DELIVERABLE 3 ADEI

Carles Cidraque i Eduard Cidraque

INDEX

3 Deliverable

Target hours.per.week

Modelling using numeric variables (covariates)

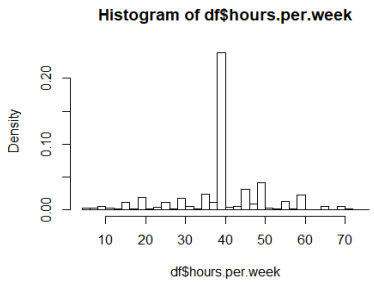
```
library(car)
library(FactoMineR)
library(effects)
library(lmtest)
library(rgl)

## Warning: package 'rgl' was built under R version 3.6.3
```

Comencem la feina de modelització, va adreçada a poder determinar en aquest cas per la variable target numerica hours.per.week quines son les variables amb les que esta relacionada i com s'han de combinar aquestes variables per tal de fer una predicció del target. És a dir del hours.per.week. Aquests models son utils per veure les variables crítiques i veure les combinacions entre elles, pero sobretot per tal de fer prediccions en el futur. Dintre de la estructura de modelatge, primer agafem la varibale target hours.per.week que ja coneixem bé gràcies al analisi descriptiu fet en les anteriors entregues.

```
summary(df$hours.per.week)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   40.00   40.00   39.92   45.00   75.00
```

Mirem la distribució d'aquest target, perque els models tinguin bones propietats, començarem pels models més simples, els de regresió. Aleshores la variable target ha de tenir condicionat a tot el que expliquen les variables explicatives ha de tenir unes característiques de normalitat. Fem un histograma per veure. El següent es el perfil de les hours.per.week i el podem recordar. Recordem que el perfil no es normal.

```
hist(df$hours.per.week,30,freq=FALSE)
```



També sabem de les anteriors entregues les relacions dos a dos entre les variables explicatives numèriques i dels factors i el target. En aquesta fase del modelatge només ens interesa tot el que sigui la relació d'explicatives numeriques amb el target. Primer treballem la relació del target amb les numeriques. Un cop tenim les numeriques que son importants posades dins del

model el que farem es veure si els factors son rellevants, és a dir, afegir-los com a més a més al model. Hi haurà moments on normalment afegirem el factor o el substituirem per el seu anàleg. I després veurem les interaccions entre variables numeriques i factors quan tenen el rol de variables explicatives.

Agafem les variables numeriques [1] “age” “education.num” “capital.gain” “capital.loss”
[5] “hours.per.week” “capital.var” , la variable target es la cinquena

```
res.condes<-condes(df[,vars_con],5)
res.condes$quanti
```

```
##               correlation      p.value
## education.num  0.14865402 1.136054e-25
## capital.gain   0.08685800 1.077558e-09
## capital.var    0.07668594 7.401486e-08
## age            0.07664351 7.525259e-08
## capital.loss   0.05119971 3.309228e-04
```

```
          correlation      p.value
```

education.num 0.14865402 1.136054e-25 capital.gain 0.08685800 1.077558e-09 capital.var 0.07668594 7.401486e-08 age 0.07664351 7.525259e-08 capital.loss 0.05119971 3.309228e-04

Veiem les relacions, veiem que com més education.num te un individu més hores treballa per setmana. Totes elles tenen una relació positiva, és a dir, a mesura que augmenten també augmenten les hours.per.week. La intensitat d’aquestes relacions pero es molt feble, tal com vam veure en la anterior entrega. Sabem que el model que es construeixi llavors serà un model predictiu no gaire bo. Ara ja podem començar, som conscients de que la variable no es normal, aixo vol dir que tindrem problemes quan haguem de fer una validació de les hipotesis del model, i l’altre cosa que sabem es que no ho tindrem facil per explicar la variable target hours.per.week gràcies a l’analisi multivariant fet en la segona entrega. El metode per calcular els models que relacionen la variable target amb un conjunt de variables explicatives es el metode “lm”. Si tenim una correlacio de 0.14865402 entre education.num i la variable target i després amb el capital.gain una de 0.08685800 aleshores un model que tingui les dues variables com a explicatives no mostrarara una correlacio de 0.235512 perque no son independents la variable eduaction.num i la variable capital.gain sino que estan relacionades entre elles. Aleshores la aportació de cadascuna de les variables explicatives numeriques que son les que estem treballant ara amb el target no es una relacio que sigui additiva. No es el mateix que explica una variable per ella mateixa, que quan està “en companyia” d’altres variables en un model. Com que tenim 5 variables numeriques en la bd, podem posar el primer model, que sera un model bastant gran en el sentit de que posarem les variables que sabem que estan relacionades amb el target. Expliquem hours.per.week amb totes les variables explicatives que ens siguin posibles, per tant el data frame que li pasem el restringim a les numeriques que es el que volem.

```
m1<-lm(hours.per.week~.,data=df[,vars_con])
summary(m1)
```

```
##
## Call:
## lm(formula = hours.per.week ~ ., data = df[, vars_con])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.364  -2.430   0.660   4.103  36.152
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.26225    0.76498  42.174  < 2e-16 ***
## age            0.04910    0.01182   4.153 3.34e-05 ***
## education.num  0.54043    0.06125   8.823  < 2e-16 ***
## capital.gain   0.06337    0.01152   5.499 4.01e-08 ***
## capital.loss  -0.06202    0.01152  -5.383 7.67e-08 ***
## capital.var   -0.06311    0.01152  -5.476 4.56e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.89 on 4906 degrees of freedom
## Multiple R-squared:  0.03788,    Adjusted R-squared:  0.0369
## F-statistic: 38.64 on 5 and 4906 DF,  p-value: < 2.2e-16
```

Si utilitzem totes les variables numeriques disponibles per explicar el target la explicavilitat que obtenim es del 3.79%, és a dir, no arribem a explicar ni el 4% de la variabilitat del target, es molt poc. Mirem si tenim redundancies en aquestes variables, perque redundancies voldrà dir variables explicatives que estiguin molt relacionades entre elles, això ho podem veure en la segona entrega, i es fatal per els models. Mirem primer quines son significatives: ens fixem en la columna Pr(>|t|), es un test d’hipòtesi, aleshores per cadascun de les variables explicatives que estan involucrades en el model veiem que es rebutja la

hipotesi nula ja que el p-value en totes elles es molt proper a 0, per sota del llindar del 5% que normalment ens marquem, per tant, totes aquestes variables son rellevants, sobretot la de education.num amb un Pr(>|t|) de < 2e-16. Fem un estudi dels efectes nets. Responem a la pregunta de si un cop tenim totes les variables menys una en el model afegir aquesta una aporta alguna millora al model? Fem per exemple una prova:

```
bro1<-lm(hours.per.week~age+capital.gain+capital.loss+capital.var,data=df[,vars_con])
```

Aquest model te totes les variables numeriques menys la education.num, i en el m1 les te totes. Podem fer una h0 on els dos models son equivalents, si els dos models son equivalents el que vol dir es que un cop tenim les 4 variables que hem posat en el bro1 aleshores afegir-li la edication.num no aportaria res de nou. Això es un tests d'efectes nets. Per fer els tests el que fem es en cas de models encaixats (nested), el model bro1 esta dins de m1, aleshores el que fem es:

```
anova(bro1,m1)

## Analysis of Variance Table
##
## Model 1: hours.per.week ~ age + capital.gain + capital.loss + capital.var
## Model 2: hours.per.week ~ age + education.num + capital.gain + capital.loss +
##     capital.var
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4907 590572
## 2   4906 581347  1    9225.3 77.852 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test de fisher que es la eina bàsica de modelització i de tractament inferencial dels models.Per veure si son o no son equivalents. el model bro1 te 4 variables, el m1 en te 5, un cop tenim afegides les 4 variables en el bro1 val la pena afegir el education.num? el test ens diu, h0 son equivalents, la h0 es rebutja ja que per < 2.2e-16, per tant, no son equivalents, el m1 es millor que bro1. Per tant education.num aporta un valor afegit.

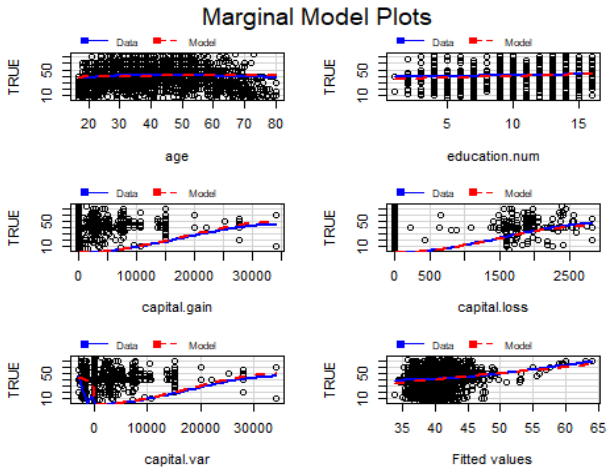
Per no haver de fer aquest procediment per cadascuna de les variables, la llibreria car ens proporciona:

```
Anova(m1)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##               Sum Sq   Df F value    Pr(>F)
## age             2044    1  17.246 3.340e-05 ***
## education.num   9225    1  77.852 < 2.2e-16 ***
## capital.gain    3583    1  30.241 4.009e-08 ***
## capital.loss    3434    1  28.976 7.668e-08 ***
## capital.var     3554    1  29.988 4.563e-08 ***
## Residuals     581347 4906
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

tests d'efectes nets per a totes les variables implicades Veiem que totes aporten al model. Aquesta conclusió seria una conclusió valida si totes les variables fòssin independents entre si, pero en aquest punt ja sabem que per exemple el capital.gain i el capital.loss no son independents del capital.var. Per tant, aixo vol dir que aquest model, com que te variables explicatives que estan relacionades i precisament aquestes, com veure estan molte correlacionades i fan gairabé la mateixa feina.

Una de les coses que hem de mirar per tant, es si les variables estan o no correlacionades, i per veure la colinealitat tenim "vif": El vif es el variance inflation factor i el que ens diu es si el coeficient es un coeficient elevat vol dir que hi ha variables que estan molt correlacionades amb la variable que te aquest coeficient tant elevat. Veiem que capital.gain, capital.loss i capital.var son valors molt elevats i indiquen gran colinealitat. Han d'estar al voltant de 1, a la que tenim valors molt grans vol dir que tenim colinealitat. A la que trobem colinealitat les variables les hem de treure del model.

```
vif(m1)

##           age education.num  capital.gain  capital.loss   capital.var
##      1.016395      1.041218  38042.890288    956.849305  39623.032376

marginalModelPlots(m1)
```



Sembla que capital.var es la que tindriem que obviar. Pero com que el capital.var es un resum del capital.gain i el capital.loss doncs podem treure les altres dues. Això ho fem gracies també a que coneixem aquest arxiu i les seves dades. Per tant actualitzem m1:

```
m1<-lm(hours.per.week~age+education.num+capital.var,data=df[,vars_con])
summary(m1)

##
## Call:
## lm(formula = hours.per.week ~ age + education.num + capital.var,
##     data = df[, vars_con])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.965  -2.472   0.541   4.112  36.331
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.170e+01  7.629e-01  41.554  < 2e-16 ***
## age           5.464e-02  1.183e-02   4.617 3.99e-06 ***
## education.num 5.953e-01  6.080e-02   9.790  < 2e-16 ***
## capital.var   2.185e-04  5.884e-05   3.713 0.000207 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.93 on 4908 degrees of freedom
## Multiple R-squared:  0.0297, Adjusted R-squared:  0.0291
## F-statistic: 50.07 on 3 and 4908 DF,  p-value: < 2.2e-16
```

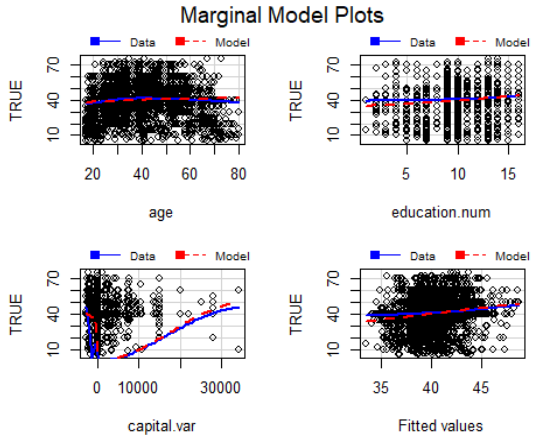
Totes les variables son explicatives i els coeficients han canviat, hem canviat el model cap a millor. Ara veiem els efectes nets, i tots seràn significatius:

```
Anova(m1)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##               Sum Sq   Df F value    Pr(>F)
## age             2547    1  21.319 3.988e-06 ***
## education.num  11450    1  95.853 < 2.2e-16 ***
## capital.var     1647    1  13.784 0.0002074 ***
## Residuals     586293 4908
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El model que tenim ara es un model que hours.per.week = 31.7+(5.464e-02*age)+(5.953e-01*education.num)+(2.185e-04*capital.var), te una explicavilitat del R-squared: 0.0297, 2.97%.

El següent punt que ens hem de fixar es que aquest model només conté relacions lineals entre les variables numèriques i el target pero la relacio entre aquestes variables es realment lineal? Llavors la eina que tenim per d'alguna manera fer aquesta diagnosi es la marginalModelPlots de la llibreria car.

```
marginalModelPlots(m1)
```



Obtenint aquests plots podem veure que hi ha una relació marginal entre les dades i el model, aleshores si mirem el blau que son les dades i el vermell que son el que diu el model podem detectar si hi ha discrepancies entre el que diuen les dades i el que diu el model, si trobem que hi ha discrepancies això suggereix que la relació lineal no està vent agafada, la relació que hi ha entre la variable expplicativa i el target es no lineal, alguna transformació hem de fer. Veiem que hi ha un petit desajust entre el blau i el vermell en els extrems, això ens fa pensar que aquest model logicament te molta variabilitat encara per explicar i es un model que d'entrada requerirà algun tipus de transformació per veure si podem millorar les propietats de explicavilitat en relació a les variables explicatives numèriques i al target.

Intentem introduir alguna transformació, la transformació més senzilla es la transformació basada en la regresió polinòmica. La regresió polinòmica, estem dient que en el model 2 volem modelar les hours.per.week com la age numerica lineal, la education.num amb els termes lineals i quadràtics i amb el capital.var lineal. Aleshores quan utilitzem la comanda poly amb 2 volem dir que la modelització de la eductaion.num com a variable explicativa inclourà els termes lineals i quadràtics. Aquesta eina no ens permet interpretar directament els pesos o coeficients en el model.

Include polynomial regressors

```
m2<-lm(hours.per.week~age+poly(education.num,2)+capital.var,data=df[,vars_con])
m2llegible<-lm(hours.per.week~age+education.num+I(education.num^2)+capital.var,data=df[,vars_con])
summary(m2)
```

```
##
## Call:
## lm(formula = hours.per.week ~ age + poly(education.num, 2) +
##     capital.var, data = df[, vars_con])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.720  -2.633   0.763   3.603  35.363
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             3.790e+01  4.874e-01  77.766  < 2e-16 ***
## age                     4.945e-02  1.200e-02   4.122 3.83e-05 ***
## poly(education.num, 2)1 1.084e+02  1.102e+01   9.834  < 2e-16 ***
## poly(education.num, 2)2 2.864e+01  1.111e+01   2.578 0.009969 **
## capital.var             2.097e-04  5.891e-05   3.559 0.000375 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.92 on 4907 degrees of freedom
## Multiple R-squared:  0.03101,    Adjusted R-squared:  0.03022
## F-statistic: 39.26 on 4 and 4907 DF,  p-value: < 2.2e-16
```

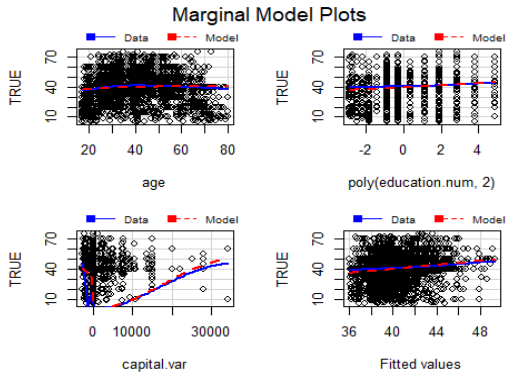
Veiem que el pvalue de poly education.num, el terme quadràtic es ortogonal al primer que es el lineal, si fem la h0 de coeficient = 0 aleshores te un pvalue de 0.009969, aleshores la probabilitat de que sigui 0 es força petita, aleshores si que li fa falta aquesta

transformació. Aquest coeficients no son directament interpretables, si el que volem fer es la equació hem de mirar el que hem construit com a m2llegible.

```
summary(m2llegible)

##
## Call:
## lm(formula = hours.per.week ~ age + education.num + I(education.num^2) +
##     capital.var, data = df[, vars_con])
```

El que veiem que es idèntic, pero el que veiem ara es que els coeficients son diferents, ara si que podem interpretar el model i dir que hours.per.week = 35.13+4.945e-02*age+(-1.351e-01)*education.num+3.821e-02*(education.num^2)+2.097e-04*capital.var Amb el m2llegible el marginalModelPlots el R no enten que education.num i el quadrat de education.num siguin la mateixa variable, i llavors les va tractant per separat i això no es el que volem. El que ens interesa es veure conjuntament si hem guanyat alguna cosa, per això fem sobre m2.

```
marginalModelPlots(m2)
```



Tenim moltes informacions i el model es força dolent, no hem millorat gran cosa. Una mica si, ja que te una explicavilitat del R-squared: 0.03101, 3.1% respecte el 2.97% a m1. Com que el model m1 i m2 son encaixats podem provar el test de fisher.

Nested models - Fisher test

```
anova(m1,m2)
## Model 1: hours.per.week ~ age + education.num + capital.var
## Model 2: hours.per.week ~ age + poly(education.num, 2) + capital.var
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1   4908 586293
## 2   4907 585500  1    792.95 6.6456 0.009969 **
```

H0 es rebutja, m2 es una mica millor que m1 La probabilitat de que siguin equivalents es molt petita.

intenem proposar alguna millora més afegint alguna transformació incloent les variables explicatives numeriques amb els seus termes quadràtics.

