Planteamiento del modelo de regresión lineal simple (RLS)

Erika R. Badillo
erika.badilloen@unaula.edu.co

Facultad de Economía
Universidad Autónoma Latinoamericana

En este tema

- Una presentación intuitiva
- El concepto de perturbación aleatoria
- Especificación del modelo (Supuestos)
- Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)
- Propiedades de los estimadores MCO

Lecturas

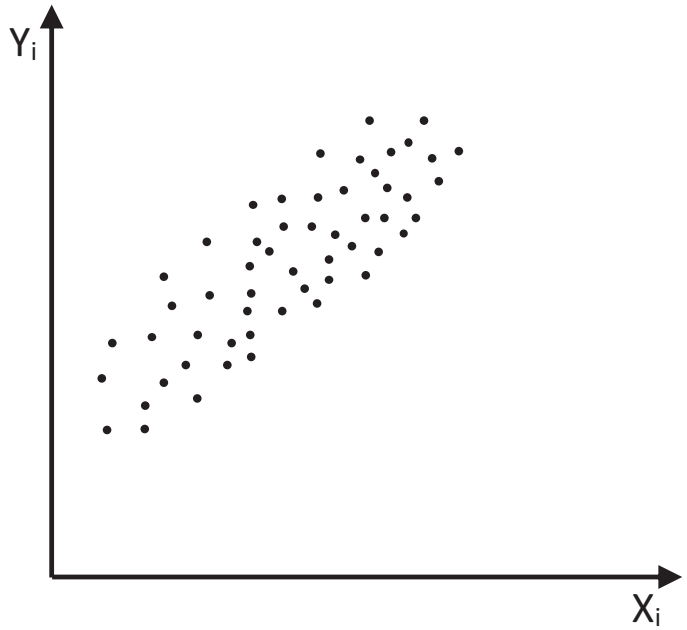
- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 2 y 3
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 2 y 3

Una presentación intuitiva

- El problema a estudiar tiene que ver con el consumo de los individuos y sus ingresos en una comunidad:

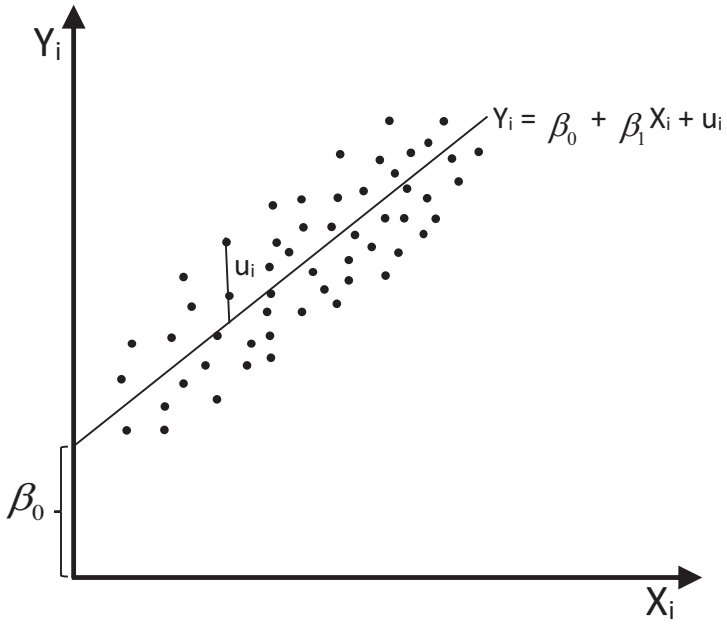
$$Y_i : \text{consumo del individuo } i$$
$$X_i : \text{ingreso del individuo } i,\ i = 1, 2, ..., n$$

- La observación de la realidad mostraría



Una presentación intuitiva

A nivel teórico qué se puede decir? Existe una relación positiva entre el consumo y el ingreso, por lo que es posible ajustar una línea recta que pase por el medio de los puntos y cada individuo se aleja positiva y negativamente de ella



La representación matemática del modelo es:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

