Modelos Estadísticos para la Regresión y la Clasificación

Clase 5: Regresión lineal

Mathias Bourel

Instituto de Matemática y Estadística Prof. Rafael Laguardia (IMERL)
Facultad de Ingeniería, Universidad de la República, Uruguay

4 de septiembre de 2024

Plan

Motivación

Recordamos que el objetivo del aprendizaje supervisado en un problema de regresión consiste en buscar una función f que minimice el riesgo teórico

$$\mathbb{E}\left[\left(Y - f(X)\right)^2\right]$$

La función f que minimiza esta función es la esperanza condicional $m(X) = \mathbb{E}(Y|X)$.

$$\mathbb{E}(Y|X) = \underset{f}{\text{Argmin}}\ \mathbb{E}\left[\left(Y - f(X)\right)^2\right]$$

La función $m(X)$ se conoce también como función de *regresión* de Y sobre X.

En efecto si $f : \mathbb{R}^d \to \mathbb{R}$ entonces:

$$\mathbb{E}\left[(f(X)-Y))^2\right] = \mathbb{E}\left[(f(X)-m(X)+m(X)-Y))^2\right] = \mathbb{E}\left[(f(X)-m(X))^2\right] + \mathbb{E}\left[(m(X)-Y))^2\right]$$

y es claro que esta expresión es mínima cuando $f(x) = m(x) = \mathbb{E}(Y|X=x)\ \forall\, x$.

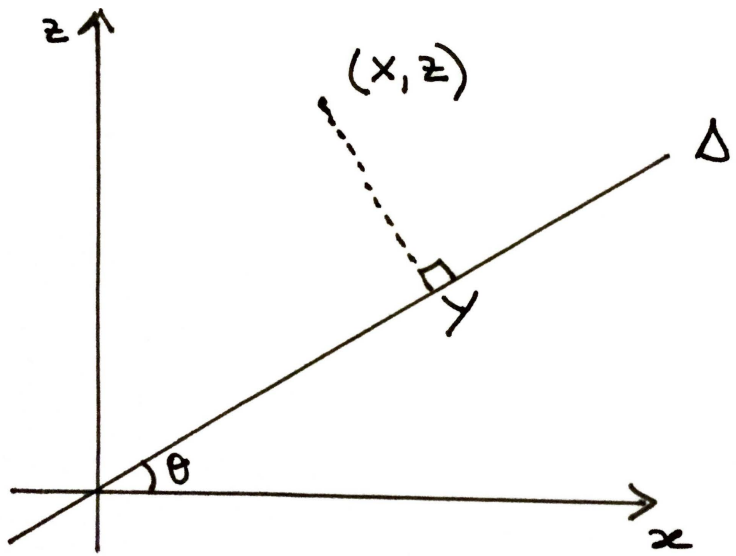
En el cálculo anterior usamos que:
$\mathbb{E}_{(X,Y)}\left[(f(X)-m(X)+m(X)-Y))^2\right] =$
$\mathbb{E}_{(X,Y)}\left[(f(X)-m(X))^2\right] + \mathbb{E}_{(X,Y)}\left[(m(X)-Y)^2\right] + 2\mathbb{E}_{(X,Y)}((f(X)-m(X))(m(X)-Y))$ y el último término se anula porque

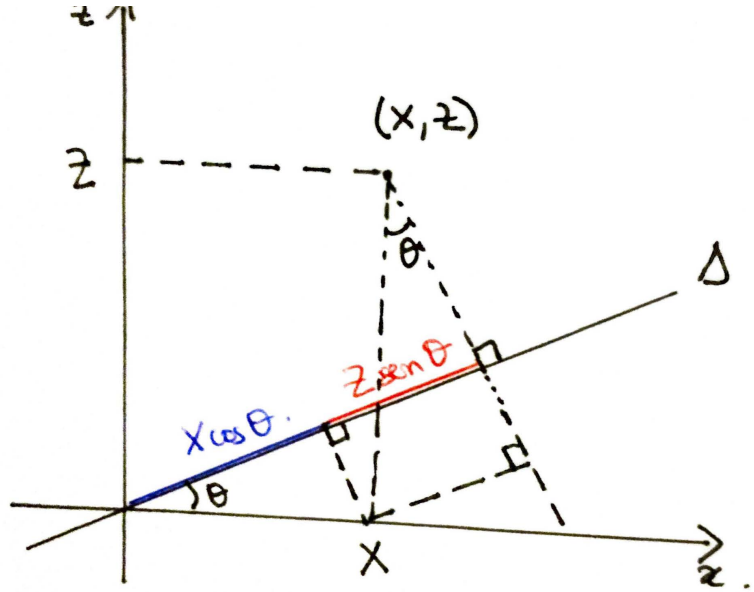
$$\mathbb{E}_{(X,Y)}((f(X)-m(X))(m(X)-Y)) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}((f(X)-m(X))(m(X)-Y)|X)\right]$$

$$= \mathbb{E}_X\left[(f(X)-m(X))\mathbb{E}_{Y|X}(m(X)-Y|X)\right] = \mathbb{E}_X\left[(f(X)-m(X))(m(X)-\underbrace{\mathbb{E}_{Y|X}(Y|X)}_{m(X)})\right] = 0$$

Motivación: enfoque probabilístico

Sean X y Z dos variables aleatorias independientes normales estadándares ($X, Z \sim \mathcal{N}(0,1)$). Entonces (X, Z) es un vector aleatorio con distribución normal estándar. Consideramos la recta Δ que pasa por el origen y que forma un ángulo de θ con el eje (Ox). Sea Y la proyección de (X, Z) sobre Δ. En realidad identificamos Y con la norma del vector OY.





