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

CURS 2016-2017 Q1 – EXAMEN REAVALUACIÓ : MODEL LINEAL GENERALITZAT

(Data: 4/07/2017 a les 15:00h Aula 002-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:	2h 30 min
Sortida de notes:	Abans del 13 Juliol a la Intranet Docent de MLGz
Revisió de l'examen:	13 Juliol a 12h a ETSEIB H- P6-67

Problema 1 (3 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 cirterio AIC.

Modelo A:

```
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:

```
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 0.05, las variables Altura2 y Baños no son significativas y por ello se deben 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. La variable Baños sí se puede eliminar del modelo.

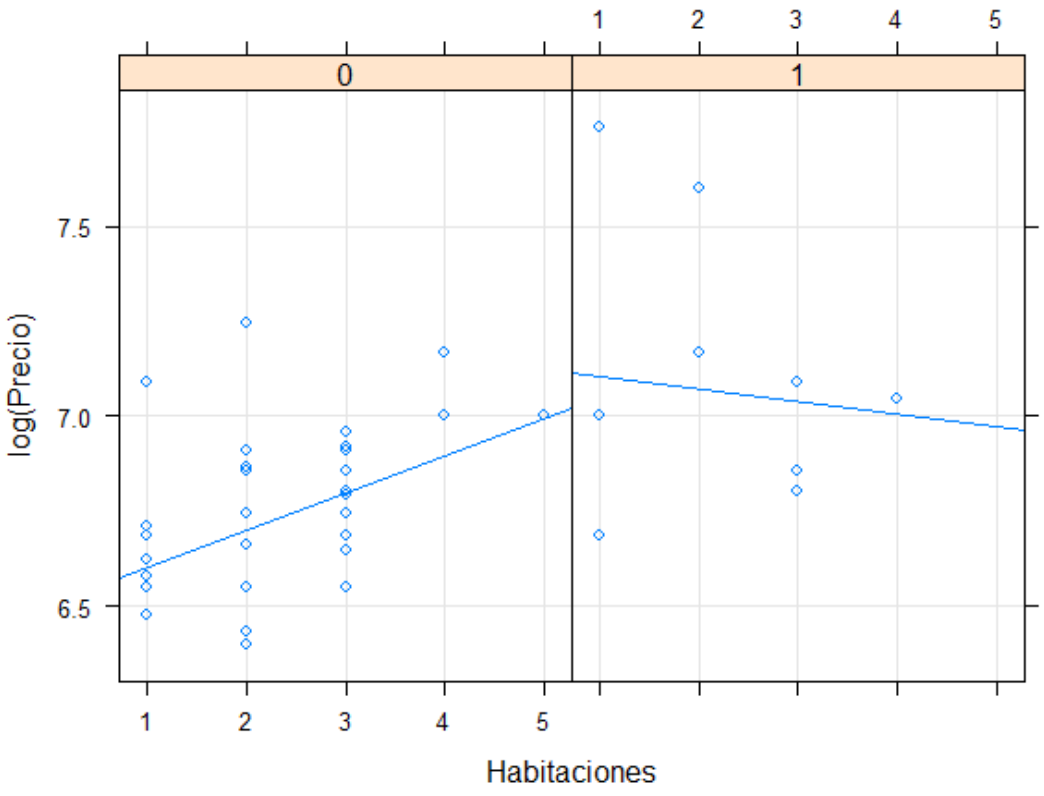
3. “Como la respuesta está en logaritmos se ha corregido el problema de heterocedasticidad de residuos, por lo que en el modelo B ya se puede afirmar que los residuos tienen varianza constante”

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 A es 0.1868 y es significativo y la respuesta está en logaritmos, podemos interpretar que si un piso tiene aire acondicionado, el precio de su alquiler es un 18.68% 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(0.1868)=1.2053, 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 20.53%

Para estos datos, ajustamos un modelo que incluye como predictores el número de habitaciones y si tiene o no Aire Acondicionado. 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 aire acondicionado o no (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
-0.42198 -0.12220 -0.01264  0.10511  0.65558

