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

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

(Data: 2/07/2014 a les 15:00h Aula 003-FME)

Nom de l´alumne:	**DNI:**
Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	Edifici C5 D217 o H6-67
Normativa:	SÓN PERMESOS APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	3h 00 min
Sortida de notes:	Abans del ¿? al Web Docent de MLGz
Revisió de l'examen:	8 de Julio 15h – Pr 2 C5-217-C Nord o Pr 1 i 3 H- P6-67

Problema 1 (4 punts): Resposta normal

Se han recogido datos de 50 pisos que se ofrecen en alquiler en Barcelona (datos de 2003). La variable Tamaño es la superficie en m2. El precio del alquiler en euros es la variable respuesta. Habitaciones es el número de habitaciones y Baños, el número de baños. Altura es la planta donde se encuentra el piso. Las variables de equipamiento (Ascensor, Calefaccion, Aire Acondicionado, Exterior, Amueblado) son binarias e indican(0=No/1=Sí).

```
> summary(pisos)
     Tamaño           Precio          Ascensor    Altura  Habitaciones   Calefaccion
 Min.   : 30.00   Min.   : 600.0   Min.   :0.00   1: 7   Min.   :1.00   Min.   :0.00
 1st Qu.: 56.25   1st Qu.: 727.5   1st Qu.:1.00   2:31   1st Qu.:1.00   1st Qu.:0.00
 Median : 77.50   Median : 850.0   Median :1.00   3:12   Median :2.00   Median :0.00
 Mean   : 76.36   Mean   : 932.4   Mean   :0.82          Mean   :2.24   Mean   :0.48
 3rd Qu.: 95.00   3rd Qu.:1009.4   3rd Qu.:1.00          3rd Qu.:3.00   3rd Qu.:1.00
 Max.   :120.00   Max.   :2350.0   Max.   :1.00          Max.   :5.00   Max.   :1.00
   AireAcond      Exterior       Baños        Amueblado
 Min.   :0.0   Min.   :0.0   Min.   :1.00   Min.   :0.0
 1st Qu.:0.0   1st Qu.:0.0   1st Qu.:1.00   1st Qu.:0.0
 Median :0.0   Median :1.0   Median :1.00   Median :0.0
 Mean   :0.2   Mean   :0.7   Mean   :1.32   Mean   :0.1
 3rd Qu.:0.0   3rd Qu.:1.0   3rd Qu.:2.00   3rd Qu.:0.0
 Max.   :1.0   Max.   :1.0   Max.   :2.00   Max.   :1.0
```

Se ajustan 2 modelos sin considerar interacciones: uno con todas las variables y otro seleccionado mediante eliminación "backward" y criterio AIC.

Modelo A:

```
Call:
lm(formula = log(Precio) ~ ., data = pisos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38828 -0.09146  0.00288  0.07087  0.41675

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.108818   0.112717  54.196  < 2e-16 ***
Tamaño        0.006143   0.001487   4.132 0.000184 ***
Ascensor      0.061183   0.066296   0.923 0.361742
Altura2      -0.030552   0.081272  -0.376 0.709013
Altura3       0.173593   0.087732   1.979 0.054945 .
Habitaciones -0.010854   0.034843  -0.312 0.757058
Calefaccion   0.042015   0.064599   0.650 0.519250
AireAcond     0.186881   0.065688   2.845 0.007041 **
Exterior     -0.028890   0.056993  -0.507 0.615074
Baños         0.099189   0.081028   1.224 0.228243
Amueblado    -0.007238   0.091750  -0.079 0.937526
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1638 on 39 degrees of freedom
Multiple R-squared:  0.7303,     Adjusted R-squared:  0.6612
F-statistic: 10.56 on 10 and 39 DF,  p-value: 2.37e-08
```

Modelo B:

```
Call:
lm(formula = log(Precio) ~ Tamaño + Altura + AireAcond + Baños,
    data = pisos)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41277 -0.09199 -0.00114  0.09718  0.41792

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.096057   0.090595  67.289  < 2e-16 ***
Tamaño       0.006164   0.001290   4.777 2.01e-05 ***
Altura2     -0.044450   0.070636  -0.629  0.53242
Altura3      0.176584   0.077857   2.268  0.02829 *
AireAcond    0.191312   0.058844   3.251  0.00221 **
Baños        0.131924   0.068591   1.923  0.06092 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1569 on 44 degrees of freedom
Multiple R-squared:  0.7209,     Adjusted R-squared:  0.6892
F-statistic: 22.73 on 5 and 44 DF,  p-value: 3.385e-11
```

