4.2 Modelo de regresión lineal

Sir Francis Galton, un experto inglés en estudios de herencia, fue uno de los primeros investigadores que estudió relaciones entre variables por los años 1800. En un conocido ejemplo, Galton investigó la relación existente entre la altura de los hijos con respecto a la de sus padres.

De acuerdo a los estudios realizados, observó que padres altos tenían hijos altos pero no tanto como sus padres. De la misma manera, padres bajos tenían hijos bajos pero no tan bajos como ellos. Las tendencias de las alturas de los hijos eran más hacia un cierto promedio de la población que hacia las alturas de sus respectivos padres. Galton expresó que las alturas de los hijos regresaban a un promedio y de allí surgió el término regresión.

En la actualidad, la palabra regresión se utiliza para definir la naturaleza de la relación entre dos o más variables. A partir del concepto de que para cada valor de la variable x se genera una distribución de valores de la variable y, el modelo de regresión estima una recta promedio que denominaremos $\mu_{y/x}$, que explicaremos a continuación.

En primer lugar estudiaremos el modelo de regresión más simple, aquel que considera la existencia de sólo dos variables y haciendo la suposición de que la forma de la relación que liga a estas dos variables responde a la ecuación de una recta: es lineal.

Una vez comprendido este modelo, se puede generalizar al caso en que se cuenta con más de dos variables que generalmente deriva en lo que se conoce como regresión lineal múltiple. También el mismo procedimiento se puede utilizar cuando la relación que liga a las variables en estudio no es lineal: regresión curvilínea.

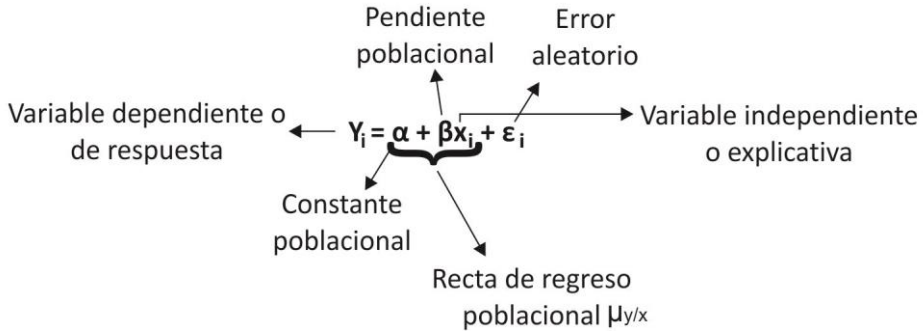
4.2.1. El modelo de regresión

La ecuación de regresión es una recta que describe la dependencia del valor promedio de una variable sobre otra.

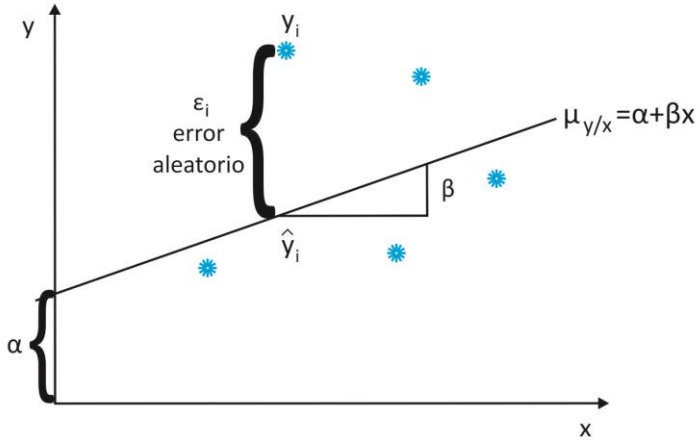
El modelo de regresión se expresa, matemáticamente, de la siguiente manera:

$$\text{siendo } y_i = \alpha + \beta \, . \, x_i + \varepsilon$$

¿Quiénes son los componentes de la ecuación de regresión poblacional?



Gráficamente se tiene:



$$\mu_{y/x}=\alpha+\beta x$$

$$V(y) = \sigma^2$$

El modelo de regresión incluye dos partes bien definidas. Por un lado tenemos una ecuación que expresa la forma de la relación que liga a la variable de respuesta con la variable x, es decir la ecuación $\mu_{y/x}= \alpha + \beta x$. A su vez, el término ε engloba a una determinada cantidad de factores que influyen sobre y pero que no están incluidos en x y están representados por el denominado error aleatorio.

En síntesis, el término ε indica en qué medida las variables x e y se apartan de la **relación lineal**.
Daremos a continuación, algunos conceptos teóricos.

