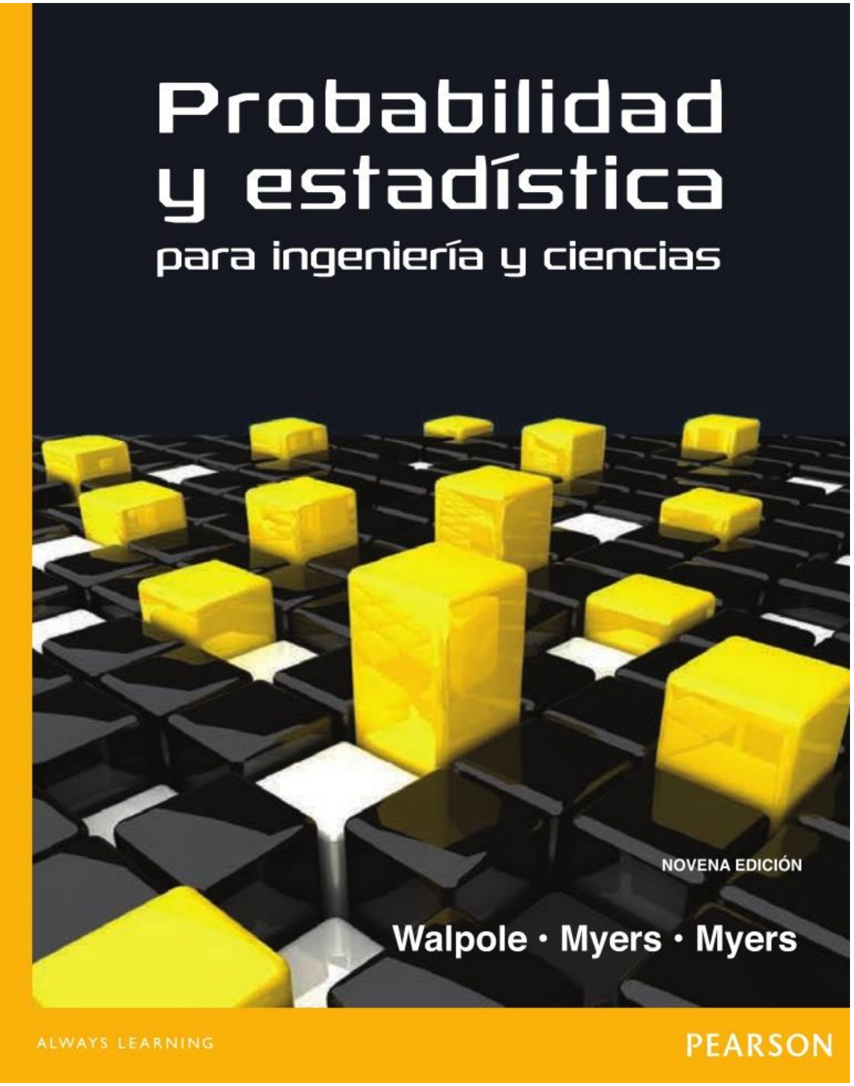
Análisis de Datos I

Unidad 7: Regresión Lineal Simple

Clase: Regresión Lineal Simple

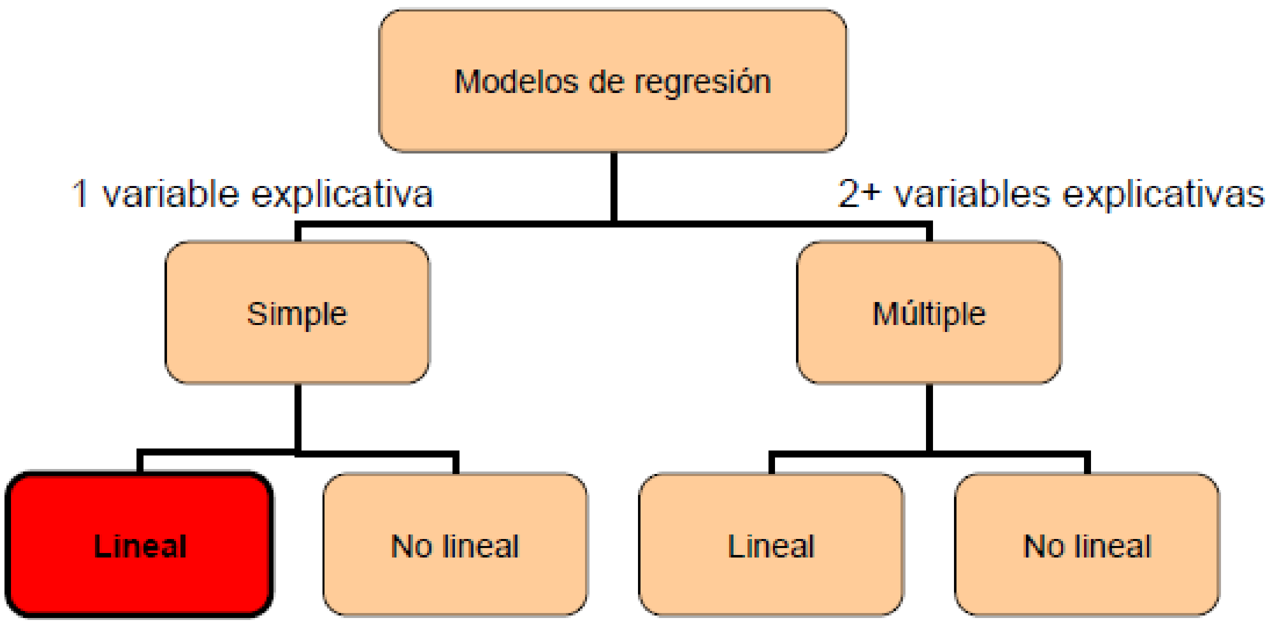
II Semestre de 2024

UNIVERSIDAD DEL NORTE


Probabilidad
y estadística
para ingeniería y ciencias
NOVENA EDICIÓN
Walpole · Myers · Myers
ALWAYS LEARNING
PEARSON

Modelos de regresión

El análisis de regresión es un estudio de las relaciones entre variables: una variable dependiente y una o más variables independientes.



Modelos de regresión

- Los problemas que involucran conjuntos de variables cuando se sabe que existe alguna relación inherente entre las variables se pueden resolver mediante modelos de regresión.

- Los modelos de regresión pueden ser lineales o no lineales dependiendo de la estructura de la ecuación y la relación entre las variables dependientes e independientes.

