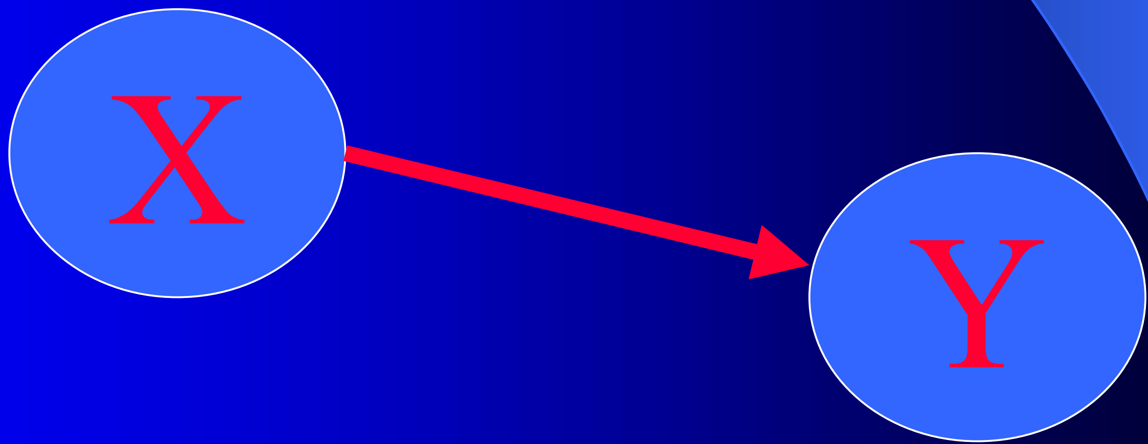
Regresión

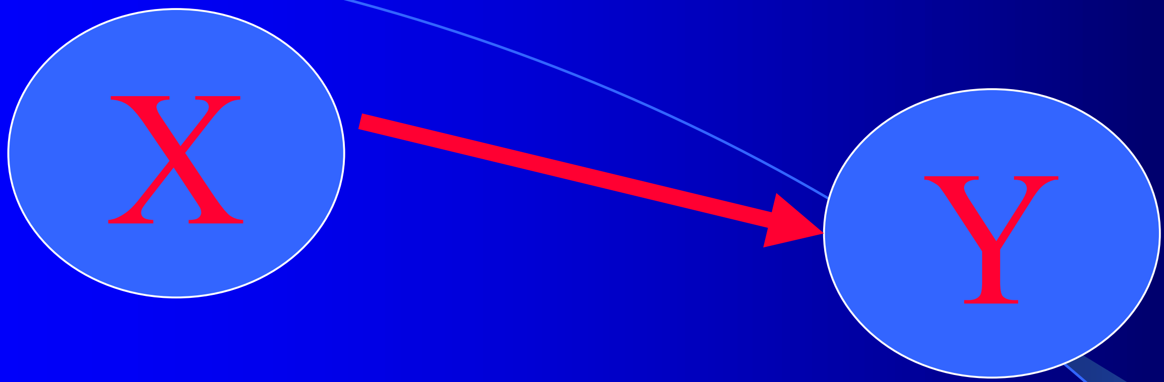
Vamos a estudiar:

- Relación entre variables.
- Cómo influye una variable X sobre otra variable Y

Vamos a estudiar:

- Relación entre variables.
- Cómo influye una variable X sobre otra variable Y





¿Para qué puede servir ésto a un jurista?

- Se puede estudiar cómo influye:
  - La renta media de los barrios (X) sobre el nivel de delincuencia (Y)
  - El número de casos entrantes en un juzgado (X) sobre el retraso medio en resolver los casos (Y)

Pensemos algunos ejemplos

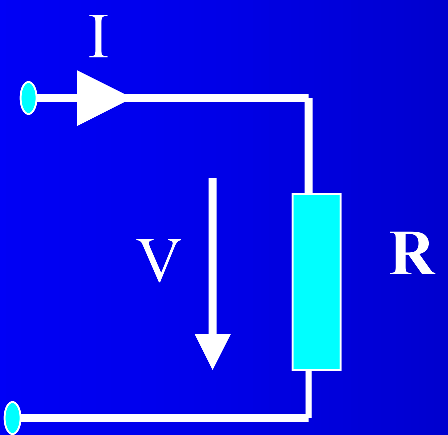
¿………………………?

Existen dos tipos de relaciones:

- **Deterministas:**
  - **Si conocemos el valor de X, el valor de Y queda perfectamente establecido.**
  - **Aparecen en ciencias.**
  - **Ejemplo:**
    - **Una resistencia de valor R ohmios.**
    - **La caída de tensión en sus bornes es: V=R.I siendo I la Intensidad en amperios**

Existen dos tipos de relaciones:

- **Deterministas:**
  - **Si conocemos el valor de X, el valor de Y queda perfectamente establecido.**
  - **Aparecen en ciencias.**
  - **Ejemplo:**
    - **Una resistencia de valor R ohmios.**
    - **La caída de tensión en sus bornes es: V=R.I siendo I la Intensidad en amperios**



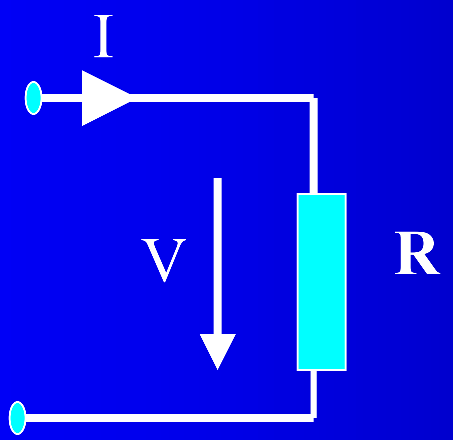
V=R.I

Existen dos tipos de relaciones:

- **Deterministas:**

  Si R=2 Ohmios

  - **Circulan 3 Amperios de intensidad**
    La caída de tensión será de 2.3=6 voltios.
  - **Circulan 4 Amperios de intensidad**
    La caída de tensión será de 2.4=8 voltios.



V=R.I

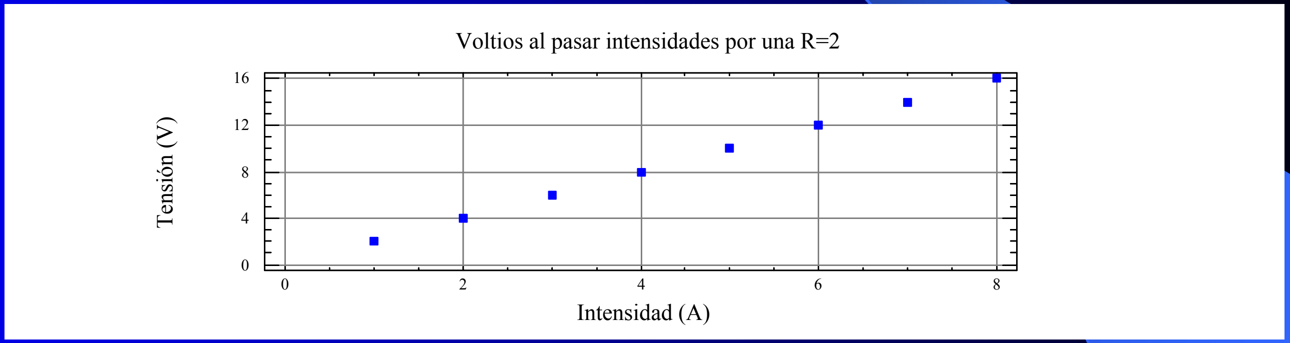
Existen dos tipos de relaciones:

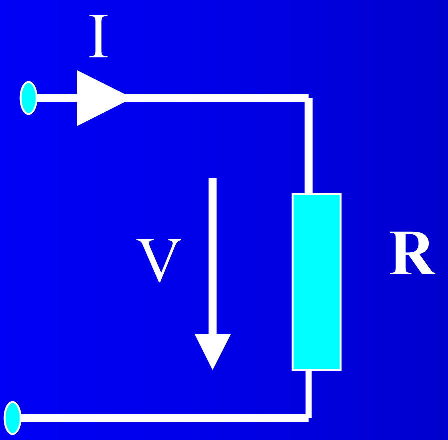
- **Deterministas:**

  Si R=2 Ohmios

  - **Circulan 3 Amperios de intensidad**
    La caída de tensión será de 2.3=6 voltios.
  - **Circulan 4 Amperios de intensidad**
    La caída de tensión será de 2.4=8 voltios.

SIEMPRE





V=R.I

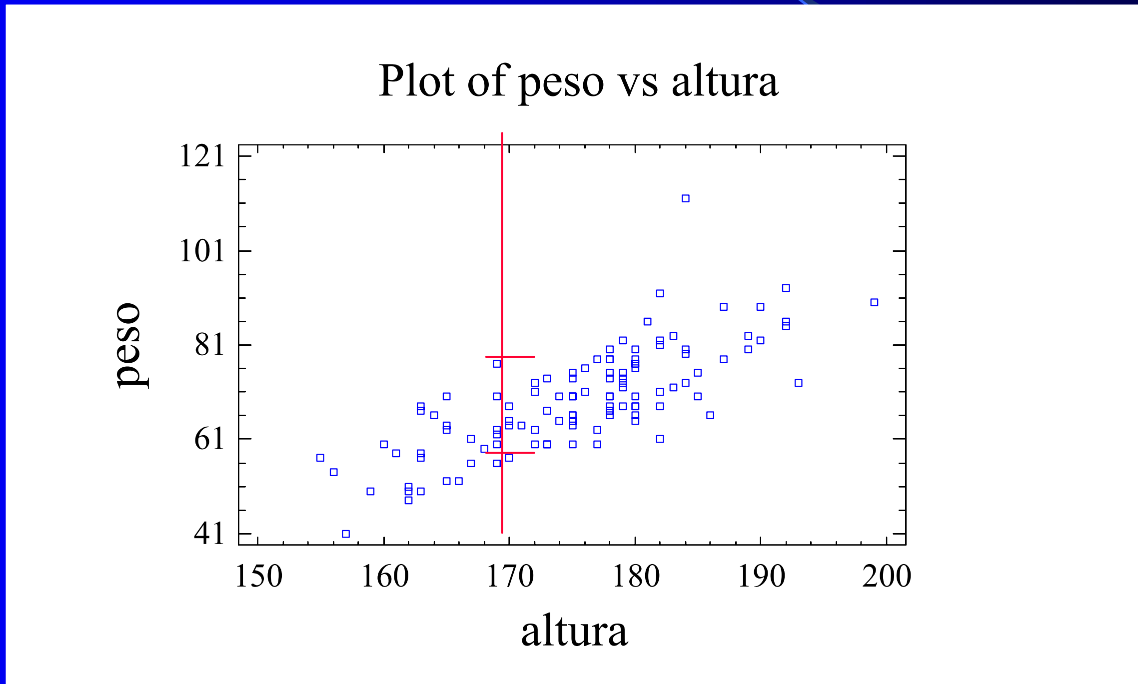
Relaciones deterministas

- SIEMPRE que x toma un valor........
- ....Y toma el mismo valor

Relaciones No Deterministas

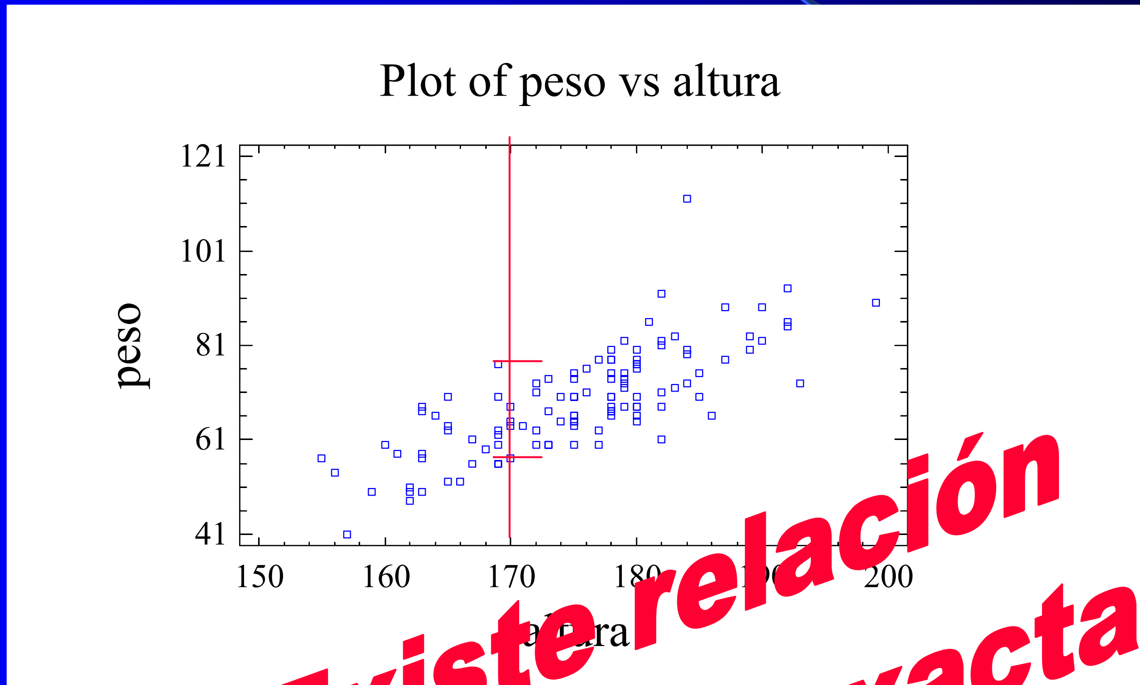
- **No Deterministas:**
  - **Si conocemos el valor de X, el valor de Y no queda perfectamente establecido. Hay una cierta variabilidad**
  - **Aparecen en ciencias, en ciencias sociales y en problemas de calidad.**
  - **Ejemplo:**
    - **Conocemos el peso y la altura de 117 estudiantes de ingeniería:**

Relaciones no deterministas



Si un estudiante mide 170cm su peso estará razonablemente entre 55 y 75 kg

Relaciones no deterministas



Existe relación

Pero no es exacta

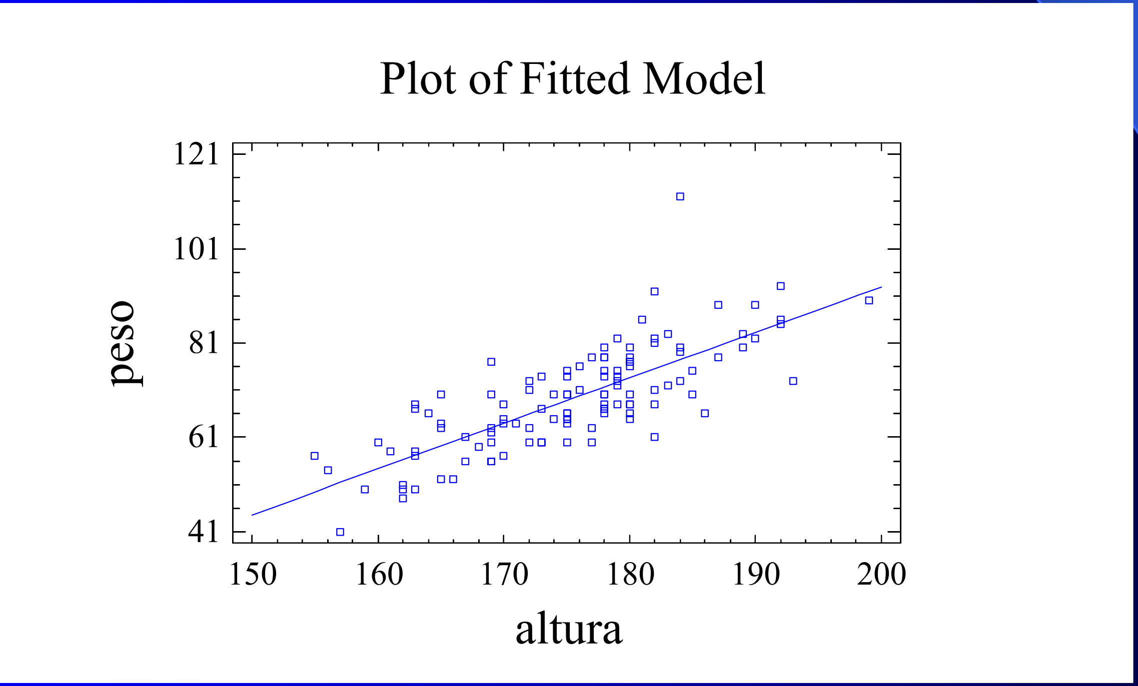
Si un estudiante mide 170cm su peso estará razonablemente entre 55 y 75 kg

Relaciones no deterministas

¿Ejemplos?

La regresión estudia las relaciones no deterministas

- **Ajusta una línea recta a la nube de puntos:**



El modelo de regresión hace básicamente eso: ajustar líneas a datos que sean razonablemente rectos.

La recta ajustada es la *recta de regresión* y explica la relación entre la variable Y (Peso) y la variable X (Altura).

El modelo de regresión hace básicamente eso: ajustar líneas a datos que sean razonablemente rectos.

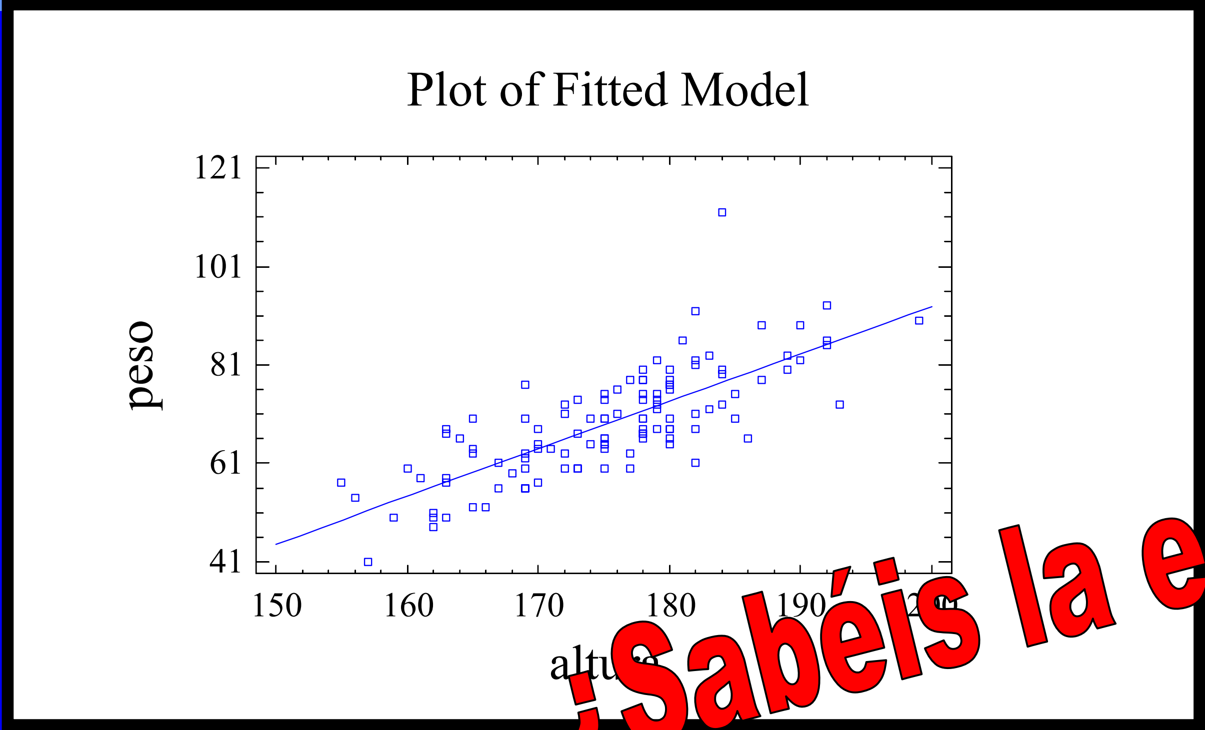
La recta ajustada es la *recta de regresión* y explica la relación entre la variable Y (Peso) y la variable X (Altura).

X → Y

Independiente
Explicativa

Dependiente
Respuesta
A explicar

La recta ajustada, ***que proporciona el ordenador***, es:



¿Sabéis la ecuación de una línea recta?

Peso = -100.22 + 0.97Altura

Ésto es un línea recta

Por si no se conoce....

$$Y = 5 + 2X$$

Pintando los puntos en el gráfico X-Y

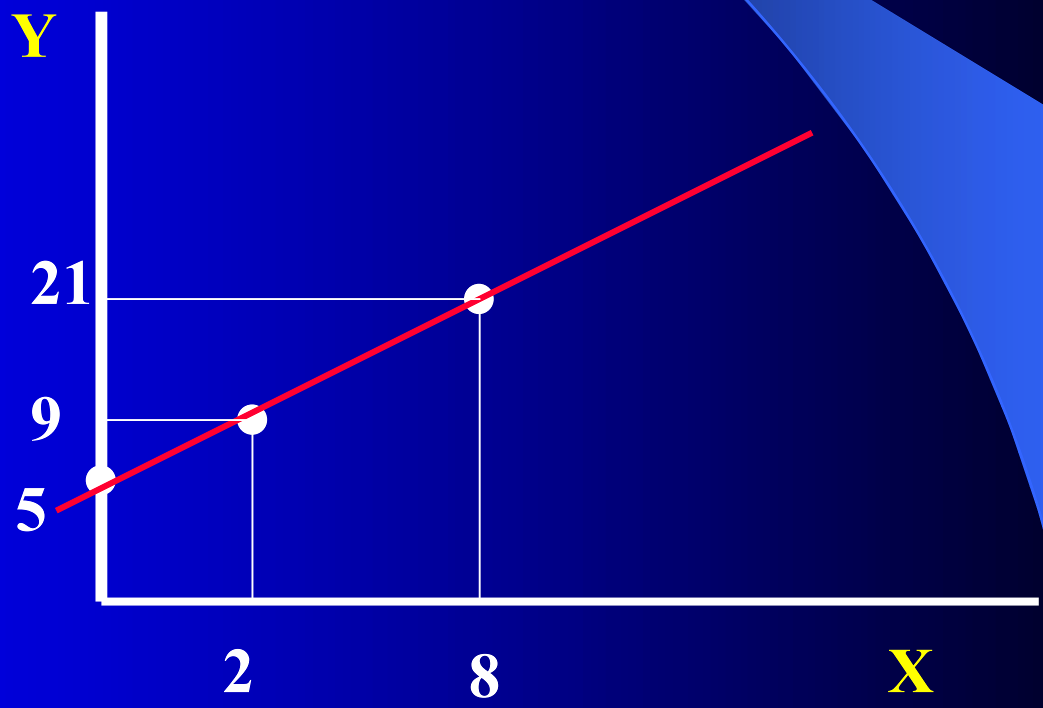
Y	X
5+2x2=9	2
5+2x8=21	8
5+2x0=5	0
....	....

Por si no se conoce....

$$Y = 5 + 2X$$

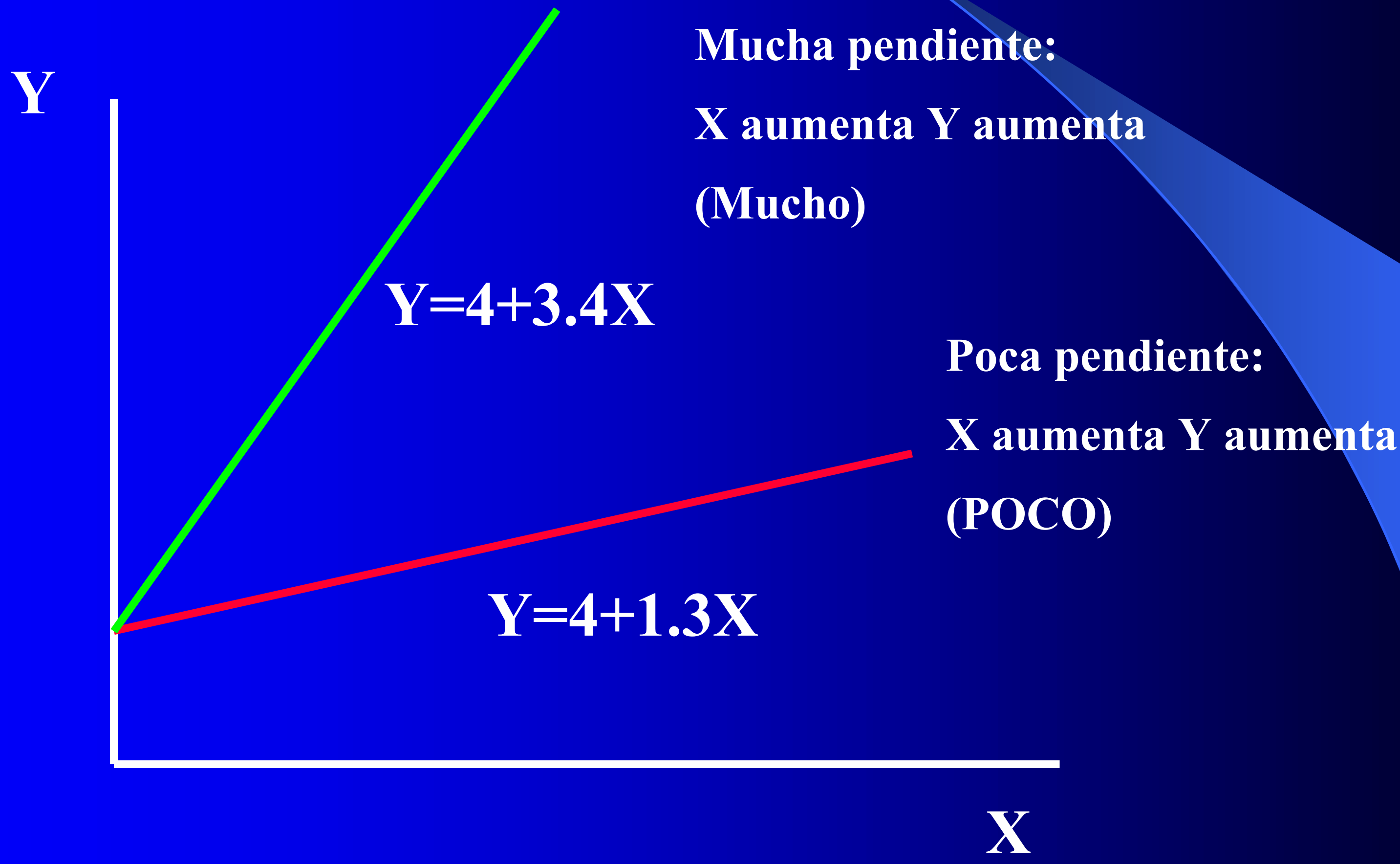
Y	X
5+2x2=9	2
5+2x8=21	8
5+2x0=5	0
....	....

Pintando los puntos en el gráfico X-Y

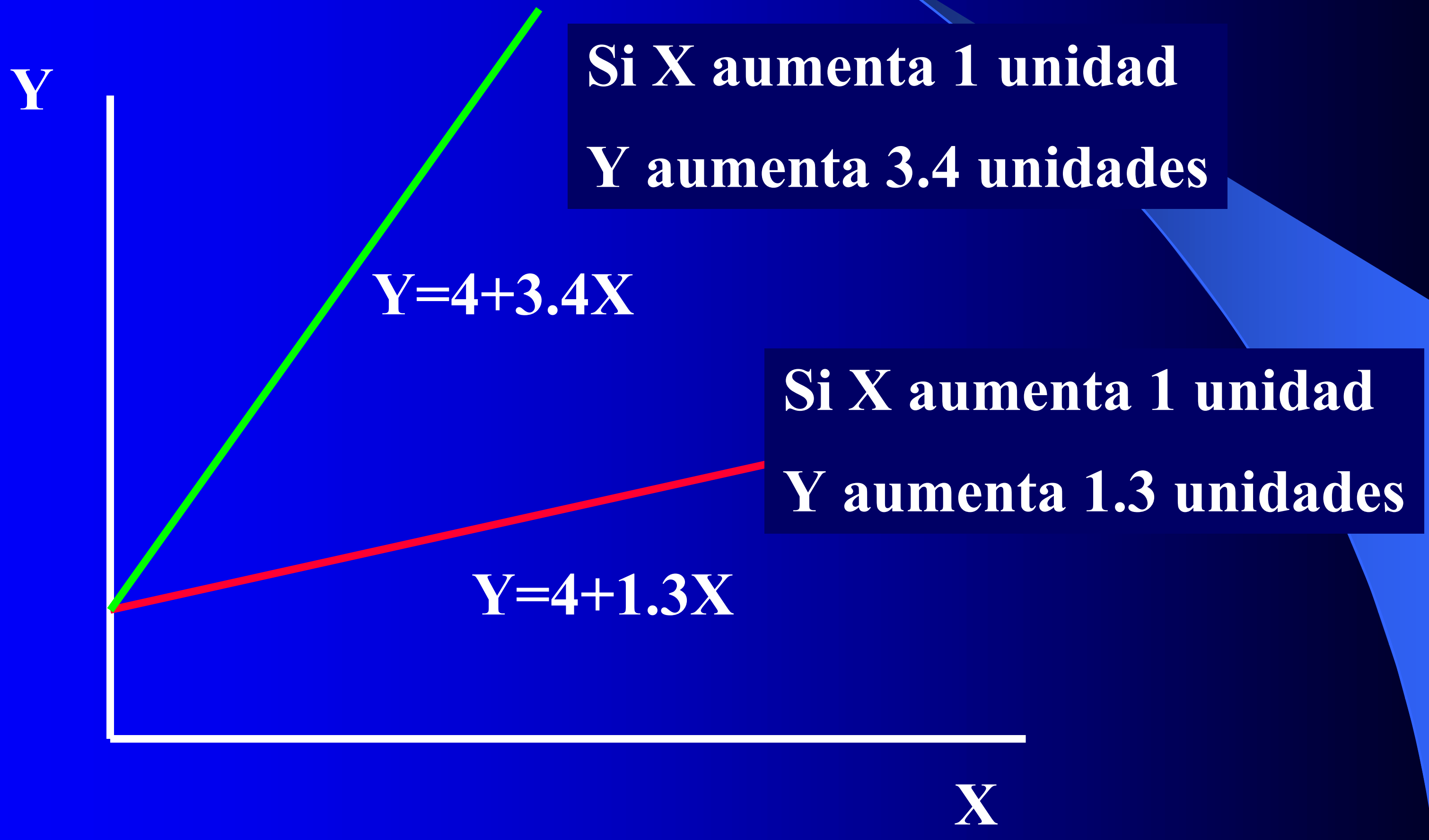


Sale una recta

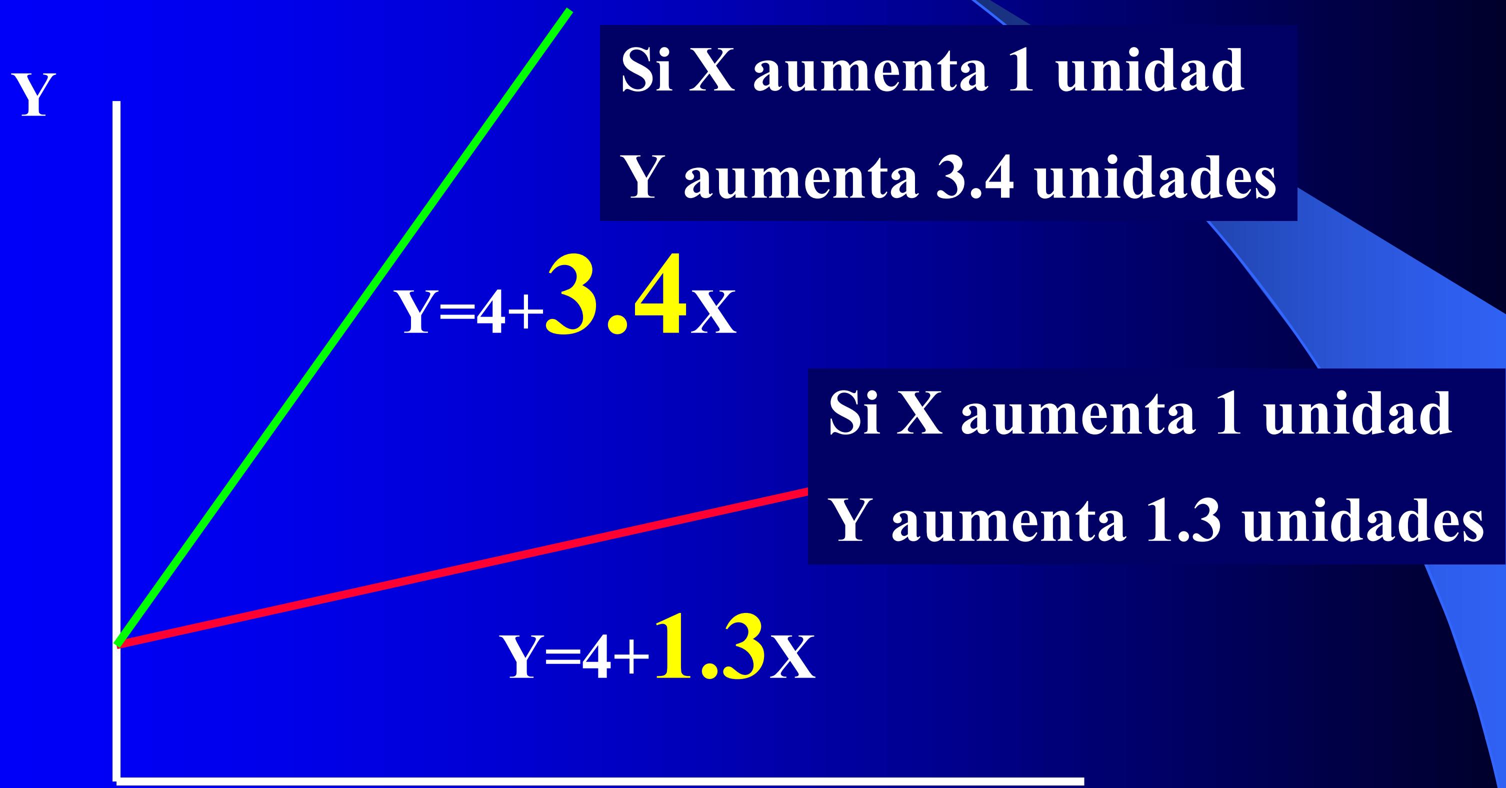
Más sobre rectas



Más sobre rectas

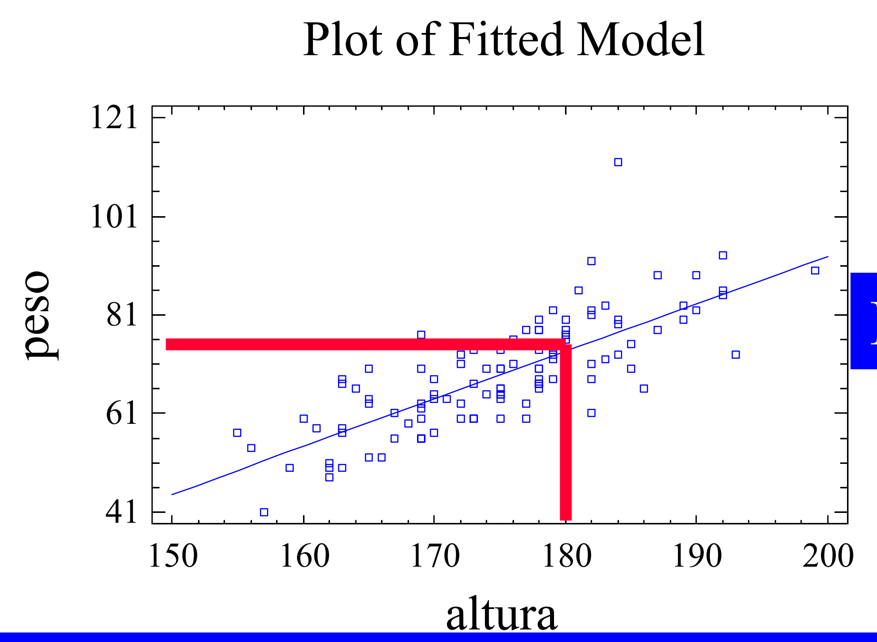


Más sobre rectas



El número que multiplica a la X se llama pendiente (slope) y da una medida cómo de rápido sube (baja) la recta

La recta ajustada, ***que proporciona el ordenador***, es:



Peso = -100,22 + 0.97Altura

- Una persona con Altura=1.80m pesará según la recta de regresión:

Peso= -100,22 + 0.97*180 = 74,38 Kg

•Indudablemente no todas las personas de 1.80 m pesan 74.38 Kg

Si un individuo de altura=1.80m tiene Peso=76 kg, el error o residuo del modelo será:

e=76-74.38 = 1.62 Kg

VAMOS A PREDECIR PESOS

Peso = -100,22 + 0.97Altura

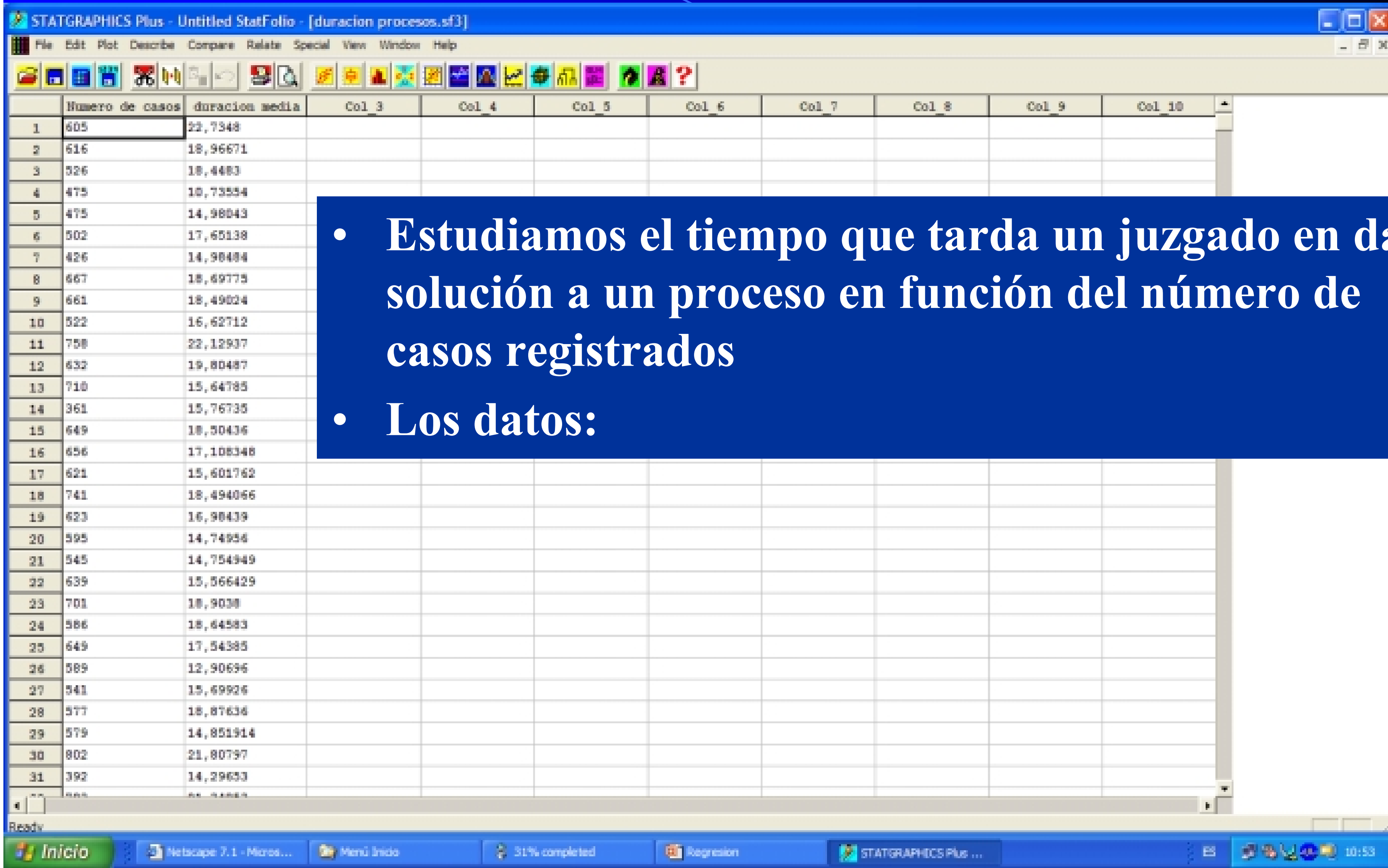
¿Cuánto debería pesar cada un@?

La recta de regresión permite:

- Estudiar cómo X influye sobre Y
- Predecir valores de Y para un valor de X
  - Una persona que mida 1.80m pesará en promedio 74.38kg
- Podemos saber si para nuestra altura estamos “gorditos” o “delgaditos”

- **ESTO ES PARA ESTUDIANTES DE 20 AÑOS !!!!!!!!!**

Veamos otro ejemplo y para qué sirve



	Numero de casos	duracion media	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Col_9	Col_10
1	605	22,7348								
2	616	18,96671								
3	526	18,4483								
4	475	10,73554								
5	475	14,98043								
6	502	17,65138								
7	426	14,98484								
8	667	18,69775								
9	661	18,49024								
10	522	16,62712								
11	758	22,12937								
12	632	19,80487								
13	710	15,64785								
14	361	15,76735								
15	649	18,50436								
16	656	17,108348								
17	621	15,601762								
18	741	18,494066								
19	623	16,98439								
20	595	14,74956								
21	545	14,754949								
22	639	15,566429								
23	701	18,9038								
24	586	18,64583								
25	649	17,54385								
26	589	12,90696								
27	541	15,69926								
28	577	18,87636								
29	579	14,851914								
30	802	21,80797								
31	392	14,29653								

- Estudiamos el tiempo que tarda un juzgado en dar solución a un proceso en función del número de casos registrados
- Los datos:

Si estudiamos los retrasos “a lo bestia”

El retraso medio es de 17 meses.

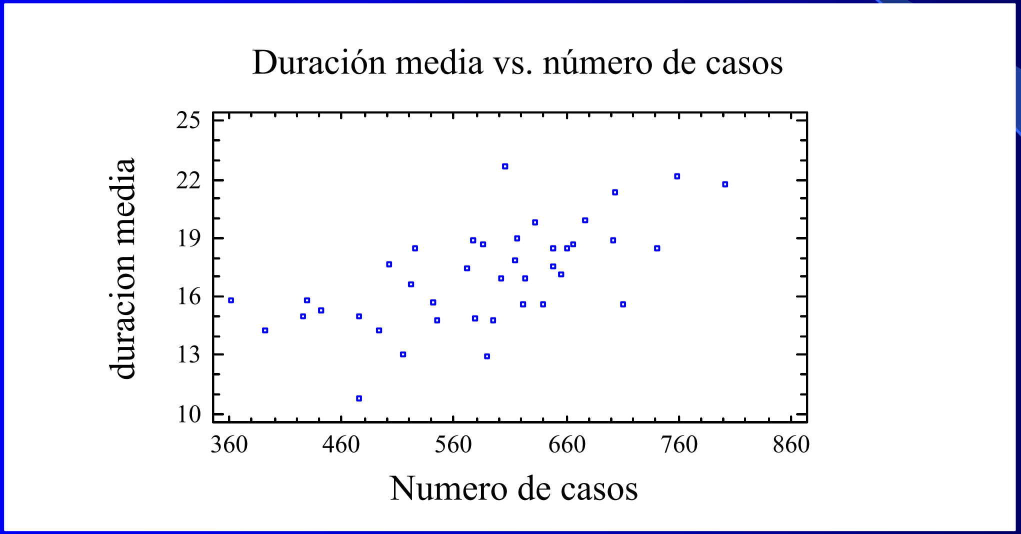
El juzgado número 18 tarda (VER DATOS): 18.49

¿Debemos tomar medidas?

