GENERACION AUTOMATICA DE MODELOS DE REGRESION LINEAL

Ing. Susana B. Chauvet, Sr Gabriel Raya, Ing Carlos A. Correa
Instituto de Ingenieria Quimica, Facultad de Ciencias Exactas y Tecnología, Universidad Nacional de Tucumán, Av Independencia 1800 , 4000 San Miguel de Tucumán, Argentina.

RESUMEN

Se propone un algoritmo para generar un modelo de regresión lineal en forma automática cuando no se dispone de una teoría fundamental que sugiera un modelo determinado.

Se presupone la falta de asociación entre las variables independientes.

El algoritmo integra el modelo progresivamente , mediante la incorporación de una variable por vez, estimando los parámetros mediante mínimos cuadrados.

En cada etapa, el análisis se efectúa con los residuos del modelo precedente, explorando las variables independientes y sus transformadas hasta maximizar el valor del coeficiente de correlación simple, condición que determina la forma funcional del nuevo término lineal.

El modelo queda concluido al momento de probar la falta de asociación lineal de la última variable incorporada.

ABSTRACT

The proposition is an algorithm to generate automatically one linear-regression model, when a principal theory which suggest a determinate model, is not disposable.

Hence, we will assume that the independent variables are not associated.

The model results by incorporating one variable at a time, and their parameters are estimated by the least-squares method. In each step, the analysis is done whith the precedent model's residues, investigating the independent variables and their transformed, up to the simple correlation coefficient be maximun, in such a manner, we determinate the funtional form of the new linear term.

The model is concluded when the non linear association of the last variable is tested.

INTRODUCCION

La regresión es una técnica estadística para determinar la relación entre una variable dependiente y una o más variables independientes .

Utiliza valores experimentales de las variables para generar una expresión funcional que muestra la influencia de las variables independientes sobre la variable dependiente del conjunto.

En Ingeniería Química , la regresión puede ser aplicada para explicar los datos de una amplia variedad de problemas , desde la simple correlación de propiedades físicas hasta el análisis de un proceso industrial complejo .

Por ejemplo , en un reactor catalítico con una reacción química compleja , métodos de regresión han sido utilizados para desarrollar una ecuación para obtener un producto deseado en función de las concentraciones , temperatura , presión y tiempo de residencia .

Frecuentemente se asume una relación lineal . En otros casos , donde no sería adecuada una relación lineal , el Ingeniero puede tratar de ajustar alguna función no lineal . Algunas relaciones funcionales pueden haber sido determinadas sobre bases teóricas del fenómeno . Por ejemplo , la teoría cinética de los gases predice que la viscosidad de estos varía con la potencia (3/2) de su temperatura absoluta .

$$\mu = b\ T^{3/2} \qquad (1)$$

Valores experimentales pueden ser usados para evaluar la constante "b" empleando métodos de regresión . En la práctica , sin embargo , esta función es útil sólo en pequeños rangos de temperatura , porque el modelo físico es muy simplificado .

En la mayoría de los softwares de regresión , debe especificarse el modelo con que se ajustará los datos experimentales . Entonces el computador evalúa los estadísticos asociados al modelo preconcebido. Si no hay una base teórica que sirva de sustentación al modelo propuesto , lo usual es explorar (de idéntico modo) distintos modelos tentativos, de entre los cuales se adopta finalmente aquel que brinde el mayor grado de explicación de la interdependencia experimental de los datos .

Otra alternativa es la integración progresiva de un modelo, mediante la incorporación de una variable por vez, hasta que se alcance la máxima reproducibilidad de los datos con las variables consideradas .

Con esta estrategia se obtiene un modelo de interdependencia para los valores experimentales, seleccionando las variables y sus transformaciones que mejor ajusten los datos .Se obtiene de este modo un modelo estrictamente empírico, sin condicionamiento teórico de ninguna especie.

Cuando el ingeniero necesita conocer la relación funcional entre dos o más variables , esta última alternativa efectúa el análisis usando el coeficiente de correlación de Pearson , el que cuantifica el grado de dependencia lineal entre dos variables.

DESCRIPCION DEL METODO

Se trata de obtener un modelo de máximo grado de correlación, sin el auxilio de ninguna teoría que predetermine la estructura funcional más adecuada.

