

---

# Análisis de Regresión Múltiple

Pedro Valero Mora-valerop@uv.es

*Metodología de las CC del Comp-Universitat de València*

*Marzo 2012*



VNIVERSITAT DE VALÈNCIA

# Contenidos

## Introducción

Introducción	1
El proceso de ajuste de modelos	2
Ejemplo Salario Actual	4
Planteamiento del modelo	5
Evaluación del Ajuste	6
Diagnóstico del modelo	7
Gráficos de residuales	10
Busqueda de nuevos modelos	14
Interpretación	18

## Explorando el salario actual

Introducción	22
Planteamiento del modelo	23
Variables en el ejemplo de Empleados	24
Variables derivadas	25
Interacciones	29
Polinomios	30
El modelo inicial	33
Evaluación del ajuste	34
Diagnóstico del modelo	35
Transformaciones	37

Resultados utilizando variables transformadas	45
Aplicaciones de las transformaciones	49
Interpretación del modelo	55
La importancia de los coeficientes	57
Coeficientes estandarizados	58
Coeficientes de correlación semiparcial y parcial	60

## Explorando el salario inicial

Planteamiento del modelo	73
Introduciendo variables una por una	74
Cambio en R cuadrado	75
Examinando varias variables	78
Medidas de colinealidad	80
Métodos automáticos	83
Stepwise	88
Advertencias sobre stepwise	90
El modelo final	93
Diagnóstico del modelo	94
Residuales frente a predicha	95
Gráficos de regresión parcial	96
Análisis de los valores individuales	97
Residuales	101
Distancias (Palanca)	103

---

Influencia	118	Variables ficticias	125
Interpretación	124	Polinomios	128

# ***Introducción***

# Introducción

- La regresión múltiple es un conjunto de técnicas estadísticas que permiten diagnosticar la relación entre una variable dependiente y varias variables independientes
- Es una técnica muy importante en muchas disciplinas y es una de las que más avances ha tenido en los últimos años a pesar de ser relativamente bien conocida previamente.
  - Eso significa que se puede profundizar hasta niveles muy grandes.
- Algunas referencias son:
  - Tabachnick and Fidell (2007). Using Multivariate Statistics. Pearson (tiene muchas otras técnicas y es una buena introducción a todas ellas)
  - Cook and Weisberg (2005). Applied Linear Regression. Wiley
  - Norusis, M. (1993) SPSS Base System. SPSS inc.

- Las técnicas de regresión admiten que las variables independientes estén correlacionadas entre sí hasta cierto punto, lo cual se ajusta mejor a datos observacionales o de encuesta (aunque también pueden utilizarse en situaciones de experimentos ya que en el fondo las técnicas de ANOVA son semejantes a la regresión).
  - Regresión es más útil para analizar problemas complejos que no es fácil reducir a diseños ortogonales tal y como se plantean en un experimento
- Las variables independientes en análisis de regresión pueden ser continuas o dicotómicas.
  - Si tenemos variables categóricas con más de dos categorías deben ser recodificadas (muchos programas hacen esto automáticamente)
  - Un ANOVA puede ser calculado mediante un programa de regresión utilizando esta codificación
- De hecho, una gran cantidad de técnicas estadísticas pueden ser entendidas como casos específicos de la Regresión Múltiple (Modelo Lineal Generalizado)

# El proceso de ajuste de modelos

- Planteamiento del modelo
- Evaluación del ajuste
- Diagnóstico
- Búsqueda de nuevos modelos
- Interpretación

# Ejemplo Salario Actual

- Prediciendo/explicando el salario de unos empleados a partir del salario inicial

El salario inicial es una variable que obviamente correlaciona con el salario actual

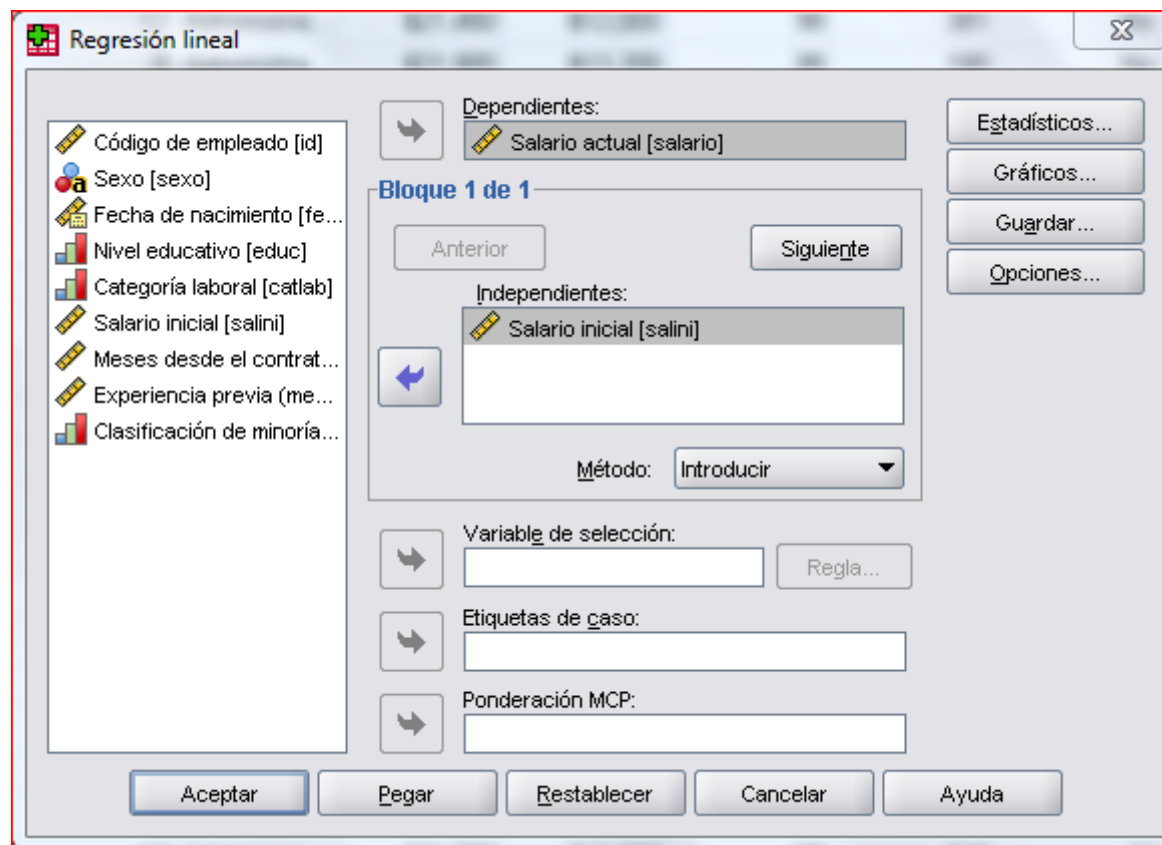
No obstante, hay una serie de factores que pueden hacer que cada sujeto avance más o menos en su carrera

Estudiar este análisis nos ayudará a entender la relación general entre salario inicial y actual pero también nos permitirá detectar casos excepcionales que pueden ser dignos de interés



# Planteamiento del modelo

- $\text{Salario Actual} = \text{Const} + \text{Salinicial} + \text{Error}$



# Evaluación del Ajuste

- Resumen del modelo

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,880 <sup>a</sup>	,775	,774	\$8,115.356

a. Predictors: (Constant), Salario inicial

- ANOVA

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,068E11	1	1,068E11	1622,118	,000 <sup>a</sup>
	Residual	3,109E10	472	6,586E7		
	Total	1,379E11	473			

a. Predictors: (Constant), Salario inicial

b. Dependent Variable: Salario actual

- Coeficientes y significación

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1928,206	888,680		2,170	,031
	Salario inicial	1,909	,047	,880	40,276	,000

a. Dependent Variable: Salario actual

- Este output es muy redundante en este caso ya que en el caso de un sólo predictor tanto  $R$ , como el ANOVA como el nivel de significación del coeficiente de la variable predictora son redundantes:

$$R = 0.88 = \sqrt{\frac{SCR}{SCT}} = \sqrt{\frac{1068 \times 10^{11}}{1379 \times 10^{11}}} = Beta$$

- El nivel de significación del ANOVA y del coeficiente para la variable predictora son iguales
  - Más adelante veremos que para varios predictores esto no es así, y cada coeficiente tiene un nivel de significación individual

- Los coeficientes no estandarizados  $B$  nos dan una idea del incremento en la variable dependiente por unidad en la variable independiente
  - Cada dolar inicial se ha convertido en 1.909 dolares de salario actual
  - Obviamente este modelo es incompleto, el tiempo trabajando es importante, ¿no? Lo veremos más adelante.

# Diagnóstico del modelo

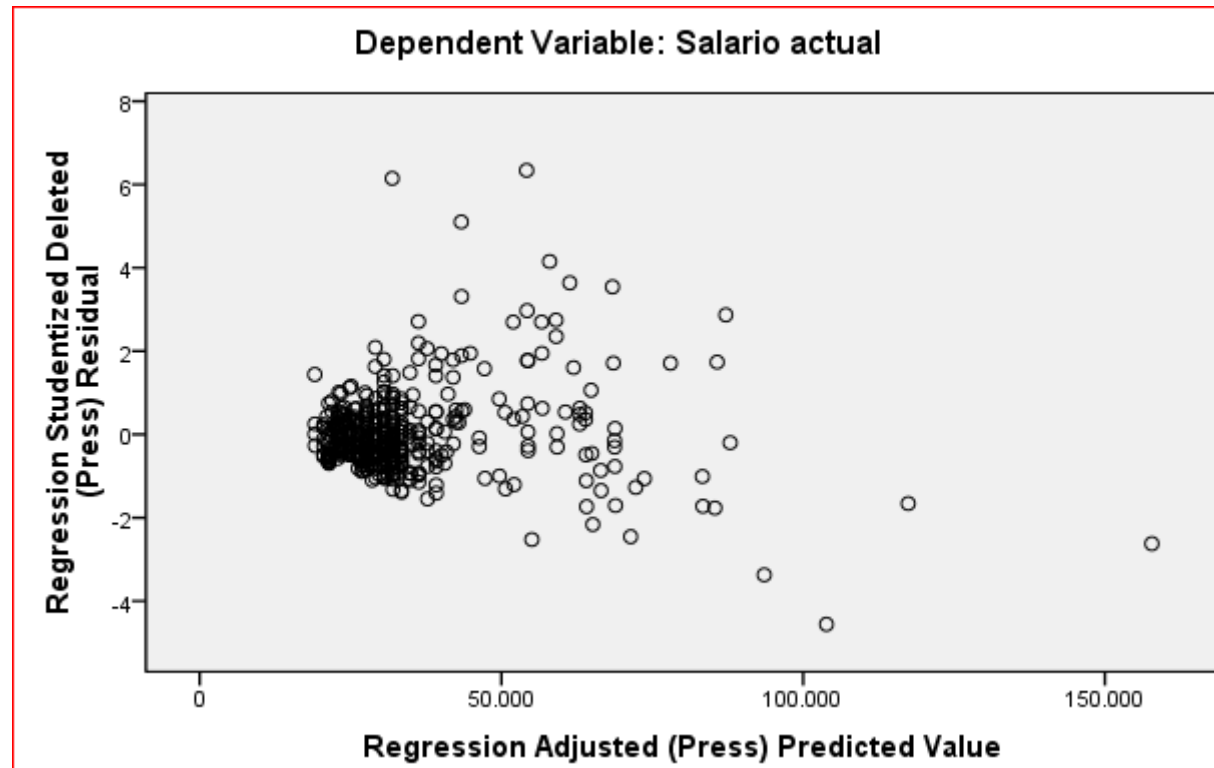
- Hay muchos aspectos a evaluar en un modelo pero generalmente el análisis de los residuales es uno de los aspectos más interesantes
- Un residual es lo que no ha sido explicado por el modelo. Es la diferencia entre el valor observado y el predicho.

$$E = Y - \hat{Y}$$

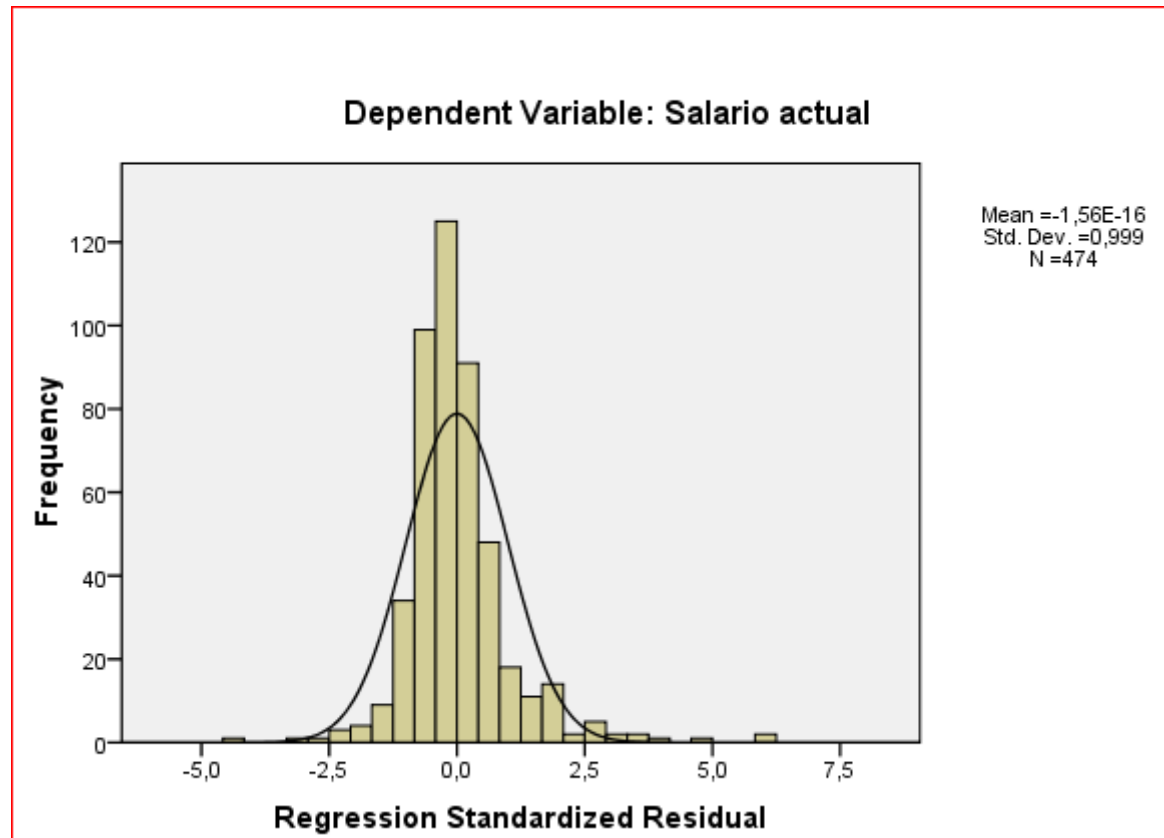
- Residuales muy grandes en valor absoluto significan que hay valores que no están bien explicados por el modelo y necesitan ser examinados individualmente

- A menudo se utiliza una versión de residuales denominada Studentizados, los cuales tienen una estimación de su desviación típica que varía de punto a punto
- Los residuales pueden utilizarse para evaluar:
  - Valores extremos con residuales altos
  - Linealidad
  - Igualdad de varianza
  - Normalidad del error
  - También, si los datos provienen de series temporales se pueden utilizar para evaluar la independencia del error pero este caso no lo veremos aquí.

- La mejor manera de hacer esta evaluación es generalmente mediante un gráfico de las puntuaciones predichas frente a los residuales. En nuestro caso vemos que no se cumple el supuesto de homogeneidad de varianza



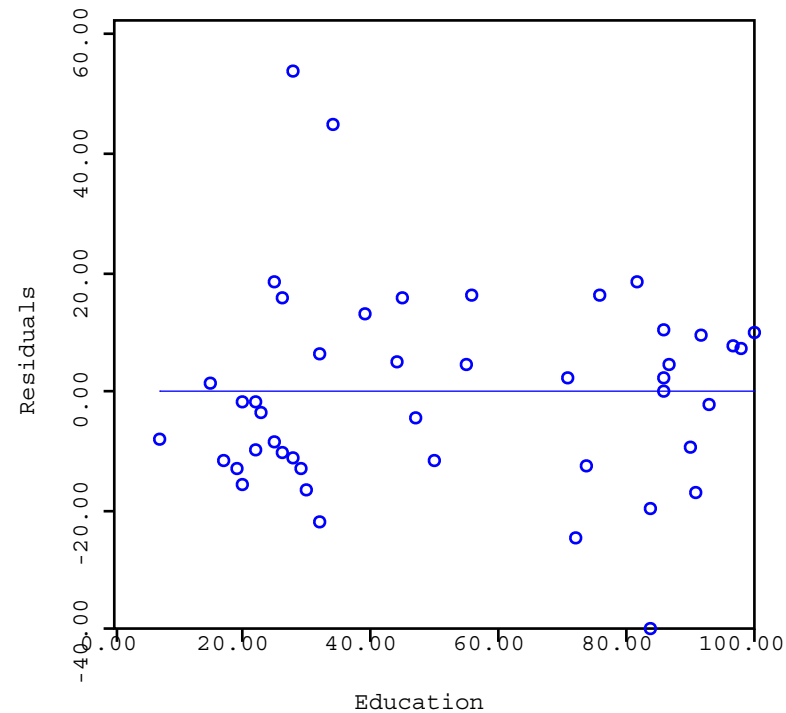
- En cambio, el histograma de los residuales no está demasiado mal (aunque se ven valores extremos)



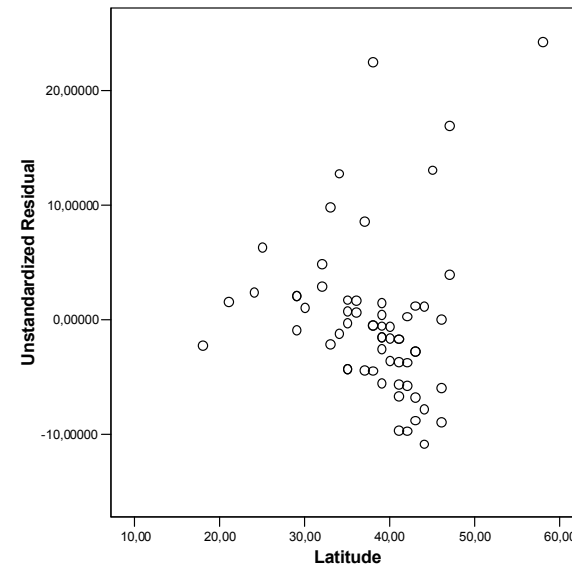
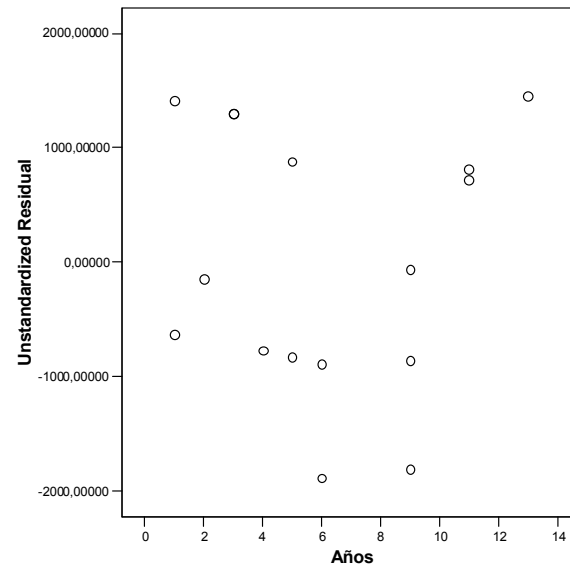


