El Modelo de Regresión Simple

Carlos Velasco[1]

[1]Departamento de Economía
Universidad Carlos III de Madrid

Econometría I
Máster en Economía Industrial
Universidad Carlos III de Madrid
Curso 2007/08

Resumen

El Modelo de Regresión Simple

Objetivos

1. Propiedades del modelo de regresión simple.
2. Estimación MCO y propiedades.
3. Estimadores MCO.

Bibliografía

Wooldridge (2006). Capítulo 2.

Goldberger (2001). Capítulos 2-7.

Greene (1999). Capítulos 3, 4.2-4.3.

1. Definición del Modelo de Regresión Simple

Definición del Modelo de Regresión Simple

Objetivo: Modelo Econométrico para explicar cómo x explica y
Problemas básicos:

- Como la relación entre x e y no es perfecta, ¿cómo se permite que otros factores afecten a y?
- ¿Cuál es la relación funcional entre x e y?
- ¿Cómo asegurarnos que está captando una relación *ceteris paribus*?

Solución sencilla a los problemas anteriores:

$$y = \beta_0 + \beta_1 x + u.$$

El supuesto es que se cumple en la población de interés.

Elementos del modelo:

- Variables y término de error.
- Relación funcional.
- Parámetros.

Modelo de Regresión lineal simple

Variables

y	x
Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control
Variable predicha	Variable predictor
Regresando	Regresor [covariable]

u : término de error o perturbación: factores distintos a x que afectan a y (y que no observamos).

Si los demás factores contenidos en u se mantienen fijos, $\Delta u = 0$, entonces x tiene un efecto lineal sobre y

$$\Delta y = \beta_1 \Delta x \quad \text{si} \quad \Delta u = 0.$$

Modelo de Regresión lineal simple

Parámetros

β_1 : parámetro de pendiente en la relación entre x e y : es el cambio en y cuando se multiplica por el cambio en x. Es el parámetro clave en aplicaciones.

β_0 : término constante (valor de y cuando x y u son cero). Menos interesante.

Modelo de Regresión lineal simple

Ejemplo: producción de soja y fertilizante

$$yield = \beta_0 + \beta_1 fertilizer + u$$

u : calidad de la tierra, lluvia, etc.

Modelo de Regresión lineal simple

Ejemplo: ecuación para el salario

$$wage = \beta_0 + \beta_1 educ + u$$

u : experiencia en el trabajo, habilidad innata, antigüedad en el empleo actual, etc.

Un cambio de una unidad en x tiene siempre el mismo efecto sobre y, independientemente del valor inicial de x.

$$\Delta x = 1 \quad \Longrightarrow \quad \Delta y = \beta_1, \quad \forall x, \Delta u = 0.$$

¿Análisis ceteris paribus?

β_1 : efecto de x sobre y, con todos los demás factores (en u) fijos.
¿Pero en qué sentido podemos mantener los otros factores para llegar a tales conclusiones?

Sólo se pueden obtener estimaciones fiables de los parámetros β_0 y β_1 a partir del muestreo aleatorio cuando establecemos supuestos que restringen el modelo en que el error no observable u se relaciona con la variable explicativa x.

Como x y u son VAs necesitamos un concepto basado en su distribución de probabilidad.

Término constante: β_0

Supuesto inicial: siempre que incluyamos el término constante β_0 en la ecuación podemos suponer que el valor medio de u en la población es cero:

$$E(u) = 0.$$

- No afirma nada sobre la relación entre x e y.
- Sólo afecta a la distribución marginal de u.
- Es simplemente una normalización: el efecto medio de los otros factores se renormaliza a cero.

Modelo de Regresión lineal simple

Relación x y u

1^a posibilidad: medir la relación con el **coeficiente de correlación**: si la correlación es cero, las variables están incorreladas, es decir no tienen relación lineal.

Pero pueden tener otro tipo de relación (no lineal): puede haber relación con x^2, etc.

2^a posibilidad: definir la independencia desde el punto de vista de la **distribución de u condicional** en x :

$$E(u|x) = E(u) = 0.$$

Para todos los posibles valores de x, la media de u siempre es la misma, 0.

Si suponemos que u es igual a la habilidad innata:

- El nivel medio de habilidad tiene que ser el mismo independientemente del número de años de formación:

$$E(habil|x = 8) = E(habil|x = 16).$$

- Si pensamos que el nivel de habilidad debe aumentar con los años de educación, el supuesto entonces debe ser falso.
- No podemos comprobarlo porque el nivel de habilidad innata no se puede observar: pero es una pregunta que hay que plantearse para interpretar el modelo.

