Investigación de Operaciones
Regresión Lineal

Nicolás Rojas Morales
nicolas.rojasm@usm.cl

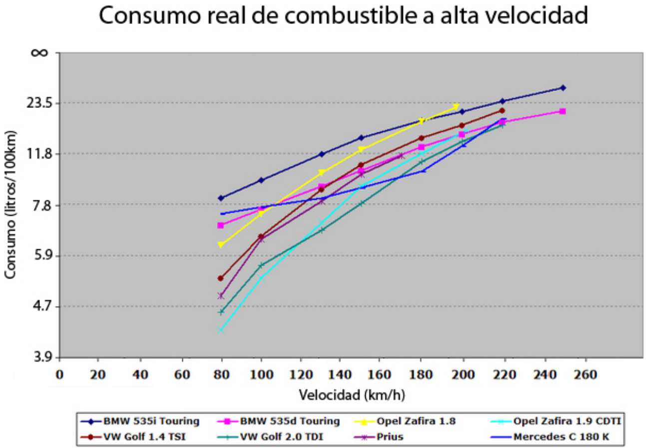
Departamento de Informática
Universidad Técnica Federico Santa María

Introducción

¿Existe una relación entre el consumo de bencina y la velocidad media de un vehículo?

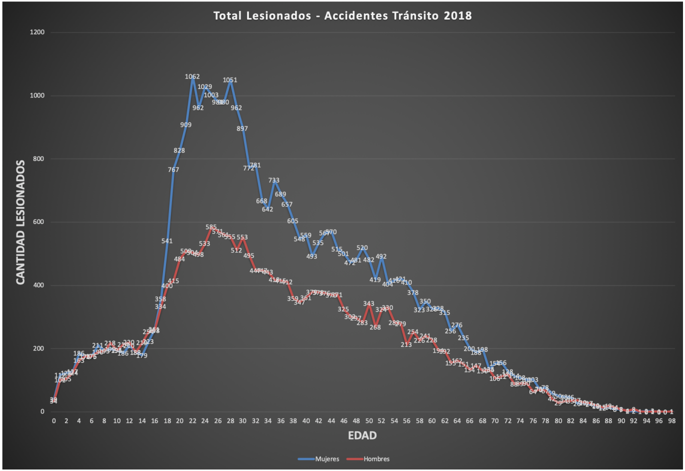
Introducción

¿Existe una relación entre el consumo de bencina y la velocidad media de un vehículo?



Introducción

¿Existe una relación entre la cantidad de accidentes de tránsito y la edad de quien conduce?



Fuente: `https://www.conaset.cl/programa/observatorio-datos-estadistica/`

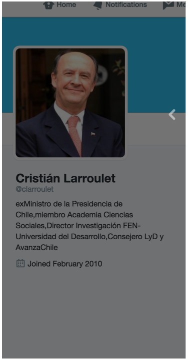
Introducción

Otras relaciones interesantes:

- ¿Existe una relación entre la calidad de la educación entregada y el (incremento del) costo de la educación?
- Asistencia a clases de una asignatura y el promedio de notas obtenido
- Duración de una paralización de actividades y personas que efectivamente participan de la movilización

Introducción

Otras relaciones interesantes:

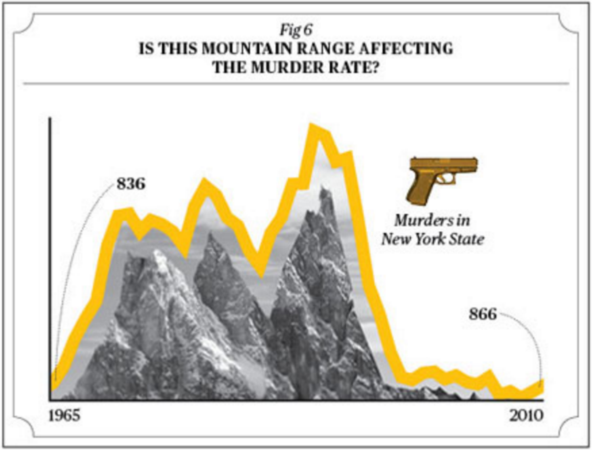


link:

- https://twitter.com/clarroulet/status/730931861384507392?lang=es
- http://www.elmostrador.cl/noticias/opinion/2016/05/16/larroulet-y-la-distorsion-de-la-realidad/

Introducción

Correlación no implica causalidad



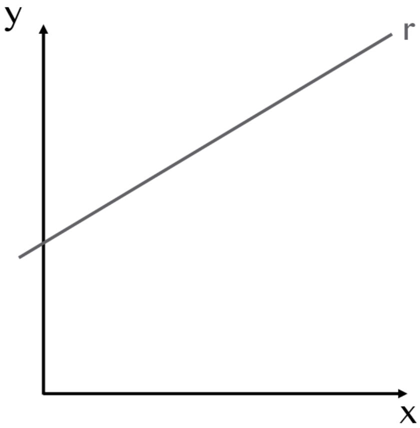
Introducción

Objetivo

- Estudiar la relación estadística que existe entre una o más variables *independientes* $(x_1, x_2, \ldots, x_k)$ y una variable *dependiente* (Y)
- Se define un modelo entre las variables, en este caso, una relación lineal
- Si consideramos una variable dependiente y una independiente, el modelo se reduce a una línea recta de la siguiente forma:

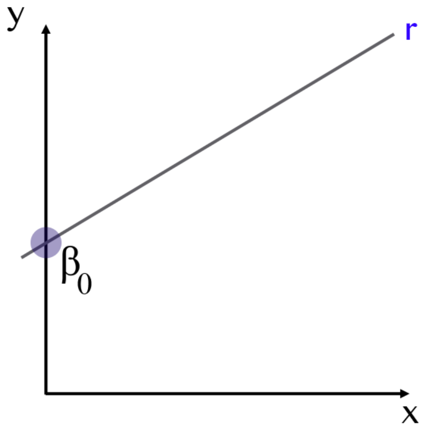
$$Y = \beta_0 + \beta_1 * X \tag{1}$$

Introducción



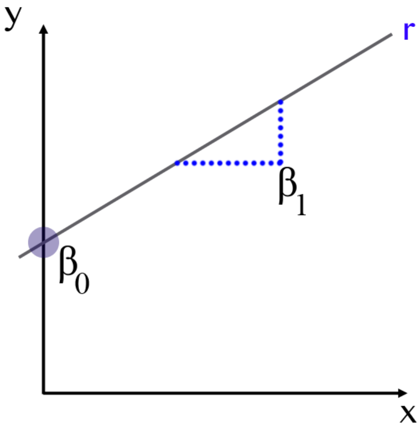
$$r : Y = \beta_0 + \beta_1 * X \tag{2}$$

Introducción



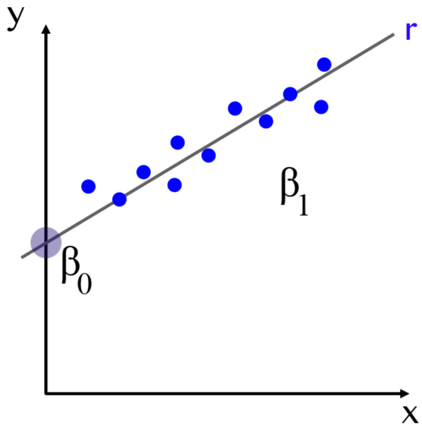
$$r : Y = \beta_0 + \beta_1 * X \tag{2}$$

Introducción



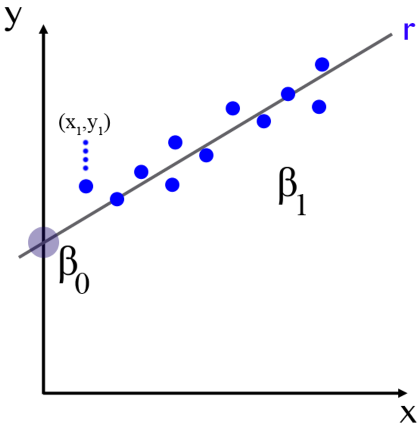
$$r : Y = \beta_0 + \beta_1 * X \tag{2}$$

Introducción



$$r : Y = \beta_0 + \beta_1 * X \tag{2}$$

Introducción



$$r : Y = \beta_0 + \beta_1 * X \tag{2}$$

Introducción

Un modelo de regresión lineal simple se formula mediante

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \quad (3) \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \quad (4)$$

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i \quad (5)$$

donde

- x_i valor de la variable independiente en el iésimo ensayo
- y_i valor de la variable dependiente en el iésimo ensayo
- β_0 y β_1 coeficientes de la regresión
- ϵ_i error aleatorio

Introducción

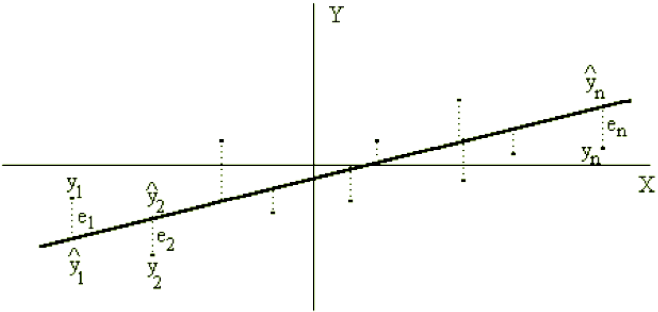
Alcances

Lo que estudiaremos en esta sección de la asignatura es:

- Entender la importancia de estudiar la Regresión Lineal
- Aprender a construir modelos de Regresión Lineal. Para ello necesitamos:
    1. Determinar los coeficientes de la Regresión
    2. Verificar aspectos sobre la relación entre las variables dependiente e independiente
    3. Realizar tests...

Construcción del Modelo

Errores Aleatorios



$$\epsilon_k = y_k - \hat{y}_k$$

Construcción del Modelo

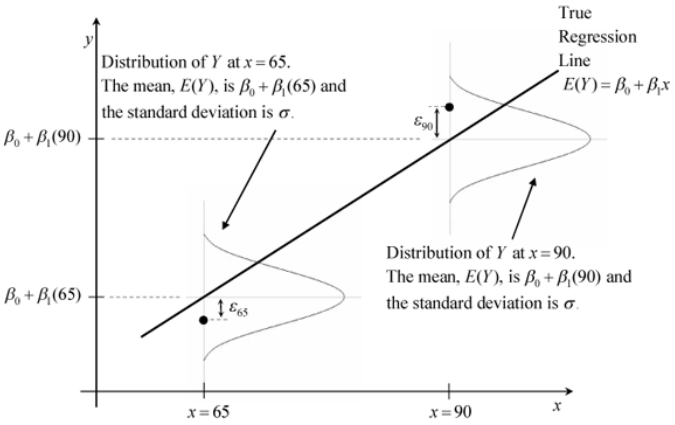
Supuestos

Para construir una regresión lineal debemos suponer que los Errores Aleatorios (ϵ) siguen una distribución Normal:

- $E(\epsilon_i) = 0$
- $VAR(\epsilon_i) = \sigma^2$
- $COV(\epsilon_i, \epsilon_j) = 0, \ \forall i, j : i \neq j$

Por ende, $\epsilon \sim N(0, \sigma^2)$

Construcción del Modelo



Construcción del Modelo

Estimación de los parámetros

Para estimar los coeficientes β_0 y β_1, utilizaremos la técnica de los Mínimos Cuadrados:

- Dado un conjunto de datos, cada uno definido como un par variable dependiente/independiente
- Se intenta encontrar una función contínua que mejor se aproxime a los datos $\rightarrow$ **una combinación lineal que con ciertos coeficientes se minimice el error**
- Utilizando el Mínimo Error Cuadrático

$$E_{cm}(f) = \sqrt{\frac{\sum_{k=1}^{n}(\epsilon_k)^2}{n}} \tag{6}$$

donde $\epsilon_k = y_k - \hat{y}_k$

Construcción del Modelo

Entonces, buscamos minimizar:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 * X_i)^2 \tag{7}$$

para ello derivamos en función de β_1 y obtenemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \tag{8}$$

Construcción del Modelo

Entonces, buscamos minimizar:

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 * X_i)^2 \tag{9}$$

para ello derivamos en función de β_0 y obtenemos:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 * \bar{X} \tag{10}$$

Nota: La recta **debe** pasar por el promedio de ambas variables

Construcción del Modelo

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (11)$$

β_1: Coeficiente de Regresión

β_1 representa el aumento o disminución en la variable dependiente por cada unidad que varía la variable independiente:

- Si $\beta_1 > 0$, las dos variables aumentan o disminuyen a la vez
- Si $\beta_1 < 0$, cuando una variable aumenta, la otra disminuye

Construcción del Modelo

Medidas de variabilidad de x:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{12}$$

Medidas de variabilidad de y:

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{13}$$

Medida de variabilidad conjunta de x e y:

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{14}$$

Ejercicio

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Asumiendo un modelo lineal $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i + \epsilon_i$, obtenga $\hat{\beta}_0$ y $\hat{\beta}_1$.

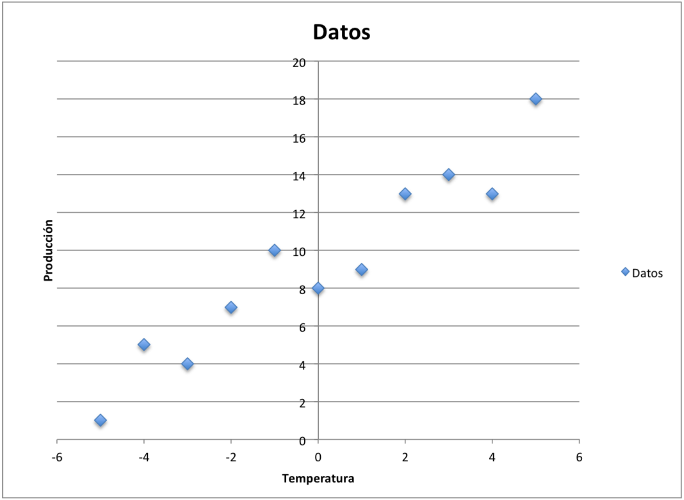
Ejercicio

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Asumiendo un modelo lineal $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i + \epsilon_i$, obtenga $\hat{\beta}_0$ y $\hat{\beta}_1$.

Ejercicio


Datos
20
18
16
14
12
10
8
6
4
2
0
-6
-4
-2
0
2
4
6
Producción
Temperatura
Datos

Ejercicio

	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	-5	25	-8,27	68,44	41,36
	-4	16	-4,27	18,26	17,09
	-3	9	-5,27	27,80	15,82
	-2	4	-2,27	5,17	4,55
	-1	1	0,73	0,53	-0,73
	0	0	-1,27	1,62	0,00
	1	1	-0,27	0,07	-0,27
	2	4	3,73	13,89	7,45
	3	9	4,73	22,35	14,18
	4	16	3,73	13,89	14,91
	5	25	8,73	76,17	43,64
Suma	**0**	**110**	**0,00**	**248,18**	**158,00**

$\bar{X} = 0, \; \bar{Y} = 9{,}27,$

Ejercicio

$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
-5	25	-8,27	68,44	41,36
-4	16	-4,27	18,26	17,09
-3	9	-5,27	27,80	15,82
-2	4	-2,27	5,17	4,55
-1	1	0,73	0,53	-0,73
0	0	-1,27	1,62	0,00
1	1	-0,27	0,07	-0,27
2	4	3,73	13,89	7,45
3	9	4,73	22,35	14,18
4	16	3,73	13,89	14,91
5	25	8,73	76,17	43,64
Suma **0**	**110**	**0,00**	**248,18**	**158,00**

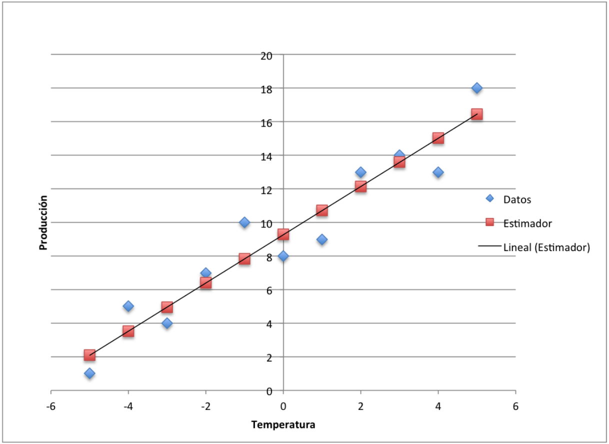
$\bar{X} = 0, \ \bar{Y} = 9{,}27, \ \beta_1 = 1{,}44,$

Ejercicio

$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
-5	25	-8,27	68,44	41,36
-4	16	-4,27	18,26	17,09
-3	9	-5,27	27,80	15,82
-2	4	-2,27	5,17	4,55
-1	1	0,73	0,53	-0,73
0	0	-1,27	1,62	0,00
1	1	-0,27	0,07	-0,27
2	4	3,73	13,89	7,45
3	9	4,73	22,35	14,18
4	16	3,73	13,89	14,91
5	25	8,73	76,17	43,64
Suma **0**	**110**	**0,00**	**248,18**	**158,00**

$\bar{X} = 0,\ \bar{Y} = 9{,}27,\ \beta_1 = 1{,}44,\ \beta_0 = 9{,}27$

Ejercicio



$$\hat{y}_i = 9{,}27 + 1{,}44 * x_i \qquad (15)$$

Ejercicio

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Análisis:

- ¿Se podría estimar $\hat{y}$ para una Temperatura de $X = 1.5$?
- ¿Se podría estimar $\hat{y}$ para una Temperatura de $X = 6$?

Bondad del Ajuste

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}} \tag{16}$$

Coeficiente de Correlación

El Coeficiente de Correlación mide el grado de relación entre las variables:

- Si $|r_{xy}| \sim 1$ - La variable y se puede calcular en base a x y viceversa
- Si $|r_{xy}| \sim 0$ - Las variables x e y no están relacionadas linealmente, por lo tanto no tiene sentido realizar un ajuste lineal

Bondad del Ajuste

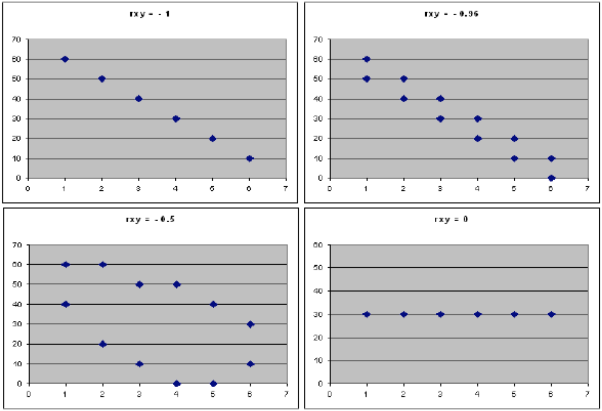
$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} * S_{yy}}} \tag{17}$$

Coeficiente de Correlación

Además, r_{xy} indica el sentido de la relación entre las variables:

- Si $r_{xy} > 0$ las variables son directamente proporcionales: cuando el valor de x aumenta, el valor de y aumenta
- Si $r_{xy} < 0$ las variables son inversamente proporcionales: cuando el valor de x disminuye, el valor de y aumenta

Bondad del Ajuste



Bondad del Ajuste

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18}$$

Coeficiente de Determinación

El coeficiente de Determinación (R^2) mide qué tanto del valor de la variable dependiente se determina a partir de la variable independiente $\sim$ Calidad del Modelo

- Toma valores entre 0 y 1
- $R^2 \sim 1$: mejor es la predicción
- $R^2 \sim 0$: puede indicar la necesidad de aplicar alguna transformación a las variables

Bondad del Ajuste

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{18}$$

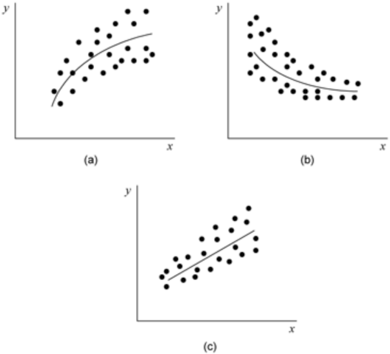
Coeficiente de Determinación

El coeficiente de Determinación (R^2) mide qué tanto del valor de la variable dependiente se determina a partir de la variable independiente $\sim$ Calidad del Modelo

- Toma valores entre 0 y 1
- $R^2 \sim 1$: mejor es la predicción
- $R^2 \sim 0$: puede indicar la necesidad de aplicar alguna transformación a las variables

Por ejemplo, $R^2 = 0{,}85$ implica que el 85 % de la variación de Y se puede atribuír a su asociación lineal con X

Bondad del Ajuste - Transformaciones



(a) $Y* = \sqrt{Y}$
(b) $Y* = \log Y$
(c) $Y* = \frac{1}{Y}$

Ejercicio

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Análisis:

- Calcular el Coeficiente de Correlación
- Calcular el Coeficiente de Determinación

Ejercicio

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Análisis:

- Calcular el Coeficiente de Correlación
- Calcular el Coeficiente de Determinación

Ejercicio

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

Análisis:

- Calcular el Coeficiente de Correlación
- Calcular el Coeficiente de Determinación

$r_{xy} = 0{,}96$ y $R^2 = 0{,}91$

Intervalos de Confianza

Intervalos de Confianza

Al estimar un parámetro, su valor puede ser parte de un Intervalo de Confianza (IC):

- Intervalo donde se estima que estará cierto valor desconocido con una determinada probabilidad de acierto
- Se calcula a partir de datos de una muestra
- La probabilidad de éxito se representa con $1 - \alpha$ (α: nivel de significancia)
- Para construir un IC es necesario conocer la distribución teórica que sigue el parámetro a estimar

Intervalos de Confianza

Estimador de la Varianza (del error)

$$\hat{\sigma}^2 = \frac{\sum e^2}{n-2} \tag{19}$$

Varianza de β_0

$$V(\hat{\beta}_0) = \frac{\hat{\sigma}^2 \sum x_i^2}{n * S_{xx}} \tag{20}$$

Varianza de β_1

$$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}} \tag{21}$$

Recordar: $e_k = y_k - \hat{y}_k$

Intervalos de Confianza

Para el parámetro β_1

$$\beta_1 \in \{\hat{\beta}_1 \pm t_{(\frac{\alpha}{2},n-2)} * \sqrt{V(\hat{\hat{\beta}}_1)}\} \tag{22}$$

Para el parámetro β_0

$$\beta_0 \in \{\hat{\beta}_0 \pm t_{(\frac{\alpha}{2},n-2)} * \sqrt{V(\hat{\hat{\beta}}_0)}\} \tag{23}$$

Nota: Los Grados de Libertad son la cantidad de variables independientes/observaciones usadas para estimar un parámetro **menos** el número de parámetros estimados (incluyendo el intercepto)

Intervalos de Confianza

Considerando que $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 * x_0$:

- Un Intervalo de Confianza del $100 * (1 - \alpha)\,\%$ para la **respuesta** de y dado que $x = x_0$

$$\hat{y}_0 \pm t_{(\frac{\alpha}{2}, n-2)} * \hat{\sigma} * \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{24}$$

- Un Intervalo de Confianza del $100 * (1 - \alpha)\,\%$ para la respuesta de y dado que $x = x_0$:[1]

$$\hat{y}_0 \pm t_{(\frac{\alpha}{2}, n-2)} * \hat{\sigma} * \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \tag{25}$$

[1] Considerando nuevas observaciones que estén dentro del dominio de la variable x

Ejercicio

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

- Calcular un Intervalo de Confianza para β_1 con $\alpha = 0{,}05$
- Calcular un Intervalo de Confianza para y cuando $X = 3$ con $\alpha = 0{,}05$

Ejercicio

Necesitamos

- $\hat{\sigma}^2$
- $V(\hat{\beta}_1)$
- $t_{(0{,}025, n-2=9)}$
- S_{xx}

Ejercicio

Necesitamos

- $\hat{\sigma}^2 = 2{,}34 \rightarrow \hat{\sigma} = 1{,}54$
- $V(\hat{\beta}_1) = 0{,}022 \rightarrow \sqrt{V(\hat{\beta}_1)} = 0{,}145$
- $t_{(1-0{,}025, n-2=9)} = 2{,}262$
- $S_{xx} = 110$

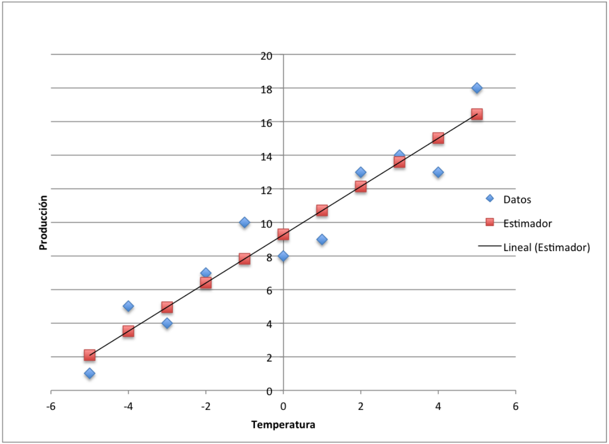
Ejercicio

Necesitamos

- $\hat{\sigma}^2 = 2{,}34 \rightarrow \hat{\sigma} = 1{,}54$
- $V(\hat{\beta}_1) = 0{,}022 \rightarrow \sqrt{V(\hat{\beta}_1)} = 0{,}145$
- $t_{(1-0{,}025,n-2=9)} = 2{,}262$
- $S_{xx} = 110$

Calcular un Intervalo de Confianza para β_1 con $\alpha = 0{,}05$:
R: $\beta_1 \in [1{,}11\ ,1{,}77]$
Calcular un Intervalo de Confianza para el verdadero valor medio de Y cuando $X = 3$ con $\alpha = 0{,}05$:
R: $x_0 = 3$, $\hat{y}_0 = 13{,}58$ luego $y \in [12{,}14\ ,\ 15, 03]$

ANOVA



ANOVA (Analysis of Variance) permite verificar la significancia de la regresión usando la varianza de los datos.

ANOVA

ANOVA

Usaremos este test para verificar si realmente existe una relación lineal entre las variables.

- En este caso, comparamos el conjunto de datos Y con el conjunto $\hat{Y}$
- Se descompone la variación total de una variable en sus diferentes fuentes de variación
  1. Debido a la Regresión $\rightarrow$ puntos en la recta.
  2. Debido al Error $\rightarrow$ ciertos puntos que no siguen la recta (variabilidad sin explicar).

ANOVA

Cada variación puede ser representada por una Suma de Cuadrados:

- Suma de Cuadrados Total:

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{26}$$

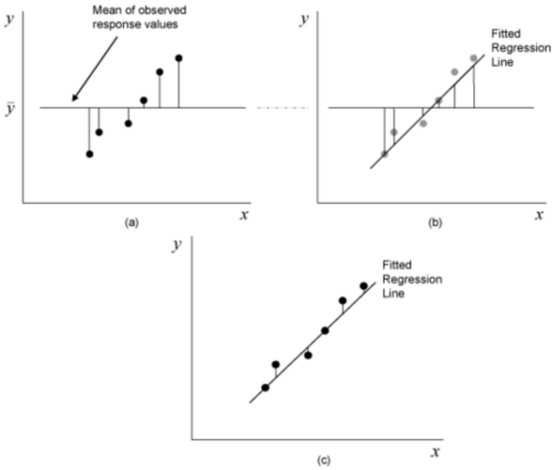
- Suma de Cuadrados de la Regresión:

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{27}$$

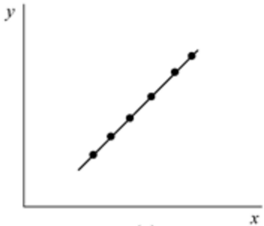
- Suma de Cuadrados del Error:

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{28}$$

ANOVA


Mean of observed
response values
y
$\bar{y}$
x
(a)
y
Fitted
Regression
Line
x
(b)
y
Fitted
Regression
Line
x
(c)

ANOVA



Entonces,

$$SCT = SCR + SCE \tag{29}$$

Caso ideal: todos los puntos de Y pasan por la recta modelada:
$\sum_{i=1}^{n} y_i - \hat{y}_i = 0 \rightarrow SCE = 0$
$SCT = SCR$

ANOVA

La idea es realizar una prueba de hipótesis para verificar si existe una relación lineal entre las variables:

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

Para probar la hipótesis nula se usa el estadístico:

$$f = \frac{\frac{SCR}{1}}{\frac{SCE}{(n-2)}} \tag{30}$$

y se rechaza H_0 con un nivel de significancia de α cuando $f > F_{1-\alpha,1,n-2}$

- Nota: Cuando se rechaza H_0 se entiende $\beta_1 \neq 0 \rightarrow$ no hay evidencia suficiente para suponer que el modelo NO es lineal

ANOVA

Construimos la Tabla ANOVA:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Estadístico Calculado
Regresión	SCR	1	$CMR = \frac{SCR}{1}$	$f = \frac{CMR}{CME}$
Error	SCE	n - 2	$CME = \frac{SCE}{n-2}$	
Total	SCT	1 + n - 2 = n - 1		

$$R_\alpha = f > F_{1-\alpha,1,n-2}$$

Ejercicio

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

- Construya la Tabla ANOVA
- Probar con $\alpha = 0{,}05$ la hipótesis $H_0 : \beta_1 = 0$ v/s $H_1 : \beta_1 \neq 0$

Ejercicio

	SCT	SCR	SCE
	68,44	51,58	1,19
	18,26	33,01	2,17
	27,80	18,57	0,93
	5,17	8,25	0,36
	0,53	2,06	4,68
	1,62	0,00	1,62
	0,07	2,06	2,92
	13,89	8,25	0,73
	22,35	18,57	0,17
	13,89	33,01	4,07
	76,17	51,58	2,39
Suma	**248,18**	**226,95**	**21,24**

ANOVA

Construimos la Tabla ANOVA:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrados Medios	Estadístico Calculado
Regresión	SCR $=$ 226,95	1	$CMR = \frac{226,95}{1}$	$f = \frac{226,95}{2,36} = 96,18$
Error	SCE $=$ 21,24	n-2 $=$ 9	$CME = \frac{SCE}{n-2} = 2,36$	
Total	SCT $=$ 248,18	n-1 $=$ 10		

$R_\alpha = f > F_{1-\alpha,1,n-2}$
$R_\alpha = 96{,}18 > 5{,}12$

Dócimas de Hipótesis en Regresión

Dócimas de Hipótesis

- Una dócima es una manera de realizar inferencia estadística con el objetivo de probar una hipótesis
- En regresión, las hipótesis están relacionadas con los parámetros β_0 y β_1
- En ambos casos, la idea apunta a verificar si los parámetros toman algún valor específico
- Por ejemplo:
  1. La variable dependiente crece el doble de la variable independiente (aproximadamente)

Dócimas de Hipótesis en Regresión

Dócimas de Hipótesis: β_1

Supongamos que deseamos saber si el valor del parámetro β_1 corresponde a un valor C. El estadístico para realizar la dócima corresponde a:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - C}{\frac{\hat{\sigma}}{\sqrt{S_{xx}}}} \tag{31}$$

Las hipótesis pueden formularse de la siguiente manera:

$H_0 : \beta_1 = C$	$H_0 : \beta_1 \leq C$	$H_0 : \beta_1 \geq C$
$H_1 : \beta_1 \neq C$	$H_1 : \beta_1 > C$	$H_1 : \beta_1 < C$
$R_\alpha = \{\lvert t_{\beta_1}\rvert > t_{(1-\frac{\alpha}{2}, n-2)}\}$	$R_\alpha = \{t_{\beta_1} > t_{(1-\alpha, n-2)}\}$	$R_\alpha = \{t_{\beta_1} < t_{(\alpha, n-2)}\}$

Nota: Se rechaza cuando se cumple la condición (Región Crítica R_α)

