

Maestría en Ciencia y Análisis de Datos- Universidad Mayor de San Andrés

Modelos lineales y modelos lineales generalizados

Rolando Gonzales Martinez, PhD

Fellow postdoctoral Marie
Skłodowska-Curie

Universidad de Groningen
(Países Bajos)

Investigador (researcher)

Iniciativa de Pobreza y Desarrollo
Humano de la Universidad de
Oxford (UK)

Recap: Contenido

(1) Introducción a los Modelos Lineales

- Definición de modelos lineales.
- Regresión lineal simple y múltiple.
- Métodos de ajuste de modelos lineales: mínimos cuadrados ordinarios (OLS).
- Laboratorio: Ajuste de modelos lineales en R/Python o el programa de preferencia de los estudiantes.

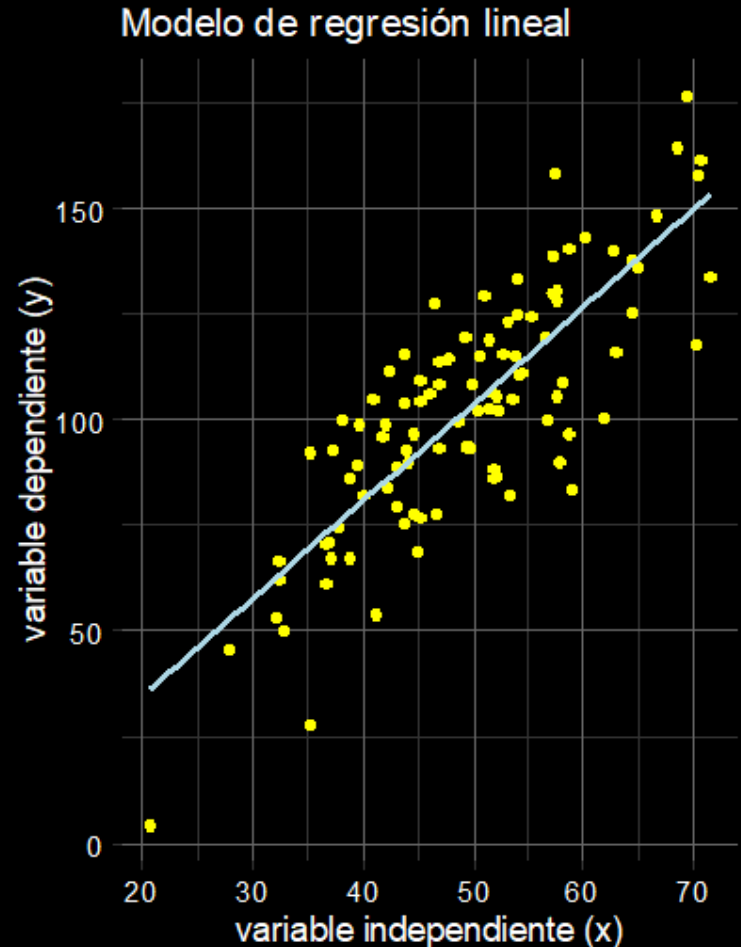
Objetivos de aprendizaje SMART de los conceptos y habilidades para las sesiones de modelos lineales

Al concluir las sesiones, se logrará:

- Conceptos:
 - Comprender qué son los modelos y qué son los modelos lineales
 - Entender los métodos de estimación de parámetros modelos lineales
 - Entender el concepto de estimador y sus propiedades
- Habilidades:
 - Estimar en la práctica modelos lineales
 - Analizar en la práctica los resultados de modelos lineales
 - Identificar en la práctica patologías en los modelos lineales

Definición de modelos lineales

- Modelos estadísticos que asumen una relación lineal entre variables.
- Los modelos lineales pueden ser simples, con una sola variable independiente, o múltiples, con varias variables independientes.
- Se utilizan en estadística y en diversas disciplinas científicas para describir y predecir relaciones entre variables.



