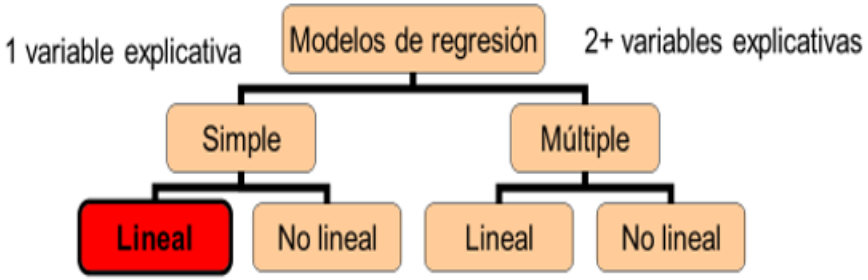
ANALISIS DE REGRESION

INTRODUCCION

El análisis de regresión es un método estadístico que permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés.

Modelos de análisis de regresión



Regresión

Regresión simple Y Múltiple

- Modelo.
- Estimación.
- Diagnosis.

OBJETIVOS

- Saber analizar las relaciones entre variables a través de un modelo de regresión lineal que describa cómo influye una variable X sobre otra variable Y.

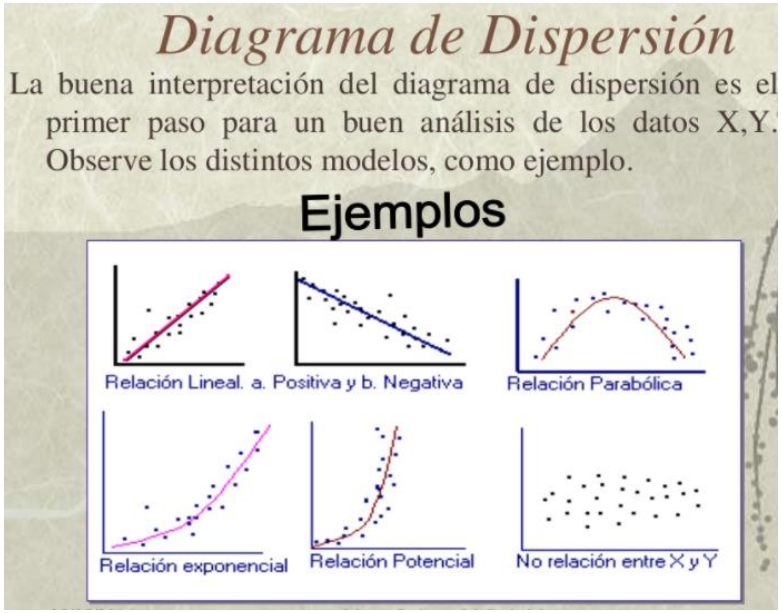
- Saber explicar la construcción de modelos usando el análisis de regresión.

Relaciones entre variables

La regresión estudia relaciones entre variables.
- Relaciones deterministas (exactas).
- Relaciones no deterministas (no exactas).

El análisis de regresión

Es una técnica para investigar y modelar la relación entre variables.
Las aplicaciones de regresión son numerosas y ocurren en casi todos los campos, incluyendo ingeniería, la física, ciencias económicas, ciencias biológicas y de la salud, como también ciencias sociales



UTILIDAD

Utilizados para varios propósitos, incluyendo los siguientes:

1. Descripción de datos
2. Estimación de parámetros.
3. Para estimación y predicción.

.

¿Cómo denominamos a las variables?

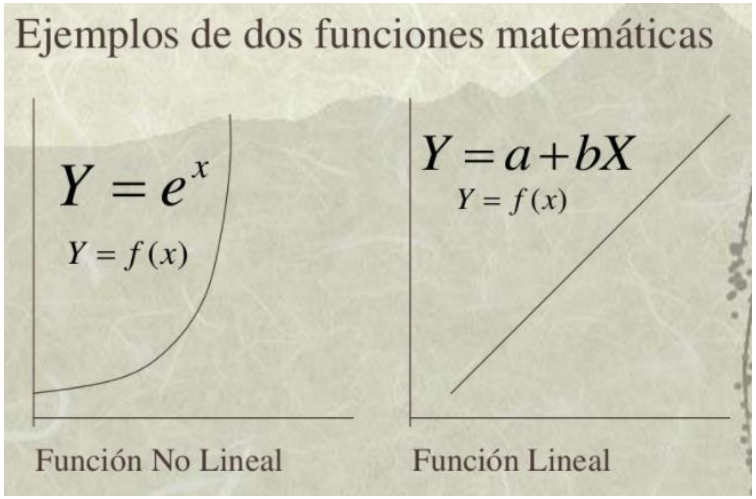
V. Independiente	V.Dependiente
Explicativa	A explicar
Es el valor que Conocemos	Es lo que queremos predecir

