
Universidad de San Andrés – Departamento de Economía – Econometría – Semestre de otoño

USOS Y EXTENSIONES DEL MODELO LINEAL CON K VARIABLES

Mariana Marchionni

mariana@depeco.econo.unlp.edu.ar

Variables explicativas binarias

- ¿Cómo incorporamos **factores explicativos cualitativos**? Género, raza, sector productivo, etc.
- Fenómenos binarios: se tiene o no cierta característica. Ej: un individuo tiene seguro médico o no lo tiene, tiene o no empleo, es hombre o mujer, etc. Hay sólo dos posibilidades y son excluyentes.
- Cualquier fenómeno binario puede representarse por una **variable binaria o dummy**: la variable =1 si para la observación en cuestión la característica está presente e igual a cero en caso contrario.
- Ejemplo: $mujer_i=1$ si el individuo i es mujer, $mujer_i=0$ si el individuo i es hombre. Notar que podíamos capturar la misma información definiendo $hombre_i=1$ si el individuo i es hombre e igual a cero si es mujer.

Ejemplo: Ecuación de salarios

$$W_i = \beta_1 + \beta_2 aedu_i + \delta \text{ hombre}_i + u_i$$

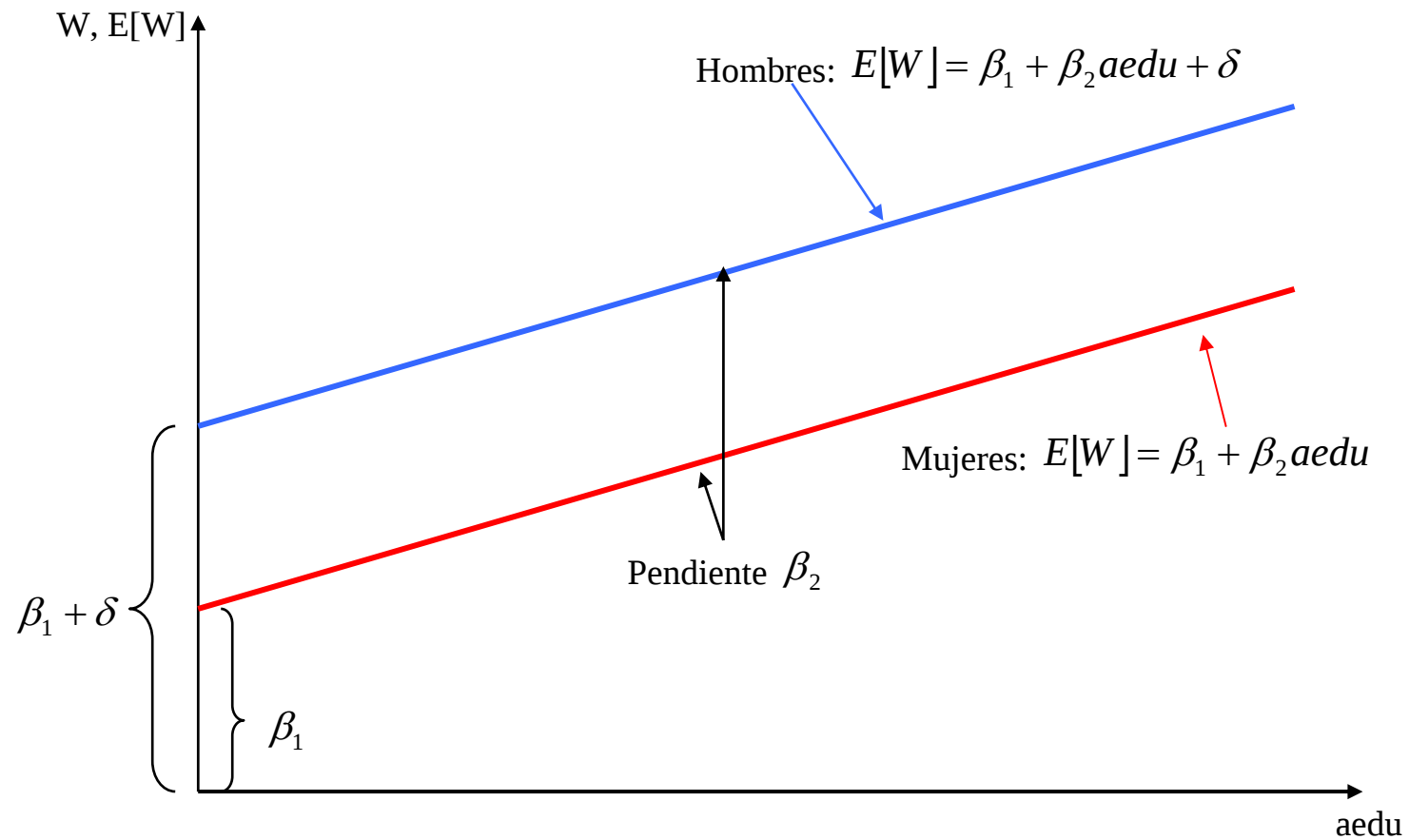
$$\text{Si } E[u_i]=0 \Rightarrow E[W_i] = \beta_1 + \beta_2 aedu_i + \delta \text{ hombre}_i$$

