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

CURS 2013-2014 Q1 – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: 21/01/2014 a les 15:00h Aula 003-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 29 de Gener al Web Docent de MLGz

Revisió de l'examen: 29 de Gener a 10:00 h a C5-217-C Nord o H- P6-67

Problema 1 (4 punts): Resposta Binària

Un estudio de los patrones de divorcio por Thornes y Collard (1979) analizó datos de 1036 sujetos. A cada persona se le pidió (a) si habían hecho el amor con alguien más antes de su matrimonio (Premarital), y (b) si tenían relaciones sexuales con otra persona después del matrimonio (Extramarital). También se refleja el género (Gender) y si su status es casado o ha solicitado el divorcio (Divorced). La tabla con los resultados es la siguiente:

			Divorced	
Gender	Premarital	Extramarital	No	Yes
Man	No	No	130	68
		Yes	4	17
	Yes	No	42	60
		Yes	11	28
Man Total			187	173
Woman	No	No	322	214
		Yes	4	36
	Yes	No	25	54
		Yes	4	17
Woman Total			355	321
Grand Total			542	494

A partir del fichero con los datos desagregados, se obtienen los siguientes resultados con R:

```
> table(divorce$Divorced,divorce$Premarital)

      No Yes
  No  460  82
  Yes 335 159
> table(divorce$Divorced,divorce$Extramarital)

      No Yes
  No  519  23
  Yes 396  98

> anova(modT)

Analysis of Deviance Table
Model: binomial, link: logit
Response: Divorced
Terms added sequentially (first to last)
                            Df Deviance Resid. Df Resid. Dev
NULL                                         1035     1434.0
```

```
Premarital                      1   42.549      1034     1391.4
Extramarital                    1   47.247      1033     1344.2
Gender                          1    4.530      1032     1339.7
Premarital:Extramarital         1   12.931      1031     1326.7
Premarital:Gender               1    0.258      1030     1326.5
Extramarital:Gender             1    0.293      1029     1326.2
Premarital:Extramarital:Gender  1    0.146      1028     1326.0

> anova(mod1,mod4)
Analysis of Deviance Table

Model 1: Divorced ~ Premarital
Model 2: Divorced ~ Premarital * Extramarital
  Resid. Df Resid. Dev Df Deviance
1      1034     1391.4
2      1032     1331.3  2   60.161

> anova(mod2,mod4)
Analysis of Deviance Table

Model 1: Divorced ~ Extramarital
Model 2: Divorced ~ Premarital * Extramarital
  Resid. Df Resid. Dev Df Deviance
1      1034     1369.6
2      1032     1331.3  2   38.304
```

Modelo seleccionado:

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0647  -1.0166  -0.8981   1.3472   1.4852

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.6997     0.1327  -5.271 1.36e-07 ***
PremaritalYes                 1.0995     0.1787   6.154 7.57e-10 ***
ExtramaritalYes               2.3960     0.3879   6.177 6.53e-10 ***
GenderWoman                   0.3089     0.1458   2.118  0.03415 *
PremaritalYes:ExtramaritalYes -1.7999     0.5130  -3.509  0.00045 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1434.0  on 1035  degrees of freedom
Residual deviance: 1326.7  on 1031  degrees of freedom
AIC: 1336.7
```

1) Calcula manualmente el modelo nulo con enlace logit para la respuesta “Divorced”

```
> log(494/542)
[1] -0.09273048

> summary(mod<-glm(Divorced~1,divorce,family=binomial))

Call:
glm(formula = Divorced ~ 1, family = binomial, data = divorce)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.138  -1.138  -1.138   1.217   1.217

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.09273    0.06220  -1.491    0.136

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1434  on 1035  degrees of freedom
Residual deviance: 1434  on 1035  degrees of freedom
AIC: 1436
```

2) Estimar manualmente el modelo para calcular la probabilidad de divorcio según se haya tenido relaciones premaritales o no con enlace logit.

```
> log(335/460)
[1] -0.317096
> log(159/82)-log(335/460)
[1] 0.9792809
> summary(mod1<-glm(Divorced~Premarital,divorce,family=binomial))
Call:
glm(formula = Divorced ~ Premarital, family = binomial, data = divorce)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.468  -1.046  -1.046   1.315   1.315

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.31710    0.07183  -4.415 1.01e-05 ***
PremaritalYes  0.97928    0.15376   6.369 1.91e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1434.0  on 1035  degrees of freedom
Residual deviance: 1391.4  on 1034  degrees of freedom
AIC: 1395.4
```

3) Estima manualmente el modelo para calcular la probabilidad de divorcio según se haya tenido relaciones extramatrimoniales o no usando el enlace probit.

