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

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

(Data: 17 d'Octubre del 2014 a les 15:00 Aula PC1-FME)

Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	ETSEIB 6a Planta 6-67
Normativa de l'examen:	ÉS POT DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	2h 00 min
Sortida de notes:	Abans del 29 d'Octubre al Web Docent de MLGz
Revisió de l'examen:	29 d'octubre a 16 h a Sala Professors FME– Campus Sud

El conjunto de datos "catalunya" contiene información de los municipios de Catalunya de más de 5.000 habitantes. Son datos recogidos del Institut Català d'Estadística (www.idescat.org)
El campo Procon indica el porcentaje de los votos emitidos en las últimas elecciones autonómicas de 2012 que corresponden a partidos pro-consulta 9N (CiU, ERC, IC i CUP). El objetivo es determinar las relaciones que se pueden dar entre el resto de campos recogidos y esta variable (respuesta).

Los campos son los siguientes:

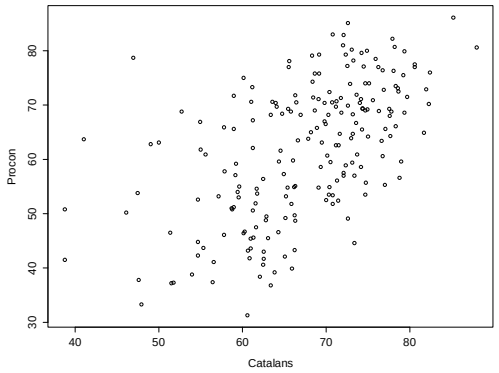
```
Municipi: Nombre de Municipio
Procon: Porcentaje de votos en 2012 a partidos pro-consulta
Poblacio: Número de Habitantes
Superficie: Superficie del municipio (Km)
PercHomes: Porcentaje de varones en la población
RatioDep: Ratio de dependencia (pobl. <16 y >65 años/pobl. 16 a 65 en porcentaje)
Catalans: porcentaje de población nacida en Catalunya
TaxaImmi: Porcentaje de población nacida en países extranjeros
RFDBpc: Renta Familiar disponible básica per cápita
PIBpc: Producto Interior Bruto per cápita
Participacio: Porcentaje de participación en las elecciones de 2012
LlarEdifici: Ratio de hogares por edificio
TaxaRecollSel: Porcentaje de residuos en recogida selectiva (tasa de reciclaje)
Atur: Habitantes apuntados en el paro registrado
DimLlar: Número medio de habitantes por hogar
Prov: Provincia de Catalunya (Barcelona, Girona, Lleida y Tarragona).
```

ANCOVA

En primer lugar se trata de analizar la relación lineal que hay entre el porcentaje de población catalana del municipio y el porcentaje de votación a partidos pro-consulta 9N (variable respuesta).

1) Obtén los resultados descriptivos (resúmenes numéricos y representaciones gráficas) para ilustrar la relación reflejada en los datos. Haz una interpretación de los resultados.

```
> plot(Procon~Catalans,dades)
```




```
> cor(dades[,c("Procon","Catalans")])
            Procon  Catalans
Procon    1.0000000 0.5623716
Catalans 0.5623716 1.0000000
```

Aparentemente la relación es directa y significativa ya que la correlación parece bastante alta. Los municipios con un mayor porcentaje de población nacida en Catalunya presentan un mayor porcentaje de votos a los partidos pro-consulta.

2) Calcula el modelo lineal correspondiente. Indica que prueba/s estadística/s permite/n establecer la significación de la relación lineal (de que test se trata, que estadístico de contraste se obtiene, que p-valor y la conclusión a la que se llega)

```
> summary(mod<-lm(Procon~Catalans,dades))

Call:
lm(formula = Procon ~ Catalans, data = dades)

Residuals:
    Min      1Q  Median      3Q     Max
-25.120  -8.481   0.099   8.026  32.916

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.89552    5.49935   1.618    0.107
Catalans     0.78426    0.08093   9.690   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.34 on 203 degrees of freedom
Multiple R-squared:  0.3163,   Adjusted R-squared:  0.3129
F-statistic:  93.9 on 1 and 203 DF,  p-value: < 2.2e-16
```

Tanto el test de significación de la pendiente con un estadístico t de 9.690 y un p-valor <2e-16, como el test ANOVA de la regresión con un estadístico F de 93.9 y el mismo p-valor confirman la significación de la relación encontrada entre ambas variables. En la regresión lineal simple ambos contrastes coinciden.

A continuación se desea ver si la relación analizada en el anterior apartado es diferente según la provincia.

3) Calcula el modelo que permite discutir diferencias en la relación entre ambas variables según la provincia. Interpreta cada uno de los coeficientes obtenidos, indicando si es o no significativo.

Para determinar si hay diferencias en la relación se debe ajustar el modelo ANCOVA para estimar las relaciones lineales por grupos.

```
> summary(mod<-lm(Procon~Catalans*Prov,dades))

Call:
lm(formula = Procon ~ Catalans * Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7308  -5.6915   0.0573   5.7802  23.3727

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -26.29378    6.38273  -4.120 5.59e-05 ***
Catalans         1.23387    0.09103  13.555  < 2e-16 ***
ProvG           57.74324   11.31342   5.104 7.81e-07 ***
ProvL           66.31244   17.58067   3.772 0.000214 ***
ProvT           22.62025   12.00831   1.884 0.061077 .
Catalans:ProvG  -0.63300    0.17362  -3.646 0.000341 ***
Catalans:ProvL  -0.77393    0.26252  -2.948 0.003584 **
Catalans:ProvT  -0.24367    0.18412  -1.323 0.187217
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.01 on 197 degrees of freedom
Multiple R-squared:  0.6017,   Adjusted R-squared:  0.5875
F-statistic: 42.51 on 7 and 197 DF,  p-value: < 2.2e-16
```

