Estadística avançada
A2 Anàlisi descriptiu i inferencial

Noemi Lorente Torrelles

11 de diciembre, 2018

Contents

1 Introducció **2**

2 Model de regressió lineal **2**
2.1 Model de regressió lineal múltiple (regressors quantitatius) 5
2.2 Model de regressió lineal múltiple (regressors quantitatius i qualitatius) 9
2.3 Efectueu una predicció de la capacitat pulmonar amb els dos models 12

3 Model de regressió logística **17**
3.1 Estimació d'un model de regressió logística 17
3.2 Predicció en el model lineal generalitzat (model de regressió logística) 20
3.3 Millora del model 22
3.4 Qualitat de l'ajust 26
3.5 La selecció dels individus fumadors 27
3.6 Corba ROC 28

4 Referències **29**

1 Introducció

En aquesta activitat s'utilitzarà el fitxer Fumadores_clean_5Y_1.csv ja preparat, és a dir, després del preprocés que s'ha realitzat en la primera activitat.

En aquesta activitat, usarem el fitxer resultant de l'activitat anterior degudament preprocessat. Recordeu que aquest fitxer emmagatzema les dades d'una investigació mèdica sobre la capacitat pulmonar de vàries persones, amb l'objectiu d'estudiar si els hàbits de salut i els hàbits com a fumadors influencien la capacitat pulmonar. Per a realitzar l'estudi es va recollir una mostra de 300 persones. A cada persona, se li va preguntar a través d'un qüestionari el seu gènere, hàbits d'esport, si era fumadora, i en cas que ho fos, quants cigarrets al dia de promig fumava i els anys que feia que fumava. A més, es va mesurar la capacitat pulmonar de cada persona a partir d'un test d'aire expulsat, des d'on es va prendre com a capacitat pulmonar la mesura FEF (forced expiratory flow), que és la velocitat de l'aire sortint del pulmó durant la porció central d'una espiració forçada. Es mesura en litres/segon. Altres dades personals recollides són: l'alçada, pes i ciutat on viu. S'inclou en el fitxer una columna addicional "PC5Y" que és la capacitat pulmonar de cada persona mesurada al cap de 5 anys de realitzar el primer test. S'assumeix que la persona no ha canviat les seves condicions personals significativament en aquest temps. D'altres dades recollides són: l'alçada, el pes i la ciutat on viu.

La base de dades conté 300 registres i 10 variables. Les variables són Sex, Sport, Years, Cig, PC, City, Weight, Age, Height, PC5Y.

2 Model de regressió lineal

Primerament, estudiarem la possible associació entre la capacitat pulmonar i algunes característiques de cada individu.

Abans de carregar el fitxer, he visualitzat el seu contingut amb *gedit* i he pogut comprobar que s'utilitza la coma "," com a separador de camps. Així m'asseguro que la lectura del fitxer es realitza de forma correcta.

```
# carrego el fitxer amb read.table, separador=',', decimal='.'
capacitat <- read.table("Fumadores_clean_5Y_1.csv", header=TRUE, sep=",", na.strings="NA",
                        dec=".", strip.white=TRUE, stringsAsFactors = FALSE)
```

Amb la funció **str** puc veure l'estructura interna del data frame *capacitat*. Veig que té 300 observacions, 10 variables i es mostra el nom de les variables.

```
# la funció str mostra l'estructura interna del data frame capacitat
str(capacitat)
```

```
## 'data.frame':    300 obs. of  10 variables:
##  $ Sex   : chr  "M" "F" "M" "M" ...
##  $ Sport : chr  "N" "N" "E" "S" ...
##  $ Years : int  25 18 0 25 0 0 33 0 0 5 ...
##  $ Cig   : int  10 32 0 14 0 0 15 0 0 12 ...
##  $ PC    : num  2.58 1.56 3.75 2.76 3.49 ...
```

```
##  $ City  : chr  "Barcelona" "Terrassa" "La Bisbal" "Blanes" ...
##  $ Weight: int  65 65 69 70 72 64 69 71 72 73 ...
##  $ Age   : int  49 35 38 55 55 42 55 44 45 35 ...
##  $ Height: int  171 166 175 176 178 165 175 177 178 179 ...
##  $ PC5Y  : num  2.53 1.44 3.73 2.67 3.49 ...
```

Com R no ha assignat correctament el tipus apropiat a les variables qualitatives nominals *Sex*, *Sport* i *City*, cal fer la conversió de caràcter a factor.

```
# Canvio a variable factor la variable Sex
capacitat$Sex <- as.factor(capacitat$Sex)
```

```
# Canvio a variable factor la variable Sport
capacitat$Sport <- as.factor(capacitat$Sport)
```

```
# Canvio a variable factor la variable City
capacitat$City <- as.factor(capacitat$City)
```

Per conèixer el tipus de variable utilitzo la funció **class**. A continuació, mostro el tipus de variable de cada variable. Per mostrar els tipus de les variables en una taula, utilitzo la funció **kable** .

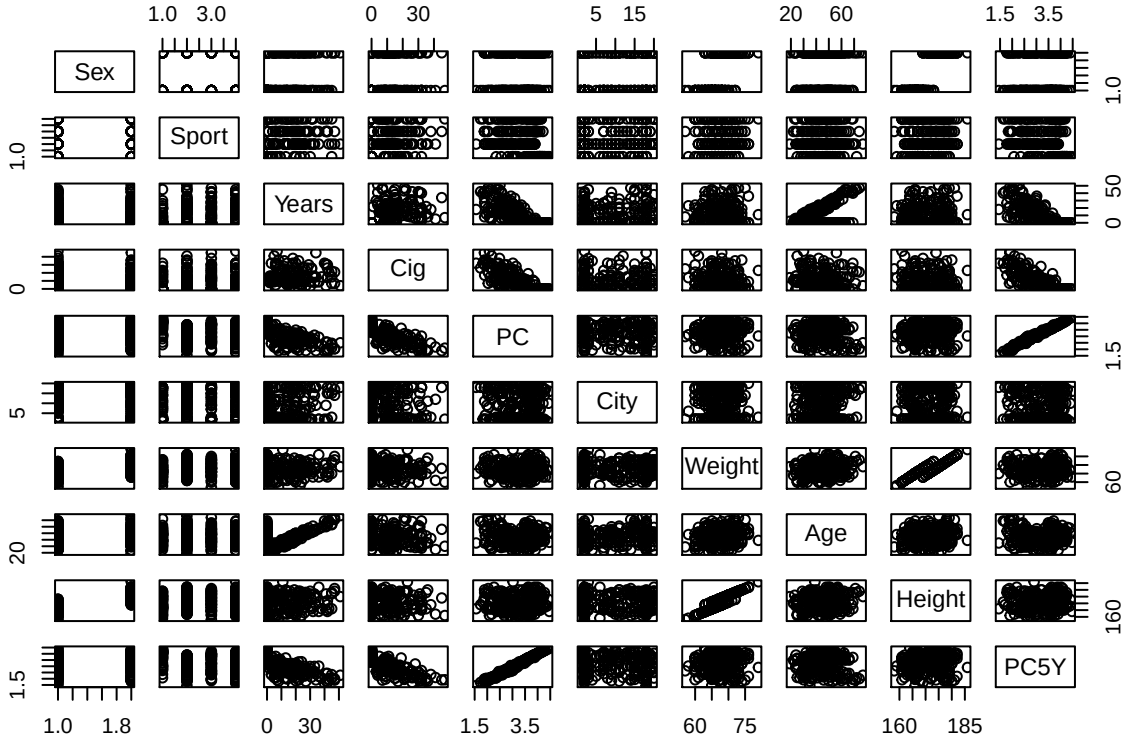
```
# Recupero el tipus de variable
tipus <- sapply(capacitat,class)
kable(data.frame(Variable=names(tipus),Classe=as.vector(tipus)), align='l',
      caption="Tipus de les variables")
```

