Econometría I

Variables Cualitativas

Carlos Yanes | Departamento de Economía | 2024-05-05

😅

Preguntas de la sesion anterior?


sky

Modelos de Regresión Múltiple: (k) Parámetros

- Recordemos que:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_{k-1} x_{(k-1)i} + \mu_i$$

- Se tiene que:

» Un número de k parámetros desconocidos que depende del número de variables de control.

» $k - 1$ regresores.

Modelos de Regresión Múltiple: K Parámetros

Recuerde que de lo que tenemos estadísticamente, podemos expresarlo de forma **matricial**:

- Una medida de variabilidad de la variable *dependiente* es:

$$y'y = \sum_{i=1}^{n} y_i^2$$

- De la descomposición **ortogonal** de (Y) podemos tener:

$$\begin{aligned} y'y =& (\hat{y} + u)'(\hat{y} + u) \\ =& \hat{y}'\hat{y} + 2\hat{y}'\widehat{u} + \widehat{u}'\widehat{u} \\ =& \hat{y}'\hat{y} + \widehat{u}'\widehat{u} \end{aligned}$$

- Lo que también puede ser establecido como:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \widehat{u}_i^2$$

Modelos de Regresión Múltiple: K Parámetros

Si de la expresión anterior restamos su media con un vector de unos (1) llamado **f** de tamaño $n \times 1$ entonces encontramos:

$$(y - \mathbf{f}\bar{y})'(y - \mathbf{f}\bar{y}) = (\hat{y} - \mathbf{f}\bar{y})'(\hat{y} - \mathbf{f}\bar{y}) + \widehat{u}'\widehat{u}$$

- De tal forma que es lo mismo que:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSE} + \underbrace{\sum_{i=1}^{n}\widehat{u}_i^2}_{SSR}$$

Los términos de la expresión **SST** hacen referencia a la suma total al cuadrado (fuente de variación principal), **SSE** es la suma al cuadrado del modelo o *suma explicada* y por último, **SSR** la suma de los residuos al cuadrado de nuestro modelo.

Modelos de Regresión Múltiple: K Parámetros

De lo anterior, para tener al **coeficiente de determinación**, la formula mas usada es:

$$R^2 = 1 - \frac{SSR}{SST}$$

Note que $SSR = \sum_i (y_i - \hat{y}_i)^2 = \sum_i e_i^2$ y que $SST = \sum_i (y_i - \bar{y}_i)^2$

Para lo cual se debe **pensar** que:

$$\text{Si} \quad SSR \downarrow \Rightarrow R^2 \uparrow$$

> A medida que se *adicionan* nuevas variables al modelo de regresión, automaticamente el **R2** aumenta.

- Al tener ese problema entonces, hay que entrar a solucionarlo.

Nuevamente.. El $R^2 - Ajustado$

UNIVERSIDAD DEL NORTE

Modelos de Regresión Múltiple: R^2 ajustado

✓ Una **solución** es **penalizar** el número de variables, *p.e* con el R^2 **ajustado**.

$$\overline{R}^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2/(n-k-1)}{\sum_i (y_i - \overline{y})^2/(n-1)}$$

Nota: R^2 Ajustado necesariamente no esta entre 0 y 1.

Mucho cuidado!!

Hay que tener en cuenta las ventajas y desventajas de añadir o quitar variables:

Menos variables

- Generalmente explican menos variación en y
- Proporcionan interpretaciones y visualizaciones sencillas (*parsimoniosas*)
- Puede ser necesario preocuparse por el sesgo de las variables omitidas

Modelos de Regresión Múltiple: R^2 ajustado

Más variables

- Es más probable que se encuentren relaciones *espúreas* (estadísticamente significativas debido a la casualidad; no reflejan una verdadera relación a nivel de la población)
- Es más difícil interpretar el modelo
- Es posible que se pasen por alto variables importantes: sigue habiendo un sesgo de variables omitidas

Modelos de Regresión Múltiple: R^2 ajustado

Tomemos los datos de *autos* :

Tabla #1: Regresión Simple					
	Estimate	Standard Error	t value	Pr(>ltl)	
(Intercept)	11,253.061	1,170.813	9.611	0.0000	***
mpg	-238.894	53.077	-4.501	0.0000	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 2624 on 72 degrees of freedom

Multiple R-squared: 0.2196, Adjusted R-squared: 0.2087

F-statistic: 20.26 on 72 and 1 DF, p-value: 0.0000

Modelos de Regresión Múltiple: R^2 ajustado

Con dos variables:

Tabla #2: Regresión Múltiple

	Estimate	Standard Error	t value	Pr(>|t|)	
(Intercept)	5,864.305	5,888.103	0.996	0.3227	
mpg	-173.702	87.723	-1.980	0.0516	.
length	21.286	22.793	0.934	0.3535	

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 2626 on 71 degrees of freedom

Multiple R-squared: 0.2291, Adjusted R-squared: 0.2073

F-statistic: 10.55 on 71 and 2 DF, p-value: 0.0001

Modelos de Regresión Múltiple: R^2 ajustado

Con tres variables:

Tabla #3: Regresión Múltiple con mas k

	Estimate	Standard Error	t value	Pr(>ltl)	
(Intercept)	14,542.434	5,890.632	2.469	0.0160	*
mpg	-86.789	83.943	-1.034	0.3047	
length	-104.868	39.722	-2.640	0.0102	*
weight	4.365	1.167	3.739	0.0004	***

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 2415 on 70 degrees of freedom

Multiple R-squared: 0.3574, Adjusted R-squared: 0.3298

F-statistic: 12.98 on 70 and 3 DF, p-value: 0.0000

Modelos de Regresión Múltiple: R^2 ajustado

- Nuestro modelo en la medida que hemos incorporado nuevas variables ha cambiado la métrica del R^2.

Variables	Coeficiente R2	Variación R	R-Ajustado	Variación Ajustado
1	0.2196		0.2087	
2	0.2291	4%	0.2073	-0.6%
3	0.3574	56%	0.3298	59%

> **Proporción** de la *variación* muestral de Y que es explicada por las X_i, o también se puede definir como la variación en (Y) que es explicada por las variables **independientes** pero con *castigo* en los grados de libertad por incluir esas nuevas variables.

- Otra forma de establecer la **formula** del R^2 ajustado es:

$$\overline{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

Sesgo por variable omitida

Sesgo por variable omitida

El sesgo de la variable omitida (SVO) surge cuando omitimos una variable que

1. Afecta a nuestra variable de resultado y
2. Se correlaciona con una variable explicativa x_j.

Como su nombre indica, esta situación provoca un sesgo en nuestra estimación de β_j.

Ejemplo

Imagine un modelo de regresión con varios individuos i de su nivel de ingresos

$$\text{Pago}_i = \beta_0 + \beta_1 \text{Escolaridad}_i + \beta_2 \text{Genero}_i + u_i$$

Donde

- Escolaridad_i son los años aprobados y cursados por el individuo i.
- Genero_i una variable *indicador* (dummy) del genero de i haciendo referencia a si este es masculino.