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             6.50053    0.09210  70.580  < 2e-16 ***
Habitaciones            0.09865    0.03706   2.662  0.01068 *
AireAcond1              0.63916    0.19024   3.360  0.00158 **
Habitaciones:AireAcond1 -0.13174    0.08007  -1.645  0.10674
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2343 on 46 degrees of freedom
Multiple R-squared:  0.349,      Adjusted R-squared:  0.3066
F-statistic: 8.222 on 3 and 46 DF,  p-value: 0.0001735
```

5. Indica los modelos que se obtienen para predecir el precio en función del número de habitaciones cuando hay aire acondicionado o no. 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 AireAcond1 vale cero y por lo tanto el modelo es:

$$\log(precio) = 6.5 + 0.09865 * Habitaciones + \varepsilon$$

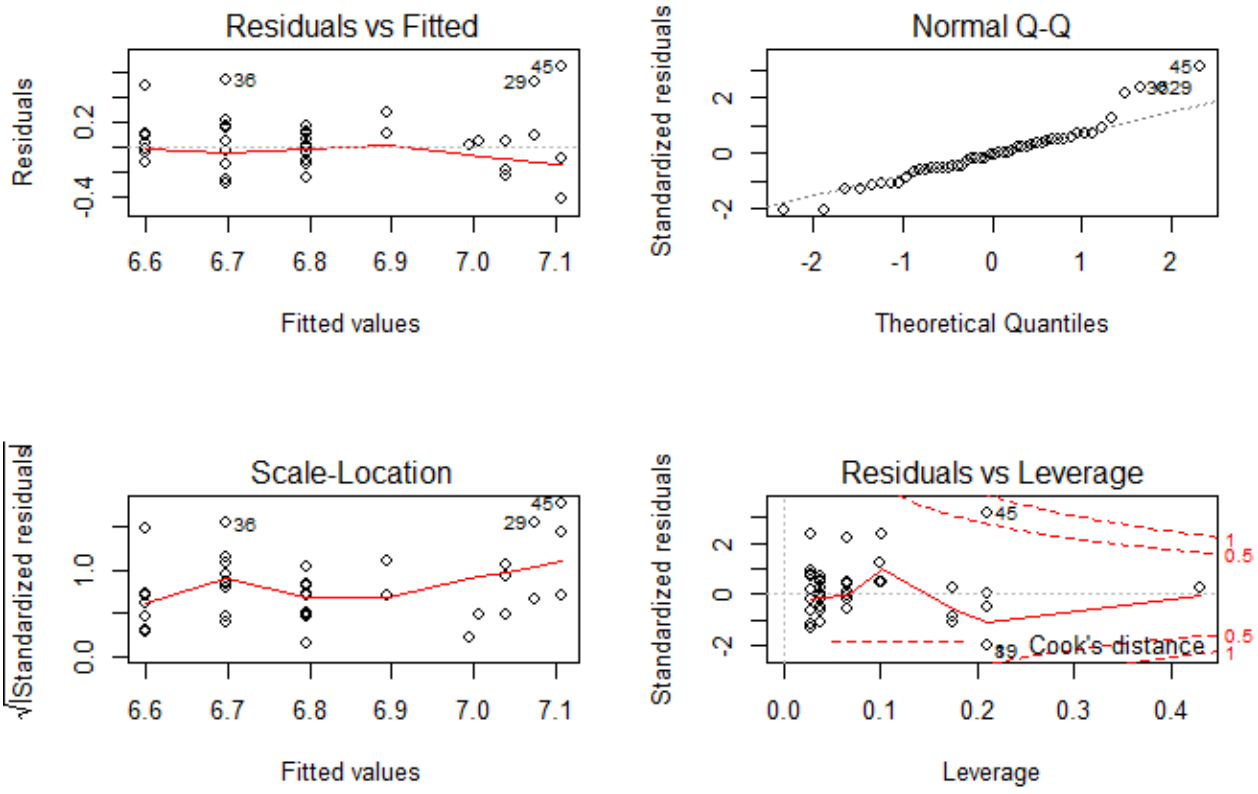
La pendiente es significativa, lo cual ya se observa en el gráfico de la izquierda, lo cual se interpreta como que si el piso no tiene aire acondicionado, el número de habitaciones es un factor que supone un incremento en el precio del alquiler.

En el caso de un piso con aire acondicionado, donde la variable AirAcond1 vale 1, el modelo obtenido es:

$$\log(precio) = (6.5 + 0.63916) + (0.09865 - 0.13174) * Habitaciones + \varepsilon$$
$$= 7.13916 - 0.03309 * Habitaciones + \varepsilon$$

No tenemos un p-valor para el coeficiente de la pendiente obtenida. Aunque la estimación es negativa, no podemos establecer su significación.
En este caso, la pendiente es negativa de forma sorprendente, ya que indica que en los pisos con aire acondicionado se espera que cuantas más habitaciones, menor sea el precio del alquiler. Por cada habitación de más que tenga, el precio desciende a un 3.26% , ya que exp(-0.03309)= 0.9674. 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 la presencia de los atípicos señalados.

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 unas 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 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
36	100	1400	1	3	2	1	0	1	2	0
45	100	2350	1	3	1	1	1	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 aire acondicionado. 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
El segundo caso (36) 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.
La tercera 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 aire acondicionado 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 a posteriori. 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.

Problema 2 (5 punts): Resposta Binària

