**ADEI: Third Deliverable**

Marc Godínez Martínez

Eduard Ortuño Garrote

07 de Juny de 2020

**Index**

[**Description**](#_heading=h.ek25fi3v98cr) **2**

[Loading data and sample selection](#_heading=h.9izsgbvxwkuq) 2

[**hours\_per\_week model construction**](#_heading=h.mqh4kro3vbz0) **4**

[Explicative Variables for modeling purposes](#_heading=h.hwgpywnxq1js) 4

[Initial modeling](#_heading=h.y8xbnuq6jwmk) 4

[Transforming variables](#_heading=h.5d8xta6zid8p) 7

[Adding factors as explanatory variables](#_heading=h.tr93jpgtfv6b) 9

[Interactions between variables](#_heading=h.mklrlmykb2um) 18

[Diagnostic of the model](#_heading=h.4bvooj195bkx) 23

[Residual analysis](#_heading=h.pdly6v6rkubl) 23

[**over\_50k model construction**](#_heading=h.mfjhz0scmzhq) **29**

[Explicative Variables for modeling purposes](#_heading=h.nbtfl4lf9gwj) 29

[Initial modeling](#_heading=h.rlem47m89jus) 29

[Transforming variables](#_heading=h.f603jw929fgj) 31

[Adding factors as explanatory variables](#_heading=h.l0vaz0mzgm7y) 37

[Interactions between variables](#_heading=h.ilmj438yijjj) 49

[Diagnostic of the model](#_heading=h.rn4o3akpuc68) 52

[Residual analysis](#_heading=h.s9g0oe8ftrao) 52

[Prediction](#_heading=h.87xph7i62umh) 55

[Work sample](#_heading=h.ioer486q11dl) 55

[Test sample](#_heading=h.yhge61wljo6n) 56

# Description

*Input variables:*

1. age: continuous.
2. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
3. fnlwgt: continuous.
4. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
5. education-num: continuous.
6. marital.status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
7. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
8. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
9. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
10. sex: Female, Male.
11. capital.gain: continuous.
12. capital.loss: continuous.
13. hours.per.week: continuous. Numeric target.
14. native.country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.
15. y.bin: Making more than $50K per year. Binary target.

## Loading data and sample selection

**if**(**!is.null**(**dev.list**())) **dev.off**()

## null device   
## 1

*# Clean workspace*  
**rm**(list=**ls**())  
  
**setwd**("/home/marc/Documentos/ADEI/DEL3/")  
  
**load**("mostra\_ME.RData")  
df**$**kmclu <- NULL  
df**$**clusterMCA <- NULL  
df**$**clusterPCA <- NULL

# hours\_per\_week model construction

## Explicative Variables for modeling purposes

Ara el que farem serà analitzar quines són les variables numèriques més relacionades amb el nostre target hours\_per\_week, per tal de decidir quines d’aquestes utilitzarem en la construcció dels diferents models fins trobar l’òptim.

**names**(df)

## [1] "age" "workclass" "final\_weight" "education"   
## [5] "education\_num" "marital-status" "occupation" "relationship"   
## [9] "race" "sex" "capital\_gain" "capital\_loss"   
## [13] "hours\_per\_week" "native\_country" "over\_50k" "f.workclass"   
## [17] "f.education" "f.education\_num" "f.marital\_status" "f.occupation"   
## [21] "f.relationship" "f.race" "f.sex" "f.native\_country"  
## [25] "f.over\_50k" "f.age" "f.capital\_gain" "f.capital\_loss"   
## [29] "f.hours\_per\_week"

vars\_num2<-**names**(df)[**c**(1,5,11**:**13)]; vars\_num2

## [1] "age" "education\_num" "capital\_gain" "capital\_loss"   
## [5] "hours\_per\_week"

*#condes(df[,vars\_num2],which(vars\_num2 == "hours\_per\_week"))*

Veiem que les variables numèriques més relacionades són education\_num, age, capital\_gain i capital\_loss, tot i que en aquestes 3 últimes la correlació que presenten és bastant baixa i poc significativa. Tot i així les considerarem també com a candidates a formar part de la construcció del model.

A partir de l’anàlisi fet fins ara, iniciarem la construcció de models, partint de un model complex de totes les variables numèriques. Realitzarem diferents anàlisis per a cada model fins a trobar el model més adient o òptim a la nostra situació o joc de dades.

## Initial modeling

m1<-**lm**(hours\_per\_week**~**age**+**education\_num**+**capital\_gain**+**capital\_loss,data=df)  
**summary**(m1)

##   
## Call:  
## lm(formula = hours\_per\_week ~ age + education\_num + capital\_gain +   
## capital\_loss, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.317 -3.045 0.517 4.310 40.683   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.069e+01 7.662e-01 40.056 < 2e-16 \*\*\*  
## age 5.489e-02 1.158e-02 4.739 2.21e-06 \*\*\*  
## education\_num 7.144e-01 6.140e-02 11.636 < 2e-16 \*\*\*  
## capital\_gain 1.258e-04 6.021e-05 2.090 0.036713 \*   
## capital\_loss 1.487e-03 3.906e-04 3.807 0.000142 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.78 on 4895 degrees of freedom  
## Multiple R-squared: 0.03985, Adjusted R-squared: 0.03906   
## F-statistic: 50.79 on 4 and 4895 DF, p-value: < 2.2e-16

Veiem que aquest model i segurament tots els que realitzarem amb el target numèric tindran una explicabilitat molt baixa (menys del 0.05 del % de les dades),i per tant serà difícil obtenir dades rellevants. Tot i així seguim amb el modeling.

No veiem cap variable que tingui un p-valor per sota del 5%, llavors ens quedem amb el mateix model.

**Anova**(m1)

## Anova Table (Type II tests)  
##   
## Response: hours\_per\_week  
## Sum Sq Df F value Pr(>F)   
## age 2610 1 22.4586 2.208e-06 \*\*\*  
## education\_num 15737 1 135.3975 < 2.2e-16 \*\*\*  
## capital\_gain 507 1 4.3661 0.0367130 \*   
## capital\_loss 1684 1 14.4930 0.0001424 \*\*\*  
## Residuals 568931 4895   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Realitzem un vif (variance inflation factor) per veure les variables explicatives del model que estan correlacionades.

**vif**(m1)

## age education\_num capital\_gain capital\_loss   
## 1.018680 1.026125 1.033347 1.017825

No veiem cap variable que tingui un vif > 3 (el que significaria un problema de col·linealitat).

m2<-**step**(m1, k=**log**(**nrow**(df)))

## Start: AIC=23339.66  
## hours\_per\_week ~ age + education\_num + capital\_gain + capital\_loss  
##   
## Df Sum of Sq RSS AIC  
## - capital\_gain 1 507.5 569439 23336  
## <none> 568931 23340  
## - capital\_loss 1 1684.5 570616 23346  
## - age 1 2610.3 571542 23354  
## - education\_num 1 15736.9 584668 23465  
##   
## Step: AIC=23335.53  
## hours\_per\_week ~ age + education\_num + capital\_loss  
##   
## Df Sum of Sq RSS AIC  
## <none> 569439 23336  
## - capital\_loss 1 1567.1 571006 23340  
## - age 1 2932.1 572371 23352  
## - education\_num 1 16710.8 586150 23469

**vif**(m2)

## age education\_num capital\_loss   
## 1.004319 1.010133 1.013011

m2<-**lm**(hours\_per\_week**~**age**+**education\_num,data=df)  
**summary**(m2)

##   
## Call:  
## lm(formula = hours\_per\_week ~ age + education\_num, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.812 -2.975 0.425 4.461 40.808   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.31007 0.76022 39.870 < 2e-16 \*\*\*  
## age 0.06023 0.01150 5.238 1.69e-07 \*\*\*  
## education\_num 0.75185 0.06074 12.379 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 10.8 on 4897 degrees of freedom  
## Multiple R-squared: 0.03635, Adjusted R-squared: 0.03595   
## F-statistic: 92.35 on 2 and 4897 DF, p-value: < 2.2e-16

Amb aquesta sortida el que podem comprovar és que les variables que són més significatives són age, education\_num i capital\_loss. Si fem el step veiem que la millor és education\_num, però un model amb només una variable és molt poc i no explicaria el suficient, llavors agafem age, education\_num i capital\_loss.

## Transforming variables

Ara el que farem serà una transformació de les nostres variables per veure si podem explicar més en el nostre model.

m3<-**lm**(**log**(hours\_per\_week)**~**age**+**education\_num**+**capital\_loss**+**capital\_gain,data=df)  
**Anova**(m3)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## age 1.50 1 12.9354 0.0003256 \*\*\*  
## education\_num 12.11 1 104.1877 < 2.2e-16 \*\*\*  
## capital\_loss 1.17 1 10.0487 0.0015339 \*\*   
## capital\_gain 0.41 1 3.4969 0.0615443 .   
## Residuals 568.99 4895   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m4 <- m3 sense capital\_gain

m4<-**lm**(**log**(hours\_per\_week)**~**age**+**education\_num**+**capital\_loss,data=df)  
**summary**(m4)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ age + education\_num + capital\_loss,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.96775 -0.03146 0.06222 0.15687 0.83553   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.383e+00 2.407e-02 140.592 < 2e-16 \*\*\*  
## age 1.399e-03 3.638e-04 3.845 0.000122 \*\*\*  
## education\_num 2.027e-02 1.927e-03 10.520 < 2e-16 \*\*\*  
## capital\_loss 3.757e-05 1.233e-05 3.048 0.002317 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.341 on 4896 degrees of freedom  
## Multiple R-squared: 0.029, Adjusted R-squared: 0.02841   
## F-statistic: 48.75 on 3 and 4896 DF, p-value: < 2.2e-16

**Anova**(m4)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## age 1.72 1 14.7828 0.0001222 \*\*\*  
## education\_num 12.87 1 110.6772 < 2.2e-16 \*\*\*  
## capital\_loss 1.08 1 9.2892 0.0023174 \*\*   
## Residuals 569.40 4896   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

m5<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+poly**(capital\_loss,2)**+poly**(capital\_gain,2),data=df)  
**summary**(m5)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + poly(education\_num,   
## 2) + poly(capital\_loss, 2) + poly(capital\_gain, 2), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.88739 -0.06796 0.02193 0.17023 1.27794   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.644806 0.004542 802.435 < 2e-16 \*\*\*  
## poly(age, 2)1 1.194290 0.325310 3.671 0.000244 \*\*\*  
## poly(age, 2)2 -8.739629 0.323392 -27.025 < 2e-16 \*\*\*  
## poly(education\_num, 2)1 1.970368 0.327833 6.010 1.99e-09 \*\*\*  
## poly(education\_num, 2)2 0.367341 0.324390 1.132 0.257519   
## poly(capital\_loss, 2)1 0.926602 0.321828 2.879 0.004004 \*\*   
## poly(capital\_loss, 2)2 0.208639 0.318460 0.655 0.512402   
## poly(capital\_gain, 2)1 0.724463 0.324262 2.234 0.025515 \*   
## poly(capital\_gain, 2)2 -0.414622 0.320224 -1.295 0.195454   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.318 on 4891 degrees of freedom  
## Multiple R-squared: 0.1568, Adjusted R-squared: 0.1554   
## F-statistic: 113.7 on 8 and 4891 DF, p-value: < 2.2e-16

**Anova**(m5)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## poly(age, 2) 75.03 2 371.1044 < 2.2e-16 \*\*\*  
## poly(education\_num, 2) 3.76 2 18.5848 9.104e-09 \*\*\*  
## poly(capital\_loss, 2) 0.88 2 4.3572 0.01286 \*   
## poly(capital\_gain, 2) 0.67 2 3.2968 0.03708 \*   
## Residuals 494.45 4891   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(m5)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(age, 2) 1.082859 2 1.020100  
## poly(education\_num, 2) 1.106229 2 1.025560  
## poly(capital\_loss, 2) 1.027797 2 1.006878  
## poly(capital\_gain, 2) 1.054807 2 1.013429

Com podem observar les nostres variables més significatives del nostre model són age y education\_num. Aplicant la logaritmes i polinomis aconseguim una explicabilitat d’un 15,6%, un número molt més gran comparat amb la explicabilitat que tenien els altres models. Per aquest motiu i per la millora significativa en els residus decidim prioritzar aquesta versió en contra de la original. Decidim llavors agafar el model m5.