Sobre esta base, un procedimiento a seguir sería analizar gráficamente la forma de la relación funcional que la variable dependiente acusa frente a cada una de las variables exógenas.

Confrontando la variable dependiente con cada una de las independientes, se obtiene un conjunto de gráficos, cada uno de los cuales sugiere un determinado patrón funcional, con mayor o menor grado de dispersión.

El patrón define un modelo simple, cuyos parámetros se estiman por regresión. Del conjunto de modelos simples, se adopta como más representativo aquél para el cual la suma de los cuadrados de los residuos acusa un valor mínimo.

Los residuos del modelo obtenido conforman una nueva variable dependiente, sobre la cual se adopta igual estrategia para definir el término que mejor explique su variabilidad.

Mediante este procedimiento recurrente se compone progresivamente una relación funcional que incorpora todos los términos necesarios hasta obtener, teóricamente, una completa aleatoriedad en los residuos.

Este es un procedimiento de tipo interactivo. Se puede lograr su automatización mediante el análisis de los coeficientes de correlación simple en forma estrictamente numérica. Los distintos patrones funcionales pueden ser abarcados introduciendo diversas transformaciones sobre las variables independientes, cuyos grados de asociación con la variable endógena se analizarían mediante los coeficientes simples respectivos.

La medida de asociación entre dos variables, por ejemplo, x e y, con una distribución normal, es el coeficiente de correlación simple p. La correlación puede ser positiva o negativa. Cuando es positiva, una variable tiende a aumentar a medida que aumenta la otra; cuando es negativa, una variable tiende a disminuir a medida que disminuye la otra. p está entre los límites -1 y +1. Un valor absoluto alto de p indica un alto grado de asociación mientras que un valor absoluto pequeño indica un grado de asociación bajo.

Cuando el valor absoluto de p es 1, la relación es perfecta. Cuando p=0, las variables son independientes.

Un estimador de p está dado por el coeficiente de correlación muestral r, que se define como:

$$r = \frac{\Sigma(x_i - x_m)(y_i - y_m)}{\sqrt{\Sigma(x_i - x_m)^2 \; \Sigma(y_i - y_m)^2}} \qquad (2)$$

x_m es la media muestral de la variable x

y_m es la media muestral de la variable y

Es posible contrastar la independencia a través del coeficiente de correlación muestral r. En este caso, para un dado nivel de significación α, se puede probar la hipótesis nula Hº de p=0, a partir de los percentiles de la distribución muestral de r.

Este método propone realizar transformaciones a cada variable independiente y estimar el coeficiente de correlación entre cada una de ellas y sus respectivas transformadas con la variable dependiente.

Con este procedimiento se obtiene una matriz de coeficientes de correlación , de la que se elige su elemento de máximo valor . Esto implica seleccionar la variable y su transformación que mejor explica la variabilidad de la variable dependiente .

Efectuado esto , se estima una regresión del tipo :

$$Y = b_0 + b_1 f(X_1) \qquad (3)$$

siendo Y variable dependiente .

X_1 la variable elegida con el mayor r

$f(X_1)$ es la transformación sobre la variable X_1 ,que acusó el máximo coeficiente de correlación simple con Y.

Se estiman los parámetros b_0 y b_1 y se calculan los residuos, o sea :

$$Z_i = Y_i - b_0 - b_1 f(X_{1i}) \qquad i=1,\ldots,n \qquad (4)$$

A partir de este punto se reconstruye una nueva matriz de coeficientes de correlación, pero entre los residuos y las variables independientes y sus transformadas. Se selecciona de modo semejante el mayor valor del coeficiente de correlación y, por lo tanto, la variable y forma funcional que deberá incorporarse al modelo precedente.

A continuación se estiman los parámetros b_0, b_1, b_2 correspondientes al modelo ampliado :

$$Y = b_0 + b_1 f_1(X_1) + b_2 f_2(X_2) \qquad (5)$$

Se determinan entonces los nuevos residuos y se inicia otra etapa de cálculo y selección del coeficiente de correlación .

Esta estrategia continúa incorporando nuevas variables y transformadas hasta que por prueba de hipótesis mediante el estadístico t, el estimador del parámetro del último término incorporado al modelo sea cero , o bien , cuando todas las pruebas de hipótesis de los coeficientes de correlación simples acusan falta de asociación, resultando una matriz de correlación nula, que implica estadísticamente que no es posible incorporar al modelo precedente ninguna variable adicional.

