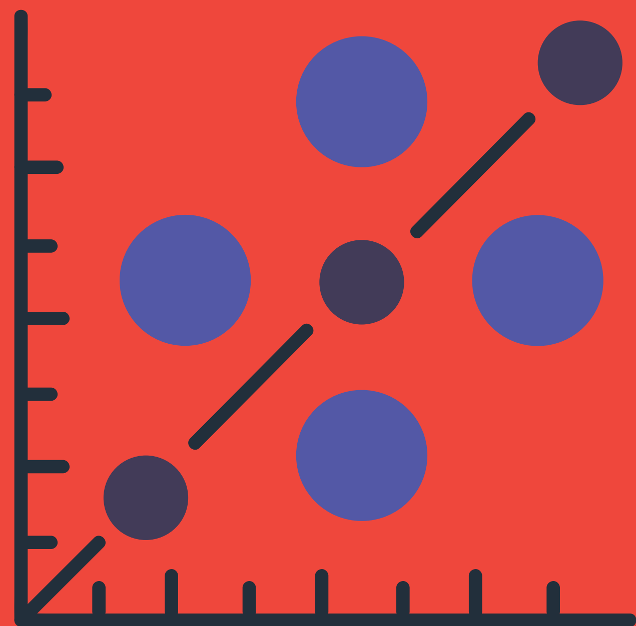
Universidad de Talca, Chile



Regresión Lineal Simple

MATEMÁTICA

Vicerrectoría de Pregrado
CIVE Dirección de Ciclo de Vida Estudiantil
CIMA Unidad de Acompañamiento Estudiantil
PACE Programa de Acceso a la Educación Superior
Ministerio de Educación

RUTA DE APRENDIZAJE

El aprendizaje esperado en este documento es conocer el modelo de regresión lineal simple, para analizar la naturaleza e intensidad de las relaciones entre las variables de un estudio.

Covarianza → Gráfico de dispersión → Correlación → Regresión lineal simple → Método de aproximación por mínimos cuadrados

ÍNDICE

Introducción

En diversas áreas frecuentemente es necesario conocer la relación existente entre variables para analizar el efecto que pudiese tener una sobre la otra. Por ejemplo, en salud, se podría tener interés en estudiar si la edad afecta la presión sanguínea, si existe relación en la ingesta de algún nutriente con el aumento de peso, si la concentración de algún medicamento altera la frecuencia cardiaca. En economía las relaciones entre variables entrega información para aumentar las ganancias, reducir costos, estimar la demanda, entre otros.

En este documento, se describe el análisis de regresión, que es útil para averiguar la forma probable de relación entre las variables, cuando se utiliza este método de análisis, el objetivo final es por lo general predecir o estimar el valor de una variable que corresponde a un valor determinado de otra variable (Wayne, 1991).

Modelo de regresión lineal simple

Es el procedimiento que desarrolla una ecuación matemática que expresa la relación lineal entre dos variables y permite estimar el valor de una con base, en el valor de la otra (Lind et al., 2012).
Para el modelo de regresión lineal simple son importantes dos variables:X, que se conoce como la variable independiente, ya que es controlable por el investigador, y la variable Y, conocida como variable dependiente que se obtiene a partir de la regresión de Y sobre X.

Suposiciones que fundamentan la regresión lineal simple

Las suposiciones que fundamentan el modelo de regresión lineal simple son las siguientes:

1. Los valores de la variable X son “fijos”, lo que significa que los valores son preseleccionados por el investigador, de modo que en la recolección de datos no se permite que estos varíen.

2. La variable X se mide sin error, lo que significa que se desprecia esta magnitud.

3. Para cada valor de X existe una subpoblación de valores para Y que debe tener una distribución normal, así son válidos los procedimientos comunes de inferencia estadística.

4. Todas las varianzas de las subpoblaciones de Y son iguales.

5. Suposición de linealidad, todas las medias de las subpoblaciones de Y están sobre la misma línea recta.

6. Los valores de Y son estadísticamente independientes. En otras palabras, al extraer la muestra, se supone que los valores de Y, obtenidos para un valor de X, de ninguna manera dependen de los valores de Y elegidos para otro valor de X (Wayne, 1991).

Ecuación de regresión de la muestra

En la regresión lineal simple, el objeto de interés del investigador es la ecuación de regresión de la población, que describe la relación entre las variables. Para encontrar la probable relación, el investigador normalmente extrae una muestra de la población de interés y según los datos obtenidos calcula una ecuación de regresión de la muestra (Wayne, 1991).
La ecuación general de una línea recta es:

$$y=a+bx$$

Donde

$a=intercepto$, valor de y para $x=0$
$b=pendiente$, cambio en y por unidad de cambio en x.