```
> qnorm(396/(519+396))
[1] -0.1692834
> qnorm(98/(23+98))-qnorm(396/(519+396))
[1] 1.046875

> summary(mod2b<-glm(Divorced~Extramarital,divorce,family=binomial(link=probit)))

Call:
glm(formula = Divorced ~ Extramarital, family = binomial(link = probit),
    data = divorce)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.822  -1.065  -1.065   1.294   1.294

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.16928    0.04165  -4.064 4.81e-05 ***
ExtramaritalYes  1.04688    0.13784   7.595 3.08e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1434.0  on 1035  degrees of freedom
Residual deviance: 1369.6  on 1034  degrees of freedom
AIC: 1373.6
```

4) Valora si el efecto de las relaciones extramaritales en el hecho de estar divorciado es significativo. Justifica la respuesta indicando las hipótesis del test, los modelos asociados, el estadístico obtenido y el p-valor asociado.

Se aplica el test de razón de verosimilitudes. Se compara el modelo nulo (sin variables predictoras, deviancia 1434, grados de libertad 1035) con el modelo que incluye sólo la variable "Extramarital" (deviancia 1369.6, grados de libertad 1034). El estadístico obtenido es la diferencia de deviancias, 64.4 y la

distribución de referencia bajo la hipòtesis nula es una chi-quadrado con 1 grado de libertad. El p-valor es <0.0001.

```
> anova(mod,mod2,test="Chi")
Analysis of Deviance Table

Model 1: Divorced ~ 1
Model 2: Divorced ~ Extramarital
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1035     1434.0
2      1034     1369.6  1   64.407 1.012e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

El efecto es significativo ya que el test de la devianza indica que el modelo con la variable "extramarital" es estadísticamente diferente del modelo nulo

5) Para el modelo seleccionado, interpreta en todas las escalas posibles el efecto de tener relaciones premaritales en el hecho de haber solicitado el divorcio.

Para simplificar, se hace la interpretación del efecto principal sin considerar la interacción, que equivale a considerar el caso en que no se ha tenido relaciones premaritales.
En la escala del predictor lineal (log-odds), si se han tenido relaciones premaritales el predictor aumenta en 1.1 (manteniendo el resto de características iguales.
En la escala del odds, el hecho de tener relaciones premaritals supone un odds-ratio de 3 (=exp(1.1)) de estar divorciado respecto a no haberlas tenido. Es decir, supone un incremento del 200% (=100*(exp(1.1)-1)%)
En la escala de la probabilidad, supone aproximadamente un incremento de la probabilidad de 0.5232*0. 4768*1.1=0.27 unidades por el hecho de haver tenido relaciones premaritales.

6) Según estos datos y para el modelo seleccionado, calcula la probabilidad de que un hombre con relaciones premaritales y extramaritales esté divorciado

```
> predict(modsel,newdata=data.frame(Gender="Man",Premarital="Yes",Extramarital="Yes"),
type="response")
        1
0.7302651
```

7) A partir del modelo seleccionado, calcula un intervalo de confianza al 95% para el odds-ratio de estar divorciado si se han tenido relaciones extramaritales respecto a no haberlas tenido.

```
> exp(2.3960+qnorm(c(0.025,0.975))*0.3879)
[1]  5.133204 23.482842
```

8) Valora la calidad del modelo seleccionado en relación a la deviancia residual.

```
> 1-pchisq(1326.7,1031)
[1] 1.044286e-09
```

La deviancia residual es sensiblement superior a los grados de libertad. Para un modelo logistico con los datos desagregados, este resultado supone que el modelo no està bien explicado por las variables explicatives incluidas.

Además, la estimación del paràmetre de dispersión a partir de la deviancia residual y los grados de libertad da un valor aproximado de 1.3, aparentmente superior al valor teórico de 1 para el modelo binomial.