Sesgo por variable omitida

Entonces

- β_1: Es el retorno económico por cada año de educación que tiene el individuo (*ceteris paribus*)
- β_2: la prima por ser hombre (*ceteris paribus*)
  Si $\beta_2 > 0$ o $\beta_2 < 0$, entonces existe una **discriminación** contra las mujeres (hombres): ya que alguno(a)s reciben menos salario por razón de su género.

Para nuestro modelo poblacional

$$\text{Pago}_i = \beta_0 + \beta_1 \text{Escolaridad}_i + \beta_2 \text{Genero}_i + u_i$$

Si nos concentramos en la estimación solo de **Escolaridad**, es decir, omitimos la variable de genero, el modelo ahora será *p.e.*,

$$\text{Pago}_i = \beta_0 + \beta_1 \text{Escolaridad}_i + (\beta_2 \text{Genero}_i + u_i)$$

Ahora vamos a tener

$$\text{Pago}_i = \beta_0 + \beta_1 \text{Escolaridad}_i + \varepsilon_i$$

Donde $\varepsilon_i = \beta_2 \text{ Genero}_i + u_i$.

Sesgo por variable omitida

La condición de **exogeneidad** ya no se cumple. pero incluso si

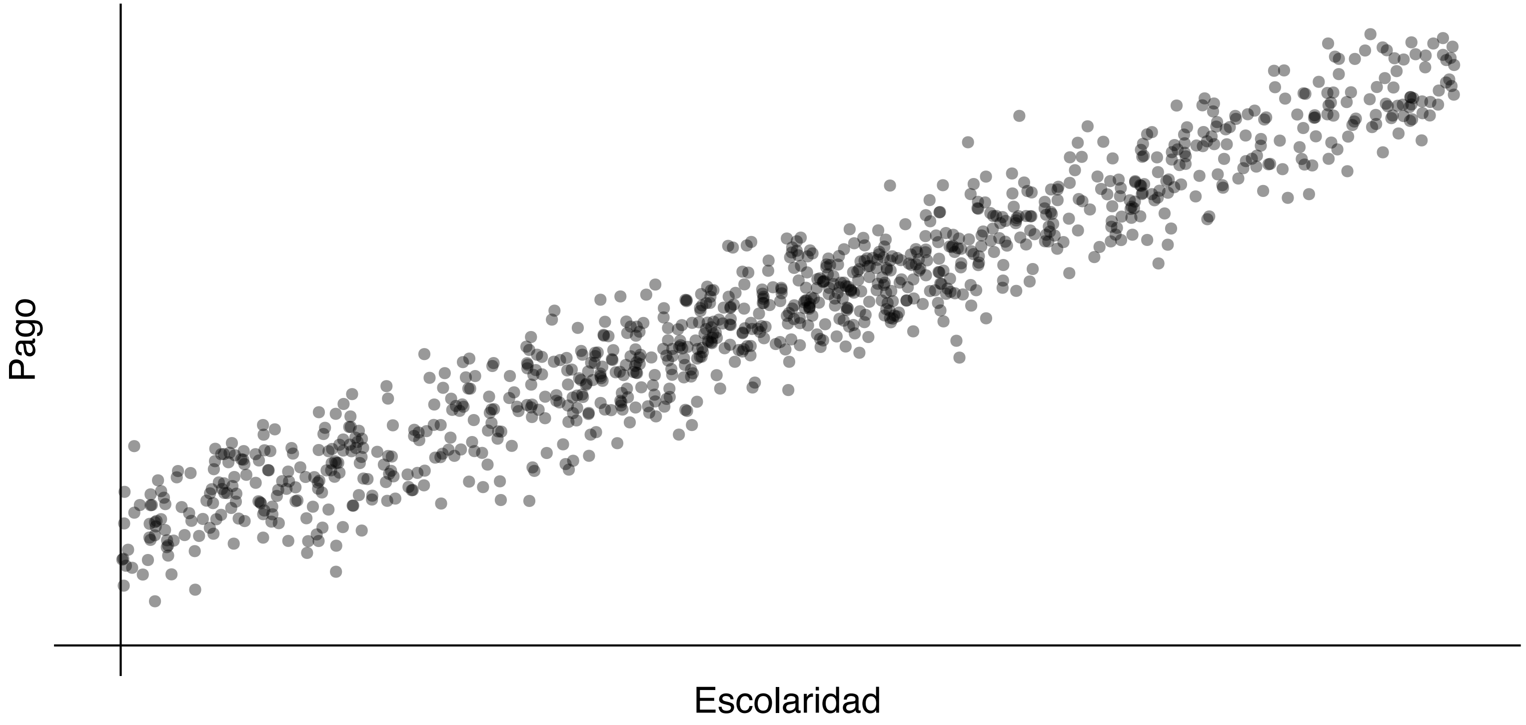
$\boldsymbol{E}[u|X] = 0$, no es cierto que $\boldsymbol{E}[\varepsilon|X] = 0$ de tal forma que $\beta_2 \neq 0$.

Esto es, $\boldsymbol{E}[\varepsilon|\text{genero} = 1] = \beta_2 + \boldsymbol{E}[u|\text{genero} = 1] \neq 0$.

Ahora entonces MCO es Sesgado.