```
m3<-lm(hours.per.week~age+poly(education.num,2)+poly(capital.var,2),data=df[,vars_con])
summary(m3)

##
## Call:
## lm(formula = hours.per.week ~ age + poly(education.num, 2) +
##     poly(capital.var, 2), data = df[, vars_con])
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              38.0198     0.4888  77.779  < 2e-16 ***
## age                       0.0491     0.0120   4.091 4.37e-05 ***
## poly(education.num, 2)1 108.4415    11.0215   9.839  < 2e-16 ***
## poly(education.num, 2)2  28.8469    11.1113   2.596 0.009455 **
## poly(capital.var, 2)1    39.4314    11.0755   3.560 0.000374 ***
## poly(capital.var, 2)2   -11.2099    10.9288  -1.026 0.305071
```

```
## Residual standard error: 10.92 on 4906 degrees of freedom
## Multiple R-squared:  0.03122,    Adjusted R-squared:  0.03023
## F-statistic: 31.62 on 5 and 4906 DF,  p-value: < 2.2e-16
```

```
Anova(m3)
```

```
## Anova Table (Type II tests)
##
## Response: hours.per.week
##                         Sum Sq   Df F value    Pr(>F)
## age                       1997    1 16.7351 4.367e-05 ***
## poly(education.num, 2)   12264    2 51.3931 < 2.2e-16 ***
## poly(capital.var, 2)      1637    2  6.8609  0.001058 **
```

```
vif(m3)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## age                    1.040213  1        1.019908
## poly(education.num, 2) 1.053163  2        1.013034
## poly(capital.var, 2)   1.029085  2        1.007193
```

Veiem que afegint el el terme quadràtic de capital.var no millorem ni un 0.1% del model. Inclòs posant el terme cúbic en education.num tampoc millora res, al voltant d'un 0.1%. Veiem que el terme quadràtic te un valor de Pr(>|t|) massa elevat, 0.305071.

Intentem fer una altre millora al nostre model posant el terme quadràtic de la variable age

```
m4<-lm(hours.per.week~poly(age,2)+poly(education.num,2)+capital.var,data=df[,vars_con])
summary(m4)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + poly(education.num,
##     2) + capital.var, data = df[, vars_con])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -39.853  -3.943  -0.550   5.310  47.280
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.983e+01  1.482e-01 268.710  < 2e-16 ***
## poly(age, 2)1            4.815e+01  1.040e+01   4.629 3.77e-06 ***
## poly(age, 2)2           -2.770e+02  1.034e+01 -26.789  < 2e-16 ***
## poly(education.num, 2)1  6.495e+01  1.042e+01   6.232 4.99e-10 ***
## poly(education.num, 2)2  3.125e+01  1.038e+01   3.011  0.00262 **
## capital.var              1.705e-04  5.504e-05   3.097  0.00196 **
## Residual standard error: 10.2 on 4906 degrees of freedom
## Multiple R-squared:  0.1547, Adjusted R-squared:  0.1538
## F-statistic: 179.5 on 5 and 4906 DF,  p-value: < 2.2e-16
```

```
anova(m4,m2) #H0 es rebutja, m4 es millor que m2.
```

```
## Analysis of Variance Table
##
## Model 1: hours.per.week ~ poly(age, 2) + poly(education.num, 2) + capital.var
## Model 2: hours.per.week ~ age + poly(education.num, 2) + capital.var
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4906 510785
## 2   4907 585500 -1    -74715 717.63 < 2.2e-16 ***
```

La probabilitat de que siguin equivalents es molt petita. Veiem que la explicavilitat millora notablement amb un total d'un 15.47%

Proposar canvia a factor de les variables numèriques presents actualment en el model: Veiem que hi ha variables com la variable capital.gain o el capital.loss, recordem que hi havia molts pocs individus que tinguèssin capital.gain o capital.loss, per tant, en el nostre model ens interesarà que capital.var posiblement ens interesi treballar-la més com a variable factor. Podem considerar el

model m1a que diem que tenim la age, education.num i considerem el capital.var com a factor (f.cvar). El que estem es mirant si ens surt més a compte utilitzar una variable explicativa factor per el concepte de capital.var.

```
m1a<-lm(hours.per.week~poly(age,2)+poly(education.num,2)+f.cvar,data=df)
summary(m1a)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + poly(education.num,
##     2) + f.cvar, data = df)
##
..
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               40.9130     0.6461  63.322  < 2e-16 ***
## poly(age, 2)1             47.5405    10.4344   4.556 5.34e-06 ***
## poly(age, 2)2           -277.0386    10.3435 -26.784  < 2e-16 ***
## poly(education.num, 2)1   64.8276    10.4421   6.208 5.80e-10 ***
## poly(education.num, 2)2   31.0476    10.3919   2.988  0.00283 **
## f.cvarf.cvar.equal        -1.1626     0.6660  -1.746  0.08091 .
## f.cvarf.cvar.gain          0.2100     0.8312   0.253  0.80054
…
## Residual standard error: 10.21 on 4905 degrees of freedom
## Multiple R-squared:  0.1545, Adjusted R-squared:  0.1534
## F-statistic: 149.3 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Veiem que te una explicavilitat del 0.1545, la empitjora una miqueta de res respecte m4, a part tant amb el summary com amb l'Anova podem veure per els seus Pr(>|t|), que son superiors al de m4 de capital.var. No ens el quedem com a factor.

Probem amb factor de age

```
m4a<-lm(hours.per.week~f.age+poly(education.num,2)+capital.var,data=df)
summary(m4a)

##
## Call:
## lm(formula = hours.per.week ~ f.age + poly(education.num, 2) +
##     capital.var, data = df)
…
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              3.693e+01  2.739e-01 134.829  < 2e-16 ***
## f.agef.age-(30,40]       5.438e+00  4.033e-01  13.484  < 2e-16 ***
## f.agef.age-(40,50]       5.081e+00  4.289e-01  11.846  < 2e-16 ***
## f.agef.age-(50,90]       1.582e+00  4.446e-01   3.559 0.000376 ***
## poly(education.num, 2)1 8.821e+01  1.087e+01   8.113 6.17e-16 ***
## poly(education.num, 2)2 2.796e+01  1.084e+01   2.580 0.009904 **
## capital.var              1.935e-04  5.763e-05   3.357 0.000793 ***
…
## Residual standard error: 10.68 on 4905 degrees of freedom
## Multiple R-squared:  0.07351,    Adjusted R-squared:  0.07238
## F-statistic: 64.86 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Veiem que la exlicavilitat baixa dràsticament respecte el millor model que teniem fins ara, el model m4. m4a obtenim un Multiple R-squared: 0.07351. Veiem que no val la pena pasar a factor la variable age. Ara amb f.education.num:

```
m4b<-lm(hours.per.week~poly(age,2)+f.education.num+capital.var,data=df)
summary(m4b)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)
…
## Coefficients:…
##
## (Intercept)                          ***
## poly(age, 2)1                        ***
## poly(age, 2)2                        ***
```

```
## f.education.numf.education.num(9)     **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                           **

## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Veiem que millora molt poc la explicavilitat, de un 0.1547 a un 0.1563. Veiem que hi ha una categoria especifica, concretament la de f.education.numf.education.num(10-12) que te un Pr(>|t|) de 0.23269, per tant, ens està dient que aquesta concreta categoria del f.education.num doncs no valdría la pena. De totes maneres si fem el test d'efectes nets veiem que si val la pena en el seu conjunt, ja que f.education.num te un Pr(>F) de 3.303e-12. Ens quedem per tant amb el f.education.num.

```
Anova(m4b)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##                 Sum Sq   Df  F value    Pr(>F)
## poly(age, 2)     73789    2 354.9685 < 2.2e-16 ***
## f.education.num   5905    3  18.9367 3.303e-12 ***
## capital.var       1035    1   9.9555  0.001613 **

m5<-lm(hours.per.week~poly(age,2)+f.education.num+capital.var,data=df)
summary(m5)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)

## Coefficients:
…

## (Intercept)                           ***
## poly(age, 2)1                         ***
## poly(age, 2)2                         ***
## f.education.numf.education.num(9)     **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                           **
## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Per comparar aquests models no podem comparar amb els test de Fisher ja que no son nested. Podem utilitzar el BIC o el AIC, per tant, comparem els BIC's de m5, m4a, m4b=m5 i m1a i amb els anteriors.

BIC / AIC : Minimum BIC is preferred

```
BIC(m5,m1a)

##     df      BIC
## m5   8 36810.96
## m1a  8 36821.43
```

Covariate use preferred Busquem el model que te el BIC o el AIC que te un valor més petit, perque es el que ens està donant millor explicavilitat i minima complicació del model. El m5 te un BIC menor.

Continuem provant BIC's: Ho mirem de manera gràfica:

```
BIC(m5,m2)

##    df      BIC
## m5  8 36810.96
## m2  6 37473.87
```

```
BIC(m5,m4)

##    df      BIC
## m5  8 36810.96
## m4  7 36811.79

BIC(m5,m4a)

##     df      BIC
## m5   8 36810.96
## m4a  8 37270.56
```

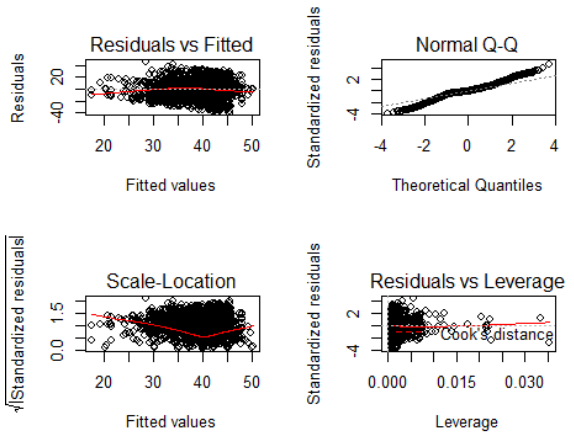
Ens quedem amb m5, perque te una major explicavilitat que totes les altres estudiades, amb una explicabilitat del 0.1563 veiem també que el BIC es el menor de tots els estudiats. El m5 reanomenem mBest.

```
mBest<-lm(hours.per.week~poly(age,2)+f.education.num+capital.var,data=df)
summary(mBest)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)
##
## Coefficients:
…


## (Intercept)                            ***
## poly(age, 2)1                          ***
## poly(age, 2)2                          ***
## f.education.numf.education.num(9)      **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                            **
## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Residual Analysis

Abans de considerar sistemàticament la addició de variables explicatives factor pasem a fer una diagnosi per veure com tenim els residus en el nostre model.

```
par(mfrow=c(2,2))
plot(mBest,id.n=0)
```




```
par(mfrow=c(1,1))
```

Mirem el gràfic residual vs fitted i en aquí hem de mirar els residus del model, es a dir, el que no queda explicat per la part sistemàtica te un soroll i que no té cap patró específic. En aquest cas es una mica dificil de veure pero la linia vermella ens ajuda a veure que si que hi ha una certa tendència descendent, és a dir, els residus son més positius quan les prediccions de les hours.per.week son petites que quan son grans. Això el que ens està dient es que els residus encara contenen informació, tenen una certa estructura, tenen un patró, no estàn centrats en el 0, no son soroll blanc. Cosa normal perque encara tenim un aprox 85% per explicar.

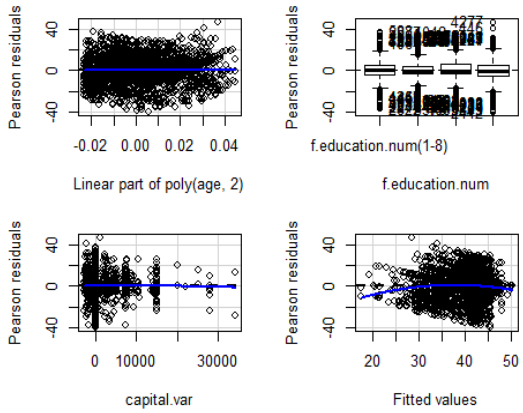
Si ens fixem en el segon gràfic en el Normal Q-Q, es el gràfic dels residus, son les observacions del target menys les prediccions del target, es a dir, el que no està explicat per el model. Aquests residus han de tenir una distribució normal. Veiem que la recta que tenim en aquest gràfic es la recta sobre la que haurien d'estar la teoria(que es la normal estandard) vs la realitat dels residus, veiem clarament que hi ha doscrepancies a les cues. Veiem que per tant els residus no son normals, per tant no estem en les millors condicions per aplicar minims quadrats perque els estimadors no son suficients, si fèssim prediccions així no tindriem suficient fiabilitat.

El Scale-Location, ens fixem principalment amb la linia vermella, si la linia vermella es plana vol dir que la variabilitat dels residus es constant al llarg de tot el rang de valors de la predicció i això es bo. En aquest cas no es així, hi ha una més variabilitat en els extrems que no pas en els valors centrals.

Finalment, l'últim gràfic es el de les observacions influents, posem en ordenades els residus normalitzats, i en el eix de les X el factor d'anclatge, una mesura de quant lluny està la observació en qüestió del centre de gravetat de les variables explicatives. Podem veure que hi ha un parell d'observacions que son unes observacions que es troben molt lluny de les altres. El fet de que estiguin lluny vol dir que tenen valors de les variables explicatives que son bastant peculiars.

Els residual plots son eienes de diagnosi de residus una mica més refinades residuals vs each explanatory variable

```
residualPlots(mBest)
```

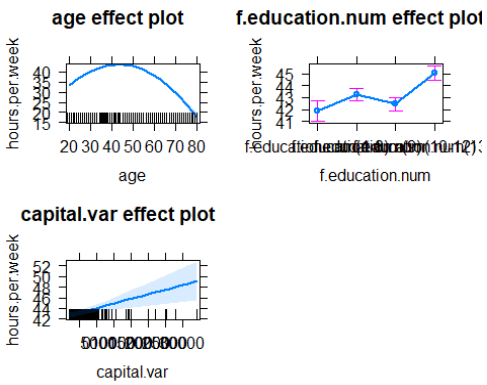



```
##                 Test stat Pr(>|Test stat|)
## poly(age, 2)
## f.education.num
## capital.var       -0.3943           0.6934
## Tukey test        -6.3775         1.8e-10 ***
```

use order 2 polynomial for age El que veiem molt clarament es que tots estàn molt centrats, vol dir que les decisions preses per construir el model han sigut força acertades, per exemple la de posar el teme quadràtic de la variable age. La relació entre els residus i la variable age la linia blava ens diu que ja està capturada, per f.education.num tampoc veiem cap patró i cap desviació, i per capital.var igual.

Si volem entendre una mica el perfil fem servir la llibreria plot effects on hi veurem dibuixat l'efecte marginal de cadascuna de les variables en el model.

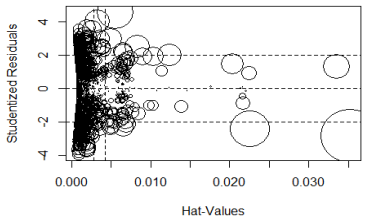
```
library(effects)
plot(allEffects(mBest))
```



El que ens diu basicament es que com més education.num més hour.per.week, com més capital.var també treballaràs més. Les sombres son la precisió que te aquest coeficient per valors molt alts de capital.var en la seva gràfica. Veiem que perdem molta precisió per aquests valors alts de capital.var perque hi ha molt poques observacions com veure a la entrega 2.

El influencePlot es un bubble plot. Podem veure quines observacions tenen un hatvalue determinat i te un residu que també son elevats. Ens indica quins son els residus student, els hat i la distancia de cook, que es el determinant si una observació es influent. En concret veiem que la 17040 es una observació molt influent amb una cookD de 0.042039239.

```
influencePlot(mBest,id=list(method="identify"))  # Click the bubble you would like to identify
```

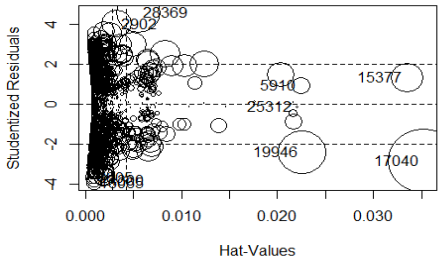



```
summary(mBest)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)
## Coefficients:
…

## (Intercept)                             ***
## poly(age, 2)1                           ***
## poly(age, 2)2                           ***
## f.education.numf.education.num(9)       **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16)   ***
## capital.var                             **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16

#Mirem que li pasa a la observació 17040
df["17040",]
```

Veiem que la variable capital.var es molt elevada, de 34095, per ser tant jove, 20 anys, es una observació influent, el fet de considerar aquesta persona dins de la mostra fa que els coeficients estimats per les variables explicatives canviin força, veiem que pertant a la categoria de f.education.num(10-12), una categoria que tenia un Pr(>|t|) més elevat. Per treure els 5 individus que tenen alguna característica més rellevant fem:

```
influencePlot(mBest,id=list(method="noteworthy",n=5))
```




```
##          StudRes          Hat        CookD
## 2902   4.0364571 0.0032579447 0.0075842322
## 5910   0.9285059 0.0224009450 0.0028222080
## 6605  -3.6808416 0.0007846633 0.0015160405
## 15377  1.3244016 0.0334396436 0.0086677677
## 16065 -3.9661308 0.0009742550 0.0021848907
## 17040 -2.8415074 0.0352137537 0.0420392391
## 19946 -2.4168971 0.0225873215 0.0192653371
## 24300 -3.8378139 0.0009566416 0.0020091894
## 25312 -0.1385842 0.0220160876 0.0000617767
## 28369  4.5927160 0.0055188240 0.0166538898
```

Ens treu els que tenen la distancia de residu més gran o el que tenen el leverage més elevat o la cookD. Els que tenen les bombolles més grans son els que tenen una cookD major. Serien a priori les observacions més influents. Mirem per exemple la observació 2902

```
df["2902",]

#Veiem que es una observació que en totes les seves categories i relacions amb les altres doncs pert
any a una minoria sempre, gent gran amb elevat numero d'estudis, treballant 75 hours.per.week, més d
e 50k, etc.
```

Ara el que farem es ampliar una mica les posibilitats de modelització, ampliarem una mica les variables explicatives que es poden posar en els models.

```
vars_con

## [1] "age"            "education.num"  "capital.gain"   "capital.loss"
## [5] "hours.per.week" "capital.var"

#m1<-lm(hours.per.week~.,data=df[,vars_con])
#mBest es el nostre model que hem fixat amb les nostres variables explicatives numèriques.
summary(mBest)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)
…
## Coefficients:
…
## (Intercept)                            ***
## poly(age, 2)1                          ***
## poly(age, 2)2                          ***
## f.education.numf.education.num(9)      **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                            **

## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Hem d'interpretar uns resultats bàsics de tipus inferencial que tenen a veure amb els contrastos d'hipòtesis que surten en la taula de coeficients del summary. Aleshores per cadascun dels coeficients independenment, el que tenim es el contarast d'hipotesi coeficient = 0.

La valoració del model ja la vam fer mirant el valor de Multiple R-squared, i vam escollir el millor. Es el percentatge de la variabilitat del target que ve explicat pel nostre model de moment, un 15,63%.

El metode step es pot utilitzar per a una motorització segons el criteri de informacio vif, el que fa aquest metode es anar reduint un model gran eliminant les variables que no son rellevants segons aquest criteri, i aquest criteri de quantitat d'informació que el que fa es ponderar el que es la explicavilitat del model amb la seva complexitat, es a dir, es busca el model que sigui el més explicatiu posible pero que sigui el més simple posible.