Lineales: Los parámetros están expresados en su menor exponente. Existen modelos lineales simples y múltiples.

Ejemplos de no lineales:

Exponencial	$\hat{y} = \beta_0 e^{\beta_1 x}$	Recíproca	$\hat{y} = \beta_0 + \beta_1 \frac{1}{x}$
Potencial	$\hat{y} = \beta_0 x^{\beta_1}$	Hiperbólica	$\hat{y} = \frac{x}{\beta_0 + \beta_1 x}$

Modelo de regresión lineal simple

El modelo básico de regresión lineal simple (RLS) es:

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

Donde:

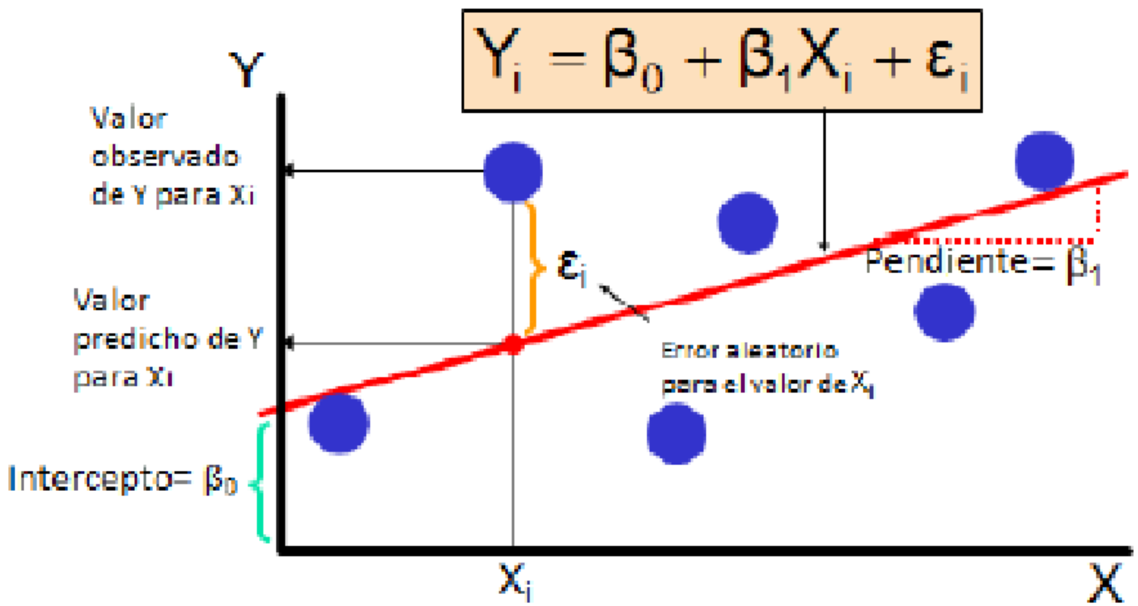
β_0 es el intercepto. Valor de Y cuando X=0.

β_1 es la pendiente. Cambio en Y por unidad de cambio en X.

Y es la variable dependiente, variable de respuesta. No se puede controlar, es aleatoria.