¿Es un juzgado ineficiente?

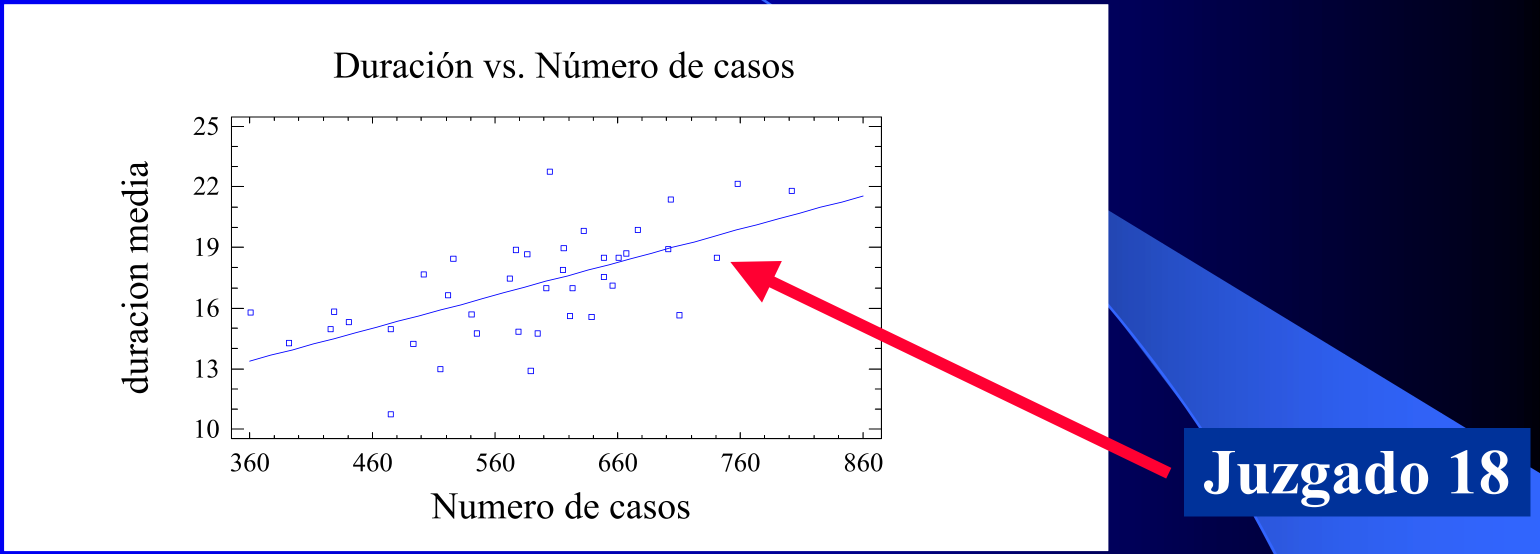
Vamos a estudiarlo de otro modo:

- **El gráfico de Duración vs. Número de casos es:**



La recta de regresión será:

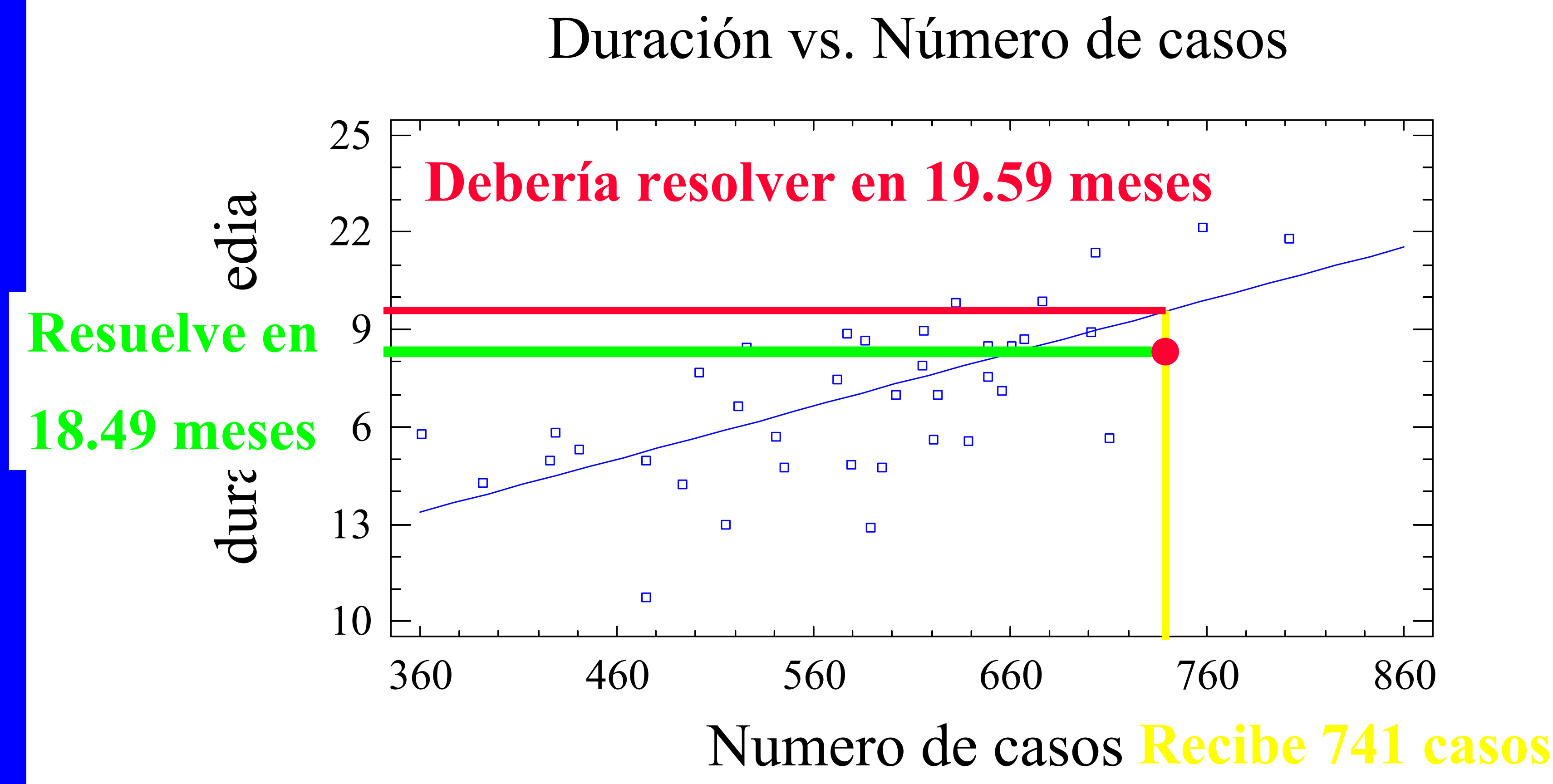
Duracion=7.5+0.016*Num Casos



El juzgado 18 recibe 741 casos. Según la recta de regresión debería

Resolverlos en una duración media de:

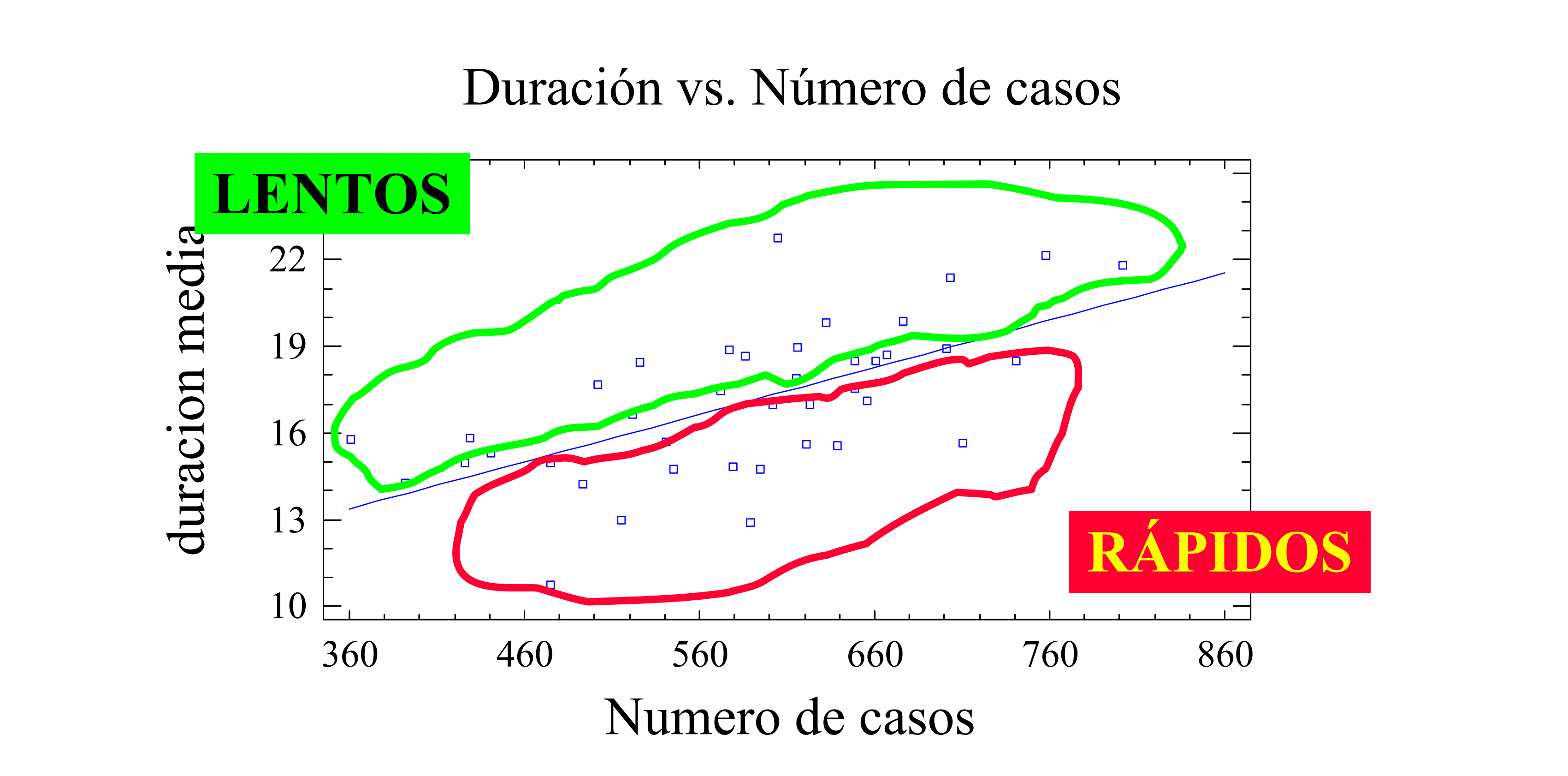
Duración=7.5+0.016*741=19.59



Es un juzgado muy eficiente:

Resuelve en menos tiempo del que le corresponde

18,49-19.59=-1.09 meses



Los juzgados con error de predicción o residuo negativo, resuelven antes de los que les corresponde por volumen de trabajo.

Los positivos (por encima de la línea) resuelven después.

La regresión nos permite:

- Clasificar los juzgados por su rapidez teniendo en cuenta el volumen de trabajo.
- Además como sabemos que un juzgado con 100 casos más tendrá una duración media de 1,6 meses más, podemos evaluar la necesidad de incrementar el personal.

$$Duracion = 7.5 + 0.016 * Num\ Casos$$

El modelo de regresión sirve para predecir Y en función de X. Pero siempre cometeremos algún error. Por ello la ecuación de la regresión será:

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

$$\mathbf{Peso_i = -100.22 + 0.97\ Altura_i + error_i}$$

$$\mathbf{Duracion_i = 7.5 + 0.016 * Num\ Casos_i + error_i}$$

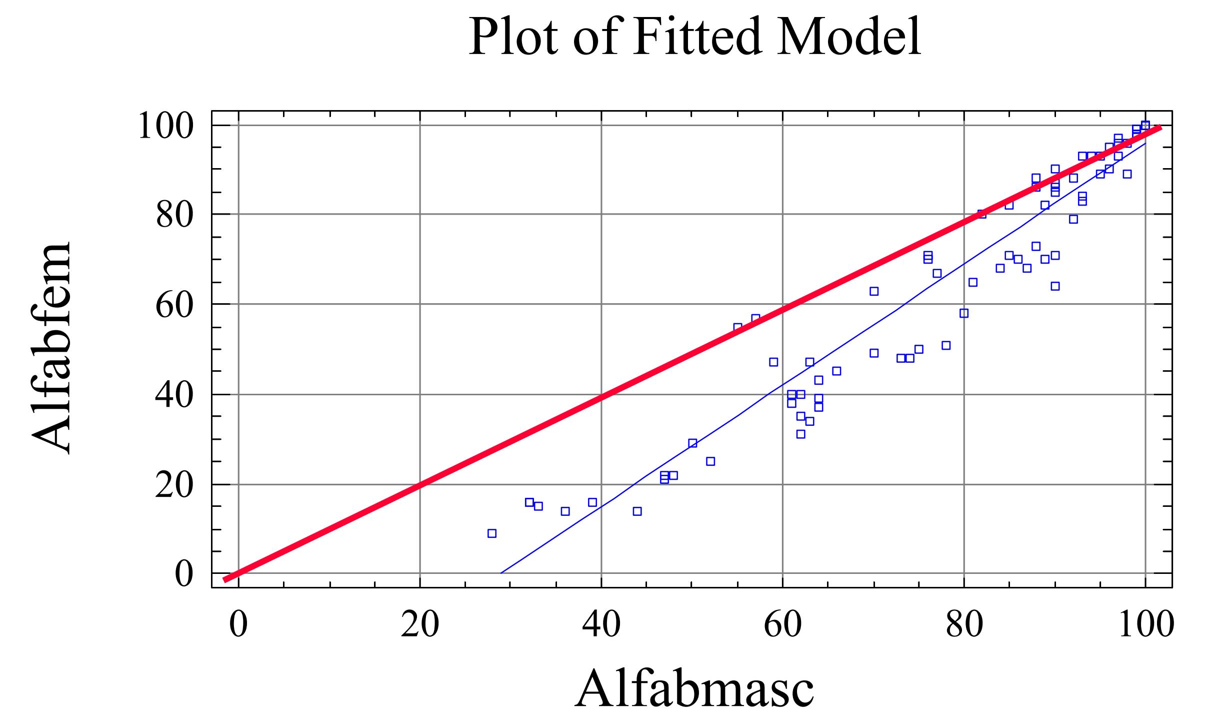
$$\mathbf{Duracion_{18} = 7.5 + 0.016 * Num\ Casos_{18} + error_{18}}$$

Estos análisis nos sirven para “ver” datos

- Y entenderlos
- Y sacar conclusiones

- ….Otros ejemplos

Otro ejemplo: Tasa de alfabetización femenina vs. Masculina en diversos paises



¿Conclusiones?

Regresión Simple

Variable Independiente → Variable Dependiente

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

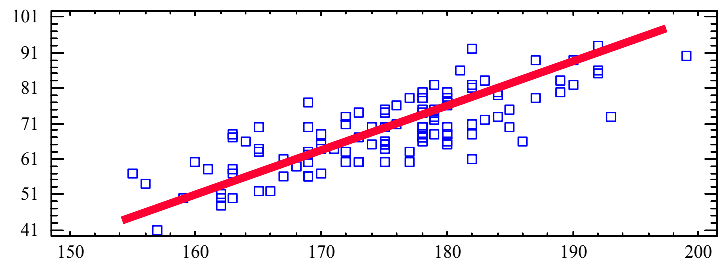
Hipótesis del modelo:

1. Linealidad
2. Homocedasticidad
3. Independencia
4. Normalidad

Resumen

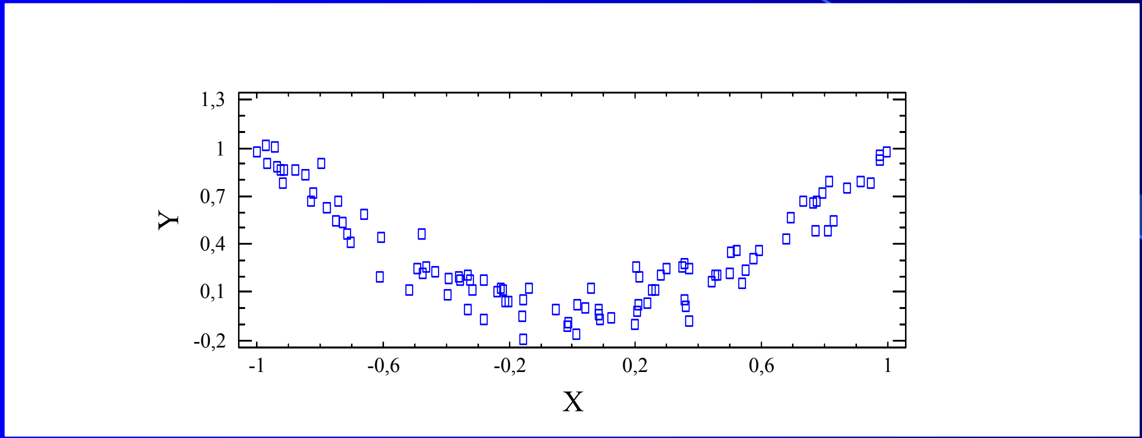
Linealidad

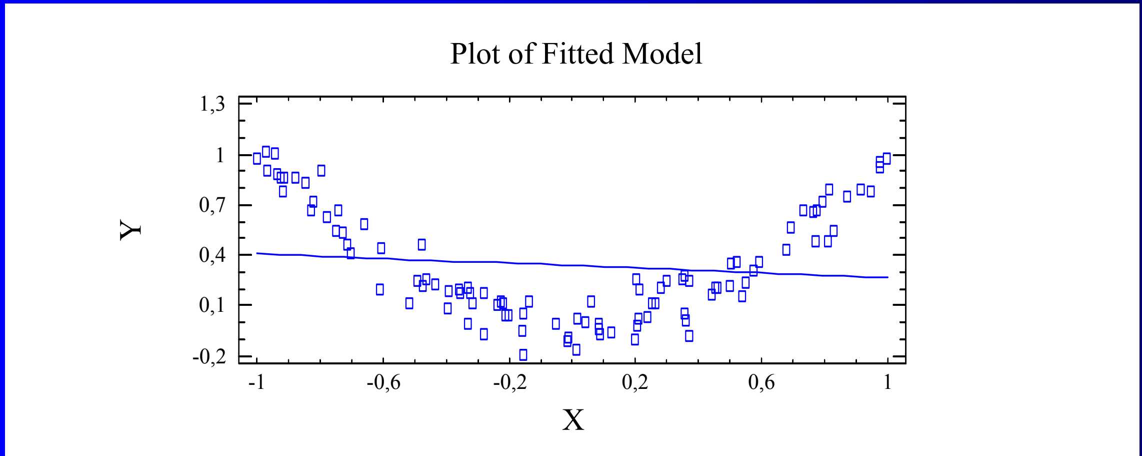
- **Fundamental: Si vamos ajustar una línea recta, los datos deben ser razonablemente rectos.**



Datos rectos:
Recta de regresión representa bien la estructura de los datos

Linealidad: Datos no lineales





Datos no rectos:

Recta de regresión no representa la estructura de los datos

Linealidad: Comprobación

Siempre:

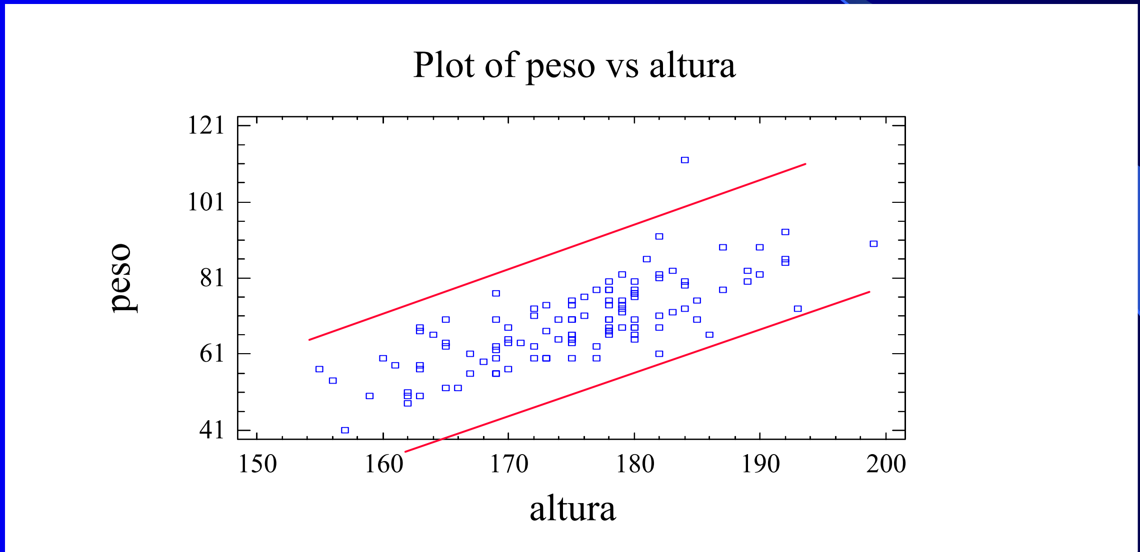
- **Haremos el gráfico X-Y de los datos y comprobaremos si son lineales.**

•Si no son lineales hay que transformarlos

•Ajustar un recta a datos curvos es un sinsentido y nos lleva a conclusiones equivocadas

Homocedasticidad

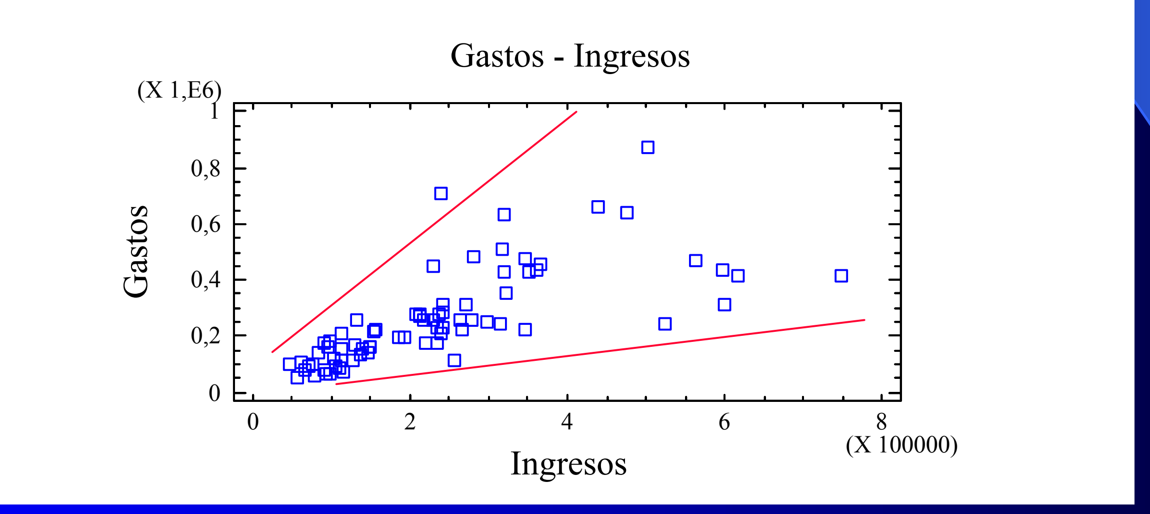
- La nube de puntos debe tener el mismo grosor
- La dispersión o varianza de los datos es constante.
- Si no se cumple se llaman HETEROCEDASTICOS.



Datos homocedásticos:
Varianza constante

Datos heterocedásticos

- **Es un fenómeno frecuente.**
- **En economía los gastos de las familias tienen variabilidad creciente a medida que aumentan los ingresos:**



Datos heterocedásticos: Varianza creciente

Homocedasticidad: Comprobación

Siempre:

- **Haremos el gráfico X-Y de los datos y comprobaremos si son homocedásticos.**

- **Si no son homocedásticos hay que transformarlos**

- **Ajustar un recta a datos heterocedásticos nos lleva a conclusiones equivocadas**

Independencia

- Fundamental: Los datos deben ser independientes. Una observación (Un punto) no debe dar información sobre las demás
- Sabremos por el tipo de datos si son adecuados para el análisis.

No sirven datos temporales

Datos independientes:
Recta de regresión representa bien la estructura de los datos

Independencia

- Datos temporales

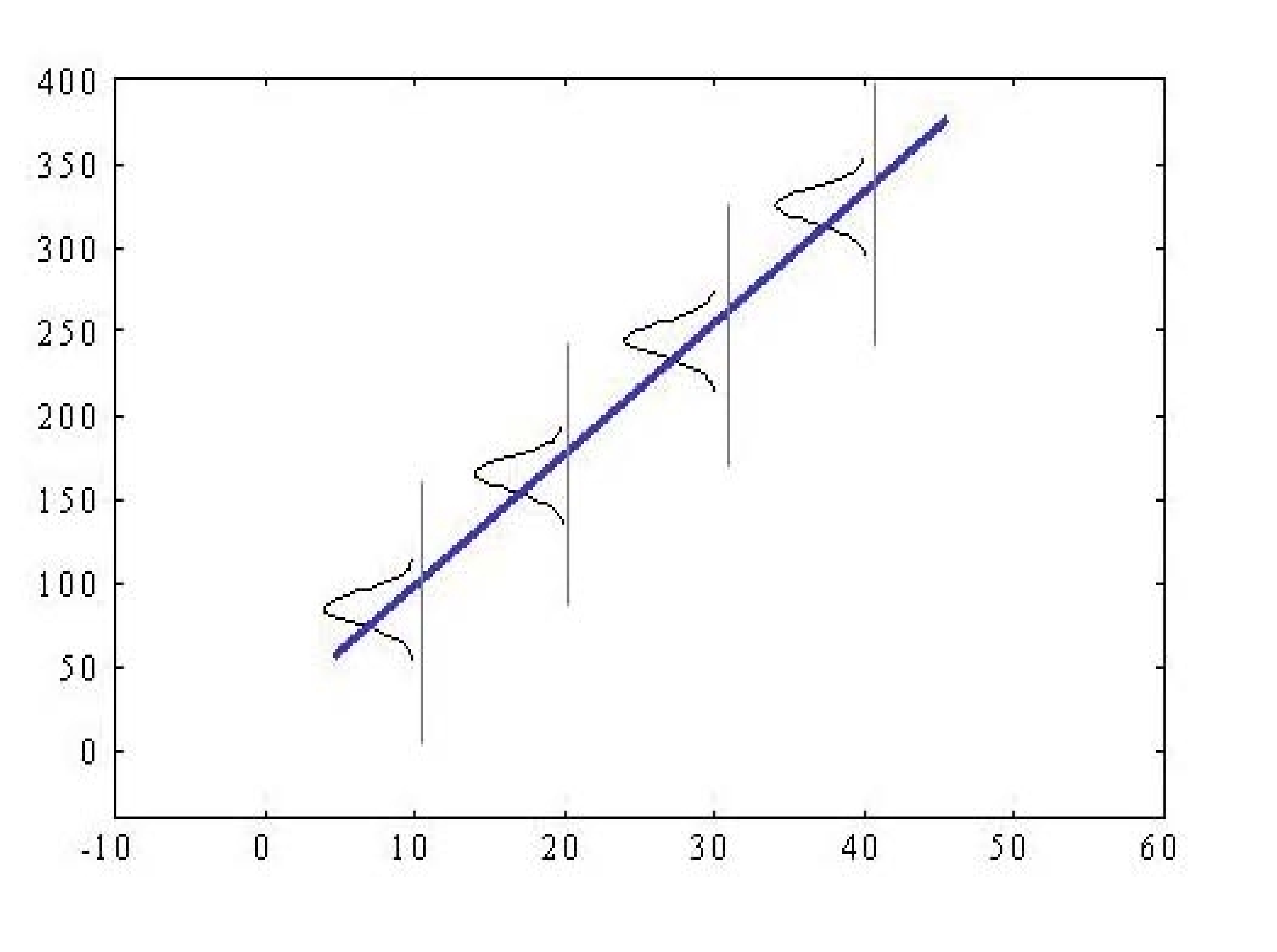
Año	Coches vendidos	Número de p[illegible]dos
1980	[illegible]56	[illegible]0.234.
1981	865	679.067
1982	905	678.890
1983	[illegible]24	6[illegible]67
1984	930	578.458
[illegible]	945	540.689
1986	980	534.485

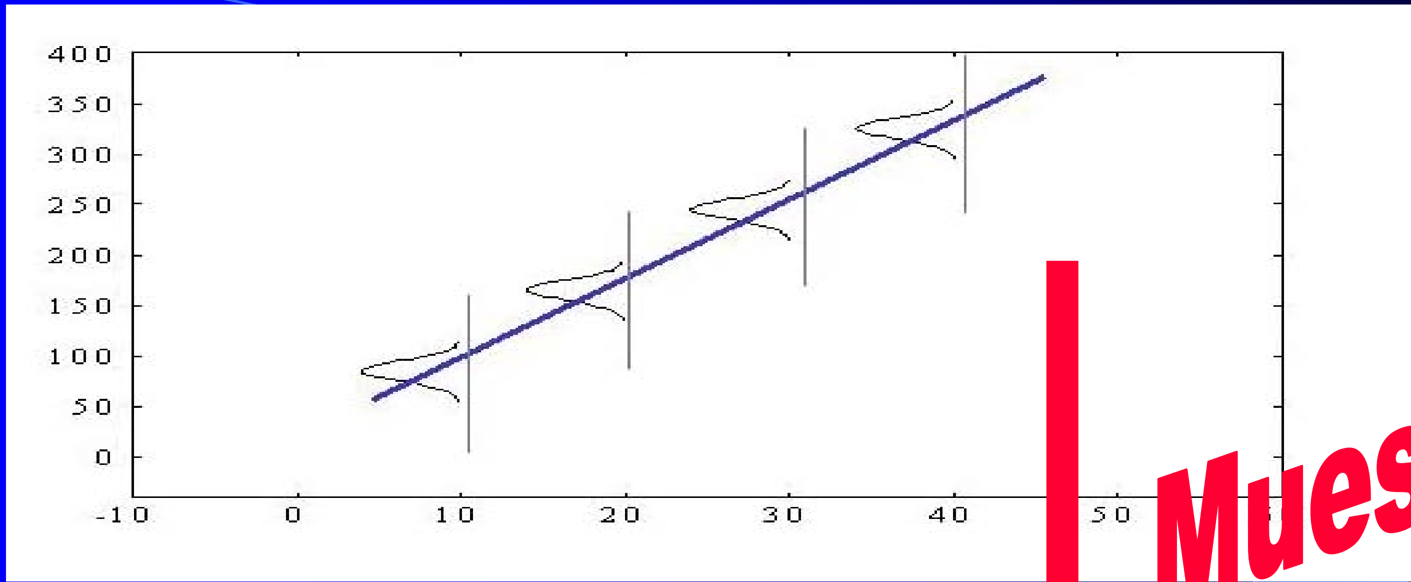
Datos dependientes:
La recta de regresión sacará falsas relaciones. Relaciones espúrias

Normalidad

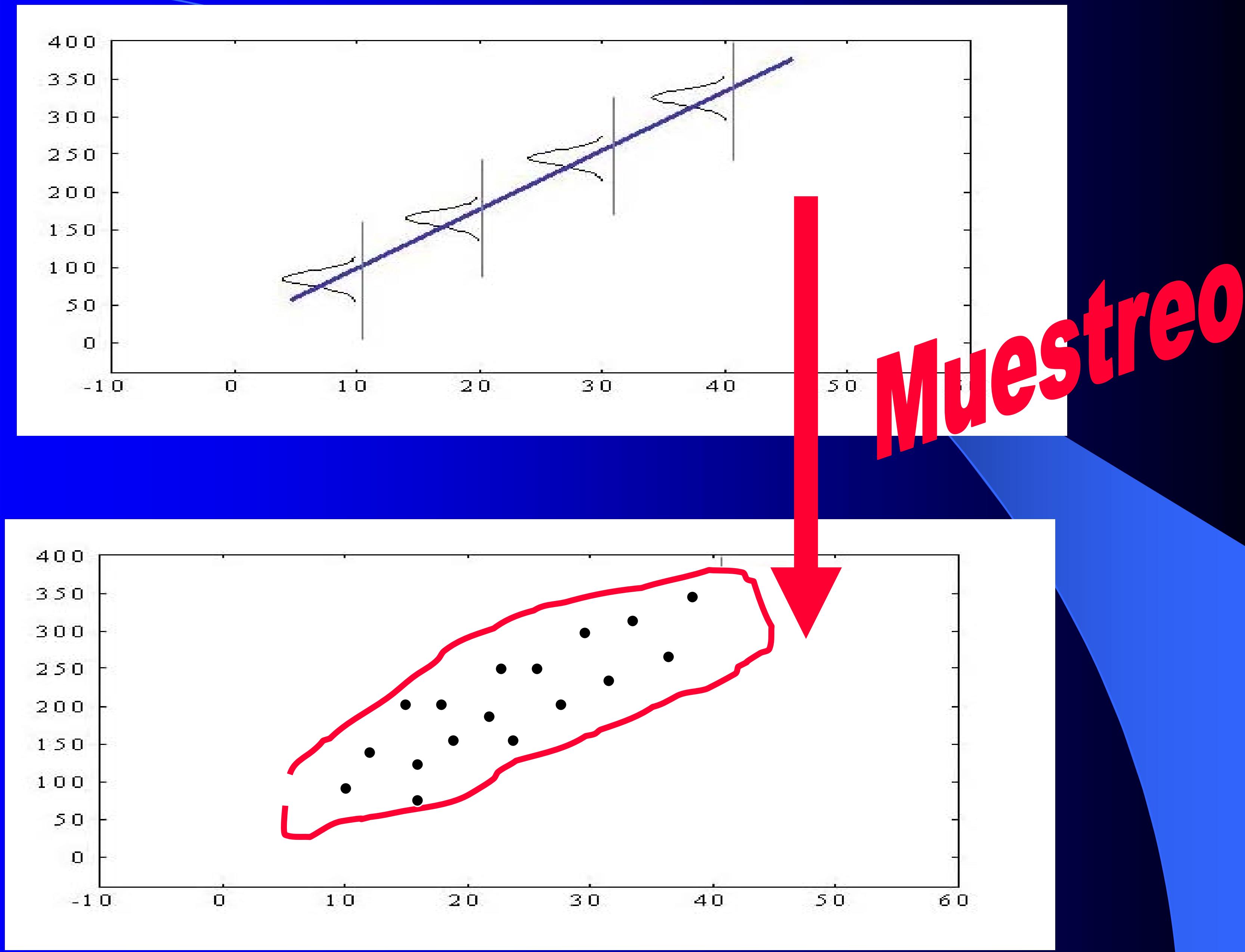
- Admitimos que los datos son normales
- No lo comprobaremos a priori

El modelo

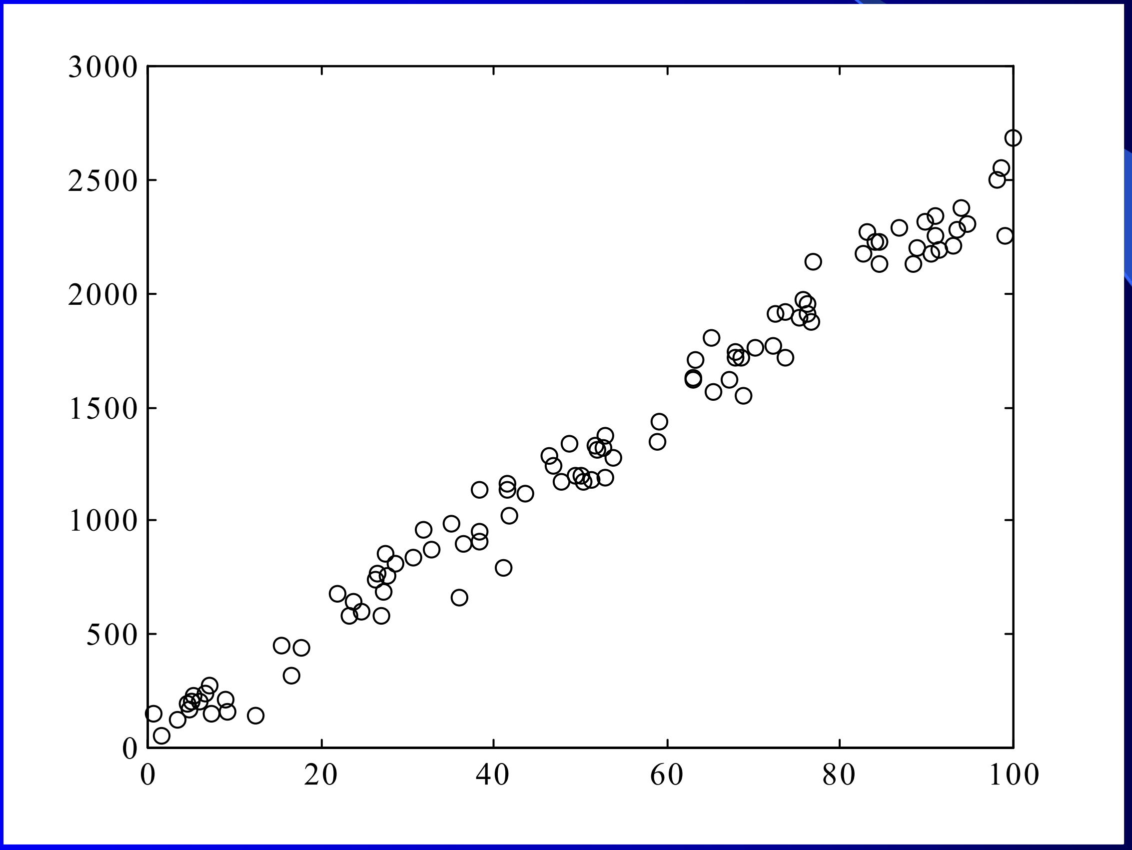




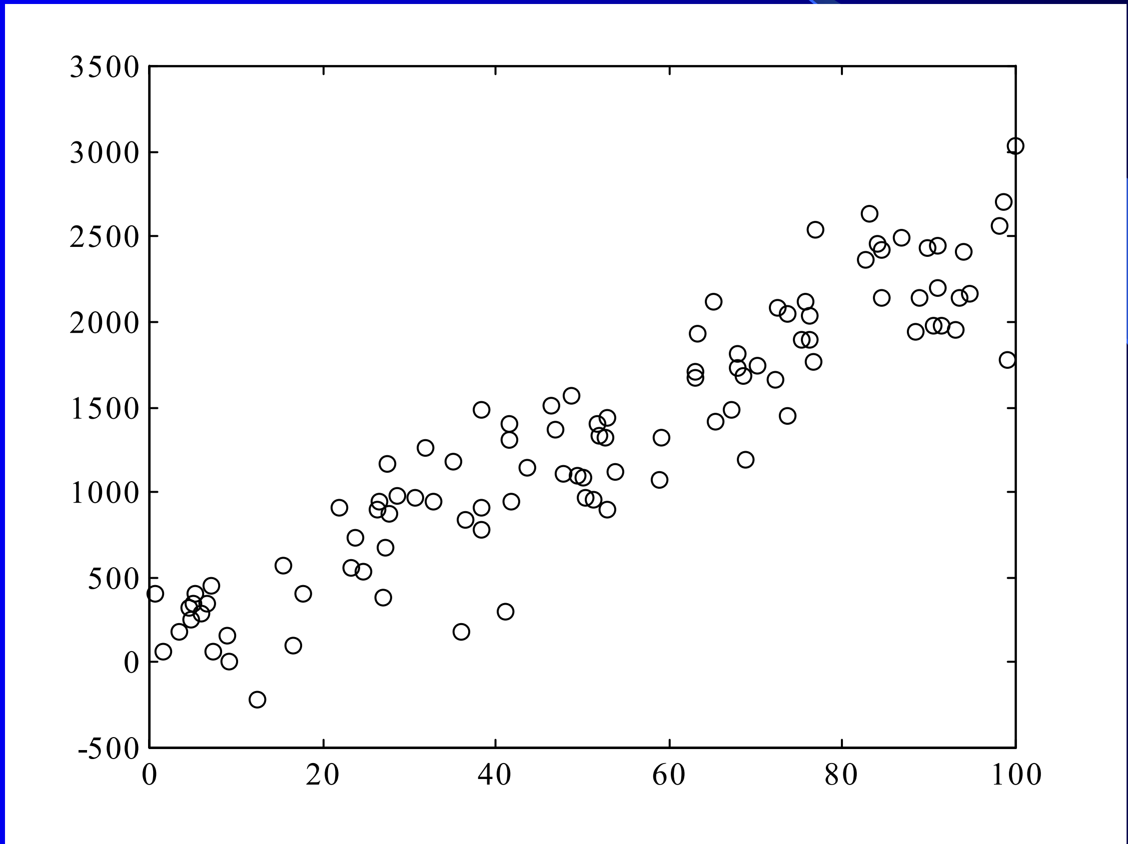
Muestreo


Muestreo

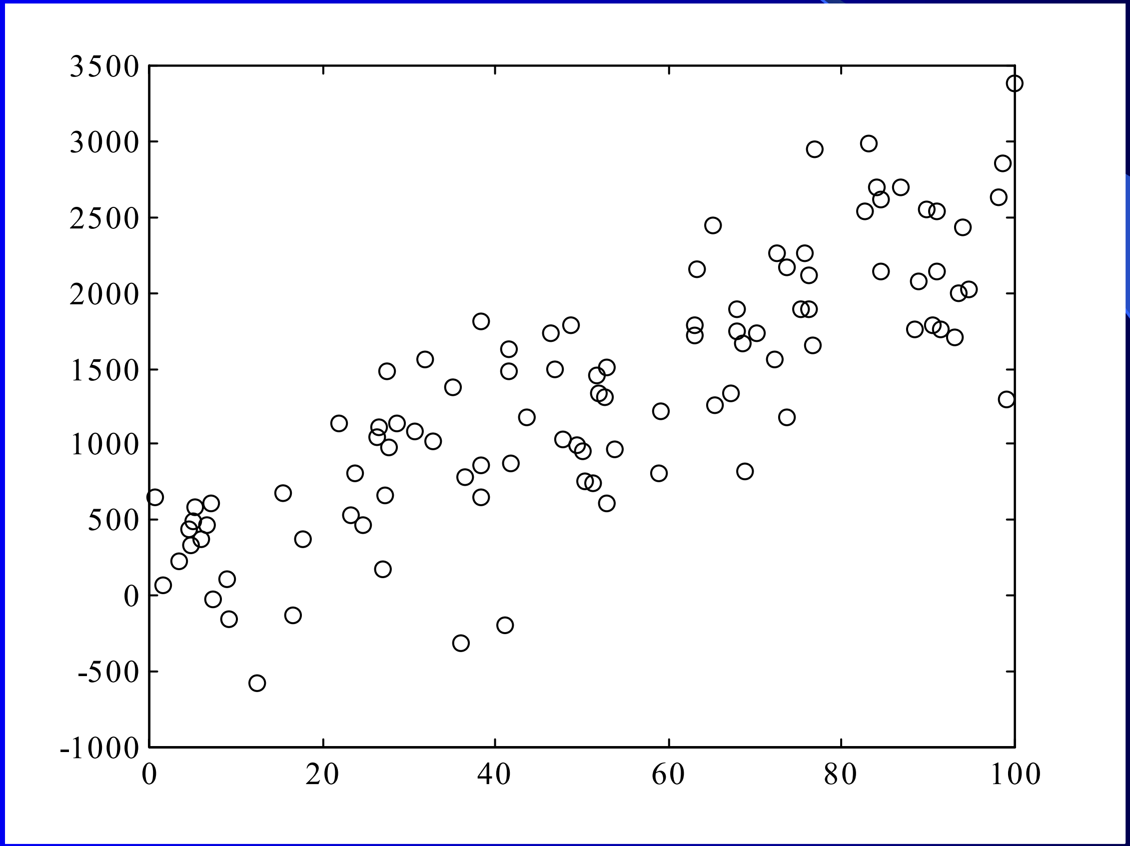
Si los datos fueran así, ¿Hay relación? ¿Cumple las hipótesis?