## Adding factors as explanatory variables

Ara el que farem és afegir variables factors com a variables explicatives. Hem de trobar les que poden ser més significatives. Aquí comencem aquest estudi.

**names**(df)

## [1] "age" "workclass" "final\_weight" "education"   
## [5] "education\_num" "marital-status" "occupation" "relationship"   
## [9] "race" "sex" "capital\_gain" "capital\_loss"   
## [13] "hours\_per\_week" "native\_country" "over\_50k" "f.workclass"   
## [17] "f.education" "f.education\_num" "f.marital\_status" "f.occupation"   
## [21] "f.relationship" "f.race" "f.sex" "f.native\_country"  
## [25] "f.over\_50k" "f.age" "f.capital\_gain" "f.capital\_loss"   
## [29] "f.hours\_per\_week"

vars\_dis2<-**names**(df)[**c**(2,4,6,7,9,10,14,15,26**:**28)];vars\_dis2

## [1] "workclass" "education" "marital-status" "occupation"   
## [5] "race" "sex" "native\_country" "over\_50k"   
## [9] "f.age" "f.capital\_gain" "f.capital\_loss"

*#condes(df[,c("hours\_per\_week",vars\_dis2)],1,proba = 0.01)*

Després de l’execució anterior el que hem vist són les variables més correlacionades amb el nostre model que són aquelles que tenen un p-valor << 0.01. Aquestes variables són: f.age+occupation+marital\_status+sex+over\_50k+education+workclass+f.capital\_loss+f.capital\_gain Llavors ara estudiarem el cas, és a dir, al nostre model li afegim aquests factors.

Primer hem de fer un rename de marital-status a marital\_status (no deixa executar la comanda)

**names**(df)[6] <- "marital\_status"  
  
m6<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+poly**(capital\_loss,2)**+poly**(capital\_gain,2)**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**education**+**workclass,data=df)  
**summary**(m6)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + poly(education\_num,   
## 2) + poly(capital\_loss, 2) + poly(capital\_gain, 2) + occupation +   
## marital\_status + sex + over\_50k + education + workclass,   
## data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.89982 -0.09416 0.03620 0.16205 1.29952   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.585493 0.031234 114.794 < 2e-16 \*\*\*  
## poly(age, 2)1 -0.239357 0.394179 -0.607 0.543728   
## poly(age, 2)2 -7.276984 0.345146 -21.084 < 2e-16 \*\*\*  
## poly(education\_num, 2)1 2.064462 0.872054 2.367 0.017955 \*   
## poly(education\_num, 2)2 1.090385 0.831323 1.312 0.189709   
## poly(capital\_loss, 2)1 0.318699 0.313559 1.016 0.309494   
## poly(capital\_loss, 2)2 0.095343 0.306563 0.311 0.755808   
## poly(capital\_gain, 2)1 -0.074425 0.324418 -0.229 0.818559   
## poly(capital\_gain, 2)2 -0.025679 0.315124 -0.081 0.935056   
## occupationCraft-repair 0.037298 0.018696 1.995 0.046101 \*   
## occupationExec-managerial 0.080906 0.018455 4.384 1.19e-05 \*\*\*  
## occupationFarming-fishing 0.088292 0.030723 2.874 0.004073 \*\*   
## occupationHandlers-cleaners -0.009314 0.025933 -0.359 0.719495   
## occupationMachine-op-inspct 0.073938 0.022411 3.299 0.000977 \*\*\*  
## occupationOther-service -0.073175 0.018377 -3.982 6.94e-05 \*\*\*  
## occupationPriv-house-serv -0.067347 0.055982 -1.203 0.229029   
## occupationProf-specialty -0.005394 0.019371 -0.278 0.780651   
## occupationProtective-serv 0.062797 0.036280 1.731 0.083536 .   
## occupationSales 0.030060 0.018236 1.648 0.099347 .   
## occupationTech-support -0.029347 0.028808 -1.019 0.308378   
## occupationTransport-moving 0.103383 0.024860 4.159 3.26e-05 \*\*\*  
## marital\_statusMarried-AF-spouse -0.569673 0.215944 -2.638 0.008365 \*\*   
## marital\_statusMarried-civ-spouse -0.026895 0.015083 -1.783 0.074625 .   
## marital\_statusMarried-spouse-absent -0.020963 0.039453 -0.531 0.595205   
## marital\_statusNever-married -0.037350 0.016382 -2.280 0.022651 \*   
## marital\_statusSeparated -0.026927 0.026806 -1.005 0.315170   
## marital\_statusWidowed -0.098204 0.029216 -3.361 0.000782 \*\*\*  
## sexMale 0.094506 0.011285 8.374 < 2e-16 \*\*\*  
## over\_50k>50K 0.056664 0.013209 4.290 1.82e-05 \*\*\*  
## education11th -0.127287 0.029319 -4.341 1.44e-05 \*\*\*  
## education12th -0.064510 0.036060 -1.789 0.073681 .   
## education1st-4th 0.048841 0.100991 0.484 0.628681   
## education5th-6th 0.052895 0.078916 0.670 0.502723   
## education7th-8th -0.020351 0.059355 -0.343 0.731714   
## education9th 0.052798 0.049963 1.057 0.290687   
## educationAssoc-acdm 0.011429 0.027488 0.416 0.677600   
## educationAssoc-voc 0.032942 0.024129 1.365 0.172249   
## educationBachelors -0.014591 0.020243 -0.721 0.471074   
## educationDoctorate -0.013854 0.059109 -0.234 0.814694   
## educationHS-grad 0.028622 0.012416 2.305 0.021197 \*   
## educationMasters -0.053214 0.030461 -1.747 0.080711 .   
## educationPreschool -0.212411 0.144883 -1.466 0.142689   
## educationProf-school NA NA NA NA   
## educationSome-college NA NA NA NA   
## workclassLocal-gov -0.004284 0.031412 -0.136 0.891531   
## workclassPrivate -0.014514 0.027153 -0.535 0.592995   
## workclassSelf-emp-inc 0.087702 0.035260 2.487 0.012906 \*   
## workclassSelf-emp-not-inc 0.016622 0.031362 0.530 0.596143   
## workclassState-gov -0.084535 0.033974 -2.488 0.012873 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3042 on 4853 degrees of freedom  
## Multiple R-squared: 0.2343, Adjusted R-squared: 0.227   
## F-statistic: 32.28 on 46 and 4853 DF, p-value: < 2.2e-16

Veiem que tenim NAs a dues categories de education, llavors decidim agafar la variable factoritzada.

m6<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+poly**(capital\_loss,2)**+poly**(capital\_gain,2)**+**relationship**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**summary**(m6)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + poly(education\_num,   
## 2) + poly(capital\_loss, 2) + poly(capital\_gain, 2) + relationship +   
## occupation + marital\_status + sex + over\_50k + f.education +   
## workclass, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.90932 -0.09515 0.03025 0.16372 1.31078   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.4734178 0.0703046 49.405 < 2e-16 \*\*\*  
## poly(age, 2)1 -1.3421647 0.4046053 -3.317 0.000916 \*\*\*  
## poly(age, 2)2 -6.4125894 0.3537244 -18.129 < 2e-16 \*\*\*  
## poly(education\_num, 2)1 0.2673458 1.7397194 0.154 0.877875   
## poly(education\_num, 2)2 1.6490560 0.7620843 2.164 0.030522 \*   
## poly(capital\_loss, 2)1 0.3648616 0.3101131 1.177 0.239435   
## poly(capital\_loss, 2)2 0.0362654 0.3032047 0.120 0.904800   
## poly(capital\_gain, 2)1 -0.1877432 0.3206998 -0.585 0.558294   
## poly(capital\_gain, 2)2 -0.1933304 0.3080687 -0.628 0.530324   
## relationshipNot-in-family 0.0781965 0.0506072 1.545 0.122371   
## relationshipOther-relative 0.0155064 0.0488421 0.317 0.750893   
## relationshipOwn-child -0.0830449 0.0508089 -1.634 0.102228   
## relationshipUnmarried 0.0455200 0.0526400 0.865 0.387223   
## relationshipWife -0.1106685 0.0241239 -4.588 4.60e-06 \*\*\*  
## occupationCraft-repair 0.0246775 0.0185380 1.331 0.183190   
## occupationExec-managerial 0.0702011 0.0182888 3.838 0.000125 \*\*\*  
## occupationFarming-fishing 0.0844750 0.0303923 2.779 0.005465 \*\*   
## occupationHandlers-cleaners -0.0075251 0.0256475 -0.293 0.769225   
## occupationMachine-op-inspct 0.0674838 0.0221517 3.046 0.002328 \*\*   
## occupationOther-service -0.0749303 0.0182043 -4.116 3.92e-05 \*\*\*  
## occupationPriv-house-serv -0.0846335 0.0556055 -1.522 0.128065   
## occupationProf-specialty -0.0180071 0.0191951 -0.938 0.348236   
## occupationProtective-serv 0.0487103 0.0359323 1.356 0.175286   
## occupationSales 0.0230418 0.0180750 1.275 0.202446   
## occupationTech-support -0.0427110 0.0285460 -1.496 0.134662   
## occupationTransport-moving 0.0962260 0.0246131 3.910 9.37e-05 \*\*\*  
## marital\_statusMarried-AF-spouse -0.5187435 0.2142925 -2.421 0.015526 \*   
## marital\_statusMarried-civ-spouse 0.0469247 0.0508074 0.924 0.355750   
## marital\_statusMarried-spouse-absent -0.0233171 0.0390076 -0.598 0.550029   
## marital\_statusNever-married -0.0140197 0.0167110 -0.839 0.401540   
## marital\_statusSeparated -0.0218209 0.0265322 -0.822 0.410872   
## marital\_statusWidowed -0.1104172 0.0289438 -3.815 0.000138 \*\*\*  
## sexMale 0.0686945 0.0129095 5.321 1.08e-07 \*\*\*  
## over\_50k>50K 0.0658630 0.0131020 5.027 5.16e-07 \*\*\*  
## f.educationAssociated 0.1197295 0.0496806 2.410 0.015990 \*   
## f.educationHS-graduated 0.1103417 0.0265149 4.162 3.22e-05 \*\*\*  
## f.educationPost-Bachelors 0.0642299 0.0896426 0.717 0.473710   
## f.educationSome-college 0.0926434 0.0340326 2.722 0.006508 \*\*   
## f.educationProf-school 0.1112193 0.1033800 1.076 0.282057   
## f.educationBachelors 0.0914990 0.0671448 1.363 0.173037   
## workclassLocal-gov -0.0009183 0.0310757 -0.030 0.976426   
## workclassPrivate -0.0148896 0.0268759 -0.554 0.579597   
## workclassSelf-emp-inc 0.0906550 0.0348888 2.598 0.009394 \*\*   
## workclassSelf-emp-not-inc 0.0205636 0.0310324 0.663 0.507587   
## workclassState-gov -0.0780980 0.0336311 -2.322 0.020263 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.301 on 4855 degrees of freedom  
## Multiple R-squared: 0.2497, Adjusted R-squared: 0.2429   
## F-statistic: 36.71 on 44 and 4855 DF, p-value: < 2.2e-16

Ara veiem que ja no tenim cap variable amb NAs.

