Regresión Lineal Múltiple

Formulación y estimación del modelo de regresión

Introducción

El proceso de investigación científica está constituido por fases a través de las cuales el investigador maneja tres elementos y sus relaciones:

a. El problema a resolver
b. El objeto a investigar y
c. Su representación

Modelo:

Representación de la realidad que reproduce o aproxima los aspectos más saltantes de la realidad y que por lo general se construye para ayudar a resolver un modelo.

CLASIFICACION:

Modelos cuantitativos
Modelos cualitativos***…***

Ejemplos de situaciones donde se aplica el modelo de regresión múltiple:

- Predecir el costo de una vivienda en función de sus características físicas y su ubicación.
- Explicar el rendimiento universitario, mediante variables sociales, familiares, económicas, enseñanza y motivación del profesor, materiales de estudio, tiempo que dedica para estudiar sus cursos.
- Rendimiento de gasolina, en % del crudo, en base a la gravedad específica del crudo, presión de vapor del crudo, temperatura de destilación al 10% del crudo (ºF), temperatura de final de destilación de gasolina (ºF)
- Se quiere explicar la presión arterial sanguínea media, en función de la edad, peso, área de la superficie corporal, duración de la hipertensión, pulso básico, medición del estrés.

Usos de la regresión:

- **Descripción de datos**
- **Estimación de parámetros**
- **Predicción y estimación**
- **Control**

1.2. Modelo Lineal General

El modelo lineal general surge de la necesidad de expresar en forma cuantitativa las relaciones entre un conjunto de variables, en la que una de ellas es denominada variable dependiente o de respuesta y las otras denominadas covariables, explicativas (regresoras) o independientes.
Y es una variable aleatoria cuya función de distribución de probabilidad pertenece a una familia de distribuciones de probabilidades, y es explicada por el conjunto "k" de variables regresoras las cuales son fijadas antes de conocer Y.

$$E(Y / X_1, X_2, ..., X_K) = \beta_0 + \beta_1 X_1 + ... + \beta_K X_K = \mu$$

Si se extrae una muestra aleatoria de tamaño "n" de una población donde Y y X_1, X_2,….., X_k , se relacionan linealmente, entonces cada observación de la muestra puede expresarse como:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i \quad i=1,2,\ldots\ldots,n \qquad (1)$$

Donde: $E(\varepsilon_i) = 0 \quad Var(\varepsilon_i) = \sigma^2 \quad \rho(\varepsilon_i \varepsilon_j) = 0$

Este modelo (1) puede representarse matricialmente como:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad (2)$$

Y_{nx1} vector de variable dependiente (variable respuesta)

X_{nxp} matriz de variables independientes (variables regresoras)

β_{px1} vector de parámetros desconocidos.

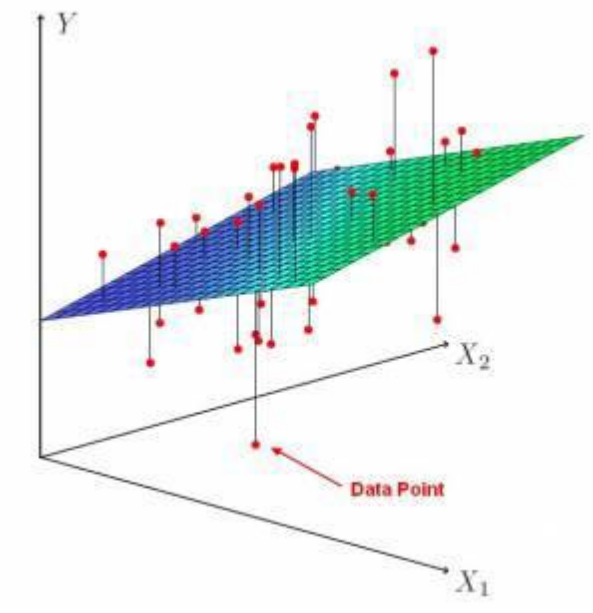
$\mathbf{X\beta}$ componente sistemática

$\boldsymbol{\varepsilon}_{nx1}$ componente aleatoria del error (perturbaciones).

Una característica distintiva del modelo lineal general, es que la **variable respuesta esta medida en escala** métrica, mientras que las variables

regresoras pueden estar medidas en escala métrica o no métrica (numéricas o categóricas); además de ser independientes entre sí.

1.2 MODELO DE REGRESIÓN LINEAL GENERAL (Modelo de Regresión Lineal Múltiple)



Sabemos que los Modelos de Regresión estudian la relación estocástica cuantitativa entre una variable de interés y un conjunto de variables explicativas.
Sea Y la variable de interés, variable respuesta o dependiente y sean $x_1,x_2,\ldots,x_k$ las variables explicativas o regresoras.
Este es un modelo similar al modelo lineal general teniendo como diferencia que en el modelo lineal general las variables independientes no son aleatorias y en el modelo de regresión lineal general las variables son aleatorias.
Cuando todas las variables regresoras son continuas el modelo (1)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k + \varepsilon_i \qquad i=1,2,\ldots\ldots,n$$

se denomina **modelo de regresión lineal múltiple**.