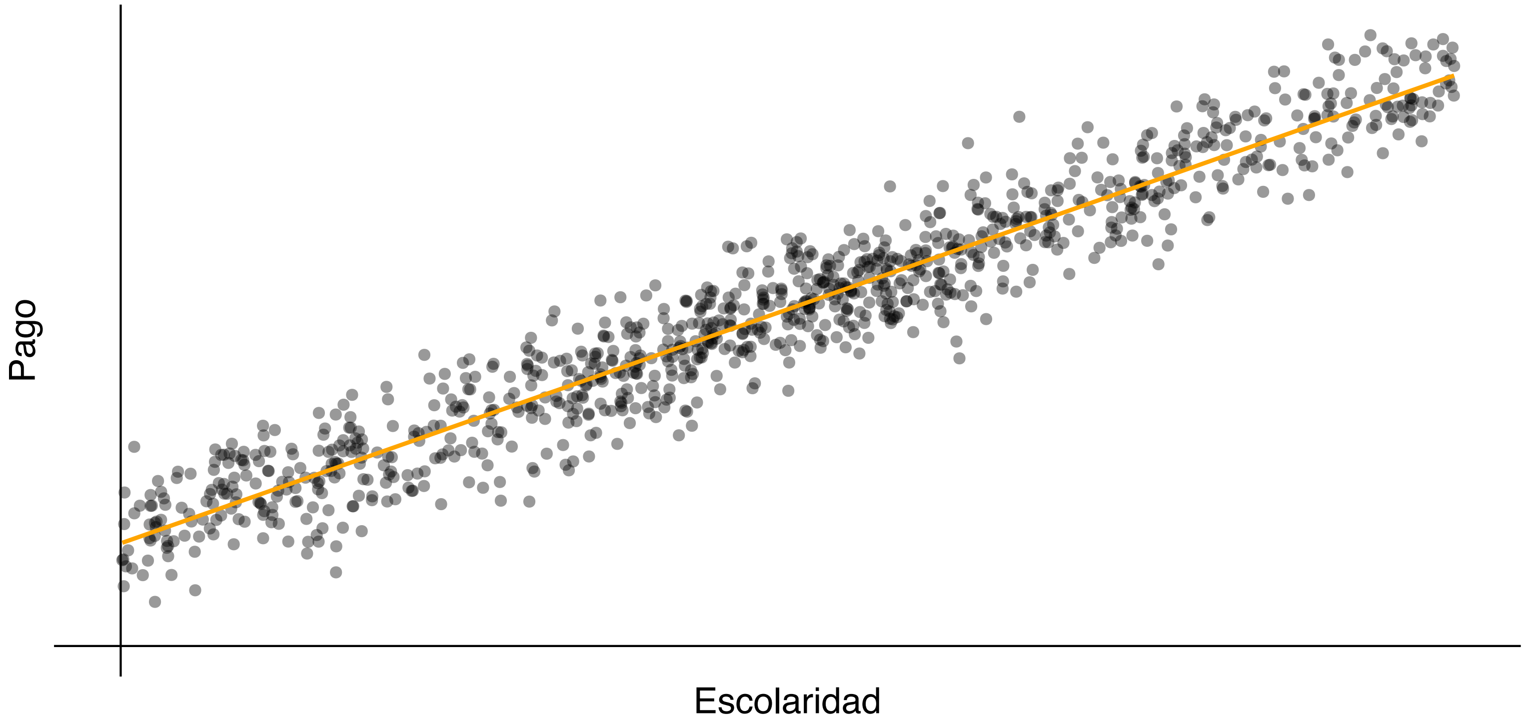
Sesgo por variable omitida

Veamos un gráfico:



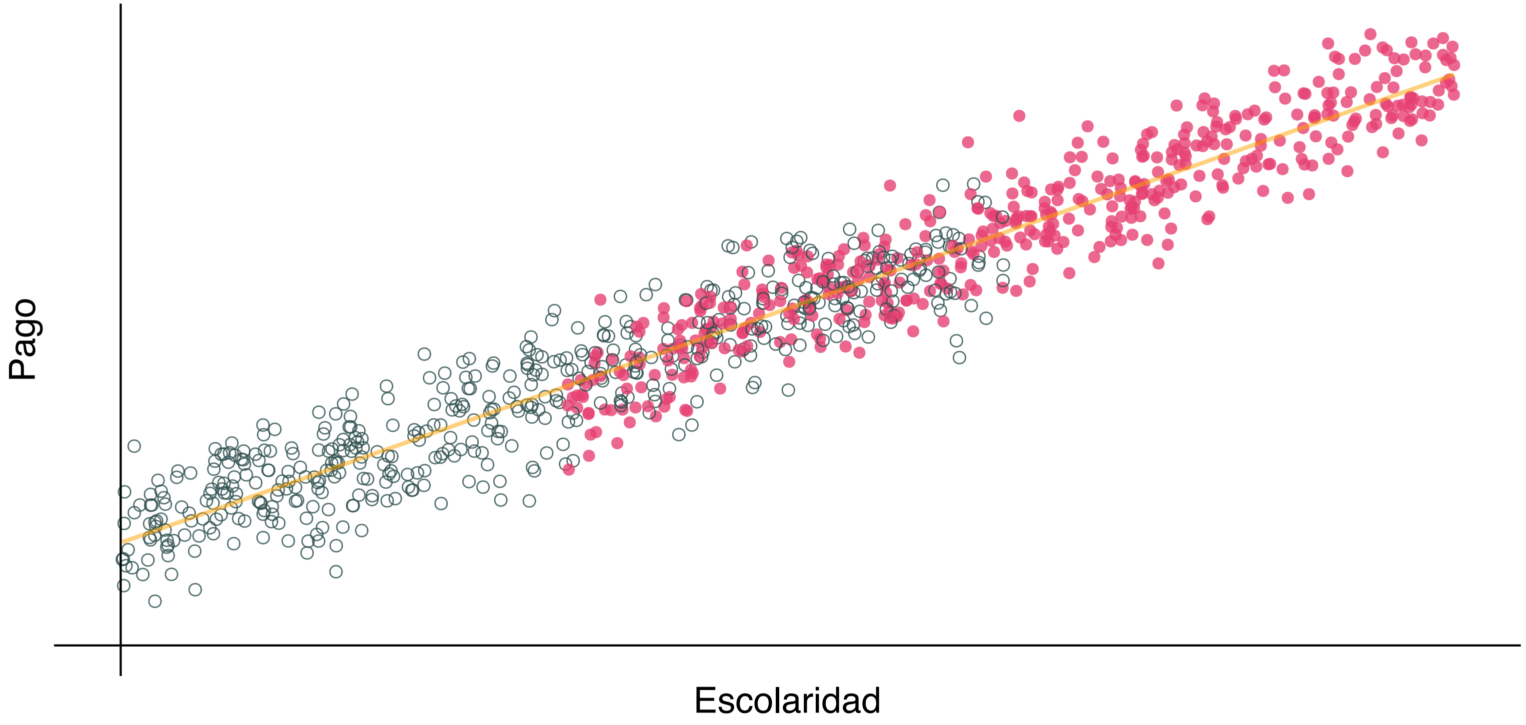
Sesgo por variable omitida

Tenemos un modelo: $\text{Pago}_i = 20.5 + 10.4 \times \text{Escolaridad}_i$



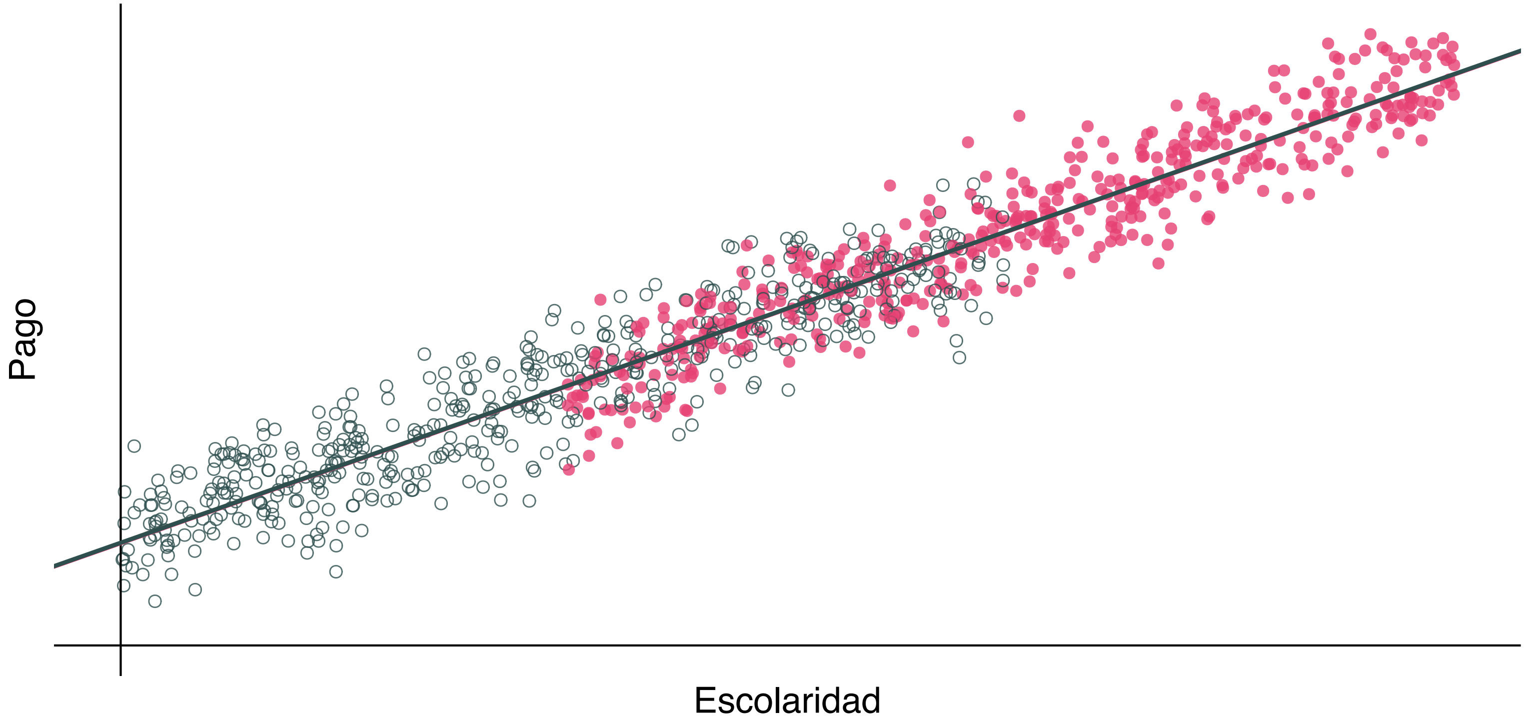
Sesgo por variable omitida

Variable omitida: Genero (**Masculino** y **Femenino**)



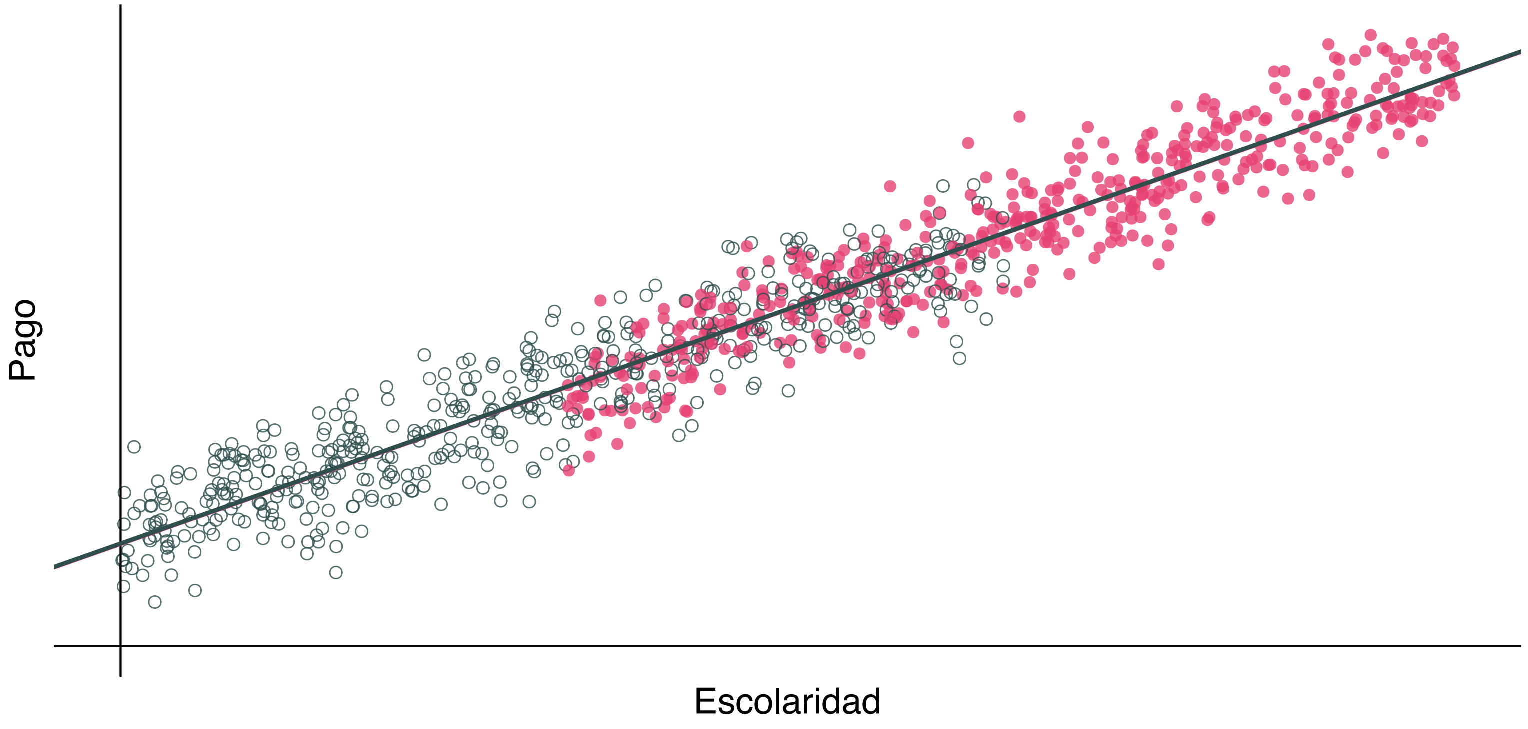
Sesgo por variable omitida

Variable omitida: Genero (**Masculino** y **Femenino**)