**Anova**(m6)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## poly(age, 2) 35.22 2 194.3307 < 2.2e-16 \*\*\*  
## poly(education\_num, 2) 0.83 2 4.5632 0.010473 \*   
## poly(capital\_loss, 2) 0.13 2 0.6989 0.497203   
## poly(capital\_gain, 2) 0.07 2 0.3815 0.682877   
## relationship 11.34 5 25.0165 < 2.2e-16 \*\*\*  
## occupation 10.16 12 9.3441 < 2.2e-16 \*\*\*  
## marital\_status 2.00 6 3.6757 0.001204 \*\*   
## sex 2.57 1 28.3153 1.077e-07 \*\*\*  
## over\_50k 2.29 1 25.2703 5.162e-07 \*\*\*  
## f.education 2.76 6 5.0766 3.332e-05 \*\*\*  
## workclass 3.01 5 6.6407 3.594e-06 \*\*\*  
## Residuals 440.01 4855   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

A partir d’executar Anova(m6) podem veure quines són les variables significatives llavors agafem el model. Hem decidit quedarnos amb la variable capital\_gain de moment fins a veure la hauríem de considerar com a numèrica o com a factor depenent de com de explicativa sigui.

Ara mirarem si les variables age, capital\_loss i capital\_gain ens expliquen més com a numèriques o com a factors. Comencem amb capital\_loss

mAux<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+poly**(capital\_gain,2)**+**f.capital\_loss**+**relationship**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**BIC**(m6,mAux)

## df BIC  
## m6 46 2486.520  
## mAux 45 2478.262

El BIC més petit és la opció recomanable. Per tant podem veure que la variable capital\_loss explica més com a factor que com a numèrica.

Seguim amb capital\_gain

mAux2<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+**f.capital\_loss**+**f.capital\_gain**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**BIC**(m6,mAux2)

## df BIC  
## m6 46 2486.520  
## mAux2 39 2550.691

Podem veure que la variable capital\_gain no explica més com a factor que com a numèrica.

Seguim amb age

mAux3<-**lm**(**log**(hours\_per\_week)**~**f.age**+poly**(education\_num,2)**+**f.capital\_loss**+**f.capital\_gain**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**BIC**(m6,mAux3)

## df BIC  
## m6 46 2486.52  
## mAux3 40 2829.12

Podem veure que la variable age ens és més explicativa com a variable numèrica que com a factor.

**Anova**(mAux2)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## poly(age, 2) 42.68 2 229.9427 < 2.2e-16 \*\*\*  
## poly(education\_num, 2) 1.12 2 6.0484 0.002379 \*\*   
## f.capital\_loss 0.05 1 0.5024 0.478489   
## f.capital\_gain 0.18 1 1.9127 0.166726   
## occupation 10.61 12 9.5264 < 2.2e-16 \*\*\*  
## marital\_status 1.84 6 3.3066 0.002997 \*\*   
## sex 6.67 1 71.8888 < 2.2e-16 \*\*\*  
## over\_50k 2.10 1 22.5891 2.064e-06 \*\*\*  
## f.education 3.18 6 5.7104 6.276e-06 \*\*\*  
## workclass 3.01 5 6.4801 5.175e-06 \*\*\*  
## Residuals 451.26 4862   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Llavors considerem que la millor solució seria el model mAux2. Pero després de fer més probes, com l’Anova d’aquest model, podem veure que tant capìtal\_loss i capital\_gain tenen un pvalor > 0.05, ja sigui en forma de factor tant numèrica, llavors hem decidit treure aquestes dues variables del model, encara que ens podriem haver quedat f.capital\_gain com a significativa perquè hi ha un espai.

Aquest seria el nostre model de moment:

m7<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**summary**(m7)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + poly(education\_num,   
## 2) + occupation + marital\_status + sex + over\_50k + f.education +   
## workclass, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.87264 -0.09368 0.03613 0.16182 1.33835   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.484550 0.047345 73.599 < 2e-16 \*\*\*  
## poly(age, 2)1 -0.202400 0.392946 -0.515 0.606519   
## poly(age, 2)2 -7.259128 0.344752 -21.056 < 2e-16 \*\*\*  
## poly(education\_num, 2)1 0.035199 1.756952 0.020 0.984017   
## poly(education\_num, 2)2 1.825239 0.767294 2.379 0.017407 \*   
## occupationCraft-repair 0.037364 0.018713 1.997 0.045917 \*   
## occupationExec-managerial 0.080560 0.018470 4.362 1.32e-05 \*\*\*  
## occupationFarming-fishing 0.085144 0.030659 2.777 0.005505 \*\*   
## occupationHandlers-cleaners -0.009731 0.025948 -0.375 0.707656   
## occupationMachine-op-inspct 0.075510 0.022392 3.372 0.000752 \*\*\*  
## occupationOther-service -0.071813 0.018393 -3.904 9.57e-05 \*\*\*  
## occupationPriv-house-serv -0.063290 0.056033 -1.130 0.258737   
## occupationProf-specialty -0.004485 0.019376 -0.231 0.816948   
## occupationProtective-serv 0.060129 0.036322 1.655 0.097900 .   
## occupationSales 0.029010 0.018254 1.589 0.112074   
## occupationTech-support -0.027610 0.028836 -0.958 0.338360   
## occupationTransport-moving 0.105480 0.024885 4.239 2.29e-05 \*\*\*  
## marital\_statusMarried-AF-spouse -0.570123 0.216298 -2.636 0.008420 \*\*   
## marital\_statusMarried-civ-spouse -0.027419 0.015076 -1.819 0.069016 .   
## marital\_statusMarried-spouse-absent -0.020954 0.039410 -0.532 0.594952   
## marital\_statusNever-married -0.038244 0.016389 -2.333 0.019665 \*   
## marital\_statusSeparated -0.023405 0.026822 -0.873 0.382918   
## marital\_statusWidowed -0.098218 0.029222 -3.361 0.000782 \*\*\*  
## sexMale 0.095692 0.011288 8.477 < 2e-16 \*\*\*  
## over\_50k>50K 0.059039 0.012636 4.672 3.06e-06 \*\*\*  
## f.educationAssociated 0.140771 0.050175 2.806 0.005042 \*\*   
## f.educationHS-graduated 0.120938 0.026788 4.515 6.49e-06 \*\*\*  
## f.educationPost-Bachelors 0.085426 0.090577 0.943 0.345660   
## f.educationSome-college 0.104364 0.034381 3.036 0.002414 \*\*   
## f.educationProf-school 0.132480 0.104465 1.268 0.204797   
## f.educationBachelors 0.113442 0.067845 1.672 0.094572 .   
## workclassLocal-gov -0.003179 0.031431 -0.101 0.919452   
## workclassPrivate -0.013725 0.027189 -0.505 0.613738   
## workclassSelf-emp-inc 0.090107 0.035229 2.558 0.010566 \*   
## workclassSelf-emp-not-inc 0.018264 0.031392 0.582 0.560727   
## workclassState-gov -0.082011 0.034019 -2.411 0.015956 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3047 on 4864 degrees of freedom  
## Multiple R-squared: 0.2301, Adjusted R-squared: 0.2245   
## F-statistic: 41.52 on 35 and 4864 DF, p-value: < 2.2e-16

**Anova**(m7)

## Anova Table (Type II tests)  
##   
## Response: log(hours\_per\_week)  
## Sum Sq Df F value Pr(>F)   
## poly(age, 2) 42.98 2 231.5036 < 2.2e-16 \*\*\*  
## poly(education\_num, 2) 1.13 2 6.0751 0.002317 \*\*   
## occupation 10.61 12 9.5235 < 2.2e-16 \*\*\*  
## marital\_status 1.83 6 3.2837 0.003169 \*\*   
## sex 6.67 1 71.8621 < 2.2e-16 \*\*\*  
## over\_50k 2.03 1 21.8285 3.062e-06 \*\*\*  
## f.education 3.17 6 5.6919 6.592e-06 \*\*\*  
## workclass 3.02 5 6.5115 4.818e-06 \*\*\*  
## Residuals 451.50 4864   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## Interactions between variables

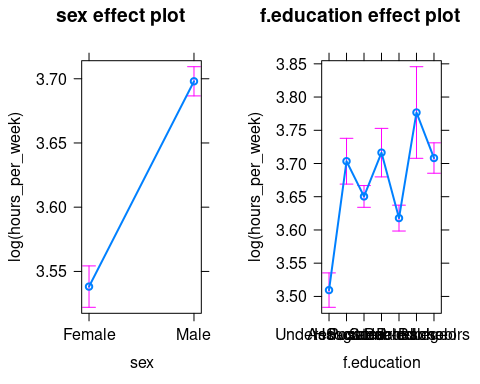
m8<-**lm**(**log**(hours\_per\_week)**~**sex**+**f.education,data=df)  
**summary**(m8)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ sex + f.education, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.97781 -0.06778 0.03969 0.16755 0.84565   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.40285 0.01483 229.402 < 2e-16 \*\*\*  
## sexMale 0.15982 0.01008 15.861 < 2e-16 \*\*\*  
## f.educationAssociated 0.19400 0.02199 8.822 < 2e-16 \*\*\*  
## f.educationHS-graduated 0.14098 0.01561 9.030 < 2e-16 \*\*\*  
## f.educationPost-Bachelors 0.20691 0.02282 9.067 < 2e-16 \*\*\*  
## f.educationSome-college 0.10836 0.01651 6.563 5.82e-11 \*\*\*  
## f.educationProf-school 0.26727 0.03756 7.116 1.27e-12 \*\*\*  
## f.educationBachelors 0.19874 0.01761 11.288 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3313 on 4892 degrees of freedom  
## Multiple R-squared: 0.08414, Adjusted R-squared: 0.08283   
## F-statistic: 64.2 on 7 and 4892 DF, p-value: < 2.2e-16

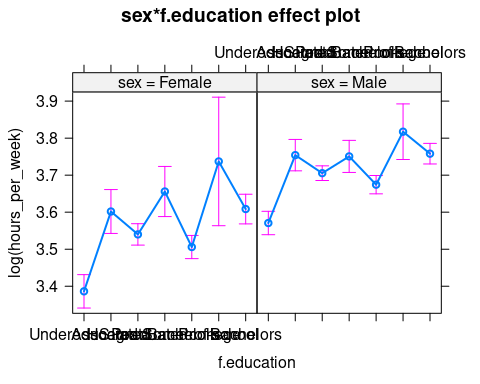
m9<-**lm**(**log**(hours\_per\_week)**~**sex**\***f.education,data=df)  
**summary**(m9)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ sex \* f.education, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.96234 -0.06522 0.03295 0.16578 0.86221   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.38629 0.02303 147.017 < 2e-16 \*\*\*  
## sexMale 0.18449 0.02811 6.563 5.82e-11 \*\*\*  
## f.educationAssociated 0.21559 0.03802 5.670 1.51e-08 \*\*\*  
## f.educationHS-graduated 0.15376 0.02737 5.619 2.03e-08 \*\*\*  
## f.educationPost-Bachelors 0.26965 0.04152 6.494 9.20e-11 \*\*\*  
## f.educationSome-college 0.11965 0.02801 4.271 1.98e-05 \*\*\*  
## f.educationProf-school 0.35081 0.09151 3.833 0.000128 \*\*\*  
## f.educationBachelors 0.22225 0.03079 7.218 6.09e-13 \*\*\*  
## sexMale:f.educationAssociated -0.03227 0.04661 -0.692 0.488835   
## sexMale:f.educationHS-graduated -0.01911 0.03332 -0.573 0.566374   
## sexMale:f.educationPost-Bachelors -0.08973 0.04972 -1.805 0.071184 .   
## sexMale:f.educationSome-college -0.01608 0.03470 -0.463 0.643178   
## sexMale:f.educationProf-school -0.10415 0.10049 -1.036 0.300057   
## sexMale:f.educationBachelors -0.03495 0.03753 -0.931 0.351815   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3314 on 4886 degrees of freedom  
## Multiple R-squared: 0.08497, Adjusted R-squared: 0.08254   
## F-statistic: 34.9 on 13 and 4886 DF, p-value: < 2.2e-16

**library**(effects)  
**plot**(**allEffects**(m8))



**plot**(**allEffects**(m9))



**anova**(m8,m9)

## Analysis of Variance Table  
##   
## Model 1: log(hours\_per\_week) ~ sex + f.education  
## Model 2: log(hours\_per\_week) ~ sex \* f.education  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 4892 537.07   
## 2 4886 536.58 6 0.48904 0.7422 0.6156

Podem veure una interacció entre el factor f.education i sex, observant els gràfics. Veiem que els homes tenen un nivell més alt d’educació i a més treballen més hores. També el que hem pogut comprovar és si les nostres interaccions són rellevants i amb la comanda “anova” fem com una comparació per veure els dos models que tenim i poder treure com a conclusió que haurem d’acceptar la hipòtesi nula, perquè el pvalor que surt és més gran que 0.05 (5%).

