Unidad II

ANÁLISIS DE REGRESIÓN LINEAL MÚLTIPLE

"Básicamente todos los modelos son erróneos, pero algunos son muy útiles"
George Box

1. Introducción

Cundo se desea explicar una variable respuesta es poco común que se utilice solo una variable predictora. Por ejemplo, si se desea el peso (en kg) de una persona en estado de crecimiento, esta puede ser explicado por su estatura (en cm), la cantidad de calorías que consume al día, el género, el tiempo (en minutos) de ejercicios que realiza al día, el peso medio (en kg) de los padres entre otras.
También nos debemos preguntar si todas las variables predictoras propuestas serán utilizadas para explicar a la variable respuesta o debemos utilizar solo algunas de ellas es decir las que son significativas

George Box (1919-2013)

El análisis de regresión lineal múltiple nos permite explicar a una variable respuesta (Y) que debe ser de tipo cuantitativa continua mediante una o muchas variables predictoras (Xs) que pueden ser cuantitativas o cualitativas

2. Modelo de Regresión Múltiple

En general, se puede relacionar la variable respuesta *y* con k regresores, o variables predictoras. El modelo

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

se llama modelo de regresión lineal múltiple con k regresores. Los parámetros β_j, j = 0, 1, ... , k se llaman coeficientes de regresión. Este modelo describe a un hiperplano en el espacio de k dimensiones de las variables regresoras X_j El parámetro β_j representa el cambio esperado en la respuesta *y* por. cambio unitario en X_j cuando todas las demás variables regresoras X_j ($i \neq j$) se mantienen constantes.

3. Estimación del modelo

Se puede aplicar el método de mínimos cuadrados para estimar los coeficientes de regresión de la ecuación. Supongamos que se dispone de n>k observaciones, y sea y_i la i-ésima respuesta observada, y x_{ij} la i-ésima observación o nivel del regresor X_j

Datos para la regresión lineal múltiple

Observación	**Respuesta**	**Regresores**			
$\boldsymbol{i}$	$\boldsymbol{y}$	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$	**...**	$\boldsymbol{x_k}$
1	y_1	x_{11}	x_{12}	...	x_{1k}
2	y_2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮	⋮		⋮
n	y_n	x_{n1}	x_{n2}	...	x_{nk}

Se supondrá que las variables regresoras X_1, X_2, ... , X_k son fijas, es decir, que son matemáticas o no aleatorias, y que se miden sin error. Sin embargo, todos los resultados obtenidos siguen siendo válidos para el caso en el que los regresores son variables aleatorias. Esto es realmente importante, porque cuando se toman datos de regresión en un estudio observacional, algunos o la mayor parte de los regresores son variables aleatorias. Cuando los datos son el resultado de un experimento diseñado es más probable que las X sean variables fijas.

Se puede escribir la ecuación

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

De la siguiente manera

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

La función de mínimos cuadrados es

$$S(\beta_0, \quad \beta_1, \quad \ldots, \beta_k) = \sum_{i=1}^{n} \varepsilon_i^2$$

$$= \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2$$

Se debe minimizar la función S respecto a β_0, β_1, ... , β_k. Los estimadores de β_0, β_1, ... , βk por mínimos cuadrados deben satisfacer

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

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

Al simplificar las ecuaciones anteriores se obtienen las ecuaciones normales de mínimos cuadrados

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

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

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

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

Hay p = k + 1ecuaciones normales, una para cada uno de los coeficientes des conocidos de regresión. La solución de las ecuaciones normales serán los estimadores por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1 \ldots, \hat{\beta}_k$.
La notación matricial del modelo es:

$$y = X\beta_0 + \varepsilon$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Se desea determinar el vector $\hat{\beta}$ de estimadores de mínimos cuadrados que minimice

$$S(\beta) = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (Y - X\beta)'(Y - X\beta)$$

$S(\beta)$ se puede expresar de la siguiente manera:

$$\begin{aligned} S(\beta) &= Y'Y - \beta'X' Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \end{aligned}$$

Los estimadores de mínimos cuadrados deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

Que se simplifica a

$$X'X\hat{\beta} = X'Y$$

Las ecuaciones anteriores son las ecuaciones normales de mínimos cuadrados. Por lo tanto, el estimador de β por mínimos cuadrados es

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

La matriz $(X'X)^{-1}$ existe si los regresores son linealmente independientes esto es, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Al escribir las ecuaciones normales con detalle se obtiene

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

X'X es una matriz simétrica de p x p, Y que X'Y es un vector columna de p X 1. Los elementos diagonales de X'X son las sumas de los cuadrados de los elementos en las columnas de X, y los elementos fuera de la diagonal son las sumas de los productos cruzados de los elementos de las columnas de X. Además, nótese que los elementos de X'Y son las sumas de los productos cruzados de las columnas de X por las observaciones y_i.