## *Gráficos de residuales*

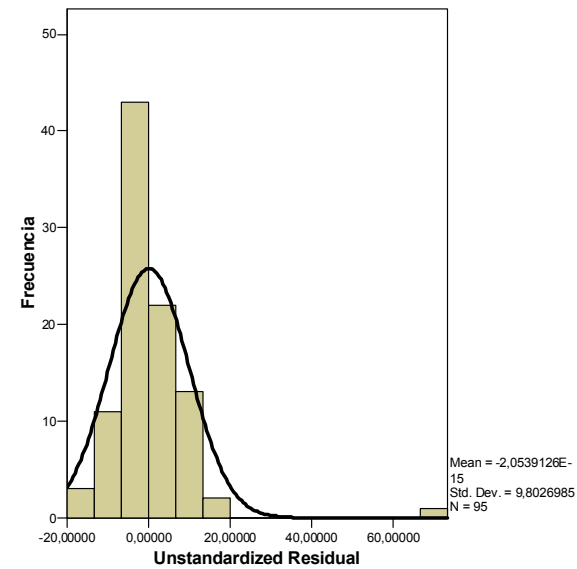
- Gráfico correcto (aunque se ven un par de residuales altos)



- Curvilinealidad



- Normalidad del error

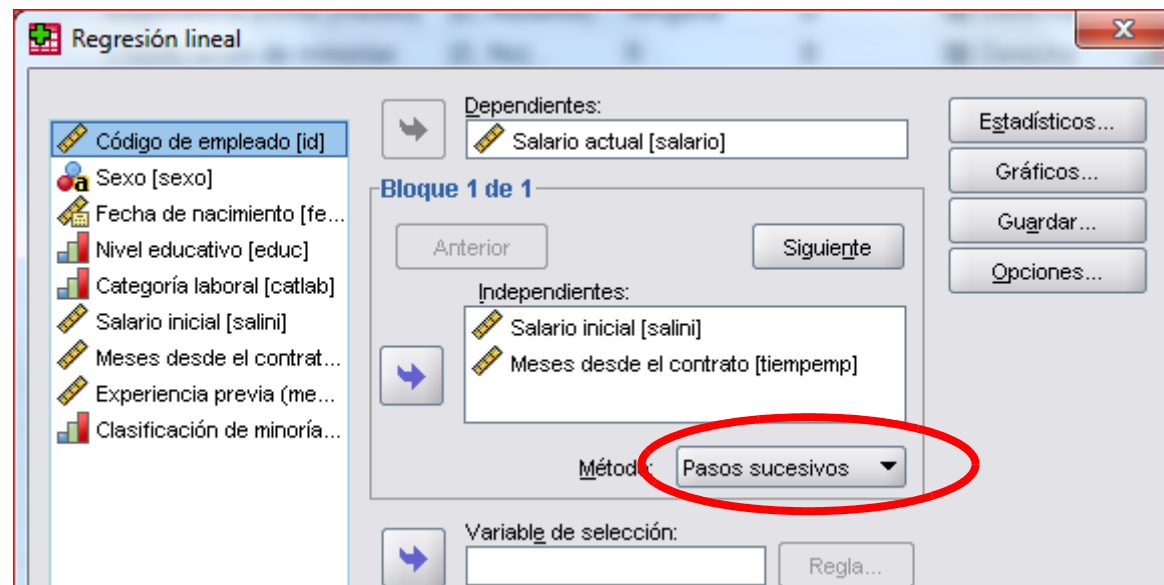


- Soluciones al incumplimiento de supuestos
  - Existen soluciones para corregir estos incumplimientos de supuestos y que hay que aplicar en cada caso. Los veremos más adelante.

# Busqueda de nuevos modelos

- Generalmente, el proceso de ajuste de modelos incluye cierta cantidad de búsqueda de alternativas con comparaciones. Estos métodos pueden ser automatizados (stepwise) aunque son recomendados por los expertos

En nuestro caso, podemos incluir el tiempo que se lleva en el trabajo para probar con un modelo más complejo y podemos pedir la comparación con el modelo anterior (Pasos sucesivos)



- El output ahora añade una sección que compara el modelo inicial con una variable con el modelo con dos variables. La columna final nos indica que añadir esta variable produce un cambio de F significativo (aunque pequeño). No obstante, si miramos el gráfico de residuales veremos que nuestro problema de cumplimiento de supuestos se mantiene y habría que encontrarle una solución

Model Summary <sup>c</sup>									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,880 <sup>a</sup>	,775	,774	\$8,115.356	,775	1622,118	1	472	,000
2	,886 <sup>b</sup>	,785	,784	\$7,936.139	,010	22,558	1	471	,000

a. Predictors: (Constant), Salario inicial

b. Predictors: (Constant), Salario inicial, Meses desde el contrato

c. Dependent Variable: Salario actual

# Interpretación

- La interpretación del modelo se refiere a valorar los coeficientes de la regresión.
- En nuestro caso, con sólo una variable independiente la interpretación es sencilla. Más adelante veremos ejemplos más complicados

Coefficients <sup>a</sup>					
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	1928,206	888,680		2,170	,031
Salario inicial	1,909	,047	,880	40,276	,000

a. Dependent Variable: Salario actual

- El salario actual es 1.91 veces el salario inicial como promedio
- El salario medio inicial es 1928\$
- A partir de esos valores podemos predecir el salario actual de un trabajador a partir de su salario inicial.  $\text{SalActual} = 1928.206 + 1909 * \text{SalInicial}$

## Actividades

---

1. Repetir el ejemplo utilizando los meses desde el contrato como variable independiente

*Examina los gráficos para ver si de esta manera se han corregido los problemas de homoscedasticidad*



***Explorando el  
salario actual***

# Introducción

- El ejemplo anterior utilizaba sólo un predictor
- No obstante, la regresión es más interesante con varios predictores, aunque también más complicada

Veremos a continuación un ejemplo con regresión múltiple

# Planteamiento del modelo

- Un modelo de regresión múltiple puede tener como variables independientes:
  - Una o varias variables independientes
  - Variables categóricas (codificadas como ficticias)
  - Interacción entre las variables independientes (es decir multiplicación)
  - Términos polinomiales

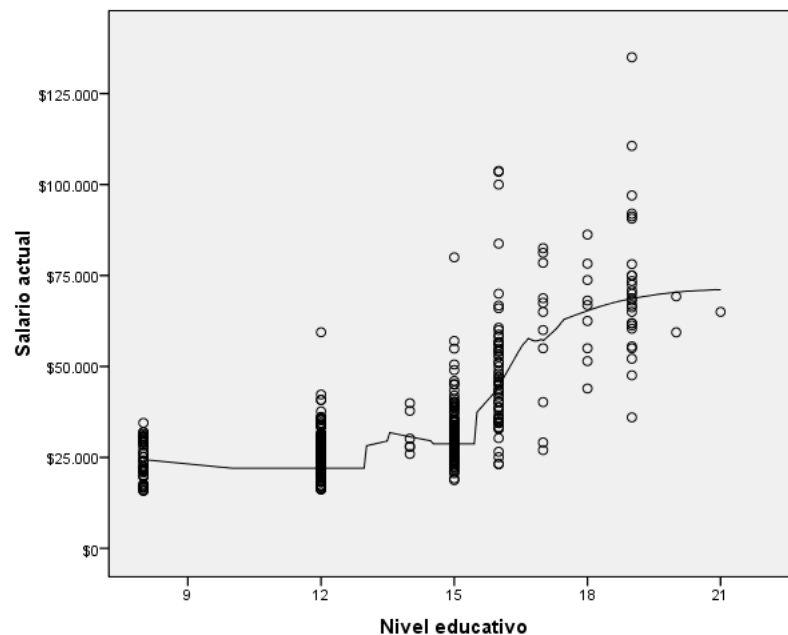
# Variables en el ejemplo de Empleados

- Estas son las variables disponibles

id	Numérico	4	0	Código de empleado	Ninguna	Ninguna	8	≡ Derecha	Escala
sexo	Cadena	1	0	Sexo	{h, Hombre}...	Ninguna	5	≡ Izquierda	Nominal
fechnac	Fecha	10	0	Fecha de nacimiento	Ninguna	Ninguna	13	≡ Derecha	Escala
educ	Numérico	2	0	Nivel educativo	{0, 0 (Ausente)...	0	8	≡ Derecha	Ordinal
catlab	Numérico	1	0	Categoría laboral	{0, 0 (Ausente)...	0	8	≡ Derecha	Ordinal
salario	Dólar	8	0	Salario actual	{\$0, Ausente}...	\$0	8	≡ Derecha	Escala
salini	Dólar	8	0	Salario inicial	{\$0, Ausente}...	\$0	8	≡ Derecha	Escala
tiempemp	Numérico	2	0	Meses desde el contrato	{0, Ausente}...	0	8	≡ Derecha	Escala
expprev	Numérico	6	0	Experiencia previa (meses)	{0, Ausente}...	Ninguna	8	≡ Derecha	Escala
minoría	Numérico	1	0	Clasificación de minorías	{0, No}...	9	8	≡ Derecha	Ordinal

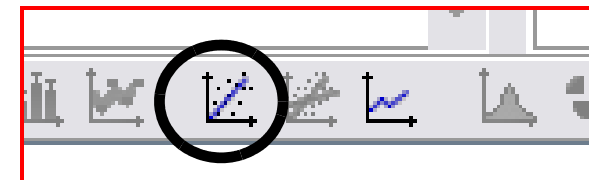
- Sexo puede utilizarse como variable dicotómica, también minoría. No obstante, Sexo está como cadena así que hay que recodificarla (el comando recodificación automática lo hace automáticamente)

- La fecha de nacimiento es una variable numérica porque cada día está codificado internamente como un número más (se puede comprobar pasando la variable de tipo fecha a tipo numérico). Eso la hace apta para ser utilizada como una variable numérica en los análisis.
- Categoría laboral está etiquetada como ordinal pero sólo tiene tres categorías. Vale la pena recodificarla en unos y ceros.



• Nivel educativo tiene muchos niveles. Convertirlas en categorías podría ser demasiado complicado así que se puede hacer el siguiente gráfico:

- Se trata de un diagrama de dispersión simple con una línea loess añadida (se obtiene haciendo doble click en el gráfico que se obtiene haciendo Gráficos>Cuadros de diálogo antiguos>Diagrama de dispersión/puntos



- En este gráfico vemos que la relación entre las variables es curvilínea pero monótona por lo que tendremos que hacer algo

- **Salario inicial, Tiempo empleado y Experiencia Previa son variables numéricas**
- **Minoría es una variable dicotómica con dos categorías.**

## *Variables derivadas*

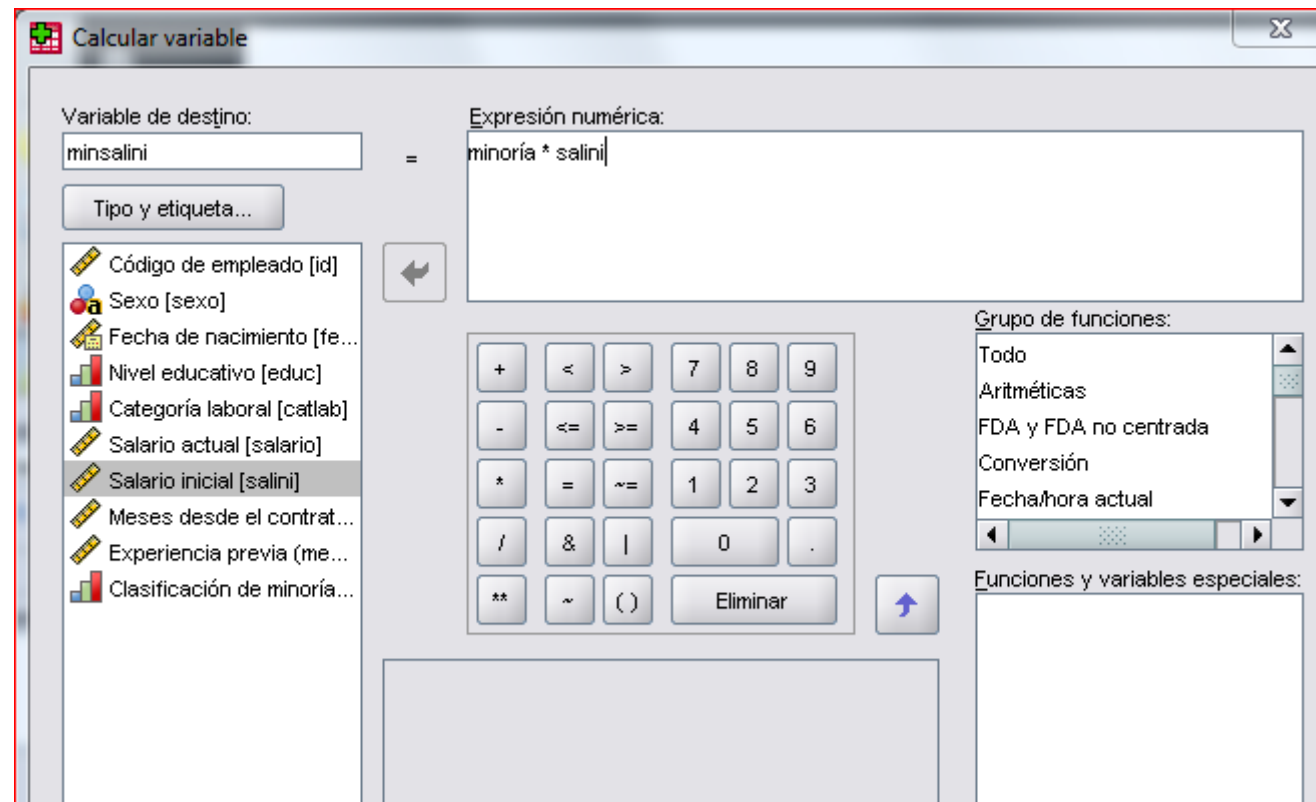
- Además de las variables originales es posible utilizar como variables independientes otras variables derivadas de las originales. Veremos dos tipos
  - Interacciones
  - Polinomios



## *Interacciones*

- La interacción de Sexo\*Salario Inicial, o de Minoría\*Salario Inicial son dos ejemplos interesantes de interacciones:
  - Es posible que haya una discriminación inicial hacia las mujeres y que sus salarios iniciales sean más bajos que los de los hombres (controlando por el resto de los factores)
  - También es posible que haya una discriminación inicial hacia las minorías y sus salarios iniciales sean más bajos

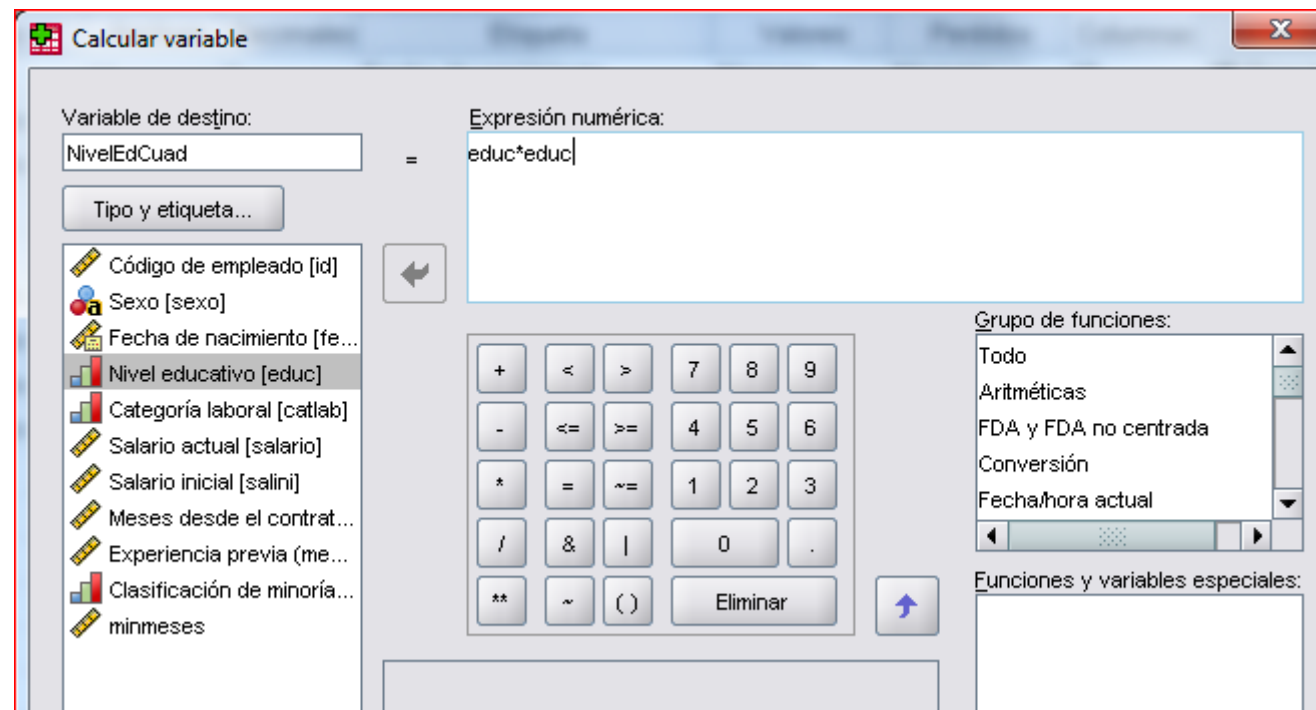
- Esas interacciones se pueden introducir en el modelo multiplicando las variables que queremos considerar de ese modo (menú Transformar>Calcular variable). Esto nos produce una variable que podemos introducir en el análisis a continuación



- Las interacciones pueden ser también entre variables numéricas (por ejemplo, produce un efecto multiplicativo la experiencia previa por los meses contratado?)

