Modelo de Regresión Lineal Clásico

Maestría en Economía
Econometría Intermedia - Módulo 2

Docente: Juan Palomino
19 de abril, 2023

1. Modelo de Regresión Lineal

La ecuación del modelo de regresión lineal es la siguiente:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_K x_{iK} + \epsilon_i \quad i = 1, ..., n \tag{1.1}$$

donde la expresión $\beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_K x_{iK}$ es conocida como la **función de regresión**, y los coeficientes $\beta' s$ son los **coeficientes de regresión** y ϵ_i es el término de error no observado.

Para poder explicar los supuestos del modelo clásico, introducimos la notación matricial. Esta notación ayudará a derivar el estimador MCO de β. Sea el vector columna $\mathbf{x_i}$ de dimensión K y β como:

$$\underset{K\times 1}{\mathbf{x_i}} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}, \quad \underset{K\times 1}{\beta} = \begin{bmatrix} \beta_{i1} \\ \beta_{i2} \\ \vdots \\ \beta_{iK} \end{bmatrix} \tag{1.2}$$

Por la definición de producto de vectores internos,

$$\mathbf{x_i}^\top \beta = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_K x_{iK} \tag{1.3}$$

La ecuación (1.1) puede ser escrito como:

$$y_i = x_i' \beta + \epsilon_i \quad (i = 1, 2, ..., n) \tag{1.4}$$

También se define:

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times 1}{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \quad \underset{n\times K}{X} = \begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix} \tag{1.5}$$

Esta ecuación puede ser escrita de la siguiente manera:

$$\underset{(n\times 1)}{y} = \underset{(n\times K)}{X} \underset{(K\times 1)}{\beta_0} + \underset{(n\times 1)}{\epsilon} \tag{1.6}$$

2 Supuestos de MCO

Supuesto 1.1: Linealidad

El primer supuesto en el modelo de regresión lineal clásico es la relación lineal entre la variable dependiente y los regresores.

Supuesto 1.1 - Linealidad. El modelo poblacional está dado por:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_K x_{iK} + \epsilon_i \quad i = 1, ..., n \tag{2.1}$$

donde β's son parámetros desconocidos a ser estimados, y ϵ_i es el término de error no observado.

Es importante entender el significado de β_K, $k = 1, ..., K$. Diferenciando totalmente el modelo (2.1), obtenemos:

$$\begin{aligned} dy &= \frac{\partial y}{x_1} dx_1 + \frac{\partial y}{x_2} dx_2 + ... + \frac{\partial y}{x_K} dx_K \\ &= \beta_1 dx_1 + \beta_2 dx_2 + \beta_K dx_K + d_\epsilon \end{aligned}$$

Sea $d\epsilon = 0$ y $dx_k = 0$ para todo $k = 2, ..., K$, obtenemos:

$$\left.\frac{dy}{dx_1}\right|_{otras\ variables\ constantes} = \frac{\partial y}{\partial x_1} = \beta_1$$

Por lo tanto, el coeficiente β_1 mide el cambio en y debido al incremento de una unidad en x_1, manteniendo las demás variables independientes y el término de error fijo (ceteris paribus).

Cabe resaltar que los parámetros deben ser lineales, es decir, β^2 o e^β violan este supuesto.

Ejemplo 1.1 - Retornos a la escolaridad

Una aplicación de economía laboral concierne a la medida del impacto de la educación sobre los salarios. La pregunta es: ¿Cuál es el efecto de educación sobre los salarios?. El modelo es el siguiente:

$$logw_i = \beta_1 + \beta_2 sch_i + \beta_3 exper_i + \epsilon_i \quad i = 1, ..., n \tag{2.2}$$

donde w denota el salario mensual, sch denota los años de escolaridad, y $exper$ denota la experiencia laboral. El subindice i refiere al individuo i en la muestra. Dado que la variable dependiente es *log wage*, el modelo es log-lineal y el coeficiente β_2 mide el cambio porcentual en los salarios asociado con un incremento de un año en educación. Formalmente,

$$\begin{aligned} \beta_2 &= \frac{\partial logw_i}{\partial sch_i} \\ &= \frac{1}{w_i} \frac{\partial w_i}{sch_i} \\ &= \frac{\frac{\partial w_i}{w_i}}{sch_i} \\ &= \frac{cambio\ relativo\ en\ w_i}{cambio\ absoluto\ en\ sch_i} \end{aligned}$$

Ejemplo 1.2 - Ingreso y obesidad

Estamos interesados en saber si el ingreso tiene un impacto en la obesidad, la cual se mide a través del Indice de Masa Corporal (BMI). El modelo poblacional es:

$$BMI = \beta_1 + \beta_2 INC_i + \beta_3 PhyAct_i + \beta_4 age + \epsilon_i \tag{2.3}$$

donde INC es el ingreso medido en soles, $PhyAct$ mide la actividad física, y age es la edad del individuo. Además, β_2 mide el impacto de ingreso sobre obesidad, mientras mantiene constante $PhyAct$ y age fijos. Matemáticamente,

$$\frac{\partial BMI}{\partial INC} = \beta_2$$

Por lo tanto, β_2 reporta el incremento o disminución sobre BMI dado un aumento de 1 sol en el ingreso.

