

a3

Carlos A. García

April 29, 2019

## Càrrega del fitxer

Establim el directori de feina

```
setwd("D:/Users/cagarcia/uoc/M2.954 - Estadística avançada/A3")
```

Llegim el fitxer proporcionat a la pràctica

```
satisfaccioLaboral <- read.csv2("rawData_clean1.csv", header = TRUE, sep = ",", dec = ".")
attach(satisfaccioLaboral)

summary(satisfaccioLaboral)
```

```
##              city      sex   educ_level job_type   happiness
## Barcelona      :640    F:960    N:480      C :960    Min.      :0.000
## Madrid          :640    M:960    P:480      PC:960    1st Qu.:1.643
## Santiago de Compostela:640      S:480      Median :2.705
##                               U:480      Mean  :2.716
##                               3rd Qu.:3.755
##                               Max.   :7.665
##
##      age      seniority      sick_leave      sick_leave_b
## Min.   :18.00  Min.    : 0.00  Min.    : 0.000  Min.    :0.0000
## 1st Qu.:33.00  1st Qu.:15.00  1st Qu.: 0.000  1st Qu.:0.0000
## Median :40.00  Median :20.00  Median : 0.000  Median :0.0000
## Mean   :40.33  Mean    :19.03  Mean    : 6.667  Mean    :0.2464
## 3rd Qu.:47.00  3rd Qu.:24.00  3rd Qu.: 0.000  3rd Qu.:0.0000
## Max.   :67.00  Max.    :35.00  Max.    :78.000  Max.    :1.0000
##
## work_hours  Cho_initial  Cho_final
## Min.   :26.20  Min.    :0.95  Min.    :0.990
## 1st Qu.:35.50  1st Qu.:1.15  1st Qu.:1.250
## Median :38.20  Median :1.25  Median :1.350
## Mean   :38.29  Mean    :1.20  Mean    :1.351
## 3rd Qu.:40.90  3rd Qu.:1.25  3rd Qu.:1.450
## Max.   :88.00  Max.    :1.45  Max.    :1.680
```

## Model de regressió lineal

### Model de regressió lineal múltiple (regressors quantitatius)

Estimeu per mínims quadrats ordinaris un model lineal que expliqui la satisfacció laboral (happiness) d'un individu en funció de tres factors quantitatius: les hores treballades a la setmana (work\_hours), l'edat (age), i els dies de baixa en el darrer any (sick\_leave).

Com a primera aproximació, obtenim la matriu de correlacions.

```
cor(satisfaccioLaboral[,c("happiness", "work_hours", "age", "sick_leave")], use="complete")
```

```
##              happiness  work_hours      age  sick_leave
## happiness      1.0000000 -0.26412337  0.13870449 -0.44804507
```

```
## work_hours -0.2641234  1.00000000 -0.13568102  0.01104576
## age         0.1387045 -0.13568102  1.00000000  0.07945524
## sick_leave -0.4480451  0.01104576  0.07945524  1.00000000
```

Es pot veure que la major correlació se produeix amb la variable sick\_leave.

```
lmFrame <- lm(formula = happiness ~ work_hours + age + sick_leave, data = satisfaccioLaboral)
summary(lmFrame)
```

```
##
## Call:
## lm(formula = happiness ~ work_hours + age + sick_leave, data = satisfaccioLaboral)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.4519 -0.8895 -0.0004  0.8833  4.1425
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.587961    0.319848  17.471  < 2e-16 ***
## work_hours  -0.089517    0.007268 -12.316  < 2e-16 ***
## age          0.022108    0.003030   7.296 4.32e-13 ***
## sick_leave  -0.050373    0.002134 -23.607  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.295 on 1916 degrees of freedom
## Multiple R-squared:  0.2877, Adjusted R-squared:  0.2866
## F-statistic: 258 on 3 and 1916 DF, p-value: < 2.2e-16
```