Por lo tanto $Y = X\cos\theta + Z\sin\theta$ y entonces al ser combinación lineal de dos normales, Y tiene distribución normal. Más aún $Y \sim \mathcal{N}(0,1)$ ya que:

$$\mathbb{E}(Y) = \mathbb{E}(X\cos\theta + Z\sin\theta) = \cos\theta\mathbb{E}(X) + \sin\theta\mathbb{E}(Z) = 0$$

$$\text{Var}(Y) = \text{Var}(X\cos\theta + Z\sin\theta) = \cos^2\theta\text{Var}(X) + \sin^2\theta\text{Var}(Z) = cos^2\theta + \sin^2\theta = 1$$

Motivación: enfoque probabilístico

El coeficiente de correlación entre X e Y es entonces:

$$\rho = \rho_{X,Y} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \text{Cov}(X,Y) = \mathbb{E}(XY) = \mathbb{E}(X(X\cos\theta + Z\sin\theta)) = \mathbb{E}(X^2)\cos\theta = \cos\theta$$

Para cada $\rho \in [-1,1]$, existe un ángulo θ tal que $\rho = \cos\theta$. Luego, para cada $\rho \in [-1,1]$ existen X e Y normales estándar con correlación ρ. Podemos escribir entonces:

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

en donde X e Z son normal estándares independientes.

Definición:

1. Decimos que el par (X,Y) tiene distribución normal estándar con correlación ρ si existe Z normal estándar independiente de X tal que

$$Y = \rho X + \sqrt{1-\rho^2}Z$$

2. Decimos que el par (X,Y) tiene distribución normal con parámetros $(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho)$ si $\left(\frac{X-\mu_X}{\sigma_X}, \frac{Y-\mu_Y}{\sigma_Y}\right)$ tiene distribución normal estándar con correlación ρ.

Motivación

De la definición anterior se desprende que existe una variable aleatoria Z normal estándar, independiente de X tal que

$$\frac{Y-\mu_Y}{\sigma_Y} = \rho\frac{X-\mu_X}{\sigma_X} + \sqrt{1-\rho^2}Z$$

es decir:

$$Y = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X) + \sqrt{1-\rho^2}Z$$

Veamos como queda en este caso la función de regresión:

$$\begin{aligned} m(x) =& \mathbb{E}(Y|X=x) = \mathbb{E}\left(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(X-\mu_X) + \sigma_Y\sqrt{1-\rho^2}Z|X=x\right) \\ =& \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) + \sigma_Y\sqrt{1-\rho^2}\mathbb{E}\left(Z|X=x\right) \\ =& \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) + \sigma_Y\sqrt{1-\rho^2}\mathbb{E}\left(Z\right) \\ =& \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X) \end{aligned}$$

ya que Z es independiente de X. Este razonamiento muestra que $m(x)$ es una función lineal en el caso que (X, Y) sea un vector normal bivariado. Observar que

$$m(x) = \beta_0 + \beta_1 x \quad \text{donde} \quad \beta_0 = \mathbb{E}(Y) - \beta_1\mathbb{E}(X) \quad \text{y} \quad \beta_1 = \frac{Cov(X,Y)}{\sigma_X^2}$$

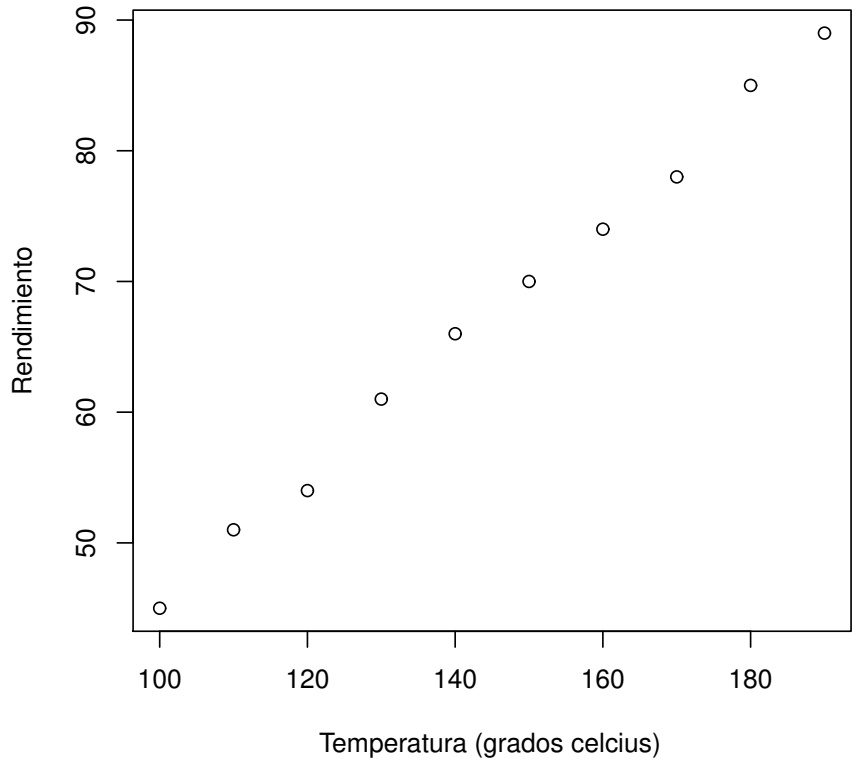
Plan

Regresión lineal simple. Primer Ejemplo

Objetivo: Establecer una relación entre una variable dependiente Y y una variable independiente x para poder hacer predicciones sobre Y cuando se conoce a x.

Ejemplo: Rendimiento de un producto químico en función de la temperatura.