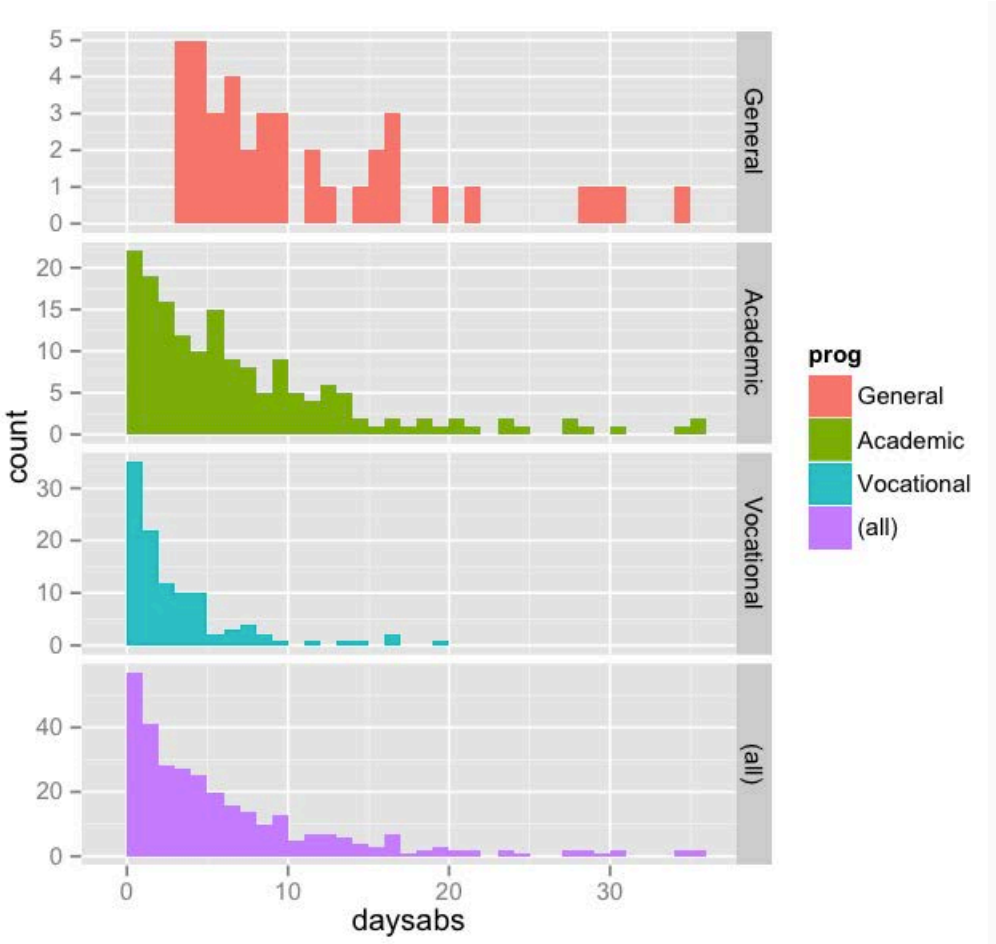
9) Para el modelo seleccionado, da una interpretación a la interacción entre haber tenido relaciones premaritales y extramaritales.

```
> exp(-1.7999)
[1] 0.1653154
> exp(1.0995)
[1] 3.002664
> exp(2.3960)
[1] 10.97917
> exp(1.0995+2.3960-1.7999)
[1] 5.449915
```

La significación de la interacción con un coeficiente negativo y teniendo en cuenta que los factores principales tienen coeficiente positivo, indica que el efecto aditivo se corrige a la baja. Es decir, si confluyen ambos factores la probabilidad de estar divorciado es menor de lo que corresponde a la adición de ambos efectos.

Problema 2 (4 Punts): Comptatges

Se dispone en el **IDRE** de UCLA (http://www.ats.ucla.edu/stat) de datos relativos a asistencia al instituto de 314 jóvenes pertenecientes a 2 institutos urbanos. La variable de respuesta es **Daysabs** (días de absentismo a clase) y las variables explicativas disponibles son el género (**gender**), la nota normalizada de matemáticas (**math**) y el programa educativo **prog** en que se matricula el estudiante (en 3 grupos -General, Académico, Vocacional). Responded a las siguientes preguntas argumentadamente.



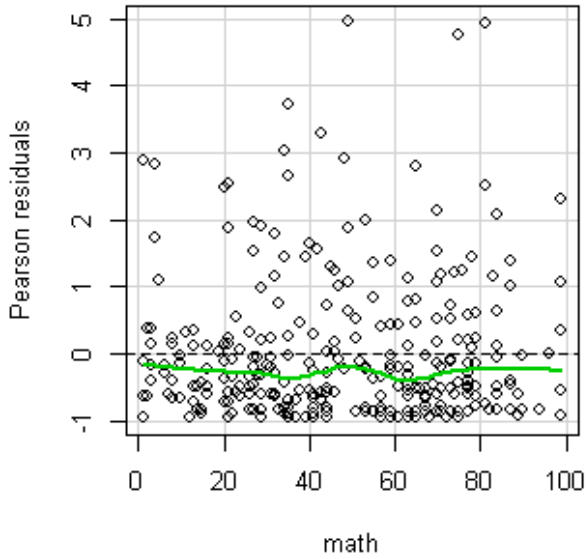

```
> summary(dat)
       id         gender         math          daysabs              prog
 1001   :  1   female:160   Min.   : 1.00   Min.   : 0.000   General   : 40
 1002   :  1   male  :154   1st Qu.:28.00   1st Qu.: 1.000   Academic  :167
 1003   :  1                Median :48.00   Median : 4.000   Vocational:107
 1004   :  1                Mean   :48.27   Mean   : 5.955
 1005   :  1                3rd Qu.:70.00   3rd Qu.: 8.000
 1006   :  1                Max.   :99.00   Max.   :35.000
 (Other):308
>
> summary(m1)

Call:
glm(formula = daysabs ~ math + prog + gender, family = "poisson",
    data = dat)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.7594786  0.0637731  43.270  < 2e-16 ***
math           -0.0069561  0.0009354  -7.437 1.03e-13 ***
progAcademic   -0.4260327  0.0567308  -7.510 5.92e-14 ***
progVocational -1.2707199  0.0779143 -16.309  < 2e-16 ***
```

