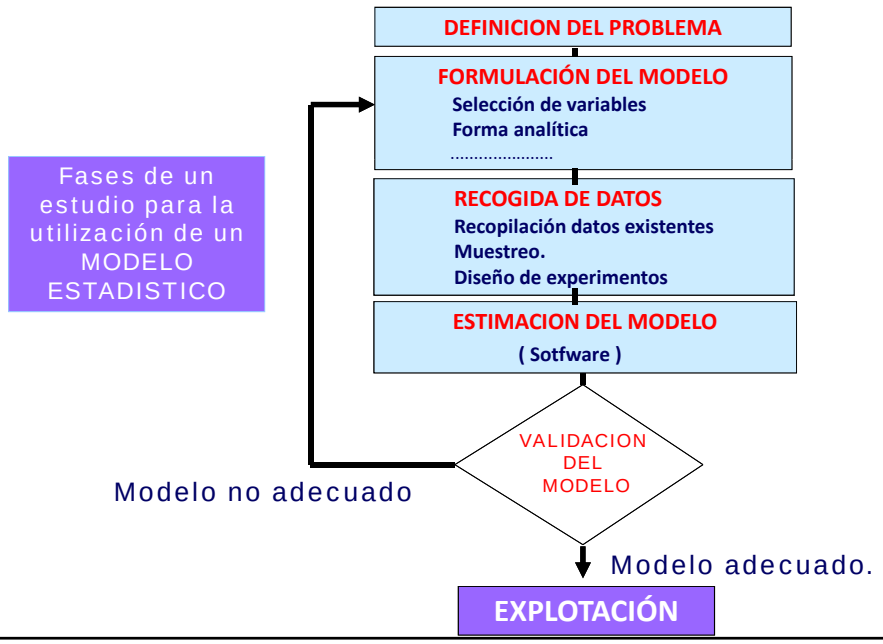
Tema 7. Regresión Lineal

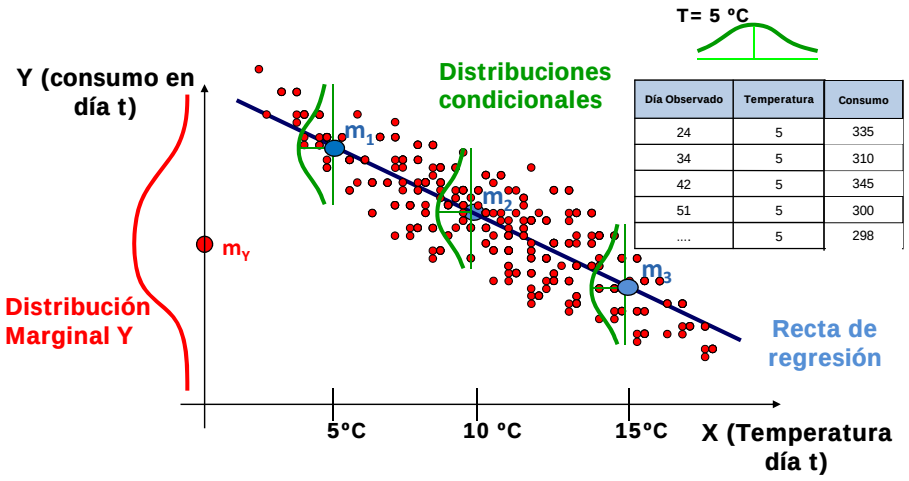
Dep. Estadística e IO Aplicadas y Calidad, Universidad Politécnica de Valencia

1. Modelos Estadísticos

Fases de un estudio para la utilización de un MODELO ESTADISTICO



2.Regresión lineal: modelo teórico



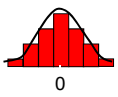
Día Observado	Temperatura	Consumo
24	5	335
34	5	310
42	5	345
51	5	300
....	5	298

$$m_{Y/(X=x_t)} = \beta_0 + \beta_1 \cdot x_t = m_t$$

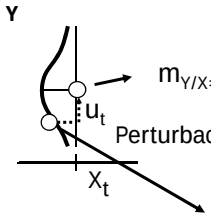
3. Regresión lineal: Perturbación

	X_t	Y_t			
Día Observado	Temperatura	Consumo	Día Semana	Codificación Día Semana	Climatizadores Encendidos
1	7,2	326,2	jueves	4	10
2	2,9	421,6	jueves	4	20
3	7,2	350,3	miercoles	3	12
4	9,1	294,9	jueves	4	9
5	10,7	233,7	jueves	4	8
6	15,6	176,0	jueves	4	5
7	7,0	282,3	jueves	4	10
8	16,0	154,0	jueves	4	5
9	17,0	136,5	jueves	4	3
10	2,0	426,7	lunes	1	20
11	0,2	473,7	lunes	1	20
12	13,0	214,9	lunes	1	6
13	13,4	231,4	lunes	1	6
14	7,5	313,2	lunes	1	10