Los datos corresponden a 935 estudiantes del Grado en Ingeniería en Tecnologías Industriales que entraron por acceso de Selectividad entre los años 2010 y 2012 en la UPC. Se recogen las siguientes variables:

Sexe:	H: Hombre, D: Mujer
AnyEntrada:	Año de entrada en el grado (2010, 2011, 2012)
Barcelona:	Vive fuera de la provincia de Barcelona (S/N)
Algebra:	Nota superior a 7 en la asignatura Algebra en la fase selectiva-Q1 (S/N)
NotaEstadistica:	Nota superior a 7 en la asignatura Estadística en tercero-Q5 (S/N)

Se quiere analizar cuáles de estos factores están asociados con el hecho de sacar más de un 7 (notable o excelente) en la asignatura de Estadística del Grado.

				Nota Estadistica >7	
Gènere	**AnyEntrada**	**Barcelona**	**Algebra>7**	**No**	**Si**
D	**2010**	N	N	12	1
D	**2010**	N	S	3	1
D	**2010**	S	N	37	5
D	**2010**	S	S	7	3
D	**2011**	N	N	16	1
D	**2011**	N	S	3	2
D	**2011**	S	N	33	9
D	**2011**	S	S	4	11
D	**2012**	N	N	5	2
D	**2012**	N	S	4	0
D	**2012**	S	N	37	6
D	**2012**	S	S	17	9
H	**2010**	N	N	18	3
H	**2010**	N	S	5	2
H	**2010**	S	N	137	13
H	**2010**	S	S	29	4
H	**2011**	N	N	33	14
H	**2011**	N	S	5	5
H	**2011**	S	N	142	29
H	**2011**	S	S	21	17
H	**2012**	N	N	15	9
H	**2012**	N	S	4	13
H	**2012**	S	N	113	23
H	**2012**	S	S	26	27

1. Determina la tabla de datos agregados necesaria para la estimación del modelo de respuesta binaria para la probabilidad de obtener una nota en la asignatura de estadística superior a 7 con el único efecto del año de entrada. ¿Cuál es la probabilidad de obtener una nota en Estadística superior a 7 que marginalmente corresponde a cada alumno?

Año de Entrada	Nota EST>7 (respuesta positiva)	Número Alumnos	Probabilidad
2010	**32**	**280**	**0.114**
2011	**88**	**345**	**0.255**
2012	**89**	**310**	**0.287**
	209	**935**	**0.223**

P('NotaEST>7')=209/935= 0,223

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

$$logit(\pi_i)=log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad \rightarrow \hat{\eta}=ln\left(\frac{^{209}/_{935}}{1-^{209}/_{935}}\right)=ln\left(\frac{0.223}{1-0.223}\right)=ln(0.2878)=-1.245$$

3. Estima manualmente a partir de la tabla del punto anterior y empleando la transformación ***probit*** cuáles son los estimadores de la constante y de los coeficiente de las *dummies* para el efecto bruto del **año de entrada** que incluye exclusivamente el factor AnyEntrada (nivel de referencia '2010').

$$\Phi^{-1}(\pi_i)=\eta+\alpha_i \quad i=2010,2011,2012 \quad \alpha_{2010}=0 \quad \rightarrow$$

$$\hat{\eta}=\Phi^{-1}(\pi_{2010})=-1.204$$
$$\hat{\alpha}_{2011}=\Phi^{-1}(\pi_{2011})-\Phi^{-1}(\pi_{2010})=0.545$$
$$\hat{\alpha}_{2012}=\Phi^{-1}(\pi_{2012})-\Phi^{-1}(\pi_{2010})=0.642$$