Temp($^\circ C$)	Rend (%)
100	45
110	51
120	54
130	61
140	66
150	70
160	74
170	78
180	85
190	89



Se quiere expresar por medio de una ecuación la relación entre las variables x e y, mediante $y = f(x)$ con f a determinar. La gráfica sugiere una relación lineal.

Regresión lineal simple. Primer ejemplo

Planteo del modelo lineal:

La obtención de una ecuación exacta $y = f(x)$ no siempre es posible e y puede depender de otros factores (*fenómenos aleatorios*). Se tendrá entonces un *error aleatorio* ϵ debido a variables y a factores no tenidos en cuenta, obteniendo de esta manera un modelo probabilístico para nuestro problema:

$$Y = f(x) + \epsilon$$

siendo ϵ el error aleatorio.

Volviendo a nuestro problema, nos proponemos hallar un modelo del tipo:

$$Y = \underbrace{\beta_0 + \beta_1 x}_{f(x)=x'\beta} + \epsilon$$

donde

- Y es la variable aleatoria dependiente, que se querrá predecir,
- x es la variable independiente, que se usa para predecir,
- β_0 y β_1 son parámetros desconocidos.
- ϵ es un error aleatorio.

Planteo del modelo lineal:

Buscamos entonces la "mejor recta" según algún criterio de manera que pase lo más cerca posible de los puntos. En este contexto, el experto elije varios valores $x_1, \ldots, x_n$ de la variable X y observa los valores correspondientes $y_1, \ldots, y_n$ de la variable aleatoria Y.
Queremos hallar $\widehat{\beta_0}$ y $\widehat{\beta_1}$, estimadores de β_0 y β_1, que minimizan la suma de los errores cometidos al cuadrado:

$$\sum_{i=1}^{n} (\underbrace{y_i - (\beta_0 + \beta_1 x_i)}_{e_i})^2$$

e_i es la diferencia entre el valor y_i observado (donde "cae el punto") y el valor $\widehat{y_i}$ predecido por el modelo (donde "tendría que haber caído").

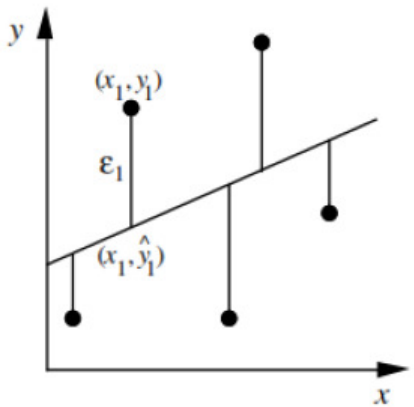
De esta manera, habiendo obtenido $\widehat{\beta_0}$ y $\widehat{\beta_1}$, para un valor x_0 de la variable independiente se podrá predecir por el modelo lineal el valor $\widehat{y_0}$ de la variable dependiente mediante

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0$$

Método de los mínimos cuadrados

Una manera de minimizar el error $e_i = y_i - \widehat{y}_i$ consiste en minimizar la suma de los errores elevados al cuadrado, o la suma de los cuadrados residuales (SCR):

$$\text{SCR} = \sum_{i=1}^{n} e_i^2$$



- Si el SCR es pequeño el ajuste es bueno, y si es grande el ajusto es malo.
- En el caso de una recta vamos a querer hallar $\widehat{\beta}_0$ y $\widehat{\beta}_1$ que minimicen

$$\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Regresión lineal simple. Primer ejemplo

Método de los mínimos cuadrados

Derivamos $\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0))^2$ respecto de β_1 y de β_0 e igualamos a 0:

$$\frac{\partial}{\partial \beta_1}\left(\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0))^2\right) = -2\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0))x_i = 0$$

$$\frac{\partial}{\partial \beta_0}\left(\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0))^2\right) = -2\sum_{i=1}^{n}(y_i - (\beta_1 x_i + \beta_0)) = 0$$

Despejamos β_0 de la primera ecuación y sustituyendo en la segunda obtenemos los estimadores MC (mínimos cuadrados) o LS (least squares):

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(x,y)}{s_x^2} = \underbrace{\frac{cov(x,y)}{s_y s_x}}_{r}\frac{s_y}{s_x} \qquad \widehat{\beta_0} = \overline{y} - \widehat{\beta_1}\overline{x}$$

donde $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$

$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$, $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$.

Es fácil ver que el punto encontrado es un mínimo.

Historia

TABLE XXII.

Father's Stature and Son's Stature.

