UNIVERSIDAD NACIONAL DE INGENIERIA

Facultad de Ingeniería Económica, Estadistíca y CC.SS.

Curso: Modelos Lineales

Modelos Lineal General

Eduardo Marcos Sánchez

Docente

Gerente de Risk Intelligence

Subgerente de CoE Pricing

Jefe en Implementación de Modelos de Riesgos Créditos Retail

Correo personal: eduardomarcos.sanchez@gmail.com
LinkedIn : https://www.linkedin.com/in/eduardo-marcos-sanchez/

Modelos

Para representar e interpretar formalmente cualquier observación utilizamos modelos.

Un modelo es una representación abstracta de cómo creemos que se comporta el fenómeno en estudio.

Los modelos son herramientas para evaluar las hipótesis de investigación.

El modelo particular que nos ocupa son los ***modelos estadísticos***, los cuales se utilizan en general en el marco de un proceso deductivo (Ver el método hipotético deductivo). Se trata de explicar la ***variabilidad*** de un fenómeno particular tratando de encontrar y comprender los componentes de esta *variabilidad*.

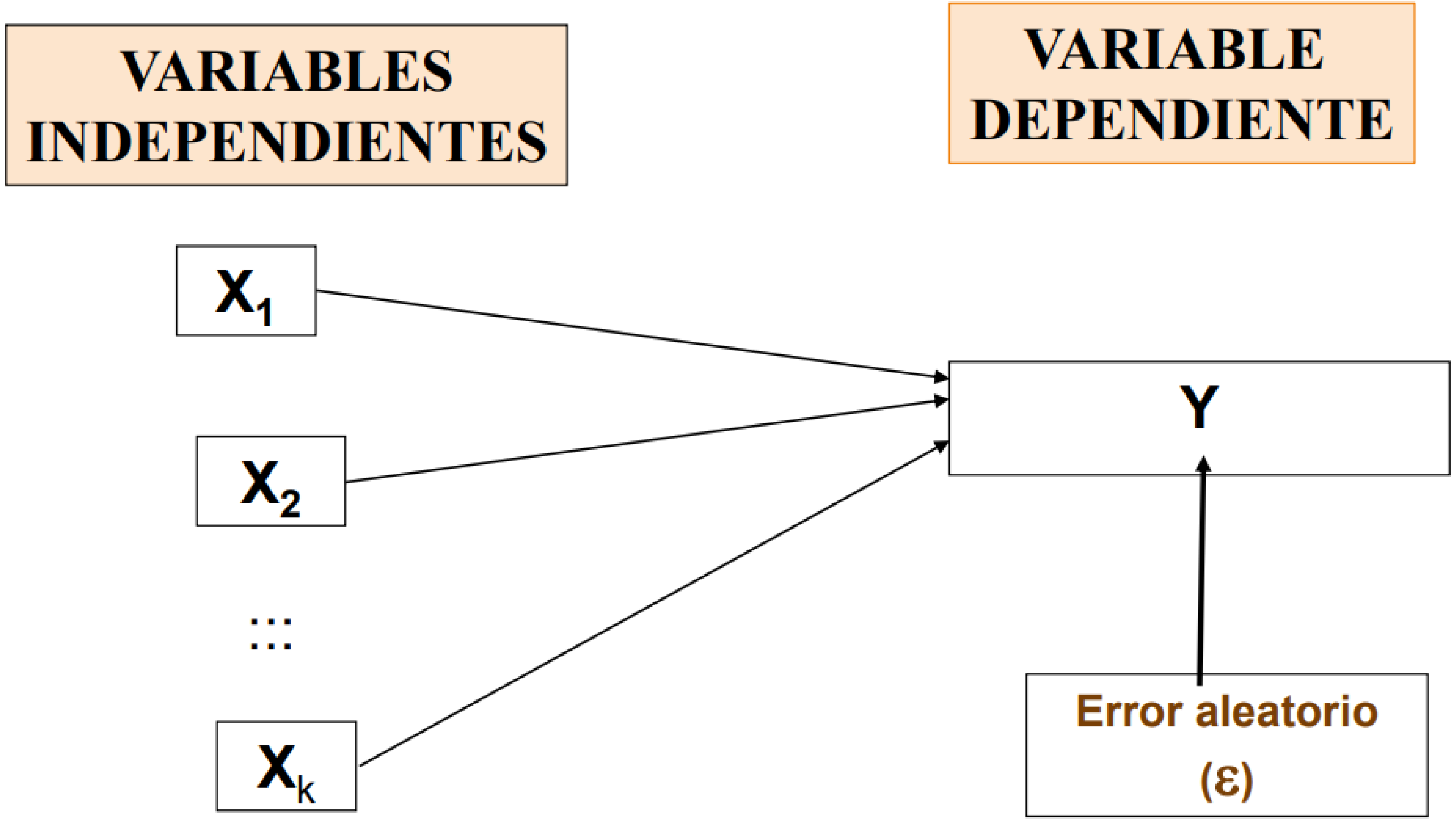
El método de análisis consiste en elaborar un cierto número de hipótesis sobre los factores que explican la variabilidad del fenómeno en estudio y estas hipótesis son verificadas a partir de los datos observados

