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

CURS 2015-2016 Q1 – EXAMEN FINAL : MODEL LINEAL GENERALITZAT

(Data: 19 de Gener 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:	2h 30 min
Sortida de notes:	Abans del 1 de Febrer al Web Docent de MLGz
Revisió de l'examen:	1 de Febrer a 9:30h a C5-217-C Nord o 10h a H- P6-67

Problema 1 (4.5 punts): Resposta Binària

Les dades contingudes a la matriu de dades en format kenya98cuse.dat procedeixen del *1998 Demographic and Health Survey for Kenya* i han estat publicades electrònicament per Germán Rodríguez. Les dades consisteixen en 7881 observacions de 14 variables de dones i estan adreçades a ***determinar el mètode anticonceptiu usat***.

Les variables contingudes a l'arxiu de dades subministrat són:

C1. Age: Edat.
C2. Tipus de lloc de residència (f.residence): 1-Urbà 2-Rural.
C3. Nivell d'Educació (f.educ): 0 Cap 1 Primària 2 Secundària 3 Superior.
C4. Nivell de Lectura (f.read): 1-Llegeix fàcilment 2-Llegeix amb dificultat 3-No sap llegir 9-*missing*).
C5. Llegeix Diaris (f.rnewsp): 0-No 1-Si 9-*missing*.
C6. Veu TV ? (f.seeTV): 0-No 1-Si 9-*missing*.
C7. Escolta Radio ? (f.lisradio): 0-No 1-Si 9-*missing*.
C8. Té Electricitat? (f.elec): 0-No 1-Si 9-*missing*.
C9. Té Radio ? (f.radio): 0-No 1-Si 9-*missing*.
C10. Té TV ? (f.TV): 0-No 1-Si 9-*missing*.
C11. Anys d'Estudi (f.yeduc): entre 0 i 19.
C12. Nombre de Fills (nsons).
C13. Embaraçada actualment? (f.pregnant): 1-Si 0-No o no està segura.
C14. Mètode anticonceptiu emprat (f.cuse): 0-Cap 1-Píldora 2-DIU 3-Injecció 4-Diafragma 5-Condó 6-Esterilització Dona 7- Esterilització Home 8 –Abstinència 9 – Abstinència Parcial 10- Altres 11- Norplant Abstinència 2-
C15. A crear, una agrupació dels mètodes anticonceptius en (f.bcuse): Cap (0) o Algun (1).

Les dades no estan depurades originàriament i s'emprarà el mètode de supressió de les observacions que presentin *missings*. Les variables afectades són les associades a disponibilitat i ús de radio i TV. Es suprimeixen també totes les observacions de dones embarassades. S'acaben obtenint 7097 observacions després del procés de depuració proposat (les dades mancants d'altres variables no afecten a la finalitat del present estudi).

1. Es vol estudiar la relació entre el target (**f.bcuse**) - Us d'alguna mesura contraceptiva (resposta positiva) i el nivell d'educació (**f.educ**). Formuleu i calculeu el model logit que modela una probabilitat idèntica d'ús d'alguna mesura de contracepció per tots els grups definits pel factor nivell d'educació (**f.educ**). Useu les dades de la taula mostrada a continuació.

```
> table(df$f.educ,df$f.bcuse)
              target.no target.yes
  fe.none           744        178
  fe.primary       3015       1188
  fe.secondary     1092        745
  fe.high            61         74
```

Heu d'estimar el model nul:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad i=1:4$$

La probabilitat marginal d'ús d'alguna mesura contraceptiva (resposta positiva) és:0.3079 i els odds 2185 a 4912 (o 0.445 a 1), per tant el logodd és -0.81, l'estimador dde la constant en el model nul.

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

 target.no target.yes
      4912       2185
> tt0<-prop.table(table(df$f.bcuse));tt0

 target.no target.yes
 0.6921234  0.3078766
> tt0[2]
target.yes
 0.3078766
> eta<-log(2185/4912);eta
[1] -0.8100654
> m0<-glm(df$f.bcuse~1, family=binomial, data=df)
> summary(m0)

Call: glm(formula = df$f.bcuse ~ 1, family = binomial, data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.81007    0.02571   -31.5   <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: 8763.2  on 7096  degrees of freedom
Residual deviance: 8763.2  on 7096  degrees of freedom
AIC: 8765.2
```

2. Es vol estudiar la relació entre el target **(f.bcuse)** - Us d'alguna mesura contraceptiva (resposta positiva) i el nivell d'educació **(f.educ)**. Formuleu i calculeu el model logit que modela una probabilitat d'ús d'alguna mesura de contracepció específica pels grups definits pel factor nivell d'educació **(f.educ).** Useu les dades de la taula mostrada anteriorment.

Heu d'estimar a partir de la taula el model Y-A on Y és el target f.bcuse i el factor A és f.educ. És pel nivell d'agregació mostrat per la taula un model saturat i per això podeu estimar-lo directament doncs serà un model que reprodueix exactament les observacions. Sigui fe.none el nivell de referència, per tant el logodd d'aquest grup constitueix l'estimador de la constant en el model Y-A, log(178/744)=-1.43:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \qquad i = 1:4 \qquad \alpha_{1\equiv none} = 0$$

Les diferències dels logodds de la resta de grups respecte el nivell de referència constituiran els estimadors dels efectes principals de cada grups:

log(1188/3015)- log(178/744)=0.499
log(745/1092)- log(178/744)=1.048
log(74/61)- log(178/744)=1.623

```
> tt<-table(df$f.educ,df$f.bcuse);tt

              target.no target.yes
  fe.none           744        178
```

```
  fe.primary        3015       1188
  fe.secondary      1092        745
  fe.high             61         74
> lodd<-log(tt[,2]/tt[,1]);lodd
     fe.none   fe.primary fe.secondary      fe.high
  -1.4302575   -0.9313286   -0.3823819    0.1931912
> lodd[2:4]-lodd[1]
  fe.primary fe.secondary      fe.high
   0.4989289    1.0478755    1.6234487
> m1<-glm(df$f.bcuse~f.educ, family=binomial, data=df)
> summary(m1)

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

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.43026    0.08344 -17.141  < 2e-16 ***
f.educfe.primary    0.49893    0.09020   5.532 3.17e-08 ***
f.educfe.secondary  1.04788    0.09602  10.913  < 2e-16 ***
f.educfe.high       1.62345    0.19201   8.455  < 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: 8763.2  on 7096  degrees of freedom
Residual deviance: 8576.6  on 7093  degrees of freedom
AIC: 8584.6
```

3. Calculeu el nombre predit d'observacions que usarien i no usarien cap mesura contraceptiva per cadascun dels nivells del factor educació (**f.educ**) sota la hipòtesi del model nul descrit al **Punt 1.**

Consisteix en aplicar la probabilitat marginal de resposta positiva 0.3079 al total d'observacions de cada grup per obtenir la predicció del nombre de respostes positives (ús contracepció) en cada grup. Aplicant la probabilitat complementària al total de grup s'obtenen les respostes negatives predites en cada grup pel model nul.

```
> tt<-table(df$f.educ,df$f.bcuse);tt

                target.no target.yes
  fe.none             744        178
  fe.primary         3015       1188
  fe.secondary       1092        745
  fe.high              61         74
> trow<-apply(tt,1,sum);trow
     fe.none   fe.primary fe.secondary      fe.high
         922         4203         1837          135
> tt0<-prop.table(table(df$f.bcuse));tt0

 target.no target.yes
 0.6921234  0.3078766
> tt0[2]
target.yes
 0.3078766
> fitm0<-round(cbind(trow*(tt0[1]),trow*tt0[2]),dig=2);fitm0
                [,1]    [,2]
fe.none       638.14  283.86
fe.primary   2908.99 1294.01
fe.secondary 1271.43  565.57
fe.high        93.44   41.56
```

4. Calculeu la deviança del model nul descrit en el punt 1 usant les dades calculades per les prediccions en el punt 3.

$$D = 2\sum_{i=1,4}\left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right)\right\} =$$

$$= 2\left(\begin{array}{l} 178\log\left(\frac{178}{283.86}\right) + 744\log\left(\frac{744}{638.14}\right) + 1188\log\left(\frac{1188}{1294.01}\right) + 3015\log\left(\frac{3015}{2908.99}\right) + 745\log\left(\frac{745}{565.57}\right) + \\ + 1092\log\left(\frac{1092}{1271.43}\right) + 74\log\left(\frac{74}{41.56}\right) + 61\log\left(\frac{61}{93.44}\right) \end{array}\right) = 186.66$$

```
> devm0a<-2*sum(tt*log(tt/fitm0));devm0a
[1] 186.6617
```

5. Valoreu si l´ús o no d'alguna mesura de contracepció està estadísticament associat al nivell d'educació (f.educ). Empreu, justifiqueu i interpreteu un test estadístic adient.

Amb els resultats dels apartats anteriors amb les dades agregades cal considerar que el model saturat Y~A té una deviança de 0 i el model nul 186.66. Per tant, es pot fer un test de bondat del model nul: H0- El model nul s'ajusta bé a les dades. La distribució de l'estadístic deviança és assimptòticament una shi quadrat amb n-p=4-1=3 graus de llibertat i el p valor del constrast és per tant P(Shi_df3>186.66)=0, per tant, hi ha evidència per rebutjar la H0 i d'aquí que el model nul no s'ajusti bé a les dades, per tant, l'efecte del factor f.educ és necessari (i significatiu).
També es podria pensar en un test de diferències de devianes 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 f.educ, doncs en aquest cas, no seria un model saturat.

```
> # Goodness of fit
> devm0a
[1] 186.6617
> 1-pchisq(devm0a,3)
[1] 0
> # Test Deviança
> anova(m0,m1,test="Chisq")
Analysis of Deviance Table

Model 1: df$f.bcuse ~ 1
Model 2: df$f.bcuse ~ f.educ
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7096     8763.2
2      7093     8576.6  3   186.66 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Model m1: $$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta + \alpha_i \quad i = 1:4 \quad \alpha_{1\equiv none} = 0$$

	Estimador dades individuals	Estimador dades agrupades per f.educ	Iguals o diferents?
Terme independent (intercept)	-1.43026	-1.43026	Iguals
Estimadors dummies per f.educ	0.49893, 1.04788, 1.62345	0.49893, 1.04788, 1.62345	Iguals

	Estimador dades individuals	Estimador dades agrupades per f.educ	Iguals o diferents?
Null deviance	8762.2	186.66	Diferents
Graus llibertat Null Deviance	7096	3	Diferents
Deviance	8576.6	0	Diferents
Graus llibertat Deviance m1	7093	0	Diferents
AIC	8584.6	36.675	Diferents
Dev(m0)-Dev(m1)	186.66	186.66	Idèntics

```
> summary(m1)

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

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.43026    0.08344 -17.141  < 2e-16 ***
f.educfe.primary    0.49893    0.09020   5.532 3.17e-08 ***
f.educfe.secondary  1.04788    0.09602  10.913  < 2e-16 ***
f.educfe.high       1.62345    0.19201   8.455  < 2e-16 ***
---

    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 8576.6  on 7093  degrees of freedom
AIC: 8584.6

Number of Fisher Scoring iterations: 4

> dft<-as.data.frame(tt)
> dft<-as.data.frame(cbind(dft[5:8,c(1,3)],dft[1:4,3]))
> dft$m<-dft[,2]+dft[,3]
> names(dft)<-c("f.educ","ypos","yneg","m")
> dft
        f.educ ypos yneg    m
5      fe.none  178  744  922
6   fe.primary 1188 3015 4203
7 fe.secondary  745 1092 1837
8      fe.high   74   61  135
> m1a<-glm(cbind(ypos,yneg)~f.educ, family=binomial, data=dft)
> summary(m1a)

Call: glm(formula = cbind(ypos, yneg) ~ f.educ, family = binomial,
    data = dft)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.43026    0.08344 -17.141  < 2e-16 ***
f.educfe.primary    0.49893    0.09020   5.532 3.17e-08 ***
f.educfe.secondary  1.04788    0.09602  10.913  < 2e-16 ***
f.educfe.high       1.62345    0.19201   8.455  < 2e-16 ***
---

    Null deviance:  1.8666e+02  on 3  degrees of freedom
Residual deviance: -2.7955e-13  on 0  degrees of freedom
AIC: 36.675
```

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

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

Clarament a la vista dels resultats el terme lineal de l'edat és significatiu (Dev(m0)-Dev(m7)=8763.2-8455.9=307.3) i a la vista dels resultats de marginalModelPlot(m7) hi ha desajust que es pot corregir afegint el terme quadràtic de l'edat centrat en la mitjana. No s'inclou informació suficient per determinar si els termes polinòmics d'ordre superior de a 2 de l'edat serien significatius. El test de la deviança entre el model lineal i el quadràtic anova(m7a,m7) té p valor de 0 indicant hi ha evidència per rebutjar la hipòtesi nul.la que són equivalents i per tant, el terme quadràtic cal afegir-lo al model un cop que el terme lineal de l'edat ja ha estat incorporat.

