.UBA económicas
FACULTAD DE CIENCIAS ECONÓMICAS

Estadística I

Modelo de Regresión Lineal Simple

Natalia SALABERRY

En muchos casos, un análisis requiere de analizar conjuntamente más de una variable dado que, existe una asociación entre las mismas. Un caso que podemos considerar es, por ejemplo, el comportamiento del incremento de las ventas de un producto y el dinero destinado a campañas publicitarias. O que el consumo de un individuo esta determinado por su nivel de ingreso, entre muchos otros ejemplos. Entonces, mediante un análisis de **regresión lineal**, se propone una función lineal que permita estimar el **valor promedio** de, por ejemplo, el consumo a partir del conocimiento del nivel de ingreso de un individuo.

Es importante, que conozcamos qué se entiende por función lineal, de manera que podamos distinguir si esta es adecuada o no para representar la relación de las variables que están siendo objeto de análisis. La existencia de una relación lineal implica que existe un cambio proporcional constante: por ejemplo, el cambio en el consumo es proporcional al cambio en el ingreso y esta proporción es siempre la misma para todo el rango de posibles valores. Es decir, si el ingreso se modifica en un determinado valor, la variación en el consumo será un porcentaje fijo de esa variación, porcentaje que se mantiene constante para todo el rango de posibles valores.

Modelo de Regresión Lineal Simple

Busca determinar la relación entre una variable aleatoria en base a los valores de otra/s variable/s con la/s cual/es presenta dependencia estadística (relación no determinística). Veremos el caso de solo dos variables aleatorias.

Modelado a través de la ecuación de recta

$$Y_i = \beta_0 + \beta_1 * X_i$$

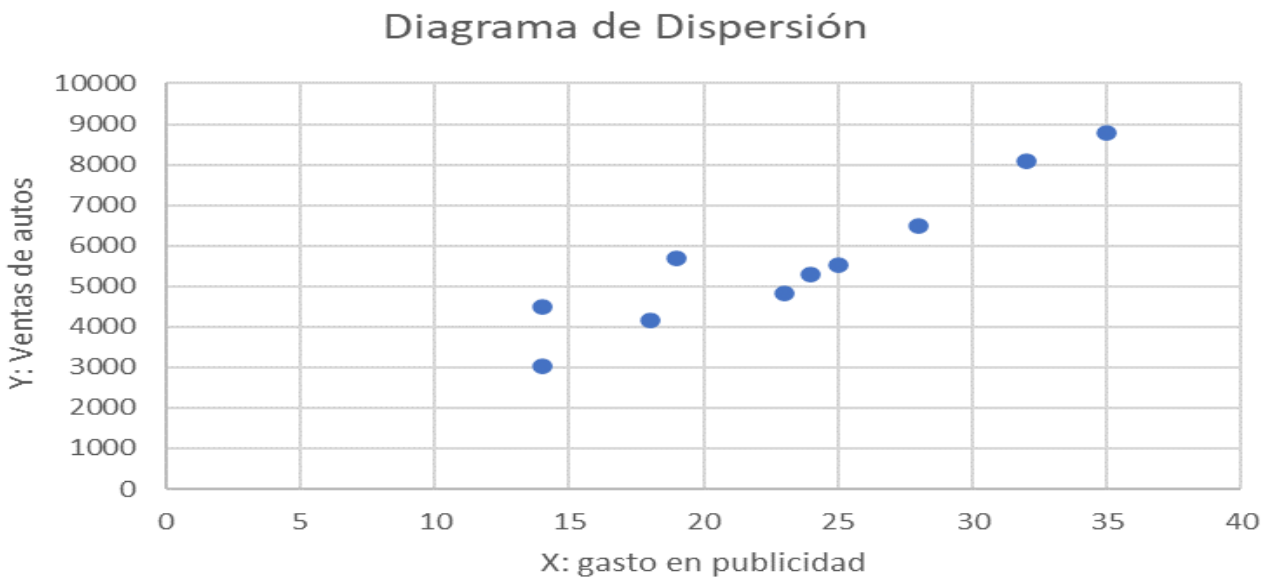
- Y_i: Variable dependiente o explicada (aleatoria)
- β_0: Ordenada al origen
- β_1: Pendiente — Cuanto varía Y ante variación de una unidad de X
- X_i: Variable independiente o explicativa (toma valores fijos)

Dado que es una recta, se tiene pares de puntos (Y_1, X_1), (Y_2, X_2),…, (Y_n, X_n)
Entonces podemos graficarlos

Primer paso: Representación gráfica para evaluar si es adecuado aplicar el modelo

Diagrama de Dispersión (o Scatter Plot)

Gasto en publicidad en miles de dólares	Ventas de autos en miles de unidades
X	Y
24	5298
32	8100
14	4506
23	4816
35	8768
28	6486
14	3022
19	5676
25	5524
18	4152



Se observa que a medida que aumenta el gasto en publicidad se incrementan las ventas. Entonces podemos pensar que existe una relación lineal directa entre las variables. Ante esta intuición podemos probar entonces de construir un modelo de regresión lineal.

