# Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

Abril 2021



#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

\_\_\_\_\_

in la muestra

1 el Muestreo étodo de mínimos

adrados oría de Muestre

(TM) sobre los parámetros

La importancia de lo Residual

Análisis visua los residuos

Anexo sobre cálculo de lo

Método analític para más de dos variables

olución altern

# 10.- Regresión lineal simple. Efectos fijos (I)

# Fundamentos en Estadística

V. Trujillo

### (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

#### Introducción

En la Població

#### En la muestra

En el Muestreo

#### todo de mínimos adrados oría de Muestreo

l'eoría de Muestreo TM) sobre los parámetros

#### La importancia de lo Residual

Análisis visual de los residuos

#### Anexo sobre cálculo de los estimadores

ara más de dos ariables

Solución alternati

# Introducción

La regresión lineal simple, múltiple, el análisis de varianza (ANOVA), el de covarianza (ANCOVA) etc. surgen del Modelo Lineal General (GLM) y por reducción de los modelos aditivos generales (GAM).

Modelo Lineal General:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$

#### Fundamentos en Estadística

V. Trujillo

#### RC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

n la Poblacioi

in la muestra

étodo de mínimos

idrados oría de Muestreo M) sobre los

La importancia le lo Residual

Análisis visual ( los residuos

Anexo sobre cálculo de los estimadores

Método analític para más de do: variables

olución alternati

La regresión se podría definir como: el estudio de la **relación** que existe entre las variables explicativas y las variables respuestas. En el caso de la regresión lineal, se basa en un modelo lineal y si además es simple, sólo existirá una variable explicativa y una variable respuesta, mal llamadas independiente y dependiente respectivamente.

La regresión (relación) y el ANOVA (diferenciación) son caras de la misma moneda. Son las técnicas de análisis mayoritariamente usadas en estadística, sobre todo la regresión y la correlación.

#### Fundamentos en Estadística

V. Trujillo

### (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

n la Poblacio

n la muestra

### En el Muestreo

odo de mínimos drados

eoría de Muestreo ΓM) sobre los

#### a importancia le lo Residual

nálisis visual de

#### Anexo sobre cálculo de los estimadores

létodo analític ara más de do: ariables

lución alternativ

# En la Población

En la regresión lineal, las diferentes posibilidades de y son  $y_i$ , para un  $x_i$  hay diferentes  $y_i$ 

Las medias de los diferentes  $y_i$  para cada x están <u>sobre</u> una recta.  $Y = media \ de \ las \ y \ para \ cada \ x$ , definida por la ecuación:

Y = A + BX

#### Fundamentos en Estadística

7. Trujillo

RC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

.....

n el Muestreo

étodo de mínimos adrados

Teoría de Muestreo (TM) sobre los

La importancia de lo Residual

Análisis visual de los residuos

Anexo sobre cálculo de los estimadores

> ara más de dos triables

Para cada valor de x la media de todos los  $y_i$  están sobre una recta, con dos suposiciones principales:

1. Para cada  $x_i$  las  $y_i$  están <u>distribuidas normalmente</u>

$$Y \cap N(\mu, \sigma^2)$$

 $\mu = media \ de \ las \ y \ para \ cada \ x \Rightarrow \mu = Y$ 

$$Y \cap N(A + BX, \sigma^2)$$
 (1)

2. Las varianzas de todos los  $y_i$  son <u>constantes</u>

$$\sigma^2 = cte$$
 (2)

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

Introduccion

'n le muestre

n el Muestreo

reoría de Muestreo (TM) sobre los

La importancia de lo Residual

> nálisis visual d s residuos

Anexo sobre cálculo de los estimadores

> ira mas de dos riables

Solución altern

Pero esto mismo es muy importante verlo bajo el "prisma" de los residuos. Otra aproximación al modelo podría ser, que para cada  $x_i$ , valor fijo de X, se cumple la ecuación:

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

donde  $\beta_0$  y  $\beta_1$  son <u>constantes</u> desconocidas.

### Fundamentos en Estadística

V. Trujillo

# (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

### introducción

n la Población

En la muestra

# n el Muestreo létodo de mínimos

uadrados 'eoría de Muestreo TM) sobre los

#### a importancia e lo Residual

Análisis visual de

#### Anexo sobre cálculo de los estimadores

ara más de do ariables

olución alterna

# Las hipótesis básicas del modelo son:

1. **Independencia en los residuos**. Sin correlación (auto-correlación) de los residuos

$$Cor(e_i, e_j) = 0$$

Cualquier par de errores  $e_i$  y  $e_j$  son independientes.

2. Esperanza nula de los residuos es:

$$E(e_i) = 0$$

3. Varianza constante de los residuos (homocedasticidad):

$$Var(e_i) = \sigma^2$$

4. Normalidad de los residuos:

$$e_i \sim N(0, \sigma^2)$$

### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroduccion

n la muestra

el Muestreo

adrados eoría de Muestreo 'M) sobre los

a importancia e lo Residual

Análisis visual de os residuos

Anexo sobre cálculo de los estimadores Método analítico

olución alterna

# De lo que se deduce:

ightharpoonup Cada valor  $y_i$  de la variable aleatoria Y tiene una distribución

$$(Y \mid X = x_i) \approx N(\beta_0 + \beta_1 x_i, \ \sigma^2)$$

Las observaciones  $y_i$  de la variable Y son independientes.

### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

#### Introducción

\_

### En el Muestreo

étodo de mínimos adrados eoría de Muestreo

#### la importancia le lo Residual

Análisis visual de los residuos

#### Anexo sobre cálculo de los estimadores

Método analític para más de dos variables

Solución alternati

Si las hipótesis del modelo son ciertas, gráficamente se tiene que:

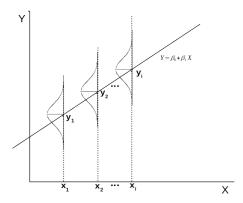


Figura 1: Modelo gráfico de la recta de regresión.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

en la Poblacioi

En la muestra

En el Muestreo

adrados coría de Muestreo

Teoría de Muestre (TM) sobre los parámetros

La importancia de lo Residual

Análisis visual d los residuos

Anexo sobre cálculo de los estimadores

Método analítico para más de do variables

lución alternativa

# En la muestra

### Criterio de selección

Las X <u>no</u> tienen criterio probabilístico, se decide arbitrariamente (por el investigador) cuánto vale. Por eso, a veces se le llama "media variable o semi-variable". Por lo tanto, X <u>NO</u> es una variable aleatoria, no tiene una probabilidad asociada. Su selección es arbitraria, tiene cierta libertad, pero una vez seleccionada las  $x_i$ , las  $y_i$  <u>SÍ</u> que se seleccionan de forma aleatoria.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

.

in la muestra

En el Muestreo

adrados oría de Muestreo

'eoría de Muestre TM) sobre los arámetros

La importancia de lo Residual

Análisis visual de os residuos

Anexo sobre cálculo de los estimadores

> létodo analítico ara más de dos ariables

olución alternati

### Fundamentos en Estadística

7. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

#### Introducción

Eli la i oblacio

ı la muestra

# En el Muestreo

odo de minime drados

l'eoría de Muestreo TM) sobre los parámetros

#### La importancia le lo Residual

nálisis visual de

### Anexo sobre cálculo de los

Método analítico para más de do variables

olución alternati

El caso más sencillo, para cada x, es que la muestra tenga un tamaño 1. Lo que quiere decir: que para cada x sólo se escoge una y de todas las posibles  $y_s$ .

Generalmente, el criterio de selección suele ser que: para cada valor arbitrario de  $x_i$  elijo un  $y_i$  aleatorio simple (a.s.), con lo cual la muestra total tiene n pares de valores.

A la X se le llamaba variable independiente, pero es un nombre antiguo y lo correcto, en este contexto, es llamarle: variable auxiliar, variable regresora o en general, **variable explicativa**.

# En el Muestreo

En el mundo del Muestreo (no confundir con el hecho de muestrear) es fundamental saber qué <u>inferencia</u> se pretende realizar; i.e. responder a: ¿Cuál es el objetivo que se quiere conocer?

En la Población para cada valor de X, la Y (media de y para cada x) está sobre una recta: es una recta. La distribución de Y para cada X es una normal con  $\mu = A + BX$  y con la misma varianza  $\sigma^2$ . También lo podemos expresar como  $y \cap (Y, \sigma^2)$ . Es decir:

$$Y = A + BX$$

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

En el Muestreo Método de mínimos

coría de Muestreo TM) sobre los arámetros

La importancia de lo Residual

Análisis visual los residuos

Anexo sobre cálculo de los estimadores

Método analític para más de do: variables

olución alternati

# Fundamentos en Estadística

Trujillo

# (CO de Vigo. IEO)

0.- Regresión neal simple.

ntroducción

ii ia robiacio

In la muestra

En el Muestreo Método de mínimos cuadrados

ruadrados Feoría de Muest TM) sobre los

a importancia e lo Residual

nálisis visual o

Anexo sobre cálculo de los estimadores

riables

Para resolver la ecuación anterior, ¿Cuántos parámetros se tienen que calcular?

Se tiene  $A y B y \cdots \sigma^2$  (común para todos los x) e... $Y_{(media)}$  que también lo desconozco, pero si se conoce A y B ya se puede calcular.

Con lo cual el objetivo que se pretende es: que a partir de la muestra, <u>estimar</u> A, B y  $\sigma^2$  para posteriormente calcular la Y.

Es a partir de este punto cuando ya se pueden hacer pruebas de hipótesis.

# Estimación de $A, B, \sigma^2$ e Y

Habra que estimar estos tres parámetros: a,b e  $\hat{Y}$ . Estimadores puntuales de A,B e Y respectivamente. Obviamente, los importantes serán los dos primeros  $(a \ y \ b)$ .

Observación		Pretensión	Estimación	
х	у	Y	$\widehat{V}$	
X <sub>1</sub>	<b>y</b> <sub>1</sub>	A + B x <sub>1</sub>	a + b x <sub>1</sub>	
X <sub>2</sub>	y <sub>2</sub>	A + B x <sub>2</sub>	a + b x <sub>2</sub>	
:	:	:	:	
X,	y <sub>i</sub>	A + B x <sub>i</sub>	a + b x <sub>i</sub>	
:	i i	i	ŧ	
x <sub>n</sub>	y <sub>n</sub>	A + B x <sub>n</sub>	a + b x <sub>n</sub>	

Figura 2: Tabla de regresión.

