

# CEF y Proyección Lineal

Ricardo Pasquini

Herramientas Econometricas. Doctorado en Economía UCA 2022

5/10/2022

# CEF y Proyección Lineal

- Motivación: Descripción del ingreso en la población
- Funcion de Esperanza Condicional (CEF) y sus propiedades
- Varianza Condicional y Varianza del Error
- Proyección Lineal
- Proyección Lineal Vs. CEF

## Ejemplo

Analizaremos la teoría junto al caso de los ingresos individuales en CABA.

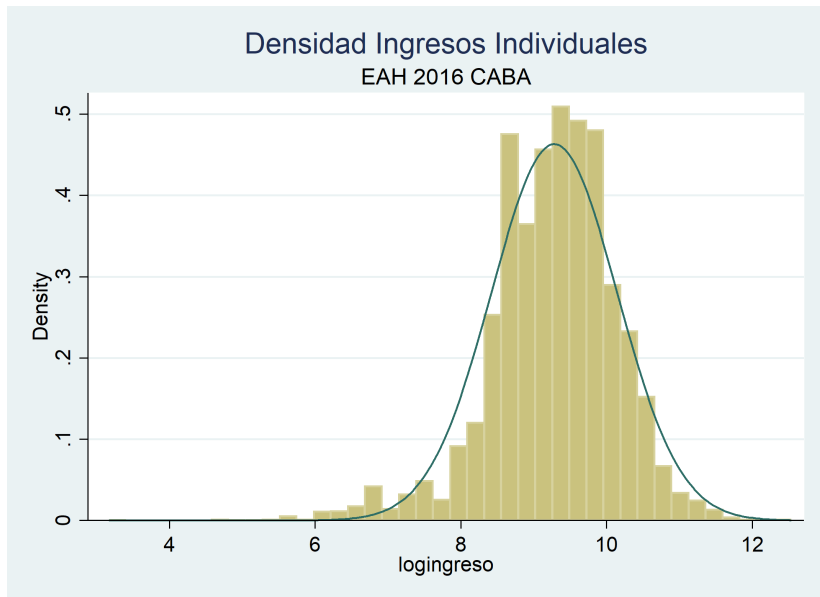
# Distribuciones Poblacionales

- ▶ Supondremos  $Y$  proveniente de una población con una CDF

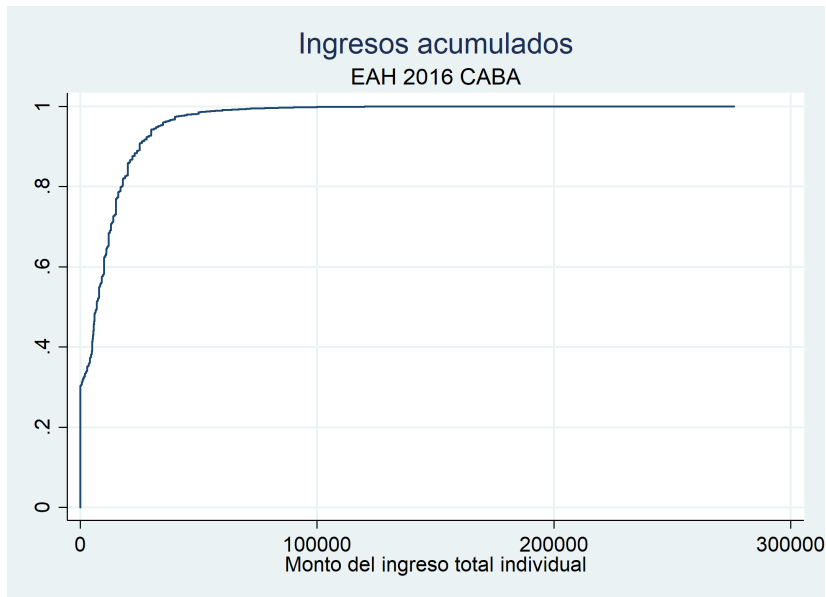
$$F(y) = Prob(Y \leq y)$$

- ▶ Supondremos factores explicativos como  $X_1, X_2, \dots, X_k$  también como variables aleatorias con sus respectivas distribuciones.

