# Tema 5. El modelo de Regresión Lineal Múltiple (RLM)

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

- El modelo
- El modelo en matrices
- Estimador MCO y propiedades
- Sesgo por variables omitidas
- El coeficiente de determinación $R^2$
- Inferencia estadística
- Ejercicio aplicado en R

# Lecturas

- Wooldridge, Jeffrey (2013). *Introducción a la econometría*. 5a edición, Cengage Learning. Cap. 3, 4 y 5
- Gujarati, D. y Porter, D. (2010). *Econometría*. 5a edición, Mc Graw Hill. Cap. 7 y 8

# El modelo

El modelo de regresión lineal múltiple tiene la siguiente estructura:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + u_i$$

El modelo entonces contiene $k$ parámetros poblacionales (desconocidos)

Se trabaja, entonces, con $n - k$ gdl

La pregunta que surge es: ¿será suficiente decir que el modelo de RLM es una nueva extensión desde el modelo RLS $\Longrightarrow$ modelo RLM?

# El modelo

Recordando las hipótesis de partida (supuestos iniciales):

- Los coeficientes $\beta_j$ con $j = 1, 2, 3, \ldots, k$ son fijos y desconocidos
- Las $X_{ij}$ son estocásticamente fijas para $j = 2, 3, 4, \ldots, k$. Este es un supuesto de partida propio del laboratorio. Un supuesto más real en economía y que lleva a resultados similares es: las variables explicatorias son exógenas. Esto implica que:

$$Cov(X_{i2}, u_i) = 0$$

$$Cov(X_{i3}, u_i) = 0$$

$$\vdots$$

$$Cov(X_{ik}, u_i) = 0$$

- El modelo esta completo: $E(u_i) = 0, \forall i = 1...n$
- Homoscedasticidad: $Var(u_i) = E(u_i - E(u_i))^2 = E(u_i^2) = \sigma_u^2$
- No autocorrelación: $Cov(u_i, u_j) = E[(u_i - E(u_i))(u_j - E(u_j))] = E(u_i u_j) = 0, \forall i \neq j$
- Normalidad: $u_i \sim NID(0, \sigma_u^2)$

# El modelo en matrices

El modelo general es un polinomio de regresión de la forma:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_k X_{ik} + u_i$$

Lo que el modelo dice es:

$$Y_1 = \beta_1 + \beta_2 X_{12} + \beta_3 X_{13} + \ldots + \beta_k X_{1k} + u_1$$

$$Y_2 = \beta_1 + \beta_1 X_{22} + \beta_3 X_{23} + \ldots + \beta_k X_{2k} + u_2$$

$$\vdots$$

$$Y_n = \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + \ldots + \beta_k X_{nk} + u_n$$

$$\mathbf{Y}_{n\text{x}1} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{u}_{n\text{x}1} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_n \end{bmatrix} \quad \mathbf{B}_{k\text{x}1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} \quad \mathbf{X}_{n\text{x}k} = \begin{bmatrix} 1 & X_{12} & X_{23} & X_{1k} \\ 1 & X_{22} & X_{23} & X_{2k} \\ 1 & X_{32} & X_{33} & X_{3k} \\ \vdots & & & \\ 1 & X_{n2} & X_{n3} & X_{nk} \end{bmatrix}$$

En esta notación $X_{ij}$ indica: fila $i$ (observación), columna $j$ (variable explicatoria). El polinomio de regresión en álgebra matricial se puede escribir como:

$$\mathbf{Y}_{n\text{x}1} = \mathbf{X}_{n\text{x}k}\mathbf{B}_{k\text{x}1} + \mathbf{u}_{n\text{x}1}$$

# El modelo en matrices

Se puede intentar aplicar MCO para obtener los estimadores del modelo. Partiendo de un modelo estimado con intercepto y tres variables explicativas:

$$Y_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_{i2} + \widehat{\beta}_3 X_{i3} + \widehat{\beta}_4 X_{i4} + \widehat{u}_i$$

$u_i$ es el residuo de la regresión

Se construye la SCR:

$$\sum \widehat{u}_i^2 = \sum (Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{i2} - \widehat{\beta}_3 X_{i3} - \widehat{\beta}_4 X_{i4})^2$$