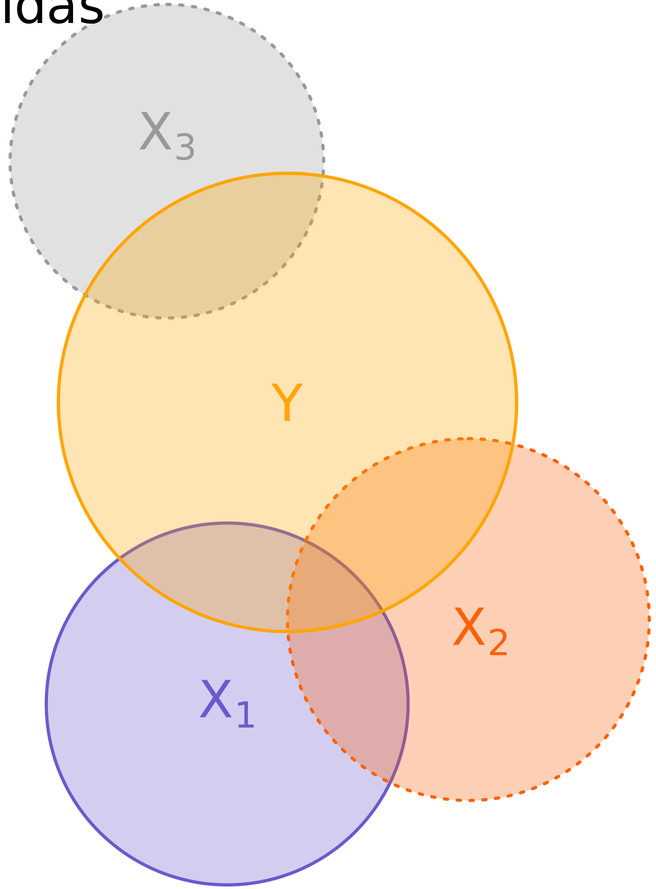


Sesgo por variable omitida

Regresión insesgada: $\text{Pago}_i = 20.4 + 10.4 \times \text{Escolaridad}_i + 0.1 \times \text{Genero}_i$



Variables Omitidas



Sesgo por variable omitida

Sesgo por variable omitida

Como corregirlo

1. No **omita variables**
2. Instrumentalizando o por MCO (dos etapas)

Modelos de Regresión Múltiple: variables irrelevantes

Se tiene:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \mu$$

- x_3 no tiene efecto sobre y una vez se controló por x_1 y x_2, es decir, $\beta_3 = 0$.
- Puede o no estar **correlacionado** x_3 con x_1 o x_2.
- $E[y|x_1, x_2, x_3] = E[y|x_1, x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- No sabemos que $\beta_3 = 0$, por lo que estimamos:

$$\hat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \mu$$

- **Sobre-especificar** el modelo, no afecta el **insesgamiento** de los estimadores de MCO.

Variables Cualitativas

Variables Cualitativas

- **Dummy** Son variables que son catalogadas como *ficticias* y su manera de abordar se hace a partir de una variable *binaria*.
- Generalmente son variables de **escala nominal** o característica.

$$D = \begin{cases} 1 & \text{si la característica está presente} \\ 0 & \text{si la característica no está presente} \end{cases}$$

- Debe escogerse un grupo **base** con la cual se hacen *comparaciones* (elección puede ser arbitraria).

Variables Cualitativas

Obs	Ingreso	Educación	Experiencia	Masculino
1	3.15	11	12	0
2	2.92	12	3	1
3	5.4	16	5	0
$\vdots$	6.00	14	7	0
324	11.2	15	6	1
325	15.3	16	12	0

Variables Cualitativas

> Son de tipo **categórico** p.e: (sexo, estrato, sector, religión, raza, etc). En los modelos debemos *transformarlas* en variables *binarias* o de tipo (0,1)

- Considere por ejemplo la siguiente estimación:

$$Ingreso_i = \beta_0 + \beta_1 \text{ Masculino}_i + \mu_i$$

Donde

› $Ingreso_i$ es una variable *continua* que mide el nivel de ingreso (pago recibido) de un individuo cualquiera.

› $Masculino_i$ es una variable *cualitativa* que mide define el sexo o genero de un individuo cualquiera.

> La interpretación de β_1 es la diferencia esperada entre hombres y mujeres en el ingreso. El parámetro β_0 es el ingreso promedio de las mujeres $(Masculino_i = 0)$ y la parte de $\beta_0 + \beta_1$ es el ingreso promedio de las **hombres**

Variables Cualitativas

Derivación

$$
\begin{aligned}
\boldsymbol{E}[\text{Ingreso}|\text{Femenino}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 0 + u_i] \\
&= \boldsymbol{E}[\beta_0 + 0 + u_i] \\
&= \beta_0
\end{aligned}
$$

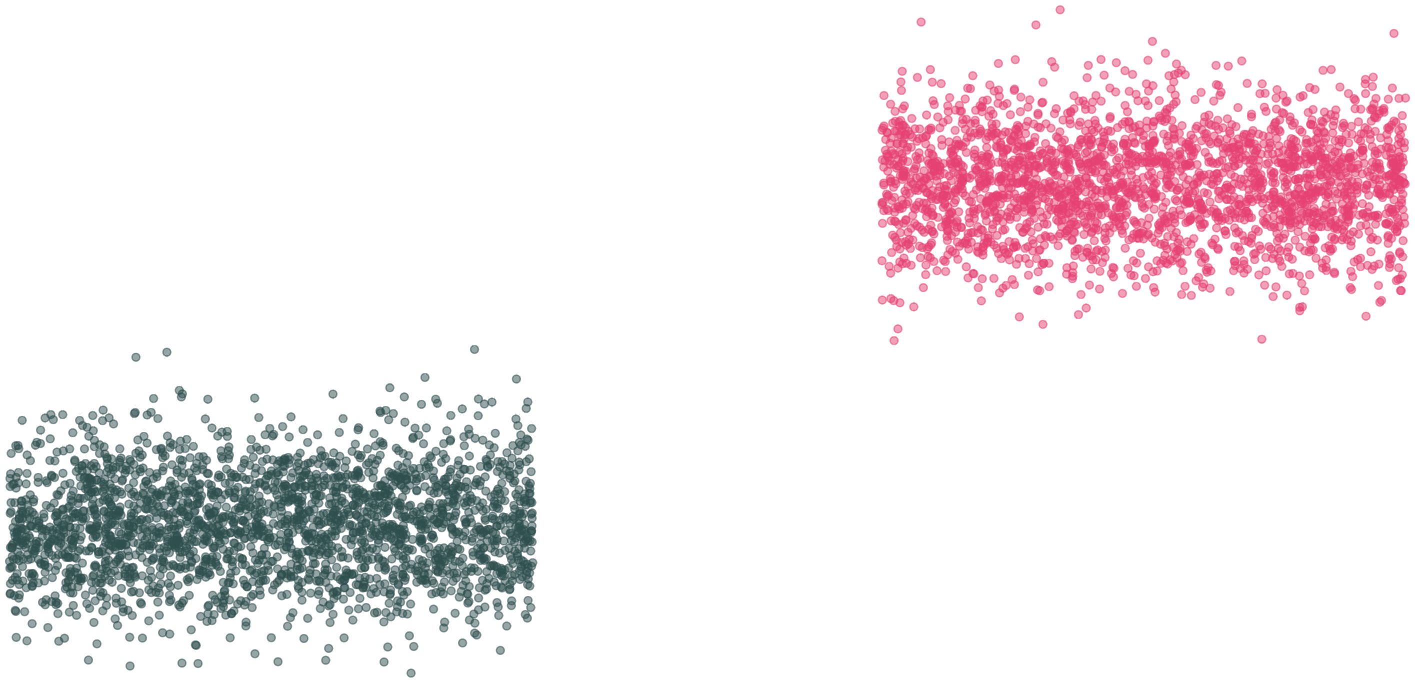
$$
\begin{aligned}
\boldsymbol{E}[\text{Ingreso}|\text{Masculino}] &= \boldsymbol{E}[\beta_0 + \beta_1 \times 1 + u_i] \\
&= \boldsymbol{E}[\beta_0 + \beta_1 + u_i] \\
&= \beta_0 + \beta_1
\end{aligned}
$$

Nota: Si no hay mas *variables* explicativas o controles, entonces $\hat{\beta}_1$ es igual a la diferencia de medias, *p.e.*, $\overline{x}_{\text{Masculino}} - \overline{x}_{\text{Femenino}}$.