β_0 : consumo autónomo (intercepto)
β_1 : propensión marginal a consumir el ingreso (pendiente)
u_i : perturbación aleatoria

Una presentación intuitiva

- El problema a resolver es encontrar una representación muestral del modelo:

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i$$

$\widehat{\beta}_0$: estima a β_0
$\widehat{\beta}_1$: estima a β_1
$\widehat{u}_i$: es la contraparte muestral de u_i

- Este ejercicio permite responder otras preguntas que subyacen de la teoría:
  - El consumo autónomo es positivo
  - El consumo autónomo es 100
  - La propensión marginal a consumir es 0.8
- El ejercicio econométrico busca ver si los datos contradicen o no las hipótesis teóricas
- No hay teorías verdaderas sino modelos útiles
- Si los datos no contradicen las hipótesis el modelo puede ser útil
- Para poder hacer este ejercicio se requiere la inferencia estadística. Esto implica hacer supuestos acerca de u_i

Una presentación intuitiva

- Otro ejemplo: Un modelo en el que se relaciona el salario y la educación

$$salario_i = \beta_0 + \beta_1 educ_i + u_i$$

donde $salario$: dólares por hora, $educ$: años de educación

entonces β_1 indica la variación en el salario por hora por cada año adicional de educación (permaneciendo todos los demás factores constantes)

Factores no observados: experiencia laboral, capacidades innatas, antigüedad en el empleo actual, ética laboral, etc

El concepto de perturbación aleatoria

El término de perturbación (inobservables o factores no observables, que en conjunto, afectan a Y):

- Sirve como sustituto de todas las variables omitidas del modelo (imprecisión de la teoría o falta de disponibilidad de datos)
- La influencia conjunta de algunas variables es muy pequeña y no se justifica su introducción explícita en el modelo. Principio de parsimonia: "Conviene mantener el modelo de regresión lo más sencillo posible". Si el comportamiento de Y se puede explicar muy bien con dos o tres variables explicativas, ¿para qué incluir más? No se deben excluir variables pertinentes sólo para que el modelo de regresión no se complique
- Errores de medición (variables proxy inadecuadas)
- Forma funcional incorrecta

El concepto de perturbación aleatoria

- Perturbación aleatoria: aquella que hace compatible la realidad y la teoría:

$$u_i = \underbrace{Y_i}_{\text{Realidad}} - \underbrace{\beta_0 + \beta_1 X_i}_{\text{Teoría}}$$

- Al considerar que u_i es una variable aleatoria tiene sentido hablar de sus características y los supuestos que se deben hacer sobre éstas

Característica	**Definición matemática**	**Contraparte muestral**
Media	$E(u_i)$	$\bar{\widehat{u}}_i = \frac{\sum \widehat{u}_i}{n}$
Varianza	$E(u_i - E(u_i))^2$	$\widehat{\sigma}^2_{\widehat{u}_i} = \frac{\sum(\widehat{u}_i - \bar{\widehat{u}})^2}{n} = \frac{\sum \widehat{u}_i^2}{n}$
Covarianza	$E[(u_i - E(u_i))(u_j - E(u_j))]$	$\frac{1}{n-1}\sum(\widehat{u}_i - \bar{\widehat{u}})(\widehat{u}_j - \bar{\widehat{u}})$

- Hay también una distribución muestral asociada, por ejemplo:

$$u_i \sim N(E(u_i); E(u_i - E(u_i))^2)$$

Especificación del modelo

El modelo de RLS tiene la siguiente especificación:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Y	X
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictora
Regresando	Regresor

Especificación del modelo: Supuestos

- El modelo RLS se especifica así:

  - Lineal en los parámetros

  - β_0 y β_1: coeficientes fijos (parámetros)

  - Modelo completo $E(u_i) = 0$

  - Homocedasticidad $Var(u_i) = E(u_i^2) = \sigma_u^2$ (este es el otro parámetro del modelo)

  - No autocorrelación $Cov(u_i, u_j) = E(u_i u_j) = 0$, $i \neq j$

- Supuestos sobre X_i:

  - X_i es estocasticamente fija, no es aleatoria, esta predeterminada antes de observar a Y_i

  - X_i no aleatoria corresponde a situaciones de laboratorio donde se puede controlar un experimento y fijar ex-ante los valores de la variable explicatoria X_i

  - Pero en economía esto no sucede, normalmente se observan Y_i y X_i al mismo tiempo

  - Lo más delicado en economía es que X_i en otro modelo pueda ser la variable a explicar

Especificación del modelo: Supuestos

- Para resolver esta situación Haavelmo (1948) formuló la hipótesis de exogeneidad: si la variable explicatoria es de naturaleza aleatoria, debe ser estadísticamente independiente de la perturbación aleatoria

$$\begin{aligned} Cov(X_i, u_i) &= E[(X_i - E(X_i))(u_i - E(u_i))] \\ &= E[(X_i - E(X_i))u_i] \\ &= E[(X_i - E(X_i))]E(u_i) \\ &= 0 \end{aligned}$$

- Hipótesis de normalidad $u_i \sim NID(0; \sigma_u^2)$

Especificación del modelo: Supuestos

- En resumen, los supuestos acerca de u_i:
  1. $E(u_i) = 0 \Longrightarrow$ modelo completo
  2. $Var(u_i) = E[(u_i - E(u_i))^2] = E(u_i^2) = \sigma_u^2 \Longrightarrow$ homocedasticidad
  3. $Cov(u_i, u_j) = E[(u_i - E(u_i))(u_j - E(u_j))] = E(u_i u_j) = 0,\ i \neq j \Longrightarrow$ no autocorrelación

  4. $Cov(X_i, u_i) = 0 \Longrightarrow$ exogeneidad
  5. $u_i \sim NID(0; \sigma_u^2) \Longrightarrow$ normalidad

Especificación del modelo: Supuestos

- Regresión: Dependencia de una variable respecto de otra(s) con el objetivo de estimar el valor promedio poblacional de Y en términos de X (valores conocidos). Por ejemplo, predecir el consumo semanal promedio (Y) de la población en su conjunto para valores dados del ingreso (X)

- Saber algo sobre X no permite saber nada sobre u, de tal forma que, el valor promedio de u no depende del valor de X:

$$E(u|X) = E(u) = 0$$

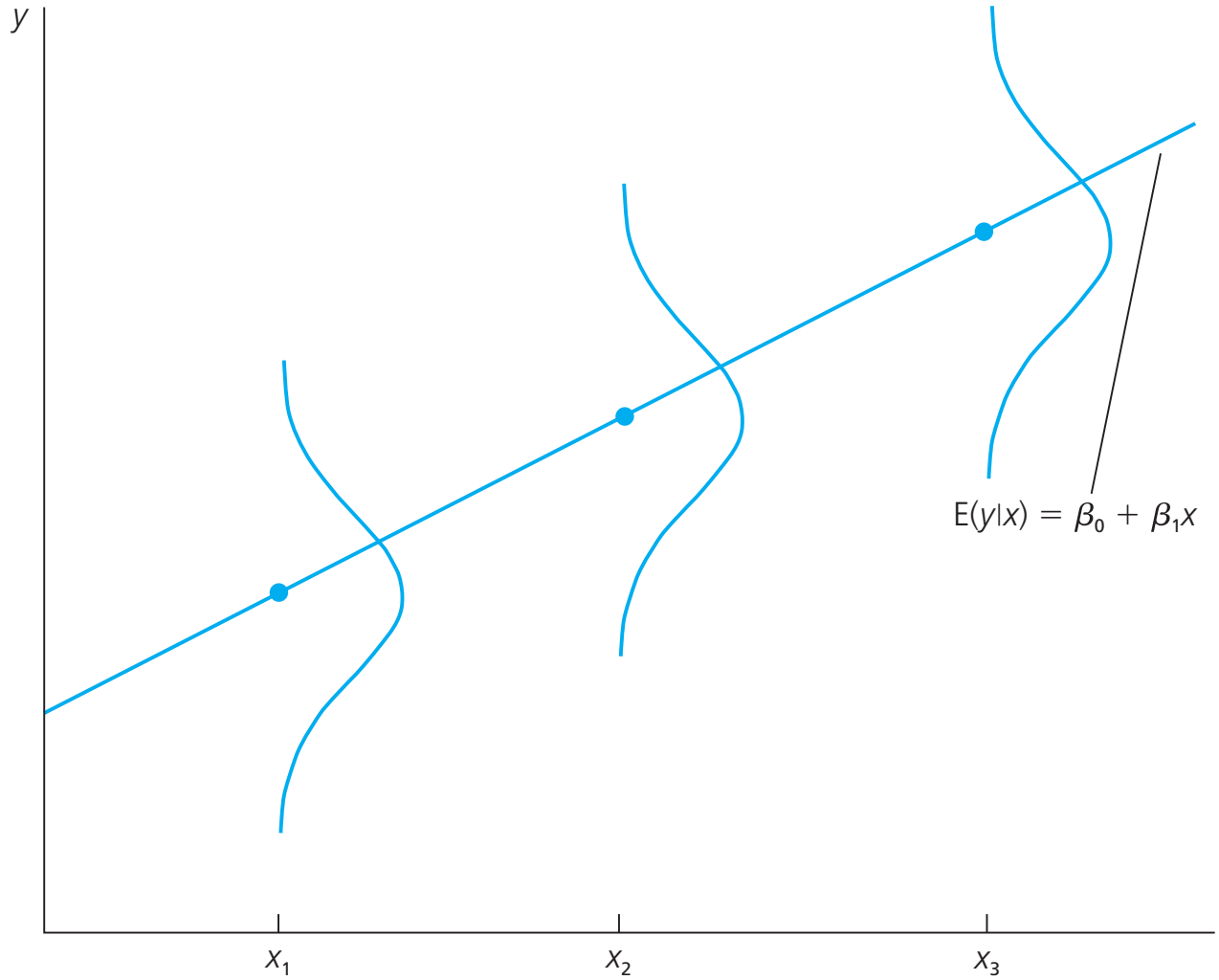
  lo cual implica que $E(Y|X) = \beta_0 + \beta_1 X$ Función de Regresión Poblacional (FRP)

- $E(Y|X)$ es una función lineal de X: por cada aumento de una unidad en X el valor esperado de Y se modifica en la cantidad β_1

Especificación del modelo: Supuestos

- Para todo X, la distribución de Y está centrada en $E(Y|X)$

FIGURE 2.1 E($y|x$) as a linear function of x.



Especificación del modelo: Supuestos

El objetivo principal de un análisis de regresión es estimar la FRP:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

con base en la FRM:

$$Y_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i + \widehat{u}_i$$

La idea es diseñar una regla o método que "acerque" la FRM lo más posible a la FRP

Obtención de los estimadores MCO

Tres opciones (existen más) para estimar β_0 y β_1 en el modelo RLS $Y_i = \beta_0 + \beta_1 X_i + u_i$:

- Minimizar la SCR (Suma de Cuadrados de los Residuales $\sum \widehat{u}_i^2$)
- Método de los momentos (usa supuestos paramétricos)
- Maximizar la función de verosimilitud (supone una distribución normal)