Se minimiza la SCR respecto a los $\beta_i, i = 1, 2, 3, 4$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_1} = -2 \sum (Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{i2} - \widehat{\beta}_3 X_{i3} - \widehat{\beta}_4 X_{i4}) = 0$$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_2} = -2 \sum (Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{i2} - \widehat{\beta}_3 X_{i3} - \widehat{\beta}_4 X_{i4}) X_{i2} = 0$$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_3} = -2 \sum (Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{i2} - \widehat{\beta}_3 X_{i3} - \widehat{\beta}_4 X_{i4}) X_{i3} = 0$$

$$\frac{\partial \sum \widehat{u}_i^2}{\partial \widehat{\beta}_4} = -2 \sum (Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{i2} - \widehat{\beta}_3 X_{i3} - \widehat{\beta}_4 X_{i4}) X_{i4} = 0$$

# El modelo en matrices

Reordenando términos se obtienen las 4 ecuaciones normales:

$$\sum Y_i = n\widehat{\beta}_1 + \widehat{\beta}_2 \sum X_{i2} + \widehat{\beta}_3 \sum X_{i3} + \widehat{\beta}_4 \sum X_{i4}$$

$$\sum Y_i X_{i2} = \widehat{\beta}_1 \sum X_{i2} + \widehat{\beta}_2 \sum X_{i2}^2 + \widehat{\beta}_3 \sum X_{i3} X_{i2} + \widehat{\beta}_4 \sum X_{i4} X_{i2}$$

$$\sum Y_i X_{i3} = \widehat{\beta}_1 \sum X_{i3} + \widehat{\beta}_2 \sum X_{i2} X_{i3} + \widehat{\beta}_3 \sum X_{i3}^2 + \widehat{\beta}_4 \sum X_{i4} X_{i3}$$

$$\sum Y_i X_{i4} = \widehat{\beta}_1 \sum X_{i4} + \widehat{\beta}_2 \sum X_{i2} X_{i4} + \widehat{\beta}_3 \sum X_{i3} X_{i4} + \widehat{\beta}_4 \sum X_{i4}^2$$

¿Cómo resolver este sistema?

Al intentar resolver el sistema para despejar los $\beta_j$ se encuentra que es "costoso" hacerlo con la notación que se tiene. Lo que se puede hacer es una agrupación en matrices

# El modelo en matrices

Entonces tenemos las 4 ecuaciones normales:

$$\sum Y_i = n\widehat{\beta}_1 + \widehat{\beta}_2 \sum X_{i2} + \widehat{\beta}_3 \sum X_{i3} + \widehat{\beta}_4 \sum X_{i4}$$

$$\sum Y_i X_{i2} = \widehat{\beta}_1 \sum X_{i2} + \widehat{\beta}_2 \sum X_{i2}^2 + \widehat{\beta}_3 \sum X_{i3}X_{i2} + \widehat{\beta}_4 \sum X_{i4}X_{i2}$$

$$\sum Y_i X_{i3} = \widehat{\beta}_1 \sum X_{i3} + \widehat{\beta}_2 \sum X_{i2}X_{i3} + \widehat{\beta}_3 \sum X_{i3}^2 + \widehat{\beta}_4 \sum X_{i4}X_{i3}$$

$$\sum Y_i X_{i4} = \widehat{\beta}_1 \sum X_{i4} + \widehat{\beta}_2 \sum X_{i2}X_{i4} + \widehat{\beta}_3 \sum X_{i3}X_{i4} + \widehat{\beta}_4 \sum X_{i4}^2$$

Matricialmente las anteriores ecuaciones se pueden resumir en:

$$\begin{pmatrix} \sum Y_i \\ \sum Y_i X_{i2} \\ \sum Y_i X_{i3} \\ \sum Y_i X_{i4} \end{pmatrix} = \mathbf{X'Y}_{4\mathrm{x}1} \quad \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{\beta}_3 \\ \widehat{\beta}_4 \end{pmatrix} = \widehat{\mathbf{B}}_{4\mathrm{x}1} \quad \begin{pmatrix} n & \sum X_{i2} & \sum X_{i3} & \sum X_{i4} \\ \sum X_{i2} & \sum X_{i2}^2 & \sum X_{i3}X_{i2} & \sum X_{i4}X_{i2} \\ \sum X_{i3} & \sum X_{i2}X_{i3} & \sum X_{i3}^2 & \sum X_{i4}X_{i3} \\ \sum X_{i4} & \sum X_{i2}X_{i4} & \sum X_{i3}X_{i4} & \sum X_{i4}^2 \end{pmatrix} = \mathbf{X'X}_{4\mathrm{x}4}$$

# El modelo en matrices

Con esta notación matricial se pueden reescribir las 4 ecuaciones normales:

$$(\mathbf{X'Y})_{4X1} = (\mathbf{X'X})_{4X4}\widehat{\mathbf{B}}_{4X1}$$

Multiplicando por $(\mathbf{X'X})^{-1}$, se obtiene

$$(\mathbf{X'X})^{-1}(\mathbf{X'Y}) = \widehat{\mathbf{B}}$$

Para que $\widehat{\mathbf{B}}$ se pueda calcular se requiere que $(\mathbf{X'X})^{-1}$ exista, esto implica que:

- $|\mathbf{X'X}| \neq 0$
- que todas las filas y columnas de $\mathbf{X'X}$ sean linealmente independientes entre si

Todo lo anterior conlleva a que el rango de $\mathbf{X'X}$ debe ser cuatro en este ejemplo:

$$\rho(\mathbf{X'X}) = 4$$

Así pues, aparece una nueva hipótesis de partida: condición de no multicolinealidad perfecta, esto es que $\rho(\mathbf{X'X}) =$completo

Queda claro que con el álgebra de sumatorias es muy engorroso obtener $\widehat{\mathbf{B}}$, además, no deja ver el supuesto de multicolinealidad. Lo económico es usar el álgebra matricial

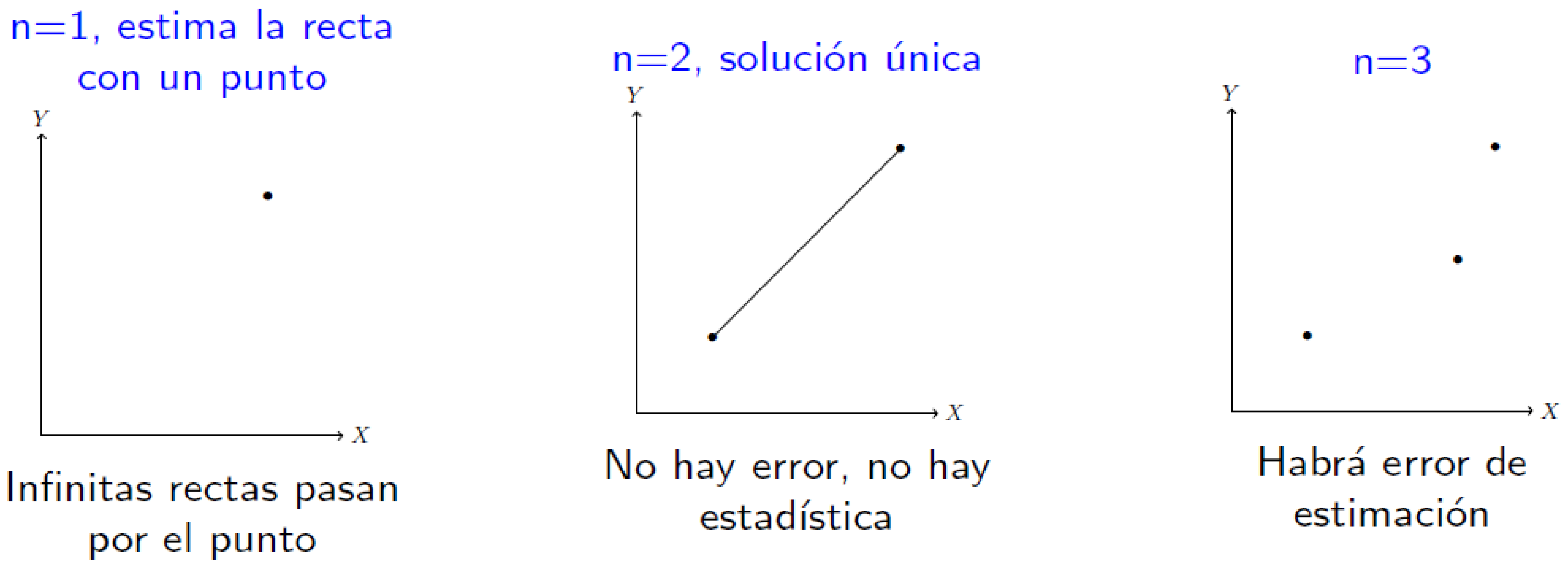
# El modelo en matrices

El supuesto de no multicolinealidad perfecta se formuló inicialmente en álgebra lineal:

$$\rho(\mathbf{X}_{n\mathrm{x}k}) = k\text{: La matriz } \mathbf{X} \text{ es de rango completo}$$

Para que el mundo de los estadístico tenga sentido se requiere que $k$ sea menor que $n$. La intuición es que si voy a estimar $k$ parámetros a partir de $n$ observaciones, el número de incógnitas debe ser menor al número de observaciones

En el caso de RLS se ve así $(k = 2)$:




Esta intuición la confirma los grados de libertad (gdl) del modelo $(n - k)$, en el RLS $(n - 2)$. Sólo para $n \geq 3$ habrá gdl positivos