REGRESIÓN LINEAL SIMPLE

La más simple relación entre dos variables, es una línea recta.
En donde se tiene pares de observaciones
Se considera un *modelo lineal* cuando los parámetros ocurren de manera lineal,

$Y = \beta_0 + \beta_1 X + \varepsilon$	Modelo simple lineal
$Y = \beta_0 . \beta_1^X + \varepsilon$	Modelo simple exponencial
$Y = \beta_0 . X^{\beta 1} + \varepsilon$	Modelo simple potencial

…..



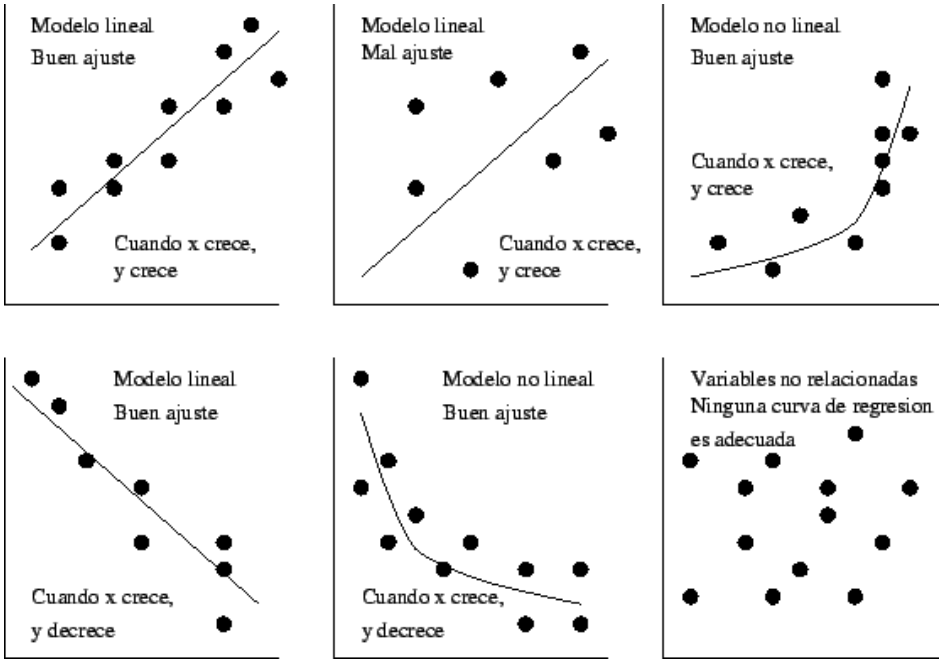
Ejms:

- El gasto en función del ingreso.
- El peso de las personas está en función de su estatura
- El número de partidos que ganarás en la liga en función del número de goles que marques.

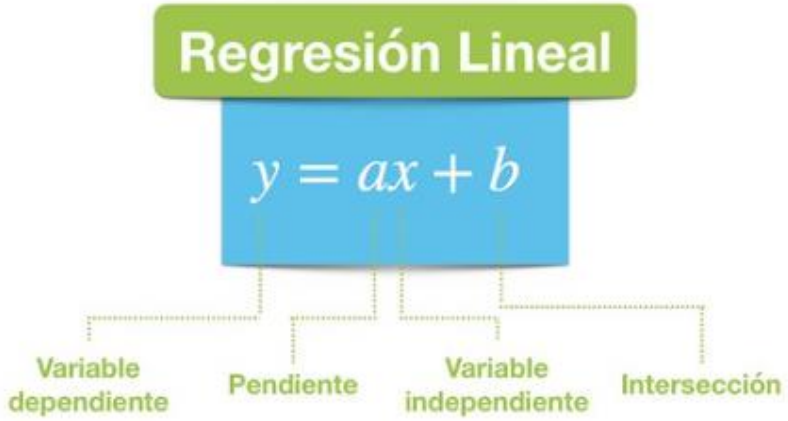
¿Cómo se analiza un modelo de regresión?