Modelo de Regresión Lineal Simple

Dado que se establece a Y como una variable aleatoria, entonces el modelo se convierte en probabilístico. Esto quiere decir que vamos a estimar Y, a partir de estimar primero los coeficientes β_i

$$\widehat{Y_i} = \beta_0 + \beta_1 * X_i + \varepsilon$$

Donde:
$\widehat{Y_i}$ son los valores estimados de Y
$\boldsymbol{\beta_0}$ es la ordenada al origen
$\boldsymbol{\beta_1}$ es la pendiente de la recta: nos indica en cuanto varía Y ante el incremento de una unidad en X.
ε Término de Error Aleatorio: en cuanto difiere $\widehat{Y_i}$ de Y_i

El término de error tiene supuestos. Estos supuestos son los de Gauss-Markov

Modelo de Regresión Lineal Simple

Supuestos de Gauss- Markov

- ε ~N(0, σ^2) El error sigue una distribución normal con media 0 y varianza constante
- Los diferentes valores que puede tomar ε son independientes
- E(ε)=0
- V(ε)= σ^2 constante

Entonces Y ~N(E(Y|x), σ^2) donde

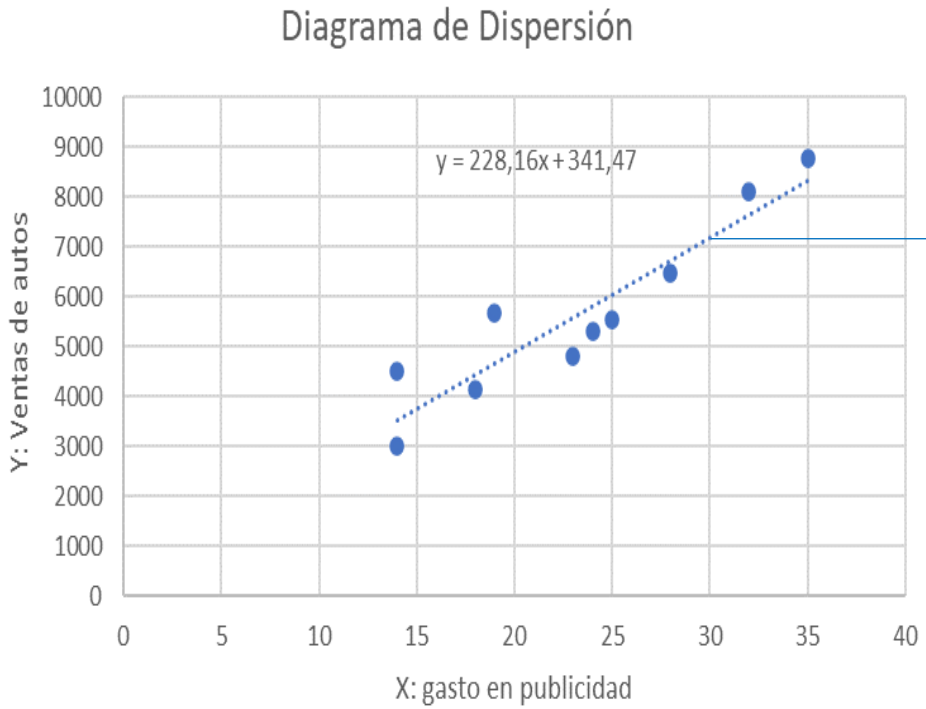
$\widehat{\boldsymbol{Y_i}}$ =E(Y|x)=E($\beta_0 + \beta_1 * X$ + ε)= $\beta_0 + \beta_1 * \mathrm{E}(X)$ siendo E(ε)=0 De aquí que la recta de regresión estimada es $\widehat{\boldsymbol{Y_i}} = \widehat{\boldsymbol{\beta_0}} + \widehat{\boldsymbol{\beta_1}} * \boldsymbol{X_i}$

V(Y|x)=V($\beta_0 + \beta_1 * X$+ ε) = ${\beta_1}^2 * \mathrm{V}(X)$= σ^2 lo cual implica que la variabilidad de cada valor de Y es la misma con cada valor de X (homogeneidad de la varianza).
A valores chicos de σ^2, los pares de puntos (X,Y) quedarán cerca de la **verdadera recta de regresión.**

Por lo tanto, si no hubiera error tendríamos

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} * X$$

Siendo la **verdadera recta de regresión, donde $\widehat{Y}$ es un valor medio**



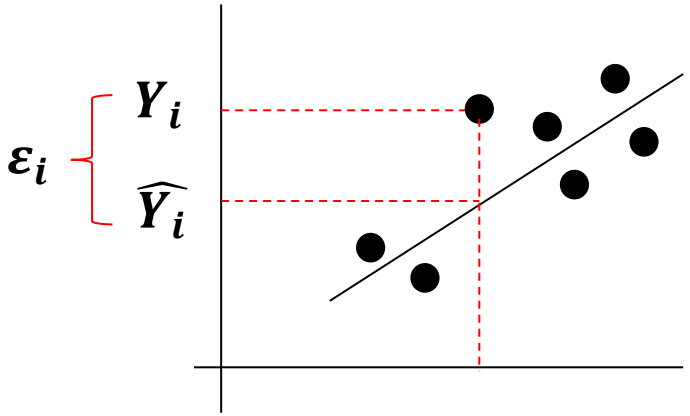
$$\widehat{Y} = 341{,}47 + 228{,}16 * X$$

Verdadera recta de regresión para el ejemplo

De esta manera, construir un modelo (en este caso de regresión lineal simple) permitirá construir escenarios futuros (Por ejemplo, estimar cuáles serán las ventas si se invierte tanto dinero en publicidad)

Estimación de parámetros del modelo: Método Mínimos Cuadrados

Cuando se trabaja con la estimación por mínimos cuadrados, los parámetros se estiman de modo tal que se minimice la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta.