Nota$_2$: El supuesto de *mantener todo lo demas constante* se aplica de igual manera para los modelos de regresión múltiple.

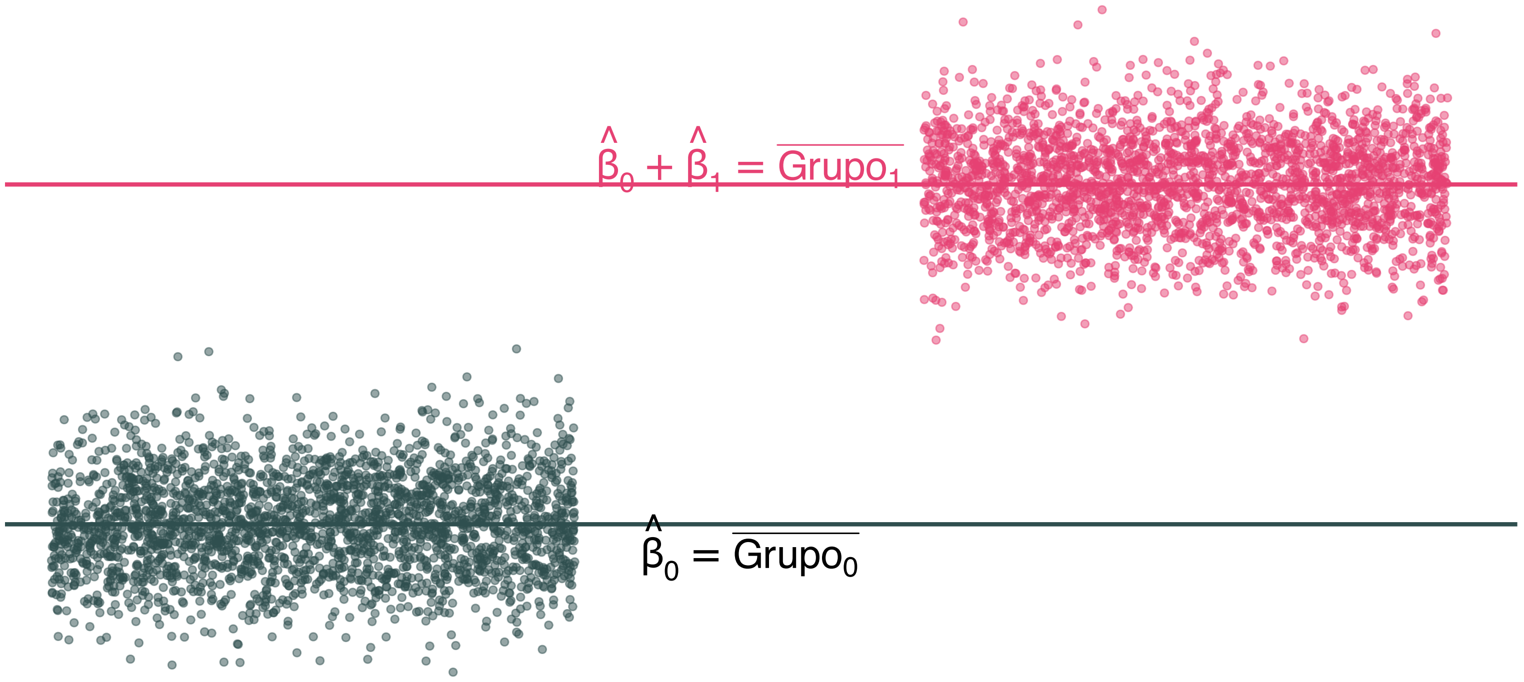
Variables Cualitativas

$y_i = \beta_0 + \beta_1 x_i + u_i$ para variable binaria $Masculino_i$ o $x_i = \{0, 1\}$



Variables Cualitativas

$y_i = \beta_0 + \beta_1 x_i + u_i$ para variable binaria $Masculino_i$ o $x_i = \{0, 1\}$



Variables Cualitativas : mas categorías

Si tuviéramos una variable *ordinal* con 3 **categorías**

- El grupo *base* es el primero por "default" *se omite por multicolinealidad* .

$$D_2 = \begin{cases} 1 & \text{para grupo 2} \\ 0 & \text{en otro caso} \end{cases} \qquad D_3 = \begin{cases} 1 & \text{para grupo 3} \\ 0 & \text{en otro caso} \end{cases}$$

- Lo que tendríamos a modo de ecuaciones:

$$\begin{aligned} y &= \beta_0 + (\beta_2 - \beta_0)\, D_2 + (\beta_3 - \beta_0)\, D_3 + \beta_i x_i + \mu \\ &= \beta_0 + \alpha_2 D_2 + \alpha_3 D_3 + \beta_i x_i + \mu \end{aligned}$$

Hay que mirar que todas son *diferencias* α_i *p.e* (α_2, α_3) son los mismos parámetros de la regresión, solo que son diferencias con respecto al grupo base.

Variables Cualitativas en R

```
#>
#> Call:
#> lm(formula = y ~ x, data = base_1)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -606.10 -249.92  -11.71  296.38  704.54
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   4916.9      161.7  30.401 2.93e-16 ***
#> xB             233.9      220.4   1.061    0.303
#> xC            -119.9      220.4  -0.544    0.593
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 396.2 on 17 degrees of freedom
#> Multiple R-squared:  0.1448,	Adjusted R-squared:  0.04422
#> F-statistic:  1.44 on 2 and 17 DF,  p-value: 0.2645
```

- **R** y muchos softwares automáticamente generan las dummies múltiple
- La variable x hace referencia a los tipos de A, B y C respectivamente.
- Note que **A** es el grupo base o de referencia, es decir, β_0 es el *promedio* de esa variable.
- Los parámetros xB y xC son en efecto β_2 y β_3 que son las diferencias con respecto al grupo de **referencia** o base.
- La significancia es utíl para decir si existe o no diferencias entre grupos.

