CAPÍTULO 2
REGRESIÓN LINEAL MULTIPLE

Edgar Acuña Fernández

Departamento de Matemáticas
Universidad de Puerto Rico
Recinto Universitario de Mayagüez

REGRESIÓN LINEAL MULTIPLE

- La regresión lineal multiple trata de explicar el comportamiento de Y con más de una variable predictora usando una funcion lineal.

Alternativas para mejorar el modelo.

- Transformar la variable predictora, o la variable de respuesta Y, o ambas y usar luego un modelo lineal.
- Usar regresión polinómica con una variable predictora.
- Conseguir más variables predictoras y usar una regresión lineal múltiple.

2.2 El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones está dado por:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots . \beta_p x_{ip} + e_i \quad \text{para i = 1,2,…,n}$$

en forma matricial : $\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ .. & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Suposiciones del modelo

1. $E(\mathbf{e})=\mathbf{0}$
2. $Var(\mathbf{e})=\sigma^2\mathbf{I_n}$

Donde:

- **e** es un vector columna aleatorio de dimensión n.
- $\mathbf{I_n}$ es la matriz identidad de orden n.

2.2.1 Estimación del vector de parámetros β por Cuadrados Mínimos

Se tiene que minimizar la suma de cuadrados de los errores.

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \mathbf{e'e} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Haciendo operaciones con los vectores y matrices

$$Q(\boldsymbol{\beta}) = \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta} = \mathbf{Y'Y} - 2\boldsymbol{\beta}'\mathbf{X'Y} + \boldsymbol{\beta}'\mathbf{X'X}\boldsymbol{\beta}$$

Derivando Q con respecto a β e igualando a cero se obtiene el sistema de ecuaciones normales

$$\mathbf{X'X}\boldsymbol{\beta} = \mathbf{X'Y}$$

resolviendo para β se obtiene:

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

2.2.2 Propiedades del estimador

- $\hat{\boldsymbol{\beta}}$ es insesgado, o sea $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- $Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1}$

- Si no se asume normalidad, el estimador minimo-cuadrático $\hat{\boldsymbol{\beta}}$ es el mejor estimador dentro de los estimadores lineales insesgados de $\boldsymbol{\beta}$.
- Si se asume normalidad de los errores entonces $\hat{\boldsymbol{\beta}}$ es el mejor estimador entre todos los estimadores insesgados de $\boldsymbol{\beta}$

2.2.3 Estimación de la varianza σ^2

Un estimado de la varianza de los errores es:

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} = \frac{\sum_{i=1}^{n} \hat{e}_i^2}{n-p-1} = \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p-1} = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1}$$

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

$$SSE = \mathbf{Y}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

Donde $\mathbf{H}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es la ***"Hat Matrix"***

la varianza estimada de los errores puede ser escrita como: $\hat{\sigma}^2 = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}}{n-p-1} = \frac{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}{n-p-1}$

Algunas Propiedades

Sea Y un vector aleatorio n-dimensional tal que $E(Y) = \mu$ y $VAR(Y) = V$ entonces

$$E(Y'AY) = traza(AV) + \mu'A\mu$$

Donde

$\mu = X\beta$ y $V = \sigma^2 I_n$

Se puede mostrar que $E[s^2] = \sigma^2$.

2.3. Inferencia en Regresión lineal múltiple

Involucra realizar

- pruebas de hipótesis eintervalos de confianza

acerca de los coeficientes del modelo de regresión poblacional.

- Intervalos de confianza de las predicciones que

se hacen con el modelo.

Suponemos que $\mathbf{e \sim NI(0, \sigma^2 I_n)}$ o equivalente que

$\mathbf{Y \sim NI(X\beta, \sigma^2 I_n)}$

Descomposición de la variación total de Y

La variación total de Y se descompone en dos variaciones: una debido a la regresión y otra debido a causas no controlables.

$$\boldsymbol{SST = SSR + SSE}$$

El coeficiente de Determinación R^2, se cálcula por:

$$R^2 = \frac{SSR}{SST}$$

Resultados para sumas de cuadrados

i) $\frac{SST}{\sigma^2} \sim \chi^2_{(n-1)}$

ii) $\frac{SSE}{\sigma^2} \sim \chi^2_{(n-p-1)}$, también que $\frac{(n-p-1)s^2}{\sigma^2} \sim \chi^2_{(n-p-1)}$

iii) $\frac{SSR}{\sigma^2} \sim \chi^2_{(p)}$

2.3.1 Prueba de hipótesis acerca de un coeficiente de regresión individual

- Ho: $\beta_i = 0$ (i=1,2,..,p),
- Ha: $\beta_i \neq 0$;

La prueba estadística es la prueba de t:

$$t = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{s\sqrt{C_{ii}}}$$ se distribuye como una tcon (n-p-1) gl.

Donde, C_{ii} es el i-ésimo elemento de la diagonal de $\mathbf{(X'X)^{-1}}$.

Los programas de computadoras, da el "P-value" de la prueba t.

2.3.2 Prueba de Hipótesis de que todos los coeficientes de regresión son ceros.

H_o: $\beta_1=\beta_2=\ldots=\beta_p=0$

H_a: Al menos uno de los coeficientes es distinto de cero.

usando propiedades de formas cuadráticas se puede mostrar que:

$E(SSR) = E[\mathbf{Y'(H-11'/n)Y}]$

$= p\sigma^2 + \boldsymbol{\beta X'(H-11'/n)X\beta}$

$= p\sigma^2 + \boldsymbol{\beta' X'(H-11'/n)X\beta}$

Donde, 1 es un vector columna de n unos.