$$Y_t = \beta_0 + \beta_1 \cdot x_t + u_t \qquad u_t \approx$$



Población | **Muestra**



$m_{Y/X=xt} = \beta_0 + \beta_1 x_t$ → $m_{Y/X = xt} = b_0 + b_1 x_t$

$\hat{\beta}_0 = b_0 \quad \hat{\beta}_1 = b_1 \quad \hat{u} = e$

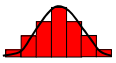
$y_t = \beta_0 + \beta_1 x_t + u_t$ → $y_t = b_0 + b_1 x_t + e_t$ Residuo

$\beta_0 \; \beta_1 \Rightarrow$ Constantes Desconocidas → $b_0 \; b_1 \Rightarrow$ Variables

Relación lineal existente entre dos componentes cuantitativas de una variable aleatoria bidimensional. (X,Y)

Estimación a partir de una muestra de la relación lineal existente entre las dos componentes cuantitativas de una variable aleatoria bidimensional

$$Y_t = \beta_0 + \beta_1 \cdot x_t + u_t \qquad u_t \approx$$



β_0 : Consumo medio los días en que la temperatura sea 0ºC

β_1 : Aumento del consumo medio de gas si la temperatura aumenta 1ºC (será <0)

4. Cálculo de la recta de predicción

Minimizar

$$\sum_{i=1}^{N}(y_i - (b_0 + b_1 \cdot x_i))^2 = \sum_{i=1}^{N} e_i^2$$

Obtenemos las **Estimaciones**:

$\hat{\beta}_0 = b_0$

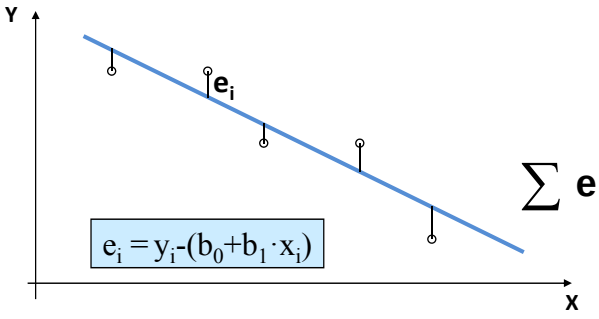
$\hat{\beta}_1 = b_1$

Precisión de las estimaciones

S_{b_0}

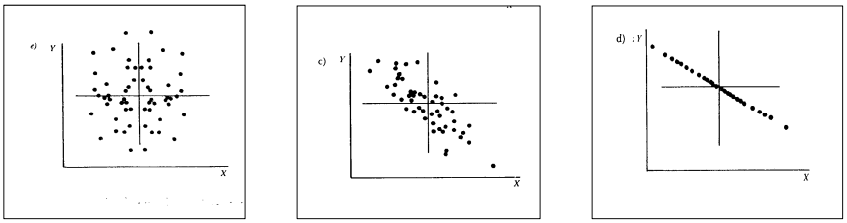
S_{b_1}

Residuos: errores de predicción de todos los individuos de la muestra



$$\sum e_i = 0$$

5.Coeficiente de Determinación R^2



↑ r (coeficiente de correlación)

↑ poder predictivo

↓ Sres (variabilidad de la distribución condicional)

En recta de regresión: $R^2 = r^2 * 100 \Rightarrow$ proporción de la variabilidad de **Y** explicada por la variable **X**

$R^2 \Rightarrow$ ANOVA $\Rightarrow$ (Variabilidad explicada Modelo /Variabilidad Total)* 100

$R^2 = (SC_{explicada} / SC_{total}) * 100$

5. Estimación del modelo

Recta de Regresión Consumo frente a Tª

Multiple Regression - CONSUMO

Multiple Regression - CONSUMO

Dependent variable: CONSUMO (consumo diario de gas)

Independent variables:

TEMPER (temperatura diaria)

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	448,913	7,63264	58,8148	0,0000
TEMPER	-18,4109	0,62714	-29,3569	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	552054,	1	552054,	861,83	0,0000
Residual	35230,8	55	640,56		
Total (Corr.)	587285,	56			

R-squared = 94,0011 percent

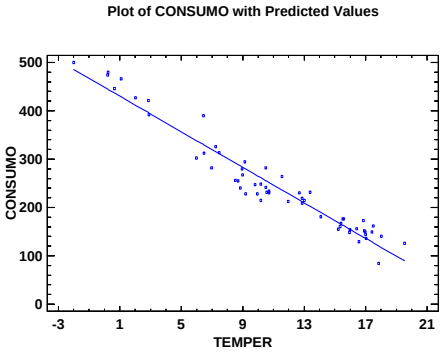
R-squared (adjusted for d.f.) = 93,922 percent

Standard Error of Est. = 25,3093

Mean absolute error = 20,5717

Durbin-Watson statistic = 1,38713 (P=0,0062)

Lag 1 residual autocorrelation = 0,292735



$$m_{Consumo/(T=T_t)} = 448.913 - 18.4109 \cdot T_t$$

$$S_{residual} = 25.3093$$

$$R^2 = \frac{SC_{explicada}}{SC_{total}} * 100 = \frac{552054}{587285} * 100 = 94.0011$$

5. Estimación del modelo

Inferencia sobre el valor de los parámetros

Multiple Regression - CONSUMO

Multiple Regression - CONSUMO

Dependent variable: CONSUMO (consumo diario de gas)

Independent variables:

TEMPER (temperatura diaria)

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	448,913	7,63264	58,8148	0,0000
TEMPER	-18,4109	0,62714	-29,3569	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	552054,	1	552054,	861,83	0,0000
Residual	35230,8	55	640,56		
Total (Corr.)	587285,	56			

R-squared = 94,0011 percent

R-squared (adjusted for d.f.) = 93,892 percent

Standard Error of Est. = 25,3093

Mean absolute error = 20,5717

Durbin-Watson statistic = 1,38713 (P=0,0062)

Lag 1 residual autocorrelation = 0,292735

¿Hasta que punto estamos seguros de que los **valores b_i estimados** para los parámetros poblacionales β_i del modelo no difieren de cero por azar del muestreo?

El valor $\beta_1 = 0$ supondría que la $\mathbf{X_1}$ no explicaría nada de la variabilidad de **Y** y se podría quitar dicha variable del modelo

$$m_{Y/(X=x_t)} = \beta_0 + 0 \cdot x_t = \beta_0$$

5. Estimación del modelo

Inferencia sobre el valor de los parámetros: Test Global

Multiple Regression - CONSUMO

Multiple Regression - CONSUMO

Dependent variable: CONSUMO (consumo diario de gas)

Independent variables:

TEMPER (temperatura diaria)

		Standard	T	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	448,913	7,63264	58,8148	0,0000
TEMPER	-18,4109	0,62714	-29,3569	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	552054,	1	552054,	861,83	0,0000
Residual	35230,8	55	640,56		
Total (Corr.)	587285,	56			

R-squared = 94,0011 percent

R-squared (adjusted for d.f.) = 93,892 percent

Standard Error of Est. = 25,3093

Mean absolute error = 20,5717

Durbin-Watson statistic = 1,38713 (P=0,0062)

Lag 1 residual autocorrelation = 0,292735

$$F_{ratio} = \frac{552054/1}{35230.8/55} = \frac{552054}{640.56} = 861,83$$

Rechazar H_0

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_I = 0$$

$$H_1 : \exists \text{ al menos una } \beta_i \neq 0$$

- Si sale aceptar H_0 :

Ninguna variable o termino explicativo incluida en el modelo **tiene un efecto poblacional real** en la variable respuesta

- Si se cumple H_0:

$$F_{ratio} = \frac{SC_{EXP}/I}{SC_{RES}/(N-1-I)} = \frac{CM_{exp}}{CM_{res}} \sim F_{I,N-1-I}$$

La H_0 se rechazará, si la F_{ratio} supera el valor en tablas $F_{I,N-1-I}(\alpha)$

5. Estimación del modelo

Inferencia sobre el valor de los parámetros: Test individual

Multiple Regression - CONSUMO

Multiple Regression - CONSUMO

Dependent variable: CONSUMO (consumo diario de gas)