Variables Cualitativas en R

```
#>
#> Call:
#> lm(formula = y ~ 0 + x, data = base_1)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -606.10 -249.92  -11.71  296.38  704.54
#>
#> Coefficients:
#>    Estimate Std. Error t value Pr(>|t|)
#> xA   4916.9      161.7   30.40 2.93e-16 ***
#> xB   5150.8      149.7   34.40  < 2e-16 ***
#> xC   4797.0      149.7   32.04  < 2e-16 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 396.2 on 17 degrees of freedom
#> Multiple R-squared:  0.9946,	Adjusted R-squared:  0.9937
#> F-statistic:  1045 on 3 and 17 DF,  p-value: < 2.2e-16
```

- Cuando se da la opción sin intercepto, todas las características de la variable **cualitativa** muestran sus respectivos promedios.

- Esta opción solo se usa mas como información que como objetivo final de la estimación de la regresión.

- El *asunto* de *omitir* una de las características de la variable **categórica** es para evitar caer en la trampa de *dummies* y entonces tener el problema de multicolinealidad.

Variables Cualitativas en R

```
#>
#> Call:
#> lm(formula = y ~ x, data = base_1)
#>
#> Residuals:
#>     Min      1Q  Median      3Q     Max
#> -606.10 -249.92  -11.71  296.38  704.54
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)   5150.8      149.7  34.399   <2e-16 ***
#> xA            -233.9      220.4  -1.061    0.303
#> xC            -353.9      211.8  -1.671    0.113
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 396.2 on 17 degrees of freedom
#> Multiple R-squared:  0.1448,	Adjusted R-squared:  0.04422
#> F-statistic:  1.44 on 2 and 17 DF,  p-value: 0.2645
```

- En **R** para cambiar el grupo de referencia solo es usar la opción `relevel`, *p.e* : `datos$x<- relevel(datos$x,"B")`.
- Recuerde que la variable categórica o cualitativa debe ser clasificada *primero* como `factor` para poder interactuar con ella *p.e* : `datos$x<-as.factor(datos$x)`. Si ya su variable de entrada es clasificada como factor, se puede omitir esta parte.
- Las diferencias ahora son con base a la opción **B**.

🔵 Y de las interacciones?

Variables Cualitativas con interacciones

Las interacciones permiten que el efecto de una variable cambie en función del nivel de otra variable.

Preguntas

1. ¿Cambia el efecto de la escolarización sobre el salario en función del genero (sexo)?
2. ¿Cambia el efecto del género en el salario según la raza?
3. ¿Cambia el efecto de la escolaridad en el salario según la experiencia?

Variables Cualitativas con interacciones

- Dos *variables* cualitativas: sexo, estado civil.
- Grupo **base**: hombre soltero.

$$Mujer = \begin{cases} 1 & \text{si es mujer} \\ 0 & \text{si es hombre} \end{cases} \qquad Casado = \begin{cases} 1 & \text{si está casado} \\ 0 & \text{si está soltero} \end{cases}$$

- La base de datos puede ser:

Obs	Ingreso	Genero (Fem=1)	E. Civil (Cas=1)	Interacción
1	3.15	0	1	0
2	2.92	1	0	0
3	5.4	0	1	0
$\vdots$	6.00	0	0	0
324	11.2	1	1	1
325	15.3	0	0	0

Variables Cualitativas con interacciones

Suponga que tiene el siguiente modelo:

$$y = \beta_0 + \beta_1 \ femenino + \beta_2 casado + \beta_3 \ femenino \times casado + \beta_i x_i + \mu_i$$

$$\begin{aligned}
E(y_i|x_i, mujer = 0, \ casado = 0) &= \beta_0 + \beta x_i \\
E(y_i|x_i, mujer = 0, \ casado = 1) &= \beta_0 + \beta_2 + \beta x_i \\
E(y_i|x_i, mujer = 1, \ casado = 0) &= \beta_0 + \beta_1 + \beta x_i \\
E(y_i|x_i, mujer = 1, \ casado = 1) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta x_i
\end{aligned}$$

Donde β_1 es el *efecto diferencial de ser mujer*; β_2 es el efecto diferencial de ser casado y β_3 es el **efecto diferencial de ser mujer casada**. Puede probarse si *diferencial* en sexo (estado civil) depende del estado civil (sexo). $H_0 : \beta_3 = 0$.

Variables Cualitativas con interacciones

Tomemos ahora otro ejemplo pero haciendo **interacción** con una variable *cuantitativa*.

Mediremos ahora en un nuevo modelo: el salario, el genero y añadiremos la escolaridad o número de años de estudio de la persona.

$$Salario_i = \beta_0 + \beta_1\ Femenino_i + \beta_2\ Escolaridad_i + \mu_i$$

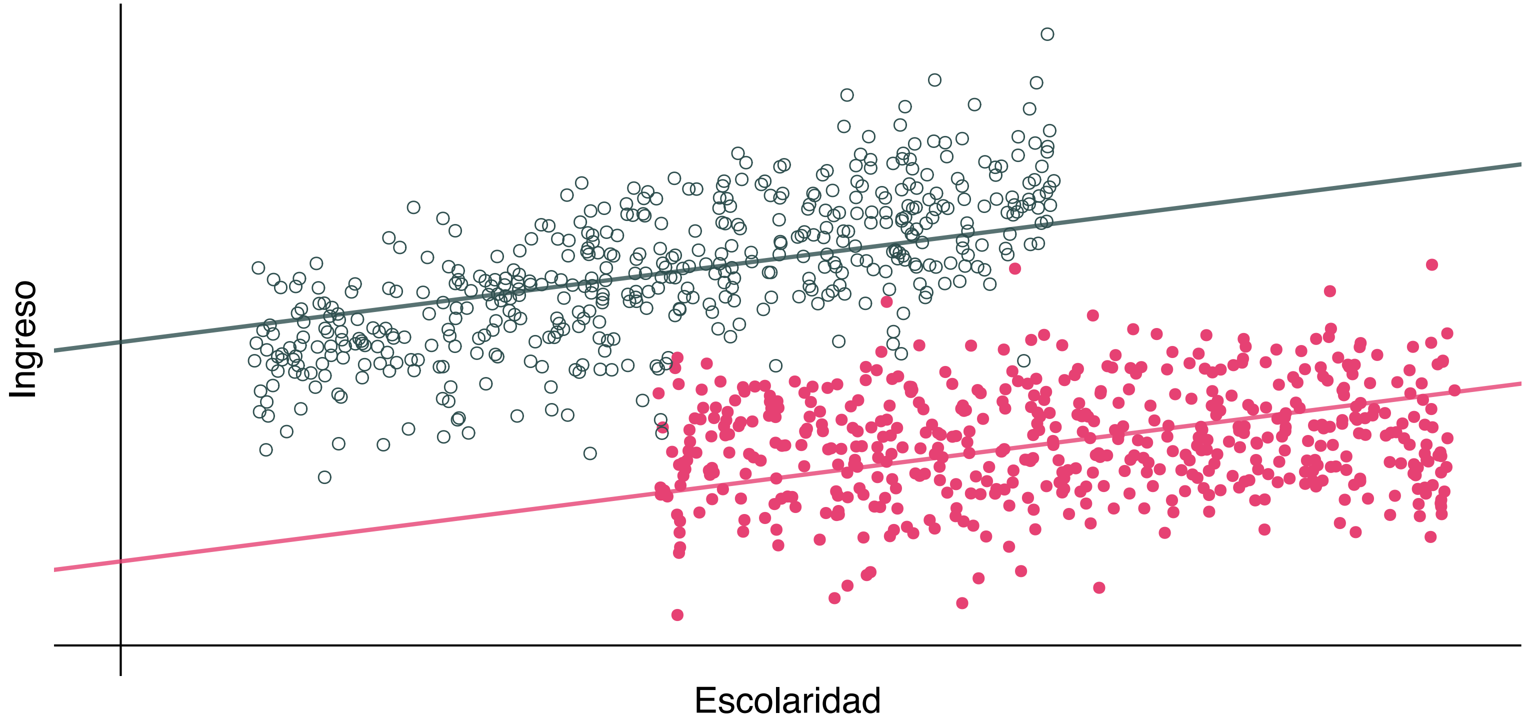
En el anterior se mira la escolaridad de igual forma o manera para todos.

