errors. L'estàndard error del model Poisson es multiplica per l'arrel quadrada del paràmetre de sobredispersió (20.32) i a partir d'aquí el constrast de coeficient lincome igual a 0, es calcularia per :

```
> (0.1037275  -0)/( 4.51 * 2.011e-02   )
[1] 1.143682
> (1-pnorm((0.1037275  -0)/( 4.51 * 2.011e-02   )))*2
[1] 0.2527555
>
```

El p valor del contrast és del 25%, per tant, considerant la sobredispersió inqüestionable que mostren les dades, l'efecte dels ingressos en el nombre mig de cigarretes diàries fumades no és estadísticament significativa. Surt directament demanant un summary sobre el model original amb ajust del paràmetre de sobredispersió tal i com es va veure a classe de laboratori:

```
> summary(pm1,dispersion=escala)
```

6. Determineu si la covariant ingressos és estadísticament significativa per contrast de deviança.

```
> anova(qm1a,qm1,test="F")
Analysis of Deviance Table

Model 1: cigs ~ lcigpric + restaurn + white + educ + age + I((age -
38)^2)
Model 2: cigs ~ lcigpric + lincome + restaurn + white + educ + age +
I((age -
    38)^2)
  Resid. Df Resid. Dev  Df Deviance      F Pr(>F)
1       800      14780
2       799      14752   1       27 1.3306 0.2490
>
```

7. Quin és el percentatge predit d'individus que fumen 0 cigarretes al dia segons aquest model? Valoreu la xifra en contrast amb les dades mostrals.

A la vista dels resultats, on es calcula la probabilitat de 0 cigarretes diàries segons el paràmetre de cigarretes diàries fumades pel model segons la llei de Poisson (zero). El valor mig és 8.705e-03, és a dir un 0.87% de la població. Aquesta predicció contrasta amb les dades mostrals on els NO fumadors representen el 61% de la mostra.

Suposeu que es vol emprar una distribució binomial negativa per la component aleatòria i s'usa el procediment glm.nb() disponible en el paquet MASS en R.

```
> summary(nbm1)

Call:
```

```
glm.nb(formula = I(cigs + 0.05) ~ lcigpric + lincome + restaurn +
    white + educ + age + I((age - 38)^2), data = tabac, init.theta =
0.208467849762761,
    link = log)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.2301969  4.0301215   0.802 0.422834
lcigpric       -0.1906393  0.9583069  -0.199 0.842315
lincome         0.1008239  0.1213381   0.831 0.406011
restaurn1      -0.4631016  0.1859570  -2.490 0.012761 *
white1         -0.1719467  0.2421239  -0.710 0.477605
educ           -0.0922206  0.0277881  -3.319 0.000904 ***
age             0.0128262  0.0062828   2.041 0.041203 *
I((age - 38)^2) -0.0015920  0.0002965  -5.369 7.92e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.2085) family taken to be 1)

    Null deviance: 912.70  on 806  degrees of freedom
Residual deviance: 878.34  on 799  degrees of freedom
AIC: 4123.6
              Theta:  0.2085
          Std. Err.:  0.0113
 2 x log-likelihood:  -4105.5790
> summary(nbm1a)

Call:
glm.nb(formula = I(cigs + 0.05) ~ lcigpric + restaurn + white +
    educ + age + I((age - 38)^2), data = tabac, init.theta =
0.208285644196159,
    link = log)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.7724603  3.9382575   0.958  0.33811
lcigpric       -0.0952176  0.9574565  -0.099  0.92078
restaurn1      -0.4593262  0.1856390  -2.474  0.01335 *
white1         -0.1846215  0.2420074  -0.763  0.44554
educ           -0.0886741  0.0269038  -3.296  0.00098 ***
age             0.0134684  0.0061676   2.184  0.02898 *
I((age - 38)^2) -0.0016422  0.0002857  -5.748 9.02e-09 ***
---
(Dispersion parameter for Negative Binomial(0.2083) family taken to be 1)

    Null deviance: 912.14  on 806  degrees of freedom
Residual deviance: 878.45  on 800  degrees of freedom
AIC: 4122.2
              Theta:  0.2083
          Std. Err.:  0.0113
 2 x log-likelihood:  -4106.2200
>
```

8. Determineu si el model s’ajusta bé a les dades.

La deviança residual és molt més reduïda que amb les tentatives Poisson anteriors i concretament amb l’ajut d’un paquet estadístic és encara força gran i no passaria el test *de goodness of fit. Hi ha variables no significatives segons l’estadístic de Wald (en el summary), tant en nbm1(lincome i white) o nbm1a (white). Per tant, segur no són models finals.*

```
> deviance(nbm1)
[1] 878.3405
> df.residual(nbm1 )
[1] 799
> 1-pchisq( deviance(nbm1),df.residual(nbm1 ))
[1] 0.02624069
>
```

9. Interpreteu el coeficient de la *dummy* corresponent al factor de restriccions per fumar als restaurants.

Usant els resultats de nbm1 i la interpretació, que és idèntica a la dels models poissonians i càlculs elementals, l'efecte de la prohibició de fumar en els restaurants habituals és d'una reducció d'un 37% en el nombre esperat de cigarretes diàries fumades.

```
> coef(nbm1)[4]
 restaurn1
-0.4631016
> (1-exp(coef(nbm1)[4]))*100
restaurn1
 37.06713
>
```

