# Activitat 3: Modelització predictiva

*Enunciat*

*26 abril de 2019*

## Índex



---

En aquesta activitat s'usarà el fitxer **rawData_clean1.csv** ja preparat, és a dir, després del preprocés que s'ha realitzat a la primera activitat.

Recordeu que aquest fitxer emmagatzema les dades d'un anàlisi de satisfacció laboral en una empresa. L'objectiu és investigar si la satisfacció laboral dels treballadors està relacionada amb el nivell d'estudis i d'altres característiques dels treballadors.

Les dades recollides per a cada treballador són:

1. La ciutat del lloc de treball
2. El sexe del treballador
3. El nivell d'estudis (N: sense estudis, P: estudis primaris, S: educació secundària o educació professional, U: universitaris)
4. El tipus de treball (C: qualificat, PC: poco qualificat)
5. La satisfacció laboral del treballador (compresa entre 0 i 10)
6. L'edat (anys)
7. L'antiguitat (en anys)
8. Els dies de baixa en el darrer any laboral
9. Si el treballador va estar de baixa o no el darrer any (variable binària)
10. Les hores que treballa a la setmana en promig
11. El nivell de colesterol en la penúltima revisió mèdica (mmol/L)
12. El nivell de colesterol en l'última revisió mèdica (mmol/L).

Aquesta base de dades conté 1920 registres i 12 variables. Les variables són city, sex, educ_level, job_type, happiness, age, seniority, sick_leave, sick_leave_b, work_hours, Cho_initial, Cho_final.

---

# 1 Model de regressió lineal

---

En primer lloc, estudiarem la possible associació entre la satisfacció laboral i algunes característiques de cada individu.

## 1.1 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). Podeu usar la instrucció d'R `lm`.

Avalueu la bondat d'ajust a través del coeficient de determinació ($R^2$) i interpreteu-lo.

A més, avalueu si algun dels regressos té influència significativa (p-valor del contrast individual inferior al 5%) i en quin sentit.

Observeu que, a diferència d'age, no s'ha afegit la variable seniority en el model de regressió. Des del punt de vista de la qualitat del model de regressiu, podeu indicar una raó que justifiqui la no inclusió de seniority?

## 1.2 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. A més dels tres anteriors (work_hours, age i sick_leave), afegiu les variables sex i educ_level. Donat que aquestes variables són categòriques haureu d'usar el que s'anomena variables **dummy** per a poder calcular el model lineal. Per a fer-ho, heu d'especificar quina és la categoria de referència amb la funció `relevel()`. En la variable Sex, la categoria de referència és "F". En la variable educ_level, la categoria de referència és "N". Quan ho feu, definiu noves variables SexR i educ_levelR, que continguin aquesta reordenació, i les afegiu al conjunt de dades.

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

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. 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?

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

Realitzeu la predicció del nivell de satifacció 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.

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

---

# 2 Model de regressió logística

---

La variable satisfacció laboral es pot convertir en una variable binària que correspongui a baixa satisfacció si té un valor menor o igual que 4, i en cas contrari serà satisfacció normal/alta.

Es vol avaluar la qualitat predictiva de vàries variables presents en l'estudi en relació a la predicció d'aquesta nova variable binària sobre la satisfacció. Per tant, s'avaluarà la probabilitat de que un individu tingui una satisfacció baixa.

Per avaluar aquesta probabilitat s'aplicarà un model de regressió logística, on la variable dependent serà la variable binària que indicarà si l'individu té una satisfacció baixa.

Seguiu els passos que s'especifiquen a continuació.

## 2.1 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.

Calculeu el model de regressió logística on la variable dependent és "low_happiness" i les explicatives són work_hours, sexR i sick_leave.

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

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

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

## 2.2 Predicció en el model lineal generalizat (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.

## 2.3 Millora del model

Cerqueu un model millor a l'anterior afegint més variables explicatives. Concretament, es faran les proves següents:

- Model regressor que afegeix la variable cityR.
- Model regressor que afegeix la variable educ_levelR.
- Model regressor que afegeix les dues variables alhora: cityR i educ_levelR.

La variable cityR prové de la variable city, on s'ha establert com a categoria de referència la ciutat "Barcelona" (fent un `relevel()`).

Decidiu si es prefereix el model inicial o bé un dels models amb cityR, amb educ_levelR, o amb les dues. El criteri per a decidir el millor model és AIC. Com més petit sigui AIC, millor és el model.

**Nota**: Si al realitzar la regressió logística s'obté un missatge similar a:

```
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

aquest missatge simplement adverteix d'una convergència lenta en el procés iteratiu de trobar les estimacions. No és necessari tenir-lo en compte.

### 2.4 Qualitat de l'ajust

Calculeu la matriu de confusió del millor model de l'apartat 2.3 suposant un llindar de discriminació del 75%. Observeu quants falsos negatius hi ha i interpreteu què és un fals negatiu en aquest context. Feu el mateix amb els falsos positius.

### 2.5 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. Per a tal efecte, podeu usar la llibreria `pROC` i la instrucció `roc` per a obtenir la corba ROC. Calculeu també AUROC usant aquesta llibreria amb la funció `auc()` on cal que passeu com a paràmetre el nom de l'objecte roc.

Interpreteu el resultat.

## 3 Puntuacions dels apartats

- Apartat 1 (30%)
- Apartat 2.1 fins 2.3 (30%)
- Apartat 2.4 fins 2.5 (30%)
- Qualitat de l'informe dinàmic i del codi R (10%)

## 4 Referències

En aquest apartat, podeu indicar les referències que useu.