Independent variables:

TEMPER (temperatura diaria)

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	448,913	7,63264	58,8148	0,0000
TEMPER	-18,4109	0,62714	-29,3569	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	552054,	1	552054,	861,83	0,0000
Residual	35230,8	55	640,56		
Total (Corr.)	587285,	56			

R-squared = 94,0011 percent

R-squared (adjusted for d.f.) = 93,892 percent

Standard Error of Est. = 25,3096

Mean absolute error = 20,5717

Durbin-Watson statistic = 1,38713 (P=0,0062)

Lag 1 residual autocorrelation = 0,292735

$$H_0: \beta_i = 0$$
$$H_i: \beta_i \neq 0$$

El test para β_0 no se tendrá en cuenta y la constante se mantendrá siempre en el modelo

❑ Si sale aceptar H_0 : **El termino i del modelo NO tiene un efecto poblacional real en la variable respuesta y puede ser quitado del modelo**

❑ Si se cumple H_0:

$$t_{calculada} = \frac{b_i}{S_i} \sim t_{gl_{residuales}} = t_{N-I-I}$$

La H_0 se rechazará si la $t_{calculada} \notin [-t_{N-I-1}(\alpha/2) \quad +t_{N-I-1}(\alpha/2)]$

$$t_{calculada} = \frac{-18.4109}{0.62714} = -29.3569$$

$$\notin [-t_{55}(2.5\%) \quad +t_{55}(2.5\%)] = [-2 \quad +2]$$

$$\beta_1 \neq 0$$

6. Generalización del modelo

Inclusión de más variables explicativas de naturaleza cuantitativa:

$$E(Y / X_1 = x_{1j}, ..., X_I = x_{Ij}) = \beta_0 + \beta_1 x_{1j} + ... + \beta_I x_{Ij}$$

Inclusión de variables explicativas cualitativas

Si tiene K variantes: Introduciendo K - 1 variables **DUMMY** (valores 0 ó 1)

Ejemplo: sea un modelo donde la variable respuesta Y (Rendimiento de un proceso) y como variable explicativa el "catalizador" (usando 3 catalizadores A,B,C) como única variable explicativa.

Como tiene 3 variantes tendré que crear dos variables Dummy (C_A y C_B) donde:

	Variables	
Catalizador	C_A	C_B
A	1	0
B	0	1
C	0	0

$$E(Y / Cat) = \beta_0 + \beta_1 C_A + \beta_2 C_B$$

6. Generalización del modelo

Inclusión de interacciones:

Se incluirán términos producto entre las variables que interactúan:

$$E(Y / X_1 = x_{1j}, X_2 = x_{2j}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{1j} x_{2j}$$

Inclusión de relaciones no lineales:

Se incluirán términos de segundo grado en las variables que actúan cuadráticamente.

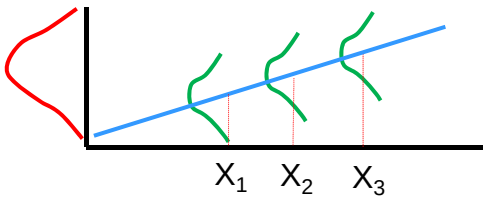
$$E(Y / X_1 = x_{1j}) = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{1j}^2$$

7. Supuestos de los modelos de regresión lineal

HOMOCEDASTICIDAD:
(Y_j /x) varianza constante y desconocida

NORMALIDAD:
Y y las (Y_j /x) siguen una distribución normal ⇒ residuos e "normales"

INDEPENDENCIA ENTRE LAS VARIABLES EXPLICATIVAS X_i



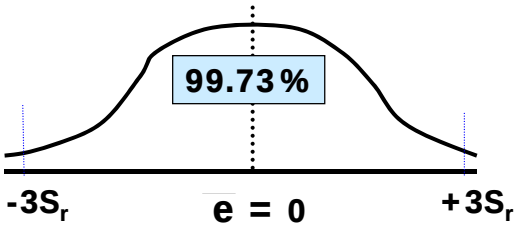
$E(Y)= \beta_0 + \beta_1 X$

7. Análisis de Residuos

Datos anómalos

Se identifican por residuos e_i mayores en valor absoluto que:

$2S_{res}$ (95%) ó $2.58S_{res}$ (98%) $3S_{res}$ (99.73%)



ALARMA !!