En cada paso , los parámetros son estimados por el método de los cuadrados mínimos .

Una vez definido el modelo se determinan los residuos y se efectúa la correspondiente prueba de normalidad .

Se adoptó a este efecto un método basado en los estadísticos de orden, de acuerdo a los lineamiento que seguidamente se exponen.

Si se tiene n datos, se los ordena de menor a mayor ,en un arreglo $x_{(i)}$,con i=1,2,...,n .

Definiendo z como una variable normal estándar , si los datos provienen de una distribución normal con media μ y desvio estándar σ, entonces deberá verificarse que:

$$x_{(i)} = \mu + \sigma\, z(i) \qquad (6)$$

Por lo tanto, se deberá probar esta dependencia lineal simple.

Los correspondientes valores de z(i) son generados a través del siguiente procedimiento.

Se define :

$$\bar{x}(z) = \frac{(i - 0.5)}{n} \qquad (7)$$

siendo i el indice posicional de los valores ordenados.

Si $\bar{x} > 0.5$ $\quad y = -\ln [2(1-\bar{x})]$ (8)

$$z = \left[\frac{(4y + 100)\, y + (205)\, y^2}{[(2y + 56)\, y + 192]\, y + 131} \right]^{*} \qquad (9)$$

Si $\bar{x} < 0.5$, $z(\bar{x}) = - z(1-\bar{x})$ (10)

Se procede luego a estimar los parámetros del modelo lineal y a contrastar la independencia de x_i vs z(i) ,concluyéndose de este análisis si los datos provienen o no de una distribución normal.

En caso de que los residuos no resultaran normales, una de las posibles causas a tomarse en cuenta será la omisión de algunas variables independientes asociadas al fenómeno analizado.

DESCRIPCION DEL ALGORITMO

1.- Ingreso de Datos

El ingreso de datos se efectúa mediante la lectura de un archivo preestructurado.

La única identificación de las variables es su orden de ingreso. Los datos se organizan en forma tabular, haciendo corresponder a cada columna una variable determinada.La información se ingresa por columnas completas. El programa no maneja juegos experimentales incompletos o datos perdidos.

2.- Selección de la variable dependiente y de su transformación

Se puede afectar como variable dependiente a cualquiera de las del juego, simplemente indicando el número de su orden de ingreso.

Efectuada esta selección, el programa ofrece varias opciones de trans-

formaciones a efectuar sobre dicha variable.

Esta variable transformada oficiará como variable dependiente definitiva para el análisis posterior, sin posibilidad de redefinición en el transcurso del análisis.

3.- Estructuración del modelo de regresión

Una vez seleccionada la variable dependiente todas, las restantes son tratadas como variables independientes, no asociadas entre si. Sobre cada una de ellas, el algoritmo efectúa una serie de transformaciones preestablecidas . Luego calcula el coeficiente de correlación simple entre la variable dependiente y las independientes y sus transformadas.

A todas aquellas transformaciones no definidas en algunos o todos los puntos experimentales , se les asigna un coeficiente de correlación simple igual a cero.

El algoritmo se sirve de una tabulación interna de la distribución muestral de r, para un nivel de significación de un 5%, para contrastar la independencia de las variables correlacionables. En caso de aceptarse la hipótesis nula $p=0$ para la correlación investigada, se asigna a la misma el valor de $r=0$.

Cuadrando los coeficientes de correlación obtenidos, queda estructurada una matriz, correspondiendo cada fila a una variable independiente y las columnas a las tranformaciones preestablecidas.

De esta matriz se selecciona el elemento de máximo valor, quedando así particularizadas la fila y la columna que determinan la variable y su transformación de máxima correlación simple.

Así definida la estructura funcional del nuevo término lineal, éste es incorporado al modelo precedente y el programa evalúa a continuación los parámetros del modelo ampliado mediante cuadrados mínimos.

Con los valores de los estimadores de los parámetros del modelo se informan los correspondientes estadísticos t para contrastar la hipótesis de independencia. Además se acompañan de una tabla de análisis de la varianza, el estadístico F para una prueba conjunta de independencia y el coeficiente de determinación R^2, o sea la proporción de variabilidad explicada por el modelo.

