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

CURS 2014-2015 Q1 – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: a les 15:00h Aula -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, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	3h 00 min
Sortida de notes:	Abans del 26 de Gener al Web Docent de MLGz
Revisió de l'examen:	26 de Gener a 10h a C5-217-C Nord o H- P6-6

Problema 1 (4.5 punts): Resposta Binària

Se estudiará la incidencia **anual** de algún **siniestro en una aseguradora de automóviles.** Datos extraídos del artículo '*Exponential Bonus-Malus Systems Integrating a priori Risk Classification' (2000)*. El riesgo de siniestralidad de los asegurados depende de su edad, género, ocupación, uso del vehículo, color del vehículo, etc. El sistema de *bonus-malus* considera en el cálculo de la tarifa anual, el número de siniestros anteriores reportados por los asegurados. En el artículo se propone modelar el número de siniestros en función de una agrupación de la edad de los asegurados Factor Edad (<36, 36 a 49, >49) y del Factor Potencia, donde la potencia del vehiculo asegurado se ha categorizado en 4 niveles: <54, 54 a 75, 76 a 118 y >118.

Edad	Potencia	Vehículos asegurados	Declaran algún Siniestro	Prob Declarar algún Siniestros
<36	**<54**	3945	736	0,1866
36-49	**<54**	9023	1418	0,1571
>50	**<54**	11758	1509	0,1283
<36	**54-75**	11947	3208	0,2685
36-49	**54-75**	25719	5862	0,2279
>50	**54-75**	27287	5420	0,1986
<36	**76-118**	8447	2527	0,2992
36-49	**76-118**	19609	4953	0,2526
>50	**76-118**	18688	4459	0,2386
<36	**>119**	1486	478	0,3217
36-49	**>119**	5762	1640	0,2846
>50	**>119**	5812	1443	0,2483
		149483	**33653**	

1. Determinar la tabla de datos agregados necesaria para la estimación del modelo de respuesta binaria para la probabilidad de declarar algún siniestro con el único efecto de la Edad del tomador del seguro. ¿Cuál es la probabilidad de declarar algún siniestro que marginalmente corresponde a cada asegurado?

Edad	Declaran algún Siniestro (resposta positiva)	Vehículos asegurados	Probabilitat
<36	**6949**	**25825**	**0.269**
36-49	**13873**	**60113**	**0.231**
>50	**12831**	**63545**	**0.202**
	33653	**149483**	**0.225**

P('Declarar Algun Sinistre)=33653/149483= 0,2251

2. Estimad manualmente a partir de la tabla del punto anterior y empleando la transformación ***logit*** cual es el estimador del término constante en el modelo nulo.

$$\text{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad \rightarrow \hat{\eta}=\ln\left(\frac{33653/149483}{1-33653/149483}\right)=\ln\left(\frac{0.2251}{1-0.2251}\right)=\ln(0.2905)=-1.236$$

3. Estimad manualmente a partir de la tabla del punto anterior y empleando la transformación ***probit*** cual es el estimador del término constante en el modelo nulo.

$$\text{probit}(\pi_i)=qnorm(\pi_i,0,1)=\boldsymbol{\eta} \quad \rightarrow \hat{\boldsymbol{\eta}}=qnorm(0.2251,0,1)=-0.7551$$

```
> qnorm(0.2251,0,1)
[1] -0.7550816
```

4. Estimad manualmente a partir de la tabla del punto anterior y empleando la transformación ***logit*** cuáles son los estimadores de la constante y de los coeficiente de las *dummies* para el efecto bruto de la edad en el modelo que incluye exclusivamente el factor Edad (nivel de referencia i=1 ≡'<36').

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta+\alpha_i \quad i=1,..3 \quad \alpha_{1\equiv<36}=0 \quad \rightarrow$$

$$\hat{\eta}=\log\left(\frac{0.269}{1-0.269}\right)=\text{-0.99929}$$

$$\hat{\alpha}_{2\equiv 36-49}=\log\left(\frac{0.231}{1-0.231}\right)-\log\left(\frac{0.269}{1-0.269}\right)=\text{-0.20461}$$

$$\hat{\alpha}_{4\equiv>50}=\log\left(\frac{0.202}{1-0.202}\right)-\log\left(\frac{0.269}{1-0.269}\right)=\text{-0.37504}$$

5. Interpretar en la escala del predictor lineal y aproximadamente en la escala de la probabilidad el efecto de la Edad sobre la declaración de algún siniestro según el modelo calculado en el punto 4.

L'efecte del factor EDAT en el logit de la incidència de sinistres de trànsit és un decrement de 0.205 unitats en el grup de EDAT de "36-49" i un decrement de 0.375 unitats en el grup de Potència de ">50", tots respecte el grup de referència (<36).

En l'escala de la probabilitat, aproximadament, la probabilitat de declarar algun sinistre en el grup de "36-49" es redueixen en $0.2251(1-0.2251)(-0.205)=0.036$ *unitat respecte el nivell de referència del grup d'edat més jove (<36). La probabilitat de declarar algun sinistre en el grup de ">50" es redueixen en* $0.2251(1-0.2251)$ *(-0.375)=0.065 respecte el nivell de referència del grup d'edat més jove (<36).*

6. Calcular el odds-ratio de los grupos de Edad sobre la incidencia de siniestros de tráfico. Interpretar el efecto de la Edad en la escala de los odds de la probabilidad de declarar algún siniestro según el modelo calculado en el punto 4.

```
> m1<-glm(cbind(y,n-y)  ~ f.edat, family = binomial, data = bm)
> summary(m1)
Call:
glm(formula = cbind(y, n - y) ~ f.edat, family = binomial, data = bm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.99929    0.01403  -71.22   <2e-16 ***
f.edat36-49 -0.20461    0.01705  -12.00   <2e-16 ***
f.edat50+   -0.37504    0.01716  -21.85   <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: 1799.7  on 11  degrees of freedom
Residual deviance: 1313.4  on  9  degrees of freedom
AIC: 1430.3
> coef(m1)
(Intercept) f.edat36-49   f.edat50+
 -0.9992935  -0.2046072  -0.3750444
> exp(coef(m1))   # Interpretació en l'escala dels odds
(Intercept) f.edat36-49   f.edat50+
  0.3681394   0.8149674   0.6872588
> (1-exp(coef(m1)))*100
(Intercept) f.edat36-49   f.edat50+
   63.18606    18.50326    31.27412
```

Els odds de patir un sinistre es redueixen en 18.5% en el grup de 36-49 anys respecte els odds del grup de referència <36 anys. Els odds de patir un sinistre es redueixen en un 31.3% en el grup de majors de 50 anys, respecte el grup d'edat dels més jovens (<36).