Modelo de Regresión lineal simple

Ejemplo: Nota exámen

$$score = \beta_0 + \beta_1 attend + u,$$

donde *score* es el resultado de un examen final, que depende de las clases a las que se ha asistido, *attend* y de otros factores no observables, u, como capacidad del estudiante que acude al examen.

¿Cuándo podremos esperar que el modelo satisfaga

$$E(u|attend) = E(u) = 0?$$

- Si las cantidades de fertilizantes se establecen independientemente de otras características de las parcelas, entonces $E(u|x) = 0$.
- Si aplicamos mayores cantidades de fertilizante en aquellas tierras de mayor calidad, entonces el valor esperado de u cambia con el nivel de fertilizante, y $E(u|x) \neq 0$.

Otra interpretación

El supuesto $E(u|x) = E(u) = 0$ conlleva otra interpretación muy útil. Tomando el valor esperado de y condicional en el valor de x,

$$E(y|x) = \beta_0 + \beta_1 x.$$

- Esta expresión proporciona el valor de la **función de regresión** poblacional, que en este caso es lineal.
- En este caso el incremento de una unidad de x provoca un aumento esperado en y de una unidad.
- También se puede escribir

$$\begin{aligned} y &= E(y|x) + u \\ &= \beta_0 + \beta_1 x + u \end{aligned}$$

donde $E(y|x) = \beta_0 + \beta_1 x$ es la parte explicada por x y u es la parte no explicada por x.

2. Derivación de las estimaciones por Mínimos Cuadrados Ordinarios

Derivación de las estimaciones por Mínimos Cuadrados Ordinarios

Estimación de β_0 y β_1.

Se necesita una muestra de la población: $\{(x_i, y_i), i = 1, \ldots, n\}$ una muestra aleatoria de tamaño n.

Motivación de la estimación MCO

- Varias alternativas.
- Usaremos los supuestos del modelo: en la población u tiene que tener media nula y no estar correlacionado con x,

$$\begin{aligned} E(u) &= 0 \\ Cov(u, x) &= 0. \end{aligned}$$

Estos supuestos implican

$$\begin{aligned} E(y - \beta_0 - \beta_1 x) &= 0 \\ Cov(y - \beta_0 - \beta_1 x, x) &= E([y - \beta_0 - \beta_1 x] x) = 0, \end{aligned}$$

que son dos ecuaciones que definen dos parámetros.

Estimación método momentos

Dados los datos, hacemos que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ satisfagan las mismas ecuaciones, pero dentro de la muestra, no en la población,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) &= 0 \\ \widehat{Cov}_n\left(y - \hat{\beta}_0 - \hat{\beta}_1 x, x\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0. \end{aligned}$$

MCO: Resolución ecuaciones estimación (1)

Operando con la primera ecuación,

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

es decir

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

lo que nos permite obtener $\hat{\beta}_0$ una vez que tenemos estimado $\hat{\beta}_1$.

MCO: Resolución ecuaciones estimación (2)

Sustituyendo la solución para $\hat{\beta}_0$

$$\sum_{i=1}^{n} \left(y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) - \hat{\beta}_1 x_i \right) x_i = 0$$

y agrupando términos,

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i.$$

Se puede usar que

$$\sum_{i=1}^{n} (y_i - \bar{y}) x_i = \sum_{i=1}^{n} (y_i - \bar{y}) (x_i - \bar{x}); \quad \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

MCO: Resolución ecuaciones estimación (3)

Suponiendo que

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0,$$

el valor estimado de la pendiente es

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\widehat{Cov}_n(x, y)}{\widehat{Var}_n(x)}.$$

Si x e y están positivamente relacionadas en la muestra, entonces $\hat{\beta}_1$ será positiva, y viceversa.

MCO: Resolución ecuaciones estimación (4)

Si

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 0,$$

- Hemos tenido mala suerte en la muestra y todas las x_i son iguales.
- O el problema no es interesante, porque la x es constante en la población.