# Distribución del ingreso - Densidad

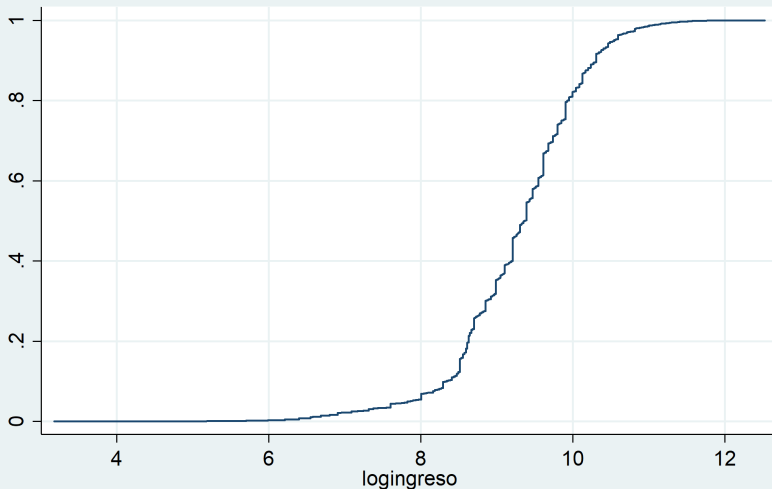


## Distribución del ingreso - Densidad Acumulada



# Distribución del ingreso- Densidad Acumulada

Ingresos acumulados - en logs  
EAH 2016 CABA

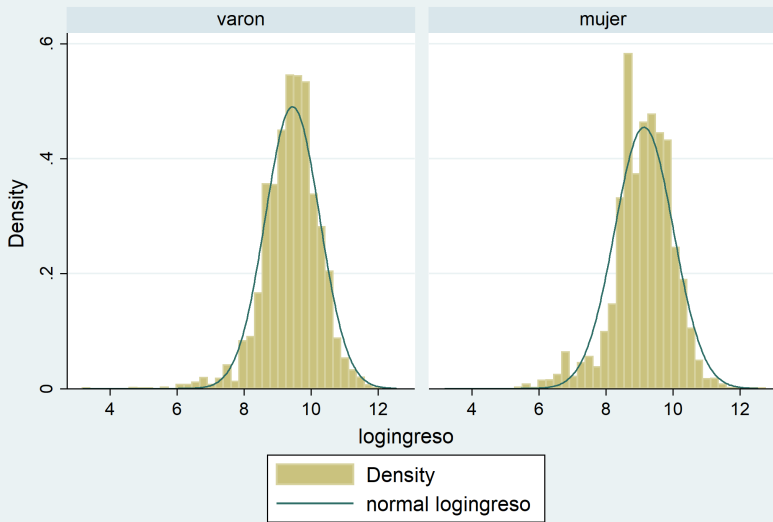


# Distribucion del ingreso por sexo

¿Explica el sexo la distribución del ingreso? ¿Varía la distribución del ingreso de acuerdo al sexo?

- ▶ Lo inspeccionaremos gráficamente
- ▶ Utilizaremos el valor esperado condicional como una aproximación

# Distribucion del ingreso por sexo



Graphs by sexo





# Función de Esperanza Condicional (CEF)

- ▶ En general, es natural que para un valor de  $x$  estemos interesados en conocer el valor esperado. Lo definimos como:

$$E[Y|x] \equiv m(x)$$

- ▶ Para explicar o predecir podríamos definir un modelo:

$$Y = m(x) + e$$

donde

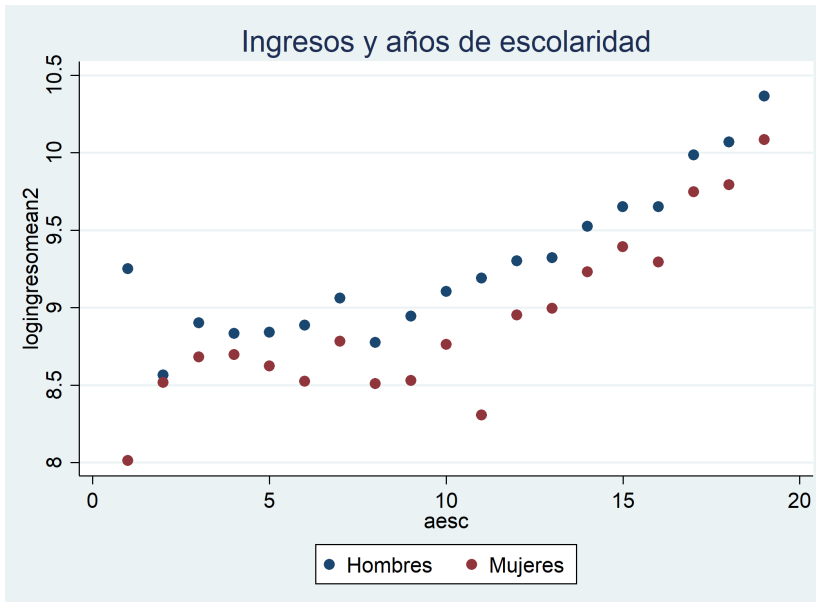
$$e \equiv Y - m(x)$$

- ▶ Algunas propiedades:

1.  $E[e|x] = 0$
2.  $E[e] = 0$

# Funcion de Esperanza Condicional

Aplicacion: Ingresos y Años de Escolaridad



# Problema de prediccion

- ▶ La función CEF tiene una propiedad teórica interesante: provee la mejor predicción en un sentido específico.
- ▶ Supongamos que dado un vector de características  $x$  queremos buscar una función  $g(x)$  que nos haga la mejor predicción posible sobre  $y$ . Una forma de definir mejor predicción, es pedir que minimice el error cuadrático esperado

$$E[(y - g(x))^2]$$

# Problema de prediccion

$g(x) = E[y|x] \equiv m(x)$  como la solución

- Se puede demostrar que la función que minimiza el error cuadrático medio es  $g(x) = E[y|x]$ .

Proof.

$$\begin{aligned} E[(y - g(x))^2] &= E[(e + m(x) - g(x))^2] \\ &= E(e^2) + 2E(e(m(x) - g(x))) + E((m(x) - g(x))^2) \\ &= E(e^2) + E((m(x) - g(x))^2) > E(e^2) = E((y - m(x))^2) \quad \square \end{aligned}$$

# Problema de prediccion

$g(x) = E[y|x] \equiv m(x)$  como la solución

- ▶ Una desventaja es que no siempre será fácil estimar  $E[y|x]$ , por ejemplo por tener pocos datos para nuestro  $x$  de interés.
- ▶ Tampoco conocemos la forma funcional de  $E[y|x]$

# Varianza Condicional

Nuestro objeto de interés es más que el valor esperado

- Definimos varianza condicional en general como:

$$\text{Var}(w|x) = E[(w - E[w|x])^2]$$

- Se sigue que la *varianza condicional del error del modelo CEF* es la esperanza condicional del error al cuadrado:

$$\sigma^2 = \text{Var}(e|x) = E[e^2|x]$$

Y definimos también el *desvío estándar condicional*:

$$\sigma = \sqrt{E[e^2|x]}$$

## CEF Lineal (Proyección Lineal)

- Un *caso particular* de un CEF es cuando el valor esperado cumple que puede expresarse como una función lineal:

$$m(x) = x_1\beta_1 + x_2\beta_2 + \dots + \beta_k$$

- Para la notación es útil resumir esta forma si usamos

$$\mathbf{x} = \begin{Bmatrix} x_1 \\ x_2 \\ \dots \\ x_{k-1} \\ 1 \end{Bmatrix} \quad \boldsymbol{\beta} = \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_{k-1} \\ \beta_k \end{Bmatrix}$$

entonces

$$m(x) = \mathbf{x}'\boldsymbol{\beta}$$



## Modelo CEF lineal o Regresion Lineal

- ▶ Por lo tanto el modelo de Regresion Lineal queda definido por:

$$y = \mathbf{x}'\boldsymbol{\beta}$$

$$E[e|\mathbf{x}] = 0$$

- ▶ Intuitivamente usar este modelo en vez del CEF supone optar por una proyección lineal
- ▶ Nuestro modelo pierde flexibilidad, pero necesitamos menos datos.

# Modelo CEF lineal o Regresion Lineal

- ▶ Un supuesto adicional define el Modelo de Regresion Homoscedastico:

$$E[e^2|\mathbf{x}] = \sigma^2$$

(La varianza es constante independiente de  $\mathbf{x}$ )

- ▶ Notar que este supuesto se utiliza para simplificar el análisis de los modelos, pero no es algo que podemos suponer en general. Al contrario, en general esperamos que exista variabilidad condicional a  $\mathbf{x}$ .

# Encontrando el Mejor Predictor Lineal

En general, también existe un mejor modelo lineal, en el sentido que existe un modelo lineal que minimiza el error cuadrático medio.