El modelo ajustado de regresión es:

$$\hat{y} = \mathrm{x}'\hat{\beta} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_j$$

El vector de valores ajustados $\hat{y}_i$ que corresponden a los valores observados y_i es:

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

La matriz $H = X(X'X)^{-1}X'$ de dimensión nxn se suele llamar matriz de sombrero.

La diferencia entre un valor y_i y el valor ajustados $\hat{y}_i$ correspondiente es el residual.
Los n residuales se pueden escribir cómodamente con notación matricial como sigue:

$$\mathrm{e} = \mathrm{y} - \hat{y}$$

Hay otras formas de expresar los residuales como:

$$\mathrm{e} = \mathrm{Y} - \mathrm{X}\hat{\beta} = \mathrm{Y} - \mathrm{HY} = (\mathrm{I} - \mathrm{H})\mathrm{Y}$$

3.1 Propiedades de los estimadores de mínimos cuadrados

$\hat{\beta}$ es un estimador insesgado de β

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{y}] = \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'(\mathrm{X}\beta + \varepsilon)] \\ &= \mathrm{E}[(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{X}\beta + (\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\varepsilon] = \beta\end{aligned}$$

La propiedad de varianza de $\hat{\beta}$ se expresa con la matriz de covarianza

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{E}\{[\hat{\beta} - \mathrm{E}(\hat{\beta})][\hat{\beta} - \mathrm{E}(\hat{\beta})]'\}$$

Que es una matriz simétrica pxp cuyo j-ésimo elemento diagonal es la varianza de $\hat{\beta}_j$ y cuyo (ij)-ésimo element fuera de la diagonal es la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$. La covarianza de la matriz de $\hat{\beta}$ es:

$$\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathrm{X}'\mathrm{X})^{-1}$$

Por lo tanto, si hacemos $C=(X'X)^{-1}$, la varianza de $\hat{\beta}_j$ es $\sigma^2 C_{jj}$ y la covarianza entre $\hat{\beta}_i$ y $\hat{\beta}_j$ es $\sigma^2 C_{ij}$.

3.2 Estimación de σ^2

Como en la regresión lineal simple, se puede desarrollar un estimador para σ^2 a partir de la suma de cuadrados de residuales

$$SS_{Res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

Se sustituye e= Y - Xβ̂, se obtiene

$$SC_{Res} = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$
$$= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$
$$= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}$$

Como $X'X\hat{\beta} = X'Y$, entonces

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

El Cuadrado Medio de Residuos es

$$CMRes = \frac{SCRes}{n - p}$$

Por lo tanto, un estimador insesgado de σ^2 es:

$$\hat{\sigma} = CM_{Res}$$

4. Prueba de Hipótesis en la Regresión Lineal Múltiple

La prueba de la significancia de la regresión es para determinar si hay una relación lineal entre la variable respuesta *y* y al menos una de las variables regresoras X_1, X_2, ... , X_k Este procedimiento suele considerarse como una prueba general o global de la adecuación del modelo. Las hipótesis son:

$H_0: \beta_1 = \beta_1 = \cdots = \beta_k = 0$
$H_0: \beta_1 \neq 0 \; al \; menos \; para \; una \; j$

El procedimiento de prueba es una generalización del análisis de varianza que se usó en la regresión lineal simple. La suma total de cuadrados SCT se divide en una suma de cuadrados debido a la regresión, SCReg y a una suma de cuadrados de residuales, SCRes. Así,

SCT = SCReg + SSRe

Para evaluar las hipótesis propuestas se puede hacer uso del estadístico

$$F_0 = \frac{SS_R / k}{SS_{Res} / (n - k - 1)} = \frac{MS_R}{MS_{Res}}$$

Se rechaza H_0 si, $F_0 > F_{\alpha,k,n-k-1}$

El procedimiento de prueba se resume normalmente en una tabla de análisis de varianza
Se parte de que

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

Y dado que

$$SCTo = SCReg + SCRes$$

Y ya que

$$SC\text{To} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \text{Y'Y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

entonces

$$SC\text{Reg} = \hat{\beta}' X' y - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

Análisis de varianza para determinar el significado en la regresión múltiple

Fuente de variación	**Suma de Cuadrados**	**Grados de libertad**	**Cuadrado medio**	$\boldsymbol{F_0}$
Regresión	SS_R	k	MS_R	MS_R/MS_{Res}
Residuales	SS_{Res}	n - k - 1	MS_{Res}	
Total	SS_T	n - 1		