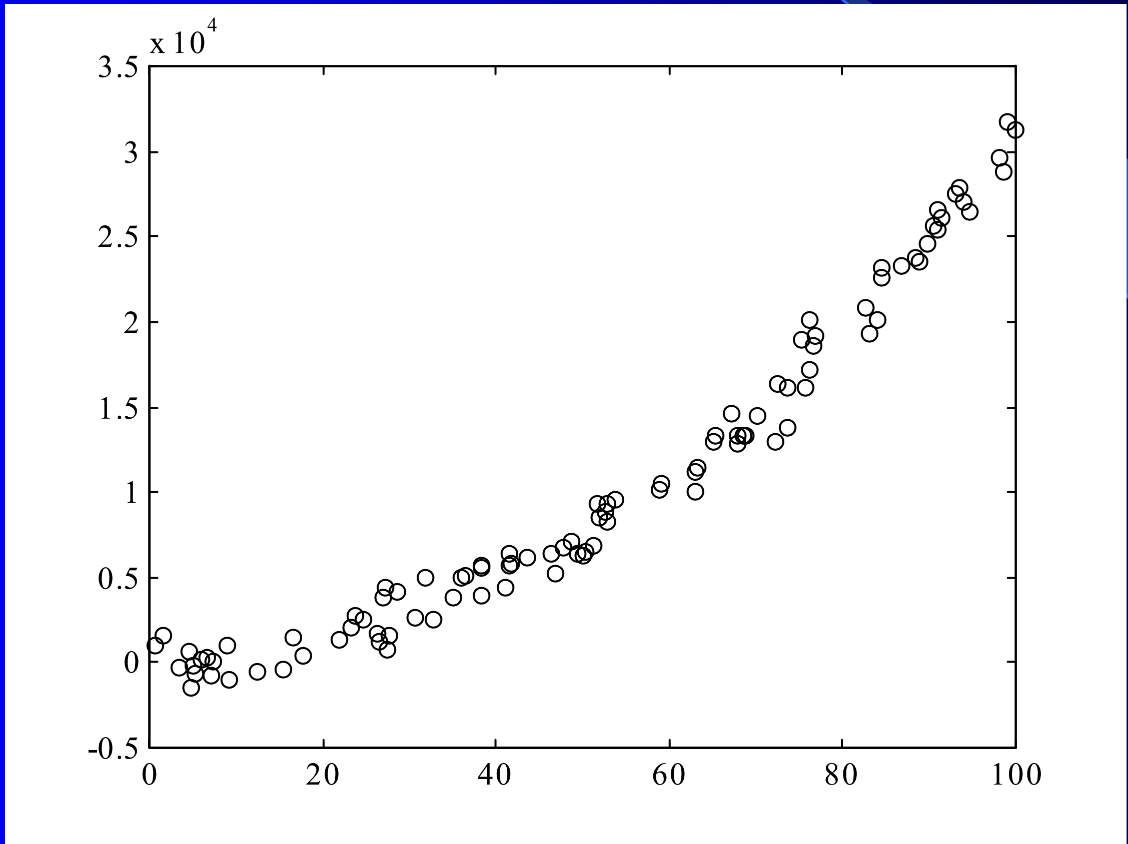
¿Hay relación?
¿Cumple las hipótesis?



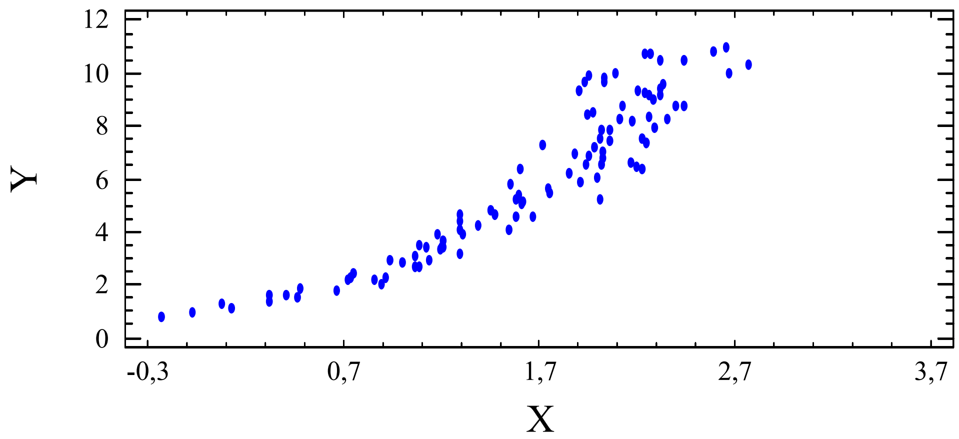
¿Hay relación?
¿Cumple las hipótesis?



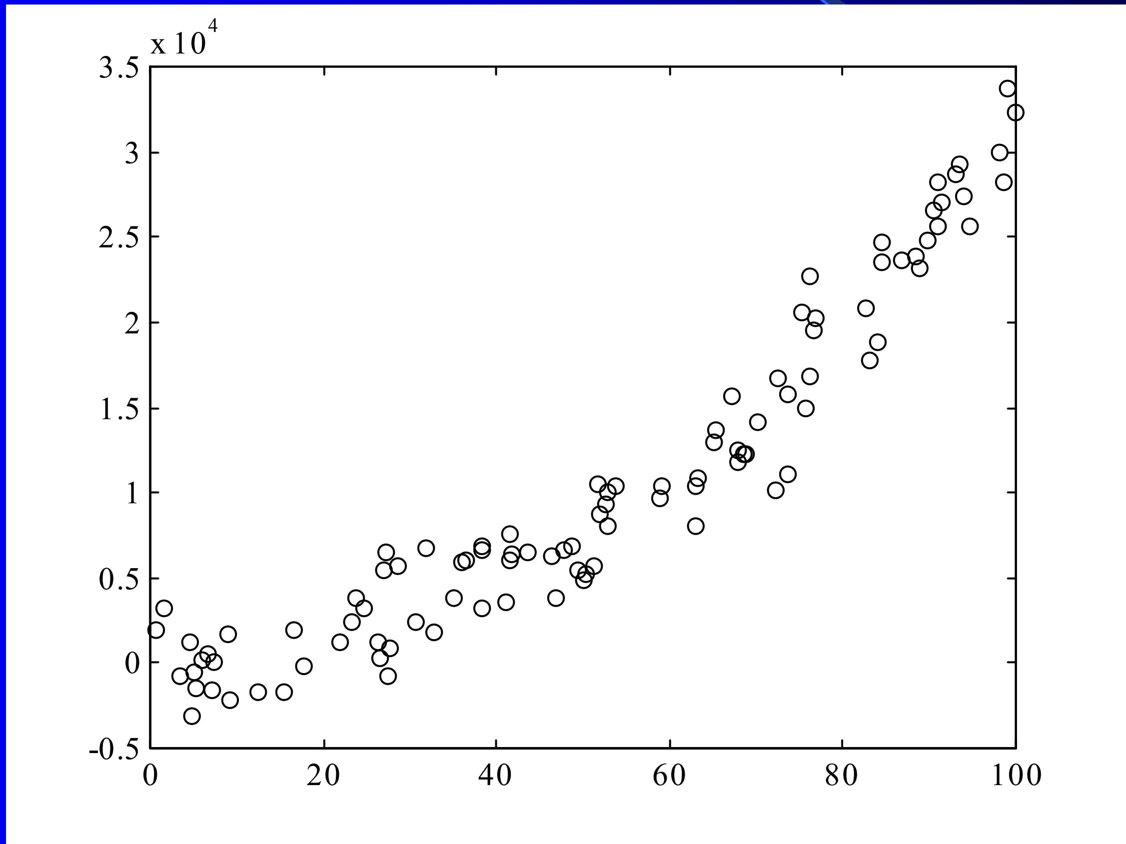
¿Hay relación?
¿Cumple las hipótesis?



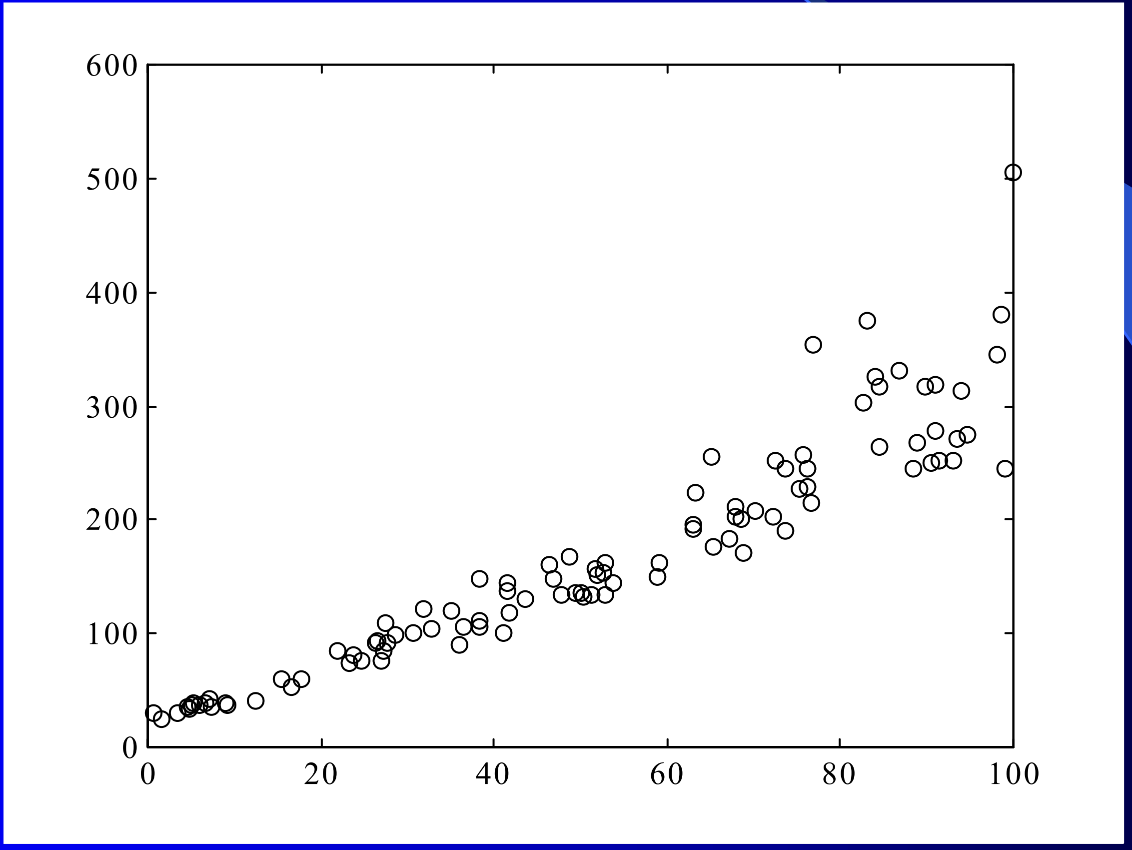
¿Hay relación?
¿Cumple las hipótesis?



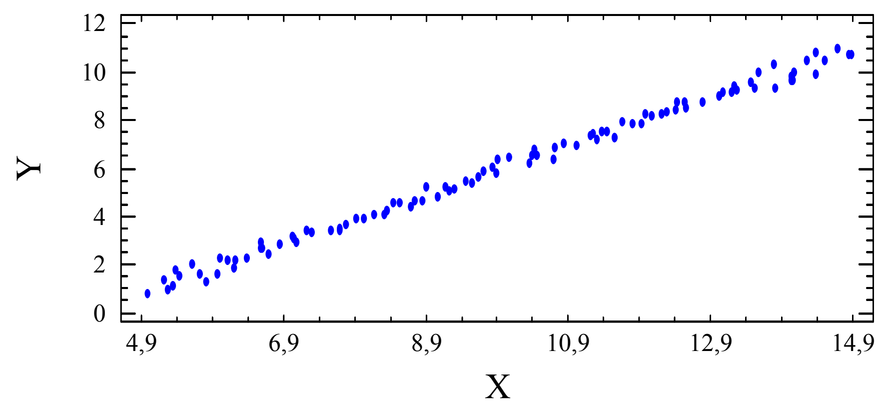
¿Hay relación?
¿Cumple las hipótesis?



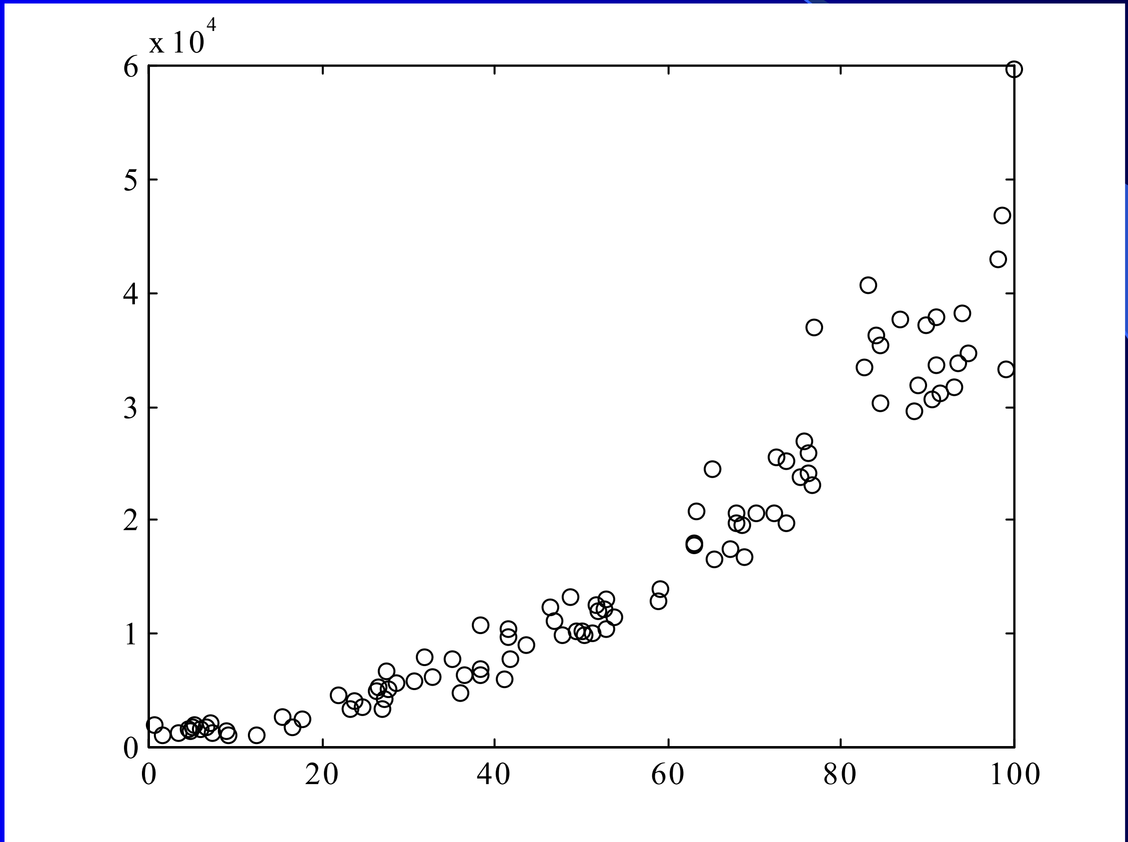
¿Hay relación?
¿Cumple las hipótesis?



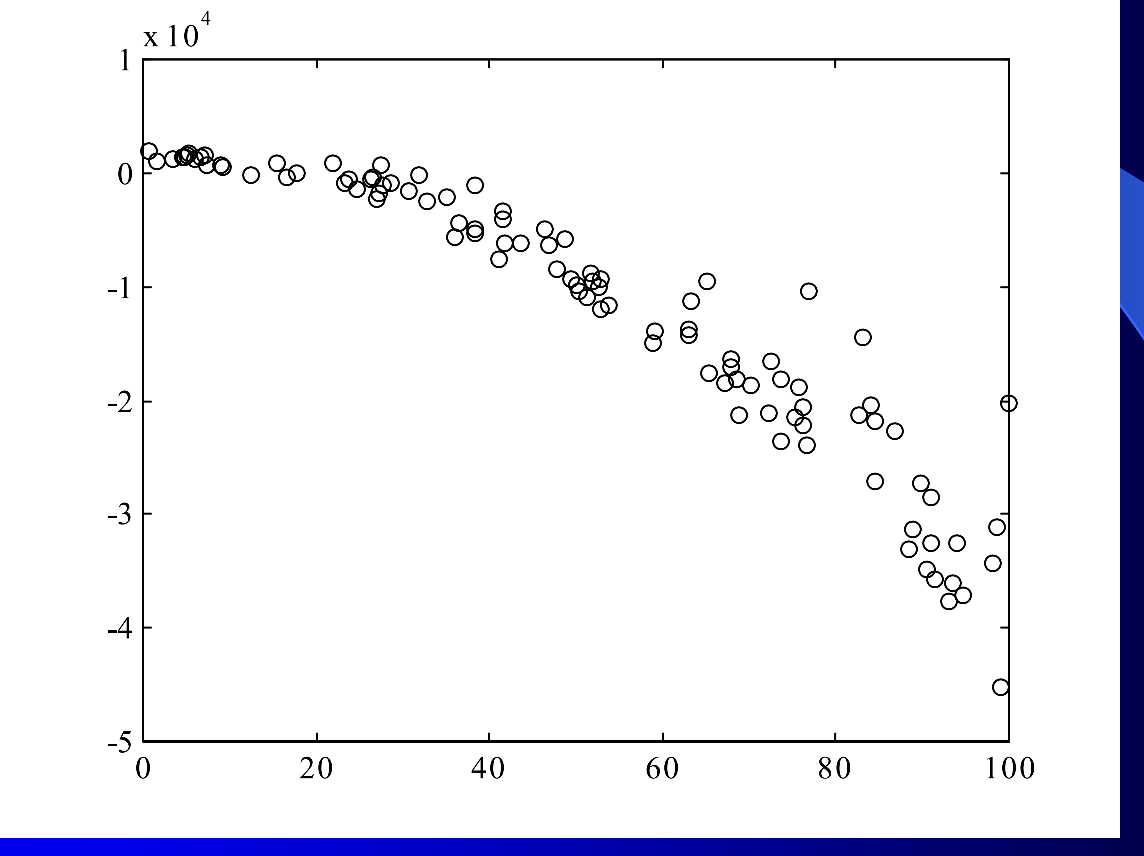
¿Hay relación?
¿Cumple las hipótesis?



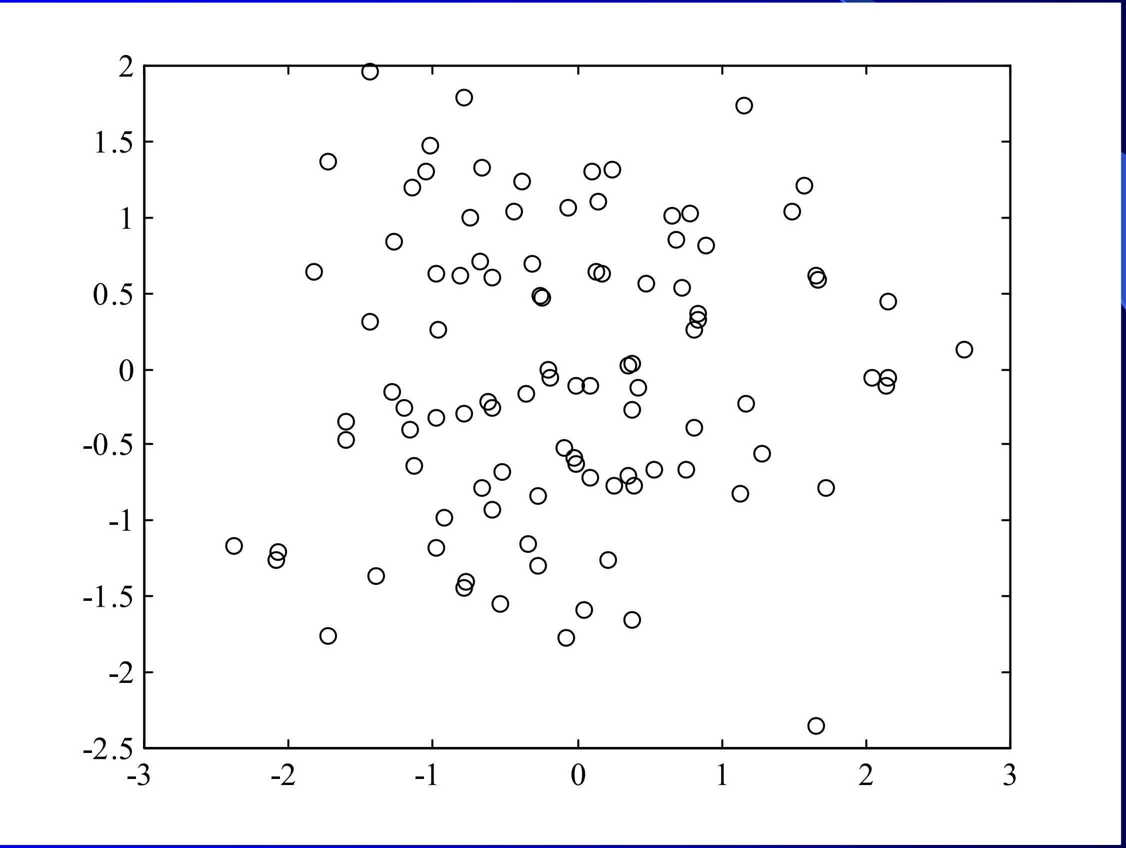
¿Hay relación?
¿Cumple las hipótesis?

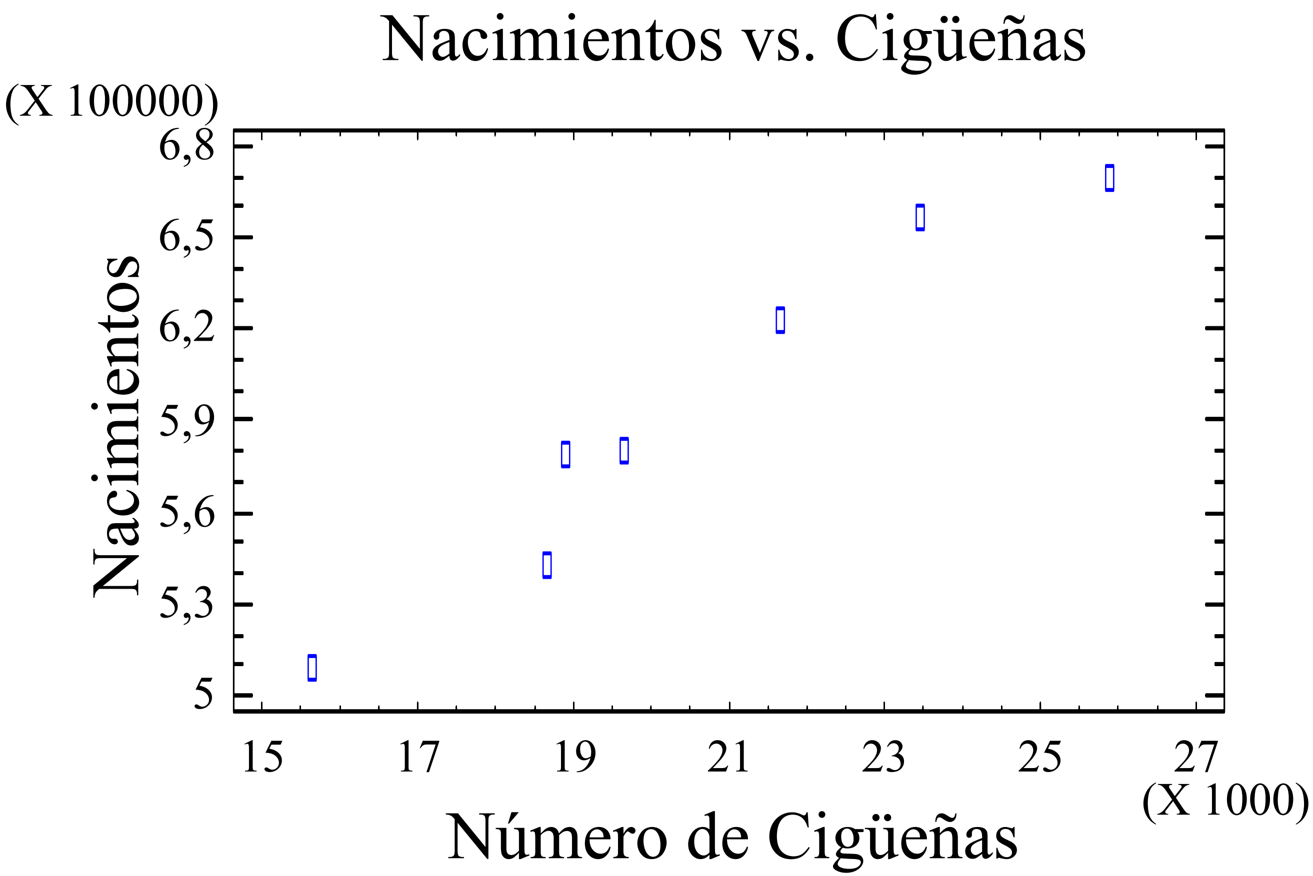


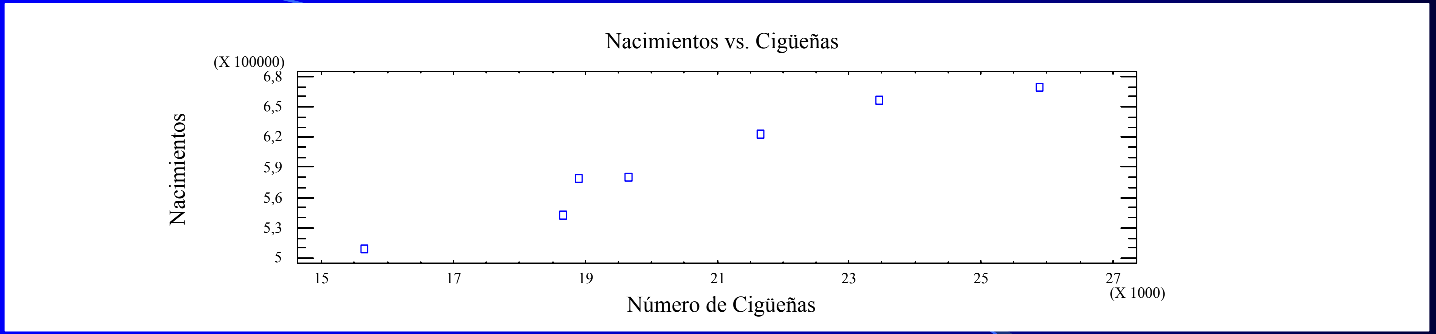
¿Hay relación?
¿Cumple las hipótesis?

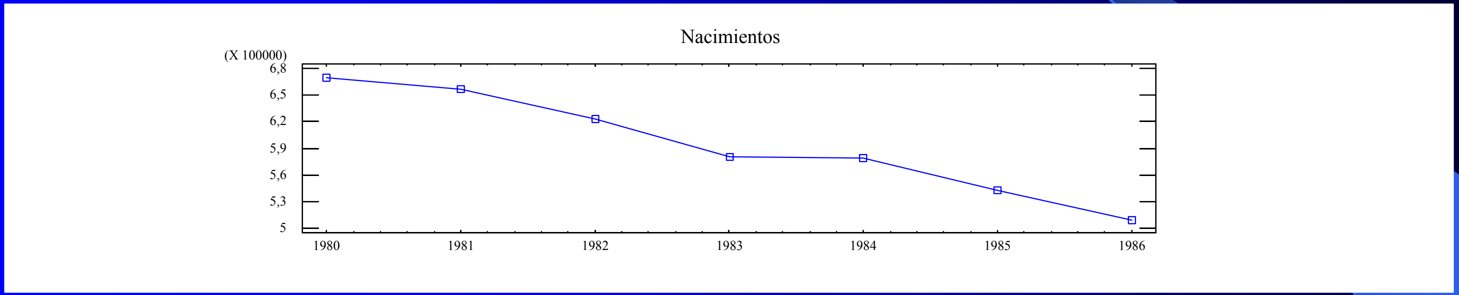


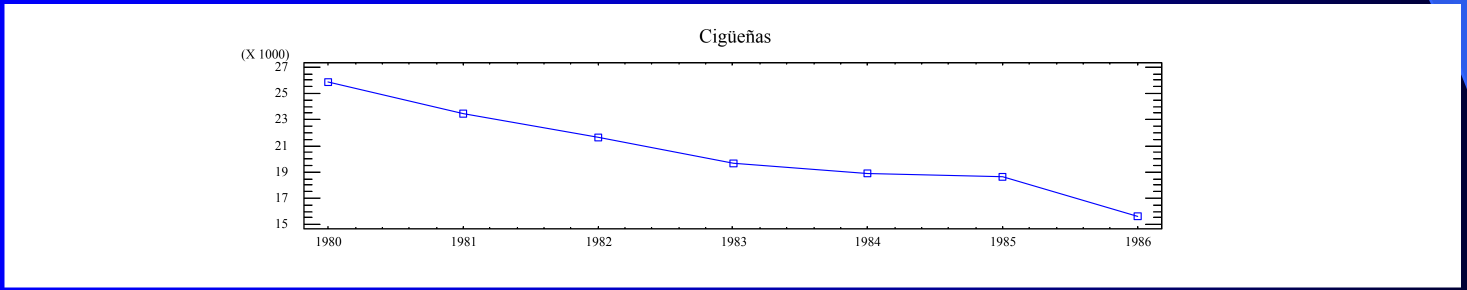
¿Hay relación?
¿Cumple las hipótesis?

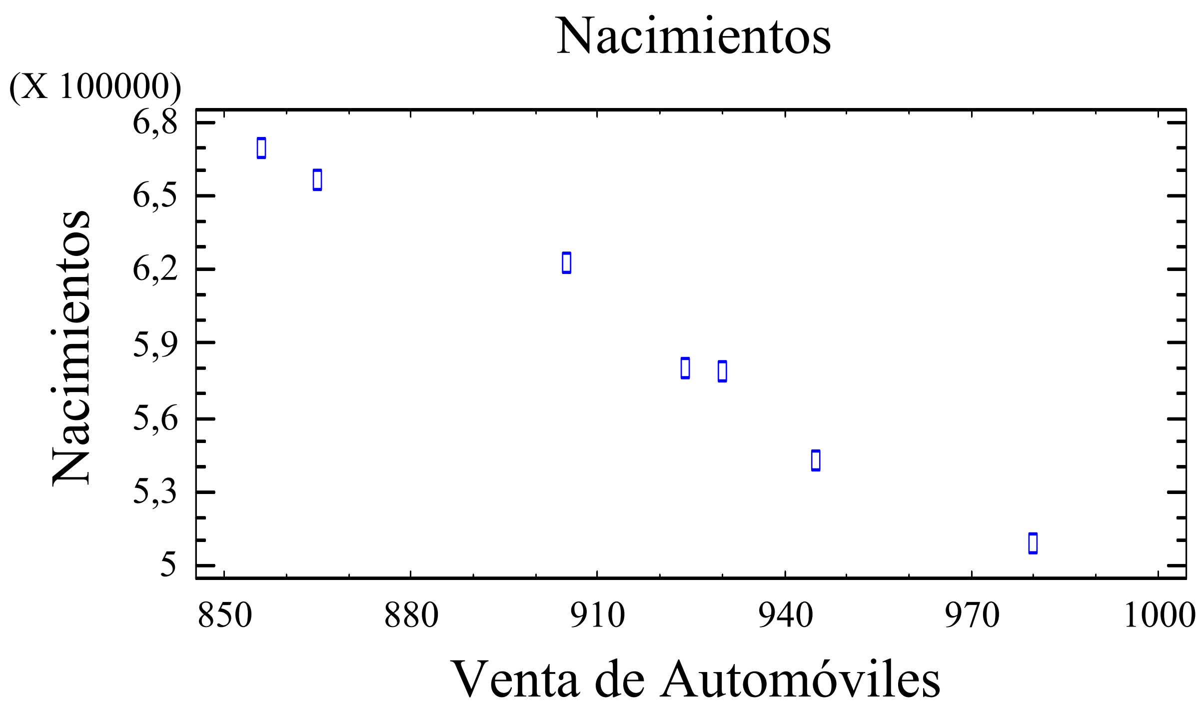



Nacimientos vs. Cigüeñas
(X 100000)
6,8
6,5
6,2
5,9
5,6
5,3
5
Nacimientos
15
17
19
21
23
25
27
(X 1000)
Número de Cigüeñas

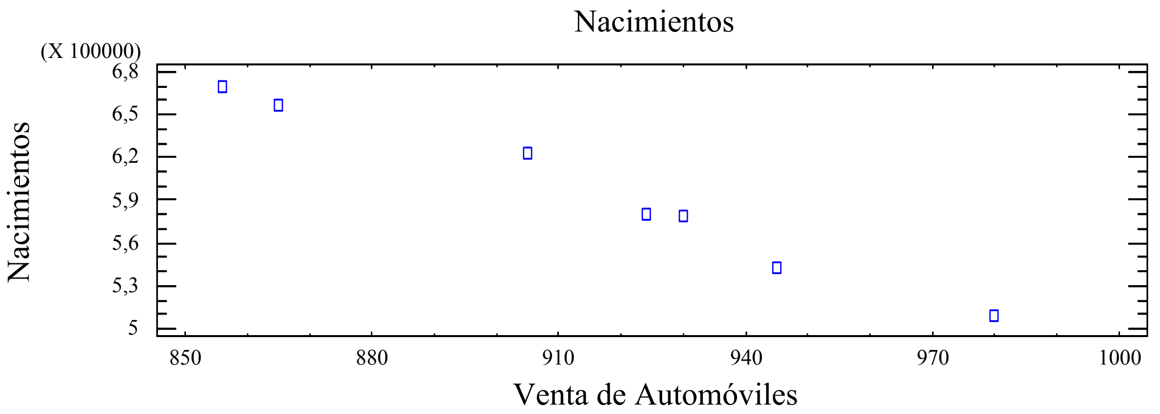

Nacimientos vs. Cigüeñas
(X 100000)
Nacimientos
6,8
6,5
6,2
5,9
5,6
5,3
5
15
17
19
21
23
25
27
(X 1000)
Número de Cigüeñas


Nacimientos
(X 100000)
6,8
6,5
6,2
5,9
5,6
5,3
5
1980
1981
1982
1983
1984
1985
1986

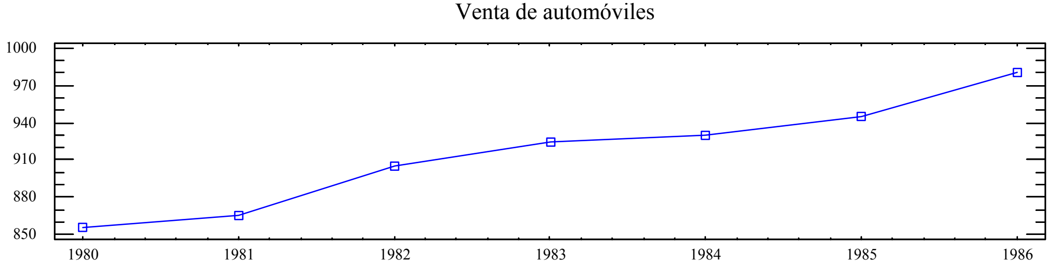

Cigüeñas
(X 1000)
27
25
23
21
19
17
15
1980
1981
1982
1983
1984
1985
1986



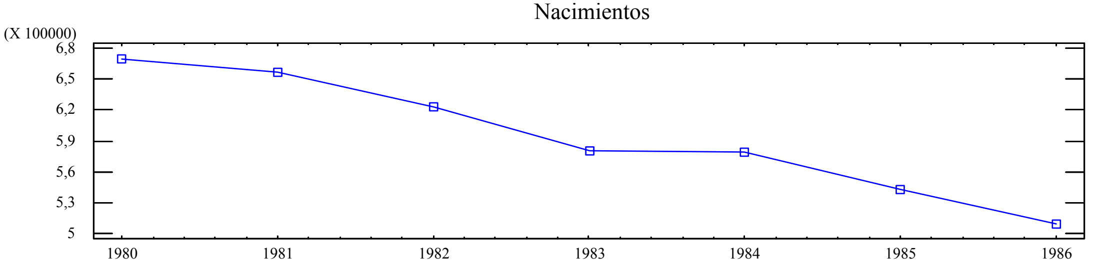
Nacimientos
(X 100000)
6,8
6,5
6,2
5,9
5,6
5,3
5
Nacimientos
850
880
910
940
970
1000
Venta de Automóviles



Nacimientos
(X 100000)
6,8
6,5
6,2
5,9
5,6
5,3
5
Nacimientos
850
880
910
940
970
1000
Venta de Automóviles



Venta de automóviles
1000
970
940
910
880
850
1980
1981
1982
1983
1984
1985
1986

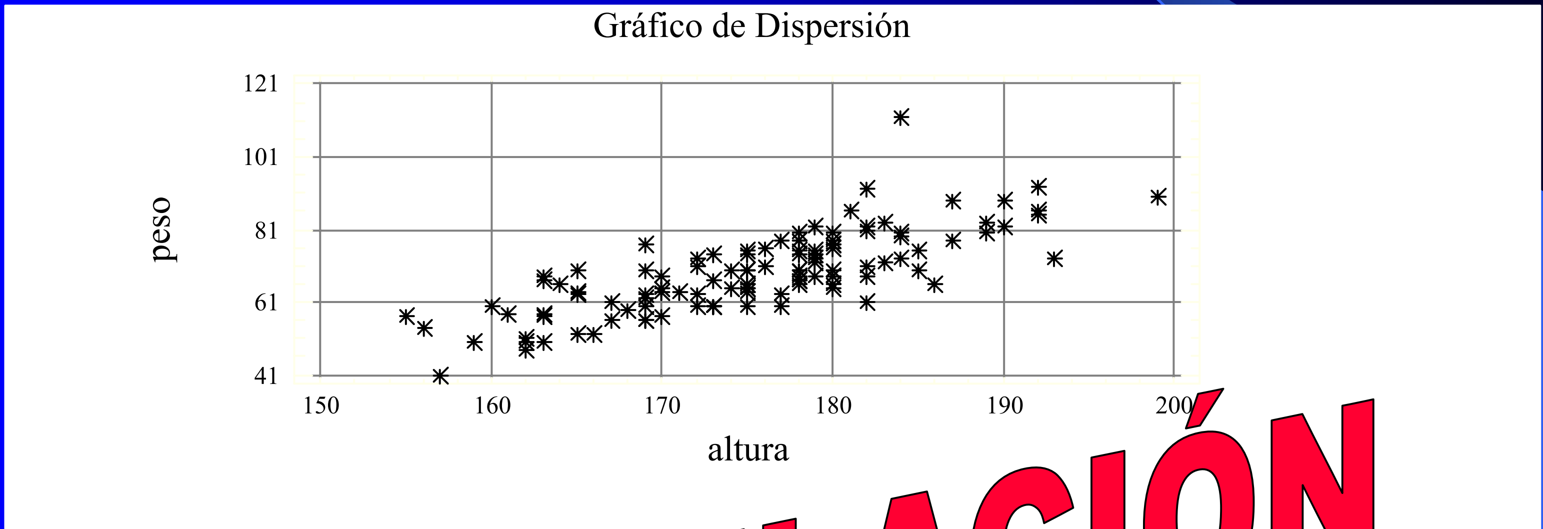


Nacimientos
(X 100000)
6,8
6,5
6,2
5,9
5,6
5,3
5
1980
1981
1982
1983
1984
1985
1986

Medir la Relación entre dos variables:

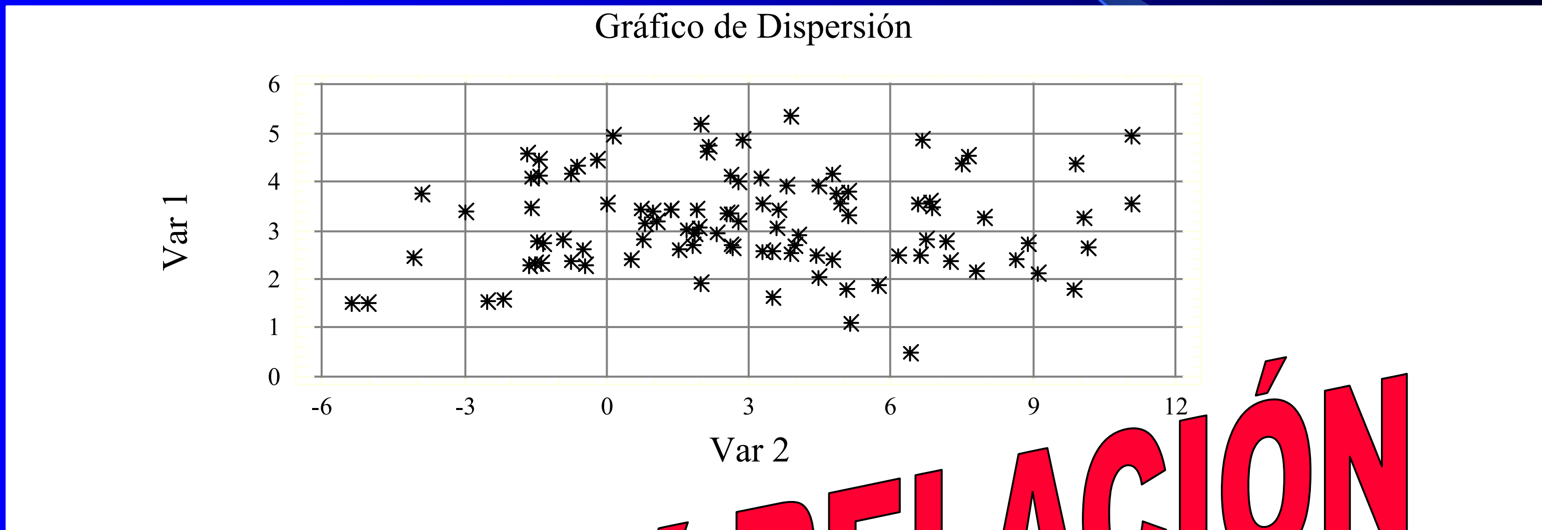
- **Gráfico de dispersión (*Scatterplot*) resulta muy útil.**
- **Correlación también muy útil**

Relación entre dos variables:
PESO Y ALTURA



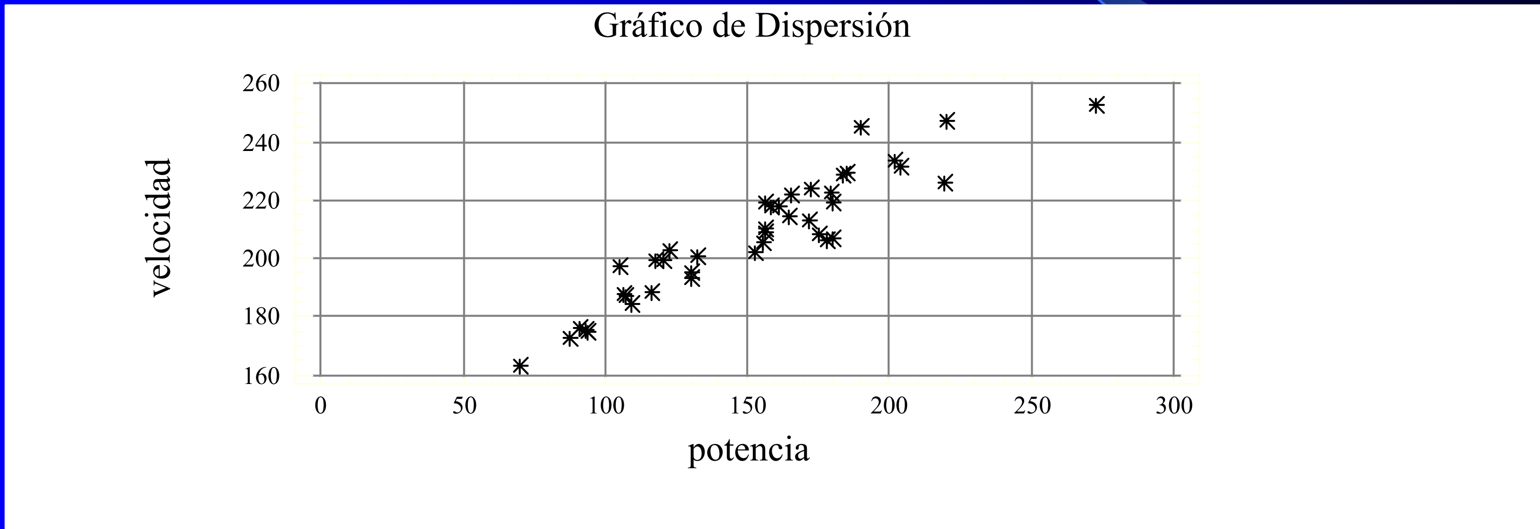
HAY RELACIÓN

No hay relación entre dos variables:



NO HAY RELACIÓN

¿Hay relación entre estas variables?



