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

CURS 2017-2018 Q1 – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: 8 de Gener a les 12) **Aula S04)**

Nom de l´alumne:	**DNI:**
Professors:	Josep Anton Sànchez, Jordi Cortés, Erik Cobo
Localització:	Edifici C5 D217 o H6-67
Normativa de l'examen:	ÉS PERMET APUNTS DE TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 30 min

Problema 1 (4.5 punts): Resposta Binària

Es vol estudiar la relació entre la mortalitat del recent nascut (neonat) amb els següents factors: (1) edat de la mare durant l'embaràs; (2) si el nen ha estat prematur; i (3) quantitat de cigarretes diàries que fumava la mare. Les dades consten de 6851 observacions amb les variables mencionades a continuació:

Mort: Mort del neonat (*NO*/*SI*)
Tabac: Quantitat de cigarretes diàries (*<5 cig/dia* / *≥5 cig/dia*)
Prematur: Si el naixement/la mort s'ha produït abans de la data a terme (*NO*/*SI*)
Edat: Edat de la mare dicotomitzada en el moment del embaràs (*≤30* / *>30 anys*)
Edat.cont: Edat en anys (variable quantitativa discreta)

[En totes les variables categòriques, la primera categoria és la de referència]

1. **Formuleu i calculeu el model logit que modela una probabilitat idèntica de mort per tots els grups definits pel factor *Tabac*. Useu les dades de la taula mostrada a continuació. (0.25 punts)**

```
> table(dades1.des$Tabac,dades1.des$Mort)

                NO   SI
  <5 cig/dia  6068  129
  >=5 cig/dia  634   20
```

El model que considera una probabilitat idèntica per tots els nivells de tabac és el model nul:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta \qquad i = 1{,}2 \qquad \pi_i: probabilitat\ de\ morir\ en\ la\ categòria\ ièsima\ del\ Tabac$$

La probabilitat marginal de mort del neonat (resposta positiva) és 0.021 (149/6851) i els odds són de 149 a 6702 (o de 0.022 a 1), per tant el logodd és **-3.81** que és l'estimador de la constant (intercept) en el model nul.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = -3.81$$

Sortida de R

```
> # Model nul
> (t1 <- table(dades1.des$Mort))          # Taula global

  NO   SI
6702  149

> (prop.t1 <- prop.table(t1))             # Probabilitats

        NO         SI
0.97825135 0.02174865
```

```
> log(prop.t1[2]/prop.t1[1])              # Logodd de morir:Intercept
       SI
-3.806215

> glm(Mort~1,dades1.des,family=binomial) # Coeficient amb glm

Call:  glm(formula = Mort ~ 1, family = binomial, data = dades1.des)

Coefficients:
(Intercept)
     -3.806

Degrees of Freedom: 6850 Total (i.e. Null);  6850 Residual
Null Deviance:     1436
Residual Deviance: 1436        AIC: 1438
```

2. Emprant la mateix taula, calcula els coeficients del model amb la variable tabac com a explicativa dins del model (0.25 punts)

Ara, el model a estimar és el següent:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \quad i = 1{,}2 \quad sent\ \alpha_1 = 0$$

Per calcular el coeficient corresponent a la categoria de *"≥5 cig/dia"*, s'han de calcular els odds de mort neonatal per a fumadores de menys de 5 cigarretes diàries i la resta per separat:

$$\left.\begin{aligned} Odd_{<5/dia} &= \frac{129}{6068} = 0.0213 \\ Odd_{\geq 5/dia} &= \frac{20}{634} = 0.0315 \end{aligned}\right\} \rightarrow OR_{Tabac} = \frac{0.0315}{0.0213} = 1.48 \rightarrow \alpha_2 = logOR_{tabac} = \mathbf{0.395}$$

El terme constant, es correspon al logodd de la categoria de *"<5 cig/dia"*

$$\eta = \log(0.0213) = \mathbf{-3.851}$$

Sortida de R

```
> # Model amb tabac
> (t2 <- table(dades1.des$Tabac,dades1.des$Mort))

                NO   SI
  <5 cig/dia  6068  129
  >=5 cig/dia  634   20

> (Odds <- t2[,2]/t2[,1])
 <5 cig/dia >=5 cig/dia
 0.02125906  0.03154574

> (logOR <- log(Odds[2]/Odds[1]))
>=5 cig/dia
  0.3946553

> (Intercept <- log(Odds[1]))
<5 cig/dia
 -3.850972

> glm(Mort~Tabac,dades1.des,family=binomial)

Call:  glm(formula = Mort ~ Tabac, family = binomial, data = dades1.des)

Coefficients:
     (Intercept)  Tabac>=5 cig/dia
         -3.8510            0.3947

Degrees of Freedom: 6850 Total (i.e. Null);  6849 Residual
Null Deviance:     1436
Residual Deviance: 1433        AIC: 1437
```

