Auxiliar 9

Resumen de conceptos

Profesor: Raimundo Undurraga
Auxiliares: Brandon Galarza, Camila Jáuregui, Leonardo Meneses, Francisca Monetta, Matías Reyes, Bastián Urzúa, Antonia Villegas.

Regresión lineal

Método estadístico que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. La forma más simple de regresión lineal es la regresión lineal simple, que implica una variable independiente, mientras que la regresión lineal múltiple involucra dos o más variables independientes.

En la regresión lineal simple, se asume que la relación entre la variable dependiente (denotada como Y) y la variable independiente (denotada como X) puede describirse mediante una ecuación de la forma:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Donde:

- Y es la variable dependiente.
- X es la variable independiente.
- β_0 es la ordenada al origen (intercepto), que representa el valor esperado de Y cuando X es igual a cero.
- β_1 es la pendiente de la recta de regresión, que representa el cambio esperado en Y por un cambio de una unidad en X.
- ϵ es el término de error, que representa la variabilidad no explicada por el modelo.

El objetivo de la regresión lineal es estimar los parámetros β_0 y β_1 de manera que la suma de los cuadrados de los residuos (ϵ) sea minimizada. Los residuos son las diferencias entre los valores observados de la variable dependiente y los valores predichos por el modelo.

La regresión lineal multivariable es una técnica estadística que extiende la regresión lineal simple para analizar la relación entre varias variables independientes y una variable dependiente. En la regresión lineal simple, se estudia la relación entre dos variables, mientras que en la regresión lineal multivariable, se consideran dos o más variables independientes.

La forma general de un modelo de regresión lineal multivariable se expresa como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Donde:

- Y es la variable dependiente.
- β_0 es la intersección o término constante.
- β_1, β_2,..., β_n son los coeficientes que representan las contribuciones de las variables independientes X_1, X_2,..., X_n a la variable dependiente Y.
- X_1, X_2,..., X_n son las variables independientes.
- ϵ es el error o residuo, que representa la variabilidad no explicada por el modelo.

El objetivo de la regresión lineal multivariable es encontrar los valores óptimos de los coeficientes (β_0, β_1, β_2,..., β_n) que minimizan la suma de los cuadrados de los residuos, es decir, la diferencia entre los valores predichos por el modelo y los valores reales observados.

A continuación se muestra un gráfico de ejemplo de regresión lineal:

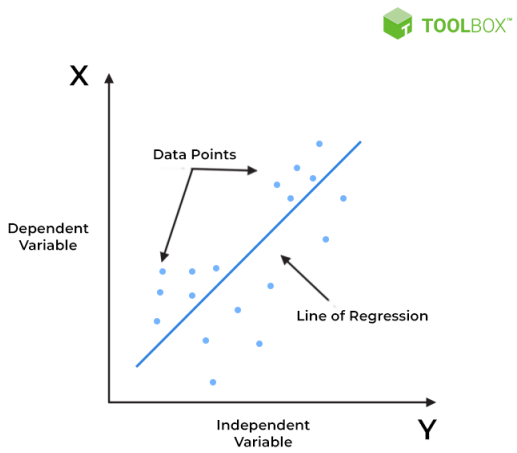

Figura 1: Gráfico de una regresión lineal

OLS

Método para estimar los parámetros de un modelo de regresión lineal. El objetivo principal de este método es encontrar los valores de los coeficientes del modelo que minimizan la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo.

Observación: es importante notar que la regresión lineal es el modelo de expresión de un conjunto de datos y el OLS/MCO es un método para estimar los parámetros del modelo de regresión lineal.

Supuestos OLS

1. **Linealidad:** El modelo de regresión es lineal en los parámetros

$$Y_i = \beta_0 + \beta_1 \cdot X_{1i} + \beta_2 \cdot \log(X_{2i}) + \beta_3 \cdot X_{3i}^8 + ... + \epsilon_i \tag{1}$$

$$Y_i = \beta_0 + \sqrt{\beta_1} \cdot X_{1i} + e^{\beta_2} \cdot X_{2i} + ... + \epsilon_i \tag{2}$$

   El modelo debe ser lineal en los parámetros, no necesariamente en las variables independientes. Esto significa que cada parámetro en el modelo debe aparecer de manera lineal en la ecuación de regresión.

