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

CURS 2017-2018 Q1 – EXAMEN REAVALUACIÓ : MODEL LINEAL GENERALITZAT

(Data: 29/01/2018 a les 12:00h Aula 005-FME)

Nom de l´alumne:	**DNI:**
Professors:	Erik Cobo, Jordi Cortés, Josep Anton Sànchez
Normativa:	SÓN PERMESOS APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	3h 00 min

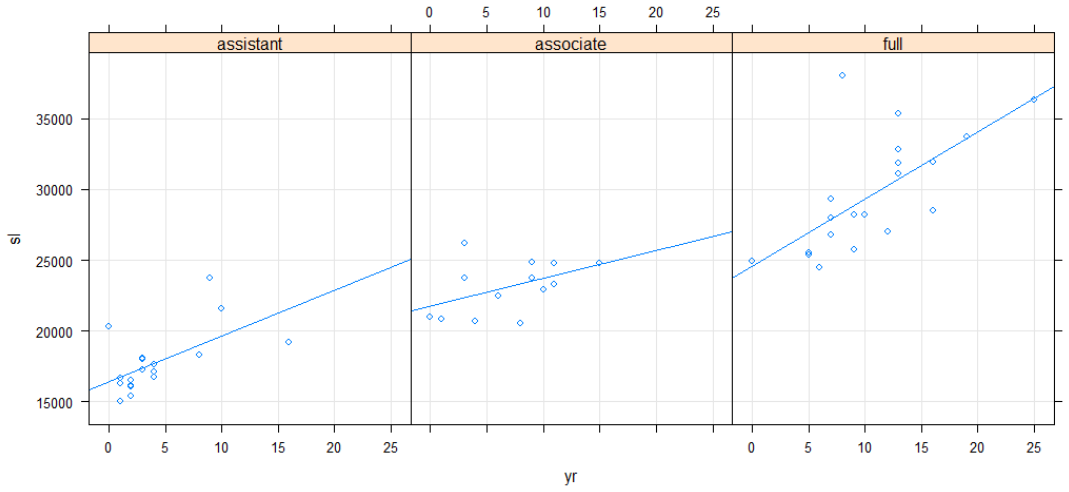
Problema 1 (3 puntos): Respuesta normal

En un estudio para analizar la discriminación salarial se recogió información de 52 profesores de universidad (Weisberg, 1985). Las variables son:

- sx : Sexo (female / male)
- rk : Rango (assistant / associate / full)
- yr : Número de años en el rango actual (antigüedad en el cargo)
- dg : Titulación (doctorate / masters)
- yd : Número de años desde la titulación
- sl : Salario académico anual, en dólares.

```
> summary(salary)
      sx              rk          yr              dg          yd              sl
 female:14   assistant:18   Min.   : 0.000   doctorate:34   Min.   : 1.00   Min.   :15000
 male  :38   associate:14   1st Qu.: 3.000   masters  :18   1st Qu.: 6.75   1st Qu.:18247
             full     :20   Median : 7.000                  Median :15.50   Median :23719
                            Mean   : 7.481                  Mean   :16.12   Mean   :23798
                            3rd Qu.:11.000                  3rd Qu.:23.25   3rd Qu.:27259
                            Max.   :25.000                  Max.   :35.00   Max.   :38045
```

La relación entre el Salario (sl) y la Antigüedad (yr) segmentando por Rango (rk) viene reflejada en el siguiente gráfico:



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

```
Residuals:
    Min      1Q  Median      3Q     Max
-3687.8 -1123.6  -392.1   720.9  9646.6

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     16416.6      816.0  20.118  < 2e-16 ***
yr                324.5      141.9   2.286 0.026887 *
rkassociate      5354.2     1492.6   3.587 0.000806 ***
rkfull           8176.4     1418.1   5.766 6.49e-07 ***
yr:rkassociate   -129.7      205.8  -0.630 0.531508
yr:rkfull         151.2      171.7   0.880 0.383307
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2386 on 46 degrees of freedom
Multiple R-squared:  0.8534,     Adjusted R-squared:  0.8375
F-statistic: 53.56 on 5 and 46 DF,  p-value: < 2.2e-16
```

1. (1p.) Indica los modelos que se obtienen para predecir el salario en función del rango y de los años de antigüedad en el cargo. Haz una interpretación de estos modelos a partir de la significación de los parámetros del modelo obtenido. ¿Es razonable el resultado obtenido?