$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left[Y_i - (\widehat{\beta_0} + \widehat{\beta_1} * X_i \,) \right]^2$$

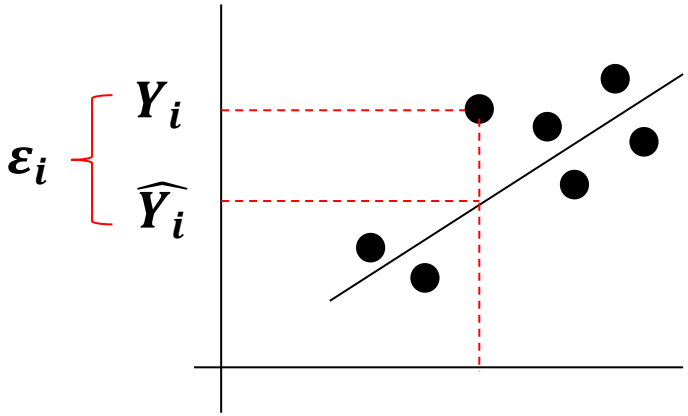
Suma cuadrada de las desviaciones (diferencia entre valor estimado y verdadero valor)

Para tener un buen ajuste del modelo, entonces debemos minimizar esta función. Es decir, hallar los valores de β_i tal que hagan mínima la diferencia. Estas diferencia no es otra cosa que la distancia entre los puntos y la recta (ε_i)

La obtención de los $\widehat{\beta_i}$ estimados requiere resolver un sistema de ecuaciones

Estimación de parámetros del modelo: Método Mínimos Cuadrados

Cuando se trabaja con la estimación por mínimos cuadrados, los parámetros se estiman de modo tal que se minimice la suma de los cuadrados de las desviaciones de las observaciones respecto de la recta.



$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left[Y_i - (\widehat{\beta_0} + \widehat{\beta_1} * X_i \,) \right]^2$$

Suma cuadrada de las desviaciones (diferencia entre valor estimado y verdadero valor)

Para tener un buen ajuste del modelo, entonces debemos minimizar esta función. Es decir, hallar los valores de β_i tal que hagan mínima la diferencia. Estas diferencia no es otra cosa que la distancia entre los puntos y la recta (ε_i)

La obtención de los $\widehat{\beta_i}$ estimados requiere resolver un sistema de ecuaciones

Estimación de parámetros del modelo: Método Mínimos Cuadrados

Objetivo

$$min \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\boldsymbol{Y_i} - \hat{Y}_i)^2 = \sum_{i=1}^{n} \left[\boldsymbol{Y_i} - (\widehat{\beta_0} + \widehat{\beta_1} * X_i)\right]^2$$

$$\text{con} \quad \boldsymbol{Y_i} - (\widehat{\beta_0} + \widehat{\beta_1} * X_i) = g$$

Método de Mínimos Cuadrados

Tomar la derivada respecto de cada parámetro e igualar a cero (es decir, encontrar el mínimo de la función g)

$$\frac{\partial g}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2 * \left(\boldsymbol{Y_i} - \widehat{\beta_0} - \widehat{\beta_1} * X_i\right) * (-1) = 0$$

Pasando dividiendo el 2 y (-1), entonces $\sum_{i=1}^{n} \left(\boldsymbol{Y_i} - \widehat{\beta_0} - \widehat{\beta_1} * X_i\right) = 0$

Distribuyendo la sumatoria, entonces $\sum_{i=1}^{n} \boldsymbol{Y_i} - \sum_{i=1}^{n} \widehat{\beta_0} - \sum_{i=1}^{n} \widehat{\beta_1} * X_i = 0$

Pasando los términos que están restando, entonces $\sum_{i=1}^{n} \boldsymbol{Y_i} = \sum_{i=1}^{n} \widehat{\beta_0} + \sum_{i=1}^{n} \widehat{\beta_1} * X_i$

Aplicando propiedades de la sumatoria, entonces $\boxed{\sum_{i=1}^{n} \boldsymbol{Y_i} = n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} * X_i}$ (1)

Estimación de parámetros del modelo: Método Mínimos Cuadrados

$$\frac{\partial g}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2 * \left(\boldsymbol{Y_i} - \widehat{\beta_0} - \widehat{\beta_1} * X_i\right) * (-X_i) = 0$$

Pasando dividiendo el 2 y (-1), entonces $\sum_{i=1}^{n}\left(\boldsymbol{Y_i} - \widehat{\beta_0} - \widehat{\beta_1} * X_i\right) X_i = 0$

Distribuyendo la sumatoria y X_i, entonces $\sum_{i=1}^{n} \boldsymbol{Y_i} X_i - \sum_{i=1}^{n} \widehat{\beta_0} X_i - \sum_{i=1}^{n} \widehat{\beta_1} * {X_i}^2 = 0$

Pasando los términos que están restando, entonces $\sum_{i=1}^{n} \boldsymbol{Y_i} X_i = \sum_{i=1}^{n} \widehat{\beta_0} X_i + \sum_{i=1}^{n} \widehat{\beta_1} * {X_i}^2$

