CEF y Proyección Lineal

Ricardo Pasquini

Herramientas Econometricas. Doctorado en Economía UCA 2020

8/10/2020

Modelos de Regresion

- 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.

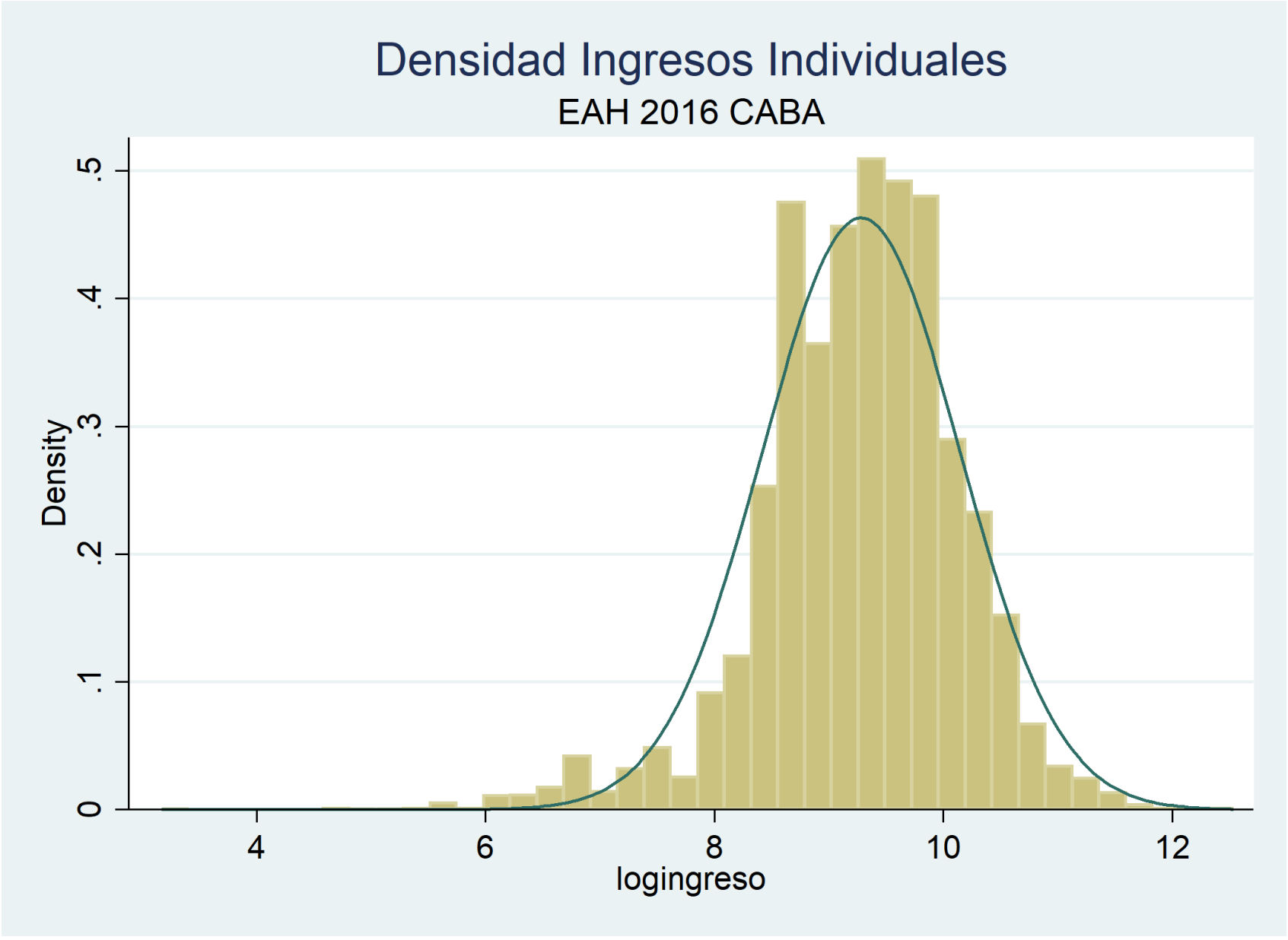
Distribuciones Poblacionales