Comenta las siguientes afirmaciones y decide si son correctas o no o si necesitan alguna matización, justificando tu respuesta:

1. "Seleccionaríamos el modelo A ya que tiene una R-squared (73.03%) superior a la del modelo B (72.09%)"

El primer modelo tiene una R-sq superior pero también tiene más parámetros, alguno de ellos no significativo. De hecho, el segundo modelo es un caso particular del primero fijando a cero una serie de coeficientes. Así pues, entre dos modelos jerarquizados siempre la R-sq será mayor con el modelo que tenga más parámetros. Por lo tanto, la R-sq no es un buen criterio para seleccionar entre dos modelos. En cambio, la R-sq ajustada es un criterio de información que penaliza la inclusión de parámetros con efecto no significativo. Cuanto mayor es la R-sq ajustada indica un mejor compromiso entre lo bien que ajusta el modelo y lo parsimonioso que es. Para seleccionar entre estos dos modelos, els criterio adecuado es la R-sq ajustada que es mayor en el modelo 2, el cual es el que debe ser seleccionado.

2. "En el modelo B, usando un nivel de significación alpha de 0.10, la variable Altura2 es la única que no es significativa y por ello se debe eliminar del modelo"

La variable Altura2 es una variable auxiliar o dummy que crea R a partir de un contraste de tipo baseline con categoría de referencia igual a la primera clase de la variable categórica. Esta variable es una variable indicadora para identificar los casos en que la altura del piso es un segundo. Esta variable está asociada a otra dummy que es Altura3 que corresponde al indicador de que el piso es un tercero. Para decidir la significación de la variable se debe realizar un test que combine ambas variables simultáneamente. El método apropiado es el método Anova que compara el modelo sin la variable categórica Altura con el modelo que la incluye y que por tanto incorpora dos coeficientes más. La significación marginal de un coeficiente ligado a un contraste solo podría implicar la posibilidad de reagrupar categorías para eliminar el parámetro, pero lo más conveniente es comprobar la significación de la variable completa.

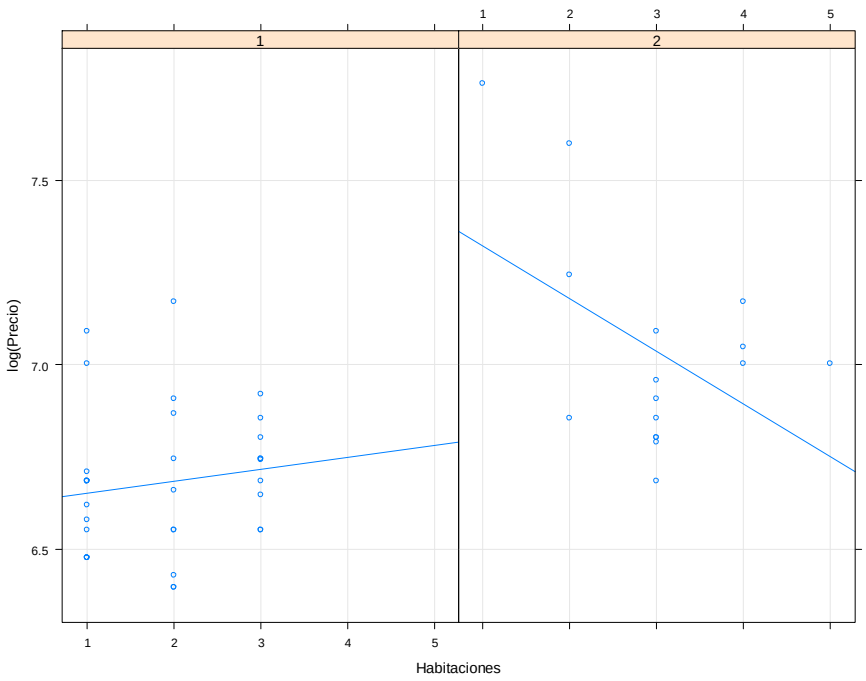
3. "Como la respuesta está en logaritmos se ha corregido el problema de heterocedasticidad de residuos, por lo que en el modelo B ya se puede afirmar que los residuos tienen varianza constante"

La transformación logaritmo, como caso particular de transformación de Box-Cox permite en algunos casos resolver el problema que se detecta cuando se observa que la varianza de los residuos crece cuando crece la predicción del modelo. Sin embargo, podría haber comportamientos de heterocedasticidad de varianza, tal vez ligado a variables explicativas, que no se hayan corregido con la transformación logaritmo. Por ello, aún cuando se ha cambiado la escala de la respuesta es necesario verificar mediante el análisis de los residuos que la varianza de éstos es constante.