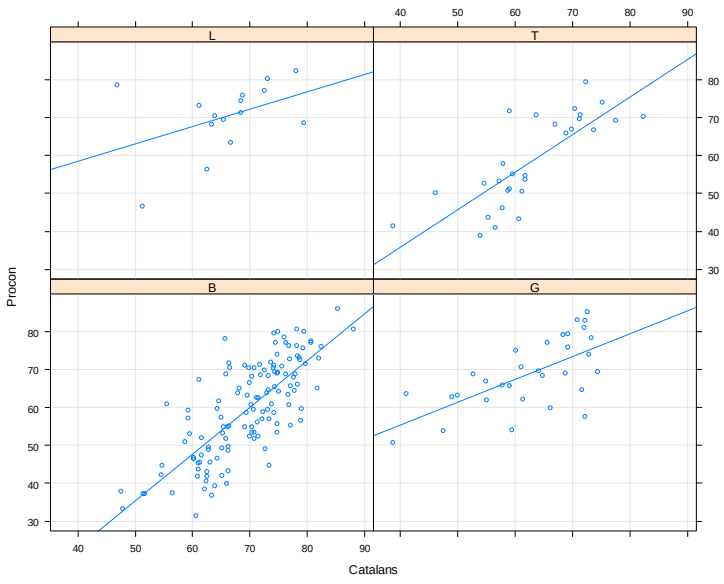
El intercept y el coeficiente de la variable Catalans permiten obtener la ecuación de la recta de la relación entre ambas variables en el grupo de municipios de referencia. Puesto que el contraste activo es de tipo baseline con la primera categoría como referencia, esto indica que se trata del grupo de municipios de Barcelona. Ambos coeficientes son significativamente diferentes de cero.

Los coeficientes ProvG, ProvL y ProvT corresponde a estimar cambios en el nivel de la recta de referencia y por lo tanto permite establecer si hay diferencias en el intercept en cada grupo en relación al grupo de Barcelona. En Girona y Lleida hay diferencias (intercepts superiores) pero en Tarragona no se pueden establecer diferencias significativas respecto a Barcelona en cuanto a nivel.

Los coeficientes de las interacciones hacen referencia a cambios en las pendientes de las rectas respecto a la del grupo de Barcelona. Nuevamente en Girona y Lleida hay diferencias (en este caso pendientes inferiores, ya que el coeficiente de la interacción es negativo) pero no se encuentran diferencias significativas entre la pendiente del grupo de Tarragona y la de Barcelona.

Gráficamente:

```
> library(lattice)
> xyplot(Procon~Catalans|Prov,dades,type=c("p","r","g"))
```



4) Indica si hay alguna diferencia significativa de forma global por provincias en el nivel y/o la pendiente de la relación a medida que aumenta el porcentaje de catalanes, incluyendo el test para el nivel y para la pendiente, los estadísticos y los p-valores que permiten determinar esta significación.

Usando el método anova obtenemos la tabla de contrates secuenciales

```
> anova(mod)
Analysis of Variance Table

Response: Procon
              Df  Sum Sq Mean Sq  F value    Pr(>F)
Catalans       1 10035.6 10035.6 156.4178 < 2.2e-16 ***
Prov           3  7857.1  2619.0  40.8212 < 2.2e-16 ***
Catalans:Prov  3  1199.9   400.0   6.2339 0.0004588 ***
Residuals    197 12639.3    64.2
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Usando el método Anova del paquete car con la suma de tipo III obtenemos la tabla de contrastes marginales (efectos netos de cada variable)

```
> Anova(mod,type="III")
Anova Table (Type III tests)

Response: Procon
              Sum Sq  Df  F value    Pr(>F)
(Intercept)   1088.8   1  16.9704 5.586e-05 ***
Catalans     11788.9   1 183.7446 < 2.2e-16 ***
Prov          2169.2   3  11.2700 7.417e-07 ***
Catalans:Prov 1199.9   3   6.2339 0.0004588 ***
Residuals    12639.3 197
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Para la interacción obtenemos el mismo resultado (F=6.2, p-valor=0.0004588) que establece que existe alguna diferencia entre las pendientes de los grupos.
Para la variable categórica Prov, ambos test dan p-valores inferiores al nivel de significación lo cual establece alguna diferencia significativa en el nivel (intercept) de los grupos.

5) ¿Hay diferencias significativas entre las pendientes de municipios de Girona y Lleida? ¿Se puede considerar que la recta de la relación en Lleida es horizontal? ¿Y en el caso de Girona? Indica los tests, valor de los estadísticos y los p-valores correspondientes. Valora los resultados obtenidos

Opción A: Para comparar las pendientes de Girona y Lleida a partir del modelo estimado podemos utilizar el método linearHypothesis del paquete car para determinar si los parámetros de las correspondientes interacciones son o no estadísticamente equivalentes:

```
> linearHypothesis(mod,"Catalans:ProvG-Catalans:ProvL=0")
Linear hypothesis test

Hypothesis:
Catalans:ProvG - Catalans:ProvL = 0

Model 1: restricted model
Model 2: Procon ~ Catalans * Prov

  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    198 12655
2    197 12639  1    15.447 0.2408 0.6242
```

La comparación de ambos modelos pone de manifiesto que se pueden considerar equivalentes ambas pendientes (F=0.24, p-valor=0.6242)

Opción B: Otra forma de hacerlo es situar Una de las dos provincias como categoría de referencia y re-estimar el modelo:

```
> dades$Prov=factor(dades$Prov,levels=c("G","B","L","T"))
> summary(mod<-lm(Procon~Catalans*Prov,dades))

Call:
lm(formula = Procon ~ Catalans * Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7308  -5.6915   0.0573   5.7802  23.3727

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     31.4495     9.3410   3.367 0.000914 ***
Catalans         0.6009     0.1478   4.064 6.96e-05 ***
ProvB          -57.7432    11.3134  -5.104 7.81e-07 ***
ProvL            8.5692    18.8572   0.454 0.650023
ProvT          -35.1230    13.8099  -2.543 0.011749 *
Catalans:ProvB   0.6330     0.1736   3.646 0.000341 ***
Catalans:ProvL  -0.1409     0.2872  -0.491 0.624207
Catalans:ProvT   0.3893     0.2179   1.787 0.075488 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.01 on 197 degrees of freedom
Multiple R-squared:  0.6017,   Adjusted R-squared:  0.5875
F-statistic: 42.51 on 7 and 197 DF,  p-value: < 2.2e-16
```

El p-valor de la interacción es el mismo que el obtenido con el anterior método (t=-0.49, p-valor=0.6242)

