SEMINARIO DE DOCTORADO

METODOLOGÍA DE INVESTIGACIÓN SOCIAL

Agustín Salvia
Santiago Poy

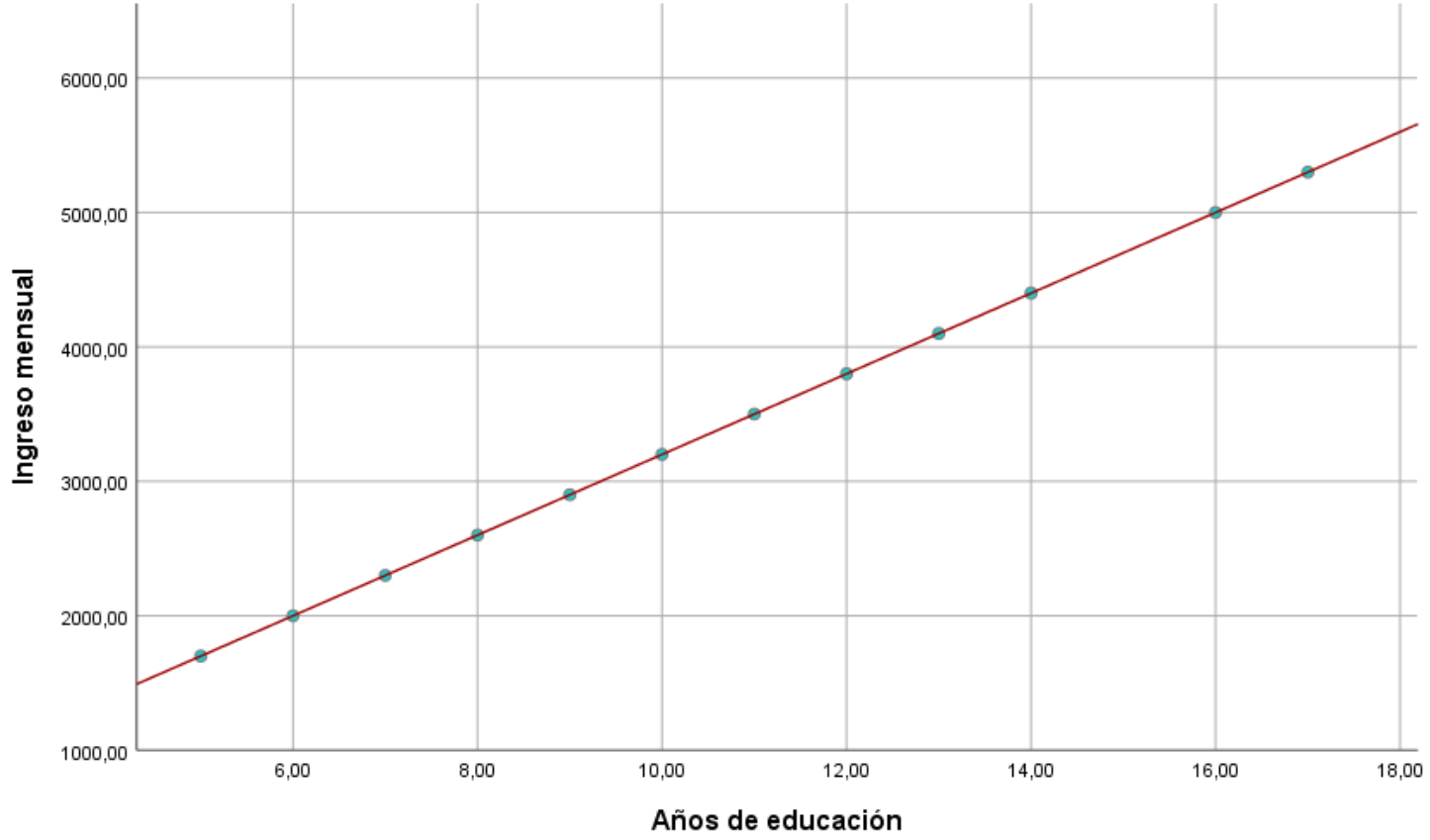
APLICACIÓN

4 REGRESIÓN LINEAL

Ejercicio 1: *análisis de regresión lineal simple*

Correlación y regresión lineal

- Llamamos ***recta de regresión*** a la recta que mejor se ajusta a la distribución conjunta de las dos variables, es decir, al diagrama de dispersión.



El análisis de regresión lineal

- Por lo tanto, cada valor de y_i, *para cada* observación es la suma del valor en la recta de regresión para cada x_i más el **error** que se comete en la predicción:

$$y_i = a + bx_i + e_i$$

- Donde e_i expresa un ***residuo***, es decir, la diferencia entre el valor pronosticado para el caso *i* por la ecuación de la recta y el valor observado.

- Se trata de buscar la recta de regresión que mejor se ajuste a la nube de puntos, es decir, la que minimiza las distancias de todos los puntos en relación a esta recta. Para ello se utiliza el ***método de mínimos cuadrados ordinarios***.

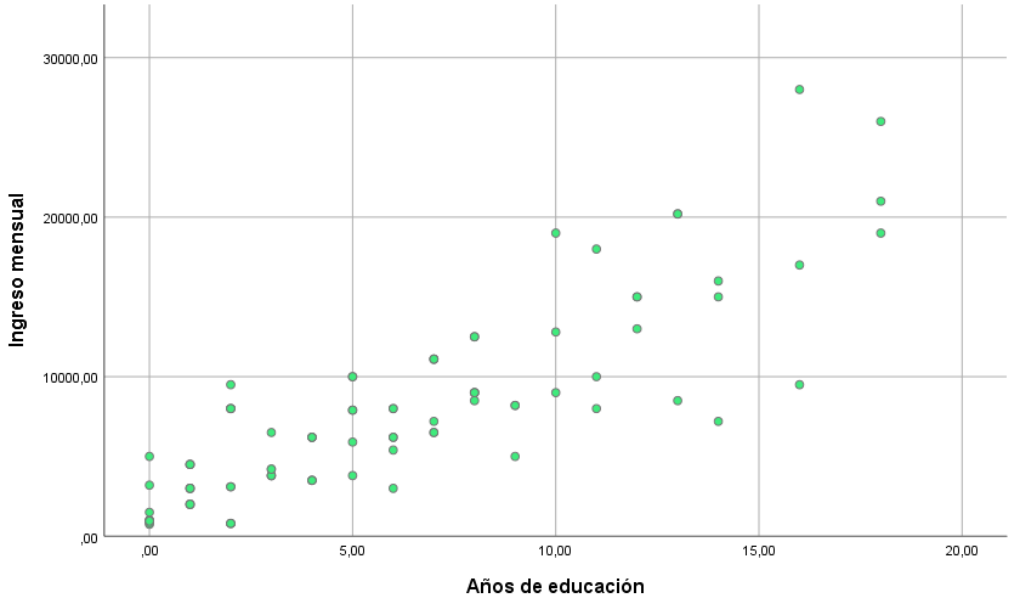
- Cuanto mayor (menor) sea la distancia entre la recta estimada y los datos observados, peor (mejor) será la ***bondad de ajuste*** del modelo de regresión.

- Cuando incorporamos más variables independientes, pasamos del análisis de regresión lineal simple al análisis de regresión lineal múltiple:

Supuestos que debe cumplir el análisis de regresión lineal

- Algunas características y ***supuestos*** del análisis de regresión lineal múltiple son los siguientes:

1) Linealidad: la relación entre X e Y debe ser lineal, lo que podemos comprobar mediante un ***diagrama de dispersión***.



Posible solución: En caso de no haber linealidad, evaluar transformar la variable dependiente en su logaritmo.

Supuestos que debe cumplir el análisis de regresión lineal

2) Distribución normal de los residuos: la distribución de los errores estandarizados debe ser normal.

Para comprobarlo, podemos utilizar dos herramientas:

- La prueba de Kolmogorff-Smirnov
- Gráficos de normalidad de tipo Q-Q (cuantiles) o P-P(proporciones)

De acuerdo con López-Roldán & Fachelli (2016), cierto incumplimiento de la normalidad no es problemática en muestras con más de 1000 observaciones.

Posible solución: eliminación de datos outliers.

Supuestos que debe cumplir el análisis de regresión lineal

3) Ausencia de autocorrelación de los errores: en el modelo de regresión lineal, se asume que los errores (es decir, los residuos) son independientes entre sí. Esto supone que los errores no siguen un patrón establecido.

La forma de evaluar la existencia de autocorrelación es mediante la prueba de ***Durbin-Watson***.

El estadístico toma valores entre 0 y 4. Valores comprendidos entre 1,5 y 2,5 indican no autocorrelación. Los inferiores indicarían autocorrelación positiva y los superiores autocorrelación negativa.

Posible solución: eliminación de datos.

Supuestos que debe cumplir el análisis de regresión lineal

- Algunas características y ***supuestos*** del análisis de regresión lineal múltiple son los siguientes:

  4) Homoscedasticidad: la varianza de los errores debe ser la misma para cada valor de la variable independiente.

  Se observa el gráfico de dispersión entre el residuo estandarizado y el valor pronosticado estandarizado.

  El gráfico relaciona ZPRED (pronósticos tipificados) y ZRESID (residuos tipificados) y deberíamos observar una distribución aleatoria.

  Posible solución: Eliminación de casos outliers, transformación de las variables independientes y/o de la variable dependiente (por ejemplo, raíz cuadrada).

