regresión apuntes

Valentina Valdez Vega

2025-04-27

Regresión

$$y = f(x_1, x_2, \ldots)$$

- y Variable de resultado, dependiente, solo tenemos a una y.
- $x_1, x_2, \ldots$, variables de control, independientes.

A partir de estas variables:

- ¿Cuál es la relación de x sobre y?
  - Lineal

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \epsilon_i$$
$$E[y_i] = E[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots]$$

$$\frac{dy}{dx_1} = \beta_1$$

Nota: Diferenciar que la regresión busca establecer relaciones basadas en los datos y no asi un proceso causal.

- Polinomial
- Etc; No lineal,
- Conocer la naturaleza de y y las variables x
  - Y es cuanti (real), X mixtas. (Modelos lineales, MCO)
  - Y es cuanti (discreta >= 0), X cuanti. (Poisson)
  - Y es cuali nominal binario, X mixtas. (LOGIT/PROBIT)
  - Y es cuali ordinal, X mixtas. (Logit/probit ordenados)

Regresión lineal

$$y \in IR$$

$$x_1, x_2, \ldots (mixtas)$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Las variables x son independientes mutuamente.

0. Pregunta de investigación, revisión de literatura
1. Base datos lista para el modelo (Unidad de investigación)
2. Establecer la relación interés
3. Definir el modelo de interés
4. Optimizar el modelo
5. Validar el modelo
6. Analizar/Predecir a partir del modelo

Causalidad vs Correlación

Paso 0: Qué determina los ingresos de una persona?

Paso 1: Base de datos

- Encuesta a hogares 2023
  - Mide todos los componentes del ingreso
  - Cuenta con información socio económica de las personas

Paso 2: Establecer la relación de interés.

- y: Ingreso laboral de la persona
- x: Educación, área, sexo, edad, experiencia, horas trabajadas
  - Priorizar alguna X
  - Explorativa
  - Predictiva
- Variables de control: Variables X observables y relacionadas con y.
- Variables de disturbio: Variables X no observables y relacionadas con y.

Algunos comentarios sobre la calidad de este modelo:

- Variables omitidas
- Se debe especificar la población objetivo de la manera más clara posible.
- Se debe identificar la naturaleza de las covariables (X)
- Se debe definir el alcance del modelo; muestral o inferencial (En el caso de encuestas)

PO: **Personas que trabajan, con 25 años o más de edad**

Paso 3: Definir el modelo a utilizar

$$y = f(x)$$

Paso 4: Optimizar el modelo

- Tratamiento sobre variables de control
  - Transformaciones
  - Definir como factor
  - Polinomios
  - Interacciones
- Tratamiento de datos atípicos
  - Bonferroni. H0: observación i es
- Stepwise: Regresión paso a paso (step)
- Backward: Regresión hacia atras

Paso 5: Validar el modelo

- Residuos
  - Normalidad
  - Varianza constante
- Colinealidad
  - VIF
  - Inclusión de interacciones y polinomios

Predicciones

Probit y Logit

Estrategia, llevar valores binarios a valores continuos. Mediante una función de enlace ($F(Y)$).

$$F(Y) = Y' = X\beta + \epsilon$$

Probit:

$$Y = \Phi(X\beta + \epsilon)$$
$$\phi^{-1}(Y) = X\beta + \epsilon$$

$$Y' = X\beta + \epsilon$$

El enlace $F(Y) = \Phi^{-1}(Y)$, es conocida como probit.

Logit:

$$logit(Y) = log\left(\frac{Y}{1-Y}\right) = X\beta + \epsilon$$

$$Y = \frac{e^{X\beta+\epsilon}}{1+e^{X\beta+\epsilon}}$$

Aplicación en R

$$Pobreza = f(sexo, edad, area, dep, ...)$$

Recomendación:

- Lo visto en este tema tiene una limitación, es el tratamiento sobre muestras autoponderadas es decir su alcance es limitado.
- Para incorporar este tratamiento en encuestas por muestreo se recomienda usar la librería survey y srvyr
- Se puede usar los métodos de agrupamiento para mejorar el rendimiento del modelo, muchas veces como una variable de estrato.
- Para garantizar que el modelo este libre de multicolinealidad se recomienda usar componentes principales