Supuestos del Modelo de Regresión

- **Supuesto 1**

La variable aleatoria ε está distribuida normalmente.

Este supuesto se establece por conveniencia debido a que la aplicación de pruebas de hipótesis y determinación de intervalos de confianza son más sencillos.

- **Supuesto 2**

La variable aleatoria ε tiene media 0: $E(\varepsilon) = 0$.

Para cualquier valor de x_i, se supone que las diferencias entre los valores de y y_i $\mu_{y/x}$ algunas veces son positivas y otras negativas. Estas diferencias se compensan y se produce que $E(\varepsilon) = 0$.

- **Supuesto 3**

Los errores ε_i y ε_j son estadísticamente independientes uno de otro.

Esto significa que el signo y tamaño de un error no condiciona el signo o magnitud del otro. En símbolos: cov $(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$. Este supuesto es violado generalmente cuando las observaciones se dan a través del tiempo.

Por ejemplo, si se está estudiando la demanda mensual de electrodomésticos, es muy probable que la demanda de un mes determinado esté condicionada a la demanda del mes anterior. Si un mes se compran muchos electrodomésticos, es probable que al mes siguiente esta demanda disminuya por producirse una cierta saturación del mercado.

- **Supuesto 4**

La variable aleatoria tiene varianza finita σ^2, que es constante para todos los valores de (x_i, y_i)

Por ejemplo, si se quisiera estudiar la distribución del consumo (y) en función del ingreso (x), es probable que para valores chicos de x, la distribución del consumo sea más homogénea pues en estos niveles de ingreso, se consume prácticamente todo lo que se gana. En cambio, para valores altos de ingreso, la dispersión del consumo será más grande ya que las decisiones con respecto al ingreso serán más dispares. Habrá familias que gastan más, otras que ahorran, otras que invierten, etc.

Ya veremos más adelante como verificamos el cumplimiento de estos importantes supuestos.

Estos supuestos también pueden aplicarse a la variable y_i de la siguiente manera:

1) La distribución de la variable y es normal.

2) El valor promedio de y (esperanza) depende linealmente de x y de los parámetros α y β:

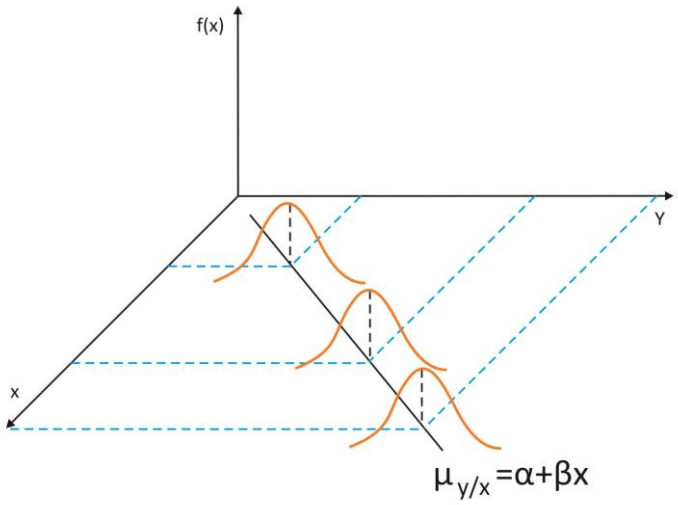
$$\mu_{y/x}=\alpha+\beta x$$

3) Las observaciones yi son independientes entre sí.

4) La varianza de y es constante:

$$V(y) = \sigma^2$$

Podemos resumir los supuestos establecidos en el siguiente gráfico:



Este gráfico expresa que para cada valor de x, la distribución de probabilidad de y es normal, con varianza constante σ^2 y promedio $\mu_{y/x}$ que varía linealmente cuando cambia x.

Cuando se toma una muestra, se tiene generalmente una observación de y para cada x. Además, se supone que las x_i son fijas, no variables aleatorias.

El conjunto de hipótesis sobre las que se basa el modelo de regresión versa sobre los siguientes aspectos:

- La forma funcional de la regresión es lineal en la variable y en los párametros.
- Se ha realizado una correcta especificación del modelo (x es la única variable independiente).
- La variable x no es una variable aleatoria.
- Los parámetros α y β se pueden estimar de una única manera.