Para un profesor que sea asistente (categoría de referencia) la fórmula que determina su salario en función de la antigüedad es:

$$salario = 16416.6 + 324.5 * \text{years} + \varepsilon$$

La pendiente es significativa, como ya se observa en el gráfico de la izquierda, lo cual se interpreta como que el incremento anual del salario es significativo.

Para un profesor que sea asociado la fórmula que determina su salario en función de la antigüedad incorpora los coeficientes correspondientes a la categoría y su interacción con antigüedad:

$$salario = (16416.6 + 5354.2) + (324.5 - 129.7) * \text{years} + \varepsilon$$

El hecho de que el coeficiente de la dummy asociada a la categoría "associate" sea significativa (p-valor=<0.001) y su interacción con yr no lo sea (p-valor= 0.53) permite interpretar que el salario de un asociado es significativamente superior al de un asistente pero su incremento anual no difiere significativamente del mismo. Esto implica que los modelos para asistente y asociado se pueden considerar rectas paralelas.

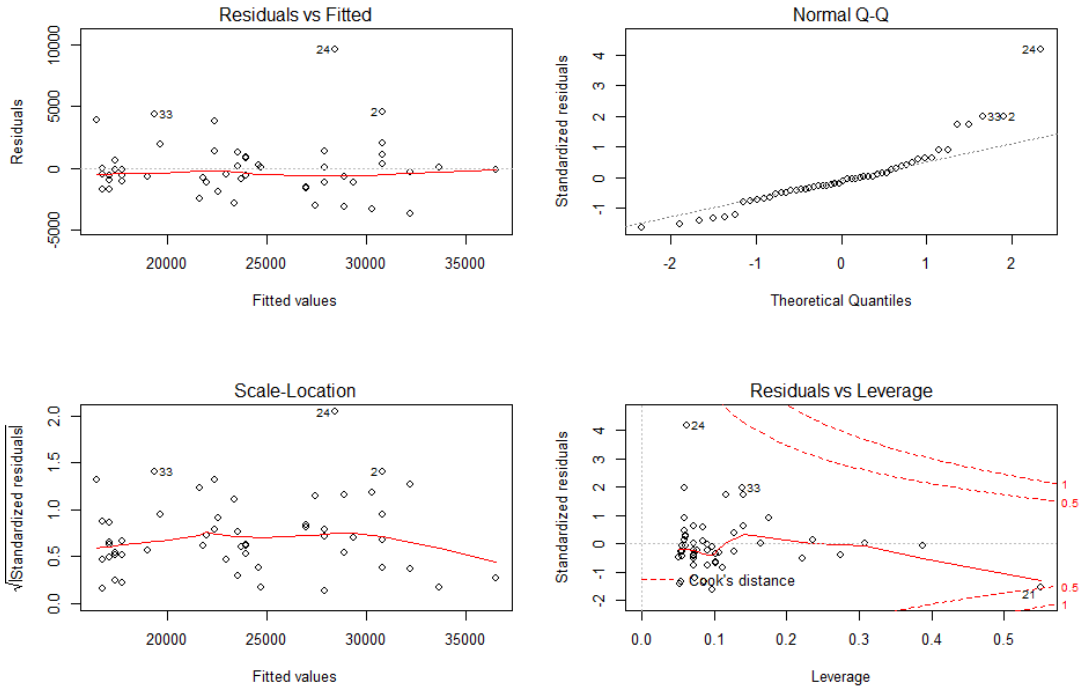
Para un full profesor (catedrático) el modelo para el salario es

$$salario = (16416.6 + 8176.4) + (324.5 + 151.2) * \text{years} + \varepsilon$$

Nuevamente, si se compara con la categoría de referencia (asistente) el salario es significativamente superior (p-valor<0.001) y sin embargo, no hay diferencias significativas en el incremento anual respecto a los asistentes (p-valor=0.38). Nuevamente podemos concluir que ambos modelos se pueden considerar rectas paralelas.