7. ¿Hay alguna evidencia estadística para afirmar que el efecto de la Potencia está relacionado con la incidencia de Siniestros? ¿Cuántos son los grados de libertad del estadístico de referencia para la realización del contraste?

```
> m2<-glm(cbind(y,n-y)  ~ f.pot,family = binomial, data = bm)
> anova(m0,m2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ 1
Model 2: cbind(y, n - y) ~ f.pot
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        11    1799.68
2         8     473.67  3     1326 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si, efectivament el contrast d'equivalència entre el model nul i el model amb l'efecte brut de la potència, mitjançant la diferència de deviances per l'estadístic de la Chi quadrat amb 3 g.ll. El p valor és pràcticament 0 i per tant, es rebutja la hipòtesi nul.la: la potència del vehicle assegurat

8. ¿Hay alguna evidencia estadística para afirmar que el efecto bruto de la Edad afecta a la incidencia de Siniestros? ¿Cuántos son los grados de libertad del estadístico de referencia para la realización del contraste?

```
> m1<-glm(cbind(y,n-y)  ~ f.edat, family = binomial, data = bm)
> m0<-glm( cbind(y,n-y) ~ 1, family = binomial, data = bm)
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ 1
Model 2: cbind(y, n - y) ~ f.edat
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        11     1799.7
2         9     1313.4  2   486.31 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

L'efecte brut de l'edat és estadísticament significatiu, l'estadístic de referència és una Chi quadrat amb 2 g.ll.

9. Determinar si el efecto NETO de la Potencia es estadísticamente significativo, cuando la Edad ya se encuentra en el modelo.

```
> anova(m2,m3,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ f.edat
Model 2: cbind(y, n - y) ~ f.pot + f.edat
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         9    1313.37
2         6      18.15  3   1295.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Si que ho és!

10. Existe evidencia para afirmar que los efectos de la Potencia no son los mismos según los distintos grupos de la Edad en la incidencia de Siniestros por tráfico?

```
> m3<-glm(cbind(y,n-y)  ~ f.pot+f.edat, family = binomial, data = bm)
> m4<-glm(cbind(y,n-y)  ~ f.pot*f.edat, family = binomial, data = bm)
> anova(m3,m4,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(y, n - y) ~ f.pot + f.edat
Model 2: cbind(y, n - y) ~ f.pot * f.edat
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         6     18.154
2         0      0.000  6   18.154  0.005858 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Demana si la relació entre el factor EDAT i la incidència de sinistres depèn del factor potència del vehicle assegurat, per tant, és un contrast entre el model complert i el model additiu, el qual té un pvalor inferior al 5% habitual i per tant, es rebutja la hipòtesi nul.la: SI, la relació entre el factor Edat i la incidència de sinistres depen dels nivells de la potència del vehicle. L'estadístic de referència és una Chi quadrat amb 6 graus de llibertat. Es pot afegir que atenen als pvalors dels coeficients de les dummies per les interaccions possiblement no totes las interaccions entre edat i potència són rellevants.

RESULTATS R

```
> summary(bm)
       m               y          f.edat     f.pot
 Min.   : 1486   Min.   : 478   <36  :4   <54   :3
 1st Qu.: 5800   1st Qu.:1437   36-49:4   54-75 :3
 Median :10390   Median :2084   50+  :4   76-118:3
 Mean   :12457   Mean   :2804             119+  :3
 3rd Qu.:18918   3rd Qu.:4582
 Max.   :27287   Max.   :5862

> summary(m1)

Call:
glm(formula = I(y/m) ~ f.pot, family = binomial, data = bm, weights = m)

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

```
(Intercept) -1.74924    0.01790  -97.71   <2e-16 ***
f.pot54-75   0.50145    0.02023   24.79   <2e-16 ***
f.pot76-118  0.67928    0.02081   32.65   <2e-16 ***
f.pot119+    0.76809    0.02658   28.90   <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: 1799.68  on 11  degrees of freedom
Residual deviance:  473.67  on  8  degrees of freedom
AIC: 592.61

Number of Fisher Scoring iterations: 3

> summary(m2)

Call:
glm(formula = I(y/m) ~ f.edat, family = binomial, data = bm,     weights = m)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)            0.01403  -71.22   <2e-16 ***
f.edat36-49            0.01705  -12.00   <2e-16 ***
f.edat50+              0.01716  -21.85   <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: 1799.7  on 11  degrees of freedom
Residual deviance: 1313.4  on  9  degrees of freedom
AIC: 1430.3

> summary(m3)

Call:
glm(formula = I(y/m) ~ f.pot + f.edat, family = binomial, data = bm,     weights = m)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.49991    0.02201  -68.15   <2e-16 ***
f.pot54-75   0.48897    0.02027   24.13   <2e-16 ***
f.pot76-118  0.66519    0.02085   31.90   <2e-16 ***
f.pot119+    0.77654    0.02664   29.15   <2e-16 ***
f.edat36-49 -0.21909    0.01715  -12.78   <2e-16 ***
f.edat50+   -0.36891    0.01726  -21.37   <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: 1799.681  on 11  degrees of freedom
Residual deviance:   18.154  on  6  degrees of freedom
AIC: 141.09

> summary(m4)
Call:glm(formula = I(y/m) ~ f.pot * f.edat, family = binomial, data = bm,     weights =
m)