- Para los hombres $E[W_i | \text{hombre}_i = 1] = \beta_1 + \beta_2 aedu_i + \delta$
- Para las mujeres: $E[W_i | \text{hombre}_i = 0] = \beta_1 + \beta_2 aedu_i$
- ¿Qué interpretación tiene δ ? Es la diferencia esperada en el salario de un hombre y una mujer con la misma educación:

$$E[W_i | \text{hombre}_i = 1] - E[W_i | \text{hombre}_i = 0] = \beta_1 + \beta_2 aedu_i + \delta \times 1 - (\beta_1 + \beta_2 aedu_i + \delta \times 0) = \delta$$

- Si $\delta > 0$ habría discriminación en contra de las mujeres. La recta de regresión de los hombres es paralela a la de las mujeres (la misma pendiente β_2) pero la ordenada al origen es mayor ($\beta_1 + \delta > \beta_1$)

Gráficamente:



La trampa de la variable binaria:

- Si bien son dos las categorías de la variable binaria (hombre o mujer, asegurado o no, etc.) sólo incluimos una variable en el modelo.
- **Intuitivamente:** necesitamos un único parámetro (δ) para diferenciar entre los dos grupos. Podemos incluir la variable *hombre* o, alternativamente, la variable *mujer*. La inclusión de ambas sería redundante.
- **Categoría base u omitida** es la que corresponde al valor cero de la variable dummy. Por ejemplo, si *hombre*=1 para los hombres y cero para las mujeres, la categoría base u omitida son las mujeres.
- **Regla:** si hay S características incluimos (S-1) variables binarias.
- Si incluimos S variables binarias caemos en lo que se conoce como **trampa de la variable binaria**.

Ejemplo: trampa de la variable binaria

Supongamos que incluimos tanto *hombre* como *mujer* como variables explicativas en el modelo de salarios visto antes. La matriz X de datos tendría esta forma:

$$X = \begin{vmatrix} 1 & aedu_1 & hombre_1 & mujer_1 \\ 1 & aedu_2 & hombre_2 & mujer_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & aedu_N & hombre_N & mujer_N \end{vmatrix} = \begin{vmatrix} 1 & aedu_1 & 1 & 0 \\ 1 & aedu_2 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & aedu_N & 0 & 1 \end{vmatrix}$$

Notar que las columnas 1, 3 y 4 **no son linealmente independientes**. ¿Qué consecuencia tiene esto sobre los estimadores MC?

Variables *dummy* para categorías múltiples

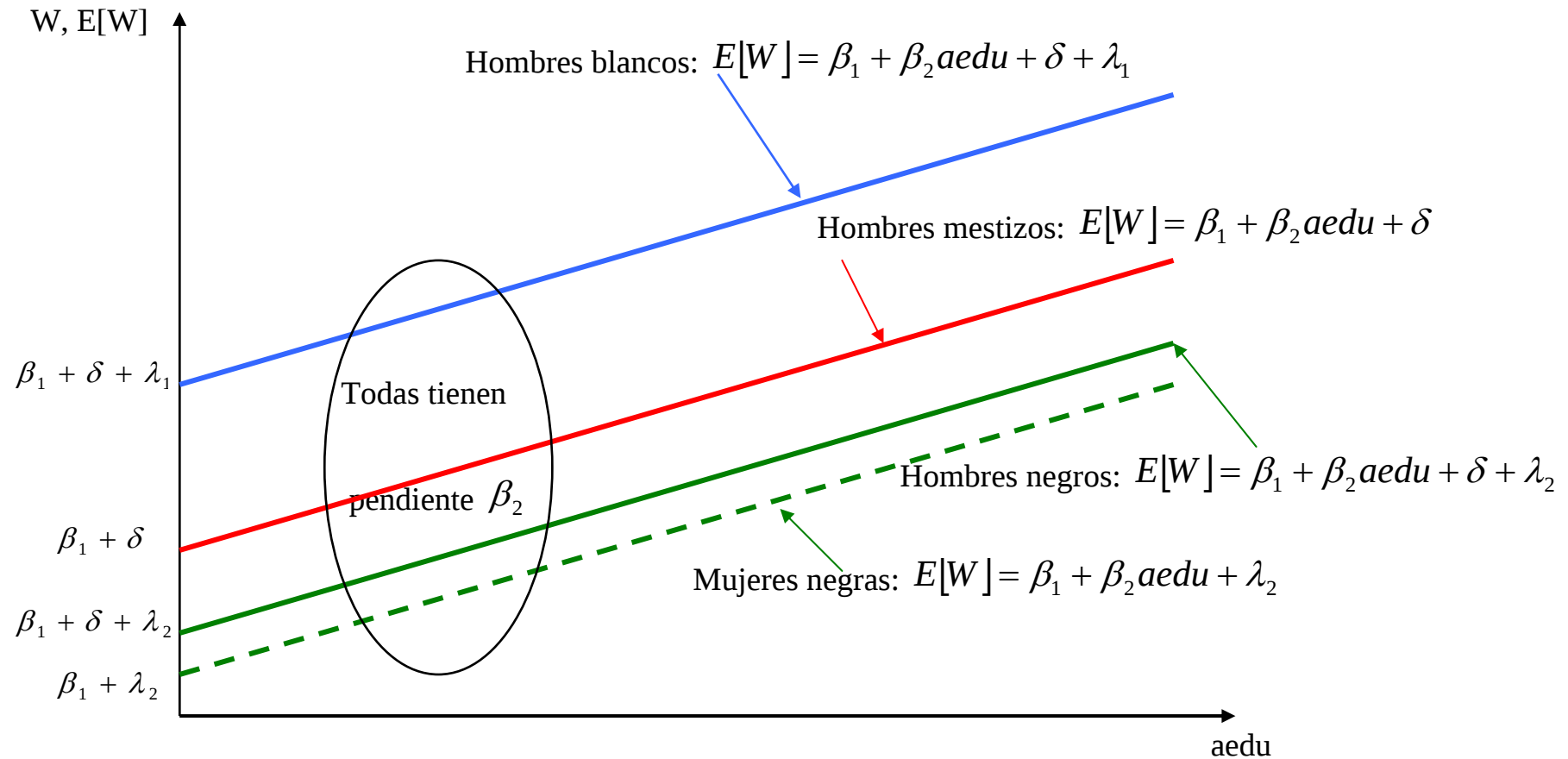
- Supongamos que la raza también afecta los salarios y que hay 3 razas distintas: blancos, negros y mestizos.
- Como son 3 categorías necesitamos 2 dummies. Definimos las variables:

$blanco_i = 1$ si el individuo es de raza blanca, $=0$ si no es blanco

$negro_i = 1$ si el individuo i es de raza negra, $=0$ en caso contrario

- La categoría omitida o base se da cuando $blanco=0$ y $negro=0$, y corresponde a los mestizos.
- El modelo queda $W_i = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 blanco_i + \lambda_2 negro_i + u_i$
- Supongamos que los hombres ganan más que las mujeres (dado todo lo demás!) $\Rightarrow \delta > 0$ y que los blancos ganan más que los mestizos ($\lambda_1 > 0$) pero los negros ganan menos que los mestizos ($\lambda_2 < 0$)

Gráficamente:



Estimación e inferencia cuando intervienen variables binarias:

- Todo igual que antes
- Lo único que cambia es la manera de interpretar los coeficientes

¿Por qué decimos modelo “lineal”?

El modelo *lineal* no es tan *lineal*

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

Notemos que este modelo puede entenderse como *lineal* en dos dimensiones:

- Y es una función lineal de X_2 y X_3 → Modelo lineal en las variables
- Y es una función lineal de β_1 , β_2 , y β_3 → Modelo lineal en los parámetros

Para los estimadores MC lo que importa es la linealidad en los parámetros

Un modelo no lineal puede escribirse en forma general como:

$$Y_i = f(X_{1i}, X_{2i}, \dots, X_{Ki}, u_i)$$

El modelo será **lineal en los parámetros** si existen funciones $g_1, g_2, \dots, g_K$ tales que:

$$g_1(Y_i) = \beta_1 + \beta_2 g_2(X_{1i}, X_{2i}, \dots, X_{Ki}) + \dots + \beta_K g_K(X_{1i}, X_{2i}, \dots, X_{Ki}) + u_i$$

que puede escribirse como:

$$Y_i^* = \beta_1 + \beta_2 X_{2i}^* + \dots + \beta_K X_{Ki}^* + u_i$$

Intuitivamente: si aplicando alguna transformación al modelo no lineal original obtenemos un modelo lineal en los parámetros, el modelo no lineal puede estimarse por MC.

Ejemplos de modelos no lineales que pueden *linealizarse*

- **Modelo logarítmico:** función de demanda exponencial.

$$Q_i = AP_i^\beta \exp(u_i)$$

donde Q es la cantidad demandada, P el precio y u el término aleatorio. A y β son los parámetros desconocidos.

Tomando logaritmos: $\ln Q_i = \ln A + \beta \ln P_i + u_i$

Que puede ser reescrito como: $q_i = \alpha + \beta p_i + u_i$ con $\alpha = \ln A$, $q = \ln Q$ y $p = \ln P$.

Luego, α y β pueden estimarse por MC regresando el $\ln Q$ en el $\ln P$.