Supuesto 1.2 - Exogeneidad Estricta

Supuesto 1.2 - Exogeneidad estricta. El valor esperado de la perturbación aleatoria debe ser cero para cualquier observación:

$$\mathbb{E}[\epsilon_i|X] = 0 \quad i = 1, ..., n \tag{2.4}$$

para cualquier valor de $x_1, x_2, ..., x_K = 0$ en la población, el promedio no observable es igual cero.

También se puede escribir como:

$$\mathbb{E}[\epsilon_i|x_1, ..., x_n] = 0 \quad i = 1, ..., n$$

Este supuesto tiene varias **implicaciones estadísticas**:

- Todas las variables que no se consideran en el modelo y capturados en ϵ no están correlacionado con cualquier variable en el modelo. Al respecto, consideremos el siguiente modelo reducido del ejemplo 1.2:

  $$BMI_i = \beta_1 + \beta_2 INC_i + \epsilon_i$$

  El supuesto de exogeneidad estricta implica que $\mathbb{E}[\epsilon_i|INC_1, INC_2, ..., INC_n] = 0$. Esto significa que aquellos factores que puedan afectar al BMI no están relacionados en promedio con INC. Esto puede o no puede ser cierto, pero ésta es la principal pregunta a fin de determinar si el método de mínimos cuadrados ordinarios pruduce estimadores insesgados y consistentes. Por tanto, cualquier problema que permite ϵ a ser correlacionado con alguna de las variables independientes causa la falla del Supuesto 1.2 - Exogeneidad Estricta

- [**Media Incondicional**] La media incondicional del término de error es cero:

  $$\mathbb{E}(\epsilon_i) = \mathbb{E}[\mathbb{E}(\epsilon_i|X] = 0 \tag{2.5}$$

- [**Ortogonalidad**] Si el momento cruzado $\mathbb{E}(xy)$ de dos variables aleatorias x e y es cero, entonces se dice que x es ortogonal a y. Bajo exogeneidad estricta, los regresores son ortogonales al término de error para todas las observaciones, es decir,

  $$\mathbb{E}(x_{jk}\epsilon_i) = 0 \quad i, j = 1, ..., n; k = 1, ..., K \tag{2.6}$$

  o

  $$\mathbb{E}(x \cdot \epsilon_i) = \begin{pmatrix} \mathbb{E}(x_j 1\epsilon_i) \\ \mathbb{E}(x_j 2\epsilon_i) \\ \vdots \\ \mathbb{E}(x_j K\epsilon_i) \end{pmatrix} = \underset{(K\times 1)}{0} \quad \textit{para todo } i, j \tag{2.7}$$

- [**Condición cero correlación**] Ya que la media del término de error es cero, las condiciones de ortogonalidad (2.7) son equivalentes a las condiciones de cero correlación.

  $$\begin{aligned} Cov(\epsilon_i, x_{jk}) &= \mathbb{E}(x_{jk}\epsilon_i) - \mathbb{E}(x_{jk})\mathbb{E}(\epsilon_i) \\ &= \mathbb{E}(x_{jk}\epsilon_i) \\ &= 0 \end{aligned}$$

En particular, para $i = j$, $Cov(x_{ij}, \epsilon_i) = 0$. Asimismo, exogeneidad estricta implica el requerimiento que los regresores deben estar contemporáneamente no correlacionados con el término de error. En otras palabras, ninguna de las variables contenidas en ϵ_i están correlacionadas con las x's. Si x_j está correlacionado con ϵ_i, x_j es **endógeno**.

Ejemplo 1.3

Considerar el siguiente modelo poblacional:

$$wage_i = \beta_1 + \beta_2 sch_i + \beta_3 habil_i + \epsilon_i \tag{2.8}$$

donde sch_i denota los años de escolaridad, y $habil_i$ denota la habilidad innata del individuo i. Sin embargo, a veces es complicado medir la habilidad, por lo que el modelo se convierte:

$$\begin{gathered} wage_i = \beta_1 + \beta_2 sch_i + \mu_i \\ u_i = \beta_3 habil_i + \epsilon \end{gathered}$$

Por lo tanto, el nuevo término del error μ_i contiene la habilidad y el término de error ϵ_i. Si $\mathbb{E}(\mu_i|sch_1, ..., sch_n) = \mathbb{E}(abil|sch_1, ..., sch_n = 0$, sch es exógeno. Sin embargo, hay razones para creer que habilidad y educación están correlacionadas (positivamente). Si esto es cierto, entonces el supuesto de exogeneidad estricta no se cumple y no podemos hacer ceteris paribus: $\mathbb{E}(\mu_i|sch_1, ..., sch_n) \neq 0$, y $Cov(\mu_i, sch_i) \neq 0$ para $i = j$.

Lemma 1 ***Media condicional de*** y_i
La media condicional de la variable dependiente es una función lineal de los regresores. Los supuestos de Linealidad y Exogeneidad Estricta implica que

$$\mathbb{E}(y_i|X) = x_i'\beta_0 \quad i = 1, ..., n \tag{2.9}$$

En notación matricial

$$\mathbb{E}(y|X) = \begin{pmatrix} \mathbb{E}(y_1|X) \\ \mathbb{E}(y_2|X) \\ \vdots \\ \mathbb{E}(y_n|X) \end{pmatrix} = \begin{pmatrix} x_1'\beta_0 \\ x_2'\beta_0 \\ \vdots \\ x_n'\beta_0 \end{pmatrix} = X\beta_0$$

Supuesto 1.3 - Rango Completo

Supuesto 1.3 - Rango Completo. Asume que ninguna de las variables en el modelo es una combinación lineal exacta de las otras.

$$X \text{ es una matriz } n \times K \text{con rango } K \tag{2.10}$$

Significa que X tiene rango de columna completo.