Dócimas de Hipótesis en Regresión

Dócimas de Hipótesis: β_0

Supongamos que deseamos saber si el valor del parámetro β_0 corresponde a un valor I. El estadístico para realizar la dócima corresponde a:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - I}{\hat{\sigma} * \sqrt{\frac{\sum x_i^2}{n * S_{xx}}}} \tag{32}$$

Las hipótesis pueden formularse de la siguiente manera:

$H_0 : \beta_0 = I$	$H_0 : \beta_0 \leq I$	$H_0 : \beta_0 \geq I$
$H_1 : \beta_0 \neq I$	$H_1 : \beta_0 > I$	$H_1 : \beta_0 < I$
$R_\alpha = \{\lvert t_{\beta_0} \rvert > t_{(1-\frac{\alpha}{2}, n-2)}\}$	$R_\alpha = \{t_{\beta_0} > t_{(1-\alpha, n-2)}\}$	$R_\alpha = \{t_{\beta_0} < t_{(\alpha, n-2)}\}$

Nota: Se rechaza cuando se cumple la condición

Ejercicio - Dócimas de Hipótesis en Regresión

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

- "La producción aumenta al doble con respecto a la temperatura". Verifique esta afirmación considerando un $\alpha = 0{,}05$

Análisis de Residuales (Supuestos)

Definimos un residual como la diferencia entre el valor observado y el valor estimado:

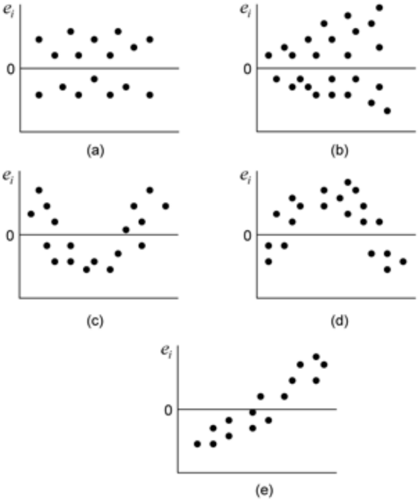
$$e_i = y_i - \hat{y}_i \tag{33}$$

Para realizar el modelo de regresión que estudiamos, consideramos los siguientes supuestos:

Los Errores Aleatorios (ϵ) siguen una distribución Normal:

- $E(\epsilon_i) = 0$
- $VAR(\epsilon_i) = \sigma^2$
- $COV(\epsilon_i, \epsilon_j) = 0, \forall i, j : i \neq j$

Análisis de Residuales (Supuestos)



Análisis de Residuales (Supuestos)

Análisis de Residuales (Supuestos)

El análisis de los residuales nos permite verificar si efectivamente se cumplen los supuestos de:

- La relación entre las variables X e Y es lineal
- Los términos de error tienen varianza constante
- **Los errores son independientes** $\rightarrow$ Test de Durbin-Watson
- **Los errores se distribuyen de forma normal** $\rightarrow$ Test de Kolmogorov-Smirnov

Durbin-Watson

Es un test estadístico utilizado para verificar la presencia de correlación en los residuos de una regresión.

Las hipótesis del Test son:

- $H_0 : \rho = 0$ (no hay correlación $\rightarrow$ son independientes)
- $H_1 : \rho > 0$ (hay correlación)

Durbin-Watson

Es un test estadístico utilizado para verificar la presencia de correlación en los residuos de una regresión.

Las hipótesis del Test son:

- $H_0 : \rho = 0$ (no hay correlación $\rightarrow$ son independientes)
- $H_1 : \rho > 0$ (hay correlación)

El estadístico se define como:

$$D = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{34}$$

donde n es el total de casos observados.
Para determinar la correlación:

- $D > d_u$ se acepta H_0: No están correlacionados
- $D < d_l$ se rechaza H_0: Los errores están correlacionados
- $d_l \leq D \leq d_u$ el test no concluye nada.

Durbin-Watson

Tabla para determinar valores:

	k'*=1		k'=2	
n	dL	dU	dL	dU
6	0.610	1.400	-----	-----
7	0.700	1.356	0.467	1.896
8	0.763	1.332	0.559	1.777
9	0.824	1.320	0.629	1.699
10	0.879	1.320	0.697	1.641
11	0.927	1.324	0.758	1.604

donde k es el número de parámetros estimados (sin el intercepto), n el número de casos considerados, con un nivel de significancia de 0.05.

Ejercicio - Durbin-Watson

Se realizó un estudio sobre los efectos de la temperatura (X) sobre la producción (Y) de un proceso químico. Se recopilaron los siguientes datos:

X	-5	-4	-3	-2	-1	0	1	2	3	4	5
Y	1	5	4	7	10	8	9	13	14	13	18

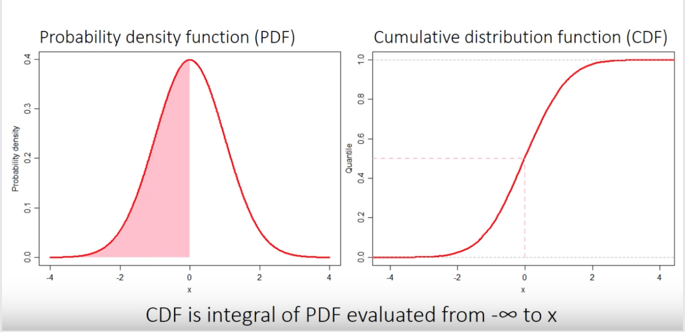
- Verificar si los errores aleatorios son independientes.

