ANALISIS DE REGRESION MULTIPLE

Dr. Porfirio Gutiérrez González

Regresión Lineal Múltiple

En muchos problemas existen dos o más variables que están relacionadas y puede ser importante modelar y explorar esta relación. Por ejemplo, el rendimiento de una reacción química puede depender de la temperatura, presión y concentración del catalizador. En este caso se requiere al menos un modelo de regresión con tres variables.

El problema general consiste en ajustar el modelo de primer orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

El problema general consiste en ajustar el modelo de primer orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

O en ajustar el modelo de segundo orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3$$

Observación	Respuesta	Regresores			
Tabla de Datos para la regresión lineal múltiple					

Se puede escribir en la siguiente forma el modelo muestral de regresión

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1,2, \ldots, n$$

La función de mínimos cuadrados es

$$L(\beta_0, \beta_1, \ \ldots, \ \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{n} \beta_j x_{ij} \right)^2$$

Se debe minimizar la función S respecto a $\beta_0, \ \beta_1, \ldots, \beta_k$. Los estimadores de $\beta_0, \ \beta_1, \ldots, \beta_k$ por mínimos cuadrados deben satisfacer

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \qquad j = 1,2, \ldots, k$$

Al simplificar la ecuación se obtienen las **ecuaciones normales de mínimos cuadrados**

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\begin{array}{cccccc}
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + & \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + & \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + & \cdots & +\hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = & \sum_{i=1}^{n} x_{i1}y_i \\
\vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + & \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + & \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + & \cdots & +\hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = & \sum_{i=1}^{n} x_{ik}y_i
\end{array}$$

Nótese que hay $p = k + 1$ ecuaciones normales, una para cada uno de los coeficientes desconocidos de regresión. La solución de las ecuaciones normales serán los estimadores por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$.

$$y = X\beta + \varepsilon$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{L} = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon'\varepsilon = \left(y - X\beta\right)'\left(y - X\beta\right)$$

$$L = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta$$

$$\mathrm{L} = y'y - 2\beta'X'y + \beta'X'X\beta$$

Ya que $\beta'X'y$ es una matriz (1 x 1), o un escalar, y su transpuesta $(\beta'X'y)' = y'X\beta$ es el mismo escalar. Los estimadores de mínimos cuadrados deben satisfacer

$$\left.\frac{\partial L}{\partial \beta}\right|_{\hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0$$

cuya simplificación es

$$X'X\hat{\beta} = X\acute{y}$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

Estimación de la varianza de regresión

Por lo general también es necesaria la varianza de regresión σ^2. Para encontrar un estimador de σ^2, considérese la suma de cuadrados de los residuales.

$$SS_E = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \qquad e = y - \hat{y}$$

$$SS_E = \sum_{i=1}^{n} e_i^2$$

$$SS_E = e'e$$

Al sustituir $e = y - \hat{y} = y - X\hat{\beta}$, se tiene

$$SS_E = \left(y - X\hat{\beta}\right)\left(y - X\hat{\beta}\right)$$

$$SS_E = y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

$$SS_E = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$$

Puesto que $X'X\hat{\beta} = X'y$

$$SS_E = y'y - \hat{\beta}'X'y$$

Se le llama la suma de cuadrados residuales o del error, y tienen $n - p$ grados de libertad asociados con ella. Puede demostrarse que

$$E(SS_E) = \sigma^2(n - p)$$

Por lo que un estimador insesgado de σ^2 esta dado por

$$\sigma^2 = \frac{SS_E}{n-p}$$

Prueba de Hipótesis en la regresión múltiple

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_a: \beta_j \neq 0 \quad \text{para al menos una j}$$

El rechazo de H_0 implica que al menos uno de los regresores $x_1, x_2, \ldots, x_k$ contribuye de manera significativa al modelo. El procedimiento de prueba incluye un análisis de varianza en el que se hace la partición de la suma de cuadrados total SS_T en una suma de cuadrados debida al modelo (o a la regresión) y una suma de cuadrados debida a los residuales (o al error) es decir

$$SS_T = SS_R + SS_E$$

Ahora bien, si la hipótesis nula $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ es verdadera, entonces SS_R/σ^2 se distribuye como una chi-cuadrada X_k^2, donde el número de grados de libertad para X^2 es igual al número de regresores del modelo *k*. Asimismo, puede demostrarse que SS_E/σ^2 se distribuye como X_{n-k-1}^2 y que SS_E y SS_R son independientes. El procedimiento de prueba para $H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ consiste en calcular

$$F_0 = \frac{\frac{SS_R}{k}}{\frac{SS_E}{(n-k-1)}} = \frac{MS_R}{MS_E}$$

Y en rechazar H_0 si F_0 excede a $F_{a,\,k,\,n-k-1}$.