Del análisis de estos estadísticos se derivará la decisión de avanzar en la estructuración del modelo o de dar por concluida la misma.

En caso de que se decidiera la prosecución del análisis, se calculan los residuos y se asimilan los mismos a la nueva variable dependiente, reiniciándose la búsqueda de la variable independiente y su transformada que convenga incorporar al modelo precedente.

Si el modelo estructurado es el definitivo , el algoritmo avanza al análisis de la normalidad de los residuos.

4.- Salida de los resultados

El programa imprime los estimadores de los parámetros del modelo

definitivo, los correspondientes estadísticos y el gráfico de normalidad de los residuos.

APLICACION A UN CASO PRACTICO

En una fábrica de papel, la planta de evaporación de licor negro acusa un ritmo muy marcado de ensuciamiento, impidiendo acceder a mayores regímenes de producción.

Con el propósito de evaluar teóricamente el resultado previsible de algunas soluciones alternativas para este problema, se decide modelar matemáticamente al sistema y aplicar la técnica de simulación por computadora.

La obtención de un modelo de adecuada flexibilidad y confiabilidad no implica mayores dificultades, excepto la disponibilidad de una correlación para la predicción de los valores del coeficiente global de transferencia de calor a partir de variables operativas tales como velocidades de flujos, viscosidades , temperaturas y algunas otras, eventualmente no previstas.

La información disponible sobre las propiedades físicas del licor negro no es suficiente para intentar aplicar algún modelo de correlación preelaborado, de modo que la única alternativa viable es la generación de un modelo particular para el caso en estudio.

A este efecto, se cuantifican todos los caudales en los circuitos de vapor y de licor negro mediante técnicas de balances sustentados en datos mínimos de planta, suficientemente confiables, tales como las temperaturas en cada uno de los cuerpos de evaporación.

El conocimiento de caudales de licor permite estimar las velocidades de flujo terminales (entrada y salida) de cada efecto , y cuantificar el coeficiente global de transferencia que se corresponde con esas mismas conjunciones locales de variables.

Mediante sucesivas evaluaciones de la planta operando bajo distintas condiciones de régimen cuasi-estacionario, se generan suficientes datos de coeficientes de transferencia para diversos niveles de velocidades , viscosidades y temperaturas.

Se tiene así una base experimental supuestamente adecuada para intentar una correlación que explique la variabilidad observada sobre el coeficiente de transferencia dentro del rango de operación de planta.

El orden asignado a las distintas variables para su procesamiento mediante el algoritmo propuesto es :

1.- Coeficiente global de transferencia de calor
2.- Velocidad lineal del licor de alimentación
3.- Viscosidad del licor de alimentación
4.- Velocidad lineal del licor de descarga
5.- Viscosidad del licor de descarga
6.- Salto térmico útil para transferencia

Advirtiendo que la experiencia demuestra que el tipo funcional de las correlaciones del coeficiente global de tranferencia no es lineal , se adopta "a priori" una transformación logaritmica para dicha variable.

Ingresados los datos en el orden referido, el programa estructura la matriz de correlación, efectuando sobre las variables independientes las siguientes transformaciones :

1.- Inversa (1/ X)
2.- Logaritmo
3.- Raiz cuadrada
4.- Raiz cúbica
5.- Potencia , exponente :2
6.- Potencia , exponente :1.5
7.- Potencia , exponente :1
8.- Exponencial

La primera matriz correlación que se obtiene es :

var.	Transformación							
	1	2	3	4	5	6	7	8
X_2	0	0	0	0	0	0	0	0
X_3	.7705	.7789	.7552	.7659	.5673	.6409	.7073	.6224
X_4	0	0	0	0	0	0	0	0
X_5	.5614	.5055	.4591	.4755	.3259	.3634	.4091	.3729
X_6	.5800	.6371	.6323	.6364	.5304	.5702	.6071	.5699

Se observa que el máximo valor corresponde a la transformación logaritmica y a la variable número 3, con lo cual la primera versión del modelo es :

$$Y = b_1 + b_2 \ln(X_3) \qquad (11)$$

Los parámetros estimados y sus correspondientes estadisticos son los siguientes:

$b_1 = 6.5166$ $b_2 = -0.7182$

Parámetro	estadistico t	Varianza del parámetro
6.5166	141.389	0.00212
-0.7182	-9.934	0.00523

Tabla de Análisis de la Varianza

Fuente Variac.	Suma de Cuad.	Grad. Libert.	Cuadrad. Medio
Variables	5.2816	1	5.2816
Residuo	1.4983	28	0.0535
Total	6.7799	29	

$R^2 = 0.77901$ $F = 98.7$

Mediante el análisis resulta evidente que la estructura funcional obtenida no es definitiva.