Nota: El rango de una matriz es igual al número de columnas linealmente independientes de la matriz. También se le conoce como **condición de identificación**.

Ejemplo 1.4 Considerar el siguiente modelo:

$$y_i = \beta_1 + \beta_2 log(x_i) + \beta_3 log(x_i^2) + \epsilon_i \tag{2.11}$$

Notar que:

$$log(x_i^2) \equiv 2log(x_i)$$

Asimismo, $log(x_i^2)$ es una combinación lineal de $log(x_i)$, por lo tanto, el supuesto 1.3 falla.

Ejemplo 1.5 - Trampa de las variables dummy Considerar la ecuación salarial y asumir que existe una variable categórica que indica el estado civil del individuo, que es igual a 1 si está casado(a), 2 si está soltero(a), 3 si está viudo(a). Ya que la variable es categórica, creamos tres variables dummies para cada categoría, por lo que el modelo es:

$$wage_i = \beta_1 + \beta_2 sch_i + \beta_3 casado_i + \beta_4 soltero_i + \beta_5 viudo_i + \epsilon_i$$

donde $casado = 1$ si el individuo está casado(a) y 0 otro caso. ¿Cuál es el problema con este modelo?

Supuesto 1.4 - Perturbaciones Esféricas

Supuesto 1.4 - Perturbaciones Esféricas. Asumir que todas las unidades tienen el mismo error de varianza.

$$E(\epsilon_i^2|X) = \sigma_0^2, \quad i = 1, 2, ..., n \tag{2.12}$$

Asumir también que no hay correlación entre las observaciones:

$$E(\epsilon_i \epsilon_j|X) = 0 \quad i = 1, 2, ..., n \quad i \neq j \tag{2.13}$$

Este supuesto señala que la varianza de todos los términos no observados es constante. Para ver esta nota:

$$\begin{aligned} Var(\epsilon_i|X) &= E(\epsilon_i^2|X) - E(\epsilon_i|X)^2 \quad \text{por definición de varianza condicional} \\ &= E(\epsilon_i^2|X) \quad \text{por exogeneidad estricta} \end{aligned}$$

Por lo tanto, es lo mismo decir $Var(\epsilon_i|X)$ o $E(\epsilon_i^2|X)$. En notación matricial, este supuesto puede ser escrito así:

$$\mathbb{E}(\epsilon\epsilon'|X) = \begin{pmatrix} \mathbb{E}(\epsilon_1\epsilon_1|X) & \mathbb{E}(\epsilon_1\epsilon_2|X) & \cdots & \mathbb{E}(\epsilon_1\epsilon_n|X) \\ \mathbb{E}(\epsilon_2\epsilon_1|X) & \mathbb{E}(\epsilon_2\epsilon_2|X) & \cdots & \mathbb{E}(\epsilon_2\epsilon_n|X) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}(\epsilon_n\epsilon_1|X) & \mathbb{E}(\epsilon_n\epsilon_2|X) & \cdots & \mathbb{E}(\epsilon_n\epsilon_n|X) \end{pmatrix} \tag{2.14}$$

$$= \begin{pmatrix} \sigma_0^2 & 0 & \cdots & 0 \\ 0 & \sigma_0^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_0^2. \end{pmatrix} \tag{2.15}$$

que puede ser resumido como $E(\epsilon\epsilon'|X) = \sigma_0^2 I_n$. Finalmente, cabe mencionar que los términos de errores que cumplen los supuestos de homocedasticidad y no autocorrelación son llamados **"errores esféricos"**

Supuesto 1.5 - Regresores No Estocásticos

Supuesto 1.5 - Regresores No Estocásticos. Las observaciones de las variables exógenas son fijos en muestras repetidas.

Asumir que los X son fijos quiere decir que, en muestras repetidas de X, los valores obtenidos $x_1, x_2, ..., x_K$ van a ser siempre los mismos; es decir, dejan de ser aleatorios. Bajo este supuesto, ya no es necesario hablar de esperanzas condicionales:

$$\begin{aligned} E(\epsilon) &= 0 \\ Var(\epsilon) &= \sigma_0^2 I_n \\ Cov(\epsilon_i, \epsilon_j) &= 0 \end{aligned}$$

Supuesto 1.6 - Normalidad de los Errores

Supuesto 1.6 - Normalidad de los Errores. La distribución de los errores siguen una distribución normal con media cero y varianza σ^2 condicionado a X

$$\epsilon|X \sim N(0, \sigma_0^2 I_n). \tag{2.16}$$

Por tanto, la distribución de ϵ condicionado sobre X no depende de X, es decir, ϵ y X son independientes.