Opción A: Para comparar la pendiente de la recta del grupo de Lleida con el valor de referencia 0, utilizando el contraste lineal, hay que tener en cuenta que para calcular la pendiente del grupo de Lleida se ha de sumar el término de la variable Catalans (pendiente del grupo de Barcelona) y la interacción correspondiente:

```
> linearHypothesis(mod,"Catalans+Catalans:ProvL=0")
Linear hypothesis test

Hypothesis:
Catalans  + Catalans:ProvL = 0

Model 1: restricted model
Model 2: Procon ~ Catalans * Prov

  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1    198 12863
2    197 12639  1    223.86 3.4891 0.06326 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

En este caso, el p-valor de 0.06326 indica que no es significativamente diferente de cero, a pesar de encontrase cerca del nivel de significación.

Opción B: También recodificando se obtiene el mismo resultado:

```
> dades$Prov=factor(dades$Prov,levels=c("L","B","G","T"))
> summary(mod<-lm(Procon~Catalans*Prov,dades))

Call:
lm(formula = Procon ~ Catalans * Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-19.7308  -5.6915   0.0573   5.7802  23.3727

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     40.0187    16.3811   2.443 0.015448 *
Catalans         0.4599     0.2462   1.868 0.063259 .
```

```
ProvB          -66.3124    17.5807  -3.772 0.000214 ***
ProvG           -8.5692    18.8572  -0.454 0.650023
ProvT          -43.6922    19.2821  -2.266 0.024542 *
Catalans:ProvB   0.7739     0.2625   2.948 0.003584 **
Catalans:ProvG   0.1409     0.2872   0.491 0.624207
Catalans:ProvT   0.5303     0.2937   1.806 0.072510 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.01 on 197 degrees of freedom
Multiple R-squared:  0.6017,   Adjusted R-squared:  0.5875
F-statistic: 42.51 on 7 and 197 DF,  p-value: < 2.2e-16
```

Hacemos lo mismo para comparar la pendiente de Girona con el valor de referencia 0 (sólo se incluye la opción A, contraste lineal):

```
> linearHypothesis(mod,"Catalans+Catalans:ProvG=0")
Linear hypothesis test

Hypothesis:
Catalans  + Catalans:ProvG = 0

Model 1: restricted model
Model 2: Procon ~ Catalans * Prov

  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    198 13699
2    197 12639  1    1059.8 16.518 6.959e-05 ***
```

El p-valor indica que no podemos considerar que la pendiente de Girona sea 0, con un p-valor<0.0001
Nota: Puede parecer contradictorio que se pueda considerar horizontal la recta de Lleida, que no se haya encontrPot semblar contradictori que podem considerar horitzontal el pendent de Lleida, que no hi hagi diferencia entre els pendents de Lleida i Girona y sin embargo la pendiente de Girona sea significativamente diferente de 0. Hay que recordar que no tiene por que cumplirse la transitividad en la significación estadística.

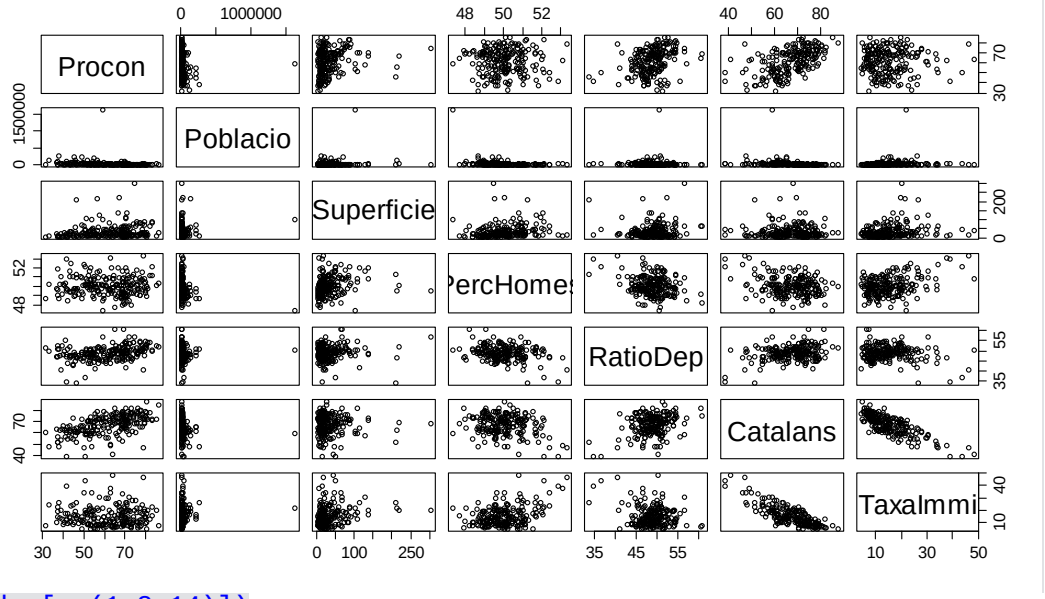
Regresión Lineal Múltiple

Se desea trabajar con las 13 variables explicativas numéricas, inicialmente sin incluir la variable categórica.

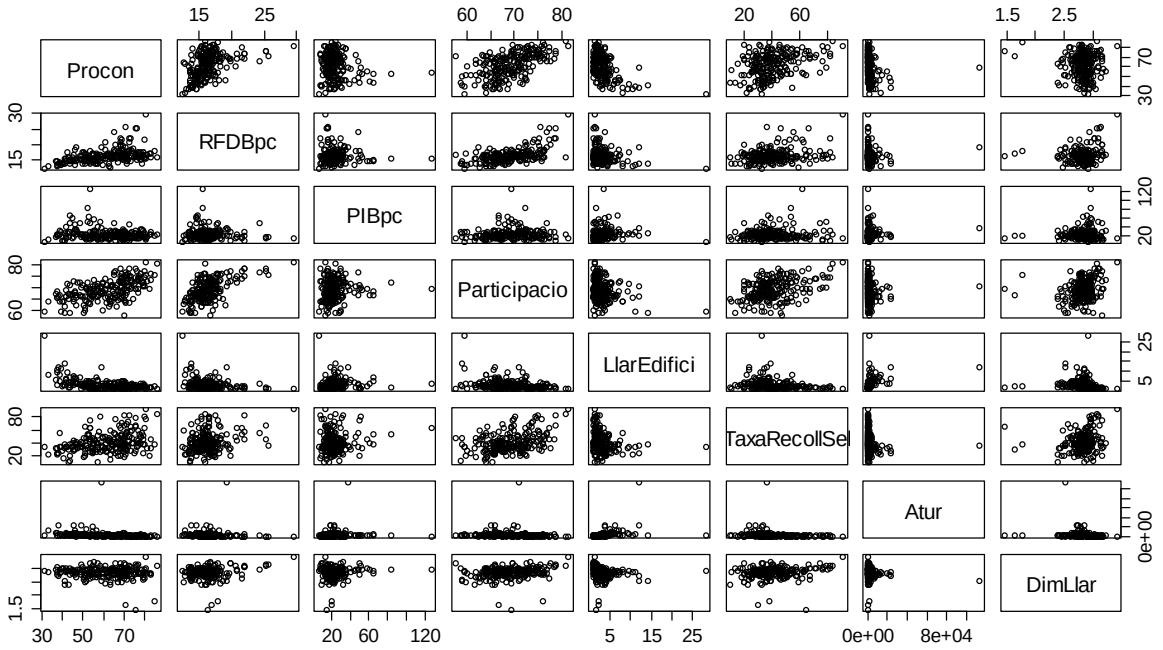
6) Realiza un análisis exploratorio preliminar para determinar si es conveniente transformar alguna variable. Utiliza el procedimiento “stepwise” (`direction="both"`) basado en el criterio BIC y partiendo del modelo que contiene todas las variables numéricas (transformadas y no transformadas) para determinar el mejor modelo.

Si hacemos los matrixplot en dos bloques (para ver claramente la disposición de los puntos):