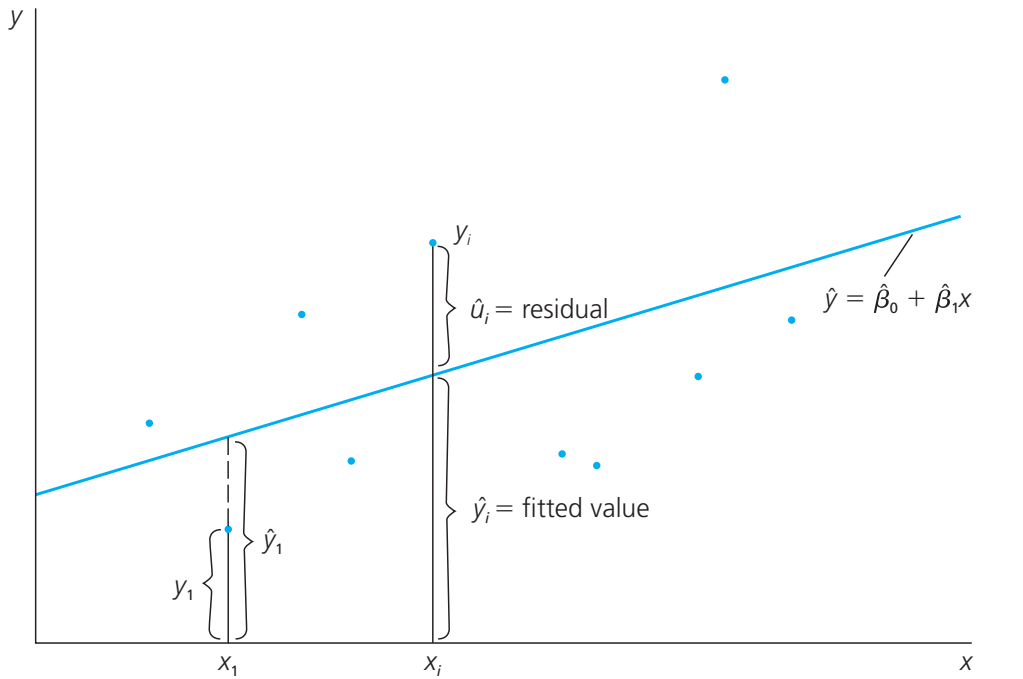
Minimizando la SCR

Es un método de ajuste de curvas, geométrico, que no establece supuestos. Lo único que establece es que existe un residuo en las estimaciones

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i \Longrightarrow \text{modelo estimado}$$

$\widehat{u}_i$: residuo en la estimación

FIGURE 2.4 Fitted values and residuals.



Minimizando la SCR

$\widehat{\beta}_0$ y $\widehat{\beta}_1$ son aquellos que resultan de minimizar libremente la SCR ($\sum \widehat{u}_i^2$)

$$\sum \widehat{u}_i^2 = \sum (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2$$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_0} = -2 \sum (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) = 0$$

$$\sum Y_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum X_i \quad (1)$$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_1} = -2 \sum (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) X_i = 0$$

$$\sum X_i Y_i = \widehat{\beta}_0 \sum X_i + \widehat{\beta}_1 \sum X_i^2 \quad (2)$$

(1) y (2) se llaman ecuaciones normales y al resolverlas aparecen los estimadores MCO

Obtención de los estimadores MCO

Minimizando la SCR

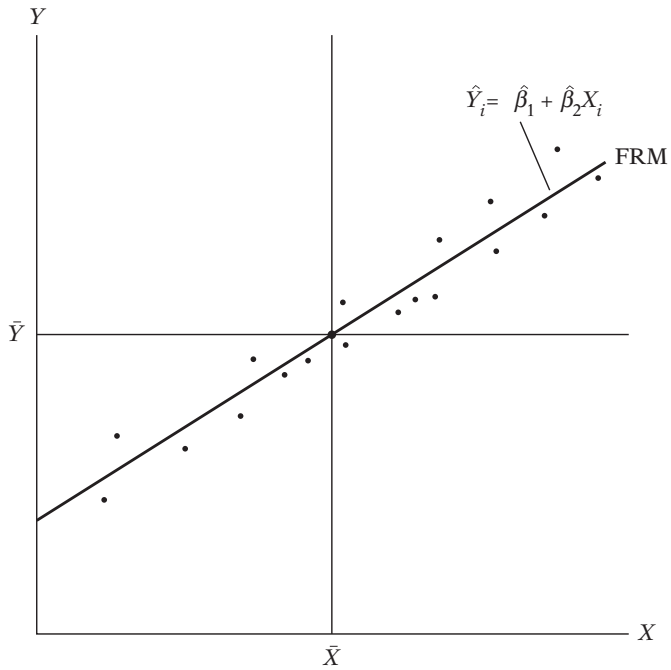
Si dividimos la ecuación (1) por n tenemos

$$\frac{\sum Y_i}{n} = \frac{n\widehat{\beta}_0 + \widehat{\beta}_1 \sum X_i}{n}$$

$$\bar{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{X}$$

Quiere decir que $(\bar{X}, \bar{Y})$ como punto esta situado sobre la recta mínima cuadrática