3. Estimación de MCO

Suponer que hemos obtenido un estimador de β_0, denominado $\hat{\beta}$, entonces calculamos el residuo como:

$$\hat{\epsilon}_i = y_i - x_i'\hat{\beta} \tag{3.1}$$

Entonces la suma de residuos cuadrados (SCR) es:

$$SCR(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = (y - X\hat{\beta})'(y - X\hat{\beta}) \tag{3.2}$$

Es una función de $\hat{\beta}$ dado que los residuos dependen de este. El estimador MCO $\hat{\beta}_MCO$ de β_0 es el $\hat{\beta}$ que minimiza la función:

$$\hat{\beta}_{MCO} \equiv \underset{\hat{\beta}}{argmin} \quad SCR(\hat{\beta}) \tag{3.3}$$

Para encontrar los valores de $\hat{\beta}$ que minimizan SCR, derivamos las condiciones de primer orden estableciendo las derivadas parciales iguales a cero. Escribimos SCR($\hat{\beta}$) como:

$$\begin{aligned}
SCR(\hat{\beta}) &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= (y' - \hat{\beta}'X')(y - X\hat{\beta}) \quad \text{ya que} (X\hat{\beta})' = \hat{\beta}'X' \\
&= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\
&= y'y - 2y'X\hat{\beta} + \hat{\beta}'X'x\hat{\beta},
\end{aligned} \tag{3.4}$$

donde el escalar $\hat{\beta}'X'y$ es igual a la transpuesta de $y'X\hat{\beta}$. El término $y'y$ no depende de $\hat{\beta}$, por lo que puede ser ignorado en la diferenciación del SCR($\hat{\beta}$). Entonces:

$$\begin{aligned}
\frac{\partial SCR(\hat{\beta})}{\partial \hat{\beta}} &= \frac{\partial(-2y'X\hat{\beta})}{\partial \hat{\beta}} + \frac{\partial \hat{\beta}'X'X\hat{\beta}}{\partial \hat{\beta}} \\
&= -2(y'X)' + 2X'X\hat{\beta}
\end{aligned}$$

Las condiciones de primer orden se obtienen igualando esta ecuación igual a cero.

$$\begin{aligned}
-2X'y + 2X'X\hat{\beta} &= 0 \\
2X'X\hat{\beta} &= 2X'y \\
\underset{K\times 1}{\hat{\beta}_{MCO}} &= \underset{K\times K}{(X'X)^{-1}} \underset{K\times 1}{X'y}
\end{aligned} \tag{3.5}$$

Las ecuaciones $K \times 1$

$$X'X\hat{\beta} = X'y$$

son llamadas **ecuaciones normales**.

El vector de residuos evaluado en $\hat{\beta} = \hat{\beta}_{MCO}$,

$$\underset{n\times 1}{\hat{\epsilon}} \equiv y - X\hat{\beta}_{MCO}, \tag{3.6}$$

es llamado el vector de residuos MCO. Su elemento ith es $\hat{\epsilon}_i = y_i - x_i'\hat{\beta}$. Reordenando (3.6) da:

$$X'(y - X\hat{\beta}_{MCO}) = 0 \quad X'\hat{\epsilon} = 0$$

que es lo mismo que:

$$\frac{1}{n}\sum_{i=1}^{n} x_i \cdot \hat{\epsilon}_i = 0 \quad o \quad \frac{1}{n}\sum_{i=1}^{n} x_i \cdot (y_i - x_i'\hat{\beta}_{MCO}) = 0$$

el cual muestra que las ecuaciones normales pueden ser interpretados como una analogía simple de las condiciones de ortogonalidad $E(x_i \cdot \epsilon_i)$.

4. Más conceptos y Álgebra

Valores Predichos

Los valores predichos para la observación i es definido como $\hat{y}_i \equiv x_i'\hat{\beta}_{MCO}$. El vector del valor predicho, $\hat{y}$, es igual a $X\beta_{MCO}$. Por lo tanto, el vector de residuos MCO puede ser escrito como $\hat{\epsilon} = y - \hat{y}$.

Matriz de Proyección y Aniquiladora

La matriz proyección P y aniquiladora M son definidos como:

$$\underset{n\times n}{P} \equiv X(X'X)^{-1}X', \tag{4.1}$$

$$\underset{n\times n}{M} \equiv I_n - P, \tag{4.2}$$

Estos tienen las siguientes propiedades:

- Tanto P y M son simétricas e idempotentes.
- $PX = X$, de ahí el término de matriz de proyección,
- $MX = 0$, de ahí el término aniquilador.

Ya que $\hat{\epsilon}$ es el vector residual, la suma de cuadrado de los residuos, SCR, es igual a $\hat{\epsilon}'\hat{\epsilon}$. Esto puede ser escrito como:

$$SCR = \hat{\epsilon}'\hat{\epsilon} = \epsilon' M \epsilon, \tag{4.3}$$

La varianza del término de error

El estimador MCO de σ^2 (la varianza del término de error), denotada como $\hat{\sigma}^2$, es la suma de residuos al cuadrado dividido por $n - K$:

$$\hat{\sigma}^2 = \frac{SCR}{n-K} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-K}, \tag{4.4}$$

donde $n - K$ son los grados de libertad.

5. Propiedades Finitas de MCO

5.1 Insesgadez y Varianza del Estimador MCO

En esta sección, estudiaremos algunas de las propiedades del estimador MCO para muestras finitas. Es decir, estudiaremos las propiedades de $\hat{\beta}_{MCO}$ que se satisfacen para cualquier muestra.