Supuestos que debe cumplir el análisis de regresión lineal

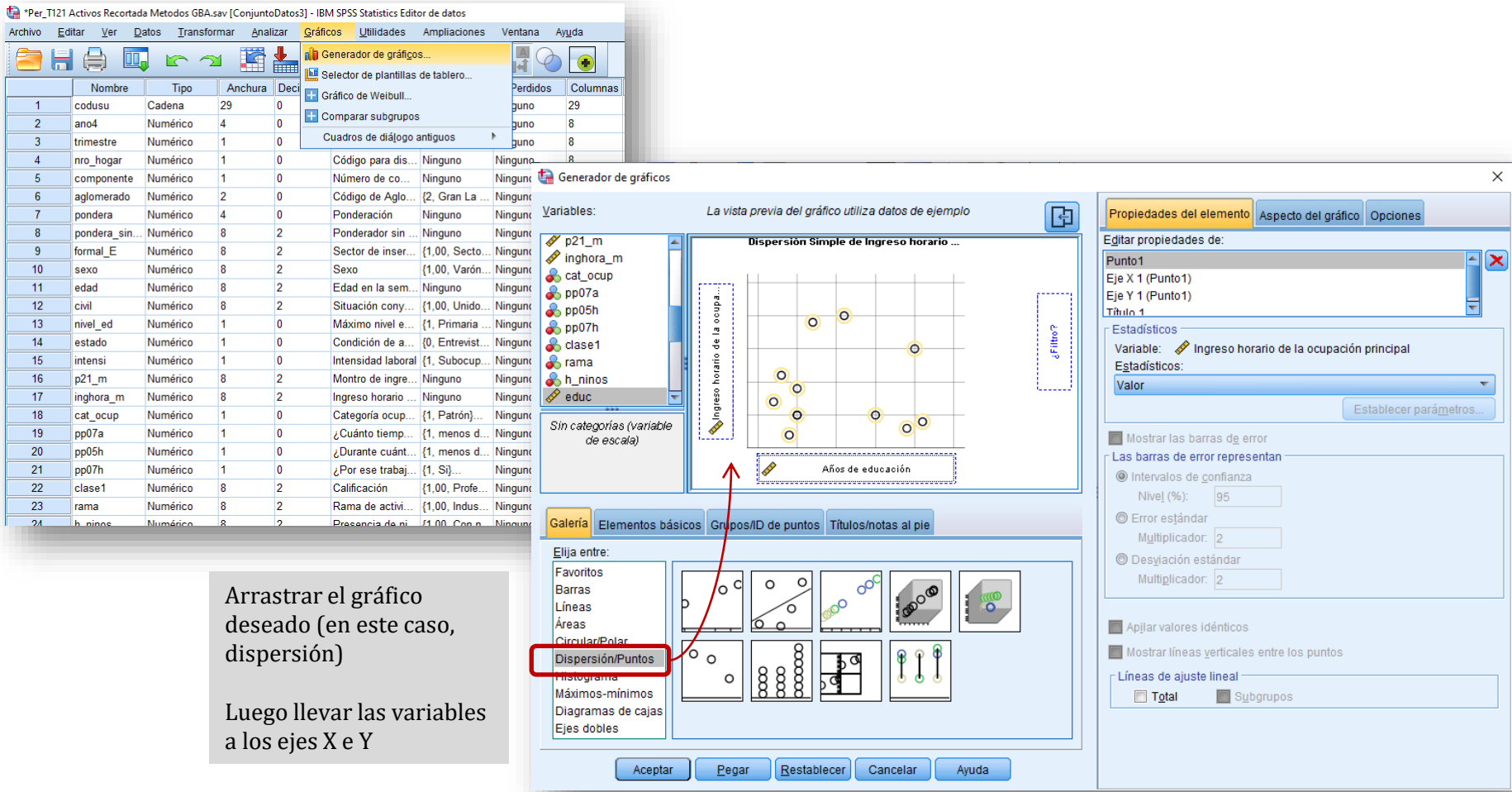
- Algunas características y ***supuestos*** del análisis de regresión lineal múltiple son los siguientes:

  5) Ausencia de colinealidad: las variables independientes no deberían estar correlacionadas entre sí.

  Posible solución: elaborar una matriz de correlaciones y descartar alguna de las variables que se encuentre altamente correlacionada con otra.

Aplicación del análisis de regresión lineal simple

- Elaboramos un diagrama de dispersión y examinamos si se cumple el supuesto de ***linealidad***:

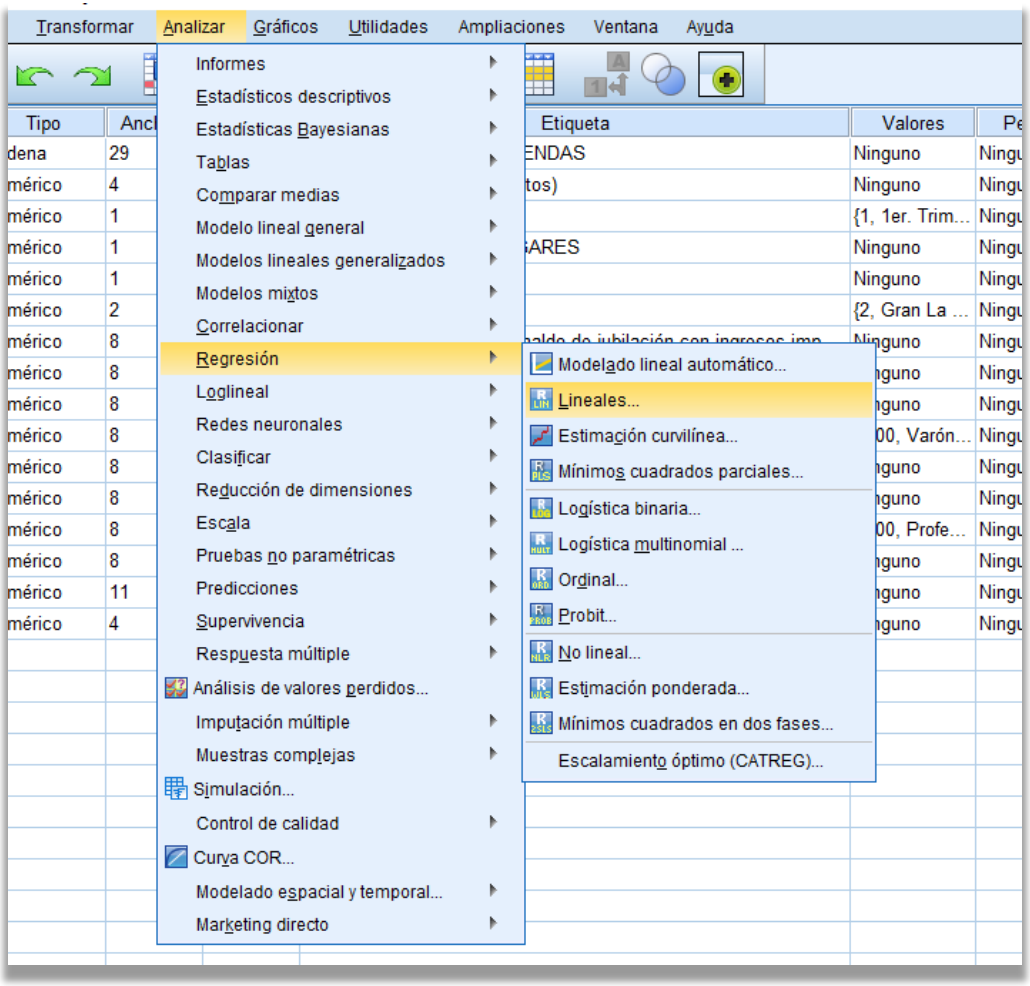


Arrastrar el gráfico deseado (en este caso, dispersión)

Luego llevar las variables a los ejes X e Y

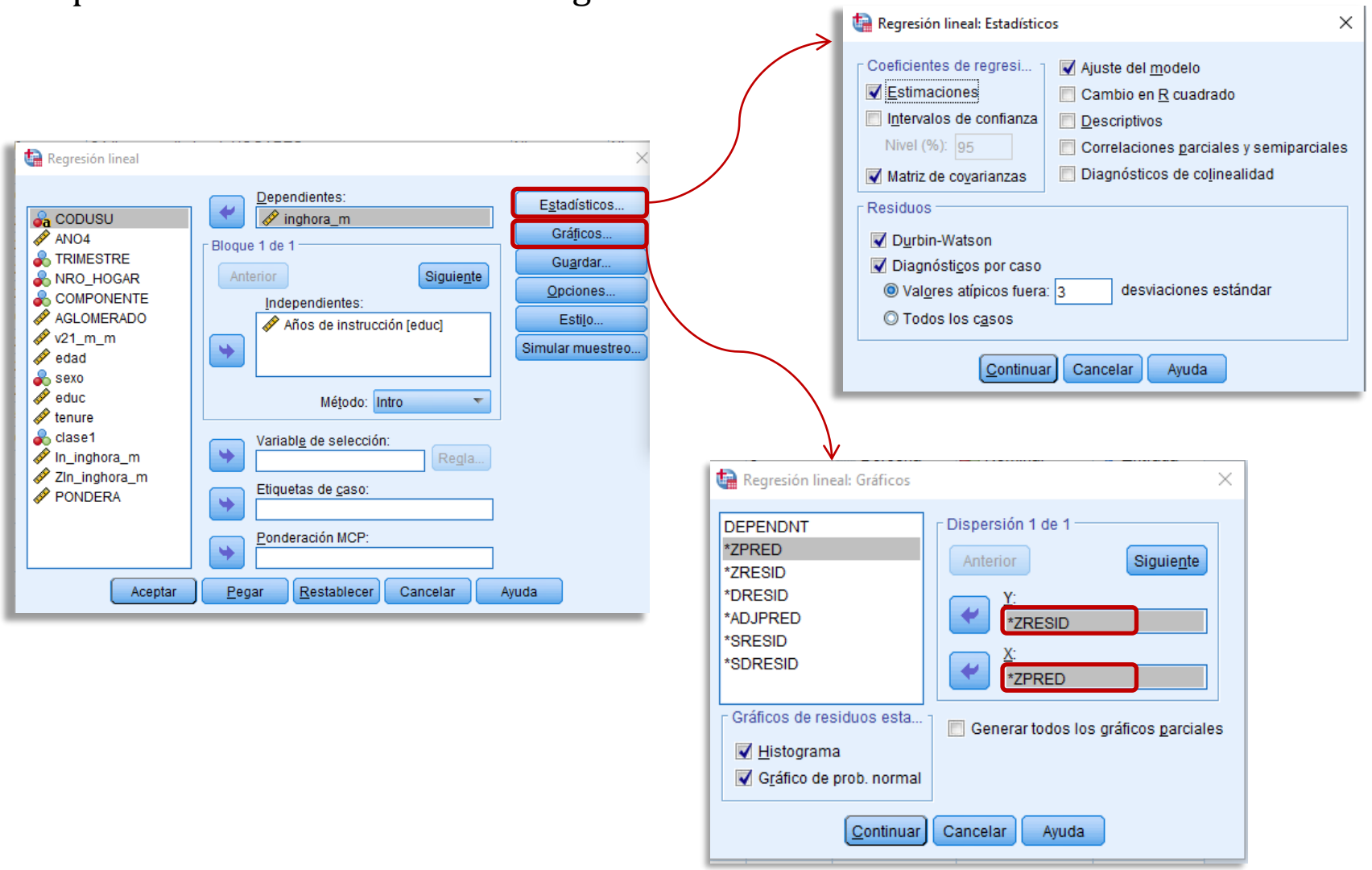
Aplicación del análisis de regresión lineal simple

- Implementamos un análisis de regresión lineal:



Aplicación del análisis de regresión lineal simple

- Implementamos un análisis de regresión lineal:



Aplicación del análisis de regresión lineal simple

Analicemos la salida:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	,372[a]	,138	,138	139,89153	1,836

a. Predictores: (Constante), educ Años de educación

b. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

R es el coeficiente de correlación y mide la asociación entre VD y VI.

