Lección 1: Análisis de regresión simple

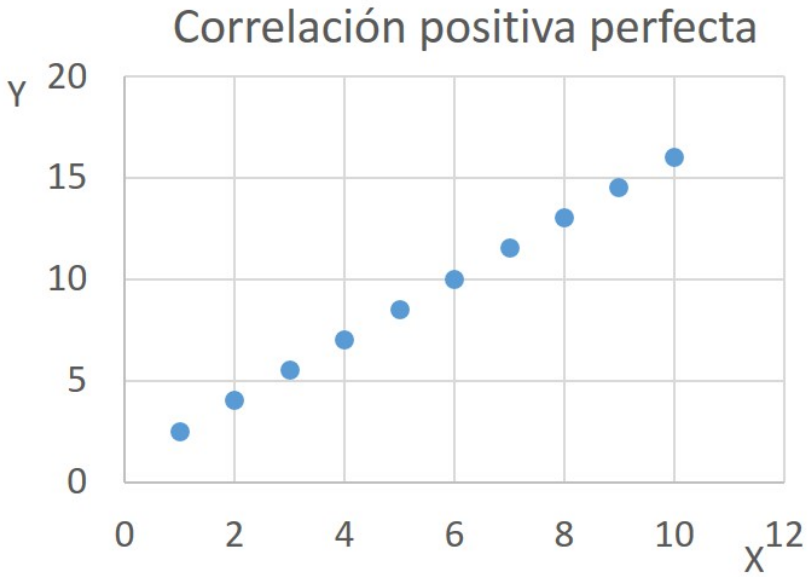
Módulo 4: Introducción al análisis de regresión