Insesgadez del Estimador MCO

Insesgadez del Estimador MCO Bajo los supuestos de Linealidad, Exogeneidad Estricta y Rango Completo:

$$E(\hat{\beta}|X) = \beta_0 \tag{5.1}$$

Por tanto, $\hat{\beta}$ es un estimador insesgado de β_0

Para estar seguros de que entendemos el concepto de estimador insesgado, considere el siguiente ejemplo. Supongamos que nos interesa estimar el efecto del ingreso sobre la obesidad. En particular, asumimos que somos Dios y sabemos que el modelo poblacional (aunque restringido) es el siguiente:

$$bmi_i = \alpha_0 + \beta_0 income_i + \epsilon_i,$$

donde β_0 es el efecto del ingreso sobre la obesidad y es igual a 1, $\beta_0 = 1$. El siguiente paso es conseguir una muestra de la población y obtener un estimador de β_0, es decir, $\hat{\beta}$. Sin embargo, imaginese que consigamos $B = 50$ muestras de 1000 individuos, y para cada muestra $(b = 1, ..., 50)$ estimamos un β. Es así que para la primera muestra conseguimos $\hat{\beta}_1 = 0.9$, para la segunda $\hat{\beta}_2 = 0.94$, para la tercera $\hat{\beta}_3 = 0.7$, y así hasta la muestra 50 que se obtiene $\hat{\beta}_{50} = 0.5$. Como pueden observar, el estimador es muy cercano al verdadero parámetro poblacional para la primera y segunda muestra, y muy lejos en los otros casos. Sin embargo, esperamos que si promediamos las 50 muestras conseguiríamos un número muy cercano a uno, es decir: $(1/B)\sum_{b=1}^{B}\hat{\beta}_b \approx 1$. La forma en que se distribuyen estas estimaciones se denomina **distribución muestral de** $\hat{\beta}$, $f(\hat{\beta})$: es la función de densidad de probabilidad de $\hat{\beta}$, aproximada usando las 50 estimaciones de $\hat{\beta}$.

Demostración: Recordar que es importante colocar el estimador en función del término de error. Así derivamos la función de error muestral, que es definido como $\hat{\beta} - \beta_0$. El error muestral es:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'y, \\
&= (X'X)^{-1}X'(X\beta_0 + \epsilon) \\
&= (X'X)^{-1}X'X\beta_0 + (X'X)^{-1}X'\epsilon \\
&= \beta_0 + X(X'X)^{-1}X'\epsilon \\
\hat{\beta} - \beta_0 &= (X'X)^{-1}X'\epsilon \\
\hat{\beta} - \beta_0 &= A\epsilon \qquad (5.2)
\end{aligned}$$

donde $A = (X'X)^{-1}X'$. Por lo tanto,

$$\begin{aligned}
E(\hat{\beta} - \beta_0|X) &= E(A\epsilon|X) \\
&= AE(\epsilon|X) \\
&= 0 \qquad (5.3)
\end{aligned}$$

Varianza del Estimador MCO

Varianza del Estimador MCO Bajo los supuestos de Linealidad, Exogeneidad Estricta y Rango Completo y Errores Esféricos:

$$Var(\hat{\beta}_{MCO}) = \sigma_0^2(X'X)^{-1} \tag{1}$$

Demostración:

$$\begin{aligned}
Var(\hat{\beta}|X) &= E(\hat{\beta} - E[\hat{\beta}]|X)^2 \\
&= E(\hat{\beta} - \beta_0|X)^2 \\
&= E(A\epsilon|X)^2 \\
&= E(A\epsilon\epsilon'A'|X) \\
&= AE(\epsilon\epsilon'|X)A' \\
&= A(\sigma_0^2 I_n)A' \\
&= \sigma_0^2 AA' \\
&= \sigma_0^2 \cdot (X'X)^{-1} \quad \text{ya que} \quad AA' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}
\end{aligned} \tag{5.4}$$

5.2 Teorema de Gauss-Markov

Ya hemos encontrado el valor esperado y la varianza muestral del estimador MCO. La pregunta es: ¿Este estimador de MCO es el mejor estimador?

Teorema Gauss-Markov Bajo los supuestos de MCO, el estimador es eficiente en la clase de estimadores lineales insesgados. Es decir, para cualquier estimador insesgado $\hat{\beta}$ que es lineal en y, $Var(\hat{\beta}|X) \geq Var(\hat{\beta}_{MCO}|X)$ en términos matriciales.

Del estimador MCO $\hat{\beta}_{MCO} = (X'X)^{-1}X'y$ se aprecia que es lineal en y. Hay muchos otros estimadores de β_0 que son lineales e insesgados. El Teorema Gauss-Markov menciona que el estimador MCO es **eficiente** en el sentido que su matriz de varianza condicional $Var(\hat{\beta}_{MCO}|X)$ es el más pequeño entre todos los estimadores lineales insesgados. Por esta razón, **el estimador MCO es llamado Mejor Estimador Lineal Insesgado(MELI)** o **Best Linear Unbiased Estimador (BLUE)**.