**Avalueu la bondat d'ajust a través del coeficient de determinació (R2) i interpreteu-lo.**

Com es pot veure  $R^2$  és molt pobre (1 representa l'ajust perfecte, 0 cap ajust). Un ajust tan baix implica que la recta de regressió no és gaire bona.

```
summary(lmFrame)$r.squared
```

```
## [1] 0.2877132
```

```
summary(lmFrame)$adj.r.squared
```

```
## [1] 0.2865979
```

**A més, avalueu si algun dels regressos té influència significativa**

Les tres variables tenen  $\Pr(>|t|) < 5\%$ , el que implica que les tres tenen una influència significativa.

```
summary(lmFrame)$coefficients[,4]
```

```
##      (Intercept)      work_hours          age      sick_leave
## 1.547270e-63 1.330140e-33 4.320362e-13 2.307357e-108
```

El signe és negatiu en el cas de les variables work\_hours i sick\_leave. Això vol dir que, com més creixen aquests dos valors, més baixa el nivell de happiness. El signe d'age és positiu; augmenta el happiness a mida que creix l'age.

```
summary(lmFrame)$coefficients[,1]
```

```
## (Intercept) work_hours          age      sick_leave
```

```
## 5.58796064 -0.08951734 0.02210791 -0.05037324
```

Des del punt de vista de la qualitat del model de regressió, podeu indicar una raó que justifiqui la no inclusió de seniority?

La correlació d'ambdues variables és molt elevada. Això implica que incloure seniority és afegir una variable redundant. Si existeix dependència entre les variables, direm que hi ha multicollinearitat.

Només com a nota, dir que el valor  $R^2$  empitjora a mida que s'afegeixen variables.

```
cor(satisfaccioLaboral[,c("seniority", "age")], use="complete")
```

```
##          seniority      age
## seniority 1.0000000 0.9657071
## age       0.9657071 1.0000000
```

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

Estimeu per mínims quadrats ordinaris un model lineal que expliqui la satisfacció laboral (happiness) d'un individu en funció de cinc regressors.

```
satisfaccioLaboral$sexR <- relevel(satisfaccioLaboral$sex, ref = "F")
satisfaccioLaboral$educ_levelR <- relevel(satisfaccioLaboral$educ_level, ref = "N")
```

Un cop creades les noves variables, calculeu el model lineal usant la funció d'R lm

```
lmFrame <- lm(formula = happiness ~ work_hours + age + sick_leave + sexR
              + educ_levelR, data = satisfaccioLaboral)
summary(lmFrame)
```

```
##
## Call:
## lm(formula = happiness ~ work_hours + age + sick_leave + sexR +
##     educ_levelR, data = satisfaccioLaboral)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.2670 -0.8662  0.0062  0.8303  4.1134
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.432211   0.311361  17.447  <2e-16 ***
## work_hours    -0.090950   0.006993  -13.006  <2e-16 ***
## age           0.024797   0.002952   8.400  <2e-16 ***
## sick_leave    -0.049948   0.002054  -24.313  <2e-16 ***
## sexRM         0.020066   0.056870   0.353   0.7242
## educ_levelRP  -0.170203   0.081503  -2.088   0.0369 *
## educ_levelRS  -0.174640   0.080714  -2.164   0.0306 *
## educ_levelRU   0.701985   0.080448   8.726  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.245 on 1912 degrees of freedom
## Multiple R-squared:  0.3428, Adjusted R-squared:  0.3404
## F-statistic: 142.5 on 7 and 1912 DF, p-value: < 2.2e-16
```

Avalueu la bondat d'ajust a través del coeficient de determinació ( $R^2$ ) i compareu el resultat d'aquest model amb l'obtingut a l'apartat 1.1. Per a fer-ho, useu el coeficient  $R^2$  ajustat en la comparació.

El coeficient  $R^2$  ajustat és lleugerament superior als obtinguts a l'apartat 1.1. Això vol dir que s'ajusta lleugerament millor, malgrat continua sent un resultat pobre.

```
summary(lmFrame)$r.squared
```

```
## [1] 0.342776
```

```
summary(lmFrame)$adj.r.squared
```

```
## [1] 0.3403698
```

Interpreteu també el significat dels coeficients obtinguts i la seva significació estadística. En particular, quina interpretació feu dels tres coeficients associats al factor `educ_levelR`?