Supondremos Y proveniente de una población con una CDF

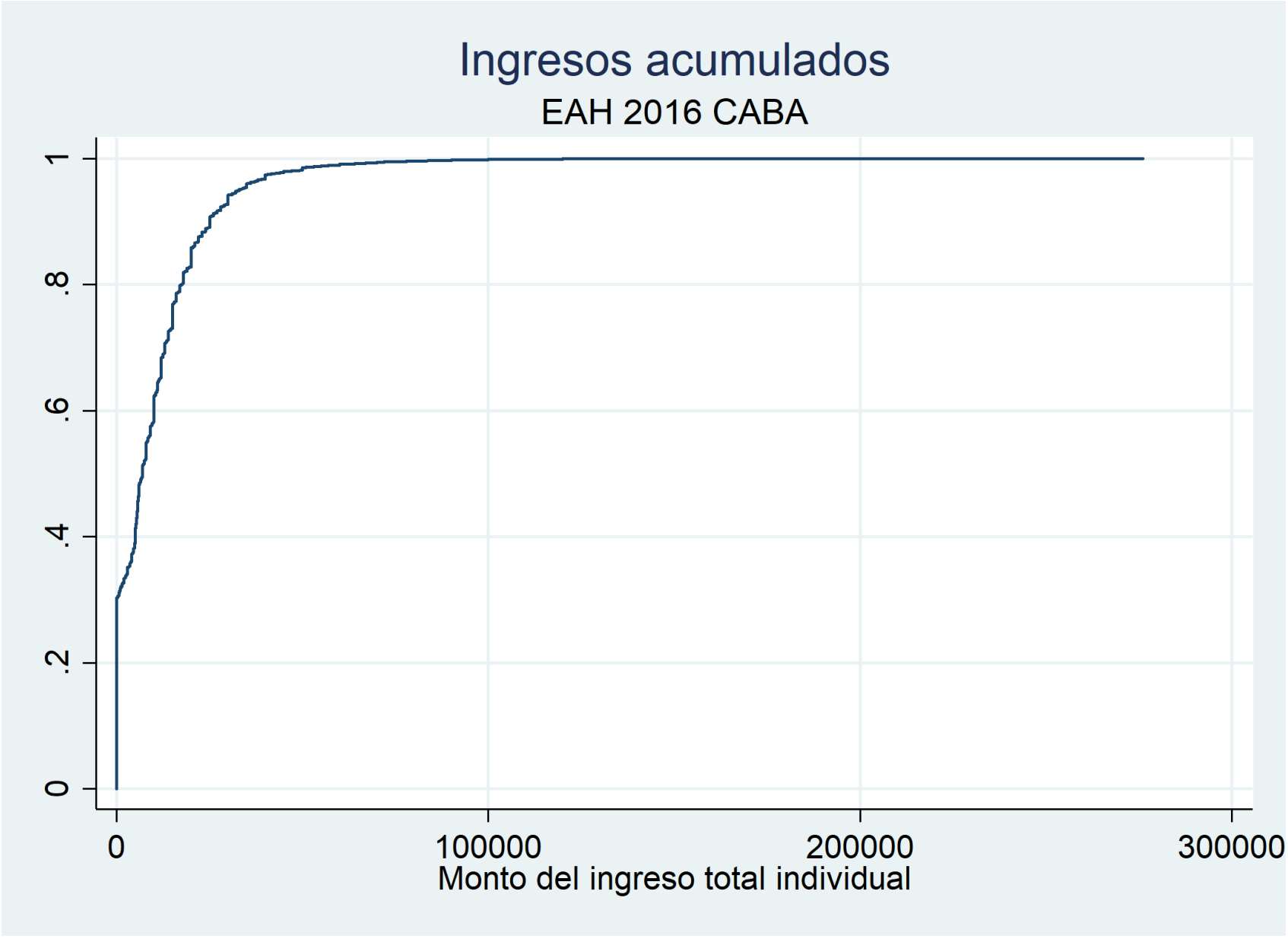
$$F(u) = Prob(Salario \leq u)$$

Supondremos factores explicativos como $X_1, X_2, \ldots, X_k$ tambien como variables aleatorias con sus respectivas distribuciones.

