T2. ANÁLISIS DE LA REGRESIÓN LINEAL. ESTIMACIÓN

- Modelo de Regresión Lineal Simple (MRLS)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

- Modelo de Regresión Lineal Múltiple (MRLM)

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_K X_{Ki}}_{\text{Parte determinista}} + \underbrace{\varepsilon_i}_{\text{Parte aleatoria}}$$

* Y_i → variable explicada

" endógena

" dependiente

* X's → variables explicativas

(K) " exógenas

" independientes

Regresores

* β's → Parámetros del modelo

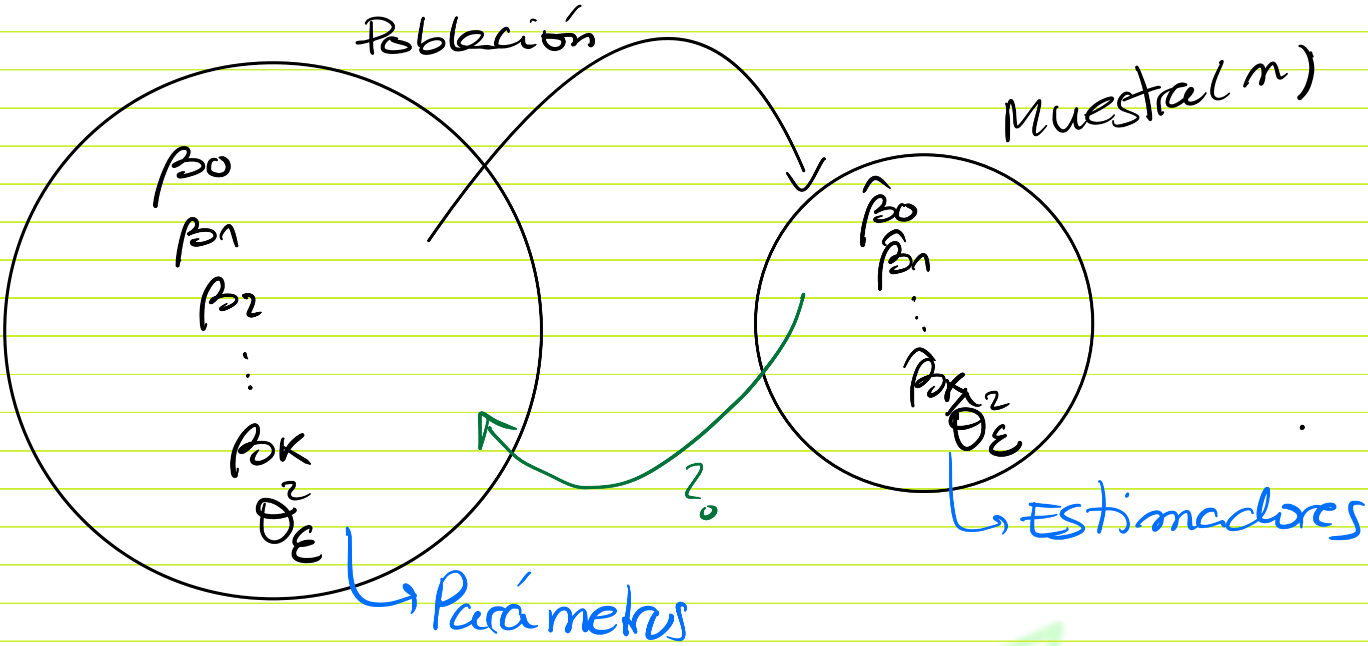
Coeficientes β_j

→ Efectos parciales de cada variable X_j sobre la Y

* ε_i → Parte aleatoria del modelo

Término de Perturbación

INFERENCIA



- **Función Esperanza condicionada (FEC)**

$$E(Y/X_1, X_2, \dots X_K) = \underbrace{\beta_0 + \beta_1 X_{1i} + \dots + \beta_K X_{Ki}}_{FRP}$$

Función de Regresión Poblacional

- **Función de Regresión Muestral (FRM)**

$$\hat{Y}_i = \hat\beta_0 + \hat\beta_1 X_{1i} + \dots + \hat\beta_K X_{Ki}$$

- ERROR DE ESTIMACIÓN. RESIDUOS

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- La variable ε se denomina **término de error** y representa todos los otros factores que además de X_1, X_2,..., X_k determinan el valor de la variable dependiente Y para una observación concreta que llamamos observación i, de manera que para cada observación i habrá un error ε_i. Este término de error es una forma de incluir el resto de factores no incluidos expresamente y que afectan a la variable regresada; por tanto tiene un papel crucial en el modelo de regresión y se tendrá que analizar su comportamiento para evaluar el modelo en todo su conjunto.