Per una banda tenim els coeficients depenent de la variable:

```
summary(lmFrame)$coefficients[,1]
```

```
## (Intercept)    work_hours          age    sick_leave      sexRM
##  5.43221120 -0.09094995  0.02479686 -0.04994849  0.02006581
## educ_levelRP educ_levelRS educ_levelRU
## -0.17020347 -0.17464006  0.70198453
```

I per un altre, el p-value que ens indica la seva significació estadística (si és major de 5%, entenem que no és representativa)

```
summary(lmFrame)$coefficients[,4]
```

```
## (Intercept)    work_hours          age    sick_leave      sexRM
## 2.270762e-63 4.043823e-37 8.578111e-17 5.304096e-114 7.242482e-01
## educ_levelRP educ_levelRS educ_levelRU
## 3.690196e-02 3.061267e-02 5.635053e-18
```

En el cas , particular de la satisfacció laboral en funció del nivell d'estudis:

1. Tant si els estudis són primaris com secundaris, el nivell de satisfacció és un 17% inferior amb un nivell de significació proper al 3%. Els treballadors amb estudis primaris o secundaris són més infeliços que els que no tenen estudis. El baix nivell de significació indica que és una dada representativa.
2. Si els estudis són universitaris, el nivell de satisfacció és un 80% inferior amb un nivell de significació molt proper a 0. Els treballadors amb estudis universitaris són més feliços que els que no tenen estudis. El baix nivell de significació indica que és una dada representativa.

## Realitzeu una predicció de la satisfacció laboral amb els dos models

Realitzeu la predicció del nivell de satisfacció laboral d'una treballadora de 30 anys d'edat, amb 10 anys d'antiguitat a l'empresa, nivell educatiu de formació professional, que treballa 37 hores a la setmana i que va estar 7 dies de baixa l'any passat. Al segon model no tenim en compte els anys d'experiència; calculat manualment tenim un happiness previst de:

```
ageVar <- 30
working_hoursVar <- 37
sick_daysVar <- 7
educ_levelVar <- "S"
sexVar <- "F"
happiness <- lmFrame$coefficients[1] + lmFrame$coefficients[7] +
```

```
ageVar * lmFrame$coefficients[3] +
working_hoursVar * lmFrame$coefficients[2] + sick_daysVar * lmFrame$coefficients[4]
happiness
```

```
## (Intercept)
##      2.286689
```

Al primer model no tenim en compte ni els anys d'experiència ni el sexe; calculat manualment tenim un happiness previst de:

```
lmFrame1 <- lm(formula = happiness ~ work_hours + age + sick_leave, data = satisfaccioLaboral)
happiness1 <- lmFrame1$coefficients[1] +
  ageVar * lmFrame1$coefficients[3] +
  working_hoursVar * lmFrame1$coefficients[2] + sick_daysVar * lmFrame1$coefficients[4]
happiness1
```

```
## (Intercept)
##      2.586444
```

Realitzeu la predicció de la satisfacció laboral (happiness) i el càlcul de l'interval de confiança amb els dos models. Interpreteu els resultats.

Podem veure que els resultats són els mateixos que els realitzats amb els càlculs manuals. Calculem amb un interval de confiança del 95%:

```
newdata = data.frame(work_hours = working_hoursVar, age = ageVar, sick_leave = sick_daysVar)
predict.lm(lmFrame1, newdata, interval = "confidence", level = 0.95)
```

```
##          fit          lwr          upr
## 1 2.586444 2.498182 2.674705
```

```
newdata = data.frame(work_hours = working_hoursVar, age = ageVar, sick_leave = sick_daysVar,
                      sexR=sexVar, educ_levelR=educ_levelVar)
predict.lm(lmFrame, newdata, interval = "confidence", level = 0.95)
```

```
##          fit          lwr          upr
## 1 2.286689 2.145544 2.427834
```

Com es pot veure, els valors i els intervals canvien depenent del model. El segon model incorpora el nivell d'educació, que és representativa. Tambè inclou el sexe, que no ho és.