Table 1: Tipus de les variables

Variable	Classe
Sex	factor
Sport	factor
Years	integer
Cig	integer
PC	numeric
City	factor
Weight	integer
Age	integer
Height	integer
PC5Y	numeric

A continuació, representa gràficament el diagrama de dispersió, ja que em pot ajudar molt en la cerca d'un model que descrigui la relació entre les variables. Utilitzo la funció *pairs*.

```
# Representa gràficament el diagrama de dispersió entre cada parell de variables
pairs(capacitat)
```



Observo que les variables *PC* i *PC5Y* es troben sobre una recta en pendent positiu, és a dir, a mesura que *PC* es fa gran, *PC5Y* també es fa gran. Passa el mateix entre les variables *Weight* i *Height*, i entre les variables *Years* i *Age*.

També observo que existeix algun tipus de relació entre les variables *PC* i *Cig*, i entre les variables *PC* i *Years*. Es veu una mena de dependència lineal amb pendent negatiu, ja que a mesura que el valor de *PC* augmenta, disminueixen *Cig* i *Years*. El mateix passa entre *PC5Y*, *Cig* i *Years*.

La resta de gràfics denoten que no existeix cap tipus de relació entre les variables, ja que el núvol de punts no presenta cap forma determinada, estan absolutament dispersos.

2.1 Model de regressió lineal múltiple (regressors quantitatius)

2.1.1. Estimar per mínims quadrats ordinaris un model lineal que expliqui la capacitat pulmonar (PC) d'un individu en funció de tres factors quantitatius: el pes (Weight), el nombre de cigarrets que fuma al dia (Cig), i el nombre d'anys que fa que fuma (Years).

Amb el model de regressió lineal multiple busco explicar la variable dependent o explicada *PC* amb les variables independents o explicatives *Weight*, *Cig* i *Years*, mitjançant la expressió següent:

$PC = \beta_0 + \beta_1 Weight + \beta_2 Cig + \beta_3 Years + e$

El mètode dels mínims quadrats per estimar un model lineal consisteix a cercar:

- la suma dels residus al quadrat
- determinar els paràmetres del model que fan que aquesta suma tingui un valor mínim

Els residus e_i són la diferència entre els valors observats en la mostra (y_i) i els valors estimats pel model ($\hat{y}_i$): $e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki})$, per a i =1, 2, 3, . . . , n

Matricialment es pot escriure com: $e = y - \hat{y} = y - X\beta$

on X és la matriu dels valors de les variables independents o explicatives i β és el vector dels paràmetres de la regressió.

El vector $\hat{\beta}$, el barret ens indica que es tracta d'un valor estimat, és el vector dels estimadors mínim quadràtics dels paràmetres: $\hat{\beta} = (X^t X)^{-1} X^t y$

Amb la funció *lm* obtinc els components principals de la regressió:

- *Residuals*: mostra el mínim, màxim i quartils dels residus de la regressió, els quals proporcionen informació sobre la seva distribució.
- *Coefficients*: informació de l'estimació dels paràmetres (o coeficients) estimats.
- *Estimate*: estimació de cada paràmetre (intercept significa constant).
- *Std.Error*: desviació (o error) estàndard de cada paràmetre estimat.
- *t value*: estadístic t de cada paràmetre estimat, otingut dividint l'estimació del paràmetre entre la seva desviació estàndard. Aquest estadístic és el que s'utilitza per a fer el contrast de significació individual dels paràmetres estimats.
- *Pr(>|t|)*: *p-valor* del contrast de significació individual de cada paràmetre estimat, el qual indica la seva significació estadística.
- *Signif. codes*: mostra, amb asteriscos i punts, per a quins nivells de significació els coeficients estimats són significatius o no. Per exemple, el valor estimat de la variable *Weight* ($\beta_1 = 0.001283$) no és significativa en el model lineal de regressió, en canvi, els valors estimats de la constant ($\beta_0 = 3.677888$), de la variable *Cig* ($\beta_2 = -0.032711$) i de la variable *Years* ($\beta_3 = -0.023139$) són significatius amb un nivell de significació del 0.01% ('***' 0.001).
- *Residual standard error*: desviació (o error) estàndard dels residus.
- *Multiple R-squared*: coeficient de determinació.

- *Ajusted R-squared*: coeficient de determinació ajustat.
- *F - statistic*: estadístic F per al contrast de la significació global o conjunta dels paràmetres estimats del model.
- *p-value*: p-valor associat al constrat anterior. En aquest model de regressió, veig que el conjunt de paràmetres estimats és significatiu amb un nivell de significació del 1% (p-valor<0.01).

$PC = 3.677888 + 0.001283 Weight - 0.032711 Cig - 0.023139 Years$

```
# aplico el model lineal amb la funció lm
model1 <- lm(data = capacitat, formula = PC ~ Weight+Cig+Years)
summary(model1)
```

```
##
## Call:
## lm(formula = PC ~ Weight + Cig + Years, data = capacitat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97891 -0.18424 -0.01939  0.19799  0.78591
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.677888   0.284701  12.918   <2e-16 ***
## Weight       0.001283   0.004178   0.307    0.759
## Cig         -0.032711   0.001923 -17.008   <2e-16 ***
## Years       -0.023139   0.001583 -14.613   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2734 on 296 degrees of freedom
## Multiple R-squared:  0.812,  Adjusted R-squared:  0.8101
## F-statistic: 426.3 on 3 and 296 DF,  p-value: < 2.2e-16
```

2.1.2. Avalueu la bondat d'ajust a través del coeficient de determinació (R^2). Podeu usar la instrucció d'R lm.

El coeficient de determinació R^2 indica el grau d'ajust de la recta de regressió als valors de la mostra, i es defineix com la proporció de variància explicada per la recta de regressió, és a dir: $R^2 = \frac{VariànciaExplicadaPelModel}{VariànciaTotalMostra}$

La bondat d'ajust és prou bona, $R^2 = 0.812$, ja que és més proper a 1 que a 0 i indica que el model de regressió lineal explica el 81.2% de la variància de les observacions.