Demostración: Ya que $\tilde{\beta}$ es lineal en y, este puede ser escrito como $\tilde{\beta} = Cy$ para alguna matriz C de dimensión $n \times K$, que posiblemente es una función de X. Sea $D = C - A$ o $C = D + A$, donde $A \equiv (X'X)^{-1}X'$. Entonces:

$$\begin{aligned}
\tilde{\beta} &= (D + A)y \\
&= Dy + Ay \\
&= D(X\beta + \epsilon) + \hat{\beta} \\
&= DX\beta + D\epsilon + \hat{\beta}
\end{aligned}$$

Tomando esperanza condicional en ambos lados, obtenemos:

$$E(\tilde{\beta}|X) = DX\beta + E(D\epsilon|X) + E(\hat{\beta}|X)$$

Ya que tanto $\hat{\beta}$ y $\tilde{\beta}$ son insesgados y dado que $E(D\epsilon|X) = DE(\epsilon|X) = 0$, resultad que $DX\beta = 0$. Para que esto sea cierto β_0, es necesario que $DX = 0$. Así, $\tilde{\beta} = D\epsilon + \hat{\beta}$ y

$$\begin{aligned}\tilde{\beta} - \beta_0 &= D\epsilon + (\hat{\beta} - \beta_0) \\ &= (D + A)\epsilon\end{aligned}$$

Así,

$$\begin{aligned}Var(\tilde{\beta}|X) &= Var(\tilde{\beta} - \beta_0|X) \\ &= Var[(D + A)\epsilon|X] \\ &= (D + A)Var(\epsilon|X)(D + A)' \\ &= \sigma^2 \cdot (D + A)(D' + A') \\ &= \sigma^2(DD' + AD' + DA' + AA')\end{aligned}$$

Pero $DA' = DX(X'X)^{-1} = 0$ ya que $DX = 0$. También, $AA' = (X'X)^{-1}$. Por tanto,

$$\begin{aligned}Var(\tilde{\beta}|X) &= \sigma^2[DD' + (X'X)^{-1}] \\ &\geq \sigma^2(X'X)^{-1} \quad \text{ya que } DD' \text{ es semidefinida positiva} \\ &= Var(\hat{\beta}|X)\end{aligned}$$

5.3 Covarianza entre $\hat{\beta}$ y ϵ

Covarianza entre $\hat{\beta}$ y ϵ Bajo los supuestos de Linealidad, Exogeneidad Estricta, Rango Completo y Errores Esféricos: $Cov(\hat{\beta}, \hat{\epsilon}|X) = 0$, donde $\hat{\epsilon} \equiv y - X\hat{\beta}$

Demostración

$$\begin{aligned}Cov(\hat{\beta}, \epsilon|X) &= E[\hat{\beta} - E(\hat{\beta}|X)][\hat{\epsilon} - E(\epsilon|X)]'|X \\ &= E[\hat{\beta} - \beta_0|X][M\epsilon - E(M\epsilon|X)]'|X \\ &= E[A\epsilon\epsilon'M'|X] \\ &= AE[\epsilon\epsilon'|X]M \\ &= AE[\sigma_0^2 I_n]M \\ &= \sigma_0^2(X'X)^{-1}X'M \\ &= \sigma_0^2(X'X)^{-1}\underbrace{(MX)'}_{0} \\ &= 0\end{aligned}$$

5.4 Propiedades Finitas de $\hat{\sigma}_2$

La varianza del estimador MCO es:

$$Var(\hat{\beta}_{MCO}) = \sigma_0^2(X'X)^{-1}$$

Sin embargo, ésta formula es desconocida, excepto en el caso que conozcamos σ_0^2. Podemos decir que es la varianza poblacional de $\hat{\beta}$. Sin embargo, podemos usar los datos para estimar σ^2, que entonces nos permite estimar $Var(\hat{\beta}_{MCO}$. Sabemos que:

$$\hat{\sigma}^2 = \frac{SCR}{n - K} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n - K}$$

Si $\widehat{\sigma}^2$ es el estimador de σ^2, un estimador natural de $Var(\widehat{\beta}|X = \sigma_0^2 \cdot (X'X)^{-1}$ es

$$\widehat{Var(\widehat{\beta}|X)} \equiv \widehat{\sigma}^2(X'X)^{-1} \tag{5.5}$$

Ya que $\widehat{\sigma}^2$ es también un estimador del parámetro poblacional σ_0^2, también nos gustaría saber si es insesgado.

Insesgadez de $\widehat{\sigma}^2$ Bajo los supuestos de MCO: $E(\widehat{\sigma}^2|X) = \sigma_0^2$ (y por tanto $E(\widehat{\sigma}^2 = \sigma_0^2)$, siempre que $n > K$, de modo que $\widehat{\sigma}^2$ está bien definido.

Demostración Ya que $\widehat{\sigma}^2 = \widehat{\sigma}'\widehat{\sigma}/(n-K)$, la prueba equivale a mostrar que $E(\widehat{\epsilon}'\widehat{\epsilon}|X) = (n-K)\sigma_0^2$. Sabemos que $\widehat{\epsilon}'\widehat{\epsilon} = \epsilon' M\epsilon$, donde M es la matriz aniquiladora. La prueba consiste en probar dos propiedades: (1) $E(\epsilon' M\epsilon|X) = \sigma_0^2 \cdot traza(M)$, y (2) $traza(M) = n - K$.