   En la primera ecuación, aunque hay funciones no lineales de las variables independientes ($\log(X_{2i})$ y X_{3i}^8), los parámetros β_0, β_1, β_2, β_3,... aparecen de manera lineal. Cada parámetro se multiplica por una función específica de una variable independiente, y no se eleva a potencias, se multiplica por funciones exponenciales, o se divide por otras variables. Por lo tanto, este modelo cumple con el requisito de linealidad en los parámetros.

   De manera similar con la ecuación (2), aunque hay funciones no lineales de los parámetros ($\sqrt{\beta_1}$ y $e^{\beta 2}$), los parámetros mismos aparecen de manera lineal en la ecuación. Cada parámetro se multiplica por una función específica de una variable independiente. Por lo tanto, este modelo también cumple con el requisito de linealidad en los parámetros.

2. **No dependencia lineal:** No hay relación lineal exacta entre dos variables independientes, para la muestra observada:

   $X^T X$ debe ser invertible, o sea, si tenemos k variables x1, x2,...,xk tienen que ser li. Esto quiere decir que ningún Xi puede ser explicado 100 % por X_i porque $\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y$

3. **Muestras aleatorias:** Los datos de la muestra son aleatorios para tener una alta variabilidad.

$$
\begin{aligned}
\text{Var}(X_k) \text{ grande } &\Rightarrow \text{Var}(\hat{\beta}_{\text{ols}}) \text{ chico} \\
\beta_{\text{ols}} &= (X^T X)^{-1} X^T Y \\
\beta_{\text{ols}} &= (X^T X)^{-1} X^T)(\beta X + \epsilon) \\
\beta_{\text{ols}} &= (X^T X)^{-1} X^T \beta X + (X^T X)^{-1} X^T \epsilon \\
\beta_{\text{ols}} &= \beta + (X^T X)^{-1} X^T \epsilon \\
\text{Var}(\beta_{\text{ols}}) &= \text{Var}(\beta + (X^T X)^{-1} X^T \epsilon)
\end{aligned}
$$

La aleatoriedad en la muestra es fundamental para obtener estimaciones válidas y eficientes de los parámetros del modelo. Asegura que las observaciones sean representativas de la población de interés y que la variabilidad en las variables proporciona información útil para la estimación de los parámetros. La aleatoriedad también juega un papel en la inferencia estadística, como la construcción de intervalos de confianza y la realización de pruebas de hipótesis.

4. **Exogeneidad:** Los errores del modelo no están correlacionados con las variables independientes: Cov(x|ϵ)=0 y también E(μ|X)=0.

5. **Homocedasticidad:** La varianza del error es independiente de los valores de las variables independientes y es constante, o sea los errores ϵ tienen igual varianza: Var(μ|X)= σ^2.

Teorema de Gauss-Markov

Bajo las hipótesis básicas de la regresión lineal, el estimador OLS de β es óptimo entre una familia de estimadores lineales e insesgados. Es decir, no es posible encontrar otro estimador de β que siendo lineal e insesgado tenga una varianza menor que el estimador OLS. MCO/OLS es el mejor estimador lineal insesgado **SÍ O SÓLO SÍ** cumple los 5 requisitos.

Observación: si bien los supuestos 1 y 2 son necesarios para la construcción del MCO, no siempre se cumple el resto. Por lo tanto es necesario identidicar qué problemas trae esta situación (diagnóstico del modelo)

Problemas en el MCO

1. Heterocedasticidad: Cuando no se cumple la homocedasticidad (Supuesto 5), o sea, la varianza de los residuos no es constante (tiene que serlo).

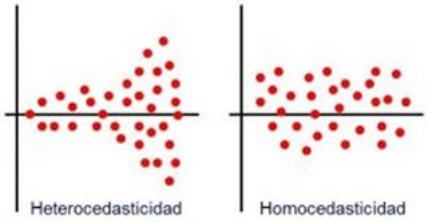

Figura 2: Ejemplo de homocedasticidad

2. Multicolinealidad: cuando existe una correlación importate entre las variables independientes. La heterogeneidad y/o multicolinealidad provocan: mayores SE($\hat{\beta}_k$) y más dificultad para encontrar $\hat{\beta}_k$ significativos provocando peores estimaciones en OLS.

3. Endogeneidad: esto ocurre cuando Cov(x|ϵ) $\neq 0$, o sea cuando falla el supuesto 4.

   ¿Por qué ocurre la endogeneidad?

   - **Sesgo por variable omitida:** $cov(x|z) \neq 0$ cuando mi modelo es $Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$, o sea el z está atrapado en ϵ
   - **Sesgo de medición:** por ejemplo una encuesta puede tener preguntas que inducen al sesgo ("¿Cuál es tu cantante favorito y por qué es Chayanne?")