# El modelo en matrices

Muchas veces el supuesto de no multicolinealidad perfecta se escribe como:

$$\underbrace{\rho(\mathbf{X}_{n\text{x}k})}_{\text{No multico. perfecta}} = \underbrace{k < n}_{\text{Sentido de la estimación}}$$

En resumen, en la especificación del modelo de Regresión Lineal Múltiple (RLM) haciendo uso de la notación matricial se tiene:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{u}$$

- Modelo completo: $E(\mathbf{u}) = \mathbf{0}$
- Exogeneidad: $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$
- Perturbaciones esféricas: $Cov(\mathbf{u}) = E(\mathbf{uu}') = \sigma_u^2\mathbf{I}_n$ (homocedasticidad y no autocorrelación)
- No multicolinealidad perfecta: $\rho(\mathbf{X}_{n\text{x}k}) = k < n$
- Normalidad: $\mathbf{u}_{n\text{x}1} \sim \mathbf{N}(\mathbf{0}_{n\text{x}1}, \sigma_u^2\mathbf{I}_n)$

# Estimador MCO y propiedades

Los estimadores MCO en términos matriciales tienen la siguiente estructura:

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Las propiedades del estimador MCO son:

- Linealidad: $\hat{\mathbf{B}}$ es lineal respecto a $\mathbf{Y}$ y $\mathbf{u}$
- Insesgadez: $E(\hat{\mathbf{B}}) = \mathbf{B}$
- Mínima varianza: $Cov(\hat{\mathbf{B}}) = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$

# Sesgo por variables omitidas

**Inclusión de variables irrelevantes en un modelo de regresión o sobrespecificación del modelo**
Suponga que se especifica un modelo como:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + u_i$$

$u_i$ satisface los supuestos estándar pero supongamos que $X_{i4}$ no tiene ningún efecto sobre $Y_i$ una vez que $X_{i2}$ y $X_{i3}$ se han controlado, lo que significa que $\beta_4 = 0$

Como no se sabe que $\beta_4 = 0$, se estima el modelo con $X_{i4}$:

$$Y_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_{i2} + \widehat{\beta}_3 X_{i3} + \widehat{\beta}_4 X_{i4} + \widehat{u}_i$$

¿Qué efecto tiene incluir un regresor irrelevante?:

- no afecta el insesgamiento de los estimadores MCO
- puede tener efectos indeseables en la varianza de los estimadores MCO

**Sesgo por variables omitidas**
Recordemos que la correlación entre una sola variable explicativa y el error, por lo general, da como resultado que todos los estimadores de MCO sean sesgados. Suponga ahora un modelo con dos variables explicativas y por error en la especificación se omite $X_{i3}$ y el modelo se estima como:

$$Y_i = \tilde{\beta}_1 + \tilde{\beta}_2 X_{i2} + \tilde{u}_i$$

Suponga que $X_{i2}$ y $X_{i3}$ están correlacionadas. Ya que $X_{i3}$ se va para el término de error y dada la correlación con $X_{i2}$, el error y $X_{i2}$ estarán correlacionados y por tanto todos los estimadores MCO estarán sesgados

# El coeficiente de determinación $R^2$

Se parte de la identidad del análisis de varianza

SCT = SCM + SCR

SCT = sumatoria de cuadrados total $\left(\sum(Y_i - \bar{Y})^2\right) \implies$ una medida de la variación total en $Y$ respecto a la media muestral
SCM = sumatoria de cuadrados del modelo $\sum(\hat{Y}_i - \bar{Y})^2 \implies$ qué parte de la variación total en $Y$ es explicada por $X$
SCR = sumatoria de cuadrados de los residuos $\left(\sum \hat{u}_i^2\right) \implies$ qué parte de la variación total en $Y$ no es es explicada por $X$

Se tiene entonces que

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum \hat{u}_i^2$$

Este es el origen de la definición del $R^2$

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

Algunas consideraciones

- Si $R^2 = 0 \implies SCR = SCT$ Quiere decir que los $X$ no agregan explicación al modelo
- Si $R^2 = 1 \implies SCM = SCT$ No hay residuos. Todos los puntos están sobre el plano de regresión, esto sucede con identidades o tautologías