Aplicando propiedades de la sumatoria, entonces

$$\sum_{i=1}^{n} \boldsymbol{Y_i} X_i = \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} * {X_i}^2 \quad (2)$$

De esta manera se obtiene un sistema de ecuaciones

Reciben el nombre de Ecuaciones normales

$$\begin{cases} \sum_{i=1}^{n} \boldsymbol{Y_i} = n\widehat{\beta_0} + \widehat{\beta_1} \sum_{i=1}^{n} * X_i \quad (A) \\ \sum_{i=1}^{n} \boldsymbol{Y_i} X_i = \widehat{\beta_0} \sum_{i=1}^{n} X_i + \widehat{\beta_1} \sum_{i=1}^{n} * {X_i}^2 \end{cases}$$

Recordemos como se resolvía un sistema de ecuaciones por el método del determinante

$$\begin{cases} 7 = 2X + 3Y \\ 5 = -X + 2Y \end{cases}$$

El determinante del sistema es: $\Delta = \begin{vmatrix} 2 & 3 \\ -1 & 2 \end{vmatrix}$

Y se resolvía multiplicando en diagonal y restando: 2*2 – (-1)*3 = 7

El determinante para X es: $\Delta X = \begin{vmatrix} 7 & 3 \\ 5 & 2 \end{vmatrix}$

Y se resolvía multiplicando en diagonal y restando: 7*2 – 5*3 = -1

Entonces $X = \frac{\Delta X}{\Delta} = \frac{-1}{7}$

El determinante para Y es: $\Delta Y = \begin{vmatrix} 2 & 7 \\ -1 & 5 \end{vmatrix}$

Y se resolvía multiplicando en diagonal y restando: 2*5 – (-1)*7 = 17

Entonces $Y = \frac{\Delta Y}{\Delta} = \frac{17}{7}$

Estimación de parámetros del modelo: Método Mínimos Cuadrados

Resolución del sistema de ecuaciones normales por el método del determinante

$$Dividiendo\ por\ n\ la\ ecuación\ (A)\ se\ obtiene \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{n\hat{\beta}_0}{n} + \frac{(\sum_{i=1}^{n} X_i)\ * \hat{\beta}_1}{n} \Rightarrow$$

$$\Rightarrow \bar{Y} = \hat{\beta}_0 + \bar{X}\hat{\beta}_1 \Rightarrow \boxed{\widehat{\boldsymbol{\beta}}_{\mathbf{0}} = \overline{\boldsymbol{Y}} - \overline{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}}$$

Por lo que repasamos para resolver el sistema de ecuaciones del ejemplo anterior (donde teníamos Y ahora es $\hat{\beta}_1$), podemos saber que $\hat{\beta}_1 = \frac{\Delta\widehat{\beta}_1}{\Delta}$

$$\Delta = \begin{vmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{vmatrix} \quad \Delta = n\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i = n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2 \ (1)$$

$$Dividiendo\ por\ n\ la\ ecuación\ (1)\ se\ obtiene \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}$$

$$que\ puede\ ser\ re\ expresado\ como \sum_{i=1}^{n} (X_i - \bar{X})^2$$

X	X_i^2	$(X_i-\bar{X})^2$
1	1	4
2	4	1
3	9	0
4	16	1
5	25	4
SUMA 15	55	10
n 5		$\sum_{i=1}^{n} (X_i-\bar{X})^2$
Media 3		

$\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} =$

$55 - \frac{15^2}{5} =$

$55 - \frac{225}{5} =$

$55 - 45 = \mathbf{10}$

Teniendo en cuenta que $\sum_{i=1}^{n}(aX_i - b) = a\sum_{i=1}^{n} X_i - \mathrm{nb}$

entonces $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \boxed{\sum_{i=1}^{n} X_i^2 - \mathrm{n}\bar{X}^2}$

Estimación de parámetros del modelo: Método Mínimos Cuadrados

$$\Delta\hat{\beta}_1 = \begin{vmatrix} n & \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} Y_i X_i \end{vmatrix} \qquad \Delta\hat{\beta}_1 = n \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i \ (2)$$

$Dividiendo\ por\ n\ la\ ecuación\ (2)\ se\ obtiene \sum_{i=1}^{n} Y_i X_i - \frac{\sum_{i=1}^{n} X_i \ \sum_{i=1}^{n} Y_i}{n}$

$que\ puede\ ser\ re\ expresado\ como \sum_{i=1}^{n} (Y_i X_i - \bar{X}\bar{Y})$

Teniendo en cuenta que $\sum_{i=1}^{n}(aX_i - b) = a \sum_{i=1}^{n} X_i - \mathrm{nb}$

entonces $= \boxed{\sum_{i=1}^{n} Y_i X_i - \mathrm{n}\bar{X}\bar{Y}}$

Estimación de parámetros del modelo: Método Mínimos Cuadrados

$$\hat{\beta}_1 = \frac{\Delta\hat{\beta}_1}{\Delta} =$$

$$\frac{\sum_{i=1}^{n} Y_i X_i - \mathrm{n}\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \mathrm{n}\bar{X}^2} = \frac{S_{XY}}{S_{XX}}$$