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i \text{ ó } Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{u}_i$$



$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} \quad (3)$$

Minimizando la SCR

Volviendo sobre la derivada de β_1 y empleando (3) para sustituir $\widehat{\beta}_0$, se obtiene

$$\sum(Y_i - (\bar{Y} - \widehat{\beta}_1\bar{X}) - \widehat{\beta}_1 X_i)X_i = 0$$
$$\sum X_i(Y_i - \bar{Y}) = \widehat{\beta}_1 \sum X_i(X_i - \bar{X})$$
$$\widehat{\beta}_1 = \frac{\sum X_i(Y_i - \bar{Y})}{\sum X_i(X_i - \bar{X})}$$

Es posible demostrar que

$$\sum X_i(Y_i - \bar{Y}) = \sum(X_i - \bar{X})(Y_i - \bar{Y})$$
$$\sum X_i(X_i - \bar{X}) = \sum(X_i - \bar{X})^2$$

Por tanto, la pendiente estimada es

$$\widehat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2} \quad (4)$$

donde $x_i = X_i - \bar{X}$ y $y_i = Y_i - \bar{Y}$

Minimizando la SCR

En resumen

- Al minimizar la SCR$=\sum \widehat{u}_i^2$ se obtuvo la primera ecuación normal:

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

- Con la segunda ecuación normal y reemplazando $\widehat{\beta}_0$ se obtuvo:

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

Minimizando la SCR

Otro parámetro a estimar en el modelo de regresión es la varianza de los residuales ($\widehat{\sigma}_u^2$). Sabemos por el supuesto de homocedasticidad que

$$\sigma_u^2 = E(u^2)$$

así que un estimador de σ_u^2 es

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{n} = \frac{SCR}{n}$$

Es posible corroborar si este estimador es insesgado o no

Propiedades de los estimadores MCO

La pregunta ahora es qué pasa con las propiedades de los estimadores MCO a la luz de los supuestos. Se trata del encuentro de dos mundos:

- Lo teórico $\Longrightarrow$ $Y_i = \beta_0 + \beta_1 X_i + u_i$, β_0 y β_1 fijos

$$E(u_i) = 0;\ Var(u_i) = \sigma_u^2;\ Cov(u_i, u_j) = 0;$$
$$Cov(X_i, u_i) = 0;\ u_i \sim NID(0; \sigma_u^2)$$

- Lo empírico $\Longrightarrow$ $\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}$

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

Se demostrará que los estimadores MCO son ELIO (o BLUE)

- Estimadores
- Lineales
- Insesgados
- Óptimos

Propiedades de los estimadores MCO

- Linealidad: los estimadores MCO son polinomios lineales en Y_i y u_i
- Insesgadez: $E(\widehat{\beta}_1) = \beta_1$ y $E(\widehat{\beta}_0) = \beta_0$
- Óptimos: dentro de la clase de estimadores lineales e insesgados del modelo, los estimadores MCO tienen la mínima varianza, dentro de los estimadores que utilizan igual cantidad de información (Teorema de Gauss-Markov)

  Mínima varianza = Máxima precisión

  Nos interesa dos cosas:
  - Encontrar la estimación de la varianza de $\widehat{\beta}_1$ y $\widehat{\beta}_0$
  - Demostrar que dicha varianza es mínima $\Longrightarrow$ Teorema de Gauss-Markov

  Varianzas:

$$Var(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum x_i^2}$$

$$Var(\widehat{\beta}_0) = \frac{\sigma_u^2 \sum X_i^2}{n \sum x_i^2}$$