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

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.47248    0.04087 -36.029  < 2e-16 ***
```

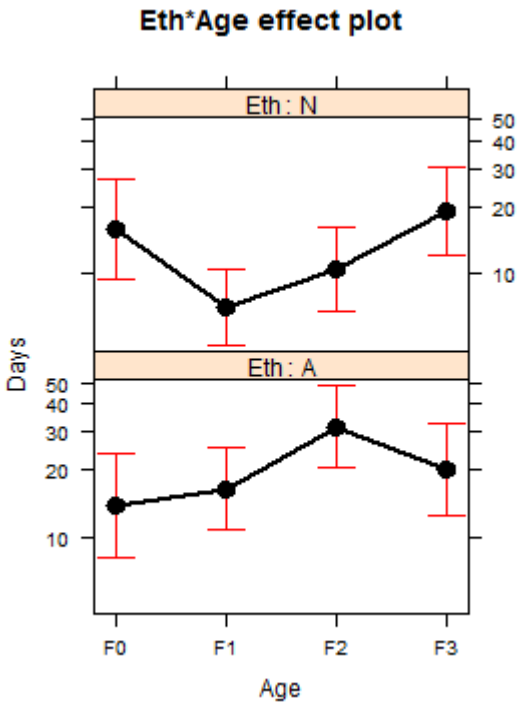
```
f.pot54-75                0.47034    0.04579  10.272  < 2e-16 ***
f.pot76-118               0.62118    0.04728  13.140  < 2e-16 ***
f.pot119+                 0.72637    0.06895  10.534  < 2e-16 ***
f.edat36-49              -0.20707    0.05007  -4.136 3.54e-05 ***
f.edat50+                -0.44325    0.04930  -8.991  < 2e-16 ***
f.pot54-75:f.edat36-49   -0.01084    0.05616  -0.193   0.8469
f.pot76-118:f.edat36-49  -0.02648    0.05781  -0.458   0.6469
f.pot119+:f.edat36-49     0.03154    0.08027   0.393   0.6943
f.pot54-75:f.edat50+      0.05051    0.05556   0.909   0.3633
f.pot76-118:f.edat50+     0.13419    0.05736   2.340   0.0193 *
f.pot119+:f.edat50+       0.08155    0.08023   1.016   0.3094
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:  1.7997e+03  on 11  degrees of freedom
Residual deviance: -2.5060e-12  on  0  degrees of freedom
AIC: 134.94
> anova(m0,m1,test="Chis")
Analysis of Deviance Table

Model 1: I(y/m) ~ 1
Model 2: I(y/m) ~ f.pot
  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1        11    1799.68
2         8     473.67  3  1326.01 3.349e-287
> anova(m0,m2,test="Chis")
Analysis of Deviance Table

Model 1: I(y/m) ~ 1
Model 2: I(y/m) ~ f.edat
  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1        11    1799.68
2         9    1313.37  2   486.31 2.508e-106
> anova(m1,m3,test="Chis")
Analysis of Deviance Table

Model 1: I(y/m) ~ f.pot
Model 2: I(y/m) ~ f.pot + f.edat
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         8     473.67
2         6      18.15  2   455.52 1.216e-99
> anova(m2,m3,test="Chis")
Analysis of Deviance Table

Model 1: I(y/m) ~ f.edat
Model 2: I(y/m) ~ f.pot + f.edat
  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1         9    1313.37
2         6      18.15  3  1295.22 1.605e-280
> anova(m3,m4,test="Chis")
Analysis of Deviance Table

Model 1: I(y/m) ~ f.pot + f.edat
Model 2: I(y/m) ~ f.pot * f.edat
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         6    18.1543
2         0 -2.506e-12  6  18.1543    0.0059
>
```

Problema 2 (4.5 puntos): Comptatges

Se analizan los datos de Aitkin (1978) relacionados con un estudio sociológico sobre comportamiento escolar de niños blancos y niños aborígenes en New South Wales, Australia. La muestra disponible contempla 146 niños de final del primer año de primaria y de los primeros 3 cursos de secundaria (notados 0 a 3). Los niños están clasificados como de rendimiento bajo o medio. La variable de respuesta a modelar es el número de días con falta de asistencia a la escuela durante el año escolar (Days).

Los datos pueden obtenerse directamente en R de la librería MASS con el comando quine. El resumen de los datos se muestra a continuación:

- Eth es "A" para aborígenes y "N" no aborígenes (mayoritariamente niños blancos).
- Sex es "M" para niños y "F" para niñas.
- Age es el grupo de edad "F0" (Primaria) "F1" a "F3" para los cursos de secundaria
- Lrn es "SL" para niños con rendimiento escolar bajo y "AL" para niños con rendi miento escolar medio.




```
> summary(df)
 Eth     Sex     Age      Lrn           Days             nage
 A:69    F:80    F0:27    AL:83    Min.   : 0.00    Min.   :0.000
 N:77    M:66    F1:46    SL:63    1st Qu.: 5.00    1st Qu.:1.000
                 F2:40             Median :11.00    Median :1.500
                 F3:33             Mean   :16.46    Mean   :1.541
                                   3rd Qu.:22.75    3rd Qu.:2.000
                                   Max.   :81.00    Max.   :3.000
```

Se calculan los modelos de Poisson que contienen los factores de etnia, género, rendimiento escolar y edad. A la edad se le da dos tratamientos distintos: como factor y como covariable con los valores numéricos 0 a 3 según corresponda a F0 a F3 (variable *nage*).

1. Interpretar los coeficientes del modelo Poisson sin interacciones con el tratamiento lineal y numérico de la edad.