3. Calculeu el nombre predit de neonats que moririen i sobreviurien per cadascun dels nivells de la variable *Tabac* sota la hipòtesi del model nul descrit al primer apartat. (0.25 punts)

El model nul considera la mateixa probabilitat de mort per ambdues categories, per tant, els efectius esperats en aquest cas serien:

$$e_{11} = n_{<5cig} \cdot P_{MortNo} = 6197 \cdot 0.978 = 6062.22$$

$$e_{21} = n_{<5cig} \cdot P_{MortSi} = 6197 \cdot 0.022 = 134.78$$

$$e_{12} = n_{\geq 5cig} \cdot P_{MortNo} = 654 \cdot 0.978 = 639.78$$

$$e_{22} = n_{\geq 5cig} \cdot P_{MortSi} = 654 \cdot 0.022 = 14.22$$

Sortida de R

```
> (t3 <- table(dades1.des$Tabac))           # Freqüencies de variable tabac

 <5 cig/dia >=5 cig/dia
       6197         654

> (fit0 <- round(cbind(t3 * prop.t1[1],
                       t3 * prop.t1[2]),2)) # Valors esperats
                [,1]   [,2]
<5 cig/dia  6062.22 134.78
>=5 cig/dia  639.78  14.22
```

4. Calcula la deviança del model nul descrit en el punt 1 usant les dades calculades per les prediccions en el punt 3. (0.25 punts)

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

$$= 2 \cdot \left(129 \cdot log\left(\frac{129}{134.78}\right) + 6068 \cdot log\left(\frac{6068}{6062.22}\right) + 20 \cdot log\left(\frac{20}{14.22}\right) + 634 \cdot log\left(\frac{634}{639.78}\right)\right)$$

$$= 2.39$$

Sortida de R

```
> 2*(129*log(129/134.78) + 6068*log(6068/6062.22)+
  20*log(20/14.22) + 634*log(634/639.78))
[1] 2.392695

> summary(glm(cbind(Mort_SI,Mort_NO)~1,dades1.tabac,family=binomial))

Call:
glm(formula = cbind(Mort_SI, Mort_NO) ~ 1, family = binomial,
    data = dades1.tabac)

Deviance Residuals:
      1       2
-0.5066  1.4604

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.80621    0.08283  -45.95   <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: 2.3894  on 1  degrees of freedom
Residual deviance: 2.3894  on 1  degrees of freedom
AIC: 15.878

Number of Fisher Scoring iterations: 4
```

5. **Valora si el fet de ser la mare fumadora té relació amb la mortalitat del neonat amb un nivell de significació α =0.05. Empra un test estadístic adient, justifica'l i interpreta. (0.25 punts)**

Amb els resultats dels apartats anteriors amb les dades agregades cal considerar que el model saturat té una deviança de 0 i el model nul de 2.39. Per tant, es pot fer un test de bondat de l'ajust del model nul:

H0: El model nul s'ajusta bé a les dades

La distribució de l'estadístic deviança segueix assimptòticament una χ^2amb 1 grau de llibertat (*gl = n-p = 2-1 = 1*) i el p valor del contrast és $P(\chi^2 > 2.39) = 0.12$. Per tant, no hi ha evidència per rebutjar hipòtesi (H0) de que el model nul s'ajusta bé a les dades: l'efecte del tabac per si sol no és significatiu.

També es podria arribar a deduir emprant el punt crític a partir de les taules:

n	*0,995*	*0,99*	*0,975*	*0,95*	*0,9*	*0,75*	*0,5*	*0,25*	*0,05*	*0,025*	*0,01*	*0,005*
1	7,879	6,635	5,024	3,841	2,706	1,323	0,455	0,102	0,004	0,001	0,000	0,000

El punt crític en una χ^2amb 1 grau de llibertat i una α =0.05 és 3.84. Com que 2.39 < 3.84, no tenim evidència per rebutjar H0.

També es podria pensar en un test de diferències de deviances entre el model nul i el model saturat, que seria equivalent amb el format de dades agregades; en canvi, en el format de dades individualitzades disponible és l'única possibilitat de fer plantejar el contrast del factor tabac, doncs en aquest cas, no seria un model saturat.