L'equació que cal interpretar és:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_1 + \beta_2 age + \beta_3 (age - 28.25)^2 = -2.83 + 0.91 age - 0.0076(age - 28.25)^2$$

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

A la vista dels resultats del model m8 que conté el tractament de l'edat descrit, l'AIC del model m8 és 7979.1 mentre que l'AIC(m7a) (tractament numèric amb termes lineal i quadràtic de l'edat) és 7921.4, el millor tractament de l'edat és com a covariant. No poden comparar-se per Test Deviances o Wald els models m8 i m7a, doncs no són encaixats: només poden comparar-se per AIC o BIC.

9. Estudiar l'efecte de la covariable Nombre de Fills (**nsons**) sobre el target **f.bcuse**. Valoreu i interpreteu els coeficients del model en l'escala lògit, dels odds i aproximadament en probabilitat. Creieu que calen termes quadràtics o cúbics?

A la vista dels resultats disponibles on només s'indica el model amb tractament com a covariant amb terme lineal del nombre de fills, l'efecte de sons és significatiu (p valor de la taula de resultats). A la vista del marginalModelPlot(m9) hi ha desajust clarissim que no sembla tant fàcil de tractar com en el cas de l'edat. El coeficient de nsons és 0.122 per tant positiu, el lògit de la probabilitat d'ús d'alguna mesura contraceptiva s' incrementa en 0.122 unitats per cada fill. Els odds de la resposta positiva s'incrementen en 100(exp(0.122)-1)=13% amb cada fill. En termes aproximats de probabilitat seria un increment absolut de 0.122*0.3079*(1-0.3079)=0.026 unitat per cada fill. Evidentment el model lineal no és adient, doncs hi ha una franja de dones que tenen molts fills i que no prenen cap mesura per previnir-ne l'arribada de més.*

```
> coef(m9)
(Intercept)       nsons
 -1.1990748   0.1221829
> exp(coef(m9))
(Intercept)       nsons
   0.301473    1.129961
> tt0[2]*tt0[1]*coef(m9)[2]
target.yes
0.02603578
```

10. Considereu ara l'agrupació del **Nb. Fills (nsons)** en 4 categories (**f.nsons**). Quin tractament considereu més adequat per la variable **nsons** amb els resultats disponibles (no s'informa de quina agrupació s'ha considerat)? *Nota: hi ha més de 100* dones *que han tingut més de 10 fills, de fet en promig les que han tingut més de 10 fills han tingut 11,7 fills, abans de res, s'ha creat una variable que per aquest grup de dones els assigna el valor 11.7.*

Comparant l'estadístic d'Akaike del model m9 (covariant, amb terme lineal només) i m10 (factor, en 4 grups), no es poden comparar ni amb el test de la deviança ni amb Wald doncs no són models encaixats, es veu que AIC(m9)=8560,2 i AIC(m10)=7921.6, per tant el tractament com a factor sembla més adient. La qüestió addicional inclosa rau en quina agrupació del nombre de fills cal emprar. A la vista dels decils i per tal d'equilibrar el nb d'observacions en cada grup, les que tenen 0 fills anirien en un grup després les que tenen menys de 3 fillls. El valor a partir d'on es considera un nb de fills exagerat estaria associat al 3er quartil i es fa difícil de precisar si és en 5 o 6 fills (no teniu possibilitat de provar-ho, però 5 dona uns grups més homogenis en tamany) i després la resta amb molts fills. La comanda emprada és:

```
df$f.nsons<-factor(cut(df$nsons,breaks=c(-1,0,2,5,15)),labels=c("lab1","lab2","lab3","lab4"))
```

11. Es necessita controlar pel Nb. De Fills si ja s'ha incorporat l'Edat en el model? Justifiqueu estadísticament les respostes.

Els resultats disponibles contenen la covariant edat (age) amb polinomi d'ordre 2 i el factor f.nsons, es presenta el model additiu (m11a) i el model amb interaccions factor-covariant (m11a) per la transformació lògit de la resposta positiva. Les interaccions són significatives per tant l'efecte principal del factor f.nsons ha de considerar-se també i la resposta és que un cop que l'edat ja està en el model si que cal controlar pel factor del nombre de fills f.nsons (l'efecte net de l'efecte principal és significatiu un cop que l'edat s'ha inclòs en el model).

```
> m11p<-glm(f.bcuse~poly(age,2)+f.nsons, family=binomial, data=df)
> Anova(m11p)
Analysis of Deviance Table (Type II tests)
Response: f.bcuse
             LR Chisq Df Pr(>Chisq)
poly(age, 2)   192.62  2  < 2.2e-16 ***
f.nsons        194.40  3  < 2.2e-16 ***
---

> anova(m11a,m11,test="Chisq")
Analysis of Deviance Table

Model 1: f.bcuse ~ age + I((age - 28.25)^2) + f.nsons
Model 2: f.bcuse ~ (age + I((age - 28.25)^2)) * f.nsons
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7091     7721.0
2      7085     7602.8  6   118.14 < 2.2e-16 ***
---
```

12. Estudieu l'efecte de la covariable Anys d'Estudi (yeduc) sobre el target f.bcuse. Valoreu i considereu d' emprar una agrupació dels Anys d'Estudi raonable en el sistema educatiu de Kènia, ajudant-vos d'una taula creuada entre Anys d'Estudi i Nivell de Lectura. Quin tractament considereu més adequat per la variable amb els resultats disponibles (no s'informa de quina agrupació s'ha considerat)? *Nota: hi ha poques dones amb 1 o 2 anys d'educació i també poques amb més de 12 que (de fet en promig les que han estudiat més de 12 anys ho han fet 15 any)s, abans de res, s'ha creat una variable que agrupi les dones amb 1 o 2 anys sota el descriptor 1.5 i amb més de 12 anys els assigni el valor 15.*

El tractament com a covariant dels anys d'educació (yeduc) mostra en la taula de resultats summary() un pvalor <0.05 per tant efecte significatiu. El tractament com a factor es pot determinar que és a partir de la diferència de devianes entre el model nul 8763.2 i el model m12f 8569.5, gairebé 200 unitats a contrastar segons una shi quadrat de 3 graus de llibertat, per tant també significatiu (no cal emprar taules). Tots 2 tractaments són significatius, però per l'argument de l'AIC es poden comparar i es selecciona el model del AIC més petit, per tant el tractament com a covariant. AIC(m12)=8561 i AIC(m12f)=8577.5. La part més interessant rau en com discretitzar la variable anys d'educació a partir de la taula: per una banda estarien les que han estudiat menys de 3 anys, la gran majoria analfabetes,

després les que llegeixen bastant bé (some) que són les que han estudiat fins a 8 anys i a partir de 9 anys o més, les que llegeixen perfectament, això serien 3 grups, però estarien desequilibrats per tant el punt d'inflexió on comencen a llegir bé sembla que està pels 5 o 6 anys i això permet per la partició en 4 grups més equilibrats.

```
df$f.yeduc<-factor(cut(df$yeduc2,breaks=c(-
1,1.5,5,8,15)),labels=c("lab1","lab2","lab3","lab4"))
```

13. Estudieu l'efecte de les variables Veu TV ? i Escolta Radio ? sobre el target f.bcuse un cop l'edat i els anys d'experiència ja estan incorporats al model. Interpreteu els seus coeficients al model i valoreu la significació.

El model m13 no es pot simplificar segons el criteri d'AIC, conté els anys d'educació (lineal), el polinomic quadràtic de l'edat en interacció amb el factor nombre de fills (f.nsons) i els dos factors f.seeTV i f.lisradio indicats. Comparant la deviança residual de m13 (m13f, són iguals) amb m13p (sense els 2 factors mediàtics): Dev(m13p)-Dev(m13)= 7317.694-7240.0=77.7 a contrastar amb una shi quadrat de 2 g.ll, per tant els dos models no són equivalents. Usant les dades del step(m13), es pot veure que tant f.seeTV com f.lisradio són significatius i no poden suprimir-se del model m13, 17 unitats de deviança per f.seeTV i 45 unitats per f.lisradio, per tant significatius. **No es poden interpretar els coeficients, doncs no están disponibles.**

14. Un problema d'usar les variables Veu TV ? i Escolta Radio ? rau en que reflecteixen un efecte socio-econòmic (no tothom que veu TV en té). Repetir la interpretació dels coeficients en el model disponible després de controlar per disponibilitat de radio i TV (Té Radio? i Té TV?).

La sortida Anova(m14) indica que totes les variables tenen un efecte net significatiu, incloent la interacció f.nsons amb l'edat (termes 1 i 2), el veu la tele està molt just doncs supera per poc el llindar habitual del 0.05, és a dir un cop s'ha controlat pere la disponibilitat de TV i de radio, el veure la televisió perd importància. En canvi, escoltar la ràdio segueix resultant molt significatiu.

```
> Anova(m14)
Analysis of Deviance Table (Type II tests)
Response: f.bcuse
                     LR Chisq Df Pr(>Chisq)
yeduc                 163.734  1  < 2.2e-16 ***
poly(age, 2)           93.469  2  < 2.2e-16 ***
f.nsons               246.993  3  < 2.2e-16 ***
f.TV                    4.610  1   0.031790 *
f.radio                 9.611  1   0.001934 **
f.seeTV                 3.664  1   0.055590 .
f.lisradio             17.167  1  3.423e-05 ***
poly(age, 2):f.nsons   87.295  6  < 2.2e-16 ***
```

El model m14 conté propietat i ús simulani de radio i TV. Comparant el model m13 Dev(m13) 7240 amb el model m14 Dev(m14)= 7223.3, un cop que l'ús de radio i TV estan en el model, la disponibilitat de ràdio i TV són factors binaris significatius doncs Dev(m13)-D(m14)= 16.7 amb 2 g.ll, per tant els test de la deviança duria a un pvalor << 0.05, indicant que els dos models no són equivalents, per tant, cal incloure la propietat, un cop l'ús ja està en el model.