```
> summary(modc<-glm(cbind(Si,No)~AnyEntrada,notes,family=binomial(probit)))

Call:
glm(formula = cbind(Si, No) ~ AnyEntrada, family = binomial(probit),
    data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2040  -0.7398   0.5258   1.3178   4.0747

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.20405    0.09839  -12.24  < 2e-16 ***
AnyEntrada2011  0.54543    0.12256    4.45 8.57e-06 ***
AnyEntrada2012  0.64216    0.12397    5.18 2.22e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.279  on 23  degrees of freedom
Residual deviance:  91.209  on 21  degrees of freedom
AIC: 171.58

Number of Fisher Scoring iterations: 4
```

4. Calcula el odds-ratio del **género** sobre la incidencia de **notas superiores a 7 en Estadística**. Interpreta el efecto del género en la escala de los odds de la probabilidad de obtener más de un 7 en EST. ¿Es significativo este efecto?

```
> p1=50/(50+178)
> p2=159/(159+548)
> log(p1/(1-p1))
[1] -1.269761
> log(p2/(1-p2))-log(p1/(1-p1))
[1] 0.03238946
> summary(moda<-glm(cbind(Si,No)~Sexe,notes,family=binomial))

Call:
glm(formula = cbind(Si, No) ~ Sexe, family = binomial, data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
-4.4912  -1.4376   0.2603   1.5235   4.7205

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.26976    0.16006  -7.933 2.14e-15 ***
SexeH        0.03239    0.18366   0.176     0.86
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.28  on 23  degrees of freedom
Residual deviance: 122.25  on 22  degrees of freedom
AIC: 200.62

> coef(mod)
(Intercept)       SexeH
-1.26976054  0.03238946
> exp(coef(mod))
(Intercept)       SexeH
  0.2808989   1.0329197
> 100*(1-exp(coef(mod)))
(Intercept)       SexeH
  71.910112   -3.291971
```

Els odds de treure una nota superior a 7 en Estadística augmenta en 3,29% en el grup dels homes respecte els odds del grup de referencia de les dones.

Aún así, este efecto es no significativo:

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

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    23     133.59
Sexe  1 0.031198        22     133.56   0.8598
```

5. Calcula i interpreta en la escala del predictor lineal i del odds-ratio del efecto bruto de haber sacado **más de un 7 en álgebra** sobre el hecho de sacar más de un 7 en Estadística.

```
> p1=115/(115+598)
> p2=94/(94+128)
> log(p2/(1-p2))-log(p1/(1-p1))
[1] 1.339923

> summary(modd<-glm(cbind(Si,No)~Algebra,notes,family=binomial))

Call:
glm(formula = cbind(Si, No) ~ Algebra, family = binomial, data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7996  -0.8052  -0.1698   0.8435   2.8634

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6487     0.1018 -16.191  < 2e-16 ***
```

```
AlgebraS      1.3399     0.1698   7.893 2.95e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.279  on 23  degrees of freedom
Residual deviance:  61.217  on 22  degrees of freedom
AIC: 139.59
```

En l'escala del predictor lineal, aquest puja 1.34 unitats si s'ha tret més d'un 7 en Àlgebra. En l'escala de l'odds-ratio, el factor d'augment és exp(1.34)=3.82 vegades respecte a la categoría base de no haver tret més d'un 7 en àlgebra.

6. Determina si el efecto NETO de la **Pertenencia a Barcelona** es estadísticamente significativo, cuando el haber sacado **más de un 7 en Algebra** y el **año de entrada** ya se encuentran en el modelo. Indica el valor del estadístico del test y la distribución de referencia que utilizas para justificar la decisión.

```
summary(modb<-glm(cbind(Si,No)~AnyEntrada+Algebra+Barcelona,notes,family=binomial))>
anova(modb,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         23    122.279
AnyEntrada  2   31.070        21     91.209 1.791e-07 ***
Algebra     1   56.274        20     34.935 6.304e-14 ***
Barcelona   1    5.411        19     29.524   0.02001 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Si que ho és. Estadístico: 5.411, distribución de referencia: Chi-sq amb 1 grau de llibertat

7. El modelo que incluye el **año de entrada**, una **nota superior a 7 en Algebra** y pertenecer a **Barcelona** ¿explica de forma adecuada los datos observados? Indica el estadístico y la distribución de referencia en que basas el test para justificar la respuesta.

Test sobre la deviància residual. Estadístic: 29.524, distribució de referència: chi-sq amb 19 graus de llibertat. El p-valor és 0.058, superior al nivel de significació habitual del 5%. Per tant, el model és satisfactori per explicar les dades observades.

```
> 1-pchisq(29.524,df=19)
[1] 0.05817742
```

8. ¿Existe evidencia para afirmar que los efectos de haber sacado **más de un 7 en Algebra** no son los mismos según los distintos **años de entrada** en la incidencia notas superiores a 7 en Estadística? Indica el valor del estadístico y la distribución de referencia del test.

