Clase auxiliar # 1

Estimación de Mínimos Cuadrados Ordinarios y densidades de probabilidad

¡Recordar!

Variables y parámetros:

- Número de observaciones: $N \in \mathbb{R}$.
- Número de variables explicativas : $K \in \mathbb{R}$.
- Vector de observaciones de la variable dependiente: $Y = (Y_1, ..., Y_N) \in \mathbb{R}^{N\times 1}$.
- Matriz de observaciones de las variables independientes: $X \in \mathbb{R}^{N\times K}$
- Vector de parámetros poblacionales: $\beta = (\beta_1, ..., \beta_N) \in \mathbb{R}^{K\times 1}$
- Vector de errores aleatorios: $U = (U_1, ..., U_N) \in \mathbb{R}^{N\times 1}$. También se suele notar como ε.

Supuestos de MCO:

1. **Linealidad de parámetros:** $Y = X\beta + U$.

   Notar que la linealidad es sobre los parámetros, no sobre las variables. Podemos tener perfectamente un modelo de la forma: $\tanh(Y_i) = \beta_1 \ln(X_{i1}) + \beta_2 \exp(X_{i2}) + \beta_3 X_{i3}^{0,7} + \beta_4 \arctan(X_{i3}) + U_i$.
2. **Matriz $X^T X$ invertible:** $rg(X^T X) = K$. Se puede verificar.
3. **Errores nulos en esperanza:** $\mathbb{E}(U) = 0$. Supuesto aceptable.
4. **Independencia y homocedasticidad:** $\mathbb{V}(U) = \sigma^2 I_N$. Se puede testear.
5. **Exogeneidad:** $\mathbb{E}(X^T U) = 0$. Supuesto fuerte y problemático. En ciencias sociales siempre hay correlación entre variables que observamos y que no observamos.

1. Estimación teórica de MCO

Considere el modelo $Y = X\beta + U$. Llamaremos $\hat{\beta}$ a un estimador genérico de β, $\hat{Y} = X\hat{\beta}$ a la predicción de la variable dependiente Y según el estimador $\hat{\beta}$ y $\hat{U} = Y - \hat{Y}$ a los residuos del modelo –diferencia entre el valor real de la variable dependiente y su predicción–.

1. Demuestre que si $X^T X$ es invertible y definida positiva, entonces $\hat{\beta}_{MCO} = (X^T X)^{-1} X^T Y$ es minimizador global de la suma de los residuos al cuadrado.

   Solución:

   Queremos demostrar que $\hat{\beta}_{MCO} = \underset{\hat{\beta}}{\text{argmin}} \sum_{i=1}^{N} \hat{U}_i^2(\hat{\beta})$.

   En primer lugar reescribiremos la suma de los residuos al cuadrado.

$$\begin{aligned}
\sum_{i=1}^{N} \hat{U}_i^2(\hat{\beta}) &= \hat{U}^T \hat{U} \\
&= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\
&= (Y^T - X^T \hat{\beta}^T)(Y - X\hat{\beta}) \\
&= (Y^T - \hat{\beta}^T X^T)(Y - X\hat{\beta}) \\
&= Y^T Y - Y^T X\hat{\beta} - \hat{\beta}^T X^T Y + \hat{\beta} X^T X \hat{\beta} \\
&= Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta} X^T X \hat{\beta}
\end{aligned}$$

   Para que $\hat{\beta}_{mco}$ sea óptimo global, debe satisfacer las condiciones de primer y segundo orden.

   Condiciones de primer orden: $\nabla_{\hat{\beta}} \left(\sum_{i=1}^{N} \hat{U}_i^2(\hat{\beta}_{MCO}) \right) = \vec{0}_{K \times 1}$.