Por otro lado, de acuerdo al modelo de regresión lineal, los promedios de estas distribuciones se disponen en una línea recta, cuya ecuación es:

$$E(x)=\alpha+\beta x$$

Donde

$E(x)=\hat{y}$ es el valor de la estimación de la variable y para un x seleccionado.
α=intersección y , valor estimado de y para x=0
β=pendiente, cambio promedio en $\hat{y}$ por unidad de cambio en x (Taucher, 2014).

Recordando

En matemática la regresión lineal entrega valores exactos de y para cada valor de x. Por otre lado en estadística, el modelo de regresión lineal entrega valores estimados de y para cada x.

Método de aproximación por mínimos cuadrados

La primera aproximación para analizar la relación entre las variables es utilizar un diagrama de dispersión, así visualizar la posición de la línea recta. La desventaja de esto, es que solo entrega una idea de la posición de la línea basada en el criterio del observador (Lind et al.,2012).

El ajuste más adecuado para obtener los valores estimados, es por el criterio de mínimos cuadrados, esta recta hace mínima la suma de las diferencias cuadráticas entre cada valor y y el valor predicho en x_i por la línea de regresión (Taucher, 2014).

Para ilustrar esta idea se trazan los mismos datos en tres gráficas de la **imagen 1**. Los puntos son los valores reales de Y, y los asteriscos son los valores predichos de Y para un valor dado de X. La recta de la gráfica 13-9 se determinó con el método de los mínimos cuadrados, es la que mejor se ajusta porque la suma de los cuadrados de las diferencias verticales es mínima.

- Suma de los cuadrados de las desviaciones verticales gráfica 13-9

$$sc = 2^2 + \ 4^2 + 2^2$$
$$sc = 4 + 16 + 4 = 24$$

- Suma de los cuadrados de las desviaciones verticales gráfica 13-10

$$sc = 2^2 + \ 6^2 + 2^2$$
$$sc = 4 + 36 + 4 = 44$$

- Suma de los cuadrados de las desviaciones verticales gráfica 13-10

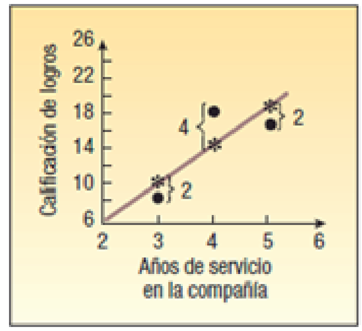
$$sc = 8^2 + \ 2^2 + 8^2$$
$$sc = 64 + 4 + 64 = 132$$

El propósito de un análisis de regresión es calcular los valores de α y β para desarrollar una ecuación lineal que se ajuste mejor a los datos, según el método de mínimos cuadrados estos parámetros se pueden estimar por medio de las siguientes fórmulas:

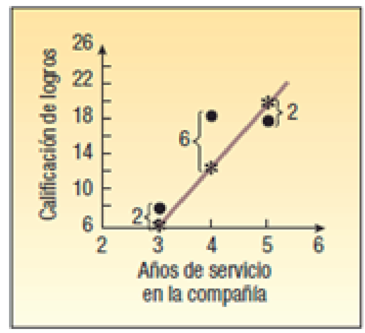
$\beta = \dfrac{n\Sigma x_i \cdot y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$	$\alpha = \bar{y} - \beta\bar{x}$

La última ecuación muestra que la recta ajustada pasa por el punto *(x ,y)*, es decir, por la media de las distribuciones (Taucher, 2014), ya que, recordemos que la fórmula de esta recta es:

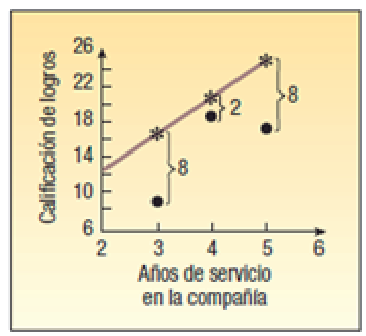
$$\hat{y} = \alpha + \beta x$$



GRÁFICA 13-9 Recta de mínimos cuadrados



GRÁFICA 13-10 Recta trazada con una regla



GRÁFICA 13-11 Recta diferente trazada con una regla

Imagen 3

Imagen 1: Diferencia de las rectas de los mismos datos graficadas por el método de los mínimos cuadrados y con regla (Lind et al., 2012).