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

in la Pobla

en ia muestra

Método de mínimos cuadrados Teoría de Muestreo

Teoria de Muestre (TM) sobre los parámetros

de lo Residual

Análisis visual los residuos

Anexo sobre cálculo de los estimadores

> Método analítico para más de do pariables

olución alternati

# Método de mínimos cuadrados

Este método de estimación es **uno** de los más habitualmente utilizados. Concebido para el cálculo de la estimación puntual de a y b, basado en reducir (minimizar) la distancia al cuadrado entre el y de la muestra menos el estimador  $(\hat{Y})$  de y:

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

esa diferencia de valores se llama  $\underline{residuo}$  y por tanto  $S^2$  es la suma de los cuadrados de los residuos. Que también se puede expresar como:

$$S^{2} = \sum [y_{i} - (a + bx_{i})]^{2}$$

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

Introducción

n la Població

n el Mue

Teoría de Muestreo TM) sobre los

La importancia de lo Residual

nálisis visual de s residuos

Anexo sobre cálculo de los estimadores

> ra más de dos riables

Solución alternativ

Para calcular a y b, simplemente se calculan las derivadas parciales de la suma de cuadrados de los residuos con respecto a cada estimador para que estas derivadas sean igual a cero, ya que lo que se busca es minimizar esa función para cada parámetro.

Este método sencillo de ajuste se utiliza en muchas ocasiones en estadística, ya que los estimadores que genera son:

- No sesgados
- ► Eficientes(Varianza pequeña en el Muestreo)
- ightharpoonup Consistentes. Si sesgados,  $\uparrow n$  tiende a ser insesgado

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroducción

n la Poblacio

n ia muestr

n el Mu

(TM) sobre los parámetros

La importancia le lo Residual

Análisis visual os residuos

Anexo sobre cálculo de los estimadores

> létodo analítico ara más de dos ariables

> > lución alterna

En resumen, para la estimación de los parámetros lo que se busca es minimizar la suma de cuadrados de los residuos  $(y_i - \hat{Y}_i)^2$  en el ajuste. También se puede expresar como:  $S = S_{residual}^2$ .

Otra forma de verlo es: que lo que se busca es reducir la diferencia entre lo que se observa  $(y_i)$  y lo que se estima  $(\hat{Y})$  o espera obtener a partir del ajuste.

Nota: Cuidado con la notación usada para las sumas de cuadrados porque a veces es confusa y depende de los autores y/o de su ubicación formal.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

n la Població

n la muestra

létodo de míni iadrados

parámetros La importancia

de lo Residual

Análisis visual os residuos

Anexo sobre cálculo de los estimadores

> Método analítico ara más de dos ariables

> > lución alternat

# Cálculo de los estimadores por mínimos cuadrados

Para el cálculo de los parámetros de la regresión lo primero que se necesita es recordar cómo se calcula la suma de cuadrados para x e y y la suma de cuadrados del producto  $x \cdot y$   $(S_{xy})$ :

$$S_{xx} = \sum (x_i)^2 - \bar{x}(\sum x_i)$$

$$S_{yy} = \sum (y_i)^2 - \bar{y}(\sum y_i)$$

$$S_{xy} = \sum x_i \cdot y_i - \bar{x}(\sum y_i)$$

$$o$$

$$S_{xy} = \sum x_i \cdot y_i - \bar{y}(\sum x_i)$$

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

O.- Regresión ineal simple. Efectos fijos (I)

ntroducción

11 10 1 0010010.

\_ - - -

cuadrados Teoría de Muestreo (TM) sobre los

a importancia le lo Residual

Análisis visua los residuos

Anexo sobre cálculo de los estimadores

> létodo analític ara más de dos ariables

> > lución alterna:

Una vez obtenidos las suma de cuadrados de las variables y obviamente sus medias aritméticas, ya se pueden calcular todas las estimaciones puntuales necesarias sabiendo que:

# 1. Estimador de **b**

$$b = \frac{S_{xy}}{S_{xx}}$$

2. Estimador de a

$$a = \bar{y} - b \ \bar{x}$$

3. Estimador de  $\hat{\mathbf{Y}}$ 

$$\hat{Y} = a + b X$$

### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

ineal simple. Efectos fijos (I

ntroducción

la muestra

n el Muest

eoría de Muestreo ΓM) sobre los arámetros

la importancia le lo Residual

Análisis visual ( los residuos

Anexo sobre cálculo de los estimadores

> ira más de dos riables

Solución alterna

# 4. Estimador para $\sigma^2 \to s^2$

$$s^2 = \frac{S_{residual}^2}{n-2}$$

Donde la  $S_{residual}^2$  es:

$$S_{residual}^{2} = S_{yy} - \frac{S_{xy}^{2}}{S_{xx}}$$

$$o$$

$$S_{residual}^{2} = S_{yy} - b^{2}S_{xx}$$

Nota: Cuidado con la notación, ya que puede aparecer tanto  $S_{xx/yy/xy}\ como\ S^2_{x/y/xy}$ 

#### Fundamentos en Estadística

V. Trujillo

### GRC-MERVEX (CO de Vigo.

lineal simple. Efectos fijos (I

ntroducción

En al Musatros

Teoría de Muestreo (TM) sobre los parámetros

La importancia de lo Residual

Análisis visua los residuos

Anexo sobre cálculo de los estimadores

Método analític para más de dos variables

olución alternati

# Teoría de Muestreo (TM) sobre los parámetros

Los estimadores de nuestros parámetros son:

Parámetros	A	В	$\sigma^2$	Y
Estimadores	a	b	$s^2$	$\hat{Y}$

Y la TM, ¿Qué es lo que dice sobre cada uno de estos estimadores?