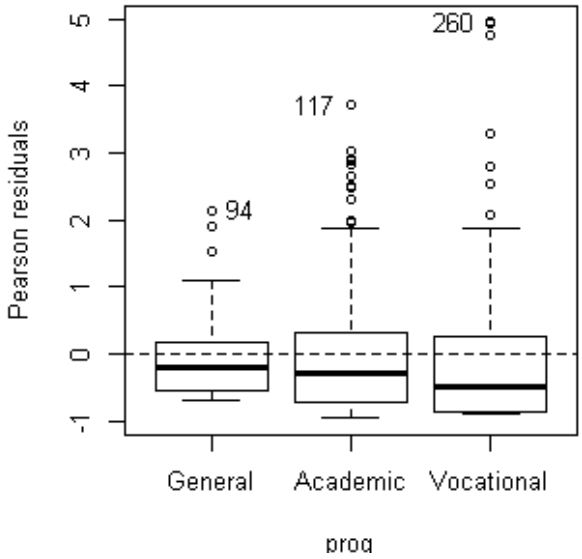
```
gendermale     -0.2424762  0.0467765  -5.184 2.18e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.7  on 313  degrees of freedom
Residual deviance: 1746.8  on 309  degrees of freedom
AIC: 2640.2

> Anova(m1)
Analysis of Deviance Table (Type II tests)

Response: daysabs
       LR Chisq Df Pr(>Chisq)
math     56.283  1  6.277e-14 ***
prog    295.690  2  < 2.2e-16 ***
gender   27.107  1  1.925e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> m2 <- glm(daysabs ~ math * ( prog + gender)+ prog * gender, family =
"poisson", data = dat)
> Anova(m2)
Analysis of Deviance Table (Type II tests)

Response: daysabs
            LR Chisq Df Pr(>Chisq)
math          56.557  1  5.459e-14 ***
prog         295.329  2  < 2.2e-16 ***
gender        27.985  1  1.223e-07 ***
math:prog      7.211  2    0.02717 *
math:gender    3.389  1    0.06564 .
prog:gender    6.767  2    0.03393 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> dispersiontest(m1,trafo=2)

        Overdispersion test

data:  m1
z = 6.6505, p-value = 1.46e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.7711299

> dispersiontest(m2,trafo=2)

        Overdispersion test

data:  m2
z = 7.0589, p-value = 8.39e-13
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.7561408

> summary(m2)

Call: glm(formula = daysabs ~ math * (prog + gender) + prog * gender,
    family = "poisson", data = dat)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           2.746489   0.097380  28.204  < 2e-16 ***
math                 -0.005474   0.001892  -2.894 0.003808 **
```

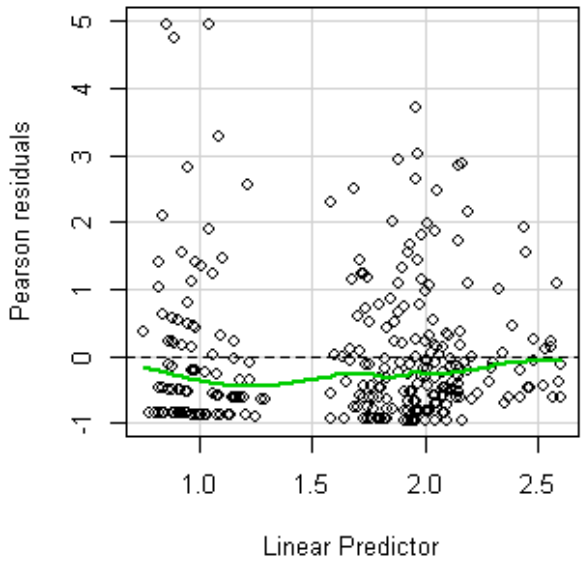
```
progAcademic              -0.416993   0.112031  -3.722 0.000198 ***
progVocational            -1.837542   0.207794  -8.843  < 2e-16 ***
gendermale                -0.227808   0.124978  -1.823 0.068335 .
math:progAcademic         -0.001241   0.002131  -0.582 0.560492
math:progVocational        0.006976   0.003321   2.101 0.035685 *
math:gendermale           -0.003511   0.001908  -1.840 0.065728 .
progAcademic:gendermale    0.099539   0.117203   0.849 0.395723
progVocational:gendermale  0.405121   0.160091   2.531 0.011388 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2217.7  on 313  degrees of freedom
Residual deviance: 1730.7  on 304  degrees of freedom
AIC: 2634
>
> m3 <- glm.nb(daysabs ~  math * ( prog + gender)+ prog * gender, data = dat)
> Anova(m3)
Analysis of Deviance Table (Type II tests)

Response: daysabs
            LR Chisq Df Pr(>Chisq)
math           6.383  1    0.01152 *
prog          47.874  2   4.02e-11 ***
gender         3.040  1    0.08123 .
math:prog      1.376  2    0.50252
math:gender    0.391  1    0.53168
prog:gender    1.381  2    0.50133
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(m3)

Call:
glm.nb(formula = daysabs ~ math * (prog + gender) + prog * gender,
    data = dat, init.theta = 1.060845607, link = log)


Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               2.7636609  0.3460635   7.986 1.39e-15 ***
math                     -0.0058472  0.0061409  -0.952 0.341006
progAcademic             -0.4552564  0.3763696  -1.210 0.226433
progVocational           -1.8357972  0.4785126  -3.836 0.000125 ***
gendermale               -0.2490624  0.3937225  -0.633 0.527006
math:progAcademic        -0.0004016  0.0065367  -0.061 0.951010
math:progVocational       0.0070179  0.0077861   0.901 0.367407
math:gendermale          -0.0031173  0.0050076  -0.623 0.533602
progAcademic:gendermale   0.1117728  0.3637728   0.307 0.758645
progVocational:gendermale 0.4030523  0.4003348   1.007 0.314037
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.0608) family taken to be 1)

    Null deviance: 435.50  on 313  degrees of freedom
Residual deviance: 358.87  on 304  degrees of freedom
AIC: 1747.3

              Theta:  1.061
          Std. Err.:  0.110

 2 x log-likelihood:  -1725.286
> logLik(m2)
'log Lik.' -1307.01 (df=10)
> logLik(m3)
```