2.1.3. A més, avalueu si algun dels regressors té influència significativa (p-valor del contrast individual inferior al 5%).

A la taula Coefficients puc veure els valors estimats dels paràmetres. També es mostra l'error estàndard, el valor de l'estadístic t-Student i un p-valor que serveix per a contrastar el nivell de significació del paràmetre.

Amb el contrast pretenc determinar si els efectes de la constant i de les variables independents són realment importants per a explicar la variable dependent o bé els efectes es poden considerar nuls.

Considero els contrasts d'hipòtesis següents:

$H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

$H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

Com el p-valor de la variable *Weight* (0.759) és major que 0.05, ens indica que cal acceptar la hipotesi nul.la, per tant, β_1 sempre tindrá valors iguals a 0. En canvi, com el p-valor de la resta de variables (<2e-16) són menors que 0.05, ens indica que cal rebutjar la hipotesi nul.la, per tant, podem suposar que els valors β_0, β_2 i β_3 són significativament diferents de 0.

Tot i que en l'enunciat no es demana, si el coeficient d'un regresor no és significatiu, es podria treure el regressor del model de regressió i es podria tornar a calcular de nou el model de regressió sense usar aquestes variables.

Habitualment el que es cerca és un model de regressió que relacioni correctament les variables explicatives amb la variable dependent (és a dir, que expliqui la variabilitat de la variable dependent en funció de les variables regressores) i a la vegada, que sigui senzill (que contingui el mínim nombre de regressors). Per tant, si es veu que hi ha regresssors que no influeixen significativament en el model, es poden eliminar del model final.

2.1.4. Observeu que, a diferència de Weight, no s'ha afegit al model de regressió la variable Height. Des del punt de vista de la qualitat del model de regressió, podeu indicar una raó que justifiqui no fer-ho?

Un dels problemes de la regressió lineal múltiple és la existencia d'algun tipus de dependència entre les variables explicatives. Si dues o més variables tenen entre elles una correlació alta, pot ser problemàtic incloure-les simultàniament com a variables explicatives. Aquest fet s'anomena multicol · linealitat.

Genero la matriu de correlacions per comprobar si existeix multicol · linealitat. Observo que les variables *Weight* i *Height* estan altament correlacionades (0.94144547) i per això no es convenient afegir la variable *Height* al model de regressió.

```r
# genero la matriu de correlacions
cor(capacitat[,c( "Weight", "Height", "Cig", "Years")], use="complete")
```

```
##              Weight      Height         Cig       Years
## Weight  1.00000000  0.94144547 -0.1307582 -0.01794329
## Height  0.94144547  1.00000000 -0.1462693 -0.04492902
## Cig    -0.13075821 -0.14626926  1.0000000  0.60189555
## Years  -0.01794329 -0.04492902  0.6018955  1.00000000
```

2.2 Model de regressió lineal múltiple (regressors quantitatius i qualitatius)

Estimar per mínims quadrats ordinaris un model lineal que expliqui la capacitat pulmonar (PC) d'un individu en funció de cinc regressors. A més dels tres anteriors (Years, Cig i Weight) ara s'hi afegeixen les variables Sex i Sport. Useu com a categoria de referència de la variable Sex la categoria "F" i de la variable Sport la categoria "N" (useu la funció relevel()). Es poden definir noves variables, SexR i SportR, per a dur a terme aquesta nova reordenació.

Amb la funció *lm* obtinc els components principals de la nova regressió:

```
# aplico el model lineal amb la funció lm
model2 <- lm(data = capacitat, formula = PC ~ Weight+Cig+Years+Sex+Sport)
summary(model2)
```

```
##
## Call:
## lm(formula = PC ~ Weight + Cig + Years + Sex + Sport, data = capacitat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73530 -0.10739 -0.00663  0.10509  0.48296
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.127129   0.233646  17.664  < 2e-16 ***
## Weight      -0.000601   0.003599  -0.167 0.867511
## Cig         -0.034035   0.001293 -26.315  < 2e-16 ***
## Years       -0.022633   0.001065 -21.249  < 2e-16 ***
## SexM         0.102336   0.027373   3.739 0.000223 ***
## SportN      -0.565597   0.032939 -17.171  < 2e-16 ***
## SportR      -0.195331   0.039491  -4.946 1.28e-06 ***
## SportS      -0.366005   0.034742 -10.535  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1827 on 292 degrees of freedom
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9152
## F-statistic: 462.2 on 7 and 292 DF,  p-value: < 2.2e-16
```

Observo que la funció *lm* afegeix automàticament una variable diferent per cada una de les categories de les variables qualitatives, menys una.

Però com vull determinar quina és la categoria base que cal considerar en la reordenació, utilitzo la funció *relevel*.

Afegeixo la variable **SexR** tenint en compte que la categoria de referència de la variable **Sex** és 'F'.

```
# afegeixo la variable SexRM amb la funció relevel
capacitat$SexR <- relevel( capacitat$Sex, ref='F')
```

També afegeixo la variable **SportR** tenint en compte que la categoria de referència de la variable **Sport** és 'N' (no practiquen esport). Recordo que el significat dels possibles valors de Sport: 'S' (practica sport algunes vegades), 'R' (regularment) i 'E' (cada dia).

```
# afegeixo la variable SportR amb la funció relevel
capacitat$SportR <- relevel( capacitat$Sport, ref='N')
```

I torno a generar el model de regressió lineal multiple però amb les noves variables afegides **SexR** i **SportR**.

```
# aplico el model lineal amb la funció lm
model2 <- lm(data = capacitat, formula = PC ~ Weight+Cig+Years+SexR+SportR)
summary(model2)
```

```
##
## Call:
## lm(formula = PC ~ Weight + Cig + Years + SexR + SportR, data = capacitat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73530 -0.10739 -0.00663  0.10509  0.48296
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.561531   0.236800  15.040  < 2e-16 ***
## Weight      -0.000601   0.003599  -0.167 0.867511
## Cig         -0.034035   0.001293 -26.315  < 2e-16 ***
## Years       -0.022633   0.001065 -21.249  < 2e-16 ***
## SexRM        0.102336   0.027373   3.739 0.000223 ***
## SportRE      0.565597   0.032939  17.171  < 2e-16 ***
## SportRR      0.370267   0.031566  11.730  < 2e-16 ***
## SportRS      0.199592   0.025917   7.701 2.11e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1827 on 292 degrees of freedom
## Multiple R-squared:  0.9172, Adjusted R-squared:  0.9152
## F-statistic: 462.2 on 7 and 292 DF,  p-value: < 2.2e-16
```

Per tant, el model de regressió lineal múltiple té la forma següent:

$$PC = \beta_0 + \beta_1 Weight + \beta_2 Cig + \beta_3 Years + \beta_4 SexRM + \beta_5 SportRE + \beta_6 SportRR + \beta_7 SportRS + e$$

$$PC = 3.561531 - 0.000601 Weight - 0.034035 Cig - 0.022633 Years + 0.102336 SexRM + 0.565597 SportRE + 0.370267 SportRR + 0.199592 SportRS$$

Avalueu la bondat de l'ajust a través del coeficient de determinació (R^2) i compareu el resultat d'aquest model amb l'obtingut en l'apartat 2.1. Podeu usar la instrucció d'R lm i usar el coeficient R-quadrat ajustat en la comparació. Interpreteu també el significat dels coeficients obtinguts i la seva significació estadística.