La representación gráfica de la descriptiva es coherente con las interpretaciones anteriores, representando lo que al parecer pueden ser consideradas rectas paralelas. Aun así, la pendiente del modelo para los asociados y para los full profesor podrían ser diferentes ya que el ajuste del modelo no incluye el p-valor que compara ambos términos para la interacción.

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



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

	sx	rk	yr	dg	yd	sl
2	male	full	13	doctorate	22	35350
21	male	assistant	16	masters	23	19175
24	female	full	8	doctorate	24	38045
33	male	assistant	9	masters	14	23713

2. (1p.) Realiza la validación del modelo indicando las premisas que se validan en cada plot. Teniendo en cuenta los plots de validación y la descriptiva del fichero, 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?

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 etiquetadas por tener 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 primer y tercer plot parecen confirmar que la varianza de los residuos es constante. 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 próximo 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 2, 24 y 33 son casos con un residuo estandarizado superior a 2. Son observaciones mal explicadas por el modelo y debido al signo del residuo se puede interpretar que cobran por encima de lo previsto teniendo en cuenta su rango y antigüedad (con 13 y 8 años de antigüedad cobran por encima de 35.000 dólares). El 33 es un asistente con un sueldo elevado para su categoría y antigüedad.

El caso 21 es el que posee un leverage más alto y a su vez, su residuo estandarizado es próximo a -2. Indica una observación influyente (su distancia de Cook lo sitúa próximo a la curva de nivel de 0.5). Si observamos la descriptiva, todos los asistentes tienen una antigüedad inferior a 10 años y en cambio, este profesor lleva 16 años siendo asistente. Su sueldo, al parecer se encuentra por debajo de lo esperado ante los años que lleva en el cargo). Si se elimina del modelo, los coeficientes del nuevo ajuste pueden diferir de ls actuales de forma significativa.

Se aplicó el mecanismo stepwise para seleccionar las variables de dos maneras diferentes, partiendo del modelo que incluye todos los predictores y sus interacciones dobles (modelo m1)

modA<-step(m1,direction="both",k=2)

Modelo A:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          16351.446   1166.479  14.018 2.23e-16 ***
sxmale                 551.671   1431.951   0.385  0.70225
rkassociate           8483.718   2735.370   3.101  0.00367 **
rkfull                6487.695   2580.245   2.514  0.01640 *
yr                    1173.755    388.850   3.019  0.00458 **
dgmasters                9.492   3753.953   0.003  0.99800
yd                    -525.933    238.387  -2.206  0.03366 *
sxmale:yr             -602.014    291.928  -2.062  0.04626 *
sxmale:dgmasters      3200.169   2297.783   1.393  0.17202
rkassociate:dgmasters -3976.114   2990.329  -1.330  0.19177
```

```
rkfull:dgmasters      -8868.116   4471.130  -1.983  0.05478 .
rkassociate:yd           13.051    211.772   0.062  0.95119
rkfull:yd               499.151    204.187   2.445  0.01939 *
yr:dgmasters           -334.334    265.506  -1.259  0.21583
dgmasters:yd            379.062    285.248   1.329  0.19202
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2315 on 37 degrees of freedom
Multiple R-squared:  0.889,      Adjusted R-squared:  0.847
F-statistic: 21.16 on 14 and 37 DF,  p-value: 1.74e-13
```

modB<-step(m1,direction="both",k=log(nrow(salary)))

Modelo B:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16203.27     638.68  25.370  < 2e-16 ***
rkassociate  4262.28     882.89   4.828 1.45e-05 ***
rkfull       9454.52     905.83  10.437 6.12e-14 ***
yr            375.70      70.92   5.298 2.90e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2402 on 48 degrees of freedom
Multiple R-squared:  0.8449,     Adjusted R-squared:  0.8352
F-statistic: 87.15 on 3 and 48 DF,  p-value: < 2.2e-16
```

3. (0.5p.) ¿Qué diferencias hay entre el método de selección de variables realizado con el modelo A y con el modelo B? Indica los criterios que se utilizan en ambos casos para decidir las variables que quedan en el modelo. ¿Cuál consideras más adecuado de los dos modelos resultantes? Justifica la respuesta

*En el primer caso, se ha aplicado el mecanismo de selección stepwise, utilizando como criterio el AIC (Akaike Information Criterion), ya que el parámetro k se ha fijado a 2 (valor por defecto). Ello quiere decir que el criterio de información penaliza el número de parámetros con un peso igual a 2 (AIC=-2*loglik+2*p).*