X es la variable independiente, regresor del modelo. Se puede controlar, no es aleatoria.

ε es la perturbación, todo elemento que afecta a Y. Es aleatoria. E[ε]=0



$$\hat{y} = b_0 + b_1 \cdot x$$

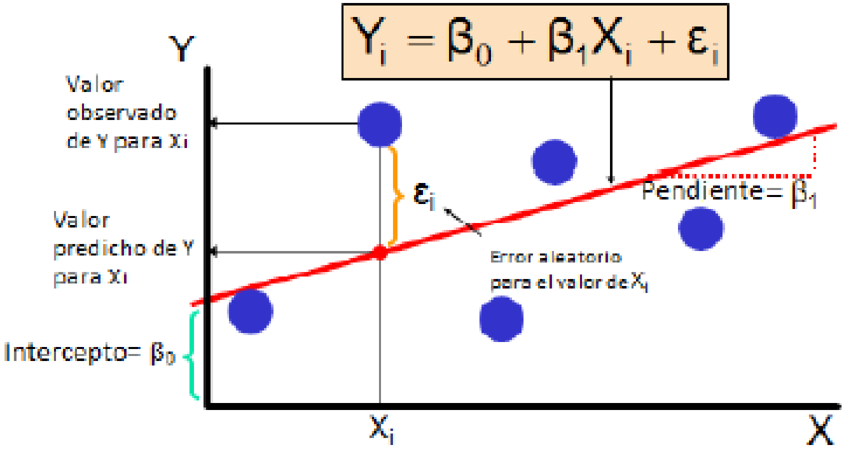
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Modelo de regresión lineal simple

Debemos encontrar los valores b_0 y b_1, que corresponde a las estimaciones de β_0 y β_1 respectivamente, basándonos en una muestra de tamaño n, construyendo un conjunto de parejas (xi,yi), de tal manera que la suma de los cuadrados de los residuos sea mínimo. A esta suma de cuadrados de los residuos se le llama suma de cuadrados de los errores (SSE, por sus siglas en inglés).

$$\hat{y} = b_0 + b_1 x$$

$$SSE = \sum_{i=1}^{n} e_i^2$$



Modelo de regresión lineal simple

La ecuación de regresión estimada será representada así: $\hat{y} = b_0 + b_1 x$

Modelo de Parámetro	Estimación
β_0	b_0
β_1	b_1
σ^2	s_e^2
E(y|x)	$\hat{y}$

Los errores e=$y - \hat{y}$ se denominan frecuentemente **residuales.**

Procedimiento para realizar una R.L.S.

1. Identificar cuál es la variable dependiente (Y) e independiente (X).
2. Estimar β_0 y β_1 a partir de b_0 y b_1 (recta ajustada preliminar $\hat{y} = b_0 + b_1 \cdot x$).
3. Determinar si β_0 y β_1 son significativos ($\beta_0 \neq 0$ y $\beta_1 \neq 0$).

   3.1. Realizar pruebas de hipótesis para β_0 y β_1 utilizando Dist. T Student.

   3.2. Realizar un ANOVA (prueba de hipótesis para β_1 utilizando Dist. Fisher)

   3.3. Estimar intervalos de confianza para β_0 y β_1.

4. Escribir la recta ajustada definitiva para $\hat{y} = b_0 + b_1 \cdot x$.

   4.1. Si $\beta_1 = 0$ no hay relación entre X y Y (no hay recta).

   4.2. Si $\beta_0 = 0$ y $\beta_1 \neq 0$ la recta no tiene intercepto ($\hat{y} = b_1 \cdot x$).

   4.3. Si $\beta_0 \neq 0$ y $\beta_1 \neq 0$ la recta sería igual a la preliminar ($\hat{y} = b_0 + b_1 \cdot x$).

Procedimiento para realizar una R.L.S.

5. Verificar qué tan buena es la recta ajustada (análisis de correlación).

   5.1. Coeficiente de determinación (R^2).

   5.2. Coeficiente de correlación (ρ, γ).

6. Verificar los supuestos realizados sobre los errores (análisis residual). Si los **tres supuestos** se cumplen, entonces la recta ajustada es válida.

   6.1. Verificar si los errores distribuyen normal con $\mu = 0$ y $\sigma^2 = s_e^2$ (realizar una prueba de bondad de ajuste).

   6.2. Verificar si los errores son independientes (Durbin-Watson).

   6.3. Verificar si los errores tienen una varianza constante (Homocedasticidad). Se realiza una gráfica de errores contra las variables X y/o Y.

