Regresión simple y correlación

Muchos estudios se basan en la creencia de que se puede identificar y cuantificar alguna relación funcional entre dos o más variables. Se puede decir que Y depende de X, en donde Y y X son dos variables cualquiera.

$$\text{Y es una función de X} \rightarrow Y = f(X)$$

Debido a que Y depende de X, Y es la **variable dependiente** y X es la **variable independiente**.

El decano de la facultad desea analizar la relación entre las notas de los estudiantes y el tiempo que pasan estudiando. Es lógico presumir que las notas dependen de la cantidad y calidad de tiempo que los estudiantes pasan con sus libros. Por lo tanto, **"notas"** es la variable dependiente y **"tiempo"** es la variable independiente.

Variable dependiente: es la variable que se desea explicar o predecir, también se le denomina variable de respuesta.

Variable independiente: es la variable que explica la dependiente, también se le denomina variable explicativa.

Existe la regresión simple y la regresión múltiple. En la regresión simple, se establece que Y es una función de sólo una variable independiente.

$$Y = f(X)$$

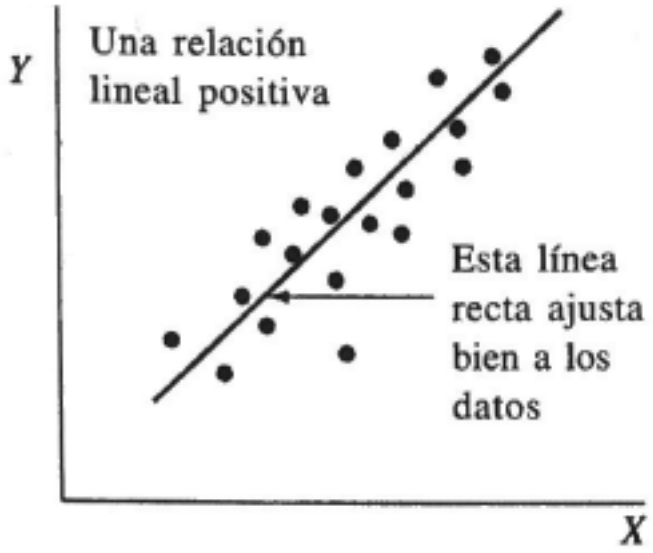
En un modelo de regresión múltiple, Y es una función de dos o más variables independientes.

$$Y = f(X_1, X_2, X_3, \cdots, X_k)$$

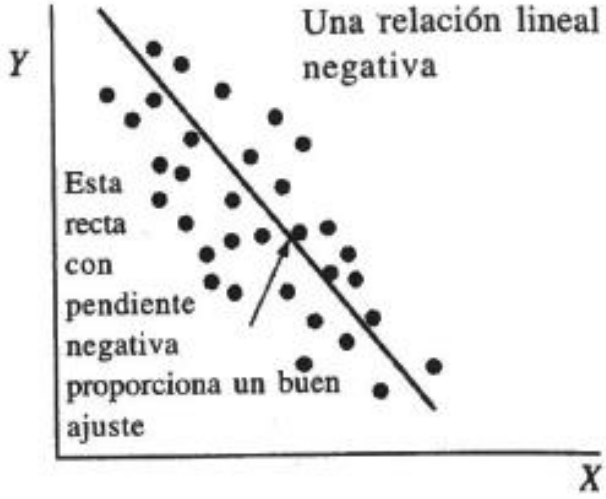
También es necesario hacer una distinción entre la regresión lineal y la regresión curvilínea (no lineal).

En el modelo de regresión lineal, la relación entre X y Y puede representarse por medio de una línea recta. Sostiene que a medida que X cambia, Y cambia en una cantidad constante.

La regresión curvilínea utiliza una curva para expresar la relación entre X y Y. Sostiene que a medida que X cambia, Y cambia en una cantidad diferente cada vez.

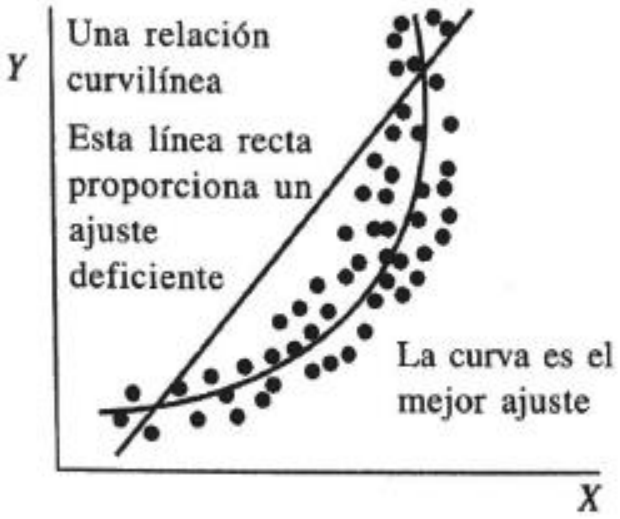


Relación positiva y lineal entre X y Y. Es positiva porque X y Y parecen moverse en la misma dirección. A medida que X aumenta (disminuye), Y aumenta (disminuye). Es lineal porque la relación puede identificarse mediante una línea recta que se dibuja entre los puntos.

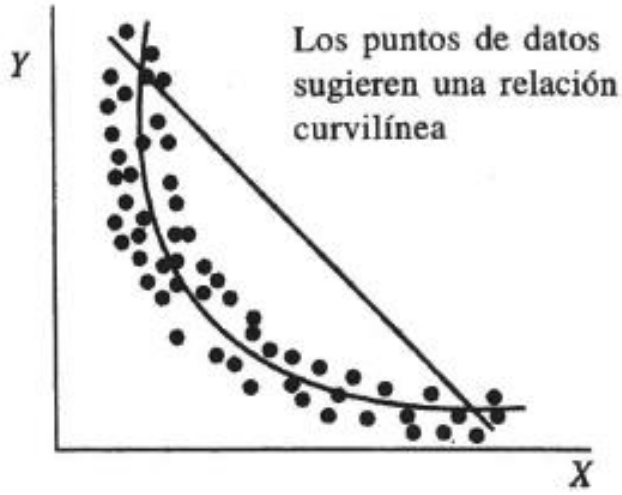


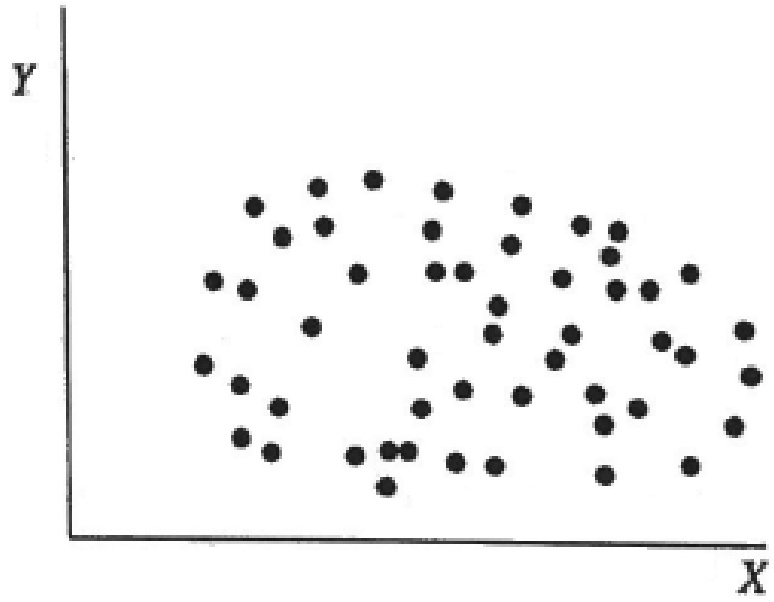
Relación lineal y negativa entre X y Y, porque las dos variables parecen moverse en direcciones opuestas.

Universidad de Medellín
Ciencia y Libertad



Los puntos de dispersión no se definen bien con la línea recta, pero se define de manera más exacta con la curva que proporciona un mejor ajuste.





Este diagrama de dispersión indica que no existe ninguna relación entre X y Y.

