Modelo lineal general (K variables)

Interpretación y usos

Mariana Marchionni
marchionni.mariana@gmail.com

Econometría I - FCE - UNLP
www.econometria1unlp.com

Temario de la clase

Vimos el modelo lineal simple bajo los supuestos clásicos

$Y_i = \alpha + \beta X_i + \mu_i, \;\; i = 1, ...n$

1. $E[\mu_i] = 0, \;\; i = 1, ..., n.$
2. $V[\mu_i] = \sigma^2, \;\; i = 1, ..., n.$
3. $Cov[\mu_i, \mu_j] = 0, \;\; i \neq j.$
4. Las X_i no son aleatorias y no son todas iguales

El modelo lineal general (K variables)

Simplemente agregamos variables explicativas:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_K X_{Ki} + \mu_i, \quad i = 1, ..., n$$

- Ahora los regresores son X_2, ..., X_K
- Desde ahora, por comodidad de notación, a la ordenada al origen del modelo la llamaremos β_1. Es como si hubiera una variable explicativa $X_{1i} = 1$ para todas las observaciones.
- Mantendremos todos los supuestos clásicos salvo una pequeña modificación en la segunda parte del supuesto 4

No multicolinealidad perfecta en el modelo lineal general

- **Requiere que no haya dependencia lineal entre las variables explicativas.**

- En otras palabras, que ninguna de las variables explicativas pueda expresarse como una combinación lineal de las demás.
    - No puede darse que existan constantes a_j, no todas iguales a cero, tales que $X_k = \sum_{j \neq k} a_j X_j \; k = 1, ...K$

- **No multicolinealidad perfecta $\neq$ no correlación entre las variables explicativas**
    - puede haber correlación siempre que no sea perfecta
    - no alcanza con que la correlación no sea perfecta de a pares de variables
    - ninguna de las variables $X_2, ..., X_K$ puede ser una constante. ¿Por qué?

Interpretación de los parámetros del modelo lineal general

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_K X_{Ki} + \mu_i, \quad i = 1, ..., n$$

- Con $E(\mu_i) = 0$ y regresores no aleatorios tenemos que:
$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_K X_{Ki}$$
- El efecto marginal de X_k viene dado por:
$$\frac{\partial E(Y_i)}{\partial X_{ki}} = \beta_k \quad \forall i$$

Interpretación de los parámetros del modelo lineal general

$$\frac{\partial E\left(Y_i\right)}{\partial X_{ki}} = \beta_k \quad \forall i$$

- β_k mide el efecto sobre $E(Y_i)$ de un cambio marginal de la k-ésima variable explicativa, **manteniendo constantes todas las demás** (derivada parcial)
- El significado de *"cambio marginal"* está atado a las unidades de medida de la variable explicativa (1 centavo, 1\$, 1 mil \$, 1 millón \$, etc.)

Interpretación de los parámetros del modelo lineal general

- ¿Y si el cambio no es **marginal** sino **discreto**?
- ¿Cuál es el efecto sobre $E(Y_i)$ cuando X_s cambia en $\triangle X_s$?

$$\triangle E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_s(X_{si} + \triangle X_{si}) + ... + \beta_K X_{Ki}$$

$$- [\beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_s X_{si} + ... + \beta_K X_{Ki}]$$

$$\triangle E(Y_i) = \beta_s \triangle X_{si}$$

- Vale la regla de 3. Notar que dejamos constantes todas las demás variables explicativas (*ceteris paribus*)

Interpretación de los parámetros del modelo lineal general

- Podemos evaluar cambios discretos en variables de naturaleza continua
  - un aumento en el ingreso de los hogares de 50.000\$ a 80.000\$.
  - un cambio en el nivel educativo de primario completo a secundario completo que implica 6 (o 5) años más de educación.

- También hay variables que son discretas por naturaleza:
  - Variables de conteo (*count variables*):
    número de hijos por familia, número de veces que un estudiante repitió un año, número de visitas al médico, etc.
  - **Variables binarias o *dummy***

Variables explicativas binarias

- ¿Cómo incorporamos al modelo **variables explicativas que son cualitativas**?
    - género
    - raza
    - sector productivo
    - región geográfica
- Primero nos concentramos en los **fenómenos binarios**: se tiene o no cierta característica
- En un fenómeno binario hay **sólo dos posibilidades y son excluyentes**.
    - una persona trabaja o no trabaja
    - un banco quebró o no quebró durante la crisis
    - un país firmó o no firmó cierto acuerdo internacional
    - una trabajadora doméstica "está en blanco" o no