```
mBest2<-step(mBest,k=log(nrow(df))) # BIC

## Start:  AIC=22862.8
## hours.per.week ~ poly(age, 2) + f.education.num + capital.var
##
##                   Df Sum of Sq    RSS   AIC
## <none>                          509815 22863
## - capital.var      1      1035 510850 22864
## - f.education.num  3      5905 515720 22894
## - poly(age, 2)     2     73789 583605 23510

summary(mBest2)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)
…
## Coefficients:
…
## (Intercept)                            ***
## poly(age, 2)1                          ***
## poly(age, 2)2                          ***
## f.education.numf.education.num(9)      **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                            **
## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Veiem que la explicavilitat es la mateixa. obtenim AIC=22862.8, es el AIC que tindria el model després de suprimir algunes de les variables, pero veiem que no hem suprimit cap de les variables, el metode step s'ha quedat amb el model mBest. Per tant, ens fa veure que ja haviem fet correctament la modelització de mBest fins el moment. El metode step proba una a una a suprimir totes les variables.

```
vif(mBest2)

##                     GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)    1.067986  2        1.016580
## f.education.num 1.075949  3        1.012275
## capital.var     1.026001  1        1.012917

vif(mBest)

##                     GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)    1.067986  2        1.016580
## f.education.num 1.075949  3        1.012275
## capital.var     1.026001  1        1.012917

#(ens surt obviament el mateix ja que es exactament el mateix model)
```

Addings factors

```
summary(mBest)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var, data = df)

…
## Coefficients:
…
## (Intercept)                          ***
## poly(age, 2)1                        ***
## poly(age, 2)2                        ***
## f.education.numf.education.num(9)     **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                          **
## Residual standard error: 10.19 on 4905 degrees of freedom
## Multiple R-squared:  0.1563, Adjusted R-squared:  0.1552
## F-statistic: 151.4 on 6 and 4905 DF,  p-value: < 2.2e-16
```

Agafem el millor model fins el moment, mBest Tenim amb mBest una explicavilitat del target del 15,63%, per intentar millorar el model mirem d'afegir variables que son factors. Fem servir el mateix mètode. La addició de factors es pot analitzar-se desde dos punts de vista, perimer es el d'afegir més informació nova al model, més variables explicatives noves, i segona de les maneres seria anem a veure si el tractament com a factor es un tractament més adequat per alguna de les variables que ja tenim en el model.

Quins factors aporten alguna cosa nova al model? La estretegia que seguim es que estem intentant fer una de les opcions mencionades anteriorment, quins factors aporten alguna cosa al model, com podem millorar el model que ja tenim a mBest afegint-hi factors.

Les variables factors més relacionades amb el target son: No coloquem aquelles variables factor que ja hem estudiat anteriorment, la f.cvar que comporta a la substitució de f.cgain,f.closs. i f.education.num i obviament f.hours.per.week perque la variable numèrica es la nostra target, a més tampoc coloquem les que ja hem estudiat si substituir la variable numèrica pel seu factor tant si ho hem fet com si no.

```
vars_fac<-names(df)[c(13,15:22)]
vars_fac
```

```
## [1] "hours.per.week"      "Y.bin"               "f.type"
## [4] "f.RelType"           "f.CountryType"       "f.EduType"
## [7] "f.MaritalStatusType" "f.OccupationType"    "f.RaceType"
```

```
#Ara tornem a recordar quins son els factors més relacionats amb el target hours.per.week:
res.condes<-condes(df[,vars_fac],num.var=1)
#Agafem directament $quali
res.condes$quali
```

```
##                              R2       p.value
## f.RelType           0.112074731 3.878248e-126
## f.OccupationType    0.100487064 4.879982e-108
## f.MaritalStatusType 0.068995345  9.460265e-76
## Y.bin               0.052895757  5.618927e-60
## f.type              0.050259699  1.406226e-53
## f.EduType           0.034852747  3.502175e-34
## f.RaceType          0.005206387  2.726590e-06
```

```
#Les que més relacionades estan son basicament: f.RelType, f.OccupationType, f.MaritalStatusType
```

Pasem a contruir i a analitzar nous models a partir de l'addició de variables factor. Comencem amb un R2 més elevat, i en el nostre cas es: f.RelType amb R2 = 0.112074731, Aquest model li diem mfprova

```
mfprova<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType, data = df)
summary(mfprova)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType, data = df)

## Coefficients:
```

```
…
## (Intercept)                              ***
## poly(age, 2)1                            ***
## poly(age, 2)2                            ***
## f.education.numf.education.num(9)        **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16)    ***
## capital.var                              **
## f.RelTypef.Reltyp-rel-WifeOther          ***
## f.RelTypef.Reltyp-rel-Child              ***
## f.RelTypef.Reltyp-Rel-NotFamily          ***
## Residual standard error: 9.895 on 4902 degrees of freedom
## Multiple R-squared:  0.2057, Adjusted R-squared:  0.2042
## F-statistic:   141 on 9 and 4902 DF,  p-value: < 2.2e-16
```

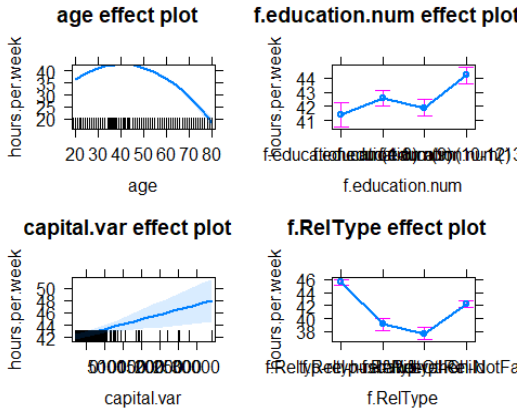
Veiem com s'han afegit les variables artificials per cadascun dels nivells del factor, hem d'anar en compte i no fer-li cas al pvalue que ens proporciona el summary d'aquestes variables, ja que realment formen part d'una de sola. Per deduïr si aquesta variable f.RelType es o no important, fem el métode Anova, un test d'efectes nets:

```
Anova(mfprova)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##                 Sum Sq   Df  F value    Pr(>F)
## poly(age, 2)     46075    2 235.2864 < 2.2e-16 ***
## f.education.num   4469    3  15.2158 7.457e-10 ***
## capital.var        866    1   8.8408   0.00296 **
## f.RelType        29853    3 101.6321 < 2.2e-16 ***
## Residuals       479962 4902
```

En el resultats que ens mostra, podem veure com estan en el model les variables segons la variable nova afegida. Veiem que tenint f.education.num, capital.var i age amb termes quadràtics al nostre model, si li afegim f.RelType veiem que val la pena; tenim un Pr(>F) de < 2.2e-16. Per tant veiem que val la pena introduir la variable factor f.RelType en el nostre model. A més a més, millorem la explicavilitat del model de un 15,63% a un 20,57%.

Per interpretar-ho primer fem una interpretació gràfica amb els efectes marginals:

```
plot(allEffects(mfprova))
```



Veiem que aproximadament a partir dels 40 anys d'edat veiem que les hours.per.week van disminuint. La eduaction.num, com més anys un individu hagi estudiat, més hours.per.week treballa. Del f.RelType effect plot podem observar que ens treu uns intervals de confiança de les hours.per.week per cadascun dels grups que tenim definits segons f.RelType, veiem que els que més hours.per.week treballen son els husband, tenen un promig de hours.per.week treballades superior a 44, els que son wife-Other tenen un promig d'entre 38 i 40 hours.per.week, els child son els que tenen el promig més baix, i finalment els Not-inFamily que tenen un promig aproximat d'entre 42 i 43 hours.per.week aproximadament. L'altre estudi que fem es la interpretació de les equacions.

```
summary(mfprova)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType, data = df)
## Coefficients:
…

## (Intercept)                            ***
## poly(age, 2)1                          ***
## poly(age, 2)2                          ***
## f.education.numf.education.num(9)      **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                            **
## f.RelTypef.Reltyp-rel-WifeOther        ***
## f.RelTypef.Reltyp-rel-Child            ***
## f.RelTypef.Reltyp-Rel-NotFamily        ***
## Residual standard error: 9.895 on 4902 degrees of freedom
## Multiple R-squared:  0.2057, Adjusted R-squared:  0.2042
## F-statistic:   141 on 9 and 4902 DF,  p-value: < 2.2e-16

#Com son models aniuats, mBest i mfprova
anova(mfprova,mBest)

## Analysis of Variance Table
##
## Model 1: hours.per.week ~ poly(age, 2) + f.education.num + capital.var +
##     f.RelType
## Model 2: hours.per.week ~ poly(age, 2) + f.education.num + capital.var
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4902 479962
## 2   4905 509815 -3    -29853 101.63 < 2.2e-16 ***

BIC(mBest)## [1] 36810.96

BIC(mfprova) ## [1] 36540.06
```

Per tant, com veiem que mfprova es millor actualitzem mBest amb aquest nou model:

```
mBest<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType, data = df)
summary(mBest)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType, data = df)

## Coefficients:
…
## (Intercept)                            ***
## poly(age, 2)1                          ***
## poly(age, 2)2                          ***
## f.education.numf.education.num(9)      **
## f.education.numf.education.num(10-12)
## f.education.numf.education.num(13-16) ***
## capital.var                            **
## f.RelTypef.Reltyp-rel-WifeOther        ***
## f.RelTypef.Reltyp-rel-Child            ***
## f.RelTypef.Reltyp-Rel-NotFamily        ***
## Residual standard error: 9.895 on 4902 degrees of freedom
## Multiple R-squared:  0.2057, Adjusted R-squared:  0.2042
## F-statistic:   141 on 9 and 4902 DF,  p-value: < 2.2e-16
```

continuem intentant afegir-hi factors que ens ajudin a millorar el nostre model, els anem afegint un a un: Ara tornem a recordar quins son els factors més relacionats amb el target hours.per.week:

```
res.condes$quali
```

```
##                              R2       p.value
## f.RelType           0.112074731 3.878248e-126
## f.OccupationType    0.100487064 4.879982e-108
## f.MaritalStatusType 0.068995345  9.460265e-76
## Y.bin               0.052895757  5.618927e-60
## f.type              0.050259699  1.406226e-53
## f.EduType           0.034852747  3.502175e-34
## f.RaceType          0.005206387  2.726590e-06
```

Les que més relacionades estan son basicament: f.RelType, f.OccupationType, f.MaritalStatusType, f.type. Ja hem afegit f.RelType on ens sortia a compte i millorava el model. Continuem amb f.OccupationType:

```
mprovaContFact<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.Occupat
ionType, data = df)
summary(mprovaContFact)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType, data = df)
##
…

## Residual standard error: 9.704 on 4895 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.2347
## F-statistic: 95.13 on 16 and 4895 DF,  p-value: < 2.2e-16

#Veiem que afegint-hi f.OccupationType millorem la explicavilitat fins al 23.72%.
vif(mprovaContFact)

##                       GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)     1.645273  2        1.132555
## f.education.num  1.610349  3        1.082646
## capital.var      1.033165  1        1.016447
## f.RelType        1.646348  3        1.086643
## f.OccupationType 1.838508  7        1.044457

#Veiem que no hi ha colinealitats, podem procedir.
anova(mprovaContFact,mBest) #veiem que no son equivalents

## Analysis of Variance Table
##
## Model 1: hours.per.week ~ poly(age, 2) + f.education.num + capital.var +
##     f.RelType + f.OccupationType
## Model 2: hours.per.week ~ poly(age, 2) + f.education.num + capital.var +
##     f.RelType
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4895 460919
## 2   4902 479962 -7    -19044 28.892 < 2.2e-16 ***

BIC(mprovaContFact)## [1] 36400.69

BIC(mBest)## [1] 36540.06

#fem test d'efectes nets, veiem que l'addició del factor f.OccupationType si que es significatiu.
Anova(mprovaContFact)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##                   Sum Sq   Df  F value    Pr(>F)
## poly(age, 2)      36441    2 193.5051 < 2.2e-16 ***
## f.education.num    1301    3   4.6050  0.003195 **
## capital.var         470    1   4.9879  0.025570 *
## f.RelType         20186    3  71.4576 < 2.2e-16 ***
## f.OccupationType  19044    7  28.8920 < 2.2e-16 ***
```

Veiem a més que el BIC es menor amb l'addició de f.OccupationType, per tant actualitzem mBest amb mprovaContFact, a més te una major explicavilitat.

```
mBest<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.OccupationType,
data = df)
```

Continuem amb f.MaritalStatusType:

```
mprovaContFact<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.Occupat
ionType+f.MaritalStatusType, data = df)
summary(mprovaContFact)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType + f.MaritalStatusType,
##     data = df)
…

## Residual standard error: 9.692 on 4892 degrees of freedom
## Multiple R-squared:  0.2394, Adjusted R-squared:  0.2365
## F-statistic: 81.06 on 19 and 4892 DF,  p-value: < 2.2e-16

#Millorem poc, menys d'1% la explicavilitat, pero la millorem fins a un 23,94%.
vif(mprovaContFact) #veiem que el vif de f.RelType i f.MaritalStatusType estàn força elevats. son di
stants a 1.

##                         GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)         2.218714  2        1.220465
## f.education.num      1.639008  3        1.085834
## capital.var          1.034689  1        1.017196
## f.RelType           16.873312  3        1.601524
## f.OccupationType     1.864223  7        1.045493
## f.MaritalStatusType 19.540695  3        1.641182

BIC(mprovaContFact)## [1] 36411.64

BIC(mBest)## [1] 36400.69
```

El BIC es major al afegir f.MaritalStatusType, a més, veiem que la explicavilitat no millora gairabé res. No actualitzem llavors mBest.

Continuem amb f.type:

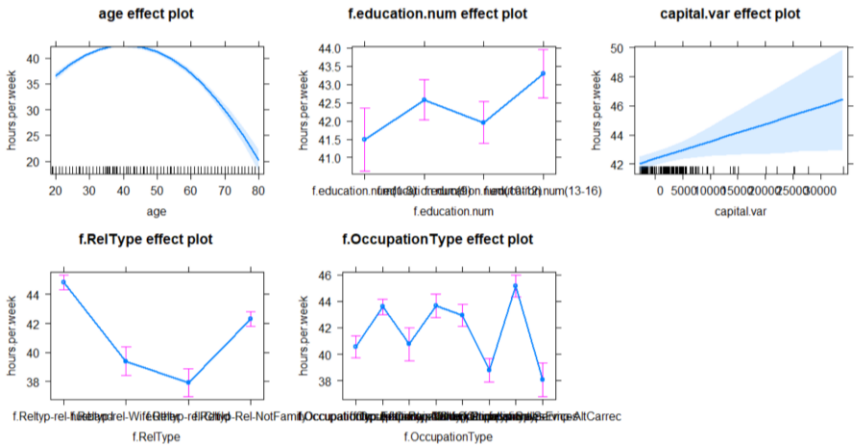
```
mprovaContFact<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.Occupat
ionType+f.type, data = df)
summary(mprovaContFact)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType + f.type, data = df)
## Residual standard error: 9.664 on 4891 degrees of freedom
## Multiple R-squared:  0.2441, Adjusted R-squared:  0.241
## F-statistic: 78.97 on 20 and 4891 DF,  p-value: < 2.2e-16

#Millorem la explicavilitat fins al 24.41%.
vif(mprovaContFact)

##                        GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)       1.694337  2        1.140906
## f.education.num    1.620801  3        1.083814
## capital.var        1.034907  1        1.017304
## f.RelType          1.667518  3        1.088960
## f.OccupationType 506.299217  7        1.560170
## f.type           323.514740  4        2.059381
```

Veiem que tenim valors força elevats per f.OccupationType i per f.type, per tant veiem colinealitat amb aquestes. No podem seguir així i per tant prescindim millor de f.type. Per tant no actualitzem amb f.type.

```
summary(mBest)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType, data = df)
## Residual standard error: 9.704 on 4895 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.2347
## F-statistic: 95.13 on 16 and 4895 DF,  p-value: < 2.2e-16

mprovaContFactStep<-step(mBest,k=log(nrow(df)))


…
## Step:  AIC=22438.5
## hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType
##                     Df Sum of Sq    RSS   AIC
## <none>                           462796 22439
## - f.OccupationType  7     23019 485815 22617
## - f.RelType         3     20910 483706 22630
## - poly(age, 2)      2     38549 501345 22815

summary(mprovaContFactStep)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType,
##     data = df)
…# Residual standard error: 9.719 on 4899 degrees of freedom
## Multiple R-squared:  0.2341, Adjusted R-squared:  0.2322
## F-statistic: 124.8 on 12 and 4899 DF,  p-value: < 2.2e-16
```

tot i que veiem que el métode step redueix el model al treure capital.var i f.education.num que el que fa es ponderar el que es la explicavilitat del model amb la seva complexitat, considerem que poden donar més a alguna explicació.

```
plot(allEffects(mBest))
```



Provem d'afegir totes les altres variables factor i procedir a executar el métode step per veure si les acabem afegint al nostre model o no.

```
mprovaContFact2<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.Occupa
tionType+f.RaceType, data = df)
summary(mprovaContFact2)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType + f.RaceType,
##     data = df)
```

```
Residual standard error: 9.702 on 4893 degrees of freedom
## Multiple R-squared:  0.2377, Adjusted R-squared:  0.2349
## F-statistic: 84.78 on 18 and 4893 DF,  p-value: < 2.2e-16

vif(mprovaContFact2)

##                       GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)     1.650677  2        1.133484
## f.education.num  1.618473  3        1.083555
## capital.var      1.033317  1        1.016522
## f.RelType        1.672976  3        1.089553
## f.OccupationType 1.858780  7        1.045275
## f.RaceType       1.045842  2        1.011269
```

Veiem que el f.EduType tindria colinealitat amb f.education.num en el nostre model i la suprimim. f.RaceType no te colinealitat amb ningú pero ens aporta molt poc al model. Tot i així fem el métode step per veure que ens proposa:

```
mprovaContFactStep2<-step(mprovaContFact2,k=log(nrow(df)))

…## Step:  AIC=22438.5
## hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType
##
##                     Df Sum of Sq    RSS   AIC
## <none>                           462796 22439
## - f.OccupationType  7     23019 485815 22617
## - f.RelType         3     20910 483706 22630
## - poly(age, 2)      2     38549 501345 22815

summary(mprovaContFactStep2)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType,
##     data = df)
## Residual standard error: 9.719 on 4899 degrees of freedom
## Multiple R-squared:  0.2341, Adjusted R-squared:  0.2322
## F-statistic: 124.8 on 12 and 4899 DF,  p-value: < 2.2e-16
```

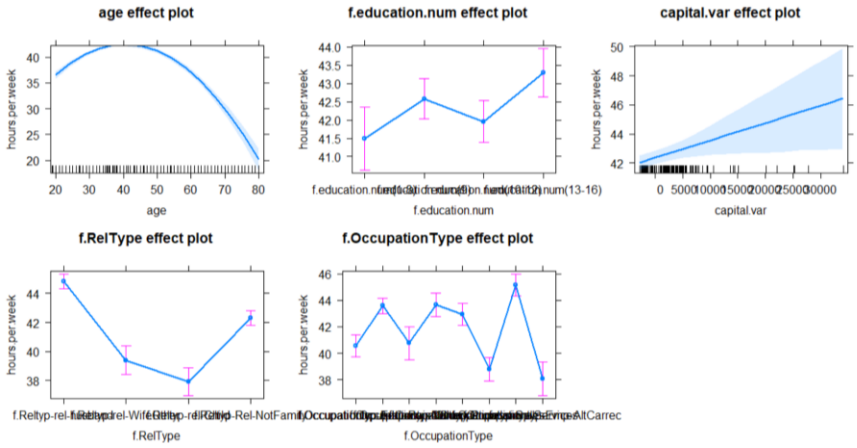
Veiem que l'step precindeix de f.RaceType, fa el procediment com anteriorment haviem analitzat.