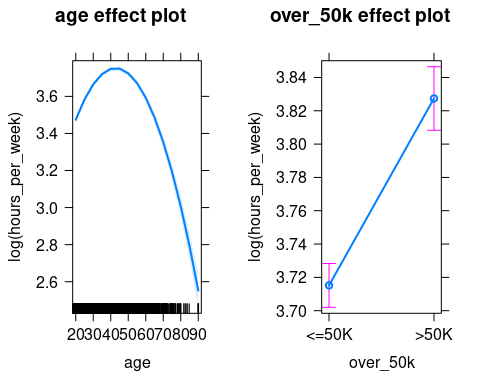
m10<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+**over\_50k,data=df)  
**summary**(m10)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + over\_50k, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.77963 -0.08122 0.02529 0.17302 1.38664   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.61831 0.00524 690.467 <2e-16 \*\*\*  
## poly(age, 2)1 0.66955 0.32687 2.048 0.0406 \*   
## poly(age, 2)2 -8.48514 0.32329 -26.247 <2e-16 \*\*\*  
## over\_50k>50K 0.11223 0.01121 10.007 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3165 on 4896 degrees of freedom  
## Multiple R-squared: 0.1634, Adjusted R-squared: 0.1629   
## F-statistic: 318.8 on 3 and 4896 DF, p-value: < 2.2e-16

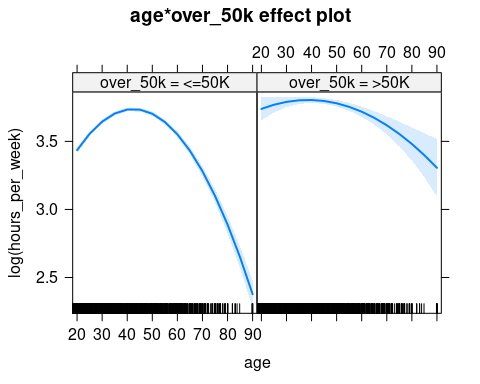
m11<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**\***over\_50k,data=df)  
**summary**(m11)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) \* over\_50k, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.79159 -0.08603 0.01644 0.17773 1.53474   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.619257 0.005222 693.102 < 2e-16 \*\*\*  
## poly(age, 2)1 0.360369 0.353658 1.019 0.308   
## poly(age, 2)2 -9.476818 0.354776 -26.712 < 2e-16 \*\*\*  
## over\_50k>50K 0.151058 0.014374 10.509 < 2e-16 \*\*\*  
## poly(age, 2)1:over\_50k>50K -1.589415 1.059278 -1.500 0.134   
## poly(age, 2)2:over\_50k>50K 6.495260 0.985319 6.592 4.8e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3151 on 4894 degrees of freedom  
## Multiple R-squared: 0.1714, Adjusted R-squared: 0.1705   
## F-statistic: 202.5 on 5 and 4894 DF, p-value: < 2.2e-16

**library**(effects)  
**library**(car)  
**plot**(**allEffects**(m10))



**plot**(**allEffects**(m11))



**anova**(m10,m11)

## Analysis of Variance Table  
##   
## Model 1: log(hours\_per\_week) ~ poly(age, 2) + over\_50k  
## Model 2: log(hours\_per\_week) ~ poly(age, 2) \* over\_50k  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 4896 490.59   
## 2 4894 485.90 2 4.6854 23.596 6.333e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Podem veure una interacció entre la variable covariant age i el factor over\_50k, observant els gràfics. Podem veure que a les edats on més hores es treballa (30-50) és on la gent guanya >50k. En aquest cas, fent la comanda anova veiem que hem de rebutjar l’hipotesi nula perquè el pvalor que surt és més petit que 0.05.

## Diagnostic of the model

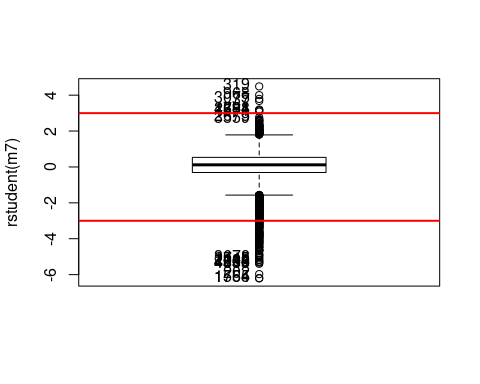
### Residual analysis

El nostre model triat és m7. Ara farem un anàlisi dels residus i després farem una predicció.

**Boxplot**(**rstudent**(m7), id.n=2)

## [1] 1565 1754 202 4835 4008 1224 2143 1745 2418 3678 319 968 3075 37 2391  
## [16] 4684 1768 1571 2675 3559

**abline**(h=**c**(3,**-**3),col="red",lwd=2)

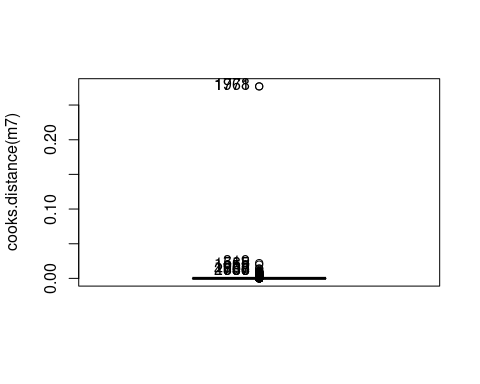


out <- **which**(**rstudent**(m7) **>=** 3 **|** **rstudent**(m7) **<=** -3);**length**(out)

## [1] 106

A partir de l’anàlisi de residus veiem que no hi han quasi possibles outliers.

infl<-**Boxplot**(**cooks.distance**(m7), id.n=4)



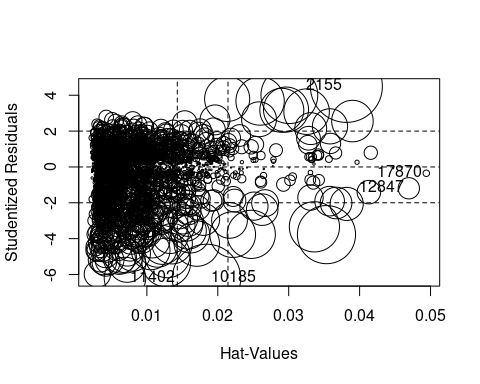
out2 <- **which**(**cooks.distance**(m7) **>=** 0.20);**length**(out2)

## [1] 2

df<-df[**-**out2,]  
  
m7<-**lm**(**log**(hours\_per\_week)**~poly**(age,2)**+poly**(education\_num,2)**+**occupation**+**marital\_status**+**sex**+**over\_50k**+**f.education**+**workclass,data=df)  
**summary**(m7)

##   
## Call:  
## lm(formula = log(hours\_per\_week) ~ poly(age, 2) + poly(education\_num,   
## 2) + occupation + marital\_status + sex + over\_50k + f.education +   
## workclass, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.87276 -0.09309 0.03623 0.16210 1.33874   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.483709 0.047307 73.641 < 2e-16 \*\*\*  
## poly(age, 2)1 -0.196362 0.392571 -0.500 0.616961   
## poly(age, 2)2 -7.264505 0.344434 -21.091 < 2e-16 \*\*\*  
## poly(education\_num, 2)1 0.051503 1.755229 0.029 0.976593   
## poly(education\_num, 2)2 1.829825 0.766541 2.387 0.017019 \*   
## occupationCraft-repair 0.038475 0.018699 2.058 0.039687 \*   
## occupationExec-managerial 0.081680 0.018457 4.426 9.83e-06 \*\*\*  
## occupationFarming-fishing 0.086426 0.030634 2.821 0.004803 \*\*   
## occupationHandlers-cleaners -0.008610 0.025927 -0.332 0.739829   
## occupationMachine-op-inspct 0.076550 0.022374 3.421 0.000628 \*\*\*  
## occupationOther-service -0.069424 0.018391 -3.775 0.000162 \*\*\*  
## occupationPriv-house-serv -0.062356 0.055982 -1.114 0.265395   
## occupationProf-specialty -0.003267 0.019362 -0.169 0.865997   
## occupationProtective-serv 0.062044 0.036294 1.709 0.087423 .   
## occupationSales 0.030000 0.018240 1.645 0.100089   
## occupationTech-support -0.026528 0.028811 -0.921 0.357229   
## occupationTransport-moving 0.106676 0.024865 4.290 1.82e-05 \*\*\*  
## marital\_statusMarried-civ-spouse -0.027331 0.015062 -1.815 0.069660 .   
## marital\_statusMarried-spouse-absent -0.020892 0.039374 -0.531 0.595722   
## marital\_statusNever-married -0.038251 0.016374 -2.336 0.019528 \*   
## marital\_statusSeparated -0.023404 0.026798 -0.873 0.382507   
## marital\_statusWidowed -0.098320 0.029195 -3.368 0.000764 \*\*\*  
## sexMale 0.095474 0.011278 8.465 < 2e-16 \*\*\*  
## over\_50k>50K 0.058947 0.012625 4.669 3.11e-06 \*\*\*  
## f.educationAssociated 0.140697 0.050129 2.807 0.005025 \*\*   
## f.educationHS-graduated 0.121021 0.026764 4.522 6.28e-06 \*\*\*  
## f.educationPost-Bachelors 0.085159 0.090494 0.941 0.346726   
## f.educationSome-college 0.104550 0.034349 3.044 0.002349 \*\*   
## f.educationProf-school 0.131826 0.104370 1.263 0.206626   
## f.educationBachelors 0.113241 0.067783 1.671 0.094855 .   
## workclassLocal-gov -0.003571 0.031402 -0.114 0.909470   
## workclassPrivate -0.013684 0.027164 -0.504 0.614461   
## workclassSelf-emp-inc 0.089953 0.035197 2.556 0.010628 \*   
## workclassSelf-emp-not-inc 0.017966 0.031363 0.573 0.566787   
## workclassState-gov -0.085536 0.034006 -2.515 0.011924 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3044 on 4863 degrees of freedom  
## Multiple R-squared: 0.2292, Adjusted R-squared: 0.2239   
## F-statistic: 42.54 on 34 and 4863 DF, p-value: < 2.2e-16

**influencePlot**(m7,id.n=3)



## StudRes Hat CookD  
## 2155 4.4932020 0.038135147 0.022779537  
## 10185 -6.2175925 0.018410979 0.020557650  
## 11402 -6.1974945 0.006911953 0.007579652  
## 12847 -1.1906289 0.046939523 0.001994645  
## 17870 -0.3526905 0.049404203 0.000184742

Decidim treure les dues observacions ja que les considerem outliers.