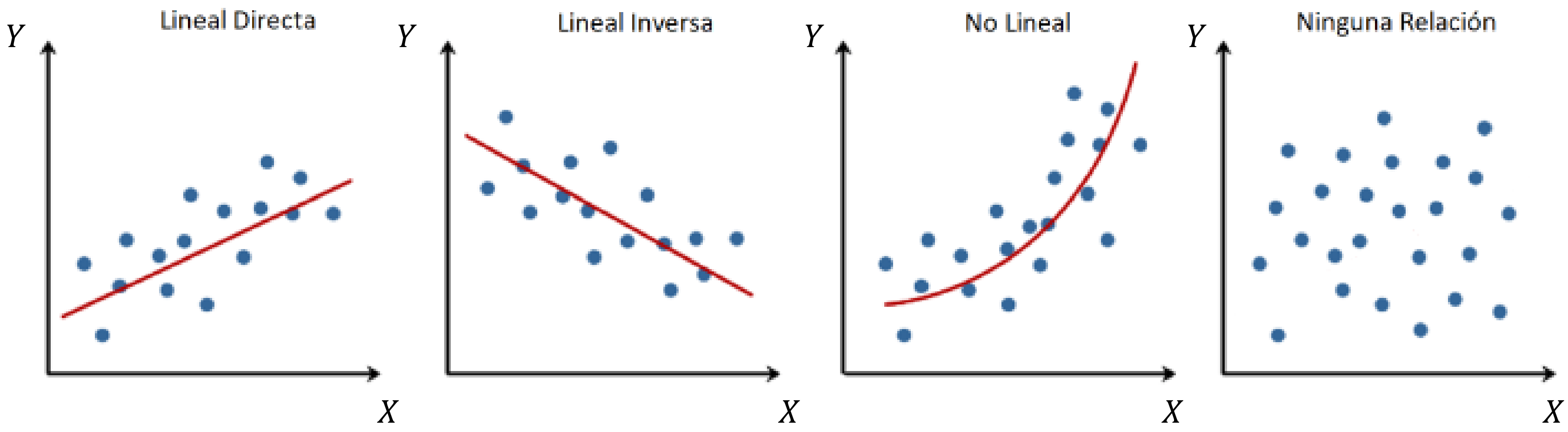
7. Inferencias para la variable dependiente (Y). Intervalos de confianza y predicción.

Supuestos para el modelo de R.L.S.

- Para cada valor de x, la variable aleatoria ε se distribuye normalmente.
- Para cada valor de x, la media o valor esperado de ε es 0; esto es E(ε)=$\mu_{\varepsilon} = 0$
- Para cada valor de x, la varianza de ε es la constante δ^2
- Los valores del término de error ε son independientes.

Paso 1. Identificar las variables

El diagrama de dispersión se usa comúnmente para mostrar cómo dos variables se relacionan entre sí.



Paso 2. Estimar β_0 y β_1 a partir de b_0 y b_1

Método de los mínimos cuadrados:

Suma de cuadrados para la variable independiente $SS_x = \sum x^2 - \frac{(\sum x)^2}{n}$

Suma de cuadrados para la variable dependiente $SS_y = \sum y^2 - \frac{(\sum y)^2}{n}$

Suma de los productos cruzados $SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$

Pendiente de la recta de regresión $b_1 = \frac{SS_{xy}}{SS_x}$

Intercepto $b_0 = \bar{y} - b_1\bar{x}$

Paso 3. Determinar si β_0 y β_1 son significativos

3.1. Realizar un ANOVA (prueba de hipótesis para β_1 utilizando D. Fisher)

El análisis de varianza divide la variación de Y en dos componentes: modelo de regresión y errores. La idea es que la variable dependiente Y sea explicada por $\beta_0 + \beta_1 X$, y que el error ε tenga una pequeña influencia en la variación.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Suma de Cuadrados Totales (SST)	=	Suma de Cuadrados de la Regresión (SSR)	+	Suma de Cuadrados de los Errores (SSE)

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

Paso 3. Determinar si β_0 y β_1 son significativos

3.1. Realizar un ANOVA (prueba de hipótesis para β_1 utilizando D. Fisher)

$SST = SS_y$	$SS_y = \sum y^2 - \frac{(\sum y)^2}{n}$
$SSR = b_1 \cdot SS_{xy}$	$SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n}$
$SSE = SS_y - b_1 \cdot SS_{xy}$	$SS_x = \sum x^2 - \frac{(\sum x)^2}{n}$

SST: variabilidad total de Y.

SSR: variabilidad de Y explicada por el modelo ajustado.

SSE: variabilidad de Y explicada por los errores.