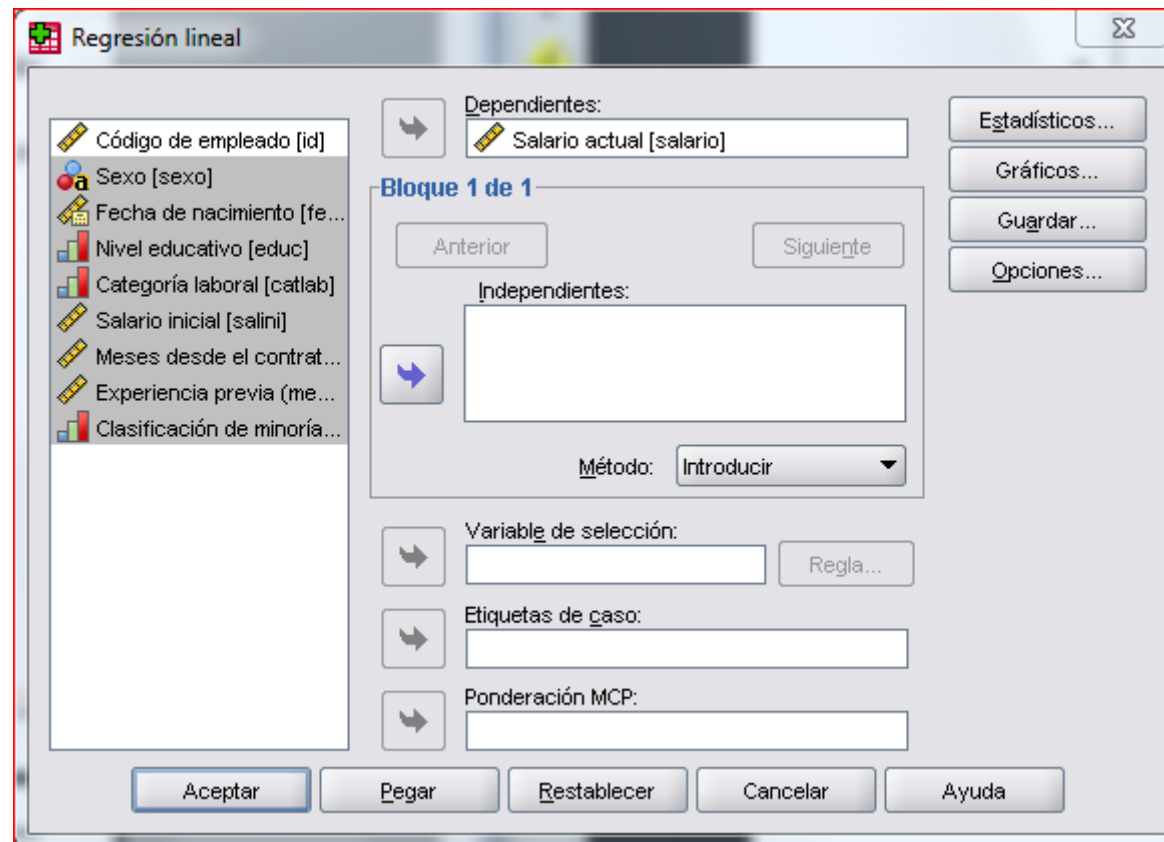
## Polinomios

- Luego veremos una aplicación para ajustar modelos que parecen curvilíneos. Se trata de calcular términos que corresponden a una variable multiplicada por sí misma. Pueden ser al cuadrado o al cubo. En nuestro caso, veremos más adelante que el nivel educativo puede ser interesante elevarlo al cuadrado.



# El modelo inicial

- En este modelo incluiremos todas las variables y veremos su efecto sobre el salario actual (en el análisis luego podemos quitar las no significativas)



# Evaluación del ajuste

- El siguiente paso de este análisis será evaluar el ajuste (en este modelo no he incluido interacciones ni polinomios)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,919 <sup>a</sup>	<b>,844</b>	,841	\$6,808.552

a. Predictors: (Constant), Experiencia previa (meses), Meses desde el contrato, Salario inicial, Sexo, Nivel educativo, Categoría laboral, Fecha de nacimiento

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1,164E11	7	1,662E10	358,590	<b>,000<sup>a</sup></b>
	Residual	2,156E10	465	4,636E7		
	Total	1,379E11	472			

a. Predictors: (Constant), Experiencia previa (meses), Meses desde el contrato, Salario inicial, Sexo, Nivel educativo, Categoría laboral, Fecha de nacimiento

b. Dependent Variable: Salario actual

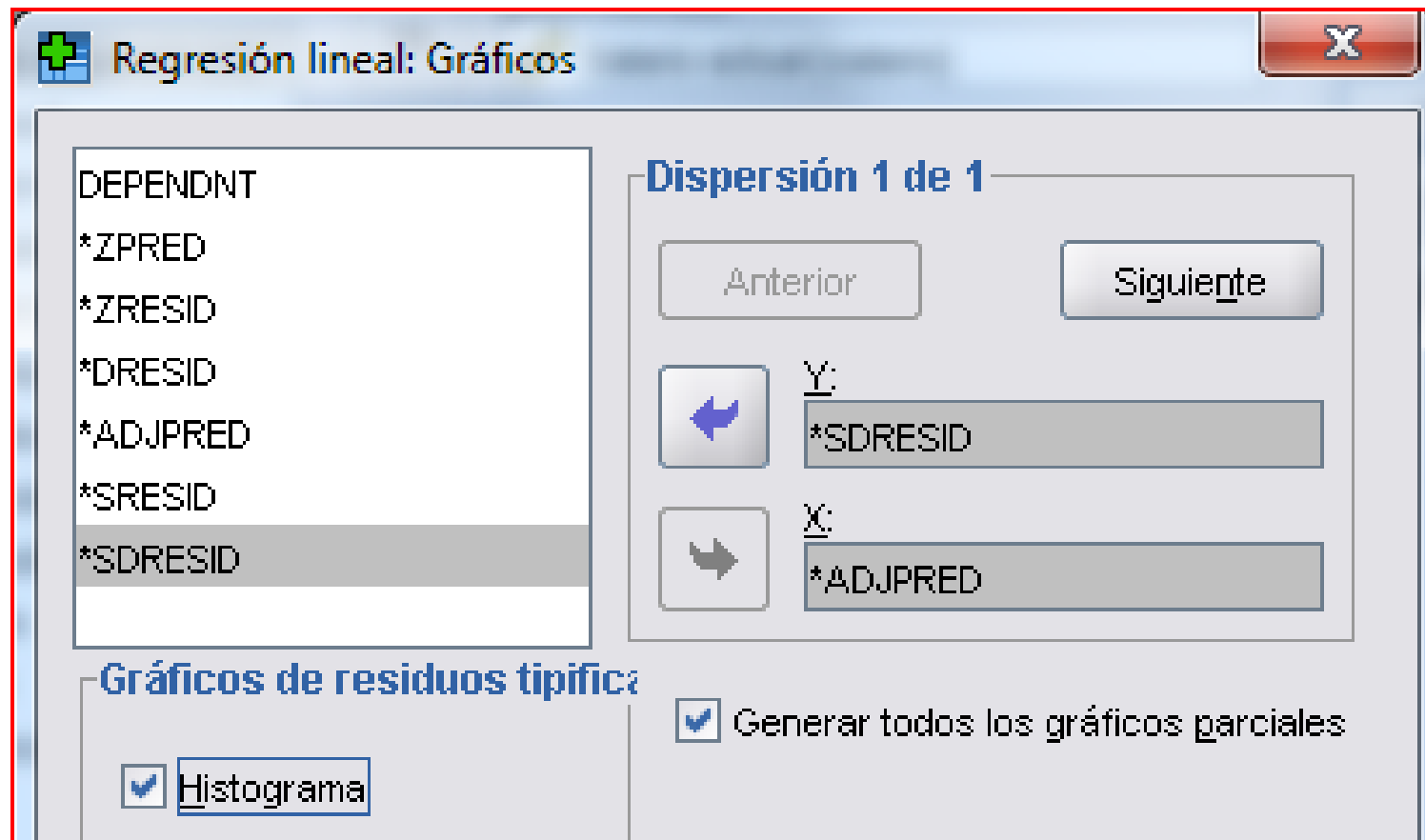
Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-36199,156	19146,093		-1,891	,059
	Sexo	-1681,168	766,574	-,049	-2,193	<b>,029</b>
	Fecha de nacimiento	2,118E-6	,000	,046	1,406	<b>,160</b>
	Nivel educativo	456,888	154,198	,077	2,963	<b>,003</b>
	Categoría laboral	5795,987	622,090	,262	9,317	<b>,000</b>
	Salario inicial	1,337	,070	,616	19,084	<b>,000</b>
	Meses desde el contrato	153,337	31,653	,090	4,844	<b>,000</b>
	Experiencia previa (meses)	-15,306	5,479	-,094	-2,793	<b>,005</b>

a. Dependent Variable: Salario actual

- He puesto en **negrita** los valores que examinaríamos en un primer momento:
  - R square (valores cercano a 1)
  - Nivel de significación de ANOVA (<0.001)
  - Niveles de significación de las variables (todos menos de 0.05 excepto la fecha de nacimiento)
- En general, todo esto sugiere un buen modelo pero antes de pasar a la interpretación necesitamos realizar el diagnóstico

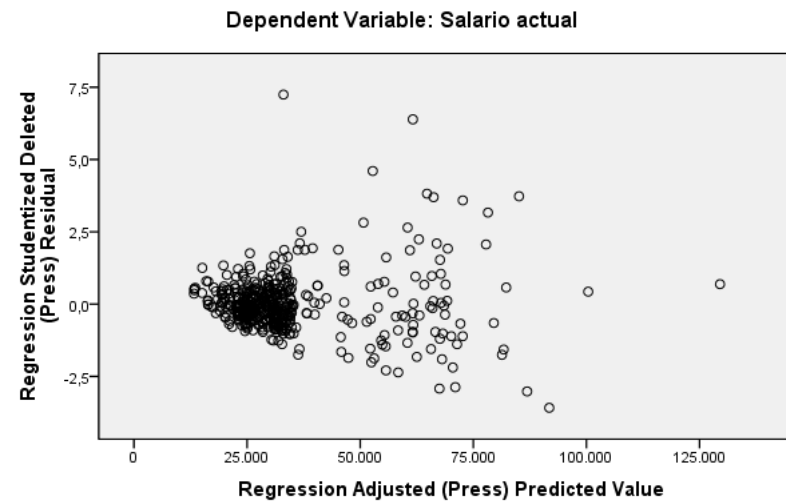
# Diagnóstico del modelo

- Un primer diagnóstico es examinar las puntuaciones predichas frente a los residuales studentizados, así como un histograma de los residuales

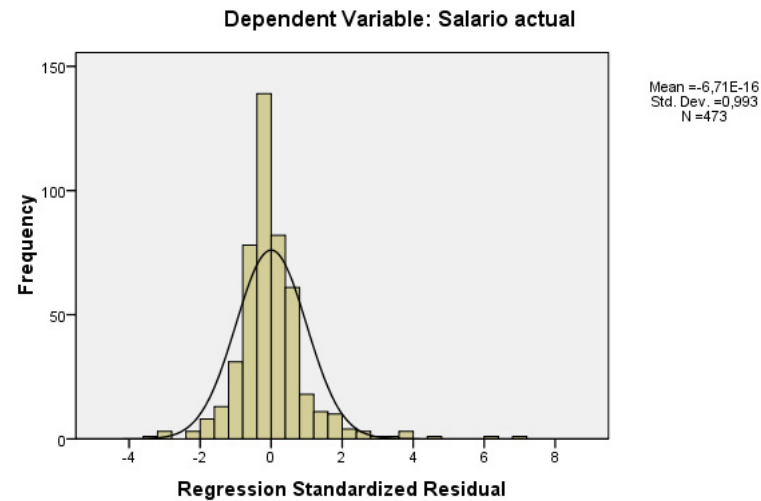




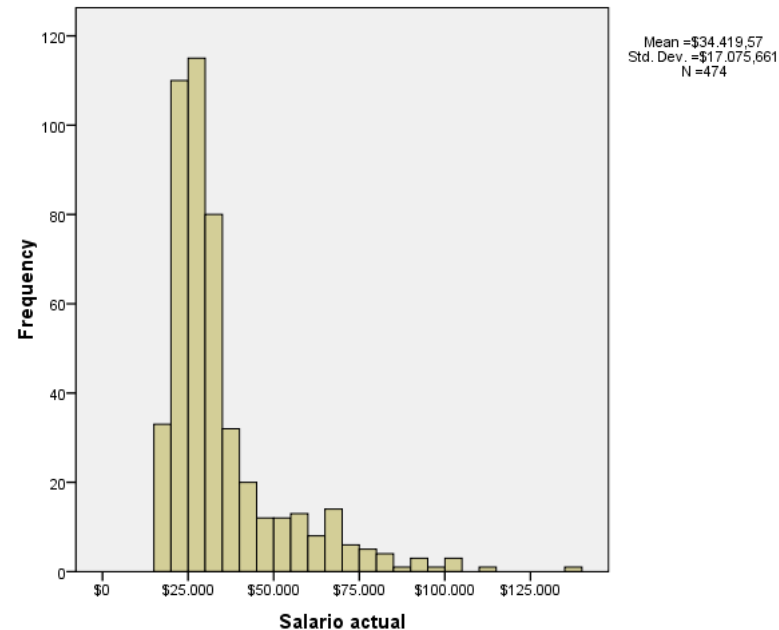
- El gráfico muestra un aspecto poco adecuado, con poca homogeneidad de varianza



- El histograma de los residuales muestra (también lo podemos ver en el de antes) que hay varios residuales muy grandes (estos valores se pueden interpretar como puntuaciones típicas)



- En realidad, un histograma de la variable dependiente ya nos habría dado las pistas de que esto iba a ocurrir



En él vemos que la variable Salario actual es muy asimétrica, con una mayoría cobrando salarios bajos y sólo unos pocos cobrando salarios más altos.

Estos gráficos sugieren que el análisis de regresión está incumpliendo el supuesto de residuales aproximadamente normales y de homogeneidad de varianza. Esto puede verse como una consecuencia de que la variable dependiente no es normal (aunque es posible en ocasiones que la variable dependiente no sea normal y se cumpla el supuesto de normalidad de los residuales).

# Transformaciones

- Transformaciones de BoxCox
  - Son una familia de transformaciones que permiten mejorar la linealidad y la normalidad de las variables

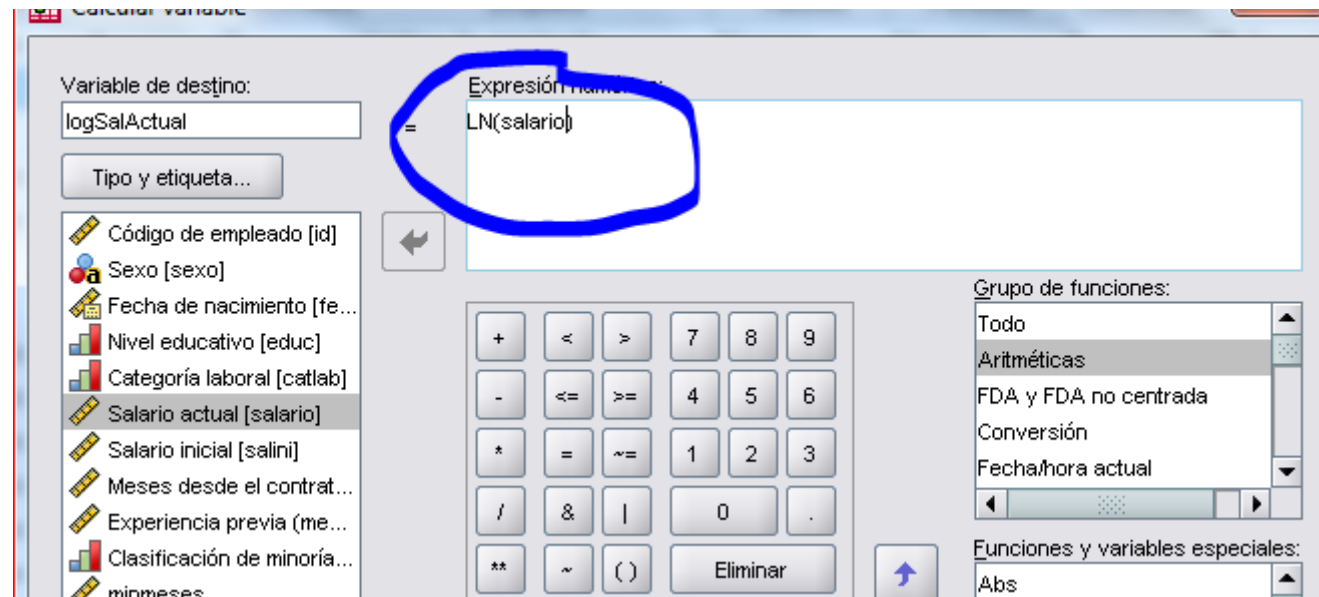
$$f(y) = \begin{cases} (y^p - 1) / p & \text{for } p \neq 0 \\ \log(y) & \text{for } p = 0 \end{cases}$$

- Algunos casos especiales de esta transformación son los siguientes

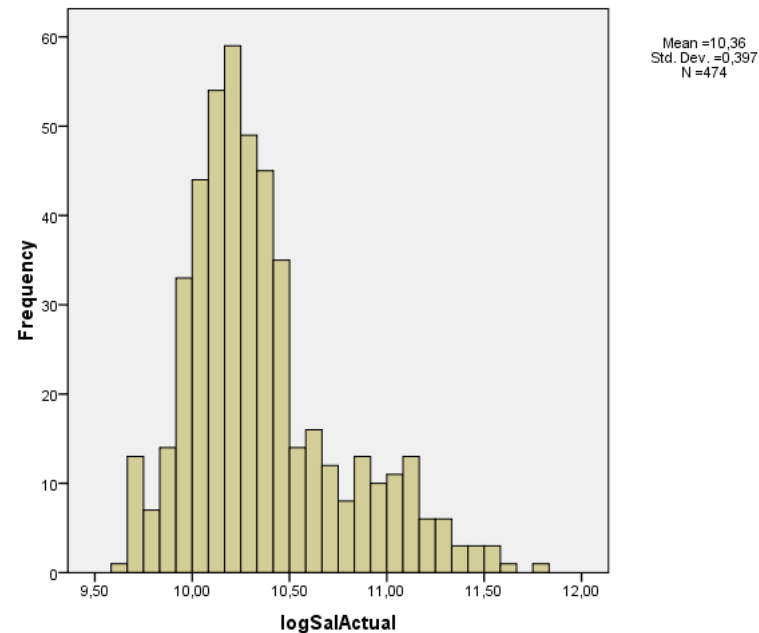
Special Members of the Box–Cox Transformation Family

Parameter Value	Transformation Name	Equation
-1.0	Reciprocal	$-1 / y$
0.0	Natural Log	$\log(y)$
0.5	Square Root	$\sqrt{y}$
1.0	Identity (no transformation)	$y$
2.0	Square	$y^2$

- La más útil de todas es la transformación logarítmica. Veamos lo que ocurre aplicada a nuestro caso.



- El logaritmo del salario actual tiene un histograma mucho más normal aunque todavía se puede observar una ligera asimetría. Quizás otra transformación sería todavía más efectiva pero nos conformaremos.



