CAPÍTULO 12
MODELOS DE REGRESIÓN

12.1 INTRODUCCIÓN

En este último capítulo se desarrollan los Modelos de Regresión Lineal, que permiten analizar las posibles relaciones entre la pauta de variabilidad de una variable aleatoria y los valores de una o más variables (aleatorias o no) de las que la primera depende, o puede depender.

Los Modelos de Regresión Lineal están estrechamente relacionados con los Modelos de Análisis de la Varianza vistos en anteriores capítulos que, de hecho, son un caso particular de los primeros.

Tras unas ideas generales, se exponen en primer lugar las diferentes fases que deben seguirse en el estudio de un problema mediante modelos de regresión. En realidad el proceso esquematizado tiene una validez mucho más general, puesto que constituye el esquema básico a seguir al abordar el análisis empírico de problemas reales mediante cualquier modelo estadístico.

Se desarrolla a continuación de forma intuitiva el modelo básico de regresión lineal simple, precisándose las hipótesis en que se basa y los elementos básicos del mismo. Se estudian seguidamente diversas generalizaciones de este modelo básico, que enriquecen sensiblemente la capacidad de las técnicas de regresión para modelar problemas reales. Estas generalizaciones incluyen los modelos de regresión múltiple, la consideración de relaciones no lineales y la inclusión en los modelos de variables cualitativas y de interacciones. En todos los casos se hace especial hincapié en la interpretación de la naturaleza de los diferentes parámetros que aparecen en dichos modelos.

Se exponen a continuación, de forma somera, las ideas básicas relativas a la estimación del modelo así como los principales resultados a utilizar en el análisis inferencial del mismo.

Se destina también un apartado a las técnicas de validación de los modelos, con especial referencia a los métodos de análisis de residuos, de tanta importancia en la práctica.

La parte más importante del capítulo la constituye la discusión detallada de un modelo real, fruto de la experiencia personal de los autores.

El estudio de los contenidos de este capítulo, debe completarse con la realización en la sesión de prácticas del análisis de diversos problemas reales.

12.2 MODELOS DE REGRESIÓN: IDEAS GENERALES

Los Modelos de Regresión Lineal permiten analizar la posible relación existente entre la pauta de variabilidad de una variable aleatoria y los valores de una o más variables (aleatorias o no), de las que la primera puede depender.

El recurso a los modelos de regresión resulta indispensable cuando no es posible fijar previamente los valores a adoptar por las variables explicativas en un determinado estudio, como sucede en particular si éstas son de tipo aleatorio (por ejemplo, efecto de la temperatura diaria en el consumo de energía de una instalación), dado que en estos casos no es posible diseñar un experimento que garantice la ortogonalidad de los efectos a investigar.

También es necesario recurrir a técnicas de regresión en el análisis de información histórica que no fue obtenida a partir de un diseño experimental, por ejemplo, los datos procedentes del control estadístico de cierto proceso recopilados el último año, o los datos resultantes de una determinada encuesta.

Desde el punto de vista computacional los Modelos de Regresión exigen cálculos mucho más laboriosos que los implicados en los Anova utilizados en Diseño de Experimentos. El recurso a un paquete estadístico es en estos casos prácticamente indispensable.

En un estudio de regresión se dispone de J observaciones de una variable aleatoria Y_j (por ejemplo, el consumo diario de energía constatado en una factoría automovilística en J días invernales) junto con los valores correspondientes de I variables (aleatorias o no) $X_{1j},...,X_{Ij}$, de las que la primera puede depender (por ejemplo, la temperatura y la producción de vehículos en dichos días).

Se trata en general de estudiar las posibles relaciones existentes entre la distribución de Y_j y los valores de las X_{ij}. A la Y se le denomina generalmente la variable **dependiente**, mientras que frecuentemente a las X_i se les llama variables **independientes** o **exógenas** del modelo, aunque nosotros preferimos la denominación de variables **explicativas**.

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

$$\mathbf{E(Y_j) = f(X_{1j}...X_{Ij})}$$
(ecuación de regresión)

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

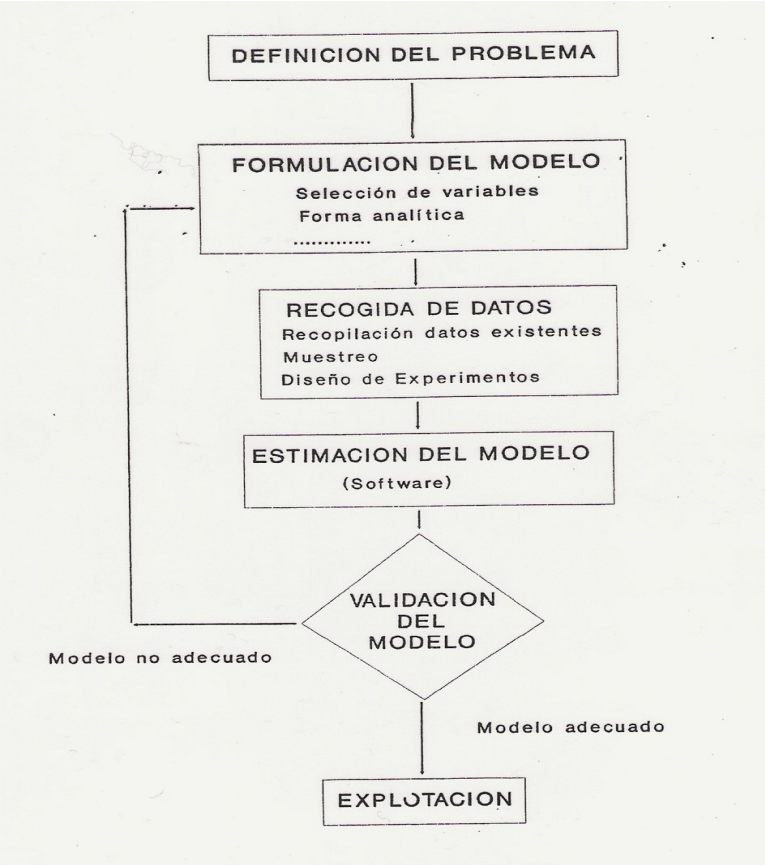
Los parámetros de la ecuación de regresión permiten precisar la naturaleza y cuantificar la magnitud de los efectos de las diferentes variables explicativas sobre el valor medio de la variable dependiente. Dichos parámetros se estiman a partir de los datos disponibles, utilizando los eficientes procedimientos estadísticos que se exponen en este capítulo, analizándose su significación mediante las técnicas de inferencia correspondientes.

Mediante técnicas, que no son más que una generalización inmediata de las expuestas en capítulos anteriores para estudiar en el Anova efectos sobre la dispersión, es posible tratar modelos más complejos, en los que se asume adicionalmente la posibilidad de que las varianzas $\sigma^2(Y_j)$ también puedan depender de las variables explicativas.

12.3 FASES DE UN ESTUDIO 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 que comentamos a continuación.



Definición del problema

Es ésta, sin duda, la fase más delicada y trascendente en el estudio de cualquier problema real. No sin razón se afirma que el error que más frecuentemente cometen los técnicos es el de **resolver correctamente el problema equivocado** (usando la jerga estadística KIMBALL bautizó este tipo de fallo como "error de 30 especie").

Una definición clara del problema que pretende analizarse y de los objetivos finales perseguidos en el estudio, es indispensable antes de empezar a recoger o analizar datos

Formulación del modelo

Esta fase implica aspectos como la definición precisa de la variable dependiente, la selección de variables explicativas, la elección de la forma analítica de la ecuación del modelo, la decisión sobre la inclusión o no de interacciones y el estudio la necesidad o conveniencia de aplicar en su caso transformaciones a las variables.

La formulación del modelo debe venir acompañada por la interpretación del sentido que tendrán los parámetros que aparecen en el mismo. (Al final de esta fase debe ser conocido "lo que es" cada uno de los parámetros del modelo, aunque todavía no se conozca "lo que vale", puesto que los valores de dichos parámetros deben ser estimados a partir de los datos).

La fase de formulación del modelo tiene también una importancia capital, y en ella deben verterse todos los conocimientos técnicos existentes sobre el problema en cuestión. No debe olvidarse que la Estadística no es nunca un sustitutivo de dichos conocimientos técnicos, sino más bien un catalizador que permite sacar más partido de los mismos.

Tal como se refleja en el esquema, la formulación de cualquier modelo tiene siempre carácter provisional y tentativo, y la adecuación del mismo para el problema en estudio debe ser revisada en una fase posterior de validación del modelo

Recogida de datos

Es fundamental darse cuenta que el diseño adecuado de la fase de recogida de datos, sólo puede elaborarse si se tiene una idea bastante clara del modelo que pretende estimarse a partir de los mismos.

En principio la situación más favorable se presenta cuando, como sucede en las ciencias experimentales, es posible fijar los valores que pueden tener en nuestras observaciones las diferentes variables explicativas. En estos casos, las técnicas de Diseño de Experimentos expuestas en anteriores capítulos permiten obtener una gran cantidad de información con un número reducido de pruebas. Adicionalmente el análisis de esta información es muy sencillo, pudiendo llevarse a cabo como se expuso en dichos capítulos, no exigiendo el recurso a las técnicas que se exponen en el presente.

En el campo de las ciencias de la observación la recogida de datos se lleva a cabo mediante el diseño de planes de muestreo adecuado. La metodología de estos planes de muestreo constituye un capítulo muy importante de la Estadística Aplicada.

Estimación del modelo

Formulado un modelo y recogidos los datos pertinentes, la fase de estimación tiene como objetivo utilizar la información existente en dichos datos para obtener los valores más verosímiles de los parámetros del modelo, así como para estimar la precisión de las estimaciones obtenidas.

La estimación de los modelos de regresión múltiple es bastante laboriosa desde el punto de vista computacional, pero la generalización de la disponibilidad de software de gran calidad desarrollado al efecto, ha hecho que deje de constituir el "cuello de botella" en este tipo de estudios. De hecho el ingeniero puede en cierto sentido despreocuparse de los detalles operativos del proceso de estimación, aunque debe tener ideas claras sobre la naturaleza de los resultados que este proceso le proporciona.

Validación del modelo

Una fase esencial, y sin embargo frecuentemente olvidada, en todo estudio estadístico es la validación crítica del modelo estimado. En síntesis se trata de aprovechar la información contenida en los datos, no sólo para estimar los parámetros del modelo postulado, sino también para cuestionarse la adecuación de las hipótesis en que se basa dicho modelo.

Cuestiones importantes que deben ser abordadas en esta fase son, entre otras, las siguientes: ¿son los datos normales?, ¿hay algún dato anómalo que pueda afectar sensiblemente a las conclusiones del estudio?, ¿es adecuada la forma analítica elegida?

Como afirma el profesor Box :

"Todos los modelos son falsos
pero...
algunos modelos son útiles"

No se puede pretender que un modelo sea el "verdadero", dado que el proceso de modelación implica necesariamente una abstracción y simplificación de la realidad. (¿Existen en la realidad péndulos matemáticos o rectángulos?). Pero el modelo debe recoger de forma razonablemente aproximada los aspectos relevantes de la misma, con el fin de que las conclusiones que se obtengan de su manipulación matemática supongan una aproximación útil a efectos prácticos de los resultados que se presentan en dicha realidad.

Las técnicas de validación del modelo, que son básicamente diferentes tipos de análisis de residuos como los vistos en anteriores capítulos, pueden ayudar además, en el caso de que el modelo en cuestión sea rechazado, a sugerir las modificaciones a introducir en el mismo para que se adapte mejor a la realidad observada.

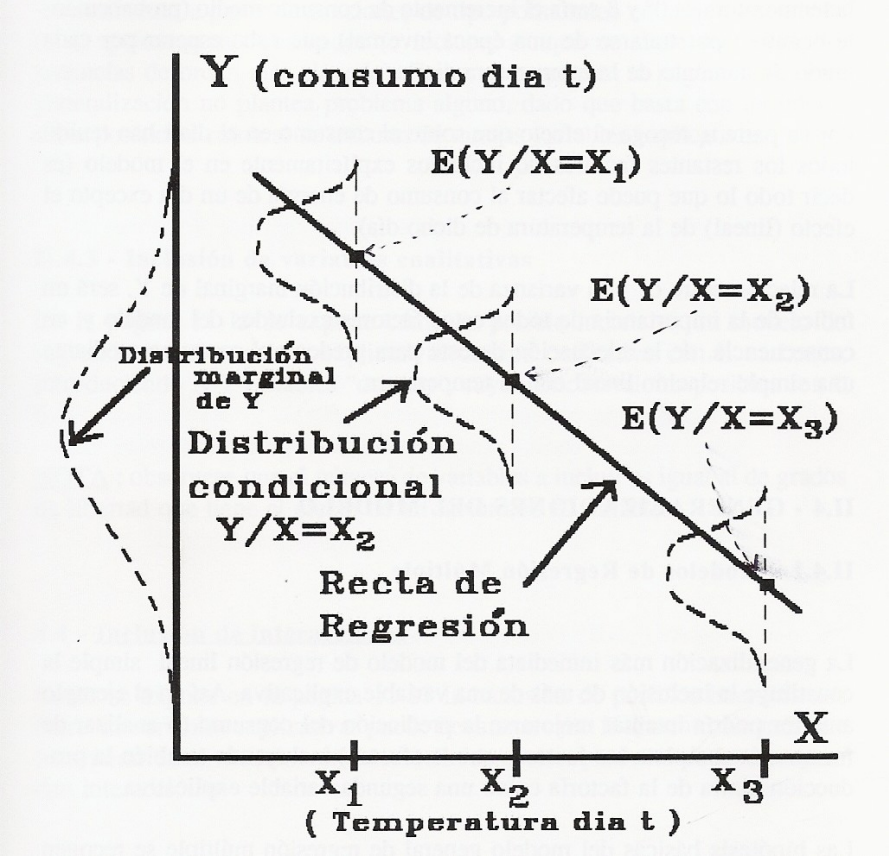
Si el resultado de esta fase de validación es negativo habrá que proceder en consecuencia a reformular el modelo, retornándose de nuevo a la fase 2.

Explotación del modelo

Una vez estimado y validado el modelo, y de acuerdo con los objetivos perseguido en el estudio, éste debe ser explotado para tomar decisiones sobre el problema en cuestión, (el objetivo de todo estudio estadístico ingenieril debe ser siempre, de forma más o menos directa, la **toma de decisiones**) y, en general, para mejorar nuestro conocimiento sobre el mismo.

12.4 MODELO DE REGRESIÓN LINEAL SIMPLE

Sea **Y** una variable aleatoria cuya distribución puede depender, en un sentido que precisaremos a continuación, de otra variable **X**. En concreto utilizaremos el ejemplo, ya manejado en el capítulo anterior, en el que Y es el consumo de energía en calefacción en los días de invierno en una factoría y X es la temperatura del día considerado.



Tal como se refleja en la figura anterior, si se considera la población de todos los días de invierno, se constatará que el consumo fluctúa sensiblemente 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.

Si consideramos exclusivamente los días invernales en los que la temperatura tiene un valor bajo x_1 (por ejemplo, 5ºC), se constata también una variabilidad en los consumos de energía (distribución condicional de **Y** cuando $\mathbf{X}=x_1$). 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 $\mathbf{X}=x_1$, tendrá un valor medio $E(\mathbf{Y}/\mathbf{X}=x_1)$, que posiblemente será superior al valor medio $E(\mathbf{Y})$ de la distribución marginal de **Y**, por estar considerándose sólo días con una temperatura baja.

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

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 $\alpha + ßx_t$ de dicho valor

$$\mathbf{E(Y/X=x_t) = \alpha + ßx_t}$$

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

Se dispone de un conjunto de J pares de valores observados x_t,y_t es decir de los valores de la temperatura y del consumo en J 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

$$u_t = y_t - (\alpha + ßx_t)$$

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

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

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

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

$$y_t = \alpha + ßx_t + u_t$$

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

En el modelo anterior:

α correspondería al consumo promedio los días en que la temperatura es 0º

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

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.

12.5 GENERALIZACIONES DEL MODELO

12.5.1 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 (o analizar de forma más completa los factores que le afectan) incluyendo también la producción diaria de la factoría como una segunda variable explicativa.

Sean :

Y_t: valor de la variable dependiente en la observación t

X_{1t}: valor de la primera variable explicativa X_1 en la observación t

.....

X_{lt}: valor de la l-ésima variable explicativa X_l en la observación t

El modelo de regresión lineal múltiple postula:

$$\mathbf{E(Y/X_1=x_{1t},....,X_l=x_{lt}) = \beta_0 + \beta_1 x_{1t} +.... + \beta_l x_{lt}}$$

Por tanto la ecuación básica del modelo es:

$$\mathbf{Y_t = \beta_0 + \beta_1 x_{1t} +.... + \beta_l x_{lt} + u_t}$$

donde las u_t son variables aleatorias $N(0,\sigma^2)$ e independientes

La interpretación de los parámetros de este modelo, será:

$ß_0$: Valor medio de Y cuando X_1=.....=X_l=0

$ß_i$: Incremento en el valor medio de Y cuando X_i aumenta en una unidad, manteniéndose constantes las restantes variables explicativas. Se asume que este incremento es constante y no depende de los valores de las x_{it}.

12.5.2 Consideración de relaciones no lineales

Los modelos expuestos hasta el momento asumen que el posible efecto de las variables explicativas 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 relaciones de tipo más general pueden aproximarse satisfactoriamente a partir de funciones de tipo polinómico, en las que además de la x_i aparezcan sus cuadrados (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

Los parámetros de este modelo tendrían la siguiente interpretación:

$ß_0$: E(Y) cuando X_1 es igual a cero

$ß_1$: Pendiente en el origen (aproximadamente igual al aumento de E(Y) cuando X_1 pasa de 0 a 1, si el rango de variación de X_1 es de varias unidades)

ß$_2$: 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)

Modelo general de 2º grado

Una ecuación muy utilizada en la práctica para modelar la relación entre E(Y) y un conjunto de variables explicativas X_1, X_2,...,X_l es la ecuación general de 2º grado

$$E(Y) = \beta_0 + \sum_i \beta_i x_i + \sum_i \beta_{ii} x_i^2 + \sum_{i<j} \beta_{ij} x_i x_j$$

La interpretación de los diferentes parámetros en este modelo es la siguiente:

ß$_0$: E(Y) en el origen (es decir para $X_1=X_2=...=X_l=0$)

ß$_i$: Pendiente en el origen del efecto de la variable X_i. Indica, aproximadamente (si el rango de valores estudiados para X_i es de varias unidades) el $\Delta E(Y)$ cuando X_i pasa de 0 a 1, estando las restantes variables a nivel cero

ß$_{ii}$: Curvatura del efecto de la variable X_i. Valores positivos indican que el efecto lineal de X_i aumenta a medida que lo hace X_i (función cóncava hacia arriba o con curvatura positiva), mientras que valores negativos indican que el efecto lineal de X_i disminuye a medida que aumenta X_i (función cóncava hacia abajo o con curvatura negativa)

ß$_{ij}$: Interacción entre los efectos lineales de X_i y X_j. Indica en cuanto aumenta el efecto lineal de una de las variables por cada unidad que aumenta la otra.

El modelo de 2º grado asume, por tanto, que los efectos lineales pueden variar al hacerlo las X_i, pero que los efectos cuadráticos y las interacciones permanecen constantes en el rango de valores estudiados para las variables explicativas.

Frecuentemente, especialmente cuando valores nulos de las variables originales X_i carecen de sentido, los valores concretos de los coeficientes se interpretan más fácilmente si las variables se centran, haciendo coincidir su origen con el valor medio observado para cada una.

12.5.3 Inclusión de variables cualitativas

Tiene un **gran interés práctico**, la posibilidad de incluir variables explicativas de naturaleza cualitativa en modelos de regresión, puesto que ello potencia enormemente la posibilidad de modelar mediante los mismos fenómenos reales.

Por ejemplo, en un estudio sobre la resistencia a la compresión (BCT) de papeles para embalajes de cartón, interesa estudiar el efecto sobre el valor medio de dicha resistencia del gramaje de los papeles (grs./m^2) y del tipo de papel (3 tipos distintos codificados como 1, 2 y 3).

Y : Resistencia a la compresión (BCT) del papel
X_1 : Gramaje, medido en exceso sobre 100 ($X_1=0 \Rightarrow$ gramaje=100 grs./cm^2)
X_2 : Tipo de papel (1: papel TIPO1; 2: papel TIPO2; 3: papel TIPO3)

Lo que ¡NUNCA! puede hacerse es formular un modelo incluyendo los códigos de una variable cualitativa (con más de dos posibles "valores") como si se tratara de una variable cuantitativa

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad \textbf{INCORRECTO}$$

Autoevaluación: comprobar que en el modelo anterior los parámetros β_0 y β_2 carecen de sentido

La solución consiste en crear, por cada variable cualitativa con K alternativas, K-1 nuevas variables, con únicos valores posibles 0 ó 1, que indiquen en cada caso cuál es la alternativa correspondiente. Estas variables binarias se denominan en inglés variables "dummy", y a veces en castellano variables "ficticias".

Por ejemplo, en el caso considerado, se definirán las dos variables siguientes:

Z_2: vale 1 si el papel es el TIPO2 y 0 en caso contrario
Z_3: vale 1 si el papel es el TIPO3 y 0 en caso contrario

	Variables Dummy	
Tipo Papel	**Z_2**	**Z_3**
TIPO1	0	0
TIPO2	1	0
TIPO3	0	1

Valores de las variables dummy según el tipo de papel

Nota: Obsérvese que el número de variables "dummy" a crear, y por tanto el número de parámetros a introducir en el modelo, para considerar el efecto del factor cualitativo, coincide con los grados de libertad asociados a ese factor en un Anova .

Nota: En Statgraphics, si la variable X2 tiene los códigos 1, 2 ó 3 del tipo de papel, para crear una de estas nuevas variables "dummy", por ejemplo Z_2, basta introducir en el cuadro de diálogo como variable explicativa la expresión X2=2, que es una operación lógica cuyo resultado es 1 si X2 es igual a 2 y es 0 en caso contrario

El modelo: $$\mathbf{E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3} \qquad (1)$$

sí que es correcto, puesto que sus parámetros tienen un significado técnico concreto, que puede deducirse particularizando la ecuación anterior para cada tipo de papel.

$E(Y/X_1, \text{TIPO1}) = \beta_0 + ß_1 X_1$
$E(Y/X_1, \text{TIPO2}) = \beta_0 + \beta_1 X_1 + \beta_2 = (\beta_0+\beta_2) + \beta_1 X_1$
$E(Y/X_1, \text{TIPO3}) = \beta_0 + \beta_1 X_1 + ß_3 = (\beta_0+\beta_3) + \beta_1 X_1$

Restando la primera ecuación de la segunda y de la tercera se deduce de forma inmediata:

β_2: diferencia del BCT medio obtenida usando el tipo de papel 2 respecto a cuando se usa el tipo de papel 1 (sea cual sea el valor del gramaje X_1)

β_3: diferencia del BCT medio obtenida usando el tipo de papel 3 respecto a cuando se usa el tipo de papel 1 (sea cual sea el valor del gramaje X_1)

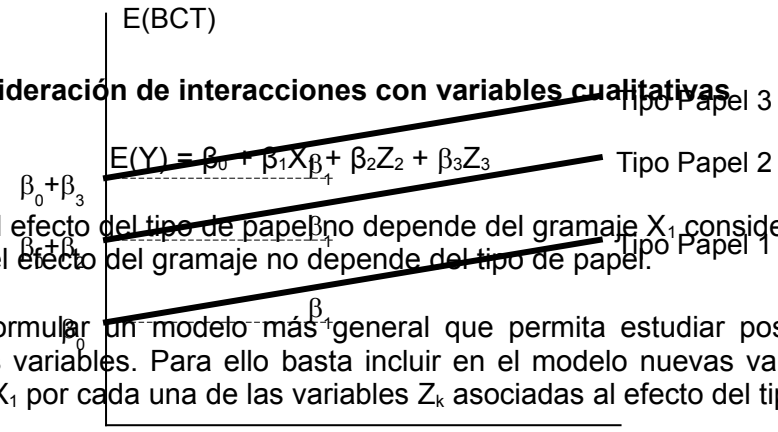
La hipótesis de que los tres tipos de papeles tienen la misma resistencia sería, por tanto, equivalente a: $\beta_2 = \beta_3 = 0$

La interpretación de los restantes parámetros del modelo (1) será:

β_0: BCT medio con el tipo de papel 1 y gramaje 100 (o sea con X_1=0)

β_1: incremento en E(BCT) por cada gramo que aumenta el gramaje. El modelo asume que dicho incremento es idéntico en los tres tipos de papeles, o sea que el efecto del gramaje no depende del tipo de papel (ausencia de interacción)

En definitiva, el modelo (1) corresponde a una situación como la representada en la siguiente figura, en la que se ha asumido $\beta_3 > \beta_2$ y positivos ambos.



12.5.4 Consideración de interacciones con variables cualitativas

El modelo: $E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3$

asume que el efecto del tipo de papel no depende del gramaje X_1 considerado o, lo que es lo mismo, que el efecto del gramaje no depende del tipo de papel.

Es posible formular un modelo más general que permita estudiar posibles interacciones entre las dos variables. Para ello basta incluir en el modelo nuevas variables que sean el producto de X_1 por cada una de las variables Z_k asociadas al efecto del tipo de papel.

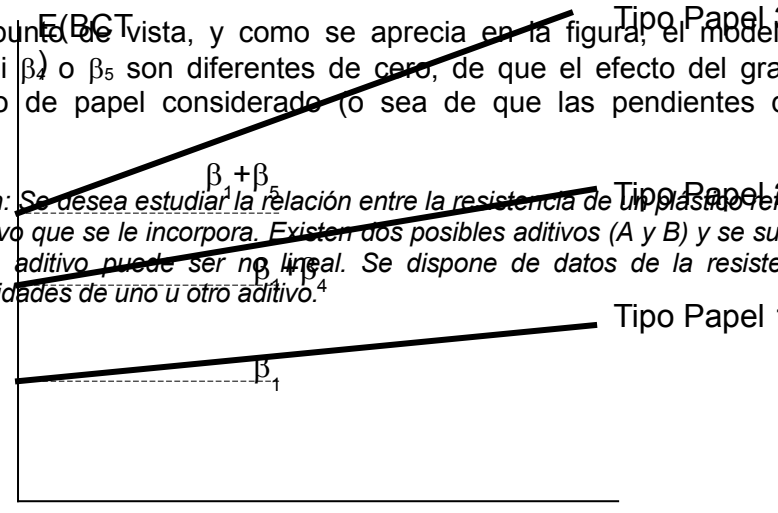
En efecto el modelo **$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 X_1 Z_2 + \beta_5 X_1 Z_3$** (2)

implica:

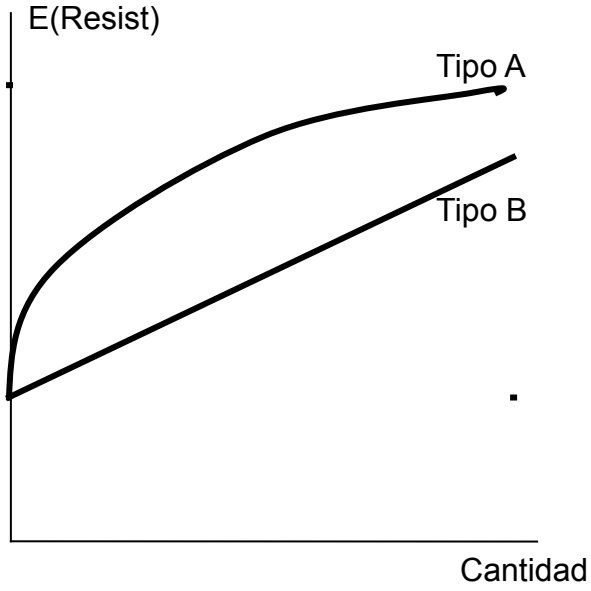
$$E(Y/TIPO1) = \beta_0 + \beta_1 X_1 + 0 + 0 = \beta_0 + \beta_1 X_1$$
$$E(Y/TIPO2) = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_4 X_1 = (\beta_0+\beta_2) + (\beta_1+\beta_4)\, X1$$
$$E(Y/TIPO3) = \beta_0 + \beta_1 X_1 + \beta_3 + \beta_5 X_1 = (\beta_0+\beta_3) + (\beta_1+\beta_5)\, X1$$

Por tanto la diferencia entre el BCT medio con el TIPO2 respecto al obtenido usando el TIPO1 será = $\beta_2 + \beta_4 X_1$, que depende del gramaje, e igual sucede con la diferencia entre el TIPO3 y el TIPO1 que resulta ser $\beta_3 + \beta_5 X_1$.

Desde otro punto de vista, y como se aprecia en la figura, el modelo (2) contempla la posibilidad, si β_4 o β_5 son diferentes de cero, de que el efecto del gramaje sea diferente según el tipo de papel considerado (o sea de que las pendientes de las rectas sean diferentes).



Autoevaluación: Se desea estudiar la relación entre la resistencia de un plástico reforzado y la cantidad y tipo de un aditivo que se le incorpora. Existen dos posibles aditivos (A y B) y se supone que el efecto de la cantidad de aditivo puede ser no lineal. Se dispone de datos de la resistencia conseguida con diferentes cantidades de uno u otro aditivo.



a) Formular un modelo de regresión lineal que permita analizar dicho datos, especificando con precisión la definición de las variables del modelo y el significado de sus parámetros.
b) Asumiendo que la relación real entre E(Resistencia) y la cantidad de aditivo es la indicada en la figura adjunta para ambos aditivos, indicar el signo que tendrían cada uno de los parámetros del modelo formulado
(Ver respuesta en el Anejo al final del Tema)

12.6 ESTIMACIÓN DEL MODELO

12.6.1 Objetivos

Dado un determinado **modelo** de regresión

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

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

y 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 de las I variables explicativas en N observaciones,

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

las estimaciones b_i de los I+1 parámetros β_i

la estimación de las desviaciones típicas s_{bi} de las estimaciones anteriores (que son una medida del margen de incertidumbre asociado a cada estimador)

la estimación s^2 de la varianza residual σ^2 del modelo

12.6.2 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 :

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

Se demuestra que los estimadores b_0 , b_1 , ... , b_I óptimos, 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.

Si se dispone de un conjunto de datos, la suma de cuadrados residual $\sum_{j=1}^{N}\left(y_j - (b_0 + b_1x_{1j} + \ldots + b_Ix_{Ij})\right)^2$ es una función de los I+1 parámetros b_i, cuyo mínimo se puede obtener de la forma habitual, igualando a cero sus I+1 derivadas respecto a dichos parámetros.

Resultados matemáticos

(La lectura de este apartado puede obviarse sin perder la comprensión del resto del capítulo)

Siendo $\vec{y}$ el vector con los valores de la variable dependiente en la N observaciones, y $\overline{\overline{X}}$ la matriz Nx(I+1) cuya fila j-ésima es [1 x_{1j} ... x_{Ij}], el vector $\vec{b}$ de los estimadores de las β_i viene dado por la expresión matricial:

$$\vec{b} = \left(\overline{\overline{X}}'\overline{\overline{X}}\right)^{-1}\overline{\overline{X}}'\vec{y}$$

La varianza residual σ^2 del modelo se estima por el cuadrado medio residual:

$$s^2 = \frac{\sum_j e_j^2}{N-1-I} = \frac{\sum_j \left(y_j - (b_0 + b_1x_{1j} + \ldots + x_{Ij})\right)^2}{N-1-I}$$

La desviación típica de cada uno de los b_i viene dada por la expresión:

$s_{b_i} = s\sqrt{c_{ii}}$

donde los c_{ii} son los elementos de la diagonal principal de $\left(\overline{\overline{X}}'\overline{\overline{X}}\right)^{-1}$

12.7 COEFICIENTE R^2. ANOVA DEL MODELO

La variabilidad total de la variable dependiente Y en el conjunto de las N observaciones viene medida por la suma de cuadrados total:

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

y tiene N-1 grados de libertad.

Parte de esta variabilidad es debida (o, al menos, está asociada) a las variables explicativas $X_1,...,X_I$. Esta parte explicada por dichas variables tiene I grados de libertad (tantos como variables explicativas haya en el modelo)

El resto estará recogido en los residuos e_j, viniendo medida su magnitud por la Suma de Cuadrados Residual

$$SC_{Residual} = \sum_j e_j^2$$

que tendrá (N-1) - I grados de libertad

La diferencia:

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

es la parte de la variabilidad de Y asociada a las variables explicativas

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 lógicamente 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 explicativas incluidas en el modelo.

Para estudiar la hipótesis nula de si al menos una de las variables explicativas estudiadas tiene un efecto real poblacional, o sea de si al menos una $ß_i$ es diferente de cero, se utiliza el siguiente resultado:

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

mientras que 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 rechazará, por tanto, de la forma habitual, 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 con el que se desee trabajar)

12.8 TESTS DE HIPÓTESIS SOBRE LAS β_I

Dado el modelo

$$E(Y_j) = \beta_0 + \beta_1 X_{1j} + ... + \beta_i X_{ij} + ... + \beta_I X_{Ij}$$

$$\text{la variable } X_i \text{ no influye sobre } E(Y) \Leftrightarrow \beta_i = 0$$

El test para contrastar la hipótesis nula H_0: $ß_i = 0$, frente a la alternativa $\beta_i \neq 0$ que implica la existencia de un efecto real poblacional de la X_i sobre E(Y), puede realizarse de la siguiente forma:

Se demuestra que si $\beta_i = 0 \Rightarrow \frac{b_i}{s_{b_i}}$ se distribuye como una t_{N-1-I}

mientras que si $ß_i \neq 0$ el cociente tiende a ser en valor absoluto mayor que una t de Student.

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

12.9 PREDICCIONES EN MODELOS DE REGRESIÓN

Sea el modelo

$$Y_t = \beta_0 + \beta_1 X_{1t} + ... + \beta_I X_{It} + u_t$$

El valor medio poblacional de Y cuando $X_1 = x_{1t} = ... = X_i = x_{It}$ será:

$$m_t = E(Y/X_1 = x_{1t} ... X_I = x_{It}) = \beta_0 + \beta_1 x_{1t} + ... + \beta_I x_{It}$$

Una vez estimado el modelo, m_t puede predecirse mediante:

$$m_t^* = b_0 + b_1 x_{1t} + ... + b_I x_{It}$$

Se demuestra que m_t^* se distribuye normalmente con media y varianza que vienen dadas por las expresiones:

$E(m_t^*) = m_t$

$$\text{varianza}(m_t^*) = \sigma^2 \vec{x}_t' \left(\overline{\overline{X}}' \overline{\overline{X}}\right)^{-1} \vec{x}_t$$

donde $\overline{\overline{X}}$ es la matriz vista en la Nota del apartado 12.6.2 y $\vec{x}_t'$ es el vector $[1 \ x_{1t} ... \ x_{It}]$

Un intervalo de confianza para m_t vendrá dado por:

$$m_t^* \pm t_{N-1-I}^{(\alpha)} s \sqrt{\vec{x}_t' \left(\overline{\overline{X}}' \overline{\overline{X}}\right)^{-1} \vec{x}_t}$$

donde s es la estimación de la desviación típica residual del modelo ($s=\sqrt{CM_{Res}}$)

12.10 VALIDACIÓN DEL MODELO. ANÁLISIS DE RESIDUOS

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.

Recordemos que el modelo se sintetiza en la ecuación

$$y_t = ß_0 + ß_1 x_{1t} + ... + ß_I x_{It} + u_t$$

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

Distintas cuestiones pueden plantearse relativas a la adecuación o no de estas hipótesis ante unos datos concretos:

¿Es admisible el que las u_t se distribuyen normalmente?

¿Hay algún dato claramente anómalo?

¿Es admisible que la varianza de las u_t no depende de los valores de las variables explicativas?

¿Es realmente lineal la relación entre E(Y) y una X_i?

Como ya se ha visto en otros capítulos a lo largo de este texto, la herramienta más poderosa para analizar estas cuestiones es el **análisis de residuos**.

Como sabemos el residuo estimado para cada observación e_t no es más que la diferencia entre el valor realmente observado (y_t) y el valor medio previsto a partir del modelo estimado para esos valores concretos de las variables explicativas ($b_0+b_1x_{1t}+...+b_Ix_{It}$)

$$e_t = y_t - (b_0+b_1x_{1t}+...+b_Ix_{It})$$

Los residuos e_t son en realidad las estimaciones de los valores de las perturbaciones aleatorias u_t en cada observación.

Nota: En Statgraphics es posible, utilizando el icono correspondiente a "salvar resultados" guardar en el fichero de datos los residuos de cualquier análisis estadístico

Determinadas representaciones gráficas de los residuos son extremadamente útiles para responder a algunas cuestiones que se plantean en la fase de validación de los modelos de regresión

1- Un gráfico de los e_t en papel probabilístico normal permite estudiar si es admisible la hipótesis de normalidad, así como detectar posibles observaciones anómalas.

2- Un gráfico de los cuadrados de los residuos frente a las diferentes x_i puede poner de manifiesto el hecho de que algunas de las variables explicativas afectan a la varianza de las y_t. En situaciones de este tipo, un análisis aproximado de los efectos sobre la varianza puede llevarse a cabo (de forma completamente análoga a la expuesta en los capítulos relativos al Análisis de la Varianza) realizando un ajuste de regresión múltiple de los valores de los cuadrados de los residuos frente a las x_{it}.

3- Gráfico de residuos frente a predicciones: Un gráfico de los e_t frente a los valores previstos para cada observación $b_0+b_1x_{1t}+...+b_Ix_{It}$ puede poner de manifiesto la existencia de relaciones no lineales, que se detectan por una configuración curvada con predominio de residuos positivos para valores extremos de la predicción y de residuos negativos para los valores intermedios, si la curvatura es positiva, o con predominio de residuos negativos para valores extremos de la predicción y de residuos positivos para los valores intermedios, si la curvatura es negativa

Nota: El gráfico 3 puede obtenerse directamente dentro de la opción "Multiple Regression" de Statgraphics. Gráficos de los tipos 1 y 2 pueden también elaborarse, mediante otras opciones ya vistas de Statgraphics, a partir de los residuos obtenidos y salvados mediante la opción anterior.

12.11 EJEMPLO DE SÍNTESIS: UN MODELO PARA EL CONTROL DEL CONSUMO DE ENERGÍA

12.11.1 Objetivo del modelo

Una factoría automovilística desea establecer un gráfico para controlar su consumo diario de energía, concretamente el de un tipo de gas utilizado para la calefacción de sus instalaciones en el periodo de octubre a abril. El objetivo del mismo, como el de cualquier gráfico control industrial, es el de detectar precozmente la presencia de cualquier anomalía (por ejemplo, una fuga de gas o un defectuoso funcionamiento de los quemadores) y ayudar a la identificación de la misma, con el fin último de eliminarla rápidamente del sistema (si es desfavorable) o de fijarla definitivamente (si es favorable).

En principio el establecimiento de un gráfico de control estándar exige la estimación previa de la media y de la desviación típica de la característica a controlar (en este caso, el consumo diario de energía) cuando el proceso funciona normalmente. Posteriormente los consumos diarios constatados se llevan a un gráfico en el que se dibuja una línea central, a la altura de dicho valor medio, y dos límites de control situados 3σ por encima y por debajo de la misma. Las salidas de control del proceso se detectan por la aparición de un punto fuera de los límites de control, o por la presencia de configuraciones especiales, como por ejemplo una racha de 7 ó más puntos consecutivos a un mismo lado de la línea central.

En el caso que nos ocupa, sin embargo, un gráfico de tipo estándar no resulta adecuado, dado que podría producir señales de falta de control cuando el proceso funciona perfectamente y no detectar, sin embargo, la presencia de anomalías importantes, al no tener en cuenta los efectos que sobre el consumo de energía pueden tener diversos factores, especialmente la temperatura diaria. Así un día muy frío el consumo puede resultar muy alto (por encima del límite superior de control) pese a no haber ninguna anomalía a corregir en el sistema de calefacción. Por el contrario es posible que si un día caluroso se observa un consumo próximo a la media diaria invernal, ello sea señal de un funcionamiento defectuoso del sistema que debe investigarse.

Para controlar el proceso es necesario por tanto establecer un modelo que permita predecir el consumo medio que cabe esperar en las condiciones concretas de cada día y la σ correspondiente, y llevar al gráfico las diferencias entre los valores realmente observados y los previstos por el modelo (o sea los residuos constatados) frente a unos límites iguales a $0 \forall 3\sigma$.

12.11.2 Modelo inicial

El primer paso en el proceso de modelación es dar una definición precisa de las variables a considerar en el modelo. De acuerdo con los objetivos perseguidos y con la información disponible, se adoptaron las siguientes definiciones operacionales:

Consumo: diferencia entre las lecturas del contador general de gas de tipo B a las 6,30 de la mañana (inicio del primer turno) de un día respecto a la realizada a la misma hora del día anterior. (Por motivos de confidencialidad en este texto se ha multiplicado por una constante, por lo que viene expresado en una unidad arbitraria a la que nos referiremos como "termias")

Temper: temperatura del día en 1C, definida como la media aritmética de las 48 medidas realizadas cada media hora entre las 6,30 de un día y la misma hora del siguiente.

En segundo lugar es necesario definir con precisión el ámbito de aplicación del modelo. En este caso se decidió que el mismo se centraría sólo sobre los días laborables normales, prescindiéndose de sábados y domingos, del periodo 15 de octubre a 15 de abril en el que funciona la calefacción.

En cuanto a las variables a considerar y a la forma analítica, en un primer modelo simplificado se decide prescindir de otras posibles variables explicativas y ensayar un modelo sencillo de regresión lineal

$$Consumo_t = ß_0 + ß_1 Temper_t + u_t \quad \text{(Modelo 1)}$$

De la definición del modelo se deduce la siguiente interpretación de los parámetros y del residuo:

$ß_0$ = consumo medio los días que la temperatura es 0 ºC

$ß_1$ = incremento del consumo medio cuando se incrementa 1 ºC la temperatura (el modelo asume que este incremento es constante y no depende de la temperatura)

u_t = diferencia entre el consumo real constatado el día t y el consumo medio que corresponde a un día de temperatura igual a la observada dicho día. Como de costumbre se asume que las u_t son independientes y siguen distribuciones $N(0,\sigma^2)$

12.11.3 Recogida de datos

En el momento del estudio se disponía de los datos correspondientes a los 57 últimos días laborables, recogidos de acuerdo con las definiciones dadas.

Se prescindió previamente de dos días en los que el consumo fue anormal por haberse realizado sendos paros por motivos laborales.

Los datos utilizados en el estudio se recogen en el archivo gas.sf3[1]. (Como se ha indicado, las cifras de consumo están todas multiplicadas por la misma constante arbitraria para respetar su confidencialidad).

12.11.4 Estimación del modelo inicial

La estimación del modelo1 a partir de los datos disponibles proporcionó el siguiente resultado:

```
Dependent variable: Consumo
-----------------------------------------------------------------------------
                                   Standard          T
Parameter              Estimate      Error       Statistic        P-Value
-----------------------------------------------------------------------------
CONSTANT                448.912      7.63267       58.8145         0.0000
Temper                 -18.4109     0.627143      -29.3567         0.0000
-----------------------------------------------------------------------------

                         Analysis of Variance
-----------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-----------------------------------------------------------------------------
Model                   552051.0      1     552051.0     861.82       0.0000
```

[1] Todos los archivos de datos que se mencionan en este texto pueden bajarse libremente de la URL http://personales.upv.es/rromero/descargas. Los alumnos de la Universidad Politécnica de Valencia pueden también bajárselos del Poliformat de la asignatura.

```
Residual                  35231.1     55      640.566
-------------------------------------------------------------------------------
Total (Corr.)            587282.0     56

R-squared = 94.001 percent
```

Los dos parámetros resultan muy significativos estadísticamente. El valor estimado de $ß_0$ indica que el consumo medio previsible los días en que la temperatura es 01C es 449 termias, mientras que el valor estimado de $ß_1$ indica que, en promedio para los valores estudiados, el consumo medio disminuye 18.4 unidades por cada grado que aumente la temperatura.

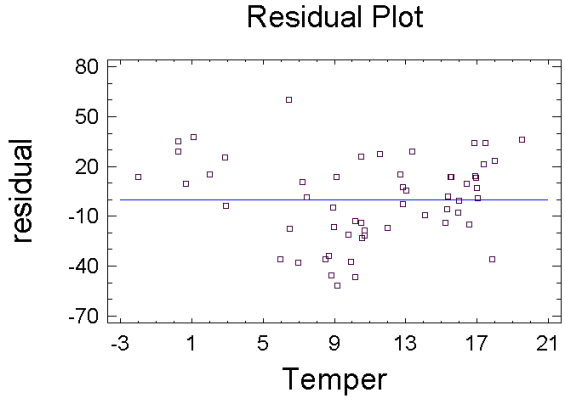
El efecto lineal de la temperatura explica ¡el 94%! de la variabilidad constatada en el consumo diario de energía. La desviación típica residual es igual a 25.

Autoevaluación: aceptando como válido el modelo anterior ¿entre qué límites cabe esperar que oscile el consumo el 95% de los días en los que la temperatura sea 10 1C?

Pese a este elevado valor de la R^2, el modelo puede ser mejorado refinando la forma de la relación funcional entre consumo de energía y temperatura e incluyendo otras variables explicativas adicionales.

12.11.5 Relación funcional entre E(Consumo) y Temperatura

Con el fin de validar la adecuación de la forma analítica adoptada en el modelo (ecuación lineal) se obtiene un gráfico de los residuos en función de los valores de Temper.



El gráfico pone de manifiesto una estructura curvilínea, con predominio de residuos positivos cuando los valores de Temper son bajos o altos, y predominio de valores negativos cuando los valores de Temper son intermedios. Esta situación indica que los valores observados se sitúan en general por encima de la recta estimada para valores extremos de Temper y por debajo de la misma para los intermedios. Se deduce, en consecuencia, que el modelo lineal no es adecuado y que es aconsejable introducir un término de segundo grado en la ecuación, para captar mejor la naturaleza del efecto que la temperatura tiene sobre el consumo diario de energía.

Se decide en consecuencia estimar el nuevo modelo :

$$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + u_t \quad \text{(Modelo 2)}$$

De acuerdo con el Modelo 2, el nuevo sentido de los parámetros $ß_i$ será :

$ß_0$ = consumo medio los días en que la temperatura es 0 1C (igual que en Modelo 1)

$ß_1$ = Pendiente en el origen. Aproximadamente igual al incremento del consumo medio cuando se incrementa la temperatura pasando de 01C a 11C.

$ß_2$ = Medida de la curvatura de la ecuación Consumo = f(Temper). (Podría definirse como la mitad de la variación de la pendiente de dicha ecuación por cada grado que aumenta la temperatura)

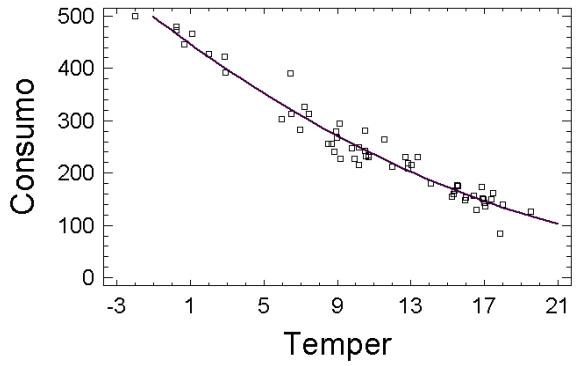
El resultado de la estimación del Modelo 2 es el siguiente:

```
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
------------------------------------------------------------------------------

                          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 (Corr.)           587282.0      56

R-squared = 95.545 percent
```

Como se aprecia el término cuadrático resulta muy significativo estadísticamente lo que confirma la conveniencia de incluirlo en el modelo. El signo positivo obtenido para dicho parámetro indica una curvatura positiva de la ecuación, lo que implica que el consumo medio aumenta cada vez más rápidamente a medida que disminuye la temperatura, tal como se refleja en la siguiente figura



Con el fin de estudiar si estaría justificado utilizar un modelo aun más complicado, que contemplara la posibilidad de que la curvatura de la ecuación no fuera constante, se ha ajustado una nueva ecuación incluyendo un término cúbico:

$$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + ß_3 Temper_t^3 + u_t \quad \text{(Modelo 3)}$$

Autoevaluación: ¿Qué interpretación tienen en el Modelo 3 los diferentes parámetros $ß_i$? (Ver respuesta en el Anejo al final del Tema)

Ajustando este nuevo modelo se han obtenido los resultados que se recogen a continuación:

Dependent variable: Consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	471.856	8.85003	53.3169	0.0000
Temper	-25.0805	3.94592	-6.35607	0.0000
Temper^2	0.26303	0.538862	0.488121	0.6275
Temper^3	0.00510361	0.0196354	0.259919	0.7959

Como se aprecia el coeficiente del término $Temper^3$ no resulta significativo estadísticamente, por lo que con su introducción se estaría complicando innecesariamente el modelo.

Nota importante:

Obsérvese que la introducción de $Temper^3$ ha hecho que tampoco resulte ahora significativo el término $Temper^2$. El origen de este fenómeno, muy frecuente en análisis de problemas reales radica, por una parte, en la estrecha correlación existente en los datos entre los valores de estas dos variables explicativas (a medida que aumenta $Temper^2$ también lo hace $Temper^3$, dado que la mayoría de las temperaturas son positivas) que se traduce en que estadísticamente no sea posible "separar" cuál de las dos es la que influye sobre Consumo.

Por otra parte, hay que tener en cuenta que el coeficiente $ß_2$ tenía en el Modelo 2 un significado diferente al que tiene en el Modelo 3, pues mientras en éste corresponde a la curvatura en el origen, en aquél estaba asociado a la curvatura media en la región estudiada. La no significación de b_2 y b_3 en el Modelo 3 indica que, a partir de los datos disponibles, es imposible diferenciar entre las dos situaciones siguientes:

- Una curvatura positiva constante en toda la región estudiada (que implicaría $\beta_2>0$ y $\beta_3=0$)

- Una curvatura nula en el origen (para Temper=0) pero creciente a lo largo de la zona estudiada (que implicaría $\beta_2=0$ y $\beta_3>0$). Puede constatarse, en efecto, que si se ajusta el modelo $E(Consumo) = \beta_0 + ß_1 Temper_t + ß_3 Temper_t^3$, el parámetro β_3 resulta muy significativo

Por tanto, como a partir de los datos no es posible asegurar que β_2 es $\neq 0$ ni tampoco que β_3 es $\neq 0$, ninguno de los dos parámetros resulta estadísticamente significativo al estimar el Modelo 3

La opción lógica, en consecuencia, es mantener el Modelo 2, que es el más sencillo de los que se ajusta bien a los datos.

12.11.6 Consideración del efecto del día de la semana

Con el fin de mejorar las predicciones obtenidas con el modelo, se decidió incluir en el mismo adicionalmente la temperatura del día anterior (variable Temp_ant) y el posible efecto del día de la semana.

Para modelar este último se incluyeron 4 variables "dummy", asociadas respectivamente a los días MARTES, MIERCOLES, JUEVES y VIERNES, proponiéndose el siguiente modelo:

$$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + ß_3 Temp\_ant_t + ß_4 MARTES_t + ß_5 MIERCOLES_t + ß_6 JUEVES_t + ß_7 VIERNES_t + u_t \quad \text{(Modelo 4)}$$

donde, por ejemplo, $MARTES_t$ vale 1 si el día t es martes y 0 en caso contrario.

Autoevaluación: ¿Qué significado exacto tienen los diferentes parámetros $ß_i$ en el modelo (4)? ¿Qué hipótesis están implícitas en la utilización de este modelo?

Nota: dado que la variable DIA contiene los códigos (1, 2, 3, 4 ó 5) de los cinco días laborables, para crear cada variable "dummy" basta con escribir en el cuadro de diálogo la expresión DIA = "código del día correspondiente"

Realizado el ajuste del Modelo 4 se obtiene el siguiente resultado:

Dependent variable: Consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	478.805	9.74108	49.1531	0.0000
Temper	-22.8388	2.42179	-9.43052	0.0000
Temper^2	0.353422	0.0926083	3.81632	0.0004
Temp_ant	-2.37723	1.4228	-1.67081	0.1011
DIA=2	-2.05894	9.83595	-0.209328	0.8351
DIA=3	-10.994	8.31683	-1.3219	0.1923
DIA=4	-0.297069	9.10169	-0.032639	0.9741
DIA=5	-22.1611	8.06962	-2.74624	0.0084

Como se aprecia el único efecto asociado al día de la semana que resulta significativo es la diferencia entre viernes y lunes, mientras que las existentes entre los restantes días y el lunes no son significativas estadísticamente.

La explicación a la diferencia constatada entre el viernes y los restantes días se debe a la forma en la que se ha definido el consumo de cada día (desde las 6,30 de la mañana de ese día hasta la misma hora del siguiente) y al hecho de que todos los días, excepto los sábados, se consume una energía adicional antes de las 6,30 para la puesta en marcha de las instalaciones.

Dado que el efecto del día de la semana parece limitarse a la diferencia entre viernes y el resto, parece innecesario mantener 4 parámetros al respecto en el modelo, y es preferible un modelo más sencillo que sólo incluya la variable VIERNES (DIA=5):

$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + ß_3 Temp\_ant_t + ß_7 VIERNES_t + u_t$ Modelo 5

Autoevaluación: ¿Cuál es ahora en el Modelo 5 el significado del parámetro $ß_7$?

El ajuste del Modelo 5 proporciona los siguientes resultados:

Dependent variable: Consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	475.304	7.9449	59.8251	0.0000
Temper	-22.3307	2.20004	-10.1501	0.0000
Temper^2	0.355978	0.086475	4.11654	0.0001
Temp_ant	-2.97109	1.30455	-2.27749	0.0269
DIA=5	-18.6483	6.40165	-2.91305	0.0053

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	565881.0	4	141470.0	343.75	0.0000
Residual	21400.6	52	411.551		

```
Total (Corr.)            587282.0     56

R-squared = 96.356 percent
Standard Error of Est. = 20.2867
```

Autoevaluación: Un técnico opinó que la diferencia entre los restantes días y el viernes debía ser mayor cuanto menor fuera la temperatura (por ser mayor, si el día es frío, el ahorro de no hacer la puesta en marcha)
Formular un modelo que permita estudiar la hipótesis anterior. ¿Cuál es el significado preciso de todos los parámetros de este nuevo modelo? ¿Qué signo cabe esperar que tenga el nuevo parámetro introducido de acuerdo con la hipótesis del técnico?
Estimar el modelo propuesto y ver si está justificado complicar el modelo para considerar la posibilidad de que la diferencia entre los viernes y el resto dependa de la temperatura.

12.11.7 Efectos sobre la varianza del Consumo

Con el fin de estudiar si la temperatura o el día de la semana, además de afectar al valor medio del consumo, también influían sobre la varianza de éste alrededor de su media, se ajusto un nuevo modelo usando como variable dependiente los cuadrados de los residuos obtenidos en el ajuste del Modelo 5.

Los resultados obtenidos del ajuste se recogen a continuación:

Dependent variable: RESIDUALS^2

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	202.394	212.177	0.953891	0.3447
Temper	57.1036	35.6764	1.6006	0.1158
Temp_ant	-56.794	36.2896	-1.56502	0.1239
DIA=2	44.5905	246.523	0.180878	0.8572
DIA=3	255.228	216.083	1.18116	0.2431
DIA=4	394.898	235.295	1.67831	0.0995
DIA=5	93.679	209.194	0.447809	0.6562

Como se aprecia no resulta significativo el efecto de ninguna variable sobre la desviación típica de los residuos, por lo que puede adoptarse para la misma un valor constante igual a 20.3, que fue el obtenido al estimar el Modelo 5.

12.11.8 Explotación del modelo

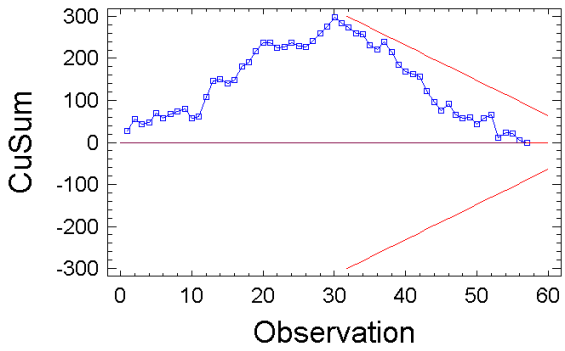
Del estudio realizado se deduce que el consumo medio previsible un día t, viene dado por la ecuación:

$$475 - 22.3Temper_t + 0.356Temper_t^2 - 2.97Temp\_ant_t - 18.6VIERNES_t$$

siendo la σ estimada en torno a dicha media aproximadamente igual a 20.

Con estos datos se estableció un gráfico de control, que se utilizó en primer lugar para analizar si el proceso había permanecido bajo control durante el periodo de toma de datos para el estudio. Se utilizó para ello un esquema CUSUM, en el que se grafican las sumas acumuladas de las desviaciones de los datos observados respecto al valor medio teórico correspondiente a cada día. En este tipo de gráfico, un disminución en el nivel medio del proceso se detecta visualmente por una serie creciente seguida de una decreciente, produciéndose el cambio de tendencia precisamente en la observación en la que se produjo dicho cambio.

En la siguiente figura se refleja el gráfico CUSUM correspondiente al periodo de 57 días en el que se tomaron los datos utilizados para ajustar el modelo.



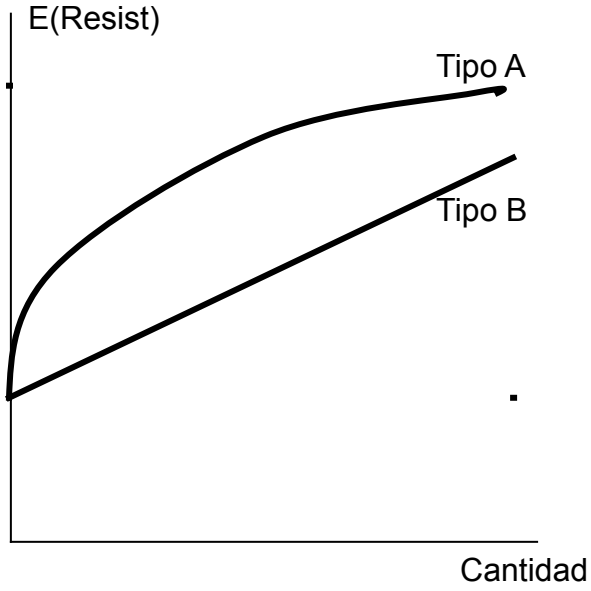
El gráfico CUSUM puso claramente de manifiesto un cambio en los niveles medios de consumo a partir del día 30, cambio que se identificó con cierta medida tomada en dicha fecha.

Autoevaluación: ¿Es posible cuantificar el efecto de la mencionada medida sobre el nivel medio del consumo? Proponer un modelo de regresión que permita estimar dicho efecto, precisando el significado de los diferentes parámetros del modelo. Estimar a partir de los datos (recogidos en gas.sf3) el modelo propuesto y cuantificar el ahorro debido a la medida adoptada el día 30. (Ver respuesta en el Anejo al final del Tema)

12.A AUTOEVALUACIONES RESUELTAS Y EJERCICIOS

12.A.1 Respuesta a algunas Autoevaluaciones

Autoevaluación: Se desea estudiar la relación entre la resistencia de un plástico reforzado y la cantidad y tipo de un aditivo que se le incorpora. Existen dos posibles aditivos (A y B) y se supone que el efecto de la cantidad de aditivo puede ser no lineal. Se dispone de datos de la resistencia conseguida con diferentes cantidades de uno u otro aditivo.



a) Formular un modelo de regresión lineal que permita analizar dicho datos, especificando con precisión la definición de las variables del modelo y el significado de sus parámetros.
b) Asumiendo que la relación real entre E(Resistencia) y la cantidad de aditivo es la indicada en la figura adjunta para ambos aditivos, indicar el signo que tendrían cada uno de los parámetros del modelo formulado

Definamos las siguientes variables:

Y: Resistencia del plástico reforzado
X_1: cantidad de aditivo (en grs/Kg)
TIPB: variable dummy que vale 1 si el aditivo es de tipo B y 0 en caso contrario

El modelo a proponer es:

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 TIPB + \beta_4 TIPBx\ X_1 + \beta_5 TIPBx X_1^2$

cuyos parámetros tienen el siguiente significado:

β_0: resistencia media si no se incorpora ningún aditivo
β_1: incremento en la resistencia media al pasar de 0 a 1 gr/Kg con el aditivo A
β_2: curvatura de la relación entre E(Y) y cantidad de aditivo A
β_3: diferencia en E(Y) usando el aditivo B respecto a cuando se usa el aditivo A, a dosis cero en ambos casos. Obviamente β_3 debe ser cero, por lo que carece de sentido incluir este término en el modelo
β_4: diferencia entre el incremento en la resistencia media al pasar de 0 a 1 gr/Kg con el aditivo B, y el incremento en la resistencia media al pasar de 0 a 1 gr/Kg con el aditivo A
β_5: diferencia entre la curvatura de la relación entre E(Y) y cantidad de aditivo B y la curvatura de la relación entre E(Y) y cantidad de aditivo A

De acuerdo con las definiciones anteriores, y a la vista de la figura, se tiene:

$\beta_0 > 0$ $\beta_1 > 0$ $\beta_2 < 0$ $\beta_3 = 0$ $\beta_4 < 0$ $\beta_5 > 0$ (pues $\beta_2+\beta_5 = 0$)

Autoevaluación:) Qué interpretación tienen en el Modelo 3 los diferentes parámetros ß$_i$?

En una ecuación de tercer grado $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ se tiene que:

la pendiente y' es igual a $\beta_1 + 2\beta_2 x_1 + 3\beta_3 x_1^2$

la curvatura será proporcional a $y'' = 2\beta_2 + 6\beta_3$

Por tanto, en el Modelo 3 los parámetros tendrían el siguiente significado:

β_0: E(Consumo) cuando Temper = 0ºC
β_1: (pendiente en el origen) incremento en E(Consumo) cuando Temper pasa de 0C a 1ºC
β_2: será proporcional a la curvatura en el origen de la función E(Consumo) = f(Temper)
β_3: indicará si la curvatura de la función de regresión aumenta ($\beta_3>0$), disminuye ($\beta_3<0$) o permanece constante ($\beta_3=0$) al aumentar Temper

Autoevaluación: ¿Es posible cuantificar el efecto de la mencionada medida sobre el nivel medio del consumo? Proponer un modelo de regresión que permita estimar dicho efecto, precisando el significado de los diferentes parámetros del modelo. Estimar a partir de los datos (recogidos en gas.sf3) el modelo propuesto y cuantificar el ahorro debido a la medida adoptada el día 30.

Habrá que incluir en el modelo una nueva variable "dummy", a la que denominaremos MEJORA, que valga 0 hasta el día 30 y 1 a partir de dicho día.

En el nuevo modelo:

$E(Consumo)=ß_0+ß_1 Temper+ß_2 Temper^2+ß_3 Temp\_ant+ß_7 VIERNES+\beta_8 MEJORA$

β_8 medirá la diferencia entre los consumos medios después y antes de la mejora, en igualdad de circunstancias para las restantes variables.

La estimación mediante Statgraphics del nuevo modelo da como resultados:

```
Dependent variable: Consumo
-----------------------------------------------------------------------------
                                     Standard          T
Parameter              Estimate        Error       Statistic        P-Value
```

CONSTANT	476.213	6.49833	73.2823	0.0000
Temper	-21.3071	1.80965	-11.7742	0.0000
Temper^2	0.34047	0.0707678	4.81108	0.0000
Temp_ant	-2.79573	1.06717	-2.61976	0.0116
DIA=5	-18.4928	5.23425	-3.53304	0.0009
MEJORA	-24.7228	4.777	-5.17537	0.0000

que pone de manifiesto que el efecto de la mejora es muy significativo estadísticamente, habiéndose traducido ésta en un ahorro promedio de casi 25 termias diarias.

12.A.2 Ejercicios resueltos

Con el fin de mejorar ciertas propiedades en el polipropileno utilizado en parachoques es conveniente aditivarlo con CO3CA. Sin embargo este aditivo empeora la resistencia al impacto del plástico. Para precisar este fenómeno se realizaron pruebas aditivando tres tipos diferentes de polipropileno (variable TIPO codificada como 1, 2 ó 3) con diferentes porcentajes de CO3Ca (variable DOSIS expresada en tanto por mil) y midiéndose la resistencia obtenida en cada caso (variable RESIST medida en newtons)
Se realizó un análisis mediante regresión múltiple de estos datos, obteniéndose mediante Statgraphics los siguientes resultados

Independent variable	coefficient	std. error
CONSTANT	41.964356	1.958876
DOSIS	-0.249608	0.06314
TIPO=2	12.742358	2.770269
TIPO=3	7.395407	2.770269
(TIPO=2)* DOSIS	-0.403074	0.089294
(TIPO=3)* DOSIS	-0.322229	0.089294

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	1730.40	5	346.079	42.0890	.0000
Error	197.34	24	8.222		
Total	1927.74	29			

*a) ¿Qué interpretación tiene el parámetro -0.403074 asociado a (TIPO=2)*DOSIS ¿Es significativo estadísticamente (para α=0.05)?*
b) ¿Cuál es el valor de la R^2 del modelo?
c) Se desea aditivar la mayor cantidad posible de CO3CA, para mejorar las otras propiedades, pero con el requisito de que la resistencia media obtenida sea mayor o igual que 26 newtons. ¿Qué cantidad de CO3CA y qué tipo de plástico deberá utilizarse?

a) El parámetro indicado mide la diferencia de pendientes de las rectas que relacionan E(RESIST) con DOSIS para el tipo de polipropileno 2 y para el tipo de polipropileno 1.
Dado que b_i/s_{bi} = -0.403/0.089 = -4.53 es, en valor absoluto, > $t_{23}^{0.05} = 2.064$, el parámetro es significativo estadísticamente (para un riesgo de 1ª especie α=0.05)

b) R^2 = 1730.4/1927.74 = 0.898 (89,8% si se expresa en porcentaje)

c) Las ecuaciones estimadas de E(RESIST) en función de DOSIS para cada uno de los tres tipos de polipropileno, y los valores máximos de DOSIS que mantienen E(RESIST) $\geq$ 26 son:

TIPO 1: E(RESIST) = 41.96 - 0.25 DOSIS máximo = (41.96-26)/0.25 = 64

TIPO 2: E(RESIST) = (41.96+12.74) - (0.25+0.40) DOSIS máximo = (54.7-26)/0.65 = 44

TIPO 3: E(RESIST) = (41.96+7.40) - (0.25-0.32) DOSIS máximo = (49.36-26)/0.57 = 41

Por tanto, deberá utilizarse el TIPO 1, aditivándolo con un 64 por mil de carbonato cálcico.

12.A.3 Ejercicios adicionales

En un estudio de regresión para investigar el efecto sobre el rendimiento de un cultivo de la variedad utilizada (variable VA codificadas como 1 ó 2) y el contenido de materia orgánica en el suelo (variable MO expresada en %), se ha estimado, a partir de los rendimientos constatados en 30 parcelas, el siguiente modelo:

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	11.81	0.937833	12.5994	0.0000
VA=2	4.55	1.326296	3.4342	0.0015
MO	1.53	0.156578	9.7941	0.0000
(VA=2)*MO	-1.09	0.221434	-4.9322	0.0000

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	437.268				
Error					146.731
Total (Corr.)	583.999				

a) Calcular aproximadamente la probabilidad de que en una parcela con un 5% de materia orgánica y cultivada con la variedad 2 el rendimiento sea superior a 22
b) ¿A partir de qué porcentaje de materia orgánica en el suelo es preferible cultivar la variedad 1 en vez de la 2?

Estudiar la relación entre PESO y ESTATURA (datos fichero curs8990.sf3) y, especialmente la influencia del SEXO en dicha relación.

Con el fin de facilitar la interpretación de los parámetros del modelo, definir dos nuevas variables que sean:
SEX = SEXO - 1 (con lo que los chicos tendrán de código 0 y las chicas 1)
EST = ESTATURA - 150 (con lo que EST=0 para una ESTATURA de 150 cms)

a) Modelo de regresión lineal simple PESO = β_0 + ß$_1$ EST
- ¿Qué interpretación tienen los parámetros β_0 y ß$_1$?
- Estimar los valores de los parámetros
- ¿Difieren significativamente de 0 los parámetros estimados?
- ¿Cuánto vale R^2?) Qué significa ese valor?

b) Modelo PESO = β_0 + ß$_1$ EST + β_2 SEX
- ¿Qué interpretación tienen ahora β_0, β_1 y β_2?
- Estimar los 3 parámetros y estudiar su significación estadística
- ¿Cuánto ha mejorado el ajuste?

c) Modelo PESO = β_0 + ß$_1$ EST + β_2 SEX + β_3 ESTxSEX
- ¿Qué interpretación tienen los 4 parámetros del modelo?
- Estimar los parámetros.
- ¿Es significativa la diferencia de pendientes entre los dos sexos?

¿Cuál sería en definitiva el modelo que propondrías para describir de forma sintética la relación del peso de un joven con su estatura y con su sexo?

Se desea comparar dos posibles configuraciones de las unidades de disco (codificadas como 0 y 1) en un sistema informático con el fin de minimizar el tiempo medio de respuesta. Dicho tiempo depende de la carga del sistema, siendo esta relación posiblemente no lineal.
A lo largo de varios días se han ensayado una u otra de ambas configuraciones, midiéndose cada día la carga media (en consultas por minuto) y el tiempo medio de respuesta (en segundos). Los resultados observados se recogen en la siguiente tabla.

Config	Carga	Tpo.Resp	Config	Carga	Tpo.Resp
1	1.0	0.9	1	1.8	1.1
0	2.0	0.3	1	2.0	1.5
1	2.4	2.0	0	2.5	0.5

0	3.1	0.8	0	3.9	1.5
1	4.0	2.7	0	4.2	1.6
1	4.3	2.6	1	5.5	3.3
0	5.8	2.5	0	6.4	3.3
0	6.6	3.2	1	7.0	3.5
0	7.5	3.7	0	8.0	4.3
1	8.0	3.9	1	8.2	4.0
0	9.0	5.3	1	9.1	4.3
1	9.2	4.2	0	9.5	5.8
1	10.2	3.9			

a) Ajustar los datos a un modelo que considere la posibilidad de que el efecto de la Carga sea no lineal, e incluya un efecto de la configuración y una posible interacción entre la Configuración y el efecto de la Carga (tanto en su efecto lineal como en el cuadrático)
b) Indicar cuáles de los efectos estudiados han resultado significativos e interpretar las ecuaciones estimadas.
c) Calcular la carga media por encima de la cual es una configuración preferible a la otra, precisando la configuración óptima en función de la carga media previsible.