```
> summary(mod1<-glm(cbind(Si,No)~AnyEntrada+Algebra,notes,family=binomial))
```

```
Call:
glm(formula = cbind(Si, No) ~ AnyEntrada + Algebra, family = binomial,
    data = notes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2784  -0.8589   0.1631   0.6314   2.4222

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.4170     0.2028 -11.918  < 2e-16 ***
AnyEntrada2011  1.0269     0.2313   4.439 9.02e-06 ***
AnyEntrada2012  1.0086     0.2328   4.333 1.47e-05 ***
AlgebraS        1.3177     0.1745   7.552 4.28e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.279  on 23  degrees of freedom
Residual deviance:  34.935  on 20  degrees of freedom
AIC: 117.3

Number of Fisher Scoring iterations: 4

> summary(mod2<-glm(cbind(Si,No)~AnyEntrada*Algebra,notes,family=binomial))

Call:
glm(formula = cbind(Si, No) ~ AnyEntrada * Algebra, family = binomial,
    data = notes)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.32094  -0.75814   0.09498   0.65041   2.32027

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.2271     0.2244  -9.924   <2e-16 ***
AnyEntrada2011          0.7857     0.2715   2.894   0.0038 **
AnyEntrada2012          0.7802     0.2850   2.737   0.0062 **
AlgebraS                0.7455     0.4160   1.792   0.0732 .
AnyEntrada2011:AlgebraS 0.7547     0.5053   1.494   0.1352
AnyEntrada2012:AlgebraS 0.6614     0.4939   1.339   0.1805
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 122.279  on 23  degrees of freedom
Residual deviance:  32.447  on 18  degrees of freedom
AIC: 118.82

Number of Fisher Scoring iterations: 4

> anova(mod1,mod2,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(Si, No) ~ AnyEntrada + Algebra
Model 2: cbind(Si, No) ~ AnyEntrada * Algebra
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        20     34.935
2        18     32.447  2   2.4871   0.2884
```

Demana si la relació entre el factor "nota>7 en álgebra" i la incidència de "notes>7 en Estadística" ha canviat per als diferents anys d'entrada, per tant, és un contrast entre el model complert i el model additiu, el qual té un pvalor superior al 5% habitual i per tant, no es pot rebutjar la hipòtesi nul.la: NO, la relació entre el factor Notaalgebra>7 i la incidència de NotaEstad>7 no depen de l'any d'entrada al grau. L'estadístic de referència és una Chi quadrat amb 2 graus de llibertat i l'estadístic val 2.4871.

RESULTATS R

```
> summary(notes)
 Sexe   AnyEntrada Barcelona Algebra       No               Si
 D:12   2010:8     N:12      N:12    Min.   :  3.00   Min.   : 0.000
 H:12   2011:8     S:12      S:12    1st Qu.:  5.00   1st Qu.: 2.000
        2012:8                       Median : 16.50   Median : 5.500
                                     Mean   : 30.25   Mean   : 8.708
                                     3rd Qu.: 33.00   3rd Qu.:13.000
                                     Max.   :142.00   Max.   :29.000
> by(notes$No, notes$Sexe,sum)
notes$Sexe: D
[1] 178
------------------------------------------------------------
notes$Sexe: H
[1] 548
> by(notes$Si, notes$Sexe,sum)
notes$Sexe: D
[1] 50
------------------------------------------------------------
notes$Sexe: H
[1] 159
> by(notes$No, notes$Algebra,sum)
notes$Algebra: N
[1] 598
------------------------------------------------------------
notes$Algebra: S
[1] 128
> by(notes$Si, notes$Algebra,sum)
notes$Algebra: N
[1] 115
------------------------------------------------------------
notes$Algebra: S
[1] 94

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

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    23     122.28
Sexe  1 0.031198        22     122.25   0.8598
> anova(modb,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)
```

```
          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                         23     122.28
Barcelona  1   7.1301        22     115.15  0.00758 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modc,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: probit

Response: cbind(Si, No)

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                         23    122.279
AnyEntrada  2    31.07        21     91.209 1.791e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(modd,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)

        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        23    122.279
Algebra  1   61.063        22     61.217 5.529e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


> anova(modb)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)

           Df Deviance Resid. Df Resid. Dev
NULL                         23    122.279
AnyEntrada  2   31.070        21     91.209
Algebra     1   56.274        20     34.935
Barcelona   1    5.411        19     29.524

> anova(mod2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Si, No)

Terms added sequentially (first to last)

                  Df Deviance Resid. Df Resid. Dev
NULL                                 23    122.279
AnyEntrada          2   31.070        21     91.209
Algebra             1   56.274        20     34.935
AnyEntrada:Algebra  2    2.487        18     32.447
```

Problema 3 (1 punto): Modelización

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

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

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

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

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

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

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