Sortida de R

```
> 1-pchisq(2.39,1)
[1] 0.1221136
> qchisq(0.95,1)
[1] 3.841459
```

6. **Es calcula amb el conjunt de dades individuals el model logístic per la mortalitat del neonat en funció del factor tabac. Ompliu la taula següent (en alguns casos, podeu emprar els resultats dels apartats anteriors) i indiqueu quins elements seran iguals o diferents en els dos models ajustats quan s'empren dades individualitzades i quan s'empren dades agrupades. (2.25 punts)**

***m0*: model nul**
***m1*: model amb variable tabac**

	Dades individuals	**Dades agrupades**	**Iguals?**
Terme independent (intercept) en *m1*	*-3.851**	*-3.851**	*Iguals*
Estimador del coeficient de la variable Tabac en *m1*	*0.394**	*0.394**	*Iguals*
Deviança de *m0*	*1435.54*	*2.389**	*Diferents*
Graus llibertat de la deviança residual en *m0*	*6850*	*1*	*Diferents*
Deviança residual en *m1*	*1433.15*	*0*	*Diferents*
Graus llibertat de la deviança residual en *m1*	*6849*	*0*	*Diferents*
AIC de *m1*	*1437.15*	*15.489*	*Diferents*
BIC de *m1*	*1450.81*	*12.875*	*Diferents*
Dev(*m0*) - Dev(*m1*)	*2.389**	*2.389**	*Iguals*

* Resultats obtinguts en apartats anteriors

Resultats que falten per dades individuals:

Deviança de m0

$$loglikelihood = l = 149 \cdot log\left(\frac{149}{6851}\right) + 6702 \cdot log\left(\frac{6702}{6851}\right) = -717.77$$

$$D_{m0} = -2l = -2 \cdot (-717.77) = 1435.54$$

Graus llibertat de la deviança residual en m0

$$gll = n - p = 6851 - 1 = 6850$$

Deviança residual en *m1*

Com que la diferencia entre la deviança del model m0 i m1 ha de ser la mateixa per dades agregades i desagregades:

$$D_{m1} = D_{m0} - 2.3894 = 1433.15$$

Graus llibertat de la deviança residual en m1

$$gll = n - p = 6851 - 2 = 6849$$

AIC de m1

$$AIC = 2k - 2l = 4 - 2 \cdot (-717.77) = 1437.15$$

BIC de m1

$$BIC = 2\log(n) - 2l = 2\log(6851) - 2 \cdot (-717.77) = 1450.81$$

Resultats que falten per dades agregades:

Graus llibertat de la deviança residual en m0

$$gll = n - p = 2 - 1 = 1$$

Deviança residual en *m1*

Al ser el model saturat, la deviança és 0

Graus llibertat de la deviança residual en m1

$$gll = n - p = 2 - 2 = 0$$

AIC de m1

$$loglikelihood = l$$
$$= log\left(\binom{6197}{129}\right) + 129 \cdot log\left(\frac{129}{6197}\right) + 6068 \cdot log\left(\frac{6068}{6197}\right) + log\left(\binom{654}{20}\right) + 20$$
$$\cdot log\left(\frac{20}{654}\right) + 634 \cdot log\left(\frac{634}{654}\right) = -5.7444$$