Modelo Lineal General

En el contexto de la Estadística, se asume que toda observación está formado por dos componentes, la ***componente sistemática***, que representa la tendencia que siguen los datos y que puede ser modelado mediante una función matemática (funciones lineales y no lineales). La segunda componente corresponde a comportamientos en los que se observan sus manifestaciones pero no son medibles y solo se pueden representar mediante modelos probabilísticos (distribución normal, poisson, binomial, etc.), denominada componente estocástica.

El ***modelo lineal general***, constituye la base de la mayoría de los modelos que han surgido a partir de este. Por ejemplo los modelos no lineales, modelos lineales generalizados, modelos de efectos mixtos, modelos lineales aditivos, entre otros

Modelo Lineal General



Modelo Lineal General

Modelo Lineal General

Modelo de Rango Completo

Def. Las columnas de la matriz son linealmente indep. (rangos)

Modelo de Rango Incompleto

Def. Las columnas de la matriz son linealmente dep. (rangos)

Modelo Lineal General

Definición.-

El modelo lineal general se define como:

$$Y = \sum^{k} \beta_j x_j + \varepsilon$$

Donde:

(1) Y es una variable aleatoria continua, observable

(2) Las variable independientes $x_1, x_2, \ldots, x_k$ son fijas y conocidas y pueden ser cuantitativas (numéricas) o cualitativas (categóricas)

(3) Los βj son parámetros (constantes) desconocidas definidas en un espacio paramétrico Ω_β;

(4) ε es una variable aleatoria inobservable con distribución de probabilidades conocida $E(\varepsilon)=0$, $V(\varepsilon)=\sigma^2$), $cov(\varepsilon_i \varepsilon_j)=0$ para todo $i \neq j$

Modelo Lineal General

Forma matricial del modelo

Definición.-

Supóngase que, se selecciona una muestra aleatoria de tamaño n dado por $\{(y_i,x_{i1},x_{i1},\ldots,x_{ik}):i=1,2,\ldots,n\}$ de una población de la forma , entonces la observación de yi para cada individuo se puede escribir como:

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \beta_k x_{1k} + \varepsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{12} + \ldots + \beta_k x_{2k} + \varepsilon_2 \\
&\vdots \quad \vdots \quad \vdots \\
y_n &= \beta_0 + \beta_1 x_{n2} + \ldots + \beta_k x_{nk} + \varepsilon_n
\end{aligned}
$$

Modelo Lineal General

Forma matricial del modelo

Expresando matricialmente el sistema de ecuaciones, tenemos

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & x_{12} & \dots & x_{1k} \\ x_{20} & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

El modelo lineal general se expresa como:

$$\mathbf{y}_{nx1} = \mathbf{X}_{nx(k+1)}\, \boldsymbol{\beta}_{(k+1)x1} + \boldsymbol{\varepsilon}_{nx1}$$

Donde:

(1) $\mathbf{y}_{nx1}$ es un vector de v.a. observables

(2) $\mathbf{X}_{nx(k+1)}$ matriz de números observables (Los elementos de X no son variables aleatorias.

(3) $\boldsymbol{\beta}_{(k+1,1}$ es un vector de parámetros desconocidos definido en un espacio

(4) $\boldsymbol{\varepsilon}_{nx1}$ es un vector aleatorio continuo, inobservable con $E(\boldsymbol{\varepsilon}_{nx1})=0$ y $cov[\boldsymbol{\varepsilon}_{nx1}]=\Sigma$

Modelo Lineal General

Forma matricial del modelo

Notas.-

(1) En la especificación(4) la $\text{cov}(\boldsymbol{\varepsilon}_{nx1})=\Sigma$, es igual a la $\text{cov}(\mathbf{y}_{nx1})$ existe y se denota como $\boldsymbol{\Sigma}$.