Determinación del modelo de regresión lineal simple

Sólo son necesarios dos puntos para dibujar la línea recta que representa esta relación lineal. La ecuación de una recta puede expresarse como:

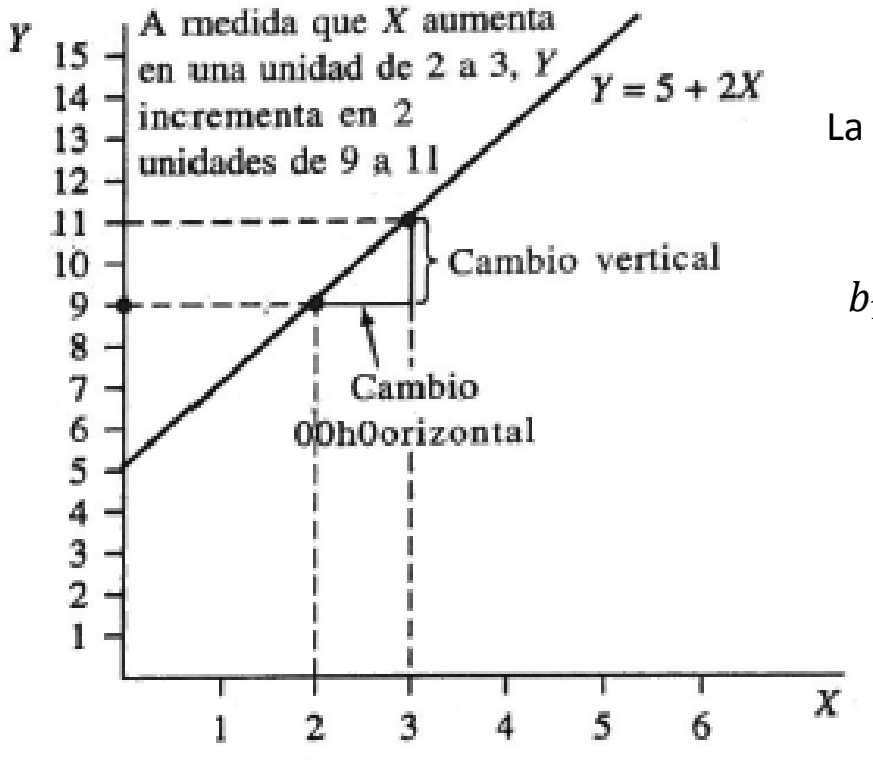
$$Y = b_0 + b_1 X$$

Donde:

b_0 es el intercepto.

b_1 es la pendiente de la recta.

$$Y = 5 + 2X$$

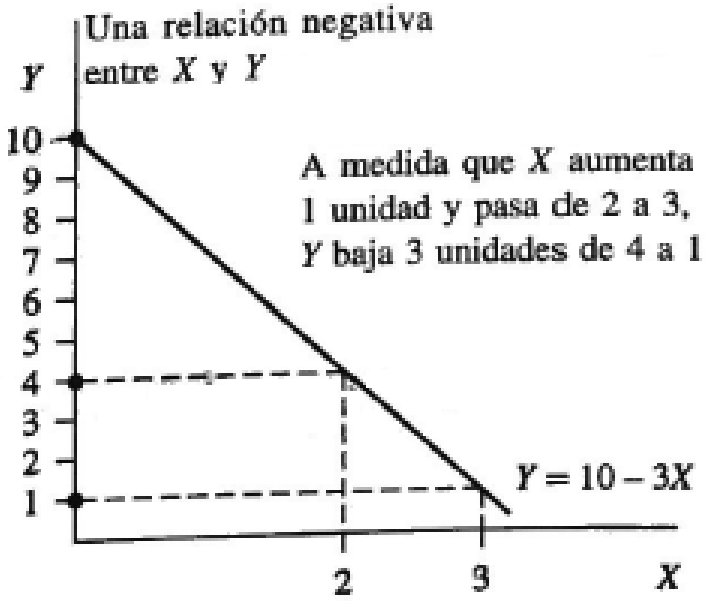


La recta intercepta el eje vertical en 5.

$$b_1 = pendiente = \frac{variación\ vertical}{variación\ horizontal} = \frac{2}{1} = 2$$

Universidad de Medellín
Ciencia y Libertad

$$Y = 10 - 3X$$

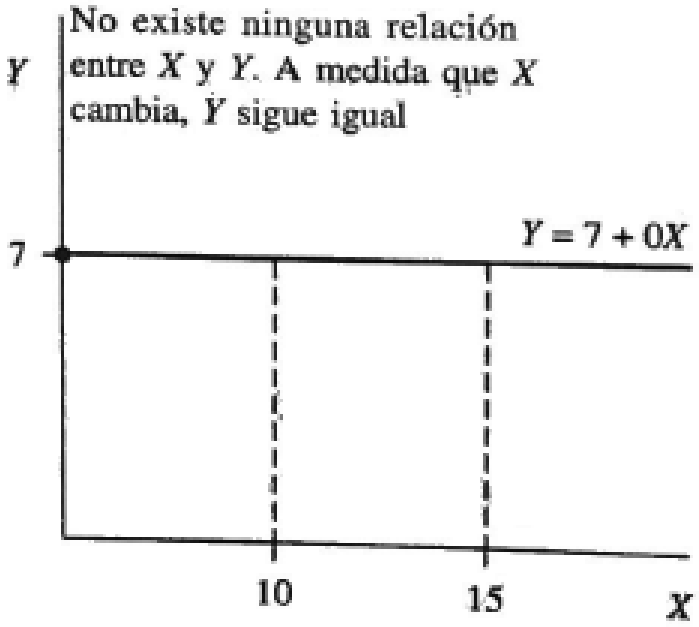


La recta intercepta el eje vertical en 10.

$$b_1 = pendiente = \frac{variación\ vertical}{variación\ horizontal} = \frac{-3}{1} = -3$$

Universidad de Medellín
Ciencia y Libertad

$$Y = 7 - 0X$$



La recta intercepta el eje vertical en 7.

$$b_1 = pendiente = \frac{variación\ vertical}{variación\ horizontal} = 0$$

X no puede utilizarse como variable explicativa de Y.

Modelo lineal estocástico

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Donde:

$\beta_0 + \beta_1 X$ es la porción determinística de la relación.

ε representa el carácter aleatorio que muestra la variable dependiente y por tanto denota el término del error en la expresión.

Los parámetros β_0 y β_1, al igual que la mayoría de parámetros, permanecerán desconocidos y se pueden estimar sólo con los datos muestrales.

Modelo lineal con base en datos muestrales

$$Y = b_0 + b_1 X + e$$

Donde:

b_0 y b_1 son estimaciones de β_0 y β_1.

e es el término aleatorio.

Debido a que e es aleatorio, Y solo puede estimarse.

Modelo de regresión estimada

$$\hat{Y} = b_0 + b_1 X$$

Donde:

$\hat{Y}$ es el valor estimado de Y.

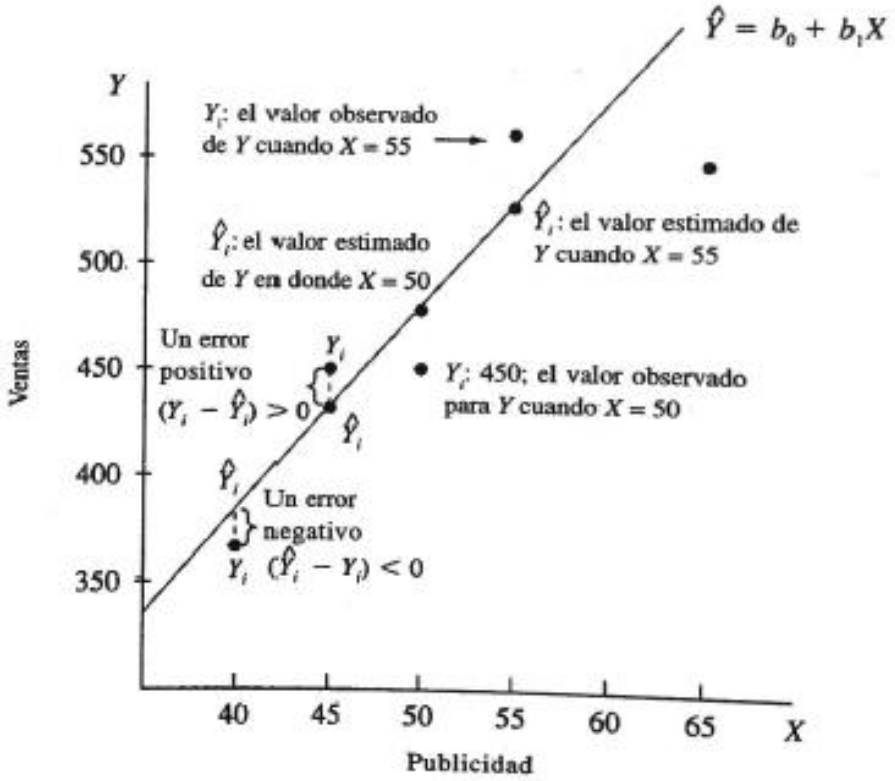
b_0 y b_1 son el intercepto y la pendiente de la recta de regresión estimada.

