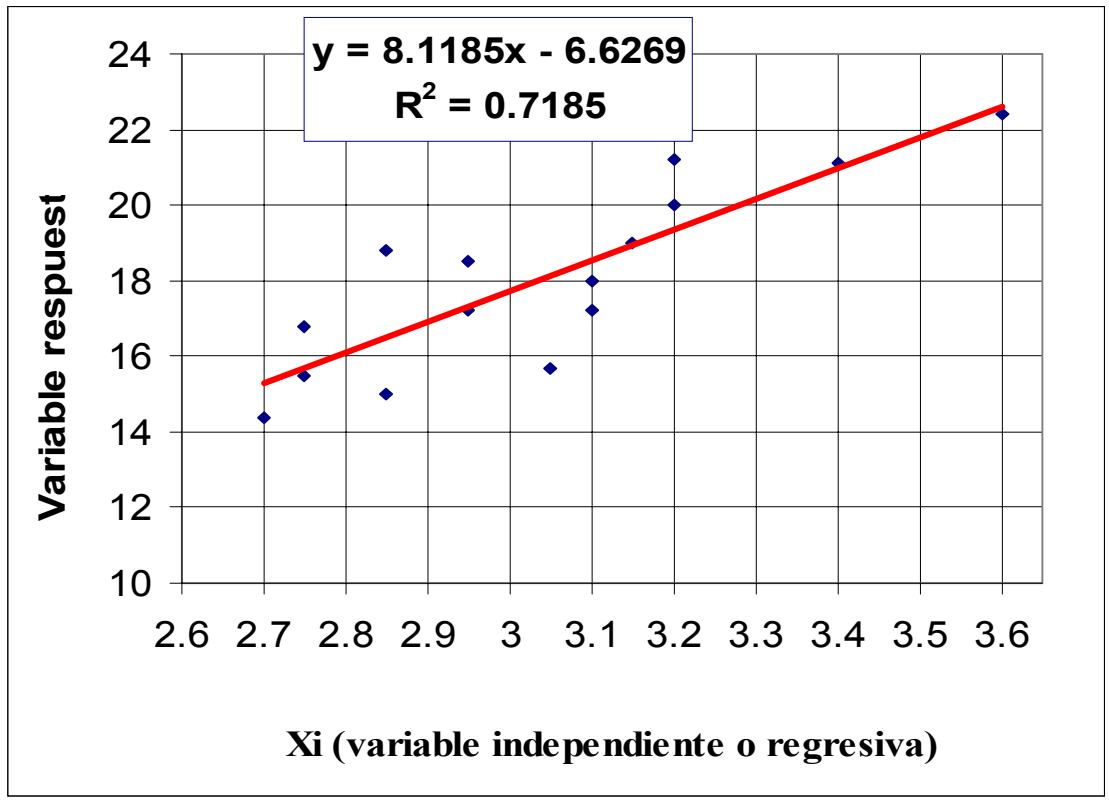
REGRESIÓN LINEAL SIMPLE

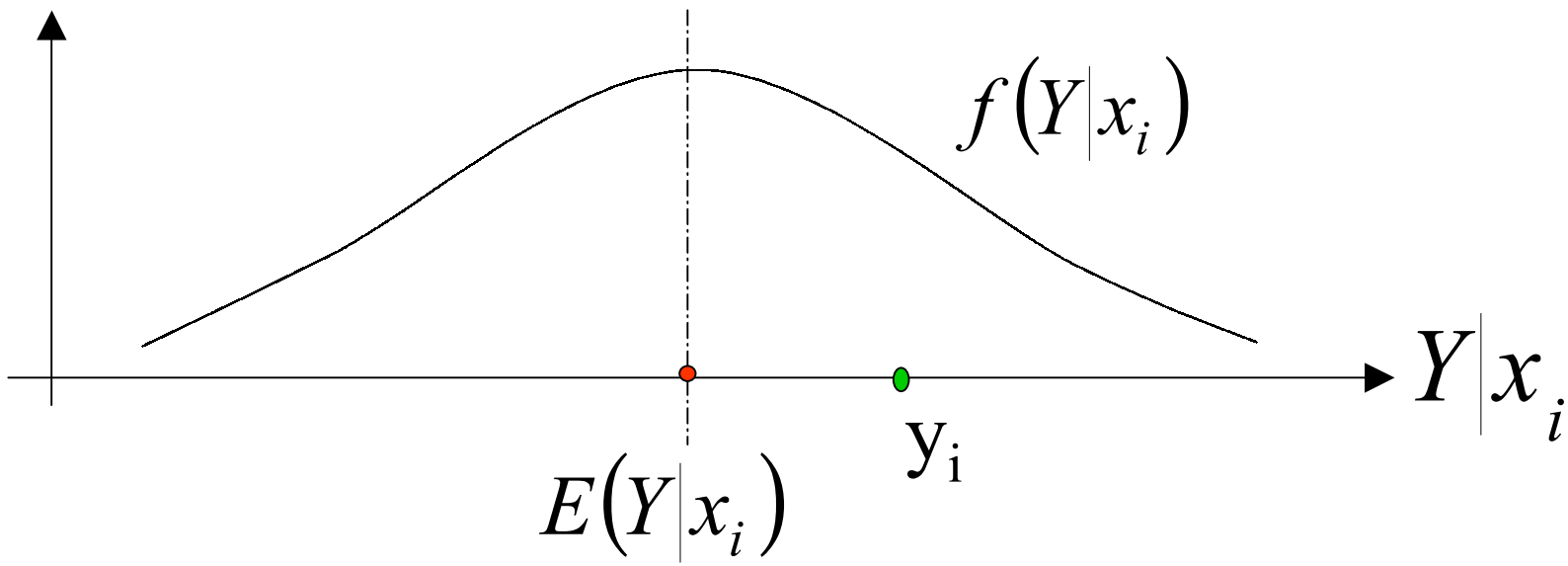
El análisis de regresión es una técnica estadística para investigar la **relación funcional entre dos o más variables**, ajustando algún modelo matemático. La regresión lineal simple utiliza una sola variable de regresión y el caso más sencillo es el modelo de línea recta. Supóngase que se tiene un conjunto de ***n*** pares de observaciones (x_i, y_i), se busca encontrar una recta que describa de la mejor manera cada uno de esos pares observados.

CP xi	SI yi
2.95	**18.5**
3.2	**20**
3.4	**21.1**
3.6	**22.4**
3.2	**21.2**
2.85	**15**
3.1	**18**
2.85	**18.8**
3.05	**15.7**
2.7	**14.4**
2.75	**15.5**
3.1	**17.2**
3.15	**19**
2.95	**17.2**
2.75	**16.8**
45.6	**270.8**