$AIC = 2k - 2l = 4 - 2 \cdot (-5.74) = 15.49$

BIC de m1

$BIC = 2\log(n) - 2l = 2\log(2) - 2 \cdot (-5.74) = 15.49$

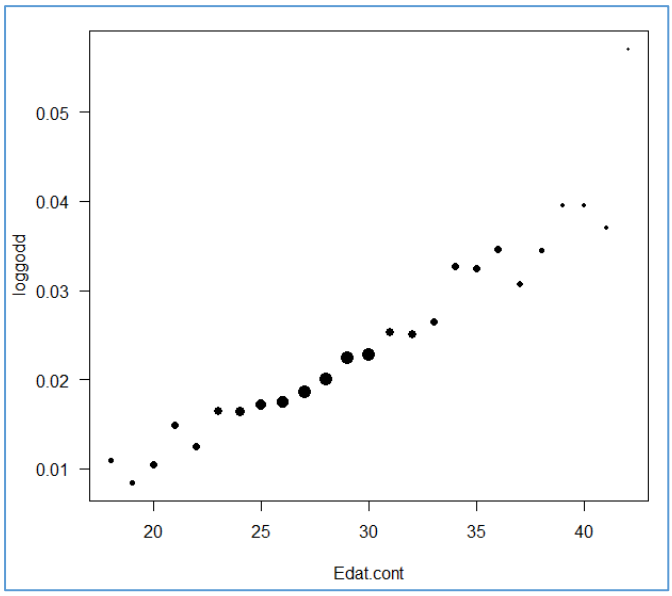
Sortida de R

```
> # [Desagregades] - Deviança m0
> (loglik.des0 <- 149*log(149/6851) + 6702*log(6702/6851))
[1] -717.7702
> logLik(mod.des.log0)
'log Lik.' -717.7702 (df=1)
> -2*(loglik.des0)
[1] 1435.54
> mod.des.log0$deviance
[1] 1435.54


> # [Desagregades] - Deviança residual en m1
> -2*(loglik.des0) - dev0
[1] 1433.148
> mod.des.log1$deviance
[1] 1433.151


> # [Desagregades] - AIC
> -2*(loglik.des0) - dev0 + 4
[1] 1437.148
> AIC(mod.des.log1)
[1] 1437.151


> # [Desagregades] - BIC
> -2*(loglik.des0) - dev0 + 2*log(6851)
[1] 1450.812
> BIC(mod.des.log1)
[1] 1450.815



> # [Agregades] - AIC
> loglik.agregades1 <- log(choose(6197,129)) + 129*log(129/6197) + 6068*log(60
68/6197) + log(choose(654,20)) + 20*log(20/654) + 634*log(634/654)
> logLik(mod.tabac.log1)
'log Lik.' -5.744419 (df=2)
> -2*(loglik.agregades1) + 4
[1] 15.48884
> AIC(mod.tabac.log1)
[1] 15.48884


 > # [Agregades] - BIC
 > -2*(loglik.agregades1) + 2*log(2)
 [1] 12.87513
 > BIC(mod.tabac.log1)
 [1] 12.87513
```

7. **Es poden comparar els AICs dels models amb dades agregades i sense agregar? (0.25 punts)**

   No. No es poden comparar perquè els conjunts de dades de partida no són els mateixos. Els AICs amb dades desagregades sortiran sempre majors perquè el nombre de sumands en el logaritme de la versemblança és molt major que en les dades agregades.

8. **El següent gràfic mostra el logodds de morir en funció de la variable *Edat.cont* (edat de la mare en anys, com a variable quantitativa discreta). La grandària dels punts és proporcional a l'arrel quadrada del nombre de mares en una determinada edat. Veient aquest gràfic, discuteix si creus si la variable edat és millor considerar-la com a quantitativa o dicotomitzada (≤30 vs. >30) (0.25 punts)**



   En el gràfic es pot veure que l'increment del logodd amb l'edat és raonablement lineal respecte a l'edat. Per tant, té sentit considerar la variable edat com a quantitativa, ja que dicotomitzar-la suposaria una pèrdua d'informació important. [nota: la grandària del punt és proporcional a l'arrel de la freqüència de cada any. sembla que en els extrems es desvia lleugerament de la linealitat, però s'ha de considerar que tenen freqüències més petites, amb més variabilitat]

9. **El següent model considera en forma additiva els factors Tabac, Prematuritat i Edat (com a quantitativa discreta). Indica quin és el Odds Rati (OR) respecte a la mortalitat del neonat corresponent a un increment de 5 anys en l'edat de la mare i interpreta'l. (0.25 punts)**