```
> pairs(dades[,1:7])
```




```
> pairs(dades[,c(1,8:14)])
```



Parece que las variables que claramente deben modificar su métrica son: población (Barcelona y Madrid aparecen como atípicos), Superfície, PIBpc y Atur también por la presencia de atípicos. También se podría considerar transformar LlarEdifici y DimLlar, pero su rango es pequeño y la transformación tampoco introduciría cambios importantes.

```
> dades$lnPIBpc=log(dades$PIBpc)
> dades$lnSup=log(dades$Superficie)
> dades$lnPoblacio=log(dades$Poblacio)
> dades$lnAtur=log(dades$Atur)
```

Con el modelo que incluye todas las variables numéricas y las cuatro transformadas, aplicamos la regresión Stepwise con el criterio BIC (k=número de datos)

```
> summary(mod<-lm(Procon~.-Prov,dades))

Call:
lm(formula = Procon ~ . - Prov, data = dades)

Residuals:
```

```
     Min       1Q   Median       3Q      Max
-12.4781  -2.3854   0.3829   2.4516   8.0142

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.968e+00  2.744e+01  -0.072  0.94290
Poblacio       1.744e-05  2.402e-05   0.726  0.46889
Superficie    -2.334e-02  1.275e-02  -1.830  0.06887 .
PercHomes     -1.805e+00  4.149e-01  -4.350 2.23e-05 ***
RatioDep       8.986e-02  9.195e-02   0.977  0.32969
Catalans       1.721e+00  8.789e-02  19.584  < 2e-16 ***
TaxaImmi       1.763e+00  8.127e-02  21.694  < 2e-16 ***
RFDBpc         8.334e-02  1.825e-01   0.457  0.64846
PIBpc         -1.198e-01  5.498e-02  -2.179  0.03059 *
Participacio   7.219e-01  1.026e-01   7.036 3.63e-11 ***
LlarEdifici   -4.405e-02  1.444e-01  -0.305  0.76061
TaxaRecollSel -2.049e-02  2.118e-02  -0.968  0.33449
Atur          -2.006e-04  3.475e-04  -0.577  0.56438
DimLlar       -4.199e+00  1.474e+00  -2.849  0.00488 **
lnPIBpc        3.153e+00  1.739e+00   1.813  0.07140 .
lnSup          1.446e+00  5.842e-01   2.476  0.01418 *
lnPoblacio    -8.925e+00  2.213e+00  -4.033 8.00e-05 ***
lnAtur         6.127e+00  2.001e+00   3.061  0.00253 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.83 on 187 degrees of freedom
Multiple R-squared:  0.9135,   Adjusted R-squared:  0.9057
F-statistic: 116.2 on 17 and 187 DF,  p-value: < 2.2e-16

> summary(m2<-step(mod,direction="both",k=log(nrow(dades))))
:     :     :     :     :     :
Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnPoblacio + lnAtur, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2743  -2.3577   0.0813   2.5110   9.3220

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.91449   22.43352  -0.219  0.82682
PercHomes    -1.65115    0.36461  -4.529 1.03e-05 ***
Catalans      1.85297    0.06883  26.920  < 2e-16 ***
TaxaImmi      1.85418    0.07101  26.110  < 2e-16 ***
Participacio  0.67302    0.08929   7.537 1.71e-12 ***
DimLlar      -5.41817    1.42962  -3.790  0.00020 ***
lnPoblacio   -7.65208    1.78576  -4.285 2.86e-05 ***
lnAtur        5.39839    1.62523   3.322  0.00107 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.914 on 197 degrees of freedom
Multiple R-squared:  0.9049,   Adjusted R-squared:  0.9015
F-statistic: 267.7 on 7 and 197 DF,  p-value: < 2.2e-16
```

7) Para el modelo obtenido anteriormente, valora si es útil para interpretar las relaciones obtenidas (en caso de no ser adecuado, indica el problema y plantea un modelo alternativo). Con el modelo finalmente seleccionado realiza las interpretaciones correspondientes para cada coeficiente.

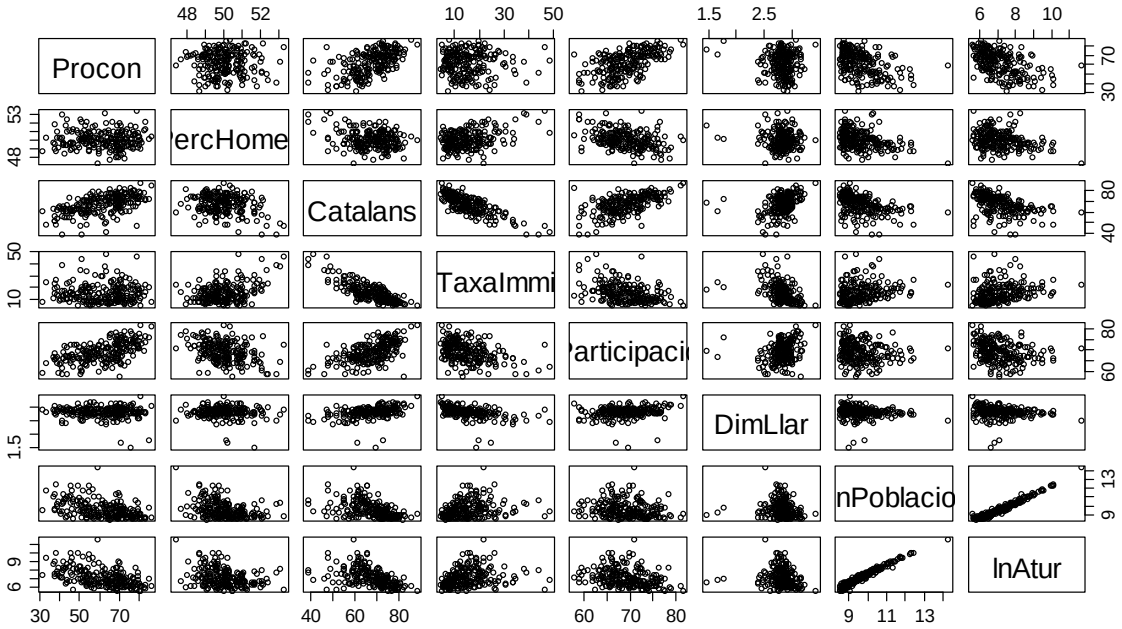
Hay que tener en cuenta la posible multicolinealidad del modelo. Si la hay, la interpretación no se puede hacer directamente sobre los coeficientes. Aunque el modelo encontrado sea ligeramente diferente, es necesario en cualquier caso revisar los VIFs, que han de ser próximos a 1 para garantizar que no nos encontramos en presencia de multicolinealidad:

```
> library(car)
> vif(m2)
```

```
   PercHomes     Catalans     TaxaImmi Participacio      DimLlar    lnPoblacio
    1.857071     5.045055     4.286212     2.300646     1.302367     37.674978
       lnAtur
   38.119471
```

Claramente, el logaritmo de la población y el paro están correlacionados, dando lugar a unos VIF que no permitirían la interpretación directa de los coeficientes. También se apunta una cierta correlación entre la proporción de catalans y la taxa de inmigración (obviamente, correlación negativa). Podemos confirmarlo con el matrixplot de los predictores.

```
> pairs(formula(m2),dades)
```



Las otras correlaciones no parecen excesivamente elevadas. Se debe seleccionar uno de los dos predictores (lnPob o lnAtur) para eliminar la multicolinealidad de los predictores. Seleccionamos el modelo con mayor R2:

```
> summary(m3a<-update(m2,.~.-lnPoblacio))

Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnAtur, data = dades)

Residuals:
    Min      1Q  Median      3Q     Max
-12.867  -2.593   0.228   2.980   8.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.84456   22.23347  -1.567 0.118662
PercHomes     -1.24365    0.36710  -3.388 0.000850 ***
Catalans       1.83413    0.07164  25.602  < 2e-16 ***
TaxaImmi       1.76590    0.07088  24.915  < 2e-16 ***
Participacio   0.50888    0.08412   6.049 7.14e-09 ***
DimLlar       -6.02198    1.48372  -4.059 7.10e-05 ***
lnAtur        -1.39857    0.36926  -3.787 0.000202 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.082 on 198 degrees of freedom
Multiple R-squared:  0.896,    Adjusted R-squared:  0.8929
F-statistic: 284.3 on 6 and 198 DF,  p-value: < 2.2e-16

> summary(m3b<-update(m2,.~.-lnAtur))

Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnPoblacio, data = dades)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.9118  -2.4807   0.3328   2.9753   9.0223

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -18.26753   22.62266  -0.807 0.420354
PercHomes     -1.45371    0.36873  -3.942 0.000112 ***
Catalans       1.81817    0.06973  26.073  < 2e-16 ***
TaxaImmi       1.78088    0.06919  25.740  < 2e-16 ***
Participacio   0.54278    0.08223   6.601 3.67e-10 ***
DimLlar       -5.87777    1.45852  -4.030 7.95e-05 ***
lnPoblacio    -1.86296    0.39877  -4.672 5.51e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.012 on 198 degrees of freedom
Multiple R-squared:  0.8995,   Adjusted R-squared:  0.8965
F-statistic: 295.5 on 6 and 198 DF,  p-value: < 2.2e-16
```

Son bastante similares, pero la mayor R2 la encontramos en el modelo en que se elimina la variable lnAtur. Veamos los VIFs resultantes:

```
> vif(m3b)
   PercHomes     Catalans     TaxaImmi Participacio      DimLlar    lnPoblacio
    1.807715     4.928193     3.872216     1.857056     1.290167     1.788067
```

No son excesivos, aunque podemos eliminar una de las dos variables con mayor VIF (Catalans o TaxaImmi) pero la pérdida en R2 es mayor. Aún así sería correcto.

Si hacemos la interpretación, podemos mirar directamente los coeficientes porque no se ha transformado la variable respuesta:

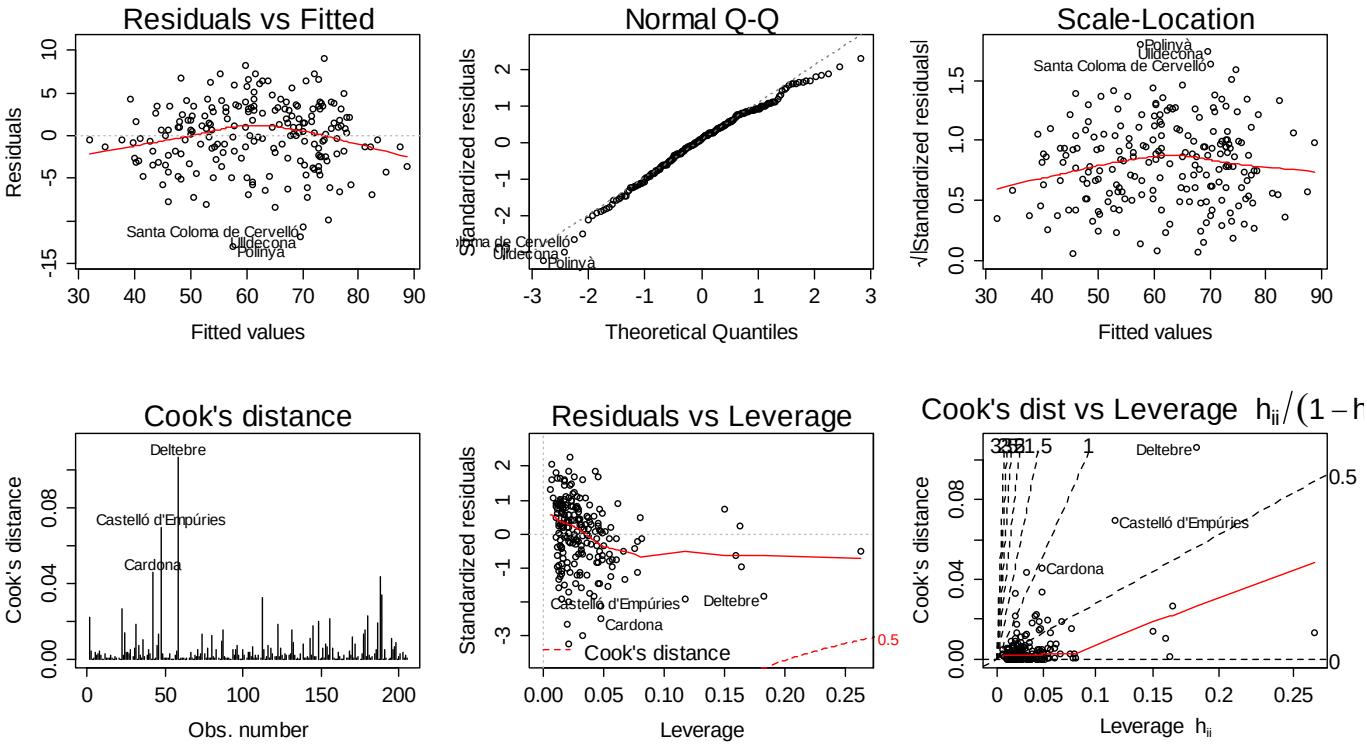
- Cada incremento en una unidad del porcentaje de hombres, disminuye en 1.45 la proporción de votos proconsulta.
- El incremento de una unidad de la proporción de catalanes, sube en 1.8 la proporción de votos proconsulta.
- El aumento de una unidad en la tasa de inmigración supone un incremento de 1.78 en la respuesta.
- Cada unidad que se incremente la participación supone un aumento de 0.5 puntos en la respuesta.
- Si aumenta en una unidad la dimensión media del hogar, se reduce en 5.87 unidades la proporción de votos proconsulta.
- Si la población en la escala logarítmica aumenta en una unidad, la respuesta disminuye en 1.86.

La variabilidad de la respuesta explicada por el modelo es del 89.95%

8) También para el modelo obtenido anteriormente, realiza la validación de las premisas, incorporando los resultados numéricos, tests y salidas gráficas necesarias. Especifica la premisa que se valida en cada gráfico. Identifica las observaciones siguientes: la más atípica, la más influyente a priori y la más influyente a posteriori.

Inicialmente, obtenemos los plots habituales:

```
> par(mfrow=c(2,3))
> plot(m3b,which=1:6)
> par(mfrow=c(1,1))
```



El primer plot permite discutir la linealidad y la varianza constante. Aunque pueda parecer que hay una cierta curvatura, en los extremos (valores de predicción extremos) hay pocas observaciones, lo cual podría cuestionar la curvatura obtenida mediante el ajuste suave.

El plot de normalidad parece validar la distribución gaussiana de los residuos. En la parte superior parece que hay algunos valores extremos, pero los que etiqueta como atípicos son 3 valores con residuos negativos (Sta. Coloma de Cervelló, Ulldecona i Polinyà). De aquí ya podemos afirmar que el dato más atípico és Polinyà con un residuo estandarizado menor que -3.

También conviene utilizar tests para validar la premisa de normalidad

```
> res=rstudent(m3b)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.985, p-value = 0.02798
```

Según este resultado, con una significación del 0.05 rechazaríamos la hipótesis de normalidad (posiblemente por la distribución de la cola superior que se aleja de la recta de referencia)

El tercer plot valida la homogeneidad de varianza como función de las predicciones.

El siguiente plot representa la distancia de Cook para cada observación y marca los valores más grandes, que son los datos más influyentes (Deltebre, Castelló d'Empúries i Cardona).

El siguiente plot representa los errores estandarizados como función del factor de apalancamiento (leverage). También indica la distancia de Cook en forma de curvas de nivel. Tanto en el anterior como en este y en el siguiente, el dato con una mayor distancia de Cook (dato más influyente a posteriori) sería Deltebre con una distancia de Cook superior a 0.1.

Para saber qué población es la más influyente a priori, hace falta determinar cuál es la que tiene mayor leverage:

```
> sort(hatvalues(m3b),decreasing=T)[1:3]
Llagostera   Deltebre   Banyoles
 0.2629989  0.1834462  0.1647402
```

El dato más influyente a priori es Llagostera con un leverage de 0.26 (es el punto que aparece más a la derecha en los dos últimos plots).

Modelo Lineal General

9) Utiliza el procedimiento "stepwise" (`direction="both"`) basado en el criterio BIC y partiendo del modelo que contiene todas las variables numéricas (transformadas y no transformadas), la variable categórica y las interacciones entre la variable categórica y cada una de las numéricas. Refina el modelo en caso de que presente algún problema para su interpretación.

Introducimos ahora los mismos predictores numéricos, pero añadimos la variable categórica Provincia y las interacciones entre ella y el resto:

```
> summary(mod<-lm(Procon~(.-Prov)*Prov,dades))

Call:
lm(formula = Procon ~ (. - Prov) * Prov, data = dades)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7470 -1.6519  0.1354  1.8628  7.9822