Siendo S_{XY} la covarianza muestral entre X e Y = Cov(X,Y)

y S_{XX} **la covarianza muestral entre X e X = Cov(X,X) =V(X)**

Finalmente se obtiene que los coeficientes estimados son

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \mathrm{n}\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \mathrm{n}\bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Para nuestro ejemplo

X	Y	X*Y	X^2	Y^2
24	5.298	127.152	576	28.068.804
32	8.100	259.200	1.024	65.610.000
14	4.506	63.084	196	20.304.036
23	4.816	110.768	529	23.193.856
35	8.768	306.880	1.225	76.877.824
28	6.486	181.608	784	42.068.196
14	3.022	42.308	196	9.132.484
19	5.676	107.844	361	32.216.976
25	5.524	138.100	625	30.514.576
18	4.152	74.736	324	17.239.104
232	56.348	1.411.680	5.840	345.225.856

n	10
Media X	23,20
Media Y	5.634,80
Sxy	104.406,40
Sxx	457,60
Syy	27.716.145,60
Beta1	228,16
Beta 0	341,47

$S_{XY} = \sum_{i=1}^{n} Y_i X_i - n\overline{XY}$=1411680 – 10*23,20*5634,80=104406,4

$S_{XX} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$=5840 – 10* $23{,}2^2$=457,6

$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{104406{,}4}{457{,}6}$=228,16

$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$= $5634{,}8 - 228{,}16 * 23{,}2 = 341{,}47$

$$\hat{Y} = 341{,}47 + 228{,}16 * X$$

Para nuestro ejemplo

A partir de $\hat{Y} = 341{,}47 + 228{,}16 * X$ calculamos los $\hat{Y}$:

Por ejemplo, para X=24 => $\hat{Y} = 341{,}47 + 228{,}16 * 24 = 5817{,}71$

Luego, podremos calcular el error:

Y - $\hat{Y}$ = 5298 – 5817,71 = - 519,31

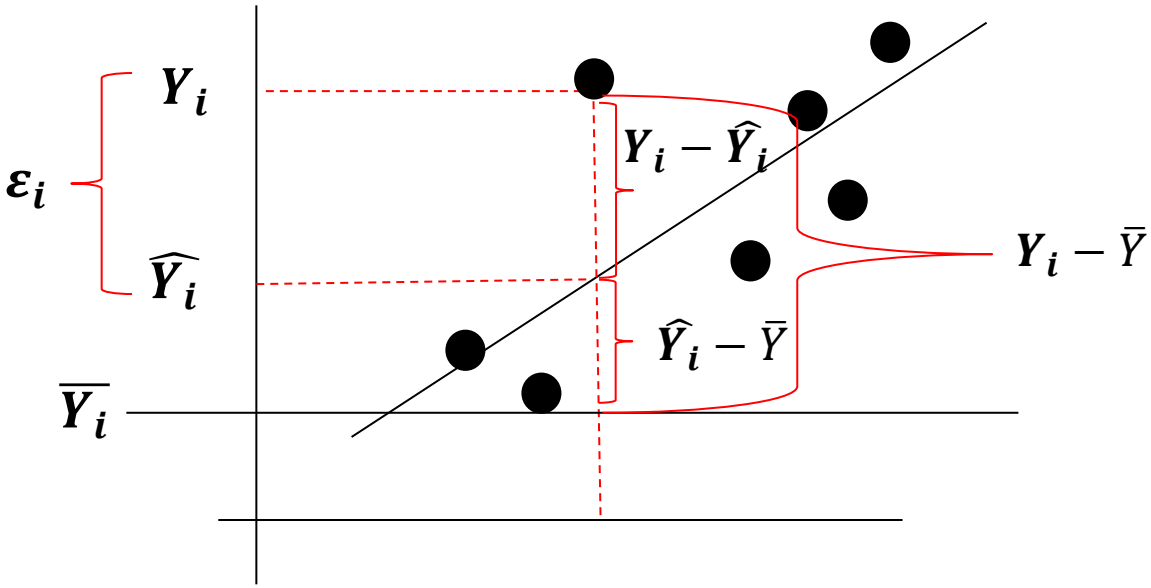
Por lo tanto, sumando toda la columna del error da 0, cumpliéndose el supuesto de Markov de que E(ε) = 0:

E(ε) = $\frac{\sum_{i=1}^{n} \varepsilon}{n} = \frac{0}{10} = 0$

Luego, al tener estimada la recta, puedo considerar valores nuevos de X y armar escenarios futuros.

X	Y	$\hat{Y}$	ε
24	5.298	5817,31	- 519,31
32	8.100	7642,59	457,41
14	4.506	3535,71	970,29
23	4.816	5589,15	- 773,15
35	8.768	8327,07	440,93
28	6.486	6729,95	- 243,95
14	3.022	3535,71	- 513,71
19	5.676	4676,51	999,49
25	5.524	6045,47	- 521,47
18	4.152	4448,35	- 296,35
232	56.348		0

Desviaciones en la estimación



$Y_i - \bar{Y}$ es la desviación total

$Y_i - \widehat{Y}_i$ es la desviación que no está explicada por la regresión, siendo la magnitud del error

$\widehat{Y}_i - \bar{Y}$ es la desviación que está explicada por la regresión

Medidas de Bondad de Ajuste

SCT (Suma de Cuadrados Total)=SCE + SCR

Esta medida considera los desvíos de cada observación respecto del promedio de la variable (la variable a estimar), sin considerar la relación que ésta tiene con la variable .