LEE Y ANALIZA LOS SIGUIENTES PROBLEMAS

Ejercicios resueltos

A continuación, se presentan ejercicios resueltos con sus procedimientos, en estos problemas se sugiere hacer lo siguiente:

- Lee comprensivamente.
- Revisa el paso a paso.
- Destaca lo que te resulte importante.
- Destaca lo que te genere dudas y luego consulta al tutor.

1. Se investiga la capacidad vital en 8 niños de diferentes edades, con los siguientes resultados:

Tabla 1: Problema página 104 de Bioestadística para carreras de la salud (Taucher, 2014).

Niño	Edad (años)	Capacidad Vital
1	4	0,79
2	5	0,93
3	6	1,15
4	7	1,29
5	8	1,47
6	9	1,71
7	10	1,87
8	11	1,99

Ajustar la recta de regresión para analizar la relación entre la edad de los infantes y su capacidad vital, interpretar los resultados.

Paso 1: Encontrar el valor de β
La fórmula que se debe utilizar es:

$$\beta = \frac{n\Sigma x_i \cdot y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

a) Obtener $\Sigma x_i \cdot yi$

Niño	**Edad x (años)**	**Capacidad Vital y**	$x \cdot y$
1	4	0,79	$4 \cdot 0{,}79 = 3{,}16$
2	5	0,93	$5 \cdot 0{,}93 = 4{,}65$
3	6	1,15	$6 \cdot 1{,}15 = 6{,}9$
4	7	1,29	$7 \cdot 1{,}29 = 9{,}03$
5	8	1,47	$8 \cdot 1{,}47 = 11{,}76$
6	9	1,71	$9 \cdot 1{,}71 = 15{,}39$
7	10	1,87	$10 \cdot 1{,}87 = 18{,}7$
8	11	1,99	$11 \cdot 1{,}99 = 21{,}89$
TOTAL	60	11,2	91,48

Luego, $\Sigma x_i \cdot y_i$ = *91,48*

b) Calcular $\Sigma x_i \Sigma y_i$, es decir, multiplicar el total de la suma de los x e y

$$\Sigma x_i \Sigma y_i = 60 \cdot 11{,}2$$

$$\Sigma x_i \Sigma y_i = 672$$

c) Determinar Σx_i^2

Niño	Edad x (años)	Capacidad Vital y	$x \cdot y$	x^2
1	4	0,79	3,16	$4^2 = 16$
2	5	0,93	4,65	$5^2 = 25$
3	6	1,15	6,9	$6^2 = 36$
4	7	1,29	9,03	$7^2 = 49$
5	8	1,47	11,76	$8^2 = 64$
6	9	1,71	15,39	$9^2 = 81$
7	10	1,87	18,7	$10^2 = 100$
8	11	1,99	21,89	$11^2 = 121$
TOTAL	60	11,2	91,48	492

Luego, Σx_i^2 =492

d) Calcula $(\Sigma x_i)^2$

$$(\Sigma x_i)^2 = 60^2$$
$$(\Sigma x_i)^2 = 3600$$

e) Reemplazar los resultados obtenidos en la fórmula:

$$\beta = \frac{n\Sigma x_i \cdot y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

En el reemplazo recordar que n=8, ya que son los datos de 8 niños, entonces se tiene:

$$\beta = \frac{8 \cdot 91{,}48 - 672}{8 \cdot 492 - 3600}$$

$$\beta = \frac{731{,}84 - 672}{3936 - 3600}$$

$$\beta = \frac{59{,}84}{336}$$

$$\beta = 0{,}18$$

Paso 2: Encontrar el valor de α

La fórmula que se debe utilizar es:

$$\alpha = \bar{y} - \beta\bar{x}$$

a) Calcular el promedio de los valores de y

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Del paso 1 tenemos que $\sum_{i=1}^{n} y_i$ =11,2, y sabemos que n=8 , reemplazando:

$$\bar{y} = \frac{11{,}2}{8}$$

$$\bar{y} = 1{,}4$$

b) Determinar el promedio de los valores de x

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Del paso 1 tenemos que $\sum_{i=1}^{n} x_i = 60$, y sabemos que n=8 , reemplazando:

$$\bar{x} = \frac{60}{8}$$

$$\bar{x} = 7{,}5$$

c) Reemplazar en la fórmula:

$$\alpha = \bar{y} - \beta\bar{x}$$

Recordar que por el paso 1 tenemos β =0,18, entonces se tiene:

$$\alpha = 1{,}4 - 0{,}18 \cdot 7{,}5$$

$$\alpha = 1{,}4 - 1{,}35$$

$$\alpha = 0{,}05$$

Paso 3: ajustar la recta con los datos obtenidos anteriormente

La fórmula en la que se debe reemplazar es:

$$\hat{y} = \alpha + \beta x$$

Por lo que la recta ajustada está dada por:

$$\hat{y} = 0{,}05 + 0{,}18x$$

Paso 4: interpretar los valores de la recta obtenida

Según los valores obtenidos se pueden establecer las siguientes interpretaciones:

- Como β=0,18, por cada año que aumente la edad de un infante entre los 4 y 11 años, su capacidad vital se incrementa en 0,18 unidades.
- Al ser α=0,05, se podría decir que la capacidad vital media para el nacimiento es de 0,05 (Taucher, 2014).