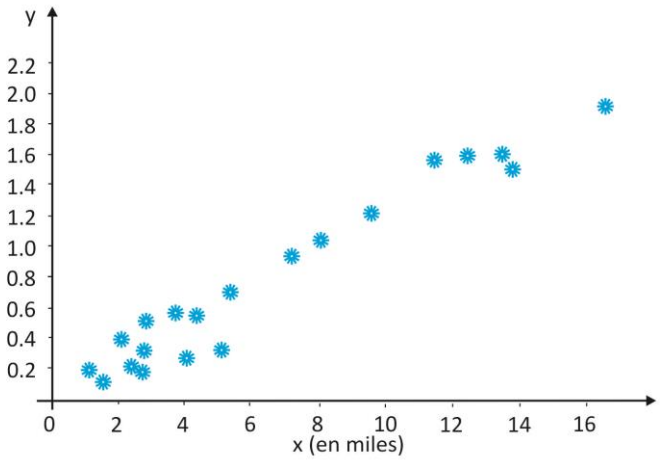
- $E(\varepsilon) = 0$.
- Cov $(\varepsilon_i , \varepsilon_j) = 0$ y $V(\varepsilon) = \sigma^2$.
- La distribución de los ε es $N(0, \sigma^2)$.

A continuación veremos cómo opera un modelo de regresión mediante un ejemplo.

Supongamos que se posee la siguiente información sobre el costo promedio mensual por alumno (y) en miles de $ y el número de alumnos (x) correspondiente a 20 universidades.

Universidad	Costo (y)	Cant. de estud. (x)
1	0,11	1564
2	0,56	3790
3	1,56	11383
4	0,69	5340
5	1,03	8028
6	0,49	2841
7	1,51	13744
8	1,58	12421
9	0,54	4348
10	0,93	7128
11	1,21	9578
12	1,58	13489
13	1,92	16545
14	0,18	1149
15	0,25	4045
16	0,31	5105
17	0,36	2102
18	0,17	2660
19	0,29	2754
20	0,22	2475

Los datos pueden representarse en un diagrama de dispersión donde en el eje de la abscisa graficamos la variable x (cantidad de alumnos) y en la ordenada la variable y (costo mensual por alumno).



El gráfico parece indicar una cierta relación positiva entre el costo y la cantidad de estudiantes inscriptos pues se observa que al aumentar la cantidad de alumnos aumenta también el costo por alumno.

Siempre se debe realizar un **diagrama de dispersión** para tener una idea previa de la forma de cualquier **relación entre variables** y para detectar **valores outliers**.

En el gráfico también puede estar representada la línea de regresión poblacional $\mu_{y/x} = \alpha + \beta x$. A su vez, la distancia de cualquier punto y a la recta de regresión poblacional es lo que hemos denominado error aleatorio y simbolizado con ε_i.

4.2.2. Estimación de la recta de regresión poblacional

El análisis de regresión es una metodología estadística que permite estimar la línea de regresión poblacional utilizando una línea de regresión calculada con los datos de una muestra.

Simbolizaremos a la regresión muestral como:

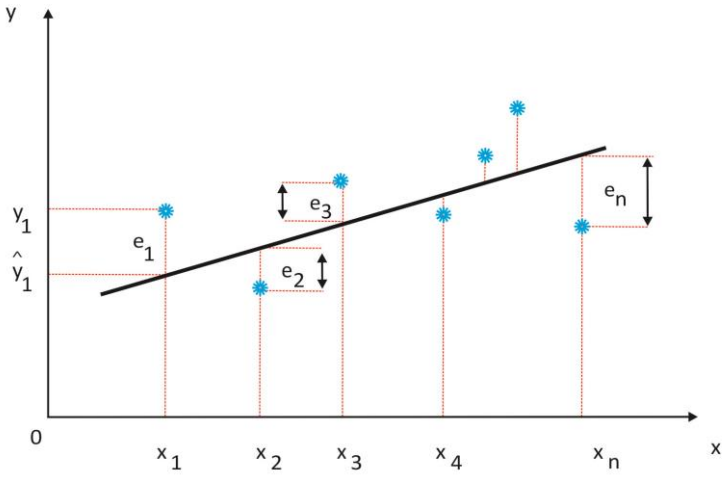
$$y = a + b\,x + e$$

Donde a será el estimador de la ordenada al origen α; b será el estimador de β, pendiente de la recta, y e representa al error aleatorio ε en cada una de las observaciones muestrales.

Para estimar α y β necesitamos utilizar un procedimiento matemático que permita obtener la recta que mejor se ajusta a los datos de la muestra. El problema radica en establecer claramente qué significa el **mejor ajuste**.

