ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN
PROFESIONAL

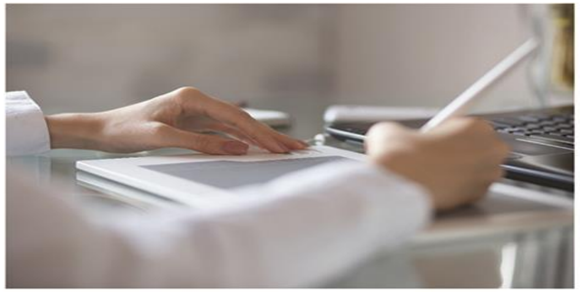
Diplomado en Big Data y Ciencias de Datos

Minería de Datos

Análisis de Regresión

Educación Profesional - Escuela de Ingeniería UC

Sebastián Raveau




Análisis de regresión

Técnica econométrica ampliamente utilizada en múltiples contextos

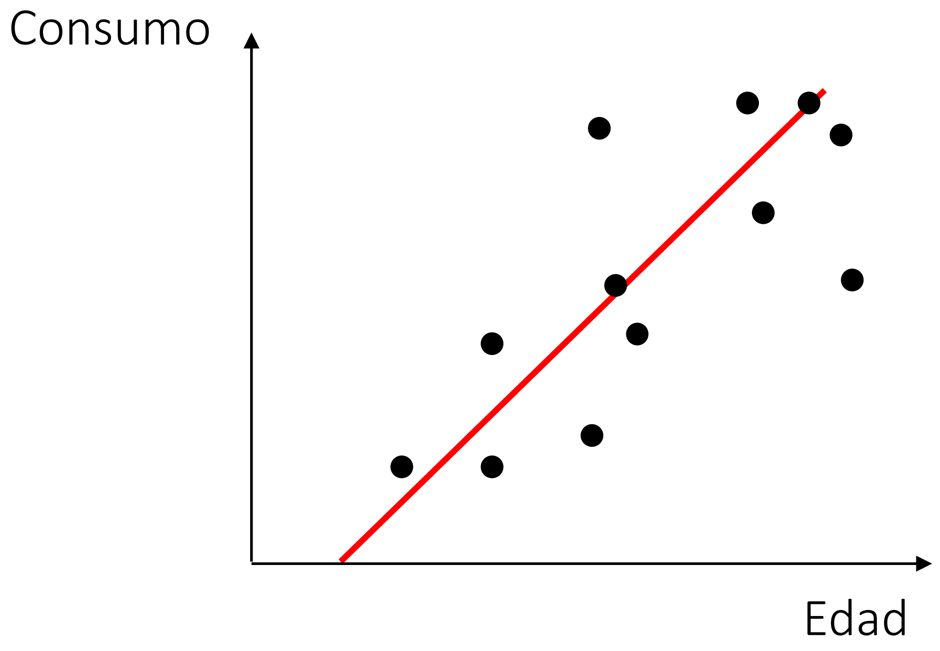
Corresponde a un método de aprendizaje supervisado

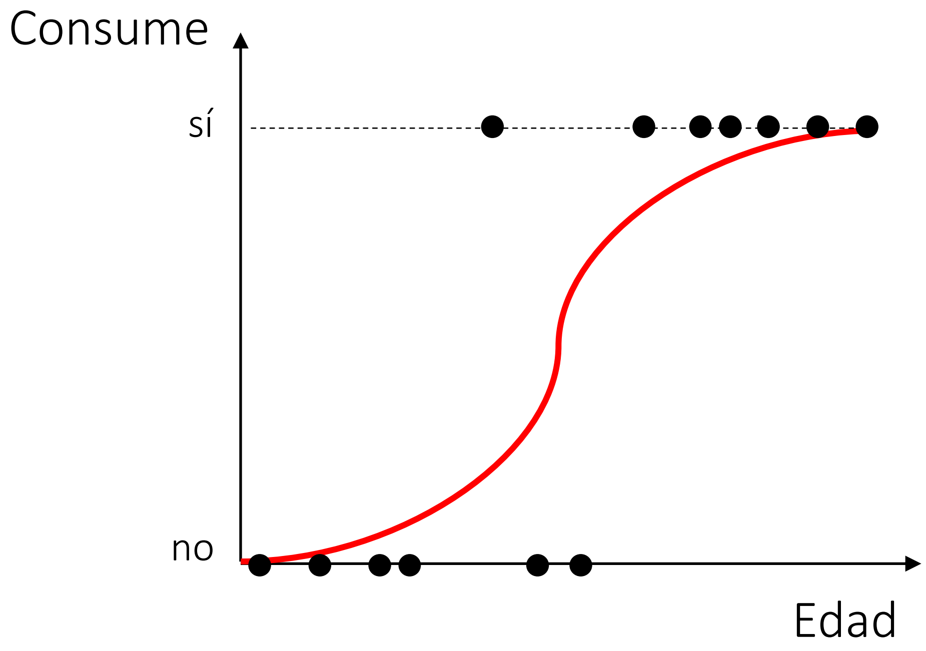
Nos interesa entender y predecir variables numéricas (continuas y discretas)