*En el segundo caso, el peso de la componente de complejidad del modelo (número de parámetros) en el criterio de información es log(n), donde n es el número de observaciones. Por ello, el criterio utilizado en este caso es el BIC (Bayesian Information Criterion), (BIC=-2*loglik+log(n)*p)*

El criterio basado en el BIC es más exigente y suele eliminar parámetros que no sean estrictamente significativos, al nivel del 0.05. Es por ello que el modelo B presenta un menor número de parámetros, porque el modelo A permite mantener en el modelo coeficientes con p-valores superiores a 0.05 (hasta un 0.15 aprox).

Si el modelo lo estamos construyendo en una etapa prospectiva (fase piloto para comenzar a determinar las relaciones presentes, puede ser adecuado usar el AIC para no eliminar predictores que en posteriores fases, con mayor potencia en el modelo puedan aparecer significativas. Sin embargo, en estudios confirmatorios donde se desea concretar las relaciones entre los predictores y la respuesta, pude ser importante depurar el modelo, evitando efectos de multicolinealidad para analizar las relaciones existentes de forma más clara. En estos casos, el criterio más adecuado es el BIC. Como en este caso no se indica cuál es la situación concreta, la respuesta es que dependerá de que interesa más, modelos

generales de tipo prospectivo (entonces el modelo A) o modelos simples para interpretar de forma neta las relaciones observadas (entonces el modelo B).

4. (0.5p.) La discriminación por género podría tener (al menos) dos componentes: (1) mayor salario para mismo trabajo; y (2) mayor posibilidad de promocionar. Discute qué modelos y coeficientes le podrían ayudar a aproximarse a ambos efectos.

Si queremos estudiar si hay discriminación por género debido a un mayor salario para el mismo trabajo, deberíamos explorar la significación del término de la dummy sxMale, que corresponde a la variación en el salario por el hecho de ser hombre, respecto a las mujeres, manteniendo el resto constante. Por otro lado, la interacción entre sexo y el resto de predictores permitiría determinar si la categoría, la antigüedad, la titulación o los años desde la titulación se valoran de forma diferente entre mujeres y hombres. Por ejemplo, en el modelo A aunque la estimación puntual del coeficiente sxMale es de 551 en positivo, indicando un incremento de 551 dólares en los sueldos de los hombres respecto a las mujeres, el p-valor asociado indica la diferencia no es significativa y se puede considerar fruto del azar (p-valor=0.70)

Para estudiar el efecto del género en la posibilidad de promocionar, habría que cambiar el modelo y establecer como variable respuesta la categoría e incluir la variable sexo entre los predictores. En este caso, como hay 3 categorías en la variable respuesta, el modelo sería multinomial y se analizaría la relación entre el sexo y la categoría para establecer si la probabilidad de estar en una categoría superior depende del género.

Problema 2 (5 puntos): Respuesta Binaria

Los siguientes datos provienen de un studio de Wilner, Walkley and Cook en 1955 sobre las actitudes raciales en la segregación e integración en viviendas públicas. Los datos recogen respuestas de 608 mujeres blancas en referencia a:

Proximity: Proximidad a una familia negra (close=cercana / distant=distante)
Contact: Frecuencia de contacto con negros (frequent / infrequent)
Norms: Posición respecto a reglas locales sobre los negros (favorable / unfavorable)
Sentiment: Actitud general respecto a los negros (favorable / unfavorable)

Se quiere analizar cuáles de estos factores están asociados con el hecho de tener un sentimiento "unfavorable" respecto a los negros (racismo)

			Sentiment	
Proximity	**Contact**	**Norms**	favorable	unfavorable
close	frequent	Favorable	77	32
close	frequent	Unfavorable	30	36
close	infrequent	favorable	14	19
close	infrequent	unfavorable	15	27

distant	frequent	favorable	43	20
distant	frequent	unfavorable	36	37
distant	infrequent	favorable	27	36
distant	infrequent	unfavorable	41	118

1. (0.5p.) Determina la tabla de datos agregados necesaria para la estimación del modelo de respuesta binaria para la probabilidad de tener **sentimiento** racista (unfavorable) con el efecto de la frecuencia del **contacto** con negros. ¿Cuál es la probabilidad de tener sentimiento racista que marginalmente corresponde a las entrevistadas?