(2) En muchos casos, como parte del modelo, se establecen suposiciones adicionales, acerca de la distribución de probabilidades de $\boldsymbol{\varepsilon}_{nx1}$ (o de $\mathbf{Y}$) acerca de la estructura de $\boldsymbol{\Sigma}$.

(3) Usualmente se asume que $\mathbf{x}_{0i}=\mathbf{1}$.

Modelo Lineal General

Ejm: Modelo de Rango Completo

Y : Calificación respecto a la satisfacción en el trabajo.

X_1: Calificación en un test de inteligencia

X_2: Índice de adaptación personal

Supongamos que el vector $(Y,X_1,X_2) \sim N_3(\boldsymbol{\mu},\boldsymbol{\Sigma})$

Recordemos que $Y/X_1=x_1,X_2=x_2 \sim N(\mu,\sigma^2)$

Además, $E(Y/X_1=x_1,X_2=x_2) = \mu = \beta_0+\beta_1 x_1+\beta_2 x_2$

El modelo de regresión lineal para cada una de las $n=31$ observaciones se expresaría como:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \qquad i=1,2,\ldots,31$$

La componente aleatoria, ε, se distribuye igual que la variable dependiente, esto es,

$E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2$, $cov(\varepsilon_i,\varepsilon_j)=0$; además, la distribución de ε es normal.

Modelo Lineal General

Ejm: Modelo de Rango Completo

$$E(Y/X_1=x_1, X_2=x_2) = \mu = \beta_1 x_1$$

Calificación respecto a la satisfacción en el trabajo (Y)	Calificación en un test de inteligencia (x_1)	Indice de adaptación personal (X_2)
54	15	8
37	13	1
30	15	1
48	15	7
37	10	4
37	14	2
31	8	3
49	12	7
43	1	9
12	3	1
30	15	1
37	14	2
61	14	10
31	9	1
31	4	5

$$\mu_1=E(Y_1/X_1=15,\ X_2=8)=\beta_0+\beta_1 15+\beta_2 8$$

$$\mu_2=E(Y_2/X_1=13,\ X_2=1)=\beta_0+\beta_1 13+\beta_2(1)$$

::::

$$\mu_{15}=E(Y_{15}/X_1=4,\ X_2=5)=\beta_0+\beta_1 4+\beta_2 5$$

Modelo Lineal General

Ejm: Modelo de Rango Completo

Un estudio fue conducido para examinar aquellas posibles variables relacionadas con la satisfacción en el trabajo de los empleados sin un grado profesional, de los hospitales. Una muestra aleatoria de 15 empleados produjo los siguientes resultados:

Calificación respecto a la satisfacción en el trabajo (Y)	Calificación en un test de inteligencia (x_1)	Indice de adaptación personal (X_2)
54	15	8
37	13	1
30	15	1
48	15	7
37	10	4
37	14	2
31	8	3
49	12	7
43	1	9
12	3	1
30	15	1
37	14	2
61	14	10
31	9	1
31	4	5

Modelo Lineal General

Ejm: Modelo de Rango Completo

El modelo expresado matricialmente es dado por:

$$\begin{pmatrix} 54 \\ 37 \\ \vdots \\ 31 \end{pmatrix}_{15x1} = \begin{pmatrix} 1 & 15 & 8 \\ 1 & 13 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 4 & 5 \end{pmatrix}_{15x3} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}_{3x1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{15} \end{pmatrix}_{15x1}$$

El rango máximo de la matriz de diseño, X es 3, (En este caso es exactamente 3)

Estimación de parámetros: Mínimos Cuadrados Ordinarios (MCO)

Estimación de β mediante el método MCO

Dado un modelo lineal general

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

El método Mínimos Cuadrados Ordinarios consiste en buscar aquel β que minimize $\sum_{i=1}^{n} e_i^2$, i.e.

se obtiene que $\hat{\beta}_{MCO}$ forma parte de un **sistema de ecuaciones normales**:

$$\mathbf{X}'\mathbf{X}\hat{\beta}_{MCO} = \mathbf{X}'\mathbf{Y}$$

Desde que $\mathbf{X}$ es una matriz de rango completo se concluye que

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

Propiedad de Mínimos Cuadrados Ordinarios (MCO)

