# 5 Regresión con más de un predictor

Métodos empíricos 2

10/05/2022

# Hoy

- Múltiples predictores
- Caso de estudios (continuado)

---

- Recolección y generación de datos
- (In)significancia estadística

# Múltiples predictores

# Misma formula, más sumandos

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

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

# Caso de estudios: tono

```
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
```

---

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

# Descripción de variables (género)

```
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
```

# Descripción de variables (contexto)

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

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

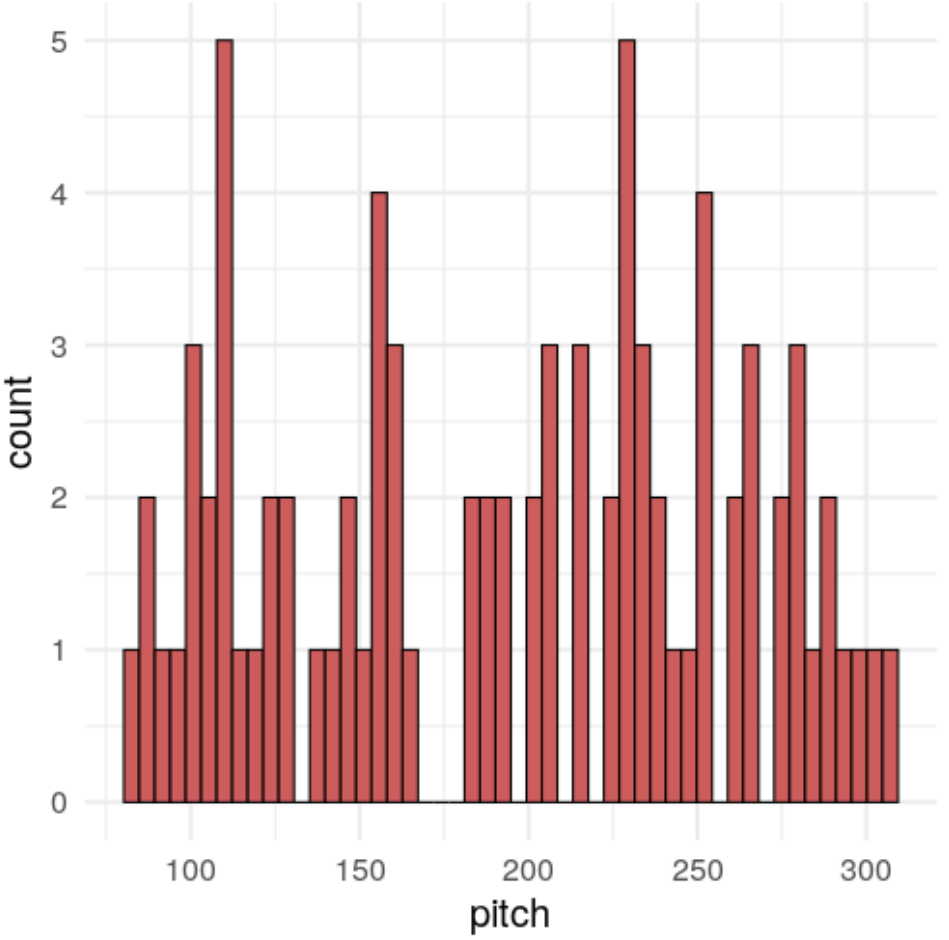
```
df_inf <- filter(df, context == 'inf')
nrow(df_inf)
```