Modelos

- **Modelo Matemático:** **Abstracción** para representar, comprender y predecir fenómenos utilizando matemáticas (ecuaciones, funciones, etc.)
 - Los modelos matemáticos son validados mediante la comparación de sus resultados/predicciones contra datos observacionales o experimentales, y pueden ser adaptados y refinados a medida que se dispone de más datos o se comprende mejor el fenómeno estudiado.
 - **Epistemológicamente**, los modelos son construcciones abstractas que permiten organizar y estructurar el conocimiento de manera sistemática y precisa.

Modelos

- **Modelo Estadístico:** Subtipo de modelo matemático que utiliza datos **empíricos** y probabilidades para entender, analizar relaciones y predecir fenómenos observables

Funciones Epistemológicas de un Modelo Estadístico:

- **Descriptiva:** Describe y resume las características principales de un conjunto de datos.
- **Inferencial:** Permite hacer inferencias sobre fenómenos de interés mediante estimaciones y pruebas de hipótesis.
- **Predictiva**
- **Explicativa:** Ayuda a entender las relaciones y mecanismos subyacentes entre variables.

Variables

La estadística analiza variables que fluctúan de una forma más o menos impredecible.

Variable aleatoria (definición). *Dado un espacio $(S, \sigma_{\mathcal{B}}, \mathbb{P})$, una variable aleatoria es una función del espacio muestral S a \mathbb{R} , $X : S \rightarrow \mathbb{R}$.*

Ejemplo: Para $S = \{C, E\}$, es posible definir una función $X(m)$ tal que,

$$X(m) = \begin{cases} 1 & \text{si } m = C \\ 0 & \text{si } m = E \end{cases}$$

Variable

Cuando se realiza un experimento, la realización es un resultado en el espacio muestral.

Para cada evento A del espacio muestral S , puede asociarse un número entre cero y uno que se llamará probabilidad de A , $\mathbb{P}(A)$.

Para una definición más precisa, es necesario definir primero el concepto de sigma álgebras.

Sigma álgebra (definición). *Una colección de subconjuntos de S se llama sigma álgebra (σ -álgebra o campo de Borel), denotado por \mathcal{B} , si satisface las siguientes propiedades:*

1. $\emptyset \in \mathcal{B}$
2. Si $A \in \mathcal{B}$, entonces $A^c \in \mathcal{B}$
3. Si $A_1, A_2, \dots \in \mathcal{B}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{B}$

Variables

Función de probabilidad (definición). *Dado un espacio muestral S y una sigma álgebra asociada \mathcal{B} , una función de probabilidad será aquella que satisfaga,*

1. $\mathbb{P}(A) \geq 0$ para todo $A \in \mathcal{B}$
2. $\mathbb{P}(S) = 1$
3. Si $A_1, A_2, \dots \in \mathcal{B}$ son disjuntos, entonces
$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Estas propiedades se denominan usualmente Axiomas de Probabilidad o Axiomas de Kolgomorov. La terna $(S, \sigma_{\mathcal{B}}, \mathbb{P})$ es un espacio de probabilidad en el que cada suceso $A \in \mathcal{B}$ recibe el nombre de probabilidad de A .

Variables

Las variables se dividen en varias categorías según su naturaleza y el tipo de valores que pueden tomar.

1. Variables Cualitativas (o Categóricas):

- Nominales: Son variables que representan categorías sin un orden inherente. Ejemplos incluyen el color de ojos (azul, verde, marrón) o el tipo de mascota (perro, gato, pájaro).
- Ordinales: Son variables que representan categorías con un orden lógico, pero sin una distancia uniforme entre categorías. Ejemplos incluyen los niveles de satisfacción (insatisfecho, neutral, satisfecho) o las clasificaciones (primero, segundo, tercero).

2. Variables Cuantitativas (o Numéricas):

- Discretas: Son variables que toman valores contables, generalmente números enteros. Ejemplos incluyen el número de hijos en una familia o el número de coches en un garaje.
- Continuas: Son variables que pueden tomar cualquier valor dentro de un rango, incluyendo fracciones y decimales. Ejemplos incluyen la altura, el peso o la temperatura.