Definimos como la recta de mejor ajuste aquella que minimiza las distancias entre los puntos realmente observados y aquellos que se posicionen exactamente sobre la recta de regresión estimada. En otras palabras, $e_i = y_i - y_i$, donde $y_i = a + b\,x$, se obtiene a partir de la muestra.

Gráficamente, tenemos:



El método de estimación más utilizado es el denominado método de **mínimos cuadrados**. El método de mínimos cuadrados permite estimar la recta de regresión minimizando la suma de los errores e.

La estimación mínimo cuadrática minimiza la siguiente suma:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Si reemplazamos y y por su equivalente, la recta de regresión muestral, tenemos:

$$\sum_{i=1}^{n} (y_i - a - b\,x_i)^2$$

LECTURA OPTATIVA

El método de mínimos cuadrados es un procedimiento matemático cuya justificación escapa a los alcances de este curso. Para aquellos participantes que tengan conocimiento del tema explicaremos brevemente el procedimiento. Para los participantes que no tengan formación matemática, les sugerimos tomar e interpretar sólo las conclusiones.

Calcular el mínimo de una función consiste en obtener la derivada primera con respecto a los parámetros que queremos estimar, igualar estas derivadas parciales a 0 y resolver el sistema de ecuaciones simultáneas que resulta.

Derivando la función con respecto a **a** e igualando a 0, se obtiene:

$$\sum_{i=1}^{k} y_i = na + b\sum_{i=1}^{k} x_i$$

Realizando el mismo procedimiento con respecto a **b** tenemos la segunda ecuación:

$$\sum_{i=1}^{k} x_i y_i = a\sum_{i=1}^{k} x_i + b\sum_{i=1}^{k} x_i^2$$

Estas ecuaciones se conocen con el nombre de **ecuaciones normales**.

Resolviendo este sistema de dos ecuaciones con dos incógnitas, obtenemos la estimación de α y β de la siguiente manera:

$$b = \frac{\text{cov}(x, y)}{V(x)} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/n}{\sum (x_i - x)^2/n} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - x)^2}$$

y reemplazando en la primera ecuación se obtiene también a, esto es:

$$a = \bar{y} - b\,\bar{x}$$

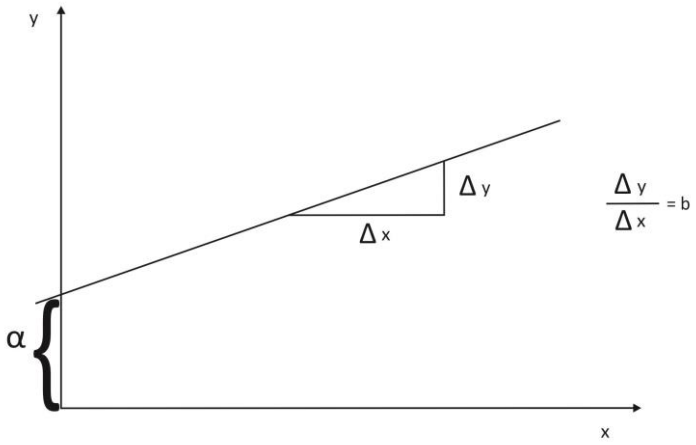
Se puede demostrar que los valores de a y b obtenidos, hacen mínima la suma de los cuadrados de errores: $\Sigma\, e_i^2$

Luego, calculando a y b podemos especificar la recta de regresión muestral:

$$\hat{y} = a + b\,x$$

Cualquier punto (x, y) sobre esta línea tiene una coordenada x (o abscisa) y una coordenada y (u ordenada) cuyos valores satisfacen la ecuación.

Cuando x = 0, y = a. Entonces, a es el punto donde la línea cruza al eje y por ello es llamada ordenada al origen. Cuando a es 0 la línea pasa por el origen. Un cambio unitario en x produce un cambio de b unidades en y por ello b es la pendiente de la recta.



Si b es positiva, ambas variables crecen o decrecen juntas y si b es negativa, una variable crece y la otra decrece.

En Matemática, este tipo de relaciones se denominan relaciones funcionales. Dado un valor de x, la relación funcional asigna un valor a y_i puesto que las dos variables están relacionadas por una fórmula matemática exacta.

En cambio, en Estadística se debe tener en cuenta que el investigador trata con observaciones sujetas a error y probablemente ninguna de ellas caiga exactamente sobre la recta. La aleatoriedad de los datos hace imposible una relación perfecta aunque exista una relación funcional entre las variables.

Estimaremos la recta de regresión utilizando los datos del ejemplo referido a una muestra de 20 universidades.