   Notemos que:

$$\begin{aligned}
\nabla_{\hat{\beta}} \left(\sum_{i=1}^{N} \hat{U}_i^2(\hat{\beta}) \right) &= \nabla_{\hat{\beta}} (Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta} X^T X \hat{\beta}) \\
&= -2\nabla_{\hat{\beta}} (\hat{\beta}^T X^T Y) + \nabla_{\hat{\beta}} (\hat{\beta}^T X^T X \hat{\beta}) \\
&= -2X^T Y + 2X^T X \hat{\beta}
\end{aligned}$$

   Considerando la hipótesis de que $X^T X$ sea invertible, se obtiene que $\hat{\beta}_{MCO}$ cumple la condición de primer orden.

$$\begin{aligned}
-2X^T Y + 2X^T X \hat{\beta} = 0 \\
\hat{\beta} = (X^T X)^{-1} X^T Y
\end{aligned}$$

Condiciones de segundo orden: la matriz $\mathcal{H}_{\hat{\beta}}\left(\sum_{i=1}^{N}\hat{U}_i^2(\hat{\beta}_{MCO})\right)$ es definida positiva.

$$\nabla_{\hat{\beta}}\left(\sum_{i=1}^{N}\hat{U}_i^2(\hat{\beta})\right) = -2X^TY + 2X^TX\hat{\beta}$$

$$\Rightarrow \mathcal{H}_{\hat{\beta}}\left(\sum_{i=1}^{N}\hat{U}_i^2(\hat{\beta})\right) = 2X^TX$$

Lo anterior se cumple $\forall\hat{\beta}\in\mathbb{R}^{K\times 1}$, en particular para $\hat{\beta}_{MCO}$.

$$\Rightarrow \mathcal{H}_{\hat{\beta}}\left(\sum_{i=1}^{N}\hat{U}_i^2(\hat{\beta}_{MCO})\right) = 2X^TX$$

Por hipótesis, X^TX es definida positiva. Es decir, $z^TX^TXz > 0, \forall z\in\mathbb{C}^n$ [1].

Es claro ver que $\mathcal{H}_{\hat{\beta}}\left(\sum_{i=1}^{N}\hat{U}_i^2(\hat{\beta}_{MCO})\right) = 2X^TX$ es definida positiva. En efecto, $z^TX^TXz > 0 \Rightarrow z^T(2X^TX)z > 0$.

2. Suponiendo que se cumplen las hipótesis anteriores, obtenga el estimador β_{MCO} para un modelo de la forma $Y_i = \beta_0 + \beta_1 X_{i,1} + U_i$.

   Notemos que el modelo se puede escribir matricialmete como $Y = X\beta + U$. Donde:

   - $Y = (Y_1, ..., Y_N)$.
   - $X = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{N,1}\end{pmatrix}$.
   - $\beta = (\beta_0, \beta_1)$.
   - $U = (U_1, ..., U_N)$.

Conocemos la fómula genérica del estimador MCO y hemos identificado sus distintas componentes para este caso particular. Lo que queda ahora es trabajo algebraico. Que comience la diversión.

Primero notemos que: $\hat{\beta}_{MCO} = (X^TX)^{-1}(X^TY) = \left(\frac{1}{N}X^TX\right)^{-1}\left(\frac{1}{N}X^TY\right)$.

$$X^TX = \begin{pmatrix} 1 & X_{1,1} \\ \vdots & \vdots \\ 1 & X_{N,1}\end{pmatrix}\begin{pmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{N,1}\end{pmatrix} = \begin{pmatrix} N & \sum_{i=1}^{N} X_{i,1} \\ \sum_{i=1}^{N} X_{i,1} & \sum_{i=1}^{N} X_{i,1}^2\end{pmatrix}$$

$$\Rightarrow \frac{1}{N}X^TX = \begin{pmatrix} 1 & \overline{X_1} \\ \overline{X_1} & \overline{X_1^2}\end{pmatrix}$$

$$\Rightarrow \left(\frac{1}{N}X^TX\right)^{-1} = \frac{1}{\overline{X_1^2} - \overline{X_1}^2}\begin{pmatrix} \overline{X_1^2} & -\overline{X_1} \\ -\overline{X_1} & 1\end{pmatrix}$$

[1] No se asusten, sólo trabajaremos con números reales.

