Último tema del curso

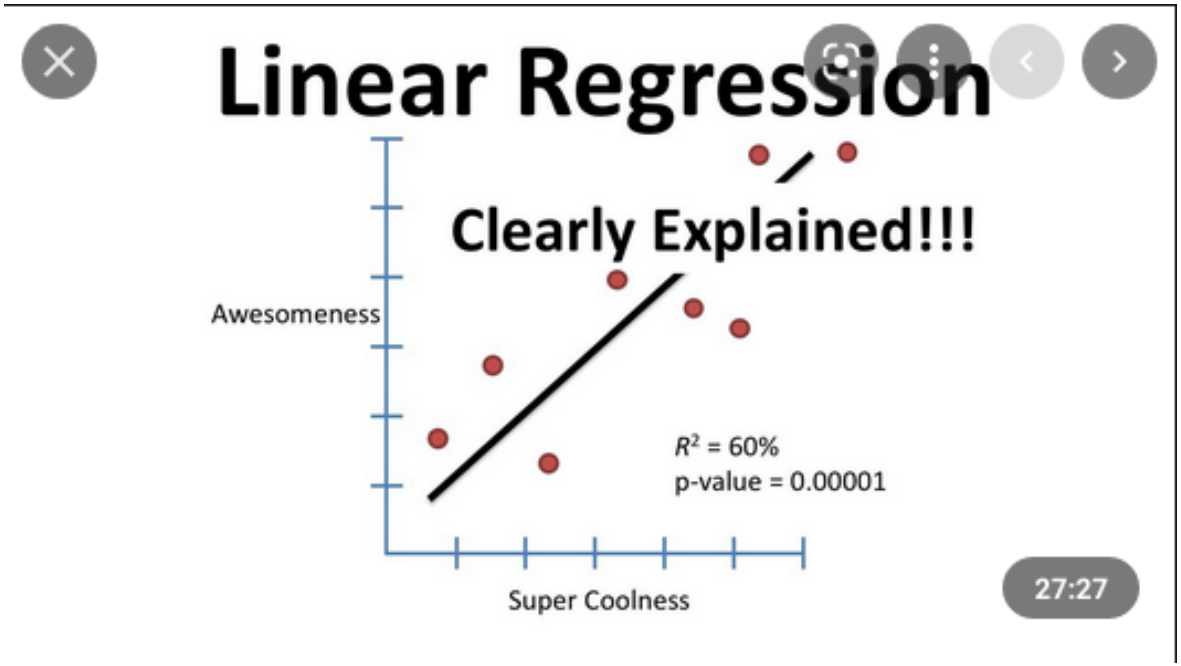


Figura: El modelo estadístico privilegiado de la Ciencia de datos

Clase 24 de Bioestadística

Modelos lineales

Ernesto Mordecki

CMAT, Facultad de Ciencias, Universidad de la República.

Uruguay

Contenidos de la clase

Regresión lineal: mínimos cuadrados

Estimación de $\hat{a}_n$ y $\hat{b}_n$

Regresión lineal: mínimos cuadrados

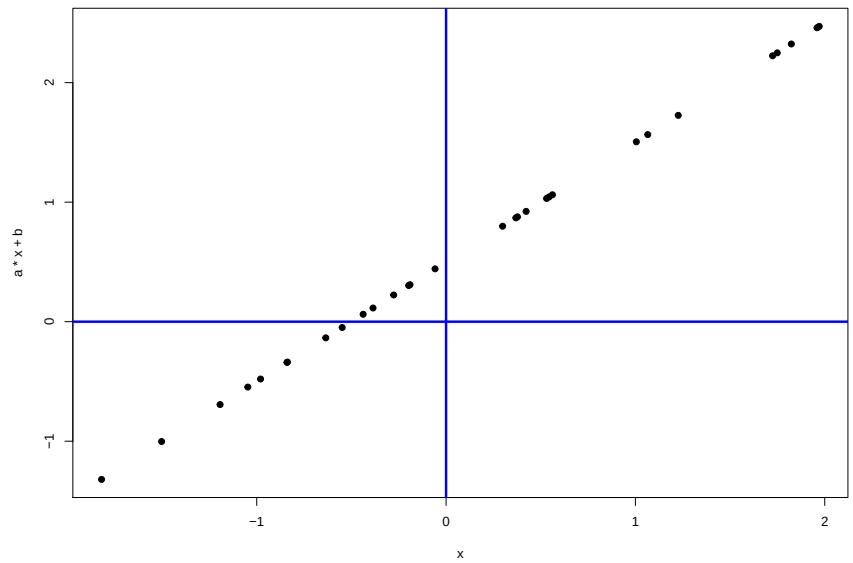
- Queremos modelar la dependencia entre dos o más variables.
- Por ejemplo, si X es una variable aleatoria muy relacionada a Y, podemos postular un modelo lineal de la forma

$$Y = aX + b,$$

- Aquí a y b son constantes que desconocemos y queremos estimar.
- La dependencia es absoluta entre las variables X e Y: sabiendo el valor de X y las constantes a y b podemos calcular sin error el valor de Y.