Son's Stature.	Father's Stature.																
	58·5—59·5	59·5—60·5	60·5—61·5	61·5—62·5	62·5—63·5	63·5—64·5	64·5—65·5	65·5—66·5	66·5—67·5	67·5—68·5	68·5—69·5	69·5—70·5	70·5—71·5	71·5—72·5	72·5—73·5	73·5—74·5	74·5—75·5	Totals
59·5—60·5	—	—	—	—	·5	·5	1	—	—	—	—	—	—	—	—	—	—	2
60·5—61·5	—	—	—	—	·5	—	—	—	1	—	—	—	—	—	—	—	—	1·5
61·5—62·5	—	·25	·25	—	·5	1	·25	·25	·5	·5	—	—	—	—	—	—	—	3·5
62·5—63·5	—	·25	·25	2·25	2·25	2	4	5	2·75	1·25	—	·25	·25	—	—	—	—	20·5
63·5—64·5	1	—	1·5	3·75	3	4·25	8	9·25	3	1·25	1·5	·75	1·25	—	—	—	—	38·5
64·5—65·5	2	·1	·5	2	3·25	9·5	13·5	10·75	7·5	5·5	3·5	2·5	—	—	—	—	—	61·5
65·5—66·5	—	·5	1	2·25	5·25	9·5	10	16·75	17·5	16	5·25	2	2·5	1	—	—	—	89·5
66·5—67·5	—	1·5	2	4.75	3·5	13·75	19·75	26·5	25·75	19·5	12·5	13·75	3·25	·5	1	—	—	148·0
67·5—68·5	—	—	1·5	2	7·5	10	10·25	24·25	31·5	23·5	29·5	13·25	8·5	9·5	2·25	—	—	173·5
68·5—69·5	—	—	1	—	5·25	5	12·75	18·25	16	24	29	21·5	10	3·5	2·25	—	1	149·5
69·5—70·5	—	—	—	—	1	2·5	5·75	18·75	11·75	19·5	22·5	19·5	14·5	6·25	3·5	1·5	1	128·0
70·5—71·5	—	—	—	—	—	3·25	5	8·75	10·75	19	14·75	20·75	10·75	8	5	1	1	108·0
71·5—72·5	—	—	—	—	—	·25	3	1.25	7	7·75	10·75	11·25	10	8·5	2·75	·5	—	63·0
72·5—73·5	—	—	—	—	—	—	·75	·75	2·5	7·5	6·5	6	7·5	6·25	3·25	·5	·5	42·0
73·5—74·5	—	—	—	—	1	—	1·5	1·5	—	5·25	2·25	2·5	6·5	3·25	3·25	—	2	29·0
74·5—75·5	—	—	—	—	—	—	—	—	—	1	2	—	2·5	·75	1·75	·5	—	8·5
75·5—76·5	—	—	—	—	—	—	—	—	—	1·25	·25	—	·5	1	1	—	—	4·0
76·5—77·5	—	—	—	—	—	—	—	—	—	1·25	·25	1	—	—	1·5	—	—	4·0
77·5—78·5	—	—	—	—	—	—	—	—	—	—	1	1	—	·25	·75	—	—	3·0
78·5—79·5	—	—	—	—	—	—	—	—	—	—	—	—	—	·25	·25	—	—	·5
Totals	3	3·5	8	17	33·5	61·5	95·5	142	137·5	154	141·5	116	78	49	28·5	4	5·5	1078

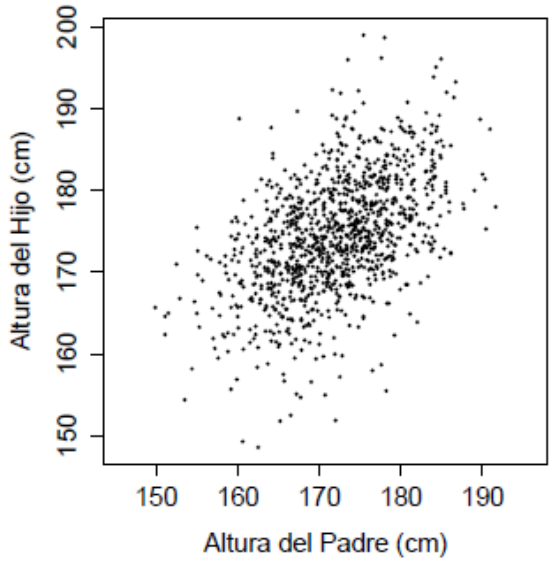
58—2

K. PEARSON AND A. LEE

415

EL matemática brítanico Karl Pearson (1857-1936) observó la estatura de 1078 padres (x) e hijos (y). Los promedios son $\overline{x} = 171,9$ cm e $\overline{y} = 174,5$ cm, los desvíos $s_x = 7$ cm y $s_y = 7{,}2$ cm, y el coeficiente de correlación $r = 0{,}5$

Historia



Observando que la recta de regresión se puede escribir como

$$y - \overline{y} = \widehat{\beta}_1(x - \overline{x})$$

se obtiene

$$y - \overline{y} = 0{,}51(x - \overline{x})$$

Si un padre tiene altura x, entonces

- Si $x > \overline{x}$ entonces $y > \overline{y}$ pero $y - \overline{y} < x - \overline{x}$.
- Si $x < \overline{x}$ entonces $y < \overline{y}$ pero $\overline{y} - y < \overline{x} - x$.

lo cual tiene la siguiente interpretación: los hijos cuyos padres tienen una estatura superior al valor medio, tienden a igualarse a éste, mientras que aquellos cuyos padres son muy bajos tienden a reducir su diferencia respecto a la estatura media, es decir, "regresan" al promedio.