- En rigor, las variables independientes no es necesario que sean normales pero todo resulta más simple si lo hacemos en algunos casos. En este caso también calcularemos el logaritmo del salario inicial y lo utilizaremos en los análisis

## Resultados utilizando variables transformadas

Model Summary<sup>b</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,927 <sup>a</sup>	,859	,857	,15029

a. Predictors: (Constant), logSalInicial, Meses desde el contrato, Experiencia previa (meses), Sexo, Nivel educativo, Categoría laboral, Fecha de nacimiento

b. Dependent Variable: logSalActual

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	64,161	7	9,166	405,798	,000 <sup>a</sup>
	Residual	10,503	465	,023		
	Total	74,664	472			

a. Predictors: (Constant), logSalInicial, Meses desde el contrato, Experiencia previa (meses), Sexo, Nivel educativo, Categoría laboral, Fecha de nacimiento

b. Dependent Variable: logSalActual

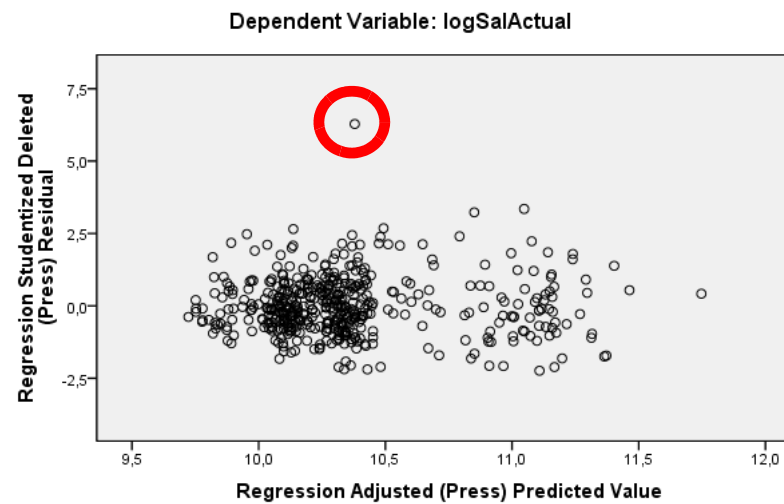


Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,312	,577		2,276	,023
	Sexo	-,042	,018	-,053	-2,369	,018
	Fecha de nacimiento	1,426E-10	,000	,133	4,298	,000
	Nivel educativo	,010	,004	,075	2,883	,004
	Categoría laboral	,123	,014	,238	8,586	,000
	Meses desde el contrato	,004	,001	,113	6,379	,000
	Experiencia previa (meses)	,000	,000	-,045	-1,414	,158
	logSalInicial	,699	,041	,621	17,252	,000

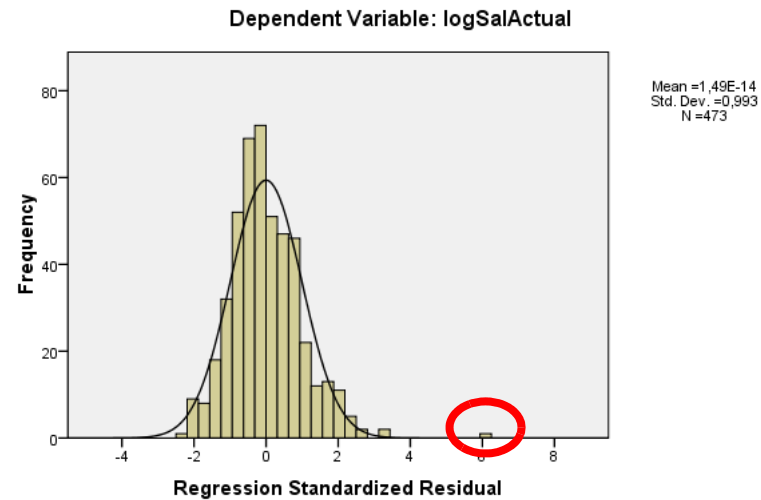
a. Dependent Variable: logSalActual

- El ajuste es un poco mejor (aunque los resultados no son comparables directamente, se trata de variables diferentes).
- Las variables que entran en el modelo han cambiado. La fecha de nacimiento es significativa y la experiencia previa no (además, el coeficiente es negativo lo que sugeriría que más experiencia lleva a menos sueldo). De todos modos, haremos la interpretación de los coeficientes más adelante.

- El gráfico de residuales frente a la variable dependiente aparece mejor (aunque todavía destaca la concentración en salarios bajos y un residual muy alto)



- El histograma de los residuales parece bastante normal (aunque el residual está ahí)



## ***Aplicaciones de las transformaciones***

- Las transformaciones permiten
  - Hacer más simétricas las variables individualmente
  - Corregir la falta de homogeneidad
  - Hacer las relaciones más lineales
- Aunque en rigor son las variables dependientes las que necesitan la transformación para cumplir los supuestos, transformar las variables independientes es también una buena idea

## Actividades

---

1. Calcular la fórmula para predecir el porcentaje de grasa corporal a partir de tres medidas simples (altura, cintura y pecho). Archivo Ch29 Body Fat. Examinar los gráficos para ver si el modelo ajusta bien y no hay valores especiales.  
*Este es un ejemplo sencillo. Ojalá todos fueran iguales.*

2. Unos estudiantes hacen una serie de exámenes intermedios, un proyecto y luego hacen un examen final, ¿se puede predecir el resultado del examen final a partir de las otras actividades? ¿qué actividades parecen no tener importancia? Este ejemplo está en Ch29 Grades.

*Se puede aplicar stepwise para elegir el modelo*

3. Infant mortality en función del número de niños recién nacidos por 1000 vivos. Es un indicador de calidad de la atención médica. Los datos están para estados de USA. Las variables disponibles son Infantmortality99, Child Deaths (muertes por 100000 para niños de 1 a 14), HSDrop porcentaje de adolescentes (16-19) que abandonan el instituto, LowBW porcentaje de bebés con peso bajo al nacer, TeenBirths (nacimientos por 100000 mujeres adolescentes entre 15-17) y TeenDeaths por accidentes, homicidio y suicidio por 100000.

*De nuevo, un ejemplo bastante sencillo.*

4. Tenemos la tasa de asesinatos por 100000 habitantes (Murder), la tasa de graduación en el instituto en porcentaje (HSGrad), ingreso per capita en dolares (Income) la tasa de analfabetismo por 1000 (illiteracy), y la expectativa de vida (lifeexpect). Encuentra un modelo de regresión para la expectativa de vida con tres variables predictoras intentando los cuatro posibles modelos. Haz las comparaciones de modelos apropiadas para demostrar que ese es el modelo correcto. Los datos están en Ch29.Fifty\_states.

*Asegurarse también que se cumplen los supuestos*



5. Carreras campo a través . Tenemos los valores de los records para los hombres y mujeres en una serie de carreras campo a través y los datos acerca de la distancia recorrida (Distance) y lo que se escala en ellas (Climb). Calcula las ecuaciones de regresión para hombres y mujeres y compara. Examina los supuestos de la regresión.

*En este caso resulta interesante una transformación. Examinar los residuales. ¿Sería interesante usar la interacción entre las variables independientes como predictor?*

# Interpretación del modelo

- La fuente principal es la tabla de coeficientes

Coefficients <sup>a</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	
1	(Constant)	1,312	,577		2,276
	Sexo	-,042	,018	-,053	-2,369
	Fecha de nacimiento	1,426E-10	,000	,133	4,298
	Nivel educativo	,010	,004	,075	2,883
	Categoría laboral	,123	,014	,238	8,586
	Meses desde el contrato	,004	,001	,113	6,379
	Experiencia previa (meses)	,000	,000	-,045	-1,414
	logSalInicial	,699	,041	,621	17,252

a. Dependent Variable: logSalActual

- Esta tabla se puede plantear como una fórmula que va sumando
  - 1,312 es el logaritmo del salario actual medio de alguien que tuviera cero en todas las demás variables
  - Ser Hombre es 1 y ser Mujer es 2 (equivale a 0 y 1), luego ser mujer baja -0.42 del salario
  - Fecha de nacimiento: cada día trabajado sube un poquito

- Nivel educativo: Cada nivel sube 0.1
- Categoría laboral: Cada nivel sube 0.123
- Meses desde el contrato 0.004
- Experiencia previa: No sube nada
- logSalInicial: sube 0.699

## *La importancia de los coeficientes*

- Lo anterior, no obstante, no aclara la importancia de cada una de las variables independientes ya que estos coeficientes están afectados por la unidad de medida. Por ejemplo, la fecha de nacimiento tiene un coeficiente muy pequeño pero si tenemos en cuenta que se refiere a cada día, vemos que no tiene mucho sentido. Si hicieramos el cálculo por año trabajado el coeficiente cambiaría bastante.
- Para determinar si una variable es importante podemos tener en cuenta dos aspectos:
  - Cómo de importante es la relación de cada variable por sí sola con la variable dependiente? Esto puede calcularse con coeficientes de correlación simples
  - Cómo de importante es la relación de cada variable con la variable dependiente cuando es usada junto con las otras variables independientes? Esto corresponde con la situación actual. Para este caso usaremos coeficientes estandarizados o coeficientes de correlación parcial

## Coeficientes estandarizados

- Una forma de hacer los coeficientes más comparables es realizar algún tipo de estandarización. Esto es lo que se hace en la tercera columna de la tabla de coeficientes:

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	1,312	,577		2,276	,023
Sexo	-,042	,018	<b>-,053</b>	-2,369	,018
Fecha de nacimiento	1,426E-10	,000	<b>,133</b>	4,298	,000
Nivel educativo	,010	,004	<b>,075</b>	2,883	,004
Categoría laboral	,123	,014	<b>,238</b>	8,586	,000
Meses desde el contrato	,004	,001	<b>,113</b>	6,379	,000
Experiencia previa (meses)	,000	,000	<b>-,045</b>	-1,414	,158
logSalInicial	,699	,041	<b>,621</b>	17,252	,000

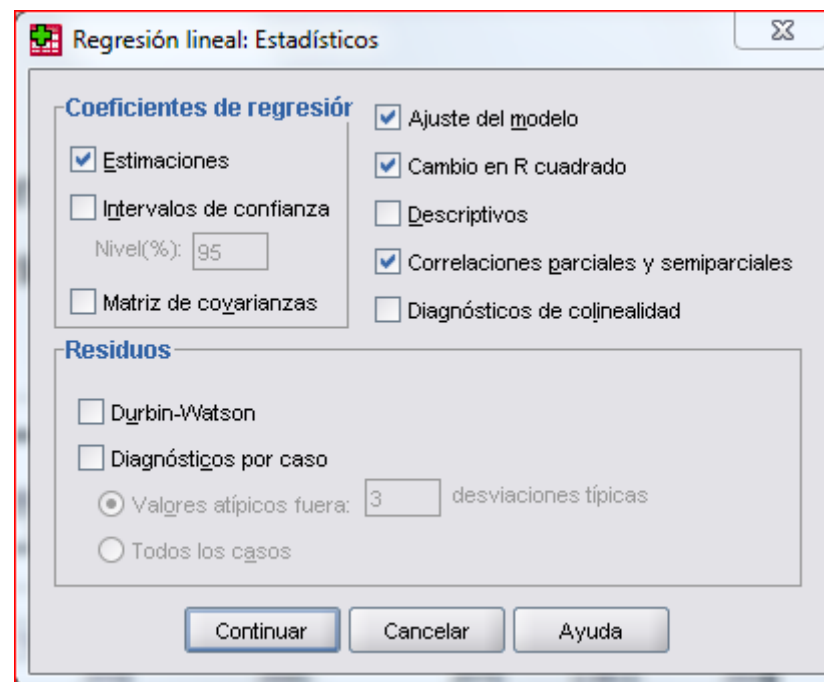
a. Dependent Variable: logSalActual

- Estos se calculan teniendo en cuenta las desviaciones típicas y se pueden interpretar como puntuaciones típicas. Así valores grandes en términos absolutos tendrían más efecto que los valores pequeños. Valores por encima de 2 o de 3 serían extraordinarios

- No obstante, los coeficientes estandarizados no tienen en cuenta las correlaciones entre las variables independientes y no reflejan de una manera absoluta la contribución de las variables independientes

Si dos variables independientes están correlacionadas entre sí, la contribución propia de cada una de ellas disminuye puesto que la contribución propia es “robada” por la otra

## *Coeficientes de correlación semiparcial y parcial*



- La tabla que obtenemos es la siguiente. Lo nuevo son las tres columnas del final

Coefficients <sup>a</sup>								
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
	B	Std. Error	Beta			Zero-order	Partial	Part
1 (Constant)	1,312	,577		2,276	,023			
Sexo	-,042	,018	-,053	-2,369	,018	-,517	-,109	-,041
Fecha de nacimiento	1,426E-10	,000	,133	4,298	,000	,213	,195	,075
Nivel educativo	,010	,004	,075	2,883	,004	,697	,132	,050
Categoría laboral	,123	,014	,238	8,586	,000	,775	,370	,149
Meses desde el contrato	,004	,001	,113	6,379	,000	,093	,284	,111
Experiencia previa (meses)	,000	,000	-,045	-1,414	,158	-,122	-,065	-,025
logInicial	,699	,041	,621	17,252	,000	,887	,625	,300

a. Dependent Variable: logSalActual



- Zero order: La correlación normal
- Part: El cuadrado de este valor indica el cambio en  $R^2$  que acarrea añadir esta variable a un modelo igual al ajustado pero sin esa misma variable. Esta variación puede considerarse la contribución única de esa variable y tiene importancia cuando las variables independientes están correlacionadas entre sí
  - El problema de este coeficiente es que no indica qué proporción de la varianza no explicada significa la variación anterior. Si el resto de las variables explican gran parte de la varianza, la varianza que queda no es mucha para la variable considerada. En resumen, comparar coeficientes de correlación semiparcial puede ser complicado.
- Partial: Los coeficientes de correlación parcial se pueden entender como la correlación entre una variable independiente  $X$  y la variable dependiente  $Y$  cuando la correlación del resto de las variables ha sido quitada tanto de  $X$  como de  $Y$ . De este modo, este coeficiente produce una medida de la relación “pura” entre  $X$  e  $Y$ .

## Actividades

---

1. Mental Health.SAV tiene datos sobre cuatro variables para un grupo de mujeres. La variable dependiente es el número de visitas a profesionales de la salud en función de síntomas de salud física, mental o acontecimientos vitales stressantes. Realiza transformaciones de las variables que sea necesario. En este caso, la raíz cuadrada puede ser una buena transformación para los datos.

*El objetivo es encontrar qué variables son buenas predictoras del número de visitas*

2. En EstadosExpVida.sav están variables de un archivo de datos que hemos visto antes. En este caso tenemos unas variables ficticias para dos estados. Si utilizamos estas variables como predictores tenemos una medida del efecto único de estas observaciones. Calcula el modelo e interpreta los resultados.

*Esta es una de las formas en las que puede tratarse la situación en la que hay casos destacados.*

3. En el archivo Tele.sav tenemos como variables predictoras el número de personas por televisión y el número de personales por médico, y como predichas la expectativa de vida. ¿Qué predice mejor la expectativa de vida, los médicos o las televisiones por país?

*En este caso es necesario hacer transformaciones.*

4. El sueldo de una profesión puede predecirse a partir de la educación necesaria y el prestigio. En este caso, se trata del tanto por ciento personas que cobran por encima de un determinado nivel de sueldo y que tienen esa profesión, el tanto por ciento de personas que consideran que esa profesión es de prestigio, y el tanto por ciento de personas que tuvo que superar un cierto nivel educativo y que trabajan en esa profesión. El archivo se llama profesiones.sav

*Este ejemplo tiene valores especiales muy interesantes y que es conveniente identificar.*

5. Abdominales. Predecir el número de abdominales a partir del peso y el tamaño de la cintura. El archivo se llama abdominales.sav

*Un ejemplo sencillo y que es interesante comprobar la interacción*

6. Bigmac. Selecciona el mejor modelo para predecir el precio de una hamburguesa Bigmac en capitales del mundo (tomado como un indicador de coste de la vida) a partir de una serie de indicadores.

*Este archivo es interesante no solo para regresión sino también para examinar algunas de las otras variables*

7. LipidData. Un grupo de sujetos se les midió en una serie de variables de salud. También hay medidas de salud tomadas hace tres años. Intenta encontrar un modelo para predecir el colesterol.

*Hay variables que en principio no están relacionadas biológicamente con el colesterol. ¿Es posible que sirvan aún así para predecirlo? El LDL y el HDL predicen muy bien, el colesterol ¿a qué se debe?*



- 
8. Repite el ejercicio anterior intentando predecir la tensión sistólica.  
*Lo mismo que el anterior con una variable dependiente diferente*

9. Repite lo mismo con la tensión diastólica

*Controla el valor destacado ¿A qué se puede deber?*

- 
10. Repite el ejercicio anterior intentando predecir los triglicéridos.  
*Examinar los residuales.*

***Explorando el  
salario inicial***

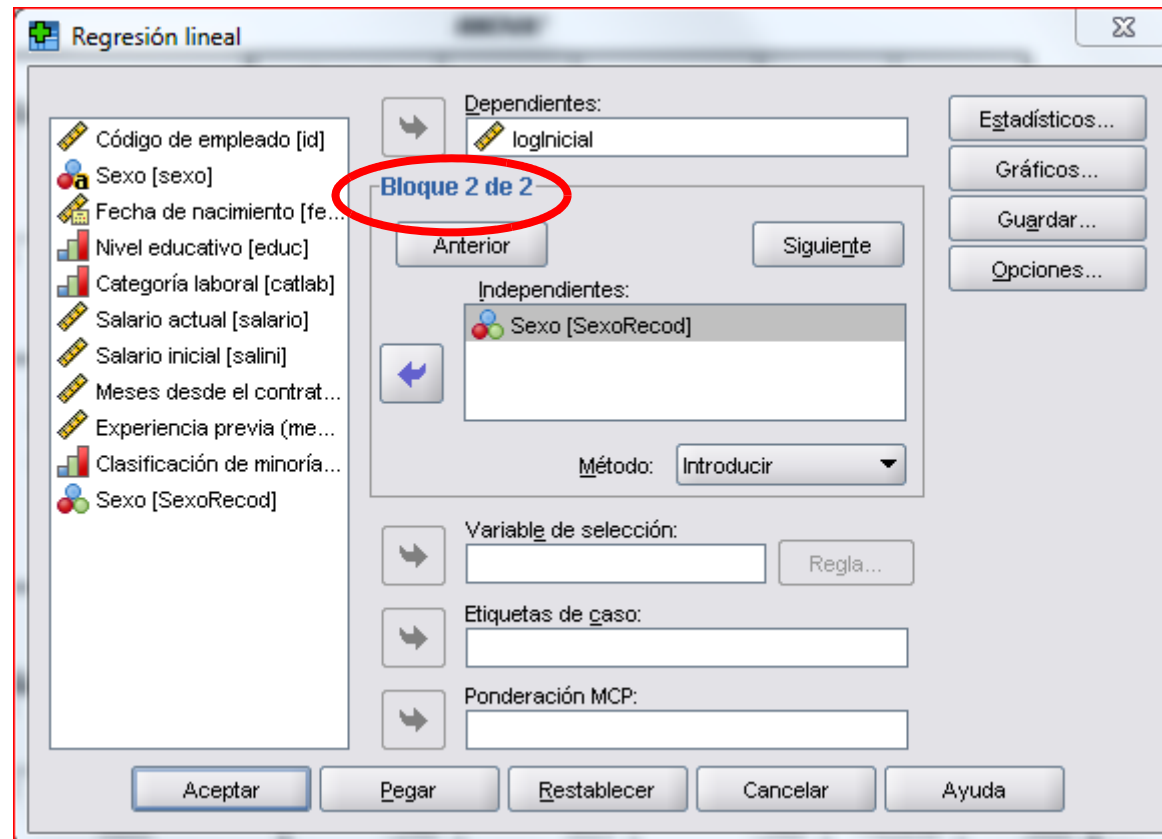
# Planteamiento del modelo

- En general, buscamos modelos que tengan pocas variables ya que son más simples
  - Hemos visto que el salario actual puede ponerse en relación con el salario inicial.
  - Es interesante a su vez ver si se puede poner en relación el salario inicial con los mismos factores
  - En este caso, además intentaremos construir un modelo que incluya un número de variables no demasiado grande. Sólo aquellas que aporten valor.