Recordem com tenim el model actual mBest:

hours.per.week ~ poly(age, 2) + f.education.num + capital.var + f.RelType + f.OccupationType, data = df)

Pasem a entendre millor aquest model obtingut: De forma gràfica amb el allEffects:

```
plot(allEffects(mBest))
```



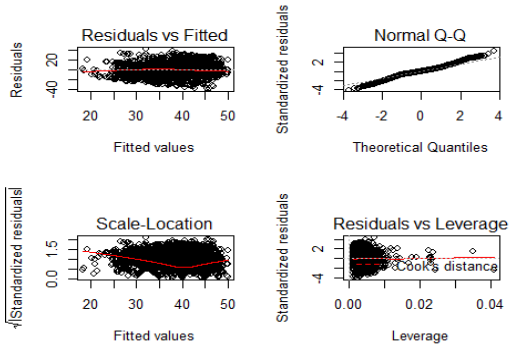
Veiem que amb l'edat, a partir dels més o menys 40 anys, com més edat, es van baixant les hours.per.week treballades. Amb el f.RelType effect plot podem observar que ens treu uns intervals de confiança de les hours.per.week per cadascun dels grups que tenim definits segons f.RelType, veiem que els que més hours.per.week treballen son els husband, tenen un promig de hours.per.week treballades superior a 44, els que son wife-Other tenen un promig d'entre 38 i 40 hours.per.week, els child son els

que tenen el promig més baix, estàn entre 0 i unes 39 hours.per.week i finalment els Not-inFamily que tenen un promig aproximat d'entre 42 i 43 hours.per.week. Els f.type que més hours.per.week treballen son els f.typ-self-emp-inc.

```
summary(mBest)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType, data = df)
…## Residual standard error: 9.704 on 4895 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.2347
## F-statistic: 95.13 on 16 and 4895 DF,  p-value: < 2.2e-16
```

Veiem que els residus sembla que segueixen tenint algun patró, tot i que veiem que hem millorat, sobretot es veu força visible en els gràfics de Residuals vs Fitted i en el de Residuals vs Leverage.Tenen una tendència positiva per valors baixos i una de negativa per valors alts en la gràfica Residuals vs Fitted. Sobre la normalitat dels residus seguim tenint un desajust, sobretot pels valors negatius. Scale-Location seguim veient la linia vermella no recta, per tant seguim tenint problemes. La validació dels residus ens diu que amb aquest model tenim problemes. Això el que ens està dient es que els residus encara contenen informació, tenen una certa estructura, tenen un patró, no estàn centrats en el 0, no son soroll blanc.

```
par(mfrow=c(2,2))
plot(mBest,id.n=0)
```




```
par(mfrow=c(1,1))
```

Residual analysis: Heterokedasticity - Non-normality

Transformations

Interactions

Ara ja hem fet la introducció de variables numèriques i de factors necessaris en el nostre model, a continuació continuem intentant millorar el model tot buscant interaccions entre variables numèriques i factors, la interacció entre els diferents factors.

mBest<-lm(hours.per.week ~ poly(age, 2) + f.education.num + capital.var+f.RelType+f.OccupationType, data = df) Interactions between numeric variables and factors:

```
summary(mBest)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     capital.var + f.RelType + f.OccupationType, data = df)
…## Residual standard error: 9.704 on 4895 degrees of freedom
## Multiple R-squared:  0.2372, Adjusted R-squared:  0.2347
## F-statistic: 95.13 on 16 and 4895 DF,  p-value: < 2.2e-16

mBestI1<-lm(hours.per.week ~ (poly(age, 2) + capital.var)*(f.RelType+f.OccupationType+f.education.nu
m),data=df)
summary(mBestI1)

##
## Call:
## lm(formula = hours.per.week ~ (poly(age, 2) + capital.var) *
##     (f.RelType + f.OccupationType + f.education.num), data = df)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.412  -4.913  -0.296   4.817  37.759
## Coefficients:
##                                                                  Estimate
## (Intercept)                                                     4.015e+01
## poly(age, 2)1                                                  -6.240e+00
## poly(age, 2)2                                                  -2.312e+02
## capital.var                                                     5.918e-04
## f.RelTypef.Reltyp-rel-WifeOther                                -5.114e+00
## f.RelTypef.Reltyp-rel-Child                                    -6.693e+00
…
## Residual standard error: 9.575 on 4856 degrees of freedom
## Multiple R-squared:  0.2633, Adjusted R-squared:  0.2549
## F-statistic: 31.55 on 55 and 4856 DF,  p-value: < 2.2e-16
```

Veiem que ara, tenim una explicavilitat millorada de un 23,72% fins a un Multiple R-squared: 0.2633 (26,33%)

Apliquem el métode step

```
mBestI12<-step(mBestI1,k=log(nrow(df)))  # BIC

## Step:  AIC=22441.63
## hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType +
##     f.education.num + poly(age, 2):f.education.num
##
##                                Df Sum of Sq    RSS   AIC
## <none>                                      455935 22442
## - poly(age, 2):f.education.num  6    5453.9 461389 22449
## - f.OccupationType              7   18595.7 474530 22579
## - f.RelType                     3   19284.8 475219 22620
```

El resultat d'aplicar el mètode step sobre el model mBestI1 es: hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType + f.education.num + poly(age, 2):f.education.num Veiem que precindeix de capital.var i age només interacciona amb f.education.num

```
summary(mBestI12)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType +
##     f.education.num + poly(age, 2):f.education.num, data = df)
…
## Coefficients:
##                                                    Estimate Std. Error
## (Intercept)                                         39.7628     0.6510
## poly(age, 2)1                                      -22.4919    23.0580
## poly(age, 2)2                                     -246.6231    23.2184
## f.RelTypef.Reltyp-rel-WifeOther                     -5.3307     0.5434
## f.RelTypef.Reltyp-rel-Child                         -6.6885     0.5207
…

## Residual standard error: 9.656 on 4890 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2422
## F-statistic: 75.74 on 21 and 4890 DF,  p-value: < 2.2e-16

#Tot i haver simplificat el model amb el mètode step, explicavilitat baixa fins a un Multiple R-squa
red:  0.2454
BIC(mBestI1)## [1] 36561.25

BIC(mBestI12)## [1] 36389.78

#Veiem que el BIC de mBestI12 es considerablement menor al de mBestI1.
anova(mBestI12,mBestI1)   # no equivalents, but mBestI12 best

## Analysis of Variance Table
        ##
        ## Model 1: hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType +
        ##     f.education.num + poly(age, 2):f.education.num
        ## Model 2: hours.per.week ~ (poly(age, 2) + capital.var) * (f.RelType +
```

```
##     f.OccupationType + f.education.num)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4890 455935
## 2   4856 445157 34     10778 3.4579 5.678e-11 ***
```

```
Anova(mBestI12)
```

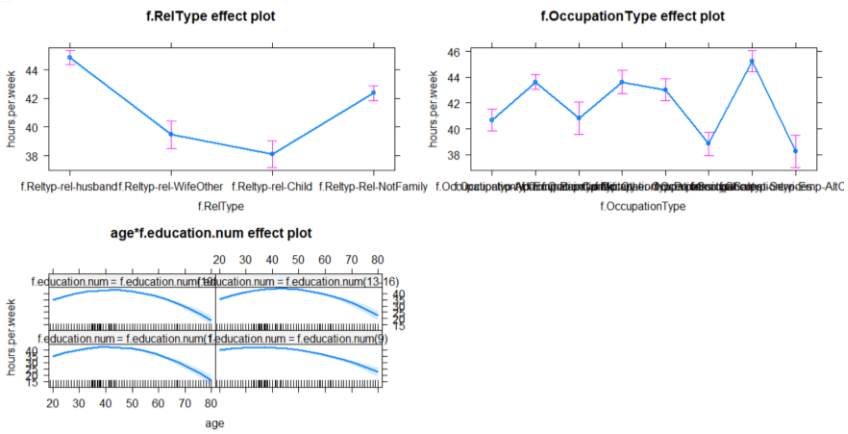
```
## Anova Table (Type II tests)
##
## Response: hours.per.week
##                              Sum Sq   Df  F value    Pr(>F)
## poly(age, 2)                  36396    2 195.1776 < 2.2e-16 ***
## f.RelType                     19285    3  68.9445 < 2.2e-16 ***
## f.OccupationType              18596    7  28.4918 < 2.2e-16 ***
## f.education.num                1407    3   5.0303  0.001759 **
## poly(age, 2):f.education.num   5454    6   9.7490 1.058e-10 ***
## Residuals                    455935 4890
```

```
#Gracies al test d'efectes nets veiem que totes les variables aporten al model.
vif(mBestI12)
```

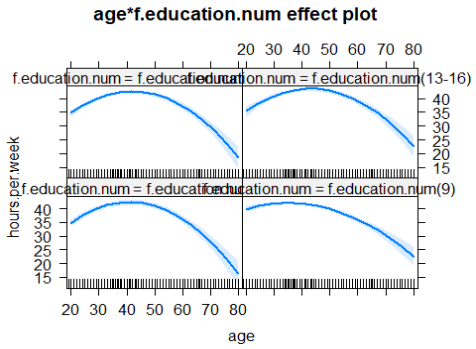
```
##                                  GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)                32.927675  2        2.395467
## f.RelType                    1.671737  3        1.089418
## f.OccupationType             1.873501  7        1.045864
## f.education.num              1.849443  3        1.107916
## poly(age, 2):f.education.num 32.734666  6        1.337367
```

```
#De moment, per tant ens quedem amb mBestI12, ja que un AIC menor.
##hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType + f.education.num + poly(age, 2):f.ed
ucation.num, de moment ens quedem amb aquest model.
```

```
plot(allEffects(mBestI12))
```

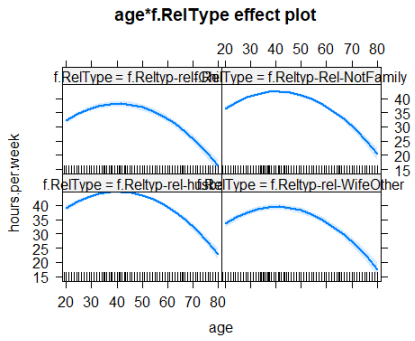



```
plot(effect("poly(age, 2)*f.education.num",mBestI12))
```

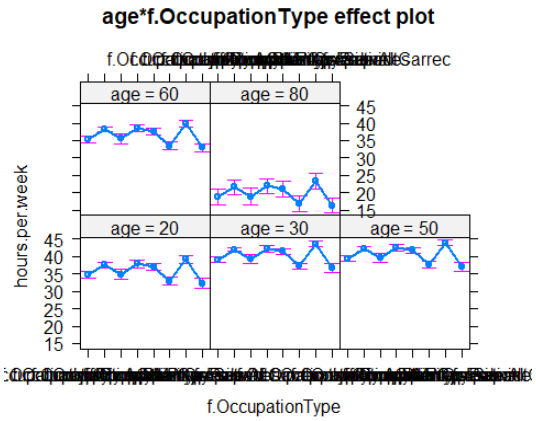


Tot i que no estiguin en el model veiem els gràfics Es veu clarament com les persones que tenen més estudis (f.education.num(13-16)), quan son joves son els que es veu clarament que més treballen, en canvi un cop es vei que superen una edat i ja superen els seus estudis acaben sent els que treballen més hores. Tot el contrari pasa en el grup de f.education.num(9), on comencen a treballar moltes més hores de joves i a mesura que es van fent grans van dismininuint més el hours.per.week respecte el grup anteriorment comentat. veiem com ja hem vist en numerosos casos, que els husband son els que més hores treballen i que segons la edat, va disminuint segons s'augmentem els anys. Veiem també que els que més hores treballen son els self-emp, i a més també els que treballen en general son els que tenen una age d'entre 30 i 50 anys. Aquest gràfics son els que no apareixen en el nostre model optimitzat per el mètode step, tot i així els posem perque trobem interesant les petites conclusions anteriorment esmentades.

```
plot(effect("poly(age, 2)*f.RelType",mBestI12))
```



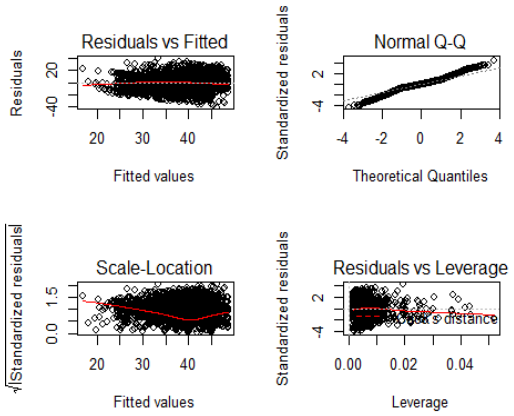

```
plot(effect("poly(age, 2)*f.OccupationType",mBestI12))
```




```
#Actualitzem mBest amb el model mBestI12
mBest<-mBestI12
# Interpretation of mBest realitzant alguna equació.
```

hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType + f.education.num + poly(age, 2):f.education.num, data = df)
equacions: Tot i com ja hem comentat anteriorment en aquest estudi algunes de les categories de f.education.num no tenen un Pr(>|t|) suficientment petit, però vàrem decidir quedarnos amb la variable en el model, a més el mètode step en aquest cas el preserva. Fem una interpretació del nostre model, on un individu tingui caracterítiques de husband, Emp-AltCarrec i education.num(13-16), la més elevada, les hours.per.week d'aquest individu tindria vindria donada per la següent equació:
y=(39.7628+0-2.4049+1.3031)+(-246.6231+26.8566)*age

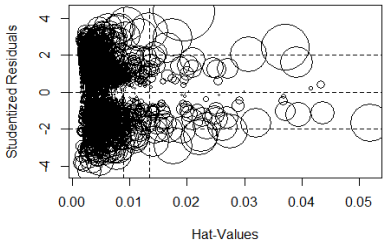
```
par(mfrow=c(2,2))
plot(mBest,id.n=0)
```




```
par(mfrow=c(1,1))
```

Veiem que els residus contenen més informació de la que contenien abans de fer la interacció entre variables numèriques i factors. Veiem que continuem tenint desviacions sobre la linia recta, degut a que hi ha uns residus estanderitzats que son més negatius del que haurien de ser. Continuen tenint un patró, no compleixen amb la normalitat. Veiem que la variança no es constant en el gràfic d'Sacle-Location. Per veure els indiviud més influents fem servir el influence plot.

```
influencePlot(mBest,id=list(method="identify"))
```




```
#Treiem dos de les observacions amb CookD més elevat:
#        StudRes        Hat       CookD
#15377  1.452627 0.03506796 0.004509978
#17040 -2.382310 0.04085416 0.014206424
```

Interactions between factors

Prenem el model mBest: hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType + f.education.num + poly(age, 2):f.education.num, data = df) mBestI1<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType + f.education.num, data = df)

```
mBestIF1<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType, data = df
)
summary(mBestIF1)

##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) * f.education.num +
##     f.RelType + f.OccupationType, data = df)
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                     39.7628     0.6510
## poly(age, 2)1                                  -22.4919    23.0580
## poly(age, 2)2                                 -246.6231    23.2184
## f.education.numf.education.num(9)                0.9440     0.4779
## f.education.numf.education.num(10-12)            0.2792     0.4997
## f.education.numf.education.num(13-16)            1.3031     0.5667
```

```
## f.RelTypef.Reltyp-rel-WifeOther                        -5.3307     0.5434
## f.RelTypef.Reltyp-rel-Child                            -6.6885     0.5207
…

## Residual standard error: 9.656 on 4890 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2422
## F-statistic: 75.74 on 21 and 4890 DF,  p-value: < 2.2e-16

BIC(mBestIF1)## [1] 36389.78
```

El que estem intentant veure es si dos a dos son factibles i de fet si ens aporten alguna cosa bona en l'explicavilitat del model aquestes interaccions entre dos factors.