La bondat d'ajust és prou bona, $R^2 = 0.9172$, ja que és més proper a 1 que a 0 i indica que el model de regressió lineal explica el 91.72% de la variància de les observacions.

Per comparar el resultat d'aquest model amb el de l'apartat 2.1, utilitzo el coeficient de determinació ajustat.

Sempre que s'afegeixen noves variables explicatives a un model, el valor de R^2 augmentarà, encara que aquestes noves variables no aportin res de nou al model. Per això mateix, el valor ajustat de R^2 inclou una penalització pel nombre de regressors que el model conté.

El model 1 té $AjustedR^2 = 0.8101$ i el model 2 té $AjustedR^2 = 0.9152$, per tant, el model 2 explica millor la variància de les observacions.

Pel que fa a la interpretació del significat dels coeficients obtinguts, em fixo en el p-valor de cada una de les variables que es mostra en l'apartat *Coefficients*.

Observo que tots els p-values són significatius amb un nivell de significació del 0.01% ('** ' *0.001), excepte per a la variable* Weight*.

Altre cop amb el contrast pretenc determinar si els efectes de la constant i de les variables independents són realment importants per a explicar la variable dependent o bé els efectes es poden considerar nuls.

Considero els contrasts d'hipòtesis següents:

$H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

$H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

$H_0 : \beta_4 = 0$ vs $H_1 : \beta_4 \neq 0$

$H_0 : \beta_5 = 0$ vs $H_1 : \beta_5 \neq 0$

$H_0 : \beta_6 = 0$ vs $H_1 : \beta_6 \neq 0$

$H_0 : \beta_7 = 0$ vs $H_1 : \beta_7 \neq 0$

Com el p-valor de la variable *Weight* (0.867511) és major que 0.05, ens indica que cal acceptar la hipotesi nul.la, per tant, β_1 sempre tindrá valors iguals a 0. En canvi, com el p-valor de la resta de variables (<2e-16) són menors que 0.05, ens indica que cal rebutjar la hipotesi nul.la, per tant, podem suposar que els valors β_0, β_2, β_3, β_4, β_5, β_6 i β_7 són significativament diferents de 0.

2.3 Efectueu una predicció de la capacitat pulmonar amb els dos models

A partir de l'observació següent: home de Lleida, 30 anys d'edat, fa esport regularment, pes 68 Kg i d'alçada 175cm, i fuma des de fa 15 anys 10 cigarrets al dia.

Realitzeu la predicció de la seva capacitat pulmonar (PC) amb els dos models. Interpreteu els resultats.

Un aspecte important a l'hora d'aplicar el model de regressió obtingut és el risc de l'extrapolació.

Per tant, cal revisar si els valors de les variables, per les quals volem estimar el valor de la variable *PC*, es troben dins el conjunt de valors que hem fet servir pera construir el model. Si no és així, és possible que no tingui sentit l'extrapolació que volem fer. Abans de fer servir el model de regressió, cal preguntar-se per allò que estem fent.

He revisat que tots els valors de les variables es troben dins el rang de valors que hem fet servir per a construir el model. Per tant, l'extrapolació té sentit.

Les variables tenen els valors següents:

Weight=68, Cig=10, Years=15, SexR='M' (SexRM=1), SportR='R' (SportRE=0, SportRR=1, SportRS=0)

```
# inicialitzo les variables
Weight<- 68
Cig<-10
Years<-15
SexR<-'M'
SexRM<-1
SportR<-'R'
SportRE<-0
SportRR<-1
SportRS<-0
```

Per calcular manualment el valor estimat de **PC**, substitueixo els valors de les variables en cada un dels models.

També calculo el valor de la predicció amb la funció *predict*. Amb el paràmetre $se.fit = TRUE$ retorna més components com:

- fit: valor predit
- se.fit: error estandard de la predicció de les mitjanes
- residual.scale: desviació estandar dels residus
- df: graus de llibertat dels residus

El model 1 és el següent:

$PC = 3.677888 + 0.001283Weight - 0.032711Cig - 0.023139Years$

```
# calculo manualment el valor de PC amb el model1
PCmodel1_manual <- 3.677888 + (0.001283 * Weight) - (0.032711 * Cig) - (0.023139 * Years)
PCmodel1_manual
```

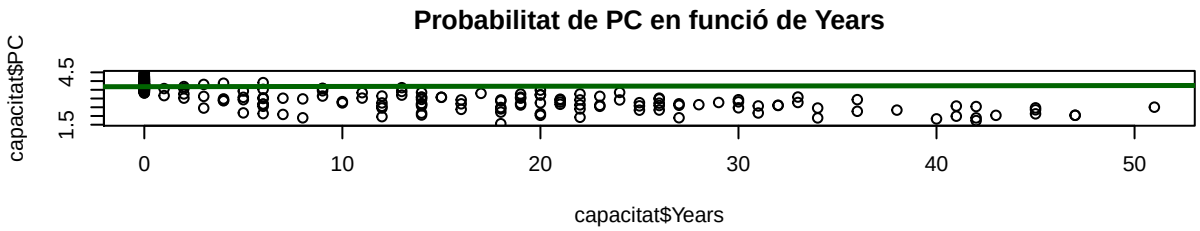
```
## [1] 3.090937
```

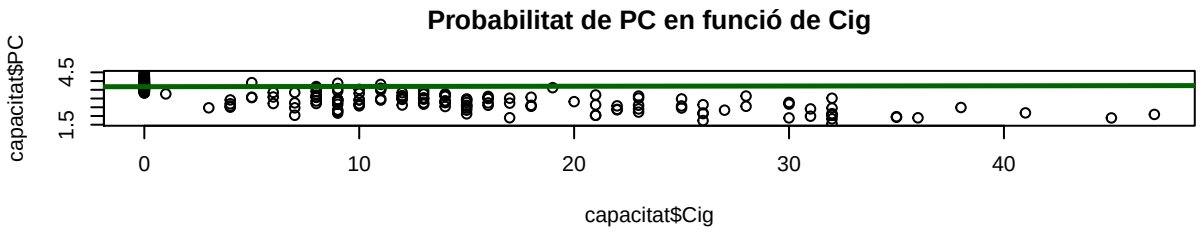
```
# calculo la predicció del valor de PC amb el model1
PCmodel1_predict <- predict (model1, newdata = data.frame(Weight, Cig, Years), se.fit = TRUE)
PCmodel1_predict
```