- ¿Cómo incorporamos un fenómeno binario en el modelo?
- Vamos a crear una **variable artificial** que toma un valor cuando la característica de interés está presente y otro valor distinto cuando no lo está.
- Los valores son arbitrarios, pero es práctico elegir **1 y 0**
- Llamamos **variable binaria o *dummy*** a esa variable artificial
- **Ejemplo**: queremos incorporar una variable que indique el género del individuo. Definamos la variable dummy $d_i = 1$ si el individuo i es hombre, $d_i = 0$ si el individuo i es mujer.

obs.	sexo	d
1	mujer	0
2	hombre	1
...	...	...
n	hombre	1

Podríamos capturar la misma información definiendo la variable d al revés: que valga 1 para las mujeres y 0 para los hombres

obs.	sexo	d (opción 1)	d (opción 2)
1	mujer	0	1
2	hombre	1	0
...	...	...	...
n	hombre	1	0

- Esta abitrariedad no modifica los resultados, sólo condiciona la manera de interpretarlos
- **Resulta útil que el nombre de la variable dummy sugiera cuál es la característica asociada con el valor 1**. En la primera opción la variable en vez de llamar d a la variable dummy, podríamos llamarla *hombre* ($hombre_i = 1$ para los hombres y $hombre_i = 0$ para las mujeres)
- **La característica asociada al valor 0 se conoce como categoría base.** En la opción 1, las mujeres son la categoría base (importante para la interpretación, ya veremos)

Ejemplo: modelo de los determinantes del salario

$$W_i = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \mu_i$$

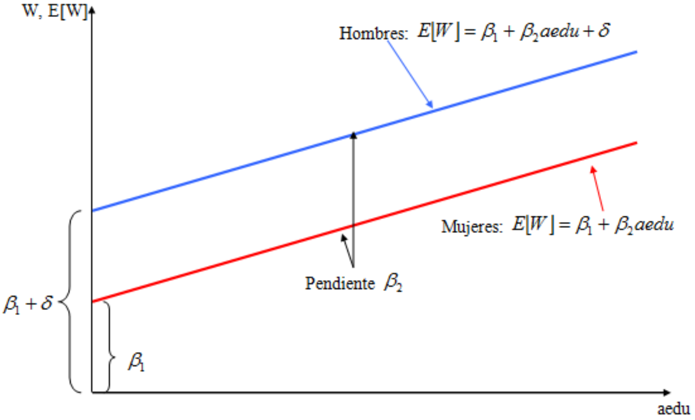
- Si $E(\mu_i) = 0 \Rightarrow E(W_i) = \beta_1 + \beta_2 aedu_i + \delta hombre_i$
- Notar que hay dos regresiones:
  - Para los hombres: $E[W_i \mid hombre_i = 1] = \beta_1 + \beta_2 aedu_i + \delta$
  - Para las mujeres: $E[W_i \mid hombre_i = 0] = \beta_1 + \beta_2 aedu_i$
- Restando miembro a miembro:

$$E[W_i \mid hombre_i = 1] - E[W_i \mid hombre_i = 0] =$$

$$\beta_1 + \beta_2 aedu_i + \delta \times 1 - (\beta_1 + \beta_2 aedu_i + \delta \times 0) = \delta$$

- Entonces δ es la diferencia entre el salario esperado de un hombre y de una mujer que tienen *la misma educación*

$$E(W_i) = \beta_1 + \beta_2 aedu_i + \delta hombre_i \text{ con } \delta > 0$$



- La recta de regresión de los hombres es paralela a la de las mujeres (misma pendiente) pero tiene una ordenada al origen mayor.

- El fenómeno binario tiene 2 categorías (hombre o mujer), pero **alcanza con una sola variable dummy** para diferenciar entre ambas.
- Dada la información de la variable dummy *hombre*, la variable *mujer* no agrega nada, es **redundante**.

obs.	sexo	d (opción 1): *hombre*	d (opción 2): *mujer*
1	mujer	0	1
2	hombre	1	0
...	...	...	...
n	hombre	1	0

- **Regla: si hay 2 categorías incluimos sólo una *dummy*.**
- ¿Qué pasa si incluimos las dos variables *dummy* en el modelo? **Trampa de la variable binaria**

- ¿Cuál de las 2 variables dummy incluir?
  - Opción 1: $W_i = \beta_1 + \beta_2 aedu_i + \delta\, hombre_i + \mu_i$
  - Opción 2: $W_i = \beta_1 + \beta_2 aedu_i + \eta\, mujer_i + \mu_i$

- Puede mostrarse que **las 2 opciones son equivalentes** con $\delta = -\eta$, pero **la interpretación cambia** con la opción (¿cuál es la categoría base en cada caso?)