Por otro lado...

$$X^TY = \begin{pmatrix} 1 & \cdots & 1 \\ X_{1,1} & \cdots & X_{N,1} \end{pmatrix} \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N Y_i \\ \sum_{i=1}^N X_{i,1}Y_i \end{pmatrix}$$

$$\Rightarrow \frac{1}{N}X^TY = \begin{pmatrix} \overline{Y} \\ \overline{X_1Y} \end{pmatrix}$$

Reemplazando ambas expresiones obtenidas en la fórmula del estimador MCO...

$$\hat{\beta}_{MCO} = \left(\frac{1}{N}X^TX\right)^{-1}\left(\frac{1}{N}X^TY\right) = \frac{1}{\overline{X_1^2} - \overline{X_1}^2} \begin{pmatrix} \overline{X_1^2} & -\overline{X_1} \\ -\overline{X_1} & 1 \end{pmatrix} \begin{pmatrix} \overline{Y} \\ \overline{X_1Y} \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \frac{1}{\overline{X_1^2} - \overline{X_1}^2} \begin{pmatrix} \overline{X_1^2}\overline{Y} - \bar{X}_1\overline{X_1Y} \\ -\overline{X_1}\bar{Y} + \overline{X_1Y} \end{pmatrix}$$

Conocemos las fórmulas poblacionales y muestrales de la covarianza y la varianza:

- $Cov(X_1, Y) = \mathbb{E}(X_1Y) - \mathbb{E}(X_1)\mathbb{E}(Y) \rightarrow \widehat{Cov(X_1, Y)} = \overline{X_1Y} - \bar{X}_1\bar{Y}$.
- $\mathbb{V}(X_1) = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 \rightarrow \widehat{\mathbb{V}(X_1)} = \overline{X_1^2} - \overline{X_1}^2$

Deléitense...

$$\hat{\beta}_1 = \frac{-\overline{X_1}\bar{Y} + \overline{X_1Y}}{\overline{X_1^2} - \overline{X_1}^2} = \frac{\widehat{Cov(X_1, Y)}}{\widehat{\mathbb{V}(X_1)}}$$

Notemos que mientras más grande sea la covarianza entre las variables dependiente e independiente, la pendiente estimada será más grande. Tiene todo el sentido, ¿no?

Cabe destacar que el desarrollo matemático nos permite medir sólo correlación, no causalidad. Necesitaremos argumentos sólidos (muchas veces teóricos) para argumentar la existencia de causalidad.

Finalmente, obtenemos $\hat{\beta}_0$:

$$\begin{aligned}
\hat{\beta}_0 &= \frac{\overline{X_1^2}\overline{Y} - \bar{X}_1\overline{X_1Y}}{\overline{X_1^2} - \overline{X_1}^2} \\
&= \frac{\overline{X_1^2}\overline{Y} - \overline{X_1}^2\overline{Y} + \overline{X_1}^2\overline{Y} - \bar{X}_1\overline{X_1Y}}{\overline{X_1^2} - \overline{X_1}^2} \\
&= \overline{Y}\left(\frac{\overline{X_1^2} - \overline{X_1}^2}{\overline{X_1^2} - \overline{X_1}^2}\right) + \left(\frac{\overline{X_1}^2\overline{Y} - \bar{X}_1\overline{X_1Y}}{\overline{X_1^2} - \overline{X_1}^2}\right) \\
&= \overline{Y} - \overline{X_1}\left(\frac{\overline{X_1Y} - \overline{X_1}\bar{Y}}{\overline{X_1^2} - \overline{X_1}^2}\right) \\
&= \overline{Y} - \overline{X_1}\hat{\beta}_1
\end{aligned}$$

Si reordenamos los términos obtenemos $\overline{Y} = \hat{\beta}_0 + \overline{X_1}\hat{\beta}_1$, lo que nos dice que los estimadores son construidos de manera que el modelo pasa por el promedio muestral.