Contact	Sentiment unfavorable	Número respuestas	Probabilidad
Frequent	**125**	**311**	**0.402**
Infrequent	**200**	**297**	**0.673**
Total	**325**	**608**	**0.535**

P('sentiment unfavorable')=325/608=0,535

2. (0.5p.) 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.

$$\text{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)=\eta \quad \rightarrow \hat{\eta}=\log\left(\frac{325/608}{1-325/608}\right)=\log\left(\frac{0.5345395}{1-0.5345395}\right)=0.1383783$$

3. (0.5p.) 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 coeficientes de las *dummies* el modelo que incluye exclusivamente el factor **Contact** (nivel de referencia 'frequent').

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

$$\hat{\eta}=\Phi^{-1}(\pi_{freq})=\Phi^{-1}(125/311)=-0.2483$$

$$\hat{\alpha}_{\text{inf }req}=\Phi^{-1}(\pi_{infreq})-\Phi^{-1}(\pi_{freq})=\Phi^{-1}(200/297)-\Phi^{-1}(125/311)=0.69768$$

```
> summary(mod<-
(glm(cbind(SentimentUnfav,SentimentFav)~Contact,housing,family=binomial(probit))))

Call:
glm(formula = cbind(SentimentUnfav, SentimentFav) ~ Contact,
    family = binomial(probit), data = housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3531  -1.4626  -0.7949   1.8290   2.3495

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.24836    0.07187  -3.456 0.000549 ***
Contactinfrequent  0.69768    0.10421   6.695 2.15e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.827  on 7  degrees of freedom
Residual deviance: 24.215  on 6  degrees of freedom
AIC: 64.925
```

```
Number of Fisher Scoring iterations: 3
```

4. (1p.) Calcula la odds-ratio de **proximity** sobre la incidencia de racismo (**sentiment unfavorable**), sin ajustar por más variables. Calcula también un intervalo de confianza al 95% para este odds-ratio. Interpreta el efecto de la proximidad a una familia de negros en la escala de los odds de la probabilidad de tener sentimientos racistas. ¿Es significativo este odds-ratio?

```
> p1=114/(114+136)
> p2=211/(211+147)
> (p2/(1-p2))/(p1/(1-p1))
[1] 1.712376

> coef(mod)[2]
   Proximitydistant
          0.5378820
> exp(coef(mod)[2])
   Proximitydistant
          1.7123762

> 0.5379+c(-1.96,1.96)*0.1663
[1] 0.211952 0.863848

> exp(0.5379+c(-1.96,1.96)*0.1663)
[1] 1.236089 2.372272
```

Odss-ratio=1.7123. Los odds de tener sentimiento racista aumentan en 71,23% si la proximidad es distante frente al odds del grupo de referencia de proximidad cercana.
Este efecto es significativo (p-valor=0.00122): el valor de 1 está fuera del intervalo de confianza construido al 95% (1.236,2.372)

5. (0.5p.) El modelo que solo incluye la frecuencia de **contacto** ¿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 deviancia residual. Estadístico: 24.215, distribución de referencia: chi-sq con 6 grados de libertad. El nivel de significación habitual del 5% fija un valor de 12.59 para una chi-sq con 6 grados de libertad. Por tanto, el modelo no es satisfactorio para explicar los datos observados.

Valor crítico:

```
> qchisq(0.95,df=6)
[1] 12.59159
```

p-valor:

```
> 1-pchisq(24.215,df=6)
[1] 0.0004767709
```

6. (0.5p.) Calcula i interpreta en la escala del predictor lineal i del odds-ratio del efecto bruto de estar en contra de las **normas** locales acerca de negros sobre el hecho de tener sentimientos racistas.

```
> p1=107/(107+161)
> p2=218/(218+122)
> log(p2/(1-p2))-log(p1/(1-p1))
[1] 0.9890495
> summary(mod<-glm(cbind(SentimentUnfav,SentimentFav)~Norms,housing,family=binomial))

Call:
```

```
glm(formula = cbind(SentimentUnfav, SentimentFav) ~ Norms, family = binomial,
    data = housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3448  -1.7712  -0.6609   2.2126   2.7529

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -0.4086     0.1247  -3.276  0.00105 **
Normsunfavorable  0.9890     0.1683   5.875 4.23e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.827  on 7  degrees of freedom
Residual deviance: 34.297  on 6  degrees of freedom
AIC: 75.007
```

En la escala del predictor lineal, éste sube 0.989 unidades si no es favorable a las normas. En la escala del odds, el factor de aumento es exp(0.989)=2.69 veces respecte a la categoría base de ser favorable a las normas.