Mínimos cuadrados ordinarios: la recta de mejor ajuste

El propósito del análisis de regresión es determinar una recta que se ajuste a los datos muestrales mejor que cualquier otra recta que pueda dibujarse. Esta recta está determinada mediante la estimación de b_0 y b_1. Un procedimiento matemático utilizado para estimar estos valores se denomina **mínimos cuadrados ordinarios (MCO)**.

La empresa Vita + Plus recolecta datos sobre los gastos publicitarios y los ingresos por ventas de 5 meses:

Mes	Ventas (x US$ 1000)	Publicidad (x US$ 100)
1	450	50
2	380	40
3	540	65
4	500	55
5	420	45



Los datos Y_i en el diagrama de dispersión son los valores de los datos observados reales para Y en la tabla. Los valores $\hat{Y}$ se obtienen mediante la recta de regresión y representan el estimado de las ventas. La diferencia entre lo que Y era realmente, Y_i, y lo que se estima que es $\hat{Y}_i$, es el error.

$$Error = (Y_i - \hat{Y}_i)$$

MCO producirá una recta tal que la suma de esos errores sea cero:

$$\sum (Y_i - \hat{Y}_i) = 0$$

MCO también asegurará que se minimice la suma de estos errores al cuadrado:

$$\sum (Y_i - \hat{Y}_i)^2 = min$$

Donde:

$(Y_i - \hat{Y}_i)$ es el error de cada dato.

min es el valor mínimo.

Para determinar esta recta de mejor ajuste, MCO requiere que se calcule:

Suma de los cuadrados de X

$$SC_x = \sum (X_i - \bar{X})^2$$

$$SC_x = \sum X^2 - \frac{(\sum X)^2}{n}$$

Suma de los cuadrados de Y

$$SC_y = \sum (Y_i - \bar{Y})^2$$

$$SC_y = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

Suma de los productos cruzados de X y Y

$$SC_{xy} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$SC_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

Dadas las sumas de cuadrados y los productos cruzados, se puede calcular la pendiente de la recta de la regresión, llamada el coeficiente de regresión y el intercepto así:

Pendiente de la recta de regresión

$$b_1 = \frac{SC_{xy}}{SC_x}$$

Intercepto de la recta de regresión

$$b_0 = \bar{Y} - b_1\bar{X}$$

Donde:

$\bar{Y}\ y\ \bar{X}$ son las medias de los valores Y y los valores X.

La agencia de Hop Scotch Airlines , la aerolínea transportadora más pequeña del mundo, considera que existe una relación directa entre los gastos publicitarios y el número de pasajeros que escogen viajar por Hop Scotch. Para determinar si esta relación existe, y si es así cuál podría ser la naturaleza exacta, los estadísticos empleados por Hop Scotch decidieron utilizar los procedimientos MCO para determinar el modelo de regresión.

Se recolectaron los valores mensuales por gastos de publicidad y número de pasajeros para los n = 15 meses más recientes.

Observación (mes)	Publicidad (en US$1000) (X)	Pasajeros (en US$1000) (Y)
1	10	15
2	12	17
3	8	13
4	17	23
5	10	16
6	15	21
7	10	14
8	14	20
9	19	24
10	10	17
11	11	16
12	13	18
13	16	23
14	10	15
15	12	16

Observación (mes)	Publicidad (en US$1000) (X)	Pasajeros (en US$1000) (Y)	XY	X^2	Y^2
1	10	15	150	100	225
2	12	17	204	144	289
3	8	13	104	64	169
4	17	23	391	289	529
5	10	16	160	100	256
6	15	21	315	225	441
7	10	14	140	100	196
8	14	20	280	196	400
9	19	24	456	361	576
10	10	17	170	100	289
11	11	16	176	121	256
12	13	18	234	169	324
13	16	23	368	256	529
14	10	15	150	100	225
15	12	16	192	144	256
	187	268	3490	2469	4960

Suma de los cuadrados de X

$$SC_x = \sum X^2 - \frac{(\sum X)^2}{n}$$

$$SC_x = \sum 2469 - \frac{(187)^2}{15} = 137.7333333$$

Suma de los cuadrados de Y

$$SC_y = \sum Y^2 - \frac{(\sum Y)^2}{n}$$

$$SC_y = \sum 4960 - \frac{(268)^2}{15} = 171.733333$$

Suma de los productos cruzados de X y Y

$$SC_{xy} = \sum XY - \frac{(\sum X)(\sum Y)}{n}$$

$$SC_{xy} = \sum 3490 - \frac{(187)(268)}{15} = 148.933333$$

Pendiente de la recta de regresión

$$b_1 = \frac{SC_{xy}}{SC_x}$$

$$b_1 = \frac{148.933333}{137.733333} = 1.0813166 \; o \; 1.08$$

Calculo de las medias

$$\bar{Y} = \frac{\sum Y}{n} = \frac{268}{15} = 17.86667$$

$$\bar{X} = \frac{\sum X}{n} = \frac{187}{15} = 12.46667$$

Intercepto de la recta de regresión

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_0 = 17.866667 - 1.08(12.46667) = 4.3865 \; o \; 4.39$$

Modelo de regresión estimada

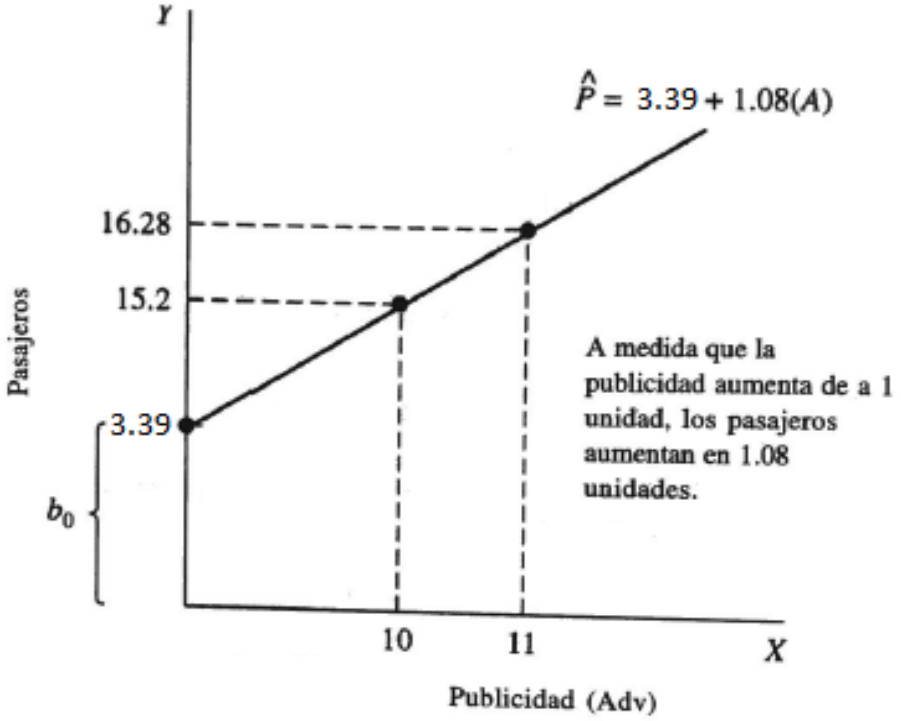
$$\hat{Y} = b_0 + b_1 X$$

$$\hat{Y} = 4.39 + 1.08 X_i$$

Si $X_i = 10$

$$\hat{Y} = 4.39 + 1.08\,(10) = 15.2$$

Debido a que tanto X como Y están expresados en miles, esto significa que si gastan US$10000 en publicidad, el modelo predice que 15200 personas decidirán viajar en Hop Scotch Airlines. El coeficiente de 1.08 significa que por cada incremento de una unidad en X, Y aumentará en 1.08 unidades.