La prueba de ***Durbin Watson*** debe dar un resultado de entre 1,5 y 2,5 para descartar autocorrelación.
En este caso, descartamos ***autocorrelación***.

El **R2 ajustado** "penaliza" cuando se incrementa el número de VI, por eso da igual con una sola VI.

R2 o ***coeficiente de determinación*** nos indica el % de varianza de VD que es explicado por VD

*Aquí vemos que se explica el **13,8%** de la varianza del ingreso horario*

Aplicación del análisis de regresión lineal simple

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	3362194,219	1	3362194,219	171,807	,000[b]
	Residuo	20939515,46	1070	19569,641		
	Total	24301709,67	1071			

a. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

b. Predictores: (Constante), educ Años de educación

El test de **ANOVA** se basa en la idea de que la variabilidad total de la muestra se descompone en la variabilidad explicada por la regresión y la variabilidad residual.

El test ofrece el estadístico **F** a partir del cual se contrasta la hipótesis nula de que VI y VD están incorrelacionadas. Por consiguiente, si el **p-valor** del test < que el nivel de significación elegido (0.05) podemos rechazar la hipótesis nula con un 95% de confianza y aceptar que las variables se encuentran asociadas.

Aplicación del análisis de regresión lineal simple

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	(Constante)	90,888	15,973		5,690	,000
	educ Años de educación	15,749	1,201	,372	13,108	,000

a. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

El **estadístico t** permite comprobar si la regresión es significativa. Si el *p-valor* es < al nivel crítico (0.05) podemos rechazar la hipótesis nula de no asociación y, por tanto, los parámetros son significativos.

La **constante** es la ordenada al origen. Su valor **B** indica el valor en el que la recta de regresión "corta" el eje de ordenadas (las Y).

El **coeficiente B** nos indica cuánto aumenta el ingreso por cada unidad que aumenten los años de estudio.
Significa que, *por cada año de escolaridad adicional*, el ingreso horario se incrementa **$15,8**

Es el error estándar que permite construir intervalos de confianza del parámetro

Beta es el coeficiente estandarizado. Permite hacer comparables los valores B de variables con distinta unidad de medida, cuando tenemos más de dos variables independientes.
Se calcula: $\beta_1 = b_1 \cdot ({}^{s_x}/_{s_y})$

En una regresión simple, coincide con el coeficiente de correlación.

Aplicación del análisis de regresión lineal simple

- Evaluamos los casos atípicos:

Diagnósticos por casos[a]

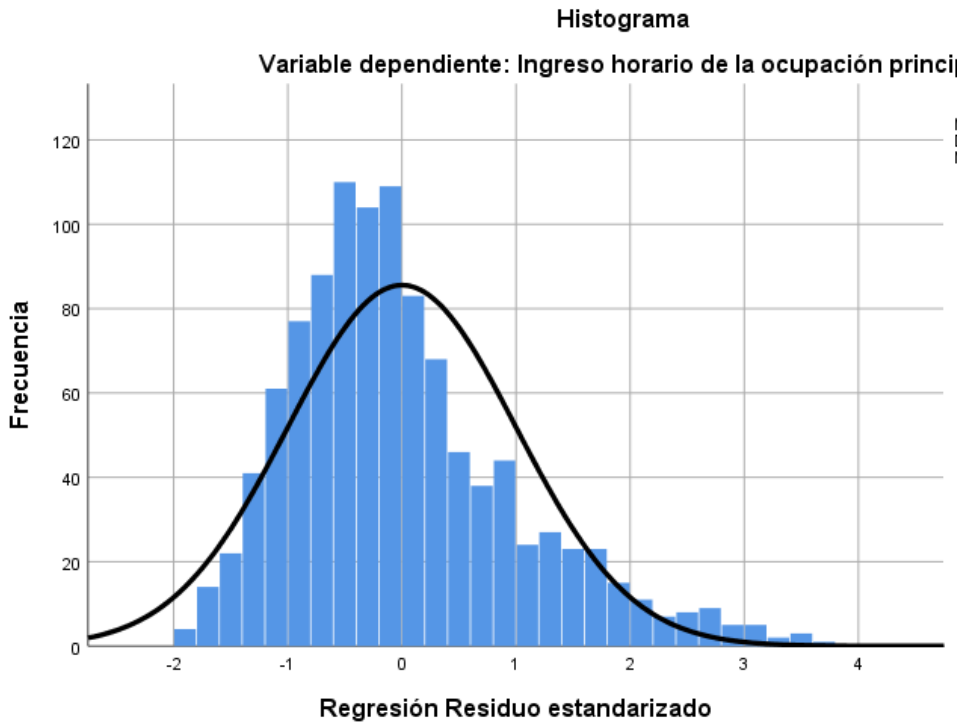
Número del caso	Desv. Residuo	inghora_m Ingreso horario de la ocupación principal	Valor pronosticado	Residuo
14	3,144	680,27	240,4995	439,77256
406	3,305	663,41	201,1280	462,28139
480	3,105	714,29	279,8711	434,41460
554	3,108	714,59	279,8711	434,71508
588	3,517	771,89	279,8711	492,02219
710	3,578	701,64	201,1280	500,51569
774	3,654	791,05	279,8711	511,18342
875	3,009	661,38	240,4995	420,87612
954	3,105	714,29	279,8711	434,41460
970	3,328	666,67	201,1280	465,53869
1171	3,468	725,68	240,4995	485,17730

a. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

Aquí vemos los casos cuyos residuos superan cierto umbral. En nuestro ejemplo, pedimos que nos indique los casos que están a **+ 3DE**

Aplicación del análisis de regresión lineal simple

- Evaluamos el supuesto de la ***distribución normal*** de los residuos:

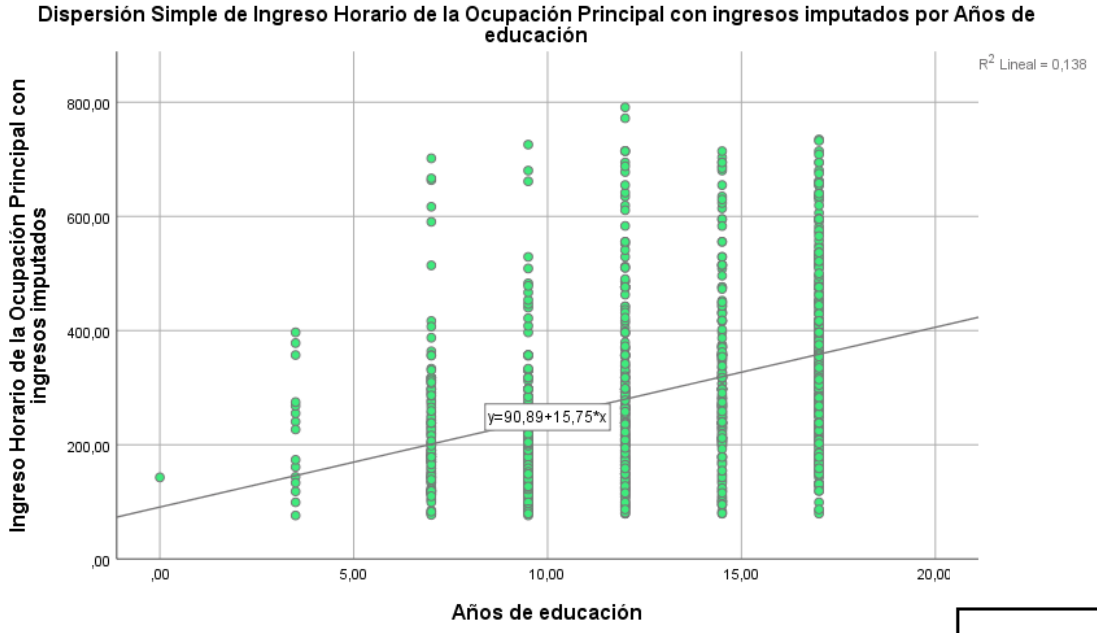


Esperamos que los residuos tengan una distribución normal.

Observamos un **sesgo** en la distribución de los residuos.

Aplicación del análisis de regresión lineal simple

- Elaboramos un diagrama de dispersión y examinamos si se cumple el supuesto de ***linealidad***:



El gráfico puede editarse haciendo doble clic en el Visor de Resultados.