Estimación Mínimo Cuadrática (1)

Valor ajustado: una vez obtenidos los estimadores,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

que es el valor predicho cuando $x = x_i$. Dada una muestra tenemos n valores ajustados.
Residuo: diferencia entre el valor verdadero y_i y el ajustado $\hat{y}_i$,

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

Suma de Cuadrados de los residuos

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2.$$

Criterio MCO: escoger $\hat{\beta}_0, \hat{\beta}_1$ tal que la suma $\sum_{i=1}^{n} \hat{u}_i^2$ sea lo más pequeña posible.

Estimación Mínimo Cuadrática (2)

Las condiciones de primer orden son

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\Big|_{b_0=\hat{\beta}_0, b_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2\Big|_{b_0=\hat{\beta}_0, b_1=\hat{\beta}_1} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0,$$

que son equivalentes a las condiciones anteriores.

¿Por qué MCO? (Y no otro método)

Se minimiza la suma de cuadrados de los residuos por varias razones:

- Es fácil obtener la fórmula de los estimadores.
- Sin técnicas de optimización numérica.
- Teoría estadística es sencilla: insesgadez, consistencia, etc.
- Solución coincide con las propiedades deducidas de la esperanza condicional.

Recta de regresión

O función de regresión muestral,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

que subraya que son valores ajustados o estimadores de $E(y|x) = \beta_0 + \beta_1 x$:

- $\hat{\beta}_0$ es el valor predicho cuando $x = 0$, lo cual puede tener sentido o no.
- $\hat{\beta}_1$ es el valor estimado de la pendiente,

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x}$$

o equivalentemente

$$\Delta \hat{y} = \hat{\beta}_1 \Delta x.$$

Ejemplos

- Salario director general y rendimiento de las acciones

$$salary = \beta_0 + \beta_1 roe + u.$$

`CEOSAL1.RAW`, $n = 209$:

$$\widehat{salary} = 963{,}191 + 18{,}501 roe.$$

- Salario y educación,

$$wage = \beta_0 + \beta_1 educ + u.$$

`WAGE1.RAW`, $n = 526$.

- Resultados electorales y gastos de campaña,

$$voteA = \beta_0 + \beta_1 shareA + u.$$

`VOTE1.RAW`, $n = 173$.

Notación

Cuando se obtiene el ajuste

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

se dice que se ha llevado a cabo la regresión de y sobre x (o hemos regresado y sobre x) : *variable dependiente sobre variable independiente.*

Esto indica que siempre se estima la pendiente y el término constante.

3. Funcionamiento del método MCO

Estudiaremos propiedades **algebraicas** de la recta ajustada por MCO, para un conjunto de datos en particular.

Luego las compararemos con propiedades **estadísticas**, que se refieren a toda la población.

Valores ajustados y residuos

Supondremos que hemos obtenido los estimadores de la ordenada en el origen y la pendiente $\hat{\beta}_0, \hat{\beta}_1$ a partir de una determinada muestra (x_i, y_i) , $i = 1, \ldots, n$.

Valor ajustado (para cada observación)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

que está sobre la recta ajustada.

Residuo (asociado a cada observación):

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i. \end{aligned}$$

- Si $\hat{u}_i > 0$, la recta ajustada infrapredice (pasa por debajo del punto (x_i, y_i)).
- Si $\hat{u}_i < 0$, la recta ajustada sobrepredice (pasa por encima del punto (x_i, y_i)).
- Si $\hat{u}_i = 0$, situación ideal, pero no ocurre casi nunca.

Propiedades Algebraicas de los Estadísticos MCO

1. La suma (y la media muestral) de los residuos MCO es cero,

$$\sum_{i=1}^{n} \hat{u}_i = 0.$$

Prueba: primera condición de primer orden de MCO.

2. La covarianza muestral de regresores y residuos MCO es cero,

$$\sum_{i=1}^{n} \hat{u}_i x_i = 0.$$

Prueba: segunda condición de primer orden de MCO.

3. La covarianza muestral de los valores ajustados y de los residuos MCO es cero,

$$\sum_{i=1}^{n} \hat{u}_i \hat{y}_i = 0.$$

Prueba: Usar (1) y (2) y que $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Propiedades Algebraicas de los Estadísticos MCO (2)

4. El punto $(\bar{x}, \bar{y})$ siempre está sobre la recta ajustada, es decir

$$\hat{y}(\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y}.$$

   Prueba: definición de $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

5. Descomposición de y_i :

$$y_i = \hat{y}_i + \hat{u}_i,$$

   en el valor ajustado y su residuo asociado, que están incorrelados. Además se deduce que:

$$\bar{y} = \overline{\hat{y}}.$$

Sumas de Cuadrados

Suma de Cuadrados Totales, **SST**:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