- Supongamos $\delta = 5\, (\Rightarrow \eta = -5)$
  - Interpretación basada en la Opción 1: el salario esperado de un hombre es **5$ más alto que el de una mujer** (categoría base) con la misma educación
  - Interpretacion basada en la Opción 2: el salario esperado de una mujer es **5$ más bajo que el de un hombre** (categoría base) con la misma educación

Variables dummy para categorías múltiples

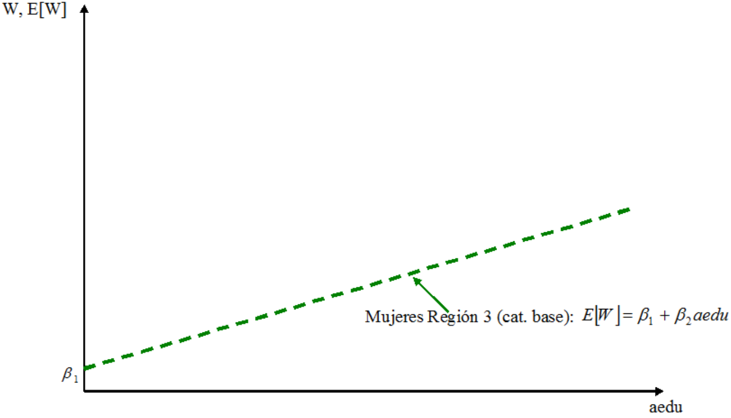
- Supongamos que los salarios no solo dependen de la educación y del género, sino también de la región; supongamos que hay 3 regiones
- Generalizamos la regla: **si hay S categorías incluimos S-1 variables *dummy***
- Hay 3 regiones $\Rightarrow$ necesitamos 2 dummies:
  - $región1_i \begin{cases} = 1 & \textit{si el individuo pertenece a la región } 1 \\ = 0 & \textit{en caso contrario} \end{cases}$
  - $región2_i \begin{cases} = 1 & \textit{si el individuo pertenece a la región } 2 \\ = 0 & \textit{en caso contrario} \end{cases}$
- La **categoría base u omitida** se da cuando $región1_i = 0$ *y* $región2_i = 0$, y corresponde a la región 3

El modelo viene dado por:

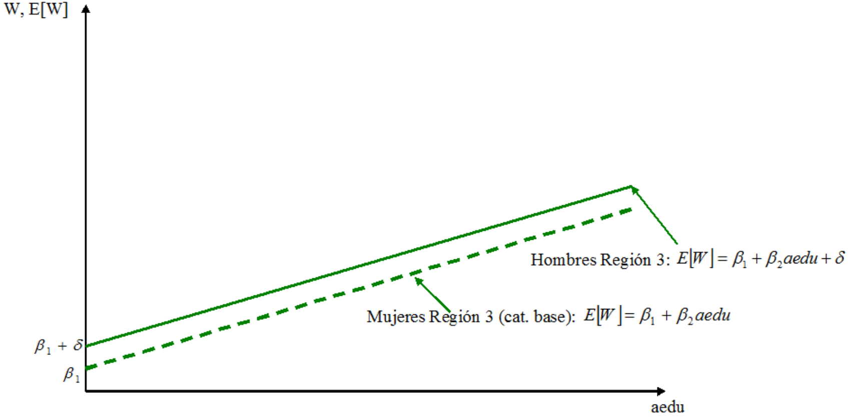
$$W_i = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i + \mu_i$$

- Supongamos que los hombres ganan más que las mujeres (manteniendo las demás variables constantes) $\Rightarrow \delta > 0$
- Además, supongamos que los trabajadores de la región 1 ganan más que los de la región 2, que a su vez ganan más que los de la región 3 ($\lambda_1 > \lambda_2 > 0$) (dado todo lo demás...¿qué es todo lo demás?)

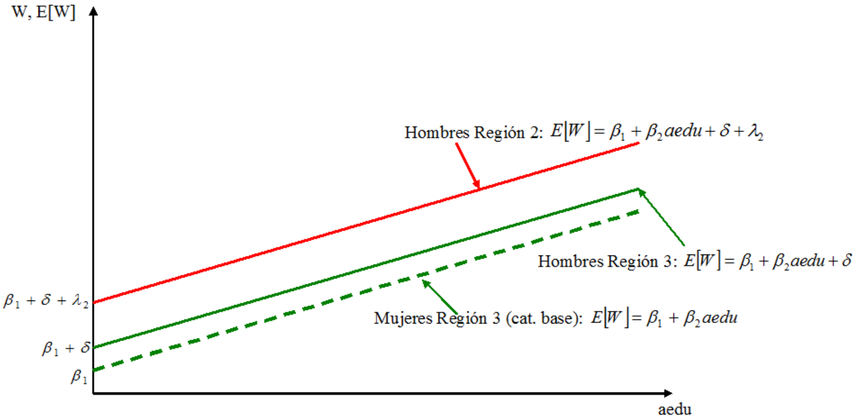
$$E[W_i] = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i$$
$$\text{con } \beta_1 > 0,\ \beta_2 > 0,\ \delta > 0 \text{ y } \lambda_1 > \lambda_2 > 0$$