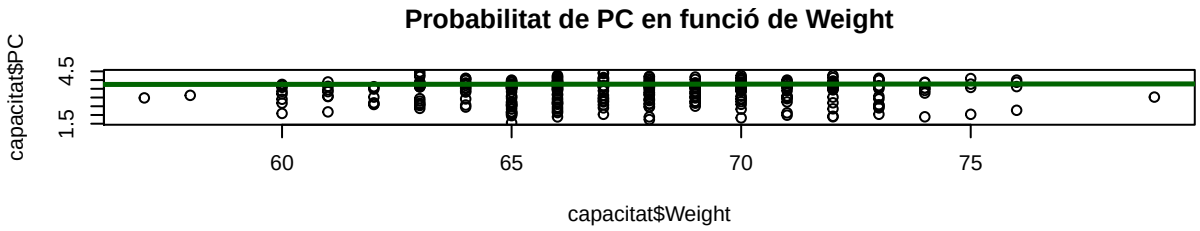
```
## $fit
##        1
## 3.090908
##
## $se.fit
## [1] 0.01787646
##
## $df
## [1] 296
##
## $residual.scale
## [1] 0.2734254
```

Represento gràficament el diagrama de dispersió i la recta de regressió del model1.

```
# Represento gràficament el diagrama de dispersió
par(mfrow=c(3,1))
plot(capacitat$Years, capacitat$PC, main="Probabilitat de PC en funció de Years")
abline(model1, lwd = 2.5, col="darkgreen")
plot(capacitat$Cig, capacitat$PC, main="Probabilitat de PC en funció de Cig")
abline(model1, lwd = 2.5, col="darkgreen")
plot(capacitat$Weight, capacitat$PC, main="Probabilitat de PC en funció de Weight")
abline(model1, lwd = 2.5, col="darkgreen")
```







El model 2 és el següent: $PC = 3.561531 - 0.000601Weight - 0.034035Cig - 0.022633Years + 0.102336SexRM + 0.565597SportRE + 0.370267SportRR + 0.199592SportRS$

```
# calculo manualment el valor de PC amb el model2
PCmodel2_manual <- 3.561531 - (0.000601 * Weight) - (0.034035 * Cig) - (0.022633 * Years) + (0
PCmodel2_manual
```

```
## [1] 3.313421
```

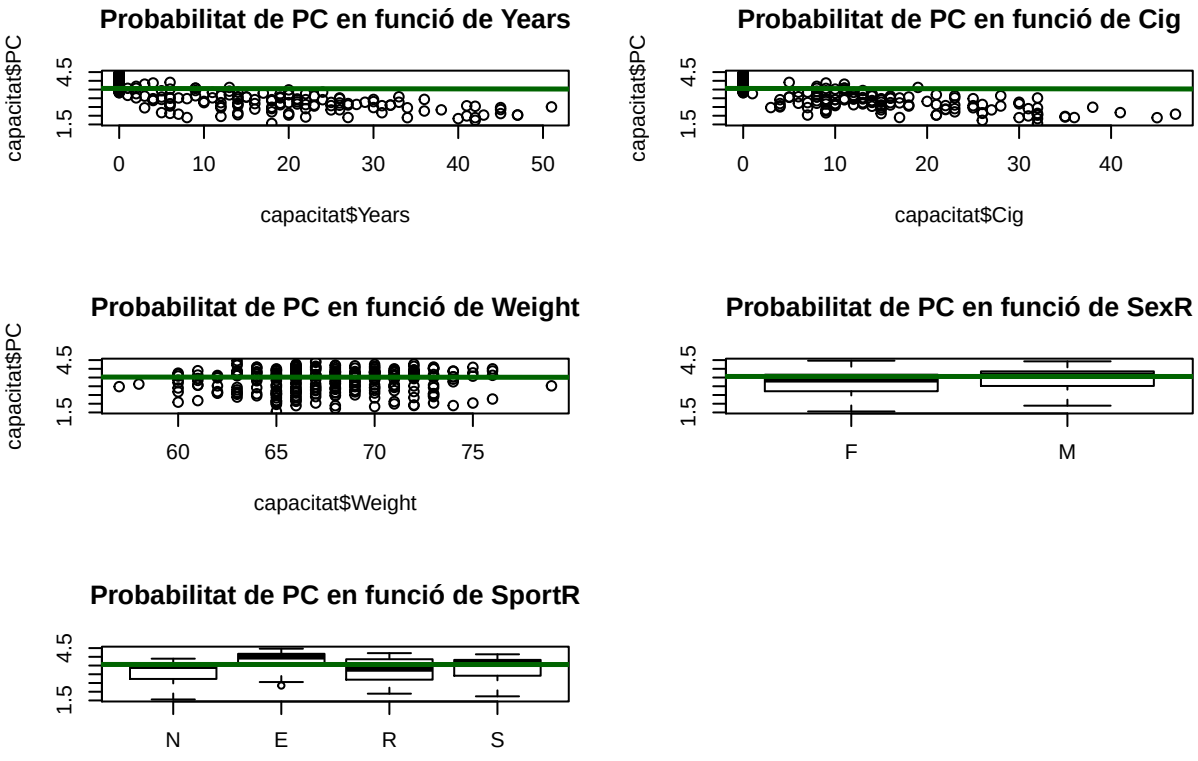
```
# calculo la predicció del valor de PC amb el model2
PCmodel2_predict <- predict (model2, newdata = data.frame(Weight, Cig, Years, SexR, SportR), se
PCmodel2_predict
```

```
## $fit
##        1
## 3.313429
##
## $se.fit
## [1] 0.02906375
##
## $df
## [1] 292
##
```

```
## $residual.scale
## [1] 0.1826956
```

Represento gràficament el diagrama de dispersió i la recta de regressió del model2.

```
# Represento gràficament el diagrama de dispersió
par(mfrow=c(3,2))
plot(capacitat$Years, capacitat$PC, main="Probabilitat de PC en funció de Years")
abline(model2, lwd = 2.5, col="darkgreen")
plot(capacitat$Cig, capacitat$PC, main="Probabilitat de PC en funció de Cig")
abline(model2, lwd = 2.5, col="darkgreen")
plot(capacitat$Weight, capacitat$PC, main="Probabilitat de PC en funció de Weight")
abline(model2, lwd = 2.5, col="darkgreen")
plot(capacitat$SexR, capacitat$PC, main="Probabilitat de PC en funció de SexR")
abline(model2, lwd = 2.5, col="darkgreen")
plot(capacitat$SportR, capacitat$PC, main="Probabilitat de PC en funció de SportR")
abline(model2, lwd = 2.5, col="darkgreen")
```



En cada un dels models observo que, tant calculada manualment com mitjançant la funció *predict*, els valors de capacitat pulmonar són pràcticament iguals:

- En el model 1, la capacitat pulmonar calculada manualment és 3.090937 i amb la funció *predict* és igual a 3.090908.
- En el model 2, la capacitat pulmonar calculada manualment és 3.313421 i amb la funció *predict* és igual a 3.313429.