Variables

3. Variables Dependientes e Independientes:

- Independientes: Son variables que se manipulan o controlan para observar su efecto sobre otras variables. En un experimento, son las que el investigador cambia para ver cómo afectan a la variable dependiente.
- Dependientes: Son variables que se miden para ver el efecto de las variables independientes. En un experimento, son las que se observan y registran para ver cómo cambian en respuesta a las variables independientes.

4. Variables Dicotómicas: variables que sólo pueden tomar dos valores posibles.

5. Variables de Escala:

- Intervalo: Son variables cuantitativas donde la distancia entre dos valores tiene significado, pero no hay un verdadero cero. Ejemplo: temperatura en grados Celsius.
- Razón: Son variables cuantitativas donde hay un verdadero cero y se pueden realizar operaciones matemáticas significativas. Ejemplo: peso, altura, ingresos.

Función de distribución de una variable aleatoria

Asociada a una variable aleatoria X existe una función, denominada función de distribución acumulada de X .

Función de distribución acumulada (definición). *La función de distribución acumulada (cdf) de una variable aleatoria X , denotada por $F_X(x)$, se define como,*

$$F_X(x) = \mathbb{P}_X(X \leq x), \text{ para todo } x$$

Esta función satisface las siguientes propiedades,

1. $\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1.$
2. $F(x)$ es una función no decreciente de x .
3. $\lim_{x \downarrow x_0} F(x) = F(x_0)$

Nótese que una variable aleatoria será continua si $F_X(x)$ es una función continua de x , y será discreta si $F_X(x)$ es una función en pasos de x .

Funciones de masa y densidad

Asociada a X y a su cdf F_X existe otra función, llamada función de densidad de probabilidad (pdf) o función de masa de probabilidad (pmf), refiriéndose al caso continuo y discreto, respectivamente. Ambas se refieren a probabilidades puntuales de variables aleatorias.

Función de masa de probabilidad (definición) . *La función de masa de probabilidad (pmf) $f_X(x)$ de una variable aleatoria discreta X es $f_X(x) = \mathbb{P}(X = x)$ para todo x*

Función de densidad de probabilidad (definición) . *La función de densidad de probabilidad (pdf) $f_X(x)$ de una variable aleatoria continua X es $F_X(x) = \int_{-\infty}^x f_X(t) dt$ para todo x (nótese que $\frac{d}{dx} F_X(x) = f_X(x)$ y en el caso discreto, de manera similar, las probabilidades puntuales $f_X(x)$ se añaden para obtener $F_X(x)$)*

La expresión " X tiene una distribución $F_X(x)$ ", se escribe $X \sim F_X(x)$.

Distribuciones discretas y continuas

Las distribuciones de probabilidad pueden ser discretas o continuas. Una variable aleatoria X tiene una distribución discreta si el rango de X , el espacio muestral, es contable (en la mayoría de las situaciones, se encuentra en \mathbb{Z}). Por ejemplo, una variable aleatoria X tiene una *distribución uniforme discreta* si,

$$\mathbb{P}(X = x|N) = \frac{1}{N}, \quad x = 1, 2, \dots, N$$

para $N \in \mathbb{Z}^+$. Esta distribución pone igual masa en todos los posibles resultados $1, 2, \dots, N$.

En usual también trabajar con *variables continuas*, definidas en \mathbb{R} , o en $\mathbb{R}^{0,+}$ (como las variables de stock), por lo que se prestará atención a las distribuciones de variables continuas en esta sección, especialmente a las utilizadas usualmente en inferencia.

Distribuciones continuas

Distribución uniforme continua (definición). *La pdf de una distribución uniforme continúa es,*

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b, \\ 0 & \text{para } x < a \text{ o } x > b \end{cases}$$

por lo que si una variable aleatoria continua $X \sim \mathcal{U}(a, b)$, su distribución queda definida por los parámetros a y b . Si $a = 0$ y $b = 1$ la distribución resultante $\mathcal{U}(0, 1)$ se denomina distribución uniforme estándar.