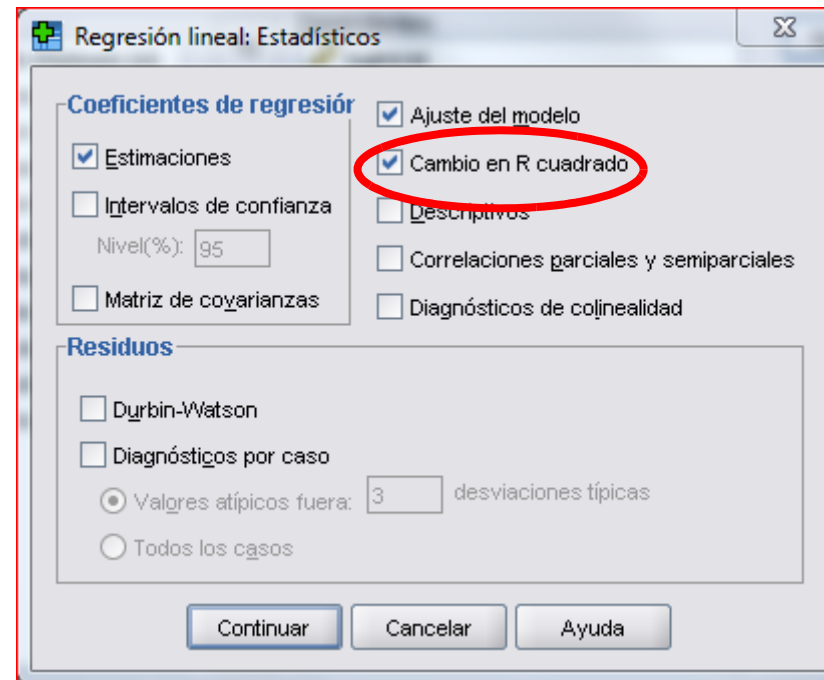
## ***Introduciendo variables una por una***

- Introducir las variables una por una según su importancia es una forma de ir construyendo el modelo. Para ello podemos usar el SPSS con la opción de comparar modelos

- Hay que introducir varios bloques



- En estadísticos hay que seleccionar cambio en R cuadrado





## *Cambio en $R$ cuadrado*

- La parte del output que es interesante en este caso es el cambio en  $R^2$ . Ese valor nos indica qué aporta el añadir una variable al modelo anterior y por tanto nos informa del efecto que tiene esa variable específica. Si el cambio es 0 entonces esa variable no aporta nada al modelo.

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,686a	<b>,470</b>	,469	,25709	,470	418,920	1	472	<b>,000</b>
2	,759b	<b>,576</b>	,574	,23025	,106	117,486	1	471	<b>,000</b>

a. Predictors: (Constant), Nivel educativo

b. Predictors: (Constant), Nivel educativo, Sexo

- Vemos que hay dos modelos, el primero con una  $R$  cuadrado de .470 y el segundo de .576.
- La significación del cambio nos indica que el segundo modelo es diferente del primero (Sig. F Change .000)

- Hay que tener en cuenta que cuando se introducen variables independientes muy intercorrelacionadas en un modelo, los resultados pueden parecer muy anómalos. La regresión en total puede aparecer significativa mientras que ninguno de los coeficientes lo es. En ese caso, lo más razonable es quitar algunas de las variables con altas intercorrelaciones.
- Otro elemento que hay que prestar atención es la R cuadrado ajustada. Esta fórmula penaliza introducir variables que no aportan nada. Por eso, añadir una variable puede aumentar la R cuadrado pero hacer disminuir la R cuadrado ajustada. En este caso vemos que la diferencia es mínima.

## *Examinando varias variables*

- Quitar y poner variables en el modelo puede ser bastante largo pero es necesario para encontrar modelos que tiene un conjunto de variables con coeficientes significativos.
  - He puesto las siguientes variables Nivel educativo, sexo, experiencia previa, clasificación de minorías, fecha de nacimiento.

Model Summary									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,547 <sup>a</sup>	,300	,298	,29575	,300	201,555	1	471	,000
2	,588 <sup>b</sup>	,346	,343	,28614	,046	33,169	1	470	,000
3	,769 <sup>c</sup>	,592	,589	,22631	,246	282,368	1	469	,000
4	,783 <sup>d</sup>	,613	,609	,22065	,021	25,350	1	468	,000
5	,784 <sup>e</sup>	,615	,611	,22022	,002	2,836	1	467	,093

a. Predictors: (Constant), Sexo

b. Predictors: (Constant), Sexo, Clasificación de minorías

c. Predictors: (Constant), Sexo, Clasificación de minorías, Nivel educativo

d. Predictors: (Constant), Sexo, Clasificación de minorías, Nivel educativo, Experiencia previa (meses)

e. Predictors: (Constant), Sexo, Clasificación de minorías, Nivel educativo, Experiencia previa (meses), Fecha de nacimiento

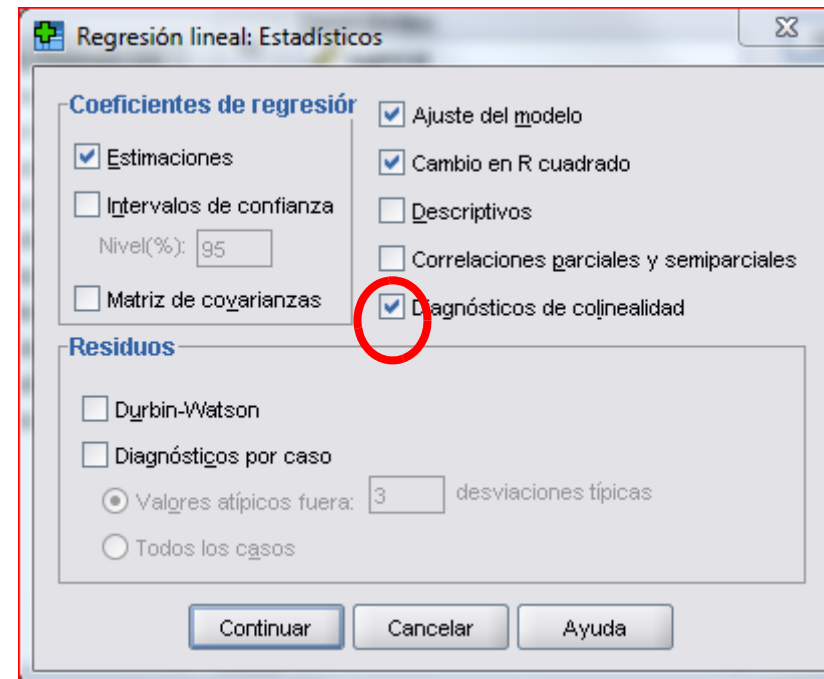
- Vemos que la última variable no entraría en el modelo por muy poco y que el aumento de la R cuadrado es mínimo. Si examinamos los modelos:

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10,233	,042		243,496	,000
	Sexo	-,388	,027	-,547	-14,197	,000
2	(Constant)	10,291	,042		245,807	,000
	Sexo	-,399	,026	-,564	-15,073	,000
	Clasificación de minorías	-,183	,032	-,215	-5,759	,000
3	(Constant)	9,183	,074		124,447	,000
	Sexo	-,259	,023	-,366	-11,503	,000
	Clasificación de minorías	-,110	,026	-,129	-4,303	,000
	Nivel educativo	,066	,004	,538	16,804	,000
4	(Constant)	9,001	,081		111,805	,000
	Sexo	-,228	,023	-,322	-9,994	,000
	Clasificación de minorías	-,120	,025	-,141	-4,812	,000
	Nivel educativo	,072	,004	,592	17,938	,000
	Experiencia previa (meses)	,001	,000	,157	5,035	,000
5	(Constant)	9,996	,596		16,765	,000
	Sexo	-,240	,024	-,338	-10,080	,000
	Clasificación de minorías	-,121	,025	-,142	-4,835	,000
	Nivel educativo	,072	,004	,592	17,984	,000
	Experiencia previa (meses)	,000	,000	,086	1,636	,102
	Fecha de nacimiento	-8,100E-11	,000	-,085	-1,684	,093

a. Dependent Variable: logInicial

- Vemos que la experiencia previa y la fecha de nacimiento no entrarían en el último modelo. Sin embargo, un modelo sólo con experiencia previa y sin fecha de nacimiento tiene un coeficiente significativo
- Esto posiblemente se debe a que la experiencia previa está relacionada con la edad (gente con más edad tiene más posibilidades de tener más experiencia previa)
- Una forma de ver la relación entre las variables independientes es calcular la correlación (en este caso es  $-0.80$ , cuantas más experiencia mas baja es la fecha de nacimiento, es decir menos edad).
- No obstante, la correlación bivariada entre variables puede no ser suficiente para diagnosticar porqué una variable entra o no en un modelo ya que puede haber casos en que la correlación múltiple es alta pero la bivariada no es tanto. Una forma de medir esto es mediante las medidas de colinealidad (otra nombre para correlación)

## *Medidas de colinealidad*



- Los diagnosticos de colinealidad aparecen de dos maneras
  - Junto a los coeficientes de la regresión
  - Como una tabla de diagnósticos de la colinealidad relacionada con eigenvalores

- Diagnósticos junto a los coeficientes (variables excluidas)

Coefficients <sup>a</sup>								
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	9,996	,596		16,765	,000		
	Sexo	-,240	,024	-,338	-10,080	,000	,732	1,367
	Clasificación de minorías	-,121	,025	-,142	-4,835	,000	,958	1,043
	Experiencia previa (meses)	,000	,000	,086	1,636	,102	,301	3,320
	Nivel educativo	,072	,004	,592	17,984	,000	,761	1,315
	Fecha de nacimiento	-8,100E-11	,000	-,085	-1,684	,093	,321	3,112

a. Dependent Variable: logInicial

- Tolerancia: Es  $1 - R^2$  cuadrado múltiple de esa variable con todas las demás. Cuanto más alto (cerca de uno) más independiente es esa variable (menos colineal).
- VIF: Es el recíproco de la tolerancia. Cuanto más grande más colinealidad.
- Vemos que la tolerancia de la experiencia previa y la fecha de nacimiento son las más bajas



- La parte de diagnosticos propiamente de la colinealidad (sólo modelo 4)

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				(Constant)	Sexo	Clasificación de minorías	Experiencia previa (meses)	Nivel educativo	Fecha de nacimiento
1	1	4,628	1,000	,00	,00	,01	,00	,00	,00
	2	,752	2,481	,00	,00	,86	,01	,00	,00
	3	,499	3,046	,00	,01	,10	,28	,00	,00
	4	,104	6,657	,00	,48	,00	,00	,10	,00
	5	,016	16,923	,00	,40	,03	,03	,89	,01
	6	,000	174,635	1,00	,11	,00	,68	,00	,99

a. Dependent Variable: logInicial

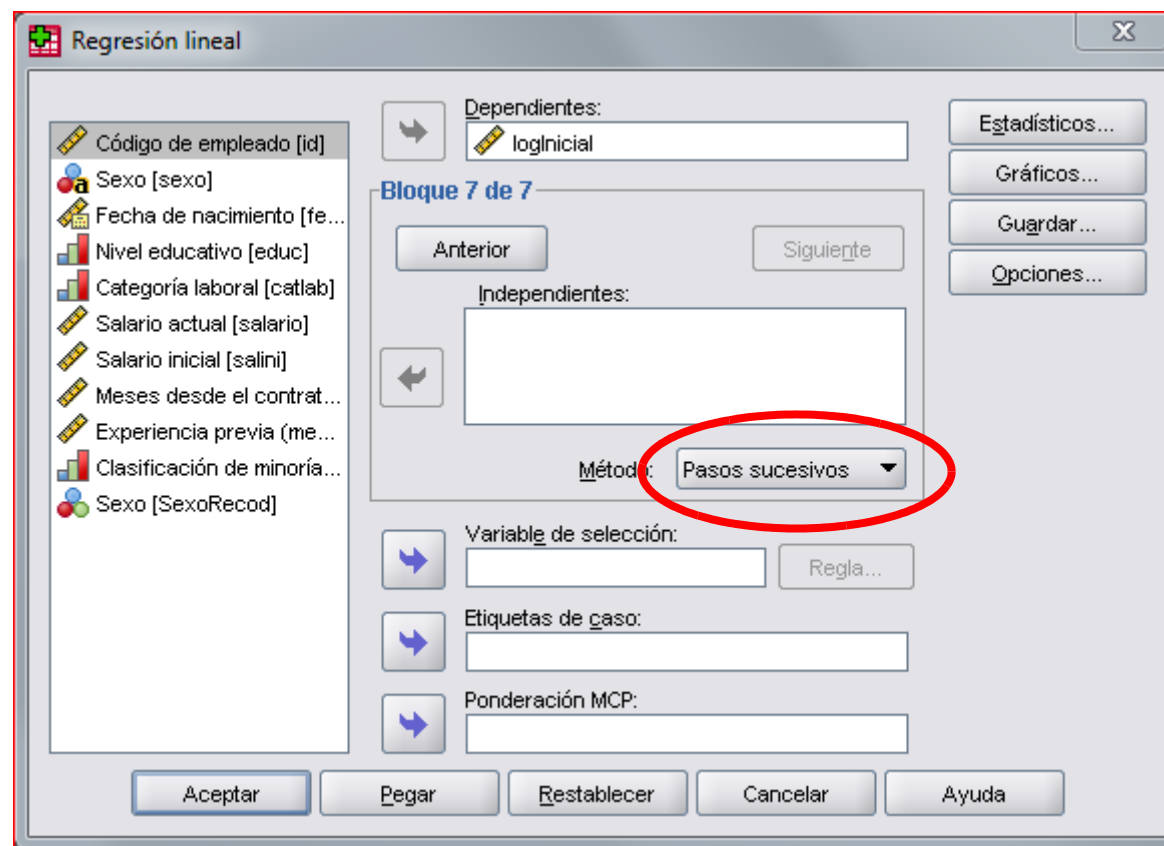
- Un resultado deseable aquí es que los eigenvalores sean aproximadamente iguales. Si algunos eigenvalores son más grandes que otros, entonces significa que la matriz de variables independientes tiene muchas redundancias (alta colinealidad)
- Una forma de medir esto es el índice de condicionalidad que se obtiene mediante la fórmula. Valores grandes son indicadores de variables redundantes

$$condindex = \sqrt{(EIGENVALOR_{max}) / (EIGENVALOR)}$$

- **Proporciones de varianza:** Esta parte de la tabla te ayuda a encontrar las variables que están muy relacionadas entre sí. En nuestro caso, la fecha de nacimiento y la experiencia previa están muy relacionadas y así aparecen en el output.

## Métodos automáticos

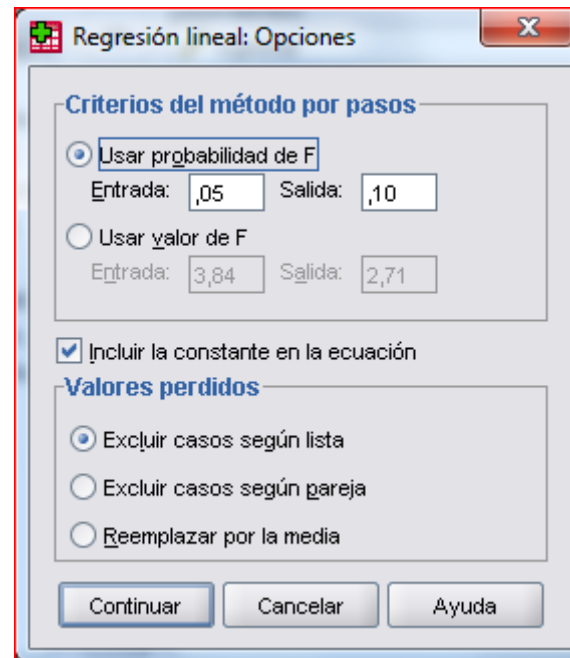
- Todo lo anterior puede parecer muy complicado pero existen métodos automáticos que se encargan de hacer la selección automáticamente entre las variables. El método más común es stepwise (pasos sucesivos)



- El método de pasos sucesivos es un método automático de selección de variables que combina la introducción y la eliminación de variables en el modelo según unos criterios de selección. Otros métodos son adelante (se va seleccionando la variable mejor en cada paso según varios criterios), hacia atrás (se van eliminando variables hacia atrás) y eliminar (se utilizan bloques y se van eliminando los bloques sucesivamente)
- Los métodos automáticos no suelen ser recomendados en la literatura pero entre ellos stepwise es el más avanzado. A continuación mostraremos un ejemplo de stepwise.

## Stepwise

- En este modelo ponemos las siguientes variables como predictoras: Género, minoría, experiencia previa, meses desde el contrato, fecha nacimiento.



- El criterio para el método es el de que una variable entra si la diferencia en la probabilidad de F es menor que 0.05 y de salida si incorporarla supone una probabilidad mayor que 0.10.