PON A PRUEBA TUS CONOCIMIENTOS

Ejercicios propuestos

A continuación, se presentan ejercicios propuestos para que puedas resolver y practicar, recuerda hacer lo siguiente:

- Resuélvelos siguiendo los pasos utilizados en el ejercicio resuelto.
- Si es necesario apóyate con los apuntes.
- Si surgen dudas, registrarlas para luego consultar con el tutor.
- ¡Buen trabajo!

1. Un grupo de profesionales especialistas en salud mental de un hospital psiquiátrico, donde los pacientes permanecen mucho tiempo, deseaba estimar el nivel de respuesta de pacientes retraídos en un programa de terapia de remotivación. Para ello, se contaba con una prueba estandarizada, pero era incosteable y tardaba para administrarla. Para superar este obstáculo, el grupo desarrolló una prueba que era mucho más fácil de aplicar. Para probar la utilidad del nuevo instrumento para medir el nivel de respuesta del paciente, el grupo decidió estudiar la relación entre las calificaciones obtenidas con la nueva prueba y las calificaciones obtenidas con la prueba estandarizada. Para esto se seleccionaron 11 pacientes que habían rendido la nueva prueba para que hicieran la prueba estandarizada, los resultados obtenidos se presentan en la siguiente tabla (Wayne 1991).

Tabla 2 Calificaciones obtenidas por los pacientes en las pruebas nueva y estandarizada.

Número de paciente	Calificación obtenida en la nueva prueba (X)	Calificación obtenida en la prueba estandarizada (Y)
1	50	61
2	55	61
3	60	59
4	65	71
5	70	80
6	75	76
7	80	90
8	85	106
9	90	98
10	95	100
11	100	114

Calcula:

a) Pendiente de la recta ajustada (β).
b) Intercepto de la recta ajustada (α).
c) Recta ajustada.
d)Interprete los valores obtenidos en a) y b).
e) ¿Qué puntaje en la prueba estandarizada se estima que obtenga un paciente que tuvo un 50 en la prueba nueva?

2. Se llevó a cabo un experimento para estudiar el efecto de cierto medicamento para disminuir la frecuencia cardiaca en adultos. La variable independiente es la dosis en miligramos del medicamento, y la variable dependiente es la diferencia entre la frecuencia cardiaca más baja después de la administración del medicamento y un control antes de administrarlo. Se reunieron los siguientes datos.

Dosis (mg) X	Disminución de la frecuencia cardiaca (latidos/min) Y
0,5	10
0,75	8
1,00	12
1,25	12
1,5	14
1,75	12
2,00	16
2,25	18
2,5	17
2,75	20
3,00	18
3,25	20
3,5	21

(Wayne 1991)

Determina
a) Pendiente de la recta ajustada (β).
b) Intercepto de la recta ajustada (α).
c) Recta ajustada.

Respuestas

1.
   - a) 1,1236
   - b) -0,9973
   - c) y =-0,9973+1,1236x
   - d) Como β=1,1236 por cada punto que aumente en la prueba nueva, un paciente, aumenta 1,1236 puntos en la prueba estandarizada.
     Por otro lado, α no tiene interpretación dentro del contexto del problema ya que sería el puntaje en la prueba estandarizada, para alguien que obtuvo 0 puntos en la prueba nueva.
   - e) 55,1827
2.
   - a) 4,3806
   - b) 6,3204
   - c) $\hat{y}$ =6,3204+4,3806x

Síntesis

El principal uso del modelo de regresión lineal simple es estudiar la relación entre las variables presentadas. Para realizar el mejor ajuste de una recta a los datos presentados se utiliza el método de mínimos cuadrados que tiene las siguientes fórmulas:

Recta ajustada

$$\hat{y} = \alpha + \beta x$$

Pendiente de la recta

$$\beta = \frac{n\Sigma x_i \cdot y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2}$$

