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
```

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 possitiu; 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 regressiu, 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. 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")
lmFrame <- lm(formula = happiness ~ work_hours + age + sick_leave + sexR + educ_levelR, data = satisfac
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
```