Importante: $\beta = \frac{\partial q}{\partial p} = \frac{\partial \ln Q}{\partial \ln P} = \varepsilon_{Q,P} \rightarrow \beta$ es una elasticidad

- **Modelo semilogarítmico:** $Y_i = \exp(\alpha + \beta X_i + u_i)$

Tomando logaritmos: $\ln Y_i \equiv y_i = \alpha + \beta X_i + u_i$

Notar que ahora $\beta = \frac{\partial y}{\partial X} = \frac{\partial \ln Y}{\partial X} \rightarrow \beta$ es una semielasticidad: es el cambio *porcentual* en Y producido por un cambio *marginal* en X .

Importante: la interpretación del efecto de X sobre Y cambia drásticamente dependiendo del modelo en que nos basamos.

Supongamos que $\beta=0.5$, y que tanto Y como X están medidas en pesos.

Modelo	Ecuación	Efecto de X en Y
Lineal en las variables	$Y = \beta_1 + \beta_2 X + u$	Si X aumenta en 1\$, Y aumenta en 50 centavos
Logarítmico	$Y = \alpha X^{\beta_2} \exp(u)$ ó $\ln Y = \beta_1 + \beta_2 \ln X + u$	Si X aumenta en un 1%, Y aumenta en un 0.5%
Semilogarítmico	$Y = \exp(\beta_1 + \beta_2 X + u)$ ó $\ln Y = \beta_1 + \beta_2 X + u$	Si X aumenta en 1\$, Y aumenta en un 50%.

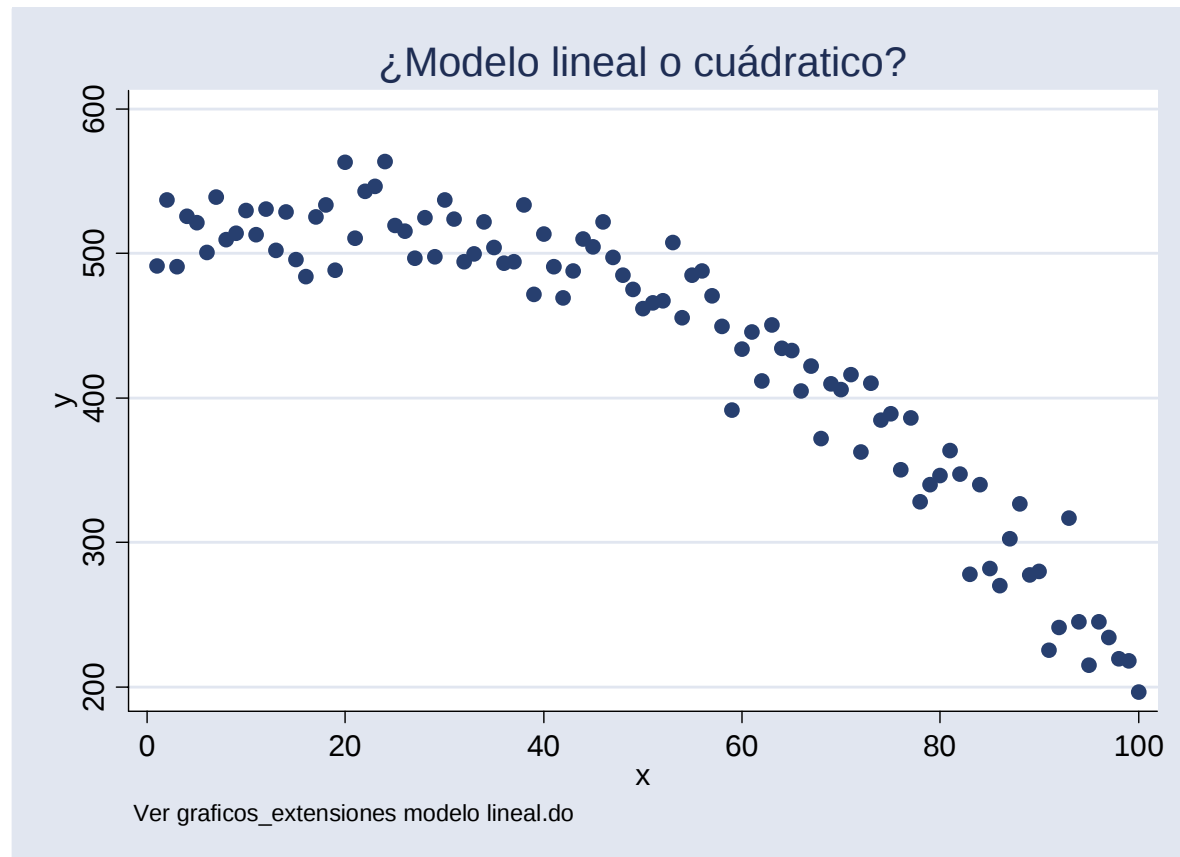
- **Modelo cuadrático en las variables:** $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$

Trivialmente, puede expresarse como un modelo lineal en los parámetros denotando $Z=X^2$.