$$\boldsymbol{SCT} = S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

SCR (Suma de Cuadrados de la regresión)

Medida de cuanta variación de Y es explicada por el modelo.

$$\boldsymbol{SCR} = \sum_{i=1}^{n}(\widehat{\boldsymbol{Y_i}} - \bar{Y})^2$$

SCE (Suma de Cuadrados del Error)

Medida de la variación total de Y que no es explicada por el modelo, es decir, no puede ser atribuida a una relación lineal

$$\boldsymbol{SCE} = = \sum_{i=1}^{n} \boldsymbol{\varepsilon_i^2} \textbf{ = SCT –SCR= } \sum_{i=1}^{n}(Y_i - \widehat{\boldsymbol{Y_i}})^2$$

Medida de Bondad de Ajuste Absoluta

Varianza Residual

Esta es una medida absoluta de que tan bien se ajusta la recta de regresión estimada a las medias de las observaciones de la variable respuesta.

$$S^2 = \frac{SCE}{n-2}$$

En general, cuanto menor sea su valor, mejor ajuste del modelo. Entonces, buscamos que sea lo más chico posible.

También, podemos calcular el desvío residual como la raíz cuadrada del anterior.

$$S = \sqrt{S^2}$$

Continuando con el ejemplo

X	Y	X*Y	X^2	Y^2	(Y - Yraya)^2	Yi estimados	Yi est - Y raya	(Yi est - Y raya)^2	(Yi - Y est)^2
24	5.298	127.152	576	28.068.804	113.434,24	5.817,3	182,5	33.316,7	269.702,3
32	8.100	259.200	1.024	65.610.000	6.077.211,04	7.642,6	2.007,8	4.031.322,6	209.200,7
14	4.506	63.084	196	20.304.036	1.274.189,44	3.535,7	- 2.099,1	4.406.135,7	941.442,7
23	4.816	110.768	529	23.193.856	670.433,44	5.589,2	- 45,6	2.082,3	597.788,5
35	8.768	306.880	1.225	76.877.824	9.816.942,24	8.327,1	2.692,3	7.248.468,0	194.394,7
28	6.486	181.608	784	42.068.196	724.541,44	6.730,0	1.095,2	1.199.401,8	59.522,4
14	3.022	42.308	196	9.132.484	6.826.723,84	3.535,7	- 2.099,1	4.406.135,7	263.908,5
19	5.676	107.844	361	32.216.976	1.697,44	4.676,5	- 958,3	918.292,0	998.951,3
25	5.524	138.100	625	30.514.576	12.276,64	6.045,5	410,7	168.665,9	271.951,3
18	4.152	74.736	324	17.239.104	2.198.695,84	4.448,4	- 1.186,4	1.407.631,2	87.831,4
232	56.348	1.411.680	5.840	345.225.856	**27.716.145,60**			**23821451,84**	**3.894.693,8**

Media X	23,20	**SCE**	**3894693,762**
Media Y	5.634,80		
Sxy	104.406,40	**SCT**	**27716145,6**
Sxx	457,60		
Syy	27.716.145,60	**SCR**	**23821451,84**
Beta1	228,16		
Beta 0	341,47		

$$SCE = \text{SCT} - \text{SCR} = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$$

$$SCT = S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$SCR = \sum_{i=1}^{n}(\widehat{Y_i} - \bar{Y})^2$$

$$SCE = \text{SCT} - \text{SCR} = 27716145{,}6 - 23821451{,}84 = 3894693{,}76$$

Medidas de Bondad de Ajuste Relativas

Coeficiente de Determinación

Brinda una medida de la proporción de variación de Y que puede ser explicada por el modelo. Cuanto mayor sea su valor, mejor ajuste del modelo. Siempre será mayor que 0 y menor que 1

$$R^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT}$$

Si R^2 = 1, ajuste perfecto del modelo
Si 0,8 < R^2 < 1, muy buen ajuste del modelo
Si 0,5 <= R^2< 0,8, buen ajuste del modelo
Si 0 < R^2 < 0,5, pobre ajuste del modelo
Si R^2 = 0, el modelo lineal no es adecuado para representar la relación entre las variables.

Coeficiente de Correlación Muestral

$$R = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

Brinda una medida de la relación lineal entre los X e Y
A valores de X grandes le corresponde valores grandes de Y (relación positiva)
A valores de X grandes le corresponde valores chicos de Y (relación negativa).

Coeficiente de correlación muestral

$$R = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} \text{ con } -1 < r < 1$$

Interpretación del coeficiente de correlación

R=1 relación **lineal** perfectamente positiva. Esto sucederá cuando X e Y estén fuertemente relacionadas de manera positiva.

R= -1 relación **lineal** perfectamente negativa. Esto sucederá cuando X e Y estén fuertemente relacionadas de manera negativa.

R=0 indica que existe una ausencia completa de relación **lineal** entre X e Y.

