Tema 3. El modelo de regresión lineal simple (RLS)

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Una presentación intuitiva
- El concepto de perturbación aleatoria
- En resumen
- Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)
- Propiedades de los estimadores MCO
- Ejercicio aplicado en R

Lecturas

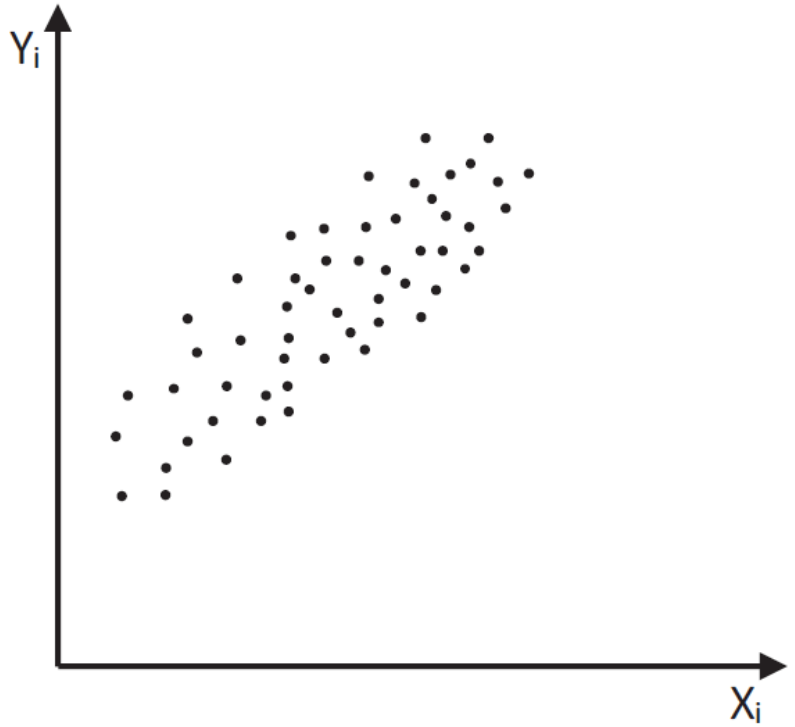
- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 2 y 3
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 2 y 3

Una presentación intuitiva

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

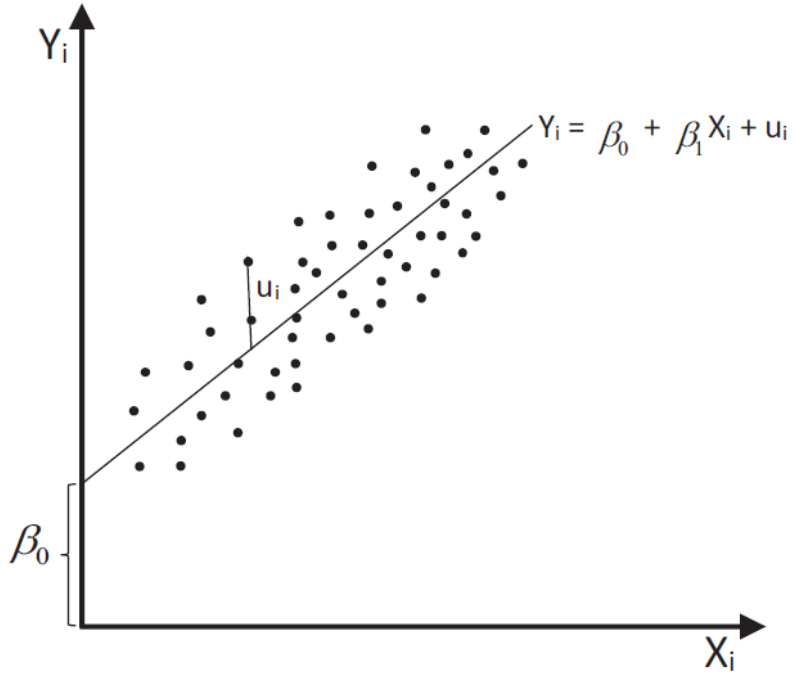
$$Y_i \text{ : consumo del individuo } i$$
$$X_i \text{ : ingreso del individuo } i,\ i = 1, 2, \ldots, n$$