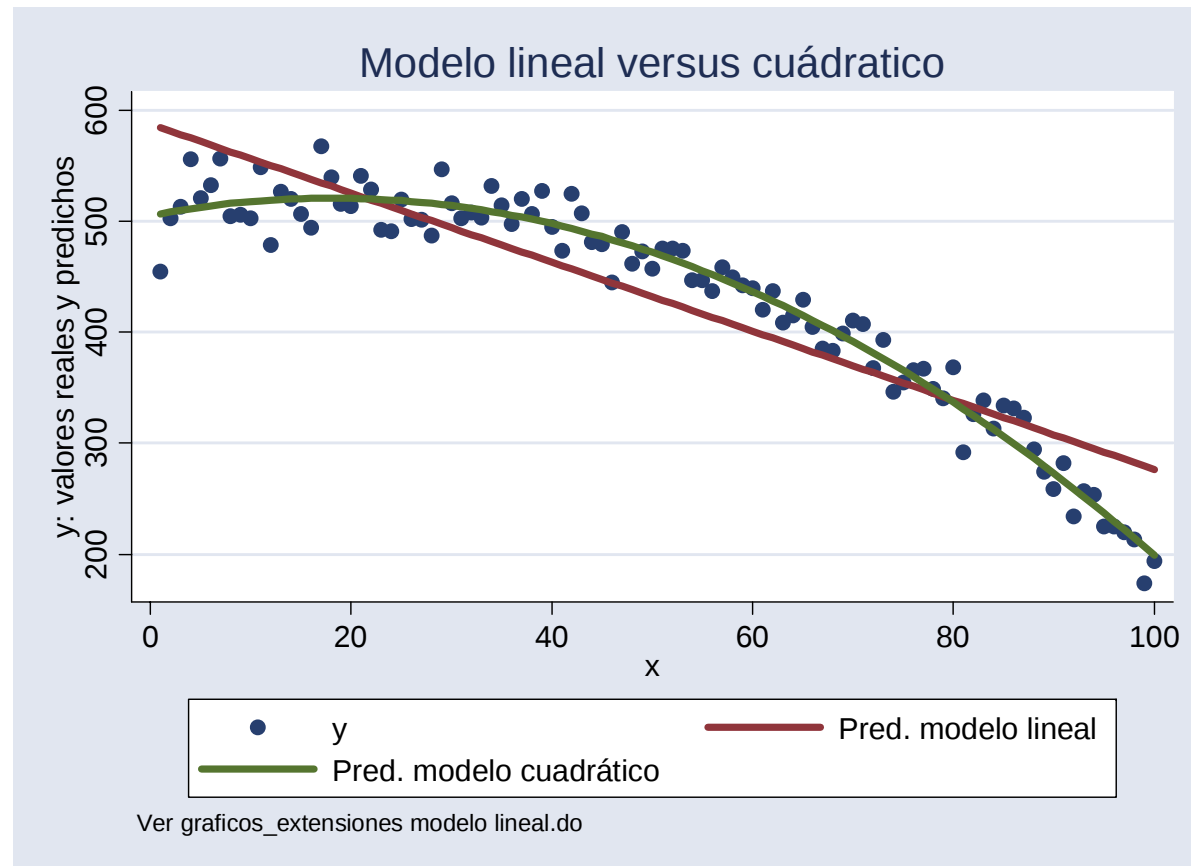
¿Cuál es el efecto de X sobre Y en este modelo?

Notar que: $\frac{dY}{dX} = \beta_2 + 2\beta_3 X$

β_2 ya no resume el efecto de X sobre Y. Este es el “costo” de utilizar un modelo más flexible: las derivadas ya no son constantes, su magnitud depende del valor que tome X.



- o Los datos sugieren una relación no lineal entre Y y X.
- o Estimamos un modelo lineal (sólo X como regresor) y uno cuadrático (X y X^2)



- o El modelo lineal predice un efecto marginal constante de X sobre Y.
- o El efecto marginal predicho por el modelo cuadrático crece con el valor de las X (pendiente).

- **Modelo con interacciones:** $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{2i} X_{3i} + u_i$

Es lineal en los parámetros para $Z = X_2 X_3$.

Notar que: $\frac{\partial Y}{\partial X_2} = \beta_2 + \beta_4 X_3$

Es importante no interpretar a β_2 como la derivada de Y con respecto a X_2 . β_2 ya no resume el efecto de esa variable sobre Y porque ese efecto también depende de β_4 y del valor que toma la variable X_3 .

Ejemplo: Y =salarios, X_2 =educación y X_3 =inteligencia

Algunos investigadores afirman que la inteligencia y la educación interactúan de modo que el efecto de la educación sobre los salarios es mayor para gente más inteligente. El término de interacción (producto de las variables X_2 y X_3) estaría captando ese efecto.

Ejemplo empírico: los determinantes de los salarios

Especificación semilogarítmica: $\ln(W_i) = \beta_1 + \beta_2 aedu_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 hombre_i + u_i$

- W es el salario por hora trabajada
- $aedu$ son años de educación formal
- $edad$ en años
- $edad^2$ es la edad al cuadrado
- $hombre$ es un indicador binario, =1 si hombre, =0 si mujer.