De manera alternativa, podría usarse el enfoque del valor de P para la prueba de hipótesis y, por lo tanto, rechazar H_0 si el valor de P del estadístico F_0 es menor que α. Por lo general la prueba se resume en una tabla de análisis de varianza como la siguiente tabla:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	
Regresión				
Error o residual				
Total				

coeficientes de determinación o R cuadrada

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

En los modelos de regresión la medida R^2 es una medida de la cantidad de reducción en la variabilidad de $\boldsymbol{y}$ que se obtiene al utilizar las variables de regresión $x_1,\ x_2,\ \ldots,\ x_k$ en el modelo.

Un valor grande de R^2 no implica necesariamente que el modelo de regresión sea adecuado.

Siempre que se agregue una variable al modelo, el R^2 se incrementará, independientemente de que la variable adicional sea estadísticamente

R cuadrada ajustada

Puesto que R^2 siempre se incrementa cuando se agregan términos al modelo, algunos constructores de modelos de regresión prefieren usar el estadístico R^2 ajustada definido como

$$R^2_{ajustada} = 1 - \left(\frac{\frac{SS_E}{(n-p)}}{\frac{SS_T}{(n-1)}} \right) = 1 - \left(\frac{n-1}{n-p} \right)(1 - R^2 \right)$$

El estadístico R^2 ajustada no siempre se incrementará cuando se agreguen variables al modelo. De hecho, si se agregan términos innecesarios, el valor de $R^2_{ajustada}$ se disminuye con frecuencia.

Pruebas de los coeficientes de regresión individuales

Las hipótesis para probar la significación de cualquier coeficiente de regresión individual, por ejemplo β_j, son

$H_0: \beta_j = 0$

$H_a: \beta_j \neq 0$

Si $H_0: \beta_j = 0$ no se rechaza, entonces esto indica que x_j puede eliminarse del modelo. El estadístico de prueba para esta hipótesis es

$$t_0 = \frac{\hat{\beta}_j}{\sqrt{\sigma^2 C_{jj}}}$$

donde $C_{jj} = (X'X)^{-1}$ $H_0: \beta_j = 0$ **se rechaza si** $|t_0| > t_{\frac{\alpha}{2},\, n-k-1}$

Métodos de regresión por selección

Pueden clasificarse en tres categorías principales: 1) **selección hacia adelante**, 2) **eliminación hacia atrás**, y 3) **regresión por segmentos**, que es una combinación muy usada de los procedimientos 1 y 2.

Selección hacia adelante

Este procedimiento comienza con la hipótesis que no hay regresores en el modelo, además de la ordenada al origen. Se trata de determinar un subconjunto óptimo insertando regresores, uno por uno, en el modelo. El primer regresor que se selecciona para entrar en la ecuación es el que tenga la máxima correlación simple con la variable de respuesta *y*.

Supóngase que ese regresor es x_1, éste también es el regresor que producirá el máximo valor de la estadística F en la prueba de significancia de la regresión. El regresor se introduce si la estadística F es mayor que un valor predeterminado de F, por ejemplo $F_{INICIAL}$ (o F para quien entra). El segundo regresor que se escoge para entrar es el que ahora tenga la máxima correlación con y, después de ajustar y por el efecto del primer regresor que se introdujo x_1.

El procedimiento termina cuando la estadística parcial F en determinado paso no es mayor que $F_{INICIAL}$, o cuándo se ha agregado el último regresor candidato al modelo

Eliminación hacia atrás

En la **eliminación hacia atrás** se comienza con un modelo que incluya todos los K regresores, a continuación se calcula la estadística parcial F para cada regresor, como si fuera la última variable que entró al modelo.