Després de controlar per la propietat de TV i radio, el logodds de sentir la radio s'incrementa en 0.301 unitats all else being equal o els odds d'ús d'anticoncepció s'incrementen en un 100*(exp(0.301)-1)=35.11% respecte el grup que no escolta la radio (ceteris paribus).
Després de controlar per la propietat de TV i radio, el logodds de veure TV s'incrementa en 0.17 unitats all else being equal o els odds d'ús d'anticoncepció s'incrementen en un 100*(exp(0.17)-1)=18.52% respecte el grup que no veu TV (ceteris paribus).

```
> summary(m14)
Call: glm(formula = f.bcuse ~ yeduc + (poly(age, 2)) * f.nsons + f.TV +
```

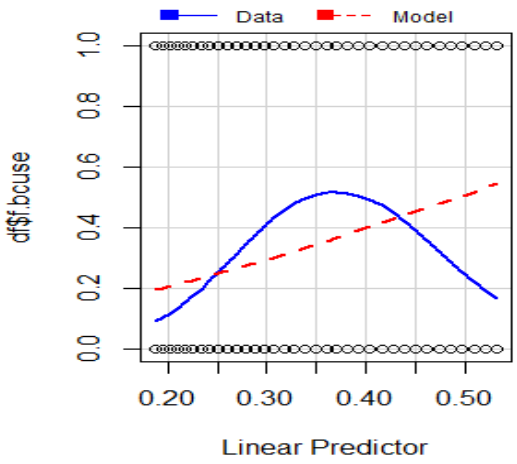
```
    f.radio + f.seeTV + f.lisradio, family = binomial, data = df)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.932488   0.289256 -13.595  < 2e-16 ***
veduc                        0.116035   0.009294  12.485  < 2e-16 ***
poly(age, 2)1              -27.516188  31.619077  -0.870  0.38417
poly(age, 2)2              -76.082787  19.421912  -3.917 8.95e-05 ***
f.nsonslab2                  1.579769   0.294051   5.372 7.77e-08 ***
f.nsonslab3                  1.941568   0.297121   6.535 6.38e-11 ***
f.nsonslab4                  0.670467   0.453496   1.478  0.13929
f.TVftv.yes                  0.224138   0.104453   2.146  0.03189 *
f.radiofr.yes                0.242089   0.078204   3.096  0.00196 **
f.seeTVfstv.yes              0.169985   0.088547   1.920  0.05490 .
f.lisradioflr.yes            0.301051   0.072835   4.133 3.58e-05 ***
poly(age, 2)1:f.nsonslab2    1.710547  33.093986   0.052  0.95878
poly(age, 2)2:f.nsonslab2    8.931662  21.519748   0.415  0.67811
poly(age, 2)1:f.nsonslab3   81.833141  32.842778   2.492  0.01271 *
poly(age, 2)2:f.nsonslab3   30.144202  20.862563   1.445  0.14849
poly(age, 2)1:f.nsonslab4  169.579609  41.862316   4.051 5.10e-05 ***
poly(age, 2)2:f.nsonslab4    7.248092  23.487198   0.309  0.75763
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7223.3  on 7080  degrees of freedom
```

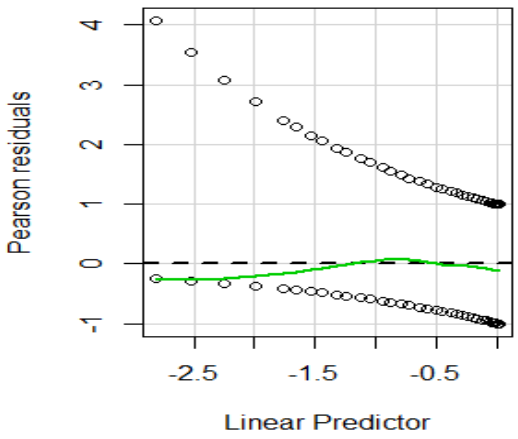
15. En el model del Punt 14 es vol fer una diagnosi per determinar la presència d'observacions influents i residuals atípics. Amb els resultats disponibles indiqueu les observacions potencialment influents, les observacions que constitueixen definitivament dades influents i aquelles que són outliers dels residus. . Què els passa a les observacions indicades?

Hi ha dues dones la 6498 i 6505 que són grans, la primera sense fills i la segona amb 2 fills i que prenen mesures de contracepció, són observacions influents segons l'indicador de la distància de Cook. Totes dues superen la cota del factor d'anclatge, leverage. Hi ha residus grans, estudentitzats per tant són interpretables, però considerant el tamany de la mostra i la magnitud en valor absolut per sota de 3, no sembla que el desajust del model sigui massa important en aquestes observacions (2434,5540), són dones amb resposta positiva malgrat no tenen cap fill, la qual cosa és comprensible doncs tenen 15 anys.

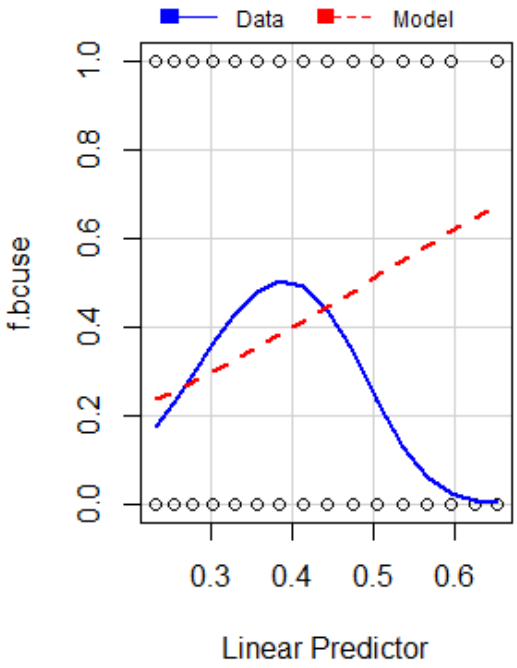
RESULTATS PEL PROBLEMA 1

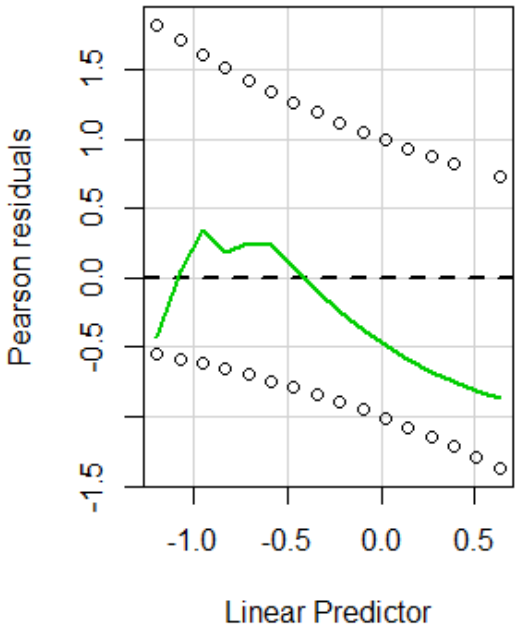
```
> m7<-glm(df$f.bcuse~age, family=binomial, data=df)
> m7a<-glm(df$f.bcuse~age+I((age-28.25)^2), family=binomial, data=df)
> summary(m7a)

Call: glm(formula = df$f.bcuse ~ age + I((age - 28.25)^2), family = binomial,
    data = df)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.8328358  0.1074796  -26.36   <2e-16 ***
age                 0.0909955  0.0038434   23.68   <2e-16 ***
I((age - 28.25)^2) -0.0076426  0.0003562  -21.46   <2e-16 ***
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7915.4  on 7094  degrees of freedom
AIC: 7921.4

> anova(m7,m7a,test="Chisq")
Analysis of Deviance Table
Model 1: df$f.bcuse ~ age
Model 2: df$f.bcuse ~ age + I((age - 28.25)^2)
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7095     8455.9
2      7094     7915.4  1   540.53 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> par(mfrow=c(1,2))
> marginalModelPlot(m7)
> residualPlot(m7a)
```






```
> summary(df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   20.00   27.00   28.25   36.00   49.00
> df$f.age<-cut(df$age,breaks=seq(10,50,by=5))
> summary(df$f.age)
(10,15] (15,20] (20,25] (25,30] (30,35] (35,40] (40,45] (45,50]
    405    1576    1288    1133     897     840     596     362
> m8<-glm(df$f.bcuse~f.age, family=binomial, data=df)
> summary(m8)

Call:glm(formula = df$f.bcuse ~ f.age, family = binomial, data = df)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.5785     0.3057 -11.706  < 2e-16 ***
f.age(15,20]   1.4854     0.3161   4.699 2.62e-06 ***
f.age(20,25]   2.9085     0.3113   9.343  < 2e-16 ***
f.age(25,30]   3.2923     0.3115  10.568  < 2e-16 ***
f.age(30,35]   3.3613     0.3130  10.740  < 2e-16 ***
f.age(35,40]   3.2908     0.3135  10.496  < 2e-16 ***
f.age(40,45]   3.1351     0.3170   9.890  < 2e-16 ***
f.age(45,50]   2.4124     0.3297   7.317 2.53e-13 ***
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7963.1  on 7089  degrees of freedom
AIC: 7979.1

> m9<-glm(f.bcuse~nsons, family=binomial, data=df)
> summary(m9)

Call:glm(formula = f.bcuse ~ nsons, family = binomial, data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.199075   0.038656  -31.02   <2e-16 ***
nsons        0.122183   0.008523   14.34   <2e-16 ***
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 8556.2  on 7095  degrees of freedom
AIC: 8560.2
>
> par(mfrow=c(1,2))
> marginalModelPlot(m9)
> residualPlot(m9)
```