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

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

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

# Descripción de variables (tono)



# Descripción de variables (tono)

```
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
```

# Posible relación entre predictor y variable (género)

# Posible relación entre predictor y variable (género)

# Posible relación entre predictor y variable (género)

```
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
```

# Posible relación entre predictor y variable (género)

```
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
```

# Modelo género

```
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
```

# Modelo género

```
coef(pitch_model1)
```

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

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

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

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

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

# Posible relación entre predictor y variable (contexto)

# Posible relación entre predictor y variable (contexto)

# Posible relación entre predictor y variable (contexto)

```
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
```

# Posible relación entre predictor y variable (contexto)

```
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
```

# Modelo contexto

```
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
```

# Modelo contexto

```
coef(pitch_model2)
```

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

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

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

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

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

# Por qué estamos creando un modelo para descubrir algo que ya nos indica la estadística descriptiva?

1. Entre otros: error estándar; $R^2$; y residuales (cf. `pitch_model1` vs. `pitch_model2`)

2. Porque se pueden expander a más predictores

# Modelo con ambos predictores

```
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
```

# Todos nuestros modelos (parámetros)

```
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
```

# Todos nuestros modelos (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
```

# Resumen: regresión lineal

- Estimación de relación lineal entre uno o más predictores y un resultado.
- Predicción de resultado a base de predictores (+ error)
- Permite relaciones más complejas

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

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

- Estimación de (i) error de predicciones; (ii) incertidumbre sobre parámetros; (iii) efecto condicional entre parámetros

# Recolección y generación de datos

# (In)significancia estadística

# Significancia estadística

Intuición: un resultado es estadísticamente significante cuando es improbable obtener un resultado así o más extremo bajo la hipotesis nula al repetir el experimento ad infinitum

- Lo que cuenta como *improbable* (debería) depende(r) del contexto de la investigación. Un número común en las ciencias sociales es bajo 5%; pero puede (y debe?) ser mucho más bajo.
- La *hipótesis nula* es, comúnmente, lo contrario a lo que uno quiere demostrar: No hay efecto diferencia entre grupo A y grupo B; el efecto de A en B es 0; el efecto de A en B es positivo/negativo; etc.
- *Así o más extremo* se refiere a: una diferencia entre grupo A y B (de un tamaño suficiente o mayor para decidir que hay una diferencia); el efecto de A en B siendo mayor (o mucho mayor) a 0; o menor (o mucho menor) a 0; el efecto de A en B siendo positivo/negativo (o muy positivo/negativo); etc.

# (In)significancia estadística: razones conceptuales para ser críticos

- En general, no nos interesa si A y B son diferentes si no que tan diferentes son
- En general, no nos interesa si el efecto de B en A es positivo, si no cuan positivo
- ...

# (In)significancia estadística: razones técnicas para ser críticos

No encontrar un efecto significante es **siempre** un problema de tamaño de muestra.

Con una muestra suficientemente grande todos los efectos son significativos

# (In)significancia estadística: razones técnicas para ser críticos

```r
set.seed(12)  #semilla aleatoria
obs <- 5      #cuantas observaciones

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)
              )

#Juntando los datos en un data frame de dos columnas
df <- data.frame(datos  = c(datos_A,datos_B),
                 grupo  = id_grupo)
```

# (In)significancia estadística: razones técnicas para ser críticos

```
summary(lm(datos ~ grupo, data = df))
```

```
##
## Call:
## lm(formula = datos ~ grupo, 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 ***
## grupoB         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)significancia estadística: razones técnicas para ser críticos

```r
set.seed(12)   #semilla aleatoria
obs <- 1000000 #cuantas observaciones

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)
              )

#Juntando los datos en un data frame de dos columnas
df <- data.frame(datos   = c(datos_A,datos_B),
                 grupo   = id_grupo)
```

# (In)significancia estadística: razones técnicas para ser críticos

```
summary(lm(datos ~ grupo, data = df))
```

```
##
## Call:
## lm(formula = datos ~ grupo, 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 ***
## grupoB       -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
```

# Próxima sesión

- Entrega de "Assignment 5" (08:00 AM 17/05)

---

- **Modelos lineales generalizados I**

---

- Ejercicio práctico: 17/05 - 24/05
- Entrega parte II de "Revisión por pares": 24/05 - 31/05
- Informe final: 28/06