El següent es model sobreparametritzat, veiem que obtenim major explicavilitat, pero major BIC que a mBestIF1, que es el anteriorment tractat, la interacció de poly(age, 2)*f.education.num es una interacció que ja hem resolt anteriorment que aporta al nostre model. Per tant, ara ens ocupem dels factors f.RelType i f.OccupationType.

```
mBestIF12<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.education.num*(f.RelType + f.Occupat
ionType), data = df)
summary(mBestIF12)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) * f.education.num +
##     f.education.num * (f.RelType + f.OccupationType), data = df)
## …
## Residual standard error: 9.587 on 4860 degrees of freedom
## Multiple R-squared:  0.2608, Adjusted R-squared:  0.253
## F-statistic: 33.62 on 51 and 4860 DF,  p-value: < 2.2e-16

BIC(mBestIF12)## [1] 36543.77

#Fem alguna prova més
mBestIF3<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType*(f.education.num + f.Occupati
onType), data = df)
summary(mBestIF3)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) * f.education.num +
##     f.RelType * (f.education.num + f.OccupationType), data = df)
##
…

## Residual standard error: 9.596 on 4860 degrees of freedom
## Multiple R-squared:  0.2593, Adjusted R-squared:  0.2515
## F-statistic: 33.36 on 51 and 4860 DF,  p-value: < 2.2e-16

BIC(mBestIF3)## [1] 36553.75 #Pitjor explicavilitat i BIC que mBestIF12, ens seguim quedant amb aques
t.
mBestIF4<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.OccupationType*(f.education.num + f.R
elType), data = df)
summary(mBestIF4)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) * f.education.num +
##     f.OccupationType * (f.education.num + f.RelType), data = df)
##
…

## Residual standard error: 9.584 on 4848 degrees of freedom
## Multiple R-squared:  0.263,  Adjusted R-squared:  0.2534
## F-statistic: 27.46 on 63 and 4848 DF,  p-value: < 2.2e-16

BIC(mBestIF4)## [1] 36631.28
```

Tot i ser molt poc millor l'explicavilitat en aquest model mBestIF4 que en el mBestIF12, el BIC es superior en mBestIF4, ens continuem quedant amb mBestIF12 per fer el mètode step. Simplifiquem el model amb el mètode step, ja que el tenim sobreparametritzat i necessitem simplificar-lo.

```
mBestIFstep<-step(mBestIF12,k=log(nrow(df)))  # BIC
```

```
## … Step:  AIC=22441.63
## hours.per.week ~ poly(age, 2) + f.education.num + f.RelType +
##     f.OccupationType + poly(age, 2):f.education.num
##
##                                Df Sum of Sq    RSS   AIC
## <none>                                      455935 22442
## - poly(age, 2):f.education.num  6    5453.9 461389 22449
## - f.OccupationType              7   18595.7 474530 22579
## - f.RelType                     3   19284.8 475219 22620

BIC(mBestIFstep)## [1] 36389.78

summary(mBestIFstep)
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) + f.education.num +
##     f.RelType + f.OccupationType + poly(age, 2):f.education.num,
##     data = df)
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       39.7628     0.6510
## poly(age, 2)1                                    -22.4919    23.0580
## poly(age, 2)2                                   -246.6231    23.2184
## f.education.numf.education.num(9)                  0.9440     0.4779
## f.education.numf.education.num(10-12)              0.2792     0.4997
## f.education.numf.education.num(13-16)              1.3031     0.5667
…
## Residual standard error: 9.656 on 4890 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2422
## F-statistic: 75.74 on 21 and 4890 DF,  p-value: < 2.2e-16
```

Veiem que el mètode step ens diu que ens quedem amb un model com el inicial amb el que hem començat l'apartat de interacció entre factors, és a dir, el model amb el que continuarem serà el següent: hours.per.week ~ poly(age, 2) + f.education.num + f.RelType + f.OccupationType + poly(age, 2):f.education.num Es un model amb un BIC de 36389.78 i una explicavilitat de 24,54%. No hem trobat per tant cap interacció entre factors nova que ens aporti al model, tot tenint en compte la solució proposada pel mètode step.

Comparem amb un test anova, el nostre mBest amb el millor que haviem pogut treure de interaccions entre factors abans d'aplicar-hi el mètode step, aquest ha sigut el model mBestIF12:

```
anova(mBest,mBestIF12) #Veiem que no son equivalents, a més el model mBest es millor segons el mètod
e step.

## Analysis of Variance Table
##
## Model 1: hours.per.week ~ poly(age, 2) + f.RelType + f.OccupationType +
##     f.education.num + poly(age, 2):f.education.num
## Model 2: hours.per.week ~ poly(age, 2) * f.education.num + f.education.num *
##     (f.RelType + f.OccupationType)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   4890 455935
## 2   4860 446656 30    9279.1 3.3655 1.716e-09 ***


Anova(mBest)

## Anova Table (Type II tests)
##
## Response: hours.per.week
##                              Sum Sq   Df  F value    Pr(>F)
## poly(age, 2)                  36396    2 195.1776 < 2.2e-16 ***
## f.RelType                     19285    3  68.9445 < 2.2e-16 ***
## f.OccupationType              18596    7  28.4918 < 2.2e-16 ***
## f.education.num                1407    3   5.0303  0.001759 **
## poly(age, 2):f.education.num   5454    6   9.7490 1.058e-10 ***
## Residuals                    455935 4890
```

Veiem que totes les variables tenen rellevància en el model, la única interacció que es conserva es la de poly(age, 2):f.education.num, alhora d'interpretar el test d'efectes nets sempre hem de tenir en compte el que tingui jerarquia més alta que corrovorarà que el de menor pes també serà necessari (sempre i quan òbviament ens estiguem referint a la mateixa variable).

```r
vif(mBest)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)                   32.927675  2        2.395467
## f.RelType                       1.671737  3        1.089418
## f.OccupationType                1.873501  7        1.045864
## f.education.num                 1.849443  3        1.107916
## poly(age, 2):f.education.num 32.734666  6        1.337367
```

```r
mBest<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType, data = df)
#Els plots effects corresponents a aquest model mBest ja han sigut per tant mostrats anteriorment, i els plots referents a l'anàlisis dels residus també.
```

Transformation of numeric target

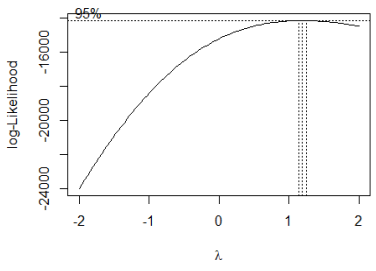
Ara pasem al tema de transformacions, que te a veure amb la transformada boxcox, la transformada boxcox es aplicable a variables de resposta que son numèriques i no negatives. Aleshores, la transformació sobre el target es una transformació que ek que intenta es millorar les propietats del model lineal, és a dir, que els residus siguin més normals, almenys que siguin menys simètrics.

El model fins ara tenim que es: (hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType, data = df)

```r
summary(mBest)
```

```
##
## Call:
## lm(formula = hours.per.week ~ poly(age, 2) * f.education.num +
##     f.RelType + f.OccupationType, data = df)
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       39.7628     0.6510
## poly(age, 2)1                                    -22.4919    23.0580
## poly(age, 2)2                                   -246.6231    23.2184
## f.education.numf.education.num(9)                   0.9440     0.4779
## f.education.numf.education.num(10-12)               0.2792     0.4997
## f.education.numf.education.num(13-16)               1.3031     0.5667
…

## Residual standard error: 9.656 on 4890 degrees of freedom
## Multiple R-squared:  0.2454, Adjusted R-squared:  0.2422
## F-statistic: 75.74 on 21 and 4890 DF,  p-value: < 2.2e-16
```

```r
library(MASS)

boxcox(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType, data = df)
```



Veiem que crea un diagrama on en aquest gràfic veiem que lambda, el que es veu es que va millorant a mesura que la tranformació que se li aplica, la lambda que li sembla recomanable està al voltant de 1,2. Es la millor transformació per assolir unes millors propietats del model. seria y^1.2. Si el model que tenim entre mans es un model que te una explicavilitat alta i veiem que encara el podem millorar doncs aleshores hem de fer tot lo posible, pero introduir en un model tant dolent com aquest una transformació per millorar una mica les propietats pero de fet sabem que no es posible explicar be aquest target aleshores ens resistim a introduir variables explicatives o transformacions que siguin extranyes.

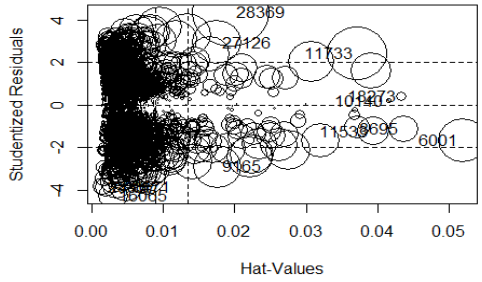
Amb això hem acabat la modelització utilitzant un patró normal del target, que es aplicable a variables amb un significat numèric.

Ara un cop tenim el model definitiu hem de fer la part del model validation, i no podem donar per vàlid cap model que tingui variables que tinguin outliers dels residus o observacions molt influents. Aquestes observacions influents venen per la distancia de cook. Tot i que ja hem estat fent validacions al llarg de la construcció del model final, en fem a continuació una última:

```
influencePlot(mBest,id=list(method="identify"))
```

Ens treu els que tenen la distancia de residu més gran o el que tenen el leverage més elevat o la cookD.

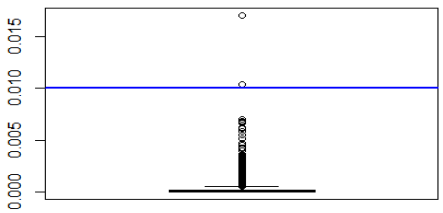
```
influencePlot(mBest,id=list(method="noteworthy",n=5))
```




```
##          StudRes         Hat        CookD
## 1471  -3.8489223 0.004927170 3.324858e-03
## 6001  -1.6567997 0.051870298 6.823597e-03
## 7776  -3.8859011 0.001945316 1.333970e-03
## 8695  -1.1331047 0.043539580 2.656498e-03
## 9165  -2.8805402 0.017555594 6.729536e-03
## 10140  0.1927311 0.041525554 7.316464e-05
## 11533 -1.2490528 0.039314640 2.901765e-03
## 11733  2.4306013 0.037191903 1.036282e-02
## 16065 -4.2793581 0.003077470 2.560544e-03
## 18273  0.3946376 0.043333838 3.207125e-04
## 27126  2.9324768 0.017599409 6.991680e-03
## 28357 -3.8482135 0.001601163 1.076471e-03
## 28369  4.3513923 0.019491412 1.704652e-02
```

Veiem que tenim observacions amb un residu estanderitzat gran, i els hatvalues, el factor d'anclatge hi ha observacions que estan fora del llindar que ens diria quan el factor d'apalancament comença a ser notable, és a dir, les particularitats que tenen les variables explicatives que estan lluny del centre de gravetat del núvol d'observacions. Veiem que hi ha clarament observacions que son problemàtiques. Per posar un exemple veiem com la observació 28369 té una distància de cook de 1.704652e-02, i te un residu StudRes molt elevat, concretament de 4.3513923. Aquestes observacions hem de veure que els hi pasa i veure si son influents. Determinar si una observació es influent en difinitiva es determinar-les utilitzant un boxplot, on les determinem mitjançant les cookD.

```
boxplot(cooks.distance(mBest))
abline(h=0.010,col="blue",lwd=2)
```



sent estrictes marquem el limit de cookD en 0.010, li posem la cota en aquest valor. Les que ens preocupen son les que estan notablement més lluny de les demés. Veim que tenim dues observacions per sobre de la distancia de cook superior a 0.010. Anem a veure aquestes dues observacions: Veiem que son les observacions 11733 que està just al límit de la cookD marcada ja que te una cookD=1.036282e-02, i l'altre es l'observació 28369 amb una cookD=1.704652e-02. Anem a veure les seves característiques:

```
df["11733",]
```

Veiem que es una dona de 76 anys, unmarried, en concret es widowed, que treballa 40 hores setmanals, es f.Occupationtyp-Emp-AltCarrec, amb poc nivell d'estudis f.education.num(1-8), f.Edutyp-Dropout. Sería un individu atípic, pero no tant com el següent.

```
df["28369",]
```

Veiem que la característica a destacar més es que es un home de 73 anys i treballa 75 hours.per.week, un número molt elevat per la seva edat, i per a qualsevol. Està Married-civ-spouse, en concret es husband, workclass Private amb un elevat nivell d'estudis f.education.num(13-16), en concret 13. Sería un individu força atípic.

Veiem que si incloem aquestes dues variables al model aleshores fem variar els coeficients del model. No hem de tolerar que dos observacions canviin les relacions entre les explicatives i el target. Les hem de suprimir.

```
qnorm(0.995)## [1] 2.575829
```

De moment tenim dues observacions a suprimir: la 28369 i la 11733, segons la cookD, ara fem una altre repasada tot utilitzant llindar per els StudRes, on ens surt com a limit el [-2.575829,2.575829] a un interval de confiança del 99%, totes les observacions que es trobin fora d'aquest rang es consideren com a misfit i seràn eliminades.

```
ll<-which(abs(rstudent(mBest))>2.575829);ll
```

Mirem el total de les observacions que tenen un StudRes superior a 2.575829 en valor absolut. Veiem que dins d'aquesta llista estàn les que haviem tret en amb la comanda influencePlot(mBest,id=list(method="noteworthy",n=5)), aquestes seràn les que eliminarem del df, en concret son: 1471, 7776, 9165, 16065, 27126, 28357 i 28369. Aquestes observacions a més, estàn molt per sobre del valor 2.575829, algunes d'elles superan el valor 4, decidim eliminar-les. De moment les observacions que eliminem son: 28369, 11733, 1471, 7776, 9165, 16065, 27126, 28357 i 28369.

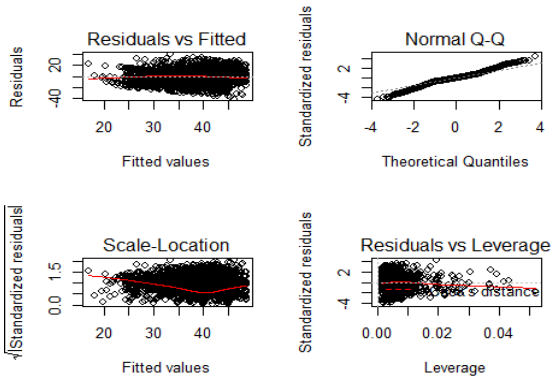
Ara continuem l'analisi de quines variables hauriem d'eliminar tot utilitzant els hatvalues: Els hatvalues son valors que estan entre 0 i 1 i son els que es coneixen com a Factor d'apalancament/anclatge o leverage. Aleshores totes les observacions que tenen el factor d'anclatge molt elevat vol dir que son extranyes i poden ser influents.

```
3*mean(hatvalues(mBest))## [1] 0.01343648

llhii<-which(hatvalues(mBest)>3*mean(hatvalues(mBest)));length(llhii)## [1] 84
```

Potential influent observations are those whose leverage is above three times the mean leverage (or 3p/n); 0.01343648. Veiem que hi ha 84 observacions cumplint aquest criteri, les mes influents del model apareixen en el resultat de la comanda influencePlot(mBest,id=list(method="noteworthy",n=5)), mirem quines d'elles apareixen en la llista dels 84 obtinguts en la llista llhii i els eliminem del nostre df, per tant al final d'aquest analisis ens queda que les observacions potencialment influents i hem d'eliminar son: 28369, 11733, 1471, 7776, 9165, 16065, 27126, 28357, 6001, 8695, 10140, 11533 i 18273. Per tant, acabem havent d'eliminar totes aquelles que ens havia dit la comanda influencePlot(mBest,id=list(method="noteworthy",n=5)) que eren les més influents segons els valors de StudRes, hatvalues i cookD.

Les eliminem:

```
par(mfrow=c(2,2))
plot(mBest,id.n=0)
```



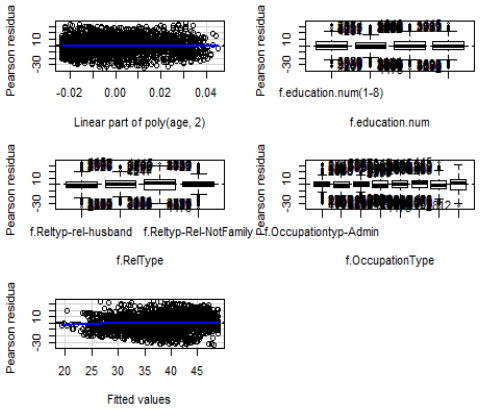

```
par(mfrow=c(1,1))
```

eliminar i actualitzar mBestEliminar creem un nou df de proba, veiem que la observació 6001 es la unica que te un fnlwgt de 84616, borrem les observacions anteriorment mencionades buscant-les pel seu fnlwgt.

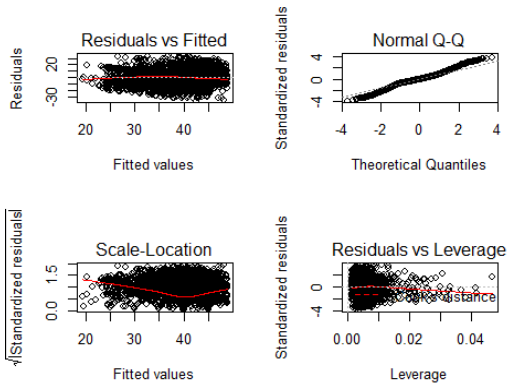
```
ll<-which((df$fnlwgt=="84616")|(df$fnlwgt=="105886")|(df$fnlwgt=="312500")|(df$fnlwgt=="120939")|(df$fnlwgt=="280169")|(df$fnlwgt=="157593")|(df$fnlwgt=="323627")|(df$fnlwgt=="107814")|(df$fnlwgt=="321824")|(df$fnlwgt=="29020")|(df$fnlwgt=="152900")|(df$fnlwgt=="142370")|(df$fnlwgt=="86111")) ;length(ll) ## [1] 13

if( length(ll)>0) dfBestEliminar<-df[-ll,]
mBestEliminar<-lm(hours.per.week ~ poly(age, 2)*f.education.num + f.RelType + f.OccupationType, data = dfBestEliminar)
```

un cop eliminades mirem la situació des residus en del nostre model mBest i mirem si ens diu es que en els residus encara contenen informació, tenen una certa estructura, tenen un patró, no estàn centrats en el 0, no son soroll blanc.

```
residualPlots(mBestEliminar)
```




```
##                  Test stat Pr(>|Test stat|)
## poly(age, 2)
## f.education.num
## f.RelType
## f.OccupationType
## Tukey test         -4.1385        3.496e-05 ***

par(mfrow=c(2,2))
plot(mBestEliminar,id.n=0)
```




```
par(mfrow=c(1,1))

#Actualitzem el nostre df amb les 13 observacions esborrades.:
df<-dfBestEliminar
```

Veiem que tot i eliminar les observacions que ens ha indicat influencePlot(mBest,id=list(method="noteworthy",n=5)), veiem que els residus segueixen tenint un patró, no estàn centrats en el 0, continuem tenint problemes, no es soroll blanc, continuen tenint informació. També es causat a que el model es dolent ja que te una baixa explicavilitat. Hem de tenir en compte que no estem eliminant totes les observacions que no cumplien amb els requisits dels limits marcats per StudRes, hatvalues i cookD, sino que hem eliminat només aquells que estaven fora dels limits i a més eren els més influents mencionats per el influencePlot.