- Esas pruebas se repiten a cada paso así que una variable puede no entrar en un paso pero hacerlo en el siguiente. En nuestro caso, el modelo final es:

Model Summary <sup>e</sup>									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,685 <sup>a</sup>	,470	,469	,25736	,470	417,151	1	471	,000
2	,759 <sup>b</sup>	,576	,574	,23048	,106	117,246	1	470	,000
3	,771 <sup>c</sup>	,594	,591	,22565	,018	21,341	1	469	,000
4	,783 <sup>d</sup>	,613	,609	,22061	,019	22,673	1	468	,000

a. Predictors: (Constant), Nivel educativo

b. Predictors: (Constant), Nivel educativo, Sexo

c. Predictors: (Constant), Nivel educativo, Sexo, Fecha de nacimiento

d. Predictors: (Constant), Nivel educativo, Sexo, Fecha de nacimiento, Clasificación de minorías

e. Dependent Variable: logInicial

- En este modelo no ha entrado la variable Experiencia previa. Una forma de ver el proceso es mediante esta tabla que muestra las variables no introducidas:

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Sexo	-,348 <sup>a</sup>	-10,828	,000	-,447	,874	1,144	,874
	Clasificación de minorías	-,083 <sup>a</sup>	-2,469	,014	-,113	,983	1,018	,983
	Experiencia previa (meses)	,228 <sup>a</sup>	6,879	,000	,302	,937	1,068	,937
	Fecha de nacimiento	-,160 <sup>a</sup>	-4,669	,000	-,211	,921	1,086	,921
2	Clasificación de minorías	-,129 <sup>b</sup>	-4,303	,000	-,195	,965	1,036	,849
	Experiencia previa (meses)	,144 <sup>b</sup>	4,549	,000	,206	,862	1,160	,775
	Fecha de nacimiento	-,142 <sup>b</sup>	-4,620	,000	-,209	,918	1,089	,805
3	Clasificación de minorías	-,140 <sup>c</sup>	-4,762	,000	-,215	,960	1,041	,788
	Experiencia previa (meses)	,075 <sup>c</sup>	1,400	,162	,065	,302	3,314	,302
4	<b>Experiencia previa (meses)</b>	<b>,086<sup>d</sup></b>	<b>1,636</b>	<b>,102</b>	<b>,075</b>	<b>,301</b>	<b>3,320</b>	<b>,301</b>

a. Predictors in the Model: (Constant), Nivel educativo

b. Predictors in the Model: (Constant), Nivel educativo, Sexo

c. Predictors in the Model: (Constant), Nivel educativo, Sexo, Fecha de nacimiento

d. Predictors in the Model: (Constant), Nivel educativo, Sexo, Fecha de nacimiento, Clasificación de minorías

e. Dependent Variable: logInicial

## *Advertencias sobre stepwise*

- Hay veces que excluir una variable que tiene un coeficiente no significativo no es buena idea porque si no lo introducimos otra variable no es significativa (no tengo un ejemplo de esto pero a veces pasa)
- Hay variables que necesitamos que entren en el modelo porque son las que tienen más importancia teórica.
- Hay variables que están intercorrelacionadas y que podemos decidir en base a nuestra opinión cuál hay que quitar (en lugar de dejar al ordenador que lo haga)
- Mirar los distintos índices nos puede ayudar a entender mejor nuestros análisis.
- Un listado de problemas puede encontrarse en este [link](#)



## *El modelo final*

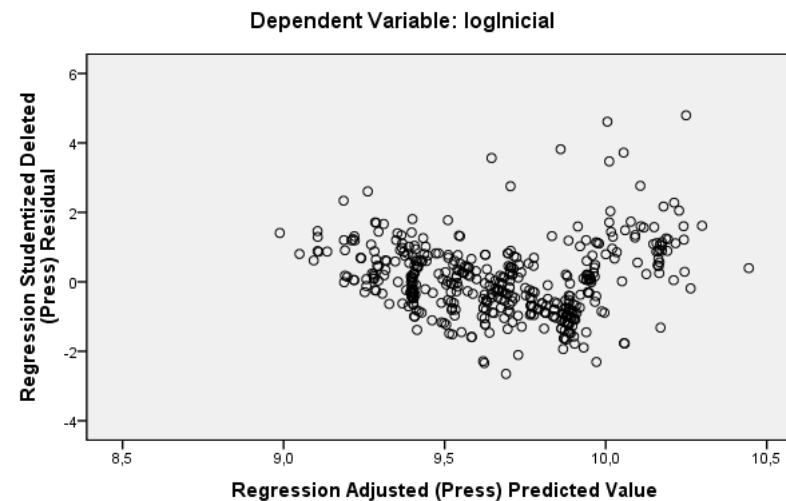
- Vamos a quedarnos con el siguiente modelo en el que todas las variables tienen coeficientes significativos
  - $\text{LogSalini} = \text{Const} + \text{Sexo} + \text{Experiencia previa} + \text{Clasificación de minorías}$
  - La variable Fecha de Nacimiento no ha sido incluida porque está muy relacionada con la Experiencia previa y ambas no podían estar simultáneamente en el modelo. He elegido Experiencia previa porque me parece una variable con más sentido que la Fecha de Nacimiento aunque están muy correlacionadas

# Diagnóstico del modelo

- El modelo anterior necesita ser diagnosticado antes de darlo por bueno. Hay muchos gráficos que pueden ser útiles:
  - El gráfico de residuales frente a la variable predicha da una visión general de los posibles problemas
  - El gráfico de regresión parcial da la información acerca de la relación específica entre una variable independiente y la variable dependiente después de haber eliminado la influencia de las otras variables
  - Histogramas o otros gráficos de residuales, puntos influyentes y distancias de Cook

## *Residuales frente a predicha*

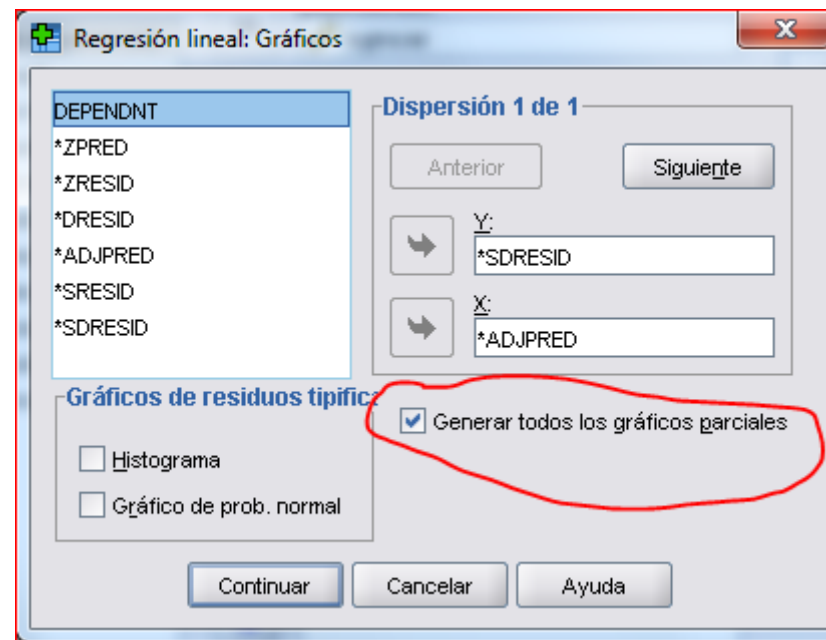
- El gráfico en nuestro caso muestra curvilinealidad



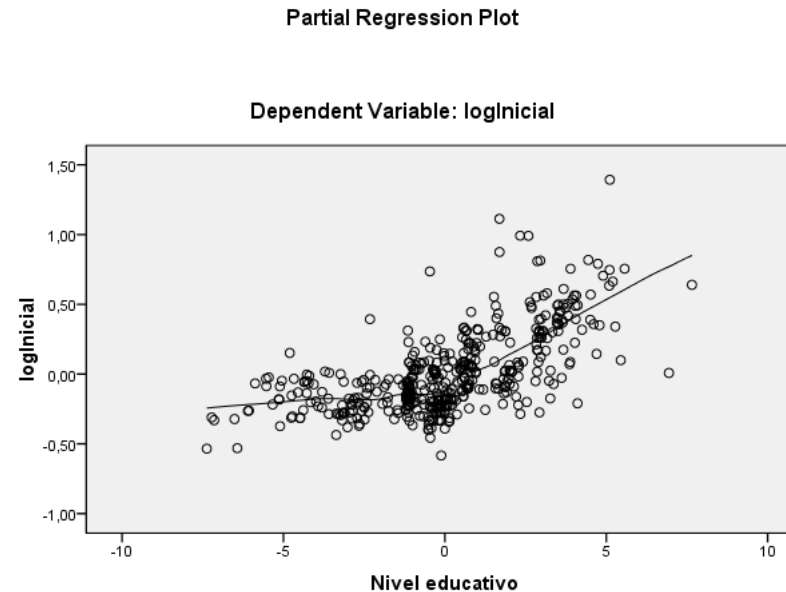
- Para explorar esta curvilinealidad se pueden utilizar los gráficos de regresión parcial.

## *Gráficos de regresión parcial*

- Esos gráficos son los residuales de una regresión que no incluye a la variable independiente considerada frente a los residuales de una regresión múltiple del resto de las variables independientes sobre la variable considerada.
- Estos gráficos es el equivalente a la correlación parcial y da una idea de la relación pura entre las variables



- De todos ellos, el más interesante es el de nivel educativo



- En este gráfico se puede ver una relación de curvilinealidad que sería conveniente ajustar.
- Una forma de tratar con esta curvilinealidad es añadir un término polinomial (nivel educativo al cuadrado) usando calcular variable

Variable de destino:	Expresión numérica:
educuad	educ * educ

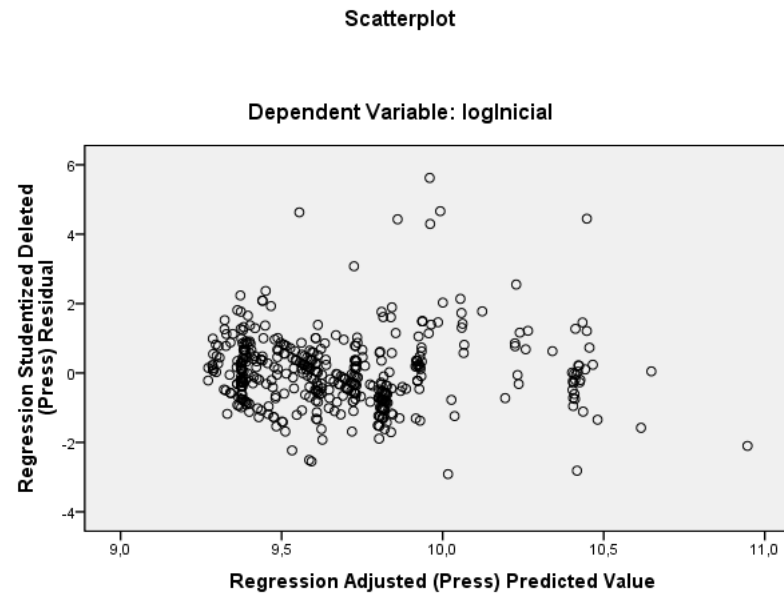
- El nuevo modelo da como resultado una R cuadrado de .711 (el anterior era .611). La tabla de coeficientes es la siguiente

Coefficients <sup>a</sup>							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	10,717	,153		69,914	,000		
Sexo	-,190	,020	-,268	-9,515	,000	,777	1,287
Clasificación de minorías	-,088	,022	-,103	-4,041	,000	,945	1,058
Experiencia previa (meses)	,000	,000	,102	3,748	,000	,834	1,198
Nivel educativo	-,206	,022	-1,687	-9,191	,000	,018	54,510
educuad	,011	,001	2,312	12,568	,000	,018	54,756

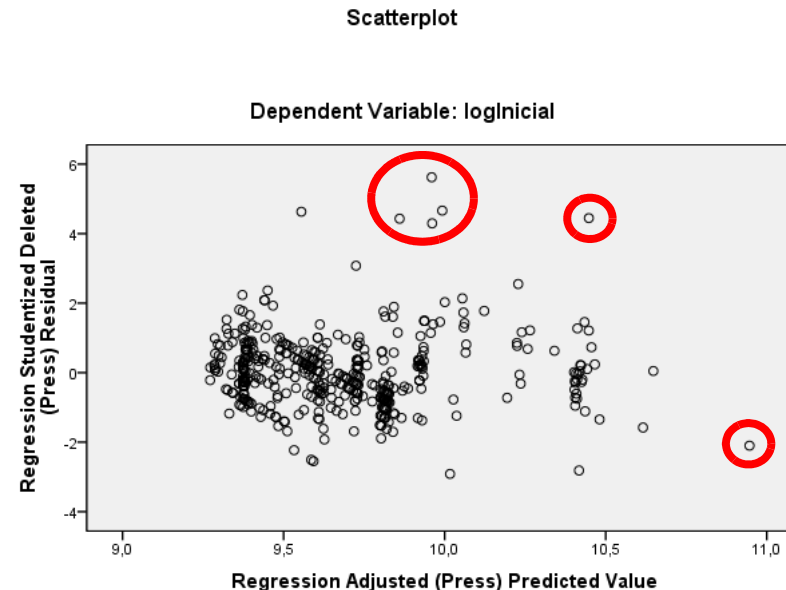
a. Dependent Variable: logInicial

- Tanto nivel educativo como educativo aparecen como significativos.
- No obstante, la tolerancia de las dos variables es muy baja. Esto se debe a que están muy correlacionadas.
- En este caso, esto no nos debe preocupar ya que trataremos ambas variables como si se tratara de un conjunto (lo veremos más adelante en interpretación)

- El gráfico de las predichas frente a los residuales no muestra curvilinealidad (aunque sí que aparecen otras cosas interesantes que estudiaremos ahora)



## *Análisis de los valores individuales*



- En este gráfico se pueden observar una serie de observaciones que destacan del resto. Esas observaciones pueden ser interesantes por sí mismas pero también para hacer diagnósticos.



- Hay tres tipos de diagnósticos de las puntuaciones individuales que son interesantes:
  - Residuales
  - Influencia
  - Palanca o efecto

## *Residuales*

- Un residual es la diferencia entre la puntuación predicha y la observada realmente
- Valores positivos significa estar por encima de lo que condicionalmente corresponde, y negativos estar por debajo.
  - Alguien con un residual positivo en salario inicial significa que le pagaron más de lo que le correspondería por su sexo, minoría, experiencia previa, etc.
  - Alguien con un residual negativo sería lo opuesto, le pagaron menos de lo que le correspondería teniendo en cuenta lo anterior.

- Hay varias versiones de residuales
  - Residuales directos: Tienen el inconveniente de que no es fácil decidir cuándo un residual es grande o pequeño
  - Residuales estandarizados: Pueden ser interpretados como puntuaciones típicas
  - Residuales studentizados: Es una estandarización que tiene en cuenta que los residuales en los extremos son más variables. Este es el más recomendado.
  - Residuales borrados: Es el residual de una ecuación calculada excluyendo esa observación (de este modo el residual es independiente)
  - Residual studentizado borrado: Es el residual studentizado pero calculado excluyendo la propia observación. Esto es todavía mejor que el residual studentizado aunque la diferencia es mínima normalmente.

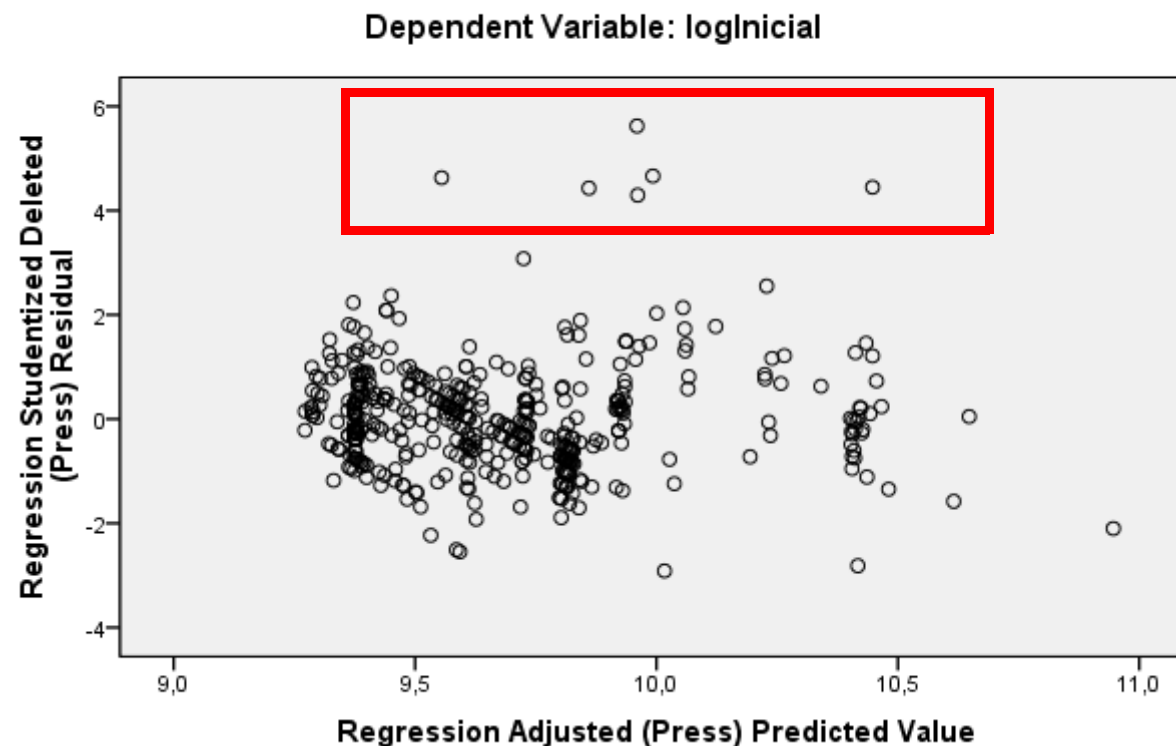
- La tabla de residuales de SPSS es la siguiente

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	9,2705	10,9170	9,6694	,29745	474
Std. Predicted Value	-1,341	4,194	,000	1,000	474
Standard Error of Predicted Value	,014	,051	,021	,006	474
Adjusted Predicted Value	9,2719	10,9461	9,6696	,29776	474
<b>Residual</b>	<b>-,54066</b>	<b>1,03593</b>	<b>,00000</b>	<b>,18979</b>	<b>474</b>
<b>Std. Residual</b>	<b>-2,834</b>	<b>5,429</b>	<b>,000</b>	<b>,995</b>	<b>474</b>
<b>Stud. Residual</b>	<b>-2,892</b>	<b>5,447</b>	<b>-,001</b>	<b>1,002</b>	<b>474</b>
<b>Deleted Residual</b>	<b>-,56324</b>	<b>1,04256</b>	<b>-,00022</b>	<b>,19245</b>	<b>474</b>
<b>Stud. Deleted Residual</b>	<b>-2,915</b>	<b>5,622</b>	<b>,001</b>	<b>1,009</b>	<b>474</b>
Mahal. Distance	1,534	32,248	4,989	3,872	474
Cook's Distance	,000	,075	,002	,007	474
Centered Leverage Value	,003	,068	,011	,008	474

a. Dependent Variable: logInicial

- Es interesante ver que hay dos residuales studentizados con valores por encima de 5 (se interpretan como puntuaciones típicas). No obstante, esta tabla es un poco limitada en que sólo informa de uno o dos valores (las medias y las desv. típicas no son muy interesantes)

- Volviendo al gráfico podemos ver que hay varios residuales con valores bastante altos y también que los valores residuales altos están sobre todo por arriba, no por abajo.,



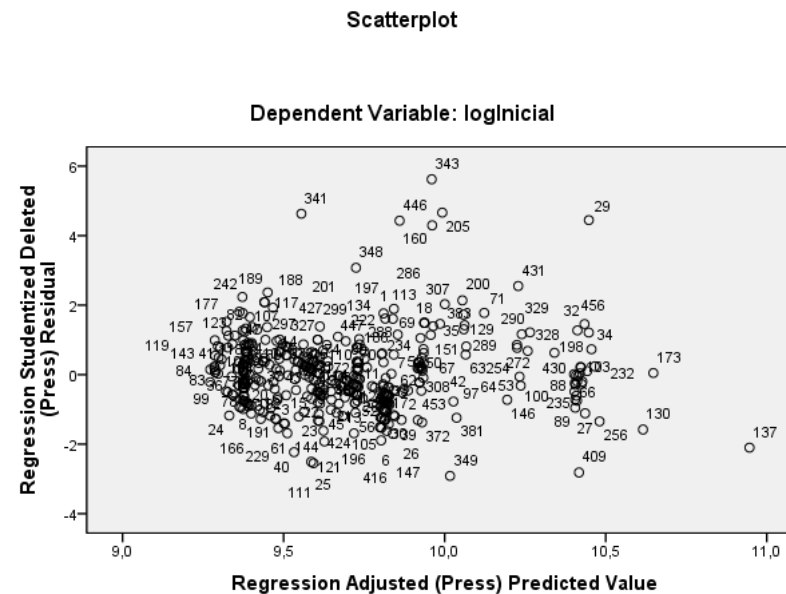
- ¿Qué se puede hacer con los residuales?
  - Se pueden explorar para ver si tienen alguna característica especial
  - Se puede ver si están asociados con alguna variable no considerada
  - Se pueden eliminar para ver su efecto sobre el ajuste del modelo (a veces el cambio puede ser interesante).