4. “Puesto que el coeficiente de Aire Acondicionado en el modelo B es 0.1913 y la respuesta está en logaritmos, podemos interpretar que si un piso tiene aire acondicionado, el precio de su alquiler es un 19.13% más caro que otro piso con las mismas características pero sin aire acondicionado”

Teniendo en cuenta que la respuesta está en logaritmos, las variables predictoras, que son lineales en la escala del logaritmo, pasan a ser multiplicativas en la escala real, pero para conocer el factor es necesario hacer la exponencial del coeficiente. Así pues, exp(0.1913)=1.2108, lo que implica que en la escala de la respuesta el aumento de precio cuando el piso tiene aire acondicionado y manteniendo el resto de características constantes es de un 21.08%

Para estos datos, ajustamos un modelo que incluye como predictores el número de habitaciones y el número de baños. Para la parte descriptiva se realizan dos gráficos por separado que relacionan el logaritmo del precio con el número de habitaciones, separando según el piso tenga uno o dos baños (se considera variable categórica):



El ajuste con R del modelo que incluye la interacción entre las dos variables es el siguiente:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.35160 -0.17205  0.00128  0.10372  0.48627

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         6.61864    0.09256  71.503  < 2e-16 ***
Habitaciones        0.03261    0.04498   0.725 0.472155
Baños2              0.84595    0.20529   4.121 0.000156 ***
Habitaciones:Baños2 -0.17540    0.07364  -2.382 0.021420 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2182 on 46 degrees of freedom
Multiple R-squared:  0.4357,     Adjusted R-squared:  0.3989
F-statistic: 11.84 on 3 and 46 DF,  p-value: 7.165e-06
```

5. Indica los modelos que se obtienen para predecir el precio en función del número de habitaciones cuando hay uno y dos baños. Haz una interpretación de estos modelos a partir de la significación de los parámetros del modelo obtenido. ¿Es razonable el resultado obtenido?

Para un piso con un único baño, la variable dummy Baños2 vale cero y por lo tanto el modelo es:

$$\log(precio) = 6.61864 + 0.03261 * Habitaciones + \varepsilon$$

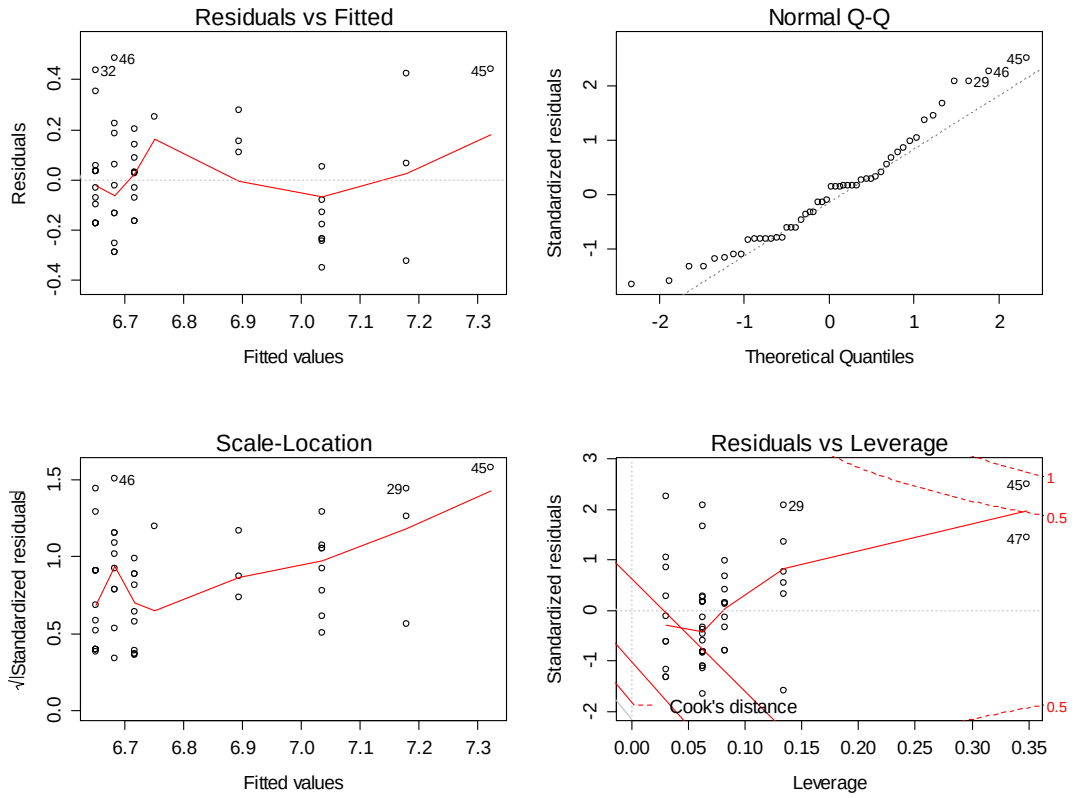
Aunque la pendiente no es significativa, lo cual ya se observa en el gráfico de la izquierda donde la recta ajustad es prácticamente horizontal, lo cual se interpreta como que si el piso tiene un único baño, el número de habitaciones no es un factor que suponga un incremento en el precio del alquiler.

En el caso de un piso con dos baños, donde la variable Baños2 vale 1, el modelo obtenido es:

$$\log(precio) = (6.61864 + 0.84595) + (0.03261 - 0.17540) * Habitaciones + \varepsilon$$
$$= 7.46459 - 0.14279 * Habitaciones + \varepsilon$$

Aunque no tenemos un p-valor para el coeficiente de la pendiente obtenida, teniendo en cuenta la significación de la interacción es muy probable que la pendiente sea significativamente negativa.
En este caso, la pendiente es negativa de forma sorprendente, ya que indica que en los pisos con 2 baños se espera que cuantas más habitaciones, menor sea el precio del alquiler. Por cada habitación de más que tenga, el precio desciende a un 88.7% , ya que exp(-0.14279)= 0.8669361. En el gráfico del modelo ya se observa esa pendiente negativa, pero hay una única observación con una habitación que presenta un precio de alquiler muy elevado (es el piso más caro), lo que supone un claro efecto palanca en la estimación del modelo.

