

Clase auxiliar # 2

Software Stata, teorema de Frisch-Waugh e interpretación

¡Recordar!

Matriz de proyección: $P_X = X(X^T X)^{-1} X^T$

Matriz de aniquiladora o de proyección ortogonal: $M_X = I - P_X$

Teorema de Frisch-Waugh: en los dos modelos que se enuncian a continuación, la estimación por MCO de β_B es equivalente. Es decir, $\hat{\beta}_B^{\text{Modelo 1}} = \hat{\beta}_B^{\text{Modelo 2}}$.

- Modelo 1: $Y = \vec{\beta}_0 + X_A \beta_A + X_B \beta_B + U$
- Modelo 2: $\underbrace{M_{X_A} Y}_{\hat{U}_{Y, X_A}} = \underbrace{M_{X_A} \vec{\beta}_0}_{\vec{\beta}_0'} + \underbrace{M_{X_A} X_B}_{\hat{U}_{X_B, X_A}} \beta_B + \underbrace{M_{X_A} U}_{U'}$

$\hat{U}_{W, Z}$ es el residuo obtenido al regresionar W en función de Z .

Juguemos con Stata

Ocuparemos la base de datos `Casen 2015.dta` y el código `Auxiliar 02 - Gasto en educación.do` (ver Material Docente). Antes que todo, presentaremos nuestras variables de interés:

- `expr`: se interpreta como la cantidad de personas en la población, que representa una persona en la muestra. Sirve para darle más peso a unas observaciones que a otras.
- `folio`: identificador único del hogar. En la base de datos, cada observación es una persona, por lo que distintas observaciones pueden tener un mismo folio.
- `edad`: edad de la persona.
- `e13b`: gasto mensual en educación parvularia, básica o media.
- `e15a`: gasto mensual en educación técnica superior, profesional o posgrado.
- `qaut`: quintil al que pertenece el hogar.
- `pco1`: parentesco con el/la jefe(a) de hogar ($pco1 = 1 \Rightarrow$ la persona es el/la jefe(a) de hogar).
- `ytoth`: ingreso total del hogar.

1. **Tratamiento de la base de datos:** mediante el comando `keep` deje sólo las variables de interés. Usando `gen` y `replace` cree variables que almacenen el gasto total en educación, la edad del jefe de hogar y el cuadrado de la edad del jefe de hogar. Finalmente, usando `collapse` haga que el nivel de agregación de la base sea por hogar y no por persona.

Solución: ver código al final del documento.

2. **Regresiones**¹: escriba matemáticamente las siguientes regresiones y ejecútelas en *Stata* considerando sólo hogares donde el/la jefe(a) tenga entre 25 y 65 años.

Regresión 1: gasto en educación del hogar en función de la edad del jefe de hogar y el cuadrado de la edad del jefe de hogar.

Regresión 2: gasto en educación del hogar en función del ingreso total del hogar, la edad del jefe de hogar y el cuadrado de la edad del jefe de hogar.

Regresión 3: gasto en educación del hogar en función del ingreso total del hogar, la edad del jefe de hogar y el cuadrado de la edad del jefe de hogar, todo esto clasificado por quintiles.

Solución:

Regresión 1: $\text{gastoeduc}_i = \beta_0 + \beta_1 \text{ejh}_i + \beta_2 \text{ejh}_i^2 + U_i$.

Regresión 2: $\text{gastoeduc}_i = \beta_0 + \beta_1 \text{ejh}_i + \beta_2 \text{ejh}_i^2 + \beta_3 \text{ingreso}_i + U_i$.

Regresión 3: $\text{gastoeduc}_i = \beta_0 + \sum_{q=1}^5 \beta_{1,q} \text{ejh}_i d_{i,q} + \sum_{q=1}^5 \beta_{2,q} \text{ejh}_i^2 d_{i,q} + \sum_{q=1}^5 \beta_{3,q} \text{ingreso}_i d_{i,q} + U_i$.

La implementación de las regresiones en *Stata* están en el código al final del documento.

A partir de ahora, sólo considere los resultados de la **segunda regresión**.

3. **Interpretación de resultados:** comente resultados relacionados a los signos de los coeficientes, significancia individual, significancia, número de observaciones y ajuste del modelo.

```
Linear regression               Number of obs   =    61.503
                              F(3, 61499)         =    186,12
                              Prob > F           =    0,0000
                              R-squared          =    0,2127
                              Root MSE       =    1,2e+05
```

gastoeduc	Robust		t	P> t	[95% Conf. Interval]	
	Coef.	Std. Err.				
ejh	8021,156	624,0248	12,85	0,000	6798,066	9244,246
ejh2	-86,3942	6,624542	-13,04	0,000	-99,37832	-73,41008
ytoth	,0496166	,0025888	19,17	0,000	,0445426	,0546906
_cons	-180994,8	14397,64	-12,57	0,000	-209214,3	-152775,4

¹Podríamos cuestionarnos la causalidad que tienen las variables consideradas, sin embargo por esta vez la asumiremos.