Coefficients: (3 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -1.402e+01  3.998e+01  -0.351  0.72646
Poblacio            3.859e-05  2.474e-05   1.560  0.12119
Superficie          3.130e-02  4.452e-02   0.703  0.48321
PercHomes          -1.592e+00  5.862e-01  -2.715  0.00748 **
RatioDep            2.267e-01  1.246e-01   1.820  0.07101 .
Catalans            1.675e+00  1.299e-01  12.893  < 2e-16 ***
TaxaImmi            1.624e+00  1.478e-01  10.988  < 2e-16 ***
RFDBpc              4.435e-01  2.440e-01   1.818  0.07131 .
PIBpc              -1.723e-01  8.824e-02  -1.953  0.05287 .
Participacio        7.686e-01  1.415e-01   5.432 2.49e-07 ***
LlarEdifici        -2.284e-01  1.689e-01  -1.352  0.17852
TaxaRecollSel       3.341e-02  2.863e-02   1.167  0.24519
Atur               -5.235e-04  3.597e-04  -1.455  0.14790
DimLlar            -4.482e+00  3.230e+00  -1.388  0.16756
lnPIBpc             3.971e+00  2.503e+00   1.586  0.11501
lnSup               5.867e-02  1.010e+00   0.058  0.95375
lnPoblacio         -1.570e+01  3.051e+00  -5.146 9.11e-07 ***
lnAtur              1.397e+01  2.862e+00   4.880 2.93e-06 ***
ProvG              -1.270e+02  8.252e+01  -1.538  0.12626
ProvL              -3.612e+02  5.597e+02  -0.645  0.51972
ProvT              -9.534e+01  1.399e+02  -0.682  0.49659
Poblacio:ProvG      1.560e-04  4.235e-04   0.368  0.71323
Poblacio:ProvL      2.084e-03  1.922e-03   1.084  0.28010
Poblacio:ProvT     -4.909e-05  4.004e-04  -0.123  0.90260
Superficie:ProvG    1.405e-01  1.760e-01   0.798  0.42605
Superficie:ProvL   -1.117e-01  8.268e-02  -1.351  0.17897
Superficie:ProvT   -1.236e-01  6.297e-02  -1.962  0.05177 .
PercHomes:ProvG     2.366e+00  1.362e+00   1.737  0.08467 .
PercHomes:ProvL     5.538e+00  6.188e+00   0.895  0.37236
PercHomes:ProvT     3.450e-01  1.804e+00   0.191  0.84863
RatioDep:ProvG      3.612e-01  4.855e-01   0.744  0.45820
RatioDep:ProvL      5.867e-01  8.998e-01   0.652  0.51545
RatioDep:ProvT     -1.581e-01  4.631e-01  -0.341  0.73336
Catalans:ProvG     -2.195e-01  5.182e-01  -0.424  0.67248
Catalans:ProvL      5.851e-01  1.278e+00   0.458  0.64780
Catalans:ProvT     -9.381e-02  3.364e-01  -0.279  0.78076
TaxaImmi:ProvG     -3.376e-01  4.122e-01  -0.819  0.41421
TaxaImmi:ProvL      9.495e-01  1.057e+00   0.898  0.37058
TaxaImmi:ProvT     -1.635e-01  3.756e-01  -0.435  0.66408
RFDBpc:ProvG        1.479e+00  9.902e-01   1.494  0.13758
RFDBpc:ProvL        4.552e+00  4.494e+00   1.013  0.31286
RFDBpc:ProvT        1.069e-02  1.410e+00   0.008  0.99396
PIBpc:ProvG        -9.329e-02  1.075e+00  -0.087  0.93097
PIBpc:ProvL        -3.808e-01  5.487e+00  -0.069  0.94477
PIBpc:ProvT         4.182e-02  1.694e-01   0.247  0.80541
Participacio:ProvG -7.900e-01  5.193e-01  -1.521  0.13056
Participacio:ProvL -8.416e-01  2.393e+00  -0.352  0.72559
Participacio:ProvT -9.438e-02  5.511e-01  -0.171  0.86426
LlarEdifici:ProvG   5.475e-01  1.338e+00   0.409  0.68312
LlarEdifici:ProvL   4.941e+00  6.102e+00   0.810  0.41954
LlarEdifici:ProvT  -1.836e-02  6.950e-01  -0.026  0.97897
```

```
TaxaRecollSel:ProvG -1.493e-01  1.088e-01  -1.373  0.17211
TaxaRecollSel:ProvL  2.834e-02  2.409e-01   0.118  0.90652
TaxaRecollSel:ProvT -3.659e-02  9.202e-02  -0.398  0.69156
Atur:ProvG          -2.457e-03  4.905e-03  -0.501  0.61725
Atur:ProvL          -2.354e-02  2.193e-02  -1.074  0.28487
Atur:ProvT          -1.110e-03  4.468e-03  -0.248  0.80419
DimLlar:ProvG        3.928e+00  4.020e+00   0.977  0.33027
DimLlar:ProvL       -1.306e+01  2.286e+01  -0.571  0.56888
DimLlar:ProvT       -4.648e-01  8.435e+00  -0.055  0.95614
ProvG:lnPIBpc        2.786e+00  2.330e+01   0.120  0.90503
ProvL:lnPIBpc        4.783e+00  9.646e+01   0.050  0.96052
ProvT:lnPIBpc        3.349e+00  8.643e+00   0.388  0.69897
ProvG:lnSup         -4.495e+00  6.068e+00  -0.741  0.46010
ProvL:lnSup                 NA         NA      NA       NA
ProvT:lnSup          4.832e+00  3.201e+00   1.509  0.13349
ProvG:lnPoblacio     1.261e+01  1.057e+01   1.193  0.23511
ProvL:lnPoblacio            NA         NA      NA       NA
ProvT:lnPoblacio     1.956e+01  1.380e+01   1.418  0.15859
ProvG:lnAtur        -1.091e+01  9.449e+00  -1.155  0.25008
ProvL:lnAtur                NA         NA      NA       NA
ProvT:lnAtur        -1.437e+01  1.359e+01  -1.058  0.29210
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.593 on 136 degrees of freedom
Multiple R-squared:  0.9447,   Adjusted R-squared:  0.917
F-statistic: 34.15 on 68 and 136 DF,  p-value: < 2.2e-16
```

Ahora aplicamos el mecanismo stepwise de selección de variables:

```
> summary(m2<-step(mod,direction="both",k=log(nrow(dades))))
```

```
:     :     :     :     :     :     :
Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnPoblacio + lnAtur + Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-12.9246  -2.2684   0.1728   2.6485   8.1776

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.33253   21.46271  -0.481 0.630762
PercHomes    -1.61364    0.34696  -4.651 6.11e-06 ***
Catalans      1.78294    0.06953  25.642  < 2e-16 ***
TaxaImmi      1.69104    0.08071  20.953  < 2e-16 ***
Participacio  0.72406    0.09184   7.884 2.24e-13 ***
DimLlar      -4.08048    1.42093  -2.872 0.004537 **
lnPoblacio   -7.67200    1.71217  -4.481 1.27e-05 ***
lnAtur        5.76175    1.55995   3.694 0.000287 ***
ProvG         4.49675    0.98026   4.587 8.04e-06 ***
ProvL         3.19643    1.23965   2.578 0.010664 *
ProvT         0.95446    0.94872   1.006 0.315646
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.724 on 194 degrees of freedom
Multiple R-squared:  0.9152,   Adjusted R-squared:  0.9108
F-statistic: 209.4 on 10 and 194 DF,  p-value: < 2.2e-16
```

No aparece ninguna interacción, pero la variable provincial sí que es significativa. Nuevamente aparece el fenómeno de multicolinealidad.