Aquí se muestra el modelo

$$Y = \frac{1}{2}X + 1$$



En este caso

$$a = \frac{Y_n - Y_1}{X_n - X_1}, \qquad b = Y_1 - bX_1$$

Pero este no es muestro problema

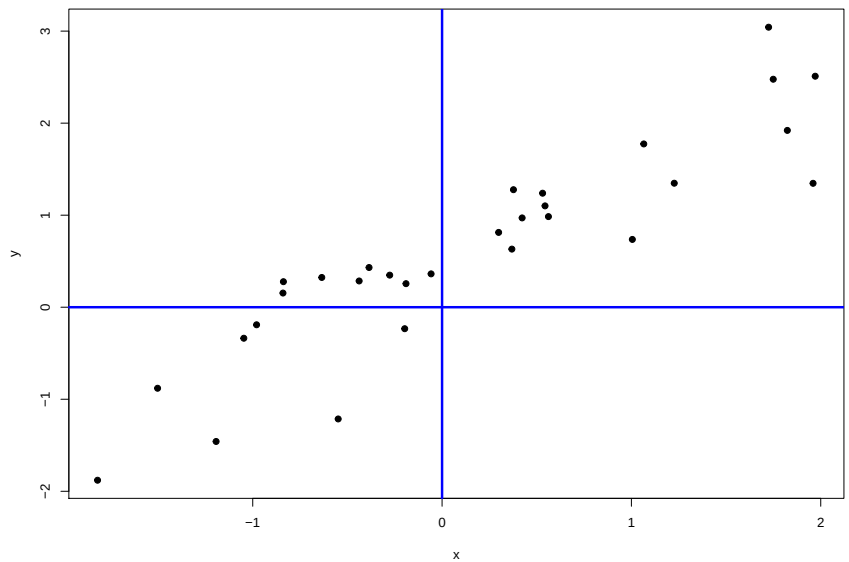
- Sin embargo, en general, el efecto de la variable X sobre la Y no sigue este modelo perfectamente lineal sino que aparece un error ε, y escribimos

$$Y = aX + b + \varepsilon$$

- El error ε es desconocido, y puede deberse
  - a errores de medición
  - Más en general, existen otras variables que no incluímos en el modelo

Aquí se muestra el modelo

$$Y = 1 + \frac{1}{2}X + \varepsilon$$



Este **es** nuestro problema

Modelo con efectos fijos

- Vamos a suponer que los X que los pensamos como no aleatorios, cada uno de los cuales tiene asociado un valor de la variable Y,
- Este es el modelo de efectos fijos
- El modelo de efectos aleatorios incluye aleatoriedad además en la X
- Por ejemplo X puede ser el género de un individuo, su edad, su peso, mientras que Y puede ser su salario en \$.

- Suponer efectos fijos tiene que ver con el diseño del experimento en si, con como son los datos, etc.
- Es el primer paso en la modelación: muchas veces es una hipótesis poco razonable,
- Suponemos entonces que las X son aleatorias con cierta distribución
- En estos casos se asume generalmente que X y ε son independientes.

El problema estadístico

- Tenemos $(X_1, Y_1), \ldots, (X_n, Y_n)$ de pares que se asumen independientes e idénticamente distribuidos
- Suponemos que todas tienen la distribución del par (X, Y),

  $P(X_k \leq a, Y_k \leq b) = P(X \leq a, Y \leq b)$ para todo $k = 1, \ldots, n$

- Suponemos entonces que las X son determinísticas
- Suponemos que los errores ε_k son normales con media 0 y varianza σ^2.
- Las constantes a y b no se conocen.

- Nuestro modelo es

$$Y_k = aX_k + b + \varepsilon_k \quad \text{para todo } k = 1, \ldots, n.$$

- El primer problema es producir estimadores $\hat{a}_n$ de a, y $\hat{b}_n$ de b basados en dicha muestra.