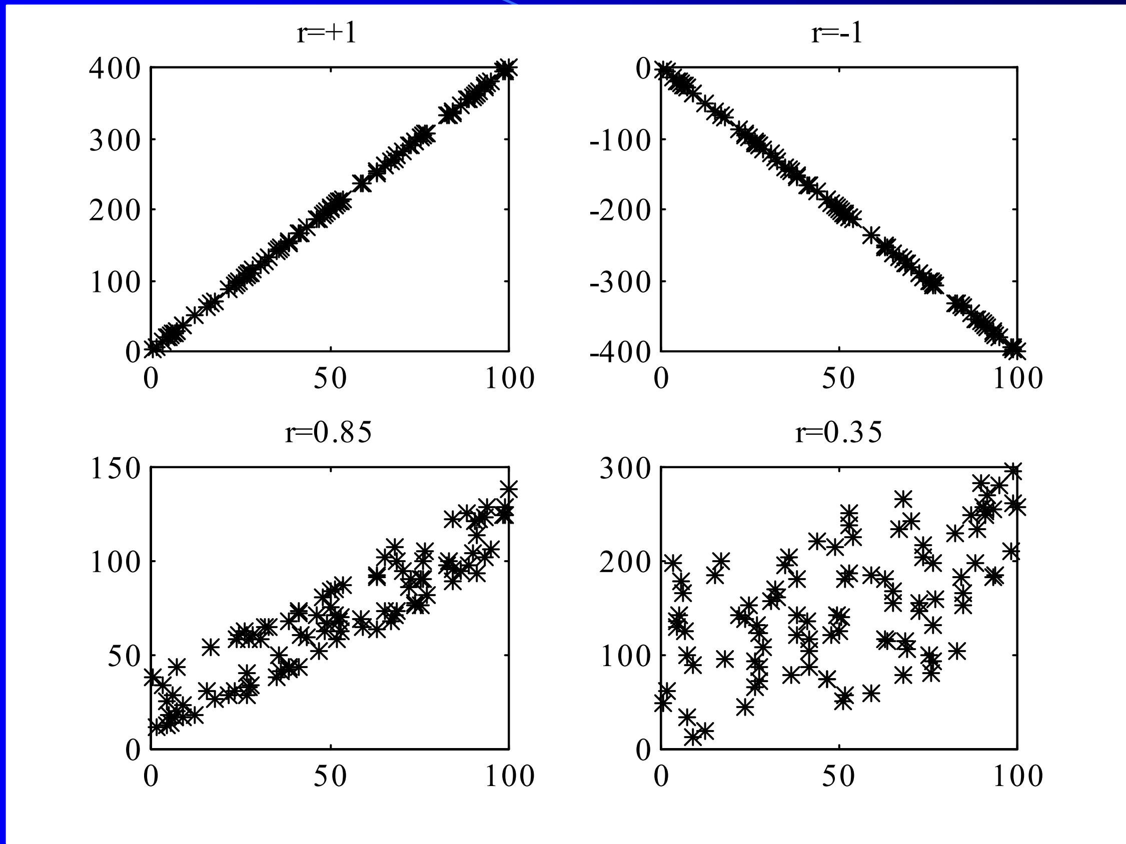
Para medir el grado de relación entre variables

- Utilizamos la correlación.
- Varía entre -1 y +1

Interpretación de la correlación:

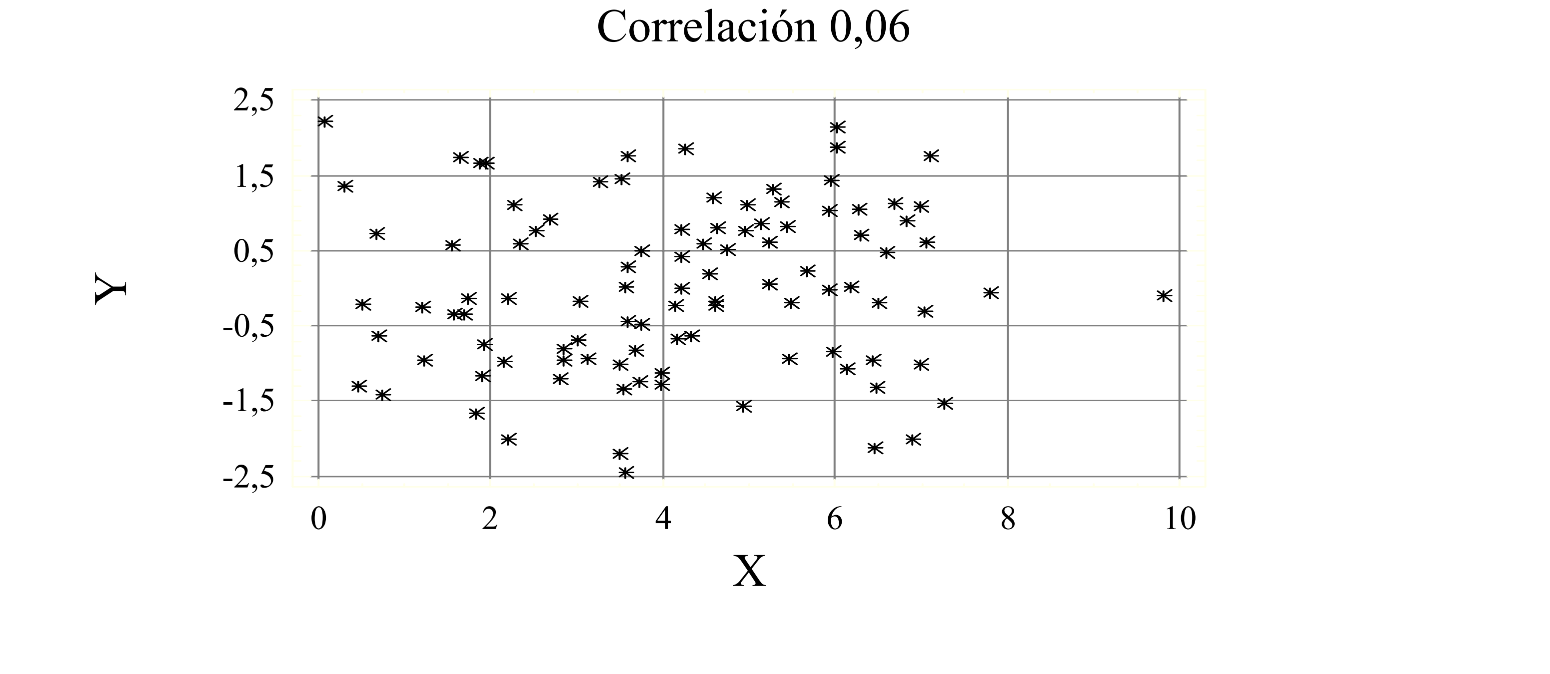
-1	0	+1
Mucha relación Decreciente	No hay relación	Mucha relación Creciente

Interpretación de la correlación



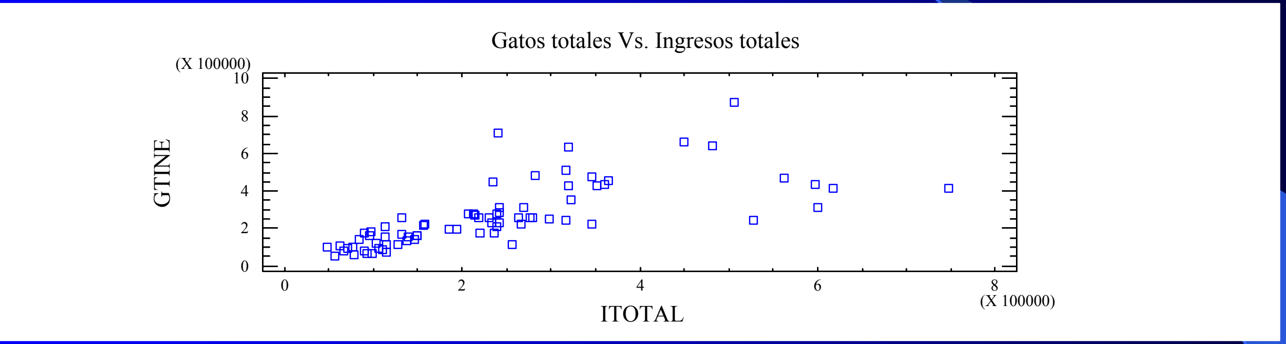
+ Relación creciente: Si una variable aumenta, la otra también
- Relación decreciente: Si una variable aumenta, la otra disminuye

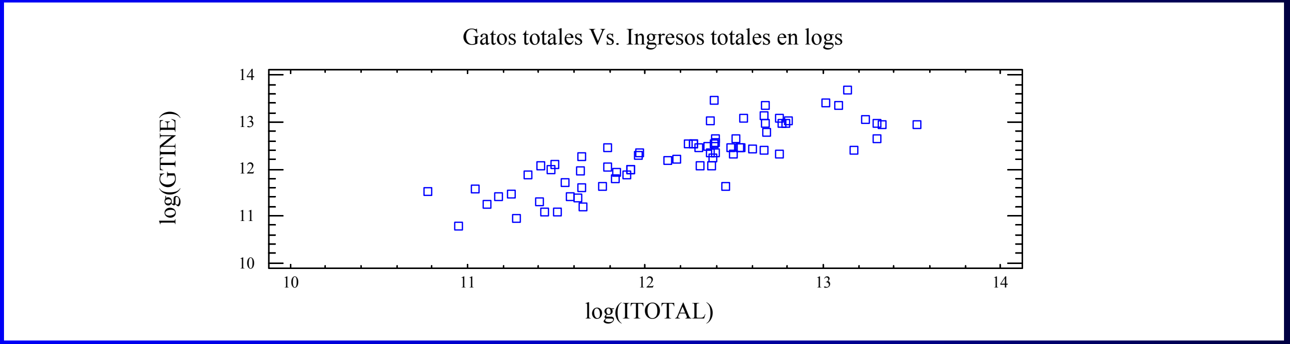
Interpretación de la correlación



Si la correlación es muy pequeña indica falta de relación entre las variables.

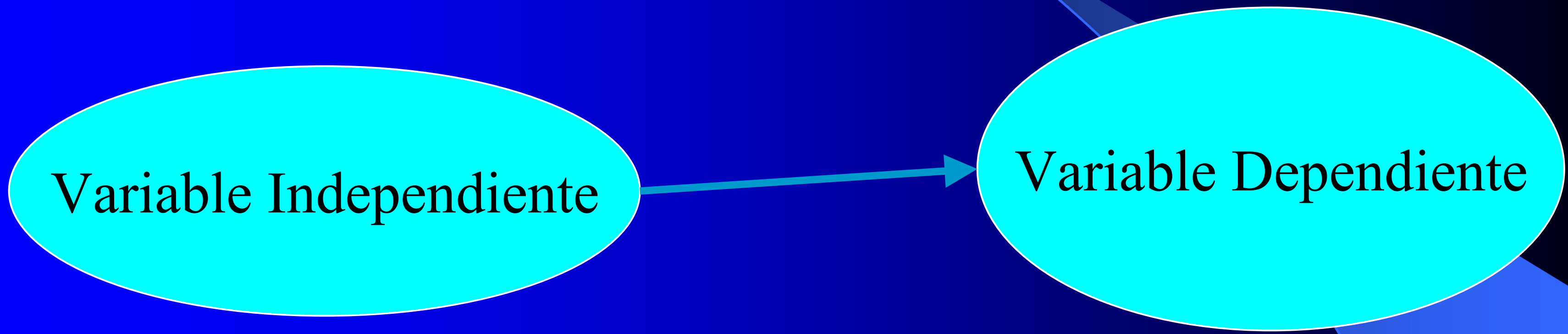
Si no se cumplen las hipótesis hay que transformar: **LOGS**





Los Logs son una transformación que indica que las variables se relacionan por su tasa de crecimiento

Sin logs



Sin logs

Altura → Peso

$$Peso = -100{,}22 + 0.97 Altura$$

Incremento de Altura de 1 cm → Incremento de Peso de 0.97x1kg

CON logs

Inflación=Tasa de variación de los precios

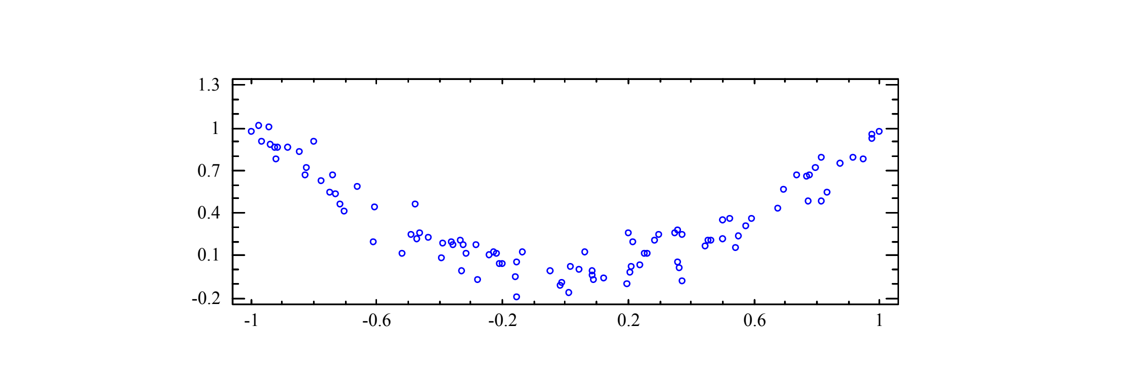
Tipos de interés= Tasa de variación del dinero

La interpretación de una regresión con logs

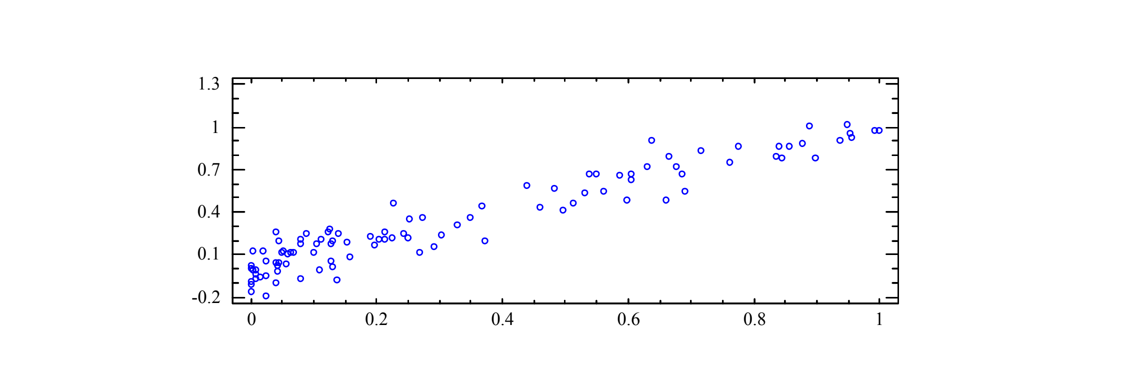
- La veremos más adelante

Si no se cumplen las hipótesis hay que transformar: Cuadrado

Y-X

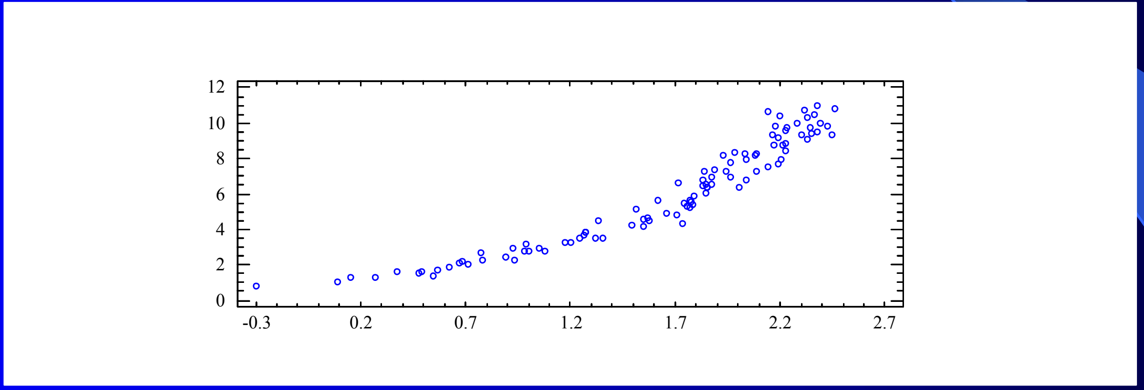


Y-X^2

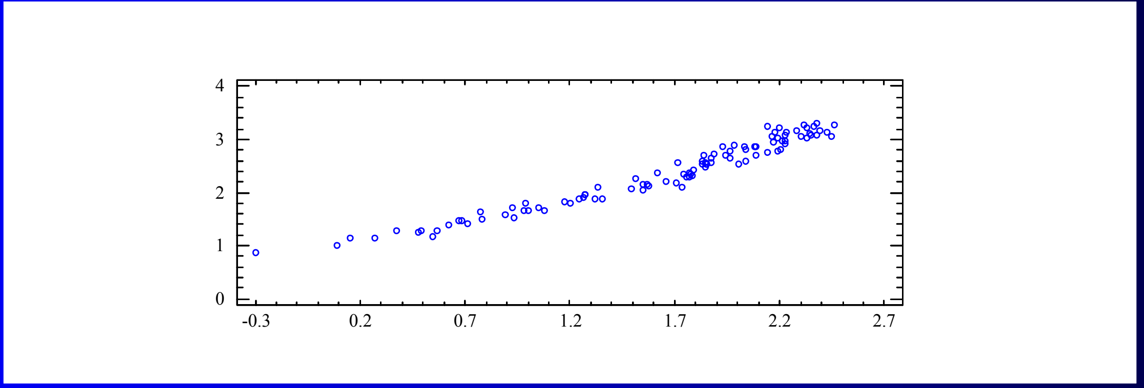


Si no se cumplen las hipótesis hay que transformar: Raíz cuadrada

Y-X

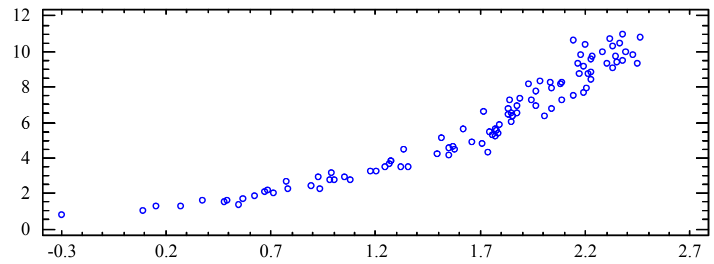


Y- $\sqrt{x}$

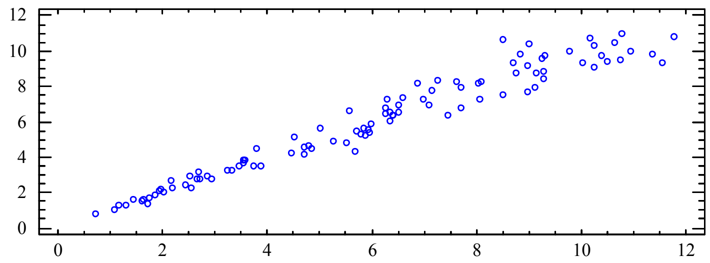


Si no se cumplen las hipótesis hay que transformar: exponencial

Y-X



Y-exp(X)



Las tranformaciones:

- Logaritmos: Muy importante
- Resto: Poco importante

Regresión: Estimación

$$\hat{\beta}_1 = \frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

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

$$\hat{S}_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

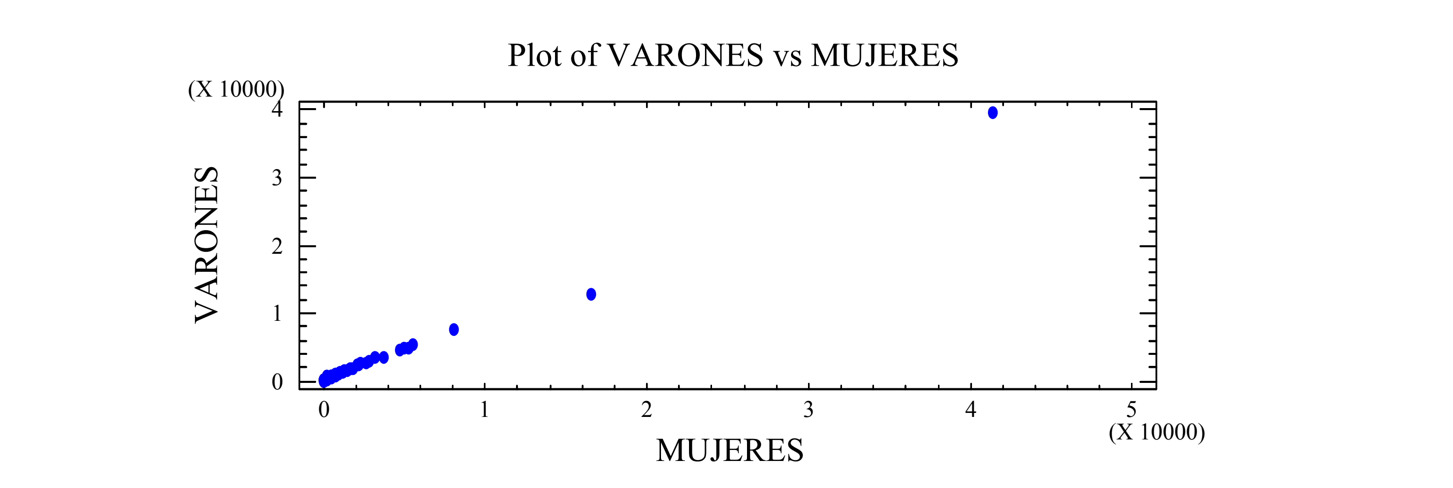
Oh! Cielos

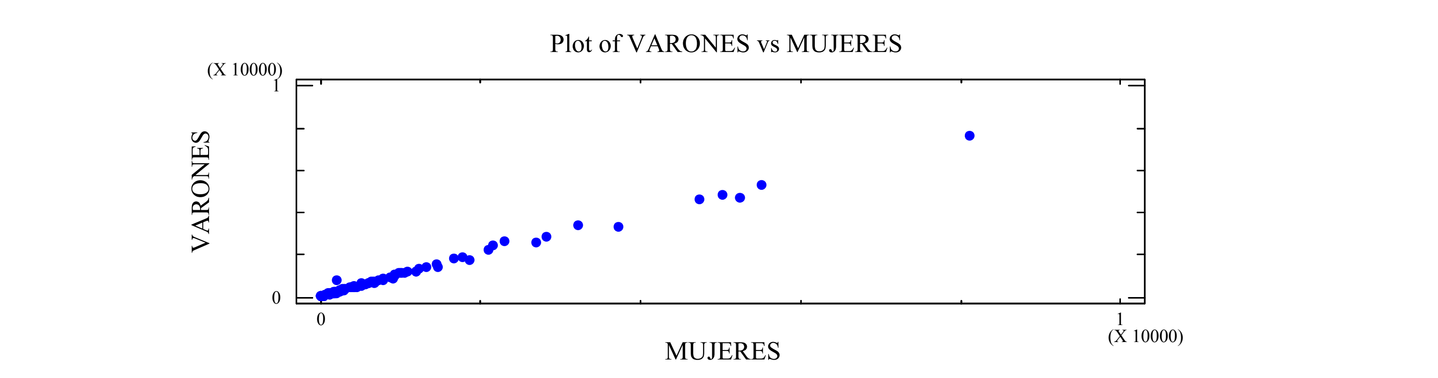
Que horror!!!!!!

Esas fórmulas tan bonitas no se usan.

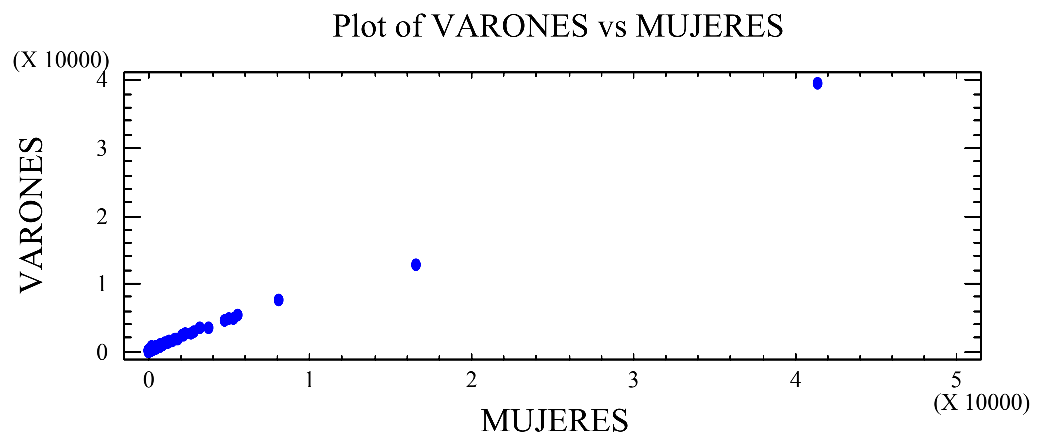
Para trabajar tenemos el ordenador!!!!!!!!!!!!!!

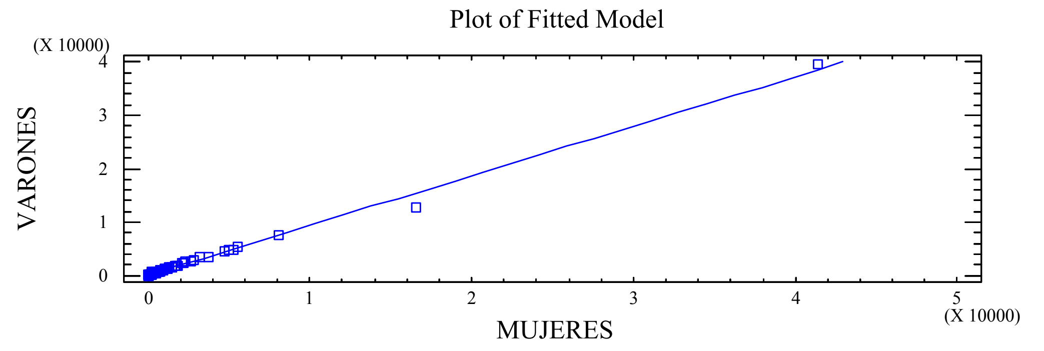
Datos: Censo de Floridablanca 1787
Provincia de Sevilla



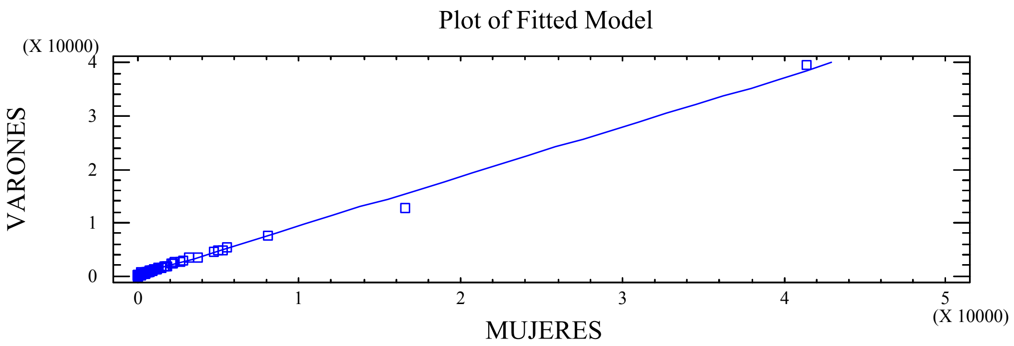


Datos: Censo de Floridablanca 1787
Provincia de Sevilla





Datos: Censo de Floridablanca 1787
Provincia de Sevilla




```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: VARONES
Independent variable: MUJERES
-----------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate       Error        Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        77,7958       31,4564        2,47313         0,0150
Slope             0,9293    0,00675621        137,547         0,0000
-----------------------------------------------------------------------------
```

$$\hat{\beta}_1 = 0.93$$

$$\hat{\beta}_0 = 77.8$$

Varones=77.8+0.93 Mujeres

Interpretación del modelo:

Varones = 77.8 + 0.93 Mujeres

1. **Signo del estimador beta$_1$:**

   Información sobre la relación entre X e Y.

   Positivo: Si X aumenta Y también aumenta.

   Negativo: Si X aumenta Y disminuye.

Al aumentar el número de mujeres aumenta el de varones

Interpretación del modelo:

Varones = 77.8 + 0.93 Mujeres

Valor del estimador beta$_1$:

Información sobre cómo se transmite el efecto de X sobre Y.

Si X aumenta, el efecto se transmite a Y multiplicado por beta$_1$

$$\Delta X \Rightarrow \Delta Y = \hat{\beta}_1 * \Delta X$$

Si un pueblo tiene 100 mujeres más

Tendrá 0.93 x 100 = 93 Hombres más

Resultados

Varones = 77.8 + 0.93 Mujeres

$Beta_1$ = 0.93   Si fuera 1, a 100 mujeres más le corresponderían 100 hombres más.

¿Puede valer 1 $Beta_1$?

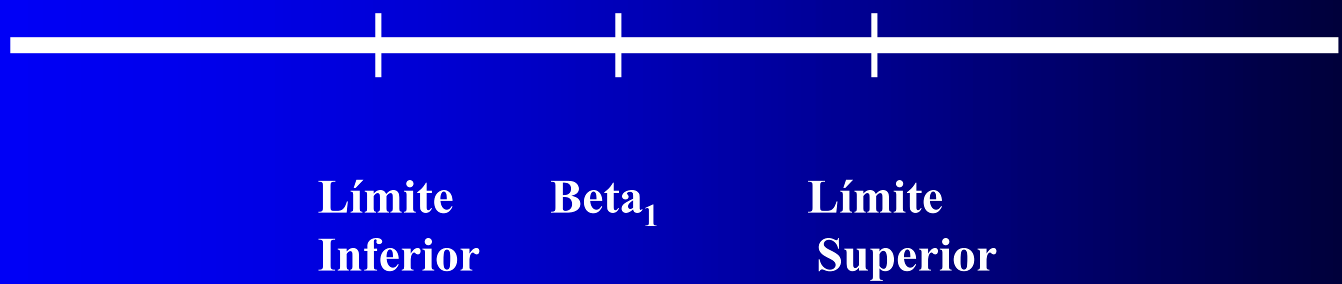
¿Puede valer 1 $Beta_1$?

<u>Solución:</u>

- Intervalo de confianza
- Contraste de hipótesis

INTERVALOS DE CONFIANZA

Un intervalo de confianza proporciona una zona en la que con una confianza predeterminada estará el auténtico valor de Beta$_1$

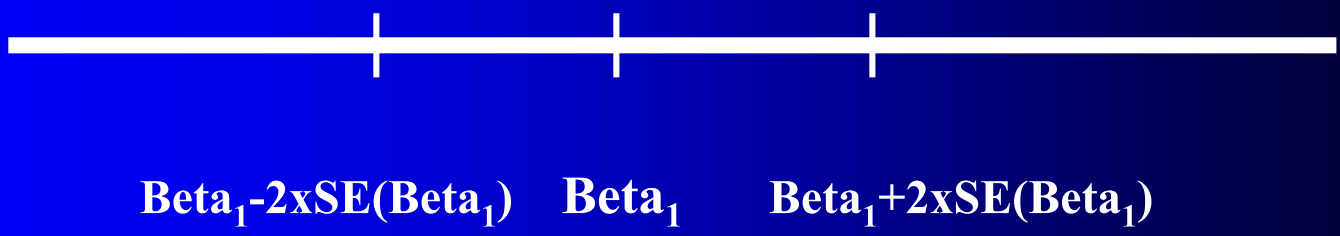


INTERVALOS DE CONFIANZA

Se calcula:

$Beta_1$-2xSE($Beta_1$) $Beta_1$+2xSE($Beta_1$)

El SE($Beta_1$) se denomina Error estándar de $Beta_1$ y lo calcula el programa



INTERVALOS DE CONFIANZA

Se calcula:

$$Beta_1\text{-}2xSE(Beta_1) \qquad Beta_1\text{+}2xSE(Beta_1)$$

El SE($Beta_1$) se denomina Error estándar de $Beta_1$ y lo calcula el programa

Regression Analysis - Linear model: Y = a + b*X

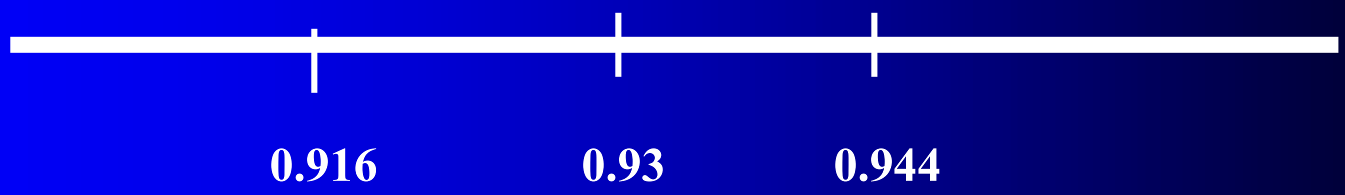
Dependent variable: VARONES
Independent variable: MUJERES

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	77,7958	31,4564	2,47313	0,0150
Slope	0,9293	0,00675621	137,547	0,0000

$$Beta_1\text{-}2xSE(Beta_1) \qquad Beta_1\text{+}2xSE(Beta_1)$$

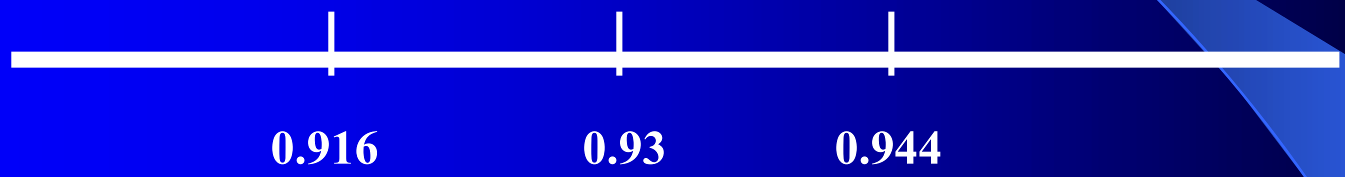
$$0.93-2x0.007 \qquad 0.93+2x0.007$$

$$0.916 \qquad 0.944$$



INTERVALOS DE CONFIANZA

$Beta_1$-2xSE($Beta_1$)	$Beta_1$+2xSE($Beta_1$)
0.93-2x0.007	0.93+2x0.007
0.916	0.944



¿Puede valer 1 Beta1?

Consideraciones sobre el error estándar SE(Beta$_1$)

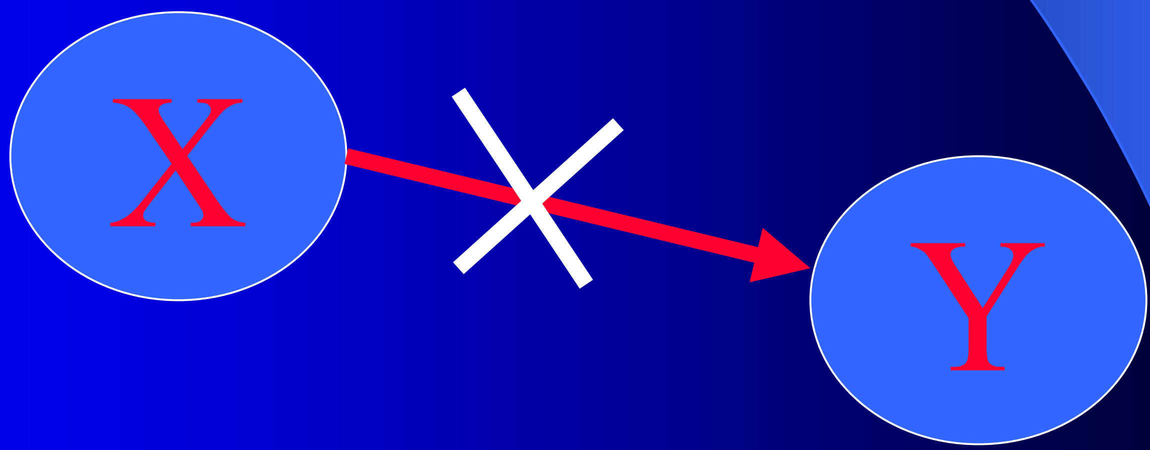
La amplitud del intervalo de confianza depende del SE:

Si SE es grande aumenta la imprecisión:

- S_R grande genera más imprecisión
- n grande más precisión
- S_x grande más precisión

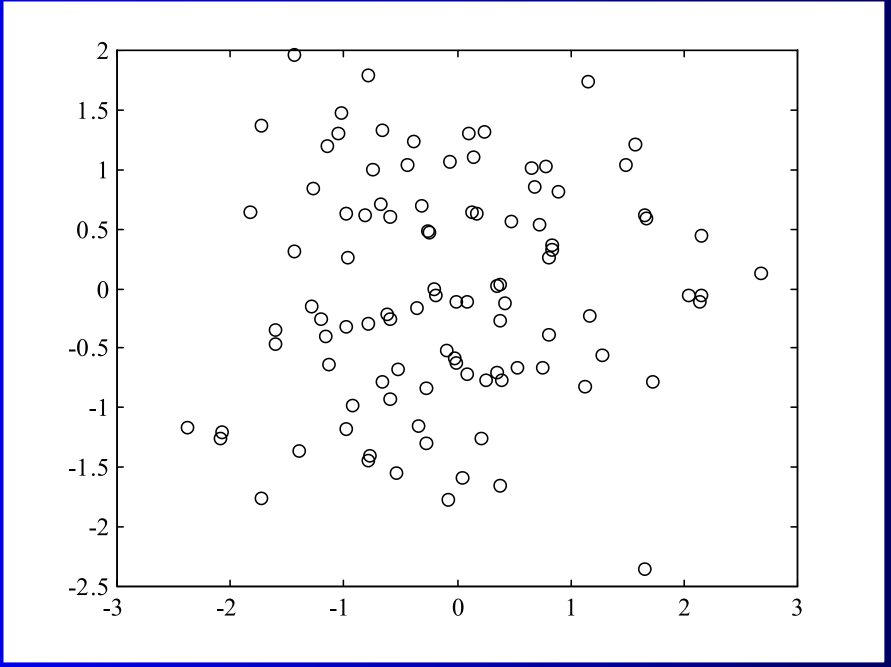
Si Beta$_1$ auténtica es cero.....
Beta$_1$=0

- Implica que Y no depende de X



Si Beta$_1$ auténtica es cero.....
Beta$_1$=0

- Implica que Y no depende de X



Si $Beta_1$ auténtica es cero.....
$Beta_1=0$

- **Implica que Y no depende de X**
- **Lo sabemos construyendo un intervalo de confianza:**
  - **Si el cero está dentro del intervalo, $Beta_1$ puede valer cero. *Variable X no es significativa***
  - **Si el cero está fuera, $Beta_1$ no puede valer cero. *Variable X es significativa***

Hay que comprobar si $Beta_1$ auténtica puede ser cero.....
$Beta_1=0$

- Mediante un intervalo de confianza
- Mediante un contraste t.

- Para los datos de Sevilla:

$Beta_1-2xSE(Beta_1)$	$Beta_1+2xSE(Beta_1)$
0.93-2x0.007	0.93+2x0.007
0.916	0.944

¿¿¿¿¿Puede valer cero?????

Contrastes de hipótesis: Contraste t

- Una forma rápida y sencilla de saber si $Beta_1$ auténtico puede valer cero es utilizar el contraste t
- El contraste t lo proporciona el ordenador.

$$H_0: \ Beta_1 = 0$$

$$H_1: \ Beta_1 \text{ distinto de } 0$$

- Si $t<2$ sin importar el signo. Nos quedamos con H_0.
- Si $t>2$ sin importar el signo. Nos quedamos con H_1.

Contraste t

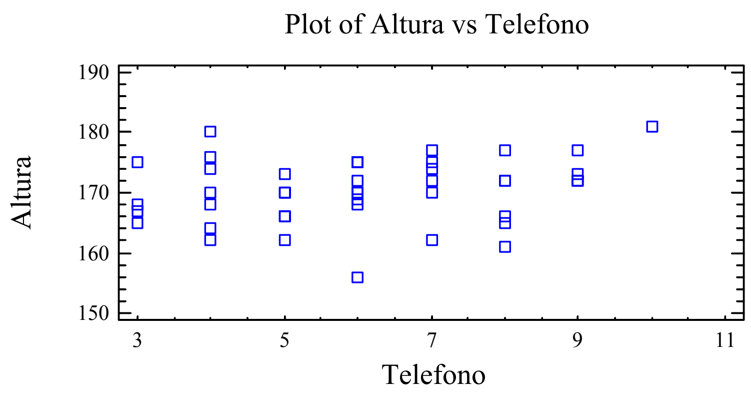
```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: VARONES
Independent variable: MUJERES
-----------------------------------------------------------------------------
                               Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept        77,7958        31,4564       2,47313          0,0150
Slope             0,9293     0,00675621       137,547          0,0000
-----------------------------------------------------------------------------
```

- T de $Beta_1$ es 137, que es mayor que 2.
- Por tanto el número de mujeres es significativo:

El número de mujeres aporta información sobre el

número de varones.

Ejemplo: Altura vs última cifra del teléfono



Regression Analysis - Linear model: Y = a + b*X

Dependent variable: Altura
Independent variable: Telefono

Parameter	Estimate	Standard Error	T Statistic	P-Value
Intercept	166,317	2,65279	62,6952	0,0000
Slope	0,66185	0,414182	1,59797	0,1175

Altura=166.3+0.66Telefono

Si el teléfono sube una unidad se crece 0.66cm

?????????????????????????????????