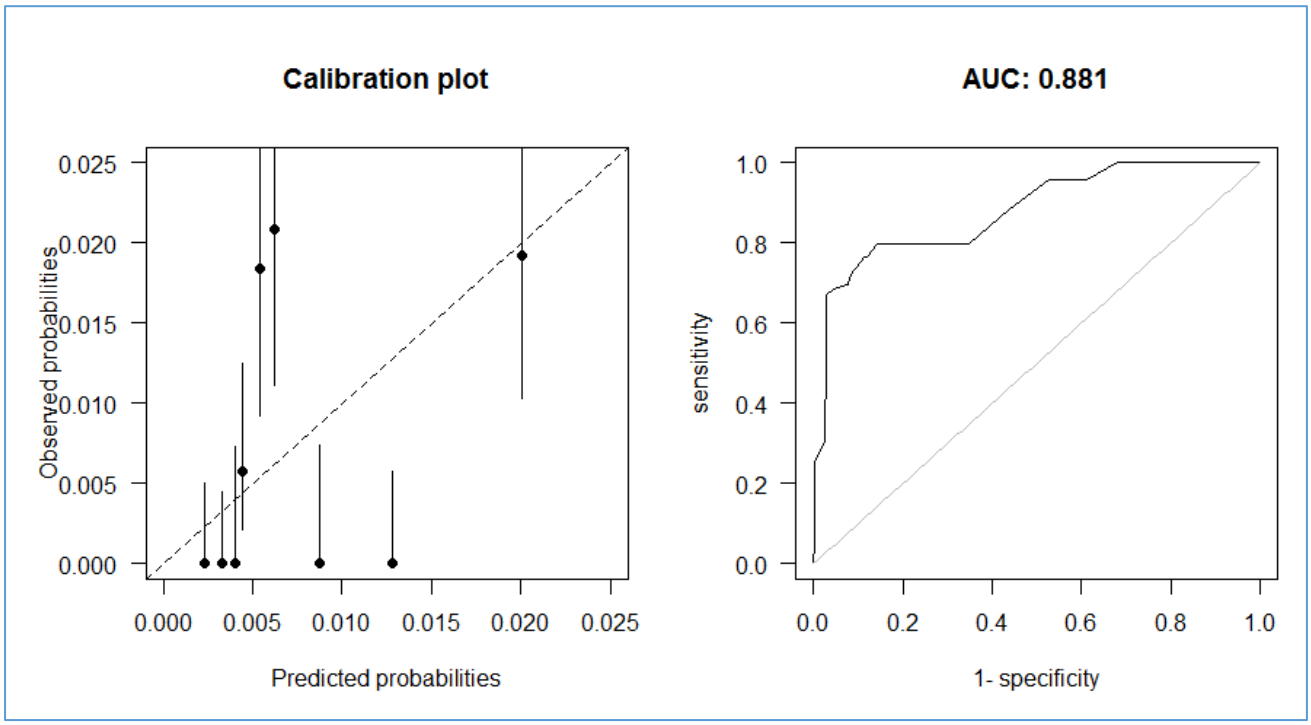
```
Call:
glm(formula = Mort ~ Tabac + Prematur + Edat.cont,
    family = binomial(link = logit),data = dades1.des)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1271  -0.1489  -0.1037  -0.0835   3.3161

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.28617    0.54718 -16.971   <2e-16 ***
Tabac>=5 cig/dia -0.56228    0.29231  -1.924   0.0544 .
PrematurSI       3.92313    0.20180  19.440   <2e-16 ***
Edat.cont        0.14515    0.01662   8.731   <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: 1435.5  on 6850  degrees of freedom
Residual deviance: 1008.5  on 6847  degrees of freedom
AIC: 1016.5
```

```
Number of Fisher Scoring iterations: 7
```

$$OR_{5anys} = e^{5 \cdot 0.14515} = 2.07$$

Un increment en l'edat de la mare de 5 anys incrementa les odds de mortalitat del neonat en **2.07**.

10. **Comenta que aporten els següents gràfics corresponents al model de l'apartat anterior. Quin fa referència a la calibració del model i quin a la seva capacitat predictiva? (0.25 punts)**



La figura de l'esquerra és un gràfic de **calibració** del model on es comparen les probabilitats observades a la mostra amb les predites. Es pot observar que el model no està prou ben calibrat ja que aquestes probabilitats discrepen molt i molts dels intervals de confiança de les probabilitats observades no creuen la bisectriu que representa la igualtat perfecta entre probabilitats predites i observades. S'hauria de reformular el model incloent factors no considerats o possibles interaccions.
La figura de la dreta és la corba ROC que representa la sensibilitat en funció de 1-especifitat per a cada punt de tall de les probabilitats predites del model. Serveix per avaluar la **capacitat predictiva** del nostre model: quan més bombada la corba, millor capacitat predictiva. Concretament, l'àrea que deixa a sota és de 0.881 que indica una capacitat predictiva molt bona, malgrat la calibració dolenta: en un 88.1% de les parelles d'observacions (mort - no mort), l'observació amb mort del neonat tindrà una probabilitat predita més alta.

Problema 2 (4.5 punts): Comptatges

Es vol estudiar la relació que puguin tenir les morts coronàries amb el tabac i l'edat en una cohort d'homes metges. Es disposa d'un seguiment global de 181469 persones/any. Les dades estan agrupades segons si la persona és fumadora o no i segons el rang d'edat. En la següent taula, estan les variables del joc de dades.

Mort: número de morts coronàries dins del rang d'edat corresponent i categoria de fumador.
Fumador: Si la persona fuma o no (*NO*/*SI*)
Edat: Edat de la persona categoritzada a l'inici del seguiment (*≤45/45-64/55-64/65-74/75-84*)
PersonesAny: Temps total de seguiment en anys

[En totes les variables categòriques, la primera categoria és la de referència]

Totes les preguntes d'aquest apartat valen 0.5 punts.

1. Compara els 2 següents models. Digues quina diferència hi ha entre ambdós i quina implicació té sobre la interpretació dels coeficients.