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroducción

n la Població

n el Muestre

Método de mínimos

a importancia e lo Residual

s residuos

Anexo sobre cálculo de los estimadores

Solución altern

# 1. Estimador de a

Esperanza

Varianza

► Error

$$E[a] = A$$

 $V[a] = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]$ 

 $var[a] = s^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{m-1}} \right]$ 

 $Error[a] = \sqrt{V[a]}$ 

 $error[a] = \sqrt{var[a]}$ 

Comportamiento (Distribución)

 $a \sim N(E, V) \rightarrow Si \sigma^2 desconocida: a \sim t_{\nu}$ 

 $\rightarrow no \ sesgado$ 

Fundamentos en Estadística

Anexo sobre

 $\rightarrow \sigma^2 desconocida$ 

 $\rightarrow \sigma^2$  conocida

 $\rightarrow \sigma^2 desconocida$ 

 $\rightarrow \sigma^2$  conocida

# 2. Estimador de b

Esperanza

$$E[b] = B$$

 $\rightarrow no \ sesgado$ 

Varianza

$$V[b] = \frac{\sigma^2}{S_{xx}}$$

 $\rightarrow \sigma^2$  conocida

$$var[b] = \frac{s^2}{S_{xx}}$$

 $\rightarrow \sigma^2$  conocida

 $\rightarrow \sigma^2 desconocida$ 

Error

$$error[b] = \sqrt{var[b]} \rightarrow \sigma^2 \ desconocida$$

 $Error[b] = \sqrt{V[b]}$ 

Si  $\sigma^2$  desconocida:  $b \sim t_{\nu}$ 

en Estadística

Fundamentos

Anexo sobre

Comportamiento  $b \sim N(E, V) \rightarrow$ 

# 3. Estimador de $\hat{Y}$

Esperanza

$$E[\hat{Y}] = Y = A + B \ X \qquad \rightarrow no \ sesgado$$

Varianza

$$V[\hat{Y}] = \sigma^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \rightarrow \sigma^2 \ conocida$$

$$var[\hat{Y}] = s^2 \left[ \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] \to \sigma^2 \ desconocida$$

► Error

$$Error[\hat{Y}] = \sqrt{V[\hat{Y}]}$$
  $\rightarrow \sigma^2 \ conocida$   $error[\hat{Y}] = \sqrt{var[\hat{Y}]}$   $\rightarrow \sigma^2 \ desconocida$ 

► Comportamiento

$$\hat{Y} \sim N(E,V)$$

Fundamentos en Estadística

RC-MERVEX

(CO de Vigo. IEO)

efectos fijos (I

ntroduccion

In la muestra

En el Muesti Método de mínic cuadrados

La importancia de lo Residual

Anexo sobre cálculo de los estimadores

Método analíti para más de do variables

# 4. Estimador $s^2$

Esperanza

$$E[s^2] = \sigma^2$$

- ➤ Varianza/Error (ahora no es relevante)
- Comportamiento

$$\frac{S_{residual}^2}{\sigma^2} \sim \chi_{(n-2)}^2$$

Nota: Recordad la ecuación para la varianza residual.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

Efectos fijos (I

ntroduccion

7-- 1- -----

n el Muestreo

todo de mínin idrados

La importancia de lo Residual

Análisis visual de los residuos

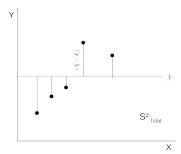
Anexo sobre cálculo de los estimadores

Solución altern

# La importancia de lo Residual

La suma de cuadrados de la "variación" total, no es más que la suma de las diferencias (distancias o dispersiones) cuadráticas de cada valor de la variable respuesta (y) y su media aritmética  $(\bar{y})$ . Se puede expresar, a partir de las **diferencias**, como:

$$(y - \bar{y}): S_{yy} = S_{Total}^2$$



Suma de cuadrados Total.

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

En la Població

n la muestra

En el Muestreo Método de mínimos

ruadrados Teoría de Muestreo (TM) sobre los

La importancia de lo Residual

Análisis visual los residuos

Anexo sobre cálculo de los estimadores

Método analíti para más de de variables

Solución alternati

Para cada valor de  $x_i$  hay un punto y sobre la recta horizontal de y, que no es más que la media aritmética de y:  $\bar{y}$ . Por lo tanto, cada segmento representa la diferencia de  $(y_i - \bar{y})$ .

La  $S_{Total}^2$  da en general una medida básica y directa de la dispersión de los puntos. En el caso de la regresión, de la dispersión de los puntos de la variable respuesta.

Precisamente, esto es lo que se quiere conocer y lo que da sentido al modelo de regresión. Saber cómo es la variación de la variable respuesta en relación (a la "influencia") de otra variable que por eso se llama variable explicativa. De ahí que se denomine **Suma de cuadrados total**.