Ejercicio 12.1. Un economista del departamento de recursos humanos de Florida State está preparando un estudio sobre el comportamiento del consumidor. Él recolectó los datos que aparecen en miles de dólares para determinar si existe una relación entre el ingreso del consumidor y los niveles de consumo.

a. Determine cual es la variable dependiente.

b. Haga un diagrama de dispersión para los datos.

c. Calcule e interprete el modelo de regresión. ¿Qué le dice este modelo sobre la relación entre el consumo y el ingreso? ¿Qué proporción de cada dólar adicional que se gana se invierte en consumo?

d. ¿Qué consumo pronosticaría el modelo para alguien que gana US$27500?

Consumidor	Ingreso(X)	Consumo (Y)
1	24.3	16.2
2	12.5	8.5
3	31.2	15
4	28	17
5	35.1	24.2
6	10.5	11.2
7	23.2	15
8	10	7.1
9	8.5	3.5
10	15.9	11.5
11	14.7	10.7
12	15	9.2

Ejercicio 12.2. Overland Group produce partes para camión que se utilizan en los semirremolques. El jefe de contabilidad desea desarrollar un modelo de regresión que pueda utilizarse para predecir los costos. Él selecciona unidades de producción fabricadas como una variable de predicción y recolecta los siguientes datos. Los costos están en miles de dólares y las unidades en cientos.

a. Determine cual es la variable dependiente.

b. Haga un diagrama de dispersión para los datos.

c. Calcule e interprete el modelo de regresión. ¿Qué le dice al contador sobre la relación entre producción y costos?

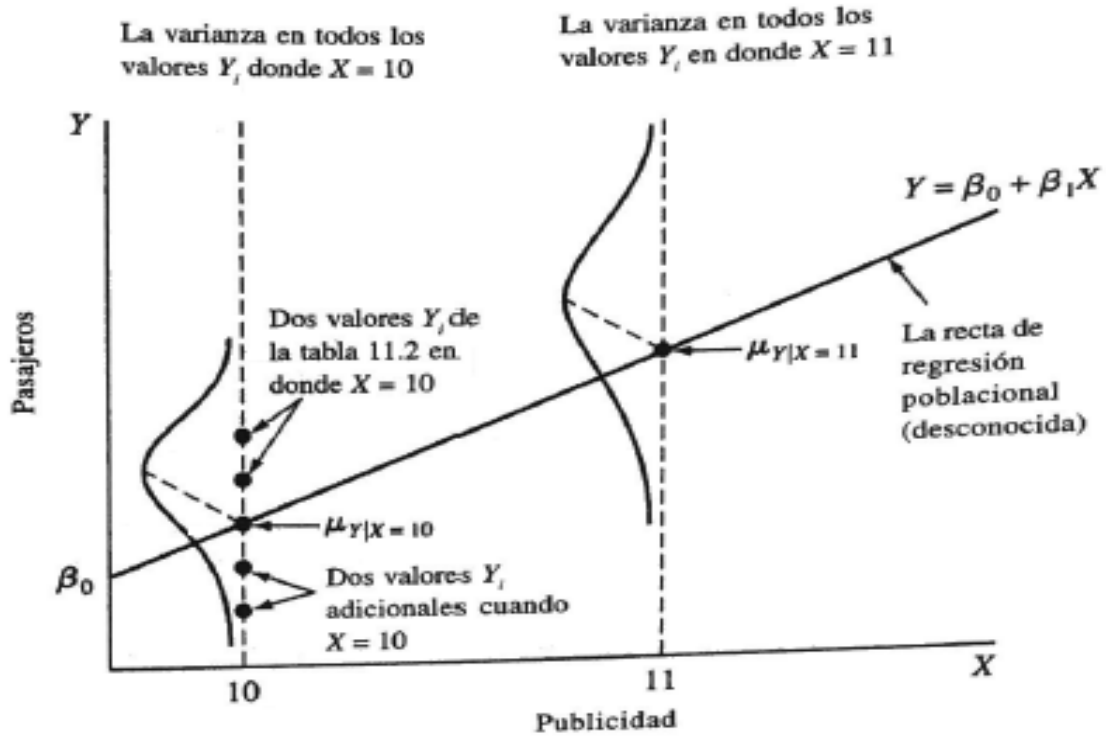
d. Según el modelo ¿Cuánto costaría producir 750 unidades?

Unidades (X)	Costo (Y)
12.3	6.2
8.3	5.3
6.5	4.1
4.8	4.4
14.6	5.2
14.6	4.8
14.6	5.9
6.5	4.2

Supuestos del modelo de regresión lineal

Supuesto 1: El término de error Ɛ es una variable aleatoria distribuida normalmente.

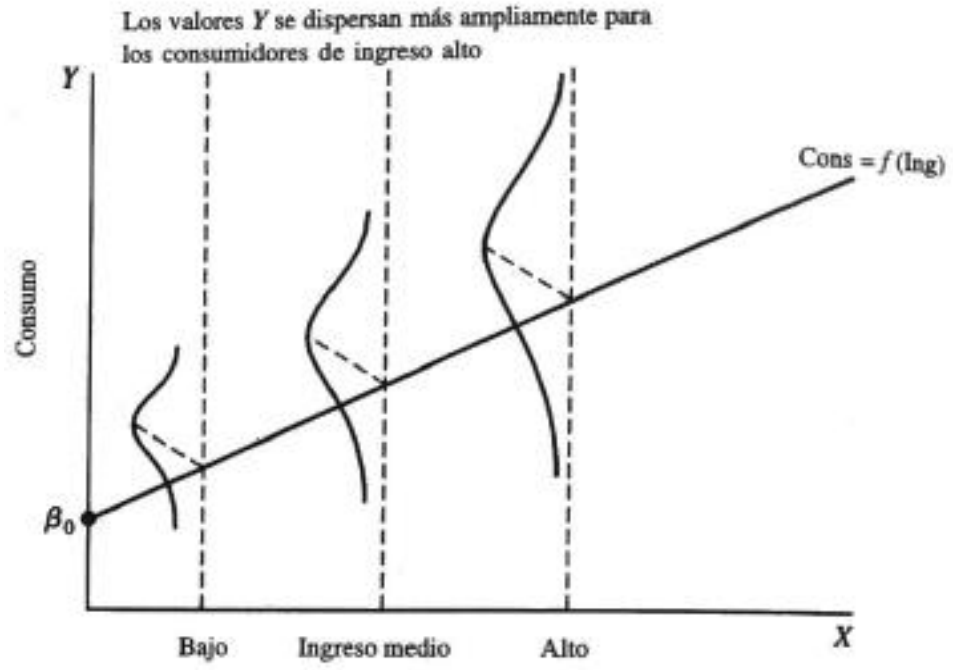
Si se fija que X es igual a un valor dado, muchas veces los valores resultantes de Y variarán. Debido a que Y_i es diferente cada vez, lo mejor que la recta de regresión puede hacer es estimar el valor promedio de Y.



Supuesto 2: Varianzas iguales de los valores Y.

Las varianzas de los valores Y son las mismas en todos los valores de X (homoscedasticidad).

Sin embargo, si se trabaja con datos de corte seccional esto no aplica. Si se recolectan datos sobre consumidores en diferentes intervalos de ingreso durante un año dado, se estarían utilizando datos de corte seccional ya que se incluyeron las observaciones a través de diferentes secciones de estrato de ingresos. Los valores de Y_i se dispersan más ampliamente a medida que el ingreso incrementa (heteroscedasticidad).