Considero que el valor predit amb el model 2 és més correcte perquè té més bondat d'ajust que el model 1. A més, la desviació estamdard dels residus és menor en el model 2 ($0.1826956 < 0.2734254$).

3 Model de regressió logística

Es desitja avaluar la qualitat predictiva de la capacitat pulmonar així com de les altres variables presents en l'estudi en relació a la predicció de ser fumador. És a dir, s'avaluarà la probabilitat de que un individu sigui fumador.

Per tal d'avaluar aquesta probabilitat s'aplicarà un model de regressió logística, on la variable dependent serà una variable binària que indicarà si l'individu és fumador. S'utilitzarà la mostra disponible per a estimar el model.

3.1 Estimació d'un model de regressió logística

El primer pas serà crear una variable binària (smoker) que indiqui la condició de fumador (smoker = 1) o no fumador (smoker = 0).

Creo la variable binària **smoker** i la inicialitzo a 0.

Comprovo que hi ha 131 persones que tenen la condició de fumador, ja que la variable $Years > 0$ i la variable $Cig > 0$.

Actualitzo el valor de la variable *smoker* a 1 en les persones amb condició de fumador i comprovo que s'han actualitzat els registres correctament: 131 registres amb valor 1 i 169 registres amb valor 0.

```
# creo variable binaria smoker i l'inicialitzo a 0
capacitat$smoker <- as.numeric(0)
# Comprovo que hi ha 131 persones fumandores (Years>0 i Cig>0)
length( which(capacitat$Years> 0 & capacitat$Cig>0) )
```

```
## [1] 131
```

```
# En aquestes 131 persones, actualitzo el valor de smoker=1
capacitat$smoker[capacitat$Years> 0 & capacitat$Cig>0] <- 1
# Comprovo que smoker conté 131 registes amb valor 1 i 169 registres amb valor 0
table(capacitat$smoker)
```

```
##
##   0   1
## 169 131
```

A continuació, estimeu el model de regressió logística on la variable dependent és "smoker" i les variables explicatives són la capacitat pulmonar (PC), Weight i SexR.

Amb la funció *glm* genero el model de regressió logística. El paràmetre *family="binomial"* indica que la funció de lligadura o link utilitzada és del tipus *logit* o regressió logística.

```
# Genero el model logistic amb la funció glm
model_logistic1 <- glm(data = capacitat, formula = smoker ~ PC+Weight+SexR, family = "binomial"
summary(model_logistic1)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR, family = "binomial",
##     data = capacitat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.47787  -0.31210  -0.04843   0.04271   3.13797
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 38.51385    7.64106   5.040 4.65e-07 ***
## PC          -9.50804    1.36436  -6.969 3.20e-12 ***
## Weight      -0.09876    0.08155  -1.211    0.226
## SexRM        0.76825    0.64182   1.197    0.231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.06  on 299  degrees of freedom
## Residual deviance: 107.87  on 296  degrees of freedom
## AIC: 115.87
##
## Number of Fisher Scoring iterations: 8
```

El model de regressió logística que obtinc és el següent:

$smoker = \beta_0 + \beta_1 PC + \beta_2 Weight + \beta_3 SexRM$

$smoker = 38.51385 - 9.50804 PC - 0.09876 Weight + 0.76825 SexRM$

Avalueu si algun dels regressors (variables explicatives) té influència significativa (valor p del contrast individual inferior al 5%).

Pel que fa a la interpretació del significat dels coeficients obtinguts, em fixo en el p-valor de cada una de les variables que es mostra en l'apartat *Coefficients*.

Observo que només dos dels p-values són significatius amb un nivell de significació del 0.01% ('***' 0.001), el valor estimat de la constant i el de la variable *PC*. Les altres dues variables *Weight* i *SexRM* no són significatives.

Amb el contrast pretenc determinar si els efectes de la constant i de les variables independents són realment importants per a explicar la variable dependent o bé els efectes es poden considerar nuls.

Considero els contrasts d'hipòtesis següents:

$H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$

$H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$

$H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$

Com el p-valor de la variable *Weight* (0.226) i de la variable *SexRM* (0.231) són major que 0.05, ens indica que cal acceptar la hipotesi nul.la, per tant, β_2 i β_3 sempre tindran valors iguals a 0.

En canvi, com el p-valor de la resta de variables (<2e-16) són menors que 0.05, ens indica que cal rebutjar la hipotesi nul.la, per tant, podem suposar que els valors β_0 i β_1 són significativament diferents de 0.

Analitzant els resultats, es pot dir que un individu amb capacitat pulmonar reduïda té major probabilitat de ser fumador?

No, al contrari. El valor estimat de la variable *PC* és negatiu, això fa que quant més alt sigui el valor de la capacitat pulmonar, menor serà diferència amb la constant β_0, per tant, més propera a zero. On smoker=0 vol dir no fumador.

Es pot dir que ser dona augmenta la probabilitat de ser fumador?

No perquè la variable *SexR* no és significativa en el model de regressió i per tant, cap dels seus valors (home o dona) influeix en la probabilitat de ser fumador.

3.2 Predicció en el model lineal generalitzat (model de regressió logística)

Usant el model anterior, calculeu la probabilitat de ser fumador per a l'observació de l'exercici anterior, sense usar la variable fumador, és clar. Podeu assumir que la seva capacitat pulmonar és 3.75 l/s. Recordeu que la resta de variables són: home de Lleida, de 30 anys, fa esport regularment, pesa 68kg i fa 175cm d'alçada.

```
# inicialitzo les variables
Weight<- 68
PC<-3.75
SexRM <-1
```

Per calcular manualment el valor estimat de **PC**, substitueixo els valors de les variables en cada un dels models.

També calculo el valor de la predicció amb la funció *predict*.