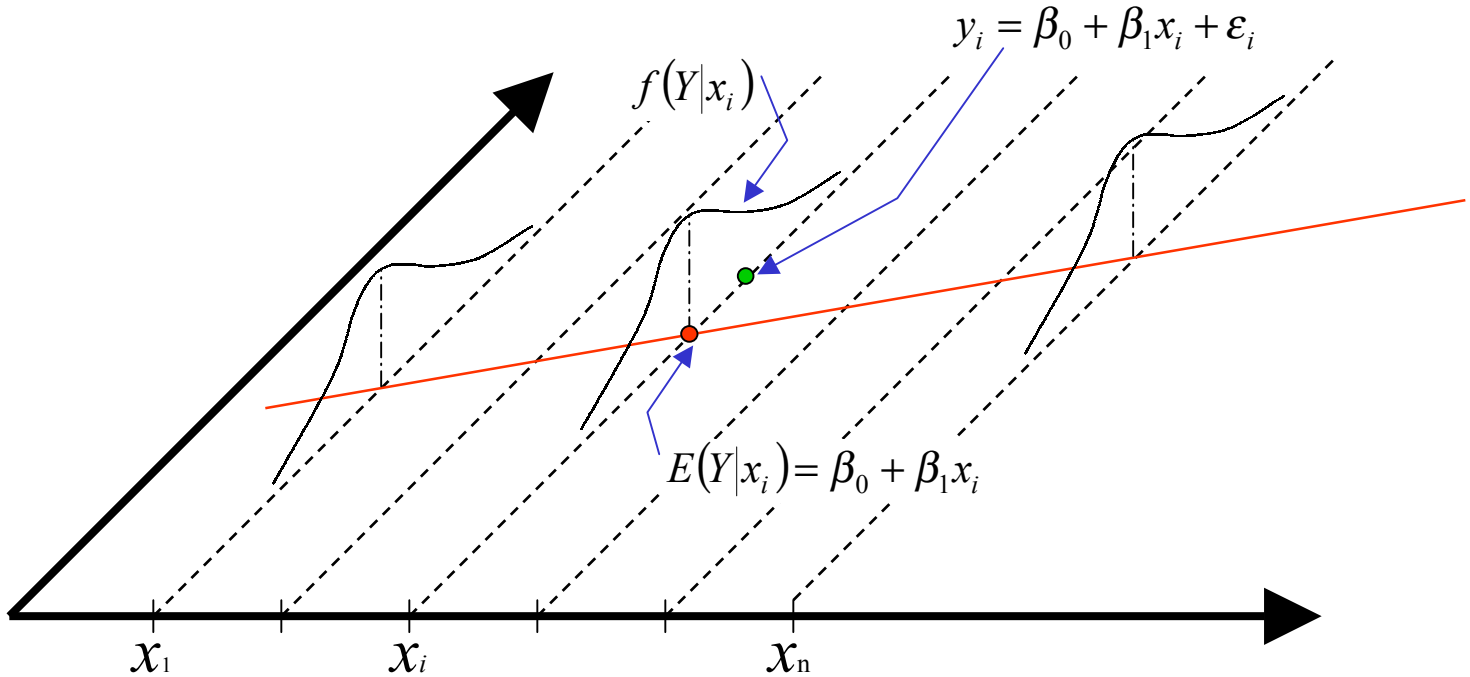


Preparado por: Irene P. Valdez y Alfaro

Se considera que la variable X es la variable independiente o regresiva y se mide sin error, mientras que Y es la variable respuesta para cada valor específico x_i de X; y además Y es una variable aleatoria con alguna función de densidad para cada nivel de X.



Regresión Lineal Simple



Si la recta de regresión es: $Y = \beta_0 + \beta_1 X$

Cada valor y_i observado para un x_i puede considerarse como el valor esperado de Y dado x_i más un error:

Modelo lineal simple : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Los ε_i se suponen errores aleatorios con distribución normal, media cero y varianza σ^2 ; β_0 y β_1 son constantes desconocidas (parámetros del modelo de regresión)

Preparado por: Irene P. Valdez y Alfaro

Método de Mínimos Cuadrados para obtener estimadores de β_0 y β_1

Consiste en determinar aquellos estimadores de β_0 y β_1 que minimizan la suma de cuadrados de los errores ε_i ; es decir, los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 respectivamente deben ser tales que:

$$\sum_{i=1}^{n} \varepsilon_i^{\,2}$$ sea mínima.

Del modelo lineal simple: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

de donde: $\varepsilon_i = y_i - \beta_0 - \beta_1 x$

elevando al cuadrado: $$\sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2$$

Según el método de mínimos cuadrados, los estimadores de β_0 y β_1 debe satisfacer las ecuaciones:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x)^2 = 0$$

Al derivar se obtiene un sistema de dos ecuaciones denominadas "ecuaciones normales":

$$\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Cuya solución es:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Ahora, el modelo de regresión lineal simple ajustado (o recta estimada) es:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Preparado por: Irene P. Valdez y Alfaro

Con respecto al numerador y denominador de B_1 suelen expresarse como S_{xy} y S_{xx} respectivamente:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \quad \Rightarrow \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Puede demostrarse que: $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

y

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i$$

Por otro lado puede demostrarse que los estimadores de β_0 y β_1 son insesgados con varianzas:

$$V(\hat{\beta}_0) = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \quad \text{y} \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$ respectivamente.