Podemos estudiar relaciones lineales y no lineales

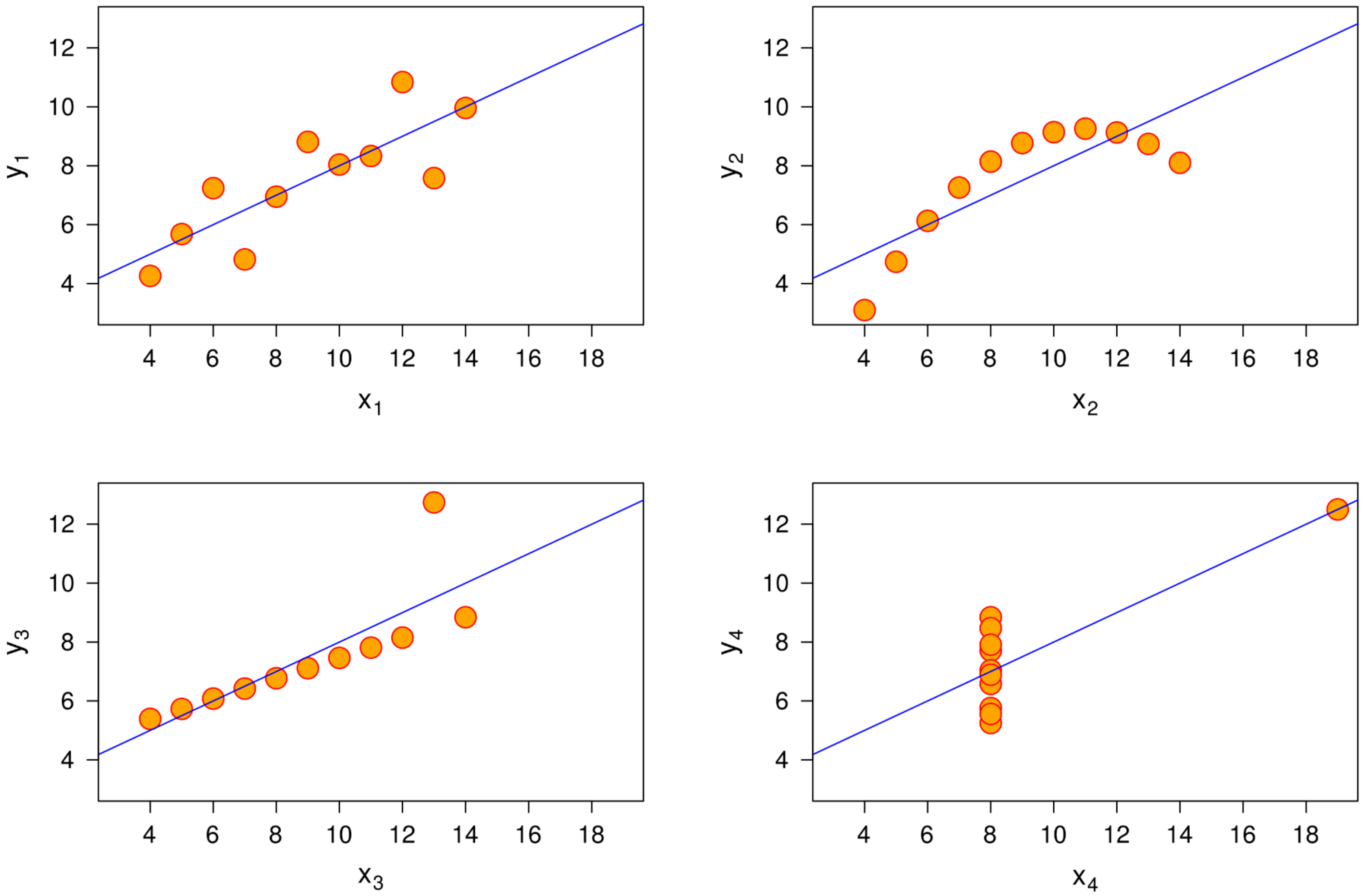
Análisis de regresión

Supongamos que queremos predecir el consumo de cierto producto

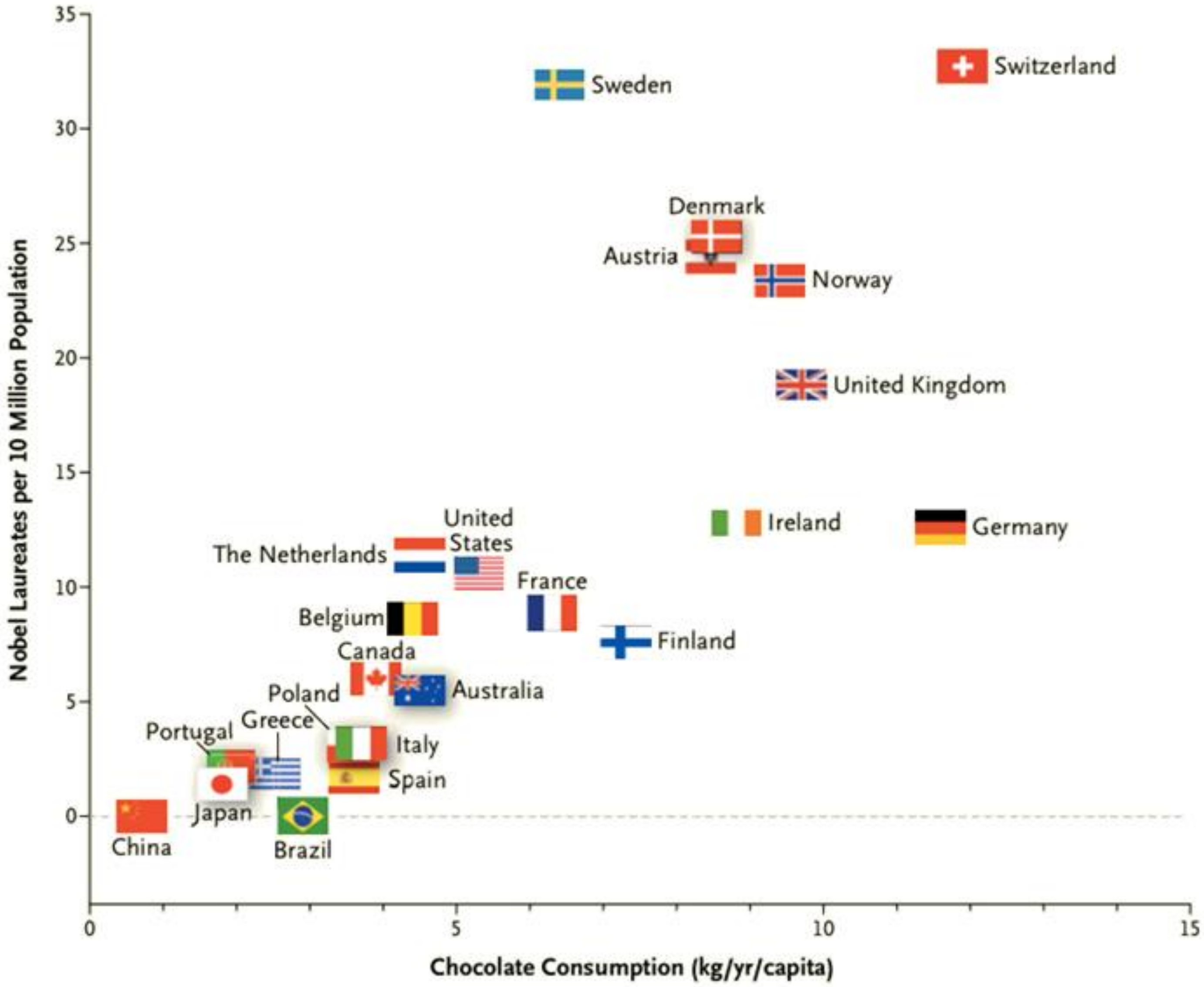




Cuarteto de Anscombe



Relaciones espurias



http://tylervigen.com/spurious-correlations

Regresión Lineal

Regresión Lineal

En esta caso la variable endógena es continua

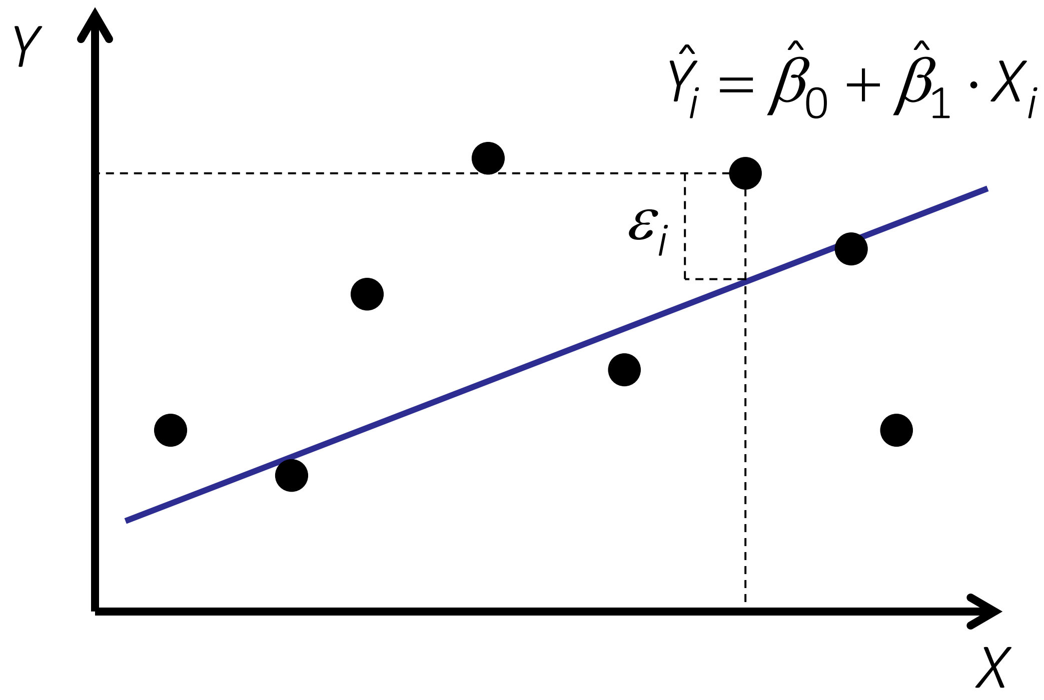
Buscamos ajustar una ecuación lineal que nos permita entender y predecir la variable endógena en función de las variables exógenas

La relación es lineal en los parámetros, no necesariamente en los atributos