¿ Cómo determinar si se debe aplicar un modelo de regresión simple?

DIAGRAMAS DE DISPERSIÓN



Esta es la ecuación matemática que resuelve un conjunto de valores



Especificación del modelo estadístico

El **modelo estadístico** de regresión lineal simple, de Y en función de X, es:

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i$$ Donde:

Y_i	=	La i-ésima observación de la variable aleatoria dependiente Y.
X_i	=	es la i-ésima observación de la variable independiente X
β_o	=	es el intercepto y es una constante (parámetro)
β_1	=	es llamado la pendiente y es una constante (parámetro)
ε_i	=	es la componente aleatoria error

βo y β_1 : Parámetros de la regresión.

Supuestos para εi

Que debe tener en cuenta

¿ Cómo obtener la ecuación de regresión o modelo ajustado ?

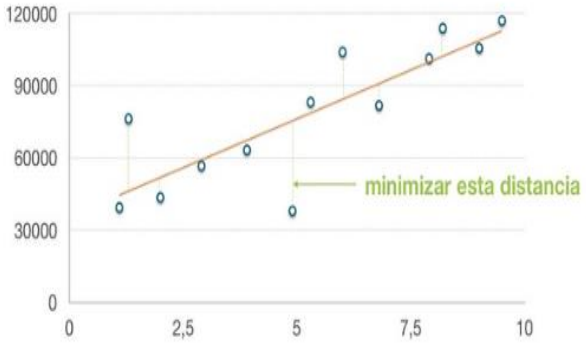
ESTIMACION DE PARAMETROS POR MINIMOS CUADRADOS

Sabemos que:

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

El método de mínimos cuadrados trata de buscar cual es la recta que más se acerca a los puntos; es decir busca la recta que haga que la distancia entre el valor real y_i, y el valor obtenido por la recta ajustada $\hat{y}_i$ sea la más pequeña,

El método de mínimos cuadrados encuentra los estimadores de los parámetros β_0 y β_1 tal que la suma de cuadrados de los residuales sea mínima. La aplicación del método de mínimos cuadrados consiste en:



Y así, la suma de todas estas distancias pueden ser simbolizadas como:

$$\text{Suma de cuadrados del error} = SC_{\text{Error}} = \sum_{I=1}^{8} (y_i - \hat{y}_i)^2$$

sea la más pequeña. Como la mejor recta está determinada por $\hat{\beta}_0$ y $\hat{\beta}_1$ entonces matemáticamente, se desea escoger los valores para $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimicen la suma de cuadrados del error

1. Escribir la suma de cuadrados del error

$$S = S(\beta_o, \beta_1) = SC_{\text{Error}} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2$$

2. Obtener la derivada de la suma de cuadrados del error con respecto a cada parámetro del

modelo;

es decir $\frac{\partial S}{\partial \beta_0}$ y $\frac{\partial S}{\partial \beta_1}$.

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i-1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i-1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i)$$

3. Igualar las derivadas a cero y simplificar (se debe sustituir β_0 y β_1 por sus respectivos estimadores b_0 y b_1).

$$-2\sum_{i-1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$-2\sum_{i-1}^{n} X_i(Y_i - b_0 - b_1 X_i) = 0$$

Simplificando obtenemos las siguientes expresiones:

$$\sum_{i=1}^{n} Y_i - nb_0 - b_1 \sum_{i=1}^{n} X_i = 0 \quad (1)$$

$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0 \quad (2)$$

(1) y (2) son llamadas **Ecuaciones normales.**

4. Solucionar el sistema de ecuaciones o ecuaciones normales.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Reemplazando el valor de b_1 en la ecuación normal (1) se obtiene la solución para b_0 $b_0 = \overline{Y} - b_1 \overline{X}$