Efecto parcial de X_j sobre Y: $\beta_j = \dfrac{\Delta Y}{\Delta X_j}$

El término β_0 no está multiplicado por ninguna variable X, entonces su interpretación es más sencilla: es el valor esperado de Y, cuando $X_1 = X_2 = \ldots = X_k = 0$.

Y	X
Variable explicada	Variable explicativa
Variable dependiente	Variable independiente
Regresada	Regresora
Endógena	Exógena
Variable respuesta	Variable de control
Predicha	Predictora

Los modelos de regresión pueden ser lineales en las variables o lineales en los parámetros:

- **Lineales en las variables**: Para ser lineal en las variables, las variables X_j no puede estar elevadas a una potencia diferente de la unidad; tampoco puede estar ni multiplicado ni divididas por otra variable.
    - Ejemplo modelos **lineales en las variables**:

      $$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

      $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

    - Ejemplo modelos **no lineales en las variables**:

      $$Y = \beta_0 + \beta_1 X_1^2 + \varepsilon \rightarrow z_1 = x_1^2 \Rightarrow Y = \beta_0 + \beta_1 z_1 + \varepsilon$$

      $$Y = \beta_0 + \beta_1 (1/X_1) + \varepsilon \rightarrow z_1 = 1/x_1 \Rightarrow Y = \beta_0 + \beta_1 z_1 + \varepsilon$$

      Los modelos **no lineales en las variables se pueden linealizar** aplicando el cambio de variable apropiado ($Z_1 = X_1^2$ ó $Z_1 = 1/X_1$) en los respectivos ejemplos.

- **Lineales en los parámetros:** cuando los coeficientes β_j están multiplicados por las variables o por alguna transformación de estas, pero sin estar multiplicándose o dividiendo entre ellos, es decir sin que exista ninguna interacción entre los diferentes parámetros del modelo.

  Un modelo es no lineal en los parámetros cuando algún β_j aparece elevado a cualquier potencia distinta de la unidad o multiplicado o dividido por otro parámetro.

<u>NO LINEAL</u>

$$\beta \cdot \beta \;;\; \frac{\beta}{\beta} \;;\; \beta^2 \;;\; \ln(\beta X) \;;\; \boxed{X^{\beta}} \; !!$$

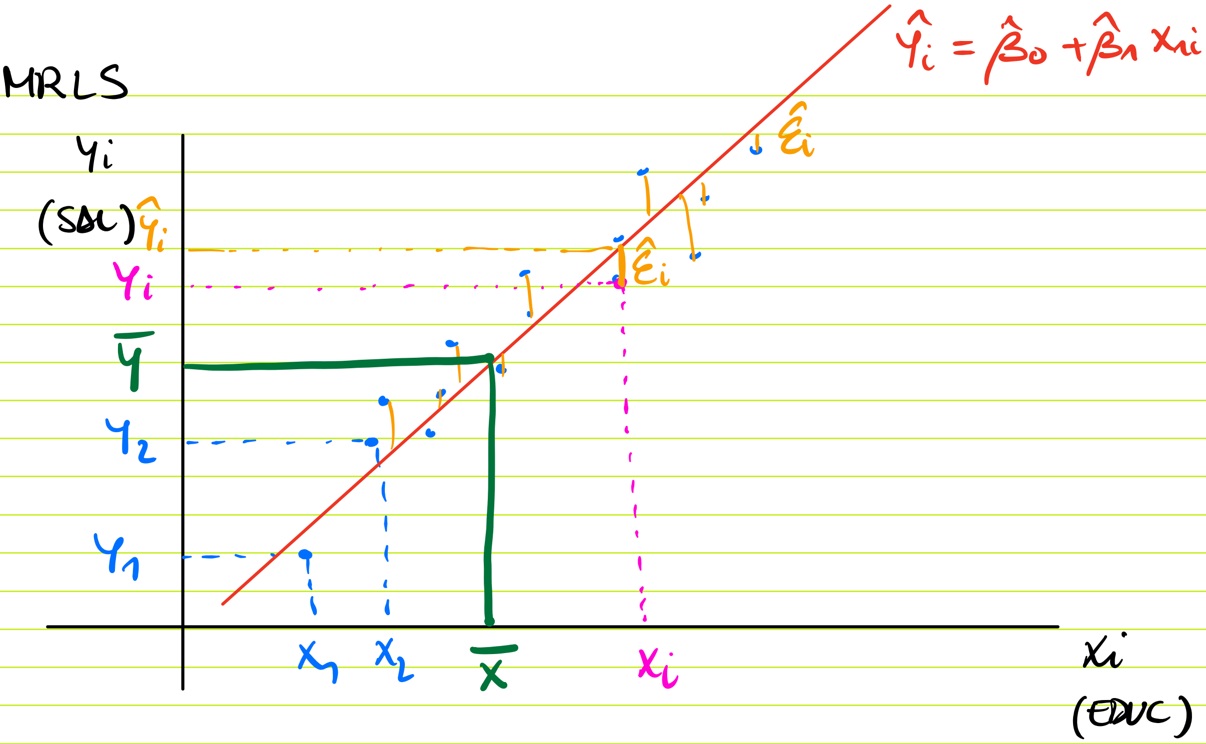
Ejemplo: $Y_i = A \cdot K^{\alpha} \cdot L^{\beta} \cdot e^{\varepsilon_i}$ → Intrínsecamente lineal