Como σ^2 (la varianza de los errores ε_i) es en general desconocida, para estimarla definimos el residuo como: $e_i = y_i - \hat{y}_i$ y la suma de cuadrados del error como:

$$SS_E = \sum_{i=1}^{n} e_i^2 \quad \Rightarrow \quad SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

que al sustituir $\hat{y}_i$ también puede expresarse como: $SS_E = S_{yy} - \hat{\beta}_1 S_{xy}$

donde: $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

Sea $MS_E = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2} = \dfrac{SS_E}{n-2}$ Entonces: $E(MS_E) = \sigma^2 \Rightarrow \hat{\sigma}^2 = MS_E$

Preparado por: Irene P. Valdez y Alfaro

Con lo anterior, las varianzas estimadas de $\hat{\beta}_0$ y $\hat{\beta}_1$ son respectivamente:

$$\hat{V}(\hat{\beta}_0) = MS_E\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad \text{y} \qquad \hat{V}(\hat{\beta}_1) = \frac{MS_E}{S_{xx}}$$

Además, si se cumplen los supuestos de que los ε_i se distribuyen normalemte con media cero y varianza σ_2, entonces, los estadísticos

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_E\left[\frac{1}{n} + \frac{x^2}{S_{xx}}\right]}} \qquad \text{y} \qquad T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_E}{S_{xx}}}}$$

tienen cada uno distribución t de Student con n-2 grados de libertad.

Lo que permite efectuar pruebas de hipótes y calcular intervalos de confianza sobre los parámetros de regresión β_0 y β_1 .

Un caso de particular interés es probar la hipótesis:

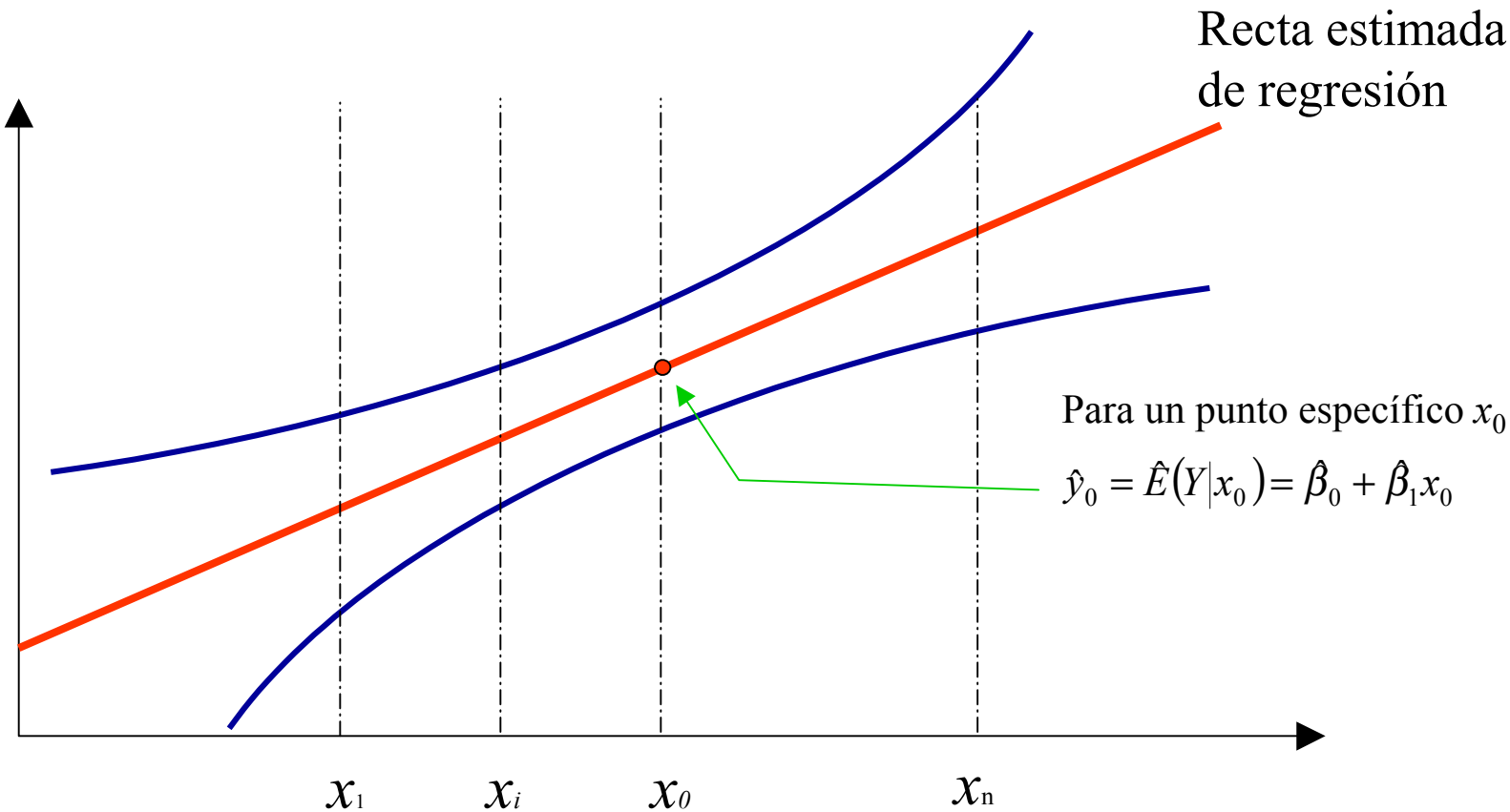
$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