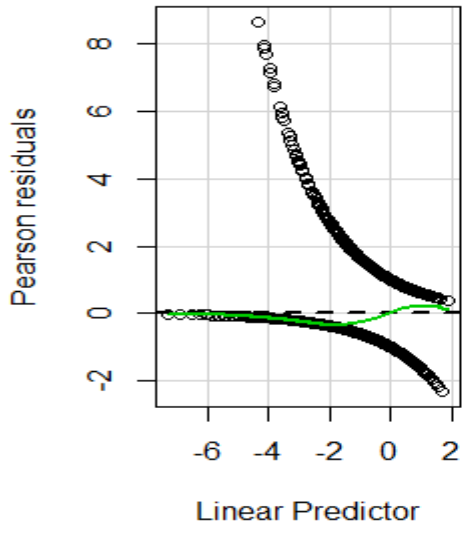


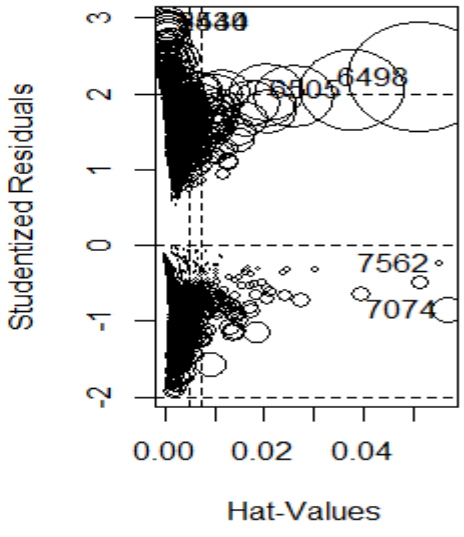



```
> table(df$nsons)
   0    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
1975  942  878  695  627  490  459  328  257  211  132   51   35   12    3    2
> quantile(df$nsons,seq(0.1,by=0.10))
  0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
   0    0    0    1    1    2    3    4    6    7   15
df$f.nsons<-factor(cut(df$nsons,breaks=[redacted],labels=c("lab1","lab2","lab3","lab4"))
> summary(m10)

Call: glm(formula = f.bcuse ~ f.nsons, family = binomial, data = df)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.49151    0.08468  -29.42   <2e-16 ***
f.nsonslab2  1.99823    0.09749   20.50   <2e-16 ***
f.nsonslab3  2.29217    0.09695   23.64   <2e-16 ***
f.nsonslab4  1.89159    0.10052   18.82   <2e-16 ***
---
 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7913.6  on 7093  degrees of freedom
AIC: 7921.6

> m11a<-glm(f.bcuse~age+I((age-28.25)^2)+f.nsons, family=binomial, data=df)
> m11<-glm(f.bcuse~(age+I((age-28.25)^2))*f.nsons, family=binomial, data=df)
> summary(m11)

Call:
glm(formula = f.bcuse ~ (age + I((age - 28.25)^2)) * f.nsons,
    family = binomial, data = df)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.661112   0.867117  -0.762 0.445806
age                          -0.016531   0.032880  -0.503 0.615128
I((age - 28.25)^2)           -0.016198   0.002831  -5.722 1.05e-08 ***
f.nsonslab2                   0.240934   0.907041   0.266 0.790527
f.nsonslab3                  -2.514005   0.956196  -2.629 0.008559 **
f.nsonslab4                  -6.311346   1.556042  -4.056 4.99e-05 ***
age:f.nsonslab2               0.033329   0.034334   0.971 0.331685
age:f.nsonslab3               0.122072   0.035655   3.424 0.000618 ***
age:f.nsonslab4               0.215119   0.051104   4.209 2.56e-05 ***
I((age - 28.25)^2):f.nsonslab2  0.004381   0.003118   1.405 0.160070
I((age - 28.25)^2):f.nsonslab3  0.008741   0.003006   2.908 0.003635 **
I((age - 28.25)^2):f.nsonslab4  0.006795   0.003315   2.050 0.040370 *
---
 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7602.8  on 7085  degrees of freedom
AIC: 7626.8
> m11p<-glm(f.bcuse~poly(age,2)+f.nsons, family=binomial, data=df)
```

```
> Anova(m11p)
Analysis of Deviance Table (Type II tests)
Response: f.bcuse
             LR Chisq Df Pr(>Chisq)
poly(age, 2)   192.62  2  < 2.2e-16 ***
f.nsons        194.40  3  < 2.2e-16 ***
---

> anova(m11a,m11,test="Chisq")
Analysis of Deviance Table

Model 1: f.bcuse ~ age + I((age - 28.25)^2) + f.nsons
Model 2: f.bcuse ~ (age + I((age - 28.25)^2)) * f.nsons
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      7091     7721.0
2      7085     7602.8  6   118.14 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> df$veduc2<-df$veduc
> sel<-which(df$veduc<3)
> df$veduc2[sel]<-1.5
> sel<-which(df$veduc>12)
> df$veduc2[sel]<-15
> table(df$yeduc2,df$f.read)

      fr.well fr.some fr.none fr.missing
  1.5      26     100     982          4
  3        23     107      84          2
  4        96     166      76          3
  5       158     206      50          1
  6       417     272      39          1
  7       913     367      37          9
  8       937     149       9          3
  9       308       0       0          0
  10      246       0       0          0
  11      479       0       0          0
  12      672       0       0          0
  15      155       0       0          0
> df$f.yeduc<-factor(cut(df$yeduc2,breaks=[redacted]labels=c("lab1","lab2","lab3","lab4")
)
>
> m12<-glm(f.bcuse~yeduc, family=binomial, data=df)
> summary(m12)

Call:
glm(formula = f.bcuse ~ yeduc, family = binomial, data = df)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.527163   0.059290  -25.76   <2e-16 ***
yeduc        0.102498   0.007334   13.98   <2e-16 ***
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 8557.0  on 7095  degrees of freedom
AIC: 8561
> m12f<-glm(f.bcuse~f.yeduc, family=binomial, data=df)
> summary(m12f)

Call:
glm(formula = f.bcuse ~ f.yeduc, family = binomial, data = df)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.32775    0.07369 -18.019  < 2e-16 ***
f.yeduclab2  0.09864    0.10632   0.928    0.354
f.yeduclab3  0.48362    0.08329   5.806 6.39e-09 ***
f.yeduclab4  1.01115    0.08737  11.573  < 2e-16 ***
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 8569.5  on 7093  degrees of freedom
AIC: 8577.5

> m13<-glm(f.bcuse~yeduc+(poly(age,2))*f.nsons+f.seeTV+f.lisradio, family=binomial, data=df)
> m13f<-step(m13)
Start:  AIC=7270.04
f.bcuse ~ yeduc + (poly(age, 2)) * f.nsons + f.seeTV + f.lisradio

                      Df Deviance    AIC
<none>                     7240.0 7270.0
- f.seeTV              1   7257.3 7285.3
- f.lisradio           1   7285.4 7313.4
- poly(age, 2):f.nsons  6   7327.5 7345.5
- yeduc                1   7432.8 7460.8
> m13p<-glm(f.bcuse~yeduc+(poly(age,2))*f.nsons, family=binomial, data=df)
> deviance(m13p)
```

```
[1] 7317.694

> m14<-glm(f.bcuse~yeduc+(poly(age,2))*f.nsons+f.TV+f.radio+f.seeTV+f.lisradio, family=binomial,
data=df)
> m14f<-step(m14)
Start:  AIC=7257.28
f.bcuse ~ yeduc + (poly(age, 2)) * f.nsons + f.TV + f.radio +
    f.seeTV + f.lisradio

                       Df Deviance    AIC
<none>                      7223.3 7257.3
- f.seeTV               1   7226.9 7258.9
- f.TV                  1   7227.9 7259.9
- f.radio               1   7232.9 7264.9
- f.lisradio            1   7240.5 7272.5
- poly(age, 2):f.nsons  6   7310.6 7332.6
- yeduc                 1   7387.0 7419.0
> summary(m14)
Call: glm(formula = f.bcuse ~ yeduc + (poly(age, 2)) * f.nsons + f.TV +
    f.radio + f.seeTV + f.lisradio, family = binomial, data = df)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -3.932488   0.289256 -13.595  < 2e-16 ***
yeduc                        0.116035   0.009294  12.485  < 2e-16 ***
poly(age, 2)1              -27.516188  31.619077  -0.870  0.38417
poly(age, 2)2              -76.082787  19.421912  -3.917 8.95e-05 ***
f.nsonslab2                  1.579769   0.294051   5.372 7.77e-08 ***
f.nsonslab3                  1.941568   0.297121   6.535 6.38e-11 ***
f.nsonslab4                  0.670467   0.453496   1.478  0.13929
f.TVftv.yes                  0.224138   0.104453   2.146  0.03189 *
f.radiofr.yes                0.242089   0.078204   3.096  0.00196 **
f.seeTVfstv.yes              0.169985   0.088547   1.920  0.05490 .
f.lisradioflr.yes            0.301051   0.072835   4.133 3.58e-05 ***
poly(age, 2)1:f.nsonslab2    1.710547  33.093986   0.052  0.95878
poly(age, 2)2:f.nsonslab2    8.931662  21.519748   0.415  0.67811
poly(age, 2)1:f.nsonslab3   81.833141  32.842778   2.492  0.01271 *
poly(age, 2)2:f.nsonslab3   30.144202  20.862563   1.445  0.14849
poly(age, 2)1:f.nsonslab4  169.579609  41.862316   4.051 5.10e-05 ***
poly(age, 2)2:f.nsonslab4   7.248092  23.487198   0.309  0.75763
---
    Null deviance: 8763.2  on 7096  degrees of freedom
Residual deviance: 7223.3  on 7080  degrees of freedom
AIC: 7257.3

> Anova(m14)
Analysis of Deviance Table (Type II tests)
Response: f.bcuse
                     LR Chisq Df Pr(>Chisq)
yeduc                 163.734  1  < 2.2e-16 ***
poly(age, 2)           93.469  2  < 2.2e-16 ***
f.nsons               246.993  3  < 2.2e-16 ***
f.TV                    4.610  1    0.031790 *
f.radio                 9.611  1    0.001934 **
f.seeTV                 3.664  1    0.055590 .
f.lisradio             17.167  1  3.423e-05 ***
poly(age, 2):f.nsons   87.295  6  < 2.2e-16 ***
---
> par(mfrow=c(1,2))
> residualPlot(m14)
> ll<-influencePlot(m14,id.n=2):ll
        StudRes          Hat      CookD
2434  2.9500592 0.0005254071 0.04819405
5540  2.9500592 0.0005254071 0.04819405
6498  2.2112508 0.0511383080 0.16586660
6505  2.0585584 0.0373157428 0.12270379
7074 -0.8533850 0.0565203186 0.03966433
7562 -0.2382045 0.0551997665 0.01008090
```





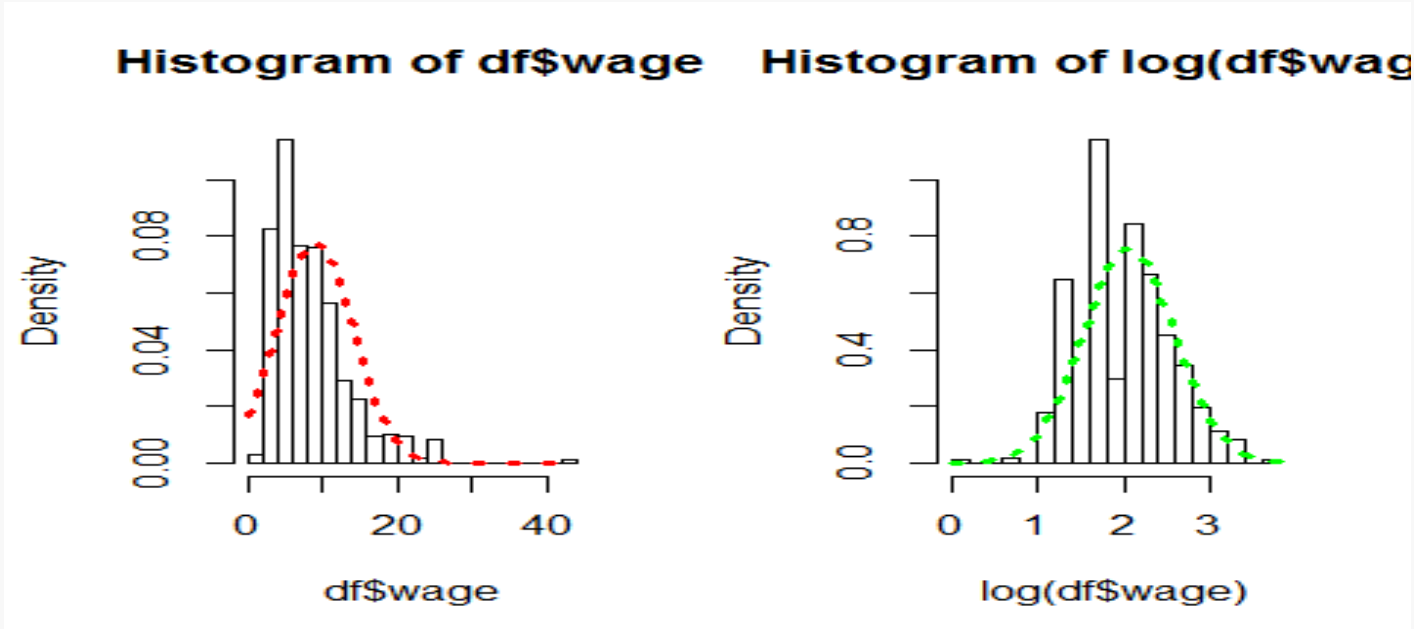

```
> df[row.names(ll),1:15]
     age f.residence       f.educ  f.read f.rnewsp  f.seeTV f.lisradio  f.elec f.radio
```

```
2434  15    fR.rural   fe.primary fr.well    fn.no  fstv.no     flr.no  fel.no   fr.no
5540  15    fR.rural   fe.primary fr.some    fn.no  fstv.no     flr.no  fel.no   fr.no
6498  40    fR.urban fe.secondary fr.well   fn.yes fstv.yes    flr.yes fel.yes  fr.yes
6505  48    fR.urban      fe.high fr.well   fn.yes fstv.yes    flr.no  fel.yes  fr.yes
7074  37    fR.urban      fe.high fr.well   fn.yes fstv.yes     flr.no fel.yes  fr.yes
7562  47    fR.urban      fe.high fr.well   fn.yes fstv.yes    flr.yes fel.yes  fr.yes
        f.TV veduc nsons f.pregnant   f.cuse    f.bcuse
2434  ftv.no     5     0       fp.no  fcu.con target.yes
5540  ftv.no     5     0       fp.no  fcu.abs target.yes
6498 ftv.yes    11     0       fp.no fcu.pill target.yes
6505 ftv.yes    19     2       fp.no  fcu.con target.yes
7074 ftv.yes    19     0       fp.no fcu.none  target.no
7562 ftv.yes    17     0       fp.no fcu.none  target.no
```

Problema 2 (4.5 punts): Determinants del Sou per hora (Wage)

Analitzarem un extracte de 534 observacions del 1985 Current Population Survey (CPS). Inclou dades per a cada treballador en anys d'educació, un indicador pels estats del sud dels EEUU, sexe, anys d'experiència laboral, un indicador de l'afiliació sindical, el salari per hora en dòlars (wage), l'edat, la raça (codificat com *Other, Hispanic, White*), l'ocupació (Gestió codificada *Management, Sales, Clerical, Service, Professional, Other*), sector (codificat *Other, Manufacturing, Construction*) i l'estat civil (codificada casat o no - *married or otherwise*).

