5 Regressió amb més d'un predictor

Mètodes empírics 2

09/05/2022

Avui

- Múltiples predictors
- Cas d'estudis (continuat)

- (In)significància estadística

Múltiples predictors

Mateixa fórmula, més sumands

$$y \sim \text{Normal}(\mu, \sigma)$$

$$\mu = \beta_0 + \beta_1 x_1 + \ldots \beta_n x_n$$

Vols saber quins factors (no) afecen el to mitjà (Hertz) d'una persona. Un col·lega t'ha donat les dades següents:

https://tinyurl.com/polite-data

1. Descriu les dades

2. Descriu com penses que les variables es podrien relacionar

Cas d'estudis: to

```
df <- read.csv("https://tinyurl.com/polite-data")
head(df)
```

```
##   subject gender sentence context pitch
## 1      F1      F       S1     pol 213.3
## 2      F1      F       S1     inf 204.5
## 3      F1      F       S2     pol 285.1
## 4      F1      F       S2     inf 259.7
## 5      F1      F       S3     pol 203.9
## 6      F1      F       S3     inf 286.9
```

Cas d'estudis: to

$$\text{pitch}_i = \beta_0 + \beta_1 \text{gender}_i + \beta_2 \text{context}_i$$

Vols saber quins factors (no) afecen el to mitjà (Hertz) d'una persona. Un col·lega t'ha donat les dades següents:

https://tinyurl.com/polite-data

1. Descriu les variants rellevants per a la teva anàlisi

Descripció de variables (gènere)

```
unique(df$gender)
```

```
## [1] F M
## Levels: F M
```

```
df_m <- filter(df, gender == 'M')
nrow(df_m)
```

```
## [1] 41
```

```
df_f <- filter(df, gender == 'F')
nrow(df_f)
```

```
## [1] 42
```

Descripció de variables (context)

```
unique(df$context)
```

```
## [1] pol inf
## Levels: inf pol
```

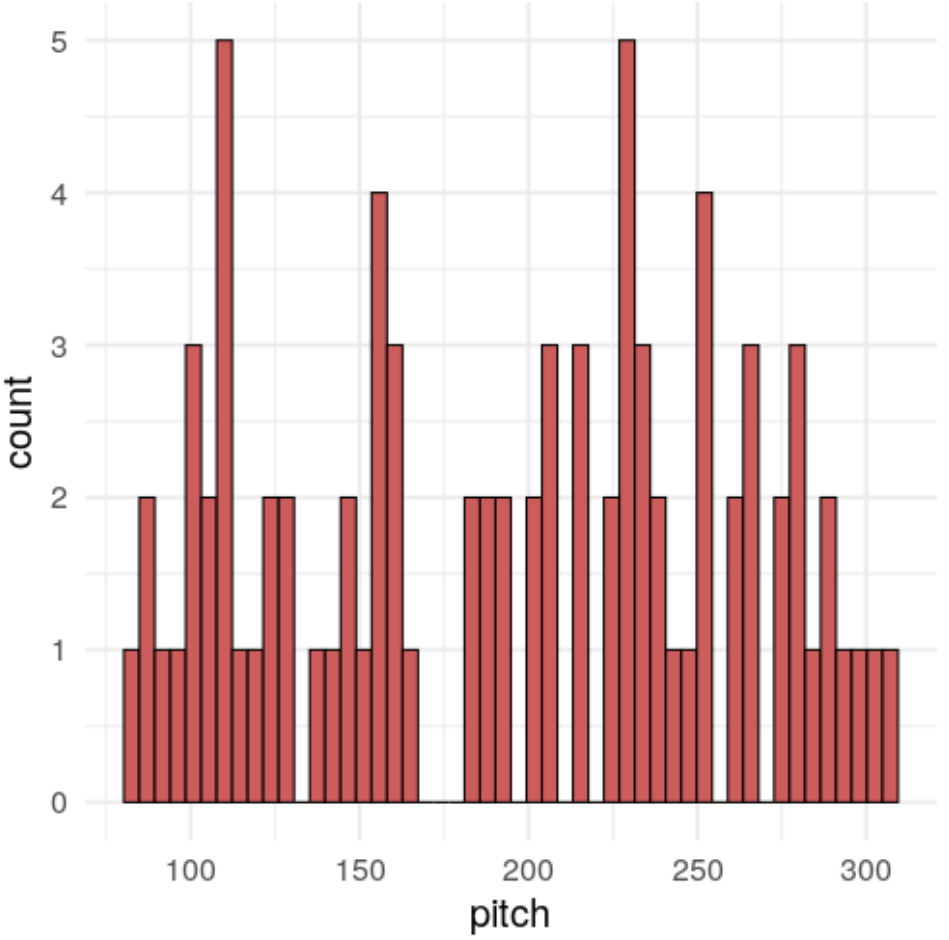
```
df_inf <- filter(df, context == 'inf')
nrow(df_inf)
```

```
## [1] 42
```

```
df_pol <- filter(df, context == 'pol')
nrow(df_pol)
```

```
## [1] 41
```