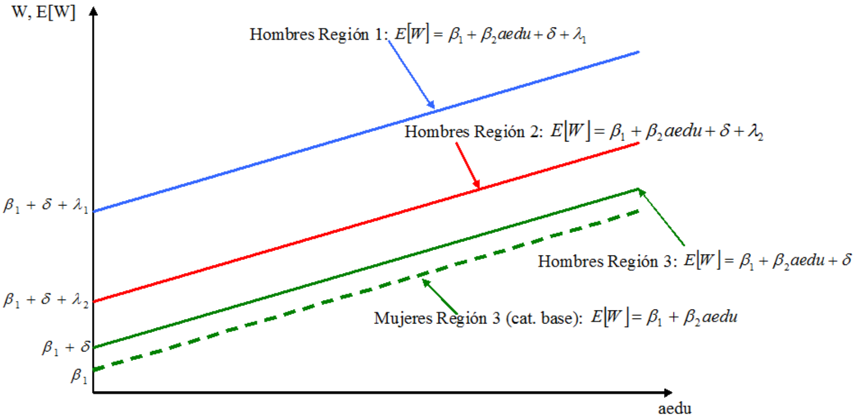
$$E[W_i] = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i$$
$$\text{con } \beta_1 > 0,\ \beta_2 > 0,\ \delta > 0 \text{ y } \lambda_1 > \lambda_2 > 0$$



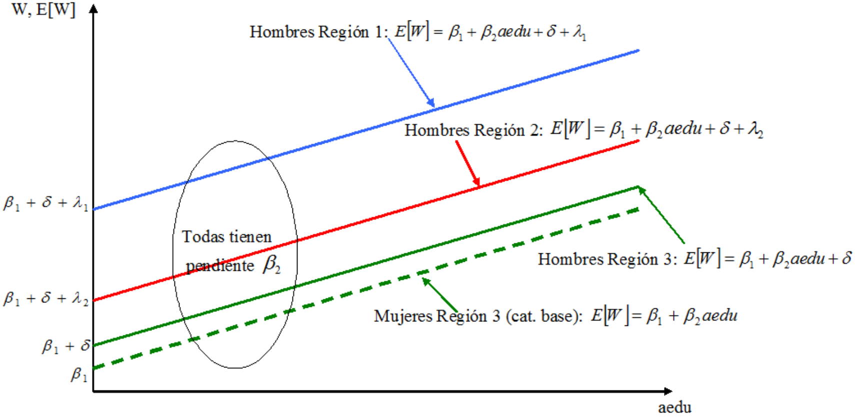
$$E[W_i] = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i$$
$$\text{con } \beta_1 > 0,\ \beta_2 > 0,\ \delta > 0 \text{ y } \lambda_1 > \lambda_2 > 0$$



$$E[W_i] = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i$$
$$\text{con } \beta_1 > 0,\ \beta_2 > 0,\ \delta > 0 \text{ y } \lambda_1 > \lambda_2 > 0$$



$$E[W_i] = \beta_1 + \beta_2 aedu_i + \delta hombre_i + \lambda_1 región1_i + \lambda_2 región2_i$$
$$\text{con } \beta_1 > 0,\ \beta_2 > 0,\ \delta > 0 \text{ y } \lambda_1 > \lambda_2 > 0$$



Estimación e inferencia con variables explicativas binarias

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

El modelo lineal no es tan lineal

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_K X_{Ki} + \mu_i$$

Notemos que este modelo es ***lineal*** **en las variables y en los parámetros**:

- Y es una función lineal de las variables $X_2, X_3, ..., X_K$ $\Rightarrow$Modelo lineal en las variables
- Y es una función lineal de $\beta_1, \beta_2, ..., \beta_K$ $\Rightarrow$Modelo lineal en los parámetros