Paso 3. Determinar si β_0 y β_1 son significativos

3.1. Realizar un ANOVA (prueba de hipótesis para β_1 utilizando D. Fisher)

Planteamiento de las hipótesis H_0: $\beta_1 = 0$ y H_1: $\beta_1 \neq 0$

Fuente de varianza	Suma de cuadrados	Grados de libertad	Cuadrados medios	$f_{calculado}$
Regresión	SSR	*1*	$MS_R = \frac{SSR}{1}$	$f_0 = \frac{SSR/1}{SSE/(n-2)} = \frac{SSR}{s_e^2}$
Error	SSE	*n-2*	$MS_E = s_e^2 = \frac{SSE}{n-2}$	
Total	SST	*n-1*		

Se rechaza H_0 si $f_0 > f_{1,n-2,\alpha}$ o si $Valor\ P < \alpha, Valor\ P = P(F_{1,n-2} > f_0)$

Si la hipótesis nula no es rechazada, el modelo no es válido para predecir Y. En el caso que la hipótesis nula sea rechazada, el modelo explica la variabilidad de Y, por lo tanto, es un modelo apropiado. Es deseable un F calculado grande para que la hipótesis nula sea rechazada en favor de la alternativa, lo que indica que $\beta_1 \neq 0$

Paso 3. Determinar si β_0 y β_1 son significativos

3.2. Realizar pruebas de hipótesis para β_0 y β_1 utilizando D. T de Student.

Pruebas de hipótesis para β_0:

$$H_0 : \beta_0 = B_0 \qquad H_0 : \beta_0 \leq B_0 \qquad H_0 : \beta_0 \geq B_0$$
$$H_1 : \beta_0 \neq B_0 \qquad H_1 : \beta_0 > B_0 \qquad H_1 : \beta_0 < B_0$$

Valor de la estadística de prueba:

$$t_0 = \frac{b_o - B_0}{\sqrt{s_e^2 \cdot \left(\frac{1}{n} + \frac{\bar{x}^2}{SS_x}\right)}}$$

Distribución *t* con n – 2 grados de libertad.

Planteamiento de las hipótesis H_0: $\beta_0 = 0$ y H_1: $\beta_0 \neq 0$

Paso 3. Determinar si β_0 y β_1 son significativos

3.2. Realizar pruebas de hipótesis para β_0 y β_1 utilizando D. T de Student.

Pruebas de hipótesis para β_1:

$$H_0 : \beta_1 = B_1 \qquad H_0 : \beta_1 \leq B_1 \qquad H_0 : \beta_1 \geq B_1$$
$$H_1 : \beta_1 \neq B_1 \qquad H_1 : \beta_1 > B_1 \qquad H_1 : \beta_1 < B_1$$

Valor de la estadística de prueba:

$$t_0 = \frac{b_1 - B_1}{\sqrt{\frac{s_e^2}{SS_x}}}$$

Distribución t con n – 2 grados de libertad.

Planteamiento de las hipótesis H_0: $\beta_1 = 0$ y H_1: $\beta_1 \neq 0$

Paso 3. Determinar si β_0 y β_1 son significativos

3.3. Estimar intervalos de confianza para β_0 y β_1.

Intervalos del $(1-\alpha)100\%$ de confianza para β_0:

$$b_0 - t_{n-2,\alpha/2} \cdot s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}} \leq \beta_0 \leq b_0 + t_{n-2,\alpha/2} \cdot s_e\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_x}}$$

Intervalos del $(1-\alpha)100\%$ de confianza para β_1:

$$b_1 - t_{n-2,\alpha/2} \cdot \frac{s_e}{\sqrt{SS_x}} \leq \beta_1 \leq b_1 + t_{n-2,\alpha/2} \cdot \frac{s_e}{\sqrt{SS_x}}$$

Paso 3. Determinar si β_0 y β_1 son significativos

Caso Especial: $\beta_0 = 0$ (β_0 no es significativo).

ANOVA presenta modificaciones

Planteamiento de las hipótesis H_0: $\beta_1 = 0$ y H_1: $\beta_1 \neq 0$

Fuente de varianza	Suma de cuadrados	Grados de libertad	Cuadrados medios	$f_{calculado}$
Regresión	SSR	*1*	$MS_R = \frac{SSR}{1}$	$f_0 = \frac{SSR/1}{SSE/(n-1)} = \frac{SSR}{s_e^2}$
Error	SSE	*n-1*	$MS_E = s_e^2 = \frac{SSE}{n-1}$	
Total	SST	*n*		

Se rechaza H_0 si $f_0 > f_{1,n-1,\alpha}$ o si $Valor\ P < \alpha$, $Valor\ P = P(F_{1,n-1} > f_0)$

Paso 3. Determinar si β_0 y β_1 son significativos

Caso Especial: $\beta_0 = 0$ (β_0 no es significativo).