Model 1

```
> summary(mod.poi0)

Call:
glm(formula = Morts ~ Fumador + Edat, family = poisson, data = dades2)

Deviance Residuals:
     1       2       3       4       5       6       7       8       9      10
 0.4910  -1.4070   0.4001  -1.0533   0.3039  -0.7797   0.1153  -0.2909  -1.2018   2.6785

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.5471     0.1948   7.942 1.99e-15 ***
FumadorSI     1.8306     0.1072  17.079  < 2e-16 ***
Edat45-64     1.2272     0.1950   6.293 3.12e-10 ***
Edat55-64     1.9290     0.1835  10.510  < 2e-16 ***
Edat65-74     1.8396     0.1846   9.964  < 2e-16 ***
Edat75-84     1.3640     0.1922   7.098 1.27e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 644.269  on 9  degrees of freedom
Residual deviance:  12.907  on 4  degrees of freedom
AIC: 79.975

Number of Fisher Scoring iterations: 4
```

Model 2

```
> summary(mod.poi1)

Call:
glm(formula = Morts ~ Fumador + Edat, family = poisson, data = dades2,
    offset = log(PersonesAny))

Deviance Residuals:
      1        2        3        4        5        6        7        8        9       10
 0.90149  -2.17960   0.51023  -1.30766   0.05178  -0.13907  -0.08749   0.22928  -0.91254   1.91944

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

```
(Intercept)  -7.9194     0.1918 -41.298  < 2e-16 ***
FumadorSI     0.3546     0.1074   3.303 0.000957 ***
Edat45-64     1.4840     0.1951   7.606 2.82e-14 ***
Edat55-64     2.6275     0.1837  14.301  < 2e-16 ***
Edat65-74     3.3505     0.1848  18.130  < 2e-16 ***
Edat75-84     3.7001     0.1922  19.249  < 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: 935.064  on 9  degrees of freedom
Residual deviance:  12.133  on 4  degrees of freedom
AIC: 79.201

Number of Fisher Scoring iterations: 4
```

El segon model inclou el temps de seguiment com a offset per considerar que diferents categories poden haver tingut un nombre més gran d'observacions.

En el primer model, els coeficients estimen les freqüències de mort coronària dins de cada combinació de categories de la resposta. En el segon model, que inclou el offset, té en compte el nombre total d'observacions i, per tant, en lloc de la proporció de casos, parla de la taxa (persona-any).

2. Usant el Model 2, troba un estimació puntual i per interval de confiança del 95% per l'increment en la taxa de morts coronàries en els fumadors i interpreta'l.

$$\widehat{IRR} = \frac{\hat{\lambda}_{FumadorSI}}{\hat{\lambda}_{FumadorNO}} = e^{0.3546} = 1.43$$

$$IC(95\%, IRR) = exp\{\widehat{IRR} \mp z_{0.975} \cdot \widehat{se}_{FumadorSI}\} = exp\{0.3546 \mp 1.96 \cdot 0.1074\} = [1.15\,, 1.76]$$

Ser fumador està associat amb un increment en la taxa per mortalitat coronària d'un 43% respecte als no fumadors. Aquest increment estarà entre un 15 i un 76% amb una confiança del 95%.

Sortida R

```
> exp(0.3546 + c(-1,1)*1.96*0.1074)
[1] 1.154995 1.759631

> exp(coef(mod.poi1)[2] + c(-1,1)*qnorm(0.975)*sqrt(diag(s$cov.unscaled)[2]))
[1] 1.155102 1.759600

> exp(confint.default(mod.poi1)[2,])
   2.5 %   97.5 %
1.155102 1.759600
```

3. Digues si el *Model 2* es pot considerar vàlid a través del test de la deviança i emprant un nivell de significació α = 0.05.

No és pot considerar vàlid ja que el punt crític per un χ2 amb 4 graus de llibertat és 9.48 i el valor de la deviança residual (12.13) queda per sobre d'aquest llindar. Per tant, no podem donar aquest model com a correcte.

Sortida R

```
> 1-pchisq(mod.poi1$deviance,df.residual(mod.poi1))
[1] 0.01638998
> qchisq(0.95, df.residual(mod.poi1))
[1] 9.487729
```

. Estima el paràmetre ϕ de sobredispersió en el segon model emprant el mètode del moments, és a dir, usant el estadístic de Pearson. Per facilitar els càlculs, es proporcionen els residus de Pearson i els residus de Pearson al quadrat.