Tabla de Análisis de Varianza

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados Medios	F
Regresión	SSR	p	MSR=SSR/p	MSR/MSE
Error	SSE	n-p-1	MSE=SSE/n-p-1	
Total	SST	n-1		

Particionamiento secuencial de la suma de cuadrados de regresión

La **suma de cuadrados de regresión** puede ser particionada en tantas partes como variables predictoras existen en el modelo. Sirve para **determinar la contribución de cada una** de las **variables** predictoras al comportamiento de Y.

$$SSR(\beta_1,\beta_2,,\ldots.\beta_p/\beta_0) = SSR(\beta_1/\beta_0) + SSR((\beta_2,/\beta_1,\beta_0) + \ldots + SSR(\beta_p/\beta_p\text{-}1,\ldots,\beta_1,\beta_0)$$

$SSR(\beta_k/\beta_{k-1},,\ldots.\beta_1,\beta_0)$ significa el **incremento en la suma de cudrados de regresión** cuando la variable $\mathbf{X_k}$ **es incluida** en el modelo, el cual ya contiene las variables predictivas $X_1,\ldots X_{k-1}$

2.3.3 Prueba de hipótesis para un subconjunto de coeficientes de regresión

Ho: $\beta_1=\ldots=\beta_k=0$. (Los k primeros coeficientes son ceros).

Ha: Al menos uno de los k primeros coeficientes no es cero.

La prueba de F parcial se calcula por:

$$Fp = \frac{\frac{SSR(C)-SSR(R)}{k}}{\frac{SSE(C)}{n-p-1}} = \frac{\frac{SSR(C)-SSR(R)}{k}}{MSE(C)}$$

k gl para el numerador y

n-p-1 gl para el denominador

Donde:

SSR(C) = SSR($\beta_1,\beta_2,\ldots.\beta_p/\beta_o$) y SSR(R) = SSR($\beta_{k+1},\beta_{k+2},\ldots,\beta_p/\beta_o$)

SSR(C) – SSR(R)=SSR($\beta_1,\beta_2,\ldots.\beta_k/\beta_{k+1},\beta_{k+2},\ldots.\beta_p$)

2.3.4 Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Se desea predecir el valor medio de la variable de respuesta Y para una combinación predeterminada de las variables predictoras X1,……Xp.

Consideremos el vector de valores observados $\mathbf{x}'_o = (1, x_{1,0}, \ldots . x_{p,0})$

El valor predicho para el valor medio de la variable de respuesta Y será $\hat{y}_o = \mathbf{x}'_\mathbf{o}\,\hat{\boldsymbol{\beta}}$ y $Var(\hat{y}_o) = \mathbf{x}'_\mathbf{o}\,\mathbf{Var}((\hat{\boldsymbol{\beta}})\mathbf{x}_\mathbf{o} = \sigma^2 \mathbf{x}'_\mathbf{o}\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathbf{o}$

Se asume que los errores están normalmente distribuidos.

2.3.4 Intervalos de Confianza y de Predicción en Regresión Lineal Múltiple.

Un intervalo del 100(1-α)% para el valor medio de Y dado que $\mathbf{x}=\mathbf{x'}_o$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{\mathbf{x'}_o (\mathbf{X'X})^{-1}\mathbf{x}_o}$$

Un intervalo de confianza (intervalo de predicción) del 100(1-α)% para el valor individual de Y dado $\mathbf{x}=\mathbf{x'}_o$ es de la forma

$$\hat{y}_o \pm t_{(\alpha/2,n-p-1)} s\sqrt{1+\mathbf{x'}_o (\mathbf{X'X})^{-1}\mathbf{x}_o}$$

2.3.5 La prueba de Falta de Ajuste

Se usa para determinar si la forma del modelo que se está considerando es adecuada.

En regresión múltiple se debe suponer que hay m combinaciones distintas de las n observaciones de las p variables predictoras y que por cada una de esas combinaciones hay n_i (i = 1,…,m) observaciones de la variable de respuesta, es decir, $\sum_{i=1}^{m} n_i = n$

La Suma de Cuadrados del Error

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\hat{y}_i)^2=\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2+\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i-\hat{y}_i)^2$$

Donde:

$\hat{y}_i$ es el valor predicho por el modelo de regresión para la i-ésima combinación de las variables predictoras y

$\bar{y}_i$ es el valor promedio de la variable predictora para la i-ésima combinación.

- **Suma de Cuadrados del Error Puro (SSPE)** Es la primera suma de cuadrados del lado derecho, tiene n-m *gl*.

- **Suma de Cuadrados de Falta de Ajuste (SSLOF)** Es la segunda suma de cuadrados tiene m-p-1 *gl*.

  también puede ser escrita como: $\sum_{i=1}^{m} n_i (\bar{y}_i - \hat{y}_i)^2$

Prueba de hipótesis

Ho: El modelo es adecuado (no hay falta de ajuste)

Ha: el modelo no es adecuado

La prueba estadística es una prueba de F dada por:

$$F = \frac{SSLOF/(m-p-1)}{SSPE/(n-m)} = \frac{MSLOF}{MSPE}$$

se distribuye como una F(m-p-1,n-m).