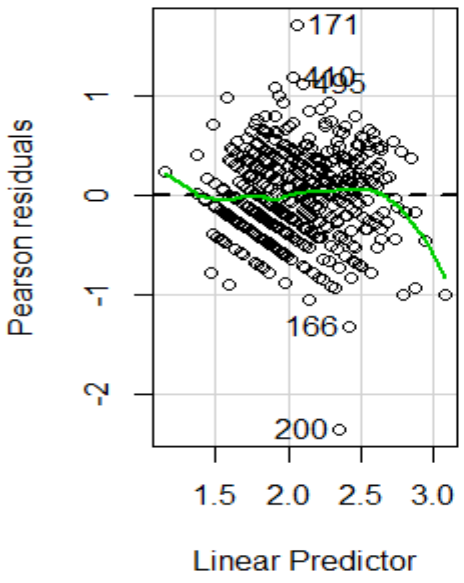
```
> summary(df)
      educ          f.south        sex          vexp          f.union       wage             age       
 Min.   : 2.00   fS.No :378   male  :289   Min.   : 0.00   fU.No :438   Min.   : 1.000   Min.   :18.00  
 1st Qu.:12.00   fS.Yes:156   female:245   1st Qu.: 8.00   fU.Yes: 96   1st Qu.: 5.000   1st Qu.:28.00  
 Median :12.00                            Median :15.00                Median : 8.000   Median :35.00  
 Mean   :13.02                            Mean   :17.82                Mean   : 9.024   Mean   :36.83  
 3rd Qu.:15.00                            3rd Qu.:26.00                3rd Qu.:11.000   3rd Qu.:44.00  
 Max.   :18.00                            Max.   :55.00                Max.   :44.000   Max.   :64.00  
       f.race          f.occup        f.sector    f.married  
 race.Other: 67   Occ.Man    : 55   fS.Other:411   fM.No :184  
 race.His  : 27   Occ.Sales  : 38   fS.Manu : 99   fM.Yes:350  
 race.white:440   Occ.Cleric : 97   fs.Const: 24               
                  Occ.Service: 83                              
                  Occ.Profl  :105                              
                  Occ.Other  :156                              
```

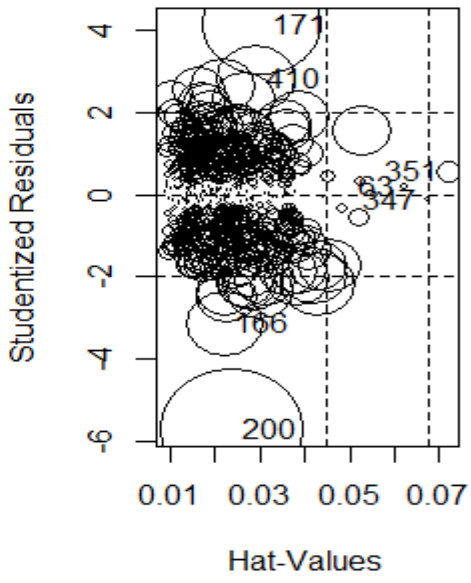



```
library(FactoMineR)
condes(df[,c("wage",expvar)],1)
## $quanti
##       correlation      p.value
## educ    0.3805239 7.654933e-20
## age     0.1843309 1.815896e-05
## yexp    0.0943312 2.928707e-02
##
## $quali
##                    R2      p.value
## f.occup    0.17867208 6.816477e-21
## sex        0.04195606 1.815476e-06
## f.union    0.02726045 1.267089e-04
## f.south    0.01969482 1.147964e-03
## f.married 0.01047830 1.797474e-02
## f.race     0.01249739 3.547364e-02
##
## $category
##                  Estimate      p.value
## Occ.Profl     2.8208115 3.110676e-11
## Occ.Man       3.5420236 1.658242e-08
## male          1.0481816 1.815476e-06
## fU.Yes        1.0963898 1.267089e-04
## fS.No         0.7869353 1.147964e-03
## race.white    1.0726728 1.230360e-02
## fM.Yes        0.5492547 1.797474e-02
## fM.No        -0.5492547 1.797474e-02
## fS.Yes       -0.7869353 1.147964e-03
## Occ.Cleric   -1.6589136 7.705670e-04
## fU.No        -1.0963898 1.267089e-04
## female       -1.0481816 1.815476e-06
## Occ.Service -2.5342086 1.627065e-06
```

S'ajusta un model lineal per explorar com el sou per hora depen de l'educació, l'experiència, l'afiliació sindical, la regió (estat del sud o no), ocupació i sexe.

```
lm1<-glm(log(wage)~educ+yexp+f.union+f.occup+f.south+sex,family=gaussian,data=df)
summary(lm1)
##
## Call: glm(formula = log(wage) ~ educ + yexp + f.union + f.occup + f.south +
##     sex, family = gaussian, data = df)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.246313   0.171754   7.256 1.44e-12 ***
## educ                0.068949   0.009945   6.933 1.22e-11 ***
## yexp                0.011037   0.001663   6.637 8.02e-11 ***
## f.unionfU.Yes       0.213027   0.051044   4.173 3.51e-05 ***
## f.occupOcc.Sales   -0.334658   0.091795  -3.646 0.000293 ***
## f.occupOcc.Cleric  -0.209785   0.076313  -2.749 0.006184 **
## f.occupOcc.Service -0.399198   0.080864  -4.937 1.07e-06 ***
## f.occupOcc.Profl   -0.039634   0.073001  -0.543 0.587417
## f.occupOcc.Other   -0.209725   0.075965  -2.761 0.005968 **
## f.southfS.Yes      -0.102890   0.041674  -2.469 0.013871 *
## sexfemale          -0.209188   0.041693  -5.017 7.19e-07 ***
## ---
##
## (Dispersion parameter for gaussian family taken to be 0.1850522)
##     Null deviance: 148.511  on 533  degrees of freedom
## Residual deviance:  96.782  on 523  degrees of freedom
## AIC: 627.39
Anova(lm1,test="F")
## Analysis of Deviance Table (Type II tests)
## Response: log(wage)
##
##               SS  Df       F    Pr(>F)
## educ       8.895   1 48.0685 1.218e-11 ***
## yexp       8.152   1 44.0499 8.021e-11 ***
## f.union    3.223   1 17.4173 3.514e-05 ***
## f.occup    6.729   5  7.2729 1.337e-06 ***
## f.south    1.128   1  6.0955   0.01387 *
## sex        4.658   1 25.1734 7.192e-07 ***
## Residuals 96.782 523
##
lm2<-glm(log(wage)~educ+yexp*sex+f.union+f.occup+f.south,family=gaussian,data=df)
summary(lm2)
##
## Call:  glm(formula = log(wage) ~ educ + yexp * sex + f.union + f.occup +
##     f.south, family = gaussian, data = df)
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.155796   0.175878   6.572 1.21e-10 ***
## educ                0.070837   0.009944   7.124 3.50e-12 ***
## yexp                0.014421   0.002249   6.413 3.20e-10 ***
## sexfemale          -0.089692   0.067892  -1.321 0.187044
## f.unionfU.Yes       0.203789   0.051022   3.994 7.42e-05 ***
## f.occupOcc.Sales   -0.329108   0.091484  -3.597 0.000352 ***
## f.occupOcc.Cleric  -0.194919   0.076320  -2.554 0.010933 *
## f.occupOcc.Service -0.377392   0.081154  -4.650 4.20e-06 ***
## f.occupOcc.Profl   -0.033201   0.072784  -0.456 0.648465
## f.occupOcc.Other   -0.194401   0.075993  -2.558 0.010804 *
## f.southfS.Yes      -0.106816   0.041555  -2.570 0.010433 *
## yexp:sexfemale     -0.006782   0.003048  -2.225 0.026500 *
##
## (Dispersion parameter for gaussian family taken to be 0.1836647)
##     Null deviance: 148.511  on 533  degrees of freedom
## Residual deviance:  95.873  on 522  degrees of freedom
## AIC: 624.35
##
par(mfrow=c(1,2))
residualPlot(lm2,id.n=5)
influencePlot(lm2,id.n=3)
```






```
> influencePlot(lm2,id.n=3)
        StudRes         Hat       CookD
63    0.23991122 0.06215793 0.017845789
166 -2.89696473 0.02212685 0.124915847
171  4.13170919 0.03073788 0.209204431
200 -5.72649164 0.02404034 0.251891796
347 -0.07018315 0.06729734 0.005447333
351  0.43023807 0.07189593 0.034594836
414 -2.45330478 0.04286220 0.149153385
```

1. Descriu els efectes de l'educació, l'experiència i la pertinença a un sindicat en el sou per hora en el model **lm1**.

El sou per hora s'incrementa per cada any addicional d'educació en un 7.1% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
El sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 1.1% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
El sou per hora en persones sindicades és un 23.75% superior al sou en persones no sindicades (base-line) ceteris paribus..

```
> exp(coef(lm1)[2:4])
        educ         yexp f.unionfU.Yes
    1.071382     1.011099     1.237418
```

2. Descriu l'efecte del gènere després de controlar pels efectes principals de la resta de variables en el model **lm1**. Indiqueu un contrast adient i interpreteu la seva significació.

El sou per hora es decrementa en les dones en gairebé un 19% (100(1-exp(-0.2092))%) respecte el sou dels homes, nivell de referència, dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).

Per contrastar la significació només cal mirar el pvalor del summary(lm1) 7.19e-07, per tant el paràmetre corresponent a la dummy necessària per modelar el grup de dones té un coeficient significativament diferent de 0, per tant, segons el test de Wald l'efecte principal és estadísticament significatiu.. Pel test de la deviança inclòs a la sortida del mètode Anova(lm1) es pot afirmar que l'efecte net del sexe és estadísticament significatiu.

```
lm1<-glm(log(wage)~educ+yexp+f.union+f.occup+f.south+sex,family=gaussian,data=df)
summary(lm1)
##
## Call: glm(formula = log(wage) ~ educ + yexp + f.union + f.occup + f.south +
##     sex, family = gaussian, data = df)
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.246313   0.171754   7.256 1.44e-12 ***
## educ               0.068949   0.009945   6.933 1.22e-11 ***
## yexp               0.011037   0.001663   6.637 8.02e-11 ***
## f.unionfU.Yes      0.213027   0.051044   4.173 3.51e-05 ***
## f.occupOcc.Sales  -0.334658   0.091795  -3.646 0.000293 ***
## f.occupOcc.Cleric -0.209785   0.076313  -2.749 0.006184 **
```

```
## f.occupOcc.Service -0.399198   0.080864  -4.937 1.07e-06 ***
## f.occupOcc.Profl   -0.039634   0.073001  -0.543 0.587417
## f.occupOcc.Other   -0.209725   0.075965  -2.761 0.005968 **
## f.southfS.Yes      -0.102890   0.041674  -2.469 0.013871 *
## sexfemale          -0.209188   0.041693  -5.017 7.19e-07 ***
## ---
Anova(lm1,test="F")
## Analysis of Deviance Table (Type II tests)
## Response: log(wage)
##               SS  Df       F    Pr(>F)
## educ       8.895   1 48.0685 1.218e-11 ***
## yexp       8.152   1 44.0499 8.021e-11 ***
## f.union    3.223   1 17.4173 3.514e-05 ***
## f.occup    6.729   5  7.2729 1.337e-06 ***
## f.south    1.128   1  6.0955   0.01387 *
## sex        4.658   1 25.1734 7.192e-07 ***
## Residuals 96.782 523
> 100*(1-exp(coef(lm1)[11]))
sexfemale
 18.87571
```

3. Determineu si els beneficis dels anys d'experiència són els mateixos en homes i en dones. Interpreteu l'efecte dels anys d'experiència pels 2 gèneres.Justifiqueu i interpreteu acuradament el test emprat.

La pregunta fa referència a la significació del terme de la interacció entre la covariant anys d'experiència (yexp) i el factor gènere (sex). Com sex és un factor binari el test de Wald per la significació de la interacció es troba directament a la sortida del summary(lm2) amb un p valor de 0.026 < 0.05 per tant hi ha evidència per rebutjar la H0 de paràmetre igual a 0. Per tant, un cop que les variables educ, f.union, f.occup i f.south estan al model, tantmateix com els efectes principals de yexp i sex aleshores els beneficis dels anys d'experiència depenen del gènere. També es pot fer el test de la deviança amb els resultats disponibles: diferència de devianes entre lm1 i lm2 igual a 4.95 unitats a contrastar amb una shi quadrat d'un grau de llibertat, per tant p valor < 0.05, es rebutja la H0 de dos models equivalents i per tant, la interacció és significativa pel test de la deviança.