$$Y = \beta_0 + \beta_1 \cdot \ln X \qquad \text{versus} \qquad Y = \beta_0 + \beta_1 \cdot X^{\beta_2}$$

Regresión lineal mediante mínimos cuadrados

Existen diversas maneras de encontrar ajustar una recta $Y_i = \beta_0 + \beta_1 \cdot X_i + e_i$ a los datos



Mínimos Cuadrados

$$\text{Min} \sum_i \varepsilon_i^2$$

Regresión lineal mediante mínimos cuadrados

Dada la regresión lineal

$$Y_i = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_k \cdot X_k + e_i$$

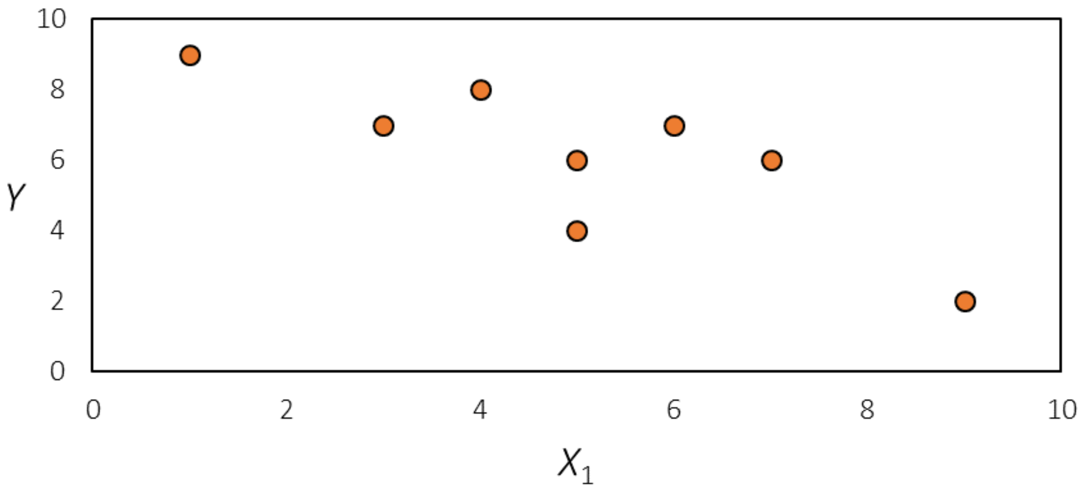
Los parámetros β que minimizan la suma cuadrática de los errores son