Los plots del análisis de residuos para la validación son los siguientes:



6. Realiza la validación del modelo indicando las premisas que se validan en cada plot.

El primer plot corresponde a los residuos frente a las predicciones del modelo. Este plot nos permite evaluar la linealidad de los datos y la varianza constante de los residuos, si observamos una disposición aleatoria de los puntos en el gráfico y sin cambios en la variabilidad. Este plot también permite detectar valores atípicos correspondientes a observaciones mal explicadas por el modelo.
El segundo plot es el plot de normalidad, para verificar el supuesto de que los residuos provienen de una distribución normal.

El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos estandarizados frente a las predicciones. Es similar al primero pero permite incidir en el análisis de la variabilidad para comprobar la hipótesis de homocedasticidad (varianza constante). Tanto este plot como el primero incluyen un ajuste suave para facilitar la interpretación.
El último plot refleja las componentes de la medida de influencia (distancia de Cook). En el eje de abscisas se refleja el factor de apalancamiento (leverage) y en el eje de ordenadas, el residuo estandarizado. Además se incluyen curvas de nivel para indicar la posición relativa de cada observación según su distancia de Cook.
En el primer y segundo plot aparecen 3 observaciones con un residuo estandarizado superior a 2, que corresponde a casos mal explicados por el modelo. El plot de normalidad apunta una posible asimetría en la distribución de los residuos, puesto que los puntos presentan una cierta curvatura en el plot.
El tercer plot parece apuntar un posible incremento de la varianza, pero hay que tener en cuenta que la observación con la predicción más alta es un valor atípico y se haya aislada, lo cual distorsiona el plot a la hora de confirmar un aumento claro de varianza. Finalmente, el último plot pone de manifiesto dos observaciones con un factor de anclaje alto y en donde una de ellas tiene un residuo estandarizado superior a 2, por lo que la distancia de Cook de esta observación es alta, indicando que es un caso muy influyente que afecta a la estimación del modelo. El modelo no es del todo válido porque los residuos no parecen normales y aparecen atípicos y observaciones claramente influyentes.

Los casos que aparecen etiquetados en alguno de los plots son los siguientes:

	Tamaño	Precio	Ascensor	Altura	Habitaciones	Calefaccion	AireAcond	Exterior	Baños	Amueblado
29	120	2000	1	3	2	1	1	1	2	0
45	100	2350	1	3	1	1	1	1	2	0
32	95	1200	1	1	1	0	0	0	1	0
47	90	1100	1	2	5	1	0	1	2	0

7. Teniendo en cuenta los plots de validación y los plots de los modelos anteriores, interpreta para cada uno de estos casos si se trata de un dato atípico i/o influyente y si lo es, si es influyente a priori o a posteriori. ¿Qué efecto tienen cada uno de ellos en la estimación del modelo? En concreto identifica cuál de ellos condiciona la interpretación obtenida en el punto 5, justificando la respuesta.

El primer caso (29) aparece en el plot de normalidad con un valor de residuo estandarizado levemente superior a 2, lo cual lo caracteriza como dato atípico. El factor de apalancamiento (leverage), sin ser extremo es de los más grandes. Su situación respecto a las curvas de nivel, lo sitúan próximo a una distancia de Cook de 0.5. Es un piso con 2 baños y 2 habitaciones que posee un precio elevado (2000€) corresponde al valor más alto de los pisos con 2 habitaciones en el gráfico del modelo lineal para pisos con dos baños. No es de los datos más influyentes pero si supone una cierta influencia a posteriori que puede explicar por qué la pendiente del segundo modelo es negativa