```
> summary(m2)
Call:
glm(formula = Days ~ Eth + Sex + Lrn + nage, family = poisson,    data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.48676    0.06153  40.418  < 2e-16 ***
EthN        -0.55117    0.04184 -13.174  < 2e-16 ***
SexM         0.22562    0.04159   5.425 5.81e-08 ***
LrnSL        0.26616    0.04457   5.971 2.35e-09 ***
nage         0.20947    0.02182   9.598  < 2e-16 ***
> exp(coefficients(m2))
(Intercept)        EthN        SexM       LrnSL        nage
 12.0222121   0.5762759   1.2531048   1.3049410   1.2330197
```

El logaritme del nombre esperat d'absències es redueix en 0.55 unitats per individus d'ètnia no aborigen respecte el grup de referencia aborigen i dins dels mateixos valors de la resta de variables del model.

El logaritme del nombre esperat d'absències s'incrementa en 0.225 unitats pels nois respecte el grup de referencia noies i dins dels mateixos valors de la resta de variables del model.

El logaritme del nombre esperat d'absències s'incrementa en 0.27 unitats pels estudiant de baix rendiment respecte el grup de referència de rendiment normal i dins dels mateixos valors de la resta de variables del model.

El logaritme del nombre esperat d'absències s'incrementa en 0.21 unitats per cada unitat de grup d'edat respecte el grup anterior normal i dins dels mateixos valors de la resta de variables del model.

El nombre esperat d'absències es decrementa en 100(1-0.58)=42% en individus d'ètnia no aborigen respecte el grup de referencia aborigen i dins dels mateixos valors de la resta de variables del model.*

El nombre esperat d'absències d'absències s'incrementa en 25% pels nois respecte el grup de referencia noies i dins dels mateixos valors de la resta de variables del model.

El nombre esperat d'absències d'absències s'incrementa en 30% pels estudiant de baix rendiment respecte el grup de referència de rendiment normal i dins dels mateixos valors de la resta de variables del model.

El nombre esperat d'absències d'absències s'incrementa en un 23% per cada unitat d'increment del grup d'edat respecte el grup anterior normal i dins dels mateixos valors de la resta de variables del model.

2. Argumentad el tratamiento más adecuado más adecuado para la edad en el modelo Poisson cuando no se utilizan las interacciones, como factor o como covariante lineal (o hasta qué orden).

Els resultats disponibles amb només efectes principals indiquen que el tractament numèric de l'edat requereix del terme quadràtic, però encara així l'AIC (2332) resulta superior a l'AIC del tractament com factor (2299) en els models sense interaccions. En considerar les interaccions entre l'ètnia i el grup d'edat com a covariant amb els termes lineal i quadràtic, el model facilita unes interaccions significatives i l'AIC del model AIC: 2190.1 és menor que l'obtingut amb la proposta del factor Age sense interaccions. En canvi si s'usa Age com a factor i les interaccions amb l'ètnia el model esdevé millor segons el criteri d'Akaike.

```
> AIC(m31)
[1] 2151.314
```

3. Argumentad si existe evidencia en la propuesta Poisson que el efecto de la etnia en los días de absentismo escolar varíe con el grupo de edad.

Tant en el model de tractament com a factor del grup d'edat com en el model quadràtic de la covariant nage, les interaccions a l'ètnia són estadísticament significatives segons els resultats del mètode Anova() per testar els efectes nets de les interaccions.

```
> Anova(m31)
Analysis of Deviance Table (Type II tests)
Response: Days
        LR Chisq Df Pr(>Chisq)
Eth      166.845  1  < 2.2e-16 ***
Age      168.324  3  < 2.2e-16 ***
Sex       14.509  1  0.0001395 ***
Lrn       48.084  1  4.084e-12 ***
Eth:Age  153.870  3  < 2.2e-16 ***
---
> Anova(m3)
Analysis of Deviance Table (Type II tests)
Response: Days
                 LR Chisq Df Pr(>Chisq)
Eth               177.766  1  < 2.2e-16 ***
poly(nage, 2)     113.325  2  < 2.2e-16 ***
Sex                28.042  1  1.187e-07 ***
Lrn                55.106  1  1.142e-13 ***
Eth:poly(nage, 2)  166.129  2  < 2.2e-16 ***
```

4. El modelo aditivo con los factores género, etnia, rendimiento y el término lineal del grupo de edad es consistente con los datos?

La deviança residual del model indicat és:

```
Residual deviance: 1768.6  on 141  degrees of freedom
```

El test de Goodness of Fit: H0-El model és consistent amb les dades, tindria un pvalor = P(X2(141)>1768.6) < 2.2e-16, per tant hi ha evidència per rebutjar la hipòtesi nul·la i per tant el model no explica bé les dades.

Amb dades desagregades, a nivell d'individu la distribució assimptòtica dels estadístics de goodness-of –fit (deviance o X2 Pearson Generalitzat) com a una Chi quadrat amb els g.ll. del model no es cumpleix. Cal valorar l'ordre de magnitud dels estadístics i els graus de llibertat, el primer milers i el segon no arriba a 2 centenes, per tant no són dels mateix ordre, el que indica que el model no és consistent amb les dades.

5. Examinad los residuos del modelo Poisson con los factores género, etnia, rendimiento y los términos lineal y cuadrático del grupo de edad y las interacciones entre el grupo de edad y la etnia. Detectad la presencia de outliers en los residuos y valores influyentes.

Clarament hi ha residus d'Student molt elevats que poden interpretar-se com si tinguessin aproximadament una distribució normal Standard i per tant valors per sobre

2-3 en valor absolut són grans i indiquen un manca d'ajust. Les observacions 36, 59, 70, 72, 104 i 126. Hi ha desajust, outliers dels residus. Els valors influents venen per la distància de Cook inusual: la obs 72 té un valor gran i per tant, és una dada influent, combina un anclatge elevat amb desajust.

```
36   9.0898034 0.05950895 1.10684829
59   8.3858825 0.05047801 0.82984001
70  -4.0794165 0.12393631 0.46707918
72   8.7829805 0.12393631 1.41955217
104  8.3939528 0.07503030 1.03588547
126  6.9557881 0.02803630 0.58070148
```

6. juzgar por los resultados del test de dispersión de la librería AER, pensaís que existe sobredispersión en los datos que requieren un tratamiento de la respuesta según un modelo probabilístico alternativo. Sugerid una relación plausible entre la varianza y la media.

El millor model disponible conté la interacció entre ètnia i el factor edat (m31).
Es disposa de l'estadístic de Pearson generalitzat pel model m3, 1661.66, que permet dividint pels graus de llibertat, 138, estimar la sobredispersió del model ϕ=12.04. No s'ha estudiat si hi ha interaccions addicionals a la inclosa (nage amb ètnia) que podessin millorar el model, amb la qual cosa no hi ha garantia que la sobredispersió no vingui causada per un model deficient (incomplert). Estrictament interpretant el test de dispersió clarament es rebutja la hipòtesi nul.la i es pot afirmar que la variança és significativament superior a la mitjana segons el Model NB2 - $h(\mu_i)=\mu_i^2 \rightarrow V[Y_i|X_i]=\mu_i+\alpha\mu_i^2=(1+\alpha\mu_i)\mu_i$.