A continuació el que fem es com modelar una variable de resposta quan aquesta variable de resposta es un factor, en el nostre cas es un factor que ens diu si guanya o no guanya més de 50k $ l'any, la variable Y.bin.

Binary target

Primer farem una dicisió, per finalitats de validació de models, aleshores el que fem es fer un split, dividir entre dos mostres, la mostra de treball i la mostra de test. Per fer-ho considerem el 75% de les observacions triades a l'atzar com la mostra de treball i el 25% restant com la mostra de test, ens ajuda a fer una validació. ## Split into 2 samples: Work and Test

```
# 75% to Working Set and 25% Test
# Útil per les confusion tables
set.seed(14031997)
ll<-sample(1:nrow(df),nrow(df)*0.75)
ll<-sort(ll)
dfwork<-df[ll,]
dftest<-df[-ll,]
```

Treballarem exclusivament amb la mostra de treball, amb dfwork, per fer el disseny de quin es el millor model per explicar el target. I després un cop tinguem el millor model aleshores per veure la capacitat predictiva i si hi ha sobreparametrització calcularem la capacitat predictiva en la mostra de treball i en la mostra de test, això ho farem mitjançant confusion tables.

Modelling

Using only covariates as explanatory vars

Dintre del modelatge no tenim cap particularitat, comencem com en el tema pasat, comencem volent veure la relació entre la variable de resposta, que ara es binaria, i abans era la numèrica hours.per.week, segons les variables que eren explicatives numèriques.

```
#Agafem les variables numèriques, les posem en la llista de vars_exp
vars_exp<-names(df)[c(1,5,11:13,26)];vars_exp

## [1] "age"            "education.num"  "capital.gain"   "capital.loss"
## [5] "hours.per.week" "capital.var"

#> vars_exp<-names(df)[c(1,5,11:13,28)];vars_exp
#[1] "age"            "education.num"  "capital.gain"   "capital.loss"
```

```
#[5] "hours.per.week" "capital.var"
summary(dfwork[,vars_exp])
```

Primeres de les coses que fem, tenim el target, que fiquem el target, que es el Y.bin. Veiem un resum de la cantitat d'observacions que guanyen <=50K i els que guanyen >50K, després també ho veiem en %, on veiem que els que guanyen >50K representen 23.43% i els que guanyen <=50K representen la resta.

```
summary(dfwork$Y.bin)

## <=50K  >50K
##  2813   861

prop.table(table(dfwork$Y.bin))
##     <=50K      >50K
## 0.7656505 0.2343495
```

Fem servir catdes per veure quines variables numèriques estan relacionades amb el target.

```
catdes(dfwork[,c("Y.bin",vars_exp)],1)
```

Veiem que globalment les variables numeriques que estan relacionades amb el target, mirem Eta2. Com més intensa es la relació entre la variable numèrica i el target més alt serà Eta2, per tant, veiem que les variables que estan més relacionades son education.num, capital.gain, capital.var, hours.per.week, age i el capital.loss. Com que en tenim poquetes el que ens està dient aquest output es que totes les que hem triat a vars_exp estàn relacionades amb la variable target binaria. Les agafem totes

Una de les coses que canviem respecte a la modelització de hours.per.week es que ara treballem amb el mètode glm. El primer paràmetre es la equació que defineix el target (Y.bin). Li posem family=binomial perque li hem de dir al model quina es la distribució, quin es el model probabilista que volem per al target. No hi posem capita.var, que te una dependencia lineal perfecte amb altres variables.

```
mYbin<-glm(Y.bin~age+education.num+capital.gain+capital.loss+hours.per.week,family=binomial,data=dfwork)
summary(mYbin)

##
## Call:
## glm(formula = Y.bin ~ age + education.num + capital.gain + capital.loss +
##     hours.per.week, family = binomial, data = dfwork)
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.466e+00  3.536e-01 -23.945  < 2e-16 ***
## age             4.386e-02  3.771e-03  11.632  < 2e-16 ***
## education.num   3.125e-01  2.039e-02  15.329  < 2e-16 ***
## capital.gain    3.032e-04  2.746e-05  11.042  < 2e-16 ***
## capital.loss    6.911e-04  9.587e-05   7.209 5.63e-13 ***
## hours.per.week  4.632e-02  4.593e-03  10.085  < 2e-16 ***
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2995.7  on 3668  degrees of freedom
## AIC: 3007.7
##
## Number of Fisher Scoring iterations: 5
```

Veiem que tenim una primera part que es la fórmula, després veiem un aparatt que te a veure amb els residus, deviance residuals, la deviance es una mesura de discrepància entre la observació i la predicció, la predicció que farà aquest model es una probabilitat, és a dir, la probabilitat de resposta positiva que es guanyar >50k $ l'any.

Ens fixem en la part corresponent als coefficients, aqui tenim la taula de resultats que per totes les variables seleccionades en el model més un terme que es el de la constant, veiem l'estimador del paràmetre, els seu standard error. A partir de tenir l'estimador i Std. Error doncs ja es pot tenir quin es el z value de la hipòtesi nula coeficient = 0.

A la part final del summary veiem que ens diu un Null deviance que seria la mesura de discrepància corresponent al model null, que son 4000.8 on 3673 degrees of freedom.

Després tenim el Residual deviance, que es una discrepància que es de 2995.7 on 3668 degrees of freedom (amb tants graus de llibertat com numero d'observacions del model menys numero de paràmetres en el model). Finalment tenim el AIC. La principal diferència amb el modelatge de hours.per.week es que ara no obtenim el coeficient de determinació del model.

Mirem si hi ha un altre model que tingui menys variables i es pugui donar la mateixa explicavilitat, intentem fer comparacions de models. El nou model es canviar capital.gain i capital.loss per la variable que abans hem tret capital.var.

```
mYbinA<-glm(Y.bin~age+education.num+capital.var+hours.per.week,family=binomial,data=dfwork)
summary(mYbinA)

##
## Call:
## glm(formula = Y.bin ~ age + education.num + capital.var + hours.per.week,
##     family = binomial, data = dfwork)
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.582e+00  3.504e-01  -24.49   <2e-16 ***
## age             4.503e-02  3.710e-03   12.14   <2e-16 ***
## education.num   3.244e-01  2.014e-02   16.11   <2e-16 ***
## capital.var     2.272e-04  2.247e-05   10.11   <2e-16 ***
## hours.per.week  4.823e-02  4.520e-03   10.67   <2e-16 ***
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 3091.8  on 3669  degrees of freedom
## AIC: 3101.8
##
## Number of Fisher Scoring iterations: 5
```

Veiem que el residual deviance es més alt, Residual deviance: 3091.8, desde aquest punt de vista sembla que la discrepancia queda millor capturada en aquest segon model. mYbinA ? nested into mYbin?

```
anova(mYbinA,mYbin,test="Chisq")  # No és correcte ja que aquest models no son encaixats, el model p
etit hauria d'estar inclòs en el model gran, i això no és així.

## Analysis of Deviance Table
##
## Model 1: Y.bin ~ age + education.num + capital.var + hours.per.week
## Model 2: Y.bin ~ age + education.num + capital.gain + capital.loss + hours.per.week
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3669     3091.8
## 2      3668     2995.7  1   96.113 < 2.2e-16 ***

#Per tant comparem els dos models que no son nested utilitzem el BIC:
BIC(mYbin,mYbinA) # Best mYbin, minimum BIC, el model que te les dues variables per separat i no cap
ital.var.

##        df      BIC
## mYbin   6 3044.961
## mYbinA  5 3132.865
```

Mirem si totes les variables incloses son significatives amb el test d'efectes nets, en aquest cas com les variables explicatives son n umèriques doncs seria un test d'efectes nets en el que el p-value que ens dona es el mateix p-value que ens surt en la taula anova, no es massa informatiu. El test Anova d'efectes nets funciona molt bé quan hi ha variables explicatives que son factors.

```
Anova(mYbin,test="LR") #Veiem que totes les variables aporten al model.

## Analysis of Deviance Table (Type II tests)
##
## Response: Y.bin
##                LR Chisq Df Pr(>Chisq)
## age             140.294  1  < 2.2e-16 ***
## education.num   275.807  1  < 2.2e-16 ***
## capital.gain    220.993  1  < 2.2e-16 ***
## capital.loss     51.974  1  5.625e-13 ***
## hours.per.week  110.400  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#Utilitzem el métode step per buscar el model que sigui el més explicatiu posible pero que sigui el
més simple posible.
mYbinB<-step(mYbin,k=log(nrow(dfwork))) # Use k=log(nrow(dfwork)) for BIC
```

```
## Start:  AIC=3044.96
## Y.bin ~ age + education.num + capital.gain + capital.loss + hours.per.week
##
##                 Df Deviance    AIC
## <none>               2995.7 3045.0
## - capital.loss    1   3047.7 3088.7
## - hours.per.week  1   3106.1 3147.2
## - age             1   3136.0 3177.0
## - capital.gain    1   3216.7 3257.7
## - education.num   1   3271.5 3312.6
```

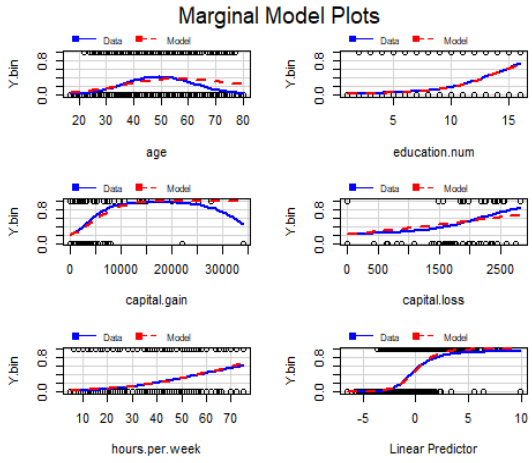
Veiem que totes les variables son útils i que no poden desaparèixer del model. Per tant, el millor model continua sent el mYbin. De totes maneres veiem si tenim colinealitat abans de deixar totes les variables:

```
vif(mYbinB)
```

```
##            age  education.num   capital.gain   capital.loss hours.per.week
##       1.020456       1.007595       1.006165       1.005288       1.016327
```

Check colinearity, sembla que ens les hem de quedar totes, pero podem utilitzar-les tal com estàn o potser seria millor aplicar alg un tipus de transformació. Per això mirem la relació marginal entre la resposta i les variables explicatives, això ho fem am el matg inalModelPlots:

```
marginalModelPlots(mYbinB)
```



Primer mirem la relació que hi ha entre la probabilitat de resposta positiva de Y.bin i la variable age, veiem que sembla que el blau (les dades), el que veiem es que tenim un increment de la probabilitat de guanyar més diners que estaria més o menys en els 50 anys. Segons el que ens diu el model (la linia vermella),aleshores veiem que en funció de l'edat veiem que el que fa el model es fixar un màxim al voltant dels 60 anys i que disminueix un cop arribat al seu màxim de manera molt més relaxada del que ho fan les dades. La relació entre la variable age i la probabilitat es no lineal. Veiem que hi ha un cert desajust en l'edat entre el model i les dades.

En el education.num sembla que no hi ha gairebé gens de desajust. Com més anys estudiin els individus més probabilitat de guanyar més diners tenen.

Sembla que també hi ha una certa correspondència, com més capital.gain més probabilitat tenen de guanyar +50k $. Tot i això veiem que hi ha observacions extranyes.

En el capital.loss també veiem que tenim un cert desajust i en les hours.per.week no en tenim gairabé gens, similar al que pasa amb education.num, és a dir, com més hores treballi un individu més probabilitat té de guanyar més diners.

Globalment veiem que entre el model i les dades hi ha una certa correspondència, això ens indica que el model no funciona del tot malament. Veiem que a mesura que augmenten de forma general els valors de les variables explicatives també augmenta la probabilitat de guanyar més de 50k $.

Intententem fer alguna transformació en les variables on tenim més desajust per així millora-la i tenir més correspondència entre el model i les dades. Fem per la variable age: El capital.loss també es una variable juntament amb la variable age que te menor corresppondencia les dades amb el model. Però capital.loss es una variable força difícil, perque només te uns quants valors que son diferents a 0, i potser interesa més utilitzar-la com a factor, que ja ho veurem més endavant. El que si que veiem es que a

major capital.loss vol dir que també es tenien més diners i que per tant tambe la probabilitat de guanyar més de 50k $ l'any es més elevada.

```
mYbinC<-glm(Y.bin ~ poly(age,2) + education.num + capital.gain + capital.loss + hours.per.week, family=binomial,data=dfwork)

summary(mYbinC)
## Call:
## glm(formula = Y.bin ~ poly(age, 2) + education.num + capital.gain +
##     capital.loss + hours.per.week, family = binomial, data = dfwork)
…
##
##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2891.1  on 3667  degrees of freedom
## AIC: 2905.1
##
## Number of Fisher Scoring iterations: 6
```

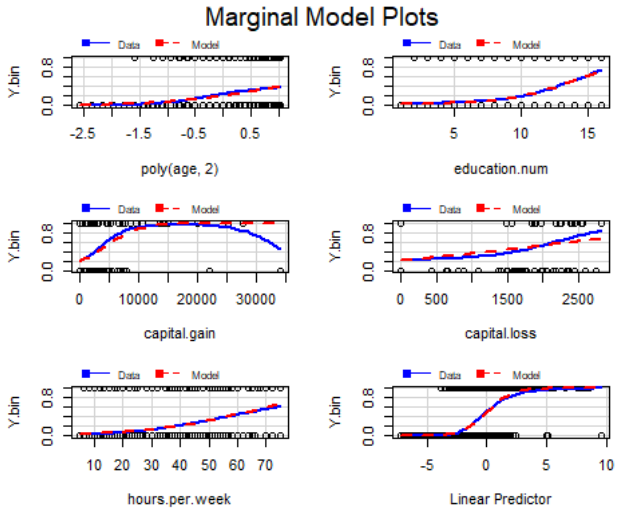
Veiem que te Residual deviance: 2891.1, apliquem ara el test de la deviance: Deviance test

```
anova(mYbin,mYbinC,test="Chisq")  # rebutjem la H0, com que la rebutjem vol dir que els dos model no son equivalents, el gran fa millor feina que el petit, ens quedem amb el model mYbinC.

## Analysis of Deviance Table
##
## Model 1: Y.bin ~ age + education.num + capital.gain + capital.loss + hours.per.week
## Model 2: Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +
##     hours.per.week
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      3668     2995.7
## 2      3667     2891.1  1   104.65 < 2.2e-16 ***
```

Analysis of Deviance Table

Model mYbin: Y.bin ~ age + education.num + capital.gain + capital.loss + hours.per.week Model mYbinC: Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week Actualitzem llavors mYbin amb mYbinC:

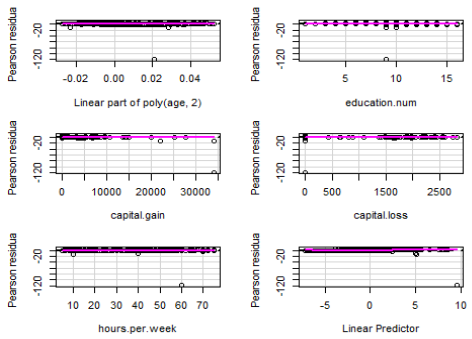
```
mYbin <- mYbinC
#  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
#1      3668     2995.7
#2      3667     2891.1  1   104.65 < 2.2e-16 ***
marginalModelPlots(mYbin)
```



Veiem que la age ara està molt més ajustada, veiem que hem millorat molt el model. La capital.loss que ara veiem que es la que més desajustada està, la preferim provar de treballar com a factor binari i no fer-li una transformació com li hem fet a la variable age per les seves característiques.

Mirem els residus: fem servir l'eina de residualPlots

```
residualPlots(mYbin)
```




```
##                Test stat Pr(>|Test stat|)
## poly(age, 2)
## education.num     0.4909           0.4835
## capital.gain     27.9129        1.269e-07 ***
## capital.loss      5.4559           0.0195 *
## hours.per.week    1.4101           0.2350
```

Veiem que hi ha observacions que son outliers dels residus. Aquestes observacions son interesants de ser detectades, ja vam detectar i eliminar observacions massa influents en el pasat tema (amb hours.per.week). Aquí no mirem patrons, a diferencia del pasat tema. Ens hem de fixar amb la linia de color rosa, quan aquesta es plana es que el model està força bé, que es el nostre cas. Veiem que el gràfic més important, el que te el predictor lineal per una banda i després els residus per una altre, veiem que la linia rosa es plana, cosa molt positiva, es un model molt satisfactori de moment utilitzant només variables explicatives numèriques.

Ens fixem que per cadascuna de les variables explicatives numèriques sense transformacions polinòmiques podem veure una h0 que acaba treient un p-value, aquesta h0 diu que els residus son prou aleatoris marginalment segons els valors de les variables explicatives, és a dir, en tots els casos que s'accepta la h0 vol dir que el tractament que s'ha donat a la variable numèrica es l'adequat per la modelització, almenys desde el punt de vista dels residus. Veiem que els que rebutjen més la h0 son el capital.gain i el capital.loss, per tant estaria bé buscar-hi una solució.

Ara ja tenim el millor dels models pel que fa a les variables explicatives que son numèriques, pasem a analitzar si l'addició de variables explicatives factor poden millorar-lo.

Adding factors

El primer que fem es substituir capital.gain i capital.loss per el seus fatcors corresponents. Recordem que un cop substituim per factors haurem de comparar per l'indicador BIC, ja que no son model encaixats.