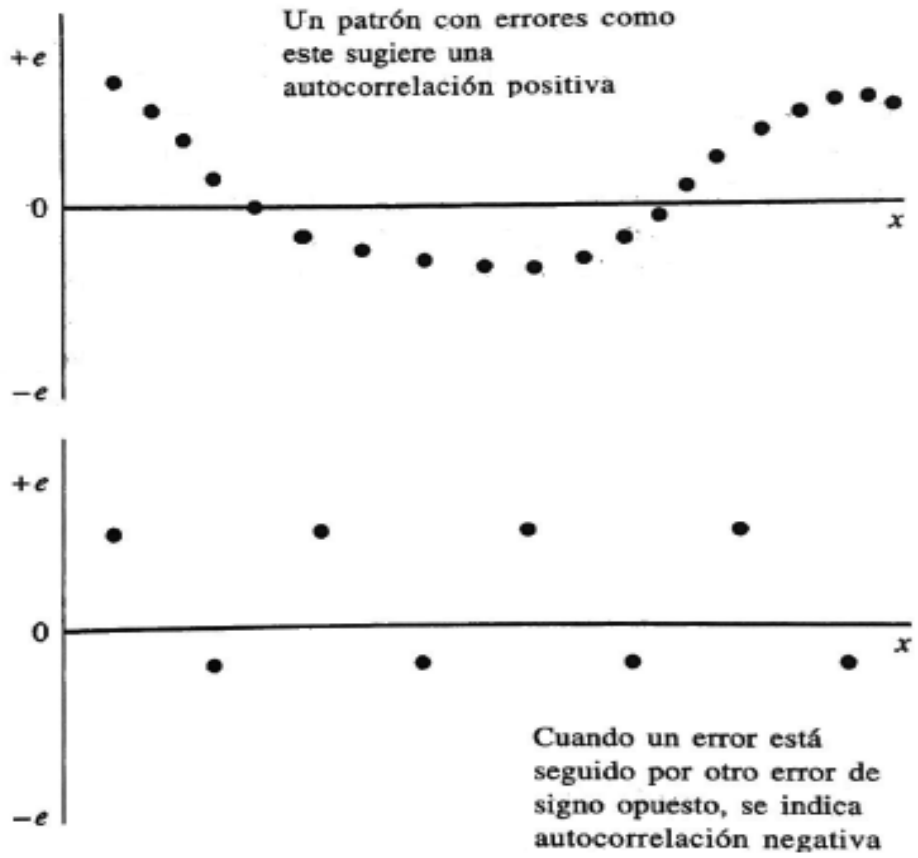
Universidad de Medellín
Ciencia y Libertad

Supuesto 3: Los términos de error son independientes uno del otro.

Autocorrelación: ocurre cuando los términos del error no son independientes.

Autocorrelación positiva: los signos iguales se agrupan.

Autocorrelación negativa: cada error es seguido de un error de signo opuesto.



Universidad de Medellín
Ciencia y Libertad

Existe una forma para detectar la autocorrelación con base en la prueba de Durbín-Watson. El estadístico que se calcula es:

$$d = \frac{\sum(e_t - e_{t-1})^2}{\sum e_t^2}$$

Donde:

e_t es el error en el periodo de tiempo t

e_{t-1} es el error en el periodo anterior.

Este valor se utiliza para probar la hipótesis de que no existe correlación entre términos de error sucesivos así:

$$Ho: \rho_{e_t, e_{t-1}} = 0 \ (no \ existe \ \ autocorrelaci \ \ ón)$$

$$HA: \rho_{e_t, e_{t-1}} \neq 0 \ (existe \ \ autocorrelaci \ \ ón)$$

El valor Durbín-Watson calculado se compara con los valores críticos tomados de la tabla para un valor de significancia determinado.

Supuesto 4: El supuesto de linealidad.

Como se expresó en el supuesto 1, si X se deja igual que un valor muchas veces, ocurrirá una distribución normal de los valores de Y. Esta distribución tiene una media $\mu_{x|y.}$ Esto es cierto para todo valor de X. MCO asume que estas medias quedan en una recta.

El error estándar de estimación: Una medida de bondad de ajuste

La recta de regresión representa la relación entre X y Y mejor que cualquier otra recta. Sin embargo, debido a que simplemente proporciona el mejor ajuste, no existe garantía de que sea buena. Para medir que tan bueno es el mejor ajuste, hay por lo menos dos medidas de bondad de ajuste:

1. El error estándar de estimación
2. El coeficiente de determinación.

Error estándar de estimación (Se)

Es una medida del grado de dispersión de los valores Y_i alrededor de la recta de regresión. El error estándar de estimación mide esta variación promedio de los puntos de datos alrededor de la recta de regresión que se utiliza para estimar Y y por ende proporciona una medida del error que se presentará en dicha estimación.

$$Se = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

Suma de cuadrados del error

$$SCE = SC_y - \frac{(SC_{xy})^2}{SC_x}$$

En un modelo de regresión simple, se imponen dos restricciones en el conjunto de datos, debido a que se deben estimar dos parámetros, $\beta_0 \; y \; \beta_1$. Por tanto hay n – 2 grados de libertad y CME es:

$$CME = \frac{SCE}{n-2}$$

El error estándar es:

$$Se = \sqrt{CME}$$

Para el caso de Hop Scotch Airlines

$$SCE = SC_y - \frac{\left(SC_{xy}\right)^2}{SC_x} = 171.7333 - \frac{(148.9333)^2}{137.73333} = 10.6893$$

$$CME = \frac{SCE}{n-2} = \frac{10.6893}{15-2} = 0.82226$$

$$Se = \sqrt{CME} = \sqrt{0.82226} = 0.90678 \; o \; 0.907$$

El error estándar siempre se expresa en las mismas unidades que la variable dependiente Y, en este caso miles de pasajeros.

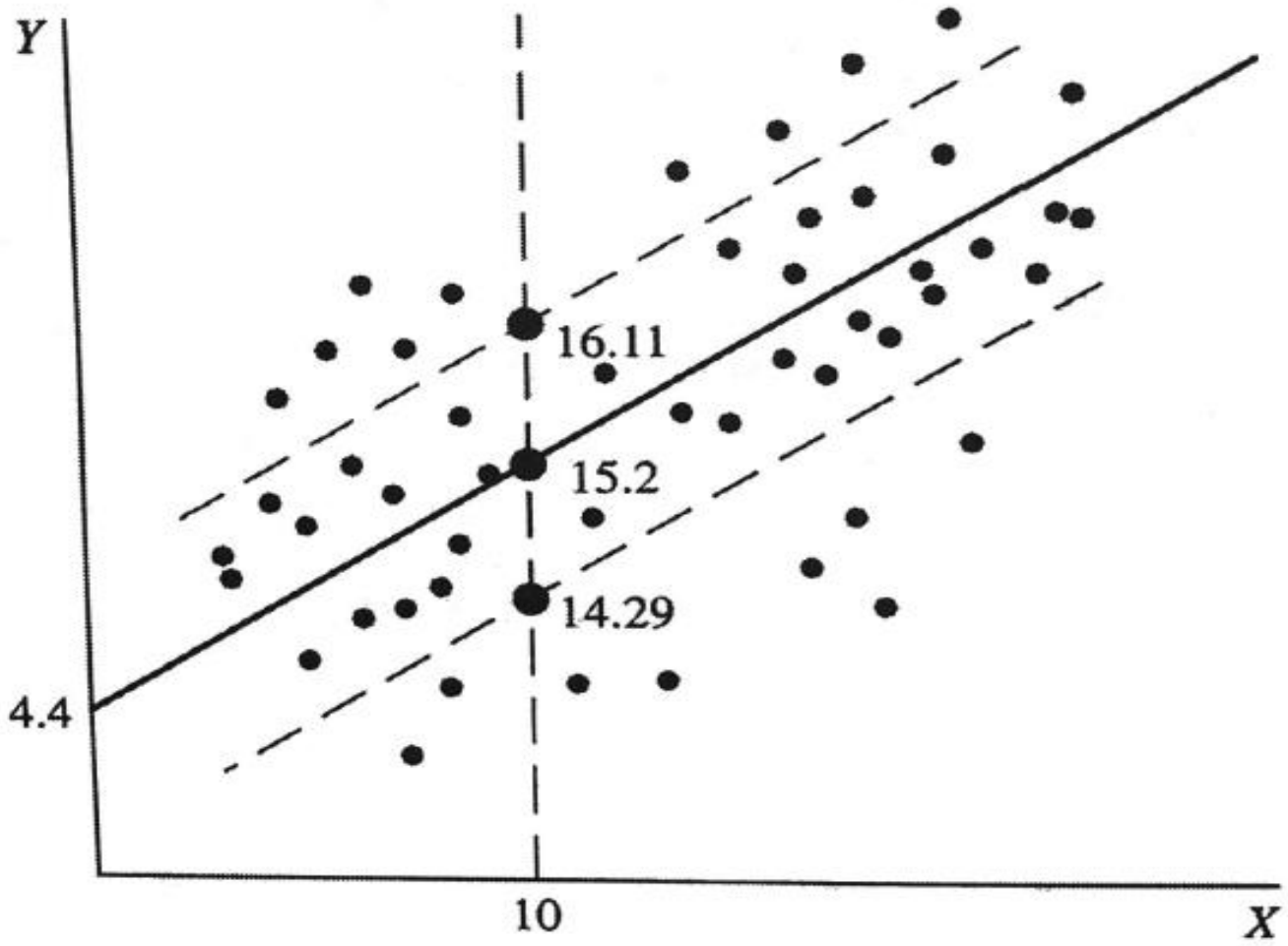
El error estándar de estimación es una medida de la dispersión de los valores Y alrededor de su media, dado un valor X específico. Al ser similar a la desviación estándar para una sola variable, puede interpretarse de manera similar (regla empírica).