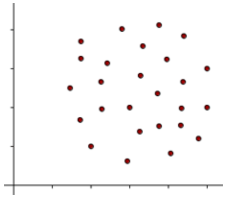
Intervalo de confianza para β_1 presenta modificaciones

Intervalos del $(1-\alpha)100\%$ de confianza para β_1:

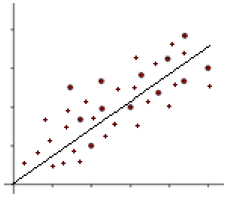
$$b_1 - t_{n-1,\alpha/2} \cdot \frac{s_e}{\sqrt{SS_x}} \leq \beta_1 \leq b_1 + t_{n-1,\alpha/2} \cdot \frac{s_e}{\sqrt{SS_x}}$$

Paso 4. Escribir la recta ajustada definitiva

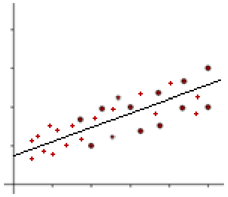
4.1. Si $\beta_1 = 0$ no hay relación entre X y Y (no hay recta).



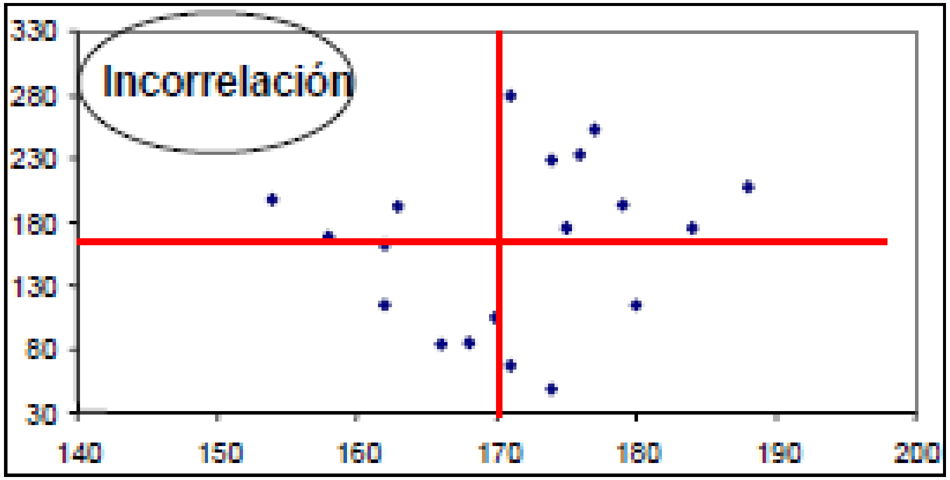
4.2. Si $\beta_0 = 0$ y $\beta_1 \neq 0$ la recta no tiene intercepto ($\hat{y} = b_1 \cdot x$).



4.3. Si $\beta_0 \neq 0$ y $\beta_1 \neq 0$ la recta sería igual a la preliminar ($\hat{y} = b_0 + b_1 \cdot x$).

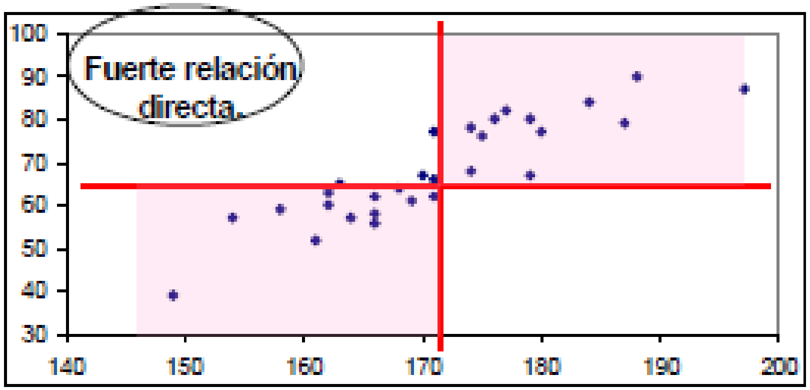


Paso 5. Verificar qué tan buena es la recta



Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares. Incorrelación.

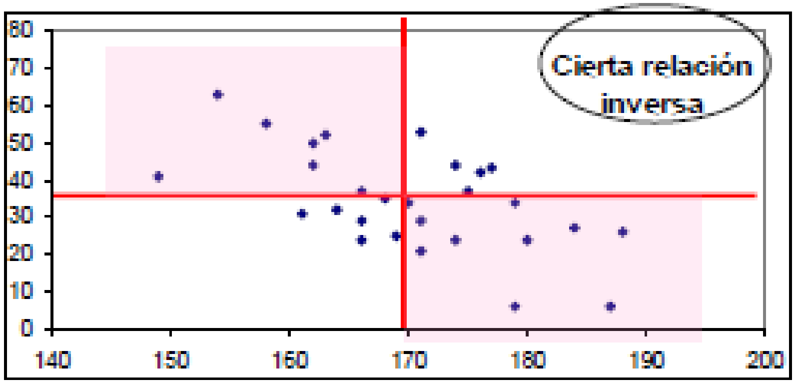
Paso 5. Verificar qué tan buena es la recta



•Para los valores de X mayores que la media le corresponden valores de Y mayores también.