$$\ln(Y_i) = \ln(A \cdot K^{\alpha} \cdot L^{\beta} \cdot e^{\varepsilon_i}) = \ln A + \ln K^{\alpha} + \ln L^{\beta} + \ln e^{\varepsilon_i} =$$

$$= \ln A + \alpha \ln K + \beta \ln L + \varepsilon_i \overbrace{\ln e}^{1} =$$

$$= \beta_0 + \beta_1 \ln K + \beta_2 \ln L + \varepsilon_i$$

MRLS



Mínimos Cuadrados Ordinarios (MCO)

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$$\sum \hat{\varepsilon}_i = 0 \rightarrow \bar{\hat{\varepsilon}}_i = 0$$

Mínimo $\sum \hat{\varepsilon}_i^2$:

$$\left.\begin{array}{l} \dfrac{\partial \sum \hat{\varepsilon}_i^2}{\partial \beta_0} = 0 \\[2ex] \dfrac{\partial \sum \hat{\varepsilon}_i^2}{\partial \hat{\beta}_1} = 0 \end{array}\right\} \begin{array}{l} \hat{\beta}_1 = \dfrac{\operatorname{cov}(X_1, Y)}{\operatorname{Var}(X_1)} \\[2ex] \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{array}$$

Ecuaciones normales:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} \right) = 0$$

$$\sum_{i=1}^{n} X_{1i} \hat{\varepsilon}_i = \sum_{i=1}^{n} X_{1i} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} \right) = 0$$

Pendiente :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_{1i} y_i}{\sum_{i=1}^{n} x_{1i}^2} = \frac{\sum_{i=1}^{n} (X_{1i}-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n} (X_{1i}-\bar{X})^2} = \frac{\widehat{\text{cov}}(X_1, Y)}{\widehat{\text{var}}(X_1)}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_{1i} Y_i}{\sum_{i=1}^{n} X_{1i}^2} = \frac{S_{X_1,Y}}{S^2_{X_1}}$$

Ordenada:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1$$

Resultados algebraicos de la regresión en el modelo de regresión simple

- Por la primera ecuación normal, deducimos que la **media de los residuos** estimados es **nula**.

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0 \longrightarrow \bar{\hat{\varepsilon}}_i = 0$$

- A partir de: $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$

  deducimos que: $\bar{Y} = \bar{\hat{Y}} + \overset{=0}{\bar{\hat{\varepsilon}}} = \bar{\hat{Y}}$ $\Rightarrow \bar{Y} = \bar{\hat{Y}}$

- A partir de: $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\varepsilon}_i$ y $\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}_1$

  $\underbrace{Y_i - \bar{Y}}_{y_i} = \hat{\beta}_1 (\underbrace{X_1 - \bar{X}_1}_{x_{1i}}) + \hat{\varepsilon}_i$