```
> dispersiontest(m3,trafo=2)
Overdispersion testdata:  m3
z = 6.1671, p-value = 3.479e-10
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0.5712361
```

Se recalculan los modelos con/sin interacciones etnia y edad, para los dos casos con tratamiento Age como factor y con tratamiento *nage* con los términos lineal y cuadrático bajo la hipótesis de respuesta binomial negativa.

7. Qué modelo os parece más adecuado en base a los resultados disponibles? Argumentad las variables que son estadísticamente significativas bajo la propuesta binomial negativa.

Els models presentats estimats amb el mètode glm() després de fixar el paràmetre theta=1.348 corresponent a la binomial negativa segons s'havia estimat al mètode específic glm.nb() de la llibreria MASS indiquen que el terme quadràtic de la covariant grup d'edat (nage) i la seva interacció amb ètnia són estadísticament significatius (p valor =0.0006243), que les interaccions entre el tractament quadràtic de la covariant grup d'edat (nage) i l'ètnia són estadísticament significatives (p valor =0.00412) i que en el tractament només lineal de la covariant nage, la

interacció amb l'ètnia no és estadísticament significativa (p valor =0.702). Per tant, sense disposar de més contrastos i examinant els pvalors de la taula de resultats del model m5:

```
glm(formula = Days ~ Eth * poly(nage, 2) + Sex + Lrn, family = neg.bin(1.348),
data = df)
```

No inclòs en els resultats finals subministrats:

```
> Anova(m5,test="F")
Analysis of Deviance Table (Type II tests)

Response: Days
                     SS  Df       F    Pr(>F)
Eth                15.465   1 15.7842 0.0001139 ***
poly(nage, 2)       9.058   2  4.6222 0.0114025 *
Sex                 1.487   1  1.5179 0.2200317
Lrn                 1.888   1  1.9274 0.1672794
Eth:poly(nage, 2)  11.202   2  5.7164 0.0041197 **
Residuals         135.211 138
```

***El gènere i el rendiment escolar han deixat de ser significatives**, per tant, el model caldria recalcular-lo considerant els efectes principals de l'ètnia, la covariant nage amb termes lineal i quadràtic i les interaccions entre ètnia-nage. Clarament el tractament binomial negatiu és molt superior la deviança residual i l'AIC disminueixen.*

```
 > summary(m60)
Call:
glm.nb(formula = Days ~ Eth * poly(nage, 2), data = df, init.theta = 1.314634409,
    link = log)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.0208     0.1085  27.846  < 2e-16 ***
EthN                -0.5846     0.1512  -3.867 0.000110 ***
poly(nage, 2)1       2.2409     1.3079   1.713 0.086644 .
poly(nage, 2)2      -2.0817     1.3080  -1.591 0.111505
EthN:poly(nage, 2)1 -0.4986     1.8171  -0.274 0.783774
EthN:poly(nage, 2)2  6.3012     1.8235   3.456 0.000549 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.3146) family taken to be 1)

    Null deviance: 200.26  on 145  degrees of freedom
Residual deviance: 167.83  on 140  degrees of freedom
AIC: 1102.9

              Theta:  1.315
          Std. Err.:  0.167

 2 x log-likelihood:  -1088.901
```

8. Examinad los residuos del modelo binomial negativo disponible. Detectad la presencia de outliers en los residuos de manera comparativa con la propuesta Poisson y valorad la presencia de valores influyentes.

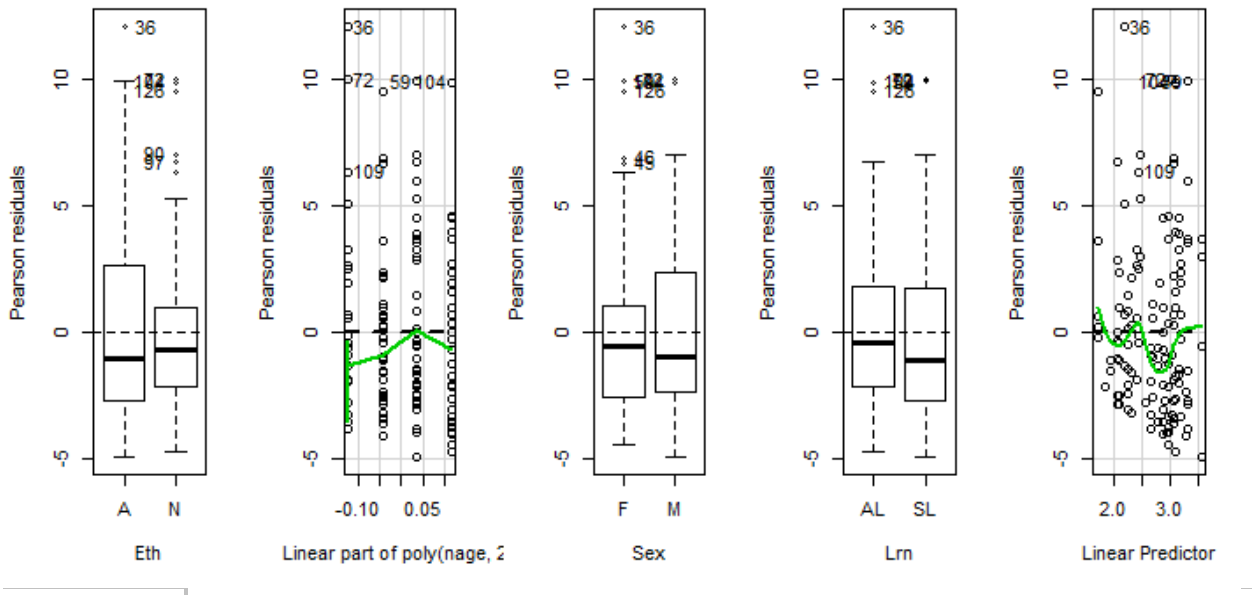
Els residus han millorat sensiblement. Els residus d'Student pel glm() m5 estan en el rang [-2,2] majorment, h'hi ha per sobre 2 en valor absolut, per l'ordre de magnitud és molt menor al que es trobava amb la proposta Poisson. Per tant, no hi ha pràcticament desajust en la proposta de tractament binomial negativa. Ara bé, el model no es pot donar per vàlid en aparèixer observacions amb una

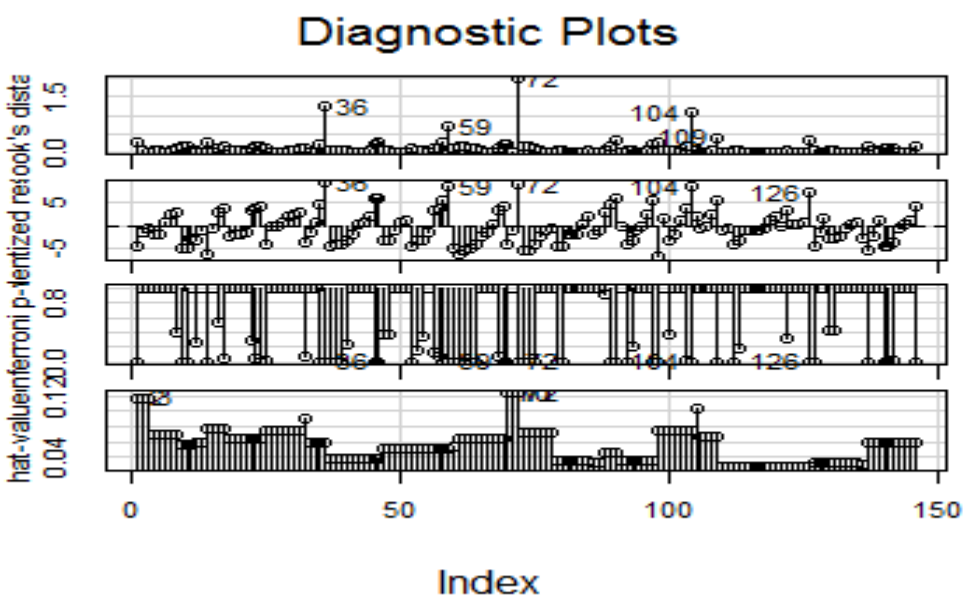
distància de Cook sospitosament gran i per tant, són observacions influents que cal examinar una a una (36, 72).

RESULTADOS PARA PROBLEMA 2