El model logístic 1 és el següent:

$smoker_1 = 38.51385 - 9.50804PC - 0.09876Weight + 0.76825SexRM$

```
# calculo manualment el valor de PC amb el model1
mlogit1_manual <- 38.51385 - (9.50804  * PC) - (0.09876 * Weight) + (0.76825 * SexRM)
mlogit1_manual
```

```
## [1] -3.08873
```

```
mlogit1_manual <- 38.51385 - (9.50804  * PC)
mlogit1_manual
```

```
## [1] 2.8587
```

```
# calculo la predicció del valor de PC amb el model1
mlogit1_predict <- predict (model_logistic1, newdata = data.frame(PC, Weight, SexRM), se.fit =
mlogit1_predict
```

```
## $fit
##         1
## -3.089031
##
## $se.fit
## [1] 0.474584
##
## $residual.scale
## [1] 1
```

```
mlogit1_predict <- predict (model_logistic1, newdata = data.frame(PC), se.fit = TRUE)
mlogit1_predict
```

```
## $fit
##         1
## -3.089031
##
```

```
## $se.fit
## [1] 0.474584
##
## $residual.scale
## [1] 1
```

```
mlogit1_predict <- predict (model_logistic1, newdata = data.frame(1), se.fit = TRUE)
mlogit1_predict
```

```
## $fit
##         1
## -3.089031
##
## $se.fit
## [1] 0.474584
##
## $residual.scale
## [1] 1
```

3.3 Millora del model

Cerqueu un model millor que l'anterior afegint al model anterior més variables explicatives. Realitzeu les proves següents:

-Model regressor que afegeix a l'anterior la variable edat (Age).

```
# Genero el model logistic amb la funció glm
model_logistic1 <- glm(data = capacitat, formula = smoker ~ PC+Weight+SexR+Age, family = "bino
summary(model_logistic1)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Weight + SexR + Age, family = "binomial",
##     data = capacitat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.05290  -0.06361  -0.00431   0.01063   2.84865
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  68.66375   14.34683   4.786 1.70e-06 ***
## PC          -14.39062    2.65981  -5.410 6.29e-08 ***
## Weight       -0.11440    0.10560  -1.083    0.279
## SexRM         1.24369    0.89143   1.395    0.163
## Age          -0.29822    0.06635  -4.494 6.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
## Residual deviance:  56.679  on 295  degrees of freedom
## AIC: 66.679
##
## Number of Fisher Scoring iterations: 9
```

```
# Genero el model logistic amb la funció glm
model_logistic2 <- glm(data = capacitat, formula = smoker ~ PC+Age, family = "binomial")
summary(model_logistic2)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Age, family = "binomial", data = capacitat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -2.19193  -0.07646  -0.00502   0.01434   2.85722
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  58.7067    10.6174   5.529 3.22e-08 ***
## PC          -13.7157     2.4549  -5.587 2.31e-08 ***
## Age          -0.2853     0.0628  -4.543 5.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
## Residual deviance:  58.802  on 297  degrees of freedom
## AIC: 64.802
##
## Number of Fisher Scoring iterations: 9
```

El model logístic que obinc al afegir la variable *Age* és el següent:

$smoker_1 = 38.51385 - 9.50804PC - 0.09876Weight + 0.76825SexRM$

-Model regressor que afegeix la variable SportR.

```
# Genero el model logistic amb la funció glm
model_logistic2 <- glm(data = capacitat, formula = smoker ~ PC+SportR, family = "binomial")
summary(model_logistic2)
```

```
##
## Call:
## glm(formula = smoker ~ PC + SportR, family = "binomial", data = capacitat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7620  -0.0965  -0.0115   0.0013   3.2515
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   59.954     11.757   5.099 3.41e-07 ***
## PC           -18.258      3.506  -5.208 1.91e-07 ***
## SportRE        9.772      2.091   4.674 2.95e-06 ***
## SportRR        7.009      1.618   4.332 1.48e-05 ***
## SportRS        3.207      1.265   2.536   0.0112 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.062  on 299  degrees of freedom
```

```
## Residual deviance:  42.989  on 295  degrees of freedom
## AIC: 52.989
##
## Number of Fisher Scoring iterations: 9
```

El mod

-Model regressor que afegeix Age i SportR.

```
# Genero el model logistic amb la funció glm
model_logistic2 <- glm(data = capacitat, formula = smoker ~ PC+Age+SportR, family = "binomial")
summary(model_logistic2)
```

```
##
## Call:
## glm(formula = smoker ~ PC + Age + SportR, family = "binomial",
##     data = capacitat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.48241  -0.05025  -0.00635   0.00299   2.59571
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  67.23180   15.15401   4.437 9.14e-06 ***
## PC          -17.15646    3.75918  -4.564 5.02e-06 ***
## Age          -0.25042    0.08816  -2.840 0.004505 **
## SportRE       7.04133    1.86924   3.767 0.000165 ***
## SportRR       4.54666    1.88536   2.412 0.015884 *
## SportRS       2.46872    1.47546   1.673 0.094290 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 411.06  on 299  degrees of freedom
## Residual deviance:  29.18  on 294  degrees of freedom
## AIC: 41.18
##
## Number of Fisher Scoring iterations: 9
```

Justifiqueu quin seria el millor model entre els quatre models proposats (l'original i els tres proposats en aquest apartat). El criteri per a decidir el millor model és l'AIC (quant més petit és AIC, millor és el model).

aic

A version of Akaike's An Information Criterion, minus twice the maximized log-likelihood plus twice the number of parameters, computed via the aic component of the family. For binomial and Poison

families the dispersion is fixed at one and the number of parameters is the number of coefficients. For gaussian, Gamma and inverse gaussian families the dispersion is estimated from the residual deviance, and the number of parameters is the number of coefficients plus one. For a gaussian family the MLE of the dispersion is used so this is a valid value of AIC, but for Gamma and inverse gaussian families it is not. For families fitted by quasi-likelihood the value is NA.

El darrer model és el que presenta un AIC més baix (41.18)

Realitzeu la interpretació dels valors d'AIC resultants en relació al model escollit.

Nota: Si al realitzar la regressió logística s'obté un missatge tal com: "glm.fit: fitted probabilities numerically 0 or 1 occurred", aquest missatge ens adverteix d'una convergència lenta en el procés iteratiu per a trobar les estimacions. Podeu prescindir del missatge.

3.4 Qualitat de l'ajust

Calcular la matriu de confusió del millor model de la secció anterior, suposant un llindar de discriminació del 70%. Analitzeu els falsos negatius i els falsos positius, i interpreteu què és un fals negatiu i un fals positiu en aquest context.

https://rpubs.com/Joaquin_AR/229736

library(vcd) predicciones <- ifelse(test = modelo_final$fitted.values > 0.5, yes = 1, no = 0) matriz_confusion <- table(modelo_final$model$spam, predicciones, dnn = c("observaciones", "predicciones")) matriz_confusion

mosaic(matriz_confusion, shade = T, colorize = T, gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))

3.5 La selecció dels individus fumadors

Establir un nivell de probabilitat (llindar de discriminació) a partir del qual penseu que l'individu té moltes probabilitats de ser fumador, per exemple, podeu escollir el llindar del 70%. Compareu el nivell de probabilitat que dóna el model amb el valor de la capacitat pulmonar de l'individu (PC). Identifiqueu els individus que no es comporten segons allò que s'espera segons el model, és a dir, que tenen elevada capacitat pulmonar i el model els classifica com a fumadors. Mostreu els valors de PC i la probabilitat de ser fumadors d'aquests individus. Utilizeu com a llindar per a declarar un individu amb elevat PC el tercer quartil de la variable PC.

Podeu realitzar aquest estudi ajudant-vos de gràfics.

3.6 Corba ROC

Realitzeu el dibuix de la corba ROC per a representar la qualitat del model predictiu obtingut. Es pot usar la llibreria pROC i la comanda roc, juntament amb el plot de l'objecte resultant. Calculeu AUROC usant també el mateix paquet amb la funció auc() on li heu de passar com a paràmetre el nom de l'objecte roc.