Suma de Cuadrados Explicada, **SSE**:

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Suma de Cuadrados de los Residuos, **SSR**:

$$\sum_{i=1}^{n} \hat{u}_i^2 .$$

Descomposición de las Sumas de Cuadrados

$$SST = SSE + SSR.$$

Prueba:

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\{y_i - \hat{y}_i\} + \{\hat{y}_i - \bar{y}\})^2 \\
&= \sum_{i=1}^{n} (\hat{u}_i + \{\hat{y}_i - \bar{y}\})^2 \\
&= \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n} \{\hat{y}_i - \bar{y}\}^2 + \overbrace{2 \sum_{i=1}^{n} u_i \{\hat{y}_i - \bar{y}\}}^{=0} \\
&= SSR + SSE.
\end{aligned}$$

Descomposición Sumas de Cuadrados

Notación

- Suma de Cuadrados Totales: SST=TSS=**SCT**=STC.
- A veces se habla de Suma de Cuadrados de la Regresión (o del Modelo) y de suma explicada de los cuadrados: SSE=**SCE**=SEC.
- También se habla de la Suma de Cuadrados de los Errores en lugar de la de Residuos, pero es muy confuso (y erróneo, son residuos, no errores): nosotros siempre **SCR**.

Bondad de Ajuste

(Suponemos que SST$\neq 0$ y que ajustamos el término constante)

- **Coeficiente de Determinación or R-cuadrado**:

$$R^2 = R_n^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- $0 \leq R^2 \leq 1$.
- $100R^2$ es el porcentaje de la variabilidad de y que explica x.
- $R^2 = 1$: ajuste perfecto.
- Propiedad:

$$R^2 = \hat{\rho}_{y,x}^2.$$

- En aplicaciones es frecuente ver R^2 bastante bajos, pero eso no significa que el modelo sea inútil o esté mal (ni que uno con R^2 alto esté muy bien).

4. Unidades de medida y forma funcional

- Los coeficientes de MCO cambian de forma totalmente predecible cuando cambiamos las unidades de medida.
- Si la variable dependiente se multiplica por c, entonces los coeficientes MCO del nuevo modelo ajustado también se multiplican por c.
- Ejemplo: $salary$ = salario en miles de dólares,

$$\widehat{salary} = 963{,}191 + 18{,}501 roe$$

que se cambia a $salarydol$, en dólares, $salarydol = 1000 * salary$

$$\widehat{salarydol} = 963191 + 18501 roe.$$

Unidades de medida y forma funcional

Unidades de Medida (2)

- Si la variable explicativa se multiplica por c, entonces el coeficiente estimado de la pendiente se divide por c (y $\hat{\beta}_0$ no cambia).
- Ejemplo: $roedec = roe/100$, entonces

$$\begin{aligned} \widehat{salary} &= 963{,}191 + 18{,}501\frac{100}{100}roe \\ &= 963{,}191 + (18{,}501 * 100)\frac{roe}{100} \\ &= 963{,}191 + 1850{,}1 roedec \end{aligned}$$

- El R^2 no cambia en ningún caso porque no depende de las unidades de medida.

Especificación de la forma funcional

No linealidades en regresión simple

- Las relaciones lineales no son suficientes para describir las relaciones económicas.
- Es importante introducir no linealidades mediante definiciones apropiadas de las variables dependiente e independientes.
- Casos más frecuentes es cuando ciertas variables aparecen en **logaritmos**.
- Pero el modelo sigue siendo lineal, en particular, lineal en los parámetros β_0 y β_1, por lo que la estimación MCO se realiza igual, pero su interpretación puede ser diferente.

Modelos No Lineales

Variable dependiente en logaritmos (log-level)

- El modelo lineal implica que el incremento de y cuando cambia x siempre es igual, independientemente del nivel de x :

$$wage = \beta_0 + \beta_1 educ + u$$

- Es más razonable suponer que es el **porcentaje** de incremento el que es constante.
- Un modelo que consigue esto, aproximadamente, es

$$\log(wage) = \beta_0 + \beta_1 educ + u.$$

Así, si $\Delta u = 0$, entonces $(100 \cdot \beta_1)$ es la **semielasticidad** de *wage* respecto a *educ*,

$$\%\Delta wage = (100 \cdot \beta_1)\, \Delta educ.$$

- Para ver la relación podríamos tomar exponenciales en ambos lados del modelo,

$$wage = \exp(\beta_0 + \beta_1 educ + u).$$

- En este caso la relación entre x e y se establece en términos de incrementos relativos.
- Ambas variables deben aparecer en logaritmos,

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + u,$$

donde β_1 es la **elasticidad** de *salary* respecto a *sales*.