A partir del segon gràfic podem observar a priori que les dades més influents són les “10185” i “11402” observant el leverage que hi ha en el plot corresponent.

### 

# over\_50k model construction

Dividim la mostra en test i work.

sample<-**sample**(1**:nrow**(df),0.8**\*nrow**(df))  
dfw<-df[sample,]  
dft<-df[**-**sample,]

## Explicative Variables for modeling purposes

Ara el que farem serà analitzar quines són les variables numèriques més relacionades amb el nostre target hours\_per\_week, per tal de decidir quines d’aquestes utilitzarem en la construcció dels diferents models fins trobar l’òptim.

**names**(dfw)

## [1] "age" "workclass" "final\_weight" "education"   
## [5] "education\_num" "marital\_status" "occupation" "relationship"   
## [9] "race" "sex" "capital\_gain" "capital\_loss"   
## [13] "hours\_per\_week" "native\_country" "over\_50k" "f.workclass"   
## [17] "f.education" "f.education\_num" "f.marital\_status" "f.occupation"   
## [21] "f.relationship" "f.race" "f.sex" "f.native\_country"  
## [25] "f.over\_50k" "f.age" "f.capital\_gain" "f.capital\_loss"   
## [29] "f.hours\_per\_week"

vars\_num2<-**names**(dfw)[**c**(1,5,11**:**13,15)]; vars\_num2

## [1] "age" "education\_num" "capital\_gain" "capital\_loss"   
## [5] "hours\_per\_week" "over\_50k"

*#catdes(dfw[,vars\_num2],which(vars\_num2 == "over\_50k"))*

Veiem que les variables numèriques més relacionades són education\_num, age, capital\_gain, hours\_per\_week i capital\_loss, tot i que en aquesta última la correlació que presenta és bastant baixa i poc significativa. Tot i així la considerarem també com a candidata a formar part de la construcció del model.

## Initial modeling

gm1<-**glm**(over\_50k**~**age**+**education\_num**+**capital\_gain**+**hours\_per\_week**+**capital\_loss,family=binomial,data=dfw)  
**summary**(gm1)

##   
## Call:  
## glm(formula = over\_50k ~ age + education\_num + capital\_gain +   
## hours\_per\_week + capital\_loss, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.6694 -0.6296 -0.3961 -0.1236 2.7198   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -9.006e+00 3.556e-01 -25.323 < 2e-16 \*\*\*  
## age 4.746e-02 3.710e-03 12.794 < 2e-16 \*\*\*  
## education\_num 3.337e-01 1.992e-02 16.750 < 2e-16 \*\*\*  
## capital\_gain 3.131e-04 2.709e-05 11.557 < 2e-16 \*\*\*  
## hours\_per\_week 4.945e-02 4.488e-03 11.017 < 2e-16 \*\*\*  
## capital\_loss 7.499e-04 9.767e-05 7.678 1.62e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 3128.5 on 3912 degrees of freedom  
## AIC: 3140.5  
##   
## Number of Fisher Scoring iterations: 5

**Anova**(gm1)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## age 171.37 1 < 2.2e-16 \*\*\*  
## education\_num 329.69 1 < 2.2e-16 \*\*\*  
## capital\_gain 236.93 1 < 2.2e-16 \*\*\*  
## hours\_per\_week 131.40 1 < 2.2e-16 \*\*\*  
## capital\_loss 60.76 1 6.457e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

No veiem cap variable que tingui un p-valor per sota del 5%, llavors ens quedem amb el mateix model.

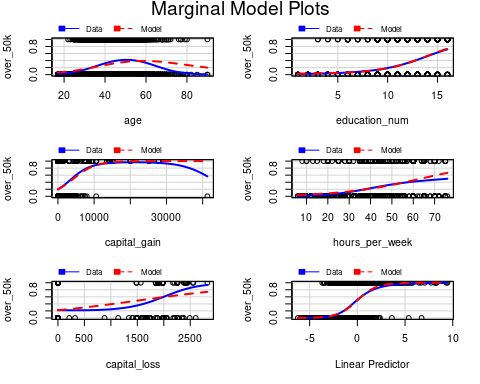
Realitzem un vif (variance inflation factor) per veure les variables explicatives del model que estàn correlacionades.

**vif**(gm1)

## age education\_num capital\_gain hours\_per\_week capital\_loss   
## 1.029500 1.014301 1.009320 1.020419 1.005188

No veiem cap variable que tingui un vif > 3 (el que significaria un problema de colinealitat).

**marginalModelPlots**(gm1)



Hem aconseguit una discrepancia baixa amb el nostre model (Residual deviance < Null deviance) i un altre indicador de que anem bé és que la Residual deviance és igual o inferior als graus de llibertat (3128.5 < 3912).

## Transforming variables

A partir del marginalPlots podem veure on hi ha un desajust entre les observacions i la predicció en algunes variables. El que farem a continuació és trobar la manera d’arreglar-ho transformant-les.

gm2<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+**capital\_gain**+**hours\_per\_week**+**capital\_loss,family=binomial,data=dfw)  
**summary**(gm2)

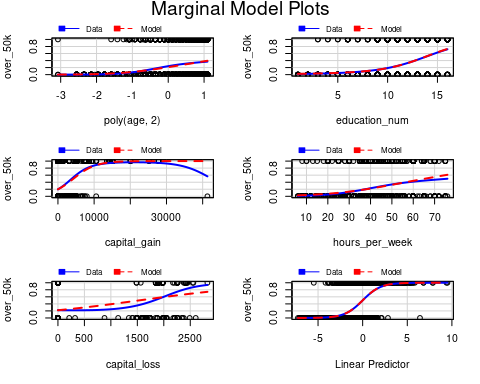
##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + capital\_gain +   
## hours\_per\_week + capital\_loss, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5845 -0.6422 -0.3546 -0.0648 3.3115   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.870e+00 3.075e-01 -22.338 < 2e-16 \*\*\*  
## poly(age, 2)1 5.127e+01 3.763e+00 13.624 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.945e+01 4.083e+00 -9.662 < 2e-16 \*\*\*  
## education\_num 3.215e-01 2.024e-02 15.884 < 2e-16 \*\*\*  
## capital\_gain 3.095e-04 2.700e-05 11.464 < 2e-16 \*\*\*  
## hours\_per\_week 4.017e-02 4.732e-03 8.488 < 2e-16 \*\*\*  
## capital\_loss 7.702e-04 1.010e-04 7.625 2.45e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 3015.5 on 3911 degrees of freedom  
## AIC: 3029.5  
##   
## Number of Fisher Scoring iterations: 6

**Anova**(gm2)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 284.408 2 < 2.2e-16 \*\*\*  
## education\_num 294.550 1 < 2.2e-16 \*\*\*  
## capital\_gain 236.379 1 < 2.2e-16 \*\*\*  
## hours\_per\_week 75.268 1 < 2.2e-16 \*\*\*  
## capital\_loss 60.296 1 8.161e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**marginalModelPlots**(gm2)

## combination



Aplicant el quadràtic a la variable age veiem que hi ha una gran millora.

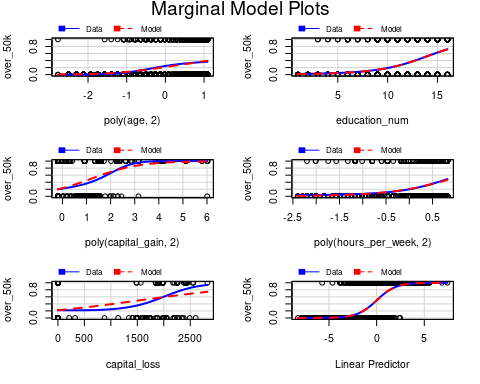
gm3<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+poly**(hours\_per\_week,2)**+**capital\_loss,family=binomial,data=dfw)  
**summary**(gm3)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + poly(hours\_per\_week, 2) + capital\_loss, family = binomial,   
## data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2587 -0.6476 -0.3403 -0.0570 3.3832   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.107e+00 2.322e-01 -21.994 < 2e-16 \*\*\*  
## poly(age, 2)1 5.146e+01 3.769e+00 13.652 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.764e+01 4.132e+00 -9.109 < 2e-16 \*\*\*  
## education\_num 3.215e-01 2.028e-02 15.851 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)1 4.526e+01 3.804e+00 11.899 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)2 -8.245e+00 2.523e+00 -3.269 0.00108 \*\*   
## poly(hours\_per\_week, 2)1 3.402e+01 4.203e+00 8.095 5.74e-16 \*\*\*  
## poly(hours\_per\_week, 2)2 -1.032e+01 3.711e+00 -2.780 0.00544 \*\*   
## capital\_loss 7.812e-04 1.015e-04 7.695 1.41e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2999.6 on 3909 degrees of freedom  
## AIC: 3017.6  
##   
## Number of Fisher Scoring iterations: 6

**Anova**(gm3)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 275.612 2 < 2.2e-16 \*\*\*  
## education\_num 293.400 1 < 2.2e-16 \*\*\*  
## poly(capital\_gain, 2) 242.842 2 < 2.2e-16 \*\*\*  
## poly(hours\_per\_week, 2) 83.655 2 < 2.2e-16 \*\*\*  
## capital\_loss 61.729 1 3.942e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**marginalModelPlots**(gm3)



Aqui, aplicant el quadràtic a capital\_gain veiem una millora en el desajust encara que no tan gran com amb age.

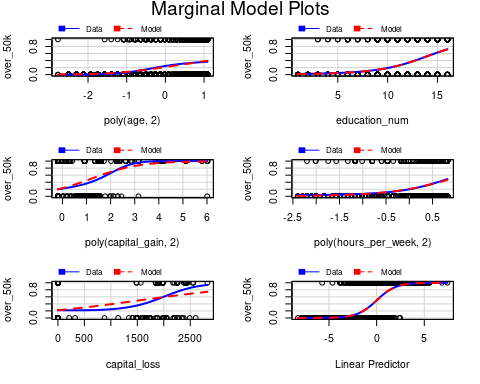
gm4<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+poly**(hours\_per\_week,2)**+**capital\_loss,family=binomial,data=dfw)  
**summary**(gm4)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + poly(hours\_per\_week, 2) + capital\_loss, family = binomial,   
## data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2587 -0.6476 -0.3403 -0.0570 3.3832   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -5.107e+00 2.322e-01 -21.994 < 2e-16 \*\*\*  
## poly(age, 2)1 5.146e+01 3.769e+00 13.652 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.764e+01 4.132e+00 -9.109 < 2e-16 \*\*\*  
## education\_num 3.215e-01 2.028e-02 15.851 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)1 4.526e+01 3.804e+00 11.899 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)2 -8.245e+00 2.523e+00 -3.269 0.00108 \*\*   
## poly(hours\_per\_week, 2)1 3.402e+01 4.203e+00 8.095 5.74e-16 \*\*\*  
## poly(hours\_per\_week, 2)2 -1.032e+01 3.711e+00 -2.780 0.00544 \*\*   
## capital\_loss 7.812e-04 1.015e-04 7.695 1.41e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2999.6 on 3909 degrees of freedom  
## AIC: 3017.6  
##   
## Number of Fisher Scoring iterations: 6

**Anova**(gm4)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 275.612 2 < 2.2e-16 \*\*\*  
## education\_num 293.400 1 < 2.2e-16 \*\*\*  
## poly(capital\_gain, 2) 242.842 2 < 2.2e-16 \*\*\*  
## poly(hours\_per\_week, 2) 83.655 2 < 2.2e-16 \*\*\*  
## capital\_loss 61.729 1 3.942e-15 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**marginalModelPlots**(gm4)



Després d’aplicar el qudràtic a la variable hours\_per\_Week, creiem que ja tenim un bon model. Llavors el model que ens quedarem serà el gm4 que és el que té menor desajust entre les observacions i les prediccions fetes.