Descripció de variables (to)



Descripció de variables (to)

```
mean(df$pitch)
```

```
## [1] 193.5819
```

```
median(df$pitch)
```

```
## [1] 203.9
```

```
sd(df$pitch)
```

```
## [1] 65.54068
```

```
quantile(df$pitch)
```

```
##      0%     25%     50%     75%    100%
##   82.20 131.55 203.90 248.55 306.80
```

Possible relació entre predictor i variable (gènere)

Possible relació entre predictor i variable (gènere)

Vols saber quins factors (no) afecen el to mitjà (Hertz) d'una persona. Un col·lega t'ha donat les dades següents:

https://tinyurl.com/polite-data

1. Fes ús de mètodes descriptius per descriure la relació entre predictor (gènere) i resultat

Possible relació entre predictor i variable (gènere)

```
mean(df_m$pitch)
```

```
## [1] 138.8756
```

```
mean(df_f$pitch)
```

```
## [1] 246.9857
```

```
median(df_m$pitch)
```

```
## [1] 126.9
```

```
median(df_f$pitch)
```

```
## [1] 248.55
```

Possible relació entre predictor i variable (gènere)

```
sd(df_m$pitch)
```

```
## [1] 38.92821
```

```
sd(df_f$pitch)
```

```
## [1] 34.61808
```

```
quantile(df_m$pitch)
```

```
##     0%    25%    50%    75%   100%
##   82.2 108.2 126.9 160.7 229.0
```

```
quantile(df_f$pitch)
```

```
##       0%      25%      50%      75%    100%
## 154.800 227.825 248.550 276.450 306.800
```

Vols saber quins factors (no) afecen el to mitjà (Hertz) d'una persona. Un col·lega t'ha donat les dades següents:

https://tinyurl.com/polite-data

1. Crea un model de regressió lineal que prediu to a base de gènere.

2. Què suggereix el model sobre la relació entre predictor i resultat?

Model gènere

```
pitch_model1 <- lm(formula = pitch ~ 1 + gender,
                   data    = df)
summary(pitch_model1)
```

```
##
## Call:
## lm(formula = pitch ~ 1 + gender, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -92.186 -28.426  -2.676  23.124  90.124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  246.986      5.680   43.48   <2e-16 ***
## genderM     -108.110      8.081  -13.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.81 on 81 degrees of freedom
## Multiple R-squared:  0.6884,    Adjusted R-squared:  0.6846
## F-statistic:   179 on 1 and 81 DF,  p-value: < 2.2e-16
```

Model gènere

```
coef(pitch_model1)
```

```
## (Intercept)     genderM
##    246.9857   -108.1101
```

```
mean(df_f$pitch)
```

```
## [1] 246.9857
```

```
mean(df_m$pitch)
```

```
## [1] 138.8756
```

Possible relació entre predictor i variable (context)

