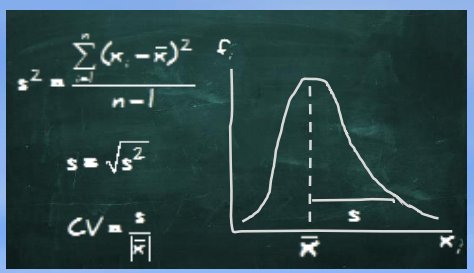
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE III. ANÁLISIS

Capítulo III.9
Análisis de regresión

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2016). Análisis de regresión. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. 1ª edición. Edición digital: http://ddd.uab.cat/record/163569.

Capítulo acabado de redactar en octubre de 2016

Contenido

PARTE III
Capítulo 9

Análisis de regresión

El análisis de regresión, o el llamado modelo lineal de la regresión, es un método estadístico basado en el estudio de la relación entre variables medidas con una escala cuantitativa. La relación que se establece es de dependencia, así, por un lado tenemos las llamadas variables criterio, dependientes o explicadas, de otro, las variables predictivas, independientes o explicativas[1]. El objetivo de esta técnica consiste en determinar la contribución de la variable independiente (análisis de regresión simple) o las variables independientes (análisis de regresión múltiple) en la explicación de la variable dependiente a través de un coeficiente para cada variable que indica la importancia relativa de cada una en la explicación de la variabilidad de la variable dependiente, como suma de efectos que se expresa en la ecuación general de un modelo lineal de la forma siguiente:

$$Y_i = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Ecuación 1

Variable dependiente — *Constante* — *Variables independientes* — *Error*

En este sentido es una técnica que razona de forma similar al análisis de varianza que vimos en capítulo anterior, si bien en el ANOVA las variables independientes son cualitativas. Con variables cuantitativas, la formalización de las relaciones entre las variables se establece a partir del concepto central de correlación (lineal), una medida que cuantifica el grado de relación (lineal) que se da entre las variables.

Como el tema anterior pretendemos proporcionar una presentación de carácter introductorio para conocer sus características principales y orientar un seguimiento posterior, planteando relaciones simples con una variable independiente y extendiendo el análisis a la regresión múltiple con dos variables independientes. Tras la presentación de la técnica de análisis a través de diversos ejemplos veremos su utilización con el software estadístico, en SPSS y en R.

[1] También veremos cómo el análisis de regresión se puede realizar con variables independientes cualitativas.

1. Presentación del análisis de regresión

La técnica del análisis de regresión ha sido fundamental en el desarrollo histórico del análisis de datos estadísticos. Francis Galton (Birmingham, 1822-1911) fue un científico multidisciplinar reconocido como una figura central para la constitución de la Estadística así como, particularmente, por inventar lo que llamó como "regresión a la media". Sus trabajos, basados en la medición cuantitativa a partir de la ley de la normal, le llevaron a desarrollar el concepto y la medida de la correlación así como los fundamentos de la regresión. En 1877, estudiando la relación existente entre la altura de los niños, tomados en relación al valor de la media de su generación, y la altura de los padres, constata que la desviación respecto de la media disminuye, es decir, regresa, e introduce el conocido como coeficiente de regresión. Por otro lado, el concepto de correlación fue estudiado y generalizado por Karl Pearson (Londres, 1857-1936), a partir de la admiración que le produjo el trabajo de Galton, y construyó las fórmulas que utilizamos en la actualidad para el cálculo del coeficiente de correlación (Doebresque y Tassi, 1990).