Si $X_i = 10$

$$\hat{Y} = 4.39 + 1.08\,(10) = 15.2$$

Para ilustrar el significado del error estándar de estimación, se localizan los puntos que están a un *Se* (0.907) por encima y por debajo del valor promedio de 15.2. Estos puntos son 14.29 (15.2 – 0.907) y 16.11 (15.2 + 0.907).

En este caso, el 68.3% de las veces cuando se invierte US$10000 en publicidad, el número de pasajeros estará entre 14290 y 16110. El 31.7% del tiempo restante, en número de pasajeros excederá de 16110 o será menor que 14290 (regla empírica).



Ejercicio 12.3. ¿Cuál es el error estándar de estimación para el departamento de recursos humanos de Florida State?

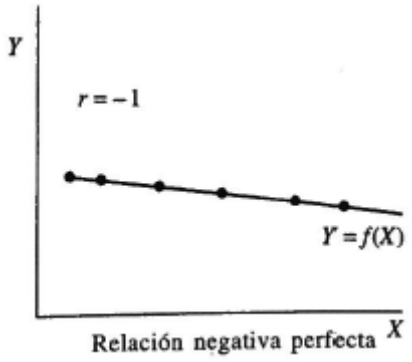
Ejercicio 12.4. Overland Group ahora desea conocer el error estándar de estimación.

Análisis de correlación

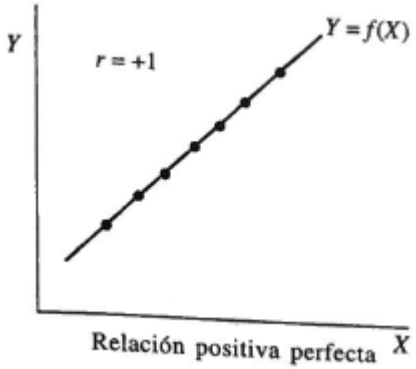
El modelo de regresión ha proporcionado un panorama claro de la relación entre variables. El valor positivo para $b1$ indica una relación directa. Ahora es útil obtener una medida de la fuerza de esa relación. Esta es la función del **coeficiente de correlación (r)**, este puede tomar valores entre -1 y 1.

$$-1 \leq r \leq 1$$

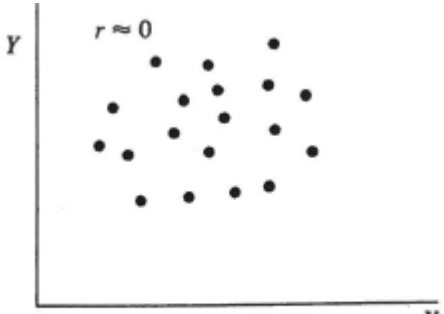
Entre mayor sea el valor absoluto de *r*, más fuerte será la relación entre X y Y.



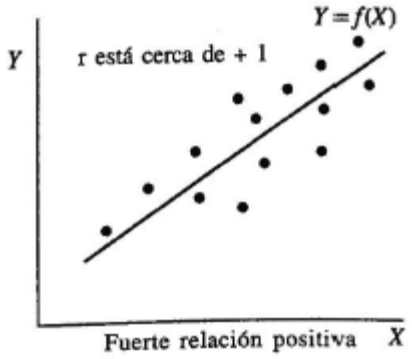
Relación negativa perfecta



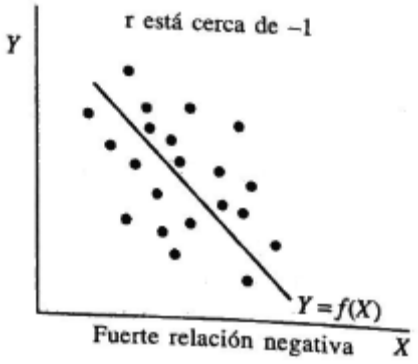
Relación positiva perfecta



Poca o ninguna relación; no se puede dibujar una recta que ajuste



Fuerte relación positiva



Fuerte relación negativa

Universidad de Medellín
Ciencia y Libertad

Para comprender plenamente lo que mide el coeficiente de correlación, se deben desarrollar tres medidas de desviación.

La **desviación total** de Y es la cantidad por la cual los valores individuales de Y, (Yi) varían de su media $\bar{Y}$ $(Yi - \bar{Y})$.

Suma de cuadrados total

$$SCT = \sum (Y_i - \bar{Y})^2$$

La desviación total puede dividirse en:

La **desviación explicada** es la diferencia entre lo que predice el modelo de regresión $\hat{Y}i$ y el valor promedio de Y, $(\hat{Y}i - \bar{Y})$.