Se le puede incluir la recta de regresión con el **ícono**

Estadísticos

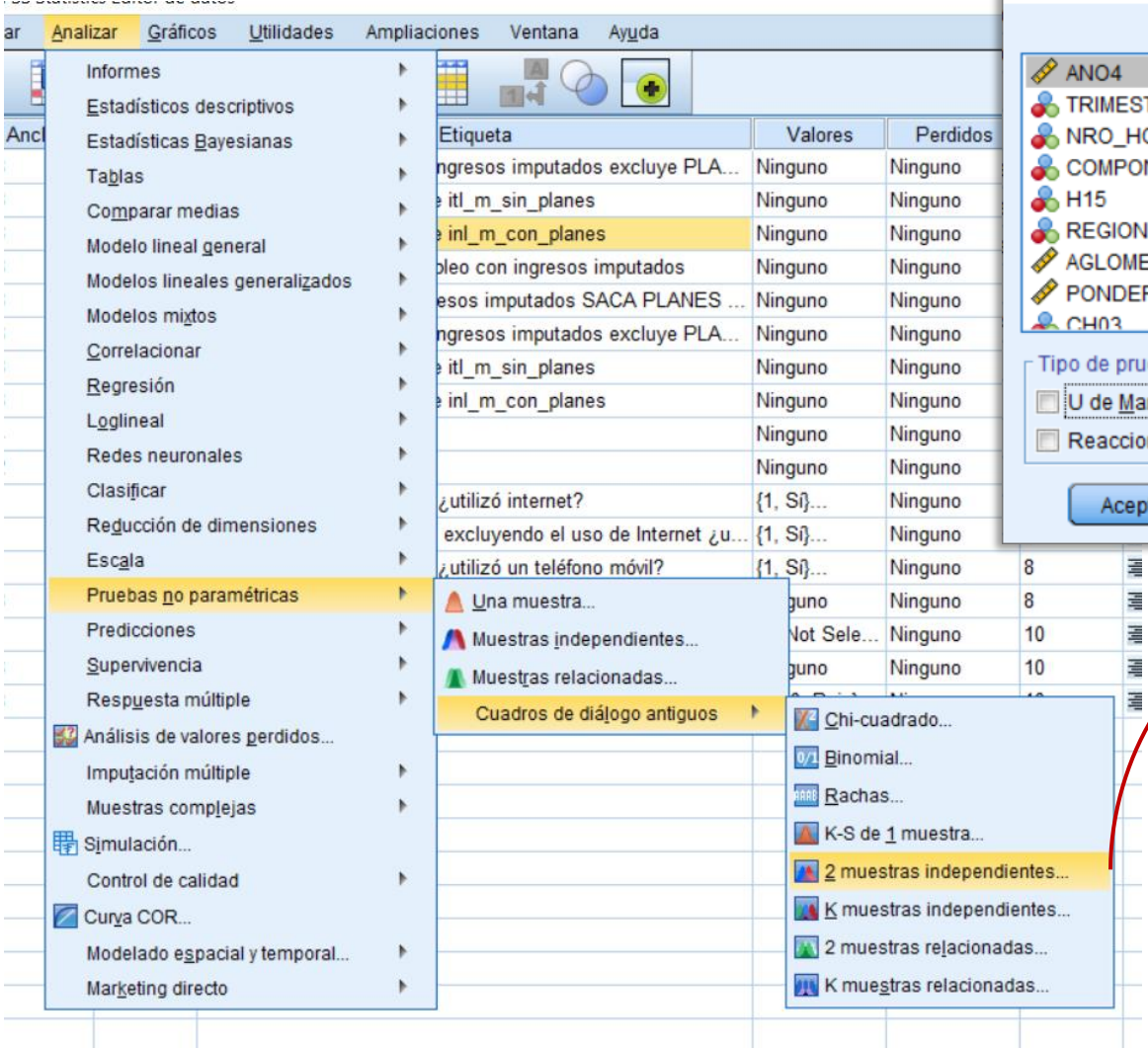
		inghora_m Ingreso horario de la ocupación principal	educ Años de educación
N	Válido	1072	1072
	Perdidos	0	0
Media		292,6228	12,8097
Desv. Desviación		150,63423	3,55774

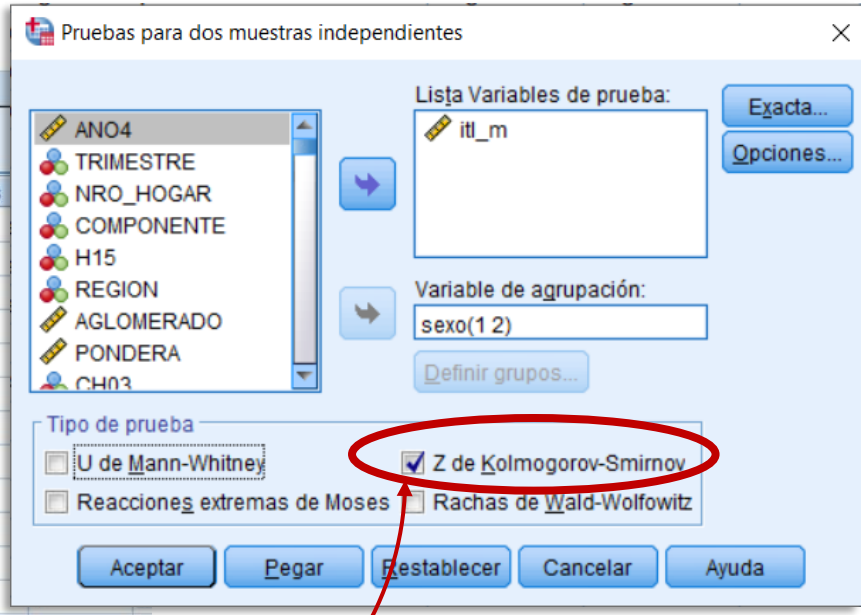
Correlaciones

		educ Años de educación	inghora_m Ingreso horario de la ocupación principal
educ Años de educación	Correlación de Pearson	1	,372**
	Sig. (bilateral)		,000
	N	1072	1072
inghora_m Ingreso horario de la ocupación principal	Correlación de Pearson	,372**	1
	Sig. (bilateral)	,000	
	N	1072	1072

**. La correlación es significativa en el nivel 0,01 (bilateral).

Prueba de Kolmogorov-Smirnov





Estadísticos de prueba[a]

		itl_m Ingreso laboral mensual
Máximas diferencias extremas	Absoluto	,151
	Positivo	,001
	Negativo	-,151
Z de Kolmogorov-Smirnov		10,229
Sig. asintótica(bilateral)		,000

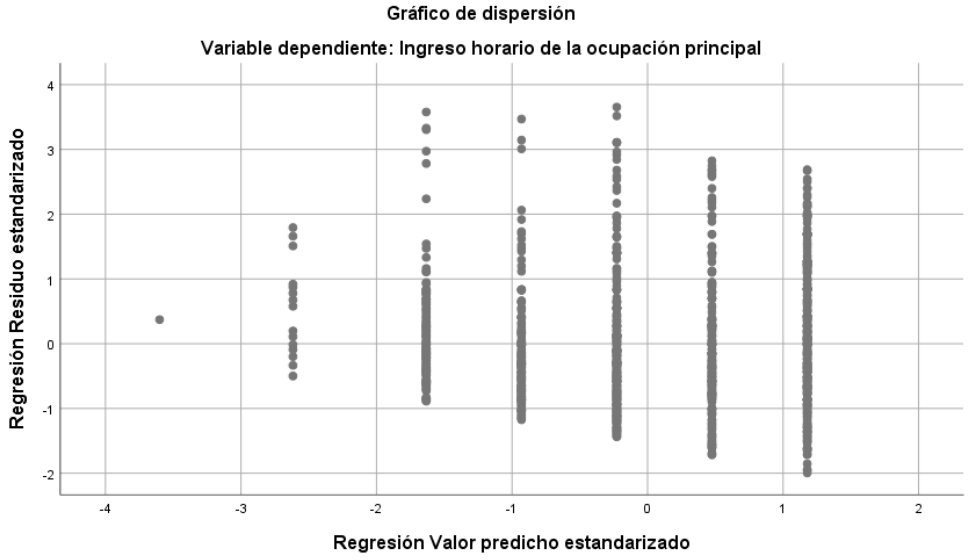
a. Variable de agrupación: sexo Sexo

No podemos concluir que haya normalidad

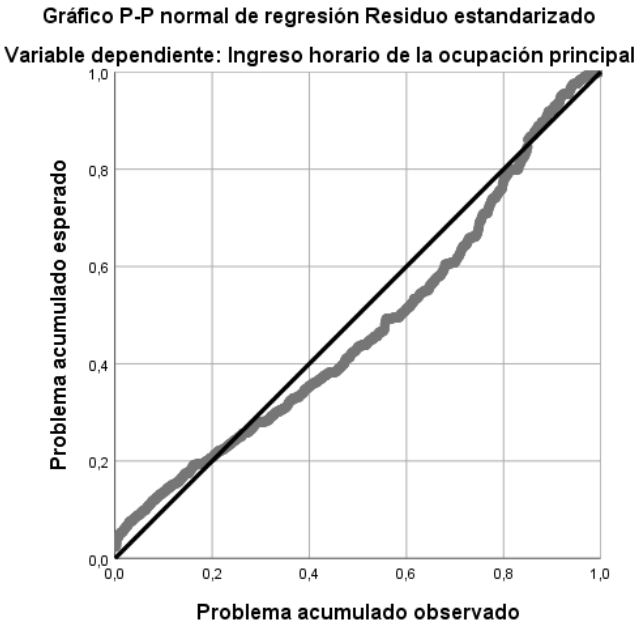
Aplicación del análisis de regresión lineal simple

- Evaluamos el cumplimiento del supuesto de ***homoscedasticidad***:

Deberíamos observar una distribución similar de las varianzas de los errores para distintos valores pronosticados. Aquí vemos cierto **alejamiento** del supuesto



La probabilidad acumulada esperada y la observada en la distribución de los residuos estandarizados deberían coincidir.