```
> m0<-glm(Days~1,family=poisson,data=df)
> m1<-glm(Days~Eth+Sex+Lrn+Age,family=poisson,data=df)
> summary(m1)

Call:
glm(formula = Days ~ Eth + Sex + Lrn + Age, family = poisson,     data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.71538    0.06468  41.980  < 2e-16 ***
EthN        -0.53360    0.04188 -12.740  < 2e-16 ***
SexM         0.16160    0.04253   3.799 0.000145 ***
LrnSL        0.34894    0.05204   6.705 2.02e-11 ***
AgeF1       -0.33390    0.07009  -4.764 1.90e-06 ***
AgeF2        0.25783    0.06242   4.131 3.62e-05 ***
AgeF3        0.42769    0.06769   6.319 2.64e-10 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1696.7  on 139  degrees of freedom
AIC: 2299.2
> Anova(m1)
Analysis of Deviance Table (Type II tests)
Response: Days
    LR Chisq Df Pr(>Chisq)
Eth  166.845  1  < 2.2e-16 ***
Sex   14.404  1  0.0001475 ***
Lrn   45.798  1  1.311e-11 ***
Age  168.324  3  < 2.2e-16 ***
> m2<-glm(Days~Eth+Sex+Lrn+nage,family=poisson,data=df)
> summary(m2)
Call:
glm(formula = Days ~ Eth + Sex + Lrn + nage, family = poisson,    data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.48676    0.06153  40.418  < 2e-16 ***
EthN        -0.55117    0.04184 -13.174  < 2e-16 ***
SexM         0.22562    0.04159   5.425 5.81e-08 ***
LrnSL        0.26616    0.04457   5.971 2.35e-09 ***
nage         0.20947    0.02182   9.598  < 2e-16 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1768.6  on 141  degrees of freedom
AIC: 2367.1
> Anova(m2)
Analysis of Deviance Table (Type II tests)
Response: Days
     LR Chisq Df Pr(>Chisq)
Eth   178.484  1  < 2.2e-16 ***
Sex    29.350  1  6.042e-08 ***
Lrn    35.676  1  2.330e-09 ***
nage   96.385  1  < 2.2e-16 ***
> m20<-glm(Days~Eth+Sex+Lrn+poly(nage,2),family=poisson,data=df)
> Anova(m20,test="Wald")
Analysis of Deviance Table (Type II tests)
Response: Days
               Df   Chisq Pr(>Chisq)
Eth             1 172.850  < 2.2e-16 ***
Sex             1  28.781  8.103e-08 ***
Lrn             1  51.086  8.840e-13 ***
```

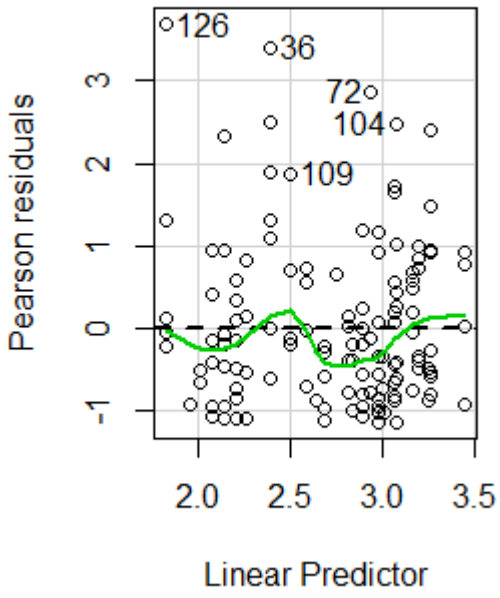
```
poly(nage, 2)   2 111.858  < 2.2e-16 ***
Residuals     140
> AIC(m20)
[1] 2352.183
> m21<-glm(Days~Eth+Sex+Lrn+nage+I((nage-1.541)^2),family=poisson,data=df)
> summary(m21)
Call:glm(formula = Days ~ Eth + Sex + Lrn + nage + I((nage - 1.541)^2),family = poisson,
data = df)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.31782    0.07483  30.973  < 2e-16 ***
EthN               -0.55005    0.04184 -13.147  < 2e-16 ***
SexM                0.22298    0.04156   5.365 8.10e-08 ***
LrnSL               0.37641    0.05266   7.147 8.84e-13 ***
nage                0.21534    0.02139  10.069  < 2e-16 ***
I((nage - 1.541)^2) 0.10237    0.02490   4.112 3.93e-05 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1751.7  on 140  degrees of freedom
AIC: 2352.2