La segunda observación (45) es claramente el caso más influyente del modelo, corresponde al leverage más alto y su residuo estandarizado lo caracteriza como atípico, lo cual implica una distancia de Cook superior a 0.5. Es claramente un caso influyente a posteriori ya que la estimación del modelo se ve altamente influida por este individuo. Corresponde a un piso con 2 baños y una única habitación y que supone el alquiler más alto de toda la muestra. Es el punto superior en el segundo gráfico de los modelos lineales. Claramente, su presencia en la muestra es influyente, claramente a posteriori. Corresponde a un piso con dos baños y una sola habitación que posee el alquiler más alto de la muestra y que influye en la estimación de la pendiente del modelo para pisos con dos baños. Su efecto de anclaje hace que la pendiente correspondiente sea más negativa de lo que se obtendría si se suprime esta observación. Esta observación parece la principal causante del sorprendente resultado obtenido en el apartado anterior.
El tercer caso (32) tiene un residuo estandarizado ligeramente superior a 2 pero no parece tener un leverage alto, ni tampoco una distancia de Cook elevada. Únicamente supone una observación atípica pero no influyente.
Por último, la observación 47 es un caso de leverage alto, pero no posee un residuo estandarizado excesivo, haciendo que sea una observación influyente a priori pero que no está mal explicada por el modelo. Es el único piso con cinco habitaciones lo que implica un factor de anclaje elevado.

Problema 2 (4 punts)

L'extensió del R amb el package MASS conté el *dataframe* `insurance` que consisteix en el nombre de sinistres entre els clients (pòlisses) d'una companyia d'assegurances d'automòbils britànica a l'any 1973. La descripció de les columnes és la següent:

`District`	district of policyholder (1 to 4): 4 is major cities (London).
`Group`	group of car (1 to 4), <1 litre, 1–1.5 litre, 1.5–2 litre, >2 litre.
`Age`	of driver in 4 ordered groups, <25, 25–29, 30–35, >35.
`Holders`	numbers of policyholders (pòlisses)
`Claims`	numbers of claims (sinistres)

Font: L. A. Baxter, S. M. Coutts and G. A. F. Ross (1980) Applications of linear models in motor insurance. *Proceedings of the 21st International Congress of Actuaries*, *Zurich* pp. 11–29

Les dades corresponen a 23359 pòlisses on s'han donat 3151 sinistres. Els autors indicats com a font, analitzen les dades usant models loglineals amb el nombre de sinistres com a resposta i el nombre de pòlisses com a offset. A continuació apareixen uns quants resultats de models estimats amb R.

1. Per què penseu que els autors usen models loglineals en comptes de tractar el nombre de sinistres rebudes del total de pòlisses per grup com una resposta binomial i usar models de resposta binària.

   La resposta binària no és apropiada en principi al no ser dicotòmica la resposta, doncs una pòlissa pot tenir més d'un sinistre en un any, malgrat a la vista del sumari de les dades el nb mig de sinistres és petit i la probabilitat de recollir més d'un sinistre en un any és a la pràctica molt petita.

2. El Factor Districte sembla que té una contribució més feble que la resta de variables explicatives del model additiu. Justifiqueu estadísticament si és possible suprimir la variable.

   L'efecte net del Districte és estadísticament significatiu amb p-valor de 0.003, per tant no pot suprimir-se:

```
> anova(baxter.m4,baxter.ma,test="Chi")
Analysis of Deviance Table

Model 1: Claims ~ offset(logtamany) + Group + Age
Model 2: Claims ~ offset(logtamany) + District + Age + Group
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        57     65.291
2        54     51.420  3   13.871     0.003
```

3. A la vista dels estimadors, valoreu la possibilitat d'agrupar el Factor District en un factor dicotòmic constituït per una categoria de gran ciutat (London) i com alternativa la resta de districtes (un districte és una demarcació geogràfica que perfectament podeu assimilar a comarca en aquestes dades) i a més considereu la definició de la variable cedat amb valors 1, 2, 3 i 4 segons el grup d'edat i el tractament de l'edat com a covariable. Determineu la bondat de l'ajust en el model additiu simplificat.

   Cal calcular la deviança residual del model additiu simplificat. Amb les dades disponibles, la deviança explicada del model additiu no simplificat és de 184.84 unitats amb 9 paràmetres i en canvi el model additiu simplificat explica 184.124, el contrast de la deviança entre els 2 models (no són jeràrquics) no es pot fer. La pregunta però demana la bondat de l'ajust, és a dir la deviança residual del model additiu simplicitat ha de ser 51.42+(184.84-184.00)=52.26 a contrastar amb una xi quadrat amb 58 graus de llibertat dona un p valor superior al llindar clàssic del 5% i per tant s'accepta la hipòtesi nul.la que el model s'adapta bé a les dades.

```
> 1-pchisq(58,52.26)
  0.2719875
```

4. Interpreteu el coeficient de la variable *dummy* corresponent al grup d'edat constituït per les persones més grans de 35 anys en el model additiu original i l'efecte de pertanyer al quart grup d'edat (cedat) en el simplificat. Determineu la consistència/inconsistència de la interpretació comparativa entre els dos models.

   Les pòlisses subscrites per persones amb edat >35 indiquen en el model additiu no simplificat una reducció de 1-exp(-0.53667)=0.4153 per u o 41.53% del nombre de sinistres respecte el grup de persones més joves (<25) dins del mateix grup per la resta de variables (District i Group).

Les pòlisses subscrites per persones amb edat >35 (covariable cedat amb valor 4) indiquen en el model additiu simplificat una reducció de 1-exp(-0.17616*3)=0.4105 per u o 41.05% del nombre de sinistres pel valor de la covariable cedat 4 respecte la referència dels més joves (grup de la covariable 1) dins del mateix grup per la resta de variables (District i Group). Cada salt en un grup de la covariable cedat implica una reducció en la sinistralitat de 1-exp(-0.17616)=0.1615 per un o 16.15%. Interpretant bé el significat dels valors de la covariable els dos models donen resultats molt semblants en interpretació.

5. Determineu l'estimació per punt i per interval (al nivell de confiança del 95%) del nombre predit de sinistres per pòlissa dins del grup de referència en el model additiu no simplificat.

A partir del llistat subministrat i com es demana el nb mig de sinistres predit no pas el nb total de sinistres pel grup de referència (District 1, Age <25 i Group <1 l) no calen més dades.

L'estimació per punt és exp(-1.82174)=0.1617 sinistres per any en el grup de referència.
En l'escala definida per la funció de link, l'estimació per interval donaria
-1.82174 ± 1,96 x 0.07679, és a dir (-1.9723,-1.6712) i per tant, exponenciant tots dos extrem de l'interval la predicció del nombre mig de sinistres per any en el grup de referència estaria entre 0.139 i 0.188 en un 95% dels casos.

```
glm(formula = Claims ~ offset(logtamany) + District + Age + Group,
    family = poisson(link = log))

...