Ya que si la pendiente es igual cero, entonces puede significar o que la variación de X no influye en la variación de Y, o que no hay regresión lineal entre X y Y.

Por otro lado, si la pendiente es diferente de cero, entonces existirá algún grado de asociación lineal entre las dos variables, es decir, la variabilidad de X explica en cierta forma la variabiliad de Y (aunque no implica que no pueda obtenerse un mejor ajuste con algún polinomio de mayor grado en X).

Nota: si se utilizara en lugar de una recta, una curva con grado mayor a 1 en X pero grado 1 en los coeficientes de X, la regresión sigue siendo lineal, ya que es lineal en los parámetros de regresión p.ej. $\mathbf{Y=\beta_o+\beta_1 x+\beta_2 x^2}$

Estimación de intervalos de confianza en torno a la línea de regresión: BANDAS DE CONFIANZA



Preparado por: Irene P. Valdez y Alfaro

Estimación de la respuesta media para un x_0 específico:

$$\hat{\mu}_{\hat{y}_0} = \hat{y}_0 = \hat{E}(Y|x_0) = \beta_0 + \beta_1 x_0$$

$$V(\hat{y}_0) = \sigma^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}\right] \Rightarrow \hat{V}(\hat{y}_0) = MS_E\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}\right]$$

$\hat{y}_0$ tiene distribución normal, por lo que: $\dfrac{\hat{y}_0 - \mu_{\hat{y}_0}}{\sqrt{\hat{V}(y_o)}}$

tiene distribución T de Student con n-2 grados de libertad, por lo que los límites de confianza superior e inferior para la respuesta media dado x_0 están dados por: $\hat{y}_0 \pm t_{\alpha/2,n-2}\sqrt{\hat{V}(y_o)}$

Graficando los limites de confianza superior e inferior de $\mu_{\hat{y}_0}$ para cada punto x_i de X pueden dibujarse las bandas de confianza para la recta de regresión.

Puede observarse que la amplitud del intervalo de confianza es mínima cuando $x_0 = \bar{x}$ mientras que es mayor en los extremos de los valores observados de X.

Preparado por: Irene P. Valdez y Alfaro

Predicción de nuevas observaciones

Nótese que $\hat{y}_0$ es la respuesta media para los valores de x_i seleccionados para encontrar la recta de regresión; sin embargo, frecuentemente es de interés predecir la respuesta futura para un x_a dado seleccionado posteriormente.

Sea Y_a la observación futura en $x = x_a.$, ; Y_a es una variable aleatoria con varianza σ^2 y por otro lado, la varianza de $\hat{y}_a = \hat{\beta}_0 + \hat{\beta}_1 x_a$ es $\hat{V}(\hat{y}_a) = MS_E\left[1 + \frac{1}{n} + \frac{(x_a - \bar{x})^2}{S_{xx}}\right]$