```
> vif(m2)
                 GVIF Df GVIF^(1/(2*Df))
PercHomes    1.858012  1        1.363089
Catalans     5.688228  1        2.385001
TaxaImmi     6.116839  1        2.473224
Participacio 2.689070  1        1.639838
DimLlar      1.421550  1        1.192288
lnPoblacio  38.267442  1        6.186068
lnAtur      38.802763  1        6.229186
Prov         2.699336  3        1.179984
```

```
> summary(m3a<-update(m2,.~.-lnPoblacio))

Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnAtur + Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5085  -2.5892   0.3996   3.0162   8.1164

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -40.87034   21.32429  -1.917  0.05675 .
PercHomes     -1.21087    0.35112  -3.449  0.00069 ***
Catalans       1.76834    0.07277  24.299  < 2e-16 ***
TaxaImmi       1.60755    0.08228  19.538  < 2e-16 ***
Participacio   0.55934    0.08818   6.343 1.54e-09 ***
DimLlar       -4.52684    1.48516  -3.048  0.00262 **
lnAtur        -1.05255    0.36408  -2.891  0.00428 **
ProvG          4.64592    1.02650   4.526 1.04e-05 ***
ProvL          2.73697    1.29443   2.114  0.03575 *
ProvT          1.12700    0.99323   1.135  0.25790
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.902 on 195 degrees of freedom
Multiple R-squared:  0.9064,   Adjusted R-squared:  0.9021
F-statistic: 209.9 on 9 and 195 DF,  p-value: < 2.2e-16

> summary(m3b<-update(m2,.~.-lnAtur))

Call:
lm(formula = Procon ~ PercHomes + Catalans + TaxaImmi + Participacio +
    DimLlar + lnPoblacio + Prov, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-12.5700  -2.5536   0.4237   2.8686   8.2579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -24.84522   21.77322  -1.141 0.255232
PercHomes     -1.40743    0.35336  -3.983 9.61e-05 ***
Catalans       1.75397    0.07129  24.602  < 2e-16 ***
TaxaImmi       1.62587    0.08127  20.007  < 2e-16 ***
Participacio   0.57990    0.08578   6.760 1.56e-10 ***
DimLlar       -4.47437    1.46213  -3.060 0.002524 **
lnPoblacio    -1.50690    0.39355  -3.829 0.000173 ***
ProvG          4.44206    1.01142   4.392 1.84e-05 ***
ProvL          2.64581    1.26992   2.083 0.038514 *
ProvT          0.96094    0.97899   0.982 0.327533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.843 on 195 degrees of freedom
Multiple R-squared:  0.9093,   Adjusted R-squared:  0.9051
F-statistic: 217.1 on 9 and 195 DF,  p-value: < 2.2e-16
```

Nos quedamos con el Segundo modelo, que es similar al obtenido anteriormente, pero se añade la variable categórica.

```
> anova(m3b)
Analysis of Variance Table

Response: Procon
             Df  Sum Sq Mean Sq   F value    Pr(>F)
PercHomes     1     7.5     7.5    0.5096 0.4761787
Catalans      1 10523.0 10523.0  712.5901 < 2.2e-16 ***
TaxaImmi      1 16775.6 16775.6 1136.0002 < 2.2e-16 ***
Participacio  1   613.5   613.5   41.5476 8.822e-10 ***
DimLlar       1   273.2   273.2   18.4980 2.687e-05 ***
lnPoblacio    1   351.4   351.4   23.7946 2.220e-06 ***
Prov          3   308.1   102.7    6.9540 0.0001806 ***
Residuals   195  2879.6    14.8
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> vif(m3b)
                 GVIF Df GVIF^(1/(2*Df))
PercHomes    1.809901  1        1.345326
Catalans     5.615837  1        2.369776
TaxaImmi     5.824455  1        2.413391
Participacio 2.203376  1        1.484377
DimLlar      1.413542  1        1.188925
lnPoblacio   1.898722  1        1.377941
Prov         2.651802  3        1.176495
```

También podría refinarse eligiendo entre Catalans y TaxaImmi.

10) Comenta las siguientes afirmaciones, indicando si son correctas o no y justificando la respuesta en base al modelo final obtenido.

Nota: Si el modelo encontrado en la cuestión 9 es diferente, la valoración de la respuesta se determina en base al modelo encontrado.

a. “Existe relación directa significativa entre la proporción de mujeres en el municipio y la votación a partidos proconsulta en las últimas elecciones autonómicas”

Verdad. Para el modelo del apartado 9, el coeficiente de PercHomes es significativamente negativo, por lo que existe relación inversa entre el porcentaje de hombres y la proporción de votos proconsulta. Esto implica relación directa significativa respecto a la proporción de mujeres, ya que PercDones=100-PercHomes.

b. “No hay diferencias significativas entre los modelos obtenidos con los municipios de Barcelona, Tarragona y Lleida”

Falso. Puesto que la variable categórica aparece en el modelo afectando al nivel pero no formando parte de ninguna interacción, quiere decir que alguna diferencia sí que existe. Si miramos los p-valores, comparar Barcelona con Tarragona no permite determinar diferencias (p-valor=0.327) pero al comparar Barcelona con Lleida obtenemos un p-valor de 0.03, que indica diferencias significativas.

c. “Los municipios con mayor votación a partidos proconsulta son municipios con mayoría de mujeres, con mayor porcentaje de catalanes y extranjeros en su población, con mayor porcentaje de participación en las elecciones, con menor población y menor número de habitantes por hogar. De entre todos los municipios con estas características, los de Girona votan más a estos partidos”

Cierto. En base al modelo obtenido y teniendo en cuenta los signos de los coeficientes, un municipio con esas características dará lugar a una mayor predicción de porcentaje de votos proconsulta. La categoría de Girona supone 4.44 puntos por encima de Barcelona, mayor incremento que el resto de provincias.