MCO sólo requiere linealidad en los parámetros

Ya vimos algunos ejemplos de no linealidades en las variables

Variables en logaritmos

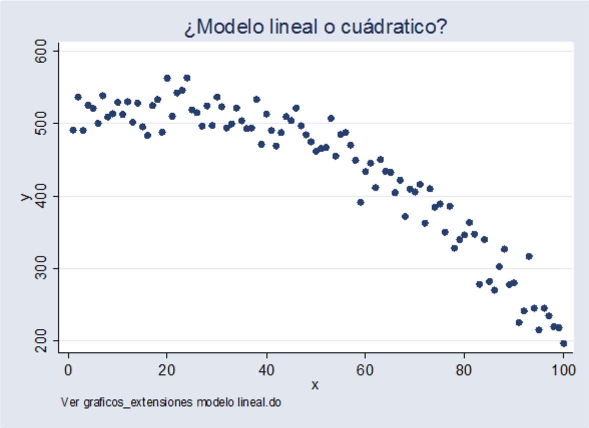
- Modelo logarítmico (log-log): $lnY_i = \beta_1 + \beta_2 lnX_i + \mu_i$
  $\rightarrow \beta_2$ es una elasticidad
- Modelo semilogarítmico (log-lin): $lnY_i = \beta_1 + \beta_2 X_i + \mu_i$
  $\rightarrow \beta_2$ es una semielasticidad

Interpretación

- Importante: siempre tenemos que pensar en **qué modelo** estamos y cuáles son las **unidades de medida** de las variables
- Supongamos que Y y X están medidas en pesos, y que $\beta_2 = 0{,}5$

Modelo	Ecuación	Efecto marginal
Lineal	$Y = \beta_1 + \beta_2 X_i + u_i$	Si X aumenta en \$1, Y aumenta en \$0,50
Log-log	$\ln Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	Si X aumenta un 1 %, Y aumenta en 0,50 %
Log-lin	$\ln Y_i = \beta_1 + \beta_2 X_i + u_i$	Si X aumenta en \$1, Y aumenta en 50 % (0,50 $\times$ 100 %)
Lin-log	$Y_i = \beta_1 + \beta_2 \ln X_i + u_i$	Si X aumenta en 1 %, Y aumenta en 0,005\$ ($\frac{0,50}{100}$)

¿Qué sugieren estos datos?



Variables cuadráticas

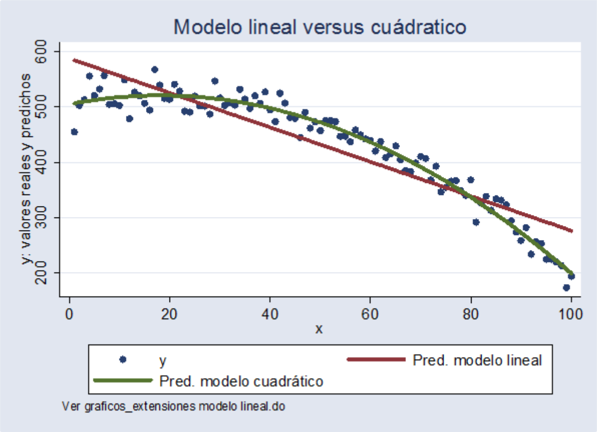
Modelo cuadrático en X:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \mu_i$$

- El efecto marginal de X ahora viene dado por:

$$\frac{\partial E[Y_i]}{\partial X_i} = \beta_2 + 2\beta_3 X_i$$

- **El efecto marginal ya no es constante**: depende de β_2, de β_3 y del valor que asignemos a X_i.
- No sólo la magnitud del efecto depende de X. Cuando β_2 y β_3 tienen distinto signo, el signo del efecto marginal también depende del valor que asignemos a X_i



- **Por construcción**, el modelo lineal predice un efecto marginal constante de X sobre Y
- Notar que el efecto marginal del modelo cuadrático cambia con el valor de las X (aquí el efecto marginal es cada vez más negativo)

Otra forma de no linealidad: interacción entre variables

- Un año más de educación ¿tiene el mismo efecto sobre la productividad/salarios para cualquiera?
- Es factible que el "talento" potencie el efecto de la educación sobre la productividad/salarios, de manera que un año más de educación "rinda" más para individuos con más talento.
- La interacción entre variables se usa para capturar situaciones donde **el efecto de una variable X se potencia o atenúa con el de otra variable Z**

Otra forma de no linealidad: interacción entre variables

- Incluimos en el modelo un **término multiplicativo** entre las variables para capturar su interacción:

$$Y_i = \beta_1 + \beta_2 X_i + \gamma(X_i \times Z_i) + u_i$$

- Ahora el efecto marginal de X viene dado por

$$\frac{\partial E[Y_i]}{\partial X_i} = \beta_2 + \gamma Z_i$$

  - Si $\beta_2 > 0$ y $\gamma > 0$, aumentos de Z potencian el efecto positivo de X sobre $E[Y]$
  - Si $\beta_2 > 0$ y $\gamma < 0$, aumentos de Z atenúan (e incluso podrían contrarrestar) el efecto positivo de X sobre $E[Y]$
  - Notar que cuando el signo de los coeficientes coincide, una variable potencia el efecto de la otra; de lo contrario, el efecto se atenúa.