Distribución Normal (Gaussiana)

Distribución gaussiana (definición). *Una variable aleatoria continua X se distribuye normalmente (sigue una distribución Gaussiana o Gauss-Laplace) con media μ y varianza σ^2 , denotado por $X \sim \mathcal{N}(\mu, \sigma^2)$, si la pdf de X es,*

$$f(x) \equiv f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}$$

Algunas propiedades útiles de $f(x)$ son,

1. $f(\mu + x) = f(\mu - x)$ (simetría alrededor de μ).
2. $\mathbb{P}(\mu - \sigma < X < \mu + \sigma) \cong 0.683$
3. $\mathbb{P}(\mu - 2\sigma < X < \mu + 2\sigma) \cong 0.954$
4. $\mathbb{P}(\mu - 3\sigma < X < \mu + 3\sigma) \cong 0.997$

Distribución Normal (Gaussiana) estandar

Distribución gaussiana estándar (definición). *Una variable aleatoria continua X sigue una distribución normal estándar si $\mu = 0$ y $\sigma^2 = 1$, por lo que $X \sim \mathcal{N}(0, 1)$.*

Distribución continua t de Student

Distribución t de Student. Sean dos variables aleatorias $Z \sim N(0, 1)$ y $Y \sim \chi^2(k)$, la variable aleatoria continua T ,

$$T = \frac{Z}{\sqrt{\frac{Y}{k}}}$$

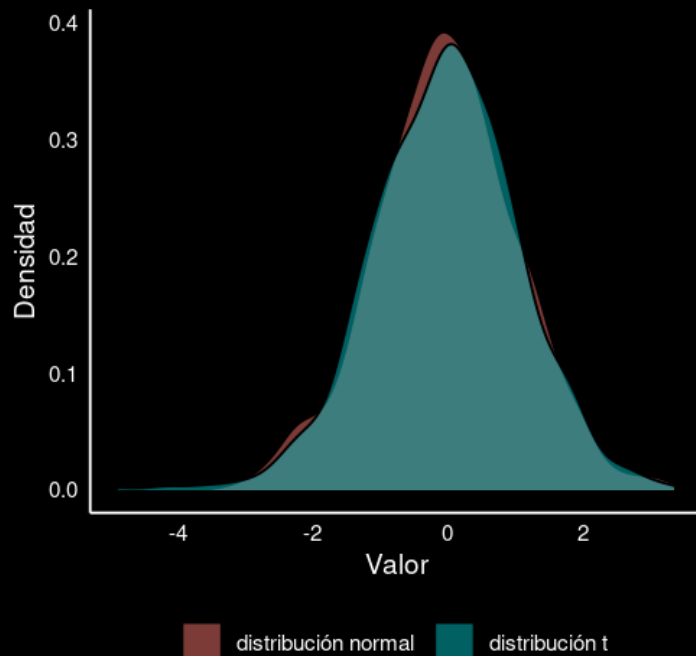
seguiría una distribución t -Student con una pdf,

$$f(T, k) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi} \Gamma(\frac{k}{2})} \left(1 + \frac{T^2}{k}\right)^{-\frac{1}{2}(k+1)}$$

Una propiedad importante e interesante de la distribución t de Student es que puede ser usada para muestras pequeñas que impliquen la distribución Gaussiana.

Sea F_k la cdf de $t(k)$ y Φ la cdf de una distribución $\mathcal{N}(0, 1)$. Entonces, $\lim_{k \rightarrow \infty} F_k(t) = \mathcal{N}(0, 1)$. De hecho para $k > 40$ se logra una buena aproximación.

Distribución t de Student vs Normal



Momentos

La esperanza de una variable aleatoria es una medida de su tendencia central, que resume el valor esperado de esta variable.