Interpreteu el resultat.

4 Referències

- Rmarkdown cheat sheet
  https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf
- Rmarkdown: The Definitive Guide
  https://bookdown.org/yihui/rmarkdown/pdf-document.html
- L. Kocbach, LaTeX Math Symbols
  http://web.ift.uib.no/Teori/KURS/WRK/TeX/symALL.html

http://wpd.ugr.es/~bioestad/wp-content/uploads/ContrasteReg.docx

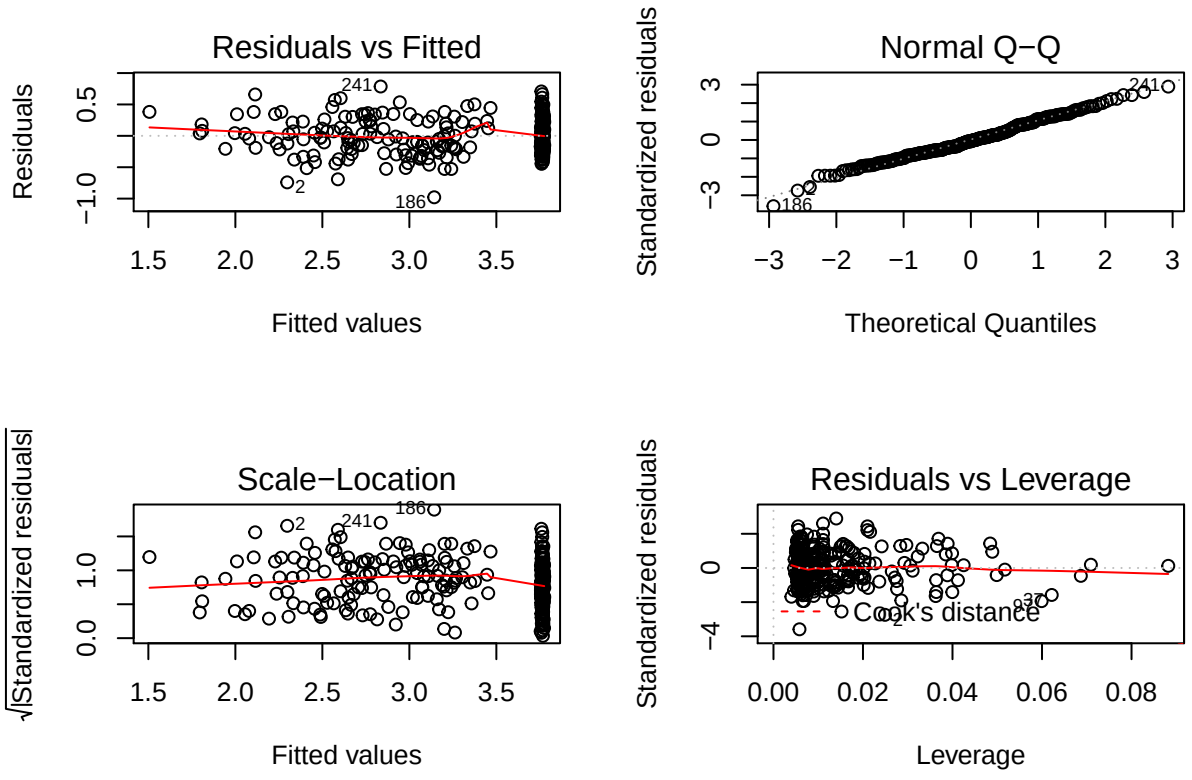
Bon exemple: http://wpd.ugr.es/~bioestad/guia-r-studio/practica-3/

https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.lm.html

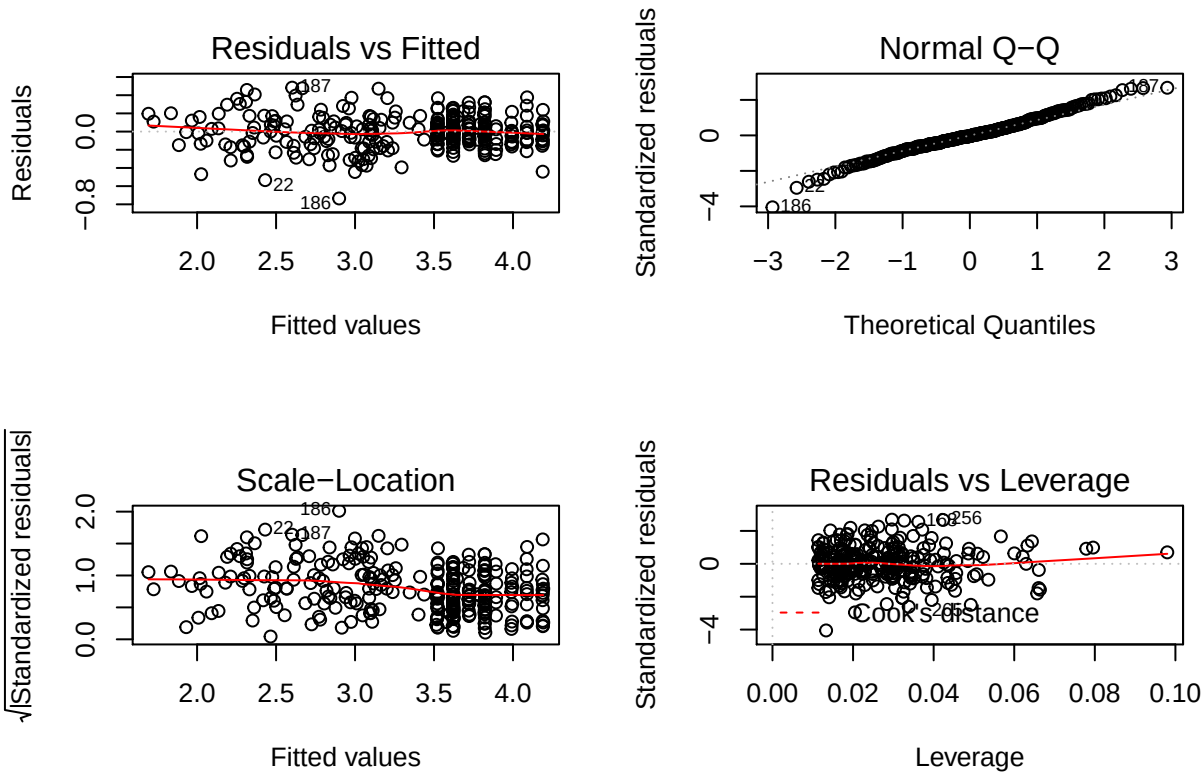
http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-in

https://www.statmethods.net/advstats/glm.html

```
# Genero els gràfics per a la validació del model1
par(mfrow=c(2,2))
plot(model1)
```




```
plot(model2)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage