UNIDAD DIDÁCTICA 5-4

INTRODUCCIÓN A LA REGRESIÓN LINEAL

1. Introducción

En esta parte del Unidad Didáctica 5 se estudia una introducción a los Modelos de Regre¬sión 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 la Unidad Didáctica 5.3 que, de hecho, son un caso particular de los primeros.

Tras unas ideas generales, se desarrolla de forma intuitiva el modelo básico de regresión lineal simple, precisándose las hipótesis en que se basa y los elementos básicos del mismo.

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.

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}.

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

(ecuación de regresión)

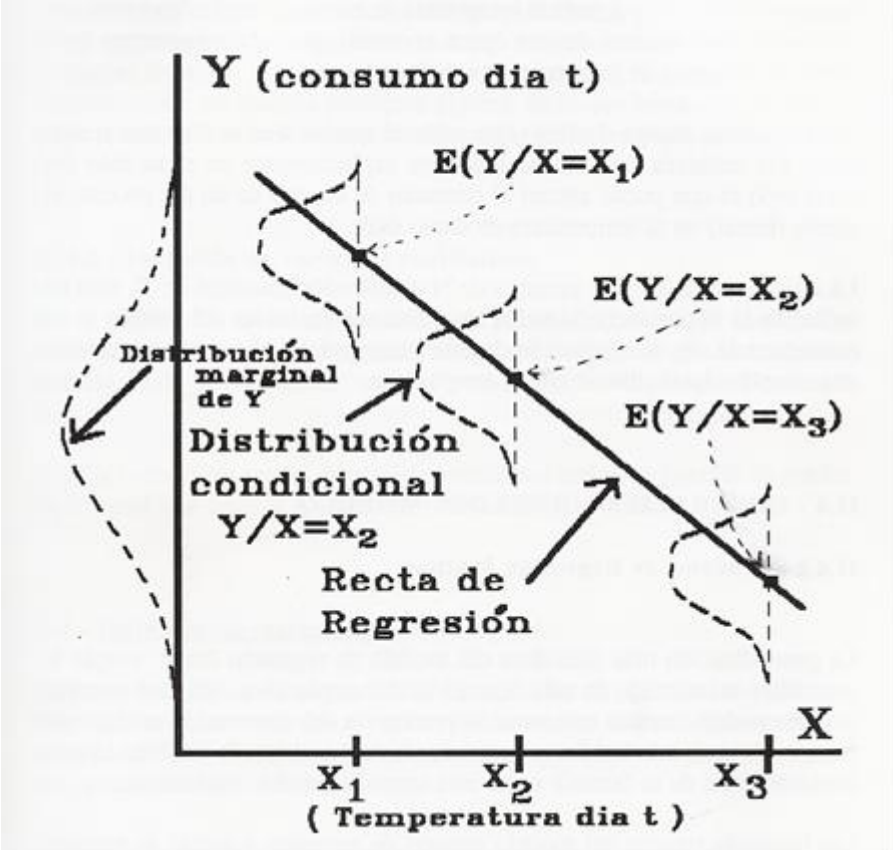
En principio, por tanto, el posible efecto de las X_i sobre la distribución de Y se concreta en modificar el valor medio de dicha variable dependiente.

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 esta unidad, analizándose su significación mediante las técnicas de inferencia correspondientes.

3. Modelo de regresión lineal simple

3.1. Planteamiento del modelo

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

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

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

$$E(Y/X=x_t) = \alpha + ßx_t$$

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

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 ut (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, ut 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.

3.2. Estimación del modelo

Dado un determinado modelo de regresión simple

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

y dados unos datos

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

constituidos por los valores de la variable dependiente y de la variable explicativa 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 a y b de los parámetros α y β

la estimación de las desviaciones típicas s_a y s_b 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

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

y_j : valor de la variable dependiente

x_j: valor de la variable explicativa

El residuo e_j que se obtendría para unos posibles valores a y b de los coeficientes se define como :

$$e_j = y_j - (a + bx)$$

Se demuestra que los estimadores a y b 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 $\sum_{j=1}^{N}\left(y_j - (a + b\ \ x_j)\right)^2$es una función de los parámetros a y b, cuyo mínimo se puede obtener de la forma habitual, igualando a cero sus 2 derivadas respecto a dichos parámetros.