7. (0.5p.) Determina si el efecto NETO de la **proximidad** es estadísticamente significativo, cuando la frecuencia de **contacto** y la actitud frente a las **normas** locales ya se encuentran en el modelo. Indica el valor del estadístico del test y la distribución de referencia que utilizas para justificar la decisión.

No lo es. Estadístico: 0.283, distribución de referencia: Chi-sq amb 1 grau de llibertat

```
> 1-pchisq(0.283,1)
[1] 0.5947416
> anova(mod,test="Chi")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(SentimentUnfav, SentimentFav)

Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          7     69.827
Contact    1   45.612         6     24.215 1.441e-11 ***
Norms      1   21.694         5      2.520 3.198e-06 ***
Proximity  1    0.283         4      2.238     0.595
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. (1p.) ¿Existe evidencia para afirmar que los efectos de tener **contacto** infrecuente no son los mismos según si se está a favor o en contra de las **normas** locales sobre negros en la incidencia de actitudes racistas? Indica el valor del estadístico y la distribución de referencia del test.

Se pregunta si la relación entre el factor "contacto" i la incidència de "sentimient unfavorable" ha cambiado para los diferentes grupos a favor y en contra de normas locales. Por lo tanto, es un contraste entre el modelo completo y el modelo aditivo, el cual tiene un pvalor superior al 5% habitual y por tanto, no se puede rechazar la hipótesis nula: NO, la

relación entre el factor Contacto y la incidencia de sentimientos racista no depende de si se está afavor o en contra de normas locales. El estadístico de referencia es una Chi cuadrado con 1 grado de libertad y el estadístico vale 0.632

RESULTATS R

```
Call:
glm(formula = cbind(SentimentUnfav, SentimentFav) ~ Proximity,
    family = binomial, data = housing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3535  -1.9347   0.5445   1.6998   4.0297

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.1765     0.1270  -1.390  0.16465
Proximitydistant   0.5379     0.1663   3.234  0.00122 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 69.827  on 7  degrees of freedom
Residual deviance: 59.289  on 6  degrees of freedom
AIC: 99.999


Analysis of Deviance Table

Model: binomial, link: logit
Response: cbind(SentimentUnfav, SentimentFav)
Terms added sequentially (first to last)


          Df Deviance Resid. Df Resid. Dev
NULL                          7     69.827
Contact    1   45.612         6     24.215
Norms      1   21.694         5      2.520
Proximity  1    0.283         4      2.238


Analysis of Deviance Table

Model: binomial, link: logit
Response: cbind(SentimentUnfav, SentimentFav)
Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev
NULL                              7     69.827
Contact           1   45.612         6     24.215
Norms             1   21.694         5      2.520
Contact:Norms     1    0.632         4      1.888
```

Problema 3 (2 puntos): Modelo log-lineal multinomial

Ahora intentamos ver la relación existente entre el tabaco, el género y la edad a través del estudio de 1050 individuos. Las variables de les que se dispone y sus categorías están a continuación:

Fumador: Si la persona no ha fumado nunca, ha fumado en el pasado o fuma habitualmente (*Mai*/*Ex*/*Fumador*)
Edat: Edad de la persona categorizada (*Jove*/*Gran*)
Sexe: *Home*/*Dona*

[En todas las variables, la primera categoría entre paréntesis es la de referencia]

A continuación, tienes la salida de R de todos los modelos posibles que incluyen las 3 variables.

Modelo 1

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     4.94238    0.06298  78.480  < 2e-16 ***
EdatGran       -0.67179    0.06524 -10.298  < 2e-16 ***
FumadorEx      -0.57982    0.08921  -6.499 8.08e-11 ***
FumadorFumador  0.36464    0.06958   5.241 1.60e-07 ***
SexeDona       -0.42535    0.06312  -6.738 1.60e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.68  on 11  degrees of freedom
Residual deviance: 183.87  on  7  degrees of freedom
```