Propiedades del método MCO

- $\hat{\beta}_{MCO}$ es un estimador insesgado de β.
- $Var(\hat{\beta}_{MCO}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
- Las variables explicativas son ortogonales al vector de residuos MCO, i.e. $\mathbf{X}'\hat{\varepsilon} = \mathbb{O}$.
- El predictor es ortogonal al vector de residuos MCO, i.e. $\hat{\mathbf{Y}}'\hat{\varepsilon} = \mathbb{O}$.
- $\hat{\beta}_{MCO}$ y $\hat{\varepsilon}$ están incorrelacionados.
- La suma de cuadrados residual se puede expresar como $\mathbf{Y}'\mathbf{Y} - \hat{\beta}'_{MCO}\mathbf{X}'\mathbf{Y}$, esto debido a que $\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \hat{\mathbf{Y}}'\mathbf{Y}$.
- El vector de residuos MCO es una transformación lineal del término
- $\mathbb{E}(\hat{\varepsilon}) = \mathbb{O}$ y $Var(\hat{\varepsilon}) = \sigma^2 M_{\mathbf{X}}$, donde $M_{\mathbf{X}} = \mathbb{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Si el modelo posee término independiente, la suma de residuos es cero.

Propiedad de Mínimos Cuadrados Ordinarios (MCO)

Propiedades del método MCO

- $s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-p-1}$ es un estimador insesgado para σ^2
- Dadas las siguientes sumas:
  - Suma de Cuadrados Total (SCT) $:= \sum_{i=1}^{n}(y_i - \overline{y})^2$
  - Suma de Cuadrados Explicada (SCE) $:= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$
  - Suma de Cuadrados Residual (SCR) $:= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

  Si el modelo lineal general posee término independiente, entonces

$$SCT = SCE + SCR$$

Teorema de Gauss - Markov

El estimador $\hat{\beta}_{MCO}$ es el mejor estimador lineal insesgado para β, i.e. cualquier otro estimador tiene matriz de covarianzas mayor que la del estimador MCO.

Medida de bonda de ajuste

- El objetivo es buscar un modelo que se ajuste correctamente a los datos y adem´as sea el más sencillo posible.
- Para el caso de la bondad de ajuste, nótese que se obtiene un buen ajuste del modelo cuando los residuales son pequeños.

Coeficiente de Determinación

Dado un modelo lineal general de la forma **Y = Xβ + ε** con término independiente, el coeficiente de determinación se define como

$$R^2 = 1 - \frac{SCR}{SCT} = \frac{SCE}{SCT}$$

Si el modelo no posee término independiente (β_0= 0) entonces se utiliza una modificación del R^2 de la forma

$$R_0^2 = 1 - \frac{SCR}{\sum_{i=1}^{n} y_i^2}$$

Medida de bonda de ajuste

El R^2 no indica:

- Las variables regresoras son una causa de los cambios en la variable dependiente.
- Existe sesgo por variables regresoras omitidas.
- Ha sido escogido el conjunto más apropiado de variables explicativas.
- Hay presencia de colinealidad en los datos, pero sí te da una primera pista.
- El modelo puede ser mejorado por transformaciones en las variables regresoras.
- Hay suficientes observaciones muestrales para dar una conclusión solida

Coeficiente de Determinación ajustado

- El coeficiente de determinación ajustado es una generalización del R^2 tradicional. Este indicador de bondad de ajuste además de tomar en cuenta el ajuste del modelo penaliza el uso de más variables regresoras.

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

- El R^2_{adj} es una medida para comparar modelos. Además el $R^2_{adj} \leq R^2$

Modelo lineal general

Ejercicio:

- Pruebe que dado un modelo lineal general con intercepto de la forma $Y = X\beta + \varepsilon$, entonces la suma de cuadrados explicada se puede representar como
$$SCE = \hat{\boldsymbol{\beta}}_{\mathbf{1}} X_c' X_c \hat{\boldsymbol{\beta}}_{\mathbf{1}} = \hat{\boldsymbol{\beta}}_{\mathbf{1}} X_c' \mathbf{Y}$$
- Pruebe que la suma de cuadrados residual de un modelo lineal general no cambia si usamos la versión centrada del mismo modelo.
- Sea el modelo lineal general $Y = X\beta + \varepsilon$ con p variables explicativas y n observaciones. Suponiendo que ε posee distribución normal multivariada, halle la distribución de
$$W = \frac{\hat{\boldsymbol{\beta}}_{\mathbf{1}} X_c' Y / p}{(Y'(\mathbb{I}_n - \frac{1}{n}\mathbb{J}_n)Y - \hat{\boldsymbol{\beta}}_{\mathbf{1}} X_c' Y)/(n-p-1)}$$
- Pruebe que si agregamos una nueva variable explicativa a un modelo lineal general, entonces el R^2 mejora. ¿Qué ocurriría con el R^2_{adj}?