Esto da como resultado que los estimadores de a y b son:

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

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

3.3. Coeficiente de determinación 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:

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

y tiene N-1 grados de libertad.

Parte de esta variabilidad es debida (o, al menos, está asociada) a la variable explicativa X. Esta parte explicada por dichas variables tiene 1 grado 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) - 1 grados de libertad (N-2)

La diferencia:

$$SCExplicada = SCTotal - SCResidual$$

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 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 la variable explicativa X incluida en el modelo. El coeficiente de determinación se suele multiplicar por 100, expresando entonces el porcentaje de variabilidad de Y explicada por X. En el modelo de regresión lineal simple coincide con el valor del coeficiente de correlación al cuadrado r^2x100.

Para estudiar la hipótesis nula de si la variable explicativa X estudiada tiene un efecto real poblacional, o sea de si ß es diferente de cero, se utiliza el siguiente resultado:

Si β = 0

F-ratio=$\frac{SCExplicada/1}{SCResidual/N-2} = \frac{CMExplicado}{CMResidual} \sim F_{1,N-2}$

mientras que si ß es diferente de cero, el cociente anterior es, en promedio, mayor que una $F_{1,N-2}$. La hipótesis de que la variable X tiene un efecto real sobre E(Y) se rechazará, por tanto, de la forma habitual, si el cociente supera el valor en tablas $F_{1,N-2}(\alpha)$ (o, lo que es equivalente, si el correspondiente p-value es inferior al riesgo de 1ª especie α con el que se desee trabajar)

3.4. Test de hipótesis sobre los parámetros α y β

Dado el modelo

$$E(Yj) = \alpha + \beta Xj$$

la variable X no influye sobre E(Y) si β= 0

El test para contrastar la hipótesis nula H_0: ß = 0, frente a la alternativa β≠0 que implica la existencia de un efecto real poblacional de la X sobre E(Y), puede realizarse de la siguiente forma:

Se demuestra que si β = 0

t-calculada=$\frac{b}{s_b}$ se distribuye como una t_{N-2}

mientras que si ß≠ 0 el cociente tiende a ser en valor absoluto mayor que una t de Student.

Por tanto si $\left|\frac{b}{s_b}\right| > t_{N-2}(\alpha)$ se rechaza la hipótesis β= 0 y se deduce que X influye sobre E(Y).

De forma análoga para estudiar la significación de la ordenada α del modelo,

Si $\left|\frac{a}{s_a}\right| > t_{N-2}(\alpha)$ se rechaza la hipótesis α= 0 y se deduce que la ordenada α≠0.

3.5. Predicciones en modelos de regresión

Sea el modelo

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

El valor medio poblacional de Y cuando X= x_t será:

$$m_t = E(Y/X = x_t) = \alpha + \beta x_t$$

Una vez estimado el modelo, m_t puede predecirse mediante:

$$m_t^* = a + bx_t$$

(Y/X = x_t) se distribuye normalmente con media m_t^* y desviación típica residual $s_{residual}$ que se puede calcular $s_{residual}=s_y\sqrt{(1-r^2)} = \sqrt{CMResidual}$.

4. Validación del modelo de regresión. 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

$$y_t = \alpha + ßx_t + u_t$$

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 ut se distribuyen normalmente?

¿Hay algún dato claramente anómalo?

¿Es admisible que la varianza de las u_t no depende de los valores de la variable explicativa?

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

La herramienta más poderosa para analizar estas cuestiones es el análisis de residuos.

Como sabemos el residuo estimado para cada observación et 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 la variable explicativa ($a+bx_t$)

$$e_t = y_t - (a+bx_t)$$

Los residuos e_t son en realidad las estimaciones de los valores de las perturbaciones aleatorias ut 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 la x puede poner de manifiesto el hecho de que la variable explicativa afectan a la varianza de las y_t.

3- Gráfico de residuos frente a predicciones: Un gráfico de los e_t frente a los valores previstos para cada observación $a+bx_t$ 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.

Ejemplo de síntesis: un modelo para el control de consumo de energía.