## Model de regressió logística

### Estimació d'un model de regressió logística

El primer pas serà crear una variable binària (low\_happiness) que indiqui la condició de satisfacció baixa (low\_happiness = 1) si happiness <= 4 o satisfacció normal/alta (low\_happiness = 0) si happiness > 4.

```
satisfaccioLaboral$low_happiness <- ifelse(satisfaccioLaboral$happiness<=4, 1, 0)
```

Calculeu el model de regressió logística on la variable dependent és “low\_happiness” i les explicatives són work\_hours, sexR i sick\_leave

```
glmFrame <- glm(formula = low_happiness ~ work_hours + sexR + sick_leave, data = satisfaccioLaboral)
summary(glmFrame)
```

```
##
## Call:
```

```
## glm(formula = low_happiness ~ work_hours + sexR + sick_leave,
##      data = satisfaccioLaboral)
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -0.95698  -0.03036   0.17595   0.25234   0.46124
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0544052  0.0838470   0.649   0.517
## work_hours   0.0184868  0.0021658   8.536 <2e-16 ***
## sexRM        -0.0160463  0.0177890  -0.902   0.367
## sick_leave   0.0058589  0.0006401   9.154 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1517143)
##
##      Null deviance: 314.91  on 1919  degrees of freedom
## Residual deviance: 290.68  on 1916  degrees of freedom
## AIC: 1834.1
##
## Number of Fisher Scoring iterations: 2
```

Avalueu si algú dels regressores té influència significativa (p-valor del contrast individual inferior al 5%).

Sí. Les variables `work_hours` i `sick_leave` tenen una influència significativa. El seu p-value és molt proper a 0.

Avaluant els resultats, es pot dir que un individu amb moltes hores de treball té major probabilitat de tenir “low.happiness”?

Sí, ja que el coeficient és positiu (en aquest cas, 1 és low.happiness; més elevat significa “més trist”) i la variable té una influència significativa.

Es pot afirmar a partir del model que ser dona augmenta la probabilitat de tenir “low.happiness”?

No. Sí que a primera vista ho sembla (el coeficient de home és negatiu, el que vol dir menys infelicitat). No ho podem afirmar perquè la variable no té una influència significativa.

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

Usant el model anterior, calculeu la probabilitat de tenir un baix happiness (low\_happiness) per a un home que treballa 40 hores a la setmana i va estar 15 dies de baixa l'any passat.

Podem veure que fent els càlculs, tant de forma automàtica (funció `predict.glm`) com manual, el resultat és el mateix (proper a 1).

```
working_hoursVar <- 40
sick_daysVar <- 15
sexVar <- "M"
glmFrame$coefficients
```

```
## (Intercept)  work_hours      sexRM  sick_leave
##  0.054405207  0.018486764 -0.016046349  0.005858882
```

```

happiness <- glmFrame$coefficients[1] + glmFrame$coefficients[3] +
  working_hoursVar * glmFrame$coefficients[2] + sick_daysVar * glmFrame$coefficients[4]
happiness

## (Intercept)
## 0.8657127

newdata = data.frame(work_hours = working_hoursVar, sexR = sexVar, sick_leave = sick_daysVar)
predict.glm(glmFrame, newdata)

## 1
## 0.8657127

```

## Millora del model

Decidiu si es prefereix el model inicial o bé un dels models amb cityR, amb educ\_levelR, o amb les dues.