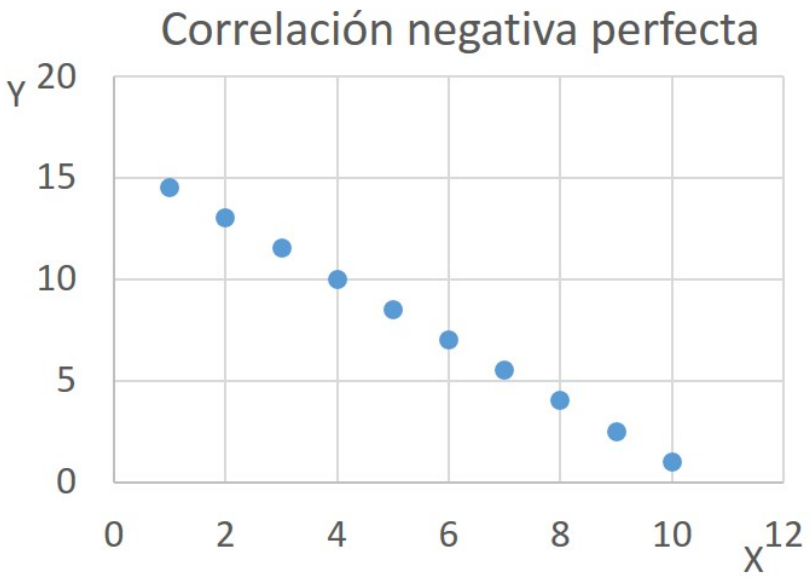
Magdalena Cornejo

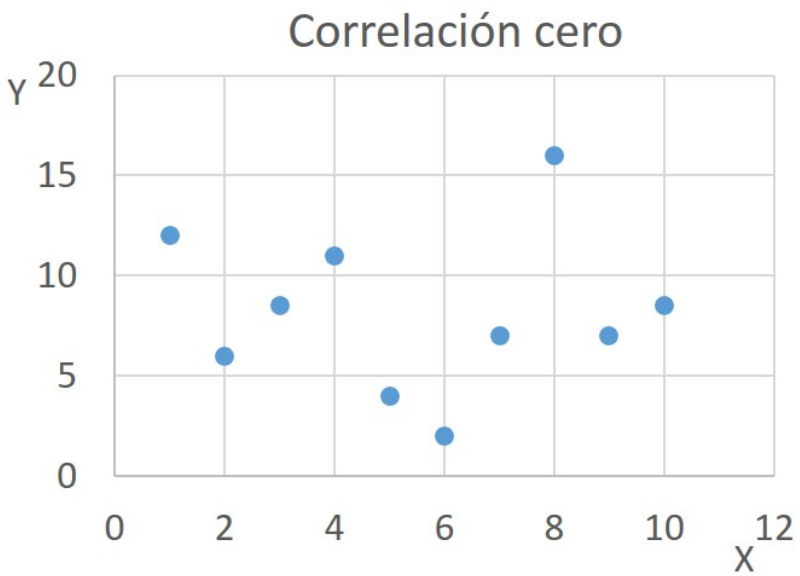
Análisis de Correlación

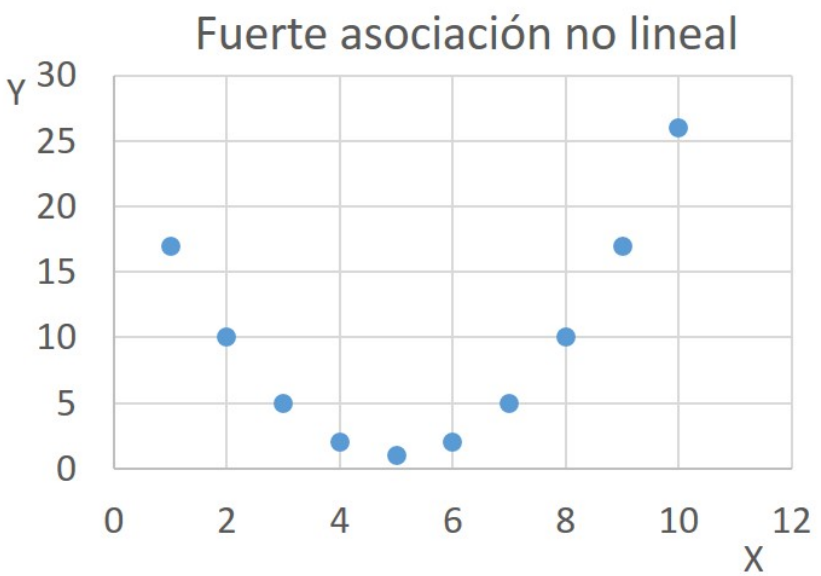
- Un análisis de correlación expresa la relación entre dos variables utilizando un único número.
- El coeficiente de correlación es una medida de asociación lineal entre dos variables.
- El coeficiente de correlación nos da una pauta de cuál alineada está la nube de puntos (gráfico de dispersión) en torno a una recta.
- Pero **no** no dice cuál es esa relación (función), ni puede capturar relaciones no lineales.

Análisis de Correlación









Fitting line

- Utilice la base de datos en Excel: `ventas.xls` que contiene datos de ventas e inversión en publicidad de una compañía entre 1993 y 2007.
- Compruebe que el coeficiente de correlación da 0.904.
- Construya un gráfico de dispersión entre las ventas y los gastos de publicidad.
- Prediga cuánto espera vender el próximo año si va a gastar 10 mil en publicidad.

Modelo lineal

Asume:

$$y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

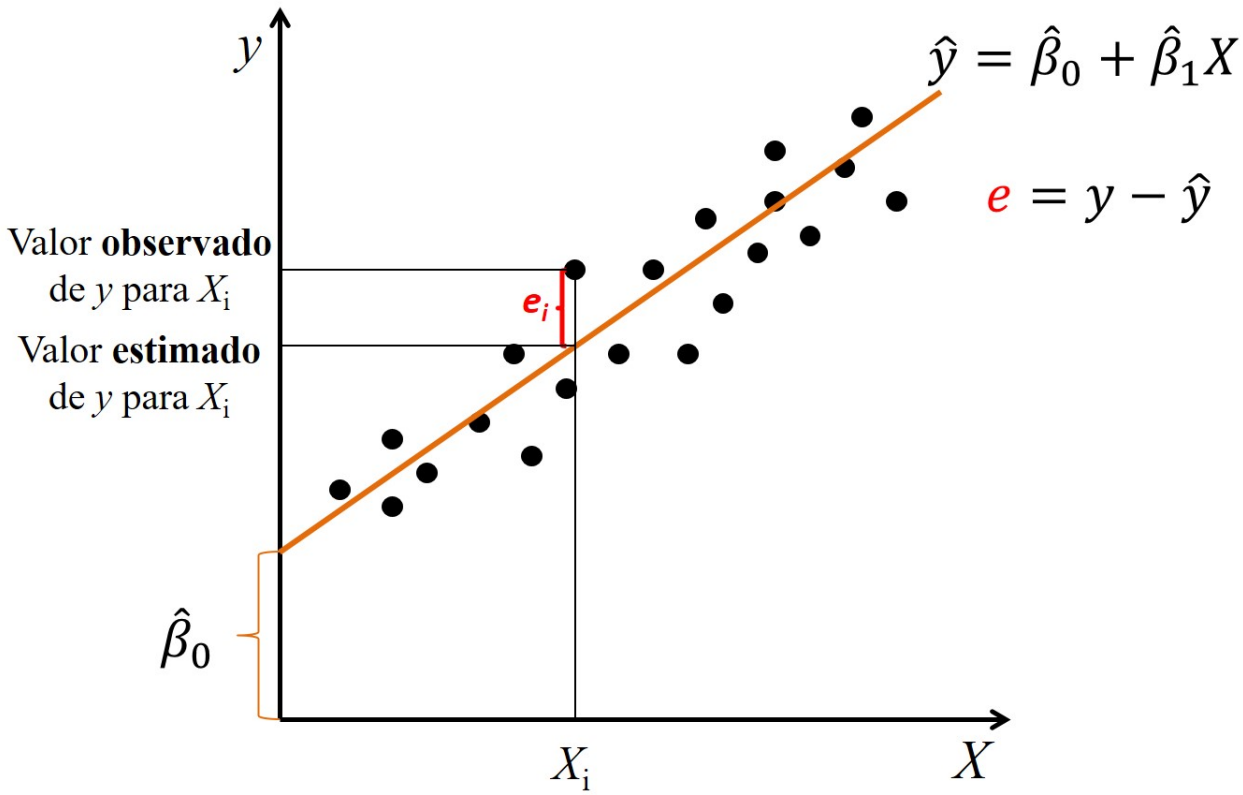
- y: variable dependiente o explicada
- X: variable explicativa o regresor
- β_0 y β_1: parámetros desconocidos
- ε: error aleatorio