- En este caso

$$\%\Delta wage = \beta_1 \%\Delta sales.$$

- En este caso controlamos incrementos relativos de x,

$$y = \beta_0 + \beta_1 \log(x) + u,$$

- Significado de β_1,

$$\Delta y = \frac{\beta_1}{100} \%\Delta x.$$

5. Valores esperados y varianzas de los estimadores MCO

Valores esperados y varianzas de los EMCO

Supuestos

RLS.1 (*Modelo lineal en parámetros*). En el modelo para la población, la variable dependiente y se relaciona con la variable independiente x y el error u mediante

$$y = \beta_0 + \beta_1 x + u,$$

donde β_0 y β_1 son los parámetros del término constante y la pendiente.
Esta expresión es importante para interpretar β_0 y β_1.

RLS.2 (*Muestreo aleatorio*). Para estimar los parámetros podemos usar una muestra de tamaño n, (x_i, y_i) , $i = 1, \ldots, n$ del modelo poblacional,

$$y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, \ldots, n.$$

Esta expresión es importante para deducir las propiedades de los EMCO de β_0 y β_1.

RLS.3 (*Media Condicional cero*).

$$E(u|x) = 0.$$

- Para la muestra aleatoria esto implica que

$$E(u_i|x_i) = 0, \quad i = 1, \ldots, n.$$

- Este supuesto permite deducir las propiedades de los EMCO condicionales en los valores de x_i en nuestra muestra. (Similar supuesto a situaciones donde x_i son fijos en muestras repetidas, aunque no es la forma habitual de recoger datos en Economía).

RLS.4 (*Variación muestral en la variable independiente*). En la muestra, las variables independientes x_i, $i = 1, \ldots, n$, no son todas iguales a una misma constante. Esto requiere alguna variación de x en la población.

- Es decir, necesitamos que la **variación total** en x_i, s_x^2, sea positiva,

$$s_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 > 0.$$

- Si este supuesto falla no se pueden calcular los EMCO, por los que su análisis estadístico no tiene sentido.

Valores esperados y varianzas de los EMCO

Insesgadez estimadores MCO

- $\hat{\beta}_1$ es una muestra aleatoria:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

- Sustituyendo el modelo $y_i = \beta_0 + \beta_1 x_i + u_i$ tenemos que

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{s_x^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i)}{s_x^2}.$$

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x}) (\beta_0 + \beta_1 x_i + u_i) &= \sum_{i=1}^{n} (x_i - \bar{x}) (\beta_1 x_i + u_i) \\
&= \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum_{i=1}^{n} (x_i - \bar{x}) u_i \\
&= \beta_1 s_x^2 + \sum_{i=1}^{n} (x_i - \bar{x}) u_i.
\end{aligned}$$

Valores esperados y varianzas de los EMCO

Insesgadez estimadores MCO (2)

- Por tanto, definiendo $d_i = (x_i - \bar{x})$,

$$\begin{aligned} \hat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})\, u_i}{s_x^2} \\ &= \beta_1 + \left(\frac{1}{s_x^2}\right) \sum_{i=1}^{n} d_i u_i \end{aligned}$$

- $\hat{\beta}_1$ es igual a la pendiente poblacional, β_1, más un término que es una combinación lineal de los errores u_i.
- Condicional en x_i, la aleatoriedad de $\hat{\beta}_1$ depende solamente de los errores $\{u_1, \ldots, u_n\}$.
- La razón de que $\hat{\beta}_1$ difiera de β_1 es debido a que los errores $\{u_1, \ldots, u_n\}$ no son cero.