Aplicación del análisis de regresión lineal simple

- Podemos reescribir la ecuación de regresión a partir de los parámetros estimados:

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \hat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{1}}$$

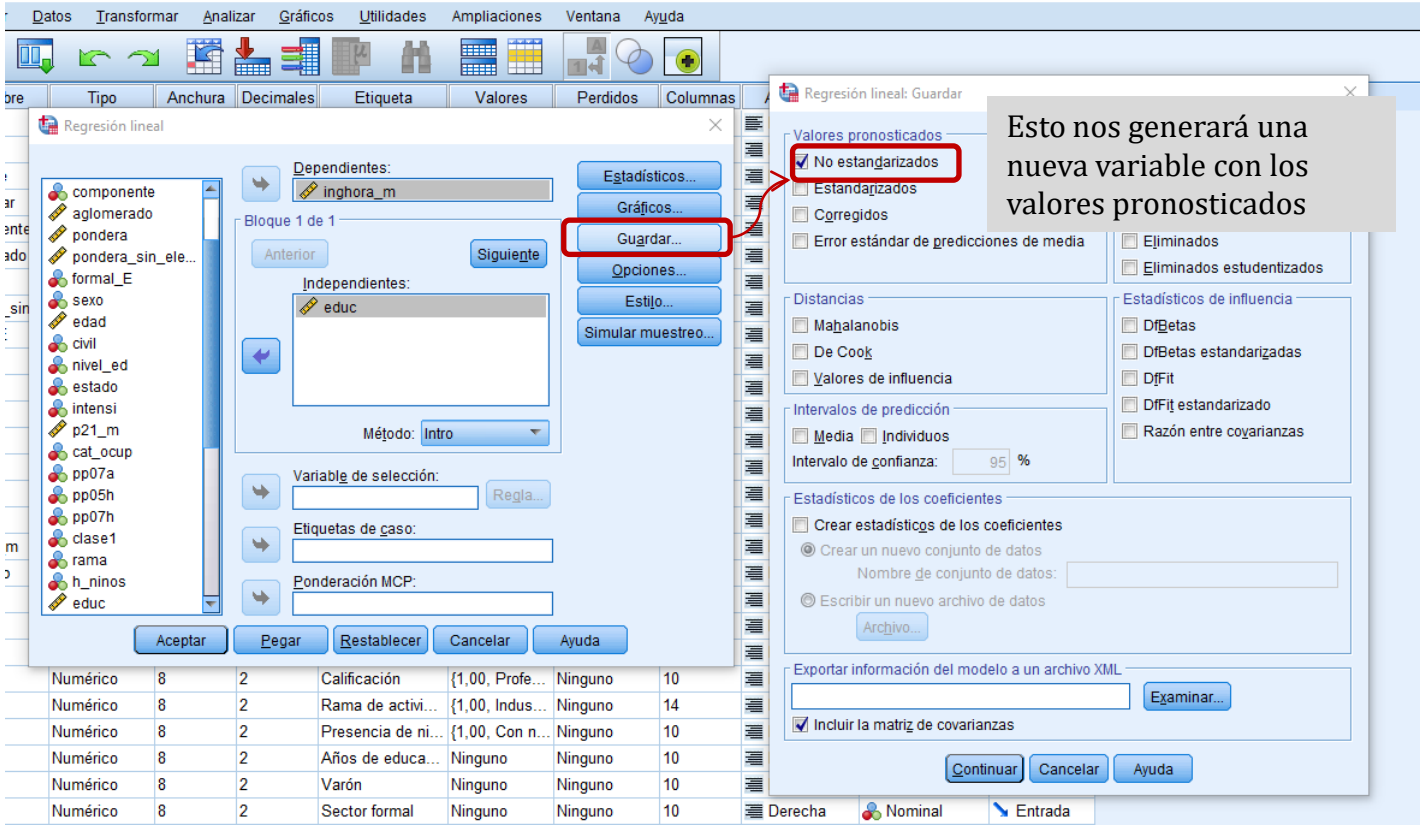
$$\$\boldsymbol{Ingreso\ horario} = \$\mathbf{90,9} + \$\mathbf{15,75} * \boldsymbol{a}\text{ño}$$

- Por ejemplo, para un individuo con 17 años de escolaridad, su ingreso pronosticado será:

$$\$\boldsymbol{Ingreso\ horario} = \$\mathbf{90,9} + \$\mathbf{15,75} * \mathbf{17} = \mathbf{\$358,65}$$

Aplicación del análisis de regresión lineal simple

- Para que veamos los resultados predichos por el modelo:



Ejercicio 2: *análisis de regresión lineal múltiple*

Análisis de regresión lineal múltiple

Incorporamos el sexo, la edad, la educación y el sector de inserción

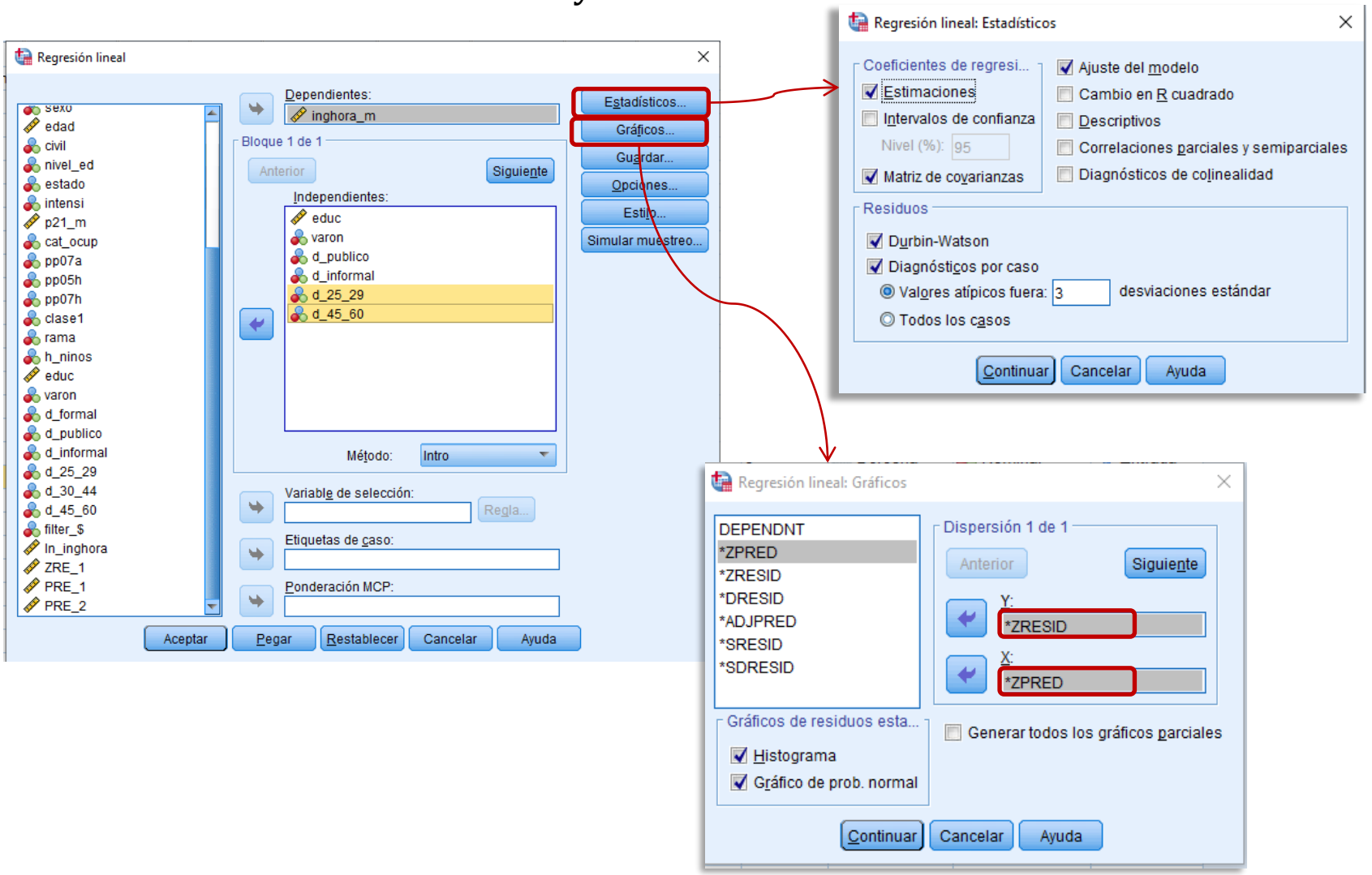
- Generamos ***variables dummies*** *para las variables cualitativas.*
- La **regla** para la cantidad de dummies es: *cantidad de categorías de la variable – 1.*
- Por ejemplo, si la variable edad quiero introducirla de forma categórica para representar a tres grupos o intervalos, debo generar **dos dummies.** La categoría omitida será la referencia o comparación de las demás:

```
recode edad (25 thru 29=1) (else=0) into d_25_29.
execute.
variable labels d_25_29 '25 a 29 años'.

recode edad (45 thru 60=1) (else=0) into d_45_60.
execute.
variable labels d_45_60 '45 a 60 años'.
```

Aplicación del análisis de regresión lineal múltiple

- Los comandos son similares a los ya vistos:



Análisis de regresión lineal múltiple

Incorporamos el sexo, la edad, la educación y el sector de inserción:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	,417[a]	,174	,169	137,29997	1,837

a. Predictores: (Constante), d_45_60 45 a 60 años, varon Varón, d_publico Sector público, d_25_29 25 a 29 años, educ Años de educación, d_informal Sector informal

b. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4225095,791	6	704182,632	37,355	,000[b]
	Residuo	20076613,88	1065	18851,281		
	Total	24301709,67	1071			

a. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

b. Predictores: (Constante), d_45_60 45 a 60 años, varon Varón, d_publico Sector público, d_25_29 25 a 29 años, educ Años de educación, d_informal Sector informal

Análisis de regresión lineal múltiple

Incorporamos el sexo, la edad, la educación y el sector de inserción

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones			Estadísticas de colinealidad	
		B	Desv. Error	Beta			Orden cero	Parcial	Parte	Tolerancia	VIF
1	(Constante)	91,781	21,208		4,328	,000					
	educ Años de educación	15,077	1,295	,356	11,642	,000	,372	,336	,324	,829	1,206
	varon Varón	16,782	8,781	,055	1,911	,056	-,028	,058	,053	,940	1,064
	d_publico Sector público	22,795	12,148	,057	1,876	,061	,166	,057	,052	,845	1,183
	d_informal Sector informal	-36,479	9,636	-,118	-3,786	,000	-,227	-,115	-,105	,796	1,256
	d_25_29 25 a 29 años	-32,940	13,782	-,070	-2,390	,017	-,073	-,073	-,067	,899	1,113
	d_45_60 45 a 60 años	28,074	9,176	,092	3,060	,002	,016	,093	,085	,852	1,174

a. Variable dependiente: inghora_m Ingreso horario de la ocupación principal

El coeficiente estandarizado nos va a permitir evaluar la **importancia relativa** de cada regresor