Pels homes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 1.4% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
Per les domes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 0.7% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).

```
> anova(lm1,lm2,test="F")
Analysis of Deviance Table

Model 1: log(wage) ~ educ + yexp + f.union + f.occup + f.south + sex
Model 2: log(wage) ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1       523     96.782
2       522     95.873  1  0.90935 4.9511 0.0265 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

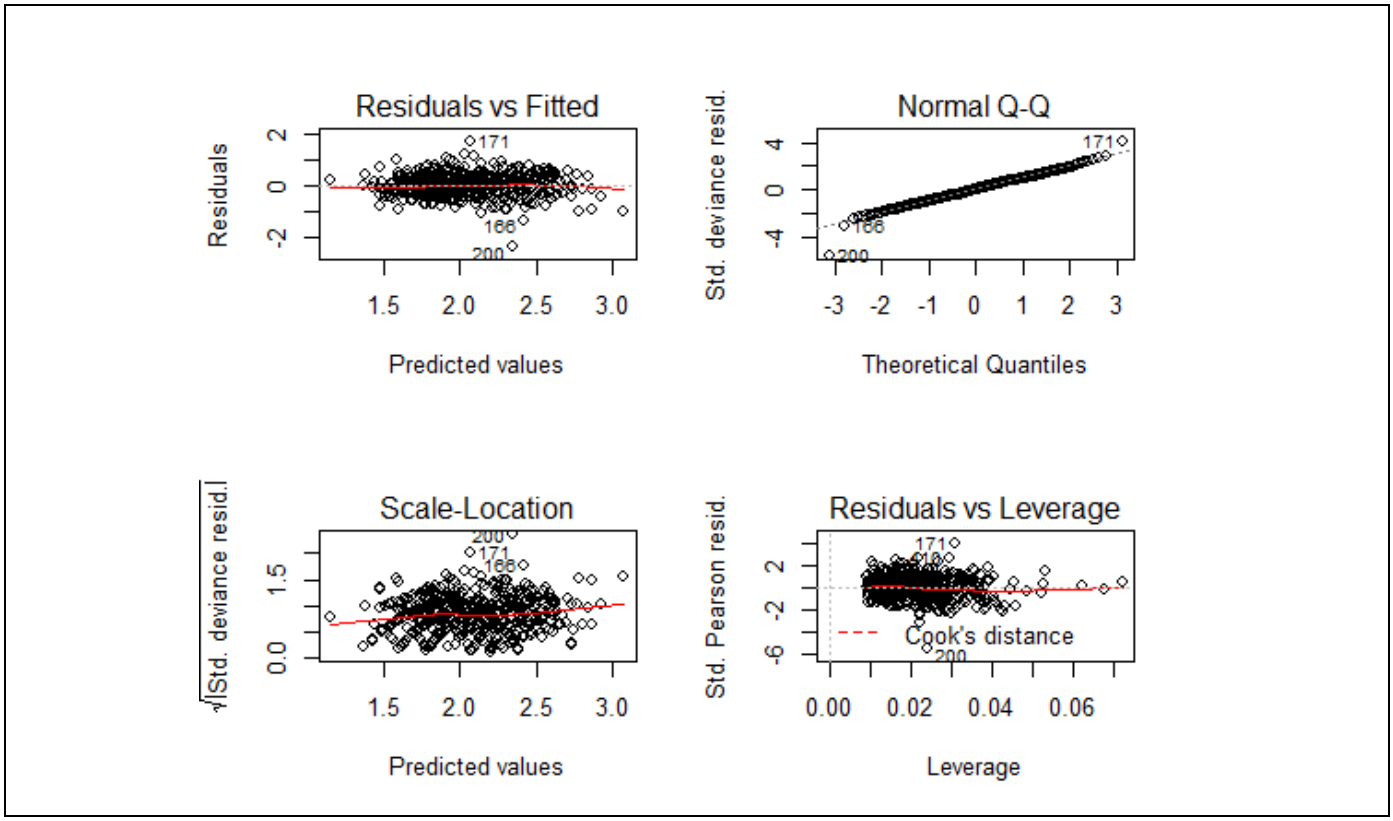
```
summary(lm2)
##
## Call:  glm(formula = log(wage) ~ educ + yexp * sex + f.union + f.occup +
##     f.south, family = gaussian, data = df)
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.155796   0.175878   6.572 1.21e-10 ***
## educ                0.070837   0.009944   7.124 3.50e-12 ***
## yexp                0.014421   0.002249   6.413 3.20e-10 ***
## sexfemale          -0.089692   0.067892  -1.321 0.187044
## f.unionfU.Yes       0.203789   0.051022   3.994 7.42e-05 ***
## f.occupOcc.Sales   -0.329108   0.091484  -3.597 0.000352 ***
## f.occupOcc.Cleric  -0.194919   0.076320  -2.554 0.010933 *
```

```
## f.occupOcc.Service -0.377392   0.081154  -4.650 4.20e-06 ***
## f.occupOcc.Profl   -0.033201   0.072784  -0.456 0.648465
## f.occupOcc.Other   -0.194401   0.075993  -2.558 0.010804 *
## f.southfS.Yes      -0.106816   0.041555  -2.570 0.010433 *
## yexp:sexfemale     -0.006782   0.003048  -2.225 0.026500 *
##
## (Dispersion parameter for gaussian family taken to be 0.1836647)
##     Null deviance: 148.511  on 533  degrees of freedom
## Residual deviance:  95.873  on 522  degrees of freedom
> 100*(exp(coef(lm2)[3])-1)
    yexp
1.452544
> 100*(exp(coef(lm2)[3]+coef(lm2)[12])-1)
     yexp
0.7668513
```

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

*Hi ha residus estudentitzats superiors a 3 o 4 en valors absolut, per tant, el model no s'adapta bé a bastantes observacions. L'observació 200 i la 171 són outliers dels residus (n'hi ha d'altres), hi ha alguna observació potencialment influent (lluny del centre de gravetat de les variables, a una cota d'anclatge superior a 3p/n=3*12/534=0.067) com la 351, però no és influent doncs els residu és baix. A totes, totes l'observació 200 i la 171 també són dades influent (dist Cook >0.20). El model mostra un desajust en les prediccions de sou per hora més elevats doncs facilita unes prediccions superiors als valors observats, possiblement perquè calgui incorporar alguna no linealitat en les covariants. Resaltant els outliers dels residus 171 i 200, per motius diferents el primer observació molt superior a la predicció del model i el segon al contrari.*

Podríem haver calculat i interpretar aquest model usant la transformació logarítmica de la resposta i el mètode lm(), la inferència resultaria idèntica, però aleshores podríem haver interpretat amb més facilitat els residus.



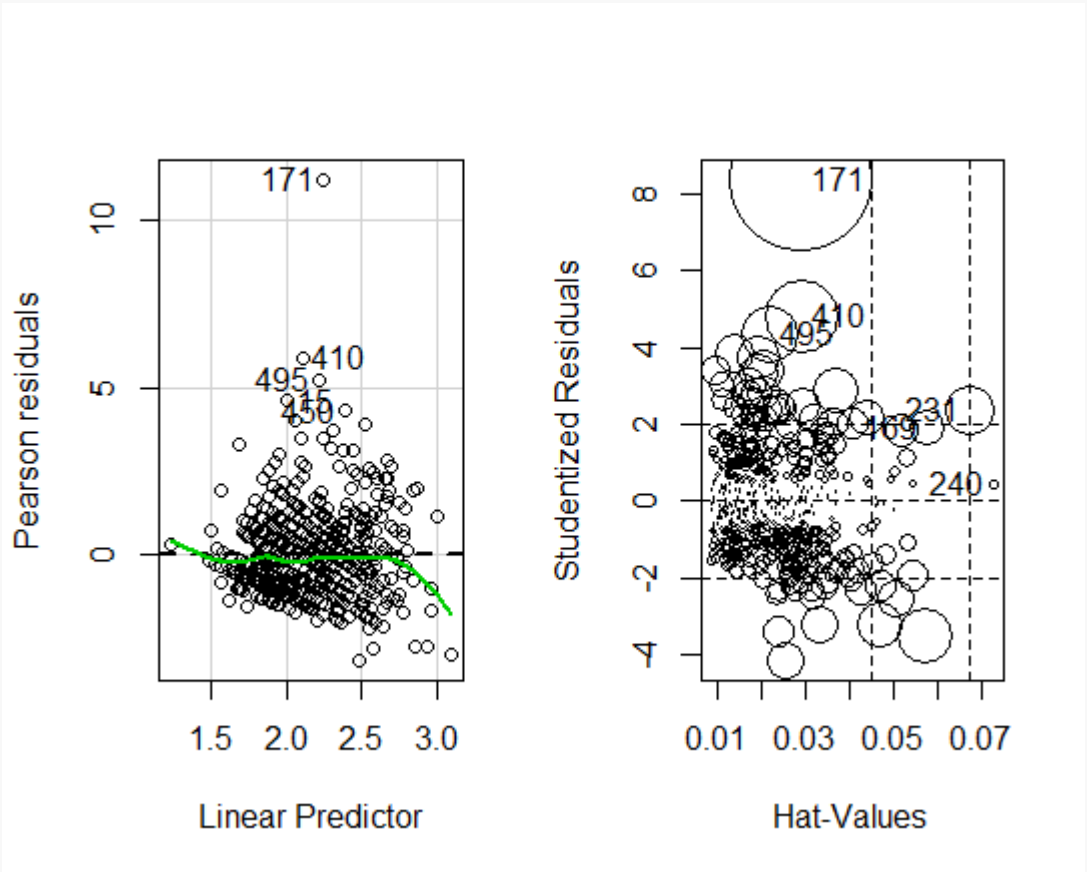
Donat que la variable de resposta té una experança no negativa, es consideren la família de models loglineals per una resposta poissoniana.

```
pm0<-glm(df$wage~1,family=poisson, data=df)
pm1<-glm(df$wage~educ+yexp+sex+f.union+f.occup+f.south,family=poisson, data=df)
pm2<-glm(df$wage~educ+yexp*sex+f.union+f.occup+f.south,family=poisson, data=df)

Anova(pm1)
## Analysis of Deviance Table (Type II tests)
##
## Response: df$wage
##          LR Chisq Df Pr(>Chisq)
## educ       86.797  1  < 2.2e-16 ***
## yexp       63.790  1  1.384e-15 ***
## sex        38.699  1  4.946e-10 ***
## f.union    19.274  1  1.132e-05 ***
## f.occup    65.509  5  8.789e-13 ***
## f.south     6.636  1   0.009996 **
## ---
Anova(pm2)
## Analysis of Deviance Table (Type II tests)
##
## Response: df$wage
##           LR Chisq Df Pr(>Chisq)
## educ        90.900  1  < 2.2e-16 ***
## yexp        63.790  1  1.384e-15 ***
## sex         38.699  1  4.946e-10 ***
## f.union     17.237  1  3.299e-05 ***
## f.occup     60.180  5  1.116e-11 ***
## f.south      7.428  1   0.006422 **
## yexp:sex     8.661  1   0.003250 **
##
summary(pm2)
##
## Call:
## glm(formula = df$wage ~ educ + yexp * sex + f.union + f.occup +
##     f.south, family = poisson, data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1311  -0.9237  -0.2238   0.6470   8.1384
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.278670   0.136719   9.353  < 2e-16 ***
## educ             0.073672   0.007765   9.488  < 2e-16 ***
## yexp             0.013492   0.001665   8.105 5.29e-16 ***
## sexfemale       -0.070035   0.054958  -1.274  0.20254
```

```
## f.unionfU.Yes       0.156196   0.037187   4.200 2.67e-05 ***
## f.occupOcc.Sales   -0.374281   0.070753  -5.290 1.22e-07 ***
## f.occupOcc.Cleric  -0.296235   0.056832  -5.212 1.86e-07 ***
## f.occupOcc.Service -0.416512   0.062694  -6.644 3.06e-11 ***
## f.occupOcc.Profl   -0.109421   0.048735  -2.245  0.02475 *
## f.occupOcc.Other   -0.234029   0.054920  -4.261 2.03e-05 ***
## f.southfS.Yes      -0.091048   0.033632  -2.707  0.00679 **
## vexp:sexfemale     -0.007141   0.002429  -2.940  0.00328 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1351.12  on 533  degrees of freedom
## Residual deviance:  865.45  on 522  degrees of freedom
## AIC: 2983.8
##
## Number of Fisher Scoring iterations: 4
```

```
sum(resid(pm2,type="pearson")^2)
## [1] 957.6935
dispersiontest(pm2,trafo=2)
##
##  Overdispersion test data:  pm2
## z = 3.56, p-value = 0.0001855
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## 0.09636096
```

```
> influencePlot(pm2,id.n=3)
      StudRes        Hat      CookD
169 1.9145851 0.05767642 0.14582264
171 8.3708037 0.02949818 0.57400476
231 2.3799095 0.06746539 0.20301309
      StudRes        Hat      CookD
240 0.4362551 0.07262683 0.03588193
410 4.8150746 0.02914968 0.29795447
495 4.3605005 0.02208798 0.22866511
```



5. Valoreu pros i contres de la modelització del sou com una variable de Poisson.

La resposta no és un recompta, a banda el sou per hora pot tenir valors reals, no en aquest cas en que s'ha forçat l'arrodoniment a un enter de la variable original wage. Resulta molt forçat. No sembla adient a priori, millor tantejar una opció de model de la resposta segons un model

gamma. Per la banda positiva, les prediccions no seran mai negatives amb la proposta Poisson.

6. **Compareu el millor model proposat (entre lm1 i lm2) en relació amb el model loglineal Poisson en termes de les variables significatives i el coeficient de determinació generalitzat.**

El millor model disponible és l'homònim del lm2, és a dir pm2, pm2<-glm(df$wage~educ+yexp*sex+f.union+f.occup+f.south,family=poisson, data=df). Segons els resultats del mètode Anova(pm2) totes les variables i termes són estadísticament significatius.

El goodness of fit de la H0: El model pm2 ajusta bé les dades, calculat a partir de la deviança del model pm2 mostra un pvalor=P(Shi2(522)>865.45)=0, per tant, es rebutja la H0: el model no ajusta bé les dades.

El goodness of fit de la H0: El model pm2 ajusta bé les dades, calculat a partir de l'estadístic de Pearson generalitzat del model pm2 mostra un pvalor=P(Shi2(522)>965.69)=0, per tant, es rebutja la H0: el model no ajusta bé les dades.

El coeficient de determinació generalitzat de pm2 és del 36% aproximadament.
El coeficient de determinació generalitzat de lm2 és del 34% aproximadament.