Solución:

- **Signos de los coeficientes:** $\hat{\beta}_{ejh} = 8021,146 > 0$ y $\hat{\beta}_{ejh2} = -86,156 < 0$. Lo anterior implica que el retorno marginal de la edad del jefe de hogar es creciente en primera instancia y luego se vuelve decreciente. Por otro lado $\hat{\beta}_{ytoh} = 0,0496166 > 0$. Esto último significa que a mayor ingreso, la gente tenderá a gastar más en educación.
- **Significancia individual:** se puede observar que todas las variables son significativas a un nivel de significancia $\alpha = 0,05^2$. Podemos verificar ésto de 3 maneras equivalentes:
 - El p-valor representa la probabilidad de error tipo I³. Según Wikipedia, es “la probabilidad de obtener un resultado al menos tan extremo como el que realmente se ha obtenido, suponiendo que la hipótesis nula es cierta”. Si el p-valor $< \alpha$, entonces se puede rechazar la hipótesis nula (la variable X_i es estadísticamente significativa para explicar la variable dependiente). En caso contrario, no se puede rechazar la hipótesis nula ni concluir la significancia individual.
 - Se puede concluir lo mismo observando el estadístico $t = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}}$. Para valores suficientemente grandes de N (número de observaciones), la región de rechazo para la hipótesis nula estará dada por los t tales que $|t| > 1,96 = t_{\alpha/2}$. Si t está en esta región de rechazo, entonces se puede rechazar la hipótesis nula (la variable X_i es estadísticamente significativa para explicar la variable dependiente). En caso contrario, no se puede rechazar la hipótesis nula ni concluir la significancia individual.
 - Finalmente, podemos observar de la tabla el intervalo de confianza para $\hat{\beta}_i$. Si el cero está contenido en dicho intervalo, significa que no se puede rechazar la posibilidad de que $\hat{\beta}_i = 0$. Sino, el caso $\hat{\beta}_i = 0$ pertenecerá a la región de rechazo del test, lo cual permite concluir la significancia individual de la variable X_i .
- **Significancia global:** en el test F, la región de rechazo se encuentra en la lado derecho del gráfico y la hipótesis nula consiste en la insignificancia conjunta de los coeficientes asociados a todas las variables incluídas en la regresión ($H_0 : \beta_1 = \dots = \beta_K = 0$). Como podemos ver en la tabla, el valor del estadístico es $F(K - 1 = 3, N - K = 61,503) = 186,12$, mientras que para $\alpha = 0,05$, $F_{crítico} \approx 2,6^4$. Como $F > F_{crítico}$, se rechaza la hipótesis nula y se concluye que el modelo es globalmente significativo. Otra forma de concluir lo mismo es notar que $p\text{-valor} = \text{Prob} > F = 0,000 < 0,05 = \alpha$.
- **Número de observaciones:** si bien no existe una regla definida para el número de observaciones, podría considerarse que 61.503 son suficientes como para realizar inferencia en esta ocasión.
- **Bondad de ajuste:** la bondad de ajuste de este modelo, puede observarse mediante el coeficiente $R^2 = 0,2127$. Se puede concluir que gran parte de la varianza de la variable dependiente no puede ser explicada mediante los regresores que se han utilizado, lo cual es común en estudios sociales. Podría darse que variables no observadas ayuden a explicar mejor el comportamiento del gasto en educación o que quizás otras formas funcionales provean un mejor ajuste.

²Este nivel de significancia es arbitrario, pero por alguna razón es socialmente aceptado.

³Rechazar la hipótesis nula incorrectamente del test ($H_0 : \beta_i = 0$)

⁴Esta información se puede verificar en una tabla o en Excel, con la función $\text{INV.F}(0,95;3;61499)$

4. **Gráfico:** bosqueje el modelo mediante un gráfico. Notar que se trata de una función de $\mathbb{R}$ a $\mathbb{R}^2$.

