Análisis de Regresión Lineal Múltiple

ANÁLISIS DE REGRESIÓN

Métodos Estadísticos

MET. Alejandra Cerda

Regresión Lineal Múltiple

Un modelo de regresión donde interviene más de una variable regresora, supongamos k, se llama modelo de regresión múltiple; un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos $\beta_0, \beta_1, \dots, \beta_k$

En general, se puede relacionar la respuesta y con los k regresores, o variables predictivas bajo el modelo:

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

Donde los parámetros $\beta_0, \beta_1, \dots, \beta_k$ se llaman coeficientes de regresión. Este modelo describe un hiperplano en el espacio de k dimensiones de las variables regresoras $x_0, x_1, \dots, x_k$.

Otros modelos con estructura más complicada que se pueden analizar con técnicas de regresión lineal múltiple son por ejemplo un modelo polinómico o modelos que incluyan efectos de interacción.

Datos para regresión lineal múltiple

Supongamos que se dispone de n>k observaciones, con y_i la i-ésima respuesta observada, y x_{ij} la i-ésima observación o nivel del regresor x_j, con j=0,1,…, k. Suponiendo además que el termino de error del modelo tiene median cero, varianza constante y que los errores no están correlacionados.

observación	respuesta	Regresores			
i	y	x_1	x_2	…	x_k
1	y_1	x_{11}	x_{12}	…	x_{1k}
2	y_2	x_{21}	x_{22}	…	x_{2k}
3	y_3	x_{31}	x_{32}	…	x_{3k}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
n	y_n	x_{n1}	x_{n2}	…	x_{nk}

Estimación por mínimos cuadrados

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

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i$

De tal manera que la función de mínimos cuadrados es

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

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

$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij}) = 0$

$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij})\, x_{ij} = 0$

Ecuaciones normales

Al simplificar $\frac{\partial S}{\partial \beta_0}$ se obtienen las ecuaciones normales de mínimos cuadrados

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

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

$$\vdots$$

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

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 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

Notación matricial del modelo

Es más cómodo manejar modelos de regresión múltiple cuando se expresan de forma matricial. La notación matricial del modelo es $y = X\beta + e$ en donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Las ecuaciones normales de mínimos cuadrados quedan dadas por $X'X\hat{\beta} = X'y$, para resolverlas se multiplica ambos lados por la inversa de X´X. Así el estimador de mínimos cuadrados es $\boldsymbol{\hat{\beta} = (X'X)^{-1}X'y}$ siempre y cuando exista la matriz inversa $(X'X)^{-1}$; es decir, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Multicolinealidad

REGRESIÓN LINEAL MÚTIPLE

Métodos Estadísticos

MET. Alejandra Cerda

¿Qué es la multicolinealidad?

Un problema serio que puede influir mucho sobre la utilidad de un modelo de regresión es la multicolinealidad, o dependencia casi lineal entre las variables de regresión. La multicolinealidad implica una dependencia casi lineal entre los regresores, los cuales son las columnas de la matriz X, por lo que es claro que una dependencia lineal exacta causaría una matriz X´X singular.

Los elementos de la diagonal principal en la inversa de la matriz X´X en forma de correlación se llaman con frecuencia factores de inflación de varianza (VIF, de *Variance Inflation Factors*), y son un diagnóstico importante de la multicolinealidad.

¿Cómo medir la multicolinealidad?

Se puede demostrar que, en general, el factor de inflación de varianza para el j-ésimo coeficiente de regresión se puede escribir como sigue:

$$VIF_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación múltiple obtenido haciendo la regresión x_j sobre las demás variables regresoras. Si x_j depende casi linealmente de alguno(s) de los demás regresores, entonces R_j^2 será casi la unidad, y VIF_j será grande.

Los factores VIF mayores que 10 implican problemas graves de multicolinealidad.