Modelo sin intercepto

Modelo con intercepto

Formulación del Modelo de Regresión Lineal Múltiple

$$y_i = f(x_1, x_2, ..., x_k) + \underbrace{g(x_{k+1}, x_{k+2}, ..., x_n)}$$

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon_i \qquad i=1,2,\ldots\ldots,n \quad (1)$$

$$\text{Con} \quad E(\varepsilon_i) = 0 \qquad Var(\varepsilon_i) = \sigma^2 \qquad \rho(\varepsilon_i \varepsilon_j) = 0$$

Donde:

Y_i:	v. respuesta
$X_{1,}X_{2,\ldots,}X_{Kj}$:	v. independientes
ε_{ij} :	v.a. error

Es decir, tenemos un sistema de ecuaciones donde cada una establece la relación entre la endógena y las exógenas en un momento del tiempo. Matricialmente se escribe:

1.
$$\begin{pmatrix} y_1 \\ y_2 \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & . & x_{1k} \\ x_{21} & x_{22} & . & x_{2k} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ \varepsilon_n \end{pmatrix}$$

2.
$$\begin{pmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{pmatrix} = \begin{pmatrix} 1\, x_{11} & x_{21} \ldots x_{k1} \\ 1\, x_{12} & x_{22} \ldots x_{k2} \\ \ldots & \ldots \quad .. \quad . \\ 1\, x_{1n} & x_{2n} \ldots . x_{kn} \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \ldots \\ u_n \end{pmatrix}$$

La expresión matricial del modelo de regresión múltiple es la siguiente:

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

El modelo estimado puede expresarse en forma matricial:

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

$$y$$

$$Y - \hat{Y} = e \quad residuales$$

SUPUESTOS BÁSICOS

La siguiente tabla presenta los supuestos del MRLM.

HIPÓTESIS del Modelo de Regresión Lineal General	
En base a la var. error ε_i	**En base a la var. respuesta *Y***
$E^{(\varepsilon_i)} = 0$	$E^{(Y/x_{i1}, x_{i2}, \ldots, x_{ik})} =$ $\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik}$
Homocedasticidad $Var^{(\varepsilon_i)} = \sigma^2$	Homocedasticidad $Var^{(Y/x_{i1}, x_{i2}, \ldots, x_{ik})} = \sigma^2$
Independencia, $Cov^{(\varepsilon_i; \varepsilon_j)} = 0$ los errores, ε_i, son independientes	Independencia las observaciones, y_i, son independientes
Normalidad $\varepsilon_i \in N(0, \sigma^2)$	Normalidad $Y/x_{i1}, x_{i2}, \ldots, x_{ik} \sim N^{(\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik}, \sigma^2)}$
$n > k + 1$	$n > k + 1$
Las variables regresoras son linealmente independientes	Las variables regresoras son linealmente independientes

Los errores tienen los siguientes supuestos:

1) $E(\varepsilon_i) = 0 \qquad i = 1, 2, ..., n$

2) $Var(\varepsilon_i) = \sigma^2 \qquad i = 1, 2, ..., n$

3) $\rho(\varepsilon_i \varepsilon_j) = 0 \qquad \forall \quad i \neq j$

4) $\varepsilon_i \sim N(0, \sigma^2)$ Matricialmente se tiene que: $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$

5) n>k+1

6) Ninguna de las variables regresoras es una combinación lineal exacta de las demás, es decir son linealmente independientes.

La siguiente tabla presenta los supuestos del MRLM.

1.3. ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO

Estimación de β.

Sabemos que:

$$e_i = y_i - \hat{y}_i$$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \ldots \hat{\beta}_k x_{ki})$$

Son los parámetros estimados del modelo

1.3.1. MÉTODO MÍNIMOS CUADRADOS

El estimador de mínimos cuadrados de β, denotado por βˆ, es el valor de β que minimiza

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

Por lo tanto, lo que se debe hacer es derivar la expresión anterior y buscar el valor de β que la hace igual a cero. Se puede escribir como

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

Derivando e igualando a cero

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$$

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

se obtiene

$$\mathbf{X'X}\hat{\boldsymbol{\beta}} = \mathbf{X'y}$$

Estas son las **ecuaciones normales** de mínimos cuadrados

Por lo tanto el estimador de β por mínimos cuadrados es:

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

Interpretación de los coeficientes de regresión:

1.3.2 METODO MAXIMA VEROSIMILITUD

En el modelo RLM:
El error se distribuye como una normal p variante y se distribuye normalmente….
Entonces