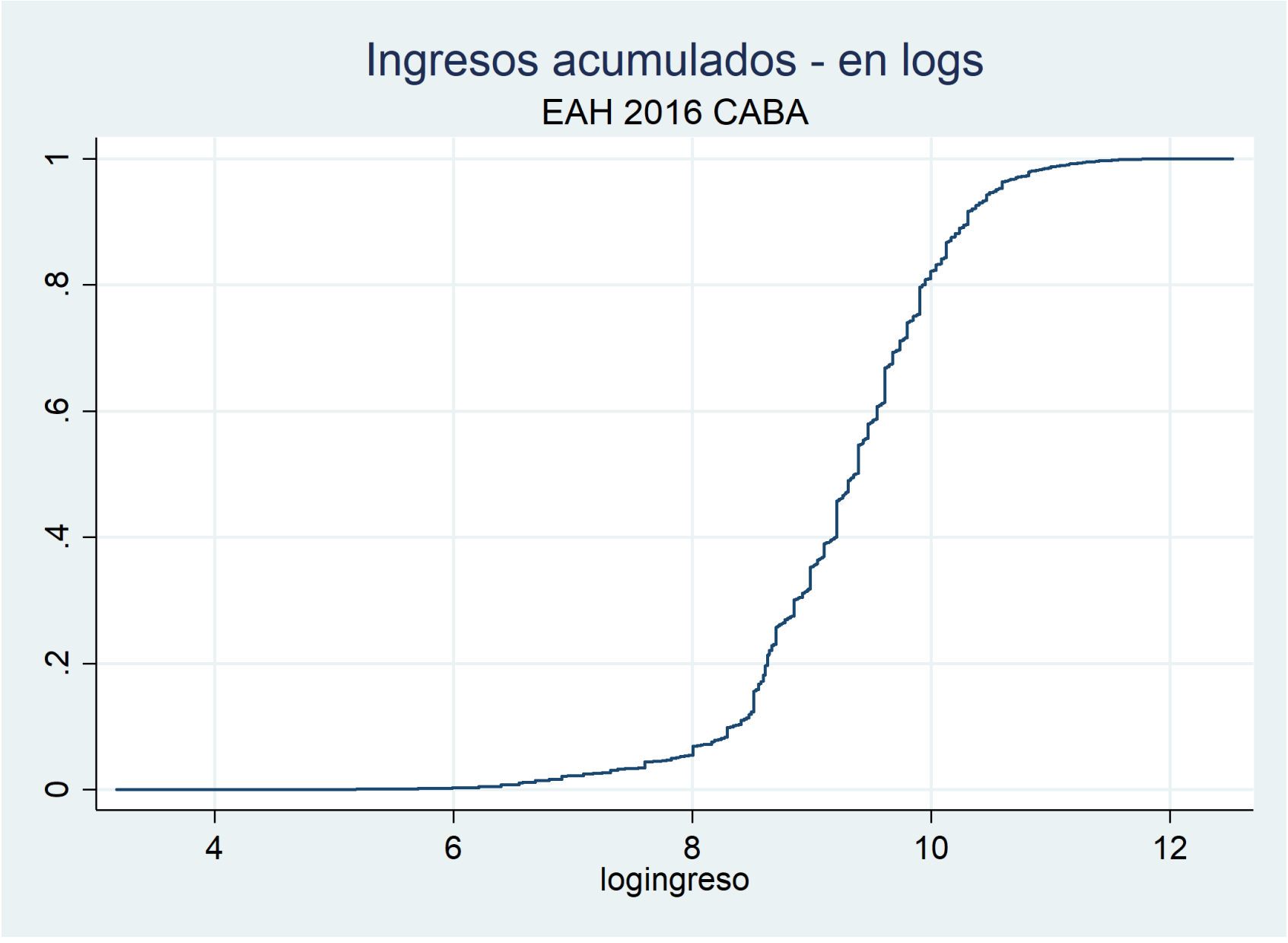
Distribución del ingreso



Distribución del ingreso



Distribucion del ingreso



Distribucion del ingreso

Aproximación: Ingreso esperado

```
. mean logingreso

Mean estimation                     Number of obs    =     10,113
```