Vemos si hay regresores que tengan un $p<0.1$ como umbral máximo

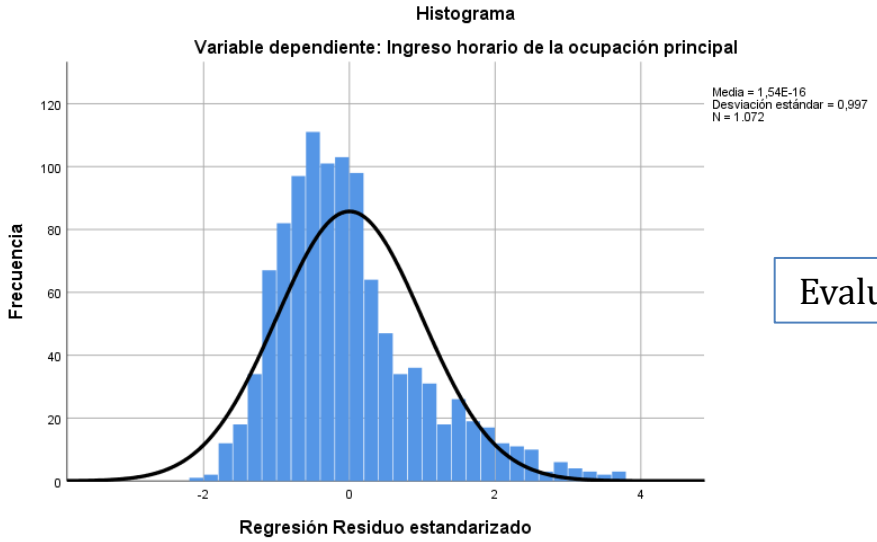
Los coeficientes β se interpretan como **efectos parciales** o **efectos *ceteris paribus***

$$\Delta\hat{y} = \hat{\boldsymbol{\beta}}_1 \Delta \boldsymbol{x}_1$$

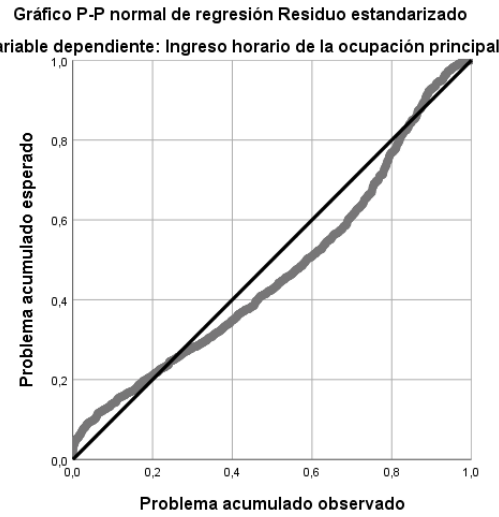
Un **VIF** por debajo de 10 implica que no hay problemas de colinealidad.

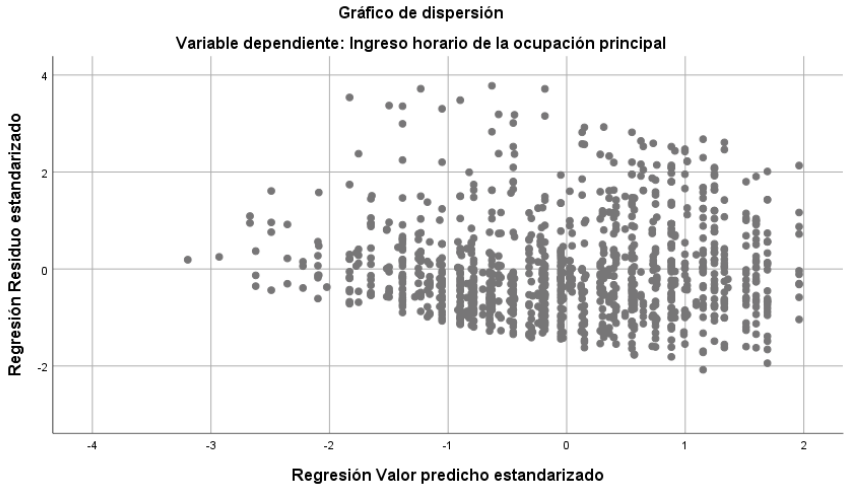
Análisis de regresión lineal múltiple

Incorporamos el sexo, la edad, la educación y el sector de inserción



Evaluar normalidad





Evaluar Homoscedasticidad

Análisis de regresión lineal múltiple

- Reconstruimos la ecuación:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \ \ldots + \hat{\beta}_n x_n$$

$$\widehat{inghora} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 varon + \hat{\beta}_3 publico + \hat{\beta}_4 informal + \hat{\beta}_5 25a29 + \hat{\beta}_6 45a60$$

- Por ejemplo, un varón con 15 años de educación que trabaja en el sector público y tiene 45 a 60 años:

$$\widehat{inghora} = 91,8 + 15,1 * 15 + 16,8 * 1 + 22,8 * 1 + 28,1 * 1 = \$\,386$$

- En cambio, una mujer con similares características:

$$\widehat{inghora} = 91,8 + 15,1 * 15 + 16,8 * 0 + 22,8 * 1 + 28,1 * 1 = \$\,369$$

Ejercicio 3: *regresión múltiple alterando la forma funcional*

¿Qué pasa si alteramos la forma funcional?

- Detectamos algunos problemas de ajuste en la distribución de residuos. La transformación de la variable dependiente puede mejorar el ajuste del modelo.

- Una sencilla transformación logarítmica la implementamos mediante el comando **Transformar > Calcular variable**. O con la siguiente sintaxis:

```
compute ln_inghora=ln(inghora_m).
execute.
variable labels ln_inghora 'Logaritmo del ingreso
horario'.
```

¿Qué pasa si alteramos la forma funcional?

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	,448[a]	,201	,197	,46724	1,828

a. Predictores: (Constante), d_45_60 45 a 60 años, varon Varón, d_publico Sector público, d_25_29 25 a 29 años, educ Años de educación, d_informal Sector informal

b. Variable dependiente: ln_inghora Logaritmo del ingreso horario

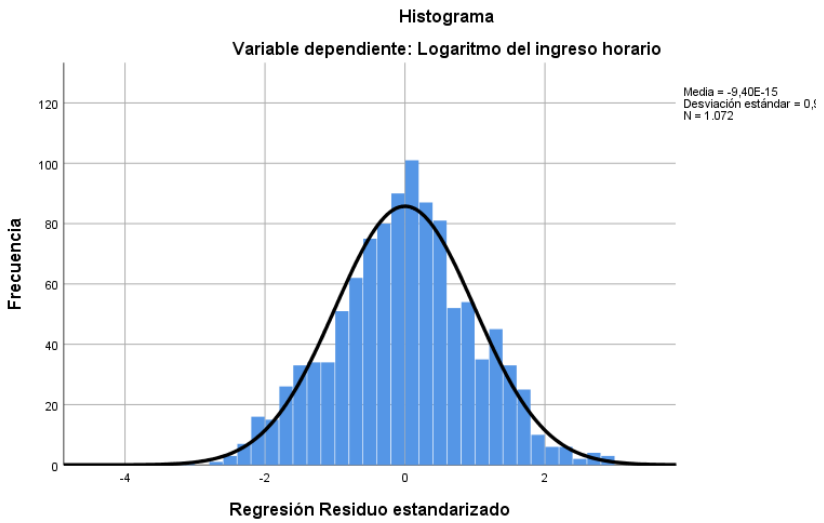
Un aumento de 1 año de educación se traduce en un incremento aproximado de **5,4%** en el ingreso horario

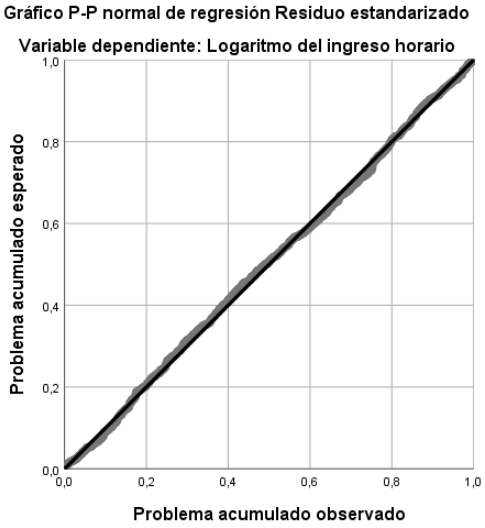
Coeficientes[a]

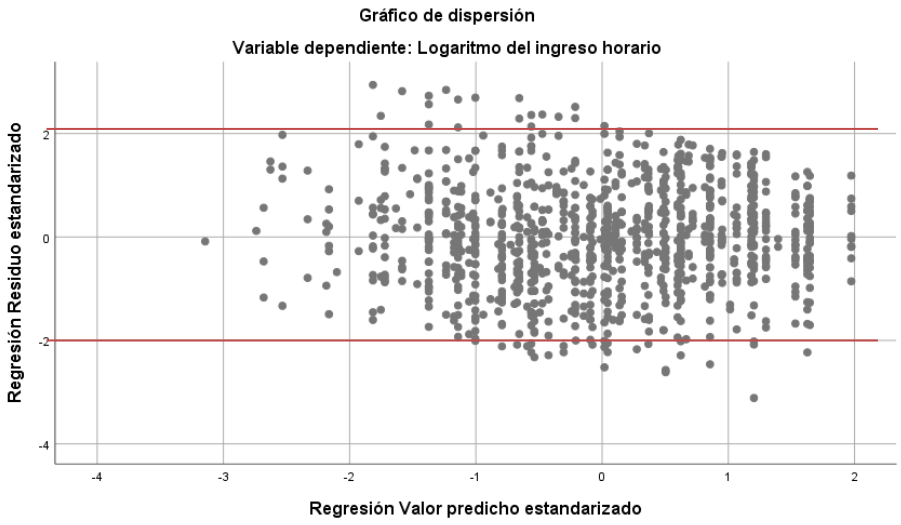
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	(Constante)	4,826	,072		66,863	,000
	educ Años de educación	,054	,004	,370	12,303	,000
	varon Varón	,081	,030	,077	2,724	,007
	d_publico Sector público	,077	,041	,055	1,855	,064
	d_informal Sector informal	-,164	,033	-,153	-4,988	,000
	d_25_29 25 a 29 años	-,121	,047	-,074	-2,578	,010
	d_45_60 45 a 60 años	,104	,031	,098	3,316	,001

a. Variable dependiente: ln_inghora Logaritmo del ingreso horario

¿Qué pasa si alteramos la forma funcional?







