UNIDAD DIDÁCTICA 5

INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

(4ª parte: Introducción a la regresión lineal)

Objetivos

1. Estudiar métodos descriptivos para analizar la relación entre las dos componentes de una variable bidimensional numérica.

2. Estudiar el modelo de regresión lineal simple.

Contenidos

1. Estadística descriptiva bidimensional

    1.1 Diagrama de dispersión

    1.2 Covarianza

    1.3 Coeficiente de correlación

Contenidos

2. Modelo de regresión lineal simple
   2.1 Introducción
   2.2 Planteamiento y estimación del modelo
   2.3 Recta de regresión y residuos
   2.4 Coeficiente de determinación R^2. ANOVA del modelo
   2.5 Test de hipótesis sobre los parámetros del modelo
   2.6 Validación del modelo: análisis de residuos

3. Modelo de regresión con efecto lineal y cuadrático

Estadística descriptiva bidimensional

- En la mayor parte de los problemas reales se estudia más de una característica aleatoria sobre cada individuo de la población.

- Cuando se observan los valores de <u>dos características</u>:

$$\Downarrow$$

Variable aleatoria bidimensional

<u>Ejemplo 1</u>: En un sistema informático en red se estudia el **TIEMPO** que tarda en ejecutarse un programa prueba (*benchmark*) y el **NÚMERO DE USUARIOS** conectados a una determinada hora.

Estadística descriptiva bidimensional

Ejemplo 2:

- Para el control del consumo de energía en calefacción de una factoría durante los meses de invierno, se anota diariamente el **CONSUMO** (termias) y la **TEMPERATURA** (ºC a las 12):

$$\Downarrow$$

Muestra de esta variable bidimensional

Todos los pares de valores (**TEMPERATURA**, **CONSUMO**) durante los días laborables de los meses de invierno.

Estadística descriptiva bidimensional

Ejemplo 3: En la población constituida por los estudiantes universitarios españoles se observa la ESTATURA (cms) y el PESO (kgs) de cada estudiante.

Una muestra de esta variable bidimensional puede estar formada por los 130 pares de valores constatados en los 130 alumnos que contestaron la encuesta.

Estadística descriptiva bidimensional

- En el estudio de variables aleatorias bidimensionales numéricas no basta con conocer las características (posición, dispersión, forma,...) de cada una de las dos componentes.

- Tienen también interés práctico estudiar si hay relación entre dichas componentes.

- Y si la hay:
  - Cuantificar el grado de relación
  - Obtener un modelo que describa dicha relación

Estadística descriptiva bidimensional

- Método inferencial: Regresión
  - Modelo que refleje la relación lineal existente
  - Predicción
  - Ejemplo: relación entre TIEMPO que se tarda en ejecutar un programa prueba y el NºUSUARIOS.

Estadística descriptiva bidimensional

- Cuando la variable bidimensional es numérica se puede estudiar a nivel descriptivo la relación entre sus dos componentes con
  - Gráficos: Diagrama de dispersión
  - Parámetros
    - Covarianza
    - Correlación

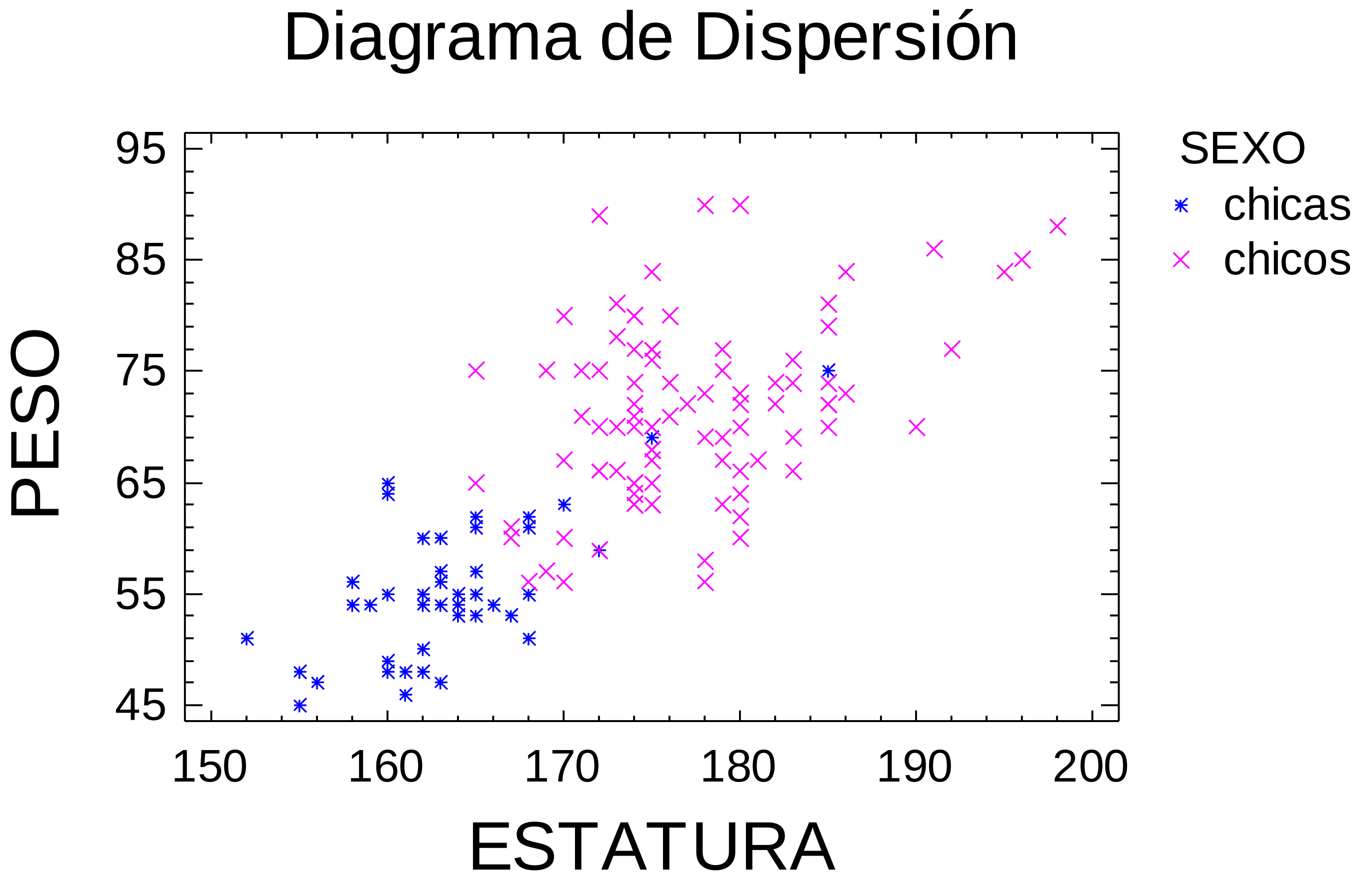
Diagramas de dispersión

- Una forma sencilla de describir gráficamente las relaciones constatadas entre las dos componentes, consiste en representar cada observación por un punto en un plano, cuya abcisa es el valor de la primera componente y la ordenada es el de la segunda.

- A este tipo de gráfico se le denomina **diagrama de dispersión**.

Diagramas de dispersión

Ejemplo



Relación positiva y lineal

Diagramas de dispersión

- El diagrama pone claramente de manifiesto una **relación positiva** entre las dos variables estudiadas, que se refleja en una nube de puntos cuyo eje principal tiene un sentido creciente, como consecuencia del hecho de que, en términos generales, los individuos más altos pesan más que los más bajos.
- El diagrama también pone de manifiesto que las chicas tienen en general valores menores de ambas variables que los chicos, pero que la relación entre PESO y ESTATURA es **lineal positiva** en ambos sexos.

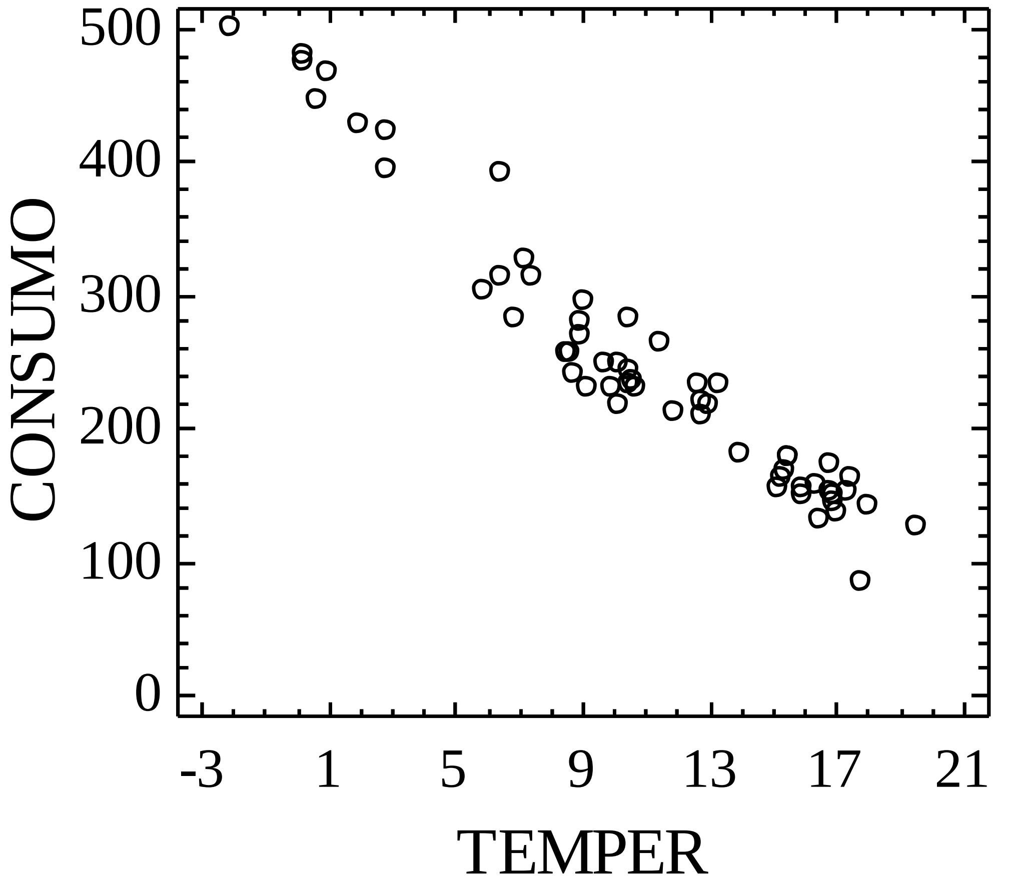
Diagramas de dispersión

Ejemplo

EJERCICIO 1*: Para estudiar un ejemplo en el que el Diagrama de Dispersión pone claramente en evidencia una relación negativa entre dos variables, obtener el diagrama para las variables TEMPER y CONSUMO del fichero gas.sf3*

Diagramas de dispersión

Ejemplo



Covarianza. Coeficiente de correlación

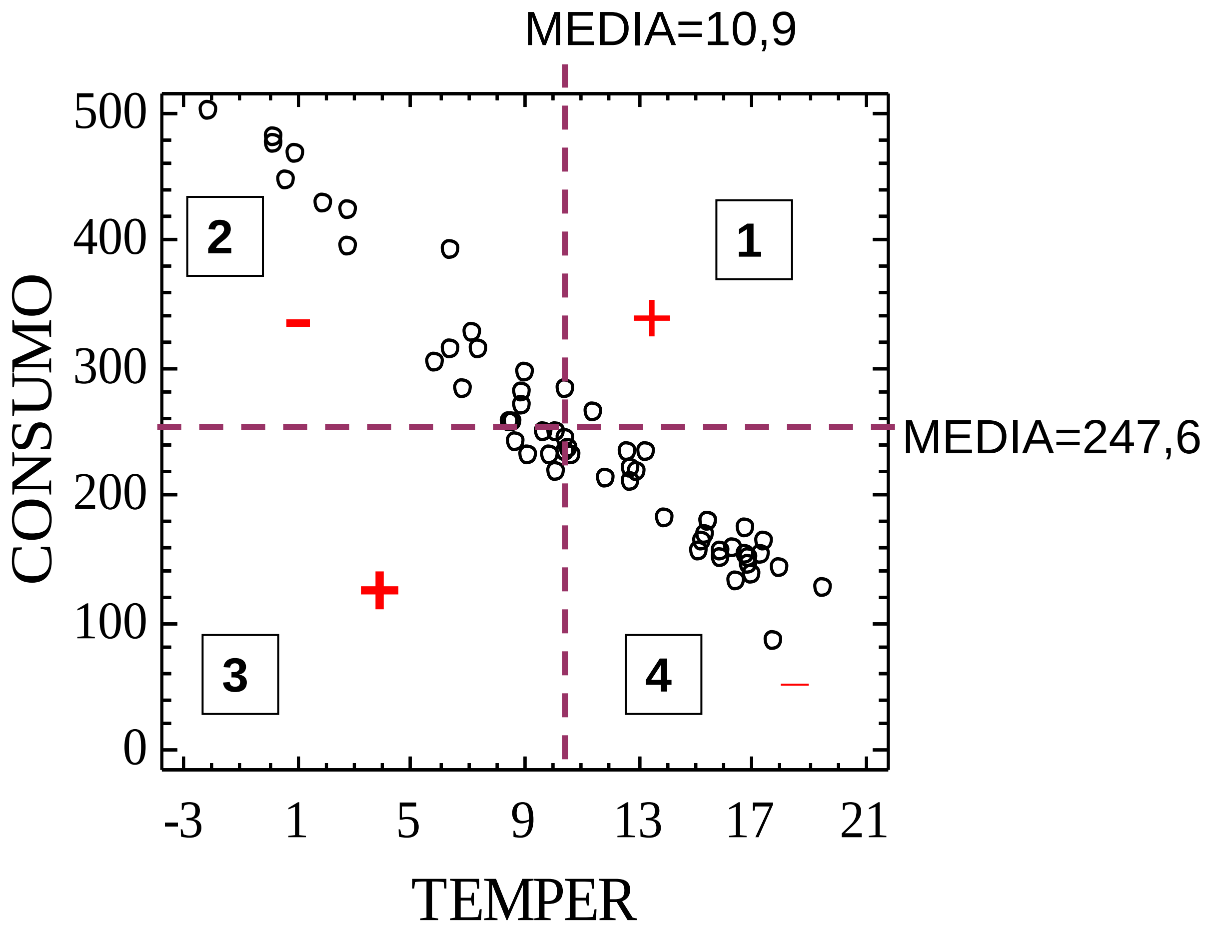
- Cuanto más estrechamente se agrupen los puntos del diagrama de dispersión alrededor de una recta, más fuerte es el grado de relación lineal existente entre las dos variables consideradas.

- Con el fin de cuantificar en un índice numérico el grado de relación lineal existente entre dos variables, se utilizan en Estadística dos parámetros:

Covarianza

Coeficiente de correlación

$$\sum^{N}(X_i - \overline{X})(Y_i - \overline{Y}) < 0$$



Covarianza

$$\mathrm{Cov}(X,Y) = S_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{N-1}$$

- Inconveniente: Depende de las unidades de medida de las componentes X e Y.

- Ejemplo: La covarianza entre ESTATURA y PESO será 100 veces mayor si la primera variable se mide en cm que si se mide en m.

Matriz de varianzas-covarianzas

- Caso bidimensional: En la diagonal aparecen las varianzas de X y la de Y. Es simétrica, pues fuera de la diagonal aparece la Cov(X,Y)

$$\overline{\overline{V}} = \begin{Bmatrix} s^2{}_X & Cov(X, Y) \\ Cov(X, Y) & s^2{}_Y \end{Bmatrix}$$

Matriz de varianzas-covarianzas

- Ejemplo: Matriz de varianzas-covarianzas de TEMPER y CONSUMO

Tamaño muestra

Varianza TEMPER

	TEMPER	CONSUMO
TEMPER	29,0832	-535,449
	(57)	(57)
CONSUMO	-535,449	10487,2
	(57)	(57)

Covarianza

Varianza CONSUMO

Coeficiente de correlación lineal

- Para obviar este problema se utiliza universalmente en Estadística el coeficiente de correlación lineal, que no es más que la covarianza dividida por el producto de las desviaciones típicas de las dos variables X e Y.

$$r(X,Y) = \frac{Cov(X,Y)}{s_X s_Y}$$

Coeficiente de correlación lineal

Propiedades:

- Se mantiene inalterable si cualquiera de las dos variables X e Y sufre una transformación lineal.
  - Ejemplo: El coeficiente de correlación r entre CONSUMO y TEMPERATURA no se modifica por el hecho de que esta última se exprese en grados Farenheit en vez de en grados centígrados.
- r(X,Y) está siempre comprendido entre -1 y +1

Coeficiente de correlación lineal

Propiedades:

- Los valores extremos ,-1 y +1, sólo se alcanzan si existe relación lineal exacta entre X e Y, o sea, si los puntos del diagrama de dispersión están exactamente alineados en una recta. (Valor +1 si la recta es creciente y -1 si es decreciente).

- No hay relación lineal entre X e Y $\Rightarrow$ coeficiente de correlación=0. (En la práctica en una muestra de dos variables independientes r(X,Y) estará "cercano" a cero debido al azar de muestreo)

Coeficiente de correlación lineal

Propiedades:

- Cuanto más estrecho es el grado de relación lineal existente entre dos variables, más cercano a 1 es el valor de r (o a -1 si la relación es decreciente)

- Un valor de r nulo o cercano a cero indicará una relación lineal inexistente o muy débil.

- El cuadrado del coeficiente de correlación mide la proporción (o porcentaje si se multiplica por 100) de la varianza de Y que está asociada a la variabilidad de X.

Coeficiente de correlación lineal

EJERCICIO 2*: Calcula los coeficientes de correlación entre ESTATURA y PESO, entre EDAD y ESTATURA, y entre TEMPER y CONSUMO.*

¿Hasta qué punto las diferencias de peso entre los alumnos están asociadas a las diferencias de estatura entre ellos?

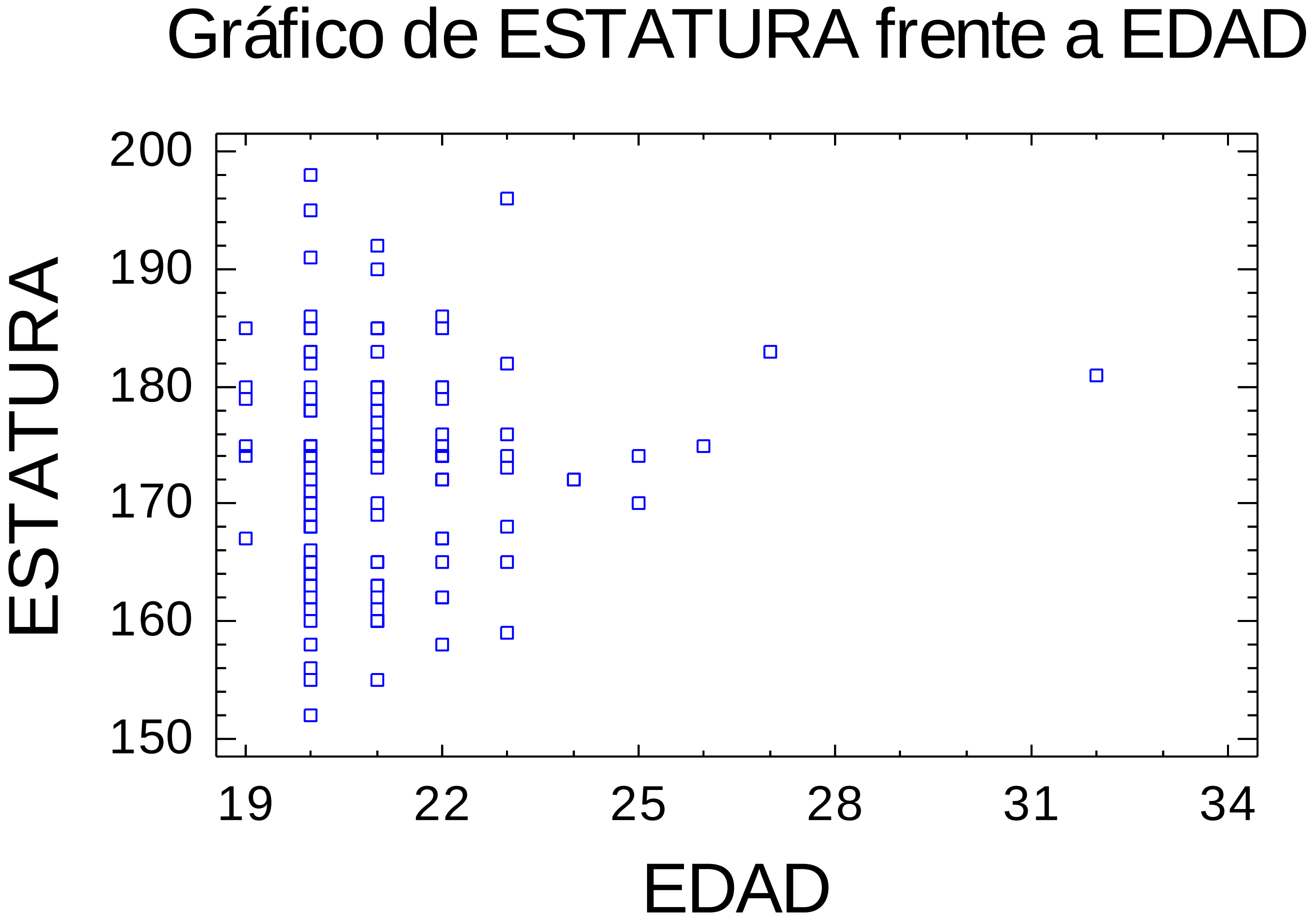
r(ESTATURA, PESO)= 0,7404

r(EDAD, ESTATURA)= 0,0868

r(TEMPER, CONSUMO)=-0,9695

r^2 (ESTATURA, PESO)=$0{,}7404^2$ =0,548 $\Rightarrow$ El 54,8% de las diferencias de peso entre alumnos están asociadas a las diferencias de estatura entre ellos.

Coeficiente de correlación lineal



Coeficiente de correlación lineal

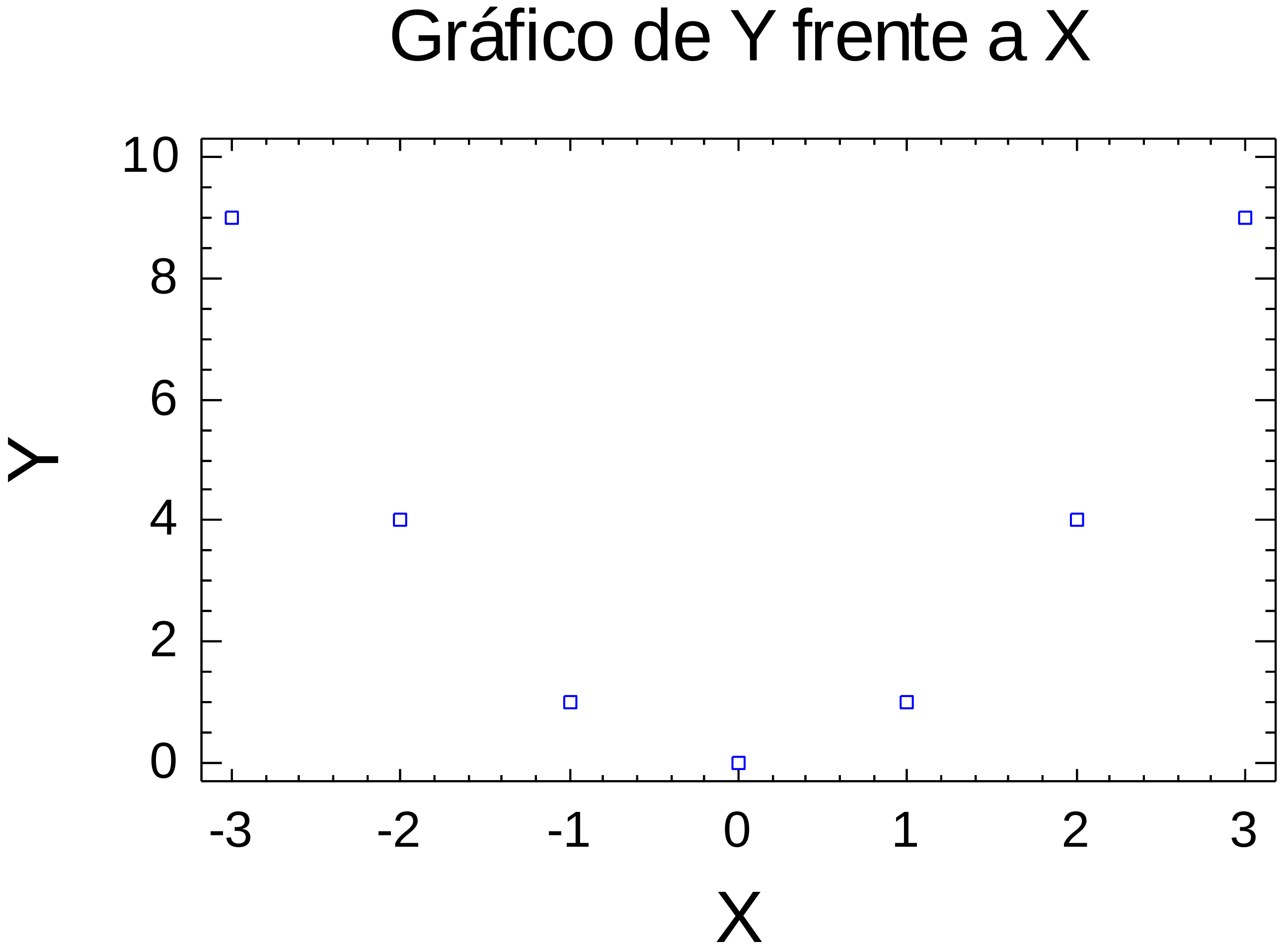
- Es importante resaltar que tanto la <u>covarianza</u> como el <u>coeficiente</u> de correlación <u>miden sólo el grado de relación lineal</u> existente entre dos variables. Dos variables pueden tener una relación estrecha y sin embargo resultar r cercano a cero por ser dicha relación no lineal.

EJERCICIO 3: *Introducir en Statgraphics dos variables: una X con valores -3, -2, -1, 0, 1, 2, 3 y otra Y de valores 9, 4, 1, 0, 1, 4, 9.*

Dibujar el diagrama de dispersión y hallar el coeficiente de correlación entre ambas.

¿Están relacionadas las variables? ¿Lo están linealmente?

Coeficiente de correlación lineal



$$r_{X,Y} = 0$$

Matriz de correlaciones

- Caso bidimensional: En la diagonal aparecen el valor 1 (r(X, X) y r(Y, Y)). Es simétrica pues fuera de la diagonal está r(X,Y)

$$\overline{\overline{R}} = \begin{Bmatrix} 1 & r(X,Y) \\ r(X,Y) & 1 \end{Bmatrix}$$

Matriz de correlaciones

- Ejemplo: Matriz correlaciones de TEMPER y CONSUMO.

r(TEMPER, TEMPER)

r(TEMPER,CONSUMO)

	TEMPER	CONSUMO
TEMPER	1	-0,9695
		(57)
		0,0000
CONSUMO	-0,9695	1
	(57)	
	0,0000	

Tamaño muestra

r(CONSUMO,CONSUMO)

r(CONSUMO, TEMPER)

Regresión lineal simple. Introducción

- **Objetivo de los Modelos de Regresión Lineal:**

  analizar la posible relación existente entre la pauta de variabilidad de una variable aleatoria Y y

  los valores de una o más variables X_1, X_2, X_3,....,X_l, (aleatorias o no), de las que la primera puede depender.

Regresión lineal simple. Introducción

Los modelos de regresión se utilizan:

1. Cuando no es posible fijar los valores a adoptar por las variables explicativas en un estudio.

   Por ejemplo:

   el efecto de la temperatura diaria en el consumo de energía de una instalación,

   o el efecto de la carga en el tiempo de respuesta de un sistema informático.

Regresión lineal simple. Introducción

2. Cuando se analiza información histórica, que no se ha obtenido con un diseño experimental. Por ejemplo:

los datos procedentes del control estadístico de un proceso recopilados el último año,

o los datos de una encuesta.

Regresión lineal simple. Introducción

- En un estudio de regresión se dispone de:

N observaciones de una variable aleatoria Y_j (por ejemplo, el consumo diario de energía constatado en una factoría automovilística en N días invernales),

junto con los valores correspondientes de I variables (aleatorias o no) $X_{1j},\ldots,X_{Ij}$, de las que la primera puede depender.

(por ejemplo, la temperatura y la producción de vehículos en dichos días).

Regresión lineal simple. Introducción

Variable aleatoria Y

Variables explicativas X_1, X_2, X_3,....,X_I

Datos:

y_1	$x_{11}, x_{21}, \cdots, x_{I1}$
y_2	$x_{12}, x_{22}, \cdots, x_{I2}$

....

y_N	$x_{1N}, x_{2N}, \cdots, x_{IN}$

Regresión lineal simple. Introducción

- Se trata de estudiar las relaciones entre la distribución de Y_j y los valores de las X_{ij}.
- A la Y se le denomina variable **dependiente**, **explicada**, **endógena** o **respuesta**,
- mientras que a las X_i se les llama variables **independientes**, **explicativas**, **exógenas** o **regresores**.

Regresión lineal simple. Introducción

- **Los modelos clásicos de regresión asumen que cada observación y_j es el valor observado de una variable aleatoria Y_j normal, de varianza $\sigma^2(Yj)$ constante desconocida, y cuyo valor medio es una función de los valores constatados de las X_{ij}.**

$$E(Y_j) = f(X_{1j} \ldots X_{Ij})$$

(ecuación de regresión)

- En principio, por tanto, el posible efecto de las X_i sobre la distribución de Y se concreta en modificar su valor medio.

Regresión lineal simple. Introducción

- Los parámetros de la ecuación de regresión permiten precisar la naturaleza y cuantificar la magnitud de los efectos de las variables explicativas, sobre el valor medio de la variable dependiente.
- Dichos parámetros se estiman a partir de los datos disponibles, utilizando un procedimiento estadístico que se expone en este tema, y se analiza su significación mediante las técnicas de inferencia correspondientes.

Regresión lineal simple. Introducción

Ejemplo:

El responsable del control de consumo de energía de una factoría, desea saber si el consumo de 290 termias constatado el día anterior, puede considerarse “normal” sabiendo que la temperatura fue de 10ºC.

Dado que el consumo no depende sólo de la temperatura sino de otros factores (humedad, viento, volumen de producción, etc), es de esperar que aún no habiendo anomalías, el consumo en la población constituida por los días en que la temperatura es 10ºC fluctuará aleatoriamente.

Pero, en promedio ¿cuánto se consumirá los días en que la temperatura sea 10ºC? Con toda seguridad menos que lo que se consumirá en promedio los días en que la temperatura sea de 5ºC.

Pero ¿cuánto menos?

Regresión lineal simple. Introducción

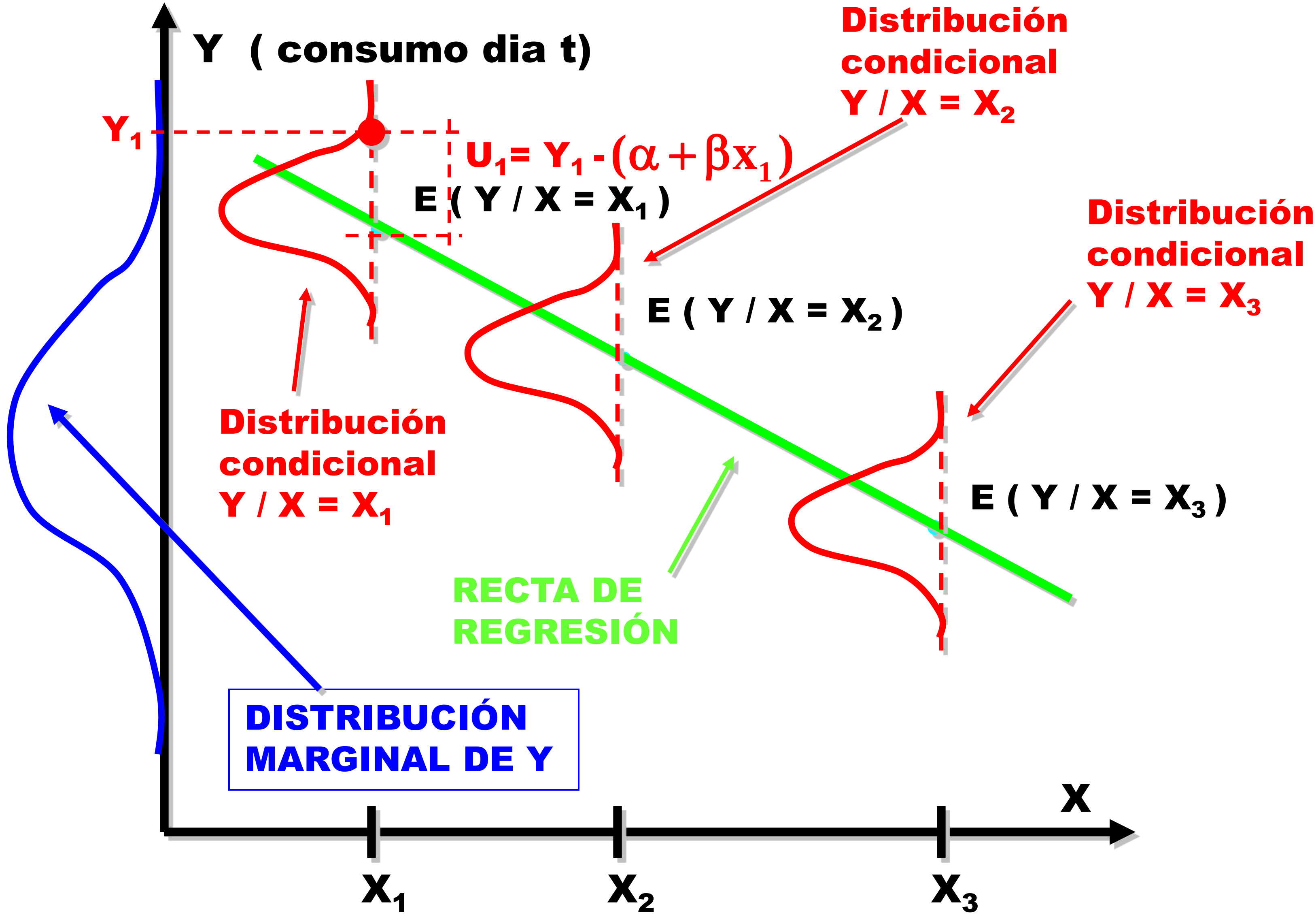
- Para responder a preguntas como la anterior se utiliza en Estadística la **recta de regresión**.

- Mediante esta recta se pretende predecir el valor que en promedio corresponde a una variable **Y**, cuando otra variable **X** tiene una valor determinado $\Rightarrow$ E(Y/X).

Regresión lineal simple. Introducción

- En la figura siguiente, si se considera la población de todos los días de invierno, se constata que el consumo fluctúa de unos días a otros (distribución marginal de **Y**).

- Estas fluctuaciones se deben a muchas causas; una de ellas, cuyo efecto queremos cuantificar mediante el modelo de regresión, es la variabilidad de la temperatura de unos días a otros.


Y (consumo dia t)
Distribución condicional Y / X = X2
Y1
U1= Y1 - (α + βx1)
E (Y / X = X1)
E (Y / X = X2)
Distribución condicional Y / X = X3
Distribución condicional Y / X = X1
E (Y / X = X3)
RECTA DE REGRESIÓN
DISTRIBUCIÓN MARGINAL DE Y
X
X1
X2
X3

Regresión lineal simple. Introducción

- En los días invernales con temperatura x_1 baja, (por ejemplo, 5ºC), se constata también una variabilidad en los consumos de energía (distribución condicional de **Y** cuando **X**=x_1).
- Esta variabilidad será, con, toda seguridad, menor que la existente en la distribución marginal de **Y**, porque en ella no estará influyendo el efecto de la variabilidad en la temperatura diaria, puesto que ésta es fija (x_1) todos estos días.
- La distribución condicional de **Y** cuando **X**=x_1, tendrá un valor medio E(**Y**/**X**=x_1), que posiblemente será superior al valor medio E(**Y**) de la distribución marginal de **Y**, por estar considerándose sólo días con una temperatura baja.

Regresión lineal simple. Introducción

- De forma análoga, podríamos definir para otros valores posibles de la temperatura **X**, por ejemplo x_2 y x_3, las distribuciones condicionales (**Y**/**X**=x_2) e (**Y**/**X**=x_3), cada una de ellas con su correspondiente valor medio.

Recta de regresión. Planteamiento del modelo

- Básicamente el modelo de regresión lineal simple asume que la distribución condicional del consumo los días en que la temperatura es x_t, es una **variable aleatoria normal** cuya varianza σ^2 no depende de x_t, pero cuya media es una función lineal $\alpha + ß\, x_t$ de dicho valor:

$$E(Y/X=x_t) = \alpha + ß\, x_t$$

$$\sigma^2(Y/X=x_t) = \sigma^2 \text{ (constante)}$$

Recta de regresión. Planteamiento del modelo

- Se dispone de un conjunto de N pares de valores observados (x_t, y_t) es decir de los valores de la temperatura y del consumo en N días diferentes.
- Denominando u_t a la diferencia entre el consumo observado el día t (y_t) y el consumo correspondiente en promedio a los días cuya temperatura es x_t :

$$u_t = y_t - (\alpha + \beta\, x_t)$$

Recta de regresión. Planteamiento del modelo

- Se deduce inmediatamente de las hipótesis anteriores que las u_t (a las que se denomina **perturbaciones** aleatorias) tienen todas distribuciones normales, con media nula e idéntica varianza σ^2 :

$$\mathbf{E(u_t) = 0 \qquad\qquad \sigma^2(u_t) = \sigma^2}$$

- Adicionalmente, se asume que las u_t correspondientes a diferentes observaciones son independientes entre sí.

Recta de regresión. Planteamiento del modelo

- El modelo puede, en consecuencia, escribirse también de la forma alternativa:

$$y_t = \alpha + ß\, x_t + u_t$$

donde u_t son valores de variables $N(0,\sigma^2)$ independientes.

- En el modelo anterior,

α= consumo promedio los días en que la temperatura es 0ºC

ß= incremento del consumo medio (probablemente negativo por tratarse de una época invernal) que cabe esperar por cada grado de aumento de la temperatura diaria.

Recta de regresión. Planteamiento del modelo

- Por su parte, u_t recoge el efecto que sobre el consumo en el día t han tenido todos los restantes factores no incluidos explícitamente en el modelo (es decir todo lo que puede afectar al consumo de energía de un día excepto el efecto (lineal) de la temperatura de dicho día).
- La relación entre σ^2 y la varianza de la distribución marginal de Y, será un índice de la importancia de todos estos factores excluidos del modelo y, en consecuencia, de la adecuación de éste para predecir el consumo mediante una simple relación lineal con la temperatura.

Recta de regresión. Estimación del modelo

Objetivos

Una vez planteado el modelo

$$y_j = \alpha + \beta x_j + u_j$$

donde se asume que u_j son variables aleatorias $N(0, \sigma^2)$ independientes,

en la siguiente etapa del análisis se estiman los 2 parámetros $\alpha^*=a$, $\beta^*=b$,

la precisión de estas estimaciones s_a y s_b,

y la varianza residual $\sigma^{2*} = s^2$

Recta de regresión. Estimación del modelo

Dados unos datos:

y_1	x_1
...	....
y_j	x_j
...	....
y_N	x_N

y mediante el recurso a un software adecuado, el procedimiento de estimación por mínimos cuadrados, consiste en obtener los paramétros a y b que minimizan la suma de los residuos al cuadrado.

Recta de regresión. Estimación del modelo

- La recta resultante es, de todas las rectas posibles, la que minimiza la suma de cuadrados de las desviaciones en el sentido vertical de la variable Y, que es la que se desea predecir

$$\min \sum_{j=1}^{N} \left(y_j - \left(\alpha + \beta x_j\right)\right)^2 = \min \sum_{j=1}^{N} u_j^2$$

- Valores a y b (estimaciones) que minimizan dicha suma de cuadrados:

$$b = r \frac{s_y}{s_x}$$

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

Recta de regresión y residuos

- Una herramienta muy útil para complementar cualquier estudio de regresión es el análisis de los residuos.
- Se denomina **residuo** de un dato a la diferencia entre el valor y_i del mismo y el valor $a + b\, x_i$ que se predice para el valor medio de Y en los individuos de la población en los que la variable X vale x_i.

$$\textbf{Residuo}_i = y_i - (a + b\, x_i)$$

Recta de regresión y residuos

- Ejemplo: En el estudio para controlar el consumo de energía, el residuo para un día determinado

Residuo día i = consumo día i-(a + b temperatura día i)

- Dicho residuo recoge el efecto que en el día i han tenido otras variables que influyen sobre el consumo, incluyendo las posibles anomalías que se hayan producido.

- Para controlar el consumo de energía, un procedimiento adecuado sería, por tanto, calcular el residuo cada día y ver si su valor es o no admisible.

Recta de regresión y residuos

<u>Propiedades de los residuos</u>:

- El valor medio de los residuos para todos los datos utilizados en un estudio de regresión es siempre cero.

- La varianza de los residuos permite estimar el orden de magnitud del efecto conjunto de todos los restantes factores no considerados al calcular la recta de regresión.

$$s^2_{residual} = \text{Varianza de Y} - \text{Varianza explicada por la recta}$$

$$\Downarrow$$

$$s^2_{residual} = s^2_Y - s^2_Y\, r^2_{XY} = s^2_Y\,(1 - r^2_{XY})$$

Recta de regresión y residuos

Propiedades de los residuos:

- El coeficiente de correlación lineal al cuadrado se interpreta como la proporción de variabilidad de Y asociada a la variabilidad de X.

- Se denomina también **coeficiente de determinación**.

- Este coeficiente mide la calidad del ajuste de regresión.

Recta de regresión y residuos

EJERCICIO 4: *Obtener con el Statgraphics la recta de regresión del consumo en función de la temperatura.*

¿Qué consumo cabe esperar en promedio los días en los que la temperatura es 10ºC?

¿El consumo de 290 termias constatado un día en que la temperatura fue 10ºC, puede considerarse anormalmente elevado e indicador de que algo ha funcionado mal?

Recta de regresión y residuos

- Salida que se obtiene mediante Statgraphics del ajuste de regresión lineal simple del consumo frente a la temperatura (opción *Relate…Simple Regression*):

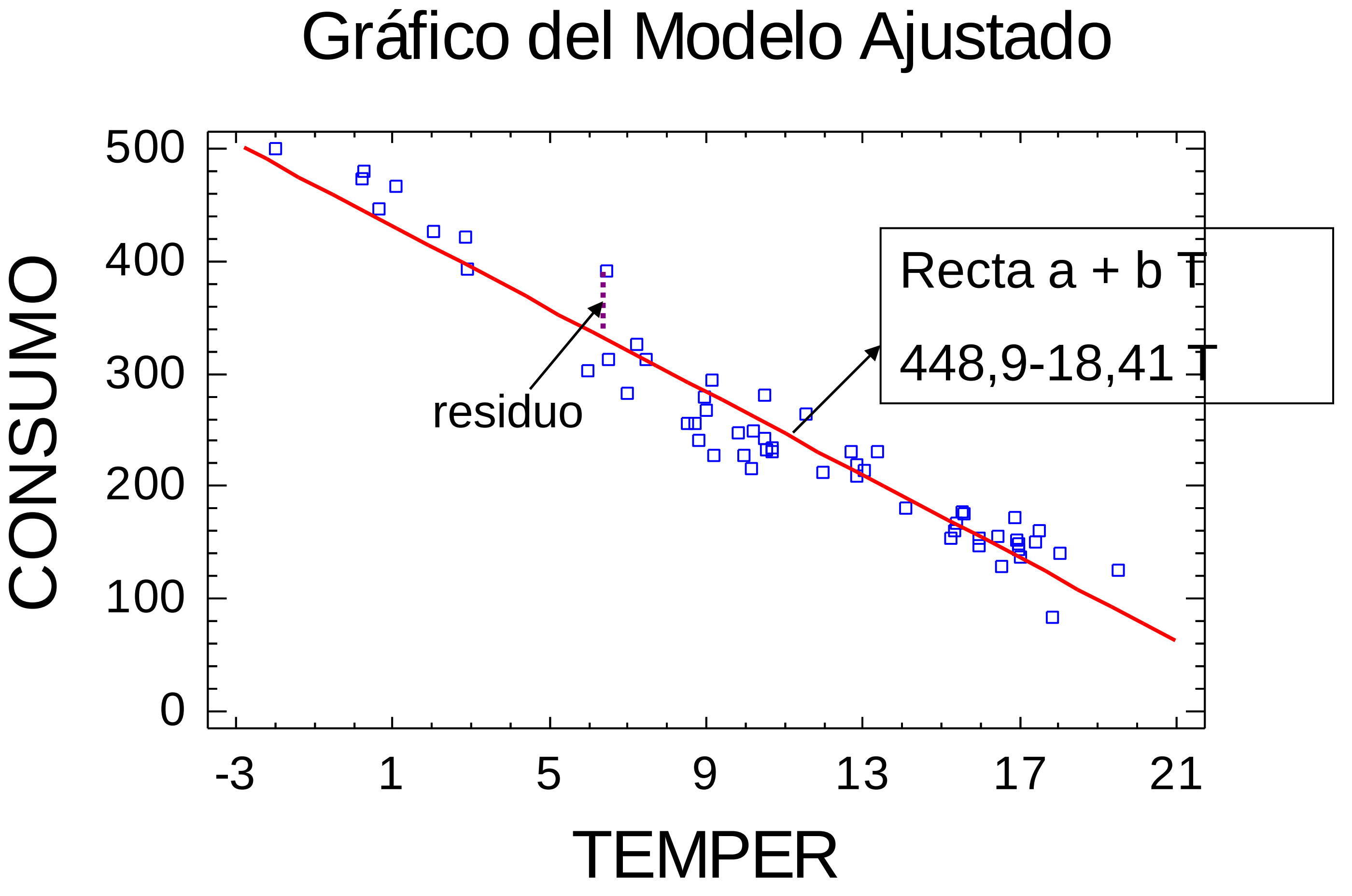
```
Dependent variable: CONSUMO
Independent variable: TEMPER
--------------------------------------------------
                          Standard      T
Parameter    Estimate       Error    Statistic P-Value
--------------------------------------------------
Intercept     448,912     7,63267     58,8145 0,0000<0,05
Slope        -18,4109    0,62713    -29,3567 0,0000<0,05
--------------------------------------------------
Standard Error of Est.= 25,3094
```

a: ordenada

b : pendiente

Desviación típica residual

Recta de regresión y residuos



Recta de regresión y residuos

- La ecuación de la recta de regresión es por tanto

$$E(Y/Temp) = 448{,}9-18{,}41 \times Temp$$

Interpretación de a y b:

- a = 448,9= ordenada en el origen = consumo medio cuando T sea 0ºC.
- b = -18,41= pendiente= lo que disminuye el consumo medio cuando T aumenta 1ºC

Recta de regresión y residuos

- *¿Qué consumo cabe esperar en promedio los días en los que la temperatura es 10ºC?*

E(Consumo/T=10ºC)=448,9-18,41x10= 264,8

- *¿El consumo de 290 termias constatado un día en que la temperatura fue 10ºC, puede considerarse anormalmente elevado e indicador de que algo ha funcionado mal?*

  290 no coincide con la media 264,8 porque hay otros factores que influyen en el consumo (residuo= 290-264,8= 25,2).

Recta de regresión y residuos

- Aplicamos una propiedad de la desviación típica en poblaciones normales*:*
  - El 95% de los datos difieren de la media menos de 2 veces la desviación típica

$$\Downarrow$$

Los días con temperatura 10ºC, el 95% de los consumos diferirán de 264,8 menos de 2 veces la desviación típica residual

$$\Downarrow$$

La desviación típica residual vale 25,3

$$\Downarrow$$

290 difiere de 264,8 menos de 2x25,3=50,6

$$\Downarrow$$

NO es un consumo anormalmente elevado

Recta de regresión y residuos

Ejercicio 5:

- *En un estudio para establecer un sistema para controlar el consumo Y de la energía utilizada en una factoría para climatizar sus instalaciones durante los meses invernales, se han obtenido los siguientes resultados a partir de los valores del consumo diario Y y de la temperatura diaria X constatados en una muestra de muchos días:*

Media de X: 12 grados *Desviación típica de X: 4 grados*
Media de Y: 300 unidades *Desviación típica de Y: 60 unidades*

Coeficiente de correlación entre X e Y: -0,95

- *¿Entre qué límites fluctuará aproximadamente el consumo de energía en el 95% de los días en los que la temperatura sea 5 grados?*

Recta de regresión y residuos

- Aplicamos una propiedad de la desviación típica en poblaciones normales*:*
  - El 95% de los datos difieren de la media menos de 2 veces la desviación típica

$$\Downarrow$$

Los días con temperatura 5 grados, el 95% de los consumos fluctuarán entre los límites

$$E(\text{consumo}/X=5) \pm 2s_{\text{residual}}$$

$$\Downarrow$$

Calculamos E(consumo/X=5) con la recta de regresión

Recta de regresión y residuos

- Obtención de la recta de regresión Y= a + b X:

$$b = r\frac{s_y}{s_x} = -0{,}95\frac{60}{4} = -14{,}25$$

$$a = \bar{y} - b\,\bar{x} = 300\text{-}(\text{-}14{,}25)12 = 471$$

E(consumo/T=5)=471-14,25 x 5 = 399,75

Recta de regresión y residuos

Los días con temperatura 5 grados, el 95% de los consumos fluctuarán entre los límites

$$399{,}75 \pm 2s_{residual}$$

La dispersión alrededor de 399,75 vendrá dada por la desviación típica residual

$$s_{residual} = \sqrt{s^2{}_y\left(1 - r^2{}_{xy}\right)} = \sqrt{60^2\left(1-\left(-0{,}95\right)^2\right)} = 18{,}735$$

Recta de regresión y residuos

- El intervalo

$$E(\text{consumo}/T=5 \text{ grados}) \pm 2s_{\text{residual}}$$

será por tanto

$$399{,}75 \pm 2 \times 18{,}735$$

$$\Downarrow$$

$$[362{,}28 \quad 437{,}22]$$

Coeficiente de determinación R^2. ANOVA del modelo

Variabilidad Total

$$\mathbf{SC_{Total}} = \sum_{j=1}^{N} \left(y_j - \overline{y} \right)^2$$

Con N-1 grados de libertad

Variabilidad explicada

Debida (o asociada) a X

Tiene 1 grado de libertad

Coeficiente de determinación R^2. ANOVA del modelo

Variabilidad residual

$$\mathbf{SC_{Residual}} = \sum_j [y_j - (a + bx_j)]^2$$

Tiene N-2 grados de libertad

La diferencia:

$$\mathbf{SC_{Explicada} = SC_{Total} - SC_{Residual}}$$

Es la parte de variabilidad de Y asociada a la variable explicativa X

Coeficiente de Determinación R^2. ANOVA del modelo

$$R^2 = \frac{SC_{Explicada}}{SC_{Total}} = 1 - \frac{SC_{Residual}}{SC_{Total}}$$

Comprendido entre 0 y 1

Cuanto más cercano a 1 esté, mayor parte de la variabilidad constatada de Y estará asociada a la variable explicativa X del modelo.

Coincide con el coeficiente de correlación lineal al cuadrado r^2.

Coeficiente de determinación R^2. ANOVA del modelo

Ejercicio 7:

La tabla del ANOVA del modelo

E(consumo/temperatura)= 448,9-18,41 x temperatura

es la siguiente:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	552051,0	1	552051,0	861,82	0,0000
Residual	35231,1	55	640,566		
Total (Corr.)	587282,0	56			

Calcula e interpreta el valor del Coeficiente de Determinación del modelo.

Coeficiente de Determinación

$$R^2 = \frac{SC_{Explicada}}{SC_{Total}} = 1 - \frac{SC_{Residual}}{SC_{Total}} = 1 - \frac{35231{,}1}{587282} = 0{,}94$$

R^2 sale muy cercano a 1.

Indica que la recta de regresión que incluye como variable explicativa la temperatura, explica el 94% de la variabilidad observada en el consumo.

ANOVA del modelo

- Para estudiar la hipótesis de si la variable explicativa tiene efecto real poblacional, se utiliza el siguiente resultado:

$$\text{Si } \beta=0 \Rightarrow F_{ratio} = \frac{SC_{Explicada}/1}{SC_{Residual}/N-2} = \frac{CM_{Explicado}}{CM_{Residual}} \sim F_{1,\,(N-2)}$$

Si β difiere de cero, el cociente anterior es mayor que $F_{1,\,(N-2)}$

La hipótesis $\beta=0$ se rechazará si el cociente F_{ratio} supera el valor en tablas $F_{1,\,(N-2)}$ (α) (o el *p-value* es $< \alpha$)

ANOVA del modelo

Ejercicio 7:

La tabla del ANOVA del modelo

E(consumo/temperatura)= 448,9-18,41 x temperatura

es la siguiente:

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	552051,0	1	552051,0	861,82	0,0000
Residual	35231,1	55	640,566		
Total (Corr.)	587282,0	56			

¿Es admisible $\beta=0$?

Respuesta: La hipótesis $\beta=0$ se rechaza porque $F_{ratio}=861,82$ supera el valor en tablas

$861,82>>>F_{1,55}(0,05)\approx 4$ (o el *p-value* es < 0,05).

ANOVA del modelo

- La Tabla del ANOVA de la recta de regresión también proporciona una estimación de la varianza residual a través del cuadrado medio residual.

- Así en el ejemplo de la recta que relaciona el consumo con la temperatura, la desviación típica residual s estimada resulta= $\sqrt{CMResidual} = \sqrt{640.56} = 25{,}3$

- Este valor coincide con el que proporciona el programa Statgraphics en la salida de la opción de regresión en el campo **`Standard Error of Est. = 25,3094`**

Test de hipótesis sobre los parámetros del modelo

Dado el modelo:

$$E(Y_j) = \alpha + \beta X_j$$

La variable X no influye sobre E(Y) $\Leftrightarrow \beta = 0$

El test para contrastar H_0: $\beta = 0$ frente a la alternativa H_1: $\beta \neq 0$, se realiza dividiendo el coeficiente estimado (b) por el margen de incertidumbre asociado a su estimación (s_b).

También se puede aplicar el mismo test para analizar si es admisible la hipótesis nula $\alpha = 0$

Si $\beta=0 \Rightarrow t_{calculada} = \dfrac{b}{s_b}$ se distribuye como una t_{N-2}

Si $\beta \neq 0$ el cociente tiende a ser en valor absoluto mayor que t_{N-2}

Por tanto si $\left|\dfrac{b}{s_b}\right| > t_{N-2}(\alpha)$ se rechaza H_0: $\beta=0$

y se deduce que <u>X influye sobre E(Y) ($\beta \neq 0$)</u>

(donde α es el riesgo de primera especie)

De forma equivalente si *p-value* <α $\Rightarrow$ $\beta \neq 0$

Test de hipótesis sobre los parámetros del modelo

Ejercicio 8: La tabla siguiente da la estimación del modelo E(consumo/Temper)=$\alpha+\beta$ Temper

Multiple Regression Analysis

Dependent variable: CONSUMO

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	448,912	7,63267	58,8145	0,0000
TEMPER	-18,4109	0,627143	-29,3567	0,0000

¿Son significativos α y β? Utiliza un riesgo de primera especie del 5%.

Test de hipótesis sobre los parámetros del modelo

Respuesta:

El *p-value* de α es < 0,05 $\Rightarrow \alpha \neq 0$ (es significativo).

También se llega a la misma conclusión con el valor de la t calculada=448,912/7,633=58,81, que es en valor absoluto mayor que la t de tabla con 55 grados de libertad (2,004)

El *p-value* de β es < 0,05 $\Rightarrow \beta \neq 0$ (es significativo)

La t calculada=-18,4109/0,627=-29,3567, es en valor absoluto >2,004 $\Rightarrow \beta \neq 0$

Validación del modelo. Análisis de residuos

- El modelo se sintetiza en la ecuación

$$Y_t = \alpha + \beta X_t + u_t$$

donde los residuos u_t se suponen $N(0,\sigma^2)$ e independientes

¿Es admisible que u_t se distribuyen normalmente?

¿Hay algún dato claramente anómalo?

¿Es admisible que la varianza de u_t no depende de X?

¿Es realmente sólo lineal la relación entre E(Y) y X?

$$\Downarrow$$

RESPUESTA: ANÁLISIS DE RESIDUOS

- Los residuos se estiman como la diferencia entre el valor observado y_t y el valor medio previsto:

$$e_t = y_t - (a+b\,x_t)$$

- Se analizan gráficamente para contestar las cuestiones planteadas.

- Gráfico de los e_t en papel probabilístico normal permite:

  Estudiar si es admisible la hipótesis de normalidad.

  Detectar posibles observaciones anómalas.

- Gráfico de los e_t frente a los valores de X permite:

  Estudiar si X influye en la varianza de Y.

  Detectar si X tiene además de efecto lineal sobre Y, un efecto de tipo no lineal (por ejemplo cuadrático).

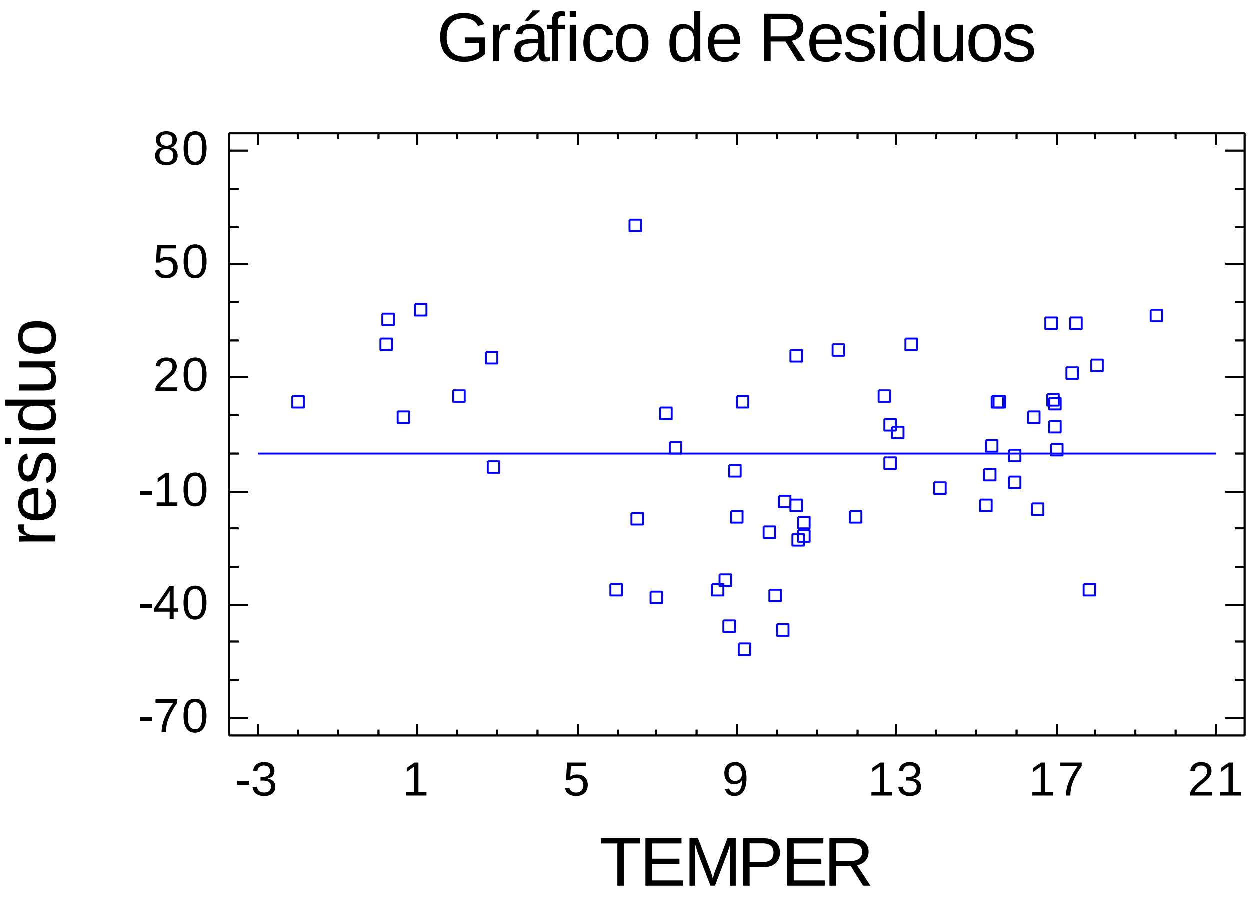
Validación del modelo. Análisis de residuos

Ejemplo:

La siguiente figura representa el diagrama de dispersión de los residuos de la recta de regresión del consumo en función de la temperatura, frente a dicha temperatura.

Se observa que con temperaturas bajas los residuos tienden a ser positivos, con temperaturas intermedias tienden a ser negativos, y con temperaturas elevadas son otra vez positivos.

Validación del modelo. Análisis de residuos



Validación del modelo. Análisis de residuos

- Esto indica que el efecto de la temperatura sobre el consumo no es sólo lineal, sino que tiene además una componente cuadrática.

Efecto lineal y cuadrático

En algunos casos el efecto de X puede ser además de orden 2 ó superior (cuadrático, cúbico, etc)

Un efecto de X de tipo cuadrático, por ejemplo, se incluye en el modelo con un término en el que la variable explicativa es X^2.

Ejemplo:

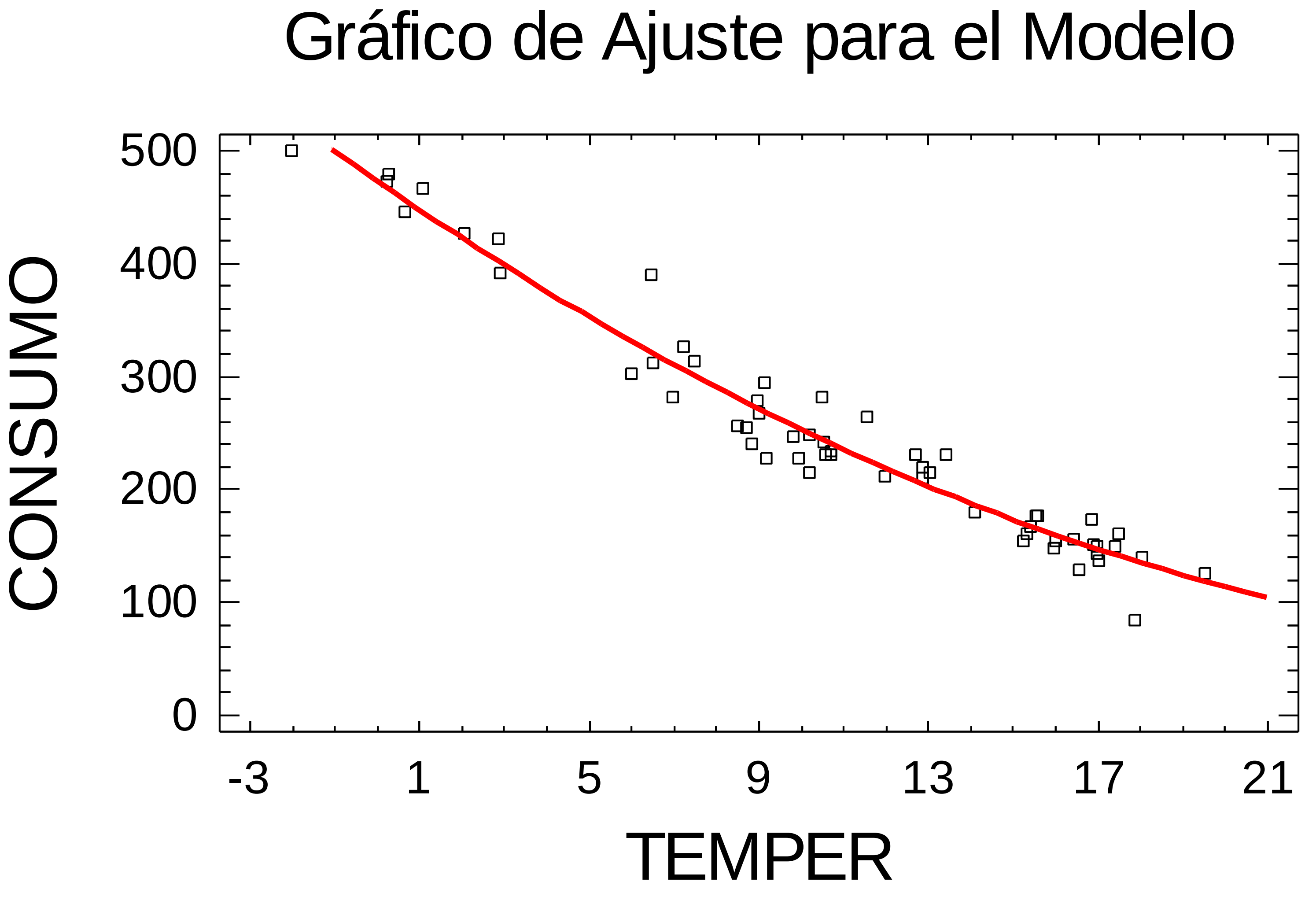


Gráfico de Ajuste para el Modelo
500
400
300
200
100
0
CONSUMO
-3
1
5
9
13
17
21
TEMPER

- El modelo de regresión lineal múltiple con el efecto cuadrático de T queda

$$E(CONSUMO/T=Tt)= \beta_0 + \beta_1 T_t + \beta_2 T_t^2$$

β_0 es el consumo medio cuando la temperatura es 0ºC

β_1 es la pendiente para T=0

β_2 es el efecto cuadrático de T sobre el consumo medio. En este caso $\beta_2>0$.

- Pendiente del modelo:

$$\frac{d(CONSUMO/T)}{dT} = \beta_1 + 2\beta_2 T$$

- Por tanto, la pendiente cuando T=0 será β_1

Estimación del modelo con Statgraphics:

```
Multiple Regression Analysis
-----------------------------------------------------------------
   Dependent variable: CONSUMO
-----------------------------------------------------------------
                            Standard          T
Parameter   Estimate         Error        Statistic        P-Value
-----------------------------------------------------------------
CONSTANT      472,35        8,56846        55,1266         0,0000
TEMPER      -25,9864        1,83412       -14,1683         0,0000
TEMPER^2    0,400966       0,092686        4,32607         0,0001
-----------------------------------------------------------------
```

Son significativos los tres parámetros del modelo (*p-value*<0,05)

La Tabla de ANOVA confirma este resultado

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	561118,0	2	280559,0	579,06	0,0000
Residual	26163,6	54	484,51		
Total	587282,0	56			

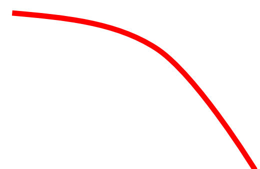
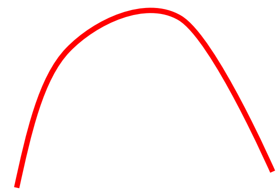
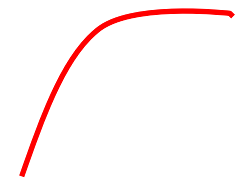
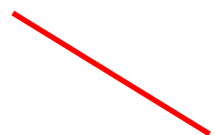
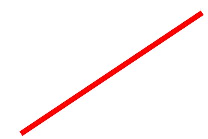
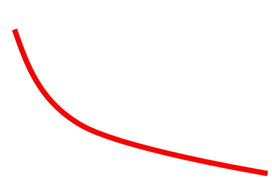
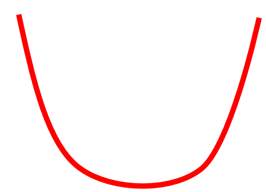
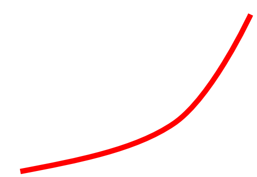
R-squared = 95,545 percent

Standard Error of Est. = 22,0116

El coeficiente de determinación vale ahora 95,5% (es el porcentaje de varianza del consumo que explica el modelo)

La desviación típica residual vale 22,01.

Efecto lineal y cuadrático

Posibles respuestas		Efecto lineal β_1		
		<0	0	>0
Efecto cuadrático β_2	<0			
	0			
	>0			

Ejercicios autoevaluación

1.- Dada la siguiente salida del Statgraphics:

Covariances

	X	Y
X	11.2	15
	(50)	(50)
Y	15	20.5
	(50)	(50)

Covariance

(sample size)

Ejercicios autoevaluación

Indica, justificando la respuesta, cuáles de las siguientes sentencias son verdaderas:

a) No existe relación lineal entre las variables X e Y porque la *covarianza* entre ellas es mucho mayor que 1 ($Cov_{XY}=15$)

b) Existe una fuerte relación lineal positiva entre X e Y porque el *coeficiente de correlación* entre ellas es muy cercano a +1 ($r_{XY}=0.989$)

c) Existe una fuerte relación lineal positiva entre X e Y porque la *covarianza* no es muy elevada ($Cov_{XY}=15$)

d) El resultado $Cov_{XY}=15$ se ha obtenido a partir de una muestra de tamaño 50.

Ejercicios autoevaluación

2.- En un estudio para optimizar un sistema informático, se ha observado durante varios días la carga media X (en consultas por minuto) y el tiempo medio de respuesta Y (en segundos). Se han obtenido los siguientes resultados en la muestra:

Media de X:5,6 Desviación típica de X:3,2

Media de Y:2,9 Desviación típica de Y:1,2

Coeficiente de correlación entre X e Y: 0,96

¿Entre qué límites fluctuará aproximadamente el tiempo medio de respuesta en el 95% de los días en los que la carga media sea de 6 consultas por minuto?