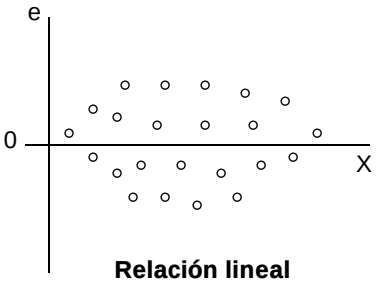
8. Análisis de Residuos

No normalidad de los datos

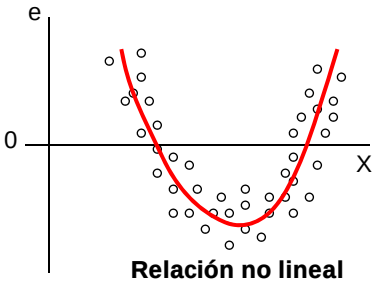
Puede estudiarse representando los residuos e_i en papel probabilístico normal

No linealidad de la relación entre E(Y) y X

Puede estudiarse representando los e_i en función de X_i



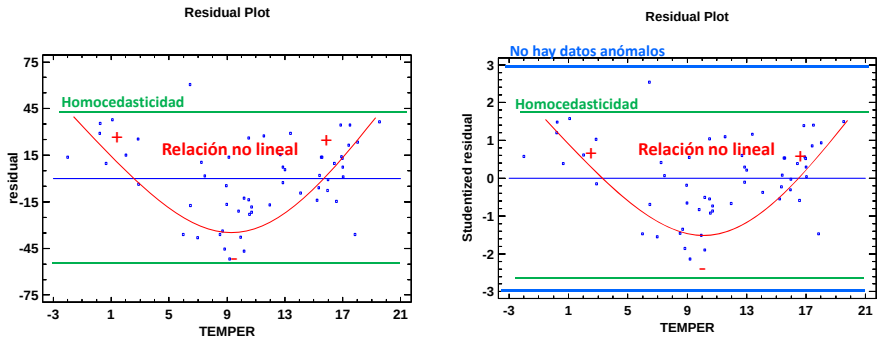
Relación lineal



Relación no lineal

8. Análisis de Residuos

Gráfico de "Residuos frente a X"



El modelo queda **validado**: Se cumplen los supuestos del modelo y no se han usado datos anómalos para su estimación

Ejercicio

Ejercicio. Hallar un intervalo centrado donde se encontrará el consumo del 95% de los días en donde la Tª es de 10ºC

Multiple Regression - CONSUMO

Multiple Regression - CONSUMO

Dependent variable: CONSUMO (consumo diario de gas)
Independent variables:
TEMPER (temperatura diaria)

		Standard	*T*	
Parameter	*Estimate*	*Error*	*Statistic*	*P-Value*
CONSTANT	448,913	7,63264	58,8148	0,0000
TEMPER	-18,4109	0,62714	-29,3569	0,0000

Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	552054,	1	552054,	861,83	0,0000
Residual	35230,8	55	640,56		
Total (Corr.)	587285,	56			

R-squared = 94,0011 percent
R-squared (adjusted for d.f.) = 93,892 percent
Standard Error of Est. = 25,3093
Mean absolute error = 20,5717
Durbin-Watson statistic = 1,38713 (P=0,0062)
Lag 1 residual autocorrelation = 0,292735

Ejercicio

Ejercicio. Hallar un intervalo centrado donde se encontrará el consumo del 95% de los días en donde la Tª es de 10ºC usando un modelo donde se ha introducido la Tª con efecto cuadrático

¿El efecto cuadrático de la Tª será significativo? (justifica la respuesta)

¿Cuál será el R^2 del nuevo modelo?

Regresión Múltiple - CONSUMO

Variable dependiente: CONSUMO (consumo diario de gas)

Variables independientes:

TEMPER (temperatura diaria)

TEMPER^2

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	472,351	8,56841	55,127	0,0000
TEMPER	-25,9865	1,83411	-14,1685	0,0000
TEMPER^2	0,400966	0,0926855	4,32609	0,0001

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	561122,	2	280561,	579,07	0,0000
Residuo	26163,3	54	484,505		
Total (Corr.)	587285,	56			

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Tema 7. Regresión Lineal

Dep. Estadística e IO Aplicadas y Calidad, Universidad Politécnica de Valencia

Fuentes:
Material docente : S. Vidal y F. Villa

20