Luego la ecuación de regresión o modelo ajustado es $\hat{Y} = b_0 + b_1 X$

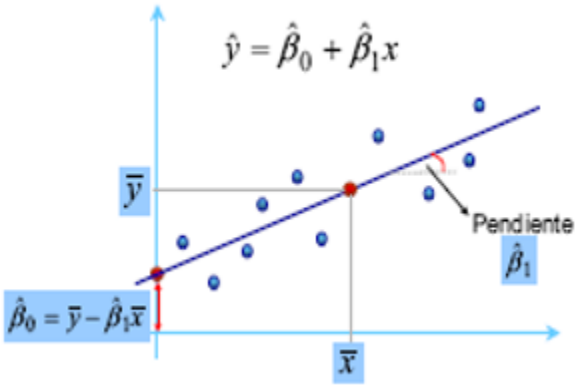
$$\hat{Y} = \hat{\beta}o + \hat{\beta}1X$$

Propiedades de los residuos de Mínimos Cuadrados

Nótese que :

1. $\sum \square_i = 0$
2. $\sum x_i \square_i = 0$ son dos propiedades del estimador de mínimos cuadrados:

Gráficamente estimación de los coeficientes de la recta



Interpretación de Coeficientes

Sabemos que : $\hat{Y} = \hat{\beta}o + \hat{\beta}1X$

$\hat{\beta}o$
$\hat{\beta}1$

Estimación de la varianza σ^2

La desviación típica de la perturbación, σ, mide la precisión del ajuste de la recta de regresión. Para medir la variabilidad de los puntos alrededor de la recta utilizaremos la desviación típica residual (estimador de σ). Se define la varianza residual como:

$$S^2_R = \sum e^2_i \; / \; (n-2)$$

donde $e_i = y_i - \hat{y}_i$ son los ***residuos*** del modelo.

Entonces $S^2_R = \sum e^2_i \; / \; (n-2)$

Ejemplo

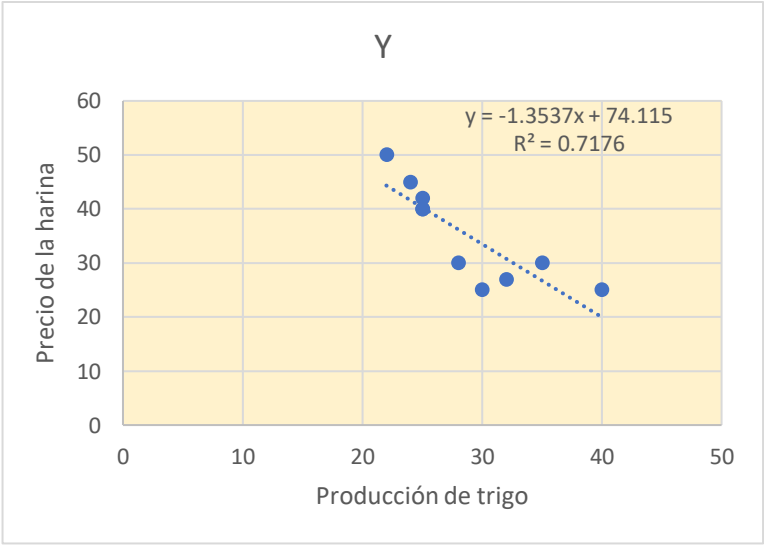
Los datos de la producción de trigo en toneladas (X) y el precio del kilo de harina en pesetas

(Y) en la década de los 80 en España fueron:
Producción de trigo:
Precio de la harina :
Ajusta la recta de regresión por el método de mínimos cuadrados

a. Calcula la varianza residual.

Resultados

a. Para determinar de manera inicial la relación lineal entre las dos variables se debe laborar un diagrama de dispersión



Interpretación : $\hat{\beta}_0$ *:* $\hat{\beta}_1$*:*.

b. Sabemos que la *varianza residual* esta definida como:

$S^2_R = \sum e^2_i \,/\, (n-2)$ *Interpretar* S_R

…

Análisis de la significancia

Para analizar si β_1 es cero, tenemos tres herramientas:

- Intervalos de confianza.
- Contrastes de Hipótesis:
  - Estadístico t.
  - p-valor.

Intervalos de confianza

Calcularemos un rango donde estará la estimación del verdadero valor del parametro por ejemplo $\beta 1$, cualquiera que sea la muestra que tomemos.
Esto lo aseguramos con una cierta probabilidad (generalmente el 95%).