Siempre nos fijamos en la t

- Si es menor que 2 (No importa el signo)…..
- …….la variable X o influye sobre Y

R^2

Indica cuánto de Y es explicado por X

Datos de Sevilla:

R^2=99.4%

Resumen

- Estudiamos los datos y vemos si cumplen las hipótesis.
- Si no las cumplen transformamos.
- Ajustamos el modelo.
- Intervalos y contrastes para ver si X es significativa (INFLUYE) sobre Y

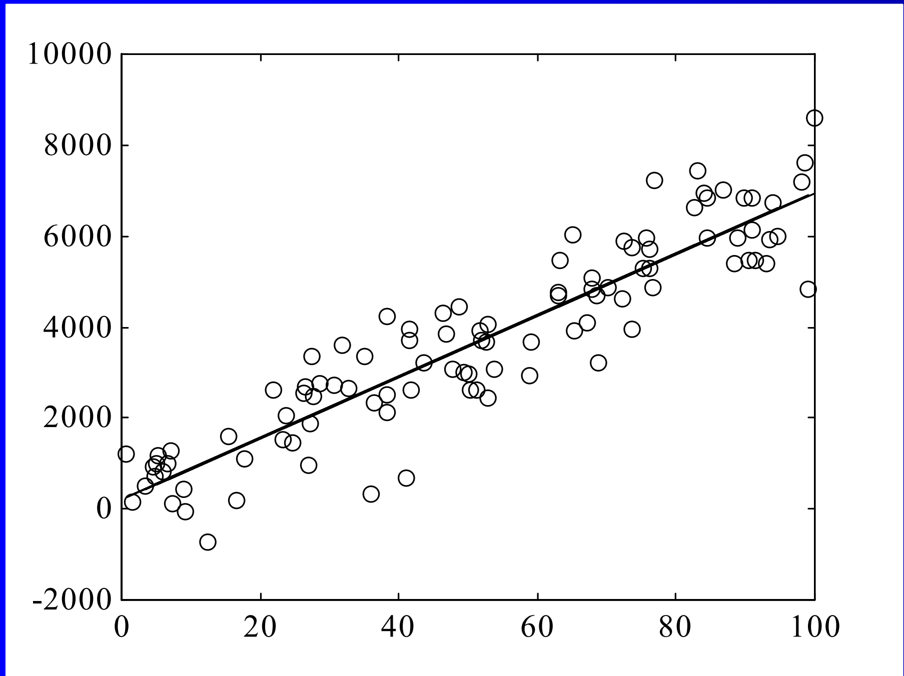
Diagnosis

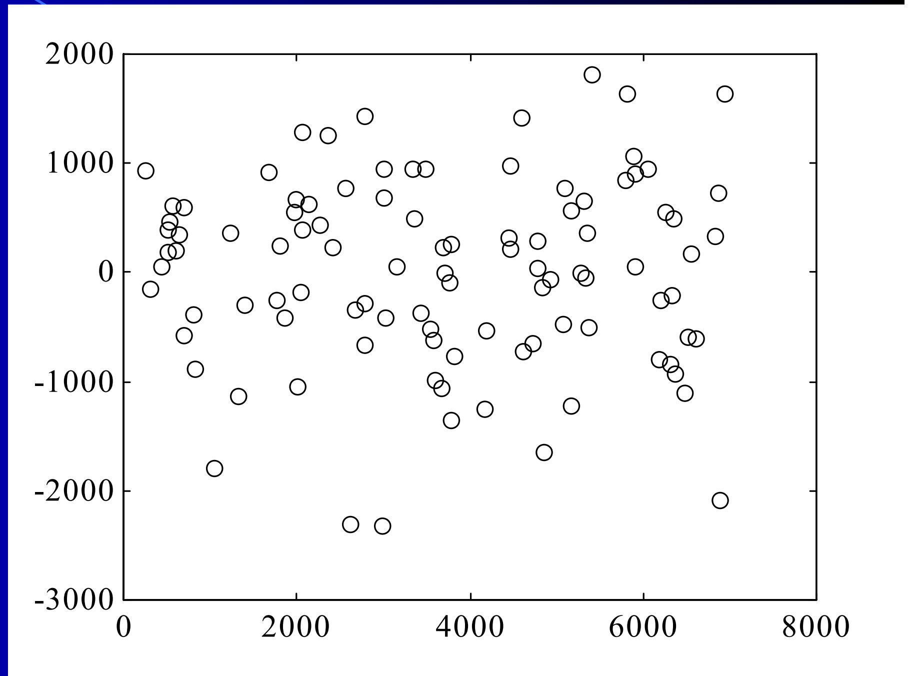
- La diagnosis sirve para ver si se cumplen las hipótesis del modelo a posteriori. Hay que comprobar básicamente:
  - Linealidad
  - Homocedasticidad

Lo hacemos mediante

- *Gráfico de residuos frente a valores predichos*
- **Este gráfico si el modelo está bien ajustado no debe presentar ninguna estructura.**

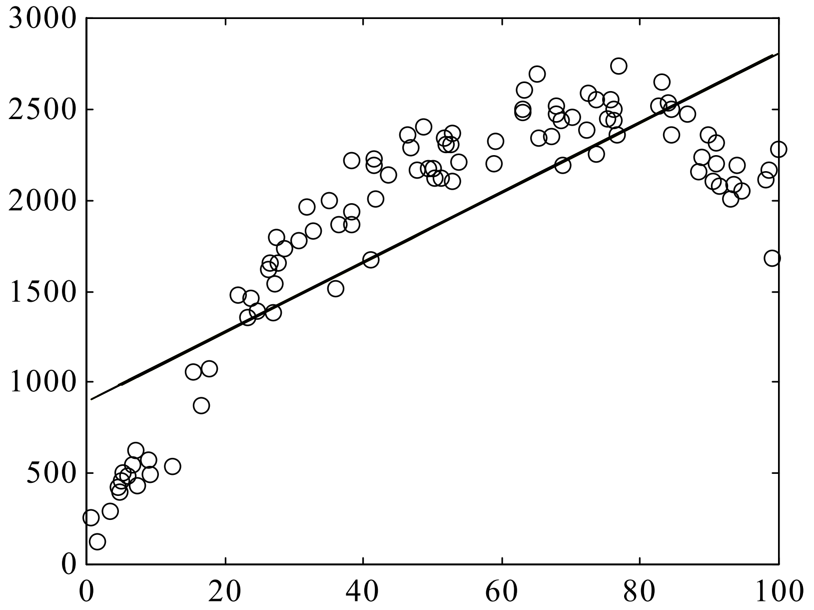
Ejemplo: Datos Lineales

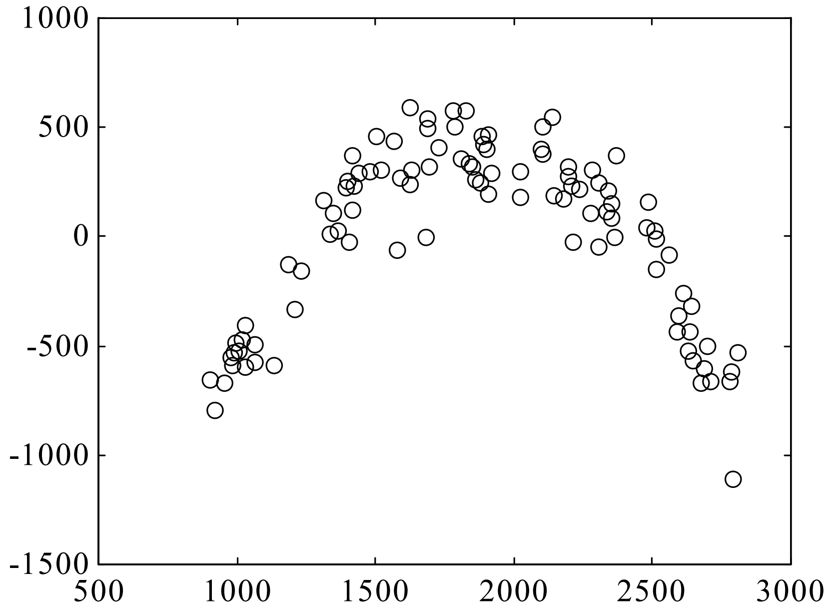




Residuos "Gotelet"

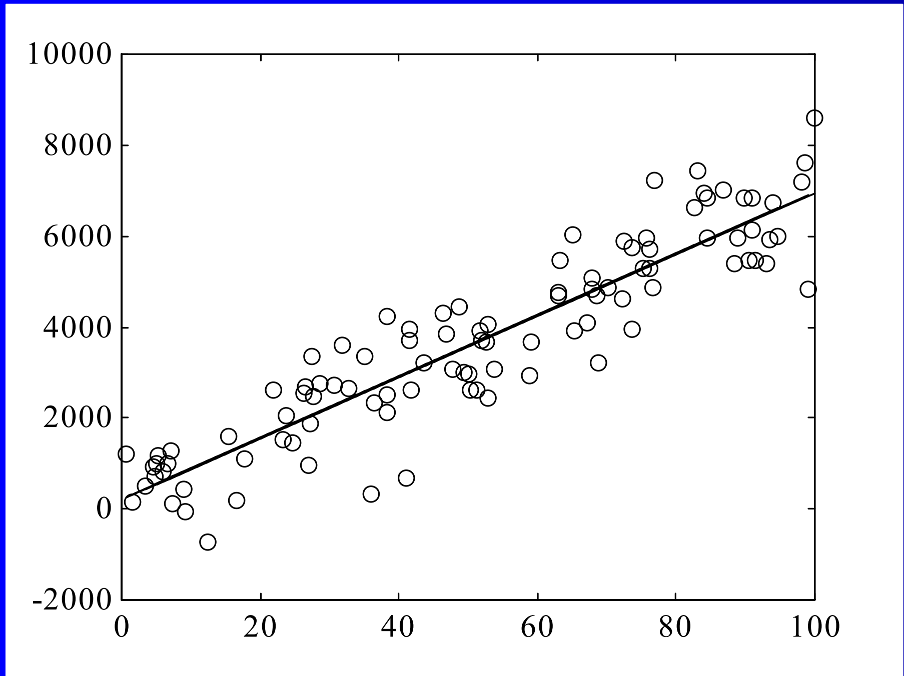
Ejemplo

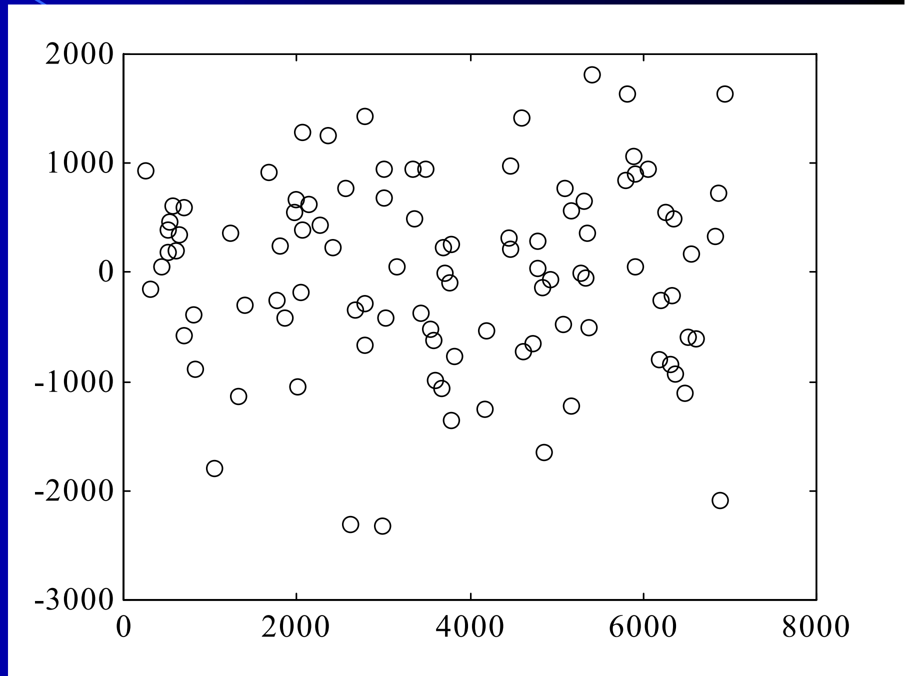




Residuos no “Gotelet”

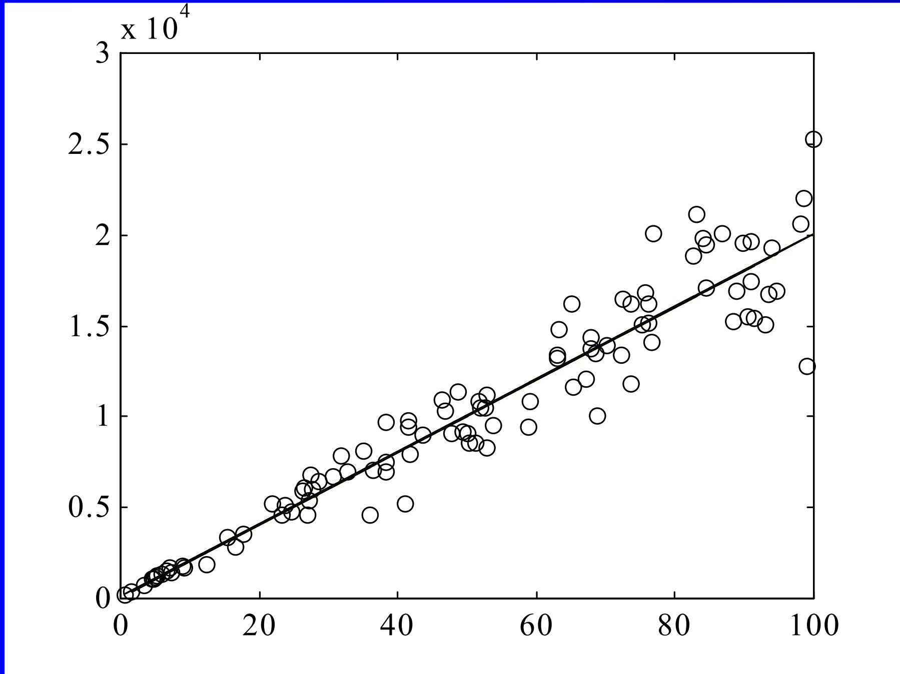
Ejemplo: Datos Homocedásticos

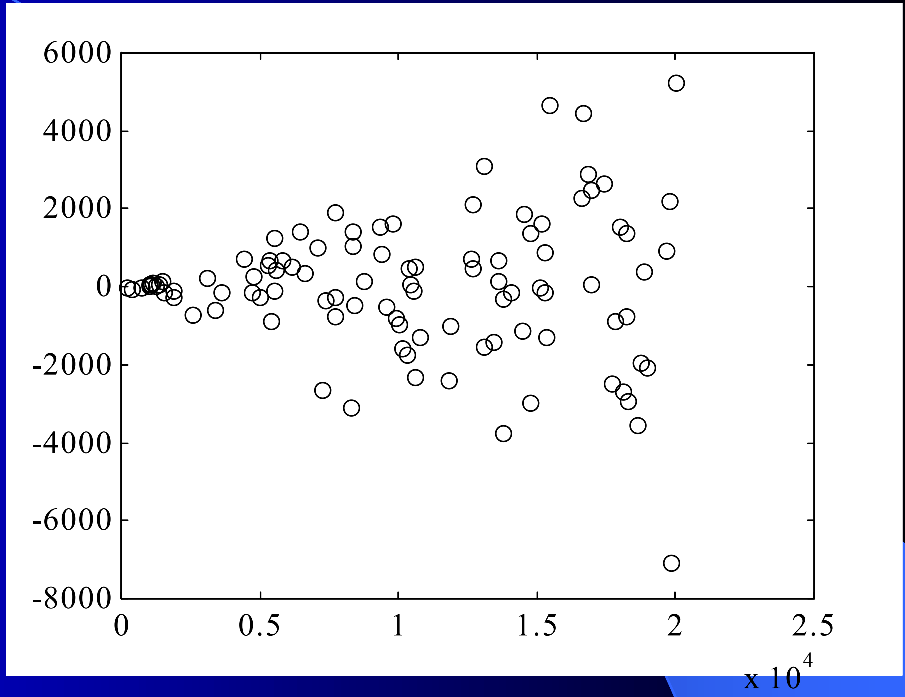




Residuos "Gotelet"

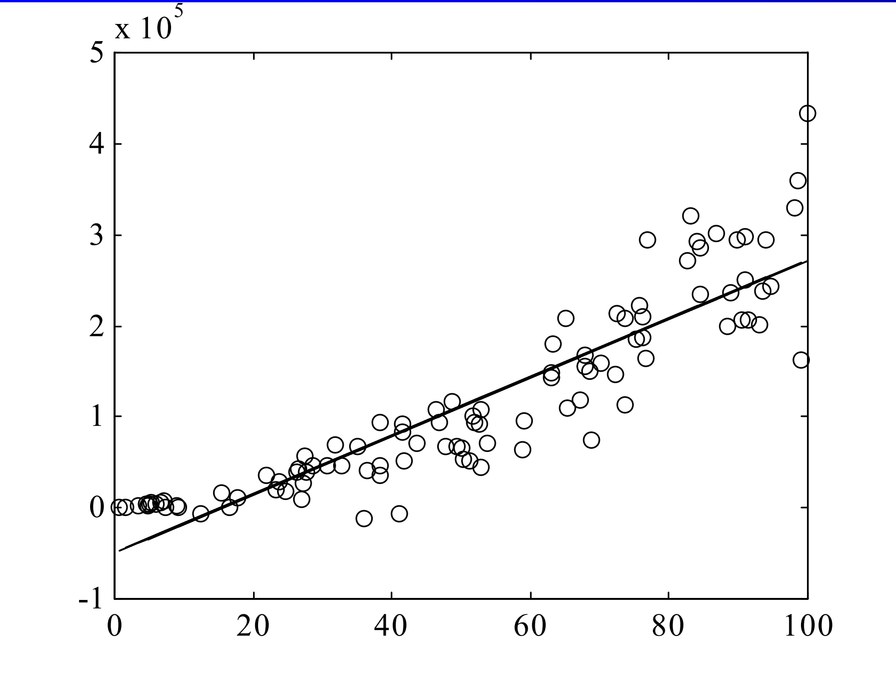
Ejemplo

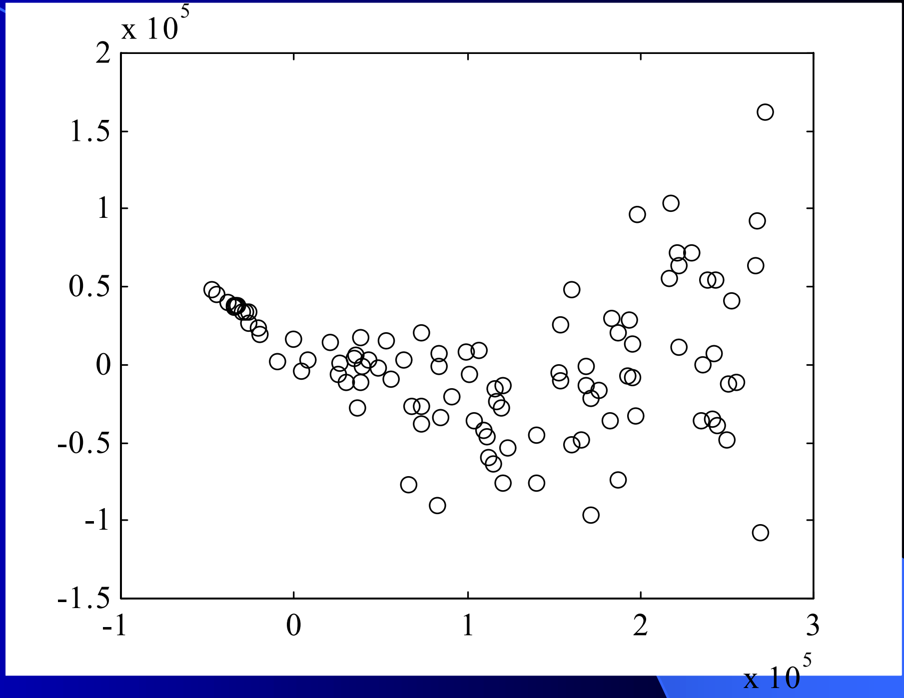




Residuos no “Gotelet”

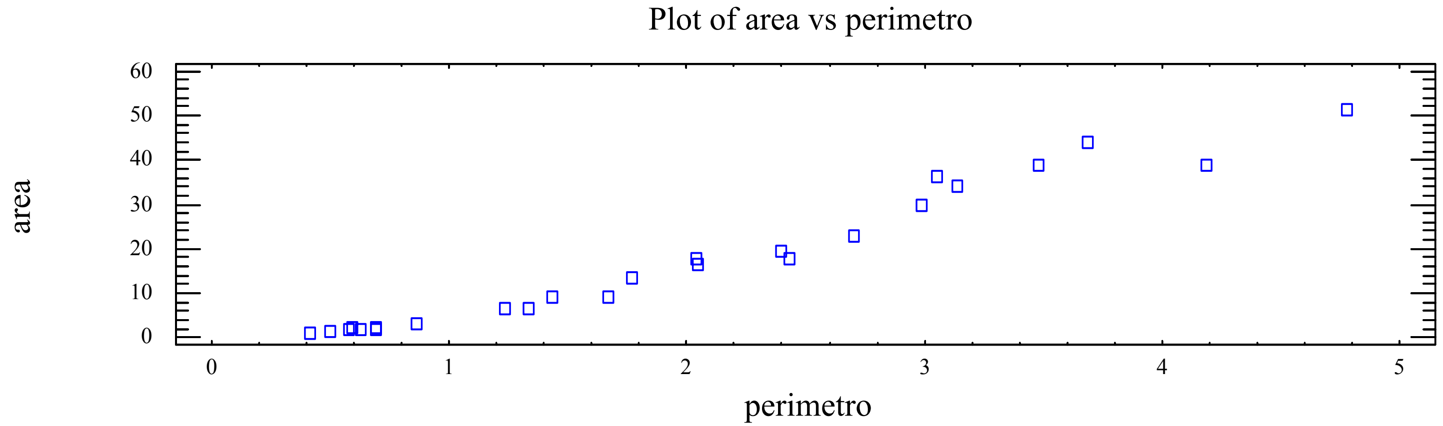
Ejemplo

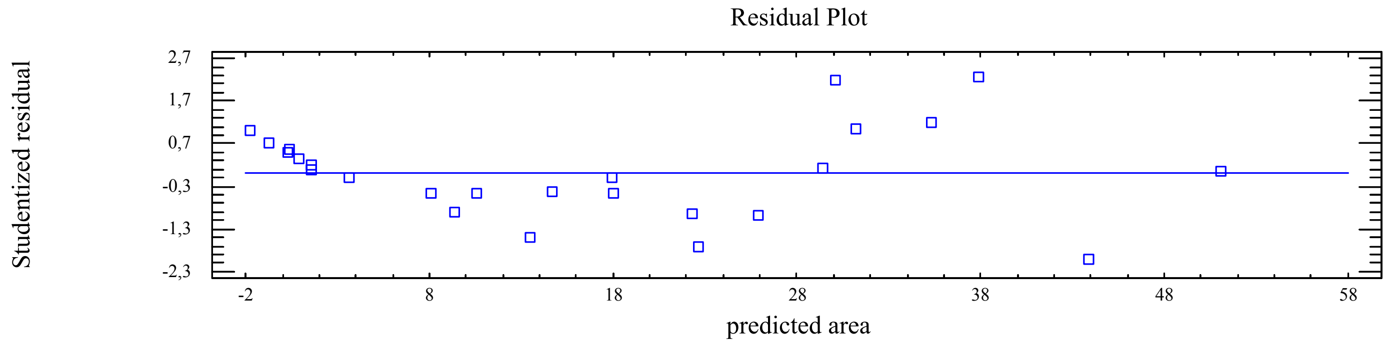


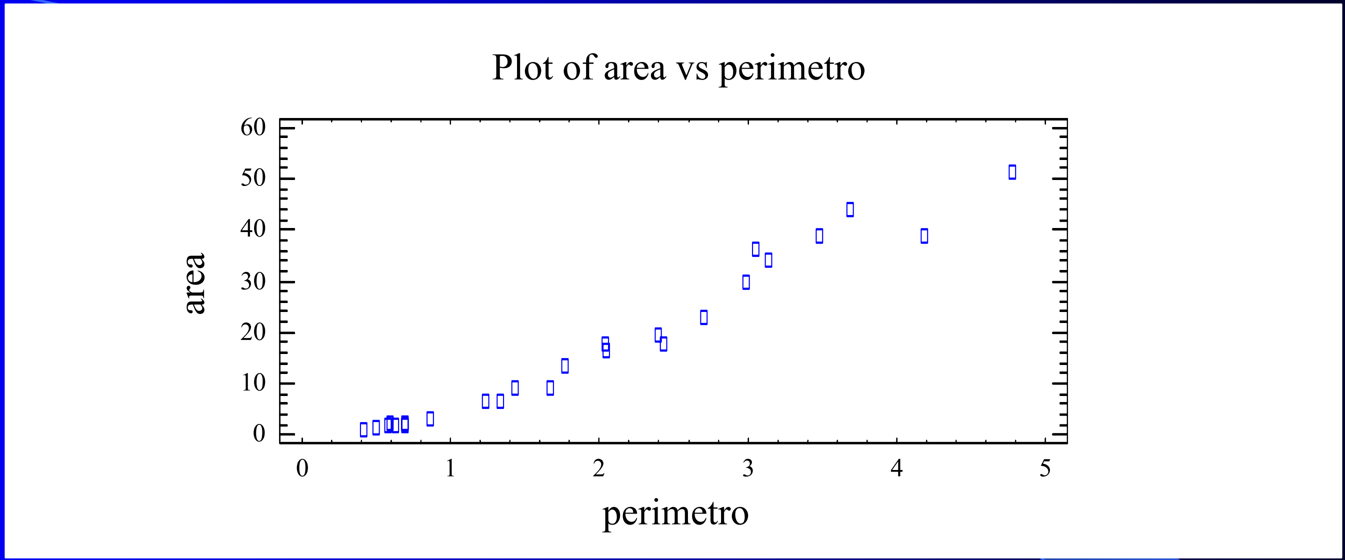


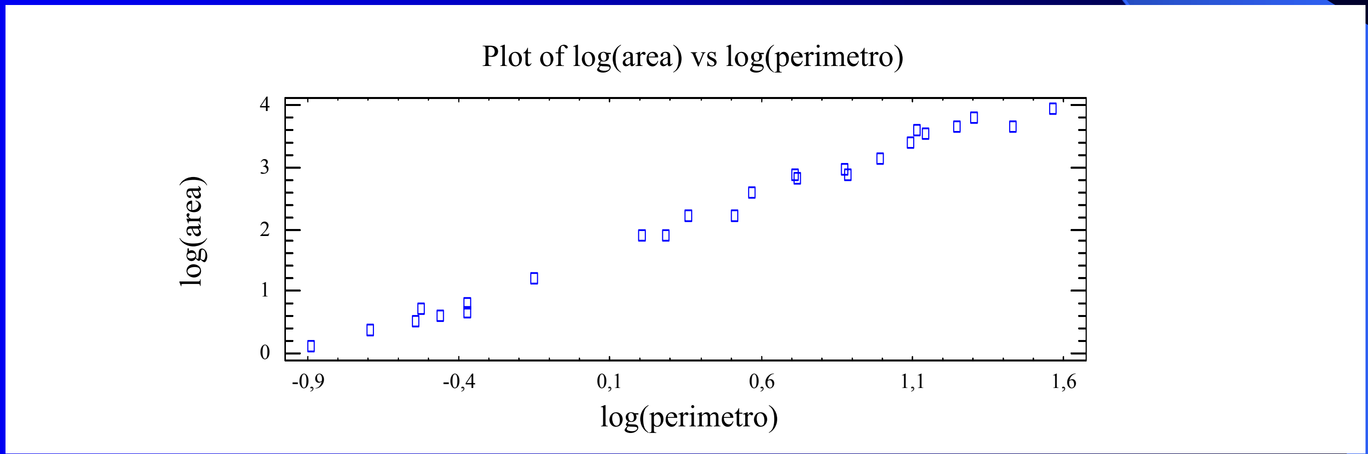
Residuos no "Gotelet"

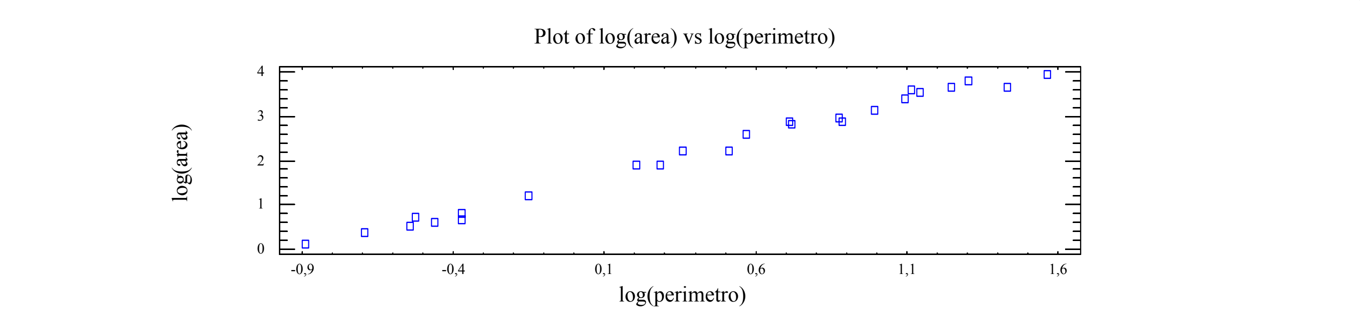
Perímetro y área de iglesias

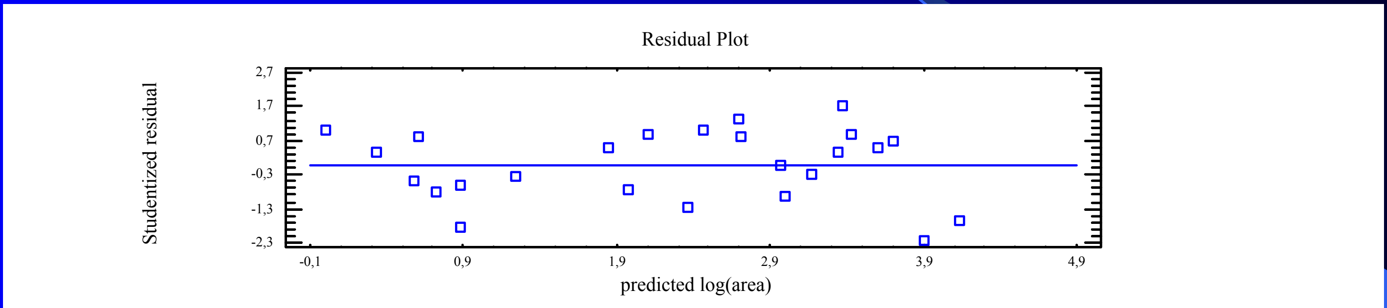





Plot of area vs perimetro
area
perimetro
0
10
20
30
40
50
60
0
1
2
3
4
5


Plot of log(area) vs log(perimetro)
log(area)
log(perimetro)
0
1
2
3
4
-0,9
-0,4
0,1
0,6
1,1
1,6






```
Regression Analysis - Linear model: Y = a + b*X
-----------------------------------------------------------------------------
Dependent variable: log(area)
Independent variable: log(perimetro)
-----------------------------------------------------------------------------
                                Standard          T
Parameter        Estimate         Error       Statistic        P-Value
-----------------------------------------------------------------------------
Intercept         1,49907       0,0306395       48,9261         0,0000
Slope             1,68335        0,035831       46,9802         0,0000
-----------------------------------------------------------------------------
```

Log(area)=1.5 + 1.68 Log(perimetro)

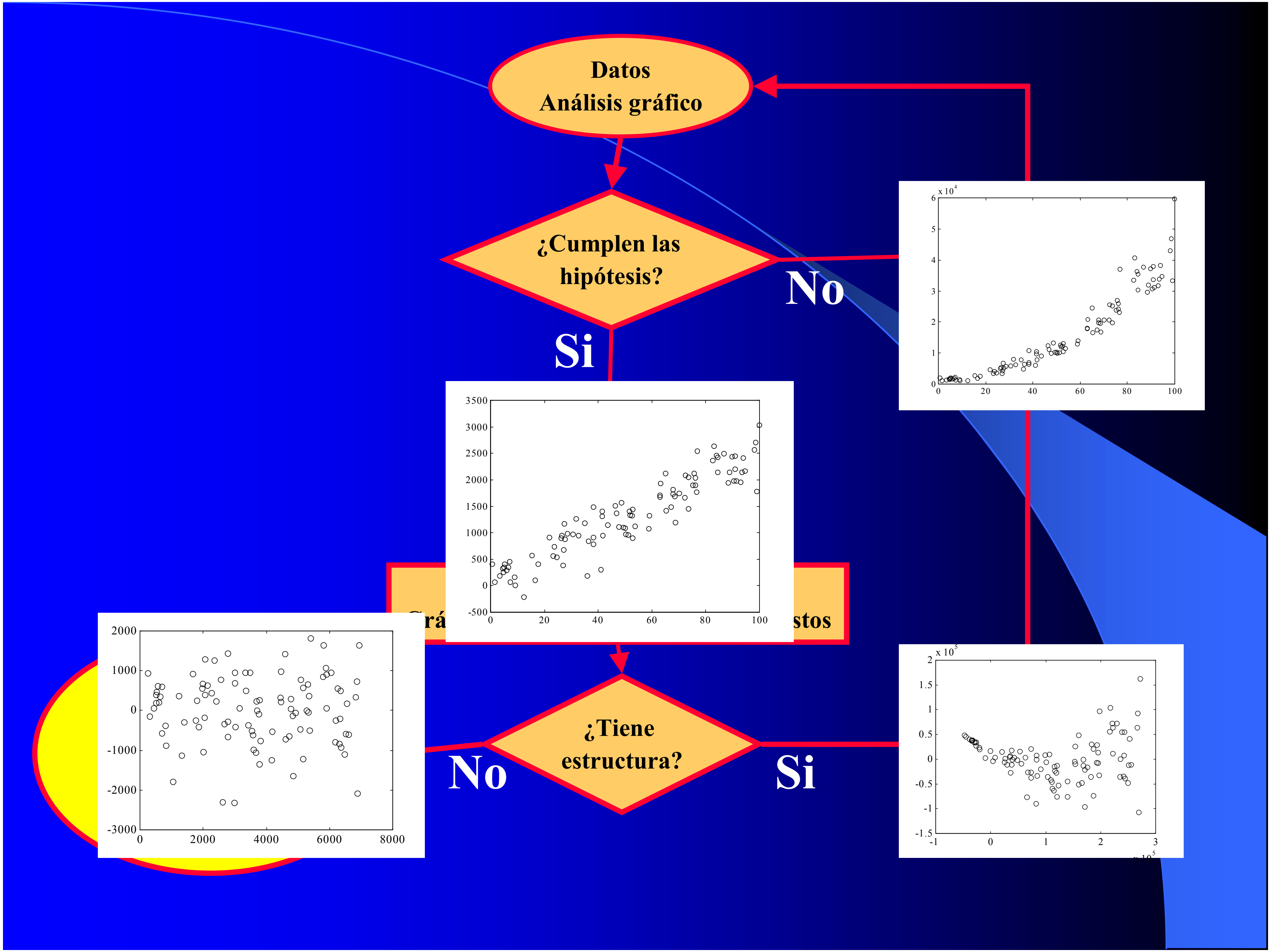
(48.9) (46.9) **R^2=98.9%**

$$\text{Log(area)} = 1.5 + 1.68\ \text{Log(perimetro)}$$
$$(48.9) \quad (46.9)$$

Cuando hay logaritmos:

Incrementos porcentuales. Si el perímetro se incrementa un 10%, el area se incrementará un

$$1.68 \times 10 = 16.8\ \%$$



Datos
Análisis gráfico
¿Cumplen las hipótesis?
No
Si
¿Tiene estructura?
No
Si

FIN de Regresión Simple

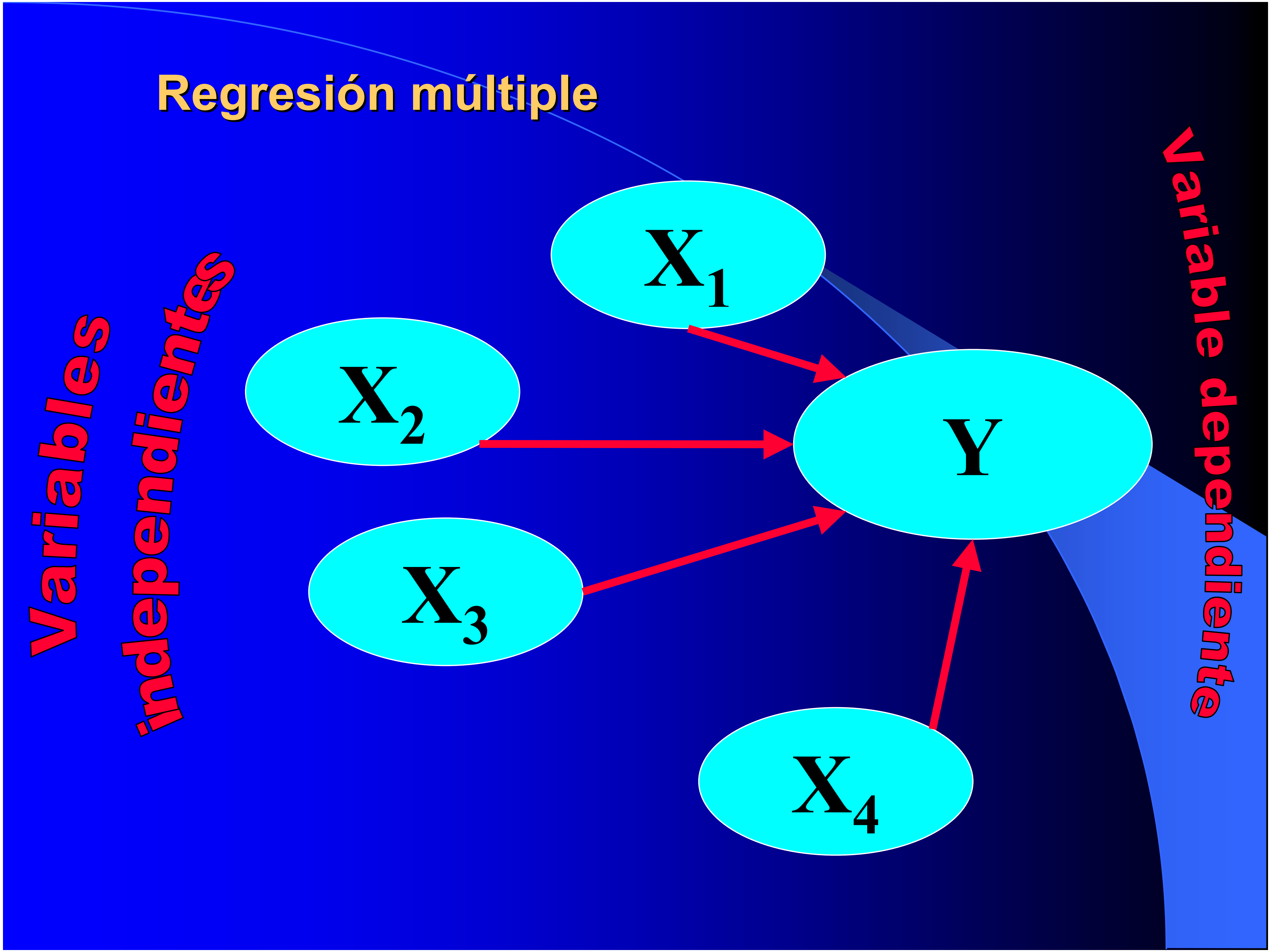
Regresión Múltiple

En regresión simple:

- Y es explicada por una sola variable

En regresión múltiple:

- Y es explicada por más variables


Regresión múltiple
Variables independientes
X_1
X_2
X_3
X_4
Y
Variable dependiente

La ecuación de regresión será:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_k X_k + u$$

Las β representan la influencia (PESOS) que las variables X tienen sobre Y

Las hipótesis en regresión múltiple

Son idénticas a las de regresión simple:

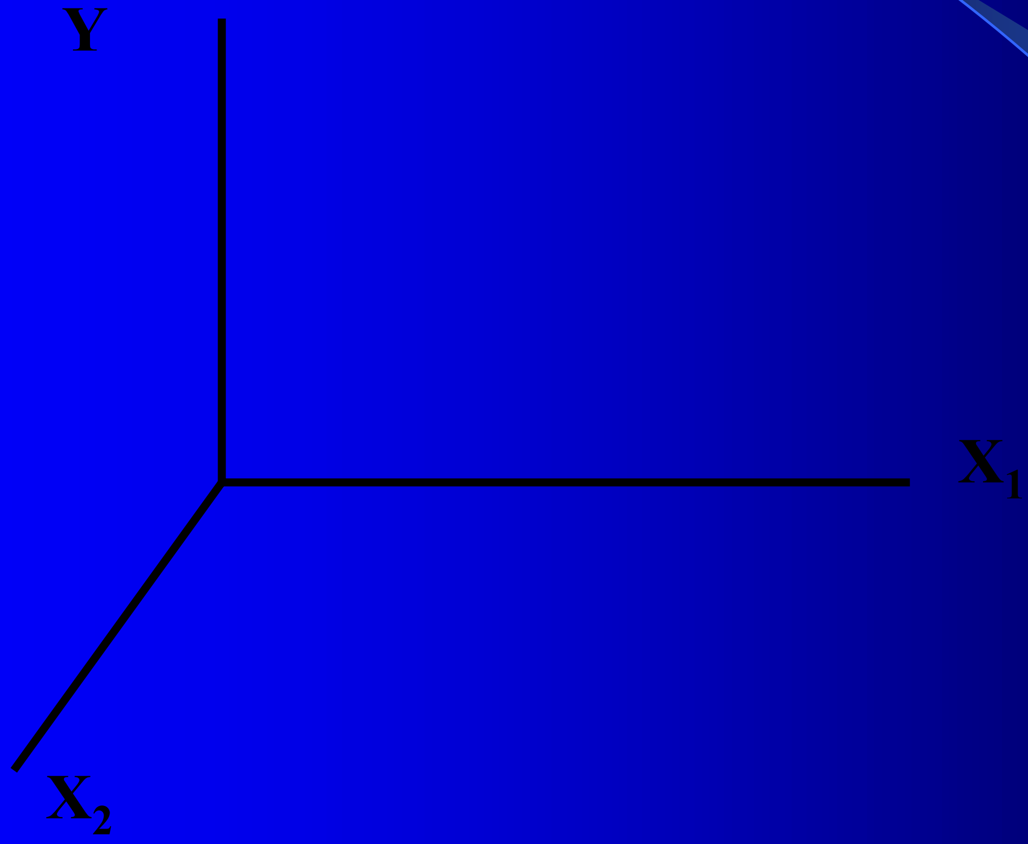
1. **Linealidad**
2. **Homocedasticidad**
3. **Independencia**
4. **Normalidad**

Si tenemos dos variables explicativas:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

Las hipótesis indican:

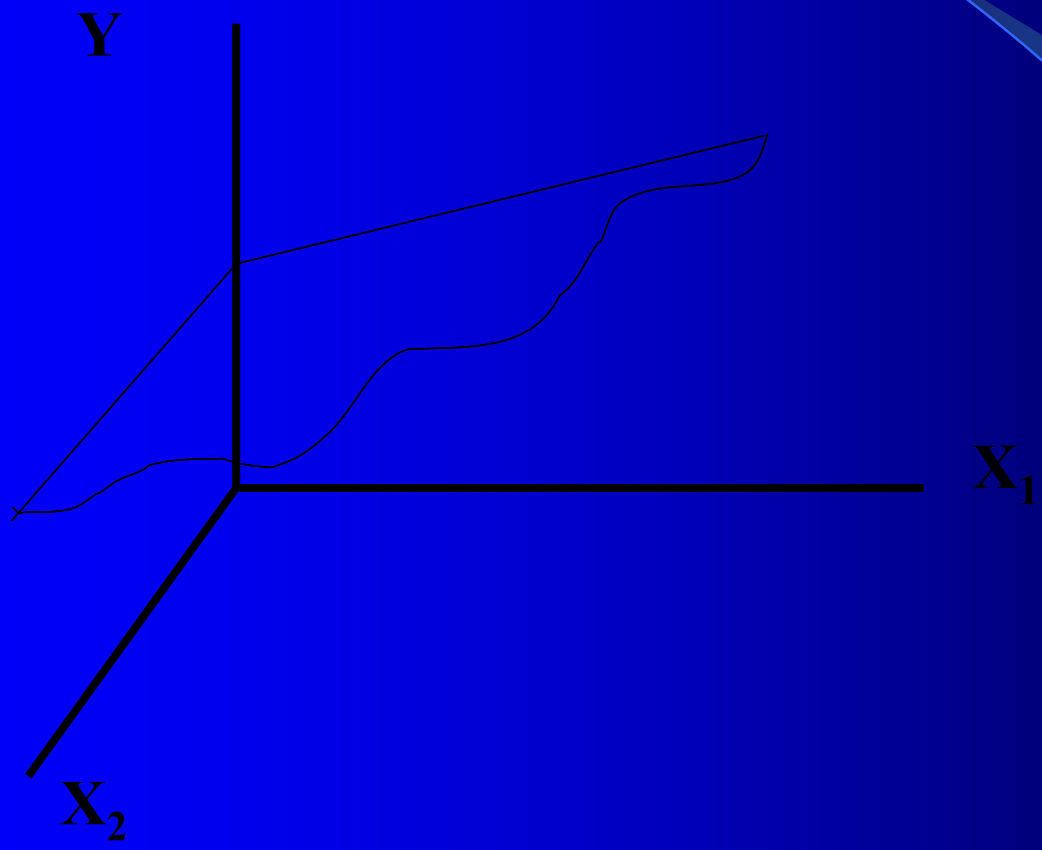
Hipótesis



Hipótesis

Linealidad:

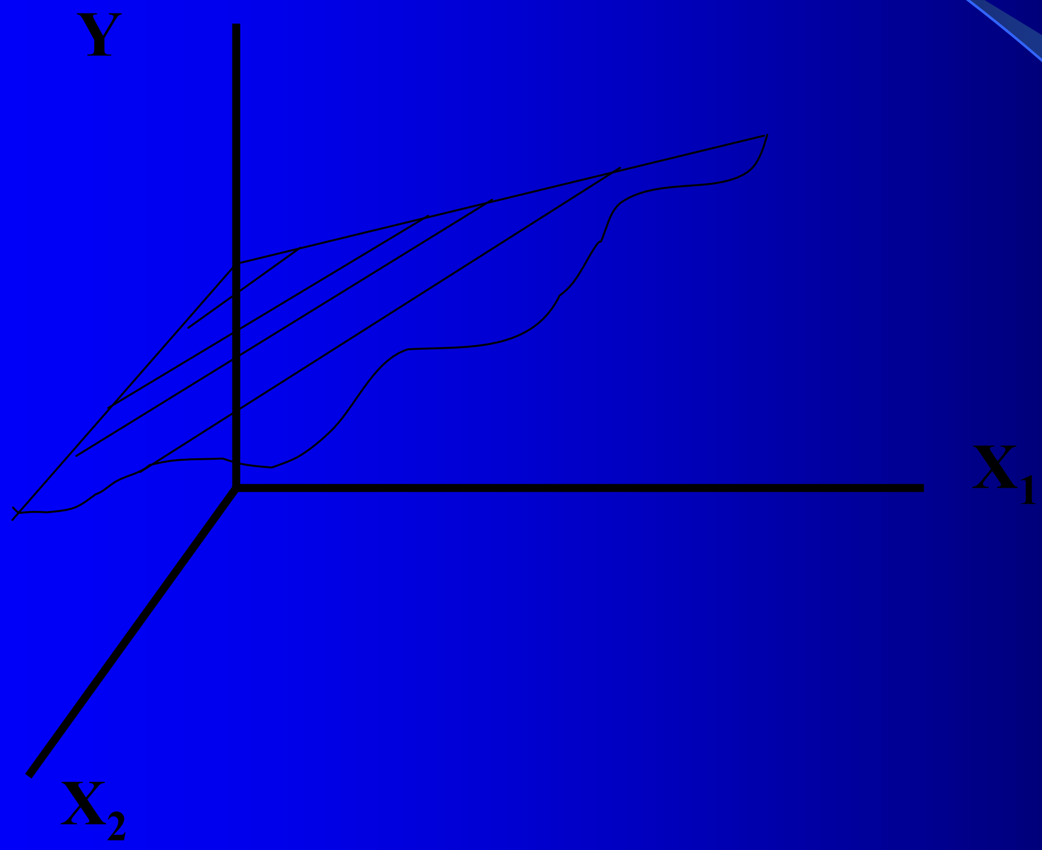
Los puntos se ajustan a un plano



Hipótesis

Linealidad:

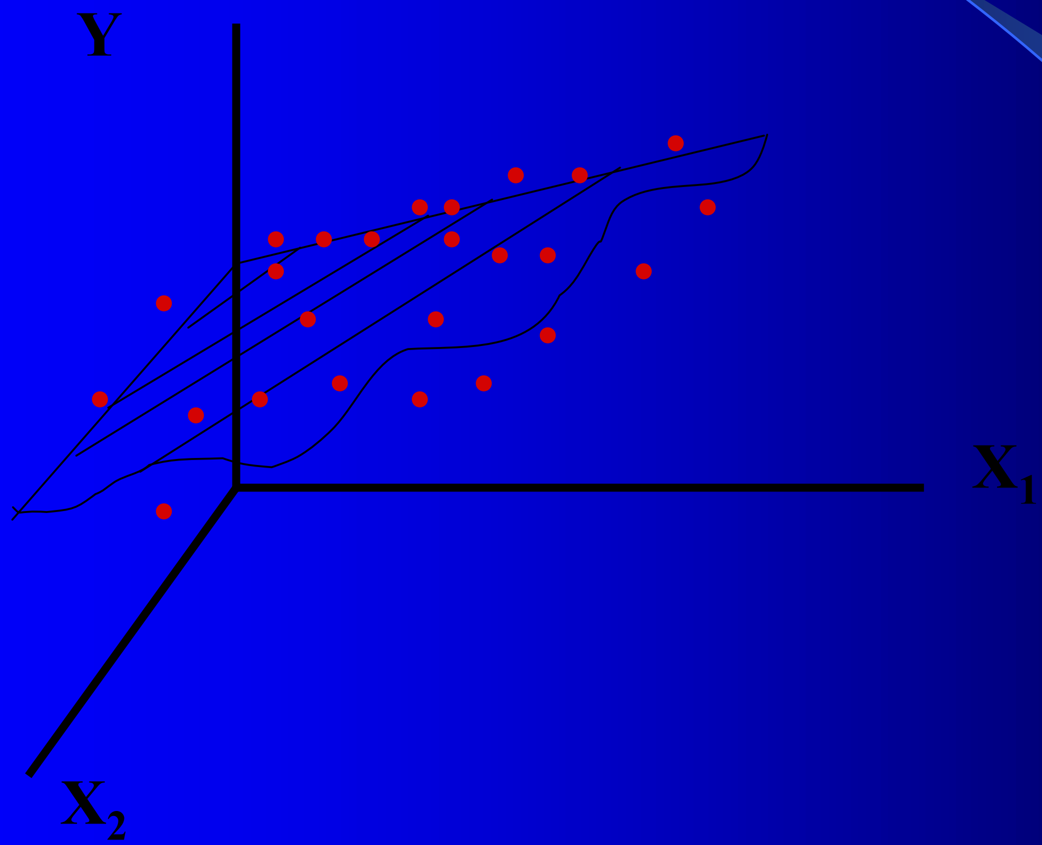
Los puntos se ajustan a un plano



Hipótesis

Linealidad:

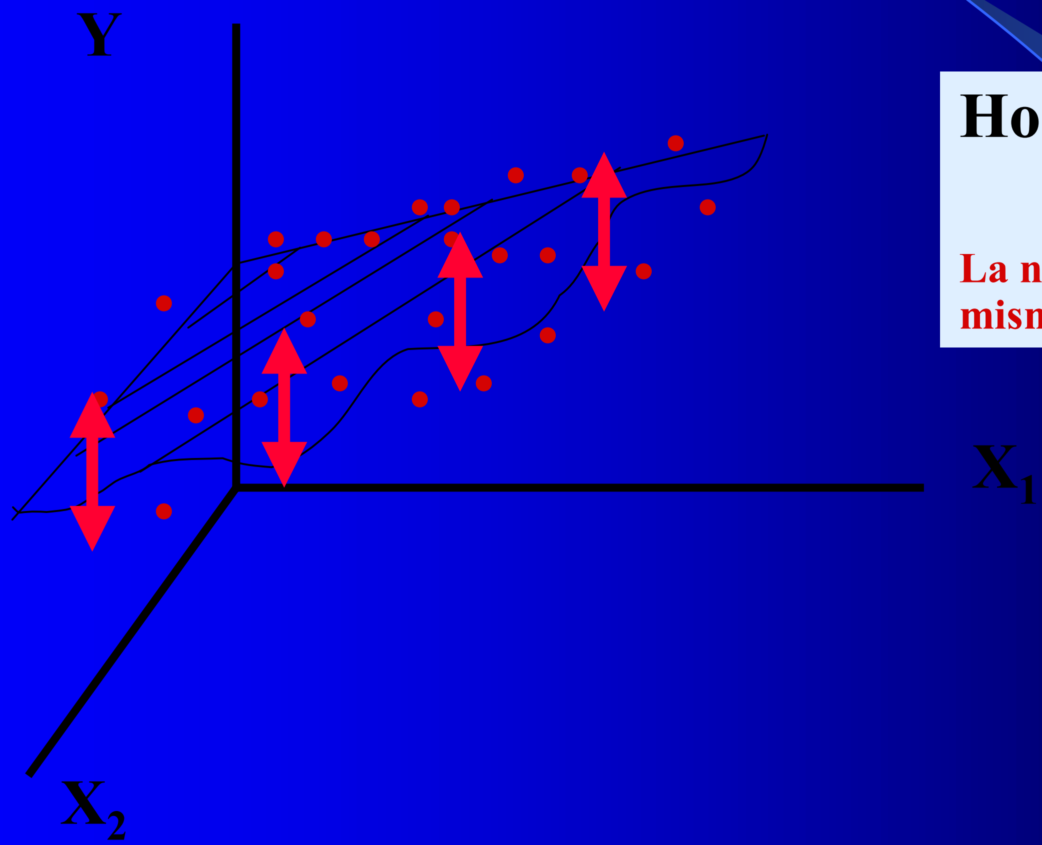
Los puntos se ajustan a un plano



Hipótesis

Linealidad:

Los puntos se ajustan a un plano



Homocedasticidad:

La nube de puntos tiene el mismo grosor

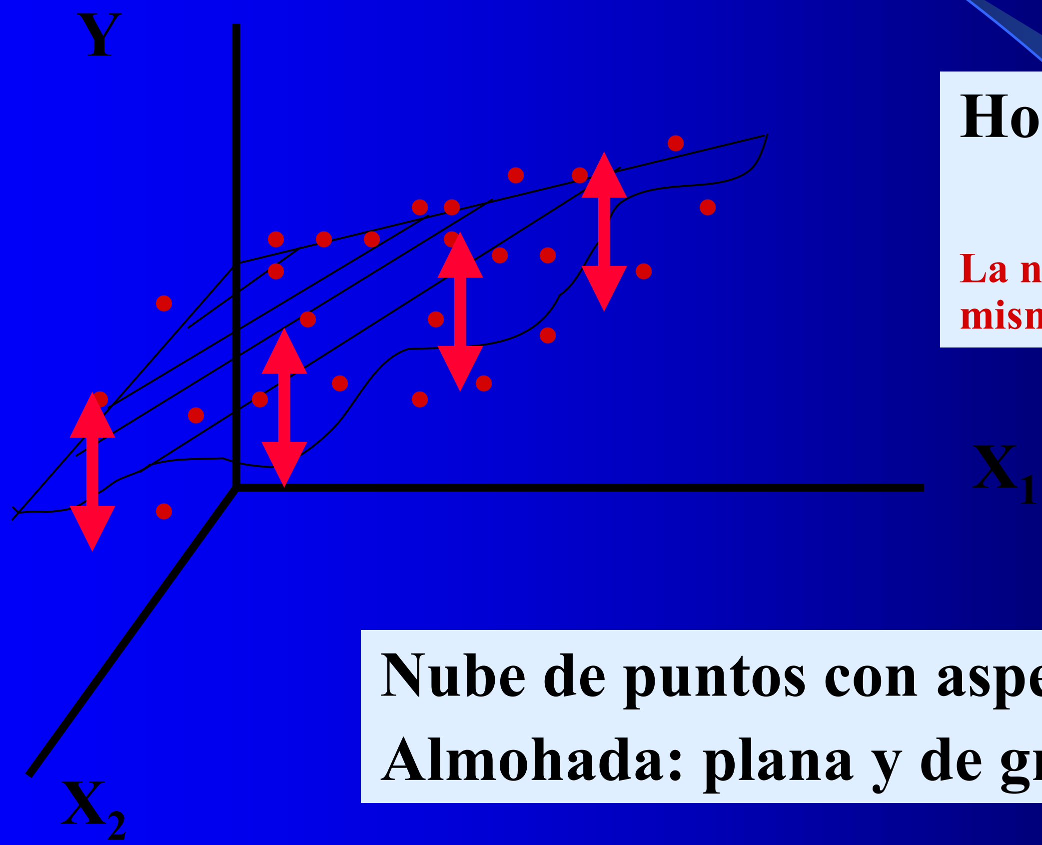
Hipótesis

Linealidad:

Los puntos se ajustan a un plano

Homocedasticidad:

La nube de puntos tiene el mismo grosor



Nube de puntos con aspecto de Almohada: plana y de grosor constante

Estimación

- Lo hace el ordenador.
- Proporciona:
  1. Estimadores: valores de las Betas
  2. Errores estándar: medida de la precisión. Sirve para construir intervalos de confianza
  3. Estadísticos t: para ver si las variables son significativas (t>2) o no (t<2)
  4. R^2: Cuánto de Y es explicado por las X

Interpretación de regresiones

$$MpG_i = 61.69 - 0.27\ Pot_i - 0.64\ Acel_i + 0.00035\ Precio_i$$
$$(-16.22) \qquad (-3.95) \qquad (2.02)$$

$R^2 = 65.6\%$

MpG= Millas recorridas por galón
Pot= Potencia (CV)
Acel= Aceleración (Espacio recorrido en 10 seg)
Precio= Precio

Interpretación de regresiones

$$\text{Log } R_i = 3.4 + 0.95 \text{ Log Ventas}_i - 0.5 \text{ Log Num Empleados}$$
$$\quad (4.2) \quad (5.6) \qquad\qquad\qquad (-7.4)$$

R^2=67%

R= Remuneración del director general de la empresa

Ventas= Ventas de la empresa

Num Empleados= Número de empleados de la empresa

Interpretación de regresiones

$$\underset{(1.2)}{\text{Log } A_i = 2.1} + \underset{(6.6)}{0.8} \log S_i - \underset{(-4.4)}{0.5} \text{ Log Años}_i - \underset{(-3.7)}{0.02} \text{ Log Edad}_i$$

R^2=54.3%

A = Precio del alquiler de viviendas
S = Superficie del piso
Años= Antigüedad del contrato
Edad= Edad del edificio

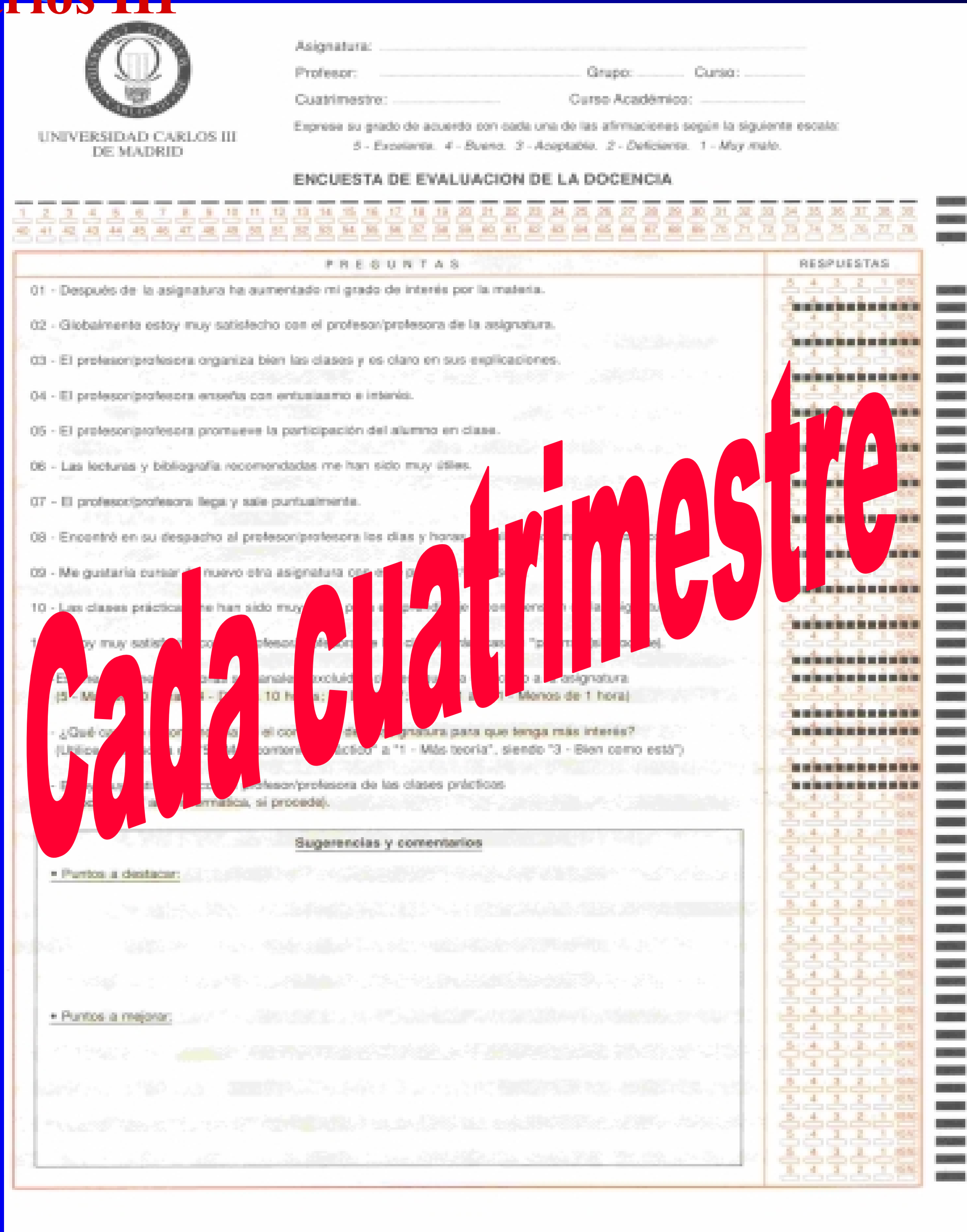
Interpretación de regresiones

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: log(Esp vida Fem)
-----------------------------------------------------------------------------
                                        Standard          T
Parameter               Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                 2,83884        0,81341         3,49004         0,0008
log(Calirias)           0,271963      0,0892364         3,04767         0,0033
log(Casos SIDA)       -0,0145695     0,00476294        -3,05893         0,0031
log(TasaNata)          -0,209542      0,0379077         -5,5277         0,0000
-----------------------------------------------------------------------------
```

Log Esp Vida Femenina

Para países en 1995

Universidad Carlos III


UNIVERSIDAD CARLOS III DE MADRID

Asignatura: ....

Profesor: .... Grupo: .... Curso: ....

Cuatrimestre: .... Curso Académico: ....

Exprese su grado de acuerdo con cada una de las afirmaciones según la siguiente escala:
5 - Excelente. 4 - Bueno. 3 - Aceptable. 2 - Deficiente. 1 - Muy malo.

ENCUESTA DE EVALUACION DE LA DOCENCIA

PREGUNTAS	RESPUESTAS
01 - Después de la asignatura ha aumentado mi grado de interés por la materia.	5 4 3 2 1 NS/NC
02 - Globalmente estoy muy satisfecho con el profesor/profesora de la asignatura.	
03 - El profesor/profesora organiza bien las clases y es claro en sus explicaciones.	
04 - El profesor/profesora enseña con entusiasmo e interés.	
05 - El profesor/profesora promueve la participación del alumno en clase.	
06 - Las lecturas y bibliografía recomendadas me han sido muy útiles.	
07 - El profesor/profesora llega y sale puntualmente.	
08 - Encontré en su despacho al profesor/profesora los días y horas [illegible]	
09 - Me gustaría cursar de nuevo otra asignatura con [illegible]	
10 - Las clases prácticas me han sido muy [illegible]	
11 - [illegible]	
12 - [illegible] a la asignatura. [illegible] 10 horas; [illegible] Menos de 1 hora)	
13 - ¿Qué [illegible] el contenido de la asignatura para que tenga más interés? (Utilice [illegible] contenido práctico" a "1 - Más teoría", siendo "3 - Bien como está")	
14 - [illegible] profesor/profesora de las clases prácticas [illegible] si procede).	

Sugerencias y comentarios

- Puntos a destacar:

- Puntos a mejorar:

Cada cuatrimestre

Universidad Carlos III

No todas las preguntas son igualmente importantes.

Variables Globales:

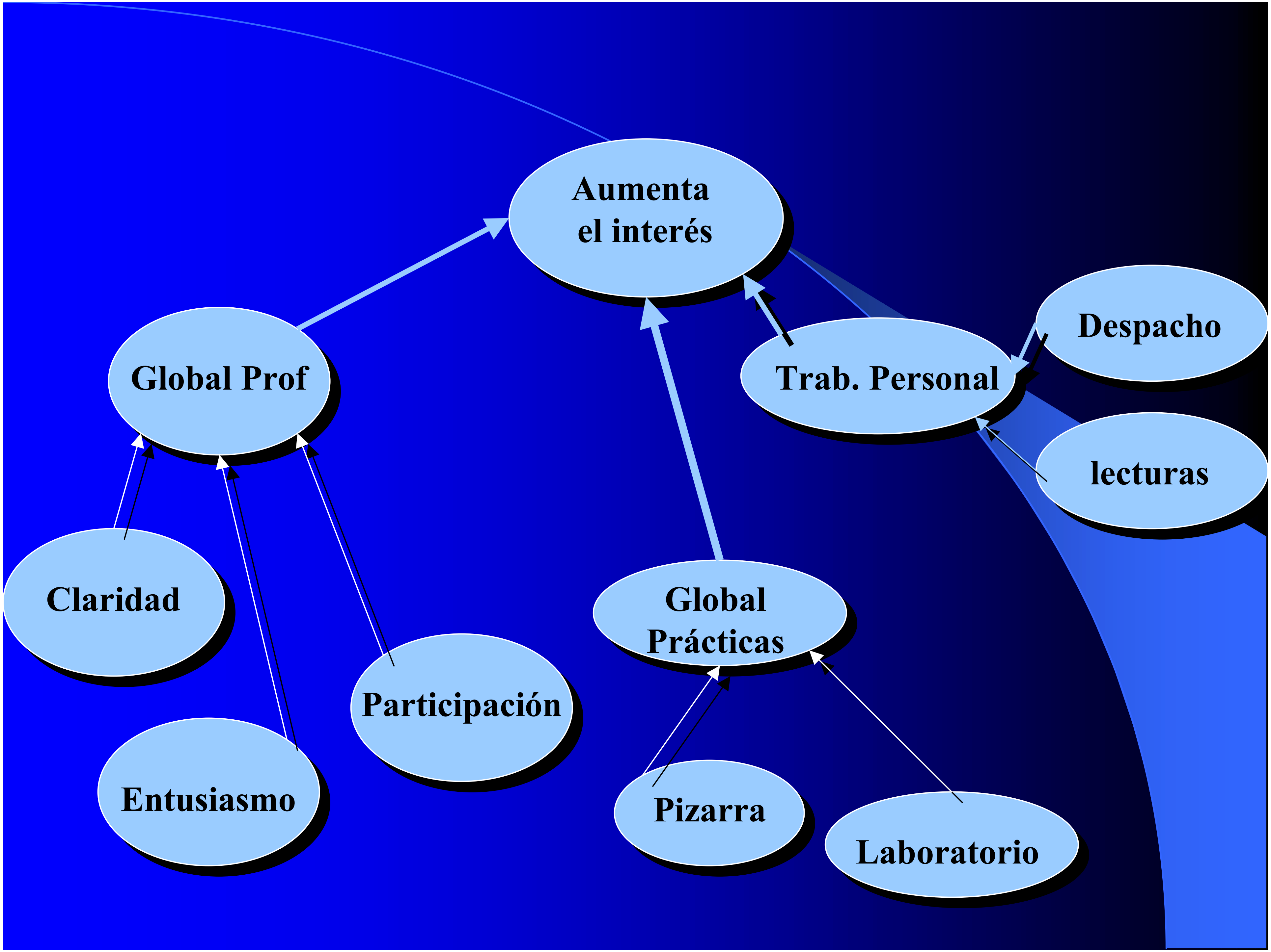
1. Aumenta interés por la materia
2. Satisfacción global con el profesor
3. Me gustaría cursar otra asignatura con este profesor

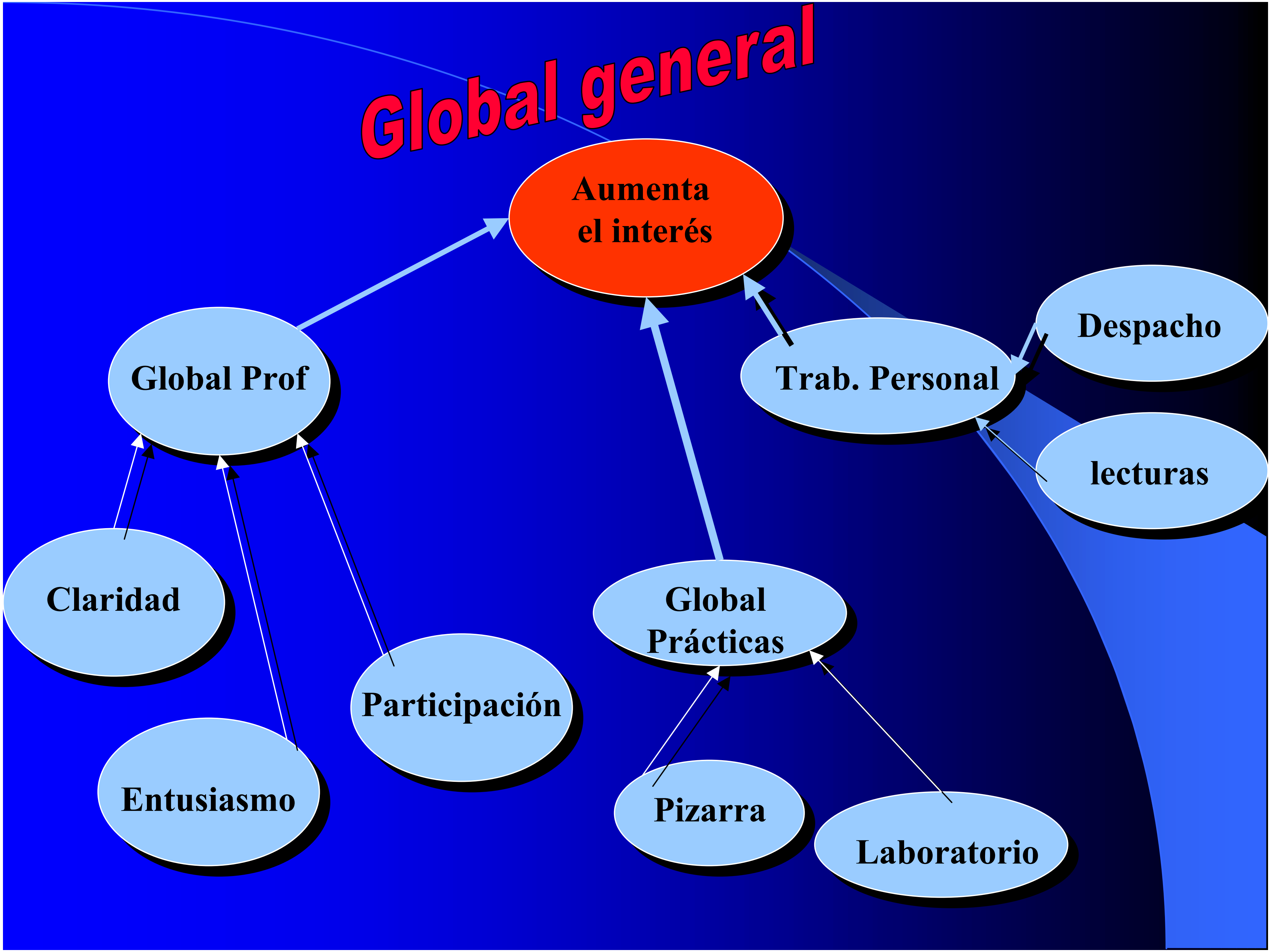
Variables parciales:

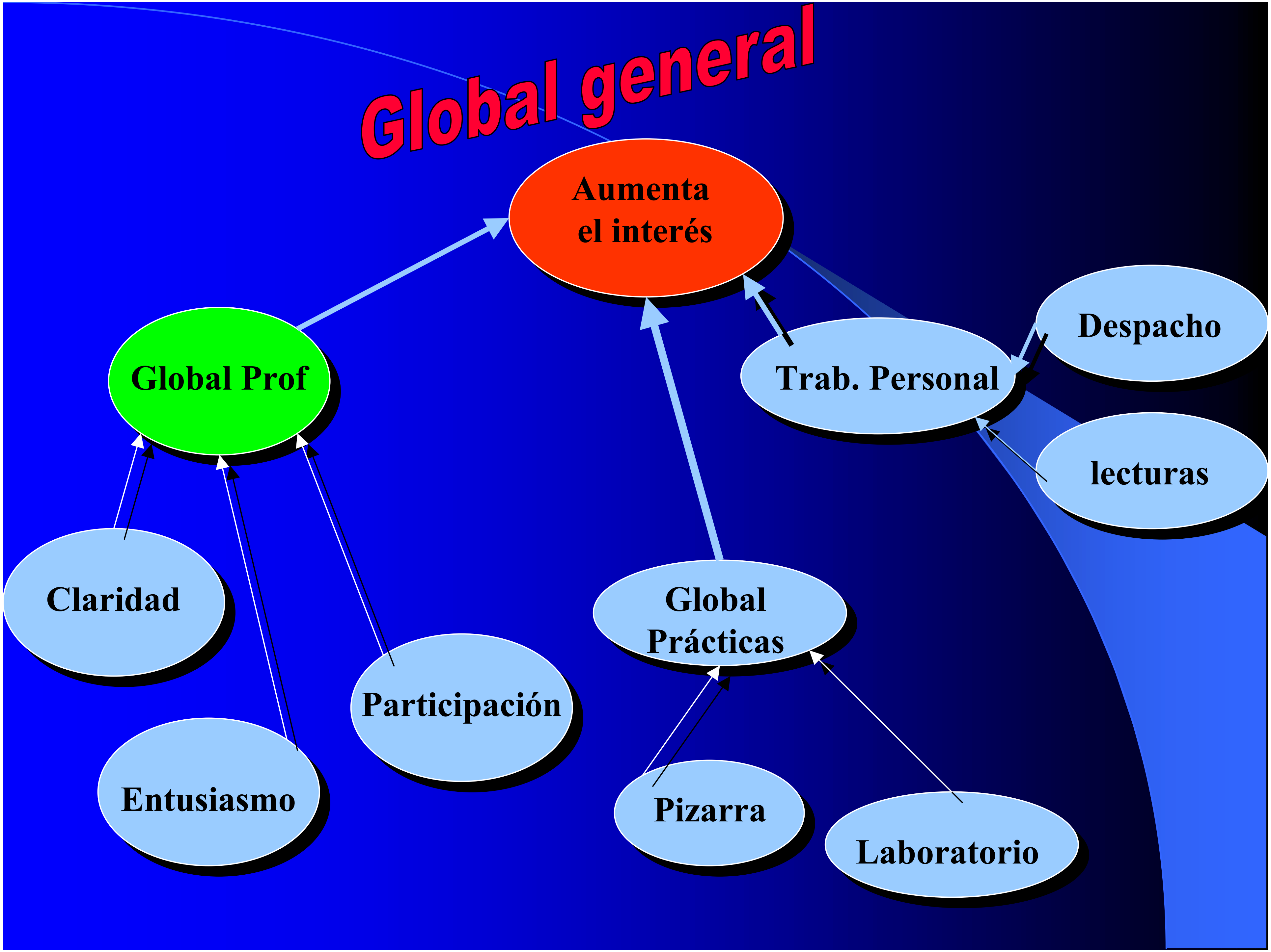
- CLARIDAD
- ENTUSIASMO
- PARTICIPACION
- LECTURAS
- PUNTUALIDAD
- DESPACHO

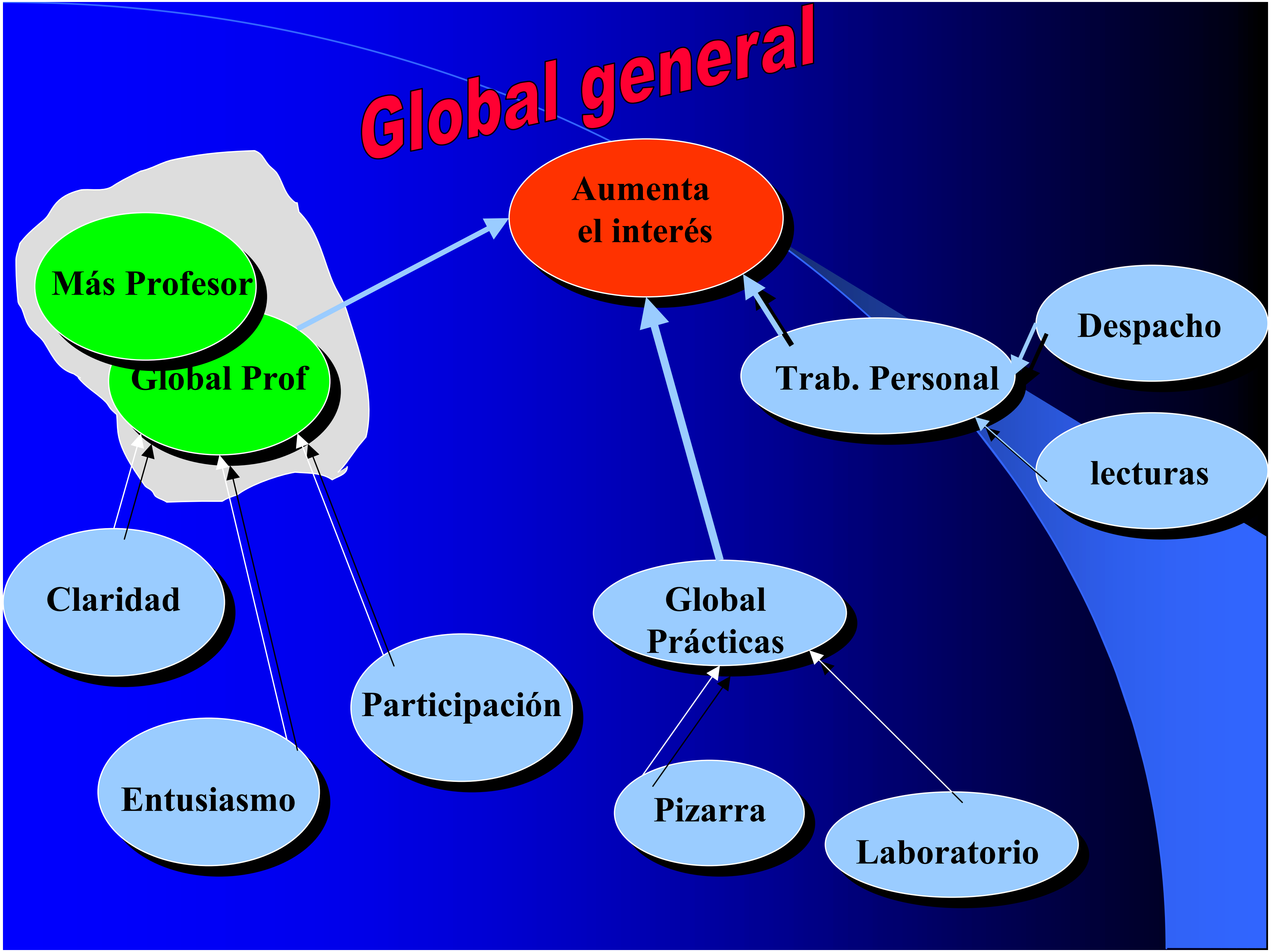
Variables de Prácticas

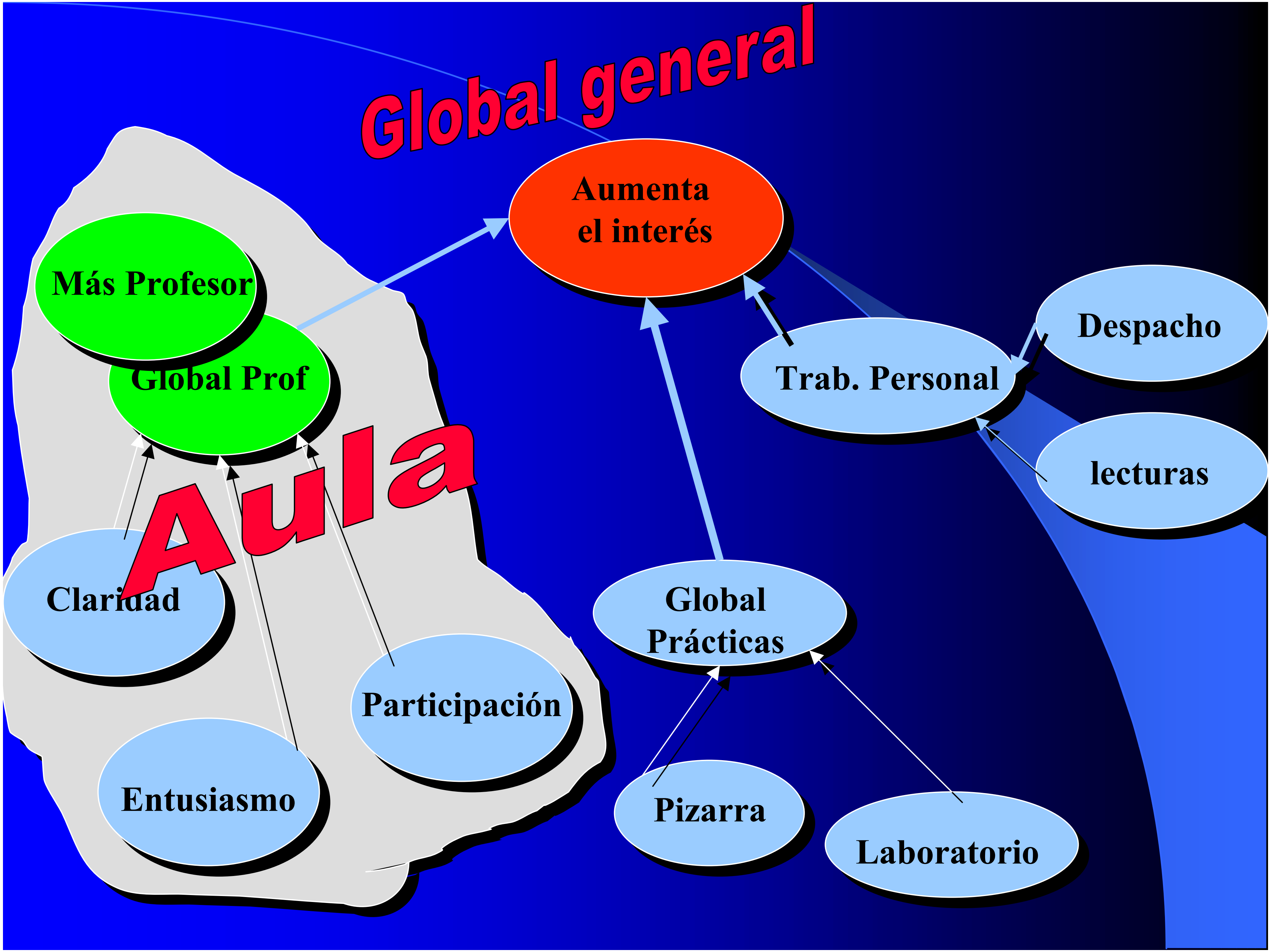
- Utilidad
- Satisfacción Pizarra
- Satisfacción Laboratorio


Aumenta el interés
Global Prof
Trab. Personal
Despacho
lecturas
Claridad
Global Prácticas
Participación
Entusiasmo
Pizarra
Laboratorio


Global general
Aumenta el interés
Global Prof
Trab. Personal
Despacho
lecturas
Claridad
Global Prácticas
Participación
Entusiasmo
Pizarra
Laboratorio


Global general
Aumenta el interés
Global Prof
Trab. Personal
Despacho
lecturas
Global Prácticas
Claridad
Entusiasmo
Participación
Pizarra
Laboratorio


Global general
Aumenta el interés
Más Profesor
Global Prof
Claridad
Entusiasmo
Participación
Global Prácticas
Pizarra
Laboratorio
Trab. Personal
Despacho
lecturas



Global general
Aumenta el interés
Más Profesor
Global Prof
Aula
Claridad
Entusiasmo
Participación
Global Prácticas
Pizarra
Laboratorio
Trab. Personal
Despacho
lecturas

Interpretación de regresiones

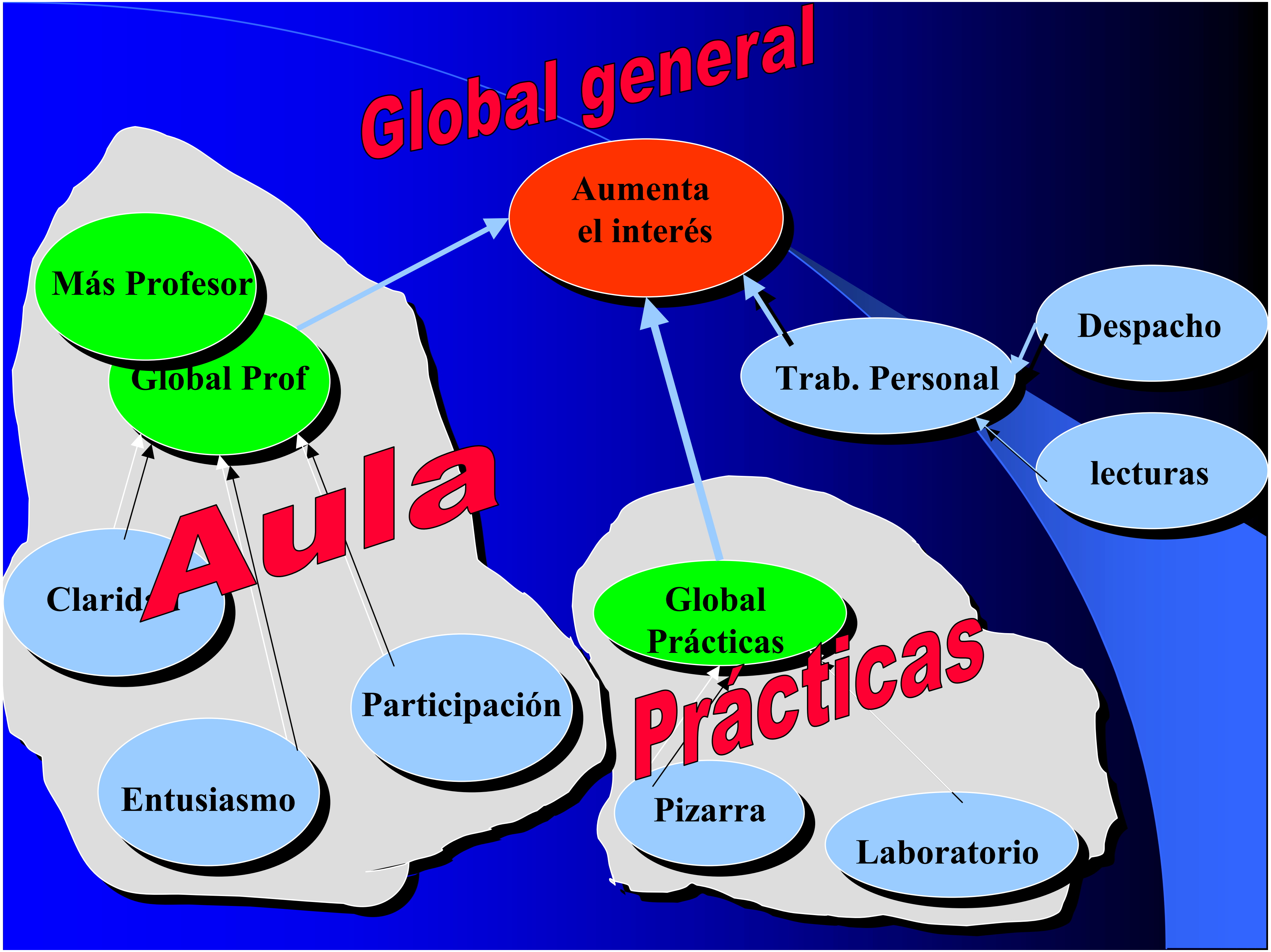
Multiple Regression Analysis

Dependent variable: Satis_profesor

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-0,177384	0,0555375	-3,19395	0,0016
Entusiasmo_intere	0,243065	0,0311795	7,79567	0,0000
Organiza_clases	0,724773	0,0234401	30,9202	0,0000
Prom_participacio	0,0984656	0,0252624	3,89772	0,0001

$$\text{Sat Global} = -0.18 + \underset{(7.8)}{0.24\ \text{Entus}_i} + \underset{(30.9)}{0.72\text{Organiza}_i} + \underset{(3.9)}{0.1\ \text{Part}_i}$$

R^2=95.4%


Global general
Aumenta el interés
Más Profesor
Global Prof
Aula
Claridad
Entusiasmo
Participación
Global Prácticas
Prácticas
Pizarra
Laboratorio
Trab. Personal
Despacho
lecturas

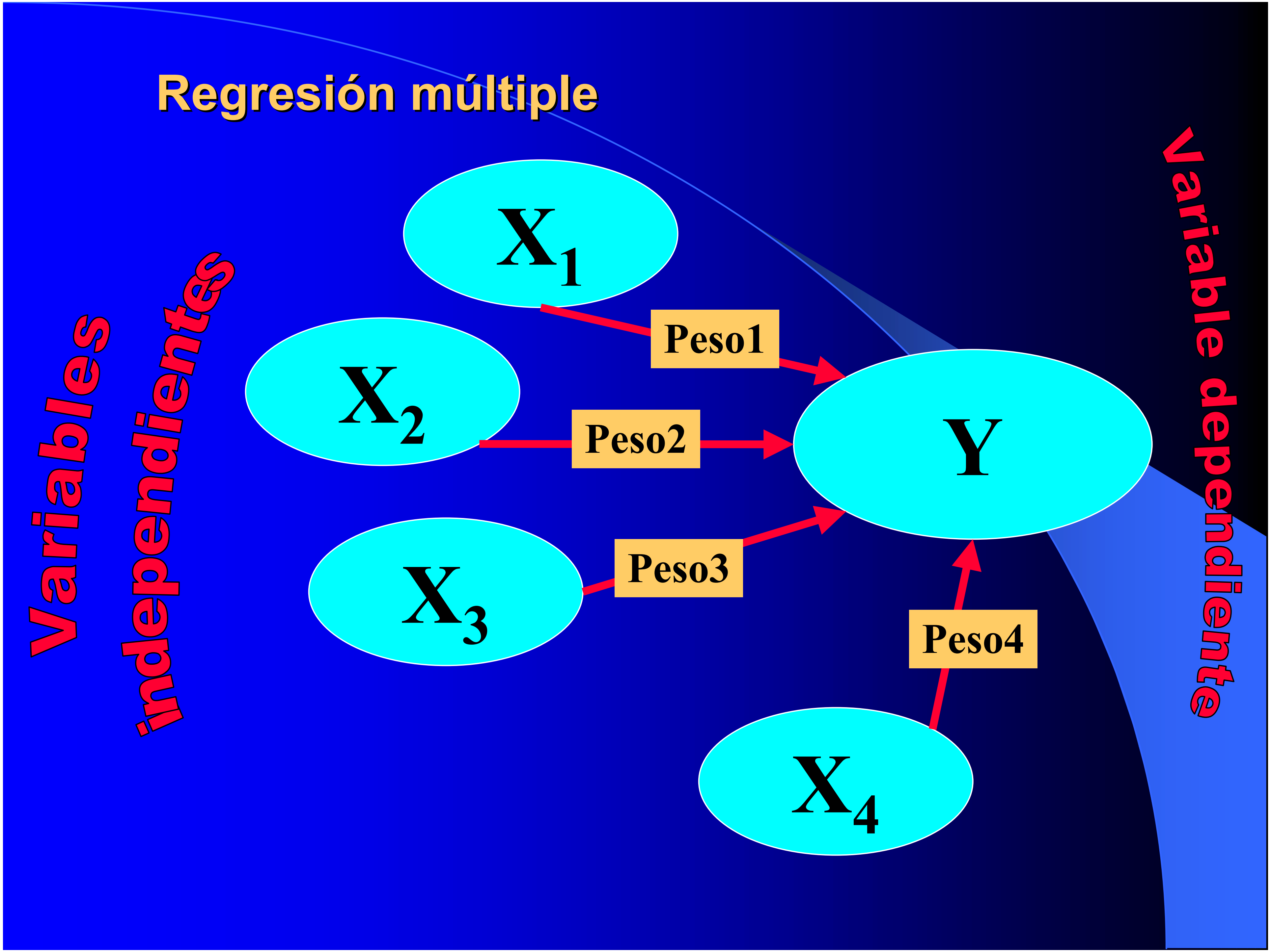
Interpretación de regresiones

Multiple Regression Analysis

Dependent variable: Clase_Practica

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	0,590115	0,143453	4,11365	0,0001
Pof_ Laboratorio	0,557256	0,0651018	8,55975	0,0000
Prof_Pizarra	0,243002	0,0579318	4,19462	0,0000

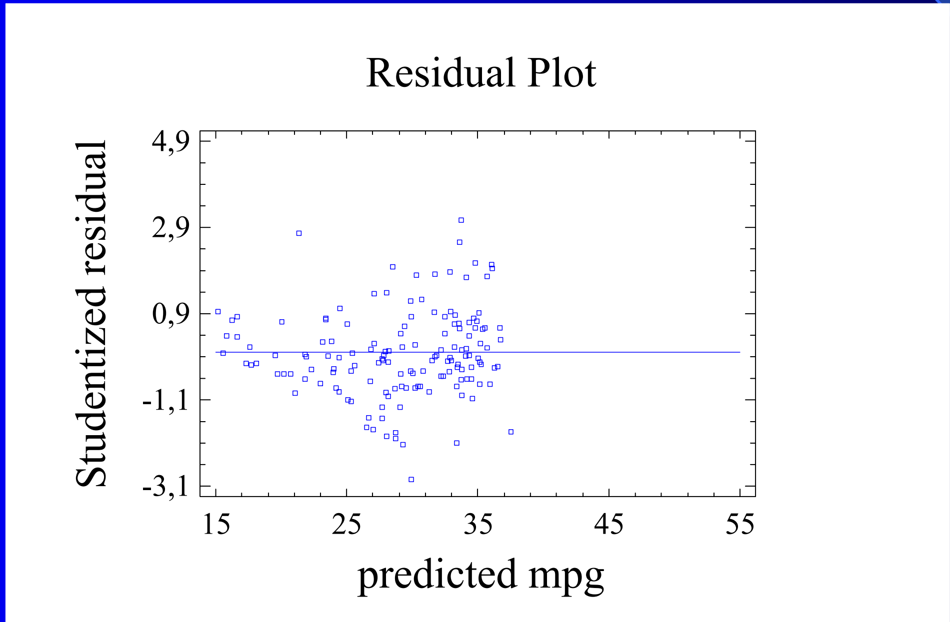
$$\text{Sat Prac} = 0.59 + \underset{(8.55)}{0.56}\,\text{Lab}_i + \underset{(4.19)}{0.24}\,\text{Pizarra}_i \qquad R^2 = 95.4\%$$



Regresión múltiple
Variables independientes
Variable dependiente
X_1
X_2
X_3
X_4
Y
Peso1
Peso2
Peso3
Peso4

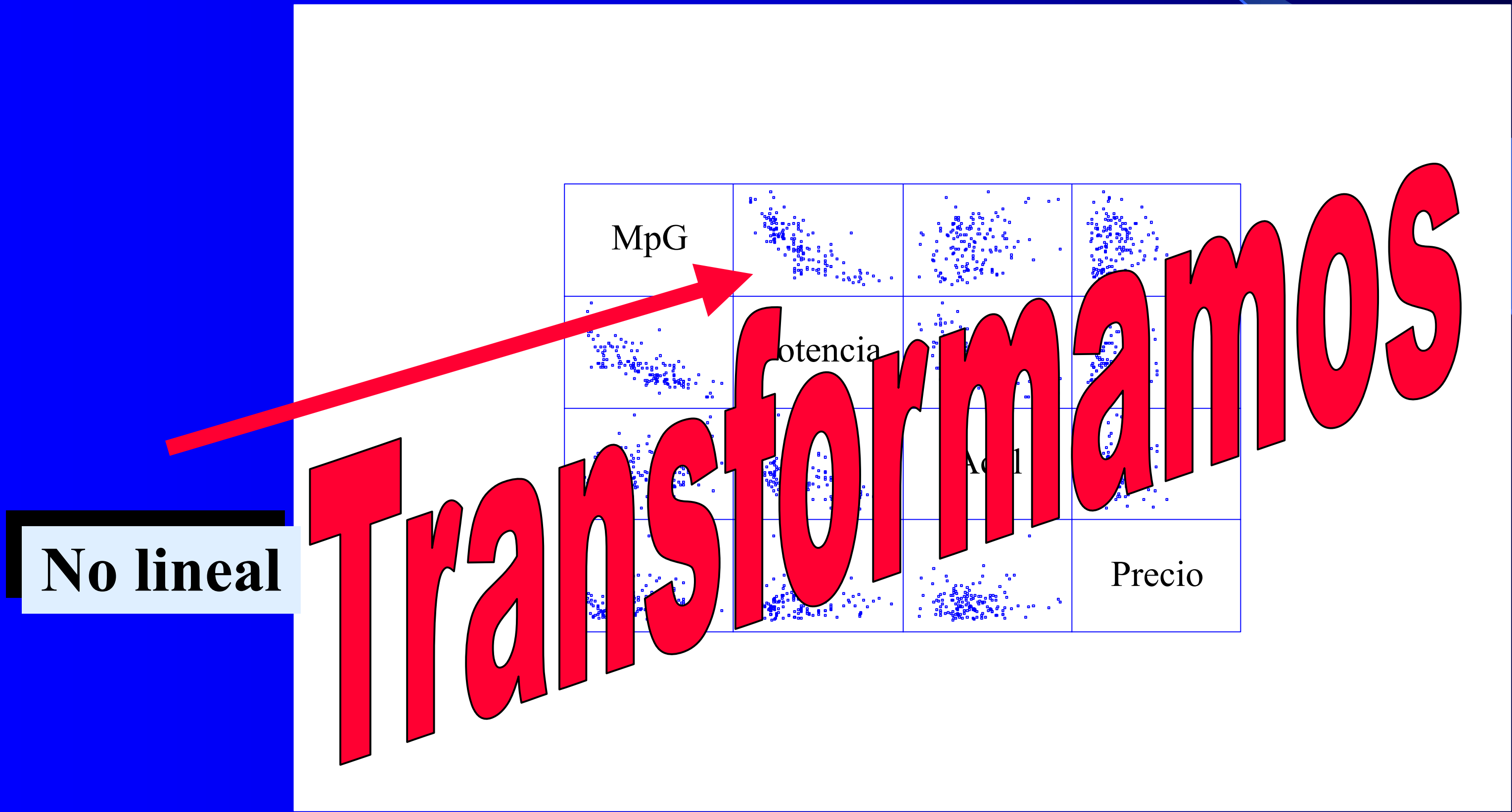
Diagnosis

Gráfico de residuos vs. Valores ajustados

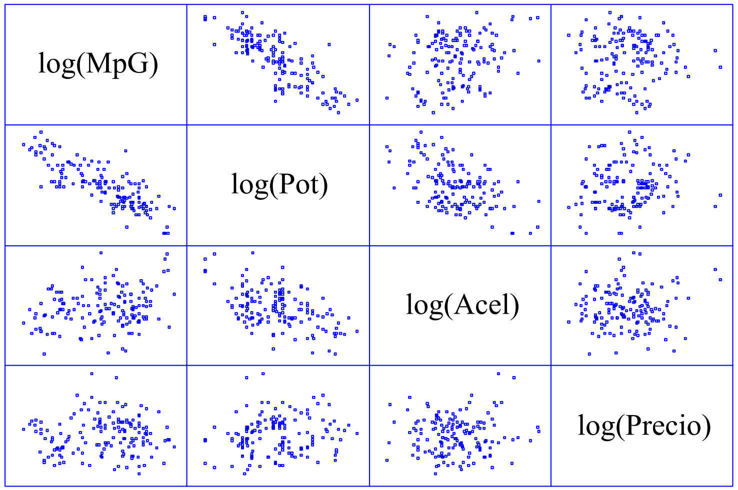


¿¿¿¿Bien????

Miramos los datos:



Tomando logaritmos



Regresión en Logaritmos:

$$\text{Log } MpG_i = 8.2 - 1 \log Pot_i - 0.5 \log Acel_i + 0.13 \log Precio_i$$

$$(-20.6) \qquad (-6.2) \qquad (4.4)$$

$R^2 = 75.3\%$

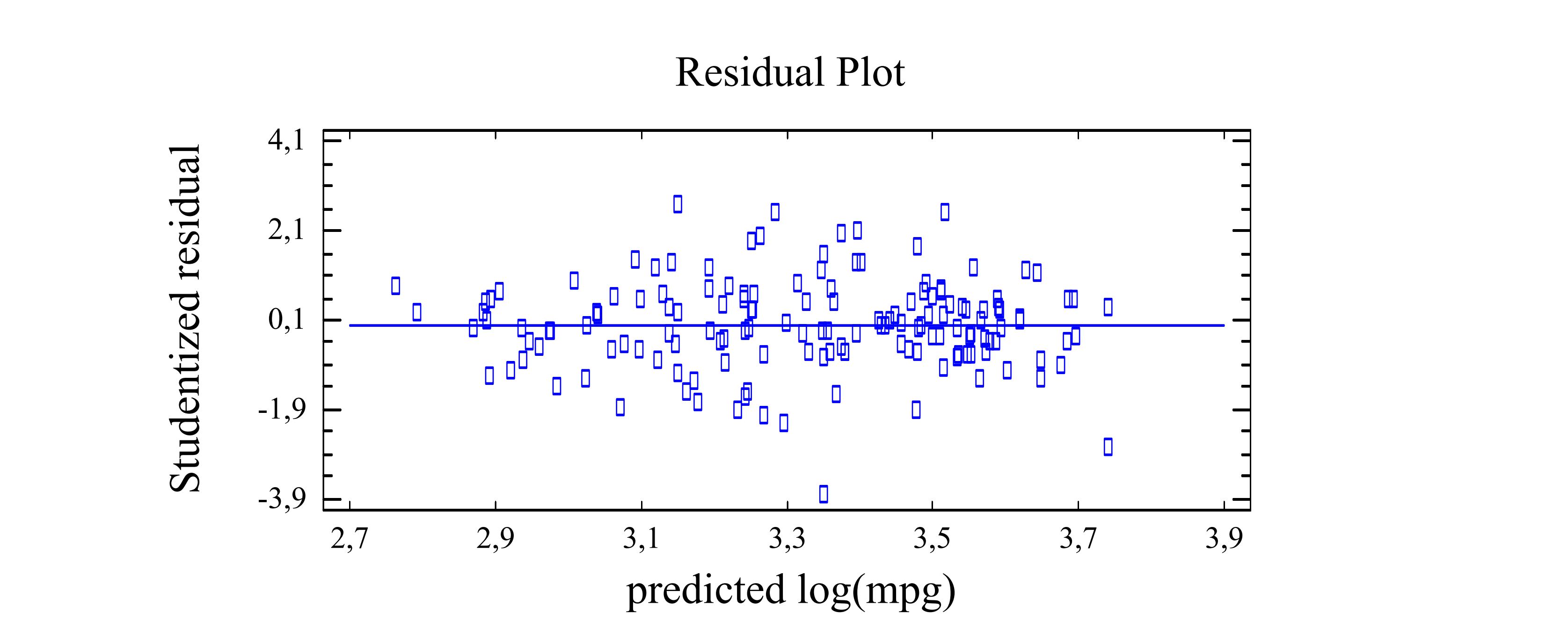
MpG= Millas recorridas por galón

Pot= Potencia (CV)

Acel= Aceleración (Tiempo de 0-100Km)

Precio= Precio

Residuos



Residual Plot
Studentized residual
4,1
2,1
0,1
-1,9
-3,9
2,7
2,9
3,1
3,3
3,5
3,7
3,9
predicted log(mpg)

R^2 corregido

- R^2 tiene un problema:
  - Si metemos variables en el modelo R^2 aumenta ***aunque las variables sean no significativas***

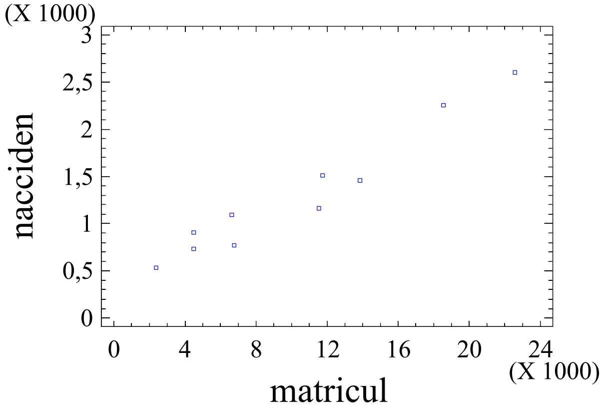
Para evitarlo se define: $\overline{R}^2$

R^2 corregido por grados de libertad.
Al introducir nuevas variables en el modelo no aumenta

FIN de Regresión Múltiple

Multicolinealidad

Ejemplo



Numero de accidentes en provincias españolas en función del número de vehículos matriculados

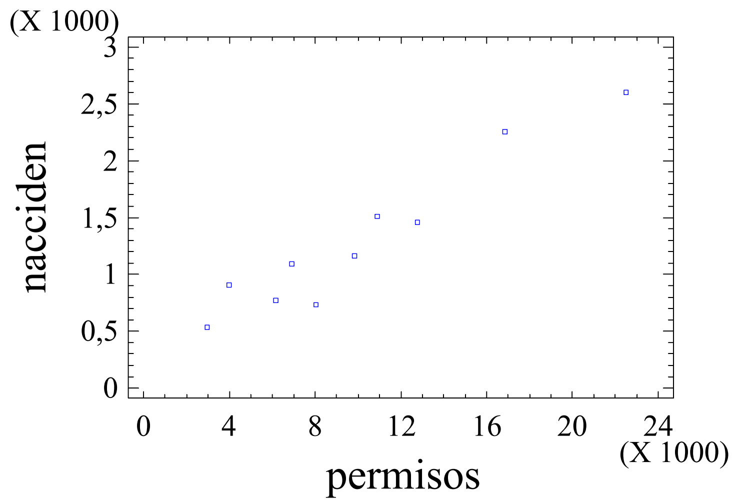
```
-----------------------------------------------------------------------------
Dependent variable: nacciden
-----------------------------------------------------------------------------
```

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	278,24	102,518	2,71406	0,0265
matricul	0,0993373	0,00850344	11,682	0,0000

R-squared (adjusted for d.f.) = 93,7703 percent

Ejemplo

Numero de accidentes en provincias españolas en función del número de permisos de conducir