Possible relació entre predictor i variable (context)

Possible relació entre predictor i variable (context)

```
mean(df_pol$pitch)
```

```
## [1] 184.3561
```

```
mean(df_inf$pitch)
```

```
## [1] 202.5881
```

```
median(df_pol$pitch)
```

```
## [1] 193.4
```

```
median(df_inf$pitch)
```

```
## [1] 209.05
```

Possible relació entre predictor i variable (context)

```
sd(df_pol$pitch)
```

```
## [1] 63.55659
```

```
sd(df_inf$pitch)
```

```
## [1] 66.94803
```

```
quantile(df_pol$pitch)
```

```
##     0%   25%   50%   75%  100%
##   82.2 126.5 193.4 232.6 289.4
```

```
quantile(df_inf$pitch)
```

```
##      0%     25%     50%     75%    100%
##   99.10 138.65 209.05 259.70 306.80
```

Vols saber quins factors (no) afecen el to mitjà (Hertz) d'una persona. Un col·lega t'ha donat les dades següents:

https://tinyurl.com/polite-data

1. Crea un model de regressió lineal que prediu to a base de context

2. Què suggereix el model sobre la relació entre predictor i resultat?

Model context

```
pitch_model2 <- lm(formula = pitch ~ 1 + context,
                   data    = df)
summary(pitch_model2)
```

```
##
## Call:
## lm(formula = pitch ~ 1 + context, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.488  -62.122    9.044   51.178  105.044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   202.59      10.08  20.107   <2e-16 ***
## contextpol    -18.23      14.34  -1.272    0.207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 65.3 on 81 degrees of freedom
## Multiple R-squared:  0.01958,    Adjusted R-squared:  0.007475
## F-statistic: 1.618 on 1 and 81 DF,  p-value: 0.2071
```

Model context

```
coef(pitch_model2)
```

```
## (Intercept)  contextpol
##    202.5881    -18.2320
```

```
mean(df_pol$pitch)
```

```
## [1] 184.3561
```

```
mean(df_inf$pitch)
```

```
## [1] 202.5881
```

Per què estem creant un model per descobrir una cosa que ja ens indica l'estadística descriptiva?

1. Entre altres: error estàndard; R^2; i residuals (cf. `pitch_model1` vs. `pitch_model2`)
2. Perquè es poden expandir a més predictors

Model amb ambdós predictors

```
pitch_model3 <- lm(formula = pitch ~ 1 + gender + context,
                   data    = df)
summary(pitch_model3)
```

```
##
## Call:
## lm(formula = pitch ~ 1 + gender + context, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.409 -26.561  -4.262  24.690 100.140
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  256.762      6.756  38.006   <2e-16 ***
## genderM     -108.349      7.833 -13.832   <2e-16 ***
## contextpol   -19.553      7.833  -2.496   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.68 on 80 degrees of freedom
## Multiple R-squared:  0.7109,    Adjusted R-squared:  0.7037
```

Tots els nostres models (paràmetres)

```
summary(pitch_model3)$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  256.76238   6.755918  38.005550 5.752326e-53
## genderM     -108.34856   7.832968 -13.832376 6.398784e-23
## contextpol   -19.55332   7.832968  -2.496285 1.460499e-02
```

```
summary(pitch_model2)$coefficients
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 202.5881   10.07528 20.107444 3.125044e-33
## contextpol  -18.2320   14.33521 -1.271833 2.070720e-01
```

```
summary(pitch_model1)$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  246.9857   5.679855  43.48451 6.397703e-58
## genderM     -108.1101   8.081359 -13.37771 3.271928e-22
```

Tots els nostres models (goodness of fit)

```
summary(pitch_model3)$r.squared
```

```
## [1] 0.7109337
```

```
summary(pitch_model2)$r.squared
```

```
## [1] 0.01957888
```

```
summary(pitch_model1)$r.squared
```

```
## [1] 0.6884175
```