  Restando: $y_i = \hat{\beta}_1 x_{1i} + \hat{\varepsilon}_i$

  De manera que. en desviaciones a las medias, la variable estimada por la regresión es $\hat{y}_i = \hat{\beta}_1 x_{1i}$

  Multiplicando ambas partes por los errores estimados y sumando desde i=1 hasta n y con la segunda ecuación normal se obtiene:

$$\sum_{i=1}^{n} \hat{y}_i \hat{\varepsilon}_i = \hat{\beta}_1 \sum_{i=1}^{n} x_{1i} \hat{\varepsilon}_i = 0$$

- De manera que la covarianza $\text{cov}(\hat{Y}_i, \hat{\varepsilon}_i) = 0$. Y a partir de la segunda ecuación normal tenemos que:

$$\text{cov}(X_{1i}, \hat{\varepsilon}_i) = 0$$

Entonces la variable independiente X_{1i} y los residuos $\hat{\varepsilon}_i$ están incorrelados.

A partir de: $Y_i = \hat{Y}_i + \hat{\varepsilon}_i$. La varianza es:

$$\mathrm{var}(Y_i) = \mathrm{var}(\hat{Y}_i) + \mathrm{var}(\hat{\varepsilon}_i)$$

- $SCT = \sum (y_i - \bar{y})^2 = \sum y_i^2 - n(\bar{y})^2$
- $SCE = \sum (\hat{y}_i - \bar{y})^2 = \sum \hat{y}_i^2 - n(\bar{y})^2$
- $SCR = \sum (y_i - \hat{y})^2 = \sum \hat{\varepsilon}_i^2$

$$SCT = SCE + SCR$$

Coeficiente de Determinación R^2

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

El Coeficiente de Determinación es una medida estadística de bondad de ajuste o fiabilidad del modelo estimado a los datos. Este coeficiente indica cuál es la proporción de la variación total en la variable dependiente Y, que es explicada por el modelo de regresión estimada, es decir, mide la capacidad explicativa del modelo estimado.

Sabemos: $\mathrm{var}(Y_i) = \mathrm{var}(\hat{Y}_i) + \mathrm{var}(\hat{\varepsilon}_i)$

Entonces el R^2 se define como la proporción de la varianza explicada por la regresión respecto de la varianza que queremos explicar:

$$R^2 = \frac{\mathrm{var}(\hat{Y})}{\mathrm{var}(Y)} = \frac{\mathrm{var}(Y) - \mathrm{var}(\hat{\varepsilon})}{\mathrm{var}(Y)} = 1 - \frac{\mathrm{var}(\hat{\varepsilon})}{\mathrm{var}(Y)} \qquad \left(0 \le R^2 \le 1\right)$$

$R^2\ [0, 1] \rightarrow R^2(\%) \rightarrow$ Explicación de la variable X_1 sobre la Y

$$R^2=(r_{xy})^2=\left(\frac{Cov(x,y)}{Sx\cdot Sy}\right)^2$$

Este coeficiente siempre es positivo y menor o igual que 1.

- Si R^2 = 1, la regresión explica completamente la variación de la variable dependiente, es decir, todas las observaciones estarían sobre la recta de regresión.
- Si R^2 = 0, la regresión no explicaría nada sobre el comportamiento de la variable dependiente.

El R^2 multiplicado por 100 se interpreta como el porcentaje de la variable dependiente explicado por la regresión. En ningún caso un alto coeficiente de determinación garantiza que el modelo de regresión tenga necesariamente buenas características.

El coeficiente de determinación es igual al coeficiente de correlación lineal al cuadrado.

$$R^2=(r_{XY})^2$$

$$r_{XY}=\frac{Cov(X,Y)}{S_XS_Y}=\frac{n\sum X_iY_i-(\sum X_i)(\sum Y_i)}{\sqrt{[n\sum X_i^2-(\sum X_i)^2][n\sum Y_i^2-(\sum Y_i)^2]}}$$

- Si r_{XY} = 1 tendremos una relación perfecta directa entre X e Y.
- Si r_{XY} = 0 no existirá relación lineal entre X e Y pero podrá existir otro tipo de relación no lineal entre ellas.
- Si r_{XY} = -1 tendremos una relación lineal inversa perfecta entre X e Y.
- Cuanto más cerca este el valor del coeficiente de los valores 1 ó -1 más intensa será la relación lineal que existe entre X e Y.