Intercepto de la recta

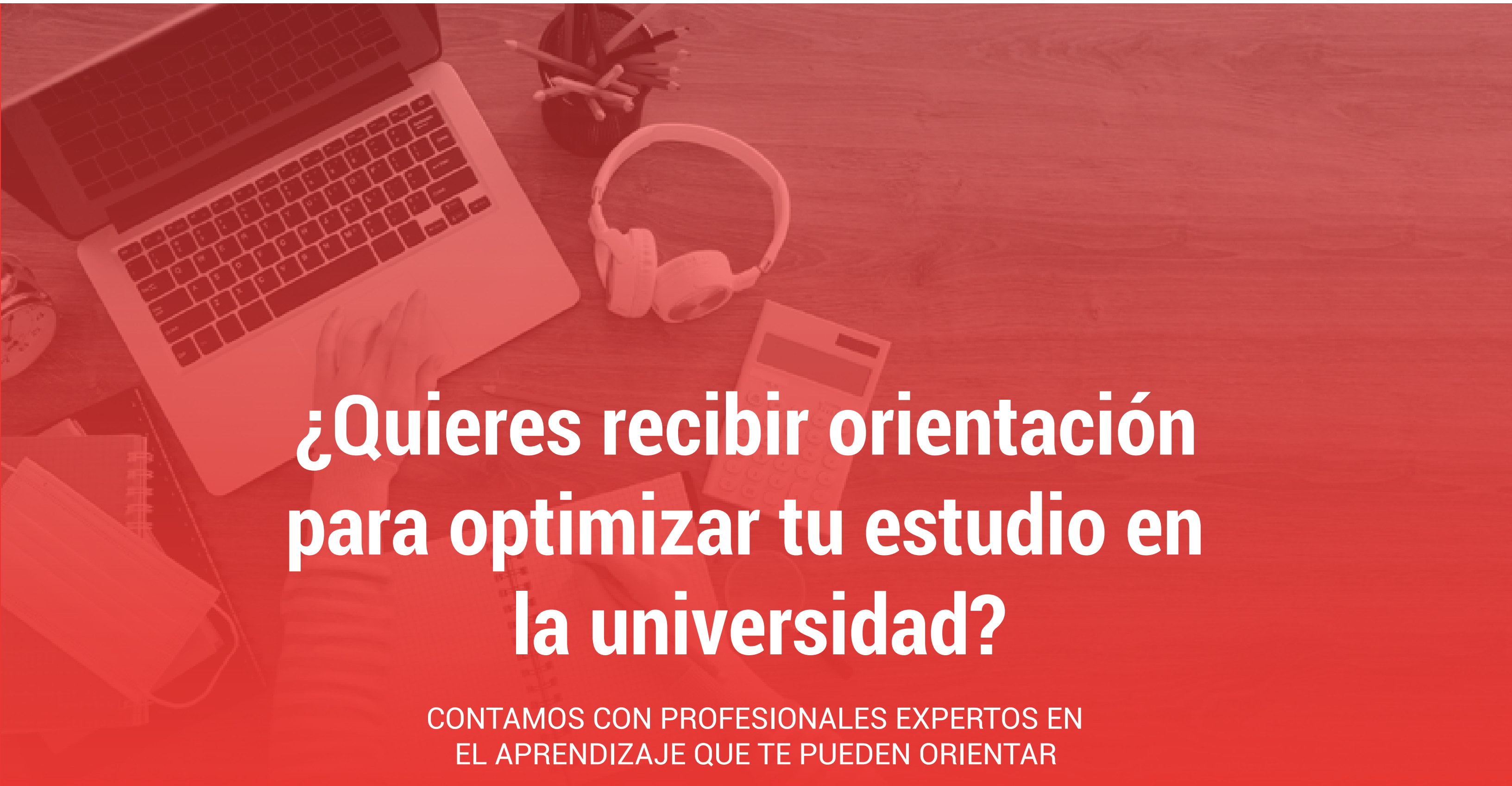
$$\alpha = \bar{y} - \beta \bar{x}$$

Referencia bibliográfica

Lind, D., Wathen, S., & Marchal , W. (2012). Estadística aplicada a los negocios y la economía . México: The McGraw-Hill Companies, Inc.

Taucher, E. (2014). Bioestadística . Ocho Libros Editores Ltda.

Wayne, D. (1991). Bioestadística base para el análisis de las ciencias y la salud. México : Limusa S.A. .



¿Quieres recibir orientación para optimizar tu estudio en la universidad?

CONTAMOS CON PROFESIONALES EXPERTOS EN
EL APRENDIZAJE QUE TE PUEDEN ORIENTAR

SOLICITA NUESTRO APOYO

Sitio Web de CIMA

Ver más fichas

Solicita más información