Comprobamos una **mejora** en el cumplimiento de los supuestos del modelo.

Análisis de regresión lineal múltiple

- Ahora debemos cambiar la interpretación:

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 educ + \hat{\beta}_2 varon + \hat{\beta}_3 publico + \hat{\beta}_4 informal + \hat{\beta}_5 25a29 + \hat{\beta}_6 45a60$$

- Por ejemplo, un varón con 15 años de educación que trabaja en el sector público y tiene 45 a 60 años:

$$\widehat{loginghora} = 4,826 + 0,054 * 15 + 0,081 * 1 + 0,077 * 1 + 0,104 * 1 = 5,901$$

- El resultado está expresado en términos de logaritmo del ingreso horario:

$$Exp(5,901) = \$365$$

- Con los mismos atributos, pero si es mujer:

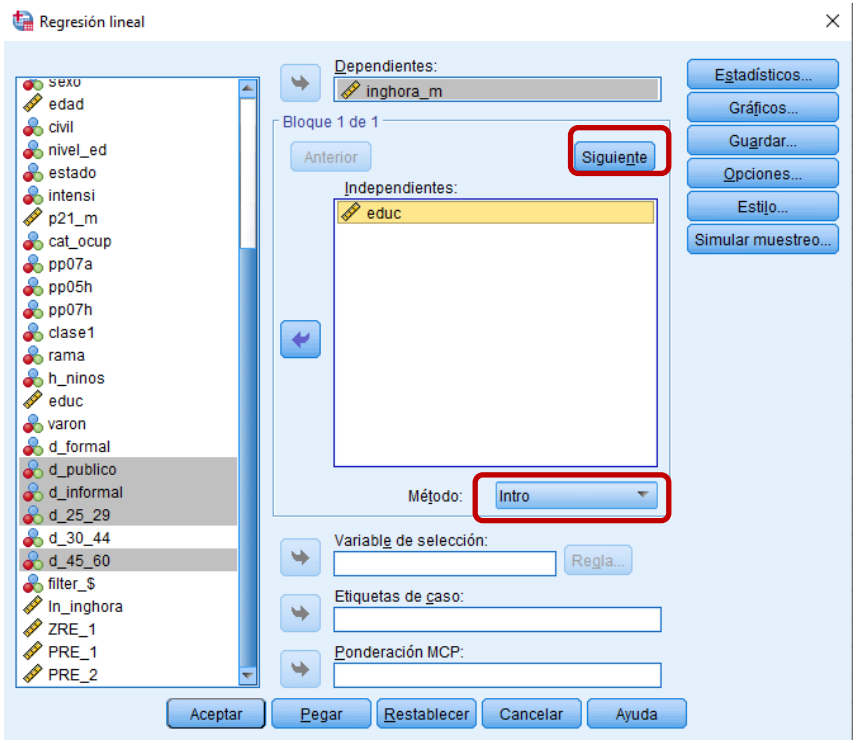
$$\widehat{loginghora} = 4,826 + 0,054 * 15 + 0,081 * 0 + 0,077 * 1 + 0,104 * 1 = 5,819 = 337$$

- La brecha de género sería **8,4%**, que es similar al coeficiente asociado a varón.

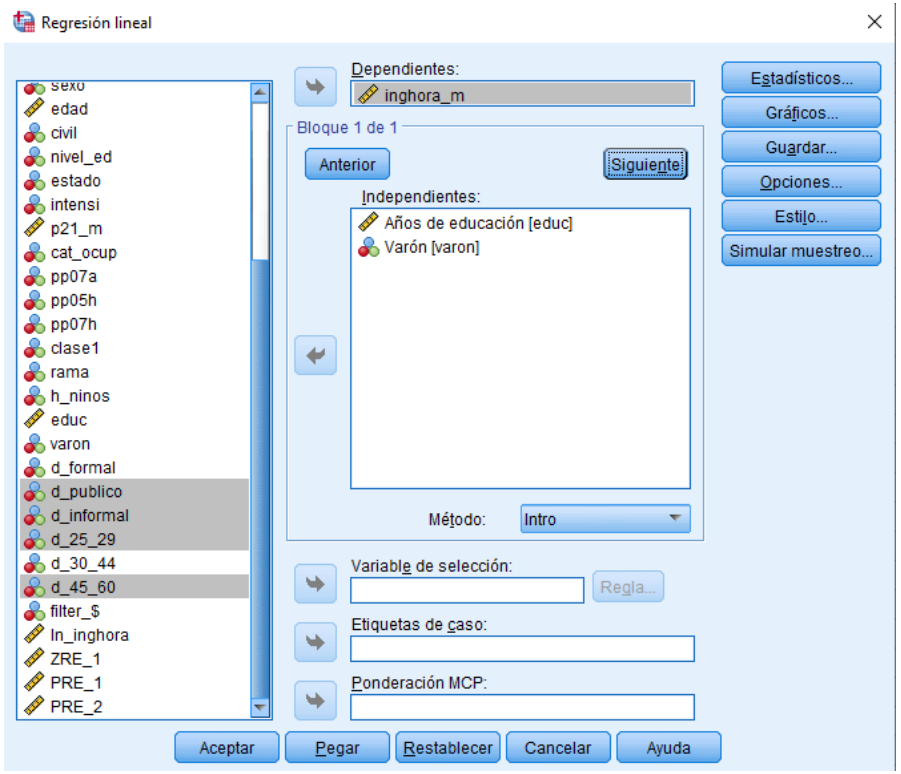
Ejercicio 3: *métodos de incorporar las covariables*

Método Enter

- Vamos incorporando las covariables de a una y evaluamos cómo cambian los coeficientes.



Primer paso



Segundo paso… y así hasta incluir todas las covariables

Método Enter

- Con este método, vamos incorporando los regresores de a uno y evaluamos cómo cambian los coeficientes.

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	educ Años de educación[b]	.	Introducir
2	varon Varón[b]	.	Introducir
3	d_25_29 25 a 29 años, d_45_60 45 a 60 años[b]	.	Introducir
4	d_publico Sector público, d_informal Sector informal[b]	.	Introducir

a. Variable dependiente: ln_inghora Logatirmo del ingreso horario

b. Todas las variables solicitadas introducidas.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,390[a]	,152	,152	,48016
2	,397[b]	,158	,156	,47889
3	,415[c]	,172	,169	,47511
4	,448[d]	,201	,197	,46724

a. Predictores: (Constante), educ Años de educación

b. Predictores: (Constante), educ Años de educación, varon Varón

c. Predictores: (Constante), educ Años de educación, varon Varón, d_25_29 25 a 29 años, d_45_60 45 a 60 años

d. Predictores: (Constante), educ Años de educación, varon Varón, d_25_29 25 a 29 años, d_45_60 45 a 60 años, d_publico Sector público, d_informal Sector informal

Vemos cómo mejora el R2

Método Enter

Correlación semiparcial si lo elevamos al cuadrado nos dice cuánto mejora el modelo: $0,072^2 = 0,005$ Que es la diferencia entre el R2 sólo con *educ* y con *educ* y *sexo= 0,158-0,152=0,005*

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Desv. Error	Beta			Orden cero	Parcial	Parte
1	(Constante)	4,815	,055		87,824	,000			
	educ Años de educación	,057	,004	,390	13,869	,000	,390	,390	,390
2	(Constante)	4,741	,062		76,721	,000			
	educ Años de educación	,059	,004	,405	14,140	,000	,390	,397	,397
	varon Varón	,078	,030	,074	2,583	,010	-,009	,079	,072
3	(Constante)	4,672	,067		69,508	,000			
	educ Años de educación	,063	,004	,427	14,656	,000	,390	,409	,408
	varon Varón	,087	,030	,082	2,880	,004	-,009	,088	,080
	d_25_29 25 a 29 años	-,115	,048	-,071	-2,421	,016	-,075	-,074	-,067
	d_45_60 45 a 60 años	,085	,032	,081	2,700	,007	,015	,082	,075
4	(Constante)	4,826	,072		66,863	,000			
	educ Años de educación	,054	,004	,370	12,303	,000	,390	,353	,337
	varon Varón	,081	,030	,077	2,724	,007	-,009	,083	,075
	d_25_29 25 a 29 años	-,121	,047	-,074	-2,578	,010	-,075	-,079	-,071
	d_45_60 45 a 60 años	,104	,031	,098	3,316	,001	,015	,101	,091
	d_publico Sector público	,077	,041	,055	1,855	,064	,178	,057	,051
	d_informal Sector informal	-,164	,033	-,153	-4,988	,000	-,265	-,151	-,137

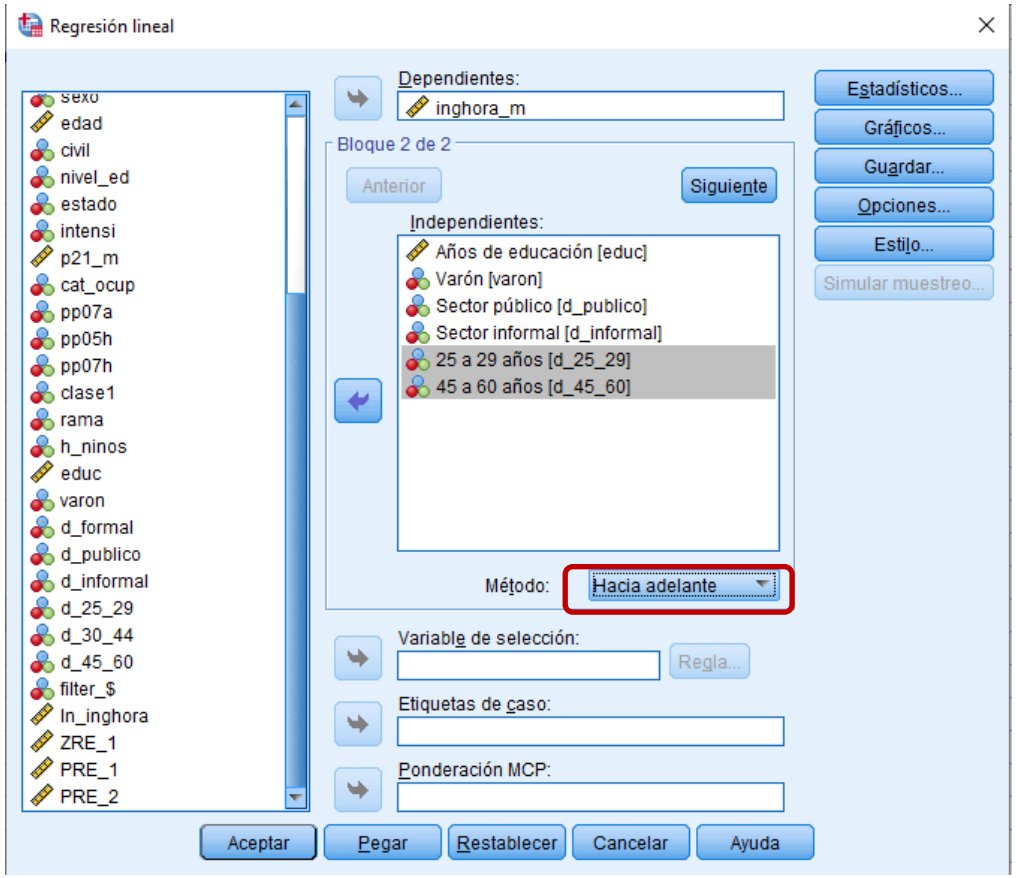
a. Variable dependiente: ln_inghora Logaritmo del ingreso horario