## Adding factors as explanatory variables

Primer el que farem serà mirar si les variables numèriques que tenim ara mateix en el nostre model gm4 són més explicatives com a variables numèriques o com a factors. Utilitzarem la comanda BIC. Començem per la variable age.

gmAux<-**glm**(over\_50k**~**f.age**+**education\_num**+poly**(capital\_gain,2)**+poly**(hours\_per\_week,2)**+**capital\_loss,family=binomial,data=dfw)  
**BIC**(gm4,gmAux)

## df BIC  
## gm4 9 3074.031  
## gmAux 10 3103.253

Podem veure que el model amb la variable age com a factor en comptes de com a numèrica té un BIC més gran i per tant és menys explicativa. Ens quedem amb age com a numèrica. Contiuem amb capital\_gain.

gmAux<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+**f.capital\_gain**+poly**(hours\_per\_week,2)**+**capital\_loss,family=binomial,data=dfw)  
**BIC**(gm4,gmAux)

## df BIC  
## gm4 9 3074.031  
## gmAux 8 3151.647

Podem veure que el model amb la variable capital\_gain com a factor en comptes de com a numèrica té un BIC més gran i per tant és menys explicativa. Ens quedem amb capital\_gain com a numèrica. Continuem amb hours\_per\_week.

gmAux<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss,family=binomial,data=dfw)  
**BIC**(gm4,gmAux)

## df BIC  
## gm4 9 3074.031  
## gmAux 9 3069.369

Podem veure que el model amb la variable hours\_per\_week com a factor en comptes de com a numèrica té un BIC més petit i per tant és més explicativa. Ens quedem amb hours\_per\_week com a factor. Continuem amb capital\_loss.

gmAux<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**f.capital\_loss,family=binomial,data=dfw)  
**BIC**(gm4,gmAux)

## df BIC  
## gm4 9 3074.031  
## gmAux 9 3088.784

Podem veure que el model amb la variable capital\_loss com a factor en comptes de com a numèrica té un BIC més gran i per tant és menys explicativa. Ens quedem amb capital\_loss com a numèrica.

Llavors definim el nostre model que més ens explica fins ara, agafant hours\_per\_week com a factor en comptes de com a numèrica:

gm5<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss,family=binomial,data=dfw)

Ara el que farem és afegir les variables factors com a variables explicatives. Per decidir quines hem d’afegir hem de trobar les que són més significatives.

**names**(dfw)

## [1] "age" "workclass" "final\_weight" "education"   
## [5] "education\_num" "marital\_status" "occupation" "relationship"   
## [9] "race" "sex" "capital\_gain" "capital\_loss"   
## [13] "hours\_per\_week" "native\_country" "over\_50k" "f.workclass"   
## [17] "f.education" "f.education\_num" "f.marital\_status" "f.occupation"   
## [21] "f.relationship" "f.race" "f.sex" "f.native\_country"  
## [25] "f.over\_50k" "f.age" "f.capital\_gain" "f.capital\_loss"   
## [29] "f.hours\_per\_week"

vars\_dis2<-**names**(dfw)[**c**(2,4,6,7,9,10,14)];vars\_dis2

## [1] "workclass" "education" "marital\_status" "occupation"   
## [5] "race" "sex" "native\_country"

*#catdes(dfw[,c("over\_50k",vars\_dis2)],1)*

Veiem que tenim alguns factors interesants que podem afegir al model gm5; aquestes variables són: marital\_status+education+occupation+sex+workclass+race.

gm6<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss**+**marital\_status**+**education**+**occupation**+**sex**+**workclass**+**race,family=binomial,data=dfw)

**summary**(gm6)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + f.hours\_per\_week + capital\_loss + marital\_status + education +   
## occupation + sex + workclass + race, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.2833 -0.4678 -0.1720 -0.0165 3.8749   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -4.905e+00 1.258e+00 -3.901 9.59e-05 \*\*\*  
## poly(age, 2)1 3.843e+01 4.661e+00 8.244 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.733e+01 5.129e+00 -7.279 3.36e-13 \*\*\*  
## education\_num 1.013e-01 1.019e-01 0.994 0.320137   
## poly(capital\_gain, 2)1 2.561e+02 4.151e+01 6.168 6.92e-10 \*\*\*  
## poly(capital\_gain, 2)2 1.906e+02 3.657e+01 5.211 1.88e-07 \*\*\*  
## f.hours\_per\_week[40] 6.209e-01 1.799e-01 3.452 0.000557 \*\*\*  
## f.hours\_per\_week[41-78] 1.037e+00 1.871e-01 5.543 2.97e-08 \*\*\*  
## capital\_loss 6.665e-04 1.139e-04 5.850 4.92e-09 \*\*\*  
## marital\_statusMarried-civ-spouse 2.494e+00 2.000e-01 12.468 < 2e-16 \*\*\*  
## marital\_statusMarried-spouse-absent 6.548e-01 5.776e-01 1.134 0.256997   
## marital\_statusNever-married -6.271e-02 2.499e-01 -0.251 0.801890   
## marital\_statusSeparated 1.335e-01 4.764e-01 0.280 0.779298   
## marital\_statusWidowed 4.344e-02 5.131e-01 0.085 0.932534   
## education11th -1.234e+00 6.379e-01 -1.934 0.053063 .   
## education12th -8.014e-01 6.309e-01 -1.270 0.203984   
## education1st-4th -1.391e+01 4.212e+02 -0.033 0.973653   
## education5th-6th -3.914e-01 1.072e+00 -0.365 0.715075   
## education7th-8th -1.745e+00 9.479e-01 -1.841 0.065566 .   
## education9th -1.024e+00 8.896e-01 -1.151 0.249755   
## educationAssoc-acdm -4.619e-01 3.759e-01 -1.229 0.219163   
## educationAssoc-voc 1.600e-01 2.835e-01 0.565 0.572347   
## educationBachelors 5.852e-01 3.731e-01 1.568 0.116794   
## educationDoctorate 1.085e+00 7.843e-01 1.383 0.166742   
## educationHS-grad -4.727e-02 1.613e-01 -0.293 0.769442   
## educationMasters 8.709e-01 4.976e-01 1.750 0.080099 .   
## educationPreschool -2.023e+02 4.357e+02 -0.464 0.642503   
## educationProf-school 1.010e+00 6.517e-01 1.549 0.121372   
## educationSome-college NA NA NA NA   
## occupationCraft-repair -6.474e-02 2.303e-01 -0.281 0.778655   
## occupationExec-managerial 6.669e-01 2.230e-01 2.991 0.002785 \*\*   
## occupationFarming-fishing -8.261e-01 4.094e-01 -2.018 0.043639 \*   
## occupationHandlers-cleaners -4.657e-01 4.085e-01 -1.140 0.254179   
## occupationMachine-op-inspct -7.389e-01 3.114e-01 -2.373 0.017664 \*   
## occupationOther-service -9.308e-01 3.270e-01 -2.847 0.004417 \*\*   
## occupationPriv-house-serv -1.401e+01 4.215e+02 -0.033 0.973479   
## occupationProf-specialty 4.092e-01 2.342e-01 1.748 0.080538 .   
## occupationProtective-serv 7.732e-01 3.918e-01 1.973 0.048452 \*   
## occupationSales 1.510e-01 2.445e-01 0.618 0.536852   
## occupationTech-support 6.813e-01 3.248e-01 2.098 0.035937 \*   
## occupationTransport-moving -4.494e-01 2.940e-01 -1.529 0.126289   
## sexMale -4.638e-02 1.558e-01 -0.298 0.765987   
## workclassLocal-gov -8.983e-01 3.156e-01 -2.847 0.004420 \*\*   
## workclassPrivate -8.581e-01 2.616e-01 -3.280 0.001038 \*\*   
## workclassSelf-emp-inc -3.743e-01 3.520e-01 -1.063 0.287621   
## workclassSelf-emp-not-inc -1.280e+00 3.146e-01 -4.067 4.77e-05 \*\*\*  
## workclassState-gov -1.420e+00 3.553e-01 -3.997 6.41e-05 \*\*\*  
## raceAsian-Pac-Islander 1.164e+00 7.373e-01 1.579 0.114338   
## raceBlack 8.565e-01 7.083e-01 1.209 0.226588   
## raceOther 1.512e+00 1.019e+00 1.484 0.137940   
## raceWhite 1.106e+00 6.788e-01 1.629 0.103213   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2346.5 on 3868 degrees of freedom  
## AIC: 2446.5  
##   
## Number of Fisher Scoring iterations: 15

**Anova**(gm6)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 98.51 2 < 2.2e-16 \*\*\*  
## education\_num 0   
## poly(capital\_gain, 2) 246.36 2 < 2.2e-16 \*\*\*  
## f.hours\_per\_week 34.49 2 3.234e-08 \*\*\*  
## capital\_loss 35.60 1 2.420e-09 \*\*\*  
## marital\_status 383.25 5 < 2.2e-16 \*\*\*  
## education 68.26 14 3.977e-09 \*\*\*  
## occupation 72.28 12 1.198e-10 \*\*\*  
## sex 0.09 1 0.7660   
## workclass 27.06 5 5.565e-05 \*\*\*  
## race 4.75 4 0.3141   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Veiem que tenim NAs a dues categories de education, llavors decidim agafar la variable factoritzada.

gm7<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss**+**marital\_status**+**f.education**+**occupation**+**sex**+**workclass**+**race,family=binomial,data=dfw)  
**summary**(gm7)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + f.hours\_per\_week + capital\_loss + marital\_status + f.education +   
## occupation + sex + workclass + race, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6151 -0.4781 -0.1779 -0.0231 3.6137   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.748e+00 1.008e+00 -6.693 2.19e-11 \*\*\*  
## poly(age, 2)1 3.926e+01 4.609e+00 8.518 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.568e+01 4.697e+00 -7.595 3.07e-14 \*\*\*  
## education\_num 1.930e-01 1.121e-01 1.721 0.085257 .   
## poly(capital\_gain, 2)1 4.449e+01 3.906e+00 11.390 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)2 -8.829e+00 2.645e+00 -3.338 0.000845 \*\*\*  
## f.hours\_per\_week[40] 6.489e-01 1.770e-01 3.667 0.000245 \*\*\*  
## f.hours\_per\_week[41-78] 1.069e+00 1.833e-01 5.832 5.47e-09 \*\*\*  
## capital\_loss 6.700e-04 1.129e-04 5.936 2.93e-09 \*\*\*  
## marital\_statusMarried-civ-spouse 2.420e+00 1.924e-01 12.575 < 2e-16 \*\*\*  
## marital\_statusMarried-spouse-absent 7.213e-01 5.490e-01 1.314 0.188909   
## marital\_statusNever-married -8.228e-02 2.396e-01 -0.343 0.731258   
## marital\_statusSeparated 9.075e-02 4.721e-01 0.192 0.847558   
## marital\_statusWidowed -4.776e-02 5.056e-01 -0.094 0.924745   
## f.educationAssociated 2.240e-01 6.882e-01 0.325 0.744834   
## f.educationHS-graduated 5.371e-01 4.320e-01 1.243 0.213731   
## f.educationPost-Bachelors 1.030e+00 9.990e-01 1.031 0.302632   
## f.educationSome-college 5.292e-01 5.333e-01 0.992 0.321010   
## f.educationProf-school 1.054e+00 1.107e+00 0.952 0.341320   
## f.educationBachelors 8.389e-01 8.428e-01 0.995 0.319560   
## occupationCraft-repair -7.345e-02 2.262e-01 -0.325 0.745430   
## occupationExec-managerial 6.133e-01 2.193e-01 2.796 0.005173 \*\*   
## occupationFarming-fishing -8.835e-01 4.089e-01 -2.161 0.030729 \*   
## occupationHandlers-cleaners -5.424e-01 4.015e-01 -1.351 0.176736   
## occupationMachine-op-inspct -7.321e-01 3.040e-01 -2.408 0.016043 \*   
## occupationOther-service -9.017e-01 3.194e-01 -2.823 0.004756 \*\*   
## occupationPriv-house-serv -1.300e+01 2.568e+02 -0.051 0.959634   
## occupationProf-specialty 3.820e-01 2.289e-01 1.669 0.095176 .   
## occupationProtective-serv 6.656e-01 3.829e-01 1.738 0.082168 .   
## occupationSales 8.203e-02 2.390e-01 0.343 0.731406   
## occupationTech-support 6.244e-01 3.191e-01 1.957 0.050363 .   
## occupationTransport-moving -5.071e-01 2.901e-01 -1.748 0.080425 .   
## sexMale -9.892e-03 1.526e-01 -0.065 0.948303   
## workclassLocal-gov -9.224e-01 3.094e-01 -2.981 0.002870 \*\*   
## workclassPrivate -8.537e-01 2.566e-01 -3.327 0.000877 \*\*\*  
## workclassSelf-emp-inc -3.847e-01 3.462e-01 -1.111 0.266471   
## workclassSelf-emp-not-inc -1.257e+00 3.084e-01 -4.076 4.57e-05 \*\*\*  
## workclassState-gov -1.439e+00 3.529e-01 -4.077 4.56e-05 \*\*\*  
## raceAsian-Pac-Islander 1.226e+00 7.050e-01 1.739 0.082014 .   
## raceBlack 9.740e-01 6.740e-01 1.445 0.148422   
## raceOther 1.520e+00 9.915e-01 1.533 0.125232   
## raceWhite 1.141e+00 6.451e-01 1.769 0.076949 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2405.2 on 3876 degrees of freedom  
## AIC: 2489.2  
##   
## Number of Fisher Scoring iterations: 14