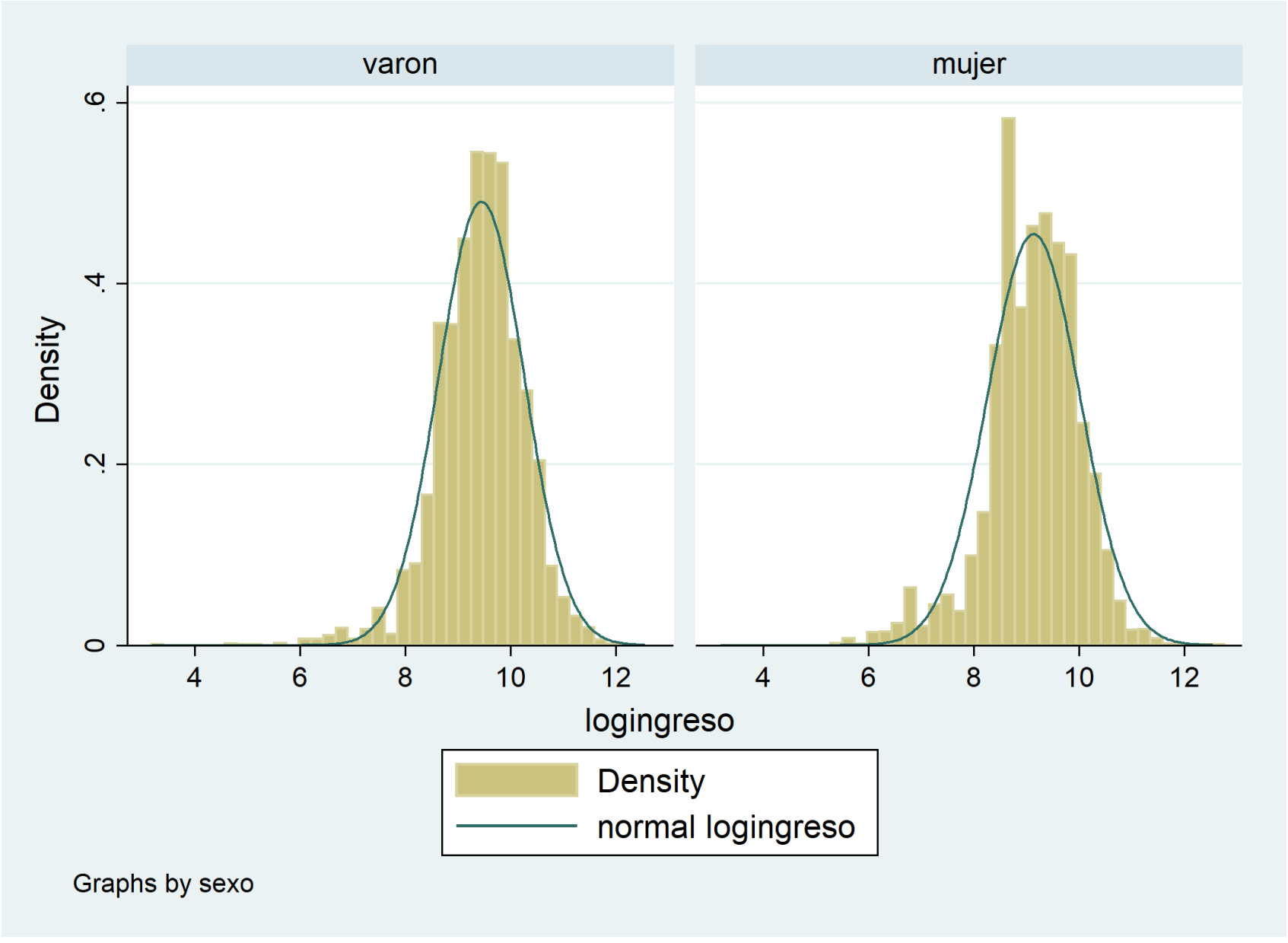
	Mean	Std. Err.	[95% Conf.	Interval]
logingreso	9.28212	.0085636	9.265334	9.298907

Distribucion del ingreso por sexo

¿Varía la distribución del ingreso de acuerdo al sexo?

- Graficamente
- Valor esperado como aproximación

Distribucion del ingreso por sexo



Distribucion del ingreso por sexo

Aproximación: Esperanza Condicional

$$E[Y|sexo = "varon"] = 9.44$$

$$E[Y|sexo = "mujer"] = 9.13$$

```
. mean logingreso, over(sexo)

Mean estimation                   Number of obs   =     10,113

        varon: sexo = varon
        mujer: sexo = mujer
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
logingreso				
varon	9.444136	.0117493	9.421105	9.467167
mujer	9.136385	.0120212	9.112822	9.159949

Función de Esperanza Condicional

Es natural que para un valor de x estemos interesados en

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

Para explicar o predecir podríamos definir un modelo:

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

donde

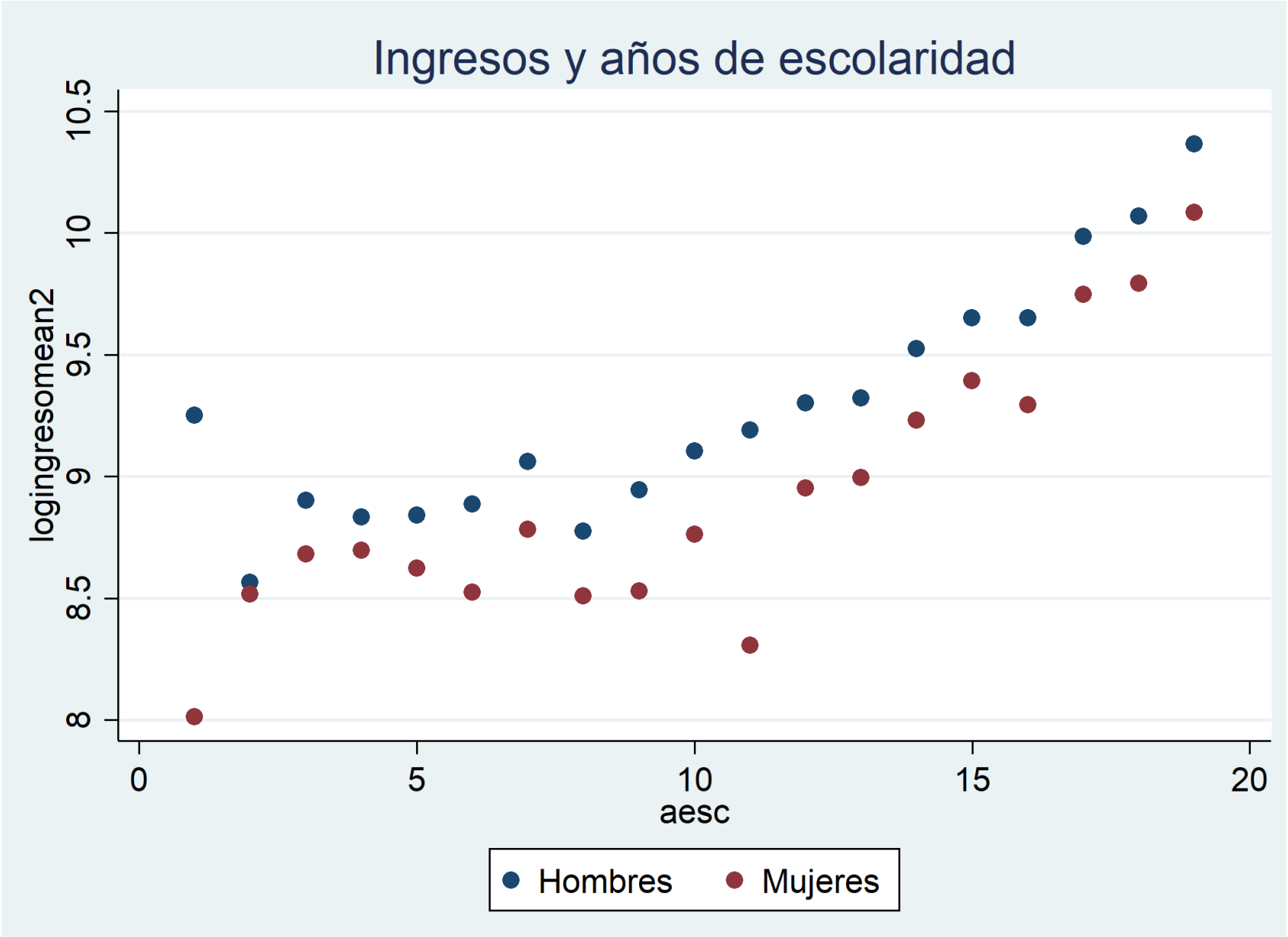
$$e \equiv Y - m(x)$$

Algunas propiedades:

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

Funcion de Esperanza Condicional

Aplicacion: Ingresos y Años de Escolaridad



Problema de prediccion

Supongamos que dado un vector de caracteristicas x queremos buscar una funcion $g(x)$ que nos haga la mejor prediccion posible sobre y. Una forma de definir mejor predicción, es pedir que minimice

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

Problema de prediccion

$m(x)$ como la solucion

Supongamos que dado un vector de caracteristicas x queremos buscar una funcion $g(x)$ que nos haga la mejor prediccion posible sobre y. Una forma de definir mejor predicción, es pedir que minimice

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

- Se puede demostrar que la funcion que minimiza el error cuadratico medio es $g(x) = m(x)$.

Varianza Condicional

Definimos varianza condicional en general como:

$$Var(w|x) = E[(w - E[w|x])^2]$$

Nos permite facilmente ver que la *varianza condicional del error del modelo CEF*:

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

Y el desvio estandar condicional:

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

Varianza Condicional

La varianza incondicional es la varianza condicional promedio:

$$\sigma^2 = E(e^2) = E(E(e^2|x)) = E(\sigma(x))$$

Homoskcedasticidad y Heteroskcedasticidad

Definicion: El error es homocedástico si

$$E(e^2|x) = \sigma^2$$

y es heteroscedastico si

$$E(e^2|x) = \sigma^2(x)$$

CEF Lineal (Proyección Lineal)

Un *caso particular* de un CEF es una función lineal:

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

Para la notación es útil resumir esta forma si usamos

$$\mathbf{x} = \left\{ \begin{array}{c} x_1 \\ x_2 \\ \ldots \\ x_{k-1} \\ 1 \end{array} \right\} \boldsymbol{\beta} = \left\{ \begin{array}{c} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_{k-1} \\ \beta_k \end{array} \right\}$$

entonces

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

Modelo CEF lineal o Regresion Lineal

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

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e$$

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

Un supuesto adicional define el Modelo de Regresion Homoscedastico:

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

(La varianza es constante independiente de x)

Encontrando el Mejor Predictor Lineal

¿Existe un mejor CEF lineal?

- $$argmin_{b \epsilon \Re^k} S(\beta) = E[(y - x'\beta)^2]$$
- $$= E[y^2] - 2\beta' E[xy] + \beta' E[xx']\beta$$
- $$0 = \frac{\partial S(\beta)}{\partial \beta} = -2E[xy] + 2E[xx']\beta$$

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

Completamos el Modelo de Proyeccion Lineal

El modelo lineal de menor error será:

$$y = \boldsymbol{x'}\boldsymbol{\beta} + e$$

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

Propiedades del Modelo de Proyeccion Lineal

$$E[\boldsymbol{xe}] = 0$$

Si x contiene una constante entonces

$$E[\boldsymbol{e}] = 0$$

CEF lineal y variables dummys

El CEF es lineal si los regresores toman valores finitos

Consideremos primero un ejemplo simplificado donde el sexo es el único atributo. El modelo lineal recupera exactamente los dos valores condicionales.

$$E(y|sexo) = \begin{Bmatrix} \mu_0 & \text{si sexo=hombre} \\ \mu_1 & \text{si sexo=mujer} \end{Bmatrix}$$

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

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

Variables Categoricas, Dummys y Modelos no-lineales

Incorporando más de una característica y la posibilidad de interacciones

$$E(y|sexo, casado) = \begin{cases} \mu_{00} & \text{si hombre soltero} \\ \mu_{01} & \text{si hombre casado} \\ \mu_{10} & \text{si mujer soltera} \\ \mu_{11} & \text{si mujer casada} \end{cases}$$

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

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

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		
Model	369.089695	3	123.029898	Prob > F	0.0000
Residual	7129.97081	10,107	.705448779	R-squared	0.0492
				Adj R-squared	0.0489
Total	7499.06051	10,110	.741746835	Root MSE	.83991

Number of obs	=	10,111
F(3, 10107)	=	174.40

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

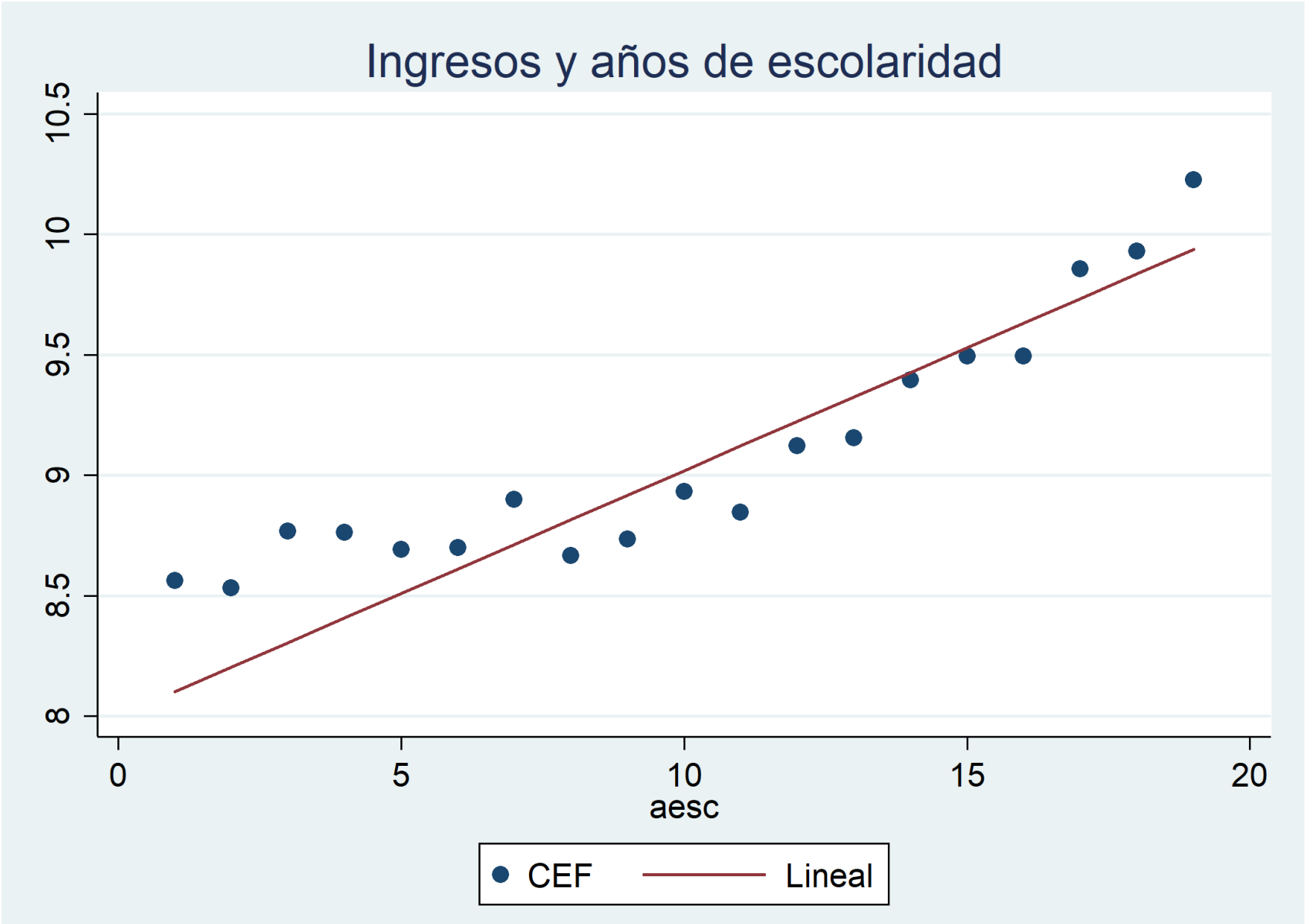
logingreso	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
mujer	-.3025795	.016733	-18.08	0.000	-.3353795	-.2697795
inmigrante	-.25841	.0190282	-13.58	0.000	-.295709	-.221111
_cons	9.508862	.0130398	729.22	0.000	9.483302	9.534423

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 Lineal Vs CEF. Ejemplos

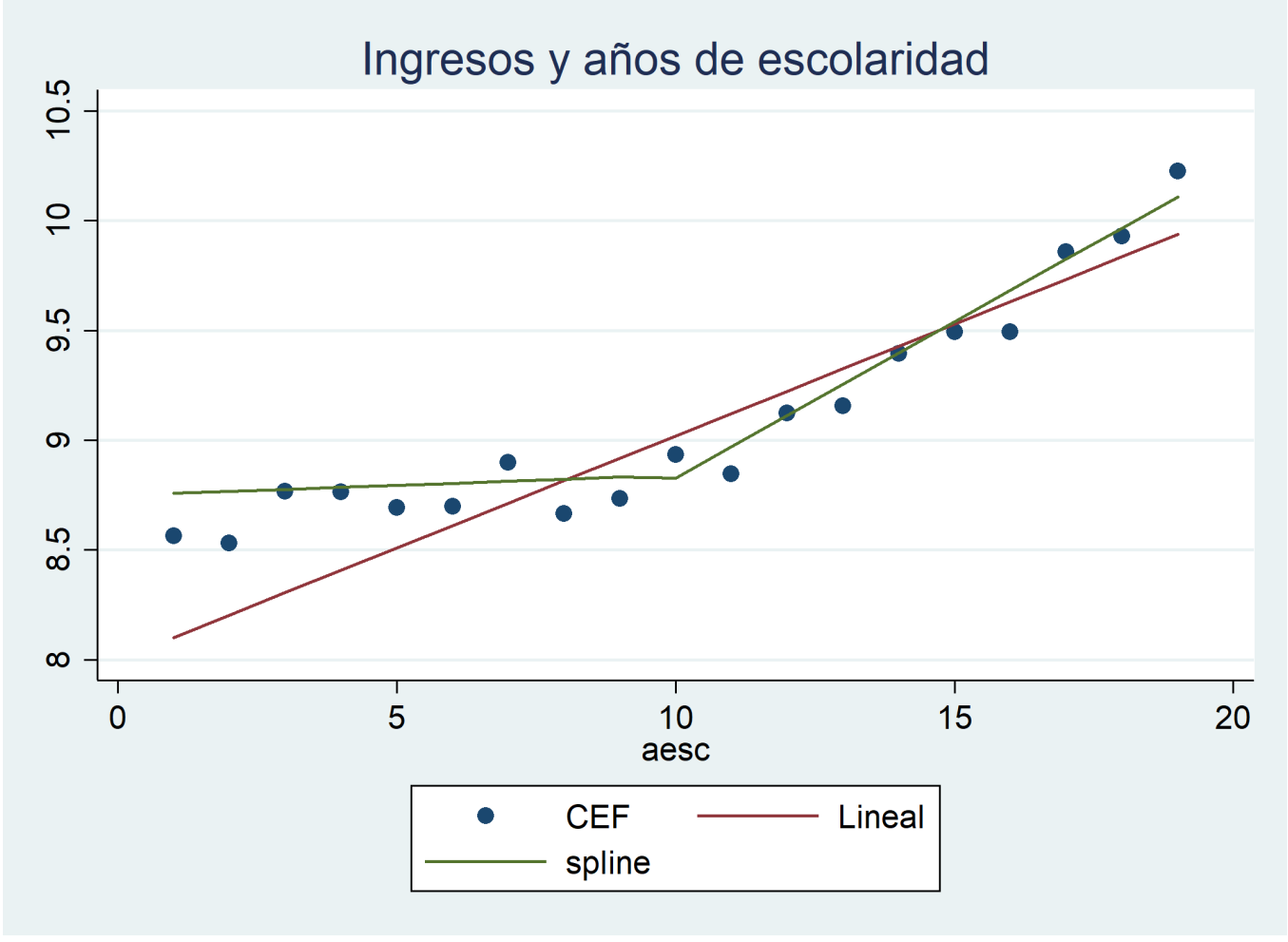
2. Lineal cuando el ajuste no es bueno

A veces la relacion lineal es buena solo en un segmento



Proyeccion Lineal Vs CEF. Ejemplos

2. Lineal cuando el ajuste no es bueno



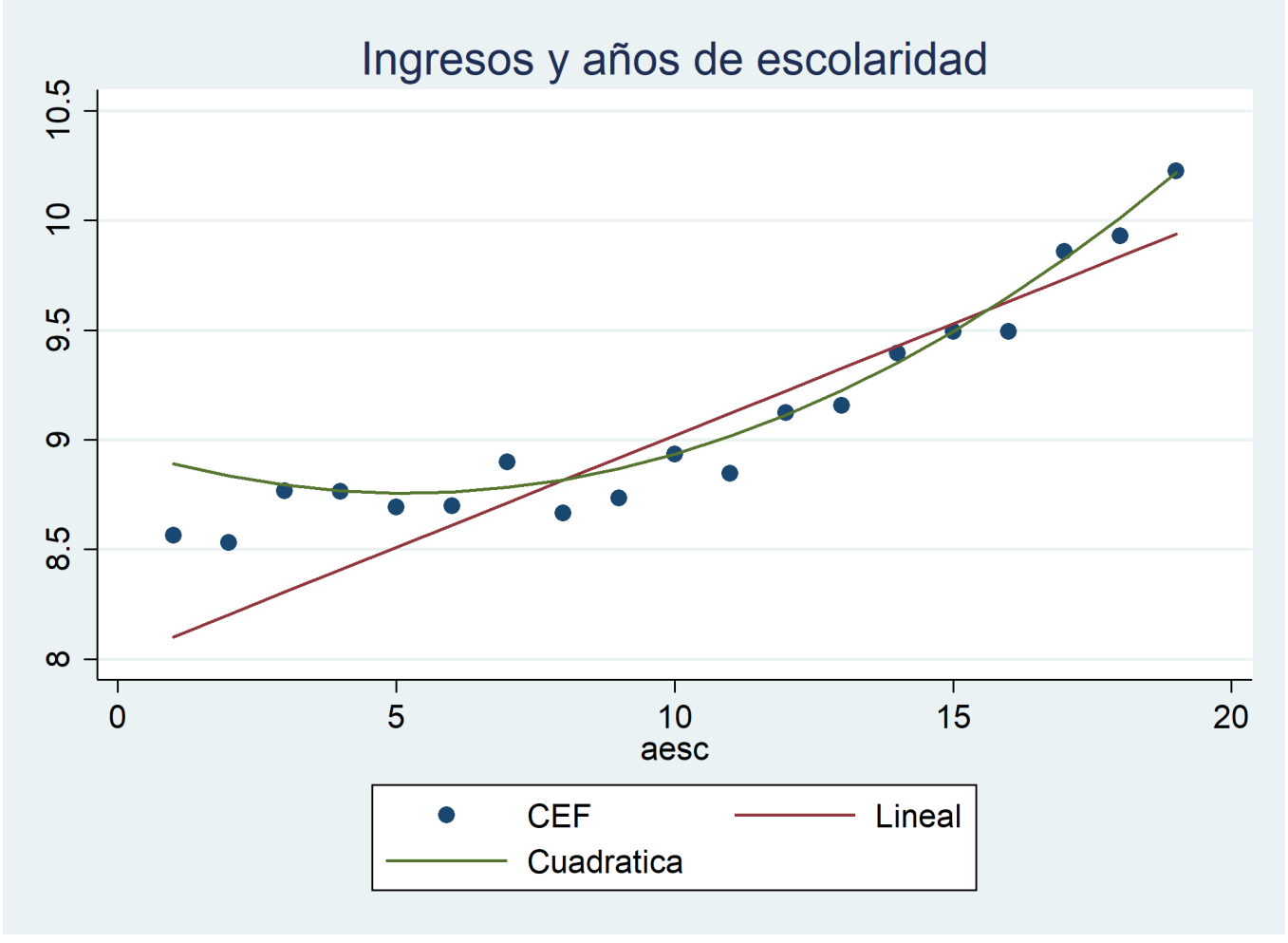
$$P(log(ingreso)|esc) = \beta_1 + \beta_2 esc + \beta_3 (esc - 9) * 1(esc >= 9)$$

Proyeccion Lineal Vs CEF. Ejemplos

3. Proyeccion Cuadratica

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

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



Práctica

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.