Ejemplo: efecto de la educación sobre los salarios

- Supongamos que el salario (Y) depende de los años de educación (X) y de los años de experiencia laboral (Z).
- También creemos que la educación tiene un efecto positivo sobre los salarios que se potencia con la experiencia laboral.
- Entonces propondríamos un modelo con interacción entre X y Z y esperaríamos que $\beta_2 > 0$ y $\gamma > 0$
- ¿Y si Z fuera una variable dummy?

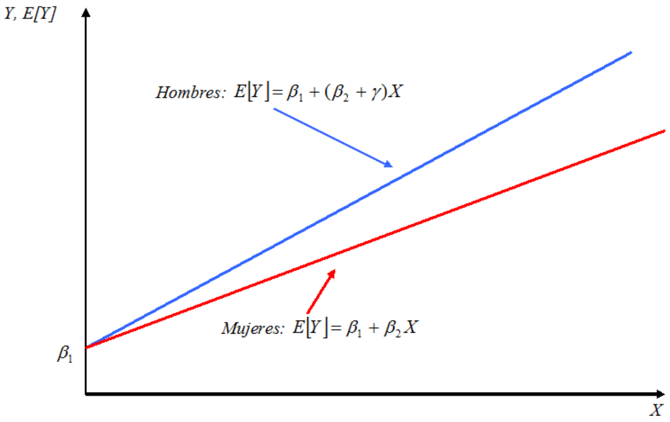
Interacción con una dummy

Volvamos al modelo con interacción, pero reemplacemos Z por la variable dummy *hombre* (=1 si es hombre, =0 si es mujer)

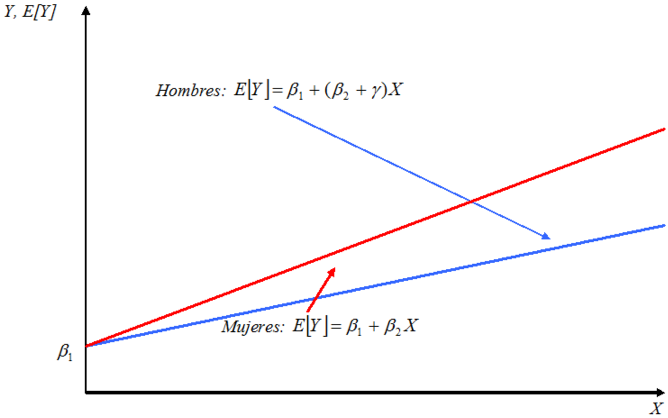
$$Y_i = \beta_1 + \beta_2 X_i + \gamma(X_i \times hombre_i) + \mu_i$$

- Tenemos dos modelos en realidad:
  - Para las mujeres: $Y_i = \beta_1 + \beta_2 X_i + \mu_i$
  - Para los hombres: $Y_i = \beta_1 + \beta_2 X_i + \gamma X_i + \mu_i$
- El efecto marginal de X también difiere:
  - Para las mujeres: $\frac{\partial E[Y_i]}{\partial X_i} = \beta_2$
  - Para los hombres: $\frac{\partial E[Y_i]}{\partial X_i} = \beta_2 + \gamma$

$$E[Y_i] = \beta_1 + \beta_2 X_i + \gamma(X_i \times hombre_i) \text{ con } \beta_2 > 0 \text{ y } \gamma > 0$$



$$E[Y_i] = \beta_1 + \beta_2 X_i + \gamma(X_i \times hombre_i) \text{ con } \beta_2 > 0 \text{ y } \gamma < 0$$