Teorema. (Insesgadez de los EMCO). Usando los supuesto RLS.1-RLS.4,

$$E\left(\hat{\beta}_0\right) = \beta_0, \quad E\left(\hat{\beta}_1\right) = \beta_1.$$

- Prueba: los valores esperados son condicionales en los valores muestrales de x_i,

$$\begin{aligned} E\left(\hat{\beta}_1\right) &= \beta_1 + E\left[\left(\frac{1}{s_x^2}\right)\sum_{i=1}^{n} d_i u_i\right] \\ &= \beta_1 + \left(\frac{1}{s_x^2}\right)\sum_{i=1}^{n} d_i E\left(u_i\right) \\ &= \beta_1 + \left(\frac{1}{s_x^2}\right)\sum_{i=1}^{n} d_i \cdot 0 = \beta_1, \end{aligned}$$

ya que $E\left(u_i\right) \equiv E\left(u_i|x_i\right) = 0$.

Valores esperados y varianzas de los EMCO

Insesgadez estimadores MCO (4)

- Prueba para $\hat{\beta}_0$. Usando que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, y usando RLS.1

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0 = \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x}.$$

- Entonces, tomando esperanzas condicionales en x_i, $i = 1, \ldots, n$,

$$\begin{aligned} E\left(\hat{\beta}_0\right) &= \beta_0 + \left(\beta_1 - E\left(\hat{\beta}_1\right)\right)\bar{x} + E\left(\bar{u}\right) \\ &= \beta_0 + \left(\beta_1 - \hat{\beta}_1\right)\bar{x} + 0 \\ &= \beta_0 \end{aligned}$$

porque $E\left(\bar{u}\right) = 0$ por RLS.3.

Valores esperados y varianzas de los EMCO

Insesgadez estimadores MCO (5)

- La propiedad de Insesgadez no dice nada sobre el valor que se obtiene para una muestra en particular.
- Esta propiedad no se cumple si alguno de los **supuestos falla**.
  - Si RLS.4 falla, no se pueden computar los EMCO.
  - RLS.1 se puede hacer cumplir eligiendo x e y incluso si la relación original es no lineal.
  - RLS.2 no se cumplirá para datos de series temporales.
  - El supuesto clave es RLS.3, si falla los EMCO no serán insesgados. La posibilidad de que x esté correlada con u siempre está ahí con datos no experimentales.
- **Correlación espúrea**: si hay factores en u que afectan a y y que también están correlados con x.

Valores esperados y varianzas de los EMCO

Ejemplo: rendimiento en matemáticas y el programa de comidas en el colegio

- Modelo para explicar el efecto de un programa de comidas subvencionadas en el colegio, *cp*, sobre el rendimiento:

$$math10 = \beta_0 + \beta_1 lnchprg + u.$$

- *math*10 : % que aprueban un test de matemáticas en la high school (`MEAP93`)
- *lnchprg:* % de estudiantes elegibles para el programa de comida subvencionadas.
- ¿Qué signo esperarías para β_1 si midiese un efecto cp?

Valores esperados y varianzas de los EMCO

Varianza de los estimadores MCO (1)

- Además de saber que la distribución muestral de $\hat{\beta}_1$ está centrada, es importante saber su variabilidad alrededor de β_1.
- La varianza de $\hat{\beta}_1$ se puede deducir a partir de RLS.1-RLS.4, pero su expresión es complicada. Por eso añadimos un nuevo supuesto:

RLS.5 (*Homocedasticidad Condicional*): u tiene varianza, condicional en x, constante,

$$Var\,(u|x) = \sigma^2.$$

- Este supuesto simplifica el análisis y hace que los EMCO tengan ciertas propiedades de eficiencia.
- Es, junto con RLS.4, más débil que el supuesto de independencia.
- σ^2 se le llama la varianza del error o perturbación.

Valores esperados y varianzas de los EMCO

Varianza de los estimadores MCO (2)

- Los supuestos sobre u, condicionales en x,

$$\begin{aligned} E(u|x) &= 0 \\ Var(u|x) &= E\left(u^2|x\right) = \sigma^2, \end{aligned}$$

- Se escriben frecuentemente en función de la distribución de y, condicional en x,

$$\begin{aligned} E(y|x) &= \beta_0 + \beta_1 x \\ Var(y|x) &= \sigma^2. \end{aligned}$$

- Si $Var(y|x) = Var(u|x)$ dependiese de x se hablaría de **Heterocedasticidad** (condicional).
- Ejemplo: heterocedasticidad en salarios.

Teorema 2. (Varianza de los EMCO). Bajo los supuesos RLS.1-5

$$\begin{aligned} Var\left(\hat{\beta}_1\right) &= \frac{\sigma^2}{s_x^2} \\ Var\left(\hat{\beta}_0\right) &= \frac{\sigma^2 \frac{1}{n}\sum_{i=1}^{n} x_i^2}{s_x^2}, \end{aligned}$$

condicionales en los valores muestrales $\{x_1, \ldots, x_n\}$.