-1<R<0 indica que la relación es **lineal negativa:**

- **Si** -1<R≤ -0,8 entonces relación ***lineal*** **fuertemente** negativa
- **Si** -0,8<R ≤ -0,5 entonces relación ***lineal*** **moderadamente** negativa
- **Si** -0,5<R< 0 entonces relación ***lineal*** **débilmente** negativa

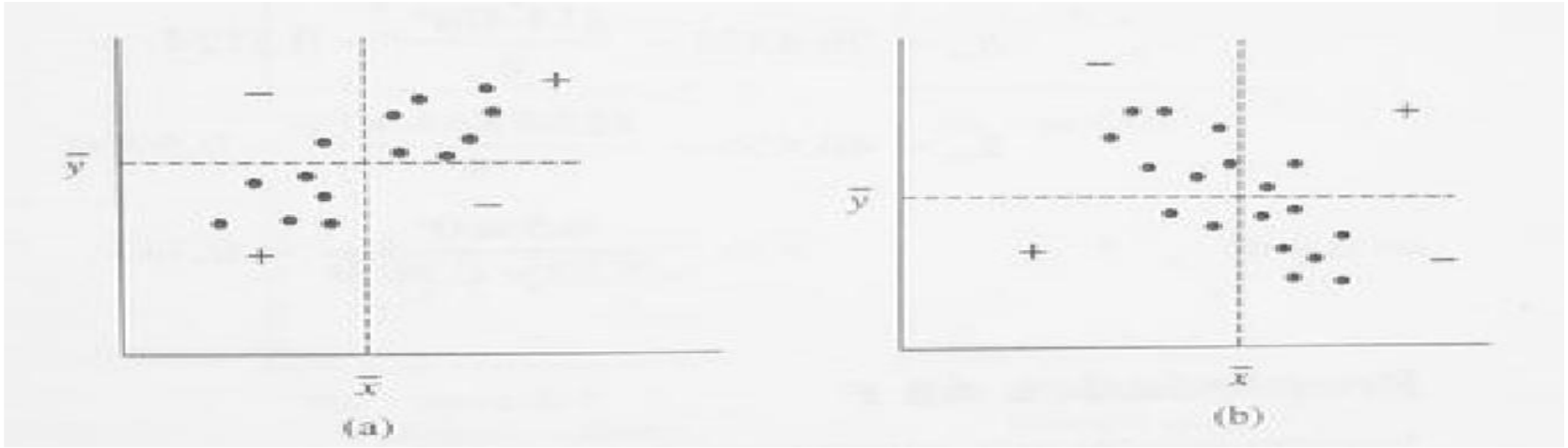
0<R<1 indica que la relación es **lineal positiva**:

- **Si** 0<R< 0,5 entonces relación ***lineal*** **débilmente** positiva
- **Si** 0<R< 0,8 entonces relación ***lineal*** **moderadamente** positiva
- **Si** 0,8<R< 1 entonces relación ***lineal*** **fuertemente** positiva

Medida de Bondad de Ajuste Relativas

Si la relación es positiva entonces
$S_{XY}>0$

Si la relación es negativa entonces
$S_{XY}<0$



entonces
R será $positivo$

$$R = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}}$$

entonces
R será $negativo$

Coeficiente de correlación muestral

$$R = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} \text{ con } -1 < R < 1$$

- El cuadrado del coeficiente de correlación muestral da el valor del coeficiente de determinación que resultaría de ajustar el modelo de regresión lineal simple. Por lo tanto, es posible calcular a R como la raíz cuadrada de R cuadrado

$$R = \sqrt{R^2}$$

Pero como -1 < R < 1, entonces hay que definir el signo (ya que la raíz cuadrada da un número siempre el resultado es positivo).

Para definir el signo observamos el **signo que posee la pendiente de la recta**, es decir, $\widehat{\beta_1}$. Si es positivo entonces R es positivo. Si es negativo entonces al resultado obtenido de hacer $\sqrt{R^2}$ le agregamos un menos y obtenemos un R negativo.

Continuando con el ejemplo

$$\hat{\sigma}^2 = S^2 = \frac{\mathbf{3894693,76223779}}{\mathbf{10-2}} = 486836{,}72$$

$$\sigma = S = \sqrt{486836{,}72} = 697{,}74$$

$$R^2 = \frac{SCR}{SCT} = \frac{23821451{,}8}{27716145{,}6} = 0{,}8594 = 86\%$$

El modelo presenta buen ajuste

$$R = \frac{104406{,}4}{\sqrt{457{,}6}\sqrt{27716145{,}6}} = \frac{104406{,}4}{21{,}3915871313935 * 5264{,}61257833851} \approx 0{,}93 = 93\%$$