•Para los valores de X menores que la media le corresponden valores de Y menores también.

•Esto se llama relación directa o creciente entre X e Y.



Para los valores de X mayores que la media le corresponden valores de Y menores. Esto es relación inversa o decreciente.

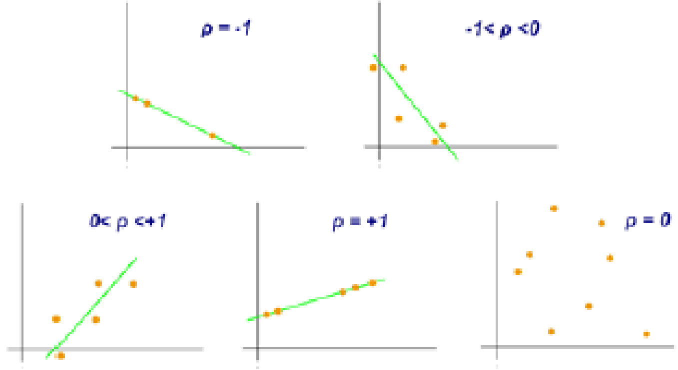
Paso 5. Verificar qué tan buena es la recta

5.1. Coeficiente de correlación

El coeficiente de correlación de Pearson mide el grado de asociación lineal entre X y Y. Se puede estimar por medio del coeficiente de correlación muestral o estimado.

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \beta_1 \frac{\sigma_X}{\sigma_Y} \qquad -1 \leq \rho \leq 1$$

$$\gamma = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} \qquad -1 \leq \gamma \leq 1$$



Paso 5. Verificar qué tan buena es la recta

5.1. Coeficiente de correlación

- $\rho = 0$ cuando $\beta = 0$: No existe regresión lineal, es decir, la recta de regresión es horizontal y cualquier conocimiento de X es inútil para predecir Y.
- $\rho = \pm 1$ cuando $\sigma^2 = 0$: Existe relación lineal perfecta entre las dos variables
- Los estimadores muestrales de ρ con valores cercanos a la unidad implican una buena correlación entre X y Y.
- Los estimadores muestrales de ρ con valores cercanos a cero implican poca o ninguna correlación entre X y Y.

Paso 5. Verificar qué tan buena es la recta

5.2. Coeficiente de determinación

El coeficiente de determinación, denotado R^2 es la proporción de la varianza en la variable dependiente que es predecible a partir de la(s) variable(s) independiente(s)

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}$$

0.9 – 1.0 Excelente
0.8 – 0.9 Muy bueno
0.6 – 0.8 Bueno
0.5 – 0.6 Regular
< 0.5 Malo

SST: variabilidad total de Y.

SSR: variabilidad de Y explicada por el modelo ajustado.

SSE: variabilidad de Y explicada por los errores.

Determina la calidad del modelo para replicar los resultados, y a su vez representa la proporción de variación de los resultados que puede explicarse por el modelo.

Paso 6. Verificar supuestos sobre los errores

6.1. Verificar si los errores distribuyen normal con $\mu = 0$ y $\sigma^2 = s_e^2$

H_o : Los resíduos se distribuyen normalmente con media $\mu = 0$ y varianza σ^2 estimada con s_e^2

H_1 : Los resíduos **NO** se distribuyen normalmente con media $\mu = 0$ y varianza σ^2 estimada con s_e^2

Se determinan los residuos $e_i = y_i - \widehat{y}_i$ para las observaciones en la muestra y se aplica la prueba de bondad de ajuste chi-cuadrado.

Paso 6. Verificar supuestos sobre los errores

6.2. Verificar si los errores son independientes

La independencia de los residuos se puede probar mediante el análisis de gráfico de residuos, que en el caso que exista un patrón, se rechaza Ho. Otra forma de probarlo es utilizando la prueba Durbin-Watson, que es la más empleada para detectar auto correlación en los errores.

H_0:Los residuos son independientes.

H_1:Los residuos son dependientes.

El estadístico Durbin-Watson tomará valores entre 0 y 4, y si este valor es cercano a 2 los residuos se asumen independientes. Para cierto nivel de significancia se buscan los valor dl y du de la tabla de Durbin-Watson.