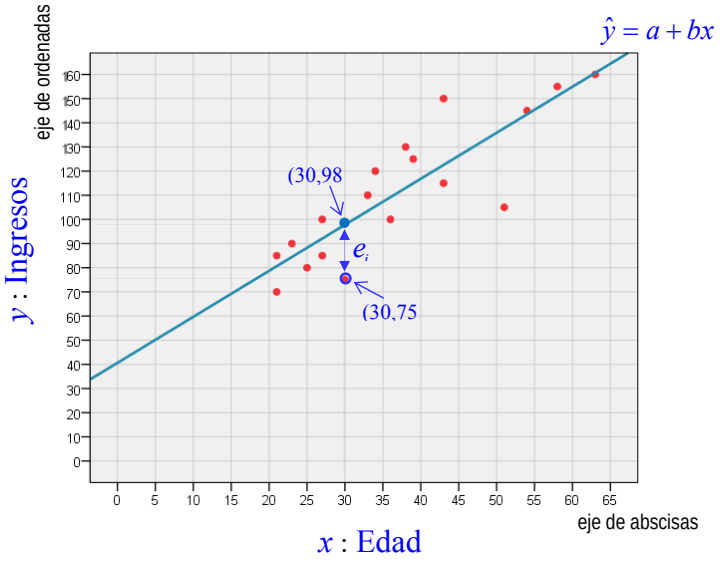
En un análisis de regresión lineal simple se puede poner de manifiesto la relación de dependencia sencilla y directa, de tipo lineal, que se da entre dos variables cuantitativa, donde una de ellas es considerada como dependiente y la otra como independiente. Por ejemplo, considerando los años de escolaridad para explicar los ingresos; o estableciendo una relación entre el número de horas realizadas de trabajo doméstico en función de la edad o los ingresos; para dar cuenta de la satisfacción en el trabajo explicándola en función de factores como la antigüedad en la empresa o el tamaño de la misma teniendo en cuenta el número de trabajadores; para determinar la esperanza de vida de un país en función de su nivel de desarrollo o riqueza, considerando en concreto como indicador la renta per cápita; para explicar la tasa de delincuencia como variable que influye en percepción de la seguridad de la ciudadanía, etc.

De forma similar al análisis de varianza el objetivo consiste en alcanzar la mayor precisión en el conocimiento de la variable dependiente *Y* a través del conocimiento y de la dependencia que tiene en relación a la variable independiente *X*. Es decir, se busca reducir la variabilidad de *Y* mediante los valores de *X*. Esto equivale a afirmar que la variable *X* explica la variable *Y*, y también que la predice, y según la teoría de que estemos considerando o las características del diseño de análisis que *X* causa *Y*[2]. En un análisis de regresión simple se evalúa el efecto de una sola variable; en un análisis de regresión múltiple se valora tanto el efecto individual de cada variable independiente como el efecto conjunto de todas ellas para determinar el comportamiento de la dependiente. En ambos casos, simple y múltiple, una vez establecida la ecuación de regresión se realiza un ejercicio de pronóstico: si la realidad social estudiada se explica según el modelo de relaciones obtenido y, *cæteris paribus* (si el resto permanece constante), es posible predecir un comportamiento futuro o esperado de la variable dependiente dados unos valores de la(s) variable(s) independiente(s).

[2] Con hemos destacado en otra ocasiones en este manual la causalidad no se deriva necesariamente de distinguir y formalizar relaciones entre una variable dependiente y una o más independientes. Hablamos de dependencia y de factores que son determinantes de una variable de respuesta entendiendo que existe una relación funcional (la ecuación de regresión) que establece el vínculo estadístico entre ellas sin que ello presuponga establecer una relación de causalidad. Para ello debemos sustentarnos en una teoría que así lo razone y en un diseño de análisis que así lo establezca.

El modelo algebraico que adoptamos es de la relación lineal, la que tiene una imagen gráfica de línea como se comenta a continuación y que denominamos recta de regresión. A partir de la consideración de dos variables, una dependiente y una independiente, la relación entre ambas se puede representar gráficamente a través de un diagrama de dispersión donde cada punto representa un individuo o un caso con las coordenadas sobre los ejes cartesianos, de abscisas (eje horizontal) y de ordenadas (eje vertical). Estas coordenadas son las puntuaciones o valores en las variables consideradas. El Gráfico III.9.1 adjunto ilustra la representación de un conjunto de 18 puntos a partir de dos variables: *y*, los ingresos (en euros, por día trabajado), y *x*, la edad (en años). Así, por ejemplo, se destaca en el gráfico en punto del individuo que tiene de coordenadas el par (30,75), es decir, 35 años de edad y un valor de ingresos de 75 euros diarios. El conjunto de todos ellos configura una nube de puntos con una disposición específica. Cuando esta disposición de los puntos es alargada con la forma representada en el Gráfico III.9.1, la relación entre las dos variables se puede expresar y formalizar ajustando la nube de puntos a la función matemática de una recta: expresamos y substituimos la nube de puntos por una línea recta. Esta función es la llamada recta de regresión, es la regresión de *y* sobre *x*, de la variable dependiente (criterio o explicada) *y* sobre la variable independiente (predictiva o explicativa) *x*, y se fundamenta en un modelo matemático que persigue, primero, describir linealmente la relación entre *y* y *x*, segundo, hacerlo con una determinada capacidad explicativa del comportamiento de *y* en función de *x* (la cuantificación de su grado de relación), y, tercero, predecir, para un valor de *x* el valor de *y*.