```
> residuals(mod.poi1,"pearson")
          1           2           3           4           5
 0.92711854 -1.84873670  0.51457296 -1.23678799  0.05181536
          6           7           8           9          10
-0.13846726 -0.08739936  0.23096959 -0.89929957  2.04772941

> residuals(mod.poi1,"pearson")^2
          1           2           3           4           5
0.859548782 3.417827376 0.264785333 1.529644532 0.002684832
          6           7           8           9          10
0.019173182 0.007638647 0.053346951 0.808739722 4.193195753
```

$$\hat{\phi} = \frac{\chi^2}{n-p} = \frac{\sum e_i^2}{n-p}$$
$$= \frac{0.8595 + 3.4178 + 0.2648 + 1.5296 + 0.0027 + 0.0192 + 0.0076 + 0.0533 + 0.8087 + 4.1932}{10-6}$$
$$= 2.789$$

Sortida R

```
> pr <- residuals(mod.poi1,"pearson")
> (phi <- sum(pr^2)/df.residual(mod.poi1))
[1] 2.789146
```

5. Digues quin seria el coeficient de la variable "*Fumador*" per la categoria "SI" en el model quasipoisson equivalent al *Model 2* i quin seria el seu error estàndard.

El model amb la familia quasipoisson proporciona els mateixos coeficients que el model amb la família Poisson; per tant, el coeficient seria exactament el mateix: **0.3546**

En canvi, l'error estàndard es veuria incrementat per l'arrel quadrada del paràmetre de sobre dispersió:

$$se_{qp} = \sqrt{\hat{\phi}} \cdot se_p = \sqrt{2.789} \cdot 0.1074 = \mathbf{0.179}$$

Sortida R

```
> summary(mod.qpoi1)

Call:
glm(formula = Morts ~ Fumador + Edat, family = quasipoisson,
    data = dades2, offset = log(PersonesAny))

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10
 0.9015  -2.1796  0.5102  -1.3077  0.0518  -0.1391  -0.0875  0.2293  -0.9125  1.9194

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.9194     0.3203 -24.728 1.59e-05 ***
FumadorSI     0.3546     0.1793   1.978 0.119121
Edat45-64     1.4840     0.3258   4.554 0.010383 *
Edat55-64     2.6275     0.3068   8.563 0.001021 **
Edat65-74     3.3505     0.3086  10.856 0.000409 ***
Edat75-84     3.7001     0.3210  11.526 0.000324 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for quasipoisson family taken to be 2.789156)

    Null deviance: 935.064  on 9  degrees of freedom
Residual deviance:  12.133  on 4  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

6. Comenta si amb les dades disponibles, podries calcular el AIC del model de la família quasipoisson equivalent al Model 2 i justifica-ho.

No es podria calcular l'AIC ja que la família quasipoisson no és una família de probabilitats i per tant, no és pot calcular la versemblança ni, òbviament, el AIC.

7. A continuació, tens la sortida de R equivalent al Model 2 però emprant un model binomial negativa. Digues perquè els coeficients i els errors estàndards d'aquest model coincideixen a quan s'empra el model de Poisson.

```
Call:
glm.nb(formula = Morts ~ Fumador + Edat + offset(log(PersonesAny)),
    data = dades2, link = log)

Deviance Residuals:
      1       2       3       4       5       6       7       8       9      10
0.90149 -2.17960  0.51023 -1.30766  0.05184 -0.13907 -0.08748  0.22928 -0.91254  1.91944

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.9194     0.1918 -41.298  < 2e-16 ***
FumadorSI     0.3546     0.1074   3.303 0.000957 ***
Edat45-64     1.4840     0.1951   7.606 2.82e-14 ***
Edat55-64     2.6275     0.1837  14.301  < 2e-16 ***
Edat65-74     3.3505     0.1848  18.130  < 2e-16 ***
Edat75-84     3.7001     0.1922  19.249  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(22206930039) family taken to be 1)

    Null deviance: 935.064  on 9  degrees of freedom
Residual deviance:  12.133  on 4  degrees of freedom
AIC: 81.202

Number of Fisher Scoring iterations: 1
```

El paràmetre θ estimat és molt gran (22206930039), el que indica que el terme quadràtic dins de la sobre-dispersió no té rellevància i que, per tant la variància de la resposta es considera igual a l'esperança. Això implica que el model de Poisson i de binomial Negativa siguin equivalents.