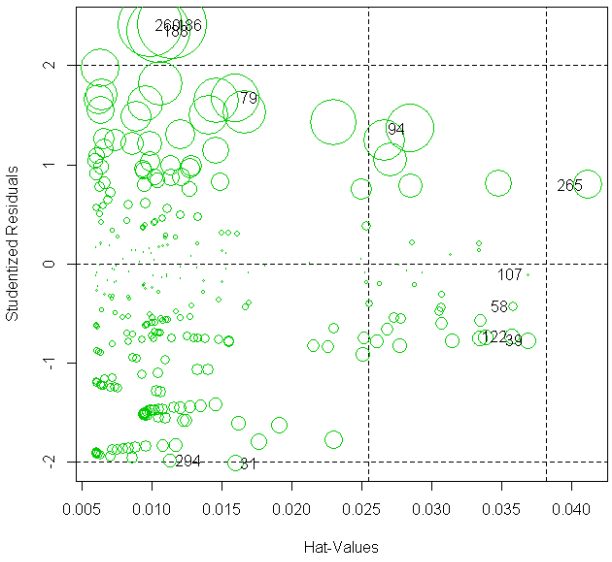
```
'log Lik.' -862.6429 (df=11)
> m4 <- glm.nb(daysabs ~  math + ( prog ), data = dat)
> summary(m4)

Call:
glm.nb(formula = daysabs ~ math + (prog), data = dat, init.theta = 1.032713156,
    link = log)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.615265   0.197460  13.245  < 2e-16 ***
math           -0.005993   0.002505  -2.392   0.0167 *
progAcademic   -0.440760   0.182610  -2.414   0.0158 *
progVocational -1.278651   0.200720  -6.370 1.89e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.0327) family taken to be 1)

    Null deviance: 427.54  on 313  degrees of freedom
Residual deviance: 358.52  on 310  degrees of freedom
AIC: 1741.3

              Theta:  1.033
          Std. Err.:  0.106

 2 x log-likelihood:  -1731.258
>
> anova(m4,m3)
Likelihood ratio tests of Negative Binomial Models

Response: daysabs
                                     Model    theta Resid. df    2 x log-lik.
1                           math + (prog) 1.032713       310       -1731.258
2 math * (prog + gender) + prog * gender 1.060846       304       -1725.286
    Test    df LR stat.   Pr(Chi)
1
2 1 vs 2     6 5.972043 0.4263291
> logLik(m4)
'log Lik.' -865.6289 (df=5)
> m5<-glm(daysabs ~  math + ( prog ),family=neg.bin(1.0327),data = dat)
> summary(m5)

Call:
glm(formula = daysabs ~ math + (prog), family = neg.bin(1.0327),
    data = dat)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.615264   0.206757  12.649  < 2e-16 ***
math           -0.005993   0.002623  -2.285   0.0230 *
progAcademic   -0.440760   0.191208  -2.305   0.0218 *
progVocational -1.278650   0.210170  -6.084 3.45e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial family taken to be 1.096364)

    Null deviance: 427.54  on 313  degrees of freedom
Residual deviance: 358.52  on 310  degrees of freedom
AIC: 1739.3

> sum(resid(m4,type="pearson")^2)
[1] 339.8771
> sum(resid(m2,type="pearson")^2)
[1] 1961.478
```

```
> residualPlots(m5)
     Test stat Pr(>|t|)
math     0.283    0.595
prog        NA       NA
```










```
 > influencePlot(m5, labels=row.names(dat),id.n=5, col=3)
      StudRes        Hat      CookD
31  -2.0169568 0.01595749 0.05891289
39  -0.7764599 0.03689353 0.05897027
58  -0.4362228 0.03584191 0.03711326
79   1.6729647 0.01595749 0.17686665
94   1.3642395 0.02848060 0.17778442
107 -0.1125886 0.03689353 0.01091051
122 -0.7405747 0.03579274 0.05609456
186  2.4112252 0.01143427 0.25496203
188  2.3477770 0.01045154 0.23471314
260  2.4126195 0.00985926 0.23733565
265  0.7958090 0.04120000 0.10775617
294 -1.9891253 0.01133823 0.04925039
```