Los intervalos de confianza para los parámetros se definen como

$$IC_{1-\alpha}(\beta_1) = \left(\hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{S}_R}{S_x\sqrt{n}}\right)$$

Error típico

$$IC_{1-\alpha}(\beta_0) = \left(\hat{\beta}_0 \pm t_{n-2,\alpha/2} \frac{\hat{S}_R}{\sqrt{n}}\sqrt{1+\frac{\bar{x}^2}{S_x^2}}\right)$$

Contrastes de hipótesis

- Una alternativa para asegurar que β_1 no es cero es plantear un contraste según la forma estándar:

H_0: $\beta_1 = 0$,

H_1: $\beta_1 \neq 0$.

TABLA ANOVA

Fuente de variación	Suma de cuadrados	g.l.	Varianza	Test F	$p-$val
Explicada	SCE	1	$\hat{S}_e^2 = \frac{SCE}{1}$	$F = \frac{\hat{S}_e^2}{\hat{S}_R^2}$	¿?
Residual	SCR	$n-2$	$\hat{S}_R^2 = \frac{SCR}{n-2}$		
Total	SCT	$n-1$			

La tabla ANOVA se utiliza para hacer el contraste de la regresión

H_0 : **El modelo de regresión lineal NO sirve para explicar la respuesta**

H_1 : **El modelo de regresión lineal SI sirve para explicar la respuesta**

A nivel de significación α, rechazamos cuando

$$F > F_{1,n-2,\alpha}$$

Rechazamos Ho

La regresión es significativa.

Contrastes de hipótesis: Prueba de la t

- Aún tenemos una alternativa al p-valor para resolver el contraste:

H_0: $\beta 1 = 0$,

H_1: $\beta 1 \neq 0$.

|t|>2 rechazamos Ho,

|t|<2 aceptamos Ho.

|t|>2 Rechazamos Ho. La regresión es significativa.

Coeficiente de determinación – R^2

¿Cómo evaluamos la fuerza del ajuste de una recta regresión?

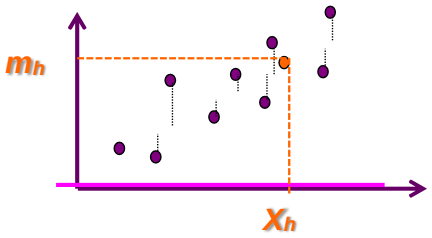
El **COEFICIENTE DE DETERMINACIÓN** es la proporción de variabilidad explicada por la regresión

En REGRESIÓN SIMPLE el COEFICIENTE DE DETERMINACIÓN coincide con el COEFICIENTE DE CORRELACIÓN

El coeficiente de determinación indica cuanto de Y es explicado por X

Estimación de la media de Y

¿ Cuál es la respuesta media para un valor fijo de $x = X_h$?



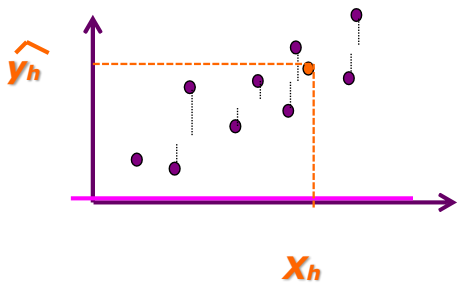
Como no conocemos la media, proponemos la respuesta media que hemos estimado con el modelo – la recta

$$m_h = \beta_0 + \beta_1 x_h$$

$$IC_{1-\alpha}(\text{media}) = \left(m_h \pm t_{n-2,\alpha/2}\, \hat{S}_R \sqrt{\frac{1}{n} + \left(\frac{x_h - \bar{x}}{S_x\sqrt{n}}\right)^2} \right)$$

Predicción de Y

¿ Qué respuesta predecimos para un nuevo valor de $x = X_h$?



La mejor propuesta es la media de las y cuando $x = X_h$. Como no conocemos la media, proponemos la respuesta media que hemos estimado con el modelo – la recta

$$\hat{y}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$$

$$IC_{1-\alpha}(\text{prediccion}) = \left(\hat{y}_h \pm t_{n-2,\alpha/2}\ \hat{S}_R \sqrt{1 + \frac{1}{n} + \left(\frac{x_h - \bar{x}}{S_x \sqrt{n}} \right)^2} \right)$$

Diagnosis

- Necesitamos estar seguros de que nuestras conclusiones son correctas.