Instituto Valenciano de Investigaciones Agrarias

Seminario

MÉTODOS ESTADÍSTICOS PARA LA INVESTIGACIÓN AGRONÓMICA

Tema 8

MODELOS DE REGRESIÓN (I)

MODELOS DE REGRESIÓN LINEAL (I)

1. Introducción a los modelos de regresión.
2. El Modelo de Regresión Clásico. Ejemplos
3. Fases de un estudio de regresión
4. Modelo de Regresión Lineal Simple
    - 4.1 Hipótesis básicas
    - 4.2 Residuos del modelo
    - 4.3 Interpretación de lo parámetros del modelo
5. Modelo de Regresión Lineal Múltiple
6. Inferencia en modelos de regresión
    - 6.1 Estimación del modelo.
    - 6.2 Coeficiente R2. Análisis de la Varianza
    - 6.3 Test de hipótesis sobre una β
    - 6.4 Validación del modelo. Análisis de residuos
    - 6.5 Análisis de efectos sobre la varianza
    - 6.6 Ejemplo: modelo para el control del consumo de energía
7. Consideración de relaciones no lineales
    - 7.1 Inclusión de un término de 2º grado. Ejemplo
    - 7.2 Uso de transformaciones
    - 7.3 Ejemplo: Pérdida de chlorina en el almacenamiento
    - 7.4 Ejemplo: modelación del efecto de la temperatura la liberación de nutrientes de un abono “slow release”

INTRODUCCIÓN A LOS MODELOS DE REGRESIÓN

- Los Modelos de Regresión constituyen, posiblemente, la herramienta estadística más general y poderosa para analizar la información contenida en unos datos observados o experimentales.
- Los Modelos de Regresión permiten analizar las posibles relaciones entre la pauta de variabilidad de una variable aleatoria **Y** (variable dependiente) y los valores de una o más variables explicativas $\mathbf{X_i}$ (aleatorias o no, cuantitativas o cualitativas) de las que la primera depende, o puede depender.
- Esta relación se estudia mediante el i**establecimiento de un modelo matemático**! (la ecuación de regresión) que, en general, expresa el valor medio de la variable dependiente Y como una función de los valores de las variables explicativas:

$$E(Y_j) = f(X_{1j}...X_{lj})$$

(ecuación de regresión)

La ecuación de regresión

$$E(Y_i) = f(X_{1i} \ldots X_{li})$$

- La ecuación de regresión contiene un conjunto de parámetros, en principio desconocidos, cuyos valores permitirán precisar la naturaleza y cuantificar la magnitud de los efectos de las diferentes variables explicativas sobre la variable dependiente.
- Dichos parámetros se estiman a partir de los datos disponibles, utilizando eficientes procedimientos estadísticos, que proporcionan también los márgenes de incertidumbre asociados y su significación estadística.
- Aunque los cálculos para llevar a cabo la estimación de estos modelos son bastante complejos, el usuario puede despreocuparse de esta fase del análisis, que se lleva a cabo mediante un software adecuado.
- Sin embargo, para utilizar eficientemente estos modelos, interpretar correctamente los resultados que le proporcionan y sacarles el máximo partido en sus estudios, el investigador debe comprender bien la naturaleza del modelo y el sentido de sus diferentes parámetros

TRES TIPOS DE MODELOS DE REGRESIÓN

Tres tipos de modelos de regresión, muy importantes por sus aplicaciones en la investigación en Biología

1. El **Modelo Clásico de Regresión (MRC)**: que asume que la variable dependiente **Y** es una variable continua distribuida normalmente

2. El **Modelo de Regresión Logística (MRL)**: en el que la variable dependiente **Y** es una variable binaria, cuyo valor medio es la probabilidad de que se presente un determinado suceso

3. Los modelos de regresión utilizados en **Análisis de Supervivencia**, en los que la variable dependiente **Y** está frecuentemente censurada porque en muchos individuos no se conoce su valor, sino sólo una cota inferior del mismo:

a. Modelo Log-lineal o de vida acelerada

b. Modelo de Cox de Azar Proporcional

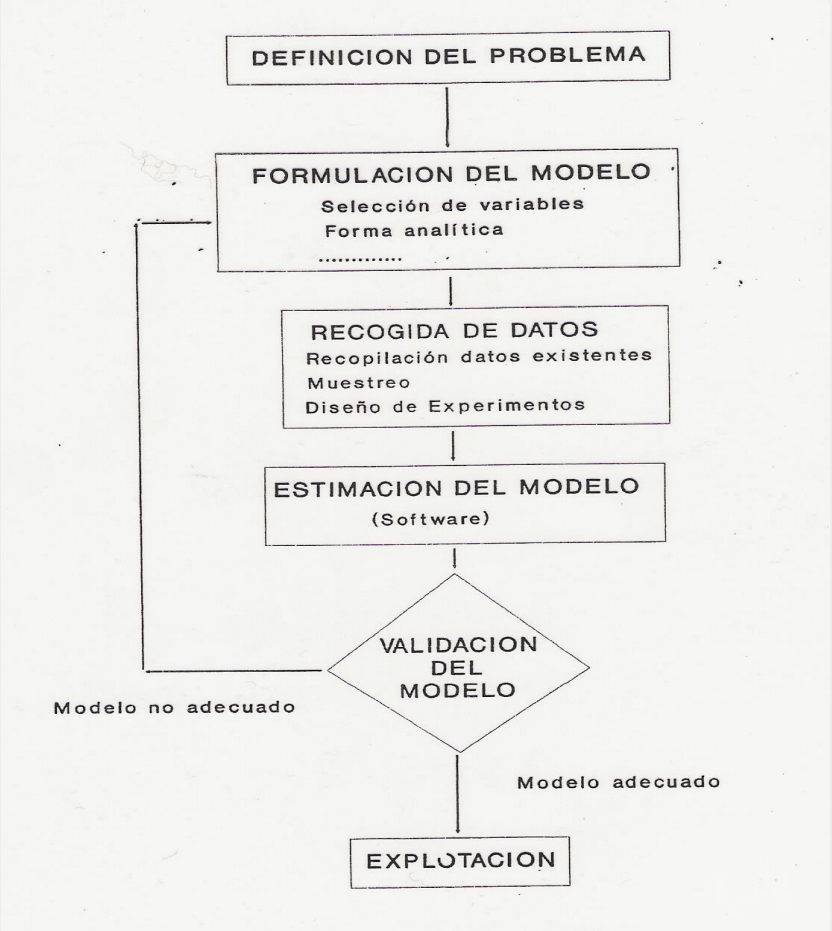
En este curso básico inicial nos centraremos primero en el MRC, pues un buen manejo del mismo es un prerrequisito antes de abordar modelos más avanzados.

EL MODELO CLÁSICO DE REGRESIÓN: EJEMPLOS

- Los modelos clásicos de regresión (MRC) asumen que cada observación y_j es el valor observado de una variable aleatoria $\mathbf{Y_j}$ normal, de varianza $\sigma^2(\mathbf{Y_j})$ constante desconocida, y cuyo valor medio es una función de los valores en dicha observación de las variables explicativas $\mathbf{X_{ij}}$.
  - **Ejemplo 1:** se desea estudiar la relación entre el consumo de energía para calefacción en una instalación industrial y la temperatura diaria y el día de la semana
  - **Ejemplo 2:** se desea estudiar la relación entre el % de nutrientes liberado por un abono del tipo "slow release" y el número de días transcurridos desde el abonado, así como la dependencia de esta relación con la temperatura ambiental.
  - **Ejemplo 3:** se desea estudiar si existe una relación entre la precocidad de un conjunto de variedades de maíz y el número medio de mazorcas por planta
  - **Ejemplo 4:** se desea estudiar la relación entre el descenso en la tensión arterial en personas hipertensas (variable **Y**) y el fármaco suministrado (variable $\mathbf{X_1}$ con 2 alternativas, A y B) y la edad, el sexo y la tensión inicial del paciente (variables $\mathbf{X_2}$ $\mathbf{X_3}$ y $\mathbf{X_4}$)
- En principio el Modelo de Regresión Clásico asume que el efecto de las $\mathbf{X_i}$ sobre la distribución de **Y** consiste en modificar su valor medio.
- Es posible, sin embargo, abordar también mediante este Modelo de Regresión situaciones más complejas, que contemplen adicionalmente la posibilidad de que las varianzas $\sigma^2(\mathbf{Y_j})$ también dependan de las variables explicativas.

FASES DE UN ESTUDIO MEDIANTE MODELOS DE REGRESIÓN

Las fases que deben seguirse en un estudio mediante modelos de regresión, que son en el fondo las mismas para la utilización de cualquier modelo estadístico en el estudio empírico de un fenómeno real, se sintetizan en el esquema siguiente:

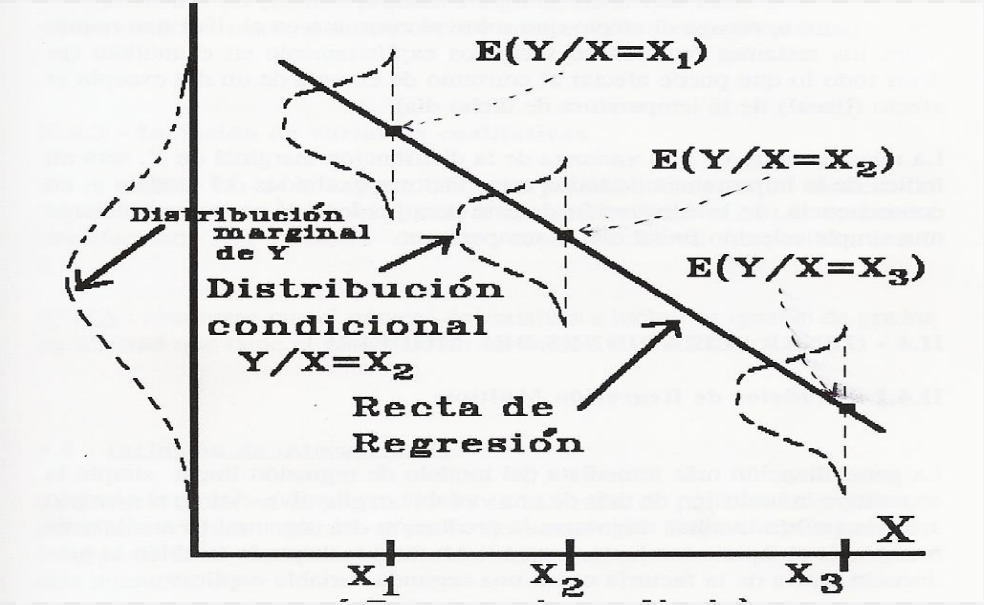


Modelo de Regresión Lineal Simple: Supuestos básicos

El Modelo de Regresión Lineal Simple es el caso más sencillo dentro de los MRC y se presenta cuando sólo hay **una variable explicativa X** y se puede, además, asumir que la **relación entre E(Y) y X es lineal**.

Ejemplo: consumo de energía en calefacción (**Y**), en los días de invierno en función de la temperatura **X** del día

¿Cuál sería la población considerada en este estudio?



Supuestos básicos

Las distribuciones condicionales **(Y/X= x_j)** son normales

Los valores medios **E(Y/X= x_j)** son una función lineal $\beta_0 + \beta_1 x_j$ de x_j

Las varianzas de las distribuciones condicionales **σ^2(Y/X= x_j) son constantes**, no dependiendo de x_j (varianza residual del modelo)

Modelo de Regresión Lineal Simple: residuos del modelo

Se dispone de los valores observados en J días de invierno de x_j (temperatura del día) y de y_j (consumo de energía en calefacción)

Denominando $\mathbf{u_j}$ a la diferencia entre el consumo observado en el día j y el consumo correspondiente en promedio a todos los días en los que la temperatura fuera x_j

$$\mathbf{u_j = y_j - (\beta_0 + \beta_1 x_j)}$$

Se deduce inmediatamente de las hipótesis anteriores que los u_j (a las que se denomina **residuos**) constituyen una variable aleatoria que se distribuye en la población estudiada como una Normal de media cero e idéntica varianza

$$E(u_j) = 0 \qquad \sigma^2(u_j) = \sigma^2 \text{ (\textbf{varianza residual})}$$

Adicionalmente, se asume que las u_j correspondientes a diferentes observaciones son **independientes** entre sí.

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

$$\mathbf{y_j = (\beta_0 + \beta_1 x_j) + u_j}$$

donde la u_j son valores de variables $N(0,\sigma^2)$ independientes, que recogen el efecto sobre el consumo de energía del día de otras variables (lluvia, viento, mal funcionamiento de la instalación de calefacción,…)

Interpretación de los parámetros del modelo

- **Cuando se formula un modelo de regresión es fundamental interpretar el significado práctico de los parámetros que aparecen en el mismo**
- Sea el modelo anterior $y_j = (\beta_0 + \beta_1 x_j) + u_j$ (siendo las $u_j \sim N(0,\sigma^2)$)
- Que también puede formularse como $E(Y/X=x_j) = \beta_0 + \beta_1 x_j$
- β_0: **E(Y/X=0)** consumo medio los días en que la temperatura es 0ºC
- β_1: **E(Y/X=x_j+1)** - **E(Y/X=x_j)** incremento del consumo medio (probablemente será negativo) que cabe esperar por cada grado que aumente la temperatura
- Por otra parte σ^2 será la varianza existente del consumo entre días que tienen la misma temperatura (o sea la varianza debida a otras causas)
- ***¡Cuando se formula un modelo se sabe "lo que son" los parámetros; para saber "lo que valen" esos parámetros hay que disponer de datos que permitan estimarlos!***

Modelo de Regresión Lineal Múltiple

La generalización más inmediata del modelo de regresión lineal simple la constituye la inclusión de más de una variable explicativa. Así, en el ejemplo anterior podría intentar mejorarse la predicción del consumo de energía (o analizar de forma más completa los factores que le afectan) incluyendo también otras variables, como las horas de sol, el día de la semana, etc.

Sean:

y_j: valor de la variable dependiente en la observación j

x_{1j}: valor de la primera variable explicativa X_1 en la observación j

.....

x_{Ij}: valor de la I-ésima variable explicativa X_I en la observación j

El modelo de regresión lineal múltiple postula:

$$E(Y/X1=x_{1j},....,X_I=x_{Ij}) = \beta_0 + \beta_1 x_{1j} +.... + \beta_I x_{Ij}$$

o, de forma equivalente:

$$y_j = \beta_0 + \beta_1 x_{1j} +.... + \beta_I x_{Ij} + u_j$$

donde las u_j son valores independientes de una variable aleatoria $N(0,\sigma^2)$

Interpretación de los parámetros en el Modelo de Regresión Lineal Múltiple

- En el modelo anterior los parámetros tendrían la siguiente interpretación
- β_0: Valor medio de Y cuando todas las X_i son iguales a cero
- Nota: cuando un valor de $x_i = 0$ carece de sentido, β_0 se interpreta mejor si esa x_i se expresa como diferencia respecto a un valor central o de referencia (por ejemplo, edad: años por encima de 30)
- β_i: Incremento en el valor medio de Y cuando X_i aumenta en una unidad, **manteniéndose constantes las restantes variables explicativas**.
- En este modelo lineal se asume que este incremento es constante y no depende de los valores de las restantes variables. Como veremos esta hipótesis se puede relajar planteando modelos no lineales.

Estimación de los modelos de regresión

Postulado un determinado modelo de regresión:

$$\mathbf{y_j = \beta_0 + \beta_1 x_{1j} + \ldots + \beta_I x_{Ij} + u_j}$$

(en el que las x_i pueden ser, como hemos visto, cuadrados o productos de otras variables, variables dummy asociadas a factores cualitativos, etc…)

Dados unos datos

y_1	$x_{11} \ldots x_{i1} \ldots x_{I1}$
……	……
y_j	$x_{1j} \ldots x_{ij} \ldots x_{Ij}$
…	……
y_N	$x_{1N} \ldots x_{iN} \ldots x_{IN}$

constituidos por los valores de la variable dependiente Y y de las I variables explicativas X_i en N observaciones

La estimación es un **proceso de cálculo**, que exige en general el recurso a un software adecuado, que utiliza la información contenida en los datos para obtener:

1. Las **estimaciones $\mathbf{b_i}$** de los I+1 parámetros β_i

2. La **estimación de las desviaciones típicas $\mathbf{s_{bi}}$** de las estimaciones anteriores (que son una medida del margen de incertidumbre asociado a cada estimador)

3. La **estimación $\mathbf{s^2}$** de la varianza residual σ^2 del modelo

Fundamento del proceso de estimación

Conocidas para cada observación j (j=1,...,N)

y_j : valor de la variable dependiente

x_{1j} , ... , x_{ij} , ... , x_{Ij}: valores de las variables explicativas

El residuo e_j que se obtendría para unos posibles valores $b_0,...,b_I$ de los coeficientes se define como :

$$e_j = y_j - (b_0 + b_1x_1 + \ldots + b_Ix_{Ij})$$

Se demuestra que los estimadores b_0 , b_1 , ... , b_I <u>óptimos</u>, desde el punto de vista de sus propiedades estadísticas, son los que conducen a un valor mínimo de la suma de los cuadrados de dichos residuos.

$$\text{minimizar} \quad \sum_{j=1}^{N} e_j^2$$

Si se dispone de un conjunto de datos, la suma de cuadrados residual es una función $H(b_0, \ldots, b_I)$ de los valores que se adopten para los I+1 parámetros desconocidos b_i.

Los valores $b_0, \ldots, b_I$ que minimicen esta suma de cuadrados residual (que se pueden obtener de la forma habitual, igualando a cero las I+1 derivadas de $H(b_0, \ldots, b_I)$ respecto a dichos parámetros), serán los estimadores óptimos de los parámetros desconocidos del modelo $\beta_0, \ldots, \beta_I$

Coeficiente R^2

- La variabilidad total de la variable dependiente Y en el conjunto de las N observaciones viene medida por la Suma de Cuadrados Total:

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

- y tiene N-1 grados de libertad
- Parte de esta variabilidad será debida (o, al menos, estará asociada) a las I variables explicativas $X_1 \ldots X_I$. Esta parte tiene I grado de libertad.
- El resto estará recogido en los residuos e_j, viniendo medida su magnitud por la Suma de Cuadrados Residual

$$SC_{Residual} = \sum_{j} e_j^2$$

- La diferencia:
- $$SC_{Explicada} = SC_{Total} - SC_{Residual}$$
- es la parte de la variabilidad de Y asociada a la variable explicativa
- Se define el Coeficiente de Determinación R^2 como el cociente

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

- que estará comprendido entre 0 y 1. Cuanto más cercano a 1 sea este coeficiente, mayor parte de la variabilidad constatada de Y estará asociada a las variables incluidas en el modelo

Anova del modelo de regresión

- Cuadro Resumen del Análisis de la Varianza del modelo estimado

Origen de la variabilidad	Suma de cuadrados	grados de libertad	Cuadrado Medio	F_{ratio}	P-value
Modelo	SC_{modelo}	I	CM_{modelo}	F_{rat}	p-val
Residual	$SC_{residual}$	N-I-1	$CM_{residual}$		
Total	SC_{total}	N-1			

- Se desea contrastar la hipótesis nula H_0: al menos una de las variables explicativas estudiadas tiene un efecto real poblacional, o sea, al menos una $ß_1$, …, β_I es diferente de cero. Se demuestra que:

$$\text{Si } \beta_1 = \beta_2 = \ldots = \beta_I = 0 \Rightarrow \frac{SC_{Explicada}/I}{SC_{Residual}/(N-1-I)} = \frac{CM_{Explicado}}{CM_{Residual}} = F_{ratio} \square F_{I,N-I-1}$$

- donde $F_{I,N-I-1}$ es una distribución F dse Fisher conI e (N-I-1) grados de libertad.
- Por el contrario, si alguna de las $ß_i$ es diferente de cero, el cociente anterior es, en promedio, mayor que una $F_{I,N-1-I}$.
- La hipótesis de que ninguna de las variables tiene un efecto real sobre E(Y) se rechaza si el cociente supera el valor en tablas $F_{I,N-1-I}(\alpha)$ (o, lo que es equivalente, si el correspondiente p-value es inferior al riesgo de 1ª especie α)

Significación del efecto de un parámetro β_i

- Dado el modelo : $\mathbf{E(Y_j) = \beta_0 + \beta_1 X_{1j} + \ldots + \beta_i X_{ij} + \ldots + \beta_l X_{lj}}$
- la variable X_i no influye sobre E(Y) ←→ $\beta_i = 0$
- Hipótesis nula $\mathbf{H_0: \beta_i = 0}$
- Se demuestra que si $\beta_i = 0$ (o sea, si H_0 es cierta)

$$\frac{b_i}{s_{b_i}} \text{ se distribuye como una } t_{N-1-l}$$

- Mientras que si H_0 es falsa el cociente tiende a ser en valor absoluto mayor que la correspondiente t de Student

Por tanto si $\left|\frac{b_i}{s_{b_i}}\right| > t_{N-1-l}(\alpha)$ se rechaza la hipótesis $\beta_i = 0$ y se deduce que X_i influye sobre E(Y)

Validación del modelo

- Todo el análisis inferencial expuesto se realiza bajo la hipótesis de que el modelo postulado es correcto. Resulta por tanto esencial utilizar adicionalmente la información contenida en los datos para cuestionarse la adecuación de dicho modelo.
- Distintas cuestiones pueden plantearse relativas a la adecuación o no de las hipótesis ante unos datos concretos:
  - ¿Es admisible el que las distribuciones $(Y/x_1,\ldots,x_I)$ son normales?
  - ¿Hay algún dato claramente anómalo?
  - ¿Es admisible que la varianza de las u_j no depende de los valores de las variables explicativas?
  - ¿Es realmente lineal la relación entre $E(Y)$ y las X_i?
- La herramienta más poderosa para analizar estas cuestiones es el **análisis de residuos**.

Análisis de residuos

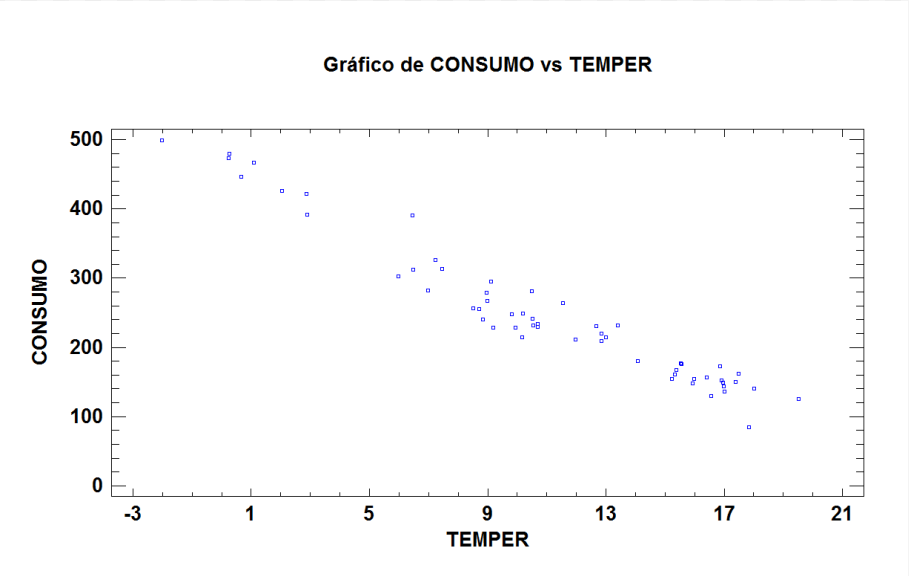
- Tras ajustar un modelo, para cada observación j se dispone de:
  - el valor observado y_j de la variable dependiente
  - el valor medio previsto a partir del modelo estimado para esos valores concretos de las variables explicativas $y_j^* = (b_0+b_1x_{1j}+...+b_Ix_{Ij})$
  - el valor del residuo estimado $e_j = y_j - y_j^*$
- Determinadas representaciones gráficas de los residuos son muy útiles para responder a algunas de las cuestiones que se plantean en la fase de validación de los modelos de regresión.
  - Un gráfico de los residuos en papel probabilístico normal permite estudiar si es admisible la hipótesis de normalidad, así como detectar posibles observaciones anómalas
  - Un gráfico de los residuos frente a los valores previstos para cada observación puede poner de manifiesto la existencia de relaciones no lineales, que se detectan por una configuración curvada con predominio de residuos de un determinado signo para valores extremos de la predicción y de residuos de signo contrario para los valores intermedios
  - En este último gráfico, el hecho de que los valores absolutos de los residuos aumenten al hacerlo las y_j^* puede ser indicativo de que algunas de las variables explicativas afectan a la varianza de las y_j

Análisis de efectos sobre las varianzas

- El MRC asume que el posible efecto de las variables explicativas X_i sobre la distribución de la variable dependiente Y consiste en influir sobre su valor medio E(Y), pero que las varianzas de las distribuciones condicionales $(Y/X_1=x_{1j},\ldots,X_I=x_{Ij})$ no dependen de los valores de las x_{ij}
- Es posible, sin embargo, que en una situación real esta hipótesis no sea cierta, y que las X_i, al margen de su posible efecto sobre E(Y), influyan también sobre las varianzas de las $(Y/X_1=x_{1j},\ldots,X_I=x_{Ij})$ en torno a sus medias. La detección de estos posibles efectos sobre las varianzas puede resultar de gran interés en muchas situaciones.
- Una forma aproximada, pero extremadamente sencilla y poderosa, de estudiar los posibles efectos de las X_i sobre la varianza de Y, consiste en ajustar un nuevo modelo de regresión, usando como variable dependiente los cuadrados de los residuos obtenidos en el ajuste realizado para estudiar sus efectos sobre la media.
- (Este procedimiento es análogo al que ya se expuso en el tema sobre Análisis de la Varianza)

Ejemplo: Relación entre el consumo de energía en calefacción y la temperatura diaria

- En el archivo *gas.sgd* se recoge para 57 días de invierno los valores de:
  - CONSUMO: consumo de energía (expresado en una determinada unidad) en una factoría para climatizar sus instalaciones
  - TEMPER: temperatura media registrada en el día considerado



- Se desea construir un modelo con el fin de controlar el consumo diario de energía, detectando precozmente consumos excesivos que puedan deberse a incidencias (fugas de gas en los depósitos, bajo poder calorífico del gas, …)

Resultados del ajuste del modelo $E(Consumo) = \beta_0 + \beta_1 xTemper$

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	448.913	7.63264	58.8148	0.0000
Pendiente	-18.4109	0.62714	-29.357	0.0000

Análisis de Varianza

Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	552054.	1	552054.	861.83	0.0000
Residuo	35230.8	55	640.56		
Total (Corr.)	587285.	56			

R-cuadrada = 94.0011 porciento
Desv.Típica Residual = 25.3093

Ecuación del modelo ajustado: **E(Consumo) = 448.9 – 18.4xTemper**

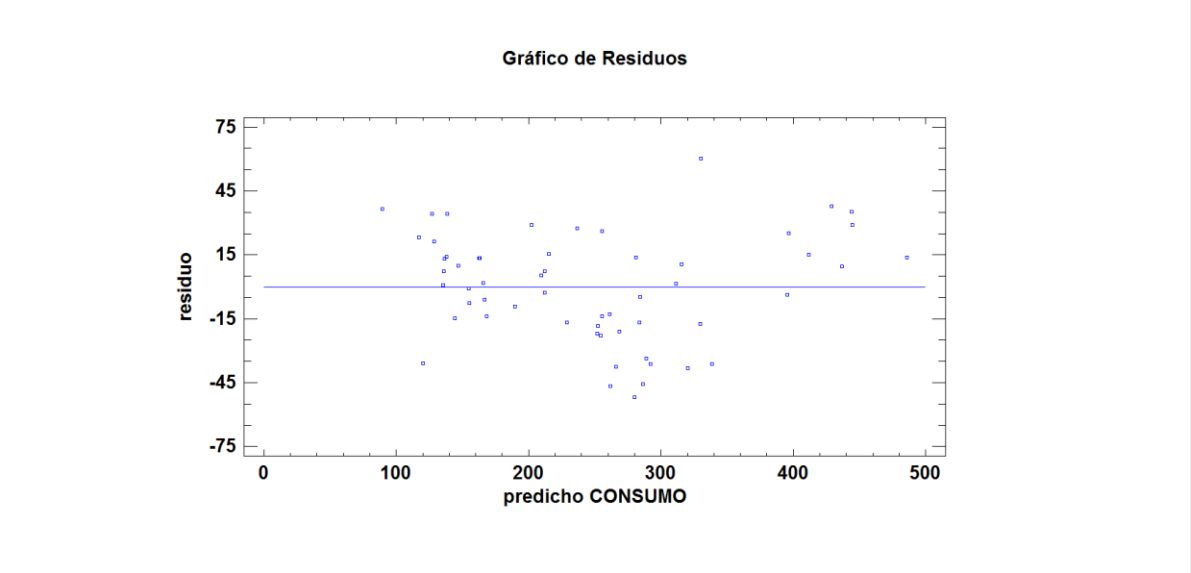
Desviación típica residual = $640.56^{1/2}$ = 25.3

Significación del parámetro β_1: p-value=0.0000

(En el modelo de regresión lineal simple el test t para β_1 y el test F del Anova son equivalentes, dando siempre idénticos p-values)

El modelo explica el 94% de la variabilidad constatada en la variable Consumo

Ejemplo: análisis de los residuos



- El gráfico de los residuos frente a los valores previstos pone de manifiesto una no linealidad en la relación entre el consumo de nergía y la temperatura

Consideración de relaciones no lineales: inclusión de un término de 2º grado

- El modelos expuesto asume que el posible efectos de la variable explicativa sobre el valor medio de Y es de tipo lineal. Una generalización, necesaria en muchos casos, es la de poder considerar en el modelo efectos más generales.

- En muchos casos estas relaciones de tipo más general pueden aproximarse satisfactoriamente a partir de funciones de tipo polinómico, en las que además de la x_i aparezca su cuadrado (o incluso, aunque no suele ser necesario, potencias de orden más elevado).

- Esta generalización no plantea problema alguno, dado que basta con introducir estas potencias como si se trataran de nuevas variables explicativas.

- Así, una relación no lineal puede aproximarse frecuentemente por una curva de 2º grado:

$$\mathbf{y = a + bX + cX^2}$$

- Si definimos una "nueva" variable $X_2 = X_1^2$, el siguiente modelo de regresión "lineal":

  - $\mathbf{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (= \beta_0 + \beta_1 X_1 + \beta_2 X_1^2)}$

- en el fondo asume una posible relación no lineal entre E(Y) y X_1

Interpretación de los parámetros en el modelo cuadrático

- ¿Qué son las β_i en el modelo: $\mathbf{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2}$?
- **β_0**: E(Y) cuando X_1 es igual a cero. Consumo medio de energía los días en los que la temperatura es 0ºC (igual que antes)
- **β_1**: <u>Pendiente **¡en el origen!**</u>: aproximadamente igual al aumento del consumo medio cuando la temperatura X_1 pasa de 0ºC a 1ºC
- **β_2**: Medida de la curvatura (positiva o negativa) de la relación entre E(Y) y X_1. Así un valor positivo de β_2 indicaría que la pendiente de la curva que relaciona E(Y) con X_1 aumenta al hacerlo X_1 (curvatura positiva), mientras que si β_2 fuera negativo dicha pendiente disminuiría al aumentar X_1 (curvatura negativa). Un valor de $\beta_2=0$ implicaría que la relación entre E(Y) y X_1 es lineal.

Modelo cuadrático para el control del consumo de energía: resultados del ajuste

		Error	*Estadístico*	
Parámetro	*Estimado*	*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	0.092685	4.32609	0.0001

El término cuadrático resulta muy significativo estadísticamente (p-val=0.0001)

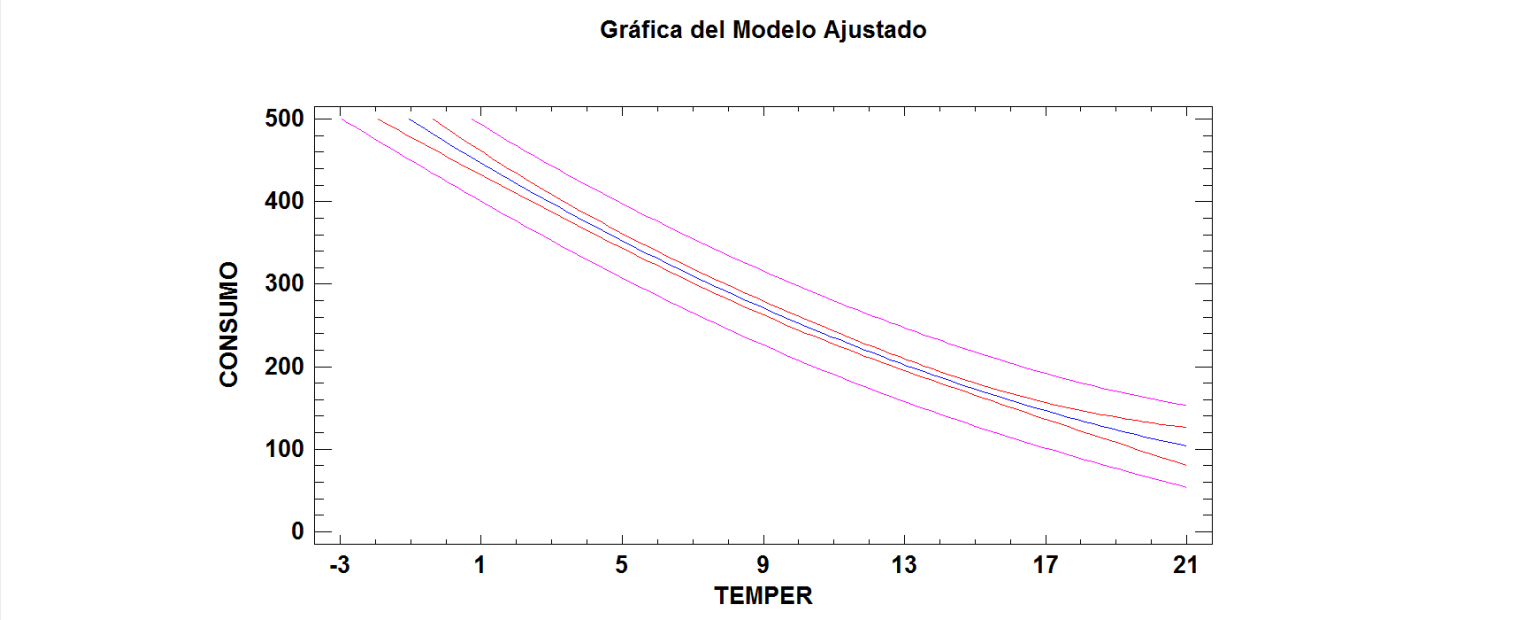
Fuente	*Suma de Cuadrados*	*Gl*	*Cuadrado Medio*	*Razón-F*	*Valor-P*
Modelo	561122.	2	280561.	579.07	0.0000
Residual	26163.3	54	484.505		
Total (Corr.)	587285.	56			

R-cuadrada = 95.5451 porciento
Error estándar del est. = 22.0115

El modelo explica el 95.5% de la variabilidad observada en el consumo

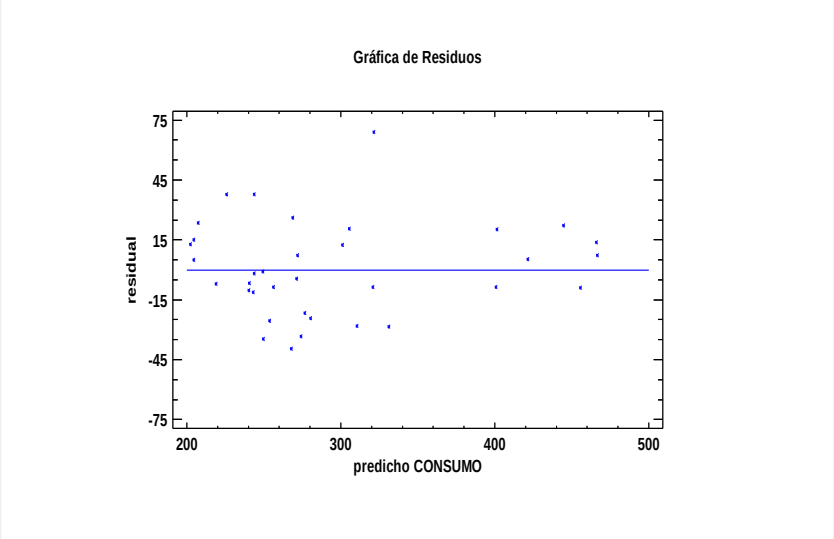
La desviación típica residual es 22

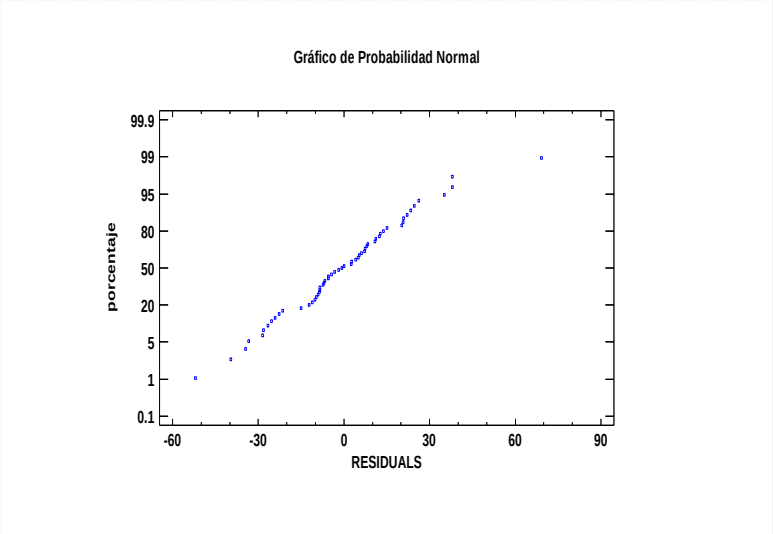
Modelo cuadrático para el control del consumo de energía: intervalos de confianza para las predicciones



- Una vez estimado el modelo es posible obtener, para cualquier valor X, el intervalo de confianza (por ejemplo, para un nivel de confianza del 95%)
  - Para el consumo **medio** previsto los días de temperatura X
  - Para el consumo previsible un día de temperatura X
- Un día en el que la temperatura fue 15ºC se constató un consumo de 300 ¿cabe sospechar que se ha producido alguna incidencia especial?

Modelo cuadrático para el control del consumo de energía: validación del modelo





Los gráficos de residuos no detectan ninguna anomalía relevante, aunque parece existir un día en el que el consumo fue sensiblemente más elevado que lo previsible para la temperatura correspondiente.

(Se constata analizando los datos que se trata de un día que, aunque la temperatura no fue excesivamente baja, estuvo precedido de 3 días muy fríos)

¿Influye la temperatura en la varianza del consumo diario?

Variable dependiente: RESIDUALS^2
Variable independiente: TEMPER (temperatura diaria)

		Error	*Estadístico*	
Parámetro	*Estimado*	*Estándar*	*T*	*Valor-P*
CONSTANTE	624.364	236.378	2.64139	0.0107
TEMPER	-15.1237	19.4221	-0.778683	0.4395

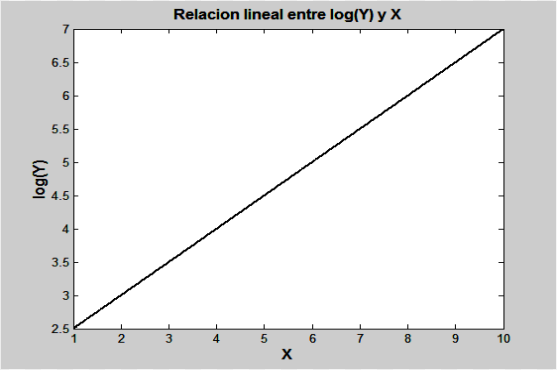
- El ajuste del cuadrado de los residuos a TEMPER no resulta significativo (p-val = 0.4396)
- Puede aceptarse que la temperatura diaria no influye sobre la varianza del consumo diario de energía (aunque, como se ha visto, sí que influye sobre su media)

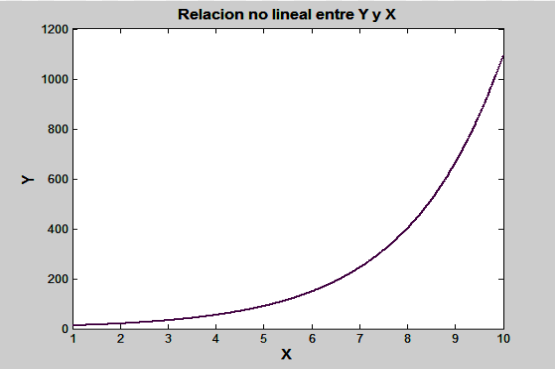
Uso de transformaciones

En ciertas ocasiones es posible modelizar una relación no lineal entre las variables primitivas X e Y, postulando una relación lineal entre las variables X' e Y' que resultan de aplicar a las primeras una determinada transformación. La transformación logarítmica es la más utilizada al respecto.

Transformación logarítmica de la variable respuesta

Una relación lineal entre Y' = log(Y) y X implica, como se ve en las siguientes figuras, una relación curvilínea entre Y y X





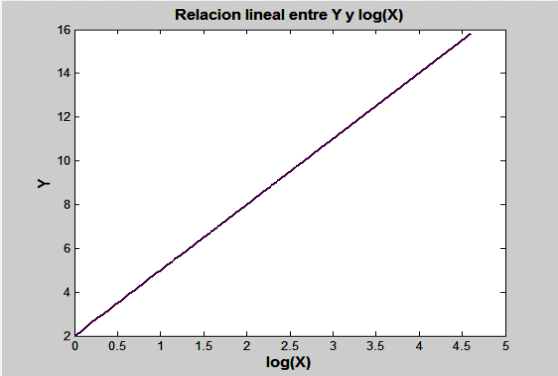
Adicionalmente, en el caso de que las distribuciones condicionales (Y/X=x) sean muy asimétricas positivas, la anterior transformación permite "normalizar" la variable dependiente

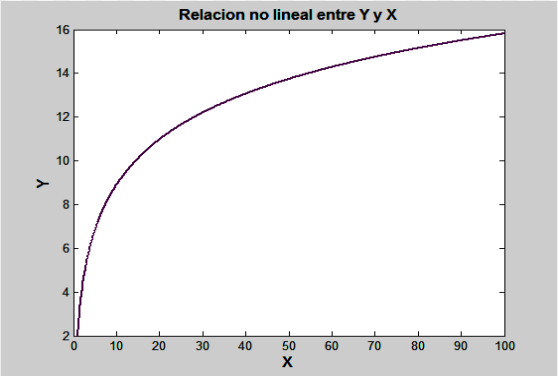
Transformaciones (continuación)

Transformación logarítmica de la variable explicativa

El modelo parabólico ($y = a + bx + cx^2$) puede presentar a veces el problema de que la curva ajustada presente en la zona estudiada máximos o mínimos carentes de sentido desde un punto de vista técnico

Una relación lineal entre Y y el logaritmo de X, $E(Y) = a + b\log(X)$ implica, como se ve en las siguientes figuras, una relación curvilínea entre Y y X, que es monótonamente creciente ($b>0$) o monótonamente decreciente ($b<0$), no presentando nunca máximos ni mínimos

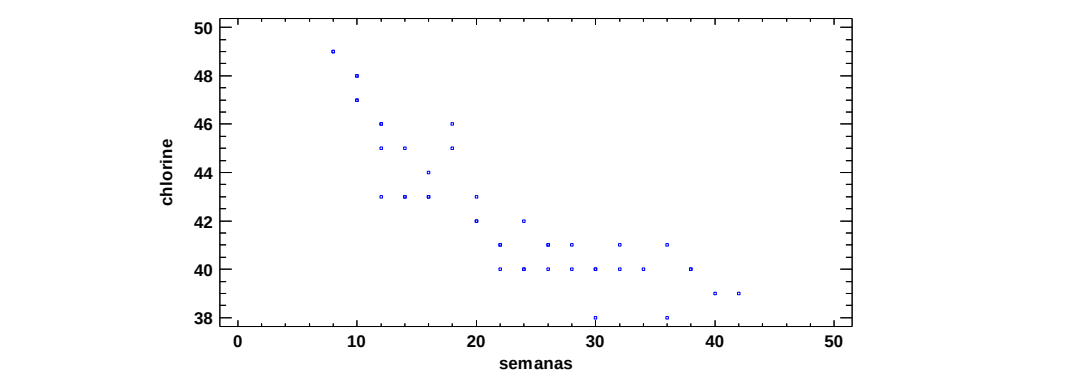




Otras transformaciones: en ocasiones se utilizan otras transformaciones de la variable explicativa como la raíz cuadrada de X o la inversa 1/X

Ejemplo de un modelo no lineal: Pérdida de chlorina en el almacenamiento

El archivo *chlorina* contiene datos de chlorina (un desinfectante) disponible en muestras de un producto como una función del numero de semanas desde que fue producido. Los datos, de Draper y Smith (1998), consisten en n = 44 muestras y se grafican a continuación



Variable dependiente Y: contenido de chlorina (expresado en %)

Variable explicativa X: semanas transcurridas desde la fabricación

Modelo propuesto;

$$\mathbf{E(Y) = \beta_0 + \beta_1 X + \beta_2 X^2}$$

o, de forma equivalente: $\mathbf{y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + u_j}$ donde $\mathbf{u_j} \sim N(0,\sigma^2)$

Interpretación de los parámetros del modelo

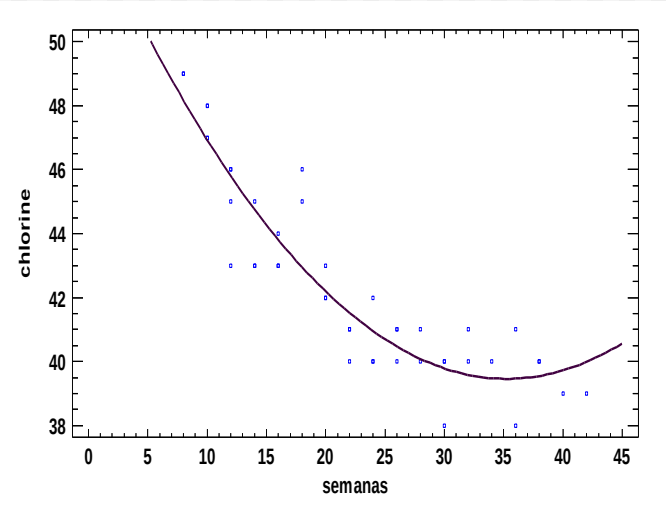
- β_0: contenido medio en chlorina de las muestras recién fabricadas
- β_1 : disminución (pues se espera que resulte negativo) del % medio de chlorina en las muestras tras la 1ª semana de almacenamiento
- β_2 : curvatura de la relación entre E(Y) y X (se espera que resulte positiva porque la pérdida de chlorina es más marcada en las primeras semana)
- σ^2: varianza residual, mide la variabilidad en los % de chlorina entre muestras que llevan el mismo tiempo almacenadas

Resultados de la estimación del modelo

- La estimación del modelo da los siguientes resultados:

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	**54.0367**	1.02386	52.7775	0.0000
semanas	**-0.826602**	0.095363	-8.66795	0.0000
semanas^2	**0.0117085**	0.0019757	5.92628	0.0000

- Los 3 parámetros estimados (b_0=54.04, b_1=-0.8267 y b_2=0.0118) resultan muy significativos estadísticamente siendo sus signos acordes con lo que cabría esperar
- En el gráfico se observa el ajuste del modelo estimado a los puntos experimentales.



Un modelo alternativo

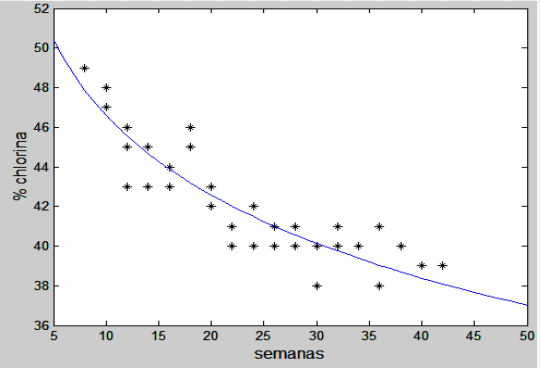
- Aunque el modelo expuesto se ajusta bien a los datos, presenta un problema si se desea extrapolar sus predicciones más allá del periodo estudiado. En efecto, el modelo y = 54.04 - 0.8267x + 0.0118x2 presenta un mínimo para x = 35, por lo que, obviamente, no puede utilizarse para predecir más allá de ese periodo
- Un modelo alternativo sencillo, que se ajusta bien a los datos y no presenta este inconveniente, puede obtenerse utilizando como variable explicativa el logaritmo de X+1 (que vale 0 cuando X=0)

$$E(Y) = \beta_0 + \beta_1 \ln(X+1)$$

- Los resultados del ajuste de este nuevo modelo se recogen a continuación

		Error	*Estadístico*	
Parámetro	*Estimación*	*Estándar*	*T*	*Valor-P*
CONSTANTE	61.6107	1.28049	48.1149	0.0000
LOG(1+weeks)	-6.25409	0.414768	-15.0785	0.0000

- Se aprecia en la figura que el modelo se ajusta también bien a los datos, permitiendo su extrapolación a valores más altos de X.



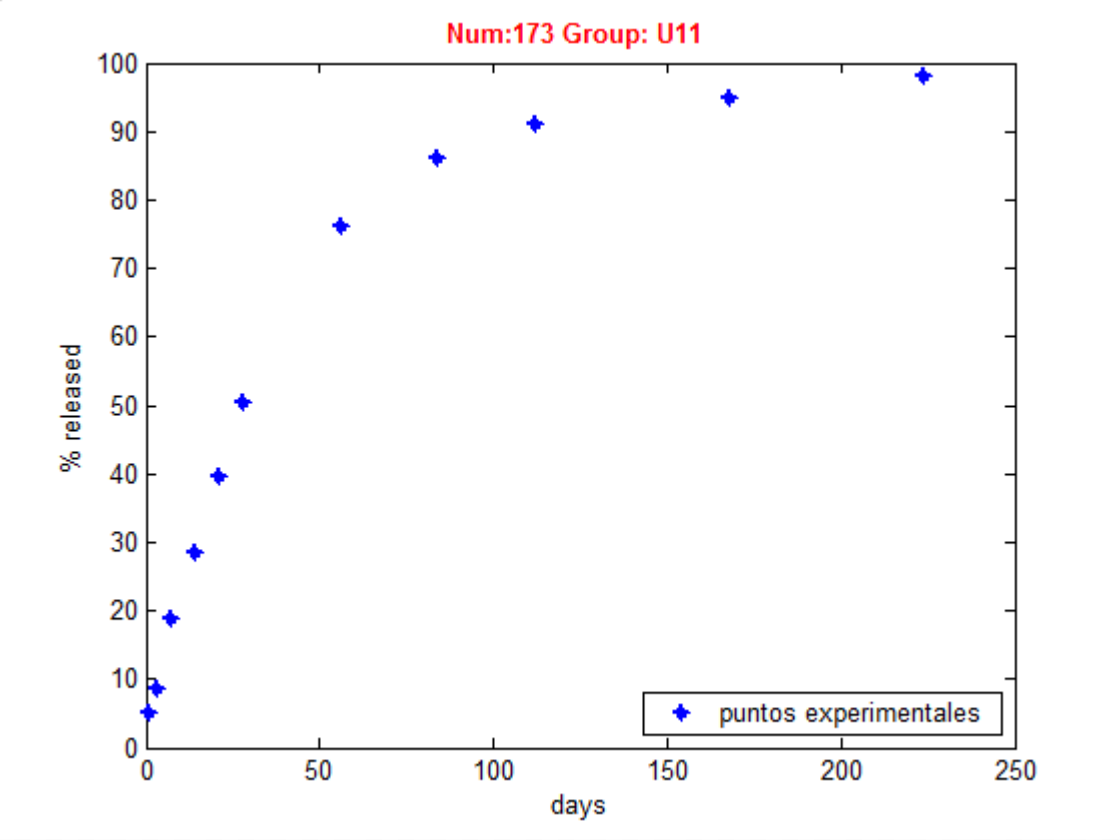
Ejemplo: efecto de la temperatura en la liberación de urea de un abono “slow release”

- El archivo Excel ensayos_*U.xls* contiene los resultados de 28 ensayos realizados por una empresa con el objetivo de modelizar la liberación de urea de unos abonos “slow release” y el efecto de la temperatura del terreno sobre la velocidad de liberación

id	id Producto	Temper(ºC)	1	3	7	14	21	28	56	84	112	168	224	280	336	392	448	504	560
74	U02	21.00	0.1617189	0.1900000	0.2234014	0.2807345	0.3179022	0.3649548	0.5227197	0.7334683	0.7457257	0.8200000	0.8700000	0.9100000	0.9500000	0.9800000			
75	U02	21.00	0.2135164	0.2700000	0.3005045	0.3325320	0.4005409	0.4500000	0.6000000	0.7200000	0.8000000	0.8800000	0.9400000	0.9800000					
76	U02	21.00	0.1399718	0.1700000	0.1874199	0.2676863	0.3309504	0.3914467	0.5622598	0.7160707	0.7935692	0.8400000	0.8900000	0.9300000	0.9700000	0.9900000			
77	U02	21.00	0.1882107	0.2100000	0.2344726	0.3088079	0.3669318	0.3736537	0.5476300	0.6717858	0.7492843	0.8200000	0.8700000	0.9100000	0.9500000	0.9800000			
78	U02	21.00	0.1486707	0.1800000	0.2111440	0.2973413	0.3495342	0.3886789	0.5563288	0.6994638	0.7378177	0.8200000	0.8700000	0.8100000	0.9500000	0.9800000			
71	U02	35.00	0.1447167	0.2000000	0.2522656	0.3242286	0.4104259	0.4558970	0.6025907	0.7757762	0.8335047	0.9200000	0.9700000	0.9900000					
72	U02	35.00	0.1075890	0.1600000	0.2016544	0.2787575	0.3855157	0.4100305	0.5859838	0.7718222	0.8271783	0.9200000	0.9700000	0.9900000					
73	U02	35.00	0.1024088	0.1500000	0.1917693	0.2550335	0.3542790	0.3756307	0.5385158	0.7560061	0.7821026	0.9000000	0.9500000	0.9800000					
174	U11	21.00	0.0234429	0.0600000	0.0921547	0.1103432	0.1786508	0.2396831	0.4094418	0.5500000	0.6600000	0.7900000	0.8500000	0.9600000	0.9400000	0.9700000			
175	U11	21.00	0.0299099	0.0600000	0.0776040	0.1289358	0.1754173	0.2364496	0.3924659	0.5400000	0.6500000	0.7800000	0.8400000	0.8900000	0.9300000	0.9600000	0.9900000		
176	U11	21.00	0.0400146	0.0600000	0.0780082	0.1277232	0.1515703	0.2188654	0.4005497	0.5500000	0.6600000	0.7900000	0.8500000	0.9600000	0.9400000	0.9700000	1.0000000		
171	U11	35.00	0.0400000	0.0670951	0.1700000	0.2602967	0.3435593	0.4898751	0.7214745	0.8200000	0.8700000	0.9100000	0.9400000	0.9700000	0.9900000				
172	U11	35.00	0.0300000	0.0638616	0.1700000	0.2728265	0.3722566	0.4947254	0.7530011	0.8500000	0.9000000	0.9400000	0.9700000	0.9900000					
173	U11	35.00	0.0500000	0.0873045	0.1900000	0.2845479	0.3961036	0.5032133	0.7602765	0.8600000	0.9100000	0.9500000	0.9800000	1.0000000					
4	U01	21.00	0.0654784	0.1000000	0.1193161	0.1487814	0.1618770	0.1807930	0.2346308	0.3441251	0.5154602	0.5800000	0.6151328	0.6722445	0.6726082	0.7200000	0.7500000	0.8000000	0.9000000
1	U01	35.00	0.0541519	0.0700000	0.1508813	0.1963524	0.2571183	0.3058964	0.5394524	0.7134826	0.7900000	0.8465888	0.8317074	0.8887529	0.9400000	0.9800000			
2	U01	35.00	0.1240120	0.1500000	0.2017262	0.2496776	0.2988690	0.3517808	0.5729156	0.7403519	0.8200000	0.8821390	0.8379080	0.8999140	0.9400000	0.9800000			
3	U01	35.00	0.1409603	0.1600000	0.1860181	0.2298356	0.2864678	0.3364860	0.5890572	0.7234036	0.8200000	0.8928867	0.8317074	0.9000000	0.9400000	0.9800000			
101	U04	21.00	0.0731175	0.1100000	0.1284103	0.1695162	0.1900000	0.2095307	0.2700000	0.3200000	0.3659513	0.4608949	0.5067297	0.6515096	0.7000000	0.7300000	0.7600000	0.7900000	0.9000000
102	U04	21.00	0.0803929	0.1300000	0.1662423	0.2138960	0.2500000	0.2750091	0.3200000	0.3500000	0.3790469	0.4629862	0.5518370	0.6656966	0.7200000	0.7500000	0.7800000	0.8100000	0.9000000
108	U05	21.00	0.0480175	0.1200000	0.1678973	0.2346308	0.3113860	0.3433976	0.4754456	0.5500000	0.6000000	0.6773372	0.6911604	0.6838850	0.7200000	0.7500000	0.7800000	0.8000000	0.9000000
109	U05	21.00	0.0683885	0.1300000	0.1880684	0.2477264	0.3055657	0.3452164	0.4696253	0.5400000	0.6000000	0.6817024	0.6984358	0.7009822	0.7400000	0.7800000	0.8100000	0.8300000	0.9000000
98	U04	35.00	0.1455075	0.1800000	0.2199934	0.3195967	0.4053422	0.4204992	0.5807307	0.8037555	0.8440299	0.9100000	0.9400000	0.9700000	0.9800000				
99	U04	35.00	0.1368463	0.1600000	0.2035372	0.2918810	0.3464463	0.3732959	0.5417554	0.7400960	0.8590927	0.9200000	0.9500000	0.9700000	0.9800000				
100	U04	35.00	0.1086975	0.1600000	0.2013719	0.2637321	0.3434149	0.3620364	0.5378579	0.7617489	0.8206447	0.9000000	0.9400000	0.9700000	0.9800000				
104	U05	35.00	0.1225554	0.2000000	0.2775901	0.3581389	0.4664034	0.4858910	0.6326976	0.8059208	0.9401687	0.9700000	0.9900000	1.0000000					
105	U05	35.00	0.1182248	0.1800000	0.2394810	0.3429819	0.4101058	0.4690017	0.6387604	0.8271406	0.9072563	0.9500000	0.9800000	0.9900000					
106	U05	35.00	0.1147603	0.2400000	0.3421157	0.3797918	0.4486480	0.4741984	0.6041158	0.7738745	0.8882018	0.9500000	0.9800000	0.9900000					

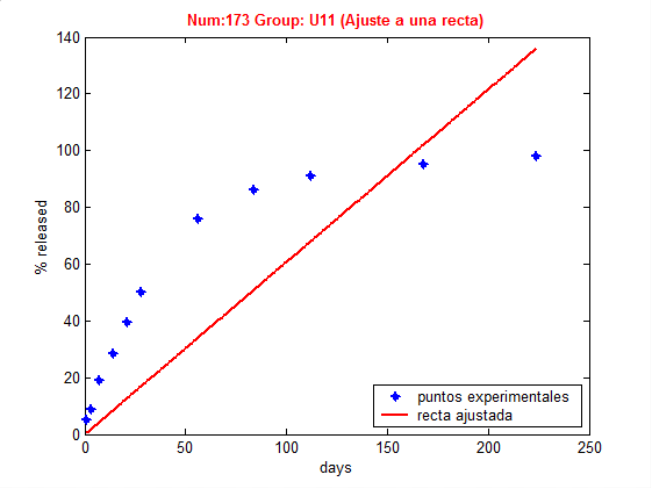
- En cada ensayo se determinó el porcentaje de urea liberado desde el inicio hasta que se consideró liberado el 100%
- En la diapositiva siguiente se recogen los resultados de un experimento concreto