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 $0 \pm 3\sigma$.

Modelos de regresión lineal simple

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

$$Consumo_t = \alpha + ßTemper_t + u_t$$

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

α = consumo medio los días que la temperatura es 0 ºC

ß = 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 $N(0,\sigma^2)$

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 . (Como se ha indicado, las cifras de consumo están todas multiplicadas por la misma constante arbitraria para respetar su confidencialidad).

Estimación del modelo de regresión simple

La estimación de la recta de regresión a partir de los datos disponibles proporcionó el siguiente resultado:

Dependent variable: Consumo

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	448.912	7.63267	58.8145	0.0000
Temper	-18.4109	0.627143	-29.3567	0.0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	552051.0	1	552051.0	861.82	0.0000
Residual	35231.1	55	640.566		
Total (Corr.)	587282.0	56			

R-squared = 94.001 percent

Los dos parámetros resultan muy significativos estadísticamente. El valor estimado de ß0 indica que el consumo medio previsible los días en que la temperatura es 01C es 449 termias, mientras que el valor estimado de ß1 indica que, en promedio para los valores estudiados, el consumo medio disminuye 18.4 unidades por cada grado que aumente la temperatura.

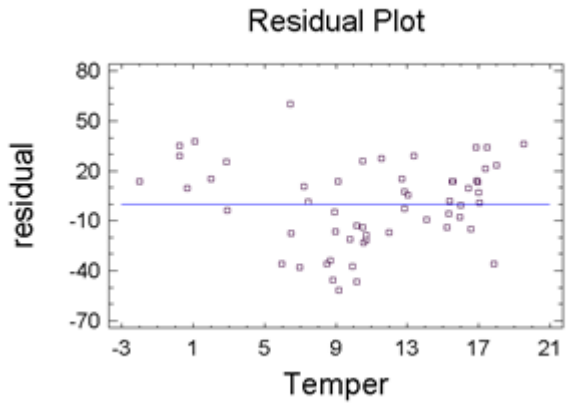
El efecto lineal de la temperatura explica ¡el 94%! de la variabilidad constatada en el consumo diario de energía. La desviación típica residual es igual a 25.

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 ºC?

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.

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 :

$$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + u_t$$

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

$ß_0$ = consumo medio los días en que la temperatura es 0 ºC (igual que en la recta de regresión)

$ß_1$ = Pendiente en el origen. Aproximadamente igual al incremento del consumo medio cuando se incrementa la temperatura pasando de 0ºC a 1ºC.

$ß_2$ = 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)

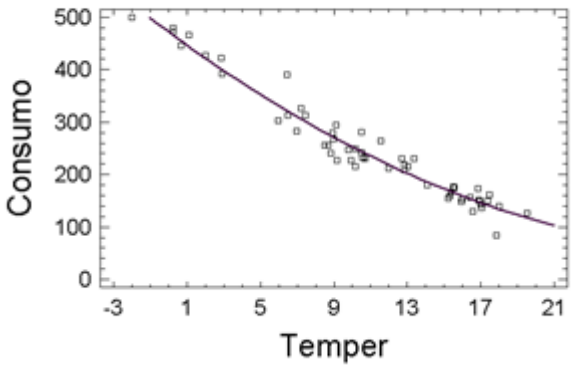
El resultado de la estimación del modelo que incluye el término cuadrático es el siguiente:

```
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:

$$Consumo_t = ß_0 + ß_1 Temper_t + ß_2 Temper_t^2 + ß_3 Temper_t^3 + u_t$$

Autoevaluación: ¿Qué interpretación tienen en el modelo que incluye $Temper^3$ los diferentes parámetros $ß_i$?

SOLUCIÓN:

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 $\beta_1 + 2\beta_2 x + 3\beta_3 x^2$

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

Por tanto, en el modelo con término cúbico 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

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 con efecto cúbico, 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 último modelo 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 último modelo.

La opción lógica, en consecuencia, es mantener el modelo que incluye Temper y $Temper^2$, que es el más sencillo de los que se ajusta bien a los datos.

Fuentes

Métodos Estadísticos en Ingeniería (Romero Villafranca, Rafael)