Primer obtenim les diferents línies logarítmiques de regressió. Concretament:

1. Incloent la ciutat, prenent com a referència Barcelona (frame glmFrameCity)
2. Incloent el nivell d'estudis, prenent com a referència els estudis bàsics (frame glmFrameEduc)
3. Incloent les ambdues anteriors (frame glmFrameCityEduc)

```

cityVar = "Barcelona"
satisfaccioLaboral$cityR <- releve(satisfaccioLaboral$city, ref = cityVar)
glmFrameCity <- glm(formula = low_happiness ~ work_hours + sexR + sick_leave + cityR,
  data = satisfaccioLaboral)
summary(glmFrameCity)

##
## Call:
## glm(formula = low_happiness ~ work_hours + sexR + sick_leave +
##   cityR, data = satisfaccioLaboral)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02495  -0.05591   0.14715   0.25534   0.49645
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.039278   0.083168  -0.472   0.6368
## work_hours     0.018912   0.002125   8.901 <2e-16 ***
## sexRM         -0.015998   0.017442  -0.917   0.3591
## sick_leave     0.005952   0.000628   9.479 <2e-16 ***
## cityRMadrid    0.047338   0.021366   2.216   0.0268 *
## cityRSantiago de Compostela 0.182935   0.021358   8.565 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1458485)
##
##      Null deviance: 314.91  on 1919  degrees of freedom
## Residual deviance: 279.15  on 1914  degrees of freedom
## AIC: 1760.4

```

```
##
## Number of Fisher Scoring iterations: 2
glmFrameEduc <- glm(formula = low_happiness ~ work_hours + sexR + sick_leave + educ_levelR,
                    data = satisfaccioLaboral)
summary(glmFrameEduc)

##
## Call:
## glm(formula = low_happiness ~ work_hours + sexR + sick_leave +
##      educ_levelR, data = satisfaccioLaboral)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.95598  -0.04813   0.14699   0.23965   0.53569
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0681609  0.0839923   0.812   0.417
## work_hours     0.0188680  0.0021211  8.895 < 2e-16 ***
## sexRM         -0.0161512  0.0173969  -0.928   0.353
## sick_leave     0.0057797  0.0006267  9.223 < 2e-16 ***
## educ_levelRP   0.0058085  0.0246412   0.236   0.814
## educ_levelRS   0.0500601  0.0245978   2.035   0.042 *
## educ_levelRU  -0.1669484  0.0245899  -6.789 1.5e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1450992)
##
##      Null deviance: 314.91  on 1919  degrees of freedom
## Residual deviance: 277.57  on 1913  degrees of freedom
## AIC: 1751.5
##
## Number of Fisher Scoring iterations: 2
glmFrameCityEduc <- glm(formula = low_happiness ~ work_hours + sexR + sick_leave + educ_levelR + cityR,
                       data = satisfaccioLaboral)
summary(glmFrameCityEduc)

##
## Call:
## glm(formula = low_happiness ~ work_hours + sexR + sick_leave +
##      educ_levelR + cityR, data = satisfaccioLaboral)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06238  -0.06772   0.12790   0.24992   0.59694
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.0255636  0.0831832  -0.307   0.7586
## work_hours     0.0192942  0.0020789   9.281 < 2e-16 ***
## sexRM         -0.0161029  0.0170405  -0.945   0.3448
## sick_leave     0.0058730  0.0006142   9.561 < 2e-16 ***
```



```
## educ_levelRP          0.0059062  0.0241365  0.245  0.8067
## educ_levelRS          0.0501662  0.0240939  2.082  0.0375 *
## educ_levelRU         -0.1669032  0.0240862 -6.929 5.76e-12 ***
## cityRMadrid           0.0471708  0.0208746  2.260  0.0240 *
## cityRSantiago de Compostela 0.1829253  0.0208666  8.766 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1392146)
##
## Null deviance: 314.91  on 1919  degrees of freedom
## Residual deviance: 266.04  on 1911  degrees of freedom
## AIC: 1674
##
## Number of Fisher Scoring iterations: 2
```

Com es pot veure, el menor AIC és el del model que inclou ciutat i nivell d'estudis:

```
AIC(glmFrameCity)
```

```
## [1] 1760.356
```

```
AIC(glmFrameEduc)
```

```
## [1] 1751.462
```

```
AIC(glmFrameCityEduc)
```

```
## [1] 1673.965
```