Datos: extraídos de EPH onda octubre 2002, todos los aglomerados encuestados.

Para reproducir estos resultados ver datos en “data_modelos_de_ingresos.dta” y rutina STATA en “estima_modelos_de_ingresos.do”

Resultados de la estimación del modelo de regresión

```
regress logw aedu edad edad2 hombre
```

Source	SS	df	MS	Number of obs = 24138		
Model	3825.0433	4	956.260826	F(4, 24133) = 2038.00		
Residual	11323.5513	24133	.469214406	Prob > F = 0.0000		
Total	15148.5946	24137	.627608839	R-squared = 0.2525		
				Adj R-squared = 0.2524		
				Root MSE = .68499		

logW	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
aedu	.0907561	.0011789	76.99	0.000	.0884454	.0930667
edad	.0622396	.0019396	32.09	0.000	.0584379	.0660413
edad2	-.0005638	.0000232	-24.31	0.000	-.0006093	-.0005184
hombre	.0712211	.0089857	7.93	0.000	.0536087	.0888336
_cons	-1.645838	.0401596	-40.98	0.000	-1.724553	-1.567122

Resultados de la estimación del modelo de regresión (cont.)

Source	SS	df	MS	Number of obs = 24138		
Model	3825.0433	4	956.260826	F(4, 24133) = 2038.00		
Residual	11323.5513	24133	.469214406	Prob > F = 0.0000		
				R-squared = 0.2525		
				Adj R-squared = 0.2524		
Total	15148.5946	24137	.627608839	Root MSE = .68499		

logW	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
aedu	.0907561	.0011789	76.99	0.000	.0884454	.0930667
edad	.0622396	.0019396	32.09	0.000	.0584379	.0660413
edad2	-.0005638	.0000232	-24.31	0.000	-.0006093	-.0005184
hombre	.0712211	.0089857	7.93	0.000	.0536087	.0888336
_cons	-1.645838	.0401596	-40.98	0.000	-1.724553	-1.567122

- Recordar que este modelo corresponde a la especificación semilogarítmica.
- Según estos resultados, un año más de educación (*aedu*) tendría el efecto de aumentar en un 9.07% el salario por hora. Esto se conoce en la literatura como **retorno a la educación**.

Resultados de la estimación del modelo de regresión (cont.)

Source	SS	df	MS	Number of obs = 24138		
Model	3825.0433	4	956.260826	F(4, 24133) = 2038.00		
Residual	11323.5513	24133	.469214406	Prob > F = 0.0000		
Total	15148.5946	24137	.627608839	R-squared = 0.2525		
				Adj R-squared = 0.2524		
				Root MSE = .68499		

logW	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
aedu	.0907561	.0011789	76.99	0.000	.0884454	.0930667
edad	.0622396	.0019396	32.09	0.000	.0584379	.0660413
edad2	-.0005638	.0000232	-24.31	0.000	-.0006093	-.0005184
hombre	.0712211	.0089857	7.93	0.000	.0536087	.0888336
_cons	-1.645838	.0401596	-40.98	0.000	-1.724553	-1.567122

- Edad: tiene un efecto *no lineal* sobre el log. de los salarios horarios (*edad* y *edad*² son significativas). Si la edad aumenta 1 año esperamos que el salario por hora aumente en $[0.06 - 2 \times 0.0005 \times \text{edad}] \times 100\%$.
- Por ejemplo, un año más en un individuo de 20 años tendría el efecto de aumentar el salario horario en un 4%.

- ¿Cómo se interpreta el coeficiente de la variable binaria *hombre*?

Denotemos $\ln \hat{W} | (\text{hombre} = 1) \equiv \ln \hat{W}_H$ y $\ln \hat{W} | (\text{hombre} = 0) \equiv \ln \hat{W}_M$

Entonces:

$$\ln \hat{W}_H = \hat{\beta}_1 + \hat{\beta}_2 aedu + \hat{\beta}_3 edad + \hat{\beta}_4 edad^2 + \hat{\beta}_5$$

$$\ln \hat{W}_M = \hat{\beta}_1 + \hat{\beta}_2 aedu + \hat{\beta}_3 edad + \hat{\beta}_4 edad^2$$

$$\ln \hat{W}_H - \ln \hat{W}_M = \ln \left[\frac{\hat{W}_H}{\hat{W}_M} \right] = \hat{\beta}_5 \quad \Rightarrow \quad \left[\frac{\hat{W}_H}{\hat{W}_M} \right] - 1 = \left[\frac{\hat{W}_H - \hat{W}_M}{\hat{W}_M} \right] = \exp(\hat{\beta}_5) - 1$$

Para $\hat{\beta}_5$ pequeño, $\exp(\hat{\beta}_5) - 1 \cong \hat{\beta}_5$. Entonces, dado que $\hat{\beta}_5 = 0.07$, deberíamos esperar que un hombre gane por hora un 7% más que una mujer con la misma edad y educación.