Si 0≤ $Dw \leq dl$, correlación negativa
Si $dl \leq Dw \leq du$, inconcluso
Si $du \leq Dw \leq 4-du$, los residuos son independientes
Si $4-du \leq Dw \leq 4-dl$, inconcluso
Si $4-dl \leq Dw \leq 4$, correlación positiva

$$D_w = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Paso 6. Verificar supuestos sobre los errores

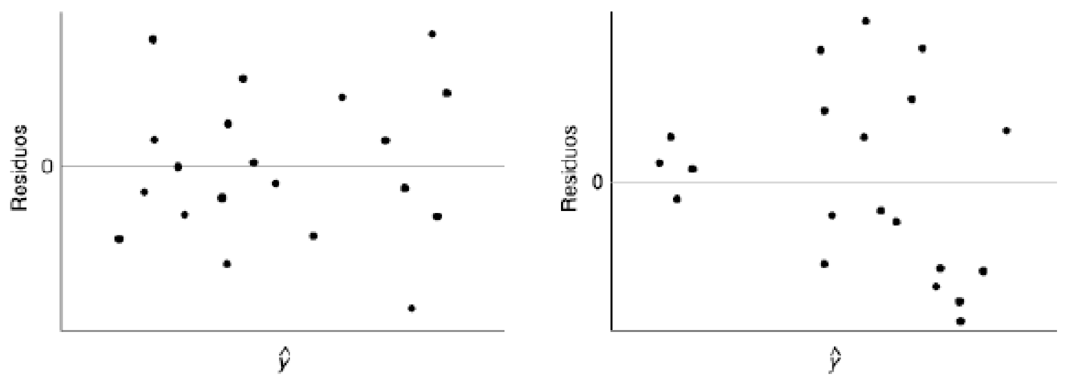
6.3. Verificar si los errores tienen una varianza constante (Homocedasticidad)

La homocedasticidad indica la homogeneidad de la varianza.

H_0:Los residuos son homocedásticos.

H_1:Los residuos son heterocedásticos.

Se realiza una gráfica de errores e_i contra $y\hat{}$ o contra x, o de e_z (residuos estandarizados) contra $y\hat{}$ o contra x. Los residuos deben estar dispersos alrededor del cero de forma aleatoria, entre -3 y 3 desviaciones estándar de la media sin mostrar ningún patrón



Paso 6. Verificar supuestos sobre los errores

6.3. Verificar si los errores tienen una varianza constante (Homocedasticidad)

La **prueba de Breusch-Pagan** se utiliza para determinar si la heteroscedasticidad está presente o no en un modelo de regresión.

La prueba utiliza las siguientes hipótesis nulas y alternativas :
Hipótesis nula (H0): la homocedasticidad está presente (los residuos se distribuyen con la misma varianza).
Hipótesis alternativa (H1): Heteroscedasticidad está presente (los residuos no se distribuyen con la misma varianza).

Si el valor p que corresponde a este estadístico de prueba de chi-cuadrado con p (el número de predictores) grados de libertad es menor que algún nivel de significancia (α), entonces rechace la hipótesis nula y concluya que la heteroscedasticidad está presente. De lo contrario, no rechace la hipótesis nula. En este caso, se asume que existe homocedasticidad.

Paso 6. Verificar supuestos sobre los errores

6.3. Verificar si los errores tienen una varianza constante (Homocedasticidad)

Usamos los siguientes pasos para realizar una prueba de Breusch-Pagan:

1. Ajuste el modelo de regresión.
2. Calcule los residuos cuadrados del modelo.
3. Ajuste un nuevo modelo de regresión, utilizando los residuos al cuadrado como valores de respuesta.
4. Calcule el estadístico de prueba de chi-cuadrado X^2 como **n*** $\mathbf{R^2}$ nuevo donde:

- **n:** el número total de observaciones
- $\mathbf{R^2_{nuevo}}$ **:** R-cuadrado del nuevo modelo de regresión que utilizó los residuos al cuadrado como valores de respuesta.

Paso 6. Verificar supuestos sobre los errores

6.3. Verificar si los errores tienen una varianza constante (Homocedasticidad)

- Si la gráfica de residuos contra valores ajustados $(\widehat{y})$no muestra una tendencia importante, es *homocedástica*, es probable pero no se tiene la seguridad que los supuestos del modelo lineal sean válidos. Sin embargo si la gráfica de residuos muestra tendencia importante o se curva, es *heterocedástica*. Si esto ocurre se realiza una transformación de la variable de interés, elegir la transformación adecuada es difícil y no siempre funcionan, si este es el caso se debe pensar en una regresión múltiple.
- Examinar datos atípicos o puntos influyentes.
- El coeficiente de correlación es una medida numérica de la fuerza de la relación lineal entre de las variables, no indica cualquier otro tipo de relación.

Paso 7. Inferencias para Y

Intervalos de confianza y predicción para la variable dependiente

Intervalos de $(1-\alpha)100\%$ de confianza para la respuesta media $E(y|x_0)$:

$$\hat{y} - t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}} \leq E(y|x_0) \leq \hat{y} + t_{n-2,\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}}$$

Intervalos de predicción del $(1-\alpha)100\%$ de confianza para y dado $x = x_0$:

$$\hat{y} - t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}} \leq y|x_0 \leq \hat{y} + t_{n-2,\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_x}}$$