Crida molt l'atenció com creix el nivell de lowHappiness si la persona fa feina a Santiago de Compostela. A més, el valor és representatiu, ja que el nivell de p-value és molt proper a 0. Com a nota, dir que AIC<sup>1</sup> és l'acrònim de “Akaike information criterion”. És un estimador relatiu de la qualitat d'un model estadístic per a un conjunt de dades. Dir que es correspon amb:

$$AIC = 2k - 2\ln(L)$$

On k és el nombre de variables i L és el màxim valor de la funció de verosimilitut per al model estimat  
##Qualitat de l'ajust ###Calculeu la matriu de confusió del millor model de l'apartat 2.3 suposant un llindar de discriminació del 75%

Obtenim la matriu de confusió on mostrem les variables:

1. falsePositive. Nombre de falsos positius; el nivell real de low\_happiness és 1, però obtenim una probabilitat menor del 75% de que així sigui segons la recta de regressió.
2. okPredictPositive. Nombre de positius que es preveuen com a positius; el nivell real de low\_happiness és 1 i obtenim una probabilitat igual o superior al 75% de que així sigui segons la recta de regressió. La previsió segons la recta de regressió és correcta.
3. okPredictNegative. Nombre de negatius que es preveuen com a negatius; el nivell real de low\_happiness és 0 i obtenim una probabilitat inferior al 75% de que així sigui segons la recta de regressió. La previsió segons la recta de regressió és correcta.
4. falseNegative. Nombre de falsos negatius; el nivell real de low\_happiness és 0, però obtenim una probabilitat igual o superior al 75% de que el low\_happiness sigui 1 segons la recta de regressió.

```
satisfaccioLaboral$low_happiness_predict <- predict(glmFrameCityEduc)
```

<sup>1</sup>[https://en.wikipedia.org/wiki/Akaike\\_information\\_criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion)

```

pThreshold <- 0.75
falsePositive <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict >= pThreshold
  & satisfaccioLaboral$low_happiness == 0])
okPredictPositive <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict >= pThreshold
  & satisfaccioLaboral$low_happiness == 1])
okPredictNegative <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict < pThreshold
  & satisfaccioLaboral$low_happiness == 0])
falseNegative <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict < pThreshold
  & satisfaccioLaboral$low_happiness == 1])
falsePositive

## [1] 91
okPredictPositive

## [1] 1043
okPredictNegative

## [1] 306
falseNegative

## [1] 480
El percentatge de valors correctament classificats és:
okClassified <- (okPredictPositive + okPredictNegative)/(nrow(satisfaccioLaboral))
okClassified

## [1] 0.7026042
El percentatge de valors classificats erròniament és:
koClassified <- (falsePositive + falseNegative)/(nrow(satisfaccioLaboral))
koClassified

## [1] 0.2973958