> m3<-glm(Days~Eth*poly(nage,2)+Sex+Lrn,family=poisson,data=df)
> sum(resid(m3,type="pearson")^2)
[1] 1661.656

> summary(m3)
Call:glm(formula = Days ~ Eth * poly(nage, 2) + Sex + Lrn, family = poisson,
    data = df)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         2.73587    0.04439  61.633  < 2e-16 ***
EthN               -0.58323    0.04413 -13.217  < 2e-16 ***
poly(nage, 2)1      3.14489    0.36726   8.563  < 2e-16 ***
poly(nage, 2)2     -1.43773    0.37154  -3.870 0.000109 ***
SexM                0.22116    0.04171   5.302 1.14e-07 ***
LrnSL               0.38583    0.05257   7.340 2.14e-13 ***
EthN:poly(nage, 2)1 -0.85116    0.49824  -1.708 0.087574 .
EthN:poly(nage, 2)2  6.62600    0.52103  12.717  < 2e-16 ***
---
 (Dispersion parameter for poisson family taken to be 1)
    Null deviance: 2073.5  on 145  degrees of freedom
Residual deviance: 1585.6  on 138  degrees of freedom
AIC: 2190.1
> m31<-glm(Days~Eth*Age+Sex+Lrn,family=poisson,data=df)
> AIC(m31)
[1] 2151.314
> Anova(m31)
Analysis of Deviance Table (Type II tests)
Response: Days
        LR Chisq Df Pr(>Chisq)
Eth      166.845  1  < 2.2e-16 ***
Age      168.324  3  < 2.2e-16 ***
Sex       14.509  1  0.0001395 ***
Lrn       48.084  1  4.084e-12 ***
Eth:Age  153.870  3  < 2.2e-16 ***
---
> Anova(m3)
Analysis of Deviance Table (Type II tests)
Response: Days
                  LR Chisq Df Pr(>Chisq)
Eth                177.766  1  < 2.2e-16 ***
poly(nage, 2)      113.325  2  < 2.2e-16 ***
Sex                 28.042  1  1.187e-07 ***
Lrn                 55.106  1  1.142e-13 ***
Eth:poly(nage, 2)  166.129  2  < 2.2e-16 ***
> residualPlots(m3, layout=c(1,5),id.method=abs(cooks.distance(m3)), id.n=5)
> influenceIndexPlot(m3, id.n=5)
```



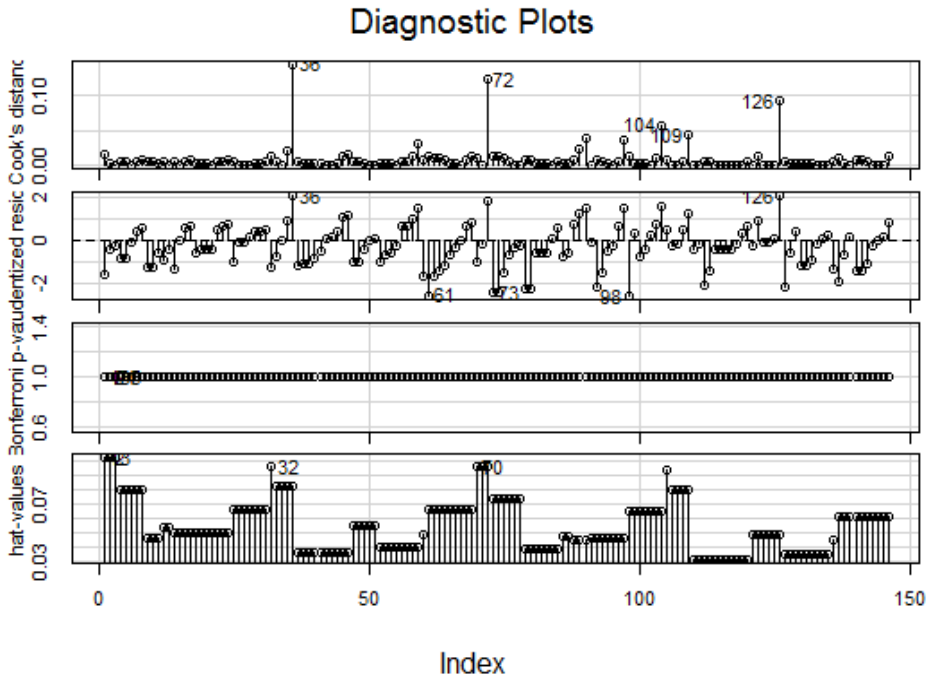



```
> m4<-glm.nb(Days~Eth*poly(nage,2)+Sex+Lrn,data=df)
> summary(m4)
Call:glm.nb(formula = Days ~ Eth * poly(nage, 2) + Sex + Lrn, data = df,
    init.theta = 1.348458226, link = log)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.8278     0.1542  18.343  < 2e-16 ***
EthN                -0.5997     0.1495  -4.010 6.07e-05 ***
poly(nage, 2)1       2.3142     1.3192   1.754 0.079395 .
poly(nage, 2)2      -1.3279     1.3877  -0.957 0.338608
SexM                 0.1862     0.1526   1.220 0.222461
LrnSL                0.2493     0.1819   1.370 0.170592
EthN:poly(nage, 2)1 -0.2879     1.7976  -0.160 0.872751
EthN:poly(nage, 2)2  6.0637     1.8032   3.363 0.000772 ***
---
 (Dispersion parameter for Negative Binomial(1.3485) family taken to be 1)
    Null deviance: 204.46  on 145  degrees of freedom
Residual deviance: 168.02  on 138  degrees of freedom
AIC: 1103.7
              Theta:  1.348
          Std. Err.:  0.173
```