```
-----------------------------------------------------------------------------
Dependent variable: nacciden
-----------------------------------------------------------------------------
                                            Standard            T
Parameter                 Estimate           Error          Statistic         P-Value
-----------------------------------------------------------------------------
CONSTANT                   216,481          127,099          1,70325          0,1269
permisos                  0,107617        0,0109657          9,81395          0,0000
-----------------------------------------------------------------------------
```

R-squared (adjusted for d.f.) = 91,3722 percent

Regresiones

Accid=278.2 +0.1 Matriculas
(11.68)

Accid=216.4 +0.1 Permisos
(9.81)

Regresión con las dos variables

Dependent variable: nacciden

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	250,63	113,216	2,21373	0,0625
matricul	0,0725492	0,0395634	1,83374	0,1093
permisos	0,0301069	0,043353	0,694461	0,5098

Regresión con las dos variables

Dependent variable: nacciden

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	250,63	113,216	2,21373	0,0625
matricul	0,0725492	0,0395634	1,83374	0,1093
permisos	0,0301069	0,043353	0,694461	0,5098

$R^2=93.3\%$

Regresiones

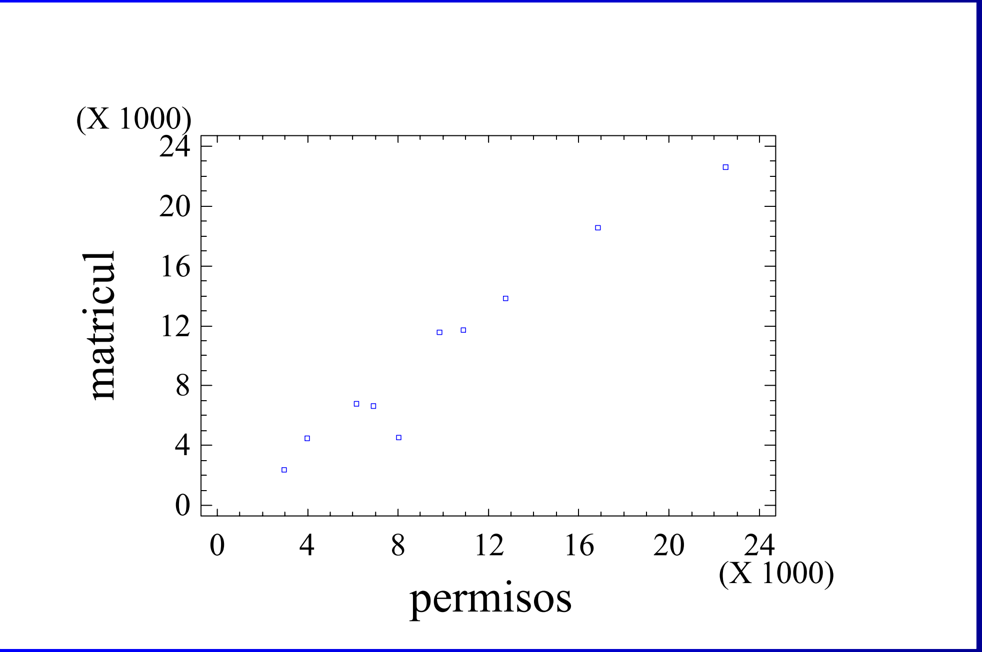
??????

Accid=278.2 +0.1 Matriculas
(11.68)

Accid=216.4 +0.1 Permisos
(9.81)

Accid=250+0.07 Matriculas +0.03 Permisos
(1.8) (0.69)

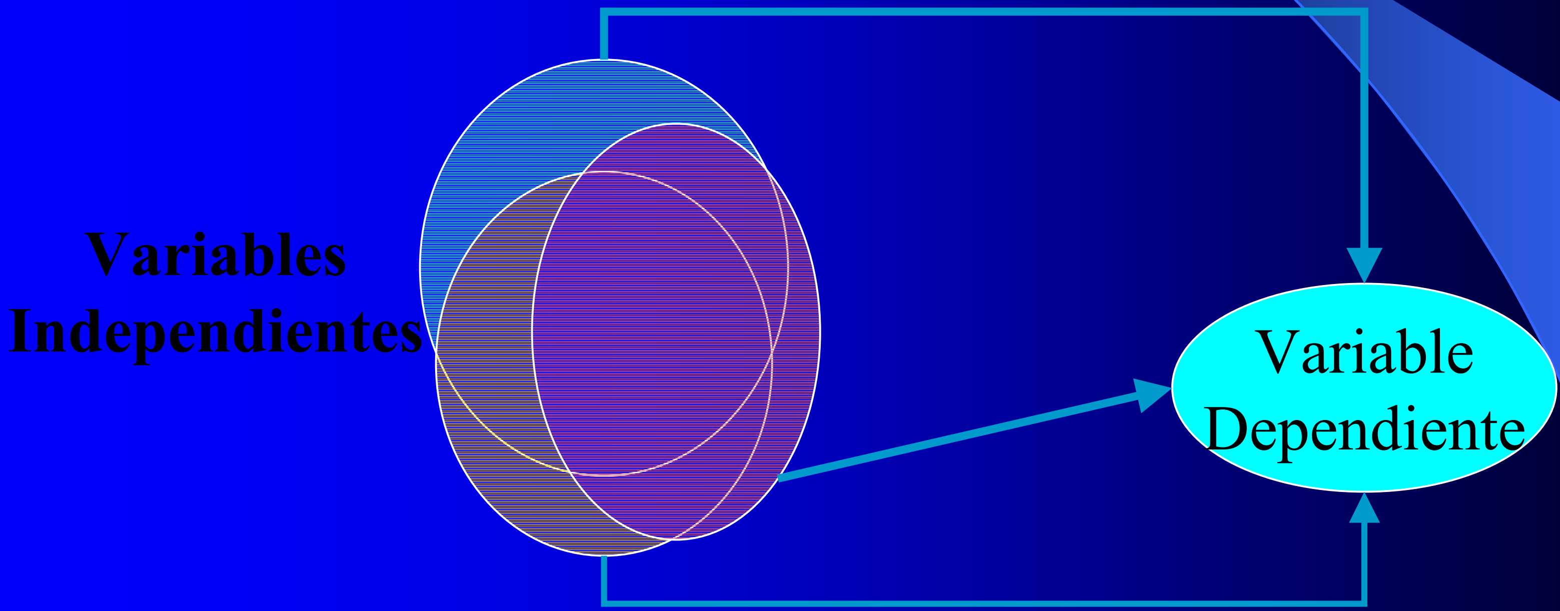
¿Qué está pasando?



Correlación=.975

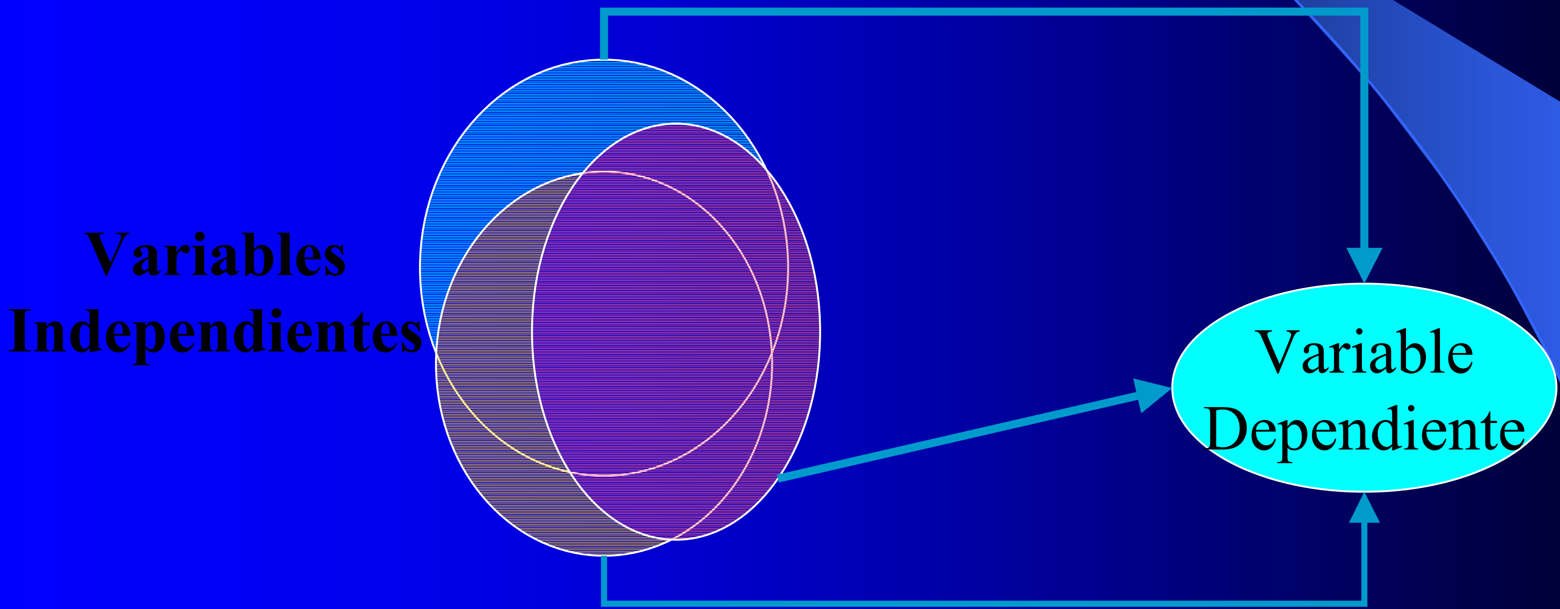
Regresión: Un problema

- A veces las variables independientes son muy parecidas: Contienen la misma información.

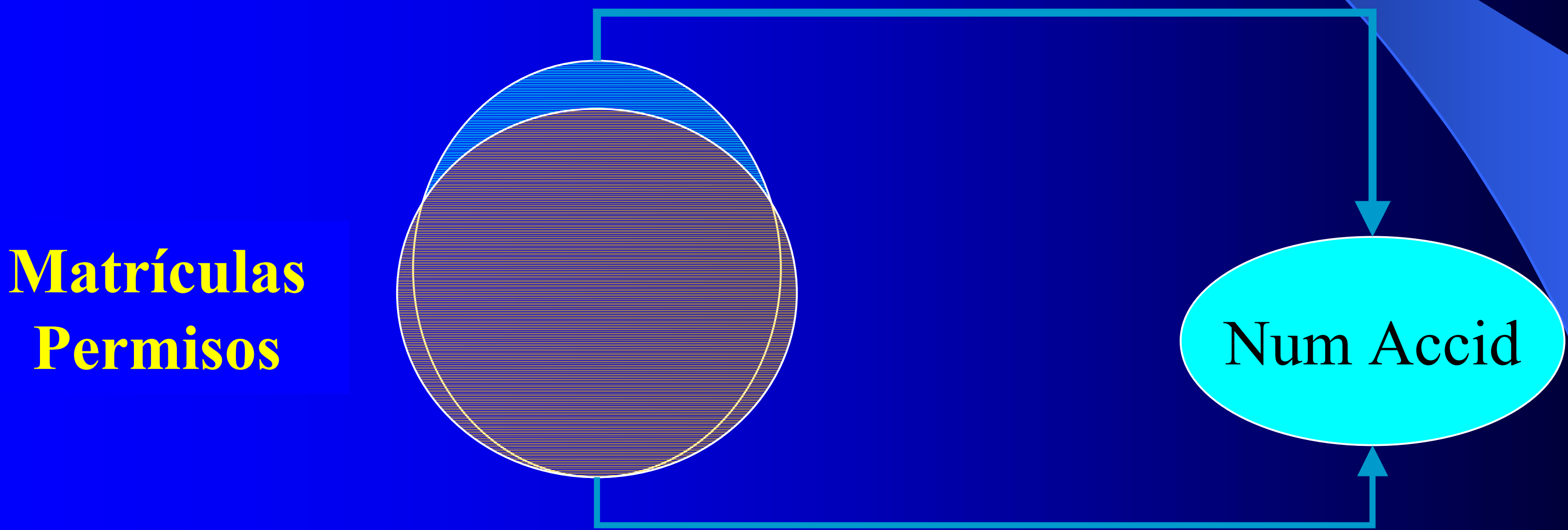


Regresión: Un problema

- El modelo no puede diferenciar entre las variables.



En nuestro ejemplo

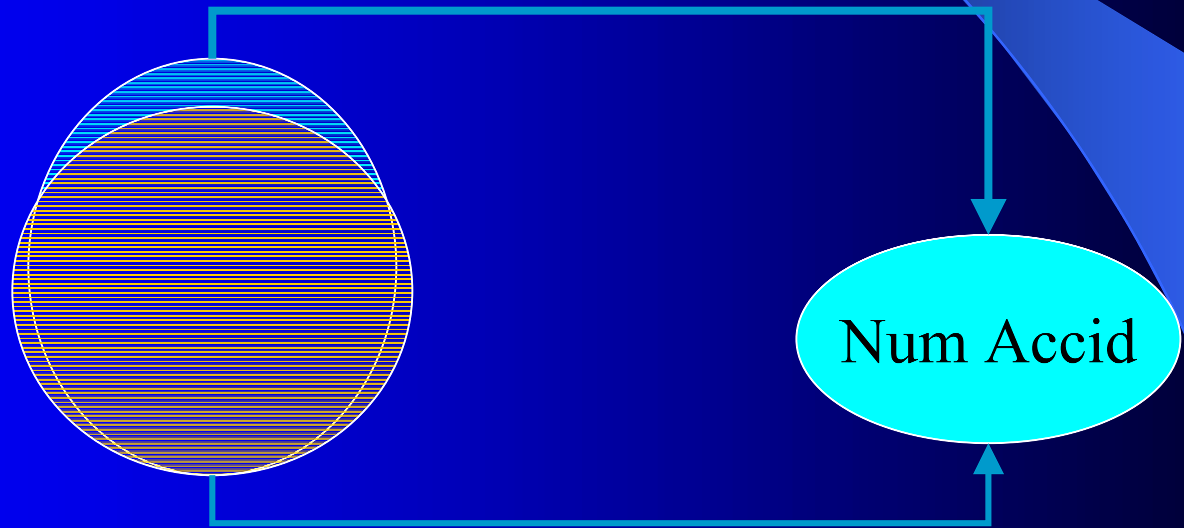


Ambas son muy parecidas para distinguir entre ellas

En nuestro ejemplo

Solución: eliminar una variable: No perdemos casi información

Matrículas
Permisos



Ambas son muy parecidas para distinguir entre ellas

En nuestro ejemplo

Solución: eliminar una variable: No perdemos casi información

Matrículas

Ambas son muy parecidas para distinguir entre ellas

- **El problema de Multicolinealidad aparece en casi todos los trabajos estadísticos.**
- **Tendemos a medir una cosa de muchas formas**
- **Se detecta:**
  - **En regresión simple las variables son significativas**
  - **Al introducir nuevas variables dejan de ser significativas**

Otro ejemplo:

Cata de quesos:

Factores:

1. **Acetico**
2. **Láctico**
3. **H_2S**

ACETICO

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-61,4986	24,8464	-2,47515	0,0196
Acetic	15,6478	4,49577	3,48055	0,0017

R-squared (adjusted for d.f.) = 27,7065 percent

LACTICO

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-29,8588	10,5823	-2,82158	0,0087
Lactic	37,7199	7,1864	5,2488	0,0000

R-squared (adjusted for d.f.) = 47,7947 percent

H_2S

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-9,78684	5,95791	-1,64266	0,1116
H2S	5,77609	0,94585	6,10677	0,0000

R-squared (adjusted for d.f.) = 55,5846 percent

ACETICO Y H_2S

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-26,9397	21,1941	-1,2711	0,2145
Acetic	3,8012	4,50534	0,843709	0,4062
H2S	5,1456	1,20928	4,25508	0,0002

R-squared (adjusted for d.f.) = 55,1227 percent

LACTICO Y H_2S

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Val
CONSTANT	-27,5918	8,98183	-3,07196	0,0048
Lactic	19,8872	7,95901	2,4987	0,0189
H2S	3,94627	1,13569	3,47477	0,0017

R-squared (adjusted for d.f.) = 62,5903 percent

ACETICO y LACTICO

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-51,366	21,1744	-2,42585	0,0222
Lactic	31,3923	8,95627	3,50506	0,0016
Acetic	5,57139	4,76133	1,17013	0,2522

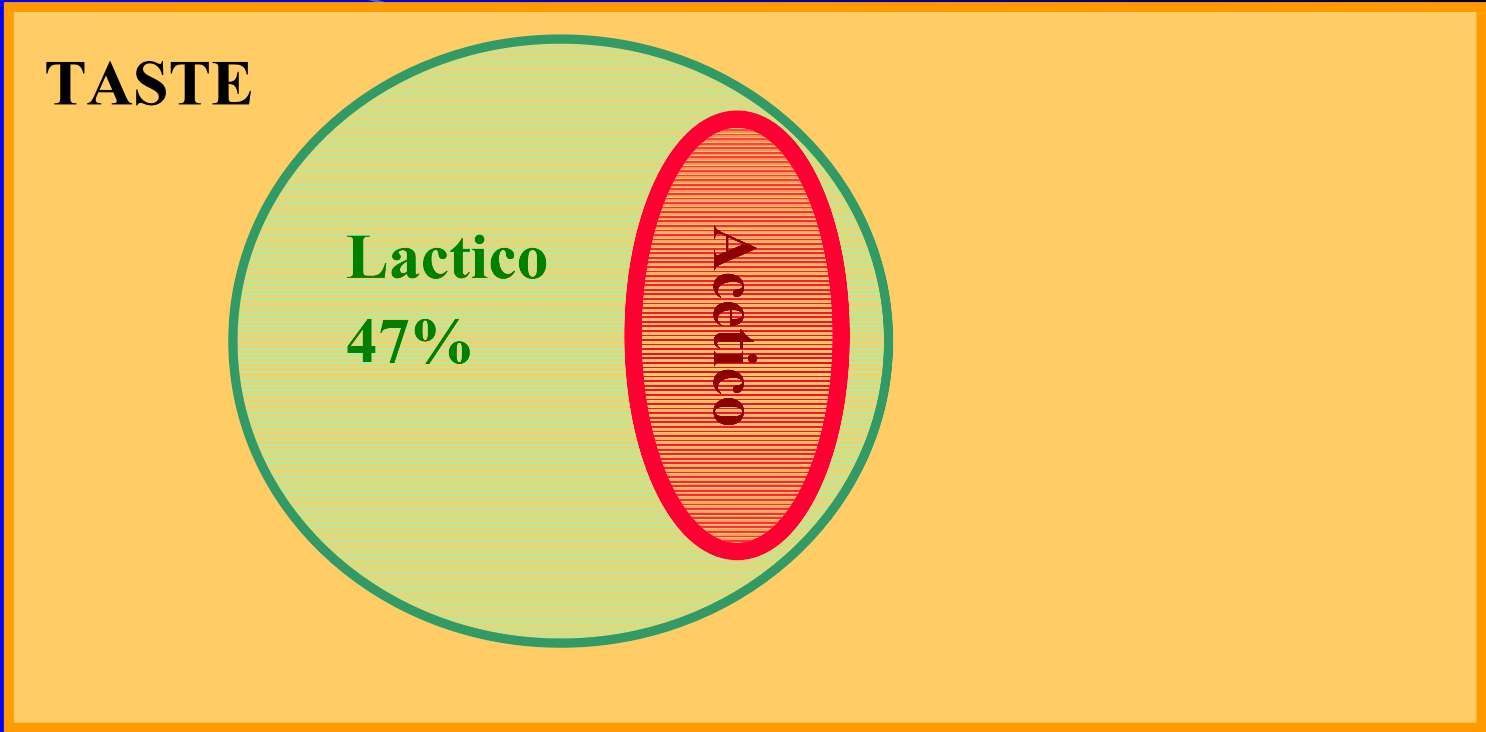
R-squared (adjusted for d.f.) = 48,4741 percent

ACETICO, LACTICO y Y H_2S

Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-28,8768	19,7354	-1,4632	0,1554
H2S	3,91184	1,24843	3,13341	0,0042
Lactic	19,6705	8,62905	2,27957	0,0311
Acetic	0,327741	4,45976	0,0734886	0,9420

R-squared (adjusted for d.f.) = 61,1595 percent

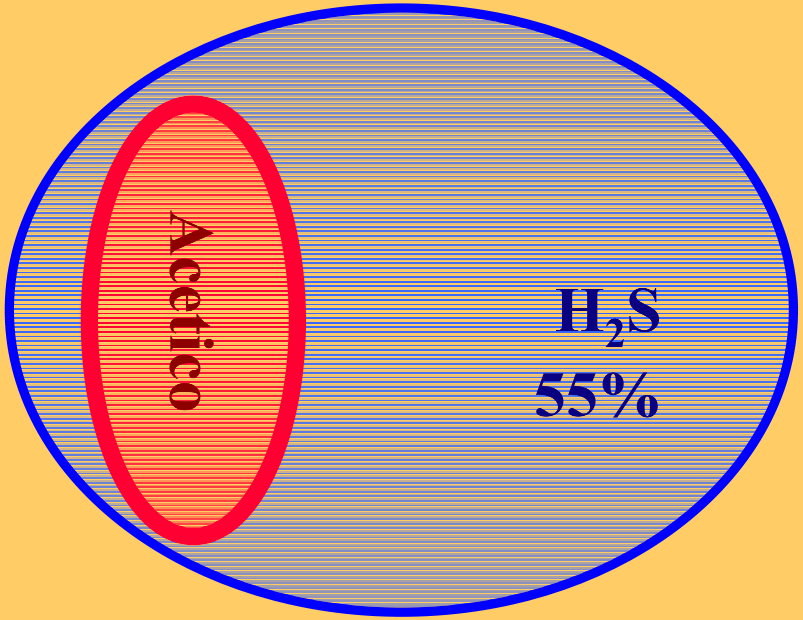


Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-51,366	21,1744	-2,42585	0,0222
Lactic	31,3923	8,95627	3,50506	0,0016
Acetic	5,57139	4,76133	1,17013	0,2522

R-squared (adjusted for d.f.) = 48,4741 percent

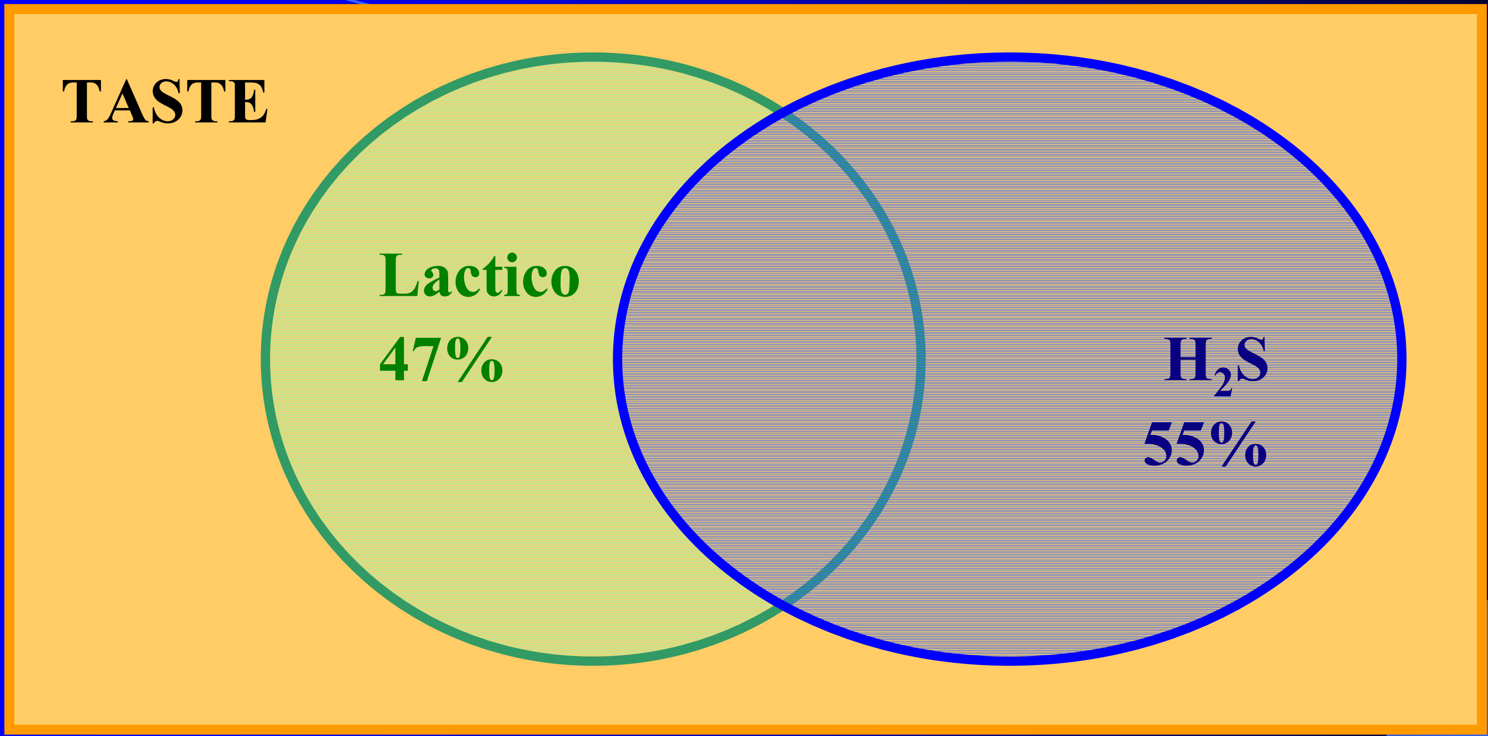
TASTE



Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-26,9397	21,1941	-1,2711	0,2145
Acetic	3,8012	4,50534	0,843709	0,4062
H2S	5,1456	1,20928	4,25508	0,0002

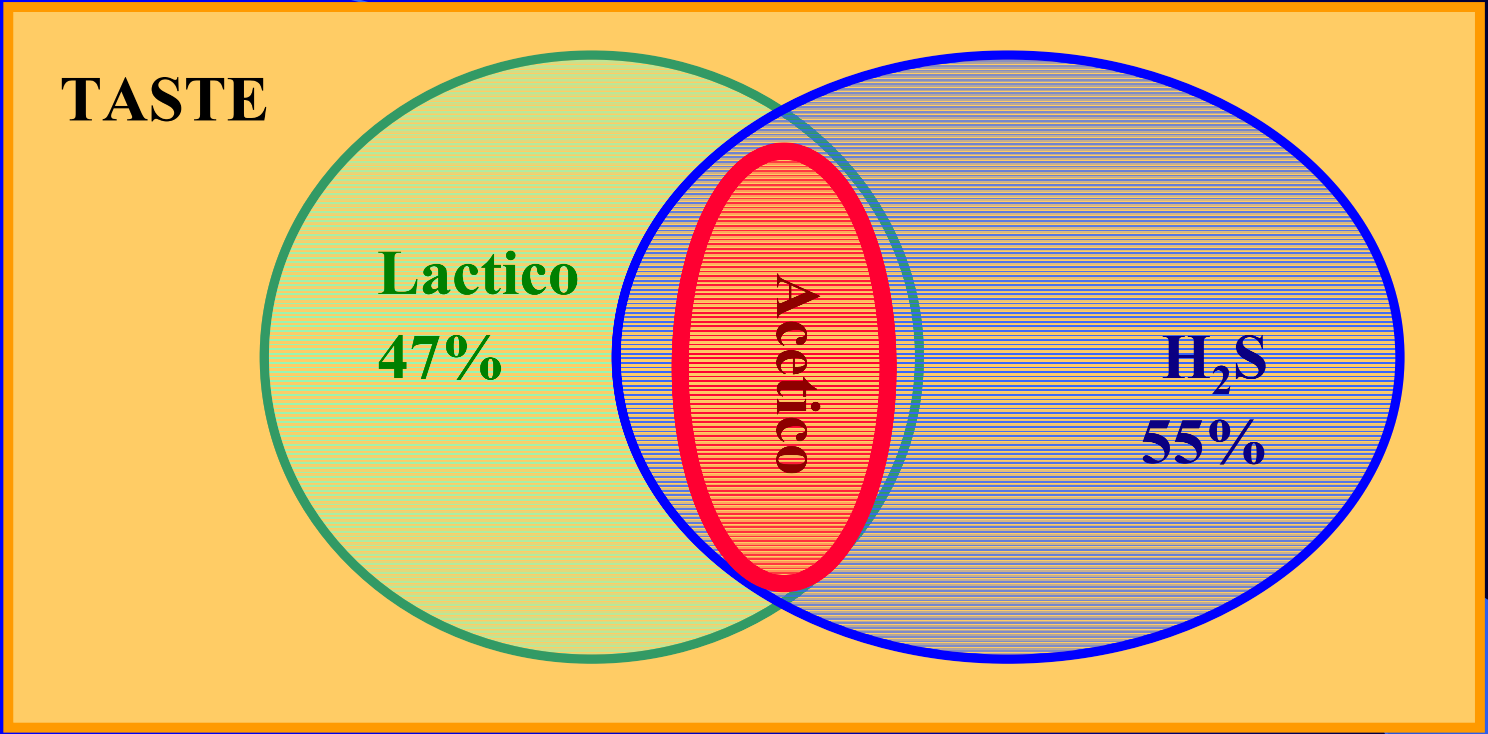
R-squared (adjusted for d.f.) = 55,1227 percent



Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-27,5918	8,98183	-3,07196	0,0048
Lactic	19,8872	7,95901	2,4987	0,0189
H2S	3,94627	1,13569	3,47477	0,0017

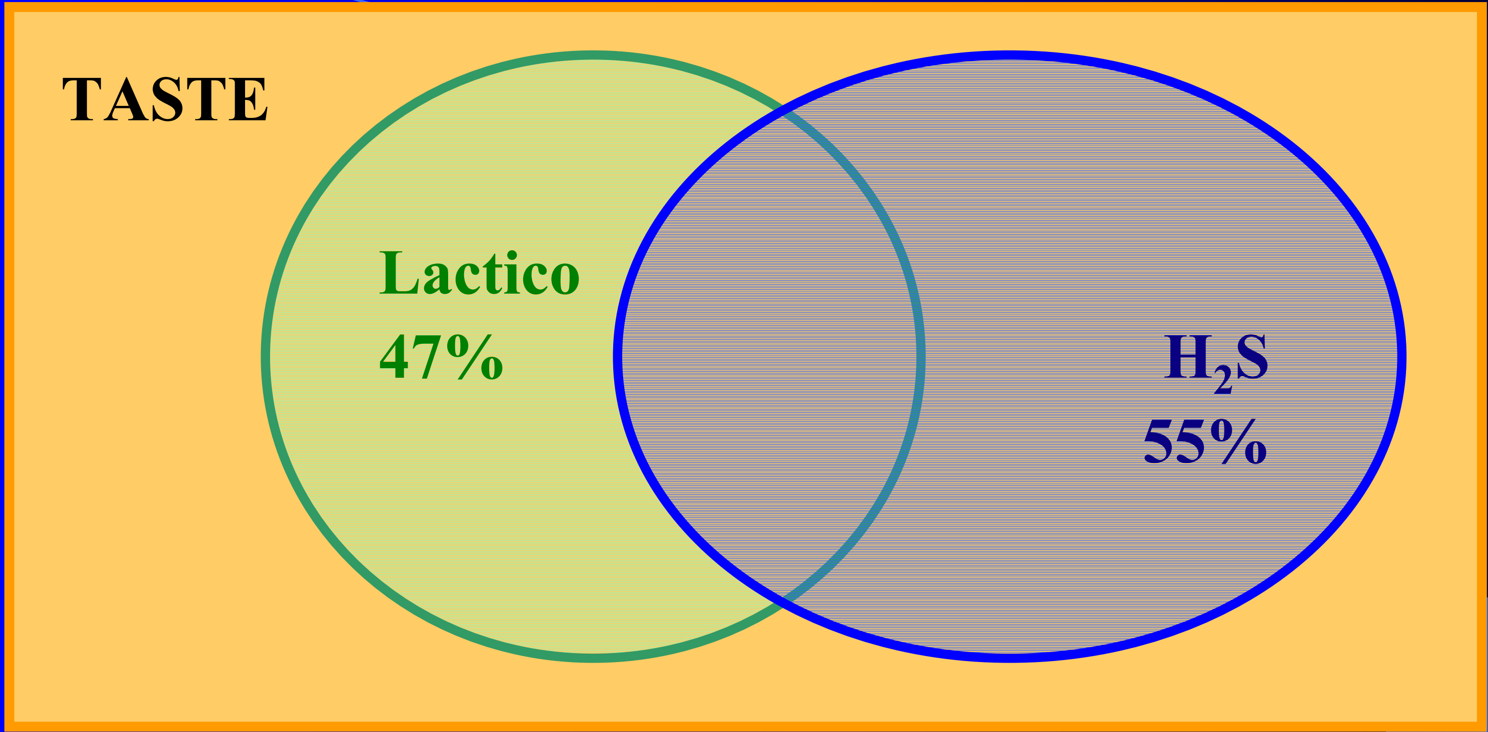
R-squared (adjusted for d.f.) = 62,5903 percent



Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-27,5918	8,98183	-3,07196	0,0048
Lactic	19,8872	7,95901	2,4987	0,0189
H2S	3,94627	1,13569	3,47477	0,0017

R-squared (adjusted for d.f.) = 62,5903 percent



Dependent variable: taste

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-27,5918	8,98183	-3,07196	0,0048
Lactic	19,8872	7,95901	2,4987	0,0189
H2S	3,94627	1,13569	3,47477	0,0017

R-squared (adjusted for d.f.) = 62,5903 percent

Estrategia para analizar problemas mediante regresiones

- Estudiar bien los datos. Puede haber errores
- Hacer gráficos X-Y
- Hacer regresiones simples:
  - Así sabemos qué variables son significativas

- Elegir la mejor regresión simple (Diagnosis+R2) e ir añadiendo variables
- Si una variable deja de ser significativa será por colinealidad

Estrategia para analizar problemas mediante regresiones

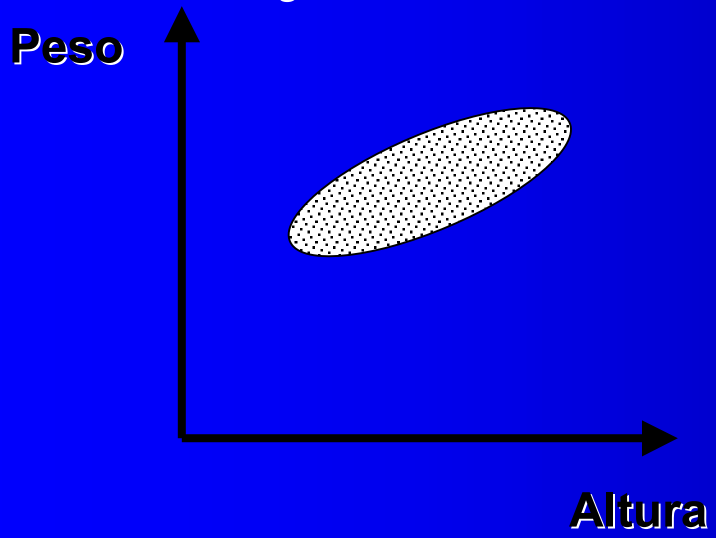
Podemos usar herramientas autmáticas: Stepwise

FIN de mlticolinealidad

Variables Cualitativas

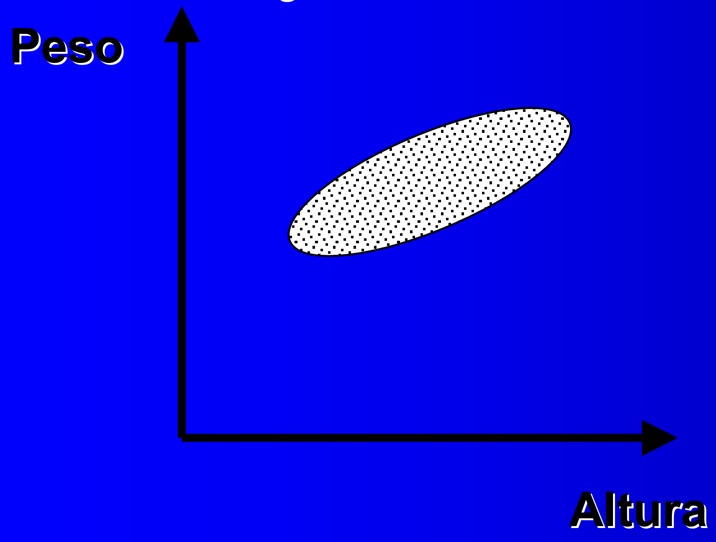
Estudiamos Pesos - Alturas

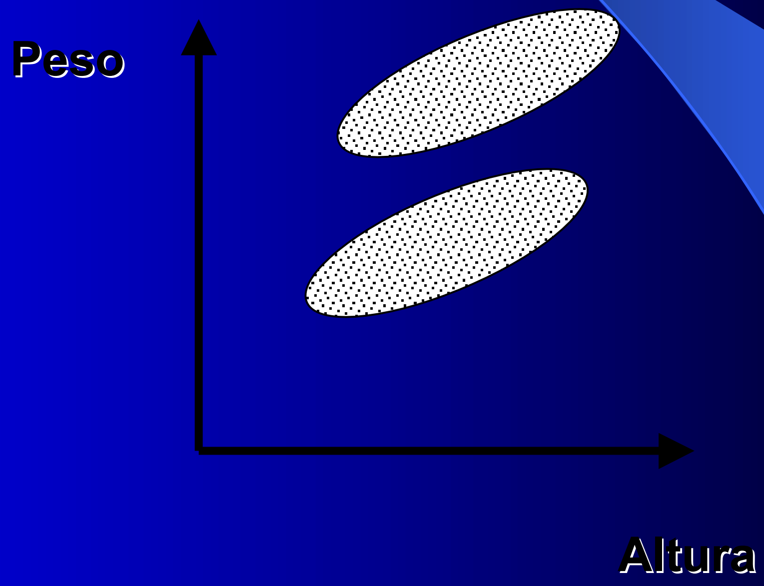
- **¿Es igual la relación para hombres que para mujeres?**



Estudiamos Pesos - Alturas

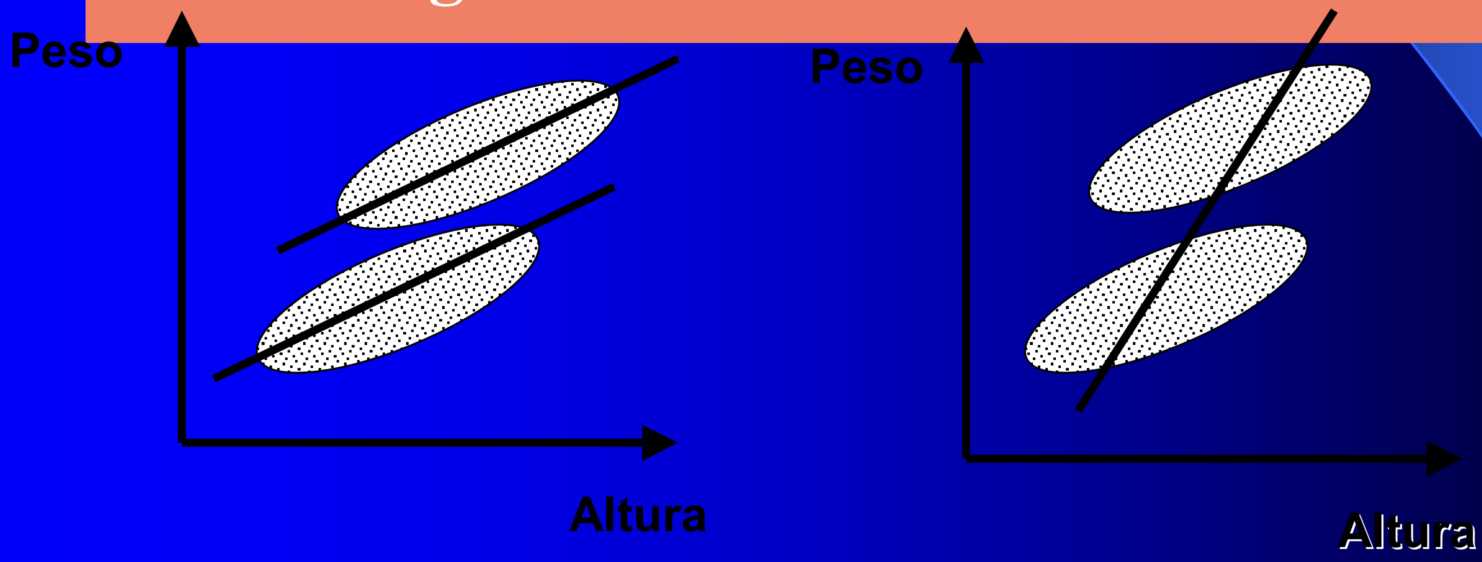
- **¿Es igual la relación para hombres que para mujeres?**





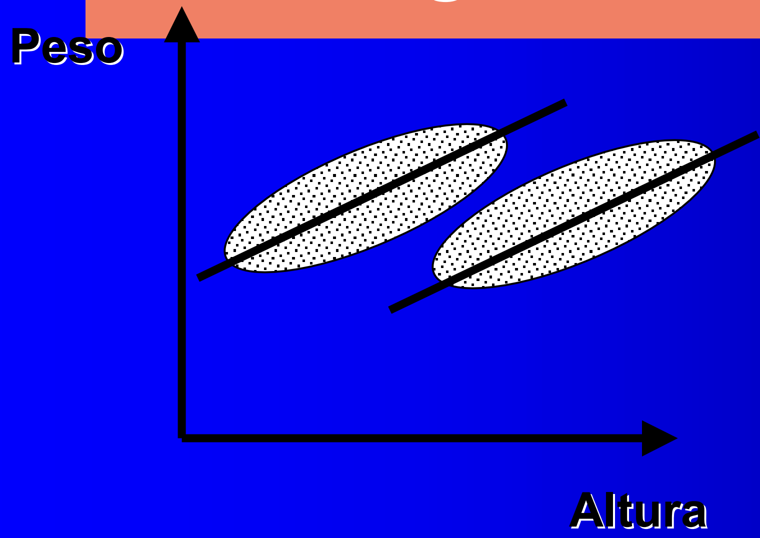
Estudiamos Pesos - Alturas

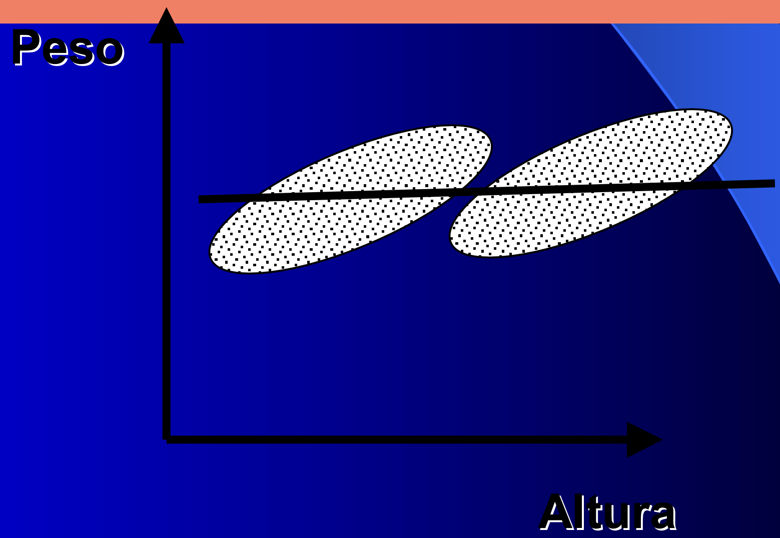
Si la relación no es igual, podemos cometer errores graves:



Estudiamos Pesos - Alturas

Si la relación no es igual, podemos cometer errores graves:

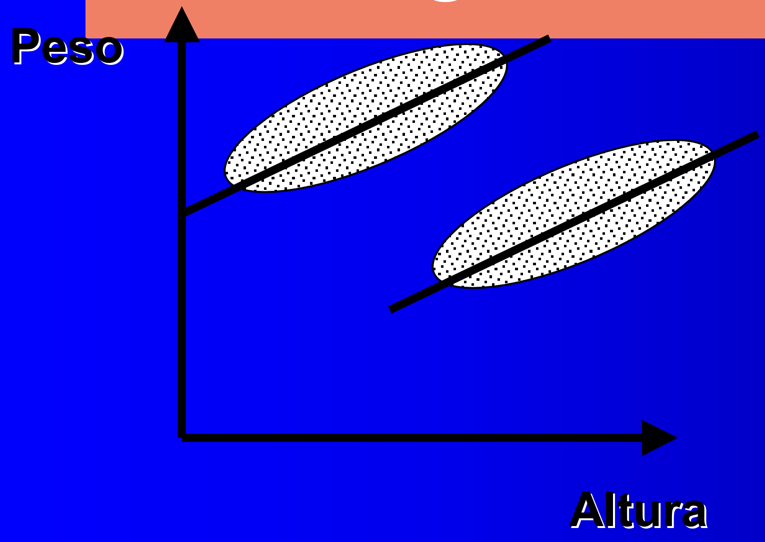


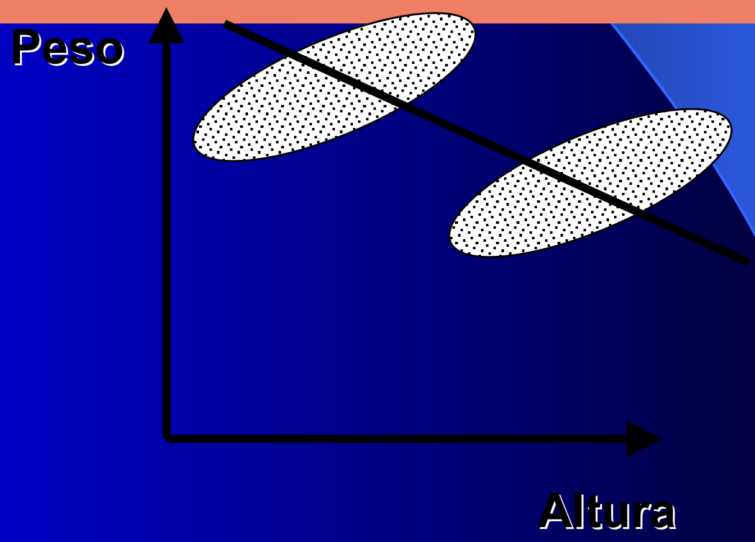


Creer que no es significativa x

Estudiamos Pesos - Alturas

Si la relación no es igual, podemos cometer errores graves:





Estimar la relación inversa

Ejemplos

Variable Y	*Variable X*	*Grupo que puede influir*
Peso	**Altura**	**Sexo: Hombre o Mujer**
Consumo de un trabajador	**Ingresos del trabajador**	**Status laboral: Paro o Empleado**
Consumo de un automóvil	**Potencia**	**Motor: Diesel o Gasolina**
Margen Ordinario de una Sucursal bancaria	**Comisiones**	**Sucursal: Urbana o Rural**