Solución:

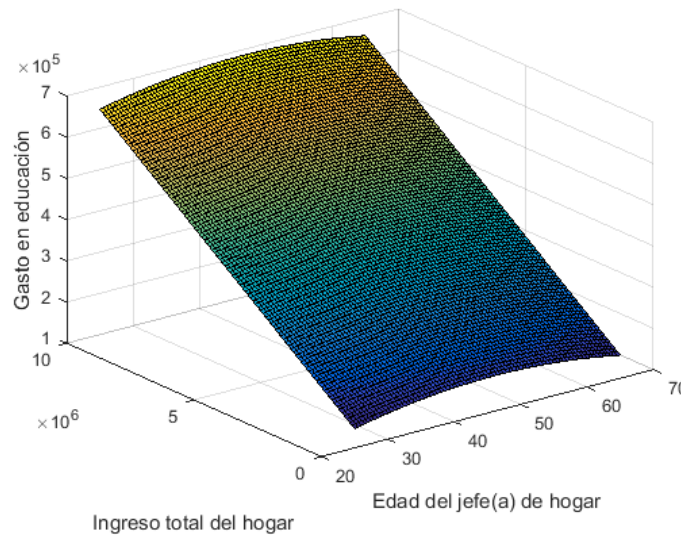


Figura 1: Dado que fracasé con mis habilidades dibujísticas, les dejo un gráfico hecho en MATLAB.

5. **Retornos marginales esperados:** obtenga el retorno marginal esperado sobre el gasto en educación causado por el ingreso del hogar y por la edad del jefe(a) de hogar cuando éste(a) tiene 30 y 40 años respectivamente.

Por conveniencia, estableceremos la siguiente notación: G será el gasto en educación, I el ingreso total del hogar y E la edad del jefe(a) del hogar.

Según el modelo de la segunda regresión: $G = -181 + 8021,156E - 86,4E^2 + 0,0496166I$. Los retornos marginales de I y E pueden ser obtenidos derivando...

$$\frac{\partial G}{\partial I} = 0,0496166$$

Por cada peso adicional que se recibe a modo de ingreso, se gastan 0.0496166 pesos en educación.

$$\frac{\partial G(E)}{\partial E} = 8021,156 - 2 \cdot 86,4 \cdot E$$

$$\frac{\partial G(30)}{\partial E} = 8021,156 - 2 \cdot 86,4 \cdot 30 = 2837,156$$

$$\frac{\partial G(40)}{\partial E} = 8021,156 - 2 \cdot 86,4 \cdot 40 = 1109,156$$

Notemos que cuando el jefe de hogar tiene 40 años de edad, el retorno marginal de dicha edad sobre el gasto en educación es menor. Podemos inferir que a los 40 años se está más cerca del *peak* de gasto en educación de la familia.

6. **Teorema de Frisch-Waugh:** usando este teorema, obtenga $\hat{\beta}_{\text{ingreso}}$.

Solución: ver código al final del documento. En material docente hay un documento explicando este teorema.

7. **Modelo log-lin:** ahora realice una regresión análoga a la segunda, donde la variable dependiente sería esta vez el logaritmo del gasto en educación. Interprete el coeficiente β_{Ingreso} y justifique su interpretación.

Solución: ver código al final del documento.

$$\ln \text{gastoeduc}_i = \beta_0 + \beta_1 \text{ejh}_i + \beta_2 \text{ejh}_i^2 + \beta_3 \text{ingreso}_i + U_i$$

$$\Rightarrow \frac{1}{\text{gastoeduc}_i} \frac{\partial \text{gastoeduc}_i}{\partial \text{ingreso}_i} = \beta_3$$

El término de la derecha se llama “semi-elasticidad” del gasto en educación con respecto al ingreso del hogar. La gracia de la semi-elasticidad, es que al multiplicarla por la variable independiente (en este caso el ingreso), se obtiene la elasticidad.

$$\Rightarrow \underbrace{\frac{\text{ingreso}_i}{\text{gastoeduc}_i} \frac{\partial \text{gastoeduc}_i}{\partial \text{ingreso}_i}}_{\text{Elasticidad}} = \beta_3 \cdot \text{ingreso}_i$$

Recordemos que la elasticidad mide la variación porcentual de una variable con respecto a la variación porcentual de otra variable. Por ejemplo: si un aumento en un 1 % de ingreso implica un 0,05 % de gasto, entonces la elasticidad del gasto en función del ingreso es $\frac{0,05\%}{1\%} = 0,05$

En términos discretos, la elasticidad se y con respecto a x se escribe como $\frac{\Delta y}{y} / \frac{\Delta x}{x}$.

La gracia del modelo log-lin es que nos permite obtener elasticidades variables. En este caso, podemos obtener la elasticidad del gasto en educación para distintos niveles de ingreso.