Correlación simple

Estos coeficientes son los que teníamos en el método anterior

Método Forward

- Vamos incorporando las covariables de a una y evaluamos cómo cambian los coeficientes.



Método Forward (Hacia adelante)

- Se selecciona en primer término a la variable independiente que tiene el mayor coeficiente de correlación con la variable a explicar. Luego se calculan los coeficientes de correlación parcial de las demás variables no incluidas. Entra en la ecuación el regresor con la mayor correlación parcial.

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,390[a]	,152	,152	,48016
2	,419[b]	,175	,174	,47379
3	,435[c]	,189	,187	,47014
4	,440[d]	,194	,191	,46897
5	,446[e]	,199	,195	,46777

a. Predictores: (Constante), educ Años de educación

b. Predictores: (Constante), educ Años de educación, d_informal Sector informal

c. Predictores: (Constante), educ Años de educación, d_informal Sector informal, d_45_60 45 a 60 años

d. Predictores: (Constante), educ Años de educación, d_informal Sector informal, d_45_60 45 a 60 años, d_25_29 25 a 29 años

e. Predictores: (Constante), educ Años de educación, d_informal Sector informal, d_45_60 45 a 60 años, d_25_29 25 a 29 años, varon Varón

El método *Forward* emplea el estadístico *F* (Fisher-Snedecor) para evaluar la significancia del *cambio* en la capacidad explicativa.

Método Forward (Hacia adelante)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Desv. Error	Beta			Orden cero	Parcial	Parte
1	(Constante)	4,815	,055		87,824	,000			
	educ Años de educación	,057	,004	,390	13,869	,000	,390	,390	,390
2	(Constante)	4,974	,061		81,045	,000			
	educ Años de educación	,050	,004	,341	11,695	,000	,390	,337	,3
	d_informal Sector informal	-,171	,031	-,160	-5,475	,000	-,265	-,165	-,1
3	(Constante)	4,888	,064		76,160	,000			
	educ Años de educación	,053	,004	,361	12,310	,000	,390	,353	,3
	d_informal Sector informal	-,186	,031	-,174	-5,982	,000	-,265	-,180	-,165
	d_45_60 45 a 60 años	,125	,030	,119	4,203	,000	,015	,128	,116
4	(Constante)	4,917	,065		75,588	,000			
	educ Años de educación	,053	,004	,359	12,263	,000	,390	,351	,337
	d_informal Sector informal	-,188	,031	-,176	-6,054	,000	-,265	-,182	-,166
	d_45_60 45 a 60 años	,101	,031	,096	3,229	,001	,015	,098	,089
	d_25_29 25 a 29 años	-,118	,047	-,073	-2,517	,012	-,075	-,077	-,069
5	(Constante)	4,838	,072		67,268	,000			
	educ Años de educación	,055	,004	,376	12,552	,000	,390	,359	,344
	d_informal Sector informal	-,183	,031	-,171	-5,895	,000	-,265	-,178	-,162
	d_45_60 45 a 60 años	,104	,031	,099	3,340	,001	,015	,102	,092
	d_25_29 25 a 29 años	-,121	,047	-,075	-2,586	,010	-,075	-,079	-,071
	varon Varón	,076	,030	,072	2,543	,011	-,009	,078	,070

a. Variable dependiente: ln_inghora Logatirmo del ingreso horario

Acá introduce primero el regresor con el mayor **coeficiente de correlación** con la variable dependiente

Sigue la que tenga el mayor **coeficiente de correlación parcial**

Con este método se excluyó un regresor que ya sabíamos que era poco significativo

Ejercicio 4: *regresión múltiple incluyendo interacciones*

Análisis de regresión lineal con interacciones

- El análisis de regresión incorpora interacciones cuando suponemos que el efecto de una variable *modula* el comportamiento de otra covariable.

$$\widehat{\boldsymbol{y}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{0}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\boldsymbol{x}_{\boldsymbol{1}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}\boldsymbol{x}_{\boldsymbol{2}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}\boldsymbol{x}_{\boldsymbol{1}} * \boldsymbol{x}_{\boldsymbol{2}} \ldots$$

- Por ejemplo: sea y el ingreso mensual, x_1 una variable *dummy* que indica que el trabajador tiene *educación inferior a secundaria completa*, x_2 otra *dummy* que indica si el trabajador es *informal*. El coeficiente de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}$ recoge el efecto de la interacción entre ambas variables.

- Si el coeficiente de la interacción es significativo, estaremos afirmando, por ejemplo, que el efecto de la *baja educación* sobre el ingreso se refuerza (o se debilita) cuando *se es informal.*

Análisis de regresión lineal con interacciones

Incorporamos una interacción entre variables cualitativas

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 mujer + \hat{\beta}_2 informal + \hat{\beta}_3 mujer * informal$$

Una mujer con trabajo informal:

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 mujer + \hat{\beta}_2 informal + \hat{\beta}_3 mujer * informal$$

Un varón con trabajo informal

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 informal + \hat{\beta}_3 * 0$$

Una mujer con trabajo formal

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 mujer + \hat{\beta}_2 * 0 + \hat{\beta}_3 * 0$$

Un varón con trabajo formal:

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 * 0 + \hat{\beta}_2 * 0 + \hat{\beta}_3 * 0$$

Análisis de regresión lineal con interacciones

Armamos las variables dummies:

```
recode sexo (1=0) (2=1) into d_mujer.
variable labels d_mujer 'Mujer'.

recode formal_E (3=1) (else=0) into d_informal.
execute.
variable labels d_informal 'Sector informal'.

if (d_mujer=1 & d_informal=1) d_mujer_informal=1.
if (d_mujer=0 & d_informal=1) d_varon_informal=1.
if (d_mujer=1 & d_informal=0) d_mujer_formal=1.
if (d_mujer=0 & d_informal=0) d_varon_formal=1.
execute.

recode d_mujer_informal d_varon_informal d_mujer_formal d_varon_formal
(missing=0).
execute.
```

Análisis de regresión lineal con interacciones

Incorporamos una interacción entre variables cualitativas

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	,274[a]	,075	,072	,50205	1,699

a. Predictores: (Constante), d_mujer_informal, d_mujer Mujer, d_informal Sector informal

b. Variable dependiente: ln_inghora Logatirmo del ingreso horario

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	(Constante)	5,629	,026		219,708	,000
	d_mujer Mujer	,070	,040	,066	1,757	,079
	d_informal Sector informal	-,219	,041	-,205	-5,333	,000
	d_mujer_informal	-,153	,064	-,108	-2,397	,017

a. Variable dependiente: ln_inghora Logatirmo del ingreso horario

Este es el coeficiente asociado al género, cuando se es formal

Esta es la "penalidad" salarial específica para mujeres en el sector informal

$$\widehat{loginghora} = \hat{\beta}_0 + \hat{\beta}_1 mujer + \hat{\beta}_2 informal + \hat{\beta}_3 mujer * informal$$

$$Mujer\ informal \rightarrow \widehat{loginghora} = 5.629 + 0.07 * 1 + (-0.219) * 1 + (-0.153) * 1$$

$$Varon\ informal \rightarrow \widehat{loginghora} = 5.629 + 0.07 * 0 + (-0.219) * 1 + (-0.153) * 0$$

$$Mujer\ formal \rightarrow \widehat{loginghora} = 5.629 + 0.07 * 1 + (-0.219) * 0 + (-0.153) * 0$$

$$Varon\ formal \rightarrow \widehat{loginghora} = 5.629 + 0.07 * 0 + (-0.219) * 0 + (-0.153) * 0$$

Análisis de regresión lineal con interacciones

Otra manera de presentar la interacción entre variables cualitativas

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Desv. Error	Beta		
1	(Constante)	5,629	,026		219,708	,000
	d_mujer_informal	-,302	,046	-,214	-6,585	,000
	d_varon_informal	-,219	,041	-,176	-5,333	,000
	d_mujer_formal	,070	,040	,058	1,757	,079

a. Variable dependiente: ln_inghora Logatirmo del ingreso horario

Esta es la "penalidad" salarial de las mujeres (respecto a varones), cuando se es informal

Este es el efecto de ser informal cuando se es varón (respecto a varones formales)

Si no hubiera interacción, *d_varon_informal* debería ser igual a *d_mujer_informal*

Este es el efecto de ser formal y de ser mujer (frente a varones formales)

Mujer trabajadora informal: $\widehat{loginghora} = 5,629 + (-0,302)mujer\_informal = 5,33 = \$206$

Mujer trabajadora formal: $\widehat{loginghora} = 5,629 + (+0,070)mujer\_formal = 5,7 = \$299$

Varón trabajador informal: $\widehat{loginghora} = 5,629 + (-0,219)varon\_informal = 5,4 = \$224$

Varón trabajador formal: $\widehat{loginghora} = 5,629 = \$278$