Gráfico III.9.1. Diagrama de dispersión. Nube de puntos y recta de regresión



Así pues, los puntos en el gráfico representan los valores observados, los pares *(x,y)*, es decir, el comportamiento observado en la distribución conjunta de una muestra de una determinada población, si bien también se podría tomar en consideración el conjunto de todos los datos de esa población. En la figura del Gráfico III.9.1 se ve que el conjunto de puntos tiene una orientación lineal lo que nos invita a suponer que esta tendencia global es la de la población observada. Pero lo que buscamos es una recta (o

función lineal) que represente mejor el comportamiento de toda la población, es decir, suponemos e imponemos que un modelo lineal (ecuación de una recta) es el que mejor se acomoda al conjunto de los datos. A esta recta la denominamos recta de regresión y es una ecuación teórica de una función matemática que es estimada. En el caso de un análisis de regresión simple la expresión es la siguiente[3]:

$$\hat{Y}_i = a + bx_i \qquad \textit{Ecuación 2}$$

El problema es encontrarla. De ello daremos cuenta más tarde, sabiendo que a es el valor de y cuando x es igual a cero, es decir, donde la recta corta el eje de ordenadas, y b es la pendiente, la inclinación de la recta.

Si observamos los puntos del gráfico, todos ellos, a excepción de los situados sobre la misma recta, guardan una distancia vertical (y perpendicular al eje de abscisas) con respecto a la recta de regresión que depende de cada punto y no de la x que da la recta de regresión. Estas distancias que dependen de otros factores que no son x, siempre aleatorios, los denominamos errores e_i. Por lo tanto cada valor de y_i, para cada observación o individuo i, es la suma del valor en la recta de regresión para cada x_i, más el error que se comete e_i tal que:

$$y_i = a + bx_i + e_i \qquad \textit{Ecuación 3}$$

y, por tanto,

$$y_i = \hat{y}_i + e_i \qquad \textit{Ecuación 4}$$

Se dice que en la predicción de y a partir de x se comete un error e_i, que es la diferencia entre cada valor de y_i observado para cada punto y el valor que se predice cuando se estima la recta de regresión $R_{y/x}$, esta diferencia son los residuos. En el modelo de regresión simple esta relación se expresa por la ecuación lineal del modelo que acabamos de detallar (*Ecuación 3*). En el Gráfico III.9.1, para el individuo de coordenadas (30,75) es la distancia entre el valor 75 de la variable y_i, y el valor dado por la recta de regresión, 98.