Es necesario introducir el grupo:

Para ello:

- **definiremos una variable *Z* que tome los siguientes valores:**

Z_i =0 si una observación pertenece al grupo A

Z_i=1 si una observación pertenece al grupo B

- **y estimaremos el siguiente modelo de regresión:**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z$$

El modelo que se estima:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z$$

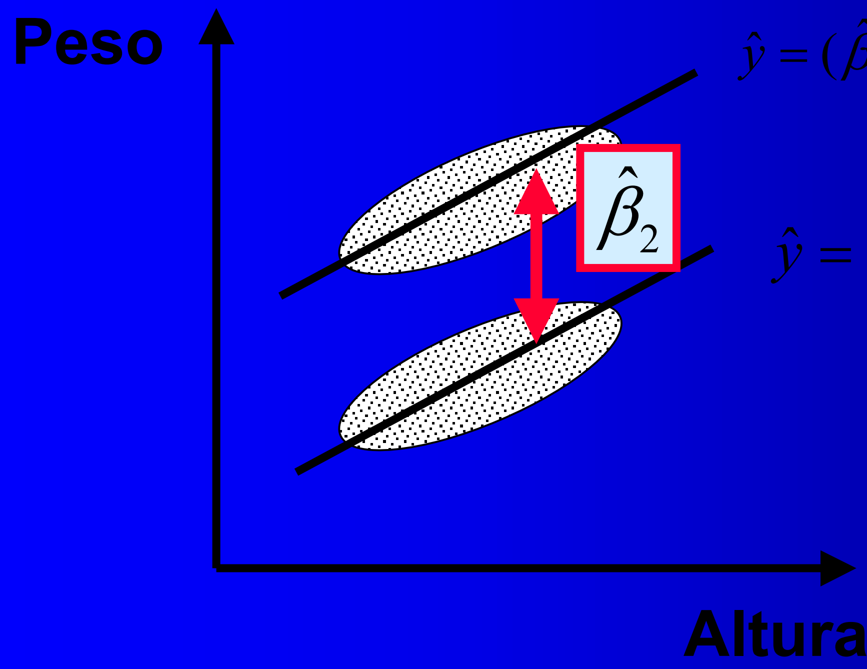
- **Mujeres: Les asignamos Z=0. Por tanto:**

$$\boxed{\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X}$$

- **Hombres: Les asignamos Z=1. Por tanto:**

$$\boxed{\hat{y} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X}$$

Por tanto:



$$\hat{y} = (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 X$$

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

El efecto es que un hombre de la misma altura pesa Beta$_2$ kilos más que una mujer de su misma altura.

¿O no?

Hagámoslo:

Dependent variable: peso

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-77,7888	16,0908	-4,83438	0,0000
altura	0,842013	0,0905752	9,29628	0,0000
sexo	-5,17748	2,20877	-2,34405	0,0208

R-squared = 60,8791 percent

R-squared (adjusted for d.f.) = 60,1927 percent

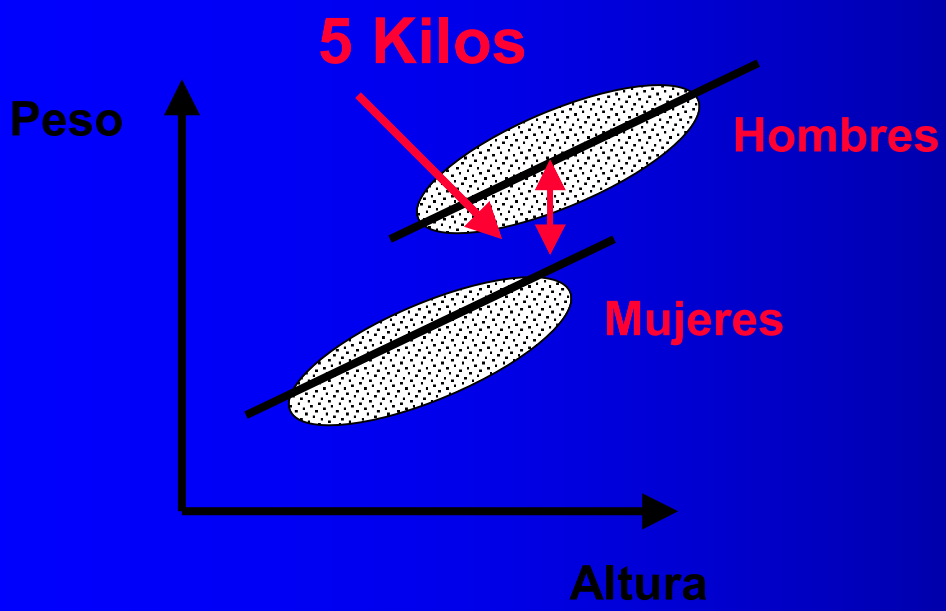
Sexo=0 Hombres
Sexo=1 Mujeres

Por tanto: un hombre que mida 180 pesará= -78+0.84x180=73 kilos

..... y una mujer de la misma altura pesará=-78+0.84x180-5.17=68 kilos

La diferencia existe porque t=-2.34 que es mayor que 2 en valor absoluto

Resultado



Millas por galón: ¿Influye el origen?

Dependent variable: mpg

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	46,1123	3,13585	14,7049	0,0000
Japon	3,82691	1,30589	2,93051	0,0071
horsepower	-0,216777	0,0363481	-5,96391	0,0000

R-squared = 66,4955 percent

R-squared (adjusted for d.f.) = 63,8152 percent

Japon=0 No Japonés
Japon=1 Japonés
1981

Millas por galón: ¿Influye el origen?

Multiple Regression Analysis

Dependent variable: mpg

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	49,0576	3,05719	16,0466	0,0000
USA	-3,56682	1,33607	-2,66962	0,0131
horsepower	-0,212448	0,0374777	-5,66864	0,0000

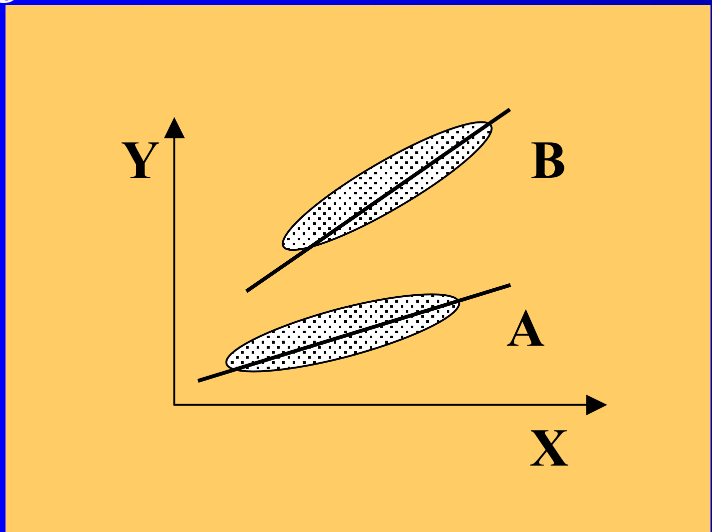
R-squared = 64,9719 percent

R-squared (adjusted for d.f.) = 62,1697 percent

USA=0
USA=1
1981

Interacciones

- Hemos supuesto que las rectas son paralelas.
- ¿Y si no lo son?



Modelización de las interacciones

La modelización de la interacción es sencilla. Hay que estimar un modelo de regresión entre:

- la variable Y
- la variable X
- la variable Z
- la interacción de X y Z que se modeliza por el producto (XZ).

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 Z + \hat{\beta}_3 XZ$$

Para el grupo con Z=0

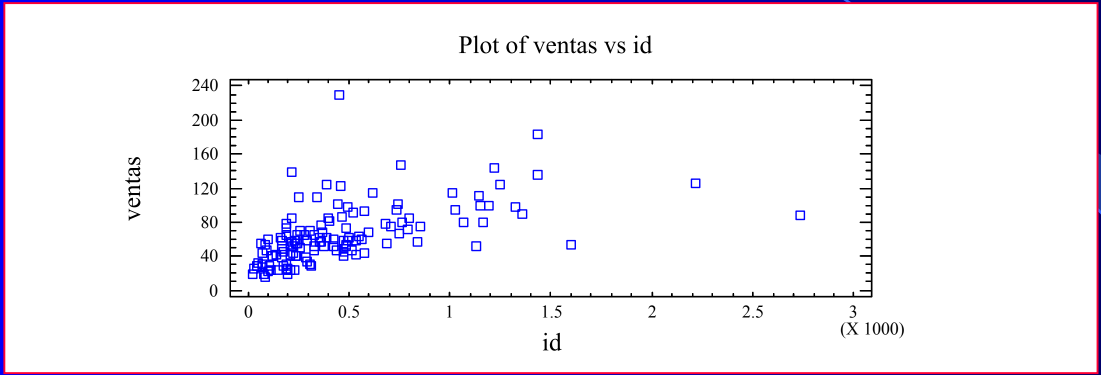
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

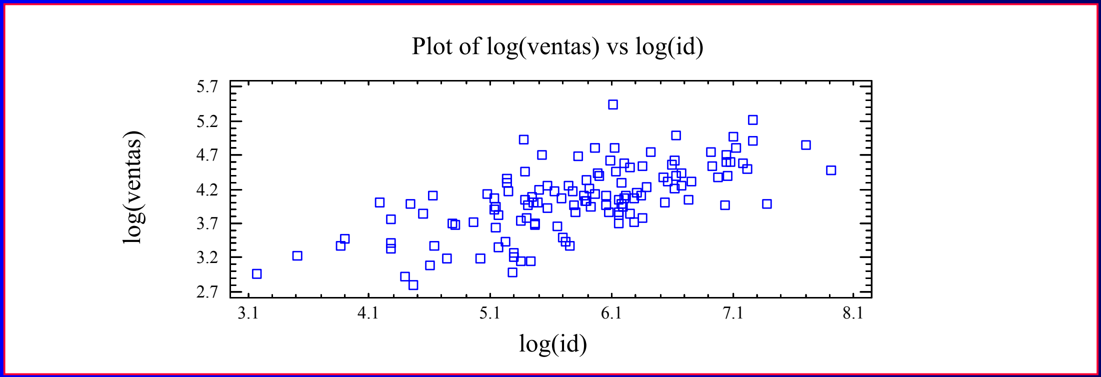
Para el grupo con Z=1

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 + \hat{\beta}_3 X = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) X$$

Por tanto, analizar si existe interacción se reduce a estimar un modelo de regresión y analizar si el parámetro es significativo (estadístico t mayor de 2) en la estimación realizada.

Ejemplo: Ventas de empresas del sector servicios en Madrid en función de su inversión en I+D





LOG(*VENTAS*) = 1.762 + 0.393 Log(ID)
(t) (7.88) (10.34) R^2 = 45.7 %

Queremos estudiar si hay diferencias por estar en el sector telecomunicaciones

z=1 Si está en el sector teleco

z=0 si no está en ese sector

$LOG(VENTAS) = 2.25 + 0.288\ Log(ID) + 0.527\ TELECO$

(t) (11.12) (8.08) (7.03) $R^2 = 61.05\%$

•Si la empresa funciona en el sector teleco:

$$Log(VENTAS) = 2.78 + 0.288\ log(ID)$$

•Si funciona en otro sector:

$$Log(VENTAS) = 2.25 + 0.288\ log(ID)$$

Estimamos la interacción:

$Log(VENTAS) = 1.99 + 0.334 Log(ID) + 1.80\ TELECO - 0.202\ TELECOxLog(ID)$

(t) (8.84) (8.40) (3.40) (-2.43) $R^2 = 62.8\%$

•Si no está en el sector teleco

$$Log(VENTAS) = 1.99 + 0.334\ log(ID)$$

•Si está en el sector teleco

$$Log(VENTAS) = 3.8 + 0.13\ log(ID)$$

Variables politómicas

Dicotómicas: Dos grupos.

Si hay más de dos grupos: Politómicas.

Variable Y	*Variable X*	*Grupos que pueden influir*
Ingresos de un trabajador	Años trabajando	Sector de actividad: •Agricultura •Industria •Construcción •Servicios
Consumo de un automóvil	Potencia	Origen del automóvil: •USA •Europa •Japón
Margen Ordinario de una Sucursal bancaria	Comisiones	Región Geográfica: •Madrid •Andalucía •Cataluña •Castilla •Extremadura

Introducción de V. Politómicas

1. Si la población puede dividirse en K grupos
2. Se generan k variables dicotómicas con criterio de pertenencia:

Ejemplo:

Estudiar ingresos de un trabajador en función de su experiencia laboral y el sector de actividad. El sector de actividad puede ser Aricultura, Industria, Construcción y Servicios

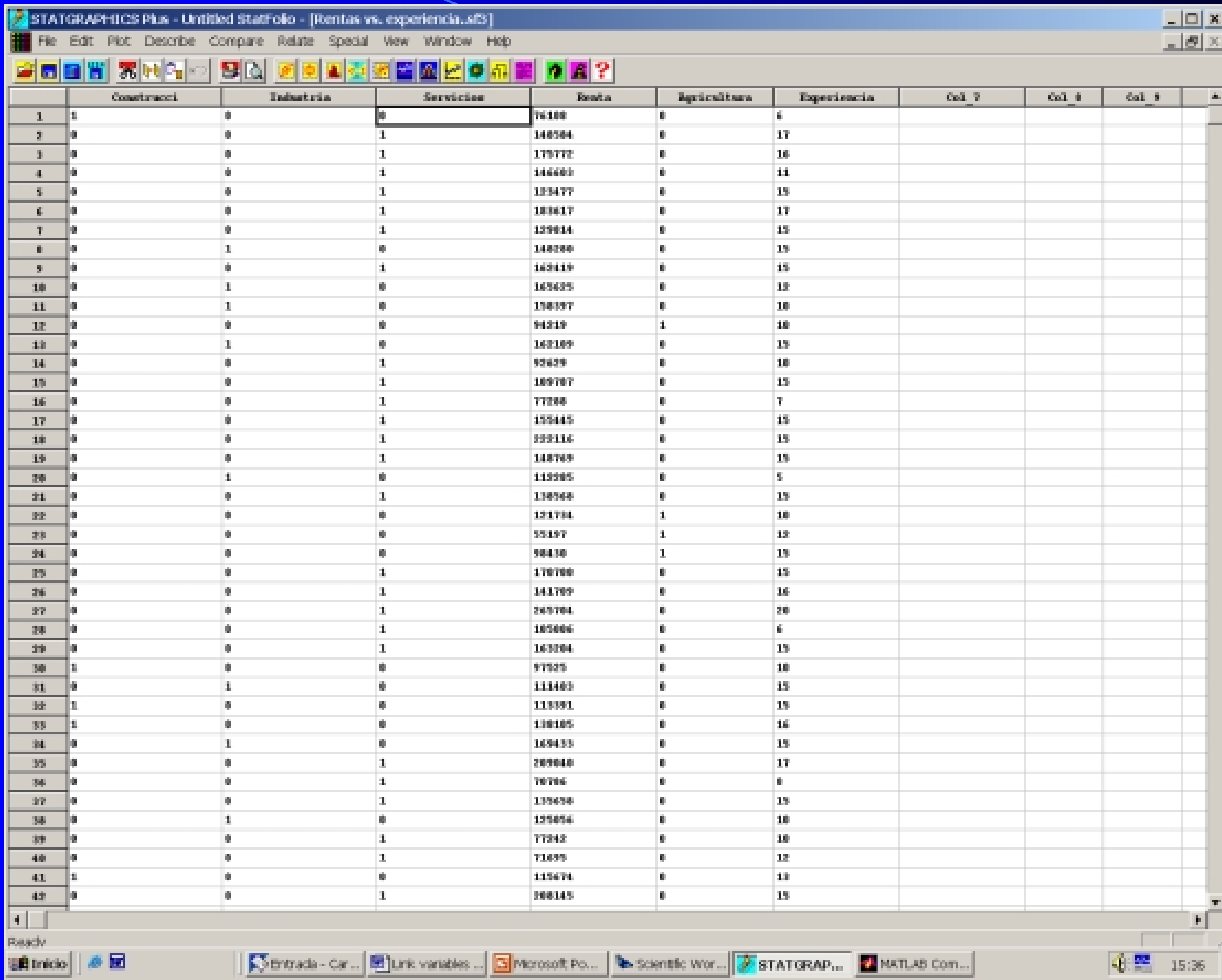
Generamos

AG=1 si el trabajador desarrolla su actividad en el sector agrícola (Si no vale 0)

IN=1 si el trabajador desarrolla su actividad en el sector industria (Si no vale 0)

CO=1 si el trabajador desarrolla su actividad en el sector construcción (Si no vale 0)

SE=1 si el trabajador desarrolla su actividad en el sector servicios (Si no vale 0)



	Construcci	Industria	Servicios	Renta	Agricultura	Experiencia	Col_7	Col_8	Col_9
1	1	0	0	76108	0	6			
2	0	0	1	148584	0	17			
3	0	0	1	179772	0	16			
4	0	0	1	146603	0	11			
5	0	0	1	123477	0	15			
6	0	0	1	183617	0	17			
7	0	0	1	129814	0	15			
8	0	1	0	148280	0	15			
9	0	0	1	162419	0	15			
10	0	1	0	165625	0	12			
11	0	1	0	150397	0	10			
12	0	0	0	94219	1	10			
13	0	1	0	162109	0	15			
14	0	0	1	92629	0	10			
15	0	0	1	109707	0	15			
16	0	0	1	77208	0	7			
17	0	0	1	155445	0	15			
18	0	0	1	222116	0	15			
19	0	0	1	148769	0	15			
20	0	1	0	112505	0	5			
21	0	0	1	130568	0	15			
22	0	0	0	121734	1	10			
23	0	0	0	55197	1	12			
24	0	0	0	98430	1	15			
25	0	0	1	170700	0	15			
26	0	0	1	141709	0	16			
27	0	0	1	265704	0	20			
28	0	0	1	105006	0	6			
29	0	0	1	163304	0	15			
30	1	0	0	97525	0	10			
31	0	1	0	111403	0	15			
32	1	0	0	113591	0	15			
33	1	0	0	130105	0	16			
34	0	1	0	169433	0	15			
35	0	0	1	209040	0	17			
36	0	0	1	70706	0	8			
37	0	0	1	135658	0	15			
38	0	1	0	125056	0	10			
39	0	0	1	77242	0	10			
40	0	0	1	71695	0	12			
41	1	0	0	115674	0	13			
42	0	0	1	208145	0	15			

Multiple Regression Analysis

Dependent variable: Renta

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	56631,1	7138,79	7,93288	0,0000
Experiencia	7075,51	506,786	13,9615	0,0000
Agricultura	-28161,0	3753,78	-7,50205	0,0000
Industria	-1626,43	4167,38	-0,390276	0,6965
Construcci	-21662,7	6061,56	-3,57378	0,0004

Trabajador de servicios (Miles de pesetas):

Renta=56+7 EXP   Si tiene 20 años de experiencia: Renta=56+140=196 mil pesetas

Trabajador de Agricultura

Renta=56+7 EXP –28   Si tiene 20 años de experiencia: Renta=56+140-28=168 mil pesetas

Trabajador de Industria (t<2) IGUAL QUE SERVICIOS

Renta=56+7 EXP   Si tiene 20 años de experiencia: Renta=56+140=196 mil pesetas

Trabajador de Construcción

Renta=56+7 EXP –22   Si tiene 20 años de experiencia: Renta=56+140-22=174 mil pesetas

FIN