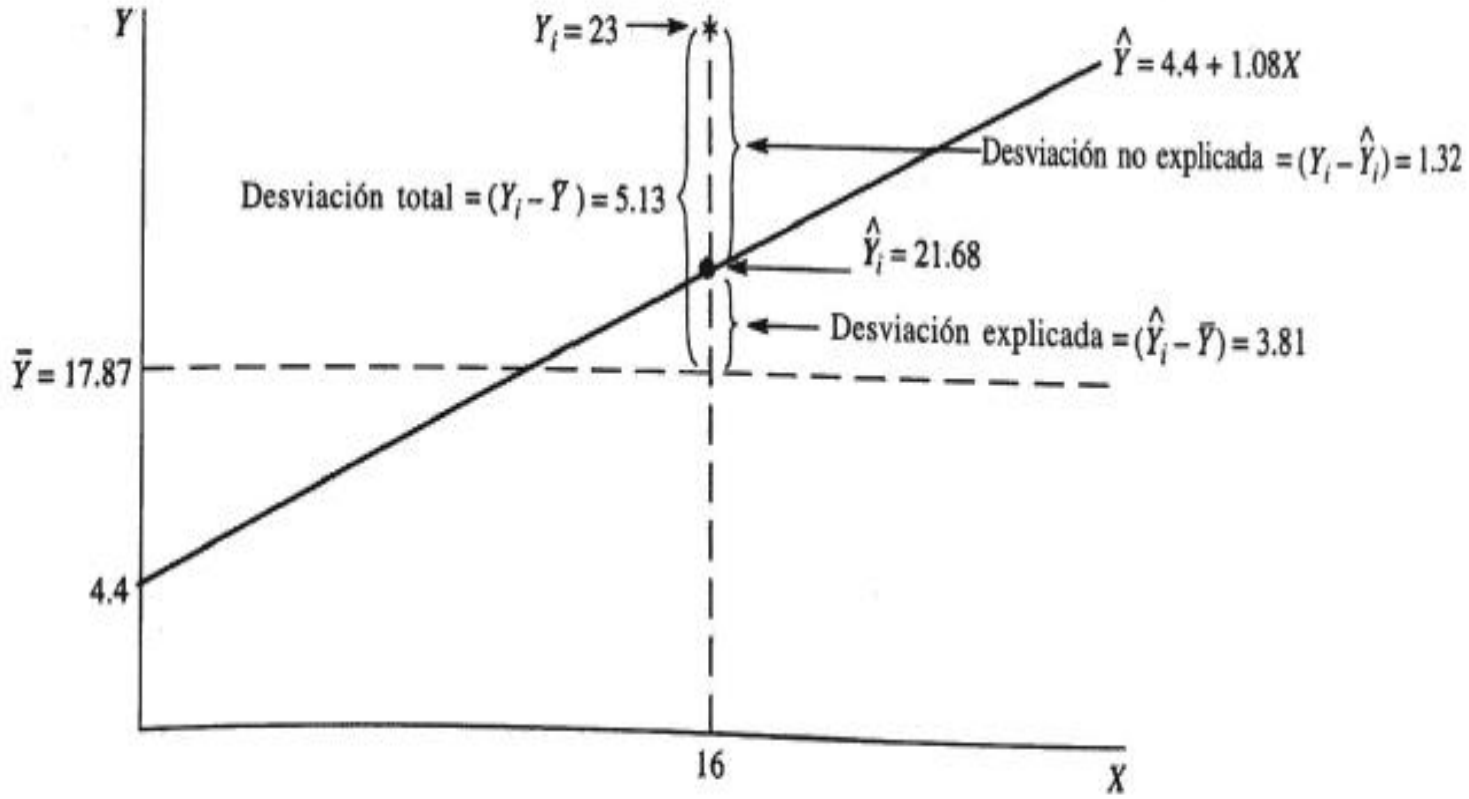
Suma de cuadrados de la regresión

$$SCR = \sum (\hat{Y}_i - \bar{Y})^2$$

La **desviación no explicada** es esa porción de la desviación total que no es explicada por el modelo de regresión. Es decir, es el error $(Yi - \hat{Y}_i)$.

Suma del cuadrado del error

$$SCE = \sum (Yi - \hat{Y}_i)^2$$



El **coeficiente de correlación** se calcula así:

$$r = \sqrt{\frac{variación\ explicada}{variación\ total}} = \sqrt{\frac{SCR}{SCT}}$$

$$r = \frac{SC_{xy}}{\sqrt{(Sc_x)(SC_y)}}$$

Para el caso de Hop Scotch se tiene:

$$r = \frac{148.93333}{\sqrt{(137.7333)(171.7333)}} = 0.9683$$

Esto indica una relación positiva entre los pasajeros y la cantidad de dinero invertido en publicidad.

Vale la pena recordar que el error estándar de estimación S_e, que se calculó anteriormente, es una medida de la bondad de ajuste. Proporciona una medida cuantificable de qué tan bien se ajusta el modelo a los datos que se están recolectando.

El **coeficiente de determinación r^2** es otra medida quizá más importante de la bondad de ajuste.

$$r^2 = \frac{desviación\ explicada}{desviación\ total} = \frac{SCR}{SCT}$$

$$r^2 = \frac{(SC_{xy})^2}{(Sc_x)(SC_y)}$$

$$r^2 = (r)^2$$

Proporciona una medida de bondad de ajuste porque revela qué porcentaje del cambio en Y se explica por un cambio en X.

El coeficiente de determinación para Hop Scotch es:

$$r^2 = \frac{(148.9333)^2}{(137.7333)(171.7333)} = 0.93776 \; o \; 0.94$$

$$r^2 = (0.9683)^2 = 0.94$$

Esto establece que el 94% del cambio en el número de pasajeros se explica mediante un cambio en la publicidad.

No se interpreta este valor como si el 94% del cambio en los pasajeros fuera causado por un cambio en la publicidad. La correlación no significa causa.

Este r^2 tiene significado solo para las relaciones lineales. Dos variables pueden tener un r^2 de cero y sin embargo estar relacionadas en sentido curvilíneo.

Ejercicio 12.5. Calcule el coeficiente de correlación y el coeficiente de determinación para el departamento de recursos humanos de Florida State.

Ejercicio 12.6. Calcule el coeficiente de correlación y el coeficiente de determinación para Overland Group.

Limitaciones del análisis de regresión

- No pueden determinar causa – efecto. La correlación no implica causalidad.

  Tasas de natalidad – número de cigüeñas. Correlación r = 0.92.

  Factor real: densidad poblacional.

- Se debe tener cuidado de no utilizar el modelo de regresión para predecir Y con valores de X que estén fuera del rango del conjunto original de datos.
- Otra falla del análisis de correlación y regresión se hace evidente cuando dos variables no relacionadas parecen presentar alguna relación.

  Número de elefantes en África – número de truchas en América.
  Correlación r = 0.91.
  No hay sustituto para el sentido común en el análisis de correlación ni en el de regresión.

Pruebas para los parámetros poblacionales

Los resultados estadísticos pueden sugerir una relación entre las variables. Sin embargo, estos resultados están basados en una muestra. Como siempre se pregunta, ¿si existe alguna relación a nivel poblacional? Podría ser que debido a un error de muestreo el estadístico calculado no es cero pero los parámetros poblacionales si lo son.

$H_o: \beta_1 = 0$

$H_A: \beta_1 \neq 0$

Pruebas para β_1

Si la pendiente de la recta de regresión poblacional real pero desconocida es cero, no existe relación entre las variables (pasajeros y publicidad) contraria a los resultados muestrales. Para esto, se debe probar la hipótesis:

$$H_o: \beta_1 = 0$$
$$H_A: \beta_1 \neq 0$$

Esta prueba emplea el estadístico t y tiene n - 2 grados de libertad.

Prueba t para el coeficiente de regresión poblacional

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

Donde:

S_{b_1} es el error estándar de la distribución muestral de b_1
b_1 es la pendiente de la recta de regresión
β_1 es el coeficiente de regresión para toda la población

Muestras diferentes dan valores diferentes para b_1, por lo tanto, si β_1 es realmente cero, estos valores para b_1 se distribuirían alrededor de cero.

Error estándar del coeficiente de regresión

$$S_{b_1} = \frac{S_e}{\sqrt{SC_x}}$$

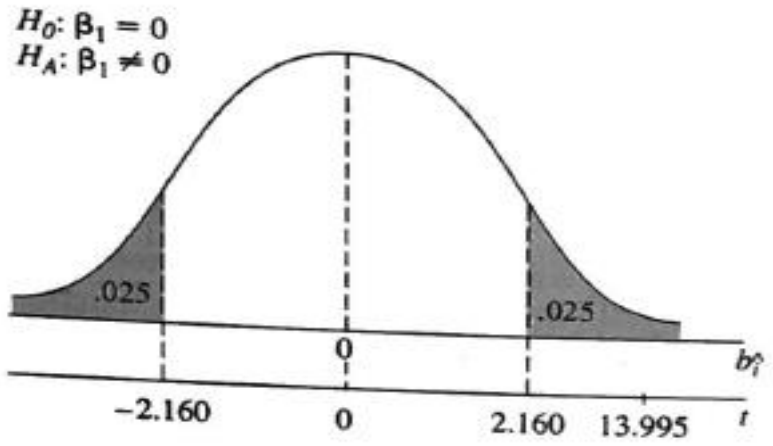
Para el caso de Hop Scotch tenemos:

$$S_{b_1} = \frac{S_e}{\sqrt{SC_x}} = \frac{0.907}{\sqrt{137.73333}} = 0.07726$$

$$t = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{1.0813 - 0}{0.07726} = 13.995$$

Si se selecciona un valor α del 5%, el valor crítico de *t* es $t_{0.05,13}$ = 2.160.