8. A continuació, tens el model de Poisson saturat que conté la interacció de l'edat amb ser fumador. Digues si els termes d'interacció han de romandre en el model o no i justifica-ho.

```
Call:
glm(formula = Morts ~ Fumador * Edat, family = poisson, data = dades2,
    offset = log(PersonesAny))

Deviance Residuals:
 [1]  0  0  0  0  0  0  0  0  0  0

Coefficients:
```

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.1479     0.7071 -12.937  < 2e-16 ***
FumadorSI             1.7469     0.7289   2.397  0.01654 *
Edat45-64             2.3574     0.7638   3.087  0.00203 **
Edat55-64             3.8298     0.7319   5.233 1.67e-07 ***
Edat65-74             4.6227     0.7319   6.316 2.69e-10 ***
Edat75-84             5.2944     0.7296   7.257 3.96e-13 ***
FumadorSI:Edat45-64  -0.9866     0.7901  -1.249  0.21174
FumadorSI:Edat55-64  -1.3625     0.7562  -1.802  0.07158 .
FumadorSI:Edat65-74  -1.4423     0.7565  -1.906  0.05659 .
FumadorSI:Edat75-84  -1.8470     0.7572  -2.439  0.01471 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 9.3506e+02  on 9  degrees of freedom
Residual deviance: 6.2172e-15  on 0  degrees of freedom
AIC: 75.068
Number of Fisher Scoring iterations: 3
```

```
Analysis of Deviance Table

Model 1: Morts ~ Fumador * Edat
Model 2: Morts ~ Fumador + Edat
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0      0.000
2         4     12.133 -4  -12.133  0.01639 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El terme d'interacció ha de romandre dins del model ja que és estadísticament significatiu (p = 0.016). També es podia haver deduït amb el fet de que el model sense la interacció no es podia considerar vàlid (veure pregunta 2)

9. Donat el model de l'apartat anterior, calcula la probabilitat que una persona fumadora en el rang d'edat de 75 a 84 anys pateixi una mort cardiovascular en el termini de 2 anys.

$$\hat{\lambda} = \exp\{-9.1479 + 1.7469 + 5.2944 - 1.8470\} = 0.01918551$$

$$F(2) = 1 - \exp\{-2 \cdot 0.01918551) = 0.037$$

Té una probabilitat de **0.037**

Problema 3 (1 punt): Modelització

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

1. Creixement d'arbres: es seleccionen 20 arbres situats en 20 parcel·les diferents. Mitjançant l'extracció d'una mostra que permet analitzar els anells s'estableix el diàmetre de l'arbre durant els últims 50 anys. També es disposa d'informació històrica del comportament climàtic en aquest període (temperatura, humitat i contaminació anual mitjana).

Model lineal mixt amb resposta de tipus normal (o gamma). La resposta és el diamant de l'arbre i les dades són longitudinals (cada arbre es segueix durant 50 anys). El factor aleatori és l'arbre. La parcel·la no és un factor aleatori ja que cada arbre pertany a una parcel·la diferent.

2. Aplicació de contactes: Una web de contactes implementa un model perquè un client pugui explorar els candidats de la base de dades i en base a les afinitats en edat, personalitat i aficions establir si la cita tindrà o no èxit. La construcció del model es realitza a partir d'una base de dades preexistent amb les primeres cites recollides amb clients d'un altre país.

Model lineal generalitzat amb resposta binaria. La resposta és si la cita tindrà o no èxit. El fet que la base de dades s'hagi construït amb primeres cites garanteix que les observacions són independents respecte als individus (cada subjecte contribueix amb una única observació).

3. Avaries en la línia de producció: en una empresa es recullen dades històriques de les màquines (3 en funcionament), operaris (5 en l'empresa) i productes (3 tipus de peces fabricades). Per a cada combinació es registra el nombre d'hores de producció i el nombre de parades a la línia causa d'incidències.

Model lineal generalitzat amb resposta de Poisson. La resposta és el nombre de parades. Les hores de treball són l'offset. Els factors considerats són tots fixos ja que és interessant determinar quins factors s'associen a un major nombre d'incidències.

4. Producció de lactis: en una granja es registren diàriament els litres de llet que produeixen 30 vaques durant un mes. De cada animal es disposa de dades fisiològics (pes, temperatura, ...) i el tipus de dieta que s'aplica (hi ha 3 tipus de dietes que es van assignar aleatòriament a cada animal).

Model lineal mixt amb resposta gaussiana. La resposta és la quantitat de llet produïda diàriament per cada animal. L'animal és un factor aleatori que actua com a bloc en un disseny longitudinal. Les covariables són els dades fisiològics i la dieta és un factor fix.

5. Viabilitat de l'oferta: Una consultora analitza les ofertes realitzades per 135 empreses en 85 obres de concurs públic. A partir del pressupost que planteja un nou client per a una adjudicació es pretén determinar si obtindrà o no la concessió.

Model lineal mixt generalitzat amb resposta binaria. La resposta és si obtindrà o no la concessió. El fet que per a 85 obres s'han presentat ofertes de 135 empreses indica que per a una mateixa obra es presenten diverses empreses. Per això, la obra actua com a factor aleatori en el model.