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroducción

in la Poblaci

in la muest

# En el Muestreo

adrados

eoría de Muestreo ΓM) sobre los arámetros

#### La importancia le lo Residual

Análisis visual d os residuos

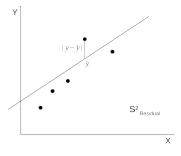
### nexo sobre álculo de los

létodo analític ara más de dos

lución alternativ

Integrando ahora la recta de regresión, se observan otros tipos de diferencias, como son las diferencias de cada valor de y con respecto a la estimación de  $\hat{y}$  derivada del modelo de regresión:  $(y-\hat{y})$ , a esta diferencia se le llama residuo. Por lo tanto, la **Suma de cuadrados residual**, es una medida de la variación residual, también llamada variación no explicada por la recta de regresión:

$$(y - \hat{y}): S_{yy} - \frac{S_{xy}^2}{S_{xx}} = S_{Residual}^2$$



Suma de cuadrados Residual.

Fundamentos en Estadística

V. Trujillo

(CO de Vigo. IEO)

ineal simple. Efectos fijos (l

Introduccion

ii ia i obiacio

En al Muaetra

étodo de mínimos adrados oría de Muestreo

(TM) sobre los parámetros

Análisis visual

Anexo sobre cálculo de los

Método analíti para más de do variables

olución alternat

# Fundamentos en Estadística

. Trujillo

(CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

1 10 1 0010010

n la muestra

n el Muestreo

uadrados 'eoría de Muest TM) sobre los

rámetros

a importancia e lo Residual

Análisis visual d os residuos

Anexo sobre cálculo de los estimadores

> ra más de dos riables

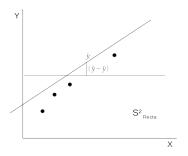
no explicada por la recta (es residual), es lo que resta del valor observado de la recta después de que se haya ajustado un modelo de regresión.

Dicho de otra manera, la variación residual es la variación

Resultando ser una herramienta muy poderosa a tener muy en cuenta porque da, entre otras cosas, una medida del ajuste.

La diferencia entre la variación total y la residual será la variación debida a la recta. Aquella parte de la variación que es *explicada* por la recta de regresión. Por lo tanto, la variación de la recta se mide a través de la **Suma de cuadrados de la recta**, que será:

$$(\hat{y} - \bar{y}) : \frac{S_{xy}^2}{S_{xx}} = S_{Recta}^2$$



Suma de cuadrados de la Recta.

Fundamentos en Estadística

V. Trujillo

(CO de Vigo. IEO)

neal simple. fectos fijos (I)

ntroduccion

n la muestra

En el Muestreo Método de mínimos cuadrados

(TM) sobre los parámetros

La importancia de lo Residual

Anexo sobre cálculo de los

Método analít para más de d variables

Solución alterr

En resumen y bajo la perspectiva de las desviaciones, tenemos que:

$$(y - \bar{y}) \equiv (y - \hat{y}) + (\hat{y} - \bar{y})$$

Las desviaciones totales para cada puntuación, son idénticas a la suma de las desviaciones hasta la recta y las desviaciones de la recta hasta la media.

Si se eleva al cuadrado y se suma cada desviación, se obtiene que la variación total es igual a la suma de la variación residual <u>más</u> la variación explicada por la recta.

$$S_{Total}^2 = S_{Residual}^2 + S_{Recta}^2 \label{eq:Stotal}$$

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

ii ia robiacio

la muestra

En el Muestreo

todo de minimo: idrados

eoría de Muestre FM) sobre los arámetros

a importancia e lo Residual

Análisis visual de os residuos

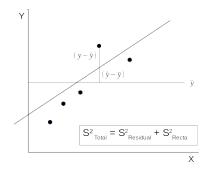
Anexo sobre cálculo de los estimadores

> fétodo analítico ara más de dos ariables

lución alterna

Esta es la descomposición fundamental de la variación para el modelo de regresión. También se expresa, como: la variación total es la suma de la variación no explicada (residual) y la variación explicada por la recta.

Evidentemente, cuánto mayor es la  $S^2_{Recta}$  menor será la  $S^2_{Residual}$  lo que implica que será mejor el ajuste a la recta.



Suma de cuadrados Total de la recta de regresión.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

In la Poblaci

En la muestr

En el Muestreo

adrados oría de Muestre

oarámetros

de lo Residua

los residuos

Anexo sobre cálculo de los estimadores

> Método analíticara más de do ariables

olución alternat

Si la relación anterior se divide por la  $S_{Total}^2$  queda:

$$1 = \frac{S_{Recta}^2}{S_{Total}^2} + \frac{S_{Residual}^2}{S_{Total}^2}$$

Con lo cual se obtiene en porcentaje (o tanto por uno), sobre cual es la importancia relativa que explica o no, la variación total de la recta de regresión.

La relación  $\frac{S_{Recta}^2}{S_{Total}^2}$  es el **coeficiente de determinación.** 

Nota: Realmente llamarlo coeficiente de determinación es un mal nombre. Sería más apropiado llamarlo: coeficiente de explicación.

Fundamentos en Estadística

V. Trujillo

RC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

itroducción

En la muestra

n el Muestreo

étodo de mínimo adrados

eoría de Muestre 'M) sobre los

a importancia

le lo Residual Análisis visual de

Anexo sobre

stimadores Método analític para más de do

Solución altern

El coeficiente de determinación (c.d.) es una medida de la bondad del ajuste, ya que indica el porcentaje de variación total que es explicado por la recta de regresión. Muchas veces se expresa como  $r^2$  y es una notación muy equívoca.

Esta denominación se suele, o se puede, relacionar erróneamente con el coeficiente de correlación.

La correlación considera que la variable explicativa es aleatoria (modelo de tipo II, o de efectos variables). Por lo tanto, asume aleatoriedad en ambas variables, lo que **no** ocurre con el modelo de regresión simple(modelo de tipo I, o de efectos fijos).

Fundamentos en Estadística

a represent

(CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

itroducción

ı la muestra

n el Muestre

do de mínim rados

eoría de Mues 'M) sobre los rámetros

importancia

lo Residual

Anexo sobre

álculo de los stimadores

ra más de dos riables

Nota: Este es un error conceptual muy extendido, donde la variable explicativa **no** es aleatoria.

# Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

£n la Població

En la muestra

En el Muestreo
Método de mínimos

cuadrados Teoría de Muestreo

a importancia

Análisis visual de los residuos

los residuos

Anexo sobre

cálculo de los estimadores

ariables Solución alternati

Con respecto a sus valores se puede interpretar que:

ightharpoonup Si  $cd|r^2=0$ 

La recta no explica nada

Si  $cd|r^2 = 1$  variación

La recta explica toda la

 $0 \le cd|r^2 \le 1$  variación

La recta explica <u>parte</u> de la

# Como resumen, se puede decir que el coeficiente de determinación, sirve para:

► Medida de la bondad de ajuste.

Cuan cercanos son los puntos a la recta.

Medida del grado de linealidad.

"Acercamiento" residual.

▶ Medida directa de la mejora en el modelo.

Mejora al introducir información al modelo: más datos, mejores datos, una u otra variable, mejores variables etc.

#### Fundamentos en Estadística

V. Trujillo

## (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

in la muestra

En el Muestreo

adrados oría de Muestre

eoría de Muestreo ΓM) sobre los arámetros

La importancia le lo Residual

nálisis visual de s residuos

Anexo sobre cálculo de los estimadores

> létodo analític ara más de do: ariables

olución alternati

## Análisis visual de los residuos

El residuo en el modelo de regresión lineal simple se había definido como:

Residuo = Error = 
$$\hat{y_i} - y_i = \mathcal{E}_{ij} = \beta_0 + \beta_1 x_1 - y_i$$

Lo primero que hay que hacer es ver si se cumple:

- 1.  $E[\mathcal{E}_i] = 0$  Alea
  - Aleatoriamente y con Esperanza nula
- 2.  $cov(\mathcal{E}_i, \mathcal{E}_j) = 0; \quad \forall i \neq j \Rightarrow \nexists rachas$  Sin auto-covariación
- 3.  $Var[\mathcal{E}_i] = \sigma^2 = cte$  Homocedasticidad
- 4.  $\mathcal{E}_i \cap N(0, \sigma^2)$  Distribución normal de los residuos

#### Fundamentos en Estadística

V. Truiillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

i ia robiaciói

n la muestra

étodo de mínimos adrados

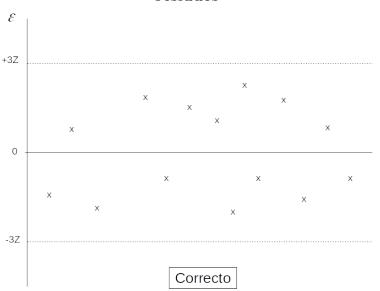
Ceoría de Muestreo TM) sobre los sarámetros

le lo Residual

Anexo sobre cálculo de los estimadores

> ara más de dos ariables

## Patrones de representación visual sobre los residuos



Representación correcta/esperable de los Residuos en un modelo de regresión.

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroducción

n ia robiacio

En la muestra

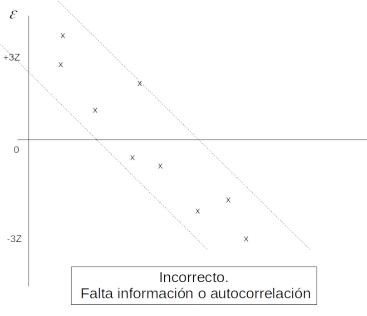
Método de mínimos cuadrados

Teoría de Muestreo (TM) sobre los parámetros

La importancia le lo Residual

Anexo sobre cálculo de los estimadores

> ara mas de dos ariables



Representación incorrecta de los Residuos en un modelo de regresión.

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

en la Poblacio

En la muestra

n el Muestreo

adrados oría de Muestreo

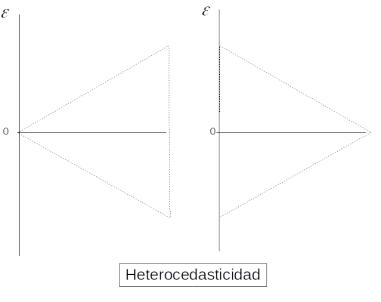
parámetros La importancia

de lo Residual

Anexo sobre cálculo de los estimadores

> létodo analítico ara más de dos ariables

> > lución alternativ



Representación con heterocedasticidad de los Residuos en un modelo de regresión.

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

En la Población

Eli la ilidestia

En el Muestreo Método de mínimos

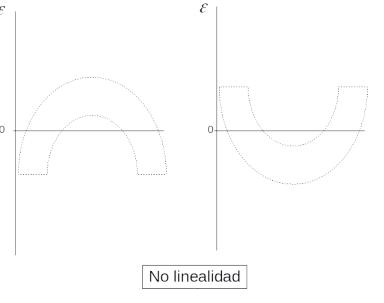
cuadrados Feoría de Muestreo (TM) sobre los

La importancia de lo Residual

Anexo sobre cálculo de los estimadores

> Método analít oara más de d ⁄ariables

> > olución alternativ



Representación con no-linealidad de los Residuos en un modelo de regresión.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

En la Población

Método de mínimos cuadrados

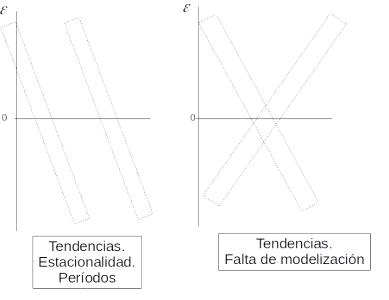
leoría de Muest TM) sobre los parámetros

La importancia de lo Residual

Anexo sobre cálculo de los estimadores

> Método analític oara más de do variables

> > olución alternati



Representación con tendencias de los Residuos en un modelo de regresión.

#### Fundamentos en Estadística

V. Trujillo

#### (CO de Vigo. IEO)

ineal simple. Efectos fijos (I

Introducción

Lii id i obidoi

En la muestra

En el Muestreo

Método de mínim cuadrados

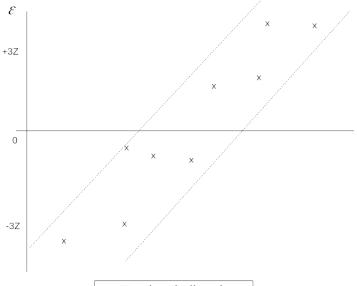
reoría de Muestreo TM) sobre los parámetros

La importancia de lo Residual

Anexo sobre cálculo de los estimadores

> Método analítico para más de do pariables

> > lución alternati



Tendencia lineal. Modelo muy incorrecto.

Representación con tendencias lineales de los Residuos en un modelo de regresión.

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo.

10.- Regresión lineal simple. Efectos fijos (I)

#### Introducción

en la Población

En la muestra

## En el Muestreo

adrados eoría de Muestreo 'M) sobre los

La importancia de lo Residual

#### Anexo sobre cálculo de los estimadores

Método analítico para más de dos variables

Solución alternativa

## Anexo sobre cálculo de los estimadores

Para el cálculo de los estimadores y sobre todo para el de dos variables, el método de estimación se puede hacer, como ya se indicó anteriormente, a través del cálculo de diferenciales parciales de la función objetivo (la suma de cuadrados de los residuos) con respecto a cada parámetro.

Como la regresión simple es una simplificación del modelo de regresión múltiple (con varias variables explicativas) y este a su vez del Modelo Lineal General (GLM, en inglés), los cálculos son matriciales. Pero el análisis matemático, también ofrece otro tipo de aproximaciones más "sencillas" para resolver este problema.

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión ineal simple. Efectos fijos (I)

ntroducción

n la Població

n la muestra

## En el Muestreo Método de mínimos

adrados

eoría de Muestreo FM) sobre los arámetros

#### La importancia de lo Residual

Análisis visua os residuos

Anexo sobre cálculo de los estimadores

> étodo analítico ara más de dos riables

lución alterna

## Método analítico para más de dos variables

Se genera un sistema de dos ecuaciones partiendo del modelo básico de regresión. A la primera ecuación: se le aplica un sumatorio; y a la segunda: se le multiplica por x y se aplica el sumatorio a la ecuación, quedando:

$$\begin{cases} 1: \sum y = n \cdot a + b \cdot \sum x \\ 2: \sum xy = a \cdot \sum x + b \cdot \sum x^2 \end{cases}$$

Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

ntroducción

in la Foblacio

n la muestra

En el Muestreo

Método de mínimos

adrados oría de Muestreo 'M) sobre los

a importancia

Análisis visual los residuos

Anexo sobr

Anexo sobre cálculo de lo estimadores

Nota: Muchas veces, por simplificación, se obvian los subíndices i en las ecuaciones.

Solución alternativa

1. Si a la primera ecuación la dividimos por n, queda:

$$\sum_{n} \frac{y}{n} = \frac{n \cdot a}{n} + \frac{b \cdot \sum x}{n} \Longrightarrow \bar{y} = a + b \ \bar{x} \Longrightarrow \boxed{a = \bar{y} - b \ \bar{x}}$$

2. Si a la segunda ecuación, ahora le restamos el término  $-\bar{y}\sum x$ , queda:

$$\sum xy - \bar{y} \sum x = a \sum x - \bar{y} \sum x + b \sum x^2 ; \quad Como \ a = \bar{y} - b\bar{x}$$

$$Sxy = (-b \ \bar{x}) \sum x + b \sum x^2$$

$$= b \ (\sum x^2 - \bar{x} \sum x) ; \quad Como \ (\sum x^2 - \bar{x} \sum x) = Sxx$$

$$= b \ Sxx \Longrightarrow \boxed{b = \frac{Sxy}{Sxx}}$$

#### Fundamentos en Estadística

V. Trujillo

(CO de Vigo.