Regla de decisión: "No rechazar si *t* está entre ± 2.160, de lo contrario rechazar"



Debido a que *t* = 13.995, la hipótesis nula de que $\beta_1 = 0$ se rechaza.
Al nivel del 5% parece existir una relación entre pasajeros y publicidad.

Universidad de Medellín
Ciencia y Libertad

Si la hipótesis nula no hubiera sido rechazada, se concluiría que la publicidad y los pasajeros no están relacionados. Se debería descartar el modelo y utilizar una variable explicativa diferente.

Debido a que se ha rechazado la hipótesis nula de que $\beta_1 = 0$, la pregunta natural es, ¿cuál es su valor?, esta pregunta se responde calculando un intervalo de confianza para β_1.

$$\text{I. C para } \beta_1 = b_1 \pm t(S_{b_1})$$

Para el caso de Hop Scotch, si se utiliza un nivel de confianza del 95% tenemos:

$$\text{I. C para } \beta_1 = 1.08 \pm 2.160(0.07726)$$

$$0.913 \leq \beta_1 \leq 1.247$$

Esto significa que se puede estar un 95% seguro de que el coeficiente de regresión para toda la población de todos los valores de X, está entre 0.913 y 1.247.

Pruebas para el coeficiente de correlación poblacional (ρ)

Puede ser que la correlación a nivel poblacional sea cero y que una muestra engañosa hizo que se asumiera equivocadamente una relación. Por consiguiente se debe probar la hipótesis

$$H_o: \rho = 0$$
$$H_A: \rho \neq 0$$

De nuevo se utiliza la prueba t.

Prueba *t* para el coeficiente de correlación poblacional

$$t = \frac{r - \rho}{S_r}$$

Donde:

S_r es el error estándar del coeficiente de correlación

r es el coeficiente de correlación

Error estándar del coeficiente de correlación

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Para el caso de Hop Scotch tenemos:

$$S_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0.93776}{15 - 2}} = 0.069$$

$$t = \frac{r - \rho}{S_r} = \frac{0.9683 - 0}{0.069} = 13.995$$

Si se selecciona un valor α del 5%, el valor crítico de *t* es $t_{0.05,13}$ = 2.160.

Regla de decisión: "No rechazar si *t* está entre ± 2.160, de lo contrario rechazar"

Debido a que *t* = 13.995 > 2.160, se rechaza la hipótesis nula. A un nivel de significancia del 5%, se concluye que el coeficiente de correlación poblacional no es cero y que los pasajeros y la publicidad están relacionados. Al igual que con la prueba para β_1, si la hipótesis nula no se rechaza se concluye que la publicidad no tiene poder explicativo y un nuevo modelo tendrá que generarse.

El hecho de que el valor *t* = 13.995 sea el mismo tanto para β_1 como para ρ no es coincidencia. Siempre se obtendrán los resultados idénticos de estas dos pruebas de hipótesis en un modelo de regresión simple.

Ejercicio 12.7. ¿ La relación entre los ingresos y el consumo que está analizando el departamento de recursos humanos de Florida State es significativa? Pruebe las hipótesis a un nivel de significancia del 1%.

Ejercicio 12.8. ¿ La relación entre los ingresos y el consumo que está analizando Overland Group es significativa? Pruebe las hipótesis a un nivel de significancia del 1%.

Intervalos de confianza para el análisis de regresión

El análisis de regresión puede pronosticar y predecir valores para la variable dependiente. Una vez que se ha determinado la ecuación de regresión, se puede desarrollar un estimador puntual para la variable dependiente sustituyendo un valor dado para X en la ecuación y despejando Y.

Sin embargo, el investigador puede estar interesado en los estimados por intervalo. Existen por lo menos dos estimados por intervalo que se relacionan comúnmente con los procedimientos de regresión.

La media de Y condicionada a un valor de X

Se supone que se desea desarrollar un estimado por intervalo para la media condicionada de Y, $\mu_{y|x}$. Esta es la media poblacional para todos los valores de Y, con la condición de que X sea igual a un valor específico.

Si se deja X igual la misma cantidad de veces, se obtendrán muchos valores diferentes de Y.

Para calcular este intervalo para el valor promedio condicional de Y, se debe primero determinar S_y, el **error estándar de la media condicionada**.

El propósito de S_y es tener en cuenta los diferentes valores de b_0 y b_1 que resultan del error de muestreo.

$$S_y = S_e \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SC_x}}$$

Donde:

S_e es el error estándar de estimación
$\bar{X}$ es la media de los valores de X
SC_x es la suma de los cuadrados de X
X_i es el valor dado para la variable independiente

Intervalo de confianza para la media condicionada

$$\text{I. C para } \mu_{y|x} = \hat{Y}_i \pm t(S_y)$$

Donde:

$\hat{Y}_i$ es el estimador puntual hallado de la ecuación de regresión.

El valor *t* se basa en *n* – 2 grados de libertas porque se deben calcular los dos valores de b_0 y b_1 de los datos muestrales.

Si Hop Scotch desea desarrollar el intervalo para la media condicionada en donde $X_i = 10$ tenemos:

$$\bar{X} = 12.47$$

$$S_y = S_e \sqrt{\frac{1}{n} + \frac{(X_i - \bar{X})^2}{SC_x}} = 0.907 \sqrt{\frac{1}{15} + \frac{(10 - 12.47)^2}{137.73333}} = 0.303$$

Si se utiliza un nivel de confianza del 95%, $t_{0.05,13}$ = 2.160.

$$\hat{Y} = 4.39 + 1.08\,(10) = 15.2$$

$$\text{I. C para } \mu_{y|x} = \hat{Y}_i \pm t(S_y)$$

$$\text{I. C para } \mu_{y|x} = 15.2 \pm 2.160(0.303)$$

$$14.55 \leq \mu_{y|x} \leq 15.85$$

Hop Scotch puede estar 95% seguro de que si se invierten US$10000 en publicidad muchas veces, existe un 95% de probabilidad de que la media de todos los valores resultantes para los pasajeros estará entre 14.55 y 15.85.

Intervalo de predicción para un valor único de Y

Ahora el objeto es predecir un valor único de Y si X se fija en una cantidad dada una sola vez.

Para calcular este intervalo de predicción, primero se debe calcular el **error estándar del pronóstico**, S_{yi}.

$$S_{yi} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SC_x}}$$

El intervalo de confianza para un valor único de Y, Y_x entonces será:

$$\text{I. C para } Y_x = \hat{Y}_i \pm t(S_{y_i})$$

Ahora se construye un intervalo de confianza para un valor único de Y cuando X = 10.

$$S_{yi} = S_e \sqrt{1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{SC_x}} = 0.907 \sqrt{1 + \frac{1}{15} + \frac{(10 - 12.47)^2}{137.73333}} = 0.956$$

$$\text{I. C para } Y_x = \hat{Y}_i \pm t(S_{y_i})$$

$$\text{I. C para } Y_x = 15.2 \pm 2.160(0.956)$$

$$13.14 \leq Y_x \leq 17.27$$

Hop Scotch puede estar un 95% seguro de que si en un mes cualquiera X_i = US$10000, el valor único resultante de Y estará entre 13.14 y 17.27.

Factores que influyen en el ancho del intervalo

Dado un nivel de confianza, es preferible minimizar la amplitud del intervalo. Entre más pequeño sea el intervalo, más precisa será la predicción de $\mu_{y|x}$ o de Y_x. Sin embargo, existen razones que trabajan en contra de producir un intervalo más estrecho.

La primera es el grado de dispersión de los datos originales. Entre más dispersos estén mayor será en S_e, un S_e mayor resulta en un intervalo más amplio.

El tamaño de la muestra es un segundo factor en la determinación de la amplitud del intervalo. Un tamaño muestral grande termina en un error estándar más pequeño, un error estándar pequeño resulta en un intervalo pequeño.

Un valor de X relativamente cercano a $\bar{X}$ producirá un intervalo pequeño porque la regresión se basa en los promedios. Por tanto, un tercer factor que influye en la amplitud del intervalo es la distancia a la cual está un valor particular de X de $\bar{X}$.