```

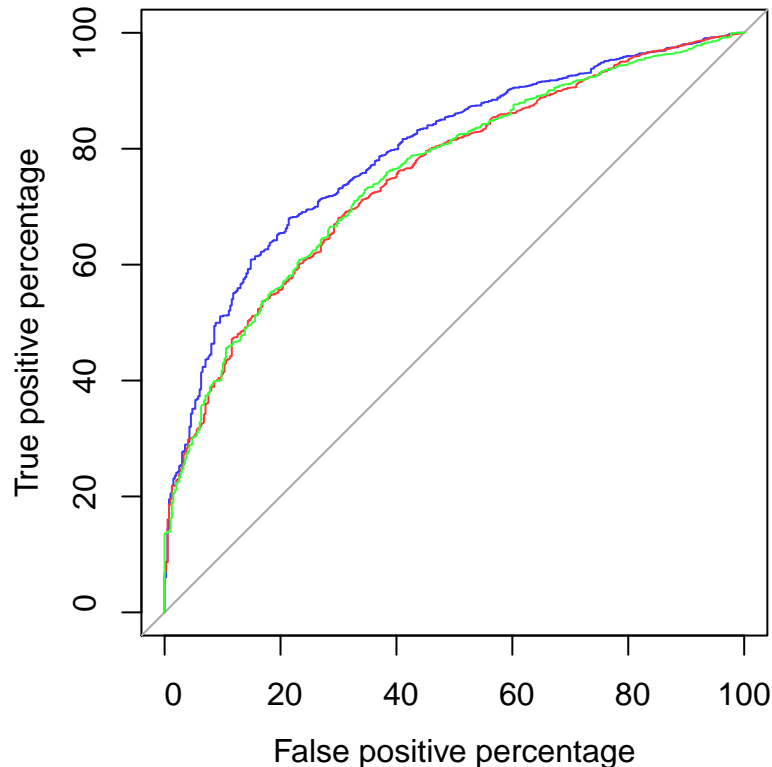
## Corba ROC

Realitzeu el dibuix de les corbes ROC per a representar la qualitat dels models predictius obtinguts a l'apartat 2.3 en un únic gràfic

```

par(pty = "s")
rocCityEduc <- pROC::roc(satisfaccioLaboral$low_happiness, glmFrameCityEduc$fitted.values, plot=TRUE, p
  legacy.axes=TRUE, lwd=1, col="#3737ff",
  xlab="False positive percentage", ylab="True positive percentage", add=FALSE, print.auc.y=40)
rocCity <- pROC::roc(satisfaccioLaboral$low_happiness, glmFrameCity$fitted.values, plot=TRUE, percent=TRUE,
  legacy.axes=TRUE, lwd=1, col="#ff3737",
  xlab="False positive percentage", ylab="True positive percentage", add=TRUE, print.auc.y=40)
rocEduc <- pROC::roc(satisfaccioLaboral$low_happiness, glmFrameEduc$fitted.values, plot=TRUE, percent=TRUE,
  legacy.axes=TRUE, lwd=1, col="#3737ff",
  xlab="False positive percentage", ylab="True positive percentage", add=TRUE, print.auc.y=40)

```



Imprimim els gràfics en percentatge per a veure-ho millor. La recta diagonal representa el cas en el que la proporció d'elements bé classificats és igual al percentatge d'elements mal classificats. És a dir: com més allunyada estigui la corba ROC de la recta, major és el percentatge d'elements bé classificats. Es veu clarament que el millor classificador és el que inclou la ciutat i el nivell educatiu (corba blava). Els altres dos classificadors són molt similars. El valor AUC se correspon amb l'àrea sota la corba. Com major sigui aquesta àrea, més allunyada està la corba i millor és el classificador. Podem veure que es confirma que el millor classificador és el que inclou la ciutat i el nivell d'educació.

```
pROC::auc(rocCityEduc)
```

```
## Area under the curve: 79.22%
```

```
pROC::auc(rocCity)
```

```
## Area under the curve: 75.27%
```

```
pROC::auc(rocEduc)
```

```
## Area under the curve: 75.44%
```

Podem veure que la corba blava s'allunya més de la recta sobre el 55%. Així, sembla que se classificaran millor més elements si prenem aquest límit. Fent la prova:

```
pThreshold <- 0.55
falsePositive <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict >= pThreshold
  & satisfaccioLaboral$low_happiness == 0])
okPredictPositive <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict >= pThreshold
  & satisfaccioLaboral$low_happiness == 1])
```

```

okPredictNegative <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict < pThreshold
  & satisfaccioLaboral$low_happiness == 0])
falseNegative <- length(satisfaccioLaboral$low_happiness_predict[
  satisfaccioLaboral$low_happiness_predict < pThreshold
  & satisfaccioLaboral$low_happiness == 1])
falsePositive

## [1] 334
okPredictPositive

## [1] 1473
okPredictNegative

## [1] 63
falseNegative

## [1] 50
okClassified <- (okPredictPositive + okPredictNegative)/(nrow(satisfaccioLaboral))
okClassified

## [1] 0.8
koClassified <- (falsePositive + falseNegative)/(nrow(satisfaccioLaboral))
koClassified

## [1] 0.2

```

Podem veure que, efectivament, el percentatge és clarament millor.