Dummies aditivas y multiplicativas

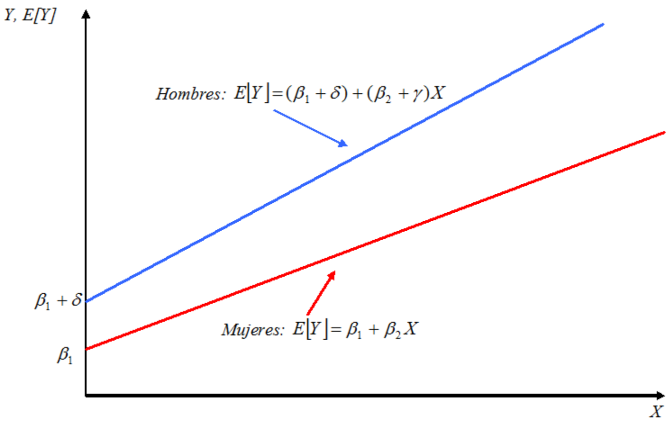
Consideremos el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_i + \delta hombre_i + \gamma(X_i \times hombre_i) + \mu_i$$

- La dummy aditiva en el término $\delta hombre_i$ permite que las ordenadas al origen difieran
- La dummy multiplicativa en el término $\gamma(X_i \times hombre_i)$ permite que las pendientes de las regresiones difieran por género

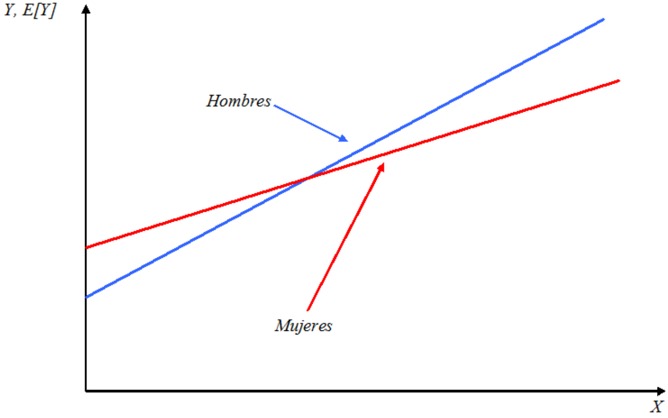
$$E[Y_i] = \beta_1 + \beta_2 X_i + \delta hombre_i + \gamma(X_i \times hombre_i)$$

supongamos todos los parámetros positivos



$$E[Y_i] = \beta_1 + \beta_2 X_i + \delta hombre_i + \gamma(X_i \times hombre_i)$$

¿Qué signo tienen los parámetros?



Ejemplo empírico: estimación de un modelo de los determinantes de los salarios en la Argentina

- **Modelo log-lin (semilogarítmico) con variables lineales, cuadráticas y dummies:**

$$ln(W_i) = \beta_1 + \beta_2 aedu_i + \beta_3 edad_i + \beta_4 edad_i^2 + \delta hombre_i + \mu_i$$

  - *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 una variable dummy =1 si hombre, =0 si mujer

- Datos de la Encuesta Permanente de Hogares para todo el país (EPH)

```
regress logW aedu edad edad2 hombre

      Source |       SS       df       MS              Number of obs =   24138
-------------+------------------------------           F(  4, 24133) = 2038.00
       Model |  3825.0433      4  956.260826           Prob > F      =  0.0000
    Residual |  11323.5513 24133  .469214406           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
------------------------------------------------------------------------------
```

Source	SS	df	MS
Model	3825.0433	4	956.260826
Residual	11323.5513	24133	.469214406
Total	15148.5946	24137	.627608839

Number of obs	=	24138
F(4, 24133)	=	2038.00
Prob > F	=	0.0000
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

- Recordar: especificación semilogarítmica.
- Un año más de educación (*aedu*) tendría el efecto de aumentar en un 9.07% el salario por hora esperado, manteniendo constantes las demás variables (***retorno a la educación***)

Source	SS	df	MS
Model	3825.0433	4	956.260826
Residual	11323.5513	24133	.469214406
Total	15148.5946	24137	.627608839

Number of obs	= 24138
F(4, 24133)	= 2038.00
Prob > F	= 0.0000
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

- Como *edad*2 es estadísticamente significativa, decimos que la edad tiene un *efecto no lineal significativo* sobre el log. de los salarios horarios.
- Si la *edad* aumenta 1 año, *ceteris paribus*, esperamos que el salario por hora aumente en [0.06 – 2 x 0.0005 x *edad*] x 100%.
- Para *edad*=20 años, el efecto estimado es de un 4%.

Source	SS	df	MS
Model	3825.0433	4	956.260826
Residual	11323.5513	24133	.469214406
Total	15148.5946	24137	.627608839

Number of obs	= 24138
F(4, 24133)	= 2038.00
Prob > F	= 0.0000
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

- El salario de un hombre es, en promedio, aproximadamente un 7% (0.0712 x 100%) mayor que el de una mujer de la misma edad y con la misma educación.
- Vamos a ver que **esta interpretación no siempre es correcta para los coeficientes de las dummies**.