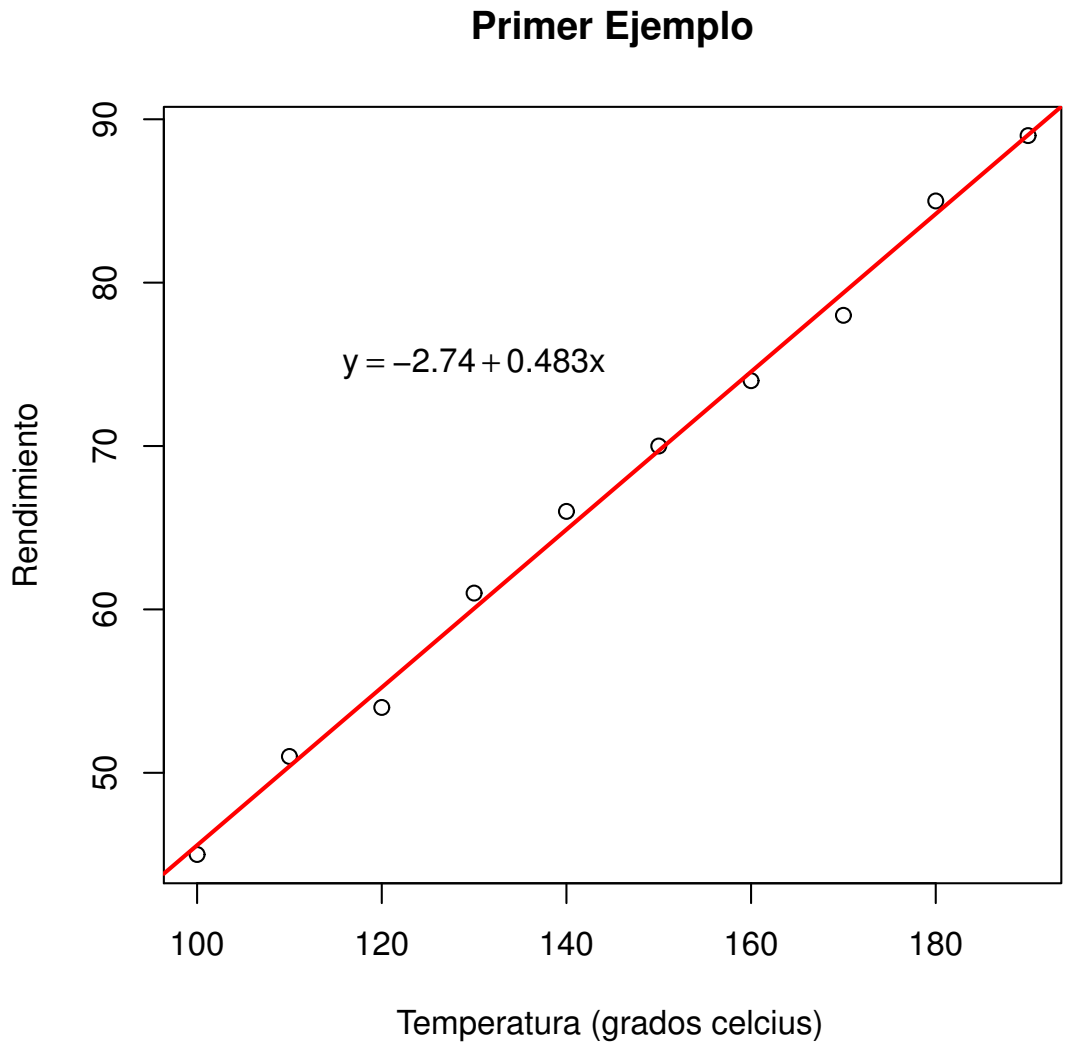
Regresión lineal simple. Primer ejemplo

```
>X=cbind(seq(100,190,10),c(45,51,54,61,66,70,74,78,85,89))
> X=as.data.frame(X)
> colnames(X)=c("Temp","Rend")
> plot(X,xlab="Temperatura (grados celcius)",ylab="Rendimiento",
main=paste("Primer Ejemplo"))
> a=lm(Rend~Temp,data=X)
> summary(a)
> abline(a,col="red",lwd=2)
Call:
lm(formula = Rend ~ Temp, data = X)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3758 -0.5591  0.1242  0.7470  1.1152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.73939    1.54650  -1.771    0.114
Temp         0.48303    0.01046  46.169 5.35e-11 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.9503 on 8 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9958
F-statistic:  2132 on 1 and 8 DF,  p-value: 5.353e-11
```

Regresión lineal simple. Primer ejemplo

La ecuación de la recta es

$$\widehat{y} = -2,74 + 0,48x$$



La función lm en R

La linealidad es sobre **los coeficientes** del modelo, es decir, el modelo es lineal en los parámetros $\beta_0, \beta_1, \ldots, \beta_d$ que se quiere hallar:

1. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$ es lineal
2. $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}^2 + \beta_4 x_{i2} x_{i4} + e_i$ es lineal
3. $y_i = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2 \cos(x_{i2}) + \beta_3 x_{i3}^2 + \beta_4 x_{i2} x_{i4} + e_i$ es lineal.
4. $y_i = \beta_0 + \beta_1 \sin(\beta_2 x_{i1}) + \beta_2 x_{i2}^{\beta_3} + e_i$ NO es lineal.

Con R, las funciones que se usan son:

```
>lm(y~x1+x2) #para el modelo y=ax1+bx2+c
>lm(y~I(x1+x2)) #para el modelo y=a(x1+x2)+c
>lm(y~poly(x,2)) #para el modelo y=ax^2+bx+c
>lm(y~x-1) #para el modelo y=ax
```

Ejemplo

Se quiere modelar la relación que existe entre el salario Y (en millones de dolares) y la cantidad de años de experiencia x de profesionales y obtener un intervalo de confianza al 95 % para Y cuando $x = 10$.
Nuestra base de datos consiste de 143 observaciones:

```
>profsalary <- read.table("profsalary.txt",header=TRUE)
>attach(profsalary)
>plot(Experience,Salary,xlab="Years of Experience", main=paste("Salary data"))
```


```
> head(profsalary,10)
   Case Salary Experience
1     1     71         26
2     2     69         19
3     3     73         22
4     4     69         17
5     5     65         13
6     6     75         25
7     7     66         35
8     8     66         16
9     9     67         16
10   10     69         16
```

Ejemplo

Claramente esta relación no es lineal y no sería adecuada el modelo de regresión lineal simple

$$Y = \beta_0 + \beta_1 x + e$$

siendo Y el salario y x la cantidad de años de experiencia. Claramente el ploteo sugiere un modelo de regresión polinomial cuadrático