Prueba: (Sólo para $\hat{\beta}_1$) Se parte de

$$\hat{\beta}_1 = \beta_1 + \left(\frac{1}{s_x^2}\right) \sum_{i=1}^{n} d_i u_i.$$

Por tanto, condicional en x_i,

$$\begin{aligned} Var\left(\hat{\beta}_1\right) &= \left(\frac{1}{s_x^2}\right)^2 Var\left(\sum_{i=1}^{n} d_i u_i\right) \\ &= \left(\frac{1}{s_x^2}\right)^2 \sum_{i=1}^{n} d_i^2 Var\left(u_i\right) \\ &= \left(\frac{1}{s_x^2}\right)^2 \sum_{i=1}^{n} d_i^2 \sigma^2 \\ &= \sigma^2 \left(\frac{1}{s_x^2}\right)^2 s_x^2 = \frac{\sigma^2}{s_x^2}. \end{aligned}$$

Valores esperados y varianzas de los EMCO

Varianza de los estimadores MCO (4)

- Estas fórmulas no son válidas en presencia de heterocedasticidad.
- Son importantes para construir intervalos de confianza y contrastes de hipótesis
- $Var\left(\hat{\beta}_1\right)$ depende de:
  - σ^2
  - s_x^2
  - (Indirectamente) n, ya que $s_x^2 = n \cdot \widehat{Var}(x)$.
- ¿Cuál podemos elegir de esos factores?
- ¿Consistencia?

Valores esperados y varianzas de los EMCO

Estimación de la Varianza del Error (1)

- Hay factores en $Var\left(\hat{\beta}_1\right)$ y $Var\left(\hat{\beta}_0\right)$ desconocidos: σ^2.
- σ^2 se puede estimar con los datos (a partir de los residuos $\hat{u}_i$, ya que los errores u_i no son observables).
- Usando $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, tenemos que

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i \\ &= \beta_0 + \beta_1 x_i + u_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= u_i + \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i, \end{aligned}$$

por lo que la diferencia entre u_i y $\hat{u}_i$ tiene un valor esperado de cero (y $\rightarrow_p 0$).

Valores esperados y varianzas de los EMCO

Estimación de la Varianza del Error(2)

- $\sigma^2 = E\left(u^2\right)$, por lo que un estimador insesgado de σ^2 es $n^{-1}\sum_{i=1}^{n} u_i^2$.
- Como los errores no son observables, un primer estimador sería $n^{-1}SSR = n^{-1}\sum_{i=1}^{n} \hat{u}_i^2$, pero éste no es insesgado.
- Los residuos satisfacen dos restricciones:

$$\sum_{i=1}^{n} \hat{u}_i = 0, \quad \sum_{i=1}^{n} \hat{u}_i x_i = 0.$$

- Por tanto los residuos tienen $n-2$ grados de libertad.
- El estimador insesgado de σ^2 que hace el ajuste por los grados de libertad es

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2}.$$

- Estimador de σ, o **error estándar de regresión**,

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}.$$

- **Error estándar** de $\hat{\beta}_1$:

$$se\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}}{\left(\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{1/2}}.$$

- Da una idea de la variación muestral de $\hat{\beta}_1$, pero es un estimador que varía de muestra a muestra.
- También son fundamentales en construir ICs y contrastes de hipótesis.

6. Regresión por el origen

Regresión por el origen

- A veces se impone la restricción de que cuando $x = 0$, el valor esperado de y es cero, es decir $\beta_0 = 0$ (renta - impuestos).
- Se quiere una función de regresión estimada que pase por $(x = 0, \tilde{y} = 0)$,

$$\tilde{y} = \tilde{\beta}_1 x.$$

- En este caso MCO minimiza

$$\sum_{i=1}^{n} \left(y_i - \tilde{\beta}_1 x_i \right)^2.$$

- $\tilde{\beta}_1$ satisface la condición de primer orden:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2} : \quad \sum_{i=1}^{n} \left(y_i - \tilde{\beta}_1 x_i \right) x_i = 0.$$

- Necesita que no todos los x_i sean cero. Sólo coincide con MCO si $\bar{x} = 0$. Estimador sesgado si $\beta_0 \neq 0$.