En cada universidad se han registrado los valores de dos variables:

y = costo promedio mensual por alumno
x = cantidad de alumnos inscriptos

Vamos a construir una tabla apropiada para facilitar los cálculos.

y_i	x_i	y_i^2	x_i^2	$x_i\, y_i$
0,11	1564	0,0121	2446096	172,04
0,56	3790	0,3136	14364100	2122,40
1,56	11383	2,4336	129572689	17757,48
0,69	5340	0,4761	28515600	3684,60
1,03	8028	1,0609	64448784	8268,84
0,49	2841	0,2401	8071281	1392,09
1,51	13744	2,2801	188897536	20753,44
1,58	12421	2,4964	154281241	19625,18
0,54	4348	0,2916	18905104	2347,92
0,93	7128	0,8649	50808384	6629,04
1,21	9578	1,4641	91738084	11589,38
1,58	13489	2,4964	181953121	21312,62
1,92	16545	3,6864	273737025	31766,40
0,18	1149	0,0324	1320201	206,82
0,25	4045	0,0625	16362025	1011,25
0,31	5105	0,0961	26061025	1582,55
0,36	2102	0,1296	4418404	756,72
0,17	2660	0,0289	7075600	452,20
0,29	2754	0,0841	7584516	798,66
0,22	2475	0,0484	6125625	544,50
15,49	130489	18,5983	1276686441	152774,13

$\bar{y} = 0{,}7745 \qquad \bar{x} = 6524{,}45$

$$S_y^2 = \frac{18{,}5983}{20} - (0{,}7745)^2 = 0{,}9299 - 0{,}5998 = 0{,}3301$$

$$S_x^2 = \frac{1276686441}{20} - (6524{,}45)^2 = 63834322{,}05 - 42568447{,}80 = 21265874{,}25$$

$S_y = 0{,}5745$

$S_x = 4611{,}49$

Calculamos la covarianza entre x e y aplicando la fórmula de trabajo:

$$\text{cov}(x, y) = \frac{\sum x_i\, y_i}{n} - \bar{x}\,\bar{y}$$

$$\text{cov}(x, y) = \frac{152774{,}13}{20} - 0{,}7745 - 6524{,}45 = 7638{,}7065 - 5053{,}1865 = 2585{,}52$$

Ya podemos calcular el valor de b.

$$b = \frac{2585{,}52}{21265874{,}25} = 0{,}0001216$$

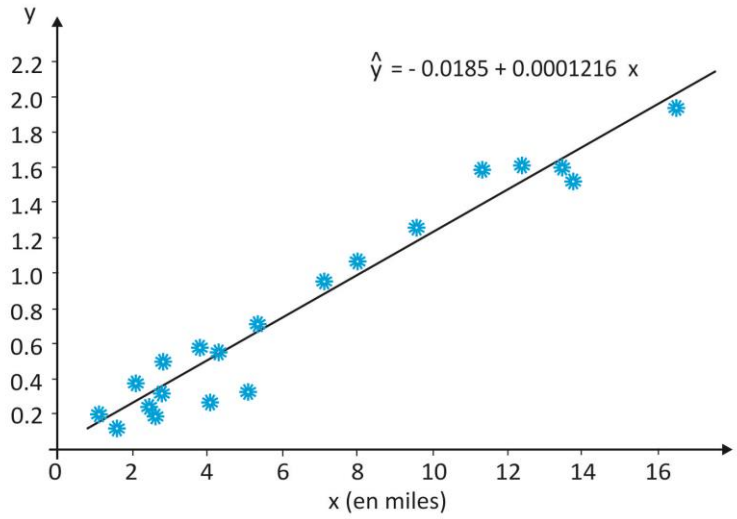
A continuación calculamos el valor de a.

$$a = 0{,}7745 - 0{,}0001216 \,.\, 6524{,}45 = 0{,}7745 - 0{,}793 = -\,0{,}0185$$

La recta de regresión estimada es, entonces:

$$\hat{y} = -\,0{,}0185 + 0{,}0001216\, x$$

Ahora podemos representar los puntos realmente observados y la recta de regresión mínimo cuadrática estimada.



¿Cómo interpretamos a y b? a es la ordenada al origen, indica el valor de la variable dependiente y cuando la variable independiente x = 0. Si no hubiera alumnos inscriptos, el costo promedio por alumno es negativo: - 0,02, b indica cuanto varía y cuando x aumenta una unidad. En este caso, por cada alumno inscripto, el costo promedio aumenta $ 0,0001.