1. Se propone un modelo de respuesta poissoniana aditivo donde se explica la respuesta Daysabs a partir del resto de variables disponibles (gender, math y prog). Determinar si los efectos netos son estadísticamente significativos y ordenarlos de más a menos importante.

A jutjar pels resultats de la comanda Anova(m1), on es fa el test d'hipòtesi dels efectes nets de les 3 variables explicatives del model additiu, totes són significatives al llindar habitual. La importancia es pot quantificar a partir del pvalor de la H0 : Efecte net no significatiu. Com més petit sigui el pvalor més evidencia per rebutjar la H0 i per tant més importancia del factor/covariant. De més a menys importants es té: prog, math i gender.

```
> Anova(m1)
Analysis of Deviance Table (Type II tests)

Response: daysabs
       LR Chisq Df Pr(>Chisq)
math     56.283  1  6.277e-14 ***
prog    295.690  2  < 2.2e-16 ***
gender   27.107  1  1.925e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Resulta necesario introducir las interacciones dobles entre las variables explicativas para mejorar la explicabilidad del modelo?

A jutjar pels resultats de la comanda Anova(m2), on es fa el test d'hipòtesis de les interaccions dobles, totes són significatives al llindar habitual condicionat a trobar-se les altres al model. Val a dir que la interacció entre math i gender té un efecte net amb un pvalor del 6% técnicament per sobre del llindar habitual, però tant just que no es considera convenient menystenir aquesta interacció doble.

```
> Anova(m2)
Analysis of Deviance Table (Type II tests)

Response: daysabs
            LR Chisq Df Pr(>Chisq)
math          56.557  1  5.459e-14 ***
prog         295.329  2  < 2.2e-16 ***
gender        27.985  1  1.223e-07 ***
math:prog      7.211  2    0.02717 *
math:gender    3.389  1    0.06564 .
prog:gender    6.767  2    0.03393 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. En base a las salidas disponibles, pensáis que los datos muestran sobredispersión? Resulta significativamente distinto de cero el parámetro según la propuesta R NB2 - $h(\mu_i)=\mu_i^2 \rightarrow V[Y_i|X_i]=\mu_i+\alpha\mu_i^2=(1+\alpha\mu_i)\mu_i$?

Clarament la H0: $\alpha = 0$ mostra evidència per ser rebutjada amb un p valor de pràcticament 0.

```
> dispersiontest(m1,trafo=2)

        Overdispersion test data:  m1
z = 6.6505, p-value = 1.46e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.7711299
```

4. Indicad en el mejor modelo de respuesta poissoniana cual sería el número esperado de ausencias para una estudiante mujer matriculada en un programa General y con una calificación estandarizada en matemáticas en la media. Cuál es la probabilidad que un individuo que satisface las condiciones anteriores tenga más de 3 ausencias en un curso?

Calculad un intervalo de confianza al 95% para la anterior predicción si pensáis que disponéis de todos los datos para hacerlo y si no es así argumentad qué falta.

*Clarament el millor model és m2 i log(μ)=2.746489-0.005474*48.27=2.48 i per tant exp(2.48)=11.968 absències.*
La probabilitat que tingui més de 3 absències és 1-P(Y≤3)= 1-ppois(lambda=11.968,3)= 0.9976509
No es pot calcular un interval de confiança directament doncs no es dona informació sobre la covariança del termes independent i de l'estimador de la pendent associada a la covariant: Var(η+α+(β+(αβ)+(γβ))x+γ+(αγ)) pel grup de referencia de prog i gender queda Var(η+(β)x)=V(η)+x²V(β)+2xCOV(η,β), però el terme de la covariança és desconegut en la sortida disponible.

5. Se tantea una propuesta binomial negativa utilizando el método adhoc glm.nb() del paquete MASS. Valorar el mejor modelo según la opción binomial negativa y los resultados disponibles.

Clarament és m4 (o m5, que és el mateix model però calculat amb el mètode genèric glm()), només conté la covariant math i el factor programa de matriculació prog.

```
> summary(m4)

Call:
glm.nb(formula = daysabs ~ math + (prog), data = dat, init.theta =
1.032713156,
    link = log)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.615265   0.197460  13.245  < 2e-16 ***
math           -0.005993   0.002505  -2.392   0.0167 *
progAcademic   -0.440760   0.182610  -2.414   0.0158 *
progVocational -1.278651   0.200720  -6.370 1.89e-10 ***
---
```

6. De acuerdo con el goodness of fit realizado a partir del estadístico de Pearson generalizado, que propuesta resulta más adecuada para modelar los datos: la propuesta Poisson o la binomial negativa?