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

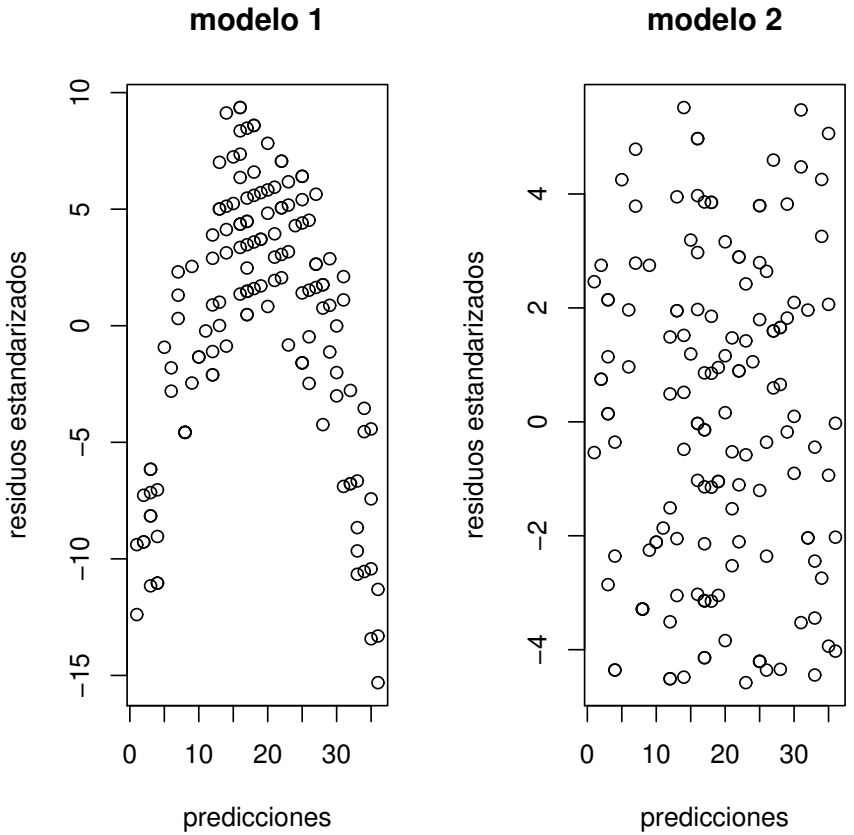
```
>m1 <- lm(Salary~Experience)
>abline(m1,col="red",lwd=2)
```


```
>m2 <- lm(Salary~Experience +
I(Experience^2))
```

Ejemplo

Acá vamos a graficar los errores (estandarizados) cometidos por cada modelo.

```
>par(mfrow=c(1,2))
>plot(Experience,m1$res,xlab="predicciones",
ylab="residuos estandarizados",main=paste("modelo 1"))
>plot(Experience,m2$res,xlab="predicciones",
ylab="residuos estandarizados",main=paste("modelo 2"))
```



El segundo modelo parecería más adecuado: no hay patrón en cuanto a los errores cometidos.

Ejemplo

```
> summary(m2)

Call:
lm(formula = Salary ~ Experience + I(Experience^2))

Residuals:
    Min      1Q  Median      3Q     Max
-4.5786 -2.3573  0.0957  2.0171  5.5176

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.720498   0.828724   41.90   <2e-16 ***
Experience       2.872275   0.095697   30.01   <2e-16 ***
I(Experience^2) -0.053316   0.002477  -21.53   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.817 on 140 degrees of freedom
Multiple R-squared:  0.9247,    Adjusted R-squared:  0.9236
F-statistic: 859.3 on 2 and 140 DF,  p-value: < 2.2e-16

>
```

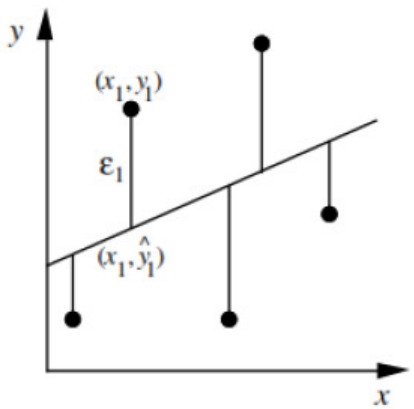
Plan

Generalización

Ecuación fundamental:

$$\text{“observación”} = \text{“modelo”} + \text{“error aleatorio”}$$
$$Y = f(x) + \epsilon$$

Los modelos de regresión utilizan la ecuación anterior suponiendo que el modelo es lineal.
En todo lo que sigue, consideramos una serie de datos $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$:



“Mejor recta” $y = \beta_1 x + \beta_0$ de manera a minimizar

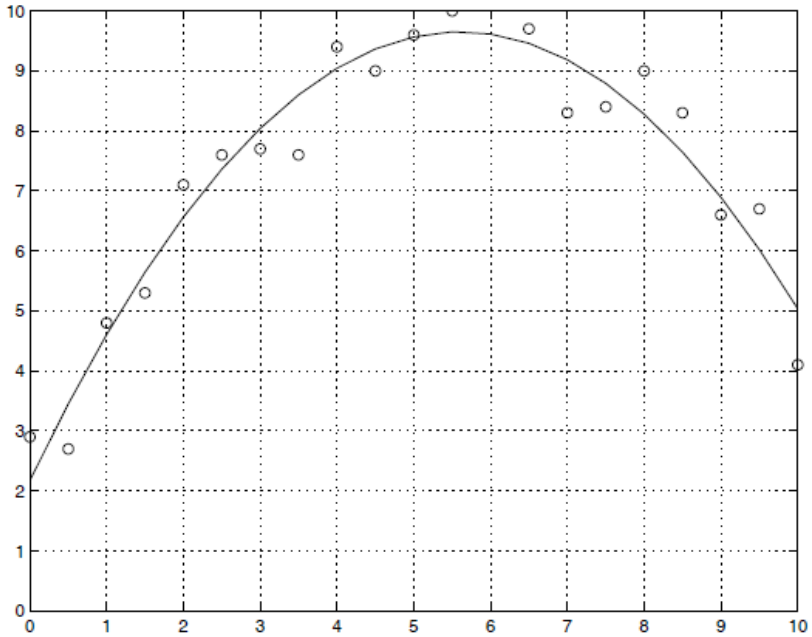
$$SCR = ||e||^2 = \sum_{i=1}^{n} e_i^2$$
$$= ||Y - X\beta||^2 \text{ donde}$$
$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}}_{e}$$