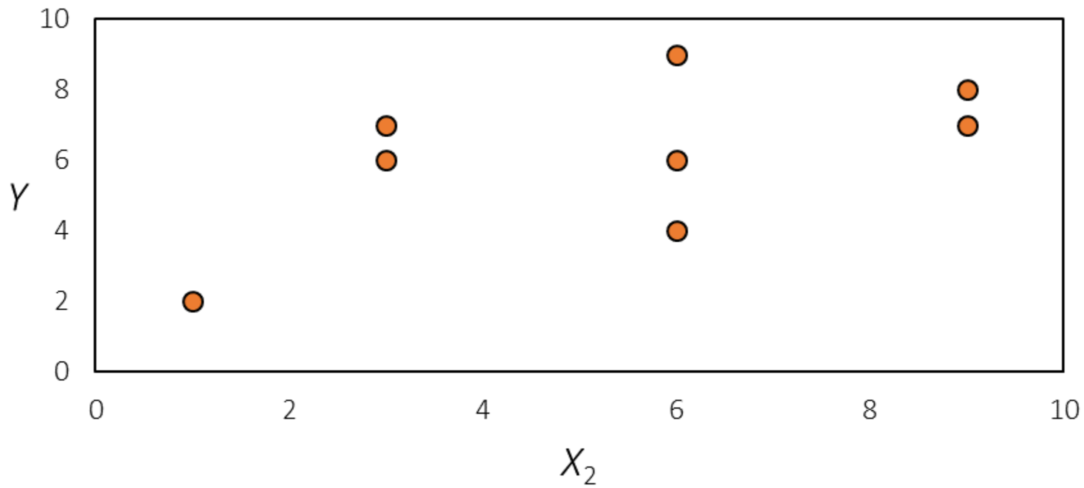
$$\hat{\beta} = \left(X^T X \right)^{-1} X^T Y$$

Consideremos los siguientes datos

Variable Endógena *Y*	**Variable Exógena X_1**	**Variable Exógena X_2**
7	6	3
4	5	6
6	7	3
7	3	9
9	1	6
2	9	1
8	4	9
6	5	6

Análisis preliminar

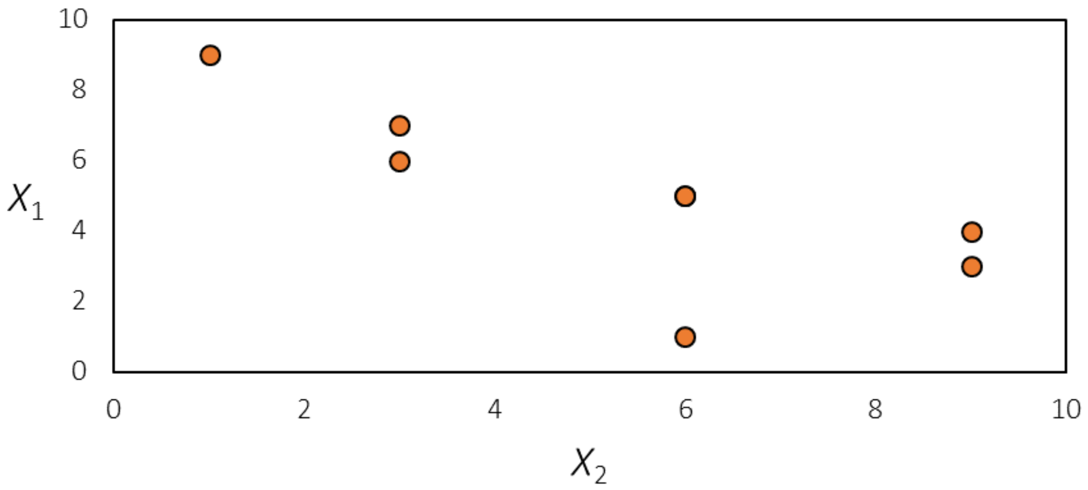




$$\rho_{Y,X_1} = -0{,}890$$

$$\rho_{Y,X_2} = -0{,}592$$

$$\rho_{X_1,X_2} = -0{,}771$$



Apliquemos Mínimos Cuadrados

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + e_i$$

$$Y = \begin{bmatrix} 7 \\ 4 \\ 6 \\ 7 \\ 9 \\ 2 \\ 8 \\ 6 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 6 & 3 \\ 1 & 5 & 6 \\ 1 & 7 & 3 \\ 1 & 3 & 9 \\ 1 & 1 & 6 \\ 1 & 9 & 1 \\ 1 & 4 & 9 \\ 1 & 5 & 6 \end{bmatrix}$$

Apliquemos Mínimos Cuadrados

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$X = \begin{bmatrix} 1 & 6 & 3 \\ 1 & 5 & 6 \\ 1 & 7 & 3 \\ 1 & 3 & 9 \\ 1 & 1 & 6 \\ 1 & 9 & 1 \\ 1 & 4 & 9 \\ 1 & 5 & 6 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 5 & 7 & 3 & 1 & 9 & 4 & 5 \\ 3 & 6 & 3 & 9 & 6 & 1 & 9 & 6 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 8 & 40 & 43 \\ 40 & 242 & 177 \\ 43 & 177 & 289 \end{bmatrix}$$

Apliquemos Mínimos Cuadrados

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$X^T X = \begin{bmatrix} 8 & 40 & 43 \\ 40 & 242 & 177 \\ 43 & 177 & 289 \end{bmatrix}$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 4{,}891 & -0{,}500 & -0{,}421 \\ -0{,}500 & 0{,}059 & 0{,}039 \\ -0{,}421 & 0{,}039 & 0{,}043 \end{bmatrix}$$

Apliquemos Mínimos Cuadrados

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$Y = \begin{bmatrix} 7 \\ 4 \\ 6 \\ 7 \\ 9 \\ 2 \\ 8 \\ 6 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 6 & 5 & 7 & 3 & 1 & 9 & 4 & 5 \\ 3 & 6 & 3 & 9 & 6 & 1 & 9 & 6 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 49 \\ 214 \\ 290 \end{bmatrix}$$