- Bondad del ajuste $R^2 = 25\%$ (¿es bueno o es malo?)

Interacción entre educación y género: los retornos a la educación pueden diferir entre varones y mujeres. Una especificación que permite capturar este efecto:

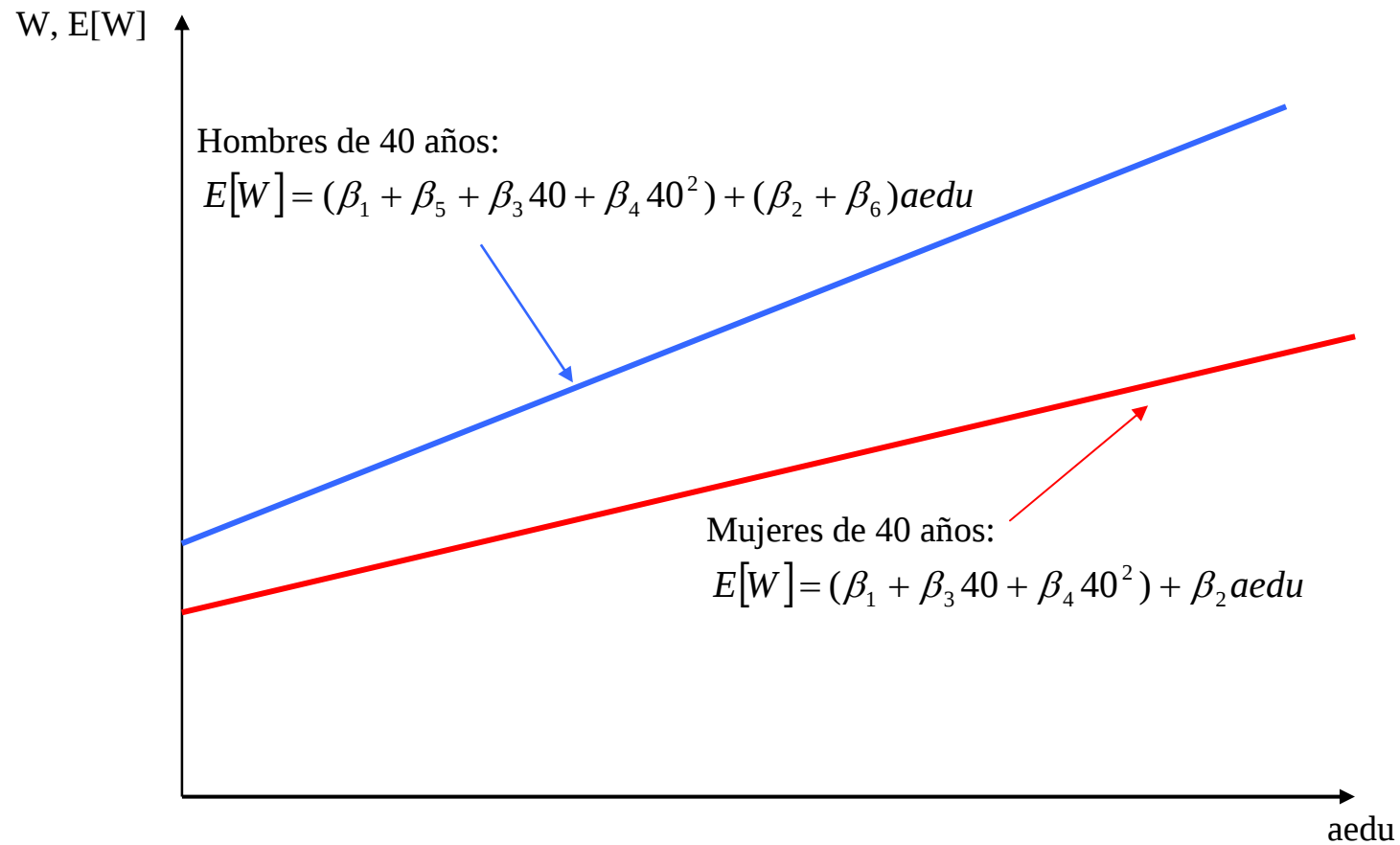
$$\ln(W_i) = \beta_1 + \beta_2 edu_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 hombre_i + \beta_6 (hombre \times aedu)_i + u_i$$

¿Cuál es el retorno a la educación bajo esta especificación?:

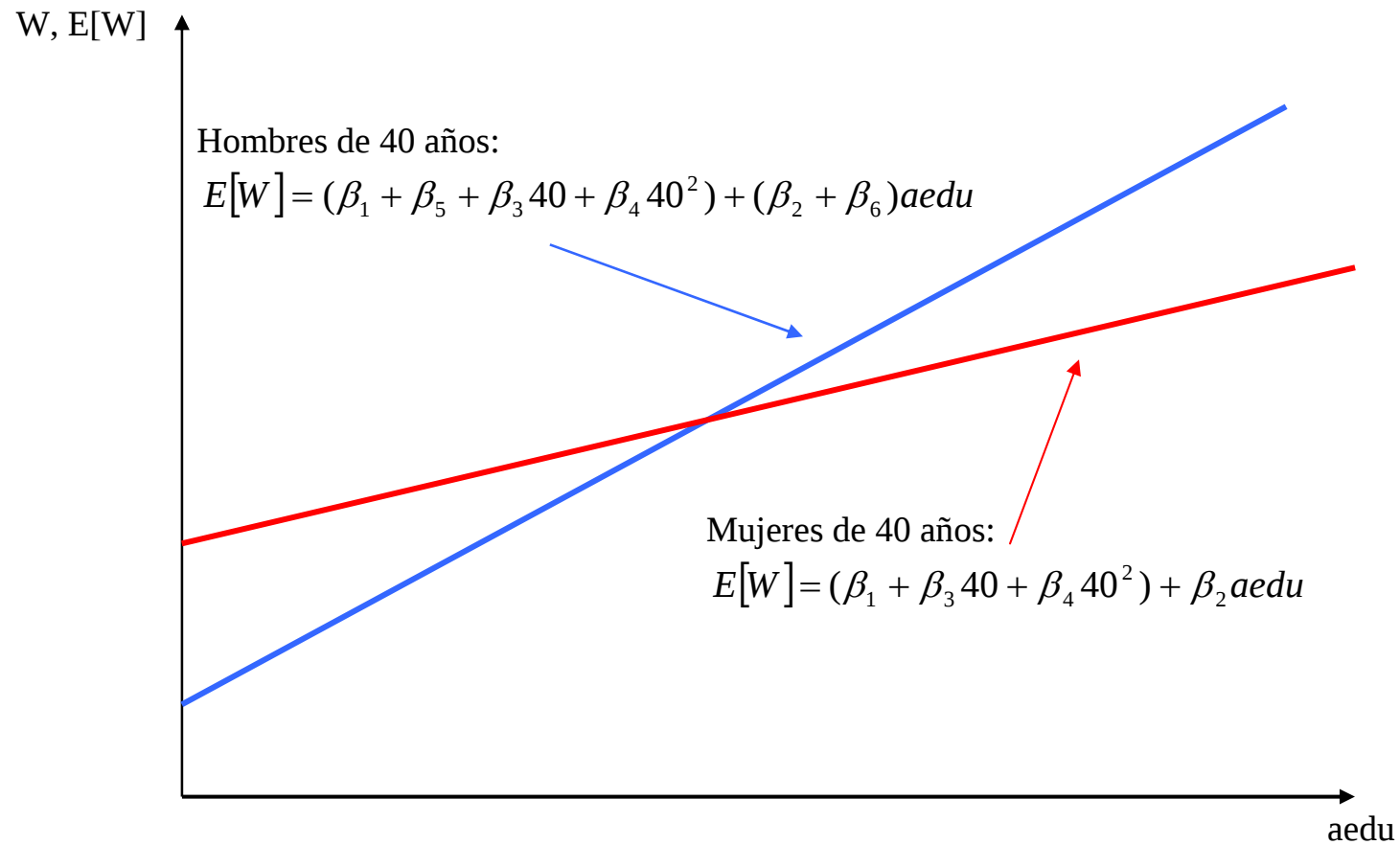
→ para mujeres $\beta_2 \times 100\%$

→ para hombres $(\beta_2 + \beta_6) \times 100\%$

Gráficamente: supongamos que $\beta_5 > 0$ y $\beta_6 > 0$



Gráficamente pero con $\beta_5 < 0$ y $\beta_6 > 0$



Resultados de estimar el modelo con interacción

```
regress logW aedu edad edad2 hombre hombre_aedu
```

Source	SS	df	MS	Number of obs = 24138		
Model	3826.06717	5	765.213433	F(5, 24132) = 1630.92		
Residual	11322.5274	24132	.469191422	Prob > F = 0.0000		
Total	15148.5946	24137	.627608839	R-squared = 0.2526		
				Adj R-squared = 0.2524		
				Root MSE = .68498		

logW	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
aedu	.0888402	.0017526	50.69	0.000	.0854049	.0922755
edad	.0622024	.0019397	32.07	0.000	.0584005	.0660044
edad2	-.0005635	.0000232	-24.30	0.000	-.000609	-.000518
hombre	.0346514	.026336	1.32	0.188	-.0169689	.0862716
hombre_aedu	.0034475	.0023337	1.48	0.140	-.0011268	.0080217
_cons	-1.623832	.0428326	-37.91	0.000	-1.707786	-1.539877

- Ahora ni la dummy *hombre* ni la interacción entre *hombre* y educación son estadísticamente significativas. ¿A qué se debe este cambio de resultados? (ver más adelante)
- Sugiere que los retornos a la educación no difieren entre géneros (generalmente en las aplicaciones sí se encuentran diferencias significativas).