H0: m2 (poisson) consistent, amb l'estadístic de Pearson generalitzat que va 339.87 distribuït segons una Chi quadrat amb 310 graus de llibertat té un pvalor 0
H0': m4 (binomial negatiu) consistent té un pvalor de 0.11 i per tant s'accepta la consistencia de les dades amb el model.

```
> sum(resid(m4,type="pearson")^2)
[1] 339.8771
> sum(resid(m2,type="pearson")^2)
[1] 1961.478
> 1-pchisq(1961.48,310)
[1] 0
> 1-pchisq(339.88,310)
[1] 0.1170121
>
```

7. Interpretar el efecto del programada de matriculación **prog** en el modelo binomial negativo **m5** disponible.

Matricular-se en un programa Academic en comptes de General suposa un decrement del logarisme de les absències per curs de 0.441 dies dins del mateix

valor de nota math. O bé un decrement del 35.66% de les absències anuals respecte la referencia.
Matricular-se en un programa Vocacional en comptes de General suposa un decrement del logarisme de les absències per curs de 1.28 dies dins del mateix valor de nota math. O bé un decrement del 72% de les absències anuals respecte la referencia (prog=General).

8. Valorar los residuos del modelo **m5: se detectan outliers en los residuos, hay observaciones influyentes a priori y a posteriori.**

Efectivament hi ha outliers dels residus de Pearson per moltes 'campanes' que es situen majorment en les categories de programa Academic i Vocacional. Les observacions 31, 186, 188, 260 tenen residus estudentitzats generalitzats majors que 2 en valor absolut. Valors d'anclatge generalitzat superior a 3p/n només es dona en la observació 265 (bubble plot). Donat el residu elevat les observacions 186, 188, 260 tenen distàncies de Cook notablement majors que la resta d'observacions i formen part d'un conglomerat de punts, caldria una anàlisi més detallada, però molt possiblement són observacions que no es capturen convenientment en el model i cal suprimir-les o buscar més variables explicatives.

```
186  2.4112252 0.01143427 0.25496203
188  2.3477770 0.01045154 0.23471314
260  2.4126195 0.00985926 0.23733565
```

Problema 3 (2 punts): Recomptes

Les dades sobre les que recau el present examen han estat extretes de la pàgina Web d'StatLib (http://lib.stat.cmu.edu) i pertanyen a 500 nens irlandesos de 11 anys, són dades del 1967, remeses per Greaney i Kelleghan (1984), del St. Patrick's College, Dublin. Les variables mostren:

1. GENERE: 1=HOME; 2=DONA.
2. DVRTEST_SCORE (Drumcondra Verbal Reasoning Test Score).
3. EDUCACIÓ - Educational level attained:

1	Primary terminal leaver
2	Junior cycle incomplete: vocational school
3	Junior cycle incomplete: secondary school
4	Junior cycle terminal leaver: vocational school
5	Junior cycle terminal leaver: secondary school
6	Senior cycle incomplete: vocational school
7	Senior cycle incomplete: secondary school
8	Senior cycle terminal leaver: vocational school
9	Senior cycle terminal leaver: secondary school
10	3rd level incomplete
11	3rd level complete

4. CERTIFICAT - Leaving Certificate. 1 if Leaving Certificate not taken; 2 if taken.
5. PRESTIGI_PARE - Prestige score for father's occupation
   (calculated by Raftery and Hout, 1985). 0 if missing.
6. ESCOLA - Type of school: 1=secondary; 2=vocational; 9=primary terminal leaver.

Les dades han estat filtrades i preprocessades de la següent manera, quedant 435 observacions de les originals:

1. Es suprimeixen les observacions amb nivell d'educació primari (variable 3 EDUCACIÓ, nivell 1).

2. Es suprimeixen les observacions amb missings d'EDUCACIÓ o PRESTIGI_PARE (els missings estan codificats amb 0).
3. Es creen variables agregades per EDUCACIÓ, PRESTIGI_PARE i DVRTEST_SCORE, anomenades G_EDUCACIÓ, G_PRESTIGE i G_DVRTEST. G_EDUCACIÓ, codificada amb 4 pels codis 2-5, 9 pels codis 6-9 i 11 pels codis 10-11. G_PRESTIGE agrupa en 4 categories definides pels quartils, els valors originals de PRESTIGI_PARE, de manera que els quartils 28, 37 , 46 i 75 (màxim) constitueixen els representants de classe. G_DVRTEST agrupa DVRTEST_SCORE en 4 categories definides pels quartils (i el màxim) 91, 102, 111 i 140.

```
MTB > Table 'G_EDUCACIÓ' 'G_PRESTIGE'  'G_DVRTEST';

Control: G_DVRTES = 91
 Rows: G_EDUCAC     Columns: G_PRESTI
           28       37       46       75      All

 4         28       26       10       12       76
 9          8       16        6        7       37
 11         0        0        0        1        1
 All       36       42       16       20      114

 Control: G_DVRTES = 102
 Rows: G_EDUCAC     Columns: G_PRESTI
           28       37       46       75      All

 4         28       23       14        6       71
 9          3       12       10        9       34
 11         0        3        2        3        8
 All       31       38       26       18      113

 Control: G_DVRTES = 111
 Rows: G_EDUCAC     Columns: G_PRESTI
           28       37       46       75      All

 4         12        9        8        7       36
 9         12        9       13       19       53
 11         1        4        3        5       13
 All       25       22       24       31      102

 Control: G_DVRTES = 140
 Rows: G_EDUCAC     Columns: G_PRESTI
           28       37       46       75      All

 4          9        4        1        6       20
 9          8       13       11       16       48
 11         5        9        7       17       38
 All       22       26       19       39      106

  Cell Contents --
                 Count
```