Experimento Nº 173



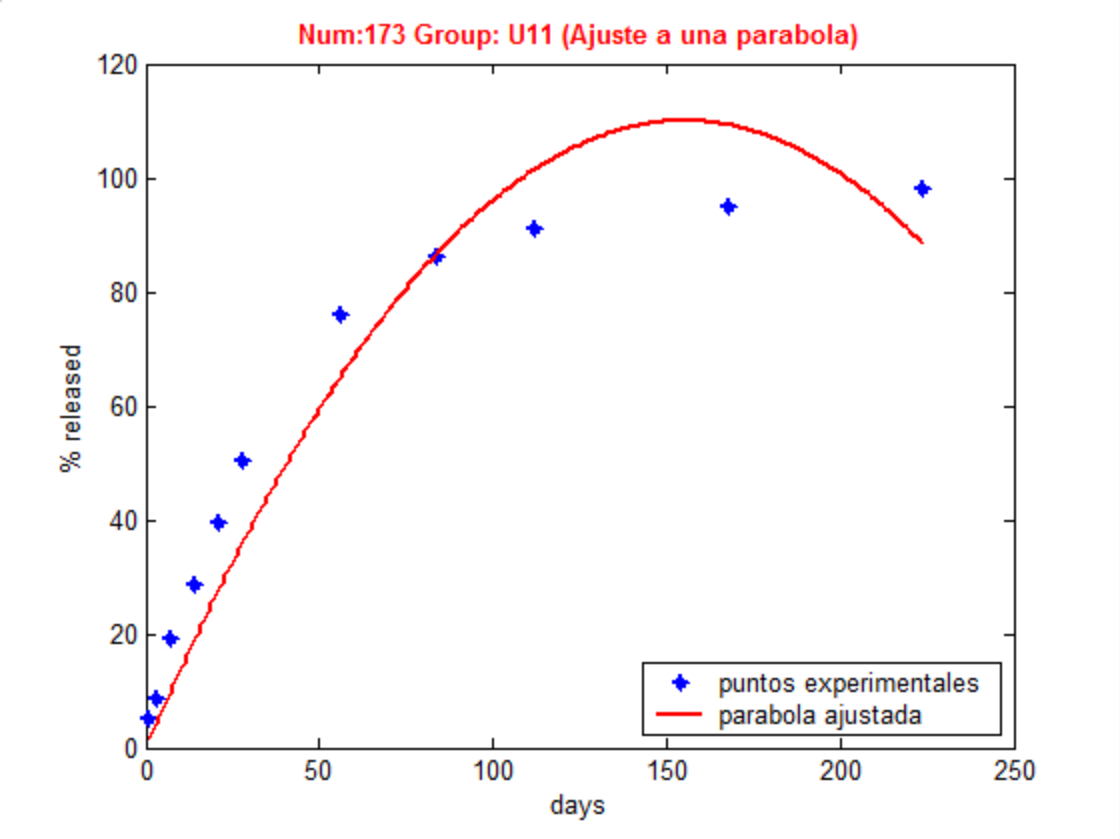
Modelación de la relación entre el %liberado y los días transcurridos

- Siendo **y**: % liberado y t: días transcurridos, el modelo **y**=f(**t**) debería cumplir las siguientes condiciones:
  - Pasar por el origen, o sea **y**=0 para **t**=0
  - Ser una función creciente en todo el rango de valores de **t**
  - No sobrepasar el valor **y**=100 y tender hacia ese valor a medida que aumenta **t**
  - Nota: como la variación del % liberado se mide mal para valores superiores al 95%, se considera suficiente que el modelo **y**=f(**t**) se comporte ben hasta este valor de **y**
- En la siguiente figura se constata que el modelo lineal **y** = β**t** ajustado a los datos no es adecuado



Modelación de la relación entre el %liberado y el tiempo: modelo cuadrático

- El ajuste de los datos a un modelo cuadrático tampoco proporciona buenos resultados, por presentar la función ajustada un máximo (superior a 100) en la zona estudiada y empezar a decrecer desde ese momento



Modelación de la relación entre el %liberado y el tiempo: modelo exponencial

- Este modelo se obtiene asumiendo que la cantidad liberada en cada instante de tiempo (dx/dt) es proporcional a la proporción de fertilizante que queda por liberar

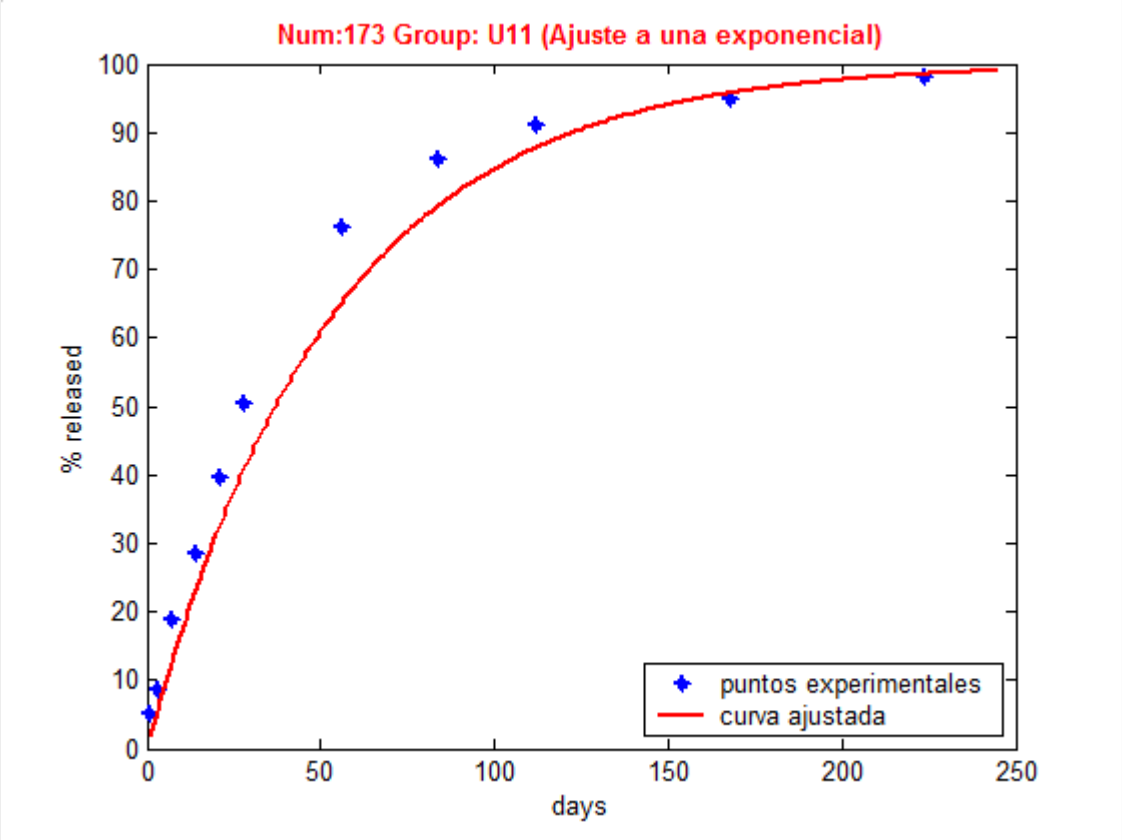
$$\frac{dx}{dt} = \alpha(100 - x)$$

- Integrando esta ecuación diferencial y poniendo la condición inicial x(0) = 0 se obtiene $x = 100(1-e^{-\alpha t})$

- La estimación de α a partir de los datos puede obtenerse ajustando por mínimos cuadrados un modelo lineal (sin constante) tomando como variable independiente a t y como variable dependiente log(1-x/100).

- Una ventaja importante del modelo exponencial es que cumple los tres requisitos exigidos, siempre que α resulte positivo (lo cual sucede siempre al ajustar resultados experimentales reales).

Ajuste de los datos al modelo exponencial



- Sin embargo, como se aprecia en la figura, el modelo exponencial no representa bien el fenómeno de la liberación del fertilizante en las fases iniciales, siendo ésta más rápida que lo que predice el modelo

Modelo exponencial cuadrático

- Este modelo se obtiene asumiendo que en la ecuación

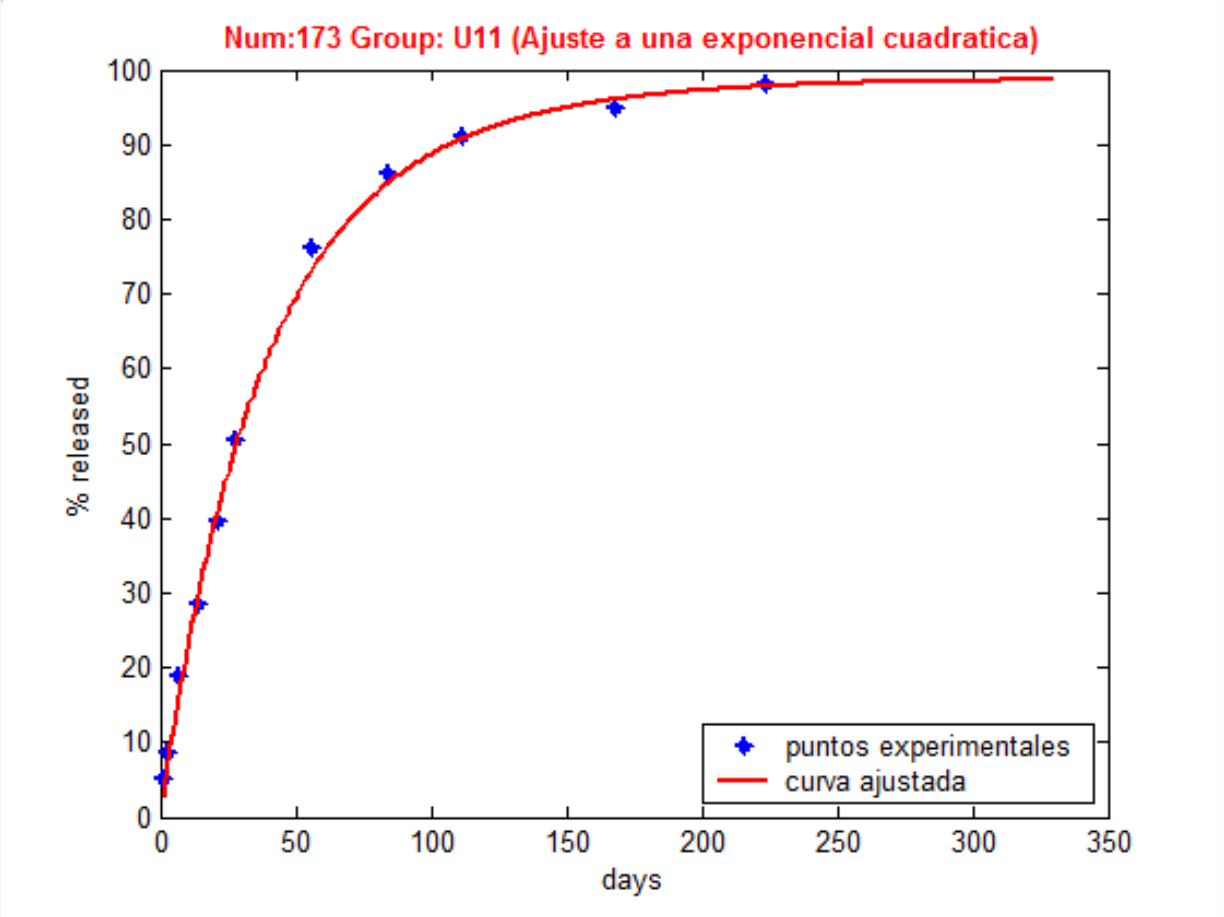
$$\frac{dx}{dt} = \gamma(100 - x)$$

- el parámetro de proporcionalidad γ es una función lineal del tiempo, siendo más elevado para valores bajos de t.
- La ecuación de este modelo es:

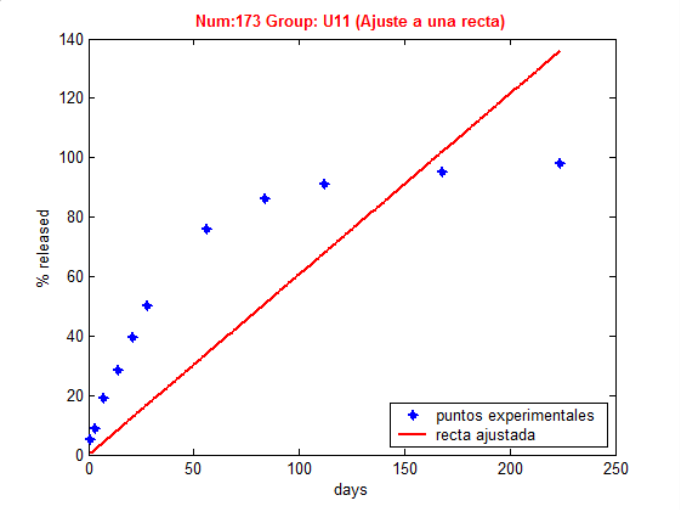
$$x = 100 \times \left(1 - e^{\alpha t + \beta t^2}\right)$$

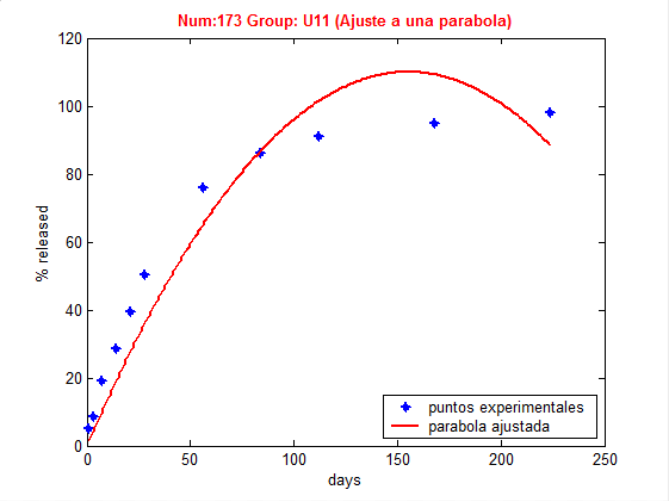
- La estimación de α y β a partir de los datos puede obtenerse ajustando un modelo de 2º grado (sin constante) tomando como variable independiente a t y como variable dependiente el log(1-x/100).
- En la siguiente diapositiva se aprecia que la curva estimada se ajusta muy bien a los datos experimentales

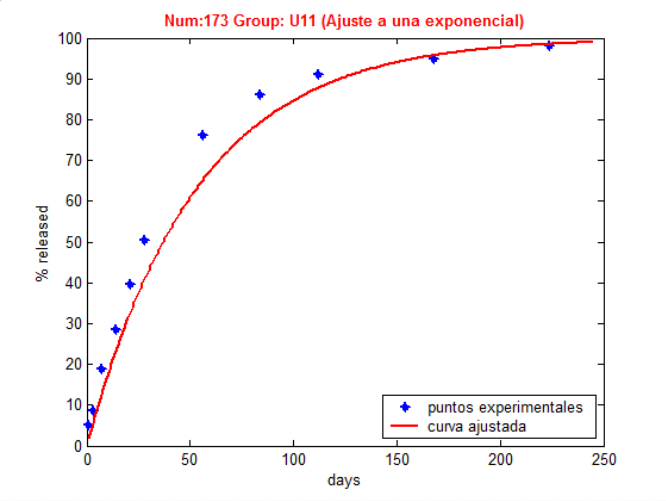
Ajuste de los datos al modelo exponencial cuadrático

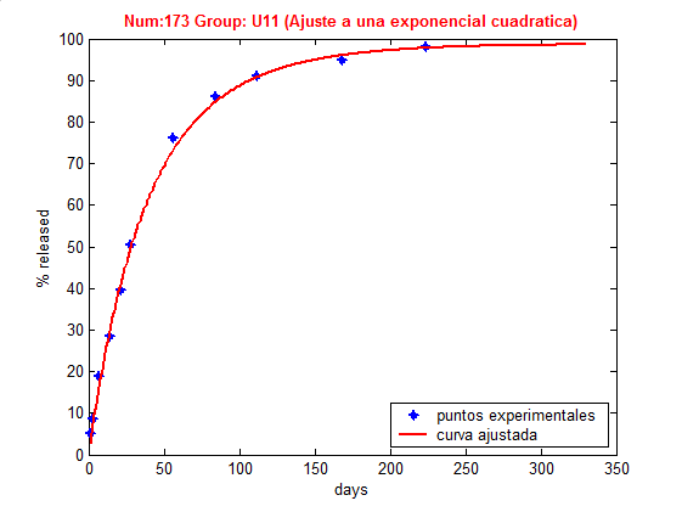


Comparación de los cuatro modelos









¿Tiene sentido estudiar la significación de los parámetros en los modelos anteriores?

- **¡NO!**
- Las observaciones sucesivas dentro de cada ensayo **¡no son independientes!**
- Realmente cada ensayo es un único individuo de la población de todos los ensayos posibles en unas determinadas condiciones (p.e. de temperatura).
- En terminología estadística estas situaciones se denominan de "**medidas repetidas"**
- La forma más sencilla de analizar este tipo de datos es asociando a cada ensayo uno (o más) resultados que sinteticen lo observado en el ensayo y analizando luego por separado cada uno de estos resultados.
- En este ejemplo, en concreto, a cada ensayo se le pueden asociar como "resultados" los valores obtenidos para los parámetros α y β que sintetizan la evolución de la liberación de urea constatada en el ensayo
- Luego es posible estudiar si estos parámetros α y β están relacionados significativamente con algunos de los factores estudiados en la investigación (en este caso, con la temperatura del suelo)

Efecto de la temperatura del suelo en la velocidad de liberación de urea

- Once de los 26 ensayos se realizaron a 21 ºC y los quince restantes a 35ºC (la empresa quería simular condiciones de países tropicales)
- En el archivo *holanda.sgd* se recogen para los 26 ensayos la temperatura del mismo y los valores de α y de β obtenidos al ajustar en cada ensayo la curva de liberación de urea.
- Una forma sencilla de estudiar el posible efecto de la temperatura sobre la velocidad de liberación es realizar sendos ajustes utilizando la temperatura como variable explicativa y α o β como variables dependientes
- Dado que sólo se han estudiado dos temperaturas sólo es posible ajustar modelos lineales.
- En la siguiente diapositiva se recoge los resultados de ambos ajustes, constatándose que en ambos casos el efecto de la temperatura es estadísticamente significativo
- Para modelar el efecto de la temperatura T del suelo en la velocidad de liberación basta sustituir α y β por las funciones de T obtenidas en la ecuación general del modelo $x = 100 \times \left(1 - e^{\alpha t + \beta t^2}\right)$

Ecuación de la liberación de urea en función de la T del suelo

Variable dependiente: alfa
Variable independiente: temper
Lineal: Y = a + b*X

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	0.00175436	0.00447117	0.392372	0.6982
Pendiente	-0.0004952	0.000149596	-3.31044	0.0029

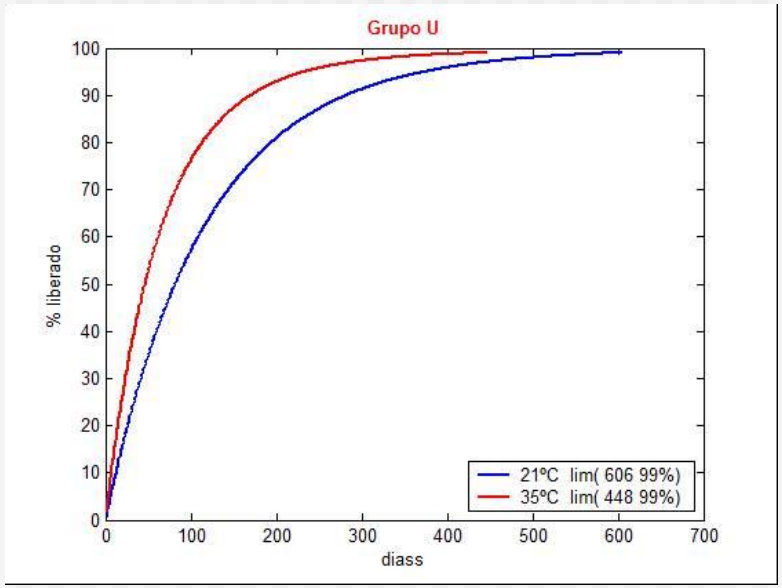
alfa = 0.00175 – 0.000495xT

Variable dependiente: beta
Variable independiente: temper
Lineal: Y = a + b*X

Coeficientes

	Mínimos Cuadrados	*Estándar*	*Estadístico*	
Parámetro	*Estimado*	*Error*	*T*	*Valor-P*
Intercepto	-0.00001345	0.00000733825	-1.83286	0.0793
Pendiente	7.21429E-7	2.45523E-7	2.93834	0.0072

beta = -0.0000134 + 0.0000007xT

Ecuación general: % liberado en función del número **t** de días transcurridos y de la temperatura **T** del suelo

$$x = 100 \times \left(1 - e^{(0.00175 - 0.000495T)t + (-0.000013 + 0.0000007T)t^2}\right)$$