Apliquemos Mínimos Cuadrados

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

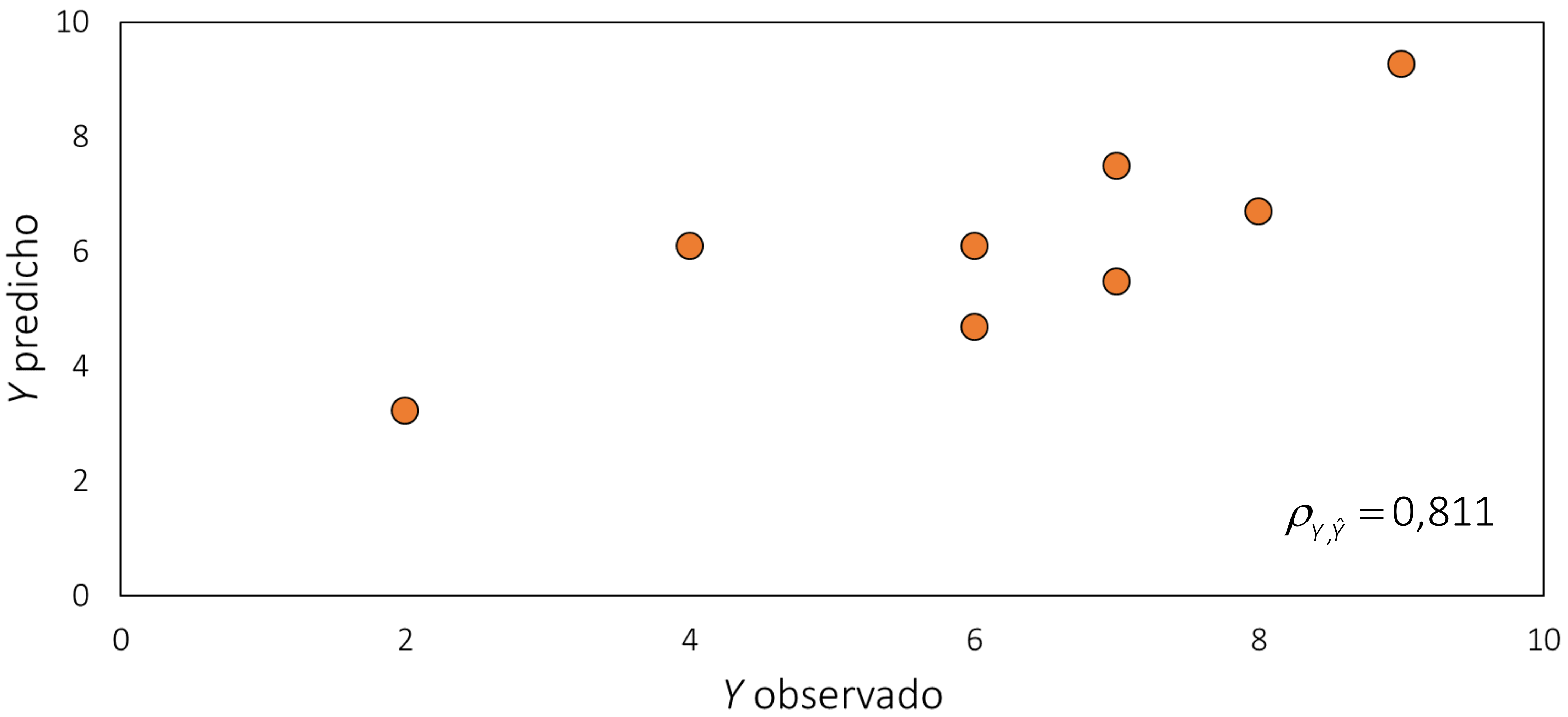
$$\left(X^T X\right)^{-1} = \begin{bmatrix} 4{,}891 & -0{,}500 & -0{,}421 \\ -0{,}500 & 0{,}059 & 0{,}039 \\ -0{,}421 & 0{,}039 & 0{,}043 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 49 \\ 214 \\ 290 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 10{,}415 \\ -0{,}793 \\ -0{,}061 \end{bmatrix}$$

$$\hat{Y}_i = 10{,}415 - 0{,}793 \cdot X_{1i} - 0{,}061 \cdot X_{2i}$$

Valores observados *vs* predichos



Índices de ajuste del modelo

Coeficiente de determinación R^2

$$R^2 = \frac{\sum_i \left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_i \left(Y_i - \overline{Y}\right)^2} = \frac{\text{Variación Explicada}}{\text{Variación Total}}$$

El índice R^2 requiere la presencia del intercepto y está acotado entre 0 y 1

Índices de ajuste del modelo

Un problema del índice R^2 es que siempre aumenta a medida que se agregan nuevas variables explicativas

Podemos ajustar por la cantidad de variables explicativas k (sin contar el intercepto) y la cantidad de observaciones n

$$\bar{R}^2 = \left(R^2 - \frac{k}{n-1} \right) \left(\frac{n-1}{n-k-1} \right)$$

En nuestro ejemplo

$$\text{Variación Total} = \sum_i \left(Y_i - \overline{Y}\right)^2 = 34{,}88$$

$$R^2 = 0{,}659$$

$$\text{Variación Explicada} = \sum_i \left(\hat{Y}_i - \overline{Y}\right)^2 = 22{,}97$$

$$\overline{R}^2 = \left(0{,}659 - \frac{2}{8-1}\right)\left(\frac{8-1}{8-2-1}\right) = 0{,}522$$

En nuestro ejemplo

$$\varepsilon = Y - \hat{Y} = \begin{bmatrix} 7 \\ 4 \\ 6 \\ 7 \\ 9 \\ 2 \\ 8 \\ 6 \end{bmatrix} - \begin{bmatrix} 5,48 \\ 6,09 \\ 4,68 \\ 7,49 \\ 9,26 \\ 3,22 \\ 6,70 \\ 6,09 \end{bmatrix} = \begin{bmatrix} 1,52 \\ -2,09 \\ 1,32 \\ -0,49 \\ -0,36 \\ -1,22 \\ 1,30 \\ -0,09 \end{bmatrix}$$

$$\sum_i \varepsilon_i = 0$$

$$\sum_i \varepsilon_i^2 = 11,91$$

¿Cuánto aporta cada variable explicativa?