Modelo 2

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        4.74493    0.06931  68.464  < 2e-16 ***
EdatGran          -0.17366    0.07967  -2.180   0.0293 *
SexeDona           0.01439    0.07587   0.190   0.8496
FumadorEx         -0.57982    0.08921  -6.499 8.08e-11 ***
FumadorFumador     0.36464    0.06958   5.241 1.60e-07 ***
EdatGran:SexeDona -1.50988    0.15681  -9.629  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.68  on 11  degrees of freedom
Residual deviance:  76.14  on  6  degrees of freedom
```

Modelo 3

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            4.74642    0.07895  60.117  < 2e-16 ***
EdatGran              -0.67179    0.06524 -10.298  < 2e-16 ***
FumadorEx             -0.16862    0.11203  -1.505 0.132270
FumadorFumador         0.59034    0.09451   6.246 4.20e-10 ***
SexeDona               0.01143    0.10691   0.107 0.914865
FumadorEx:SexeDona    -1.11004    0.19657  -5.647 1.63e-08 ***
FumadorFumador:SexeDona -0.51380    0.14099  -3.644 0.000268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.68  on 11  degrees of freedom
Residual deviance: 148.19  on  5  degrees of freedom
```

Modelo 4

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.92203    0.07091  69.416  < 2e-16 ***
EdatGran               -0.61277    0.11196  -5.473 4.42e-08 ***
FumadorEx              -0.72447    0.11617  -6.236 4.49e-10 ***
FumadorFumador          0.45558    0.08484   5.370 7.89e-08 ***
SexeDona               -0.42535    0.06312  -6.738 1.60e-11 ***
EdatGran:FumadorEx      0.36663    0.18236   2.011   0.0444 *
EdatGran:FumadorFumador -0.28416    0.14892  -1.908   0.0564 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.68  on 11  degrees of freedom
Residual deviance: 169.64  on  5  degrees of freedom
```

Modelo 5

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.72607    0.08541  55.332  < 2e-16 ***
FumadorEx              -0.31327    0.13449  -2.329 0.019838 *
FumadorFumador          0.68128    0.10625   6.412 1.44e-10 ***
EdatGran               -0.61277    0.11196  -5.473 4.42e-08 ***
SexeDona                0.01143    0.10691   0.107 0.914865
FumadorEx:EdatGran      0.36663    0.18236   2.011 0.044376 *
FumadorFumador:EdatGran -0.28416    0.14892  -1.908 0.056377 .
FumadorEx:SexeDona      -1.11004    0.19657  -5.647 1.63e-08 ***
FumadorFumador:SexeDona -0.51380    0.14099  -3.644 0.000268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.68  on 11  degrees of freedom
Residual deviance: 133.97  on  3  degrees of freedom
```

Modelo 6

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.54897    0.08409  54.097  < 2e-16 ***
SexeDona                0.45116    0.11489   3.927 8.61e-05 ***
EdatGran               -0.17366    0.07967  -2.180 0.029267 *
FumadorEx              -0.16862    0.11203  -1.505 0.132270
FumadorFumador          0.59034    0.09451   6.246 4.20e-10 ***
SexeDona:EdatGran      -1.50988    0.15681  -9.629  < 2e-16 ***
SexeDona:FumadorEx     -1.11004    0.19657  -5.647 1.63e-08 ***
SexeDona:FumadorFumador -0.51380    0.14099  -3.644 0.000268 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.684  on 11  degrees of freedom
Residual deviance:  40.465  on  4  degrees of freedom
```

Modelo 7

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             4.72458    0.07658  61.692  < 2e-16 ***
EdatGran               -0.11464    0.12094  -0.948   0.3432
FumadorEx              -0.72447    0.11617  -6.236 4.49e-10 ***
FumadorFumador          0.45558    0.08484   5.370 7.89e-08 ***
SexeDona                0.01439    0.07587   0.190   0.8496
EdatGran:FumadorEx      0.36663    0.18236   2.011   0.0444 *
EdatGran:FumadorFumador -0.28416    0.14892  -1.908   0.0564 .
```

```
EdatGran:SexeDona       -1.50988    0.15681  -9.629  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 482.684  on 11  degrees of freedom
Residual deviance:  61.913  on  4  degrees of freedom
```