10. Determineu si l'efecte dels ingressos anuals és estadísticament significatiu.

```
> anova(nbm1a,nbm1)
Likelihood ratio tests of Negative Binomial Models

Response: I(cigs + 0.05)
Model     theta Resid. df
1           lcigpric + restaurn + white + educ + age + I((age - 38)^2)
0.2082856       800
2 lcigpric + lincome + restaurn + white + educ + age + I((age - 38)^2)
0.2084678       799
2 x log-lik.   Test    df  LR stat.   Pr(Chi)
1       -4106.220
2       -4105.579 1 vs 2     1 0.6412714 0.4232508
> No n'és.
```

11. Quin és el nombre esperat de cigarrets fumats al dia per un fumador blanc que freqüenta restaurant sense prohibició de fumar i en la mediana de la resta de les variables segons el model nbm1?

```
> summary(tabac)
      educ          cigpric      white        age            income          cigs
restaurn
 Min.   : 6.00   Min.   :44.00   0: 98   Min.   :17.00   Min.   :  500   Min.   : 0.000   0:608
 1st Qu.:10.00   1st Qu.:58.14   1:709   1st Qu.:28.00   1st Qu.:12500   1st Qu.: 0.000   1:199
 Median :12.00   Median :61.05           Median :38.00   Median :20000   Median : 0.000
 Mean   :12.47   Mean   :60.30           Mean   :41.24   Mean   :19305   Mean   : 8.686
 3rd Qu.:13.50   3rd Qu.:63.18           3rd Qu.:54.00   3rd Qu.:30000   3rd Qu.:20.000
 Max.   :18.00   Max.   :70.13           Max.   :88.00   Max.   :30000   Max.   :80.000
    lincome         lcigpric     fumador      uns
 Min.   : 6.215   Min.   :3.784   0:497   Min.   :1
 1st Qu.: 9.433   1st Qu.:4.063   1:310   1st Qu.:1
 Median : 9.903   Median :4.112           Median :1
 Mean   : 9.687   Mean   :4.096           Mean   :1
 3rd Qu.:10.309   3rd Qu.:4.146           3rd Qu.:1
 Max.   :10.309   Max.   :4.250           Max.   :1
> coef(nbm1)
    (Intercept)        lcigpric         lincome       restaurn1          white1
    3.230196887    -0.190639259     0.100823861    -0.463101569    -0.171946693
           educ             age I((age - 38)^2)
   -0.092220554     0.012826225    -0.001592028
```

```
> eta=3.230196887-0.190639259 *4.112+0.100823861*9.903-0.171946693-
0.092220554*12+0.012826225 *38-0.001592028*0;eta
[1] 2.65355
> exp(eta)
[1] 14.20438  Cigarrets/dia Nb esperat
>
```

12. Valoreu la xifra en contrast amb les dades mostrals. Si trobeu discrepància indiqueu quina podria ser la causa.

La mostra recull fumadors i no fumadors, per tant, els que no fumen tenen un 0 en cigarrets diaris i això es nota en la tendència central de cigs que és 0, per tant, només tindria sentit modelar el nb de cigarrets fumats en fumadors, altrament no captura bé la diversitat de les dues subpoblacions.

Problema 4 (1 punt): Modelització

Per a les següents situacions, indica el tipus de model que faries servir, és a dir, model lineal o generalitzat, quina seria la resposta i la seva distribució i si faries servir un model mixt o no. En cas de fer servir un model mixt, indica quina variable determina l'agrupació en la mostra.

1. Abundància d'espècies: en una gran extensió de terreny es seleccionen cinc parcel·les i es compta el número d'escarabats trobats en 20 punts seleccionats a l'atzar en cada parcel·la, recollint també informació del tipus de cultiu en aquella parcel·la

Model lineal generalitzat mixt amb resposta Poisson: el número d'escarabats recollits en cada punt seria la resposta i les parcel·les seria el factor aleatori, ja que és raonable pensar que les observacions recollides en punts de la mateixa parcel·la no són independents i no té sentit incloure la parcel·la com a efecte fix.

2. Propensió als accidents: Una companyia d'assegurances de cotxe, selecciona 50 usuaris i registra el número d'accidents que han tingut en un any, així com característiques del conductor i del vehicle

Model lineal generalitzat amb resposta Poisson: el número d'accidents que han tingut seria la resposta. La mostra es pot considerar independent i per tant no té sentit que el model sigui mixt.

3. Incidència d'una malaltia infecciosa: Es seleccionen 5 classes d'una escola i 10 nens en cada classe i es registra la seva edat, el sexe, si s'han vacunat o no i si han desenvolupat o no la malaltia

Model lineal generalitzat mixt amb resposta binària: el fet d'haver desenvolupat la malaltia és la variable resposta, i està clar que les observacions recollides en nens de la mateixa classe poden estar correlacionades i per tant la classe seria el factor aleatòrio.

4. Evolució de la capacitat visual: Es seleccionen 20 individus i se'ls mesura la mitjana del número de diòptries dels dos ulls (defecte visual) en visites anuals durant 5 anys

Model lineal mixt amb resposta gaussiana: la mitjana de dioptríes seria la variable resposta i com els nens es visiten de forma repetida, les observacions dintre de cada nen estaran correlacionades. Així, el nen és el factor aleatori del model.

5. Efectivitat de la publicitat: Es seleccionen 100 individus i es recull les seves característiques si han vist o no un cert anunci i si han comprat o no el producte que s'hi anuncia

Model lineal generalitzat amb resposta binària: el fet d'haver comprat o no el producte és la variable resposta, i la mostra es pot considerar independent.