De moment el millor model fins al moment es el mYbin: (Y.bin ~ poly(age,2) + education.num + capital.gain + capital.loss + hours.per.week, family=binomial,data=dfwork)

```
#Probem amb f.cvar i amb l'altre opció de f.cgain i f.closs
mYbinFactor1<-glm(Y.bin ~ poly(age,2) + education.num + f.cvar + hours.per.week, family=binomial,dat
a=dfwork)
mYbinFactor2<-glm(Y.bin ~ poly(age,2) + education.num + f.cgain + f.closs + hours.per.week, family=b
inomial,data=dfwork)

BIC(mYbin,mYbinFactor1,mYbinFactor2)

##              df      BIC
## mYbin         7 2948.524
## mYbinFactor1  7 3020.979
## mYbinFactor2  7 3020.979

#Veiem que el model amb millor BIC segueix sent mYbin, on no utilitzem cap dels factors.
```

Probem amb f.age i f.education.num:

```
mYbinFactor3<-glm(Y.bin ~ f.age + education.num + capital.gain + capital.loss + hours.per.week, fami
ly=binomial,data=dfwork)
mYbinFactor4<-glm(Y.bin ~ poly(age,2) + f.education.num + capital.gain + capital.loss + hours.per.we
ek, family=binomial,data=dfwork)
BIC(mYbin,mYbinFactor3,mYbinFactor4)

##              df      BIC
## mYbin         7 2948.524
## mYbinFactor3  8 2991.185
## mYbinFactor4  9 2981.745
```

Veiem que continua sent millot mYbin Intentem en el model de mYbinFactor2 aplicar-li un mètode step

```
mYbinFactor2Step<-step(mYbinFactor2,k=log(nrow(dfwork)))

## Start:  AIC=3020.98
## Y.bin ~ poly(age, 2) + education.num + f.cgain + f.closs + hours.per.week
##
##                 Df Deviance    AIC
## <none>                2963.5 3021.0
## - f.closs         1   3001.1 3050.4
## - hours.per.week  1   3023.3 3072.6
## - f.cgain         1   3110.7 3160.0
## - poly(age, 2)    2   3217.0 3258.0
## - education.num   1   3240.1 3289.3
```

Veiem que el mètode step ens deixa tal qual el model, per tant deixem el capital.loss i el capital.gain com a numèriques dins del nostre model tal i com hem comprovat just abans amb els BIC's.

Continuem intentant afegir-hi factors en el millor model fins al moment, només considerem aquelles variables que son factors i els efectes principals.

```
vars_edis<-names(df)[c(10,16:25,27,28)]
vars_edis

##  [1] "sex"                "f.type"             "f.RelType"
##  [4] "f.CountryType"      "f.EduType"          "f.MaritalStatusType"
##  [7] "f.OccupationType"   "f.RaceType"         "f.age"
## [10] "f.cgain"            "f.closs"            "f.cvar"
## [13] "f.education.num"
```

Utilitzem catdes per veure quines son les variables factor amb més relació amb les categories del target (guanyar més o menys de 50k $ a l'any) i les que hauriem de provar abans d'afegir al nostre model.

catdes(dfwork[,c("Y.bin",vars_edis)],1)

Veiem que les que globalment estan més associats amb el target son: f.MaritalStatusType, f.RelType, f.education.num(tot i que no ens interesa ja que en el model tenim la variable numèrica educació), f.EduType (tampoc seria del tot necesaria perque es trepitja amb education.num), f.OccupationType, f.cvar, f.age (no el tenim en compte perque ja tenim l'edat tractant-la amb els termes lineals i quadràtics (poly(age,2))), f.cgain, sex, f.type,..

Use more significant gross effect factors

```
mYbinFactor3<-glm(Y.bin ~ (poly(age,2) + education.num + capital.gain + capital.loss + hours.per.wee
k)+(f.MaritalStatusType+f.RelType +f.EduType+f.OccupationType+sex+f.type+f.RaceType+f.CountryType),
family=binomial,data=dfwork)

summary(mYbinFactor3)

#Veiem que Residual deviance: 2257.0 , ha baixat força respecte.
#Null deviance: 4000.8  on 3673  degrees of freedom
#Residual deviance: 2257.0  on 3640  degrees of freedom
#AIC: 2325

vif(mYbinFactor3)
#Mirem si realment afegint-hi aquests factors al model implica una millora:

anova(mYbin,mYbinFactor3,test="Chisq")  #No son equivalents
#  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
```

```
#1      3667     2891.1
#2      3640     2257.0 27   634.05 < 2.2e-16 ***
```

Veiem que al afegir els factors el model millora globalment. Ara mirem si tots aquests factors afegits son realment significatius amb el test d'efectes nets:

```
Anova(mYbinFactor3,test="LR")

#Response: Y.bin
#                    LR Chisq Df Pr(>Chisq)
#poly(age, 2)          74.895  2  < 2.2e-16 ***
#education.num          1.589  1   0.207487
#capital.gain         182.885  1  < 2.2e-16 ***
#capital.loss          38.348  1  5.920e-10 ***
#hours.per.week        15.600  1  7.824e-05 ***
#f.MaritalStatusType   51.733  3  3.414e-11 ***
#f.RelType             21.140  3  9.848e-05 ***
#f.EduType              3.632  7   0.821073
#f.OccupationType      33.192  6  9.630e-06 ***
#sex                    4.243  1   0.039417 *
#f.type                 5.132  3   0.162376
#f.RaceType             2.090  2   0.351637
#f.CountryType         10.037  1   0.001534 **
```

Veiem que no, les que no son importants son les següents: f.EduType, f.type, f.RaceType.

Utilitzem el mètode step perque ens ajudi més ràpidament a netejar el model:

```
mYbinFactor4<-step(mYbinFactor3,k=log(nrow(dfwork)))
Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalStatu
sType + f.CountryType

#                      Df Deviance    AIC
#<none>                     2328.5 2418.8
#- f.CountryType        1   2345.0 2427.1
#- hours.per.week       1   2353.7 2435.8
#- capital.loss         1   2375.5 2457.6
#- poly(age, 2)         2   2419.0 2492.8
#- capital.gain         1   2535.9 2618.0
#- education.num        1   2615.4 2697.5
#- f.MaritalStatusType  3   2879.0 2944.7
```

Segons aquest criteri step, de tots els factors que hem intentat afegir en el model mYbinFactor3 veiem que només troba importants el f.MaritalStatusType i curiosament el f.CountryType. Veiem que mYbinFactor4: Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalStatusType + f.CountryType

```
anova(mYbinFactor4,mYbinFactor3,test="Chisq")  # No son equivalents,
#  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
#1      3663     2328.5
#2      3640     2257.0 23   71.505 7.108e-07 ***
```

Tot i tenir el resultat obtingut per el mètode step, fem un model amb només aquelles variables factor que tenen uns efectes nets en mYbinFactor3 que son significatius, treiem 3 variables factor respecte el model mYbinFactor3, concretament: f.RaceType, f.type i f.EduType.

```
mYbinFactor5<-glm(Y.bin ~ (poly(age,2) + education.num + capital.gain + capital.loss + hours.per.wee
k)+(f.MaritalStatusType+f.RelType+f.OccupationType+sex+f.CountryType), family=binomial,data=dfwork)

anova(mYbinFactor4,mYbinFactor5,test="Chisq")# No son equivalents
#  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
#1      3663     2328.5
#2      3652     2267.4 11   61.093 5.805e-09 ***
```

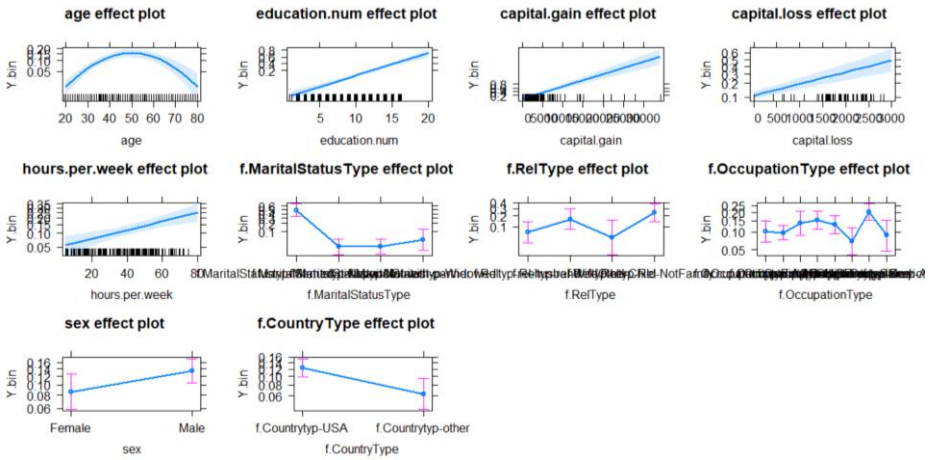
Ens quedem amb el model mYbinFactor5, tot i que les variables afegides respecte el model mYbinFactor4 tenen força categories , concretament els f.RelType+f.OccupationType, creiem que al no haver acabat la modelització poden arribar a ser útils, ens ho podem permetre i tampoc porten una exagerada complexitat al model en termes de número de categories, i molt menys obviament la variable sex.

Mirem que no tinguem colinealitats

```
vif(mYbinFactor4)
#                       GVIF Df GVIF^(1/(2*Df))
#poly(age, 2)        1.140980  2        1.033522
#education.num       1.067905  1        1.033395
#capital.gain        1.043081  1        1.021314
#capital.loss        1.016487  1        1.008210
#hours.per.week      1.057020  1        1.028115
#f.MaritalStatusType 1.199714  3        1.030812
#f.CountryType       1.007150  1        1.003568

vif(mYbinFactor5)
#                       GVIF Df GVIF^(1/(2*Df))
#poly(age, 2)         1.227012  2        1.052476
#education.num        1.414080  1        1.189151
#capital.gain         1.049536  1        1.024468
#capital.loss         1.026532  1        1.013179
#hours.per.week       1.175306  1        1.084115
#f.MaritalStatusType 32.506163  3        1.786464
#f.RelType           70.578492  3        2.032881
#f.OccupationType     1.734326  7        1.040114
#sex                  3.123928  1        1.767464
#f.CountryType        1.014009  1        1.006980
```

Veiem que no tenim colinealitats en el model que agafem finalment, el model mYbinFactor5, tot i que veiem que f.RelType i el f.MaritalStatusType tenen una certa relació com es obvi, inflen una mica la variança, pero res que sigui rellevant, no pasa de 3. Actualitzem mYbin

```
mYbin <- mYbinFactor5
summary(mYbin)

##
## Call:
## glm(formula = Y.bin ~ (poly(age, 2) + education.num + capital.gain +
##     capital.loss + hours.per.week) + (f.MaritalStatusType + f.RelType +
##     f.OccupationType + sex + f.CountryType), family = binomial,
##     data = dfwork)
…
##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2267.4  on 3652  degrees of freedom
## AIC: 2311.4
##
## Number of Fisher Scoring iterations: 8

#Per tant, de moment el model agafat fins el moment es:
#(Y.bin ~ (poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week) + (f.Marital
StatusType + f.RelType + f.OccupationType + sex + f.CountryType), family = binomial, data = dfwork)

library(effects)
plot(allEffects(mYbin))
```



Comentem una mica els diferents gràfics per entendre millor el model.

age: A mesura que s'augmenta l'edat, la probabilitat de resposta positiva i guanyar més de 50k$ doncs es va incrementant, el pic està sobre els 50 anys, a partir d'aquest valor torna a disminuir.

education.num: L'education.num pasa que a mesura que augmenten els anys d'estudis doncs aumenten la probabilitat de tenir uns ingressos per sobre de 50K

capital.gain: Si el capital.gain va creiexnt també ho fan els ingressos i com més doncs mé probabilitat de que ho siguin per sobre de 50k

hours.per.week:el capital.loss si tens pèrdues de una certa magnitud també es normal que passi perquè tens un ingressos més alts. com més capital.loss més probabilitat de guanyar 50k, tenim efecte lineal també. Pero veiem que hi han unes bandas una mica més amples i distorsionades.

sex:veiem una clara diferència en que si ets home tens més probabilitat de cobrar més de 50k de les que te una dona.

f.CountryType:Tenim que per individus que tinguin nacionalitat d'Estats Units hi han més probabilitat de que aquests guanyin més de 50k comparat amb els que no ho són

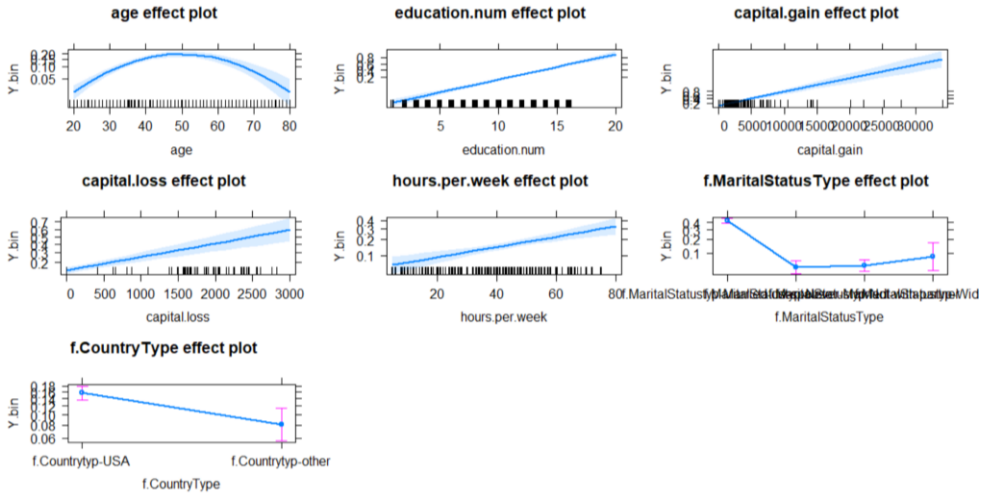
els gràfics per les variables f.MaritalStatusType, f.RelType, f.OccupationType no es veuen massa clars tal i com ens els treu el R i no podem massa cosa a partir dels gràfics que se'ns proporcionen, tot i això podem fer un petit comentari on els married comparats amb les altres categories de f.MaritalStatusType son els que tenen més probabilitats de guanyar +50k$ l'any. Per altre banda els que son f.Reltyp-Rel-NotFamily son dels que tenen més probabilitats i els Child com es obvi els que menys. Tot i així podem veure que en totes aquestes variables tenen una certa amplitud de forquilla.

Treballem el afegir interactions, seleccionarem el millor model, validarem, i veurem la seva capacitat predictiva

Per començar a afegir les interaccions, prendrem millor un model més simplificat que el últim mYbin actualitzat, que contenia més variables factors de les que proposava el mètode step. Des del punt de vista tècnic tenim una discrepància, tenim una estratègia de simplificació de models que està basada en el mètode step que ens recomana un determinat model, pero aquest model no es consistentm no es el mateix del que ens donen les eines d'inferència. Prenem la decisió de prendre el criteri de simplificació proposat per el mètode step per els pasos següents per simplificar el procès a part de ser el més aconsellat.

```
mYbin <- mYbinFactor4
summary(mYbin)

##
## Call:
## glm(formula = Y.bin ~ poly(age, 2) + f.education.num + capital.gain +
##     capital.loss + hours.per.week, family = binomial, data = dfwork)
## Coefficients:
## (Intercept)                             ***
## poly(age, 2)1                           ***
## poly(age, 2)2                           ***
## f.education.numf.education.num(9)       ***
## f.education.numf.education.num(10-12)   ***
## f.education.numf.education.num(13-16)   ***
## capital.gain                            ***
## capital.loss                            ***
```

```
## hours.per.week                          ***

##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2907.9  on 3665  degrees of freedom
## AIC: 2925.9

#Per tant, de moment el model agafat fins el moment es:
#(Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalSta
tusType + f.CountryType, family = binomial, data = dfwork)

library(effects)
plot(allEffects(mYbin))
```



Afegint interaccions

```
#(Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalSta
tusType + f.CountryType, family = binomial, data = dfwork)
mYbinI15<-glm(Y.bin ~ (poly(age,2) + education.num + capital.gain + capital.loss + hours.per.week)*(
f.MaritalStatusType + f.CountryType), family=binomial,data=dfwork)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(mYbinI15)
## Call:
## glm(formula = Y.bin ~ (poly(age, 2) + education.num + capital.gain +
##     capital.loss + hours.per.week) * (f.MaritalStatusType + f.CountryType),
##     family = binomial, data = dfwork)
##
…

##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2271.7  on 3639  degrees of freedom
## AIC: 2341.7

mYbinI16<-step(mYbinI15,k=log(nrow(dfwork)))

…Step:  AIC=2418.82
## Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +
##     hours.per.week + f.MaritalStatusType + f.CountryType
##
##                       Df Deviance    AIC
## <none>                     2328.5 2418.8
## - f.CountryType        1   2345.0 2427.1
## - hours.per.week       1   2353.7 2435.8
## - capital.loss         1   2375.5 2457.6
## - poly(age, 2)         2   2419.0 2492.8
## - capital.gain         1   2535.9 2618.0
```

```
## - education.num        1   2615.4 2697.5
## - f.MaritalStatusType  3   2879.0 2944.7
```

El millor model que obté es sense interaccions entre variables numèriques i factor, no son rellevants cap de les interaccions que hem intentat introduir, Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalStatusType + f.CountryType

Afegim interaccions entre aquells factors que son rellevants en el model simplificat son el f.MaritalStatusType+f.CountryType, segons el metode de simplificació com hem comentat just al principi de l'apartat de "Afegint interaccions":

```
mYbinI17<-glm(Y.bin ~ (poly(age,2) + education.num + capital.gain + capital.loss + hours.per.week)+(
f.MaritalStatusType+f.CountryType)^2, family=binomial,data=dfwork)
mYbinI18<-step(mYbinI17,k=log(nrow(dfwork)))
```

…

```
## Step:  AIC=2418.82
## Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +
##     hours.per.week + f.MaritalStatusType + f.CountryType
##
##                       Df Deviance    AIC
## <none>                     2328.5 2418.8
## - f.CountryType        1   2345.0 2427.1
## - hours.per.week       1   2353.7 2435.8
## - capital.loss         1   2375.5 2457.6
## - poly(age, 2)         2   2419.0 2492.8
## - capital.gain         1   2535.9 2618.0
## - education.num        1   2615.4 2697.5
## - f.MaritalStatusType  3   2879.0 2944.7
```

Veiem que el mètode step precindeix en aquest cas de les interaccions entre factors.

```
vif(mYbinI18)  # No tenim colinearitat
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## poly(age, 2)        1.140980  2        1.033522
## education.num       1.067905  1        1.033395
## capital.gain        1.043081  1        1.021314
## capital.loss        1.016487  1        1.008210
## hours.per.week      1.057020  1        1.028115
## f.MaritalStatusType 1.199714  3        1.030812
## f.CountryType       1.007150  1        1.003568
```

El model amb el que acabem l'estudi de afegir o no interaccions per tant es el següent: (Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week + f.MaritalStatusType + f.CountryType, family = binomial, data = dfwork)

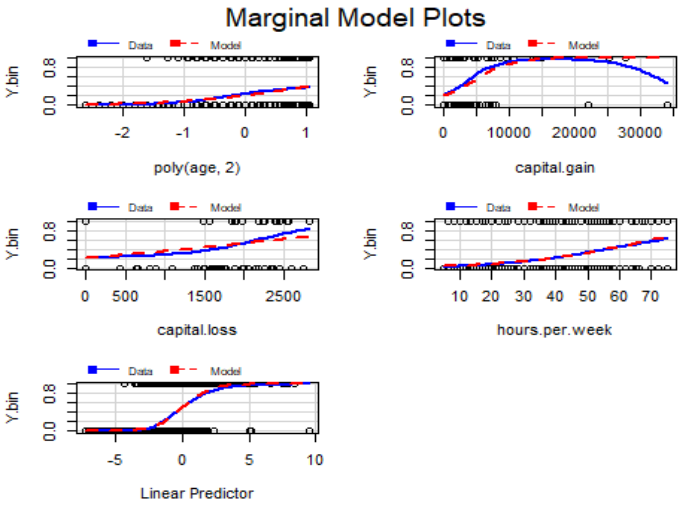
Model Validation En aquest cas no podem veure el coeficient de determinació ja que simplement no hi és.

```
summary(mYbin)
```

```
##
## Call:
## glm(formula = Y.bin ~ poly(age, 2) + f.education.num + capital.gain +
##     capital.loss + hours.per.week, family = binomial, data = dfwork)
##
…

##     Null deviance: 4000.8  on 3673  degrees of freedom
## Residual deviance: 2907.9  on 3665  degrees of freedom
## AIC: 2925.9
```

Veiem que tenim un Residual deviance: 2328.5, te una distribució de shi quadrat 3663 degrees of freedom. Si la Residual deviance de un model es de més o menys els graus de llibertat aleshores vol dir que el model s'ajusta bé a les dades. Veiem que tenim força diferencia, per tant el que ens està dient es que aquest model desde el punt de vista pràctic sembla que ajusta bé les dades, es un bon model per les dades. Primer en la validació del model mirem si tenim desajustos entre les nostres dades i el model.