Kolmogorov-Smirnov

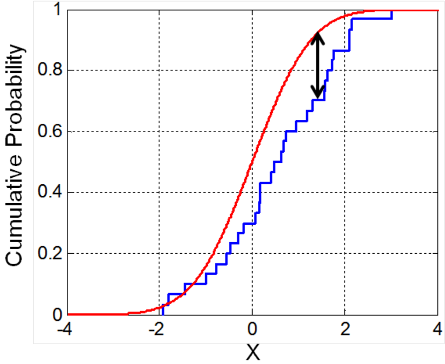
Kolmogorov-Smirnov

Test que compara la función de distribución acumulada *observada* ($F_n(x)$) de una variable con una distribución teórica determinada ($F_0(x)$).

Aquí, se utiliza para determinar si los residuos siguen una distribución normal.



Kolmogorov-Smirnov



Sus hipótesis son:

$H_0 : F_0(x) = F_n(x)$ de una $N(\mu, \sigma)$

$H_1 : F_0(x) \neq F_n(x)$ de una $N(\mu, \sigma)$

- Sea $X = x_1, x_2, ..., x_n$ una muestra aleatoria (determinada por F(x))
- Sea $X' = x_1', x_2', ..., x_n'$ la muestra ordenada

Kolmogorov-Smirnov

Kolmogorov-Smirnov

- Se define $S_n(x)$ la función de distribución obtenida en la muestra:
  - $S_n(x) = 0$ si $x < x_1$
  - $S_n(x) = \frac{k}{n}$ si $x_k \geq x \geq x_{k+1}$
  - $S_n(x) = 1$ si $x \leq x_n$
- El estadístico se obtiene a partir de:

$$|D| = \max_{1 \leq i \leq n} |F_n(x_i) - F_0(x_i)| \tag{35}$$

  donde $|D|$ es la máxima diferencia entre la frecuencia acumulada observada $F_n(x_i)$ y la frecuencia acumulada teórica $F_0(x_i)$
- Para obtener $F_n(x_i)$, se considera

$$z = \frac{x - \bar{x}}{s} \tag{36}$$

- Se rechaza H_0 si $|D| > D_n^{\alpha}$, donde D_n^{α} para $\alpha = 0,05$

Kolmogorov-Smirnov

Number of trials, n	Level of significance, α			
	0.10	0.05	0.02	0.01
1	0.95000	0.97500	0.99000	0.99500
2	0.77639	0.84189	0.90000	0.92929
3	0.63604	0.70760	0.78456	0.82900
4	0.56522	0.62394	0.68887	0.73424
5	0.50945	0.56328	0.62718	0.66853
6	0.46799	0.51926	0.57741	0.61661
7	0.43607	0.48342	0.53844	0.57581
8	0.40962	0.45427	0.50654	0.54179
9	0.38746	0.43001	0.47960	0.51332
10	0.36866	0.40925	0.45662	0.48893
11	0.35242	0.39122	0.43670	0.46770
12	0.33815	0.37543	0.41918	0.44905
13	0.32549	0.36143	0.40362	0.43247
14	0.31417	0.34890	0.38970	0.41762
15	0.30397	0.33760	0.37713	0.40420
16	0.29472	0.32733	0.36571	0.39201
17	0.28627	0.31796	0.35528	0.38086
18	0.27851	0.30936	0.34569	0.37062
19	0.27136	0.30143	0.33685	0.36117
20	0.26473	0.29408	0.32866	0.35241

Kolmogorov-Smirnov: Ejemplo

Suponga que cuenta con los siguientes residuos obtenidos en una regresión:

6.0	**2.3**	**4.8**	**5.6**	**4.4**	**3.4**	**3.3**	**1.9**	**4.6**	**4.5**

Kolmogorov-Smirnov: Ejemplo

Suponga que cuenta con los siguientes residuos obtenidos en una regresión:

6.0	**2.3**	**4.8**	**5.6**	**4.4**	**3.4**	**3.3**	**1.9**	**4.6**	**4.5**

Ordenando la muestra se obtiene:

1.9	**2.3**	**3.3**	**3.4**	**4.4**	**4.5**	**4.6**	**4.8**	**5.6**	**6.0**

Con $\bar{x} = 4{,}08$ y $s = 1{,}33$ (de los residuos).

Kolmogorov-Smirnov: Ejemplo

Datos	**Orden**	$F_0(x_i)$	**Z**	$F_n(x_i)$	$\|D\|$
1.9	1	0.1	$\frac{1,9-\bar{x}}{s}$ = -1.663	0.051	0.048
2.3	2	0.2	-1.332	0.091	**0.1082**
3.3	3	0.3	-0.583	0.281	0.019
3.4	4	0.4	-0.509	0.305	0.095
4.4	5	0.5	0.239	0.591	0.091
4.5	6	0.6	0.314	0.621	0.021
4.6	7	0.7	0.389	0.648	0.052
4.8	8	0.8	0.539	0.701	0.098
5.6	9	0.9	1.113	0.870	0.029
6.0	10	1.0	1.437	0.923	0.076

Kolmogorov-Smirnov: Ejemplo

En este caso $|D| = 0.116$ y tomando el valor de D_n^α:

n	D_n^α
>50	$\frac{1,36}{\sqrt{n}}$
20	0,29
15	0,34
10	**0.41**

La hipótesis nula se rechaza si $0{,}1082 > 0{,}41$. Por ende, $F_s(x) = F(x)$ con $N(\mu, \sigma)$.