```
> deviance(pm2)
[1] 865.4461
> 1-pchisq(deviance(pm2),522)
[1] 0
> 1-pchisq(957.6935,522)
[1] 0
> 1-(deviance(pm2)/pm2$null.deviance)
[1] 0.3594579
> 1-(deviance(lm2)/lm2$null.deviance)
[1] 0.3544401
```

7. Descriu l'efecte dels anys d'experiència en el millor model de resposta Poisson disponible.

Pels homes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 1.36% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
Per les domes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 0.64% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
Són interpretacions molt similars a les realitzades amb el model lm2 (loglineal).

```
> coef(pm2)
      (Intercept)              educ              yexp         sexfemale      f.unionfU.Yes
        1.278669737       0.073672023       0.013491966      -0.070035488       0.156196003
  f.occupOcc.Sales  f.occupOcc.Cleric f.occupOcc.Service   f.occupOcc.Profl   f.occupOcc.Other
       -0.374281417      -0.296235335      -0.416512240      -0.109421332      -0.234028812
     f.southfS.Yes     yexp:sexfemale
       -0.091047589      -0.007140831
> 100*(exp(coef(pm2)[3])-1)
    yexp
1.358339
> 100*(exp(coef(pm2)[3]+coef(pm2)[12])-1)
     yexp
0.6371346
```

8. **Estimeu el factor de dispersió. Hi ha alguna evidència per suposar que les dades presenten sobredispersió. Justifiqueu la resposta amb la interpretació adient del contrast disponible.**

El factor de dispersió s'estima segons McCullagh com l'estadístic de Pearson Generalitzat dividit pels graus de llibertat del model, això si hauríem d'assegurar-nos que no hi ha més

variables i/o termes significatius entre les variables i possibilitats que no estan incloses en el model pm2. En la sortida es facilita l'estadístic de Pearson Generalitzat 957.69 per pm2 i els g.ll són 522 el que dona un factor de sobredispersió de 1.834

.

El test de sobredispersió parametritzat com Model NB2 -

$$h(\mu_i)=\mu_i^2 \rightarrow V[Y_i|X_i]=\mu_i+\alpha\mu_i^2=(1+\alpha\mu_i)\mu_i$$

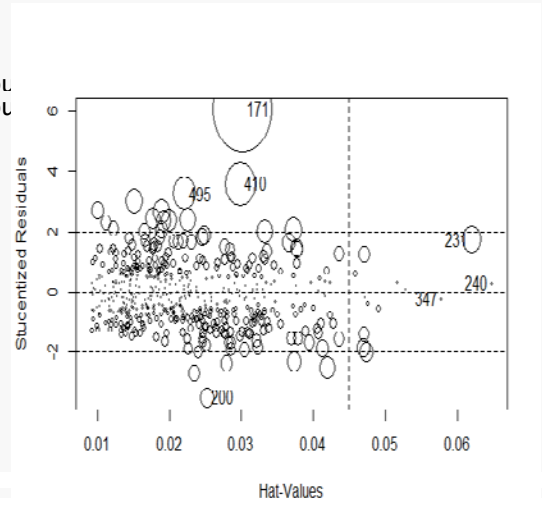
Facilita un pvalor < 0.05 per tant, es rebutja la hipòtesi nul.la alpha=0, és a dir hi ha sobredispersió. I semblaria adient emprar una modelització binomial negativa per la resposta wage (les mateixes consideracions que les aplicades al model poisson podrien plantejar-se).

```
> overdis<-sum(resid(pm2,type="pearson")^2)/df.residual(pm2);overdis
[1] 1.834662
> #library(AER)
> dispersiontest(pm2,trafo=2)

    Overdispersion test data:  pm2
z = 3.56, p-value = 0.0001855
alternative hypothesis: true alpha is greater than 0
sample estimates:
     alpha
0.09636096
```

Independentment dels resultats dels dos punts anteriors, es consideren la família de models loglineals per una resposta binomial negativa.

```
> ban2<-glm.nb(df$wage~educ+yexp*sex+f.union+f.occup+f.south,data=df)
> summary(ban2)

Call:
glm.nb(formula = df$wage ~ educ + yexp * sex + f.union + f.occup +
    f.south, data = df, init.theta = 13.58562333, link = log)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.270987   0.178662   7.114 1.13e-12 ***
educ              0.073536   0.010141   7.251 4.13e-13 ***
yexp              0.013959   0.002215   6.301 2.96e-10 ***
sexfemale        -0.074506   0.070496  -1.057 0.290562
f.unionfU.Yes     0.176540   0.049561   3.562 0.000368 ***
f.occupOcc.Sales -0.384293   0.092267  -4.165 3.11e-05 ***
f.occupOcc.Cleric -0.282168   0.074927  -3.766 0.000166 ***
f.occupOcc.Service -0.412310   0.081492  -5.059 4.20e-07 ***
f.occupOcc.Profl -0.096762   0.067495  -1.434 0.151680
f.occupOcc.Other -0.243425   0.073556  -3.309 0.000935 ***
f.southfS.Yes     -0.094089   0.043195  -2.178 0.029390 *
yexp:sexfemale    -0.007481   0.003131  -2.390 0.016869 *
---
(Dispersion parameter for Negative Binomial(13.5856) family taken to be 1)

    Null deviance: 787.62  on 533  degrees of freedom
Residual deviance: 495.48  on 522  degrees of freedom
AIC: 2878.6
              Theta:  13.59
          Std. Err.:  1.96
 2 x log-likelihood:  -2852.603
> bn0<-glm(df$wage~1,family=neg.bin(13.5856),data=df)
> bn1<-glm(df$wage~educ+yexp+sex+f.union+f.occup+f.south,family=neg.bin(13.5856),data=df)
> bn2<-glm(df$wage~educ+yexp*sex+f.union+f.occup+f.south,family=neg.bin(13.5856),data=df)
> summary(bn2)

Call: glm(formula = df$wage ~ educ + yexp * sex + f.union + f.occup +
    f.south, family = neg.bin(13.5856), data = df)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.270988   0.185587   6.848 2.11e-11 ***
educ              0.073536   0.010534   6.981 8.97e-12 ***
yexp              0.013959   0.002301   6.066 2.52e-09 ***
sexfemale        -0.074506   0.073228  -1.017 0.309406
f.unionfU.Yes     0.176539   0.051482   3.429 0.000653 ***
```

```
f.occupOcc.Sales   -0.384293   0.095843  -4.010 6.97e-05 ***
f.occupOcc.Cleric  -0.282168   0.077831  -3.625 0.000317 ***
f.occupOcc.Service -0.412310   0.084651  -4.871 1.48e-06 ***
f.occupOcc.Profl   -0.096763   0.070111  -1.380 0.168134
f.occupOcc.Other   -0.243424   0.076407  -3.186 0.001529 **
f.southfS.Yes      -0.094089   0.044870  -2.097 0.036480 *
yexp:sexfemale     -0.007481   0.003252  -2.300 0.021820 *

(Dispersion parameter for Negative Binomial family taken to be 1.079016)

    Null deviance: 787.62  on 533  degrees of freedom
Residual deviance: 495.48  on 522  degrees of freedom
AIC: 2876.6

> anova(bn1,bn2,test="F")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.south
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1       523     501.04
2       522     495.48  1   5.5618 5.1545 0.02359 *
> anova(bn1,bn2,test="Chisq")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.sou
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.sou
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       523     501.04
2       522     495.48  1   5.5618  0.02319 *
---
> par(mfrow=c(1,2))
> marginalModelPlot(bn2,id.n=5)
> influencePlot(bn2,id.n=3)
      StudRes        Hat      CookD
171  6.0654277 0.03022336 0.43572090
200 -3.5353233 0.02519865 0.10484709
231  1.7271717 0.06198854 0.13777996
240  0.2598237 0.06474770 0.01893575
347 -0.2489389 0.05777099 0.01631582
410  3.5821312 0.02982469 0.22477703
495  3.2403501 0.02214380 0.17111403
```



>

9. Contrasteu si l'efecte dels anys d'experiència depèn del gènere. Interpreteu l'efecte dels anys d'experiència en el model adient. Indiqueu el contrast emprat i interpreteu-lo.

Els test de la deviança entre el model bn1 i el bn2 contrastaria la necessitat de la interacció: cal fer el ***test usant Fisher*** *el paràmetre de dispersió no és 1 en els models binomial negatius. Concretament val 1.079016. El p valor és 2.3%, per tant es rebutja la HO, els dos models no són equivalents i la interacció és significativa.*

Pels homes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 1.40% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
Per les domes, el sou per hora s'incrementa per cada any addicional d'experiència (yexp) en un 0.65% dins del mateix grup i valors de la resta de variables explicatives (ceteris paribus).
Són interpretacions molt similars a les realitzades amb el model lm2 (loglineal) i pm2 (Poisson).

```
> anova(bn1,bn2,test="F")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.south
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
1       523     501.04
2       522     495.48  1   5.5618 5.1545 0.02359 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> coef(bn2)
      (Intercept)              educ              yexp         sexfemale      f.unionfU.Yes
       1.270987689       0.073535946       0.013958781      -0.074506337       0.176538673
  f.occupOcc.Sales  f.occupOcc.Cleric f.occupOcc.Service   f.occupOcc.Profl   f.occupOcc.Other
      -0.384292962      -0.282168048      -0.412310292      -0.096763000      -0.243424407
     f.southfS.Yes     yexp:sexfemale
      -0.094088558      -0.007481239
> 100*(exp(coef(bn2)[3])-1)
    yexp
1.405666
> 100*(exp(coef(bn2)[3]+coef(bn2)[12])-1)
     yexp
0.6498567
```

10. Quin seria el sou per hora per un professional home i dona que no està sindicat de l'estat de Washington en la mitjana de les variables explicatives numèriques?

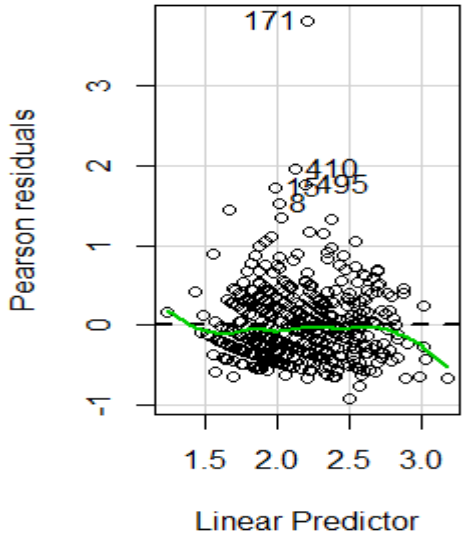
Les dades del summary inicial permeten conèixer els valors mitjos de les covariants del model. Els resultats d'aplicar els coeficients estimats pel model bn2 als valors mitjos de les covariants i als grups indicats facilita el valor del predictor lineal a on cal desfer la transformació logarítmica i permet obtenir un sou estimat pel cas d'una dona de 7.99 $/h i en cas dels homes de 9.83$/h.

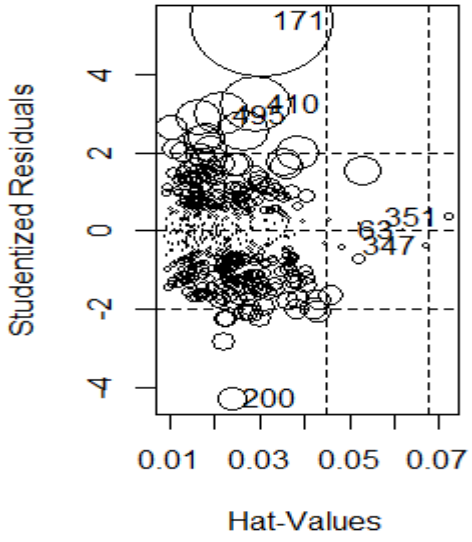
```
> newdf<-
data.frame(educ=13.02,yexp=17.82,sex="female",f.union="fU.No",f.occup="Occ.Profl",f.s
outh="fS.Yes")
> newdf<-
rbind(newdf,data.frame(educ=13.02,yexp=17.82,sex="male",f.union="fU.No",f.occup="Occ.
Profl",f.south="fS.Yes"))
> predict(bn2,newdata=newdf,type="response")
       1        2
7.992452 9.838661
```

Independentment dels resultats dels dos punts anteriors, es consideren la família de models loglineals per una resposta gamma.