Proof.

$$\begin{aligned} \operatorname{argmin}_{\beta \in \mathbb{R}^k} S(\beta) &= E[(y - x'\beta)^2] \\ &= E[y^2] - 2\beta' E[xy] + \beta' E[xx']\beta \\ 0 &= \frac{\partial S(\beta)}{\partial \beta} = -2E[xy] + 2E[xx']\beta \end{aligned}$$



$$\Rightarrow \beta = (E[xx'])^{-1} E[xy]$$

# Completamos el Modelo de Proyeccion Lineal

## Remark

El modelo lineal de menor error será:

$$y = \mathbf{x}'\boldsymbol{\beta}$$

$$E[\mathbf{x}'\mathbf{e}] = 0$$

$$\boldsymbol{\beta} = (E[\mathbf{x}\mathbf{x}'])^{-1}E[\mathbf{x}y]$$

# Variables Categoricalas, Dummys y Modelos no-lineales

¿Cómo introducir una variable categórica al modelo lineal?

¿Puede una variable categórica ser parte de un CEF lineal? Sí.

$$x_1 = \begin{cases} 1 & \text{si sexo=hombre} \\ 0 & \text{si sexo=mujer} \end{cases}$$

$$E[y|x_1] = \beta_1 x_1 + \beta_2$$

# Variables Categoricas, Dummies y Modelos no-lineales

Incorporando más de una característica y la posibilidad de interacciones

$$x_1 = \begin{cases} 1 & \text{si sexo=hombre} \\ 0 & \text{si sexo=mujer} \end{cases}, x_2 = \begin{cases} 1 & \text{si casado=si} \\ 0 & \text{si casado=no} \end{cases}$$

$$E[y|x_1] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

## CEF y Linealidad

- ▶ El CEF es lineal, siempre y cuando las variables explicativas tomen un numero finito de categorías.
- ▶ Cuando tengo una variable con  $I$  categorías puedo reducirlo a un CEF líneal, siempre y cuando traduzca las categorías en  $I - 1$  variables dummies.
- ▶ Algunas variables que toman un numero no muy grande de valores pueden categorizarse. En ese caso una proyección lineal y el CEF serían equivalente.

# Proyeccion Lineal Vs CEF. Ejemplos

## 1. Modelo con Interaccion vs. sin interaccion. Caso Sexo y condicion de inmigrante

El CEF es equivalente a la especificación con interacciones:

```
. regress logingreso mujer inmigrante mujerinmigrante
```

Source	SS	df	MS	Number of obs	=	10,111
Model	369.089695	3	123.029898	F(3, 10107)	=	174.40
Residual	7129.97081	10,107	.705448779	Prob > F	=	0.0000
Total	7499.06051	10,110	.741746835	R-squared	=	0.0492
				Adj R-squared	=	0.0489
				Root MSE	=	.83991

logingreso	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mujer	-.295399	.0194459	-15.19	0.000	-.3335168	-.2572813
inmigrante	-.2435053	.0280165	-8.69	0.000	-.2984232	-.1885875
mujerinmigrante	-.0276682	.0381717	-0.72	0.469	-.1024924	.047156
_cons	9.50513	.0140199	677.97	0.000	9.477648	9.532612

En el caso de la mujer inmigrante (versus otras mujeres) deberia sumarse -0.02 a los -0.24 del efecto inmigrante. Un total de -0.26.



# Proyeccion Lineal Vs CEF. Ejemplos

## 1. Modelo con Interaccion vs. sin interaccion

Al estimar sin interacción obtenemos:

```
. regress logingreso mujer inmigrante
```

Source	SS	df	MS	Number of obs	=	10,111
Model	368.719064	2	184.359532	F(2, 10108)	=	261.35
Residual	7130.34144	10,108	.705415655	Prob > F	=	0.0000
				R-squared	=	0.0492
				Adj R-squared	=	0.0490
Total	7499.06051	10,110	.741746835	Root MSE	=	.83989

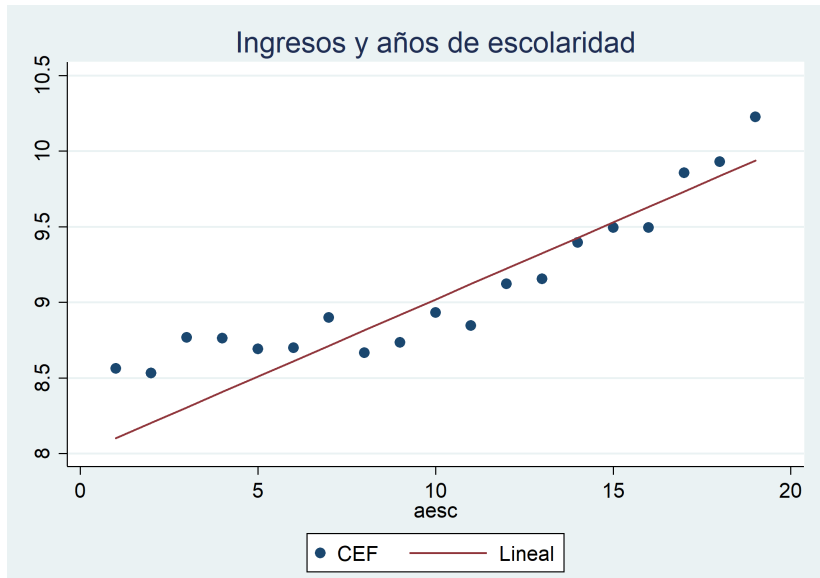
logingreso	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mujer	-.3025795	.016733	-18.08	0.000	-.3353795	-.2697795
inmigrante	-.25841	.0190282	-13.58	0.000	-.295709	-.221111
_cons	9.508862	.0130398	729.22	0.000	9.483302	9.534423