- Los residuales y otros indicadores se pueden guardar para explorar

The screenshot shows the 'Regresión lineal: Guardar' (Linear Regression: Save) dialog box. It contains several sections with checkboxes for saving different types of data:

- Valores pronosticados** (Predicted values):
  - ☐ No tipificados
  - ☐ Tipificados
  - ☐ Corregidos
  - ☐ E.T. del pronóstico promedio
- Residuos** (Residuals):
  - ☐ No tipificados
  - ☐ Tipificados
  - ☐ Método de Student
  - ☐ Eliminados
  - ☐ Eliminados estudentizados
- Distancias** (Distances):
  - ☐ Mahalanobis
  - ☐ De Cook
  - ☐ Valores de influencia
- Estadísticos de influencia** (Influence statistics):
  - ☐ DfBetas
  - ☐ DfBetas tipificadas
  - ☐ DfAjuste
  - ☐ DfAjuste tipificada
  - ☐ Razón entre covarianzas
- Intervalos de pronóstico** (Prediction intervals):
  - ☐ Media ☐ Individuos
  - Intervalo de confianza: 95 %
- Estadísticos de los coeficientes** (Coefficient statistics):
  - ☐ Crear estadísticos de los coeficientes
  - ☒ Crear un nuevo conjunto de datos
  - Nombre de conjunto de datos:

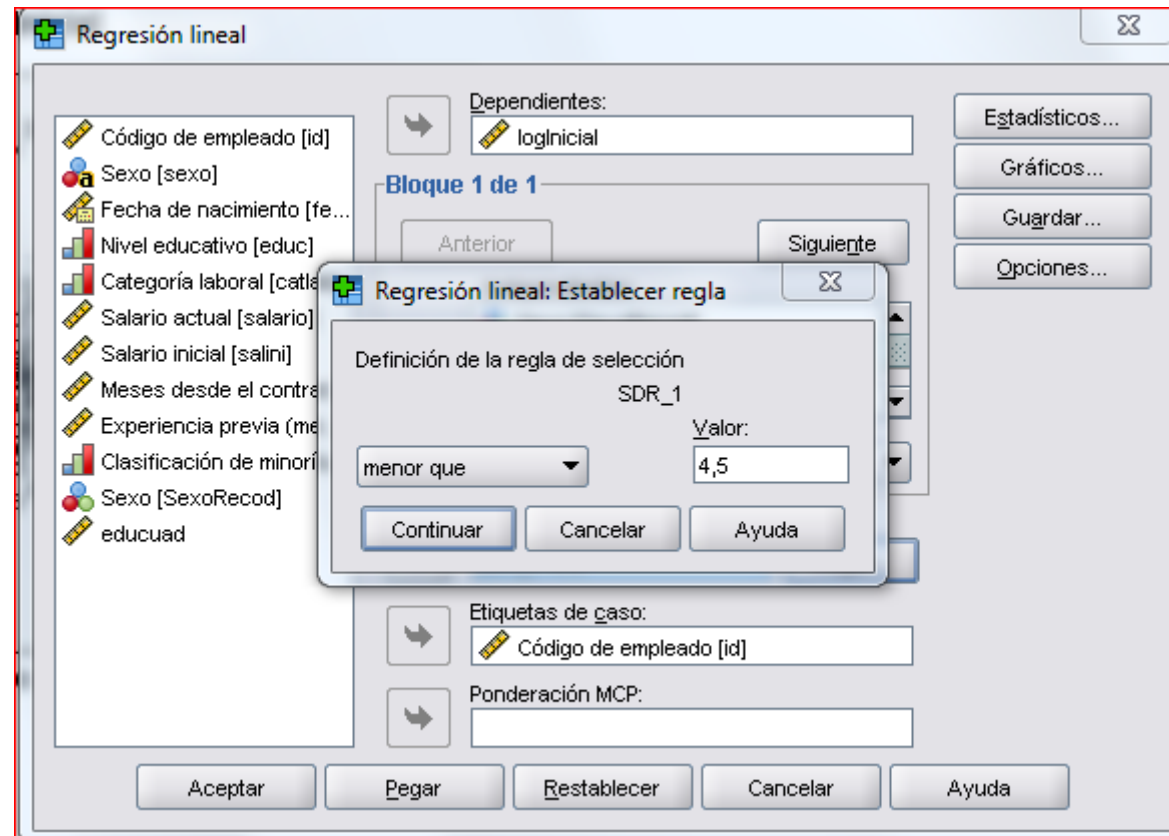
- También, si tenemos etiquetas, se pueden añadir a la regresión y aparecen en los gráficos



- Podemos ver por ejemplo que el caso 343 es el que tiene el residual studentizado más alto (por que le pagarían tanto?), pero también el 341 o el 29 son interesantes.



- Para analizar el impacto de quitar unos residuales podemos quitarlos



- El resultado muestra los resultados para el modelo sólo con los casos seleccionados

Model Summary<sup>b,c</sup>

Model	R		R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
	Studentized Deleted Residual < 4,50000 (Selected)	Studentized Deleted Residual >= 4,50000 (Unselected)				R Square Change	F Change	df1	df2	Sig. F Change
1	,859 <sup>a</sup>	,960	,738	,735	,17610	,738	262,192	5	465	,000

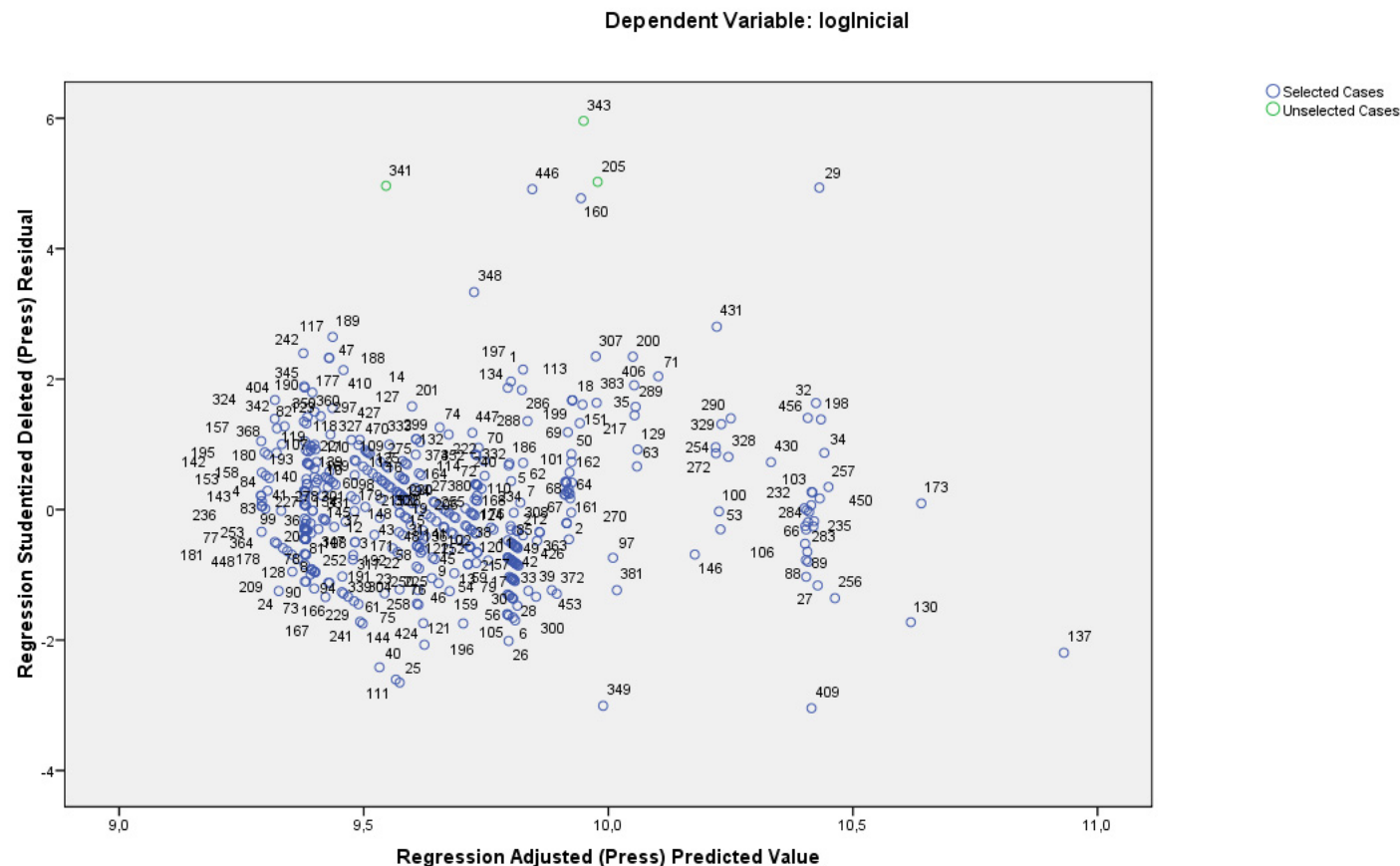
a. Predictors: (Constant), educuad, Clasificación de minorías, Experiencia previa (meses), Sexo, Nivel educativo

b. Unless noted otherwise, statistics are based only on cases for which Studentized Deleted Residual < 4,50000.

c. Dependent Variable: logInicial

- El resultado muestra una R cuadrado ajustada de .735 (con todos los casos era de .708)

- Otra forma de ver el efecto es en el gráfico de los residuales. Vemos que sólo se han eliminado tres observaciones y que hay otras cerca que podría considerarse eliminar también



## *Distancias (Palanca)*

- Un valor tiene mucha palanca si existen pocos casos con características similares a él
  - En principio, un supuesto de la regresión es que todos los puntos deberían tener una palanca semejante
  - Si un caso es inusual, su palanca crece con respecto a otros que son más comunes
  - La palanca tal y como se considera para regresión múltiple sólo se refiere a los predictores y no tiene en cuenta la variable predicha

- Hay dos medidas que indican palanca de cada caso:

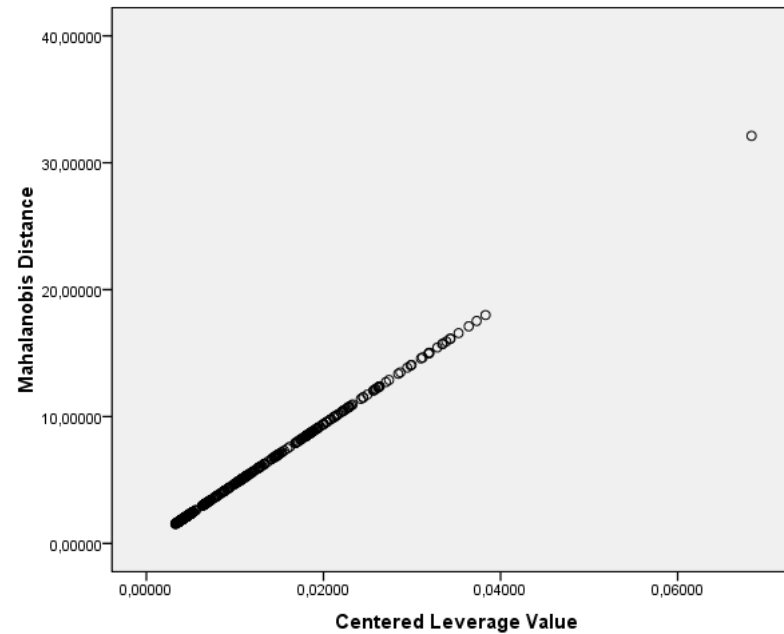
The screenshot shows a web-based interface for regression diagnostics. It contains several sections with checkboxes and a text input field:

- Top section:** Four checkboxes arranged in two columns: "No tipificados", "Tipificados", "Corregidos", "E.T. del pronóstico promedio" on the left; and "No tipificados", "Tipificados", "Método de Student", "Eliminados", "Eliminados estudentizados" on the right.
- Distancias:** A section with three checkboxes: "Mahalanobis" (checked), "De Cook", and "Valores de influencia" (checked).
- Intervalos de pronóstico:** A section with two checkboxes: "Media" and "Individuos". Below them is a text input "Intervalo de confianza:" followed by a dropdown menu showing "95" and a "%" symbol.
- Estadísticos de influencia:** A section with five checkboxes: "DfBetas", "DfBetas tipificadas", "DfAjuste", "DfAjuste tipificada", and "Razón entre covarianzas".
- Estadísticos de los coeficientes:** A section with two radio buttons: "Crear estadísticos de los coeficientes" (unchecked) and "Crear un nuevo conjunto de datos" (selected). Below the radio buttons is a text input "Nombre de conjunto de datos:" followed by an empty text box.

- Mahalanobis: Es una medida de la distancia de una observación particular respecto del centroide de los datos
- Influencia: Da unos resultados entre 0 (no influye en el ajuste) y  $(N-1)/N$  (es decir cerca de 1 es el máximo)

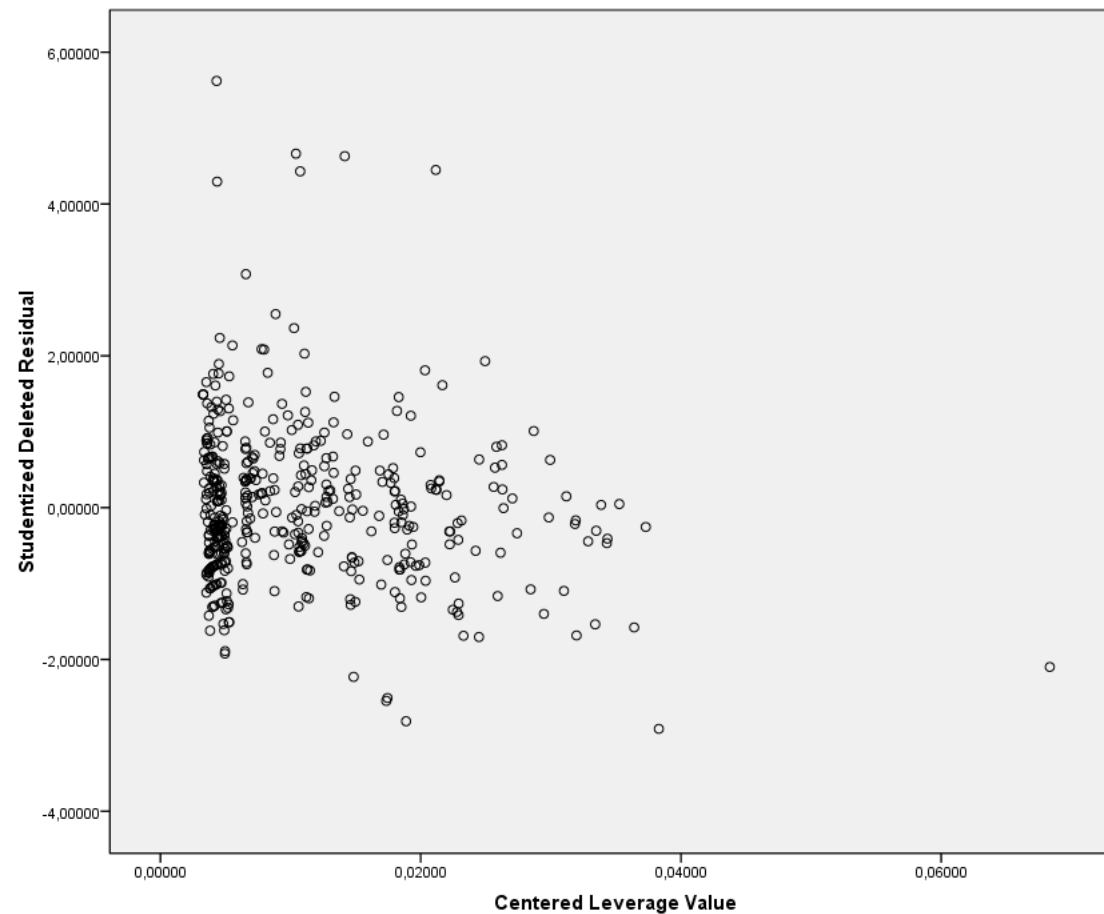
Las distancias de Cook están mal puestas, deberían estar a la derecha

- En realidad la influencia y Mahalanobis son una transformación lineal y significan lo mismo.

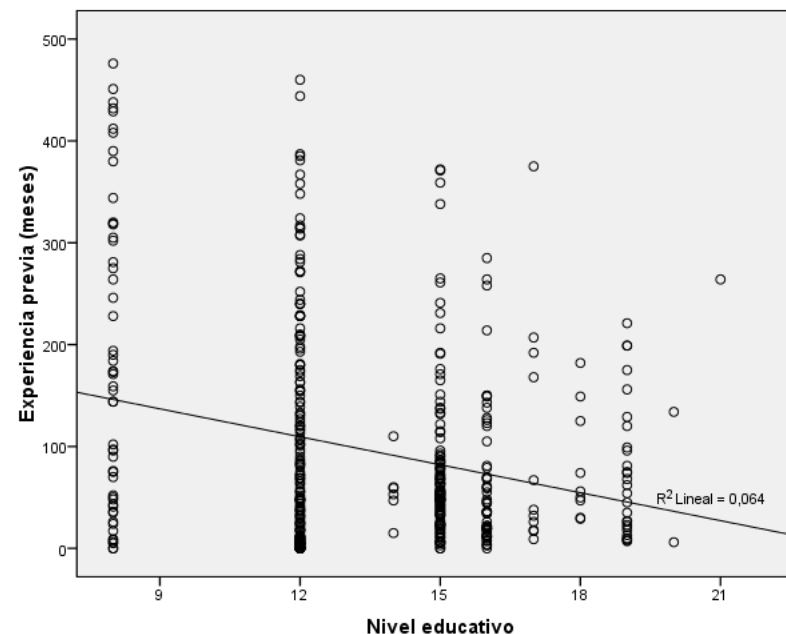


- Yo prefiero influencia

- Un gráfico interesante es el residuales studentizados frente a valores de influencia. En él vemos valores con influencia y residuales. Los peores son los que destacan en ambas cosas (no es el caso)



- El valor con más influencia es uno que pasa de 0.06. Revisando los datos vemos que ese caso es el único con un nivel educativo de 21 y que con un nivel 20 sólo hay un par.

