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.

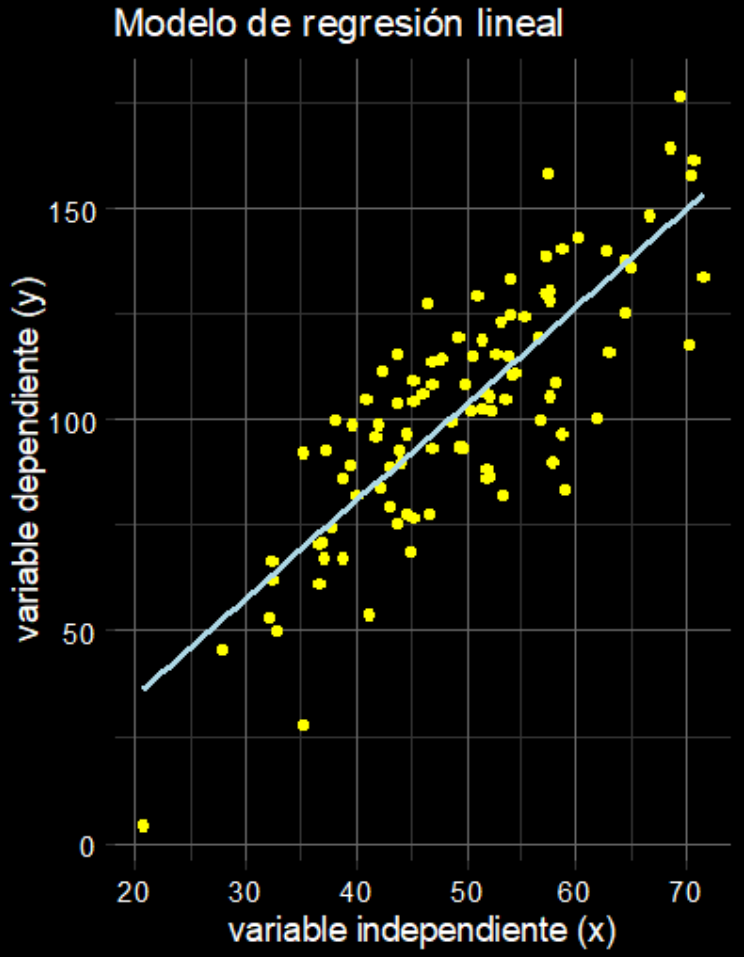
Objetivos de aprendizaje SMART de los conceptos y habilidades para las sesiones de modelos lineales

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 propiedade
- 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.
  - Epistemologicamente, 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* $\mathbb{R}$, $X : S \to \mathbb{R}$.

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

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

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-algebra 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* $A_1, A_2, ... \in \mathcal{B}$, *entonces* $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$

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* $A_1, A_2, ... \in \mathcal{B}$ *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.

Funcion de distribucion 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. $\lim_{x \to -\infty} F(x) = 0, \lim_{x \to \infty} F(x) = 1.$
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 e 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 $X \sim F_X(x)$.

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*

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

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

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

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.*

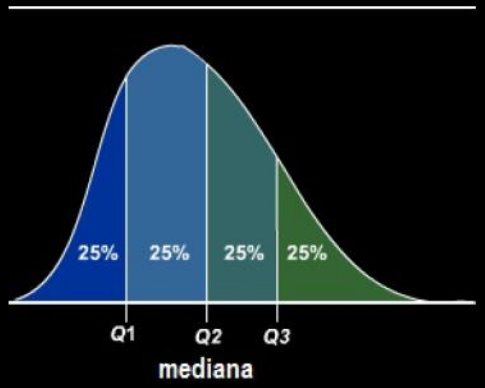
La varianza proporciona una medida del grado de dispersión de una distribución alrededor de la media. La varianza muestral de una muestra $x_1, ..., x_n$ se calcula con,

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

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.