Generalización

Más generalmente podemos querer buscar el "mejor polinomio" de grado d

$$y = \beta_d x^d + \beta_{d-1} x^{d-1} + \cdots + \beta_1 x + \beta_0$$

que se ajusta a los datos.
Por ejemplo la parábola de mínimos cuadrados que ajusta un conjunto de puntos:



$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

(¡modelo lineal en los coeficientes!)

$$\underbrace{\begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{pmatrix}}_{Y} = \underbrace{\begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & . & . \\ . & . & . \\ 1 & x_n & x_n^2 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{pmatrix}}_{e}$$

Generalización

De la misma manera que para la regresión lineal simple, si $\mathcal{L} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ se quiere hallar un vector $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$ que minimice la función

$$\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2$$

Hallamos entonces un hiperplano de regresión y podemos ver el problema como un problema de proyección ortogonal.

Observe que $\sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_d x_{id})\right)^2 = ||Y - X\beta||^2$ y por lo tanto el problema original se transforma en un problema de algebra lineal siendo:

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{pmatrix}_{n \times (d+1)}, \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{pmatrix} \in \mathbb{R}^{d+1}$$

Generalización

El estimador mínimos cuadrados se basa en la siguiente propiedad de la proyección ortgonal: si $v \in \mathbb{R}^{d+1}$ y S es un subespacio de $\mathbb{R}^{d+1}$ entonces

$$||v - P_S(v)|| \leq ||v - s|| \ \ \forall s \in S$$

Acá consideramos como subespacio S al subespacio generado por las columnas de X y notaremos $S = \langle X \rangle \in \mathbb{R}^n$.

Observar que el complemento ortogonal de S es

$$S^{\perp} = \langle X \rangle^{\perp} = \{v \in \mathbb{R}^n : X'v = 0_{\mathbb{R}^{d+1}}\} = N(X')$$

Prueba: $v \in \langle X \rangle^{\perp} \Leftrightarrow v$ es ortogonal a todas las columnas de $X \Leftrightarrow v$ es ortogonal a todas las filas de X'
$\Leftrightarrow X'v = 0_{\mathbb{R}^{d+1}} \Leftrightarrow v \in N(X')$

Buscamos entonces $\widehat{\beta} \in \mathbb{R}^{d+1}$ tal que $X\widehat{\beta} = P_S(Y)$. Entonces:

$$Y - X\widehat{\beta} = Y - P_S(Y) = P_{S^{\perp}}(Y)$$

Entonces

$$X'P_{S^{\perp}}(Y) = 0_{\mathbb{R}^{d+1}} \Leftrightarrow X'\left(Y - X\widehat{\beta}\right) = 0_{\mathbb{R}^{d+1}} \Leftrightarrow X'Y = X'X\widehat{\beta}$$

Generalización

A la expresión $X'Y = X'X\widehat{\beta}$ se le llama *ecuaciones normales*. Si X es de rango completo, es decir $rg(X) = d+1$ o $N(X) = \{0_{\mathbb{R}^{d+1}}\}$, entonces la solución por el método de los mínimos cuadrados es única, pues en este caso $X'X$ es invertible y por lo tanto

$$\widehat{\beta} = (X'X)^{-1}X'Y$$

Si el rango de X es $r < d+1$ entonces el sistema es indeterminado y la solución no es única y consideramos

$$\widehat{\beta} = (X'X)^{-}X'Y$$

donde $(X'X)^{-}$ es una pseudo inversa de $X'X$ y verifica $(X'X)(X'X)^{-}(X'X) = (X'X)$.

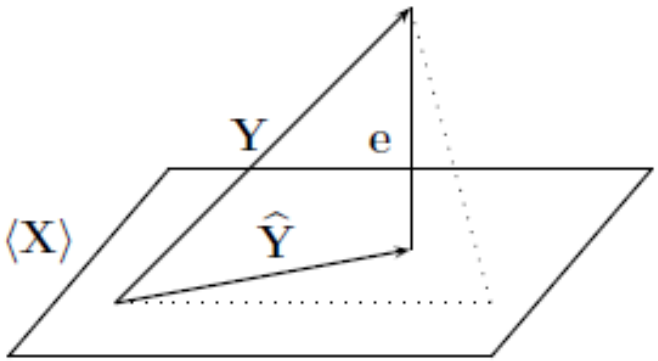
Situación geométrica

Sea $\langle X \rangle$ el subespacio generado por las columnas de X. Entonces

1. $\mathbb{E}(Y) \in \langle X \rangle$
2. El vector de residuos $e = Y - X\widehat{\beta}$ es ortogonal a $\langle X \rangle$.