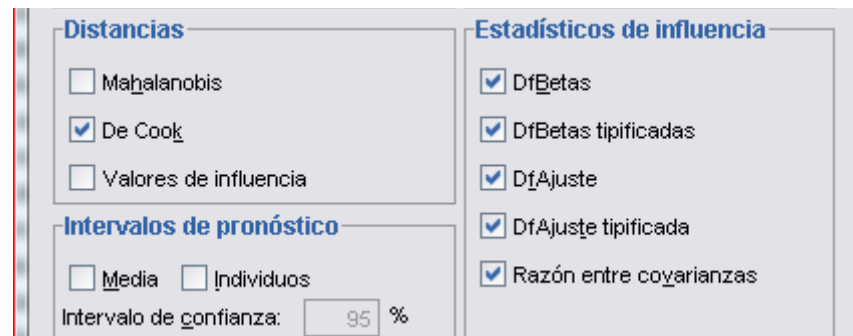


- Además, el nivel educativo está a menudo un poco asociado a menos experiencia previa pero no parece el caso.
- La influencia es importante pero la siguiente medida nos da una información adicional que es todavía más informativa.



## *Influencia*

- He llamado a esta medida Influencia porque hay un problema de traducción (en inglés se usa leverage e influence).
- Hay varias medidas para esto pero la que se suele comentar es la distancia de Cook

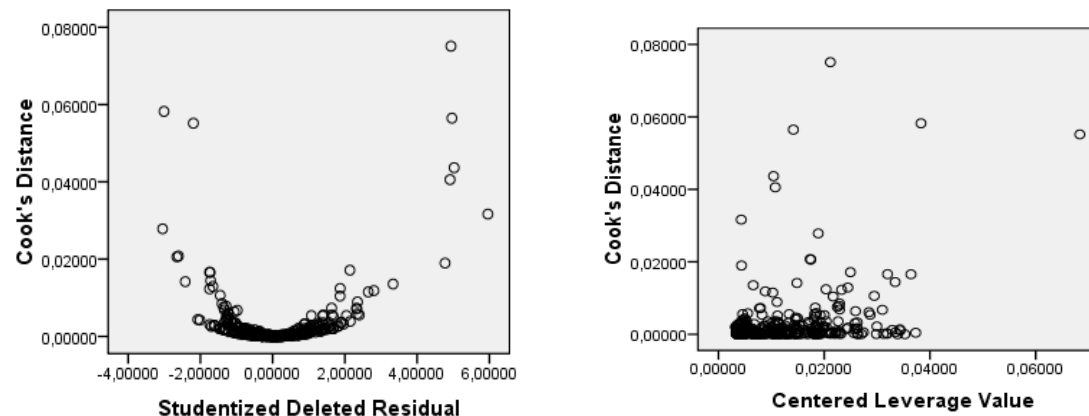


The screenshot shows a dialog box with three main sections:

- Distancias**:
  - ☐ Mahalanobis
  - ☒ De Cook
  - ☐ Valores de influencia
- Intervalos de pronóstico**:
  - ☐ Media ☐ Individuos
  - Intervalo de confianza:  %
- Estadísticos de influencia**:
  - ☒ DfBetas
  - ☒ DfBetas tipificadas
  - ☒ DfAjuste
  - ☒ DfAjuste tipificada
  - ☒ Razón entre covarianzas

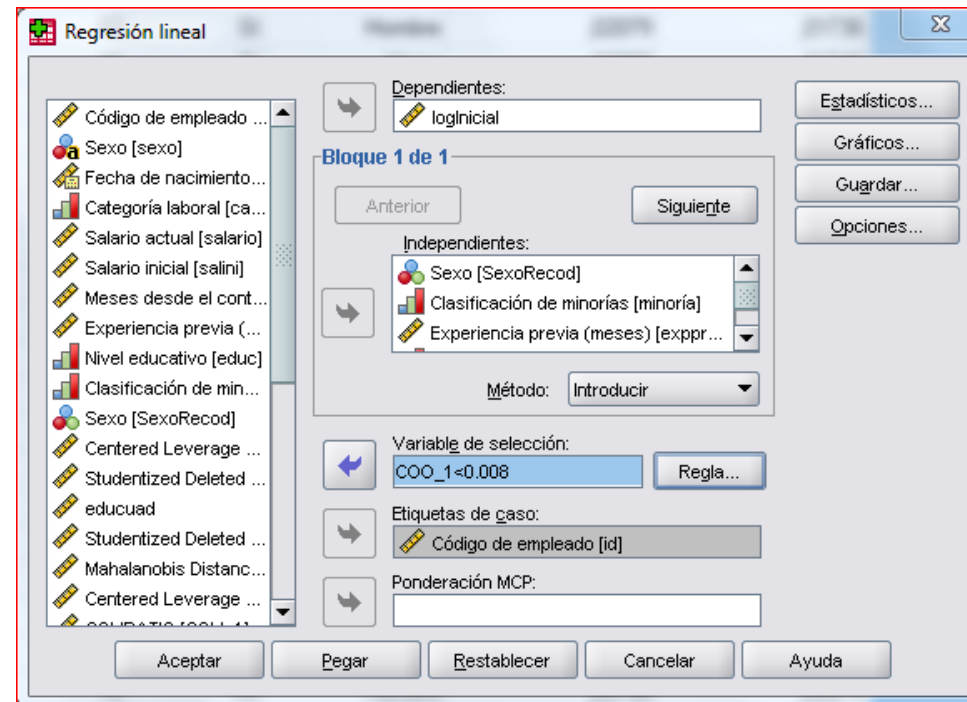
- Todas ellas tienen una idea similar, ¿qué cambio se produciría en los resultados de la regresión si eliminamos una puntuación?
  - Valores altos indican que eliminar esa puntuación cambiaría mucho los resultados
  - Son una combinación de las consecuencias de tener un residual alto y un valor de influencia alto
  - De nuevo, valores altos destacados indican valores peculiares. Estos valores tienen influencia y además el residual cambiaría mucho si ese valor fuera eliminado.
  - Aquí sólo veremos las distancias de Cook por simplificar.

- Aquí podemos ver como se relacionan las distancias de Cook con el Leverage y los residuales.

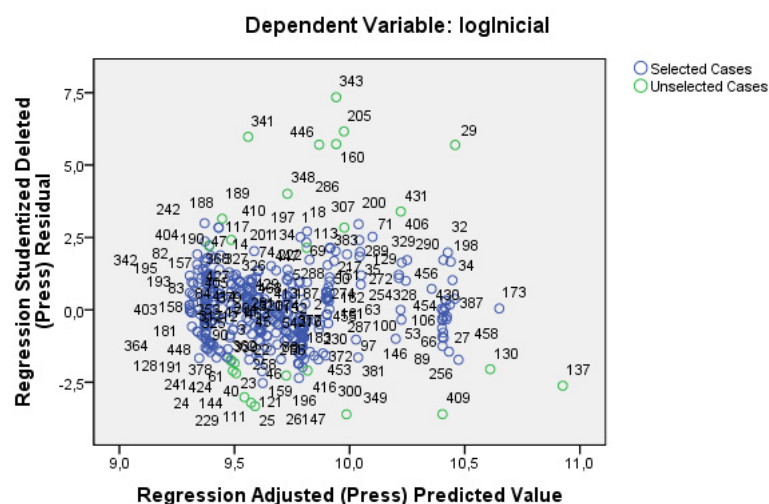


- En nuestro caso, el valor de distancia de Cook más alto es cerca de 0.07. Las distancias de Cook tienen un mínimo de 0 y se considera alto un valor de  $4/n$  (en este caso  $4/474=0.008$ ). Hay por tanto varios valores que podríamos considerar que tienen una distancia de Cook excesiva (hay 25 que pasan de este límite).

- Podemos repetir el análisis de regresión excluyendo esos caso con valores de distancia de Cook muy altos



- El resultado no incluiría muchos de los residuales más exagerados (sobre todo los positivos) y tampoco algunos valores con más palanca (137 y 130). La R cuadrado sube a .795 (antes era .708)

Model Summary<sup>b,c</sup>

Model	R		R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
	Cook's Distance < ,00800 (Selected)	Cook's Distance >= ,00800 (Unselected)				R Square Change	F Change	df1	df2	Sig. F Change
1	,892 <sup>a</sup>	,596	,795	,792	,14392	,795	342,354	5	442	,000

a. Predictors: (Constant), educuad, Clasificación de minorías, Experiencia previa (meses), Sexo, Nivel educativo

b. Unless noted otherwise, statistics are based only on cases for which Cook's Distance < ,00800.

c. Dependent Variable: logInicial

- Obviamente, eliminar 20 casos es una decisión importante y tiene que ser valorada con cuidado. Antes de eliminar un caso de los resultados habría que:
  - Valorar si hay algo especial en ese caso
  - Si hay un error en los datos
- Una buena razón para eliminar un caso con mucha influencia es que un buen análisis de regresión es aquel en que los casos tienen una influencia similar.
  - Si todo el análisis de regresión está condicionado por ese caso, eso no es bueno
  - Si un caso realmente destaca de los demás, es buena idea apartarlo y ofrecer los resultados sin ese valor (comentando que se ha eliminado un caso por las razones ofrecidas)

# Interpretación

- Ya hemos hablado antes de la interpretación de los coeficientes pero algunos de los que hemos usado tienen una interpretación especial.
  - Variables ficticias (sexo, minoria)
  - Polinomios

## *Variables ficticias*

- Las variables ficticias es la manera de utilizar variables categóricas en un análisis de regresión
  - Se necesitan  $k-1$  columnas para representar las categorías de una variable categórica
  - Generalmente se utiliza 0 y 1 para las categorías porque es más fácil de interpretar (aunque en este caso he usado 1 y 2 porque el SPSS me ha puesto eso automáticamente y no lo he cambiado)
  - En el módulo que estamos usando esta codificación se hace manualmente
  - Para variables con dos categorías no es mucho problema, pero si hay variables con más categorías resulta interesante utilizar un módulo que lo haga automáticamente (p.e. GLM)



- El coeficiente de la regresión para una variable categórica se interpreta del siguiente modo:

**Coefficients<sup>a</sup>**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	10,234	,042		244,118	,000		
Sexo	-,388	,027	-,548	-14,234	,000	1,000	1,000

a. Dependent Variable: logInicial

- El valor del coeficiente es la diferencia entre las medias en la variable dependiente de las dos categorías ( $9.458 - 9.846 = -0.388$ ) cuando sólo hay una variable independiente
- El valor de la constante puede utilizarse para calcular la media en la variable dependiente de las categorías (pero es más fácil calcular la tabla de abajo)

**Report**

logInicial

Sexo	Mean	N	Std. Deviation
Hombre	9,8461	258	,35566
Mujer	9,4583	216	,20108
Total	9,6694	474	,35284

- Cuando el modelo es más completo, los coeficientes obviamente varían pero sigue teniendo la misma interpretación (diferencia entre las categorías)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	10,717	,153		69,914	,000		
Sexo	-,190	,020	-,268	-9,515	,000	,777	1,287
Clasificación de minorías	-,088	,022	-,103	-4,041	,000	,945	1,058
Experiencia previa (meses)	,000	,000	,102	3,748	,000	,834	1,198
Nivel educativo	-,206	,022	-1,687	-9,191	,000	,018	54,510
educuad	,011	,001	2,312	12,568	,000	,018	54,756

a. Dependent Variable: logInicial

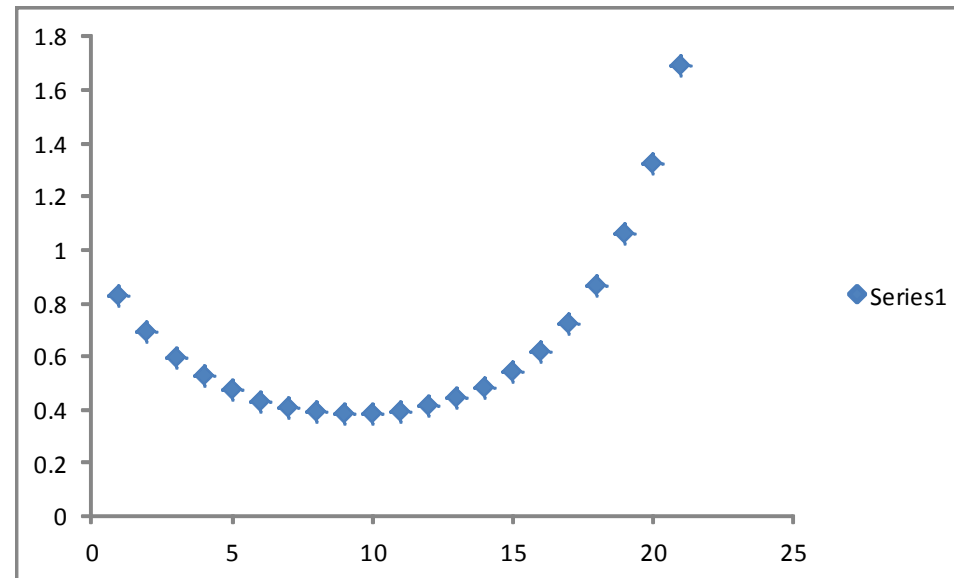
## Polinomios

- Cuando se introducen términos polinómicos en un análisis de regresión, la interpretación de cada variable por separado no tiene mucho sentido.
  - Hay que valorarlos conjuntamente
- En nuestro ejemplo, el nivel educativo tiene un valor negativo mientras que el cuadrado del nivel educativo tiene un valor positivo (aunque no hay que olvidar que la variable está en logaritmos)

Coefficients <sup>a</sup>							
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	10,717	,153		69,914	,000		
Sexo	-,190	,020	-,268	-9,515	,000	,777	1,287
Clasificación de minorías	-,088	,022	-,103	-4,041	,000	,945	1,058
Experiencia previa (meses)	,000	,000	,102	3,748	,000	,834	1,198
Nivel educativo	-,206	,022	-1,687	-9,191	,000	,018	54,510
educuad	,011	,001	2,312	12,568	,000	,018	54,756

a. Dependent Variable: logInicial

- Para ver el efecto de estos coeficientes se puede hacer un gráfico de este tipo (yo lo he hecho con Excel)



## Actividades

---

1. En este ejemplo se trata de predecir la cantidad de nitrógeno en unos ríos a partir de índices de uso agrícola, bosque, residencial, y un índice de uso comercial industrial. Se encuentra en el archivo. Se encuentra en el archivo NewYorkRivers.

*No olvidar examinar los residuales*

2. El archivo Coches tiene datos sobre el consumo, el motor, la potencia, la aceleración, el año del modelo de una serie de coches, el país de origen y la cilindrada. Prueba a hacer un modelo de regresión que prediga el consumo a partir de las otras variables.  
*Ten en cuenta que la variable País tiene tres categorías y por tanto vas a necesitar dos variables ficticias para representar todas las categorías.*

3. En el ejemplo de calaveras egipcias, tenemos las variables Year (año aproximado de formación de la calavera, negativo Antes de Cristo, positivo después), MB anchura máxima de la calavera, BL Longitud basialveolar de la calavera, y NH altura nasal de la calavera. La idea es predecir la edad de la calavera a partir de las otras variables.

*No olvides comprobar si se cumplen los supuestos*

4. Inmigración de un estado en estados unidos a otro. Los datos son de 48 de ellos (excluyendo Alaska y Hawaii). La variable dependiente es la inmigración doméstica neta lo cual representa el movimiento neto de la población dentro o fuera del estado sobre el período 1990-1994 dividido por la población del estado. Once predictores de ese movimiento que se piensa influyen en la inmigración son desempleo (tasa de desempleo en 1994), Salario (promedio de salario por hora de trabajadores de fábricas en 1994), Crime (tasa de crímenes violentos por 100000 en 1993), Income (ingresos por hogar medio en 1994), Metrop (porcentaje de población del estado viviendo en áreas metropolitanas, Poor (porcentaje de población viviendo por debajo del umbral de la pobreza), Taxes (impuestos totales y locales por cabeza en 1993), Educ (porcentaje de población de 25 años o más mayor que tienen un título de instituto o mayor en 1990), BusFail (número de negocios fallidos por la población del estado en 1993), Temp (promedio de 12 temperaturas promedio mensuales en Fahrenheit en 1993), Region (región en la que el estado se encuentra). El archivo se llama Inmigracion.sav.

*Este es un modelo con muchas variables pero muy correlacionadas. El modelo final puede incluir muy pocas variables. No obstante, hay que examinar los gráficos para ver que hay algunos problemas con los datos.*



5. Se intenta valorar el precio de las casas a partir de una serie de variables. Las variables consideradas son Y: precio de venta de la casa en miles de dólares,  $X_1$ : Impuestos en miles de dolares,  $X_2$ : número de baños,  $X_3$ : Tamaño incluyendo jardín,  $X_4$ : Tamaño habitable,  $X_5$ : Plazas de garaje,  $X_6$ : Número de habitaciones,  $X_7$ : Número de dormitorios,  $X_8$ : Antigüedad de la casa (años),  $X_9$ : Número de chimeneas. Intenta ajustar un modelo completo pero comparalo con la opinión de un experto que dice que sólo con los impuestos, el número de habitaciones y la antigüedad de la casa se puede establecer bien el precio. ¿Cómo sería de bueno un modelo que sólo incluyera los impuestos como predictor? Los datos están en el archivo PrecioCasas.sav.

*Este es un ejemplo muy sencillo en el fondo.*

6. Homicidios. Se investigó el papel de las armas de fuego para explicar el aumento en la tasa de homicidios en Detroit. Datos de 1961-1973. La variable respuesta es la tasa de homicidios (H) por 100000 y las variables son predictores que se creía que influían en ese aumento. FTP: Número de policías a tiempo completo por 100000 habitantes, UEMP, porcentaje de población desempleada, M: Número de trabajadores en la manufactura, LIC: número de licencias de pistolas por 100000, GR: número de pistolas registradas por 100000, CLEAR porcentaje de homicidios aclarados con arrestos, W: número de varones blancos en la población, NMAN: Número de trabajadores pero no en la manufactura (en miles), G: Número de trabajadores en el gobierno (en miles), HE: salario hora promedio, WE: salario semanal promedio. Los datos están en el archivo Homicidios.sav.

*En este ejemplo no es difícil encontrar modelos que ajusten. No obstante, hay que valorar si introducir demasiadas variables es apropiado o no.*