# El coeficiente de determinación $R^2$

- Si el modelo no tiene intercepto el $R^2$ pierde su significado y puede dar negativo. No es comparable con el $R^2$ de modelos con intercepto
- Si son datos de series de tiempo (crecientes o decrecientes) que se mueven en el mismo sentido, el $R^2$ tiende a 1. En datos de corte transversal el $R^2$ es bajo
- Una limitación del $R^2$ en el modelo de RLM es que al incluir regresores el $R^2$ aumenta. Si el modelo tiene $k$ regresores y se le agregan $s$ adicionales, el $R^2$ de este segundo modelo es mayor que el del primero
  En este contexto si se usa el $R^2$ para comparar dos modelos de distinto número de regresores, el que tenga menos estará en desventaja. Por lo tanto, el $R^2$ debe ser ajustado por los grados de libertad asociados a las sumas de cuadrados. Sea $\overline{R}^2$ el coeficiente de determinación ajustado:

$$\overline{R}^2 = 1 - \frac{SCR/(N-k)}{SCT/(N-1)} = 1 - \frac{N-1}{N-k}\frac{SCR}{SCT}$$

  Ya que $\frac{SCR}{SCT} = 1 - R^2$, entonces

$$\overline{R}^2 = 1 - \frac{N-1}{N-k}(1 - R^2)$$

El $\overline{R}^2$ no esta acotado entre cero y uno, por lo que puede dar negativo (en ese caso se asumiría como cero)

# Inferencia en el modelo de RLM

## Pruebas de hipótesis

$$H_0 : \beta_j = \beta_{j0}$$

$$H_A : \beta_j < \beta_{j0} \text{ ó}$$

$$H_A : \beta_j \neq \beta_{j0} \text{ ó}$$

$$H_A : \beta_j > \beta_{j0}$$

Bajo $H_0$ $\frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{N-k}$ gdl, y la regla de decisión se establece teniendo en cuenta la hipótesis alternativa, $H_A$

Por ejemplo, si $H_A : \beta < \beta_0$, la regla de decisión es rechazar $H_0$ al nivel $\epsilon$ de significancia si

$$t_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}_{\hat{\beta}_j}} < -t_{N-k}(\epsilon)$$

o cuando

$$\text{p-value} < \epsilon$$

## Intervalos de confianza

$$IC(1-\epsilon)(\beta_j) = \left[\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} t_{N-k}(\epsilon/2)\right]$$

# Inferencia en el modelo de RLM

## La significancia de la regresión globalmente considerada

En el modelo

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{2k} + u_i$$

interesa verificar si el conjunto de variables es globalmente significativas, esto es

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

$$H_A : \text{al menos un } \beta_j \neq 0$$

Aplicando la fórmula general $SCR_0 = \sum(Y_i - \overline{Y})^2 = SCT$ y $H = k - 1$, por lo tanto, Bajo $H_0$

$$F_c = \frac{(SCR_0 - SCR)/k - 1}{SCR/N - k} \sim F^{k-1}_{N-k}$$

Esta expresión es equivalente a

$$F_c = \frac{SCM/k - 1}{SCR/N - k} \sim F^{k-1}_{N-k}$$

Regla de decisión rechazar $H_0$ si $F_c > F^{k-1}_{N-k}(\epsilon)$ o si p-value $< \epsilon$

# Ejercicio aplicado en R

Algunos estados de EE.UU. han promulgado leyes que permiten a los ciudadanos llevar armas. Estas leyes son conocidas como leyes de *emisión obligatoria*, debido a que obligan a las autoridades locales a emitir un permiso para llevar armas a todos los solicitantes que sean ciudadanos, sean mentalmente competentes y no hayan sido condenados por un delito grave (algunos estados imponen algunas restricciones adicionales).

Sus defensores sostienen que si más personas llevan armas, el crimen se reducirá debido a que los criminales serán disuadidos de atacar a otras personas. Sus opositores argumentan que el crimen aumentará debido al uso accidental o espontáneo de las armas.

En este ejercicio, se analiza el efecto de las leyes sobre la tenencia de armas sobre los crímenes violentos. El archivo de datos Guns.xls contiene un panel de datos sobre 50 estados de EE.UU., más el Distrito de Columbia para los años 1977 a 1993. Se ofrece una descripción detallada en el archivo Guns_Description.pdf

- Estime (1) una regresión de la variable $ln(vio)$ sobre la variable $shall$ y (2) una regresión de la variable $ln(vio)$ sobre las variables $shall$, $incarc\_rate$, $density$, $avginc$, $pop$, $pb1064$, $pw1064$ y $pm1029$
- ¿Cuál de los dos modelos es mejor? Sobre el modelo seleccionado interprete los coeficientes estimados y analice su significancia estadística
- Repita el análisis utilizando las variables $ln(rob)$ y $ln(mur)$ como variable dependiente. ¿Qué conclusiones sacaría sobre los efectos de las leyes de tenencia de armas sobre los índices de criminalidad?