La mínima de estas estadísticas parciales F se compara con un valor preseleccionado, F_{SAL} o F_{OUT} (es decir, F que sale), por ejemplo, y si el valor mínimo de F parcial es menor que F_{OUT}, se quita ese regresor del modelo, ahora se ajusta un modelo de regresión con $K - 1$ regresores, se calculan las estadísticas F parciales para ese nuevo modelo, y se repite el procedimiento.

El algoritmo de eliminación en reversa termina cuando el valor mínimo de F parcial no es menor que F_{OUT}, el valor preseleccionado de corte.

Regresión por segmentos

Un procedimiento muy útil es el algoritmo de regresión por segmentos, de Efroymson [1960].

La regresión por segmentos es una modificación de la selección hacia adelante, en la que cada paso se reevalúan todos los regresores que habían entrado antes al modelo, mediante sus estadísticas parciales *F*.

Un regresor agregado en una etapa anterior puede volverse redundante, debido a las relaciones entre él y los regresores que ya estén en la ecuación. Si la estadística parcial *F* de una variable es menor que F_{OUT}, esa variable se elimina del modelo.

En la regresión por segmentos, se requieren dos valores de corte, F_{IN} y F_{OUT}, algunos analistas prefieren definir $F_{IN} = F_{OUT}$, aunque eso no es necesario, con frecuencia se opta por $F_{IN} > F_{OUT}$, con lo que se hace algo más difícil agregar un regresor que eliminar uno.

Ejemplo,

Un ingeniero químico se encuentra investigando el rendimiento de un proceso, del cual le interesan tres variables: temperatura, presión y concentración porcentual. Cada variable puede estudiarse a dos niveles, bajo y alto, y el ingeniero decide correr un diseño 2^3 con estas tres variables. El experimento y los rendimientos resultantes se muestran en la siguiente tabla,

X1=TEMPERATURA	X2=PRESION	X3=CONCENTRACION	Y=RENDIMIENTO
50	100	10	32
50	100	20	36
50	200	10	57
100	100	10	46
100	200	10	65
50	200	20	57
100	100	20	48
100	200	20	68

ANALISIS CON TODAS LAS VARIABLES

Parámetro	Estimación	Error Estándar	Estadístico T	Valor-P
CONSTANTE	-11.125	10.2888	-1.08127	0.4752
TEMPERATURA	0.315	0.108972	2.89064	0.212
PRESION	0.2875	0.0544862	5.27656	0.1192
CONCENTRACION	0.375	0.544862	0.688247	0.6162
TEMPERATURA*PRESION	-0.0007	0.0005	-1.4	0.3949
TEMPERATURA*CONCENTRACION	0.001	0.005	0.2	0.8743
PRESION*CONCENTRACION	-0.0015	0.0025	-0.6	0.656

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	1173.75	6	195.625	62.6	0.0952
Residuo	3.125	1	3.125		
Total (Corr.)	1176.88	7			

R-cuadrada = 99.7345 porciento
R-cuadrado (ajustado para g.l.) = 98.1413 porciento

METODO SELECCIÓN HACIA ADELANTE

Parámetro	Estimación	Error Estándar	Estadístico T	Valor-P
CONSTANTE	2.375	3.13	0.758787	0.4822
TEMPERATURA	0.225	0.0287228	7.83349	0.0005
PRESION	0.2125	0.0143614	14.7966	0.0000

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	1156.25	2	578.125	140.15	0.0000
Residuo	20.625	5	4.125		
Total (Corr.)	1176.88	7			

R-cuadrada = 98.2475 porciento

R-cuadrado (ajustado para g.l.) = 97.5465 porciento

METODO SELECCIÓN HACIA ATRAS

Parámetro	Estimación	Error Estándar	Estadístico T	Valor-P
CONSTANTE	-8.875	4.45755	-1.991	0.1405
TEMPERATURA	0.33	0.0540062	6.11041	0.0088
PRESION	0.265	0.0270031	9.81369	0.0022
CONCENTRACION	0.225	0.0853913	2.63493	0.078
TEMPERATURA*PRESION	-0.0007	0.00034157	-2.04939	0.1328

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	1172.5	4	293.125	201	0.0006
Residuo	4.375	3	1.45833		
Total (Corr.)	1176.88	7			

R-cuadrada = 99.6283 porciento
R-cuadrado (ajustado para g.l.) = 99.1326 porciento