2. Estimación computacional de MCO

En este ejercicion analizaremos cómo afecta el gasto per cápita en salud (US$) a la tasa de mortalidad de menores de 5 años (por cada 1.000) a nivel global. Para lo anterior, ocuparemos datos proporcionados por el Banco Mundial correspondientes al año 2014 y donde cada muestra corresponde a un país. La base de datos se encuentra en U-Cursos bajo el nombre de `mortalidad_gastosalud.xls`.

1. **Stata:**
   - Importe los datos y realice un gráfico *scatter* (de disperción) que muestre cómo se comportan conjuntamente ambas variables.
   - Regresione la mortalidad en función del gasto per cápita y analice los resultados. Si le parece conveniente, puede realizar transformaciones en las variables con tal de mejorar el ajuste.
2. **Matlab:** importe los datos y obtenga el estimador MCO para la misma regresión de la parte anterior. Compare los resultados con los de Stata.
3. **Solver de Excel:** proponga un estimador inicial $\hat{\beta}$ y obtenga el estimador MCO para la misma regresión de las partes anteriores. Recordar que queremos minimizar la suma de los residuos al cuadrado. Compare los resultados con los de Stata y Matlab.

Solución: ver material docente en U-Cursos.

3. Función de densidad conjunta, marginal y condicional

Considere la siguiente densidad de probabilidad conjunta entre las variables aleatorias X e Y [2].

$$f(x,y) = \begin{cases} 8xy & \text{si } 0 < x < y < 1 \\ \\ 0 & \text{en caso contrario} \end{cases}$$

1. Muestre que la función f está bien definida. Es decir: $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$ y $f(x,y) = 0 \forall (x,y) \in \mathbb{R}^2$.

   Primero veamos que $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = 1$:

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx &= \int_{0}^{1}\int_{x}^{1} 8xydydx \\
&= 8\int_{0}^{1} x\left(\int_{x}^{1} ydy\right)dx \\
&= 4\int_{0}^{1}(x - x^3)dx \\
&= 4\left(\int_{0}^{1} xdx - \int_{0}^{1} x^3 dx\right) \\
&= 4\left(\frac{1}{2} - \frac{1}{4}\right) \\
&= 1
\end{aligned}$$

[2] Referencia: Luis Rincón, Universidad Nacional Autónoma de México (`http://lya.fciencias.unam.mx/lars/0625/`)

Ahora notemos que cuando $0 < x < y < 1$, $x, y \geq 0 \Rightarrow f(x,y) = 8xy \geq 0$.
En el caso contrario: $f(x,y) = 0 \geq 0$.

2. Obtenga las densidades marginales de X e Y: $f_X(x)$, $f_Y(y)$.

   Solución:

   Calculemos primero $f_X(x)$.

   Cuando $x \in [0,1]$:

$$f(x,y) = 8xy \Rightarrow f_X(x,y) = \int_x^1 f(x,y)dy = \int_x^1 8xydy = 8\int_x^1 ydy = 4x(1-x^2)$$

   Luego

$$f(x,y) = \begin{cases} 4x(1-x^2) & \text{si } 0 < x < 1 \\ \\ 0 & \text{en caso contrario} \end{cases}$$

   Ahora calculemos $f_Y(y)$.

   Cuando $y \in [0,1]$:

$$f(x,y) = 8xy \Rightarrow f_Y(x,y) = \int_0^y f(x,y)dy = \int_x^1 8xydy = 8y \cdot \frac{y^2}{2} = 4y^3$$

3. Obtenga las densidades de probabilidad condicionales $f_{X|Y}(x|y), f_{Y|X}(y|x)$.

   Recordemos que $f_{W|Z}(w|z) = \frac{f(w,z)}{f_Z(w,z)}$.

$$f_{X|Y}(x,y) = \begin{cases} \frac{8xy}{4y^3} = \frac{2x}{y^2} & \text{si } 0 < x < 1 \\ \\ 0 & \text{en caso contrario} \end{cases}$$

$$f_{X|Y}(x,y) = \begin{cases} \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2} & \text{si } 0 < x < y < 1 \\ \\ 0 & \text{en caso contrario} \end{cases}$$