Existe una relación lineal y positiva entre X e Y. Además, se cumple que Sxy>0

Test de hipótesis para la significatividad de $\hat{\beta}_1$

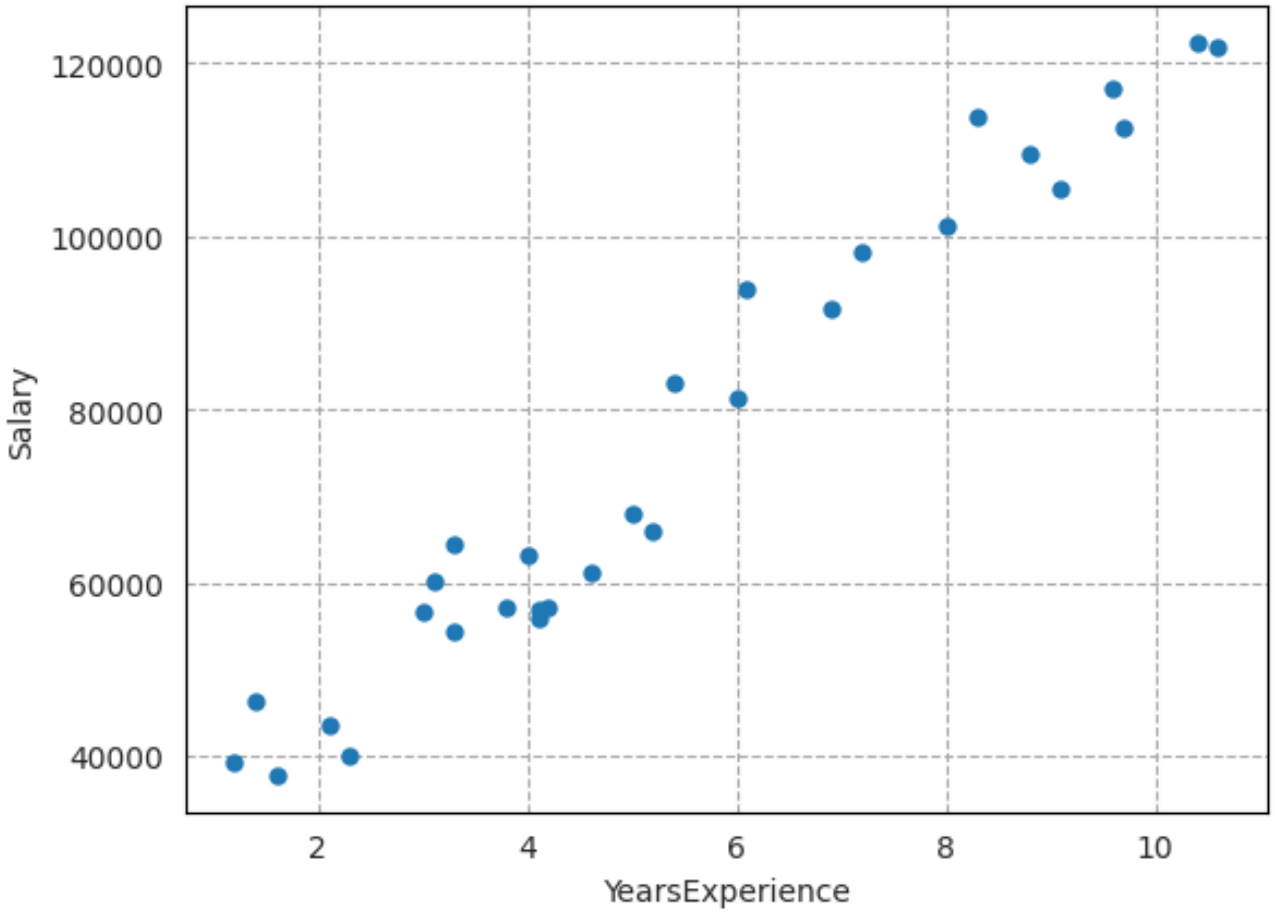
Hipótesis	Regla de decisión
H0: $\beta_1 = 0$	p-valor< α Rechazo H0
H1: $\beta_1 \neq 0$	

Test de hipótesis para la significatividad de $\hat{\beta}_0$

Hipótesis	Regla de decisión
H0: $\beta_0 = 0$	p-valor < α Rechazo H0
H1: $\beta_0 \neq 0$	

Ejemplo

Disponemos de datos de salario y años de experiencia, cuyo diagrama de dispersión y tabla de resultados son los siguientes:



¿Podría ser adecuado estimar un modelo de regresión lineal?

Ejemplo

Variable dependiente

```
                Results: Ordinary least squares
================================================================
Model:              OLS              Adj. R-squared:     0.955
Dependent Variable: Salary           AIC:                606.8823
Date:               2025-10-26 23:17 BIC:                609.6847
No. Observations:   30               Log-Likelihood:     -301.44
Df Model:           1                F-statistic:        622.5
Df Residuals:       28               Prob (F-statistic): 1.14e-20
R-squared:          0.957            Scale:              3.3505e+07
----------------------------------------------------------------
              Coef.     Std.Err.    t    P>|t|    [0.025     0.975]
----------------------------------------------------------------
const      24848.2040 2306.6537 10.7724 0.0000 20123.2380 29573.1699
x1          9449.9623  378.7546 24.9501 0.0000  8674.1187 10225.8059
----------------------------------------------------------------
```

Coeficiente de determinación

Variable independiente

Ordenada al origen

Los betas

tobs para test de significatividad

P-valor para test de significatividad

Ejemplo

Responder las siguientes preguntas:

1. Especificar el modelo obtenido.

2. Interpretar los coeficientes del modelo. ¿Son significativos?

3. ¿Cuánto se espera vender si los años de experiencia son 5?

4. Sabiendo que el SCE es 938128551.67. Calcular SCR y SCT.

5. ¿Qué porcentaje de la variación del salario es explicada por el modelo?

6. Calcular el coeficiente de correlación. ¿Es positivo o negativo? ¿Porqué?.

7. La relación lineal, ¿es directa o inversa? ¿Es débil, moderada o fuerte?

Pueden realizar de la práctica 8