El problema entonces es estimar los parámetros (β_0 y β_1) basados en datos de una muestra. Veremos el método de estimación de **mínimos cuadrados ordinarios**.

Mínimos Cuadrados Ordinarios (MCO)

Busca aproximar esta relación con nuestras observaciones de forma de:

- minimizar la suma de los residuos al cuadrado ($\sum_i e_i^2$), o
- que la nube de puntos esté lo más cerca posible de la recta estimada



Los estimadores MCO

A través de dicha minimización se obtiene la fórmula del estimador de la constante y la pendiente:

$$\widehat{\beta_1} = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sum(X - \overline{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}.\overline{X}$$

Interpretación de las estimaciones

¿Qué nos dicen los valores de los coeficiente que obtuvimos?

- $\widehat{\beta_0}$ es el **efecto autónomo** (cuando la variable explicativa es cero). Es decir, el valor medio de la variable dependiente.
- $\widehat{\beta_1}$ es el **efecto parcial**, esto es, indica cómo cambia la variable explicada cuando la explicativa cambia en una unidad.

En nuestro ejemplo:

$$\hat{y}_i = 79,324 + 102,59x$$

Bondad del Ajuste

Una vez estimado el modelo nos interesa saber:

¿Qué porcentaje de la variabilidad en y está siendo explicada por X? ¿Son las observaciones cercanas a la recta estimada?

Bondad del Ajuste

Una vez estimado el modelo nos interesa saber:

¿Qué porcentaje de la variabilidad en y está siendo explicada por X? ¿Son las observaciones cercanas a la recta estimada?

Para responder a estas preguntas podemos calcular el R^2 o **coeficiente de determinación**.

El R^2

- Es la medida más utilizada, pero no la única y se calcula como la fracción de la varianza muestral de y que está siendo explicada por X.

$$R^2 = \frac{\sum(\widehat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum(y_i - \overline{y})^2}$$

- El R^2 indica qué proporción de la variabilidad total está siendo explicada por el modelo.

- Está acotada entre 0 y 1. En nuestro ejemplo: $R^2 = 0.82$.

- En el caso de la regresión simple, $R^2 = \rho_{XY}^2$.

Aplicación en Excel

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.90
Coeficiente de determinación R^2	0.82
R^2 ajustado	0.80
Error típico	0.94
Observaciones	15

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	52.03	52.03	58.44	0.00
Residuos	13	11.57	0.89		
Total	14	63.6			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	0.35	0.70	0.50	0.63	-1.17	1.87
Ventas	0.01	0.00	7.64	0.00	0.01	0.01

Inferencia estadística

Para hacer inferencia estadística en el contexto de regresión necesitamos conocer la distribución de los estimadores de MCO ($\hat{\beta}_0$ y $\hat{\beta}_1$).

Para ello, necesitamos hacer los siguientes **supuestos**:

- **Supuesto 1**: la distribución condicional de ε_i dado X_i tiene media cero.
- **Supuesto 2:** $(X_i, Y_i), i = 1, ..., n$ son IID a través de las observaciones.
- **Supuesto 3:** los *outliers* son poco probables.

Supuestos clásicos

Bajo estos supuestos y mediante el TCL, se puede probar que:

$$\widehat{\beta}_0 \sim N\left(\beta_0, \sigma_\varepsilon^2 \frac{\sum X_i^2}{n\sum(X_i - \overline{X})^2}\right)$$
$$\widehat{\beta}_1 \sim N\left(\beta_1, \sigma_\varepsilon^2 \frac{1}{\sum(X_i - \overline{X})^2}\right)$$

Noten que en realidad las varianzas del estimador MCO son desconocidas por no se conoce σ_ε^2.

Pero como vimos antes, se puede estimar con:

$$S_\varepsilon^2 = \frac{\sum \varepsilon^2}{n-2}$$

Inferencia estadística

- **Test de significatividad individual**:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Estadístico:

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

- **Intervalos de confianza**: $\hat{\beta}_1 \pm t_c.SE(\hat{\beta}_1)$

Aplicación en Excel

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.90
Coeficiente de determinación R^2	0.82
R^2 ajustado	0.80
Error típico	0.94
Observaciones	15

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	52.03	52.03	58.44	0.00
Residuos	13	11.57	0.89		
Total	14	63.6			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	0.35	0.70	0.50	0.63	-1.17	1.87
Ventas	0.01	0.00	7.64	0.00	0.01	0.01