Modelo 8

```
Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      4.59512    0.10050  45.721  < 2e-16 ***
EdatGran                        -0.27763    0.15308  -1.814 0.069739 .
FumadorEx                       -0.43624    0.16039  -2.720 0.006532 **
FumadorFumador                   0.60889    0.12488   4.876 1.08e-06 ***
SexeDona                         0.25691    0.13384   1.920 0.054919 .
EdatGran:FumadorEx               0.53759    0.22607   2.378 0.017408 *
EdatGran:FumadorFumador         -0.04357    0.19106  -0.228 0.819602
EdatGran:SexeDona               -0.70320    0.22821  -3.081 0.002061 **
FumadorEx:SexeDona              -0.58715    0.23511  -2.497 0.012513 *
FumadorFumador:SexeDona         -0.29043    0.17056  -1.703 0.088597 .
EdatGran:FumadorEx:SexeDona     -2.28679    0.65941  -3.468 0.000525 ***
EdatGran:FumadorFumador:SexeDona -1.50702    0.37717  -3.996 6.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  4.8268e+02  on 11  degrees of freedom
Residual deviance: -1.9984e-15  on  0  degrees of freedom
```

1. (0.5p.) Utilizando la información de los modelos, indica cuantos hombres hay en la base de datos. ¿Y cuantos hombres grandes?

A través del modelo aditivo

$$\frac{\mu_{dona}}{N-\mu_{dona}} = e^{-0.42535} \rightarrow \mu_{dona} = 1050 \cdot \frac{e^{-0.42535}}{1+e^{-0.42535}} = 415 \rightarrow \mu_{home} = 1050 - 415 = 635$$

Hay 635 hombres

Mirando el modelo que contiene la interacción de género con edad:

$$\frac{\mu_{home\_gran}}{\mu_{home} - \mu_{home_{gran}}} = e^{-0.17366} \rightarrow \mu_{home\_gran} = 635 \cdot \frac{e^{-0.17366}}{1+e^{-0.17366}} = 290$$

Hay 290 hombres grandes

```
with(dades3,tapply(Freq,list(Sexe,Edat),sum))
      Jove Gran
 Home  345  290
 Dona  350   65
```

2. (1p.) Completa la tabla de devianzas y decide qué modelo es el más adecuado para reflejar los datos recogidos. Si el modelo solo puede tener una interacción doble, ¿qué modelo sería el mejor? Justifica la respuesta.

E: Edat; F: Fumador; S: Sexe

Model	DF	Deviance
Nulo	11	482.684
E+F+S	7	183.87
EF+S	5	169.64
ES+F	6	76.14
FS+E	5	148.19
E*(F+S)	4	61.91
F*(E+S)	3	133.97
S*(E+F)	4	40.47
EFS	0	0

El modelo válido seria el modelo saturado ya que el resto tiene una devianza que comporta un p-valor muy inferior a 0.05 en el contraste de la $\chi 2$. Se puede ver a simple vista ya que la devianza está muy por encima de los grados de libertad.
Si sólo se puede considerar una interacción, el modelo que incluye la interacción ES (Edad y Sexo) tiene más grados de libertad (por lo que presenta un menor número de parámetros) y su devianza residual es menor (mejor ajuste con 76.14 frente a 169.64 de la interacción EF y 148.19 de la interacción FS)

3. (0.5p.) Del último modelo, interpreta las interacciones siguientes:

```
EdatGran:FumadorFumador          -0.04357
EdatGran:FumadorFumador:SexeDona -1.50702
```

La primera interacción nos dice que la odd de ser fumador entre la gente mayor es un 5% (0.95=exp(-0.044)) inferior a la odd de ser fumador entre la gente joven. La segunda interacción, nos dice que hay una diferencia considerable en este valor entre hombres y mujeres, ya que el ratio de odds entre mayores y jóvenes respecto a ser fumador es un 78% (0.22=exp(-1.507)) inferior en las mujeres. La explicación es que hay muchas menos mujeres fumadoras entre la gente mayor ya que venimos de una sociedad donde la costumbre de fumar correspondía mayoritariamente al género masculino.