Resum: regressió lineal

- Estimació de relació lineal entre un o més predictors i un resultat.
- Predicció de resultat a base de predictors (+ error)
- Permet relacions més complexes

$$i = \beta_0 + \beta_1(\text{context} \times \text{gender})$$

$$i = \beta_0 + \beta_1 \log(\text{age})$$

- Estimació de (i) error de prediccions; (ii) incertesa sobre paràmetres; (iii) efecte condicional entre paràmetres

(In)significància estadística

Significància estadística

Intuïció: un resultat és estadísticament significant quan és improbable obtenir un resultat així o més extrem sota la hipotesi nul·la en repetir l'experiment ad infinitum

- El que compta com a *improvable* (deuria) depèn(dre) del context de la investigació. Un nombre comú a les ciències socials és sota el 5%; però pot (i deu) ser molt més baix.
- La *hipòtesi nul·la* és, comunament, el contrari al que un vol demostrar: No hi ha diferència entre grup A i grup B; l'efecte d'A a B és 0; etc.
- *Així o més extrem* es refereix a: una diferència entre grup A i B (d'una mida suficient o més gran per decidir que hi ha una diferència); l'efecte d'A a B sent més gran (o molt més gran) a 0; o menor (o molt menor) a 0; etc.

(In)significància estadística: raons conceptuals per ser crítics

- En general, no ens interessa si A i B són diferents sinó que tan diferents són
- En general, no (només) ens interessa si l'efecte de B a A és positiu, si no com de positiu
- ...

(In)significància estadística: raons tècniques per ser crítics

No trobar un efecte significant és **sempre** un problema de mida de mostra.

Amb una mostra prou gran tots els efectes són significatius

(In)significància estadística: raons tècniques per ser crítics

```r
set.seed(12)  #random seed
obs <- 5      #how many observations

datos_A <-   rnorm(n = obs, mean = 160, sd = 5)
datos_B <-   rnorm(n = obs, mean = 159, sd = 4)

id_grupo <-  c( rep('A',obs),
                rep('B', obs)
              )

#Joining data in a data frame with two columns
df <- data.frame(dades  = c(datos_A,datos_B),
                 grup  = id_grupo)
```

(In)significància estadística: raons tècniques per ser crítics

```
summary(lm(dades ~ grup, data = df))
```

```
##
## Call:
## lm(formula = dades ~ grup, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2104 -1.5996 -0.6841  0.1238 11.6636
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  156.222      2.231  70.025 1.93e-12 ***
## grupB          2.062      3.155   0.654    0.532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.989 on 8 degrees of freedom
## Multiple R-squared:  0.0507,    Adjusted R-squared:  -0.06796
## F-statistic: 0.4273 on 1 and 8 DF,  p-value: 0.5317
```

(In)significància estadística: raons tècniques per ser crítics

```
set.seed(12)  #random seed
obs <- 1000000       #how many observations

datos_A <-   rnorm(n = obs, mean = 160, sd = 5)
datos_B <-   rnorm(n = obs, mean = 159, sd = 4)

id_grupo <-  c( rep('A',obs),
                rep('B', obs)
              )

#Joining data in a data frame with two columns
df <- data.frame(dades  = c(datos_A,datos_B),
                 grup  = id_grupo)
```

(In)significància estadística: raons tècniques per ser crítics

```
summary(lm(dades ~ grup, data = df))
```

```
##
## Call:
## lm(formula = dades ~ grup, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.2465  -3.0075   0.0018   3.0035  25.4012
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 159.998657   0.004531   35313   <2e-16 ***
## grupB        -0.999789   0.006408    -156   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.531 on 1999998 degrees of freedom
## Multiple R-squared:  0.01203,    Adjusted R-squared:  0.01203
## F-statistic: 2.435e+04 on 1 and 1999998 DF,  p-value: < 2.2e-16
```

Propera sessió

- Lliurament de "Assignment 4" (08:00 AM 16/05)

- **Models lineals generalitzats I**