Por lo tanto es mímimo $\epsilon'\epsilon = ||Y - X\beta||^2$ cuando

$$X\widehat{\beta} = P_{\langle X \rangle}(Y) = PY = \widehat{Y}$$



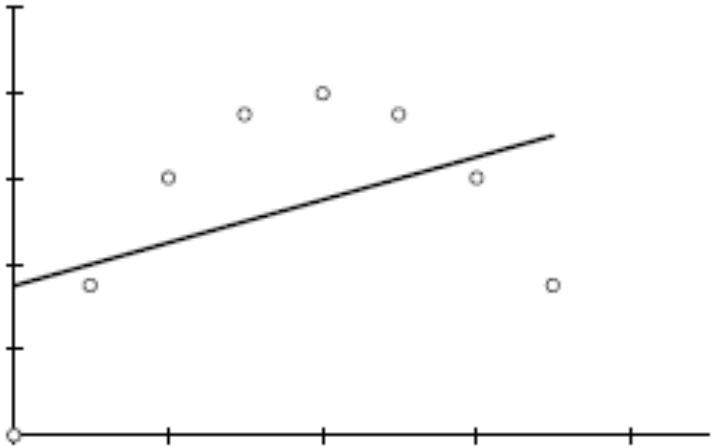
Entonces

- $e = Y - \widehat{Y}$ es ortogonal a $\langle X \rangle$
- $X'e = \mathbf{0}$ y por lo tanto $X'\widehat{Y} = X'Y$
- $SCR = e'e = (Y - PY)'(Y - PY) = Y'(I - P)Y$

Supuestos: condiciones de Gauss-Markov

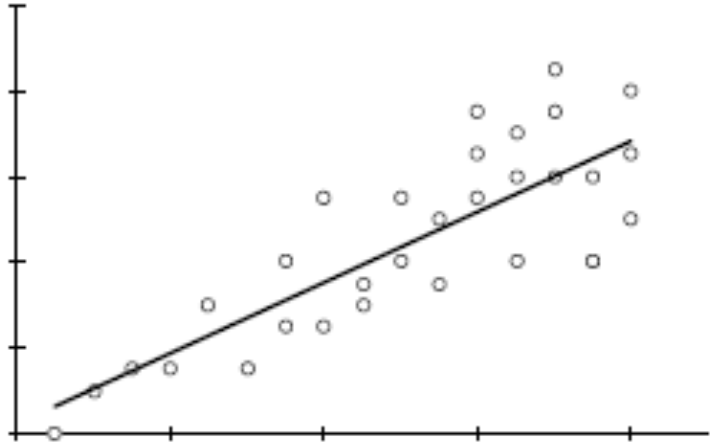
Volvemos a la regresión lineal simple. Hasta ahora el método de los mínimos cuadrados es análitico. Veamos donde interviene la estadística.
Suponemos que $x_1, \ldots, x_n$ son constantes. Supongamos que los errores e_i provienen de una variable aleatoria ϵ e impondremos que estos errores verifiquen las condiciones de Gauss-Markov:

(1) $\mathbb{E}(\epsilon_i) = 0$
$\Rightarrow \mathbb{E}(y_i) = \beta_1 x_i + \beta_0$
$\forall\, i = 1, \ldots, n$
No queremos que se dé esta situación:



Supuestos: condiciones de Gauss-Markov

(2) $Var(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \sigma^2$ (cte)
$\forall\, i = 1, \ldots, n$
(propiedad de homocedasticidad)
No queremos que se dé esta situación (heterocedasticidad):



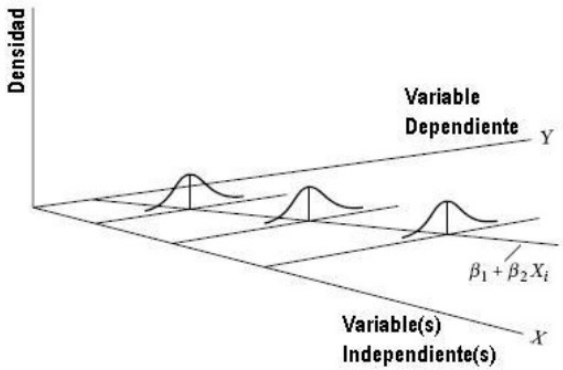


Figura: homocedasticidad

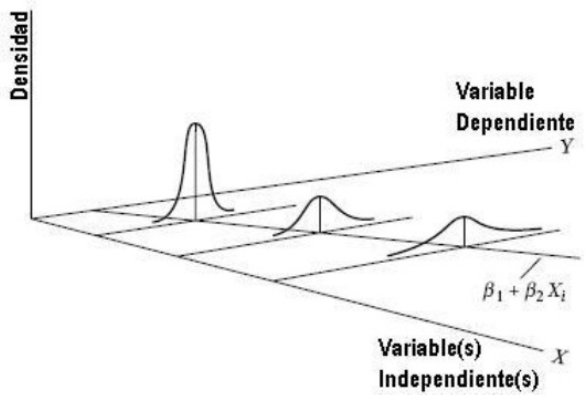


Figura: heterocedasticidad

(3) Los residuos deben ser incorrelados. Esto se puede hacer a partir del test de Durbin-Watson con el estadístico:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

que debe ser próximo a 2 si los residuos no son correlados. El estadístico no sigue ninguna ley en particular pero sus valores criticos están tabulados.

Resumen general

La expresión general del modelo lineal es:

$$Y = \underbrace{\mathbf{x}'\beta}_{f(X)} + \epsilon$$

y la estimación:

$$\widehat{Y} = \mathbf{x}'\widehat{\beta}$$

donde $\widehat{\beta}$ es la estimación del vector β obtenida por el método de los mínimos cuadrados.

Si suponemos las hipótesis de Gauss-Markov, el modelo lineal $Y = \mathbf{x}'\beta + \epsilon$ cumple que

$$\mathbb{E}(Y) = \mathbf{x}'\beta$$

Si además de suponer las condiciones de Gauss-Markov sobre los errores, se tiene que $\epsilon_i \sim N(0, \sigma^2)$ entonces decimos que el modelo es normal y se tiene que:

$$Y \sim N(\mathbf{x}'\beta, \sigma^2)$$