El programa evalúa los residuos y la matriz de coeficientes de correlación simples obtenida es :

var.	Transformación							
	1	2	3	4	5	6	7	8
X_2	0	0	0	0	0	0	0	0
X_3	0	0	0	0	0	0	0	0
X_4	0	0	0	0	.1339	.1329	0	0
X_5	0	0	0	0	0	0	0	0
X_6	.3460	.2189	.1550	.1753	0	0	0	0

Se observa que el nuevo término a incorporar corresponde a la transformación inversa de la variable X_6 .

Efectuada la incorporación, el modelo ampliado es :

$$Y = b_1 + b_2 \ln(X_3) + b_3 (1/X_6) \qquad (12)$$

Los parámetros estimados y sus correspondientes estadísticos son los siguientes :

$b_1 = 6.0860$ $b_2 = -0.5410$ $b_3 = 4.8917$

Parámetro	estadístico t	Varianza del parámetro
6.0860	67.650	0.00809
-0.5410	-8.647	0.00391
4.8917	5.153	0.90105

Tabla de Análisis de la Varianza

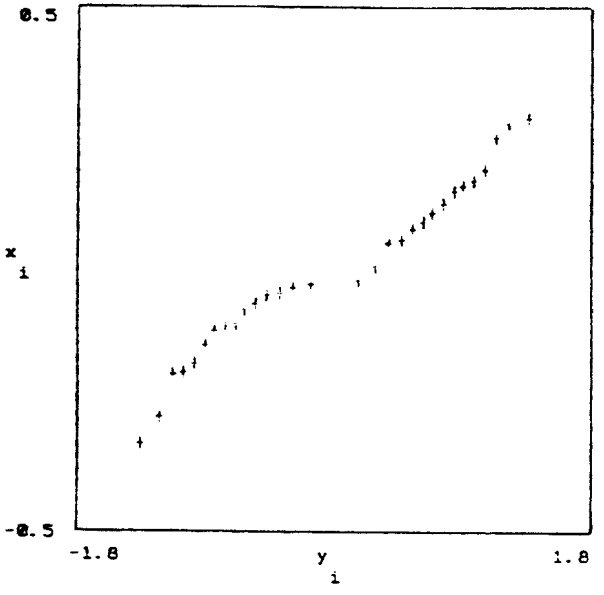
Fuente Variac.	Suma de Cuad.	Grad. Libert.	Cuadrad. Medio
Variables	6.0246	2	3.0123
Residuo	0.7552	27	0.0279
Total	6.7798	29	

$R^2 = 0.88860$ $F = 107.9$

Continuando con el proceso de integración del modelo , la nueva matriz de coeficientes de correlación resulta una matriz nula, por lo que se concluye que el modelo definitivo es :

$$\ln (X_1) = 6.0860 - 0.5410 \ln (X_3) + 4.8917 (1/X_6) \qquad (13)$$

Para culminar el análisis se efectúa el test de normalidad de los residuos, con resultados satisfactorios, según el gráfico al pie:



CONCLUSION

El algoritmo tiene la restricción de que no considera la interacción entre de las variables exógenas. Para superarla, habría que contemplar todos los productos cruzados entre las variables y sus transformadas.

Dentro de las opciones de transformadas previstas, el algoritmo genera como respuesta un único modelo que, desde un punto estadístico, es el que mejor ajuste ofrece entre todos los posibles de generación por otros métodos.

Se prevee continuar con el desarrollo del algoritmo, a fin de mejorar su flexibilidad y expandir sus alcances generales.

REFERENCIAS

1. Bowker, A.H., Lieberman, G.J., "Estadistica para Ingenieros", Prentice/Hall Internacional, 1981.

2. Johnston, J., "Métodos de Econometria", Vicens Universidad, 1979

3. Ingels, R.M., "How to Use The Computer to Analyze Test Data", Chemical Engineering, August 11, 1980, págs 145 - 156.