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82174    0.07679 -23.724  < 2e-16 ***
...
```

6. Els models estimats pressuposen una resposta poissoniana. Com podrieu validar aquesta hipòtesi? Canviarien molt les conclusions en presència de sobredispersió?

Amb el nivell d'agregació de les dades no es pot fer res per validar gràficament l'adequació de la hipòtesi poissoniana. Caldria per cada registre una descripció de la variança mostral del nb de sinistres o si es tinguessin dades individuals caldria fer un seguit de comandes com:

```
> # Gràfic per contrastar sobredispersió: amb el nivell d'agregació
> # subministrat no resulta viable
>
> m <-tapply(Claims,paste(District,Age,Group),sum)
> m <-m/tapply(Claims,paste(District,Age,Group),length)
> m2 <-tapply(Claims^2,paste(District,Age,Group),sum)
> m2 <-m2/tapply(Claims^2,paste(District,Age,Group),length)
> v <-m2-m^2
> plot(m,v)
> lines(m,m)
```

El llistat subministrat indica una estimació del paràmetre de dispersió per sota de 1 per tant sobredispersió no n'hi ha, els estimadors obtinguts per quasi versemblança són puntualment els mateixos només canvia el seu estàndard error i per tant la inferència, però ben poc en aquest cas. A més l'ajust del model additiu no simplificat (i el simplificat també) és molt satisfactori fet que suporta positivament a la hipòtesi poissoniana.

```
> options(contrasts=c("contr.treatment","contr.treatment"))
> baxter$logtamany <- log(Insurance$Holders)
> summary(baxter)
 District    Group       Age        Holders            Claims         logtamany
 1:16     <1l   :16   <25  :16   Min.   :   3.00   Min.   :  0.00   Min.   :1.099
 2:16     1-1.5l:16   25-29:16   1st Qu.:  46.75   1st Qu.:  9.50   1st Qu.:3.844
 3:16     1.5-2l:16   30-35:16   Median : 136.00   Median : 22.00   Median :4.912
 4:16     >2l   :16   >35  :16   Mean   : 364.98   Mean   : 49.23   Mean   :4.904
                                 3rd Qu.: 327.50   3rd Qu.: 55.50   3rd Qu.:5.791
                                 Max.   :3582.00   Max.   :400.00   Max.   :8.184
```

```
baxter$logtamany <- log(Insurance$Holders)
baxter.m1 <- glm(Claims~offset(logtamany)+District, family=poisson(link=log))
baxter.m2 <- glm(Claims~offset(logtamany)+Group, family=poisson(link=log))
baxter.m3 <- glm(Claims~offset(logtamany)+Age, family=poisson(link=log))
baxter.m4 <- glm(Claims~offset(logtamany)+Group+Age, family=poisson(link=log))
baxter.m5 <- glm(Claims~offset(logtamany)+District+Age, family=poisson(link=log))

> summary(baxter.m1)
Null deviance: 236.26  on 63  degrees of freedom
Residual deviance: 223.53  on 60  degrees of freedom
AIC: 548.85
> summary(baxter.m2)

Null deviance: 236.26  on 63  degrees of freedom
Residual deviance: 147.91  on 60  degrees of freedom
AIC: 473.23

> summary(baxter.m3)

Null deviance: 236.26  on 63  degrees of freedom
Residual deviance: 155.40  on 60  degrees of freedom
AIC: 480.72

> summary(baxter.m4)

Null deviance: 236.259  on 63  degrees of freedom
Residual deviance:  65.291  on 57  degrees of freedom
AIC: 396.61

> summary(baxter.m5)

Null deviance: 236.26  on 63  degrees of freedom
Residual deviance: 140.09  on 57  degrees of freedom
AIC: 471.41

> baxter.ma <- glm(Claims~offset(logtamany)+District+Age+Group, family=poisson(link=log))
> summary(baxter.ma)

Call:
glm(formula = Claims ~ offset(logtamany) + District + Age + Group,
    family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.46558  -0.50802  -0.03198   0.55555   1.94026