**Anova**(gm7)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 102.49 2 < 2.2e-16 \*\*\*  
## education\_num 3.08 1 0.0794 .   
## poly(capital\_gain, 2) 202.97 2 < 2.2e-16 \*\*\*  
## f.hours\_per\_week 37.97 2 5.679e-09 \*\*\*  
## capital\_loss 36.60 1 1.449e-09 \*\*\*  
## marital\_status 380.57 5 < 2.2e-16 \*\*\*  
## f.education 9.48 6 0.1483   
## occupation 69.58 12 3.831e-10 \*\*\*  
## sex 0.00 1 0.9483   
## workclass 27.34 5 4.894e-05 \*\*\*  
## race 4.63 4 0.3275   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Ara veiem que ja no tenim cap variable amb NAs, però amb l’anova veiem que tenim variables factors que no són gaire significatives i decidim treure-les.

gm8<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss**+**marital\_status**+**occupation**+**workclass**+**race,family=binomial,data=dfw)  
**summary**(gm8)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + f.hours\_per\_week + capital\_loss + marital\_status + occupation +   
## workclass + race, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.5873 -0.4838 -0.1801 -0.0265 3.6122   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.071e+00 7.761e-01 -9.110 < 2e-16 \*\*\*  
## poly(age, 2)1 3.951e+01 4.568e+00 8.650 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.558e+01 4.678e+00 -7.606 2.84e-14 \*\*\*  
## education\_num 2.815e-01 2.767e-02 10.172 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)1 4.421e+01 3.871e+00 11.419 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)2 -8.375e+00 2.594e+00 -3.228 0.001245 \*\*   
## f.hours\_per\_week[40] 6.435e-01 1.734e-01 3.711 0.000206 \*\*\*  
## f.hours\_per\_week[41-78] 1.061e+00 1.779e-01 5.961 2.50e-09 \*\*\*  
## capital\_loss 6.753e-04 1.120e-04 6.030 1.64e-09 \*\*\*  
## marital\_statusMarried-civ-spouse 2.408e+00 1.818e-01 13.243 < 2e-16 \*\*\*  
## marital\_statusMarried-spouse-absent 6.962e-01 5.423e-01 1.284 0.199198   
## marital\_statusNever-married -6.453e-02 2.383e-01 -0.271 0.786556   
## marital\_statusSeparated 1.069e-01 4.709e-01 0.227 0.820335   
## marital\_statusWidowed -1.396e-02 5.052e-01 -0.028 0.977963   
## occupationCraft-repair -8.316e-02 2.165e-01 -0.384 0.700859   
## occupationExec-managerial 6.185e-01 2.151e-01 2.876 0.004029 \*\*   
## occupationFarming-fishing -8.758e-01 4.048e-01 -2.163 0.030507 \*   
## occupationHandlers-cleaners -5.448e-01 3.965e-01 -1.374 0.169443   
## occupationMachine-op-inspct -7.021e-01 2.989e-01 -2.349 0.018817 \*   
## occupationOther-service -9.013e-01 3.162e-01 -2.851 0.004365 \*\*   
## occupationPriv-house-serv -1.291e+01 2.575e+02 -0.050 0.960006   
## occupationProf-specialty 4.017e-01 2.208e-01 1.820 0.068806 .   
## occupationProtective-serv 6.404e-01 3.766e-01 1.701 0.089015 .   
## occupationSales 8.471e-02 2.349e-01 0.361 0.718388   
## occupationTech-support 5.485e-01 3.159e-01 1.737 0.082460 .   
## occupationTransport-moving -4.857e-01 2.836e-01 -1.712 0.086833 .   
## workclassLocal-gov -9.035e-01 3.067e-01 -2.946 0.003219 \*\*   
## workclassPrivate -8.505e-01 2.553e-01 -3.331 0.000866 \*\*\*  
## workclassSelf-emp-inc -3.520e-01 3.455e-01 -1.019 0.308297   
## workclassSelf-emp-not-inc -1.250e+00 3.072e-01 -4.070 4.69e-05 \*\*\*  
## workclassState-gov -1.403e+00 3.493e-01 -4.015 5.94e-05 \*\*\*  
## raceAsian-Pac-Islander 1.206e+00 6.992e-01 1.725 0.084516 .   
## raceBlack 9.282e-01 6.693e-01 1.387 0.165498   
## raceOther 1.474e+00 9.941e-01 1.483 0.138044   
## raceWhite 1.102e+00 6.403e-01 1.721 0.085284 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2414.7 on 3883 degrees of freedom  
## AIC: 2484.7  
##   
## Number of Fisher Scoring iterations: 14

**Anova**(gm8)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 104.36 2 < 2.2e-16 \*\*\*  
## education\_num 115.26 1 < 2.2e-16 \*\*\*  
## poly(capital\_gain, 2) 201.59 2 < 2.2e-16 \*\*\*  
## f.hours\_per\_week 39.64 2 2.472e-09 \*\*\*  
## capital\_loss 37.79 1 7.861e-10 \*\*\*  
## marital\_status 460.60 5 < 2.2e-16 \*\*\*  
## occupation 68.95 12 5.023e-10 \*\*\*  
## workclass 27.37 5 4.835e-05 \*\*\*  
## race 4.52 4 0.3402   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(gm8)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(age, 2) 1.195048 2 1.045554  
## education\_num 1.559730 1 1.248892  
## poly(capital\_gain, 2) 1.105900 2 1.025484  
## f.hours\_per\_week 1.205177 2 1.047762  
## capital\_loss 1.016032 1 1.007984  
## marital\_status 1.256383 5 1.023086  
## occupation 2.391018 12 1.036989  
## workclass 1.486201 5 1.040418  
## race 1.103342 4 1.012369

Veiem que tenim una mica de colinealitat però no massa, així que decidim que així està bé. Veiem però que de la variable occupation tenim masses grups, així que decidim utilitzar el seu factor, que té els grups agrupats en un número més petit.

gm9<-**glm**(over\_50k**~poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss**+**marital\_status**+**f.occupation**+**workclass**+**race,family=binomial,data=dfw)  
**summary**(gm9)

##   
## Call:  
## glm(formula = over\_50k ~ poly(age, 2) + education\_num + poly(capital\_gain,   
## 2) + f.hours\_per\_week + capital\_loss + marital\_status + f.occupation +   
## workclass + race, family = binomial, data = dfw)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4896 -0.5025 -0.1982 -0.0386 3.8283   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -7.666e+00 7.942e-01 -9.652 < 2e-16 \*\*\*  
## poly(age, 2)1 3.866e+01 4.468e+00 8.653 < 2e-16 \*\*\*  
## poly(age, 2)2 -3.451e+01 4.632e+00 -7.449 9.40e-14 \*\*\*  
## education\_num 3.592e-01 2.476e-02 14.510 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)1 4.455e+01 3.854e+00 11.559 < 2e-16 \*\*\*  
## poly(capital\_gain, 2)2 -8.487e+00 2.548e+00 -3.330 0.000868 \*\*\*  
## f.hours\_per\_week[40] 6.216e-01 1.694e-01 3.670 0.000242 \*\*\*  
## f.hours\_per\_week[41-78] 1.097e+00 1.721e-01 6.374 1.84e-10 \*\*\*  
## capital\_loss 6.834e-04 1.094e-04 6.246 4.21e-10 \*\*\*  
## marital\_statusMarried-civ-spouse 2.327e+00 1.796e-01 12.957 < 2e-16 \*\*\*  
## marital\_statusMarried-spouse-absent 4.453e-01 5.420e-01 0.822 0.411321   
## marital\_statusNever-married -1.094e-01 2.371e-01 -0.461 0.644605   
## marital\_statusSeparated -4.187e-02 4.697e-01 -0.089 0.928978   
## marital\_statusWidowed -2.418e-02 5.036e-01 -0.048 0.961706   
## f.occupationAdministrativos -1.647e-01 1.514e-01 -1.088 0.276630   
## f.occupationTrabajos manuales 1.427e-01 1.328e-01 1.074 0.282749   
## f.occupationVentas 7.723e-02 1.498e-01 0.516 0.606163   
## workclassLocal-gov -9.910e-01 2.978e-01 -3.328 0.000874 \*\*\*  
## workclassPrivate -1.057e+00 2.527e-01 -4.183 2.88e-05 \*\*\*  
## workclassSelf-emp-inc -3.610e-01 3.394e-01 -1.064 0.287550   
## workclassSelf-emp-not-inc -1.458e+00 2.985e-01 -4.887 1.03e-06 \*\*\*  
## workclassState-gov -1.412e+00 3.486e-01 -4.050 5.12e-05 \*\*\*  
## raceAsian-Pac-Islander 1.188e+00 7.153e-01 1.660 0.096860 .   
## raceBlack 9.102e-01 6.865e-01 1.326 0.184902   
## raceOther 1.261e+00 9.857e-01 1.280 0.200682   
## raceWhite 1.162e+00 6.591e-01 1.764 0.077816 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 4268.5 on 3917 degrees of freedom  
## Residual deviance: 2479.4 on 3892 degrees of freedom  
## AIC: 2531.4  
##   
## Number of Fisher Scoring iterations: 6

**Anova**(gm9)

## Analysis of Deviance Table (Type II tests)  
##   
## Response: over\_50k  
## LR Chisq Df Pr(>Chisq)   
## poly(age, 2) 103.59 2 < 2.2e-16 \*\*\*  
## education\_num 256.08 1 < 2.2e-16 \*\*\*  
## poly(capital\_gain, 2) 208.27 2 < 2.2e-16 \*\*\*  
## f.hours\_per\_week 47.25 2 5.501e-11 \*\*\*  
## capital\_loss 40.36 1 2.116e-10 \*\*\*  
## marital\_status 464.06 5 < 2.2e-16 \*\*\*  
## f.occupation 4.27 3 0.2333   
## workclass 36.01 5 9.468e-07 \*\*\*  
## race 5.11 4 0.2758   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**vif**(gm9)

## GVIF Df GVIF^(1/(2\*Df))  
## poly(age, 2) 1.177529 2 1.041701  
## education\_num 1.245984 1 1.116237  
## poly(capital\_gain, 2) 1.057436 2 1.014060  
## f.hours\_per\_week 1.141331 2 1.033601  
## capital\_loss 1.012982 1 1.006470  
## marital\_status 1.216278 5 1.019772  
## f.occupation 1.205781 3 1.031679  
## workclass 1.189052 5 1.017466  
## race 1.072339 4 1.008769

Però aquí veiem que el seu factor no és gaire significatiu per tant decidim quedarnos amb l’anterior model, gm8.