Prueba de Hipótesis

REGRESIÓN LINEAL MÚLTIPLE

Métodos Estadísticos

MET. Alejandra Cerda

Prueba de significancia

La prueba de significancia de la regresión es para determinar si hay una relación lineal entre la respuesta y y cualquiera de las variables regresoras. Las hipótesis correspondientes son:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ (la regresión no es significativa, las variables involucradas no muestran relación)

$H_a: \beta_j \neq 0 \; al \; menos \; para \; una \; j$ (**al menos un regresor** contribuye de forma significativa)

El procedimiento de prueba es una generalización del análisis de varianza (ANOVA) que se usó en la regresión lineal simple; rechazando H_0 si $EP: F_0 > F_{\alpha,k,n-k-1}$

TABLA ANOVA

Fuente de Variación	Grados de libertad	Suma de Cuadrados	Cuadrado medio	F0
Regresión	k	$\hat{\beta}'X'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$	MSR=SSR/k	MSR/MSE
Residuales	n-k-1	$y'y - \hat{\beta}'X'y$	MSE=SSE/(n-k-1)	
Total	n-1	$y'y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$		

R^2 y R^2 ajustada

Otras dos maneras de evaluar la adecuación del modelo son los estadísticos R^2 y R^2 ajustada. En general, el valor de R^2 aumenta siempre, cuando se agrega un regresor al modelo, independientemente del valor de la contribución de esa variable. En consecuencia, es difícil juzgar si un aumento de R^2 dice en realidad algo importante.

Algunas personas que trabajan con modelos de regresión prefieren usar el estadístico R^2 ajustada definido como:

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SST/(n-1)}$$

En vista de que el numerador es el cuadrado medio de residuales y el denominador es constante, independientemente de cuántas variables hay en el modelo, R^2 ajustada sólo aumentará al agregar una variable al modelo si esa adición reduce el cuadrado medio residual . R^2 ajustada penaliza la adición de términos que no son útiles, además que es ventajoso para evaluar y comparar los posibles modelos de regresión.

Prueba sobre coeficientes individuales de regresión

Una vez determinado que al menos uno de los regresores es importante, la pregunta es ¿cuál(es) sirve(n) de ellos? Las hipótesis para probar la significancia de cualquier **coeficiente individual** de regresión, β_j, son:

$H_0: \beta_j = 0$ (el regresor x_j no influye significativamente en el modelo)

$H_a: \beta_j \neq 0$ (el regresor x_j influye significativamente en el modelo)

El estadístico de prueba para esta hipótesis es: $t_0 = \dfrac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$

donde C_{jj} es el elemento de la diagonal de $(X'X)^{-1}$ que corresponde a $\hat{\beta}_j$

Rechazando la hipótesis nula $H_0: \beta_j = 0$ si $|t_0| > t_{\alpha/2, n-k-1}$, sino se rechaza la hipótesis nula quiere decir que se puede eliminar el regresor del modelo

INTERVALOS DE CONFIANZA

Para construir estimados de intervalo de confianza de los coeficientes de regresión β_j, se continúa suponiendo que los errores están distribuidos normal e independientemente, con promedio cero y varianza constante. Como el estimador de mínimos cuadrados es combinación lineal de las observaciones , esta distribuido normalmente. Esto implica que cada que la distribución marginal de cualquier coeficiente de regresión $\hat{\beta}_j$ es normal con media β_j y varianza $\sigma^2 C_{jj}$ donde C_{jj} es el elemento de la diagonal de $(X'X)^{-1}$. En consecuencia, cada uno de los estadísticos $\frac{\beta_j - \hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}, j = 0,1, \ldots, k$ se distribuye como t, con n-p grados de libertad.

Así el intervalo de confianza del $100(1-\alpha)\%$ para el coeficiente de regresión β_j queda definido como:

$$\hat{\beta}_j - t_{\alpha/2, n-p}\sqrt{\hat{\sigma}^2 C_{jj}} < \beta_j < \hat{\beta}_j + t_{\alpha/2, n-p}\sqrt{\hat{\sigma}^2 C_{jj}}$$