Los parámetros estimados también tienen varianza

$$Var\left(\hat{\beta}\right) = \frac{\sum_i \varepsilon_i^2}{n-\left(k+1\right)} \cdot \left(X^T X\right)^{-1}$$

A partir de la varianza de cada parámetro, podemos obtener una medida de significancia estadística del aporte de cada variable dentro del modelo, como por ejemplo un test-*t*

En nuestro ejemplo

$$\sum_i \varepsilon_i^2 = 11{,}91$$

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 4{,}891 & -0{,}500 & -0{,}421 \\ -0{,}500 & 0{,}059 & 0{,}039 \\ -0{,}421 & 0{,}039 & 0{,}043 \end{bmatrix}$$

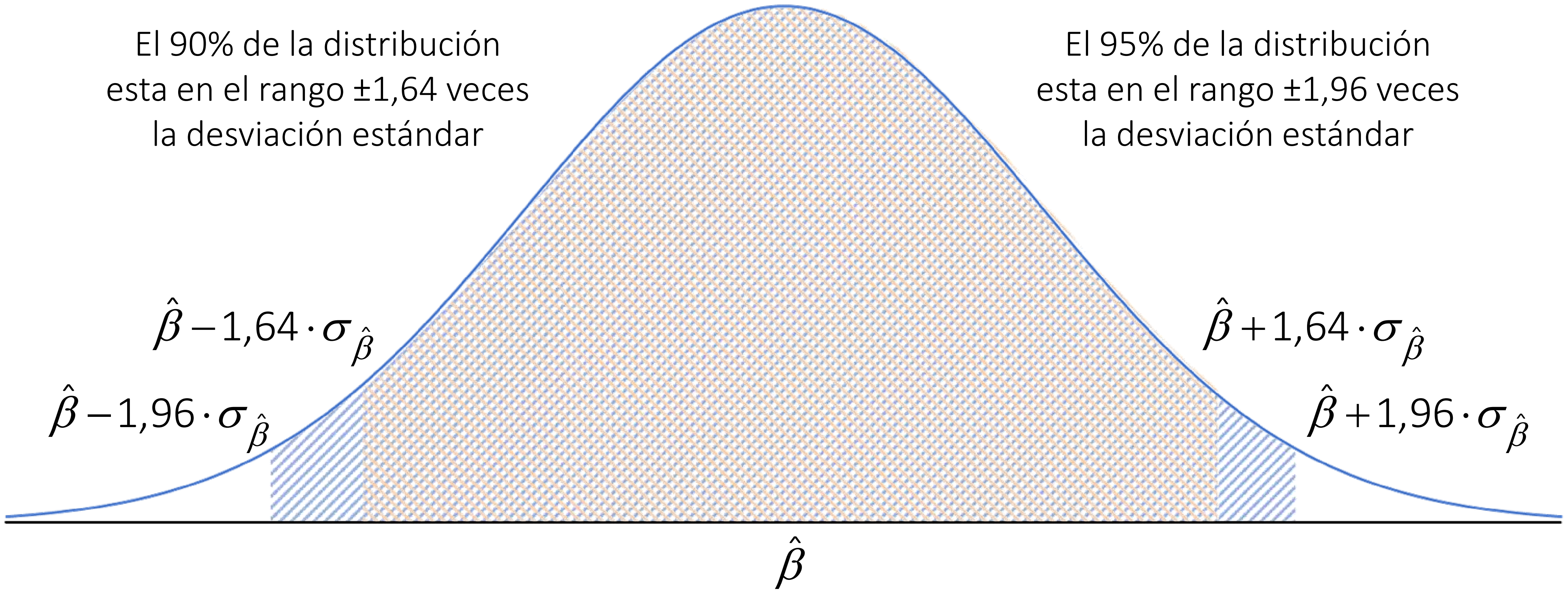
$$Var\left(\hat{\beta}_0\right) = \frac{11{,}91}{8-3} \cdot 4{,}891 = 11{,}648$$

$$Var\left(\hat{\beta}_1\right) = \frac{11{,}91}{8-3} \cdot 0{,}059 = 0{,}140$$

$$Var\left(\hat{\beta}_2\right) = \frac{11{,}91}{8-3} \cdot 0{,}043 = 0{,}101$$

¿Cuánto aporta cada variable explicativa?

Distribución Normal:



¿Cuánto aporta cada variable explicativa?

El test-*t* nos permite analizar si el valor 0 pertenece al intervalo de confianza del parámetro estimado, comparando su media con su desviación estándar

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}$$

Si está sobre 1,64 la variable es estadísticamente significativa al 90% de confianza

Si está sobre 1,96 la variable es estadísticamente significativa al 95% de confianza

En nuestro ejemplo

$$\hat{\beta}_0 = 10{,}415 \qquad Var\left(\hat{\beta}_0\right) = 11{,}648 \qquad t_{\hat{\beta}_0} = 10{,}415 \big/ \sqrt{11{,}648} = 3{,}052$$

$$\hat{\beta}_1 = -0{,}793 \qquad Var\left(\hat{\beta}_1\right) = 0{,}140 \qquad t_{\hat{\beta}_1} = -0{,}793 \big/ \sqrt{0{,}140} = -2{,}121$$

$$\hat{\beta}_2 = -0{,}061 \qquad Var\left(\hat{\beta}_2\right) = 0{,}101 \qquad t_{\hat{\beta}_2} = -0{,}061 \big/ \sqrt{0{,}101} = -0{,}190$$

La variable X_1 si es estadísticamente significativa

La variable X_2 no es estadísticamente significativa

Especificación del modelo

Con X_2 en el modelo	Sin X_2 en el modelo
$\underset{(3{,}052)}{\hat{\beta}_0 = 10{,}415}$	$\underset{(8{,}180)}{\hat{\beta}_0 = 9{,}816}$
$\underset{(-2{,}121)}{\hat{\beta}_1 = -0{,}793}$	$\underset{(-3{,}383)}{\hat{\beta}_1 = -0{,}738}$
$\underset{(-0{,}190)}{\hat{\beta}_2 = -0{,}061}$	
$R^2 = 0{,}659$	$R^2 = 0{,}656$
$\overline{R}^2 = 0{,}522$	$\overline{R}^2 = 0{,}599$