Si Σ es una matriz no singular entonces la distribución puede describirse por la siguiente función de densidad

Para encontrar los estimadores MV de los parámetros del modelo tenemos:

Las perturbaciones aleatorias o errores tienen distribución:

$$\varepsilon_i \sim N(0,\sigma^2) \Rightarrow f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}\varepsilon_i^2}$$

La función de verosimilitud conjunta es:

$$\prod_{i=1}^{n} f(\varepsilon_i) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\varepsilon_i^2} \quad \Rightarrow \text{como } \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y}-\mathbf{x}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) = \sum_{i=1}^{n}\varepsilon_i^2$$

$$ln\left[L\left(\mathbf{y},\mathbf{x},\boldsymbol{\beta},\sigma^2\right)\right] = -\frac{n}{2}ln(2\pi) - n\ln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{x}\boldsymbol{\beta})$$

Derivando encontramos

$$\boldsymbol{X'X\hat{\beta}} = \boldsymbol{X'y} \Rightarrow \quad \text{ecuaciones normales}$$

El estimador para el vector beta de los parámetros beta es:

$$\Rightarrow \quad \tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

1.3.3 ESTIMACIÓN DE LA VARIANZA RESIDUAL

Se sabe que algunas formas de expresión de la suma de cuadrados del error son:

1. **SCE = Y'Y - $\hat{\boldsymbol{\beta}}$X'X $\hat{\boldsymbol{\beta}}$**
2. **SCE = Y'(I-H) Y**

Se tiene que

$$SCE = \mathbf{e}'\mathbf{e} = \frac{1}{n}\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}$$

Además la SCE tiene (n-p) grados de libertad entonces: $CME = \frac{SCE}{n-p}$

También se sabe que $E(CME) = \sigma^2$

Un estimador insesgado de σ^2 es:

$$\hat{\sigma}^2 = CME = \frac{SCE}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$$

Tarea:

1.3.4 MINIMOS CUADRADOS GENERALIZADOS

Se considera la situación más general cuando:

La matriz de covarianzas del vector **ε** es **V** y no $\sigma^2 I$.

Supongamos que: Cov(**ε**) = V donde V es una matriz simétrica y definida positiva.

Lo que buscamos es minimizar **ε´V ε**, es decir:

$$S = SCE = (Y-X\beta)' V^{-1} (Y-X\beta)$$

$$SCE = Y'V^{-1}Y - Y'V^{-1}\beta - \beta'X'V^{-1}Y + \beta'X'V^{-1}X\beta$$

$$SCE = Y'V^{-1}Y - 2\beta'X'V^{-1}Y + \beta'X'V^{-1}X\beta$$

$$\left.\frac{\partial S}{\partial \boldsymbol{\beta}}\right|_{\hat{\boldsymbol{\beta}}} = \mathbf{0}$$

$$-2X'V^{-1}Y + 2X'V^{-1}X\beta = 0$$

$$-X'V^{-1}Y + X'V^{-1}X\beta = 0$$

$$X'V^{-1}X\beta = X'V^{-1}Y$$

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

1.3.5 TEOREMA GAUSS-MARKOV

TEOREMA: El Mejor Estimador Lineal Insesgado de varianza mínima para β en el modelo lineal $\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \qquad \text{es el MELI}$$

Tarea: Demostración

Propiedades estadísticas de los estimadores M.C.O.

1. $E(Y) = X\beta$
2. $V(Y) = V(X\beta+\varepsilon) = \sigma^2 I$
3. $E(\hat{\beta}) = \beta$ es insesgado
4. $Cov(\hat{\beta}) = V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$
5. $V(\hat{\beta}_j) = \sigma^2 C_{jj}$
6. Sea $\hat{Y} = X\hat{\beta}$ entonces $E(\hat{Y}) = XE(\hat{\beta}) = X\beta$
7. $V(\hat{Y}) = \sigma^2 X(X'X)^{-1}X' = \sigma^2 H$
8. H matriz hat, sombrero, proyección de rango p y además idempotente.
9. Sea $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X' Y = HY$

   $\hat{Y} = HY$ vector proyección

Propiedades algebraicas de los estimadores M.C.O.

1. La suma de los residuales en todo modelo de regresión que contiene el intercepto es siempre 0
2. $\sum (Y_i - \hat{Y}_i) = \sum e_i = 0$
3. $\mathbf{e}'\mathbf{e} = \mathbf{0}$ para cualquier modelo de regresión que contenga una ordenada en el origen.
4. La suma de valores observados $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$
5. El producto cruzado muestral entre cada uno de los regresores y los residuos mco es $\mathbf{X}'\mathbf{e} = \mathbf{0}$
6. $\hat{\mathbf{Y}}'\mathbf{e} = \mathbf{0}$

Tarea: Pruebe estas propiedades