Inferencia causal

¿Qué es la inferencia causal?: proceso de hacer afirmaciones sobre relaciones causales basadas en datos observacionales o experimentales. La inferencia causal en estadística busca determinar si un cambio en una variable realmente causa un cambio en otra variable.

Nunca olvidar: Correlación $\nRightarrow$ Causalidad

En este modelo de OLS, es posible identificar efectos causales siempre que se sepa reconocer los posibles sesgos y cómo tratarlos.

- **Sesgo de variable omitida:** Cuando una variable relevante se omite, el sesgo resultante puede afectar las inferencias causales. Específicamente, el sesgo por variable omitida puede llevar a conclusiones erróneas sobre la relación causal entre la variable independiente incluida y la variable dependiente. A menudo, esto se debe a la presencia de correlación entre la variable omitida y las variables incluidas en el modelo.

  Supongamos:

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i \tag{3}$$

$$Y_i = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot Z_i + \epsilon_i \tag{4}$$

Donde la ecuación (3) es un modelo simple sin variable y el de la ecuación (4) es un modelo extendido con una variable omitida (Z_i). Por OLS sabemos que:

$$\beta_1 = \frac{\text{cov}(x, y)}{\text{var}(x)}$$
$$\beta_1 = \frac{\text{cov}(x, x\beta_1^* + z\beta_2 + \epsilon)}{\text{var}(x)}$$
$$\beta_1 = \beta_1^* + \frac{\text{cov}(x, z)}{\text{var}(x)}\beta_2 + \frac{\text{cov}(x, \epsilon)}{\text{var}(x)}$$

Sabemos que $\frac{\text{cov}(x,z)}{\text{var}(x)}\beta_2 = 0$ si existe exogeneidad y por otro lado si cov(x|z) $\neq 0$:

$$\beta_1 = \beta_1^* + \frac{\text{cov}(x, z)}{\text{var}(x)}\beta_2 + \frac{\text{cov}(x, \epsilon)}{\text{var}(x)}$$
$$\beta_1 = \beta_1^* + \frac{\text{cov}(x, z)}{\text{var}(x)}\beta_2$$

En el modelo simple, estamos interesados en estimar el efecto causal de X sobre Y, representado por β_1. Sin embargo, si omitimos la variable relevante Z en el modelo extendido, nuestro modelo simple puede volverse sesgado y proporcionar estimaciones sesgadas de β_1. Esto se puede ilustrar al expresar β_1 en términos de los parámetros del modelo extendido:

$$\beta_1 = \beta_1^* + \frac{\text{cov}(x, z)}{\text{var}(x)}\beta_2 + \frac{\text{cov}(x, \epsilon)}{\text{var}(x)}$$

Donde:

- β_1* es el verdadero efecto causal de X sobre Y en ausencia de sesgo por variable omitida.
- $\frac{\text{cov}(x,z)}{\text{var}(x)} \cdot \beta_2$ es el sesgo introducido por omitir la variable Z.

La clave aquí es que si Z está correlacionada con X y afecta a Y, el término $\frac{\text{cov}(x,z)}{\text{var}(x)} \cdot \beta_2$ no será cero, lo que implica que β_1 estimado en el modelo simple será sesgado.

- **Sesgo de selección:** es otro problema potencial en la inferencia causal utilizando OLS. Este sesgo se refiere a la posibilidad de que la muestra utilizada en la regresión no sea representativa de la población total debido a la selección no aleatoria de observaciones.

Supongamos un modelo $Y_i = \beta_0 + \beta_1 \cdot D_i + \epsilon_i$ donde D_i es binaria 0,1. La diferencia de los valores esperados cuando $D_i = 1$ y $D_i = 0$ es aproximable a β_1 (el efecto de D_i sobre Y_i.

Eso implica que:

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(\beta_0 + \beta_1 + \epsilon_i|D_i = 1) - E(\beta_0 + \epsilon_i|D_i = 0)$$
$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \beta_0 + \beta_1 + E(\epsilon_i|D_i = 1) - (\beta_0 + E(\epsilon_i|D_i = 0))$$
$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = \beta_1 + E(\epsilon_i|D_i = 1) - E(\epsilon_i|D_i = 0)$$

El término $E(\epsilon_i|D_i = 1) - E(\epsilon_i|D_i = 0)$ se conoce como el sesgo de selección y se asocia cuando D_i no es independiente de ϵ_i