Ajuste estadístico y validez del modelo

Variable		Significativa	No significativa
Relevante	Signo correcto	OK	Mantener en el modelo
	Signo equivocado	Problema serio	Problema
Adicional	Signo correcto	OK	Probar si conviene sacarla
	Signo equivocado	Sacar del modelo	Sacar del modelo

Regresión Logística

Regresión logística

En esta caso la variable endógena es discreta

No podemos ajustar una ecuación lineal, pero si ajustar una función de probabilidad

Modelamos la probabilidad de que la variable endógena tome cada uno de sus posibles valores, en función de un conjunto de variables exógenas

Caso binomial

Consideremos una variable endógena con dos niveles (1 y 0)

Definimos una "función de utilidad" V_i

$$V_i = \theta_0 + \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \ldots + \theta_k \cdot X_k$$

Y probabilidades de cada nivel son:

$$\Pr(Y_i = 1) = \frac{\exp(V_i)}{\exp(V_i) + 1} \qquad \Pr(Y_i = 0) = \frac{1}{\exp(V_i) + 1}$$

Caso multinomial

Podemos extender este modelo a variables endógenas con más niveles

Por ejemplo, con cuatro niveles (3, 2, 1 y 0):

$$V_i^a = \theta_0^a + \theta_1^a \cdot X_1 + \theta_2^a \cdot X_2 + \ldots + \theta_k^a \cdot X_k$$

$$V_i^b = \theta_0^b + \theta_1^b \cdot X_1 + \theta_2^b \cdot X_2 + \ldots + \theta_k^b \cdot X_k$$

$$V_i^c = \theta_0^c + \theta_1^c \cdot X_1 + \theta_2^c \cdot X_2 + \ldots + \theta_k^c \cdot X_k$$

Caso multinomial

Podemos extender este modelo a variables endógenas con más niveles

Por ejemplo, con cuatro niveles (3, 2, 1 y 0):

$$\Pr\left(Y_i = 3\right) = \frac{\exp\left(V_i^a\right)}{\exp\left(V_i^a\right) + \exp\left(V_i^b\right) + \exp\left(V_i^c\right) + 1} \qquad \Pr\left(Y_i = 2\right) = \frac{\exp\left(V_i^b\right)}{\exp\left(V_i^a\right) + \exp\left(V_i^b\right) + \exp\left(V_i^c\right) + 1}$$

$$\Pr\left(Y_i = 1\right) = \frac{\exp\left(V_i^c\right)}{\exp\left(V_i^a\right) + \exp\left(V_i^b\right) + \exp\left(V_i^c\right) + 1} \qquad \Pr\left(Y_i = 0\right) = \frac{1}{\exp\left(V_i^a\right) + \exp\left(V_i^b\right) + \exp\left(V_i^c\right) + 1}$$

Máxima Verosimilitud

Como la función de utilidad V_i es función de los parámetros θ y las probabilidades de cada nivel son función de V_i, entonces las probabilidades son función de los parámetros θ

Buscamos los parámetros θ que le entregan una alta probabilidad de ocurrencia a los datos observados, $\Pr^*(Y_i)$

Máxima Verosimilitud

La función de verosimilitud *L* consiste en la probabilidad conjunta de ocurrencia de los datos observados

$$\text{Max}_{\{\theta\}}\ L = \prod_i \Pr^*\left(Y_i\right)$$

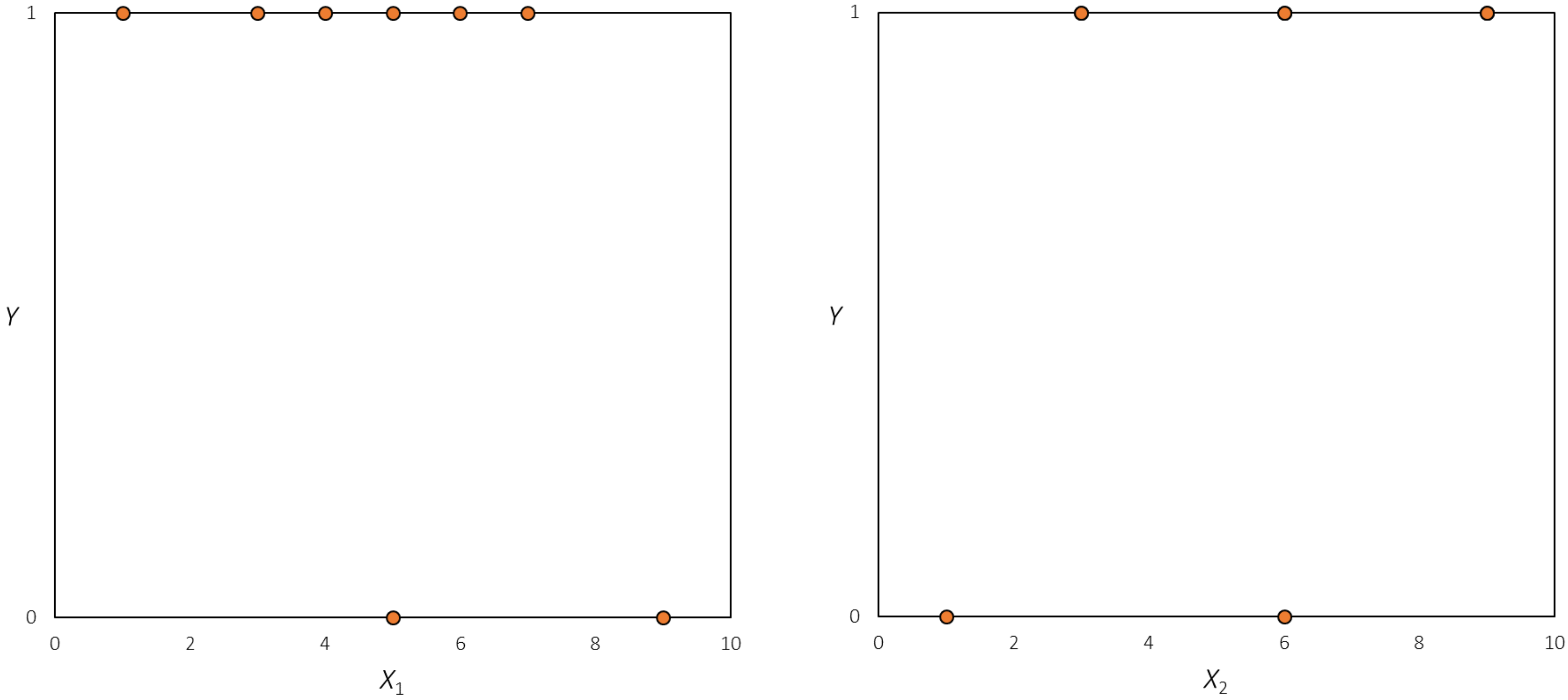
Por conveniencia computacional podemos maximizar la log-verosimilitud *l*:

$$\text{Max}_{\{\theta\}}\ l = \sum_i \ln\left[\Pr^*\left(Y_i\right)\right]$$

Consideremos los siguientes datos

Variable Endógena *Y*	**Variable Exógena X_1**	**Variable Exógena X_2**
1	6	3
0	5	6
1	7	3
1	3	9
1	1	6
0	9	1
1	4	9
1	5	6

Análisis preliminar



Apliquemos Máxima Verosimilitud

$$V_i = \theta_0 + \theta_1 \cdot X_{1i} + \theta_2 \cdot X_{2i}$$

$$\Pr(Y_i = 1) = \frac{\exp(V_i)}{\exp(V_i)+1} \qquad \Pr(Y_i = 0) = \frac{1}{\exp(V_i)+1}$$

$$L = \frac{\exp(V_1)}{\exp(V_1)+1} \cdot \frac{1}{\exp(V_2)+1} \cdot \frac{\exp(V_3)}{\exp(V_3)+1} \cdot \frac{\exp(V_4)}{\exp(V_4)+1} \cdot \frac{\exp(V_5)}{\exp(V_5)+1} \cdot \frac{1}{\exp(V_6)+1} \cdot \frac{\exp(V_7)}{\exp(V_7)+1} \cdot \frac{\exp(V_8)}{\exp(V_8)+1}$$

$$\text{Max } l = \ln L$$

Apliquemos Máxima Verosimilitud

$$\hat{\theta}_0 = 14{,}093$$

$$\hat{\theta}_1 = -1{,}675$$

$$\hat{\theta}_2 = -0{,}685$$

$$l^* = -3{,}134$$

$$L^* = 4{,}36 \times 10^{-2}$$

Y	***V***	**Pr(*Y* = 0)**	**Pr(*Y* = 1)**
1	1,985	0,121	0,879*
0	1,604	0,167*	0,833
1	0,310	0,423	0,577*
1	2,898	0,052	0,948*
1	8,305	0,000	1,000*
0	-1,669	0,841*	0,159
1	1,223	0,227	0,773*
1	1,604	0,167	0,833*

Significancia estadística

No existe una expresión cerrada para los valores óptimos de los parámetros θ

Las varianzas de los parámetros se pueden obtener a partir de la matriz de segundas derivadas de la función de log-verosimilitud

$$Var\left(\hat{\theta}\right) = -\left[\left.\frac{\partial^2 l(\theta)}{\partial \theta^2}\right|_{\hat{\theta}} \right]^{-1}$$

En nuestro ejemplo

$$Var\left(\hat{\theta}\right) = -\left[\left.\frac{\partial^2 l(\theta)}{\partial\theta^2}\right|_{\hat{\theta}}\right]^{-1} = \begin{bmatrix} 283{,}05 & -31{,}74 & -19{,}40 \\ -31{,}74 & 3{,}61 & 2{,}13 \\ -19{,}40 & 2{,}13 & 1{,}39 \end{bmatrix}$$

$$\hat{\theta}_0 = 14{,}093 \qquad t_{\hat{\theta}_0} = 14{,}093 \Big/ \sqrt{283{,}05} = 0{,}838$$

$$\hat{\theta}_1 = -1{,}675 \qquad t_{\hat{\theta}_1} = -1{,}675 \Big/ \sqrt{3{,}61} = -0{,}881$$

$$\hat{\theta}_2 = -0{,}685 \qquad t_{\hat{\theta}_2} = -0{,}685 \Big/ \sqrt{1{,}39} = -0{,}581$$

Criterios de información

De todos los criterios de información, el más utilizado es el propuesto por Akaike

$$AIC = 2 \cdot k - 2 \cdot l^*$$

Para muestras pequeñas, el AIC puede presentar sesgos

En dicho caso se puede aplicar el criterio corregido:

$$CAIC = AIC + \frac{2 \cdot k^2 + 2 \cdot k}{n - k - 1}$$

En nuestro ejemplo

$$l^* = -3{,}134 \qquad AIC = 2\cdot 3 - 2\cdot\left(-3{,}134\right) = 12{,}268$$

$$CAIC = 12{,}268 + \frac{2\cdot 3^2 + 2\cdot 3}{8-3-1} = 18{,}268$$

Predicción

Según sea necesario, podemos utilizar las probabilidades del modelo para el análisis o transformarlas en variables discretas (i.e. clasificar)

El nivel predicho será aquel con mayor probabilidad

Al predecir niveles discretos podemos aplicar diferentes enfoques de éxito predictivo, como matrices de confusión

En nuestro ejemplo

***Y* observado**	**Pr(*Y* = 1)**	***Y* predicho**
1	0,879	1
0	0,833	1
1	0,577	1
1	0,948	1
1	1,000	1
0	0,159	0
1	0,773	1
1	0,833	1

		Referencia	
		Positivo 1	Negativo 0
Predicho	Positivo 1	TP 6	FP 1
	Negativo 0	FN 0	TN 1

$$accuracy = 7/8 = 0{,}875$$