- En este caso, el valor que "predice"nuestro modelo con los parámetros estimados: dado un dato X_k de la muestra estimamos la Y_k correspondiente:

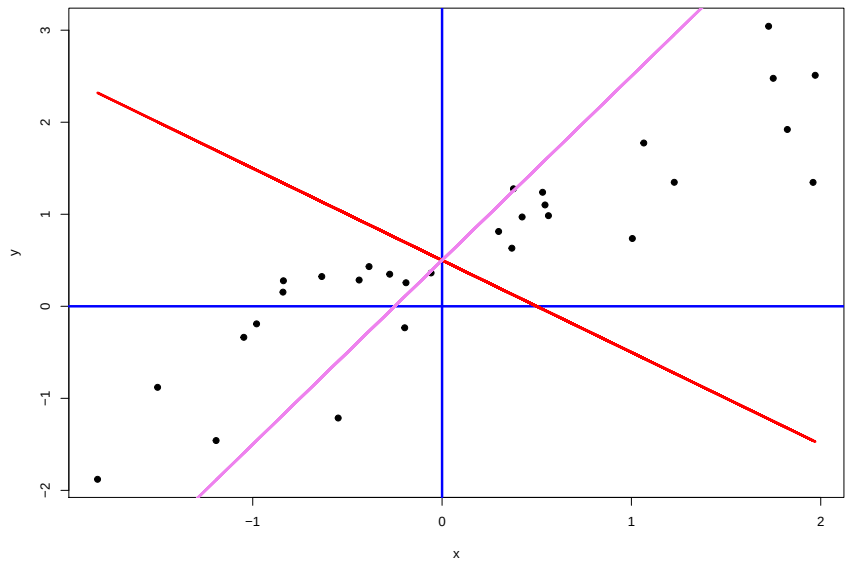
$$\hat{Y}_k = \hat{a}_n X_k + \hat{b}_n.$$

- Aquí el error ε_k no aparece, porque no se conoce,

- La diferencia entre el valor observado Y_k y el valor estimado $\hat{Y}_k$ se conoce como residuo k-ésimo.

Estimación de a y b

La idea es encontrar la recta $y = ax + b$ que mejor se ajusta a los datos:



¿Qué recta elegimos? ¿Cómo estimamos a y b?

Estimación de $\hat{a}_n$ y $\hat{b}_n$

- Buscamos la recta $y = ax + b$ que más se parezca a los datos
- Como conocemos los valores sólo en X_k, lo que queremos es que

$$Y_k \approx aX_k + b$$

- Es decir, queremos que

$$Y_k - aX_k - b \approx 0$$

- Como hicimos ya varias veces, para que los signos no compensen, pedimos

$$(Y_k - aX_k - b)^2 \approx 0$$

- En realidad, si los datos no están en una recta, no va a ser posible encontrar a y b tal que las diferencias sean cero.
- Nos resignamos con encontrar los a y b que den más cerca de cero
- Buscamos el par $(\hat{a}_n, \hat{b}_n)$ que minimice la suma de los cuadrados:

$$S(a,b) = \sum_{k=1}^{n} (Y_k - aX_k - b)^2 \tag{1}$$

- Entonces el par (a, b) va a ser un mínimo absoluto de la función

$$S(a, b)$$

- Ojo: ahora pensamos que los (X_k, Y_k) están fijos, son los datos que observamos, y a, b son las variables

- Cómo se hallan los mínimos de una función de dos variables?

- La condición necesaria de mínimo es que las derivadas parciales[1] se anulen.

- Es decir, buscamos a y b tales que

$$\frac{\partial S(a, b)}{\partial a} = \frac{\partial S(a, b)}{\partial b} = 0$$

[1] Mmm, ¿que era eso?

Las cuentas

- La derivada respecto de a:

$$\frac{\partial}{\partial a} S(a,b) = \sum_{k=1}^{n} \frac{\partial}{\partial a} (Y_k - aX_k - b)^2$$
$$= \sum_{k=1}^{n} (Y_k - aX_k - b)(-2X_k) = 0$$

- La derivada respecto de b:

$$\frac{\partial}{\partial b} S(a,b) = \sum_{k=1}^{n} \frac{\partial}{\partial b} (Y_k - aX_k - b)^2$$
$$= -2 \sum_{k=1}^{n} (Y_k - aX_k - b) = 0$$

Las cuentas

- La derivada respecto de a:

$$\begin{aligned}\frac{\partial}{\partial a}S(a,b) &= \sum_{k=1}^{n}\frac{\partial}{\partial a}(Y_k - aX_k - b)^2 \\ &= \sum_{k=1}^{n}(Y_k - aX_k - b)(-2X_k) \\ &= (-2)\left[(\sum_{k=1}^{n}X_kY_k) - a(\sum_{k=1}^{n}X_k^2) - b(\sum_{k=1}^{n}X_k)\right] = 0\end{aligned}$$