```
 2 x log-likelihood:  -1085.696

> m5<-glm(Days~Eth*poly(nage,2)+Sex+Lrn,family=neg.bin( 1.348),data=df)
> m51<-glm(Days~Eth*nage+Sex+Lrn,family=neg.bin( 1.348),data=df)
> m52<-glm(Days~Eth+poly(nage,2)+Sex+Lrn,family=neg.bin( 1.348),data=df)
> m53<-glm(Days~Eth+nage+Sex+Lrn,family=neg.bin( 1.348),data=df)
> summary(m5)
Call:
glm(formula = Days ~ Eth * poly(nage, 2) + Sex + Lrn, family =
neg.bin(1.348), data = df)
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.8278     0.1526  18.528  < 2e-16 ***
EthN                 -0.5997     0.1481  -4.051 8.49e-05 ***
poly(nage, 2)1        2.3144     1.3060   1.772 0.078590 .
poly(nage, 2)2       -1.3280     1.3739  -0.967 0.335414
SexM                  0.1862     0.1511   1.233 0.219855
LrnSL                 0.2493     0.1801   1.384 0.168513
EthN:poly(nage, 2)1  -0.2879     1.7796  -0.162 0.871700
EthN:poly(nage, 2)2   6.0638     1.7852   3.397 0.000891 ***
---
(Dispersion parameter for Negative Binomial family taken to be 0.9797957)

    Null deviance: 204.40  on 145  degrees of freedom
Residual deviance: 167.98  on 138  degrees of freedom
AIC: 1101.7
> anova(m51,m5,test="F")
Analysis of Deviance Table
Model 1: Days ~ Eth * nage + Sex + Lrn
Model 2: Days ~ Eth * poly(nage, 2) + Sex + Lrn
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1       140     183.24
2       138     167.97  2   15.261 7.7879 0.0006243 ***
---
> anova(m52,m5,test="F")
Analysis of Deviance Table
Model 1: Days ~ Eth + poly(nage, 2) + Sex + Lrn
Model 2: Days ~ Eth * poly(nage, 2) + Sex + Lrn
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1       140     179.18
2       138     167.97  2   11.202 5.7164 0.00412 **
---
> anova(m53,m51,test="F")
Analysis of Deviance Table
Model 1: Days ~ Eth + nage + Sex + Lrn
Model 2: Days ~ Eth * nage + Sex + Lrn
  Resid. Df Resid. Dev Df Deviance     F Pr(>F)
1       141     183.41
2       140     183.24  1  0.17203 0.147  0.702
> influencePlot(m5, id.n=5)
       StudRes         Hat       CookD
1   -1.5624439 0.10193136 0.12517530
2   -0.4231653 0.10193136 0.04991986
3   -0.1969005 0.10193136 0.02483471
32  -1.2248299 0.09623272 0.10702406
36   2.0956867 0.08286051 0.38067932
61  -2.5329958 0.06600363 0.11035261
70  -0.9866280 0.09584889 0.09350977
72   1.8647295 0.09584889 0.35045161
73  -2.4025216 0.07325738 0.11591334
74  -2.4025216 0.07325738 0.11591334
79  -2.1953175 0.03809878 0.07881737
98  -2.5743935 0.06452888 0.10925435
104  1.5992194 0.06452888 0.23960376
109  1.3090105 0.08014561 0.20667372
```



```
126  2.1242517 0.04847070 0.3038485
```



Problema 3 (1 punto): Modelización

Para las siguientes situaciones, indica el tipo de modelo que usarías, es decir, modelo lineal o generalizado, cuál sería la respuesta y su distribución, qué variables explicativas incluirías y si usarías un modelo mixto o no. En caso de utilizar un modelo mixto, indica qué variable determina la agrupación en la muestra.

1. Ingresos por eventos: Se organizan 24 ferias de diferente tipo (culturales, comerciales y de ocio) en cinco poblaciones. Se toma nota de características del público asistente (número de asistentes, edad media, sexo mayoritario), climatología (precipitaciones, temperatura, viento) así como de la recaudación del acto. Se desea identificar los factores relacionados con una mayor recaudación.

2. Inventario de cetáceos: Se seleccionan 10 puntos en el mediterráneo y 10 en el atlántico y durante 3 semanas se cuentan los avistamientos diarios de delfines, obteniendo 21 recuentos en cada punto. Se recogen datos de condiciones marítimas y climatológicas.

3. Resistencia a la rotura: Se preparan 30 muestras de un material y se someten aleatoriamente a tres tipos de tratamiento de interés (A, B y C). Se recoge la fuerza necesaria para romper la pieza.

4. Venta de inmuebles: En una inmobiliaria se dispone de información de 60 inmuebles con datos de sus características principales: superficie, planta, ascensor (s/n), calefacción(s/n),… y del precio asignado. Algunos de ellos se han vendido antes de los 6 meses y se desea conocer que factores determinan que un inmueble se venda antes de ese plazo.

5. Urgencias hospitalarias: Durante un año se registra el número de urgencias registradas en fin de semana en los tres hospitales que hay en una población. Se dispone de características del hospital, datos climatológicos e indicadores de brotes y epidemias de gripe. Se desea comparar los factores que determinan la asistencia a cada hospital.

1. Modelo lineal mixto con repuesta gaussiana (recaudación) y factor aleatorio de agrupación (población) a 5 niveles. Posibles factores fijos: tipo de feria, características del público y datos climatológicos.
2. Modelo lineal mixto generalizado con respuesta Poisson (número de delfines) y factor aleatorio (puntos geográficos) a 20 niveles. Corresponde a datos longitudinales (a lo largo de 21 días) y las zonas (Mediterráneo y Atlántico) pueden ser considerados fijos o aleatorios. Las covariables formarían parte del predictor lineal.
3. Modelo lineal con respuesta gaussiana (fuerza para rotura). No son datos agrupados, ya que el tratamiento es un factor fijo en el predictor lineal.
4. Modelo lineal generalizado con respuesta binaria, logística (venta antes de los 6 meses). Las características del inmueble constituyen los predictores en la parte fija del modelo.
5. Modelo lineal generalizado con respuesta Poisson. Los hospitales son los que hay en la población e interesa compararlos con lo que no es un factor aleatorio y se incluiría en la parte fija del predictor lineal, junto con el resto de covariables.