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.82174    0.07679 -23.724  < 2e-16 ***
District2    0.02587    0.04302   0.601 0.547597
District3    0.03852    0.05051   0.763 0.445657
District4    0.23421    0.06167   3.798 0.000146 ***
Age25-29    -0.19101    0.08286  -2.305 0.021149 *
Age30-35    -0.34495    0.08137  -4.239 2.24e-05 ***
Age>35      -0.53667    0.06996  -7.672 1.70e-14 ***
Group1-1.5l  0.16134    0.05053   3.193 0.001409 **
Group1.5-2l  0.39281    0.05500   7.142 9.18e-13 ***
Group>2l     0.56341    0.07232   7.791 6.65e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 236.26  on 63  degrees of freedom
Residual deviance:  51.42  on 54  degrees of freedom
AIC: 388.74

> baxter$gran.ciutat <- rep(1,length(District))
> baxter$gran.ciutat[District!=4] <- 0
> baxter$cedat<-rep(4,length(District))
> baxter$cedat[baxter$Age=="<25"] <- 1
```

```
> baxter$cedat[baxter$Age=='25-29'] <- 2
> baxter$cedat[baxter$Age=='30-35'] <- 3
> baxter.mag <- glm(Claims~offset(logtamany)+gran.ciutat+cedat+Group, family=poisson(link=log))
> summary(baxter.mag)

Call:
glm(formula = Claims ~ offset(logtamany) + gran.ciutat + cedat +
    Group, family = poisson(link = log))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.51431  -0.50310  -0.05875   0.60402   2.01176

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.63733    0.07499 -21.833  < 2e-16 ***
gran.ciutat  0.21860    0.05853   3.735 0.000188 ***
cedat       -0.17616    0.01850  -9.523  < 2e-16 ***
Group1-1.5l  0.16260    0.05048   3.221 0.001276 **
Group1.5-2l  0.39389    0.05491   7.174 7.31e-13 ***
Group>2l     0.56585    0.07216   7.842 4.44e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)
```

Deviança Explicada 184.001 unitats

```
> summary(baxter.mas)

Call:
glm(formula = Claims ~ offset(logtamany) + District + Age + Group,
    family = quasi(link = log, variance = "mu"))

...
(Dispersion parameter for quasi family taken to be 0.900543)

    Null deviance: 236.26  on 63  degrees of freedom
Residual deviance:  51.42  on 54  degrees of freedom
```

Problema 3 (2 punts): 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. Variabilidad de la humedad en el terreno: Se seleccionan 15 parcelas y en cada parcela se escogen 10 puntos donde se mide la composición (contenido calcáreo y sedimentos), la salinidad y la humedad.

Modelo lineal mixto con respuesta gaussiana (humedad). Factor aleatorio que induce agrupación en los datos: parcela (10 puntos de cada parcela). Variables explicativas: composición y salinidad.

2. Inversión en bolsa: Una entidad financiera encuesta a 50 hombres y 50 mujeres y recoge datos de edad, estado civil, nivel de estudios, ingresos y si invierten en bolsa o no.

Modelo lineal generalizado con respuesta binaria (invierte o no). Variables explicativas: Sexo, edad, estado civil, nivel de estudios e ingresos.

3. Efectos secundarios de un fármaco: Se aplica un tratamiento con un fármaco a 10 pacientes y un placebo a otros 10. De todos ellos, se recogen datos de su estado físico (edad, peso, altura, analítica en sangre) y se les pide que indiquen el número de episodios total de somnolencia a lo largo de una semana.

Modelo lineal generalizado con respuesta de tipo recuento – Poisson (número de episodios de somnolencia por semana). Variables explicativas: fármaco/placebo, edad, peso, altura, analítica.

4. Volumen de correo electrónico en un servidor: Un proveedor de servicios de internet, selecciona 100 usuarios y 20 correos de cada usuario, registrando el tamaño en bytes de cada correo. De los usuarios recoge si es empresa o particular, la antigüedad en el servicio y el número de conexiones diarias.

Modelo lineal mixto con respuesta gaussiana (tamaño en bytes del correo). Factor aleatorio que induce agrupación: usuario (20 correos de cada usuario). Variables explicativas: empresa/particular, antigüedad, número de conexiones diarias.

5. Corrección ortográfica: En 15 colegios se seleccionan 20 alumnos en cada centro y se les hace una prueba de dictado. Se cuenta el número de faltas ortográficas y se registran datos del expediente escolar del alumno: notas en matemáticas, lengua, ciencias y tecnología.

Modelo lineal generalizado con respuesta de tipo recuento-Poisson (número de faltas de ortografía en el dictado). Variables explicativas: notas en matemáticas, lengua, ciencias y tecnología.