```
> gm0<-glm(df$wage~1,family=Gamma(link=log),data=df)
> gm1<-glm(df$wage~educ+yexp+sex+f.union+f.occup+f.south,family=Gamma(link=log),data=df)
> gm2<-glm(df$wage~educ+yexp*sex+f.union+f.occup+f.south,family=Gamma(link=log),data=df)
> summary(gm2)

Call:glm(formula = df$wage ~ educ + yexp * sex + f.union + f.occup +
    f.south, family = Gamma(link = log), data = df)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.277074   0.185009   6.903 1.48e-11 ***
educ                0.072531   0.010460   6.934 1.21e-11 ***
yexp                0.014688   0.002366   6.209 1.09e-09 ***
sexfemale          -0.085795   0.071417  -1.201 0.230171
f.unionfU.Yes       0.198346   0.053671   3.696 0.000243 ***
f.occupOcc.Sales   -0.407616   0.096234  -4.236 2.69e-05 ***
f.occupOcc.Cleric  -0.268002   0.080282  -3.338 0.000903 ***
f.occupOcc.Service -0.415177   0.085368  -4.863 1.53e-06 ***
f.occupOcc.Profl   -0.085718   0.076563  -1.120 0.263408
f.occupOcc.Other   -0.261502   0.079938  -3.271 0.001141 **
f.southfS.Yes      -0.100153   0.043713  -2.291 0.022351 *
yexp:sexfemale     -0.007681   0.003206  -2.396 0.016941 *
---
 (Dispersion parameter for Gamma family taken to be 0.2032316)

    Null deviance: 149.559  on 533  degrees of freedom
Residual deviance:  94.431  on 522  degrees of freedom
AIC: 2831.9

> anova(gm1,gm2,test="F")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.south
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
```

```
1       523     95.503
2       522     94.431  1   1.0718 5.2739 0.02204 *
---
> anova(gm1,gm2,test="Chisq")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.south
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       523     95.503
2       522     94.431  1   1.0718  0.02165 *
---
> gs<-gamma.shape(gm2, it.lim = 20,verbose = F);gs
Alpha: 5.816485
> gd<-gamma.dispersion(gm2);gd
[1] 0.1719251
> par(mfrow=c(1,2))
> residualPlot(gm2,id.n=5)
> influencePlot(gm2,id.n=3)
```






```
        StudRes        Hat        CookD
63   0.01230824 0.06215793 0.0008652071
171  5.35155532 0.03073788 0.4411728262
200 -4.26571035 0.02404034 0.0933797754
> df[row.names(ll),]
    educ f.south    sex yexp f.union wage age     f.race     f.occup f.sector f.married
63     3   fS.Yes   male   55    fU.No    7  64   race.His   Occ.Other  fS.Manu    fM.Yes
171   14    fS.No female    1    fU.No   44  21 race.white     Occ.Man fS.Other     fM.No
200   12    fS.No   male   24    fU.No    1  42 race.white     Occ.Man fS.Other    fM.Yes
347    4    fS.No   male   54    fU.No    6  64 race.white Occ.Service fS.Other    fM.Yes
351    2    fS.No   male   16    fU.No    4  24   race.His Occ.Service fS.Other     fM.No
410   14    fS.No   male    4   fU.Yes   25  24   race.His Occ.Service fS.Other     fM.No
495   16   fS.Yes female    5    fU.No   25  27 race.white   Occ.Profl fS.Other     fM.No
```

```
        StudRes        Hat        CookD
347 -0.36266488 0.06729734 0.0253288415
351  0.36194319 0.07189593 0.0287240638
410  3.23659567 0.02932427 0.2211679412
495  2.96572463 0.02168265 0.1703505117
```

>

```
> AIC(lm2,gm2,bn2,gm2)
    df       AIC
lm2 13  623.7507
gm2 12 2983.8478
bn2 12 2876.6027
gm2 13 2831.9413
```

11. Contrasteu si l'efecte dels anys d'experiència depen del gènere. Indiqueu el contrast emprat i interpreteu-lo.

Si segueix essent significatiu. Un altre cop cal emprar el test de la deviança usant el test per Fisher donat que el paràmetre de dispersió de la gamma no és 1, sinó 0.203.

```
> anova(gm1,gm2,test="F")
Analysis of Deviance Table

Model 1: df$wage ~ educ + yexp + sex + f.union + f.occup + f.south
Model 2: df$wage ~ educ + yexp * sex + f.union + f.occup + f.south
  Resid. Df Resid. Dev Df Deviance      F  Pr(>F)
```

```
1       523     95.503
2       522     94.431  1   1.0718 5.2739 0.02204 *
```

12. Quin seria el sou per hora per un professional home i dona que no està sindicat de l'estat de Washington en la mitjana de les variables explicatives numèriques?

Les dades del summary inicial permeten conèixer els valors mitjos de les covariants del model. Els resultats d'aplicar els coeficients estimats pel model gm2 als valors mitjos de les covariants i als grups indicats facilita el valor del predictor lineal a on cal desfer la transformació logarítmica i permet obtenir un sou estimat pel cas d'una dona de 7.96 $/h i en cas dels homes de 9.94$/h. Molt similars als valors predits per la binomial negativa.

```
> newdf<-
data.frame(educ=13.02,yexp=17.82,sex="female",f.union="fU.No",f.occup="Occ.Profl",f.s
outh="fS.Yes")
> newdf<-
rbind(newdf,data.frame(educ=13.02,yexp=17.82,sex="male",f.union="fU.No",f.occup="Occ.
Profl",f.south="fS.Yes"))
> predict(gm2,newdata=newdf,type="response")
       1        2
7.961595 9.947219
> predict(bn2,newdata=newdf,type="response") # binomial negativa
       1        2
7.992452 9.838661
```

13. Trieu quina de les tentatives il.lustrades per modelar el sou per hora sembla més adient. Justifiqueu la resposta.

L'AIC del model gaussià no és comparable, però per la resta d'opcions poisson, binomial negativa i gamma, el millor valor, mínim, correspon a l'opció gamma que també resulta la més raonable donada la tipologia de la variable (no és un comptatge).

```
> AIC(lm2,pm2,bn2,gm2)
    df       AIC
lm2 13  623.7507
pm2 12 2983.8478
bn2 12 2876.6027
gm2 13 2831.9413
```

14. Feu pel millor model triat una diagnosi dels residus i valors influents en funció dels resultats numèrics i gràfics facilitats. Indiqueu les observacions que són outliers dels residus i/o valors influents.

Els residus de la proposta Poisson són esgarrifosos (residus studentitzat de valor absolut 8!!!), però en la resta de casos són molt similars. El model gamma aconsegueix capturar les particularitats de l'observació 200 que no queda ben reflectida en la proposta loglineal (totes les altres també la capturen). Aquesta observació té un sou per hora de 1$/h, increïblement baix, però els models no loglineal, com el gamma, el capturen. El model gamma gm2 segueix sobreajustant els salaris més elevats.
L'observació 171 és un outlier dels residus i alhora dada influent, és una dona molt jove, sense experiència que té un sou de 44 $/h, poc particular és l'observació i queda mal ajustada pel model gamma (i totes les altres alternatives provades).

De les 2 possibilitats més adients, loglineal i gamma, a la vista dels residus, jo triaria la proposta gamma per modelar wage, sou per hora.

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. Presència de petit comerç: Es seleccionen 20 barris del municipi de Barcelona i dintre de cada barri es seleccionen 5 seccions censals. Es recull informació de la superfície de la zona i del perfil de la població de cada secció censal (taules d'edat i sexe, nivell cultural, procedència de la població i renda familiar mitjana). També es prenc nota del número de petits comerços que hi ha en cada secció censal.

2. Incidència de peces defectuoses: En una fàbrica es fa un disseny experimental per determinar els factors que determinen que la peça fabricada sigui defectuosa. A l'empresa hi ha dues màquines i 3 operaris. Cada operari treballa amb cada màquina i en cada procés es registren les condicions ambientals i mesures de les propietats de la matèria primera que es fa servir, que és diferent en cada execució.

3. Efectivitat d'una dieta: Es trien 30 individus amb característiques homogènies. Aleatòriament s'assignen de forma balancejada a un grup control i a un de tractament on s'aplica una dieta alimentària. Al llarg de 6 mesos es fa un seguiment quinzenal on a més de registrar el pes de cada individu es pren nota de variables analítiques per mesurar el seu estat de salut.

4. Reclamació de garantia: Una tenda d'electrodomèstics té l'historial de venda dels aparells que ha venut, indicant el tipus d'aparell, la marca, el preu de venta i el codi de client amb dades de gènere i edat. Té un total de 2.020 articles venuts a una cartera de 1.250 clients. També per a cada aparell té registrat si s'ha fet una reclamació de garantia per avaria. Es vol veure els factors dels aparells que determinen que hi hagi reclamacions.

5. Assistència telefònica: Una companyia de telecomunicacions contracta un "*call center 24h*" per donar suport als seus clients per telèfon. Hi ha tres línies temàtiques: contractació, avaries i consultes genèriques. Cada línia té sub-línies per idioma (castellà, català i anglès). Al llarg del dia hi ha 3 torns (matí, tarda i nit). Per a cada torn i cada sub-línia es conta el número de trucades rebudes diàriament al llarg d'una setmana, per determinar els factors que fan que hi hagi més trucades.

1. *Model lineal mixt generalitzat amb resposta poisson (número de comerços en una secció censal) y factor aleatori d'agrupació (barris) a 20 nivells. Possibles factors fixos: superfície i dades del perfil de població.*
2. *Modelo lineal generalitzat amb resposta Binària (peça defectuosa o no). Com només hi ha dos operaris i dues màquines i es vol determinar els factors que fan augmentar la probabilitat de defectes, el més correcte és considerar-los fixos. L'estat ambiental i les característiques de la matèria primera són també covariants.*
3. *Model lineal mixt amb resposta gaussiana (pes). El factor d'agrupació és l'individu, a 30 nivells, ja que és un estudi longitudinal d'una mostra d'individus. Les variables analítiques poden actuar com a covariants i el tractament (dieta) és clarament el factor fix d'interès.*
4. *Model lineal mixt generalitzat amb resposta binària (reclamació) i factor aleatori seria el client (hi ha més aparells venuts que clients, indicant que hi ha clientes que han comprat més d'un aparell). Com els factors que es vol analitzar fan referència a l'aparell, els clients són factors aleatoris. El tipus, marca i preu de l'aparell serien factors fixos.*
5. *Modeli lineal mixt generalitzat amb resposta de Poisson (número de trucades rebudes) i factor d'agrupació (dia) a 7 nivells, ja que es fa un seguiment al llarg d'una setmana. Els factors fixos són les línies, idioma i la franja considerada.*