1. Probar que $E(\epsilon' M\epsilon|X) = \sigma_0^2 \cdot traza(M)$: Ya que $\epsilon' M\epsilon = \sum_{i=1}^n \sum_{j=1}^n m_{ij}\epsilon_i\epsilon_j$, tenemos:

$$\begin{aligned} E(\epsilon' M\epsilon|X) &= \sum_{i=1}^n \sum_{j=1}^n m_{ij} E(\epsilon_i\epsilon_j|X) \quad \text{porque } m_i j\text{'s son funciones de } X \\ &= \sum_{i=1}^n m_{ii}\sigma^2 \quad \text{ya que } E(\epsilon_i\epsilon_j|X) = 0 \text{ para } i \neq j \text{ por Homocedasticidad} \\ &= \sigma^2 \sum_{i=1}^n m_{ii} \\ &= \sigma^2 \cdot traza(M) \end{aligned} \tag{5.6}$$

2. Probar que $traza(M) = n - K$:

$$\begin{aligned} traza(M) &= traza(I_n - P) \quad M \equiv I_n - P \\ &= traza(I_n) - traza(P) \quad \text{el operador de traza es lineal} \\ &= n - traza(P) \end{aligned} \tag{5.7}$$

y

$$\begin{aligned} traza(P) &= traza[X(X'X)^{-1}X'] \quad \text{ya que} P = X(X'X)^{-1}X' \\ &= traza[(X'X)^{-1}X'X] \quad \text{ya que} traza(AB) = traza(BA) \\ &= traza(I_k) \\ &= K \end{aligned} \tag{5.8}$$

Por tanto $traza(M) = n - K$.

6. Ejemplo en R: Ecuación Salarial en Perú

En este apartado, mostramos cómo estimar una regresión de MCO usando R. Es así que usamos la Encuesta Nacional de Hogares (ENAHO) 2020 y corremos una regresión salarial.

6.1. Usando el comando `lm`

Lo primero es cargar la base que se encuentra en formato Stata. Existen diversas librerías para cargar este tipo de bases, usaremos en este caso la librería `rio`.

Esta librería permite cargar cualquier formato de base de datos a R:

```r
# Llamamos librerías
library(rio)
```

Cargamos la base de datos

```r
# Cargar los datos
enaho2020 <- import(paste0(inputs, "/clean_enaho_2020.dta"))

# Observar 6 primeras observaciones
head(enaho2020)
```

```
##   conglome vivienda hogar codperso ubigeo dpto area zona edad sexo sch niv_educ
## 1   005002      011    11       01 010101    1    1    2   66    0   5        4
## 2   005002      011    11       03 010101    1    1    2   47    0  14        8
## 3   005002      023    11       01 010101    1    1    2   62    1  11        6
## 4   005002      049    11       03 010101    1    1    2   51    1  11        6
## 5   005002      049    11       04 010101    1    1    2   48    0  16       10
## 6   005005      009    11       01 010101    1    1    2   57    1  11        6
##   estatus jefe ocu500      wage  fac500a
## 1       0    1      1  182.7500 51.73801
## 2       0    0      1 5272.5835 42.14626
## 3       0    1      1 2681.2500 37.88447
## 4       0    0      1  710.6667 59.99160
## 5       0    0      1 1995.4166 42.14626
## 6       1    1      1 1602.4166 28.81618
```

Si deseamos conocer el nombre de las variables:

```r
# Nombre de variables
names(enaho2020)
```

```
##  [1] "conglome" "vivienda" "hogar"    "codperso" "ubigeo"   "dpto"
##  [7] "area"     "zona"     "edad"     "sexo"     "sch"      "niv_educ"
## [13] "estatus"  "jefe"     "ocu500"   "wage"     "fac500a"
```

Podemos obtener la información sobre el número de individuos y variables usando la función `dim`:

```r
# Numero de observaciones y variables
dim(enaho2020)
```

```
## [1] 45876    17
```

El primer elemento indica el número de filas (individuos) en la base de datos, mientras que el segundo elemento indica el número de columnas (variables). Por lo tanto, nuestra muestra tiene 45876 individuos y 17 variables. Podemos hacer un resumen estadístico de las variables usando el comando `summary`:

```r
# Resumen Estadistico de variables
summary(enaho2020)
```

```
##    conglome           vivienda            hogar             codperso
##  Length:45876       Length:45876       Length:45876       Length:45876
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##     ubigeo                dpto           area             zona
##  Length:45876       Min.   : 1.0   Min.   :0.0000   Min.   :1.000
##  Class :character   1st Qu.: 7.0   1st Qu.:0.0000   1st Qu.:1.000
##  Mode  :character   Median :14.0   Median :1.0000   Median :2.000
##                     Mean   :13.1   Mean   :0.6346   Mean   :2.136
##                     3rd Qu.:19.0   3rd Qu.:1.0000   3rd Qu.:3.000
##                     Max.   :25.0   Max.   :1.0000   Max.   :4.000
##       edad            sexo             sch             niv_educ
##  Min.   :14.00   Min.   :0.0000   Min.   : 0.000   Min.   : 1.000
##  1st Qu.:32.00   1st Qu.:0.0000   1st Qu.: 6.000   1st Qu.: 4.000
##  Median :44.00   Median :1.0000   Median :11.000   Median : 6.000
##  Mean   :44.27   Mean   :0.6049   Mean   : 9.621   Mean   : 5.902
##  3rd Qu.:55.00   3rd Qu.:1.0000   3rd Qu.:13.000   3rd Qu.: 8.000
##  Max.   :98.00   Max.   :1.0000   Max.   :18.000   Max.   :12.000
##     estatus            jefe            ocu500      wage
##  Min.   :0.0000   Min.   :0.000   Min.   :1   Min.   :    2.0
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1   1st Qu.:  325.9
##  Median :1.0000   Median :1.000   Median :1   Median :  766.4
##  Mean   :0.5997   Mean   :0.566   Mean   :1   Mean   : 1176.8
##  3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:1   3rd Qu.: 1480.8
##  Max.   :1.0000   Max.   :1.000   Max.   :1   Max.   :74092.2
##     fac500a
##  Min.   :   0.3713
##  1st Qu.:  88.8874
##  Median : 182.0445
##  Mean   : 279.8328
##  3rd Qu.: 354.0065
##  Max.   :2820.7278
```