Esperanza de una variable aleatoria (definición). *La esperanza de una variable aleatoria $g(X)$, denotada por $\mathbb{E}(X)$, es*

$$\mathbb{E}(X) \begin{cases} \sum_{x \in X} g(x) \mathbb{P}(X = x) & \text{si } X \text{ es discreta} \\ \int_{-\infty}^{\infty} g(x) f(x) dx & \text{si } X \text{ es continua,} \end{cases}$$

Sea una muestra x_1, \dots, x_n . La esperanza $\mathbb{E}(X)$ está relacionada con el primer momento muestral (la media),

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Segundo momento y momentos superiores

El segundo momento central es la varianza.

Varianza (definición). *La varianza de una variable aleatoria X es su segundo momento central, $Var(X) = \mathbb{E}(X - \mathbb{E}(X))^2$. La raíz cuadrada positiva de $Var(X)$ es la desviación estándar de X .*

La varianza proporciona una medida del grado de dispersión de una distribución alrededor de la media. La varianza muestral de una muestra x_1, \dots, x_n se calcula con,

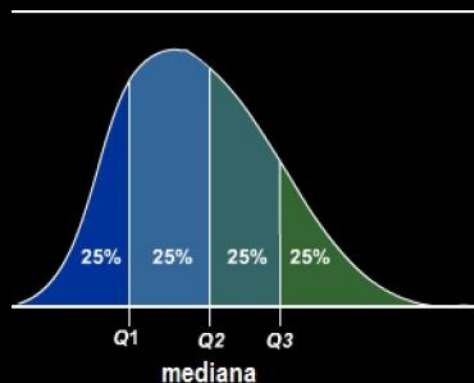
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)$$

Momentos superiores son el sesgo y la curtosis de una distribución.

El sesgo mide el grado de simetría de una distribución, y la curtosis el grado de apuntamiento de la distribución. Dependiendo del valor de la curtosis, una distribución puede ser leptocúrtica (curtosis > 3), mesocúrtica (curtosis $= 3$) o platicúrtica (curtosis < 3).

Fractiles

Fractiles (definición). Sea X una variable aleatoria continua y sea $\alpha \in (0, 1)$. Si $q = q(X; \alpha)$ es aquel tal que $\mathbb{P}(X < q) = \alpha$ y $\mathbb{P}(X > q) = 1 - \alpha$, entonces q es llamado un fractil de X .



Si se expresa las probabilidades en porcentaje, q será el 100α percentil de la distribución de X .

Estimación

La estimación es el proceso de extraer información acerca del valor de cierto parámetro poblacional a partir de la muestra x_1, \dots, x_n , utilizando algún estadígrafo como estimador, calculado con los datos x_1, \dots, x_n .

Estadígrafo.

Considérese que se obtiene una muestra de tamaño n de una población. Un estadígrafo es una función que se obtiene utilizando como datos las observaciones de la muestra x_1, \dots, x_n .

Estimación puntual e intervállica

Dado un modelo de regresión:

$$y = \beta_0 + \beta_1 x + \epsilon$$

Los estimadores puntuales son:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimación puntual e intervállica

Un intervalo de confianza $(1 - \alpha)\%$ es:

$$\beta_1 \pm t_{n-2, \alpha/2} \times SE(\beta_1)$$

Donde:

$$SE(\beta_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

$$\sigma^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} \quad \begin{array}{l} e_i = y_i - \hat{y}_i \\ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \end{array}$$

Ejercicios prácticos

R (posit cloud):

<https://posit.cloud>

R analytic flow:

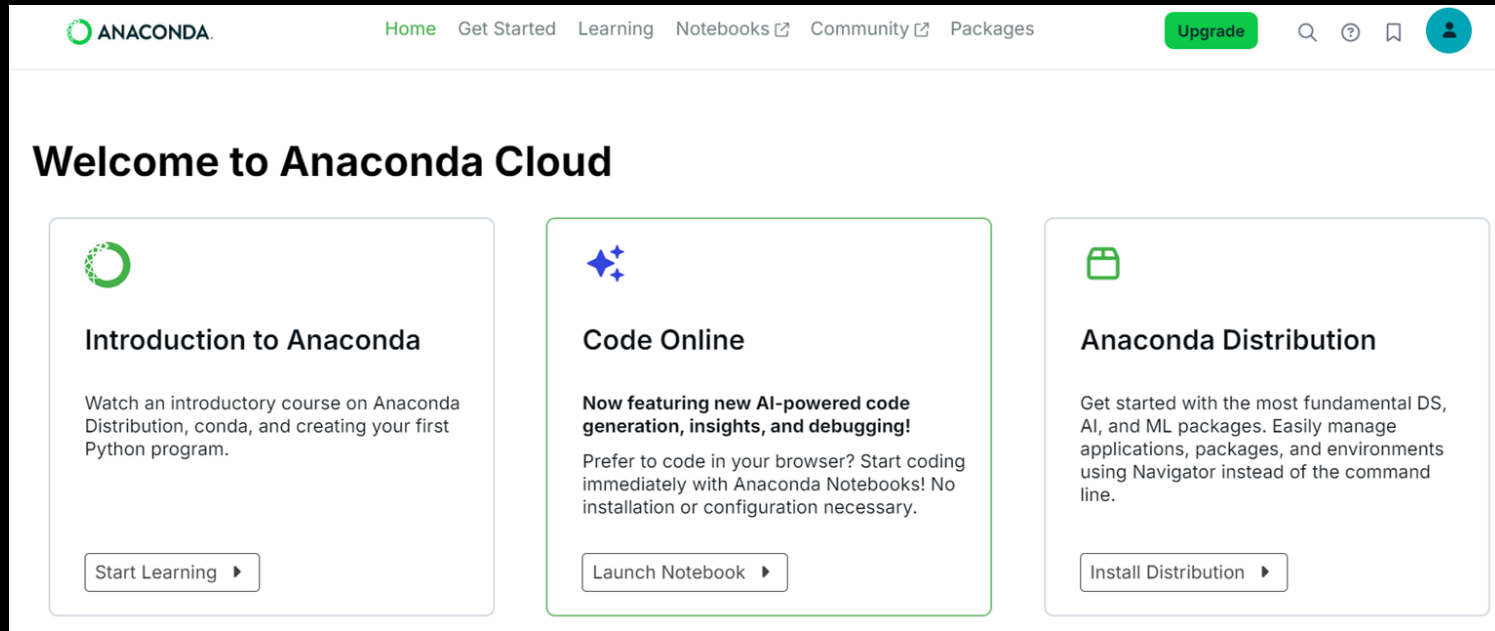
<https://r.analyticflow.com/en/>

Python (Anaconda cloud):

<https://anaconda.cloud/sign-in>

Anaconda cloud

Es necesario crear una cuenta o ingresar con un usuario de Google/GitHub y luego ingresar a Code Online:




The screenshot shows the Anaconda Cloud homepage. At the top is a navigation bar with the Anaconda logo, links for Home, Get Started, Learning, Notebooks, Community, and Packages, an Upgrade button, and search, help, and user icons. The main heading is 'Welcome to Anaconda Cloud'. Below this are three featured cards: 'Introduction to Anaconda' with a green circle icon and a 'Start Learning' button; 'Code Online' with a blue star icon, a description of AI-powered code generation, and a 'Launch Notebook' button; and 'Anaconda Distribution' with a green folder icon and an 'Install Distribution' button.

ANACONDA

Home Get Started Learning Notebooks Community Packages Upgrade


Welcome to Anaconda Cloud



Introduction to Anaconda

Watch an introductory course on Anaconda Distribution, conda, and creating your first Python program.

Start Learning ▶




Code Online

Now featuring new AI-powered code generation, insights, and debugging!

Prefer to code in your browser? Start coding immediately with Anaconda Notebooks! No installation or configuration necessary.

Launch Notebook ▶



Anaconda Distribution

Get started with the most fundamental DS, AI, and ML packages. Easily manage applications, packages, and environments using Navigator instead of the command line.

Install Distribution ▶