## Interactions between variables

Un cop afegits els factors, en aquest apartat buscarem i utilitzarem les interaccions entre algunes variables per veure si aquesta eina millora el nostre model. En el primer cas provarem amb la variable covariant age.

gmi1<-**glm**(over\_50k**~**(education\_num**+poly**(capital\_gain,2)**+**f.hours\_per\_week**+**capital\_loss**+**marital\_status**+**occupation**+**workclass**+**race)**\*poly**(age,2),family=binomial,data=dfw)

**anova**(gmi1,test="LR")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: over\_50k  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)  
## NULL 3917 4268.5   
## education\_num 1 496.11 3916 3772.4 < 2.2e-16  
## poly(capital\_gain, 2) 2 272.64 3914 3499.7 < 2.2e-16  
## f.hours\_per\_week 2 146.10 3912 3353.6 < 2.2e-16  
## capital\_loss 1 70.25 3911 3283.4 < 2.2e-16  
## marital\_status 5 646.53 3906 2636.8 < 2.2e-16  
## occupation 12 80.01 3894 2556.8 4.105e-12  
## workclass 5 32.59 3889 2524.2 4.546e-06  
## race 4 5.16 3885 2519.1 0.271254  
## poly(age, 2) 2 104.36 3883 2414.7 < 2.2e-16  
## education\_num:poly(age, 2) 2 3.60 3881 2411.1 0.165065  
## poly(capital\_gain, 2):poly(age, 2) 4 52.70 3877 2358.4 9.869e-11  
## f.hours\_per\_week:poly(age, 2) 4 3.14 3873 2355.3 0.535174  
## capital\_loss:poly(age, 2) 2 0.61 3871 2354.7 0.735559  
## marital\_status:poly(age, 2) 10 18.80 3861 2335.9 0.042937  
## occupation:poly(age, 2) 24 24.59 3837 2311.3 0.428447  
## workclass:poly(age, 2) 10 29.19 3827 2282.1 0.001162  
## race:poly(age, 2) 8 5.51 3819 2276.6 0.701734  
##   
## NULL   
## education\_num \*\*\*  
## poly(capital\_gain, 2) \*\*\*  
## f.hours\_per\_week \*\*\*  
## capital\_loss \*\*\*  
## marital\_status \*\*\*  
## occupation \*\*\*  
## workclass \*\*\*  
## race   
## poly(age, 2) \*\*\*  
## education\_num:poly(age, 2)   
## poly(capital\_gain, 2):poly(age, 2) \*\*\*  
## f.hours\_per\_week:poly(age, 2)   
## capital\_loss:poly(age, 2)   
## marital\_status:poly(age, 2) \*   
## occupation:poly(age, 2)   
## workclass:poly(age, 2) \*\*   
## race:poly(age, 2)   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**BIC**(gmi1,gm8)

## df BIC  
## gmi1 99 3095.650  
## gm8 35 2704.288

Es pot veure que hi han unes quantes interaccions que sí que són rellevants. Però el BIC del model és massa gran. Ara provarem amb el factor de hours\_per\_week.

gmi2<-**glm**(over\_50k**~**(**poly**(age,2)**+**education\_num**+poly**(capital\_gain,2)**+**capital\_loss**+**marital\_status**+**occupation**+**workclass**+**race)**\***f.hours\_per\_week,family=binomial,data=dfw)

**anova**(gmi2,test="LR")

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: over\_50k  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 3917 4268.5  
## poly(age, 2) 2 464.81 3915 3803.7  
## education\_num 1 424.40 3914 3379.3  
## poly(capital\_gain, 2) 2 227.92 3912 3151.3  
## capital\_loss 1 68.12 3911 3083.2  
## marital\_status 5 506.87 3906 2576.4  
## occupation 12 86.18 3894 2490.2  
## workclass 5 30.13 3889 2460.0  
## race 4 5.69 3885 2454.4  
## f.hours\_per\_week 2 39.64 3883 2414.7  
## poly(age, 2):f.hours\_per\_week 4 3.40 3879 2411.3  
## education\_num:f.hours\_per\_week 2 9.49 3877 2401.8  
## poly(capital\_gain, 2):f.hours\_per\_week 4 44.80 3873 2357.0  
## capital\_loss:f.hours\_per\_week 2 0.68 3871 2356.4  
## marital\_status:f.hours\_per\_week 10 24.49 3861 2331.9  
## occupation:f.hours\_per\_week 24 21.80 3837 2310.1  
## workclass:f.hours\_per\_week 10 16.39 3827 2293.7  
## race:f.hours\_per\_week 8 8.90 3819 2284.8  
## Pr(>Chi)   
## NULL   
## poly(age, 2) < 2.2e-16 \*\*\*  
## education\_num < 2.2e-16 \*\*\*  
## poly(capital\_gain, 2) < 2.2e-16 \*\*\*  
## capital\_loss < 2.2e-16 \*\*\*  
## marital\_status < 2.2e-16 \*\*\*  
## occupation 2.694e-13 \*\*\*  
## workclass 1.393e-05 \*\*\*  
## race 0.223823   
## f.hours\_per\_week 2.472e-09 \*\*\*  
## poly(age, 2):f.hours\_per\_week 0.492926   
## education\_num:f.hours\_per\_week 0.008697 \*\*   
## poly(capital\_gain, 2):f.hours\_per\_week 4.386e-09 \*\*\*  
## capital\_loss:f.hours\_per\_week 0.711223   
## marital\_status:f.hours\_per\_week 0.006407 \*\*   
## occupation:f.hours\_per\_week 0.591346   
## workclass:f.hours\_per\_week 0.089051 .   
## race:f.hours\_per\_week 0.350429   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**BIC**(gmi2,gm8)

## df BIC  
## gmi2 99 3103.836  
## gm8 35 2704.288

Aqui podem veure interaccions interessants entre el factor de hours\_per\_Week amb capital\_gain, workclass i marital status. Però després realitzem una comparació amb el nostre model previ i veiem que aquest és millor.

## Diagnostic of the model

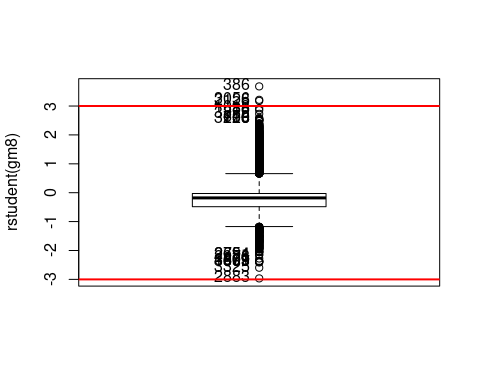
### Residual analysis

El nostre model triat és m7. Ara farem un anàlisi dels residus i després farem una predicció.

**Boxplot**(**rstudent**(gm8), id.n=2)

## [1] 2883 3325 503 3699 1009 3809 671 2978 2651 2794 386 3056 2128 1018 88  
## [16] 3222 650 3726 118 200

**abline**(h=**c**(3,**-**3),col="red",lwd=2)

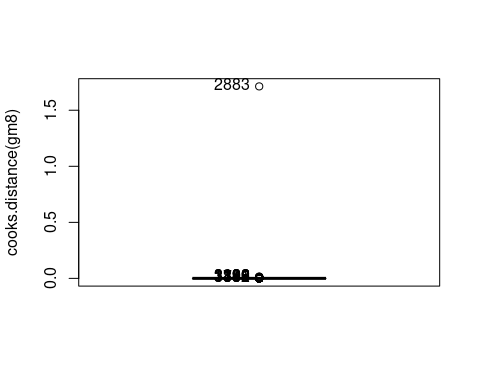


out <- **which**(**rstudent**(gm8) **>=** 3 **|** **rstudent**(gm8) **<=** -3);**length**(out)

## [1] 3

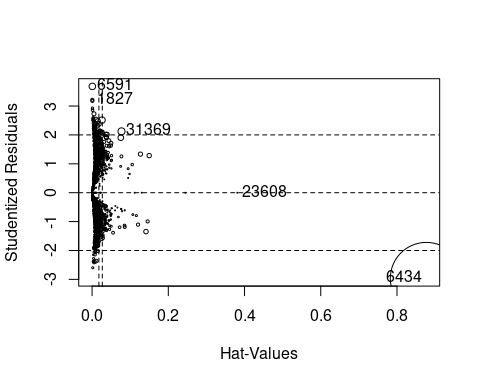
A partir d’aquest anàlisi de residus veiem que no hi han quasi possibles outliers. Ara mirarem si hi ha alguna dada influent entre l’observat.

inf<-**Boxplot**(**cooks.distance**(gm8), id.n=4)



out2 <- **which**(**cooks.distance**(gm8) **>=** 0.20);**length**(out2)

## [1] 1

**influencePlot**(gm8,id.n=3)

## StudRes Hat CookD  
## 23608 -0.002351726 0.3809855533 6.007070e-08  
## 6591 3.679308627 0.0007181840 1.398163e-02  
## 31369 2.123736277 0.0772419431 1.638049e-02  
## 6434 -2.966047063 0.8769507092 1.712768e+00  
## 1827 3.215300843 0.0006714265 3.174575e-03

A partir del segon gràfic podem observar a priori que les dades més influents són les “6434” i “23608”, observant el leverage que hi ha en el plot corresponent. Hauríem de suprimir aquests atípics/influents.

## Prediction

### Work sample

pw<-**predict**(gm8,type="response")  
pnw<- **as.numeric**(pw)  
**summary**(dfw**$**over\_50k)

## <=50K >50K   
## 2999 919

pw.y <- **factor**(**ifelse**(pnw**<**0.4,0,1),labels=**c**("pre-<50k","pred-+50k"))  
  
**length**(pw.y)

## [1] 3918

**length**(dfw**$**over\_50k)

## [1] 3918

ttwork<-**table**(pw.y,dfw**$**over\_50k);ttwork

##   
## pw.y <=50K >50K  
## pre-<50k 2707 275  
## pred-+50k 292 644

perw<-100**\*sum**(**diag**(ttwork))**/sum**(ttwork);perw

## [1] 85.52833

### Test sample

pt<-**predict**(gm8,type="response",newdata=dft)  
pnt<- **as.numeric**(pt)  
**summary**(dft**$**over\_50k)

## <=50K >50K   
## 742 238

pt.y <- **factor**(**ifelse**(pnt**<**0.4,0,1),labels=**c**("pre-<50k","pred-+50k"))  
  
tttest<-**table**(pt.y,dft**$**over\_50k);tttest

##   
## pt.y <=50K >50K  
## pre-<50k 665 74  
## pred-+50k 77 164

pert<-100**\*sum**(**diag**(tttest))**/sum**(tttest);pert

## [1] 84.59184

Aquí hem realitzat les prediccions per veure els percentatges d’encert del nostre model gm8. Hem vist que tenim una taxa d’encert del 84.591%.