Trabajamos la segunda derivada

- $$\begin{aligned}\frac{\partial}{\partial b}S(a,b) &= \sum_{k=1}^{n}\frac{\partial}{\partial b}(Y_k - aX_k - b)^2 \\ &= -2\sum_{k=1}^{n}(Y_k - aX_k - b) \\ &= -2\left[\sum_{k=1}^{n}Y_k - a\sum_{k=1}^{n}X_k - n\times b\right] = 0\end{aligned}$$

- Como las variables son a y b, lo que obtenemos es un sistema de dos ecuaciones lineales con dos incógnitas:

$$(\sum_{k=1}^{n} X_k Y_k) - a(\sum_{k=1}^{n} X_k^2) - b(\sum_{k=1}^{n} X_k) = 0$$

$$\sum_{k=1}^{n} Y_k - a\sum_{k=1}^{n} X_k - nb = 0$$

- Dividimos todo por n y usamos las notaciones

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k, \quad \overline{Y}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k$$

$$\overline{X^2}_n = \frac{1}{n}\sum_{k=1}^{n} X_k^2, \quad \overline{Y^2}_n = \frac{1}{n}\sum_{k=1}^{n} Y_k^2 \quad \overline{XY}_n = \frac{1}{n}\sum_{k=1}^{n} X_k Y_k$$

- El sistema de ecuaciones queda:

$$a\overline{X^2}_n + b\overline{X}_n = \overline{XY}_n$$
$$a\overline{X}_n + b = \overline{Y}_n$$

- Despejando

$$\hat{a}_n = \frac{\overline{XY}_n - \overline{X}_n\overline{Y}_n}{\overline{X^2}_n - (\overline{X}_n)^2}$$
$$\hat{b}_n = \overline{Y}_n - \hat{a}_n\overline{X}_n$$

Ejemplo numérico

- Calculamos

$$\overline{X}_n = 0{,}156, \quad \overline{Y}_n = 0{,}614$$

$$\overline{X^2}_n = 1{,}13, \quad \overline{Y^2}_n = 1{,}63 \quad \overline{XY}_n = 1{,}16$$

- Obtenemos como resultado

$$\hat{a}_n = 0{,}966, \qquad \hat{b} = 0{,}463$$

- En R hicimos las siguiente cuentas. Primero simulamos el modelo

```
set.seed(321)
n<-30
x<-runif(n,-2,2)
sigma<-0.5
e<-rnorm(n,0,sigma)
a<-1
b<-0.5
y<-a*x+b+e
```

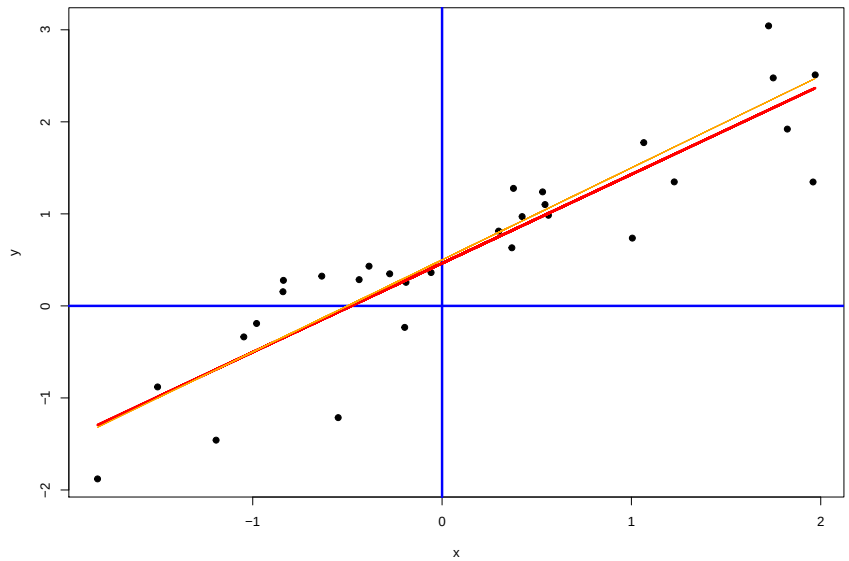
- Para calcular los promedios:

```
# promedio de las x
barx<-mean(x)
# promedio de las y
bary<-mean(y)
# promedio de las x por las y
barxy<-mean(x*y)
# promedio de los x al cuadrado
barx2<-mean(x*x)
# promedio de los y al cuadrado
bary2<-mean(y*y)
```

- Para calcular los estimadores:

```
# calculo el a estimado con la formula:
hata<-(barxy-barx*bary)/(barx2-(barx)^2)
# calculo el b estimado con la formula:
hatb<-bary-hata*barx
```

Veamos la solución:



- En rojo la recta estimada, con $\hat{a}_n = 0{,}97$ y $\hat{b}_n = 0{,}46$
- En naranja la recta teórica con $a = 1$ y $b = 1/2$