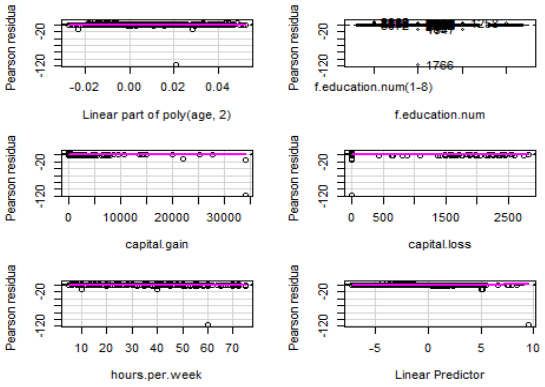
```
marginalModelPlots(mYbin)
```



Veiem que per age, education.num i hours.per.week les dades s'ajusten perfectament al model. El capital.gain i el capital.loss no ho acaben de fer. Veiem que la capital.gain tenim un individu bastant especial a la part inferior dreta, on tenim un capital.gain elevat, i potser en part, es aquest individu que causa desajustos entre el model i les dades. Per altre banda com ja hem comentat foça cops te un comportament extrany. Globalment el model te un fit adequat.

Mirem els residual plots

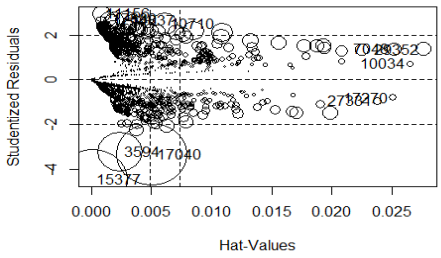
```
residualPlots(mYbin)
```




```
##                 Test stat Pr(>|Test stat|)
## poly(age, 2)
## f.education.num
## capital.gain      27.6867        1.426e-07 ***
## capital.loss       5.0550          0.02455 *
## hours.per.week     1.4539          0.22790
```

Aquí no mirem tema de patrons, cap normalitat. El que veiem es que hi ha una observació que te un residu atípic, un outlier dels residus. No sabem si serà influent o no, pero s'ha de mirar, ja que es una observació atípica. L'haurem de probablament descartar, ja que no podem construir un model amb aquestes variables que s'ajusti a la gran majoria de les observacions si aquesta observació està present.

```
res.ii<-influencePlot(mYbin,id=list(method="noteworthy",n=5))
```




```
res.ii
```

```
##          StudRes          Hat        CookD
## 3594  -3.2253683 2.375319e-03 0.0399694668
## 7048   1.3593864 2.514273e-02 0.0043267760
## 10034  0.7188606 2.652596e-02 0.0009011005
## 10710  2.4810911 6.150258e-03 0.0134529731
## 11156  2.9512339 7.997525e-04 0.0066369318
## 14837  2.6369046 2.877305e-03 0.0096403611
## 15377 -4.4610780 6.344242e-05 0.0959660792
## 17040 -3.3537663 5.059911e-03 0.0999714046
## 17270 -0.8121128 2.512266e-02 0.0011272104
## 17880  2.7047901 1.382053e-03 0.0056676834
## 27331 -0.9580124 2.379849e-02 0.0015853440
## 29352  1.3527784 2.764252e-02 0.0046973107
```

```
dfwork[rownames(res.ii),]
```
Ens fixem per exemple amb l'indiviud 17040, que es el que te una cookD major i també un dels que te major StudRes.

Veiem que te 20 anys, es un home de raça Black i on veiem clarament el perque es una observació atípica es perque principalment te uns grans guanys amb una edat molt jove, en concret te exce ssius guanys (capital.gain = 34095), f.Edutyp-Colleges, f.Occupationtyp-Emp-AltCarrec, només treballa 10 hores per setmana, veiem que es en general una observació amb lack of fit en el conjunt de les dades.

Comentem una altre, per exemple el 15377, que te un cookD el segon més elevat amb un valor de 0.088891414 i un StudRes de -4.1501389. Veiem que es una observació que es un home de 55 anys, divorciat, veiem que te un capital.gain molt elevat de 34095, basicament per aquest capital.gain tant elevat es una observació a tenir molt en compte.

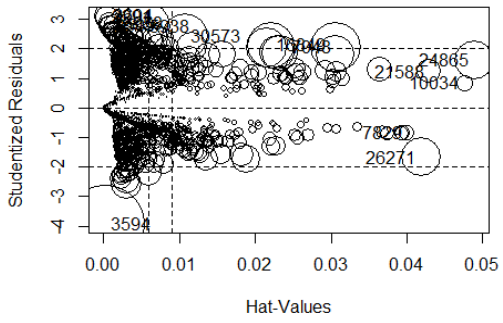
Seguim mirant i veiem que les observacions 10966 (dona de 52 anys, widowed, treballa només 20 hores i guanya més de 50k$ a l'any).

Algunes de les observacions que hem estat analitzant no les considerem a eliminar perque no les hem trobat amb valors gens extranys excessivament.

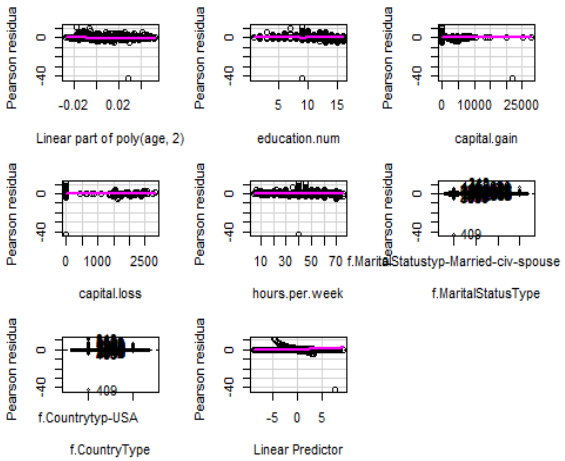
```
ll<-which(rownames(dfwork) %in% c("17040","15377","10966"))
#Recalculem el model:
mYbinI19<-glm(Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss + hours.per.week +
f.MaritalStatusType + f.CountryType, family=binomial,data=dfwork[-ll,])
mYbinI20<-step(mYbinI19,k=log(nrow(dfwork)))
```

```
## Start:  AIC=2387.92
## Y.bin ~ poly(age, 2) + education.num + capital.gain + capital.loss +
##     hours.per.week + f.MaritalStatusType + f.CountryType
##
##                       Df Deviance    AIC
## <none>                     2297.6 2387.9
## - f.CountryType        1   2314.1 2396.2
## - hours.per.week       1   2322.2 2404.3
## - capital.loss         1   2343.6 2425.7
## - poly(age, 2)         2   2384.4 2458.3
## - capital.gain         1   2534.3 2616.4
## - education.num        1   2581.3 2663.4
## - f.MaritalStatusType  3   2845.2 2910.9
```

```
res.ii2<-influencePlot(mYbinI20,id=list(method="noteworthy",n=5))
```




```
#Veiem que ja hem eliminat les observacions que voliem.
mYbin<-mYbinI20
residualPlots(mYbin)
```

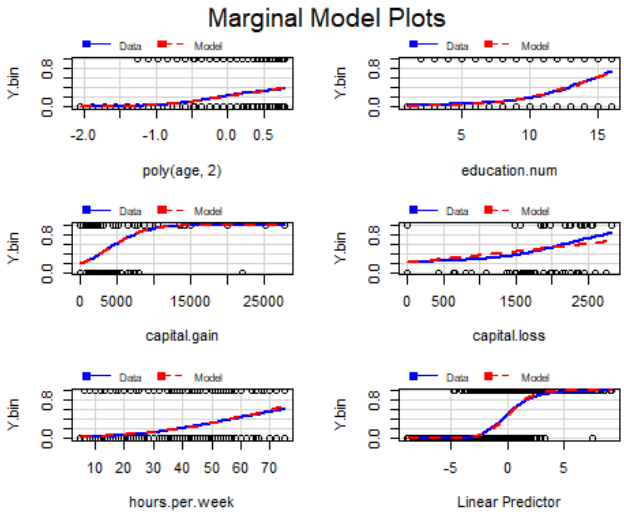



```
##                      Test stat Pr(>|Test stat|)
## poly(age, 2)
## education.num           2.3235          0.12744
## capital.gain            1.2551          0.26258
## capital.loss            6.1441          0.01318 *
## hours.per.week          3.7308          0.05342 .
## f.MaritalStatusType
## f.CountryType
```

Ara veiem que la gràfica de capital.gain millora molt clarament, ens hem de fixar amb la linia de color rosa, quan aquesta es plana es que el model està força bé, que es el nostre cas en totes les variables explicatives.

Mirem el marginalModelPlots i ara si que veiem que hem millorat molt l'ajustament de les dades i el model en capital.gain que es on teniem més desajust, ja que age, education.num i hours.per.week ja estàven molt ben ajustades.

```
marginalModelPlots(mYbin)
```



Confusion table

Ara el que fem es un cop tenim el millor model validat i veiem que les dades s'ajusten bé a aquest el que fem es quantificar-ho, la millor manera de fer-ho es utilitzar confusion table. Fem una confusion table amb la qual cosa el que fem es la predicció per totes les observacions que tenim en la mostra de treball dfwork de les probabilitats tipo resposta que tenen de guanyar més de 50k$ a l'any.

millor i final model: mYbin

Confusion table per la mostra dfwork utilitzant el millor model fins al moment mYbin:

Agafem les 10 primeres observacions de la nostra mostra de treball i mirem quina predicció fa el nostre model final mYbin sobre aquestes:

```
predict(mYbin,type="response")[1:10]

##           1           2           3          20          37          40
## 0.206493323 0.548593136 0.027476397 0.206229598 0.005747426 0.628794236
##          42          65          85          95
## 0.707247241 0.029336326 0.897523450 0.559212057
```

Hem de definir un llindar a partir del qual transformem aquestes probabilitats en una resposta positiva o negativa. Diem que els que tenen una probabilitat inferior de 0.5 aleshores son observacions que no tenen resposta positiva, per tant guanyen menys de 50k, $elsquetenenunaprobabilitatmésgrande0.5pertanyenalgrupde + 50k$. Creem un factor amb tantes observacions com dimensió te la dfwork.

```
premYbin<-factor(ifelse(predict(mYbin,newdata=dfwork,type="response")<0.5,0,1),labels=c("<50k$/any P
red","+50k$/any Pred"))
tYbin<-table(premYbin,dfwork$Y.bin)
tYbin
## premYbin         <=50K >50K
##   <50k$/any Pred  2643  367
##   +50k$/any Pred   170  494
```

Veiem que en columnes tenim la realitat de la mostra i les files son les prediccions del nostre model mYbin. Veiem que tenim 2643 observacions que guanyen en la realitat menys de 50k$ i el model també prediu que guanyen menys de 50k, $pertant, l'encerta. Desprésveiemlasegonadelescolumnesquesonlesqueguanyenmésde50k$ i tenim que el model també diu que guanyen més de 50k$ en 494 observacions, també l'encerta. El problema està fora d'aquesta diagonal, el 367 son gent que guanya realment més de 50k pero que el nostre model prediu que en guanyen menys, per tant s'ha equivocat, el mateix pasa amb els 170 que son gent que guanya menys de 50k$ l'any i que el nostre model prediu que en guanyen més.

La qualitat predictiva del model la podem obtenir suman les diagonals i dividint per el nombre d'observacions: La capacitat predictiva del model mYbin es del 85.38378%

```
capamYbin<-100*(sum(diag(tYbin))/nrow(dfwork));capamYbin ## [1] 85.38378
```

Tot i això el que ens interesa saber es si aquesta capacitat predictiva es casualitat o de fet es una conseqüència de que hem treballat bé el model.

Veiem quin seria el valor corresponent al encertar en el cas del model null. És aquell model que fa la predicció per tots els individus que no guanyen més de 50k$ l'any (representen la majoria).

```
mYbinNULL<-glm(Y.bin~1,family=binomial,data=dfwork)
predict(mYbinNULL,type="response")[1:10]

##         1         2         3        20        37        40        42        65
## 0.2343495 0.2343495 0.2343495 0.2343495 0.2343495 0.2343495 0.2343495 0.2343495
##        85        95
## 0.2343495 0.2343495

premYbinNULL<-factor(ifelse(predict(mYbinNULL,type="response")<0.5,0,1),labels=c("<50k$/any Pred"))
tYbinNULL<-table(premYbinNULL,dfwork$Y.bin)
tYbinNULL
## premYbinNULL     <=50K >50K
##   <50k$/any Pred  2813  861
```

Veiem que tenim 2813 que guanyen a la realitat menys de
50k
, $ilapredicciódelmodelnullencerta, ensdiuelmateix, peroperaltrebanda, totselsqueguanyenmésestanmalpredits. premYbinNULL <= 50K > 50K < 50k$/any Pred 2813 861

Mirem la capacitat predictiva d'aquest model null, calculem els que encerta respecte el total:

```
capamYbinNULL<-100*(tYbinNULL[1,1]/nrow(dfwork));capamYbinNULL ## [1] 76.56505
```

Veiem que encerta el 76.56505% dels casos. Concloem que després de la feina de modelització hem aconseguit introduir una millora en la capacitat predictiva del model del 10%.

Veiem com es comportaria el nostre model si haguès de fer la predicció sobre el data frame dftest. Fem el exactament el mateix procediment que anteriorment pero amb dftest en comptes de dfwork.

Confusion table for work sample, using Final Model

```
predict(mYbin,newdata=dftest,type="response")[1:10]

##           18           28           53           56           76           94
## 0.0151614358 0.3855026292 0.9428950655 0.4191781385 0.0042328306 0.2100186559
##          111          142          161          221
## 0.0260966805 0.0128673637 0.0007949537 0.2424252865

premYbinTEST<-factor(ifelse(predict(mYbin,newdata=dftest,type="response")<0.5,0,1),labels=c("<50k$/a
ny Pred","+50k$/any Pred"))
tYbinTEST<-table(premYbinTEST,dftest$Y.bin);tYbinTEST

##
## premYbinTEST     <=50K >50K
##   <50k$/any Pred   871  122
##   +50k$/any Pred    71  161
capamYbinTEST<-100*(sum(diag(tYbinTEST))/nrow(dftest));capamYbinTEST ## [1] 84.2449
```

Veiem que encerta el 84.2449% dels casos, ha minvat al voltant del 1%. Ens fixem que ara el percentatge d'encert en la mostra dftest que no s'ha utilitzat en la construcció del model, si hi ha grans diferencies entre la mostra de test i la mostra de treball vol dir que el model està sobreparametritzat, en el nostre cas no es així.

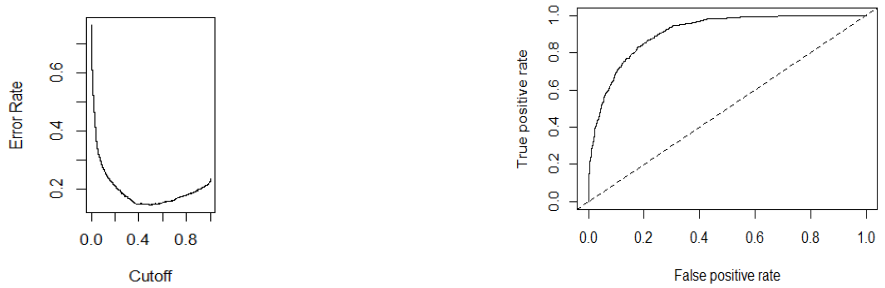
Després de l'anàlisi de capacitat predictiva veiem les ROC curve, ens permeten evaluar la capacitat predictiva d'un model en termes de discriminar un individu positiu i un de negatiu.

Utilitzem ROC curve

```
# ROC Curve

library("ROCR")
#Transformem les dades per tenir-les llestes per aplicar els mètodes perfomance de la library ROCR.
dadesROC<-prediction(predict(mYbin,newdata=dfwork,type="response"),dfwork$Y.bin)
```

Volem obtenir el plot performance de l'error segons la probabilitat, el cutoff es el llindar que ens diu quan determinem que es resposta positiva o negativa. Veiem que es en el valor sobre el 0.4 on obtenim la millor de les clasificacions. Nosaltres anteriorment hem posat 0.5, que igualment a simple vista tindria un error rate baix, molt semblant al de 0.4.

```
par(mfrow=c(1,2))
plot(performance(dadesROC,"err"))
```



Ara el que volem veure es el gràfic que es veritablament la corva ROC, que bàsicament el que ens fa es que relaciona els falsos positius amb els true positius. Dóna una idea de la capacitat discriminant del model. L'àrea sota la ROC curve indica la capacitat discriminant del model, finalitzem dient que veiem que en el nostre cas es força bo.

```
plot(performance(dadesROC,"tpr","fpr"))
abline(0,1,lty=2)
```