- 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:

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

$\widehat{\beta}_0$: 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

El concepto de perturbación aleatoria

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

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

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

- Al considerar que u_i es una variable aleatoria tiene sentido hablar de sus características y los supuestos que se deben hacer sobre éstas
- Hay también una distribución muestral asociada, por ejemplo:

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

El concepto de perturbación aleatoria

Para completar la especificación del modelo de RLS se requiere hacer supuestos acerca de u_i:

- $E(u_i) = 0 \Longrightarrow$ modelo completo
- $Var(u_i) = E(u_i - E(u_i))^2 = E(u_i^2) = \sigma_u^2 \Longrightarrow$ homocedasticidad
- $Cov(u_i, u_j) = E[(u_i - E(u_i))(u_j - E(u_j))] = E(u_i u_j) = 0, i \neq j \Longrightarrow$ no autocorrelación
- $u_i \sim NID(0; \sigma_u^2) \Longrightarrow$ normalidad

En resumen

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

En resumen

- El modelo RLS se especifica así:
  - 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. Esto en econometría se refiere a que X_i es endógena, violando uno de los supuesto importantes que X_i debe ser exogena

En resumen

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

Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)

Los tres métodos (existen más) más comúnmente utilizados para estimar β_0 y β_1 en le modelo RLS $Y_i = \beta_0 + \beta_1 X_i + u_i$ son:

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

Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)

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

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

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

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

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

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

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

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

Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)

MCO: Minimizando la SCR

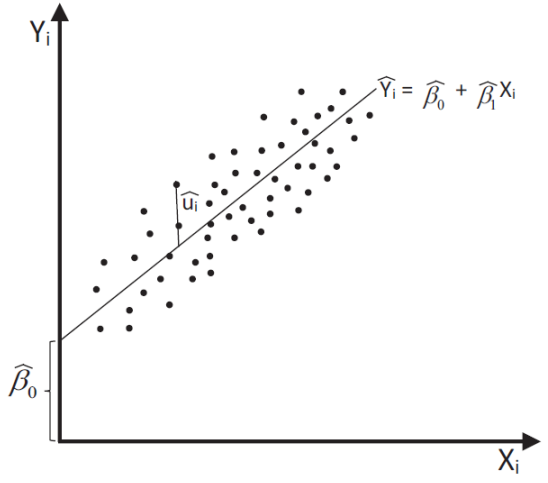
Si dividimos la ecuación (1) por n tenemos

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

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

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

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



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

Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)

MCO: Minimizando la SCR

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

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

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

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

Es posible demostrar que

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

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

Por tanto, la pendiente estimada es

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

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

Obtención de los estimadores Mínimos Cuadrados Ordinarios (MCO)

MCO: 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 $\hat{\beta}_0$ se obtuvo:

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

Que representan los estimadores MCO de β_0 y β_1

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

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \beta_0, \beta_1 \text{ son fijos}$$

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

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

- Lo empírico

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

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

Es posible demostrar 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: $E(\hat{\beta}_1) = \beta_1$ y $E(\hat{\beta}_0) = \beta_0$
- 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

Ejercicio aplicado en R

Los accidentes de tráfico son la principal causa de muerte de los estadounidenses entre los 5 y los 32 años de edad. Mediante distintas políticas de gasto, el gobierno federal ha alentado a los estados a instituir normativas de obligatoriedad de uso del cinturón de seguridad para reducir el número de muertes y lesiones graves.

En este ejercicio se investigará la eficacia de estas leyes para el aumento del uso del cinturón de seguridad y la reducción de víctimas mortales. El archivo SeatBelts.xls contiene un panel de datos sobre 50 estados de los EE.UU., además del distrito de Columbia para los años 1983-1997. Se ofrece una descripción detallada en el archivo SeatBelts_Description.pdf

- Estime el efecto del uso del cinturón de seguridad sobre las muertes mediante la regresión de la variable $fatalityrate$ sobre la variable sb_useage. ¿La regresión estimada sugiere que un mayor uso del cinturón de seguridad reduce las muertes?
- Interprete los resultados
- ¿Cuántas vidas se salvarían si el uso del cinturón de seguridad aumentara de 52% a 90%?