Amb ajut d'un paquet estadístic s'estimen els següents models log-lineals:

MODEL	Deviança	Paràmetres ? *p*	Graus de llibertat ? desagregats *N-p*	Graus de llibertat ? Agregació a 3 factors *n-p*
A+B+C	173.4376	9	435-9	39
A+B*C	147.9923	18	435-18	30
A*C+B*C	44.7311	24	...	24
A*B+A*C+B*C	14.3250	30	...	18
A*B*C	0	48	435-48	0

1. Ompleneu les dades que falten a la taula parcial d'anàlisi de la deviança il.lustrada (llevat darrera columna). Sigui A el factor nivell d'educació dels nens, B factor del nivell de prestigi professional dels pares i C factor del nivell del test DVR. Si les dades estiguessin agregades de

manera que només es consideressin les classes de les covariants definides per A, B i C, quins valors haurien d'apareixer a la darrera columna?

Per mantenir l'equivalència entre la proposta d'estudi com a model log-lineal actual i com a model de regressió logística amb resposta politòmica, cal tenir en compta totes les 48 cel.les (classes de la covariable), sense descomptar les 4 cel.les amb 0 observacions.

2. Contrasteu la hipòtesi d'independència complerta entre les 3 variables que defineixen la taula de contingència de dimensió 3.

El model additiu en l'escala definida pel logaritme del nombre d'observacions a cada cel.la (categoria creuada), 4x4x3=48 cel.les en total, representa el model d'independència total de les 3 variables. Té per deviança 173.43 deixant 39 graus de llibertat, per tant el p-valor és $P\left(\chi^2_{39} > 173.43\right) \approx 0.0$ *el que indica que el model no és estadísticament bo i per tant, la hipòtesi d'independència total no s'acceptaria doncs no s'adapta bé a les dades observades.*

3. Són consistents les dades amb la hipòtesi que el nivell d'educació dels nens és independent de les altres 2 variables?

*Per contrastar la hipòtesi d'indepedència per blocs de l'educació i les altres 2 variables, s'hauria de considerar el valor de l'estadístic deviança obtingut després d'adjustar el model log-lineal amb tots els efectes principals més la interacció del factor prestigi de la feina dels pares (B) i resultats del test DVR (C). La deviança del model A+B*C és de 147.9923 i deixa 30 graus de llibertat, per tant el p_valor torna a ser 0,* $P\left(\chi^2_{30} > 147.9923\right) \approx 0.0$ *, i la hipòtesi nul.la quedaria rebutjada doncs no s'adapta a les dades.*

4. Contrasteu la hipòtesi que condicionat al factor DVR, el nivell educatiu dels nens és independent del prestigi de la feina dels pares.

*Per contrastar la hipòtesi d'independència condicional, cal considerar els resultats de l'estadístic deviança després d'adjustar el model log-lineal que conté els efectes principals dels 3 factors i les interaccions de l'educació amb el prestigi de la feina dels pares i del prestigi amb els resultats del test DVR, és a dir, A*C més B*C. La deviança del model A*C +B*C és de 44.7311 i deixa 24 graus de llibertat, per tant el p_valor és* $P\left(\chi^2_{24} > 44.7311\right) = 0.0063$ *, i la hipòtesi nul.la quedaria rebutjada doncs no s'adapta a les dades.*

5. Hi ha alguna evidència estadística per afirmar que l'associació entre el Test DVR i l'educació dels nens depen del prestigi de la feina dels pares ?

*Per respondre a la qüestió cal estimar el model log-lineal d'associació constant, és a dir, el que conté tots 3 efectes principals més totes les 3 interaccions d'ordre 2 i contrastar la deviança. La deviança del model A*B+B*C+A*C és de 14.325 i deixa a18 graus de llibertat, per tant el p_valor és* $P\left(\chi^2_{18} > 14.325\right) = 0.7077$ *, i la hipòtesi nul.la quedaria acceptada. La hipòtesi d'associació constant, estadísticament s'adapta bé a les dades.*