¿Cómo se interpreta el coeficiente de la variable binaria *hombre* en un modelo semilogarítmico?

- Para hombres tenemos: $ln\hat{W}_H = \hat{\beta}_1 + \hat{\beta}_2 aedu + \hat{\beta}_3 edad + \hat{\beta}_4 edad^2 + \hat{\delta}$
- Para mujeres tenemos: $ln\hat{W}_M = \hat{\beta}_1 + \hat{\beta}_2 aedu + \hat{\beta}_3 edad + \hat{\beta}_4 edad^2$
- Restando miembro a miembro:

$$ln\hat{W}_H - ln\hat{W}_M = ln\left[\frac{\hat{W}_H}{\hat{W}_M}\right] = \hat{\delta} \Longrightarrow \left[\frac{\hat{W}_H}{\hat{W}_M}\right] - 1 = \left[\frac{\hat{W}_H - \hat{W}_M}{\hat{W}_M}\right] = exp\left(\hat{\delta}\right) - 1$$

Entonces, la interpretación correcta es que el salario horario estimado para un hombre es $\left[exp\left(\hat{\delta}\right) - 1\right] \times 100\%$ mayor que el de una mujer con la misma edad y educación.

Source	SS	df	MS
Model	3825.0433	4	956.260826
Residual	11323.5513	24133	.469214406
Total	15148.5946	24137	.627608839

Number of obs	= 24138
F(4, 24133)	= 2038.00
Prob > F	= 0.0000
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

- Entonces, deberíamos esperar que un hombre gane por hora $[exp\,(0.07) - 1] \times 100\% = 7.25\%$ más que una mujer con la misma edad y educación.
- Resultado muy similar al que obteníamos antes mirando directamente el coeficiente estimado $\hat{\delta}$, porque para $\hat{\delta}$ pequeño (digamos, $\hat{\delta} < 0.20$), $exp\left(\hat{\delta}\right) - 1 \cong \hat{\delta}$.
- Si $\hat{\delta}$ es grande, también lo es diferencia entre $\hat{\delta}$ y $exp\left(\hat{\delta}\right) - 1$.

Agregamos una interacción entre educación y género

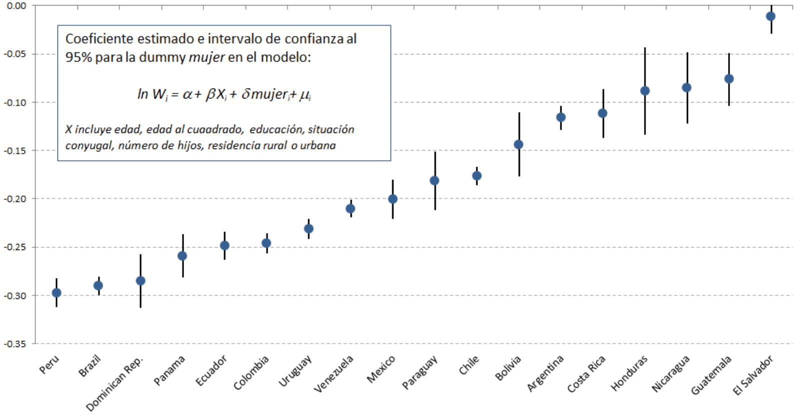
$$ln(W_i) = \beta_1 + \beta_2 aedu_i + \beta_3 edad_i + \beta_4 edad_i^2 + \delta hombre_i$$

$$+\gamma\,(hombre \times aedu)_i + \mu_i$$

- ¿Cuál es el retorno a la educación bajo esta especificación?
  $\rightarrow$ para mujeres $\beta_2 \times 100\%$
  $\rightarrow$ para hombres $(\beta_2 + \gamma) \times 100\%$

- ¿Cómo evaluarían la hipótesis de que el retorno a la educación no difiere entre hombres y mujeres?

Brechas salariales de género en América Latina



Fuente: Marchionni, Gasparini & Edo (2019), estimaciones basadas en encuestas de hogares de 2015.
https://scioteca.caf.com/handle/123456789/1401

Bibliografía para esta clase

- Notas de clase, capítulo 3 secciones 3.2, 3.3 y Apéndice del capítulo 3
- Wooldridge, capítulo 3 sección 3.1, capítulo 6 sección 6.2, capítulo 7 secciones 7.1 a 7.4.
- Marchionni. M., L. Gasparini y M. Edo (2019). Brechas de género en América Latina, un estado de situación. Caracas: CAF. Disponible en http://scioteca.caf.com/handle/123456789/1401.