
PUCP

DEPARTAMENTO DE ECONOMÍA
LABORATORIO DE ECONOMETRÍA: STATA
1ECO31

Sesión 12
Modelos Multivariados

Docente: Juan Palomino

PUCP

Índice

1 Modelo con Covariante Continuo

2 Modelo con Covariante Polinómico

3 Modelo con Covariante Dicotómico

4 Modelo con Covariante Multicategórico

5 Modelo con Covariantes de Interacción

1. Modelo con Covariante Continuo

Modelo con Covariante Continuo

El primer modelo contiene solo variables continuas como control:

$$\ln(wage)_i = \beta_0 + \beta_1 edad_i + \epsilon_i$$

Donde β_0 es la constante del modelo y β_1 es el coeficiente estimado para la variable edad.

Modelo con Covariante Continuo

Para estimar este modelo y obtener los estimadores con un intervalo de confianza al 95%, se ejecuta:

```
reg lnwage edad, level(95) cformat(%6.3fc)
```

Source	SS	df	MS	Number of obs	=	41,217
				F(1, 41215)	=	99.11
Model	121.243606	1	121.243606	Prob > F	=	0.0000
Residual	50417.0179	41,215	1.22326866	R-squared	=	0.0024
				Adj R-squared	=	0.0024
Total	50538.2615	41,216	1.22618065	Root MSE	=	1.106

Coeficientes

p-value

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	0.004	0.000	9.96	0.000	0.003	0.005
_cons	6.395	0.018	353.26	0.000	6.360	6.431

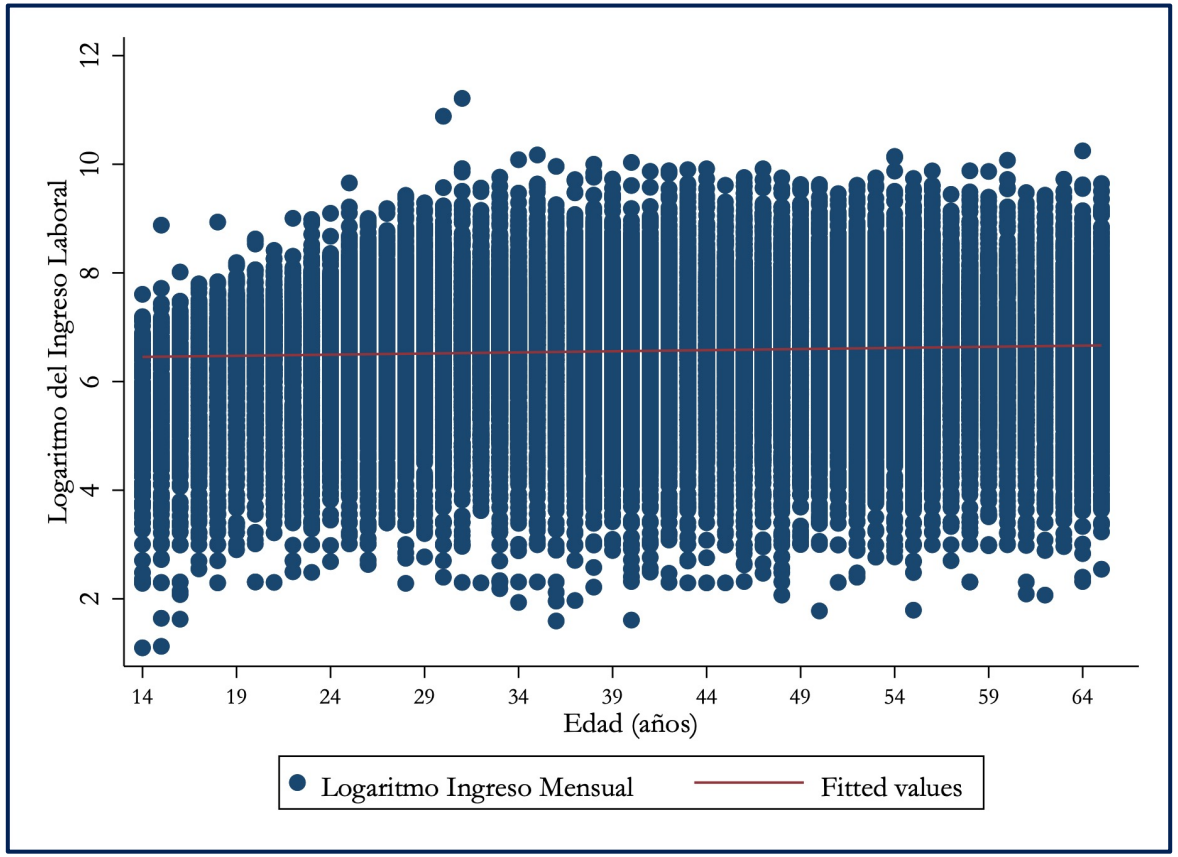
El coeficiente estimado se calcula como una semielasticidad: un año adicional de vida le asegura al individuo un incremento porcentual de 0.004*100=0.4% en los ingresos laborales.

Interpretación de Coeficientes

Especificación	Expresión	Interpretación de β_1
Nivel - Nivel	$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$	Incremento de unidades en "y" cuando aumenta una unidad la ”x" (ambas en sus unidades de medida originales)
Log - Nivel	$log(y_i) = \beta_0 + \beta_1 x_i + \epsilon_i$	Incremento porcentual de "y" cuando aumenta una unidad la ”x"
Nivel - Log	$y_i = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$	Incremento en unidades de "y" cuando aumenta un 1% la ”x"
Log - Log	$\log(y_i) = \beta_0 + \beta_1 \log(x_i) + \epsilon_i$	Incremento porcentual de "y" cuando aumenta un 1% la ”x"

Modelo con Covariante Continuo

Podemos ver los valores predichos mediante una línea de ajuste lineal



2. Modelo con Covariante Polinómico

Modelo con covariante continuo y polinómico

El segundo modelo considera variables polinómicas:

$$\ln(wage)_i = \beta_0 + \beta_1 edad_i + \beta_2 (edad_i)^2 + \epsilon_i$$

Donde β_0 es la constante del modelo, β_1 es el coeficiente estimado para la variable edad y β_2 es el coeficiente estimado para la variable edad al cuadrado

Modelo con covariante continuo y polinómico

Objetivo: estimar el impacto de la edad sobre los ingresos e identificar a qué edad empiezan a decaer los ingresos.

Primera forma

```
reg lnwage edad edad_sq, level(95) cformat(%6.3fc)
```

Source	SS	df	MS		
Model	900.096631	2	450.048316	Prob > F	0.0000
Residual	49638.1649	41,214	1.20440056	R-squared	0.0178
				Adj R-squared	0.0178
Total	50538.2615	41,216	1.22618065	Root MSE	1.0975

Number of obs	=	41,217				
F(2, 41214)	=	373.67				

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	0.069	0.003	26.71	0.000	0.064	0.074
edad_sq	−0.001	0.000	−25.43	0.000	−0.001	−0.001
_cons	5.210	0.050	104.28	0.000	5.112	5.308

Segunda forma

```
reg lnwage c.edad##c.edad, level(95) cformat(%6.3fc)
```

Source	SS	df	MS		
Model	900.096631	2	450.048316	Prob > F	0.0000
Residual	49638.1649	41,214	1.20440056	R-squared	0.0178
				Adj R-squared	0.0178
Total	50538.2615	41,216	1.22618065	Root MSE	1.0975

Number of obs	=	41,217				
F(2, 41214)	=	373.67				

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	0.069	0.003	26.71	0.000	0.064	0.074
c.edad#c.edad	−0.001	0.000	−25.43	0.000	−0.001	−0.001
_cons	5.210	0.050	104.28	0.000	5.112	5.308

Modelo con covariante continuo y polinómico

Nótese que esta vez el efecto de edad no viene dado sólo por el parámetro $\hat{\beta}_1$ sino por $\hat{\beta}_2$.

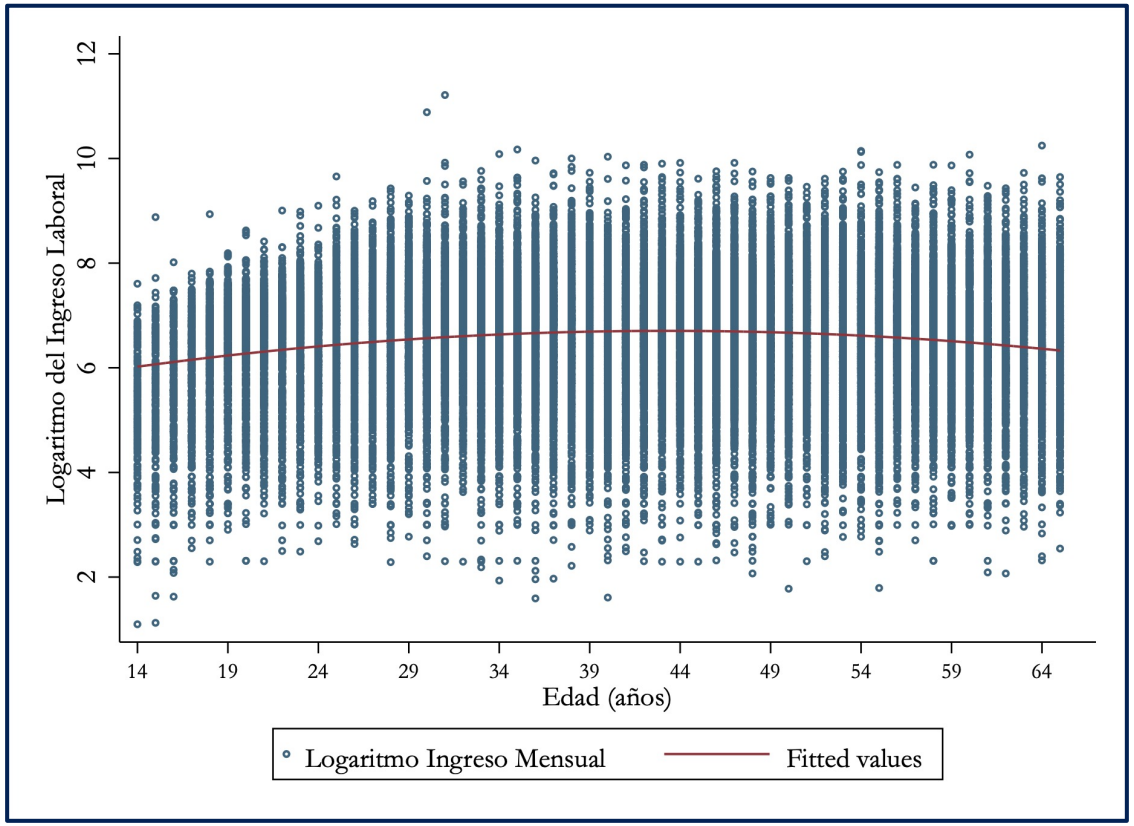
$$\frac{d\ln(wage)_i}{d(edad_i)} = \beta_1 + 2\beta_2 edad_i = 0$$

La relación entre edad e ingresos viene representada por una curva cóncava, tal que al inicio es positiva y, a partir de un punto $edad^* = \left|\frac{\beta_1}{2\beta_2}\right|$ su pendiente cambia a negativa:

```
. display "Punto Máximo= " abs(_b[edad]/(2*_b[c.edad#c.edad]))
Punto Máximo= 43.299973
```

Modelo con Covariante Polinómico

Para generar las predicciones del modelo empleamos ”predict”:



Modelo con covariante continuo y polinómico

No basta con ver las pruebas de significancia individuales para afirmar que es razonable ingresar a edad como variable cuadrática.

Test Wald

```
. testparm c.edad##c.edad

( 1)  edad = 0
( 2)  c.edad#c.edad = 0

       F(  2, 41214) =  373.67
            Prob > F =    0.0000
```

Ambas variables tienen importancia simultánea.

Operadores

Existen operadores para las especificaciones de los modelos:

N	Operador	Significado	Ejemplo	Efecto: crea virtualmente
1	c.	continua	c.edad	$edad$
2	#	interacción	c.edad#c.edad	$edad^2$
3	i.	indicador	i.mujer	1: mujer y 0: hombre (omite una)
			i.educ	6 niveles educativos (omite una)
4	##	interacción factorial	i.educ##c.edad	las variables creadas con (1) y (3)

3. Modelo con Covariante Dicotómica

Modelo con Covariante Dicotómica

Objetivo: ver si existe una brecha salarial entre sexos.

$$\ln(wage)_i = \beta_0 + \beta_1 edad_i + \beta_2 mujer + \epsilon_i$$

Donde β_0 es la constante del modelo, β_1 es el coeficiente estimado para la variable edad y β_2 es el coeficiente estimado para la variable mujer, que representa una variable dummy, donde 1 si es mujer y 0 si es hombre.

Modelo con Covariante Dicotómica

El incluir una variable dicotómica parte la línea de regresión:

- Las mujeres tienen una línea de regresión dada por:

$$\ln(wage)_{mujeres} = \beta_0 + \beta_1 edad_i + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 edad_i$$

- Los hombres tienen una línea de regresión dada por:

$$\ln(wage)_{hombres} = \beta_0 + \beta_1 edad_i + \beta_2(0) = (\beta_0) + \beta_1 edad_i$$

Así, la diferencia entre ambas radica en el intercepto con el eje de abscisas.

Modelo con Covariante Dicotómica

Es buena idea estimar el modelo incluyéndola como factoriales:

```
reg lnwage c.edad i.mujer, level(95) cformat(%6.3fc)
```

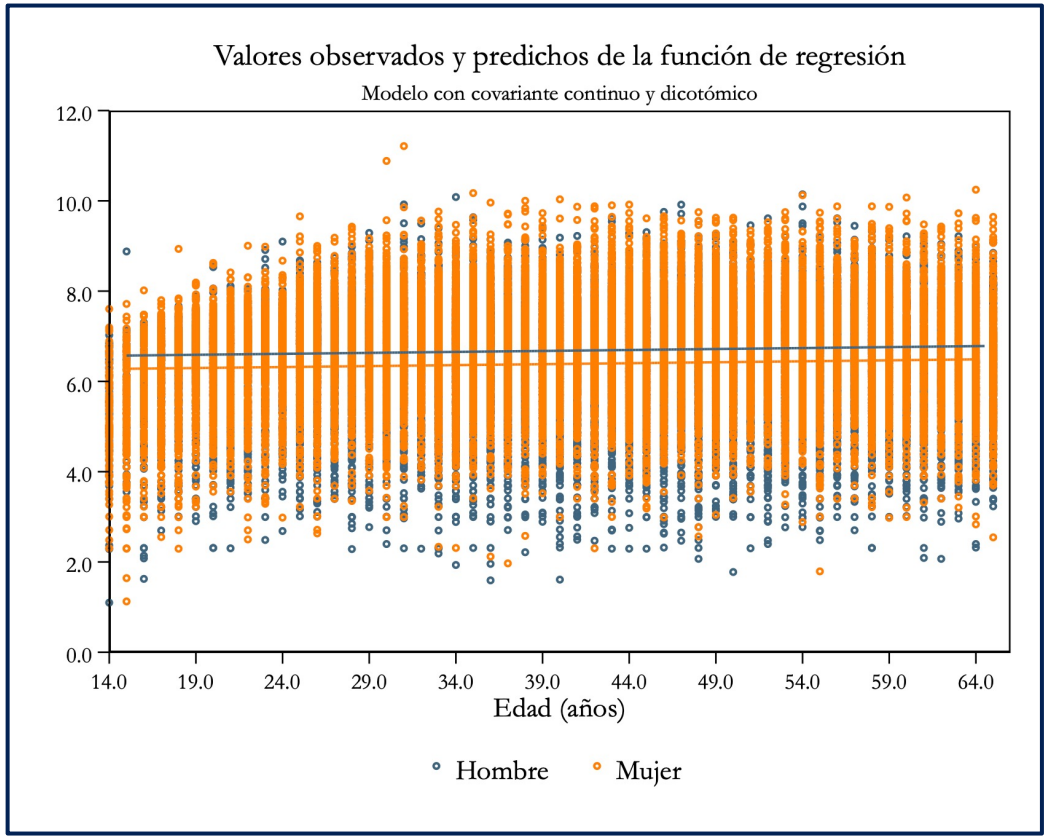
Source	SS	df	MS	Number of obs	=	41,217
				F(2, 41214)	=	403.88
Model	971.457927	2	485.728964	Prob > F	=	0.0000
Residual	49566.8036	41,214	1.20266908	R-squared	=	0.0192
				Adj R-squared	=	0.0192
Total	50538.2615	41,216	1.22618065	Root MSE	=	1.0967

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	0.004	0.000	10.28	0.000	0.003	0.005
mujer						
Mujer	-0.293	0.011	-26.59	0.000	-0.315	-0.271
_cons	6.509	0.018	352.74	0.000	6.473	6.545

Nota: Stata asume que las variables sin operador son continuas.

Modelo con Covariante Dicotómica

Para generar las predicciones del modelo de cada línea de regresión empleamos ”predict” y luego usamos “separate” para separarlas entre ambos sexos.



Modelo con Covariante Dicotómica

Siempre el operador i. genera variables dummy omitiendo una categoría.

El operador i tiene una serie de sufijos para indicarle el número de la categoría:

- **ib#**: si queremos usar la categoría # como base.
- **ib(first)**: si queremos usar la primera categoría como base.
- **ib(last)**: si queremos usar la última categoría como base.
- **ib(freq)**: si queremos usar la categoría más frecuente como base.
- **ibn**: si no queremos usar una categoría base.

4. Modelo con Covariante Multicategórica

Modelo con Covariante Multicategórica

Asumamos que el modelo poblacional difiere según su nivel educativo.

```
. label list educ
educ:
           1 Sin Nivel/Inicial
           2 Primaria
           3 Secundaria
           4 Superior no universitaria
           5 Superior universitaria
           6 Maestria/Doctorado
```

Dado que tenemos 6 niveles educativos, para evitar incurrir en la trampa de dummies debemos incluir sólo 5 (6-1) dummies.

Modelo con Covariante Multicategórica

El modelo a estimar es:

$$\ln(wage)_i = \beta_0 + \beta_1 edad_i + \delta_1 educ_{prim} + \delta_2 educ_{secund} + \delta_3 educ_{nouniv} + \delta_4 educ_{univ} + \delta_5 educ_{posgr} + \epsilon_i$$

Donde las líneas de regresión dependen del nivel educativo del individuo:

$$\ln(wage)_{prim} = \beta_0 + \beta_1 edad_i + \delta_1(1) = (\beta_0 + \delta_1) + edad_i$$

$$\ln(wage)_{secund} = \beta_0 + \beta_1 edad_i + \delta_2(1) = (\beta_0 + \delta_2) + edad_i$$

$$\ln(wage)_{nouniv} = \beta_0 + \beta_1 edad_i + \delta_3(1) = (\beta_0 + \delta_3) + edad_i$$

$$\ln(wage)_{univ} = \beta_0 + \beta_1 edad_i + \delta_4(1) = (\beta_0 + \delta_4) + edad_i$$

$$\ln(wage)_{posgr} = \beta_0 + \beta_1 edad_i + \delta_5(1) = (\beta_0 + \delta_5) + edad_i$$

Modelo con Covariante Multicategórica

Estimación del modelo en Stata:

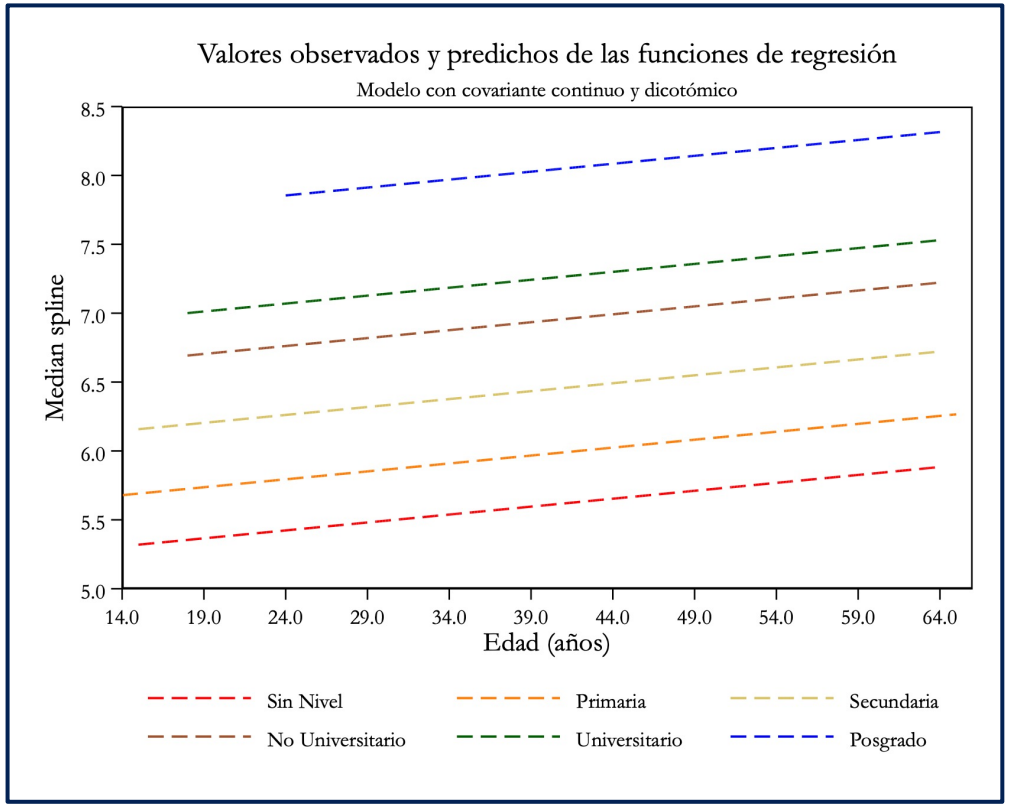
```
reg lnwage c.edad ib1.educ, level(95) cformat(%6.3fc)
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	0.012	0.000	29.21	0.000	0.011	0.012
educ						
Primaria	0.371	0.034	10.86	0.000	0.304	0.438
Secundaria	0.838	0.034	24.73	0.000	0.772	0.905
Superior no universitaria	1.338	0.035	37.94	0.000	1.269	1.408
Superior universitaria	1.647	0.035	46.56	0.000	1.577	1.716
Maestria/Doctorado	2.431	0.046	53.31	0.000	2.342	2.521
_cons	5.147	0.039	133.34	0.000	5.072	5.223

Especificar que queremos la categoría número 1 (sin nivel) como base: **ib1.educ**

Modelo con Covariante Multicategórica

Para generar las predicciones del modelo de cada línea de regresión empleamos ”predict” y luego usamos “separate” para separarlas entre los 6 niveles educativos.



Modelo con Covariante Multicategórica

¿Cómo podemos saber si es que la línea para los individuos con secundaria completa es diferente a los que tienen secundaria incompleta?

Se usa el comando “testparm” en su forma de restricción lineal (dado que estamos especificando coeficientes y no variables):

```
. testparm ib1.educ

 ( 1)  2.educ = 0
 ( 2)  3.educ = 0
 ( 3)  4.educ = 0
 ( 4)  5.educ = 0
 ( 5)  6.educ = 0

       F(  5, 41199) = 2060.58
            Prob > F =    0.0000
```

5. Modelo con Covariantes de Interacción

Modelo con Covariante de Interacción

Se puede realizar interacciones entre distintos covariantes:

$$\ln(wage)_i = \beta_0 + \beta_1 edad_i + \delta_1 educ_{prim} + \delta_2 educ_{secund} + \delta_3 educ_{nouniv} + \delta_4 educ_{univ} + \delta_5 educ_{posgr} + \phi_1 edad_i educ_{prim} + \phi_2 edad_i educ_{secund} + \phi_3 edad_i educ_{nouniv} + \phi_4 edad_i educ_{univ} + \phi_5 edad_i educ_{posgr} + \epsilon_i$$

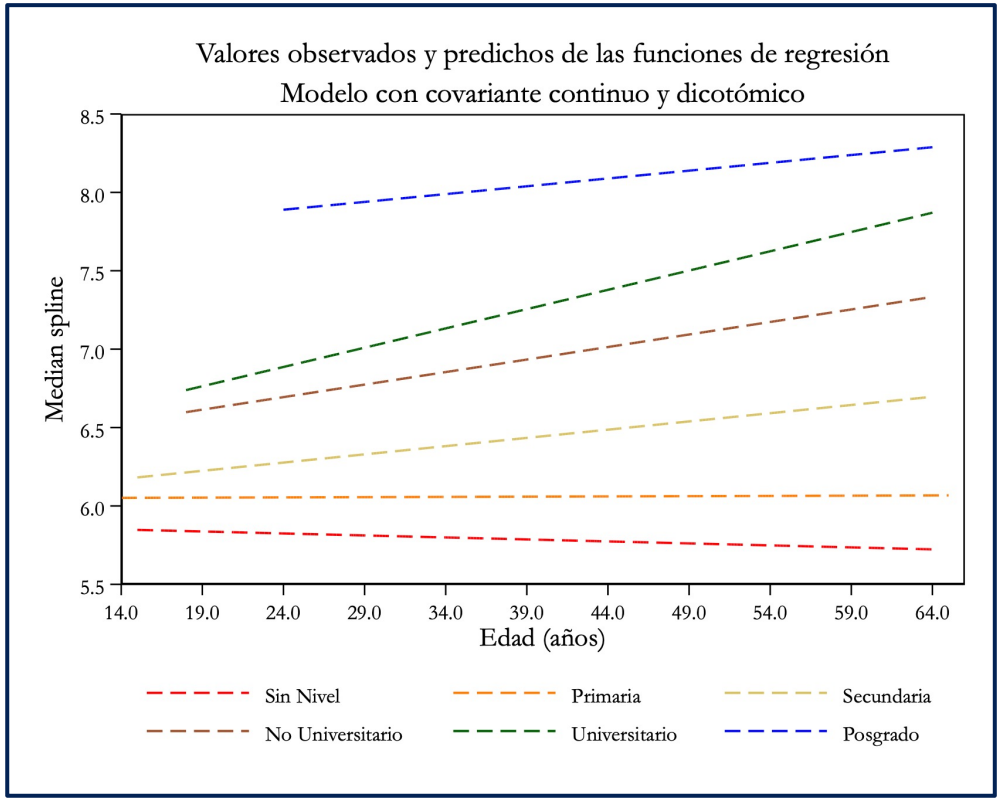
Modelo con Covariante de Interacción

```
reg lnwage c.edad##ib1.educ, noheader cformat(%6.3fc)
```

lnwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edad	-0.003	0.004	-0.71	0.479	-0.010	0.004
educ						
Primaria	0.162	0.196	0.83	0.409	-0.222	0.545
Secundaria	0.139	0.192	0.72	0.469	-0.238	0.516
Superior no universitaria	0.425	0.195	2.18	0.030	0.042	0.808
Superior universitaria	0.411	0.195	2.11	0.035	0.029	0.793
Maestria/Doctorado	1.765	0.244	7.23	0.000	1.286	2.243
educ#c.edad						
Primaria	0.003	0.004	0.77	0.440	-0.004	0.010
Secundaria	0.013	0.004	3.59	0.000	0.006	0.020
Superior no universitaria	0.019	0.004	4.97	0.000	0.011	0.026
Superior universitaria	0.027	0.004	7.32	0.000	0.020	0.034
Maestria/Doctorado	0.013	0.005	2.61	0.009	0.003	0.022
_cons	5.885	0.191	30.82	0.000	5.511	6.260

Modelo con Covariante de Interacción

Para generar las predicciones del modelo de cada línea de regresión empleamos ”predict” y luego usamos “separate” para separarlas entre los 6 niveles educativos.



Modelo con Covariante de Interacción

Podemos ver la significancia de los coeficientes de las pendientes:

```
. testparm ib1.educ#c.edad

 ( 1)  2.educ#c.edad = 0
 ( 2)  3.educ#c.edad = 0
 ( 3)  4.educ#c.edad = 0
 ( 4)  5.educ#c.edad = 0
 ( 5)  6.educ#c.edad = 0

       F(  5, 41194) =   76.09
            Prob > F =    0.0000
```

Los resultados nos permiten afirmar que, al 1%, las pendientes de las líneas de regresión son estadísticamente diferentes entre sí.

Modelo con Covariante de Interacción

También podemos ver la significancia de los coeficientes de los interceptos:

```
. testparm ib1.educ

 ( 1)  2.educ = 0
 ( 2)  3.educ = 0
 ( 3)  4.educ = 0
 ( 4)  5.educ = 0
 ( 5)  6.educ = 0

       F(  5, 41194) =   32.31
            Prob > F =    0.0000
```

Los resultados nos permiten afirmar que, al 1%, los interceptos de las líneas de regresión son estadísticamente diferentes entre sí.


PUCP