La intuición es que la proyeccion lineal promedia los casos que no surgen en la interaccion. En este caso el efecto inmigrantes (sin diferenciar sexo es -0.25, aun cuando ya incorporamos sexo como explicativa)

# Proyeccion Linear Vs CEF. Ejemplos

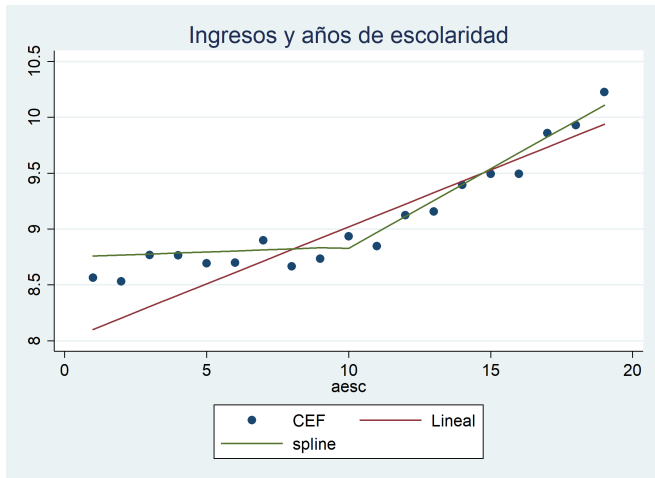
## 2. Lineal cuando el ajuste no es bueno

A veces la relacion lineal es buena solo en un segmento



# Proyeccion Lineal Vs CEF. Ejemplos

## 2. Lineal cuando el ajuste no es bueno



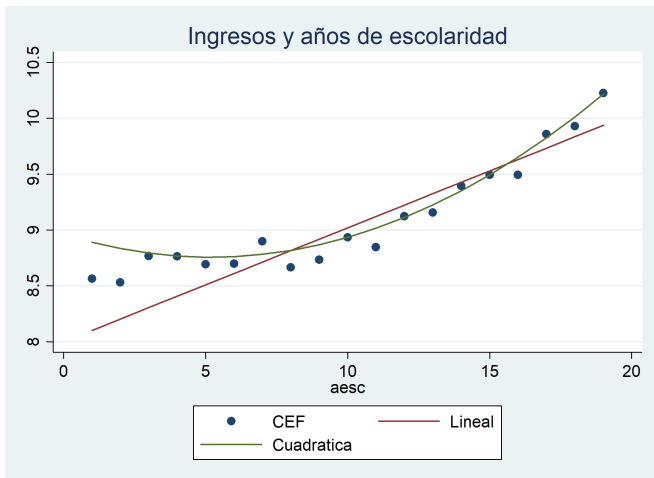
$$P(\log(\text{ingreso})|\text{esc}) = \beta_1 + \beta_2 \text{esc} + \beta_3 (\text{esc} - 9) * 1(\text{esc} \geq 9)$$

# Proyeccion Linear Vs CEF. Ejemplos

## 3. Proyeccion Cuadratica

$$P(\log(\text{ingreso})|\text{experiencia}) = \beta_1 + \beta_2 \text{experiencia}$$

$$P(\log(\text{ingreso})|\text{experiencia}) = \beta_1 + \beta_2 \text{experiencia} + \beta_3 \text{experiencia}^2$$



# Práctica (en Stata)

Usando datos de la EAH CABA:

1. Analizar (gráficamente) la distribución del ingreso y
2. Aproximar el CEF  $E[y|escolaridad]$
3. Aproximamos el CEF  $E[y|escolaridad, sexo]$
4. Graficar  $E[y|escolaridad]$  v.s. su mejor Proyeccion Lineal
5. Estimar y graficar una proyeccion cuadrática para  $E[y|escolaridad]$
6. Estimar y graficar un modelo tipo spline.