Podemos estimar diferentes versiones del siguiente modelo de regresión:

$$log(wage_i) = \beta_1 + \beta_2 sch_i + x_i'\delta + \epsilon_i,$$

donde $wage_i$ es el ingreso salarial mensual para el individuo i. El comando `lm` implementa la regresión MCO en R. Considerar los siguientes modelos:

```
# Modelos con diferentes controles
ols_1 <- lm(log(wage) ~ sch,
            data = enaho2020)
ols_2 <- lm(log(wage) ~ sch + sexo,
            data = enaho2020)
ols_3 <- lm(log(wage) ~ sch + sexo + edad,
            data = enaho2020)
ols_4 <- lm(log(wage) ~ sch + sexo + edad + estatus,
            data = enaho2020)
```

Para crear una tabla con todas las regresiones, usamos el paquete `memisc`:

```
library("memisc")
table_1 <- mtable("Model 1" = ols_1,
                  "Model 2" = ols_2,
                  "Model 3" = ols_3,
```

```
                   "Model 4" = ols_4,
                   summary.stats = c("N", "R-squared"),
                   coef.style = "default")
table_1
```

Los resultados se muestran en la Tabla 1:

Tabla 1 - Ecuación Salarial usando ENAHO

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Intercepto	5.4465***	5.2631***	4.8954***	4.8642***
	(0.0107)	(0.0120)	(0.0214)	(0.0215)
Escolaridad	0.1102***	0.1100***	0.1179***	0.1181***
	(0.0010)	(0.0010)	(0.0010)	(0.0010)
Sexo		0.3054***	0.3031***	0.2710***
		(0.0095)	(0.0095)	(0.0097)
Edad			0.0066***	0.0055***
			(0.0003)	(0.0003)
Estado civil				0.1632***
				(0.0099)
R^2	0.2114	0.2286	0.2357	0.2402
Num. obs.	45876	45876	45876	45876

$^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$.

6.2. Usando operaciones matriciales

Ahora, estimamos MCO y la matriz de varianza-covarianza usando operaciones matriciales. Primero, creamos la matriz X de variables independientes y el vector y:

```
# Crear X e y
X <- cbind(1,                # Incluye columna de unos
           enaho2020$sch,
           enaho2020$sexo,
           enaho2020$edad,
           enaho2020$estatus)
colnames(X) <- c("intercepto",
                 "escolaridad", "sexo", "edad", "estado civil")
y <- log(enaho2020$wage)
```

Para obtener el estimador $\hat{\beta}$:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'y$$

Computamos:

```
b_hat <- solve(crossprod(X)) %*% crossprod(X, y) # (X'X)^-1 X'y
b_hat
```

```
##                     [,1]
## intercepto   4.864180026
## escolaridad  0.118109938
## sexo         0.271037276
## edad         0.005513234
```

```
## estado civil 0.163247777
```

Los resultados son los mismos que se reportan en el modelo 4 de la Tabla 1. Ahora, estimamos $\hat{\sigma}^2$ usando la ecuación:

$$\hat{\sigma}^2 = \frac{SCR}{n-K} = \frac{\hat{\epsilon}'\hat{\epsilon}}{n-K}$$

```r
# Creamos sigma squared hat
y_hat <- X %*% b_hat
e_hat <- y - y_hat
sigma2_hat <- crossprod(e_hat) / (length(y) - ncol(X))  # n-K
sigma2_hat
```

```
##           [,1]
## [1,] 0.9844928
```

Finalmente, computamos la matriz de varianza-covarianza de $\hat{\beta}$ junto con los errores estándar.

$$\widehat{Var(\hat{\beta}|X)} \equiv \hat{\sigma}^2(X'X)^{-1}$$

```r
# Creamos errores estándar.
V_hat <- drop(sigma2_hat) * solve(crossprod(X))
se <- sqrt(diag(V_hat))
se
```

```
##   intercepto  escolaridad         sexo         edad estado civil
## 0.0214679441 0.0010466643 0.0096734886 0.0003273275 0.0099023595
```

Son los mismos errores estándar que los reportados por la función `lm`.