- Al añadir un término de **interacción**, se hace con el objeto de ver como varia la escolaridad por genero o grupo. El modelo entonces es:

$$Salario_i = \beta_0 + \beta_1\ Femenino_i + \beta_2\ Escolaridad_i + \beta_3\ Femenino_i \times Escolaridad_i + \mu_i$$

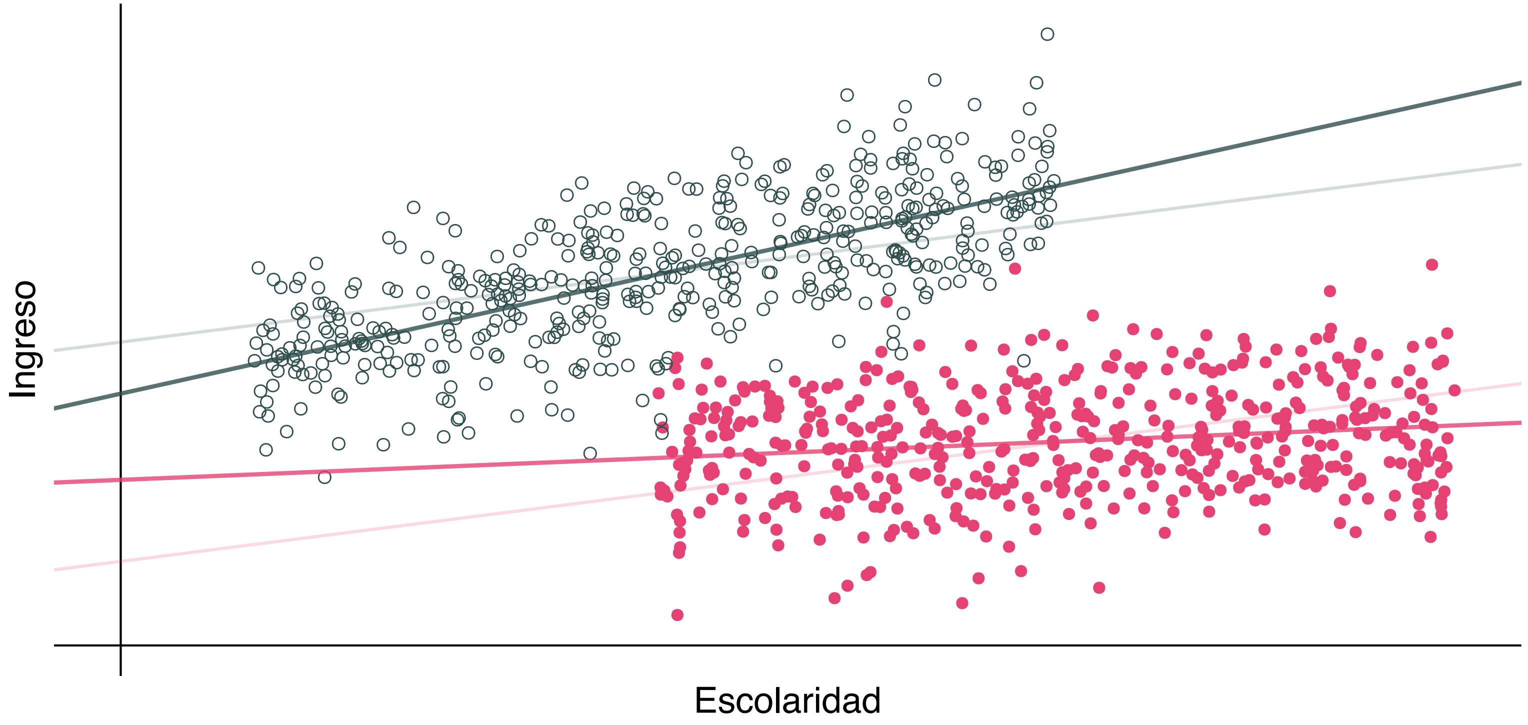
Variables Cualitativas con interacciones

La *escolaridad* tiene el mismo efecto para todos (**F**) y para (**M**)



Variables Cualitativas con interacciones

La *escolaridad* tiene distinto efecto para los grupos de (**F**) y (**M**)



Variables Cualitativas con interacciones

> La interpretación del *efecto de la interacción* puede ser un poco complejo, pero la clave* es entender la parte matematica.

$$\text{Ingreso}_i = \beta_0 + \beta_1 \,\text{Femenino}_i + \beta_2 \,\text{Escolaridad}_i + \beta_3 \,\text{Femenino}_i \times \text{Escolaridad}_i + u_i$$

Rendimiento esperado de un año adicional de escolarización para las mujeres:

$$\begin{aligned} \boldsymbol{E}[\text{Ingreso}_i | \text{Femenino} \wedge \text{Escolaridad} = \phi + 1] - \boldsymbol{E}[\text{Ingreso}_i | \text{Femenino} \wedge \text{Escolaridad} = \phi] = \\ \boldsymbol{E}[\beta_0 + \beta_1(\phi + 1) + \beta_2 + \beta_3(\phi + 1) + u_i] - \boldsymbol{E}[\beta_0 + \beta_1\phi + \beta_2 + \beta_3\phi + u_i] = \\ \beta_1 + \beta_3 \end{aligned}$$

Del mismo modo, β_1 da el rendimiento esperado de un año adicional de escolarización para los hombres. Así, β_3 da la **diferencia en los rendimientos de la escolarización** para mujeres y hombres.

* Como suele ocurrir con la econometría.

Bibliografía

- Gujarati, D. N., & Porter, D. C. (2011). *Econometria Básica*. Ed. Porto Alegre: AMGH..
- Stock, J. H., Watson, M. W., & Larrión, R. S. (2012). *Introducción a la Econometría*.
- Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Cengage learning.

Gracias por su atención!

Alguna pregunta adicional?

Carlos Andres Yanes Guerra

cayanes@uninorte.edu.co

keynes37