10.- Regresión lineal simple. Efectos fijos (I)

Introducción

E-- 1- -----

En el Muestreo

Método de mínimos

Teoría de Muestreo (TM) sobre los parámetros

La importancia de lo Residual

Análisis visual d los residuos

Anexo sobre cálculo de lo El sistema de ecuaciones anterior, también se puede resolver aplicando la *Regla de Kramer*, quedando:

$$a = \frac{\left| \sum_{xy} y \sum_{x^2} x \right|}{\left| \sum_{xy} \sum_{x^2} x^2 \right|}$$

$$b = \frac{\left| \sum_{x} x \sum_{x^2} x \right|}{\left| \sum_{x} x \sum_{x^2} x \right|}$$

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

ineal simple. Efectos fijos (

Introducción

JII IO I ODIOCI

n la muestra

En el Muestreo

adrados

Teoría de Muestreo (TM) sobre los parámetros

La importancia de lo Residual

Análisis visual los residuos

Anexo sobre cálculo de los estimadores

Método analís para más de o variables

## Fundamentos en Estadística

Trujino

RC-MERVE (CO de Vigo.

0.- Regresión ineal simple.

troducción

En la Población

En la muestra

n el N

uadrados 'eoría de M TM) sobre

(TM) so parámet

La importancia de lo Residual

Análisis visual de los residuos

Anexo sobre

stimadores

Método analí para más de

## 11.- Correlación. Efectos variable (II)

#### Fundamentos en Estadística

V. Trujillo

#### GRC-MERVEX (CO de Vigo. IEO)

10.- Regresión lineal simple. Efectos fijos (I)

#### Introducción

En la Poblacio

En la muestra

En el Muestrec

cuadrados Feoría de Muestreo TM) sobre los

parámetros

de lo Residual

Análisis visual de

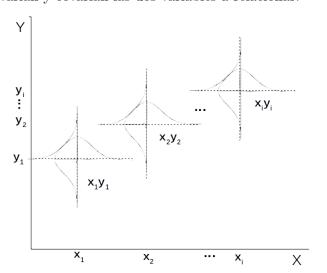
Análisis visual de los residuos

Anexo sobre cálculo de los estimadores

> ara más de dos ariables

Solución alternativ

La diferencia fundamental entre la regresión (modelo I) de efectos fijos y la correlación (modelo II) de efectos variables es que en este último los pares de valores son aleatorios, i.e. varían y covarían las dos variables a relacionar.



Correlación. Modelo II de efectos variables.

Fundamentos en Estadística

RC-MERVEX

).- Regresión

Introducción

En la 1 oblacio

n el Muestreo étodo de mínimos ladrados

importancia lo Residual

Análisis visual de los residuos

Anexo sobre cálculo de los estimadores

variables Solución alternati Se puede decir que la correlación mide la fuerza de la relación (covariación) entre un conjunto de variables cuantitativas continuas a través de un estadístico al efecto, que en el caso más sencillo, se denomina: coeficiente de correlación simple (r).

### Notas:

- ► Covariación no implica necesariamente causalidad.
- ▶ Ahora sólo hablaremos del coeficiente de correlación de Pearson.
- ▶ En el caso de más variables, se hablará de coeficiente de correlación múltiple y de coeficientes de correlación parciales.

#### Fundamentos en Estadística

V. Trujillo

GRC-MERVEX (CO de Vigo. IEO)

0.- Regresión ineal simple. Efectos fijos (I)

ntroducción

.

En el Muestre

Método de mínimo uadrados

eoría de Muestre ГМ) sobre los

a importancia e lo Residual

Análisis visual os residuos

Anexo sobre cálculo de los estimadores

> létodo analític ara más de do ariables

olución altern

La correlación no deja de ser una medida normalizada de la covariación entre dos variables, que está comprendida entre  $-1 \le r \le +1$ .

Se define como:

ightharpoonup varianza  $(s^2)$ 

$$s^{2} = \frac{S_{xx}}{n-1} = \frac{\sum (x_{i} - \bar{x})^{2}}{n-1} = \frac{\sum (x_{i} - \bar{x})(x_{i} - \bar{x})}{n-1}$$

ightharpoonup covarianza cov(x,y)

$$cov(x,y) = \frac{S_{xy}}{n-1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

ightharpoonup coeficiente de correlación de Pearson (r)

$$r = \frac{cov(x,y)}{s_x s_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Fundamentos en Estadística

RC-MERVEX

).- Regresión neal simple. fectos fijos (I)

ı la Poblaciór

n la muestra

létodo de mínimos nadrados eoría de Muestreo FM) sobre los

La importancia le lo Residual

Anexo sobre cálculo de los estimadores

> ara más de do: ariables



#### Fundamentos en Estadística

En la Población

Anexo sobre