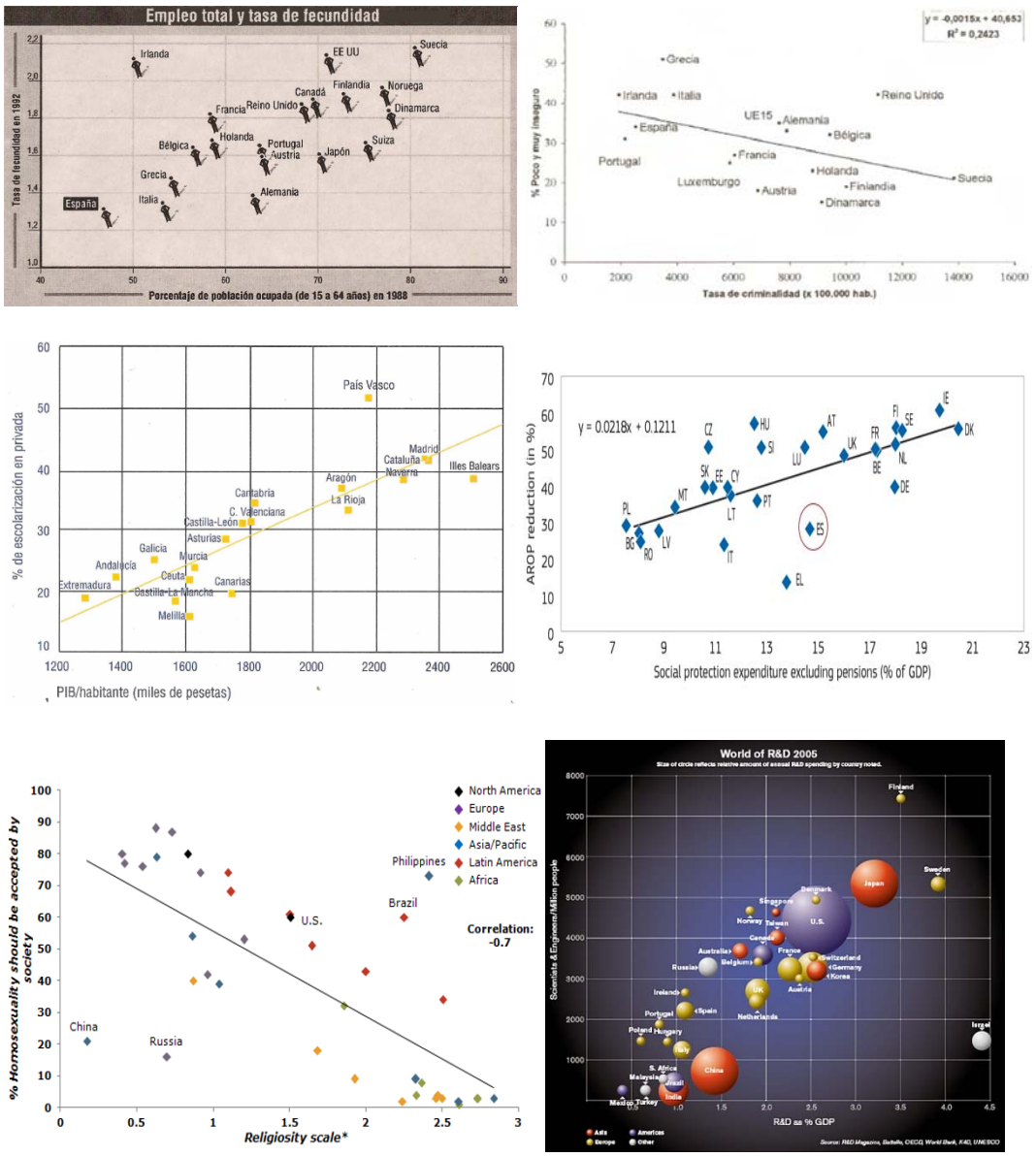
Se trata pues de buscar la recta de regresión que mejor se ajuste a la nube de puntos, es decir, la que minimiza las desviaciones o las distancias de todos los puntos en relación a esta recta. Una vez obtenida podremos describir la relación entre ambas variables y cuantificarla, establecer un grado de terminación de y en función del conocimiento de x y, por tanto, predecir cuáles serán los valores probables de y cuando se conoce x.

Cuando realizamos un análisis de regresión obtenemos resultados, por ejemplo, como este: $y=40+2x$, que se expresa en los términos siguientes: existe una cantidad mínima de ingresos de 40 que se espera obtener como promedio, y por cada año adicional se generan unos ingresos adicionales de 2. El valor de y estimado que se obtiene se interpreta como la media de ingresos que obtendrían todos aquellos que tuvieran x años de edad. Es decir, sencillamente observa como varía y ante una variación de x.

[3] La estimación se expresa simbólicamente con el sombrero sobre la variable $\hat{Y}_i$ (Y en mayúscula) que hace referencia a que los datos de toda la población que se estiman a partir de los datos muestrales y_i (y en minúscula).

Existen numerosos ejemplos que pueden mostrar este tipo de relaciones, como los que se muestran en el Gráfico III.9.2. Se puede observar la relación positiva entre la tasa de ocupación y la tasa de fecundidad: cuanto mayor es la ocupación mayor es la fecundidad en los diversos países desarrollados considerados. Se puede ver la relación inversa entre la tasa de criminalidad y la percepción de inseguridad: curiosamente, cuanto mayor es la tasa de criminalidad por 100.000 habitantes menor es el porcentaje de los que declaran sentirse inseguros.

Gráfico III.9.2. Ejemplo de relaciones con gráficos de dispersión



Encontrar la recta de regresión es la forma que tiene este procedimiento de expresar y describir la relación y la dependencia entre las dos variables. Este resultado es un paso que va más allá de un requisito fundamental y previo que se debe constatar: la existencia de relación y el grado de esta relación. En este sentido el análisis de regresión está

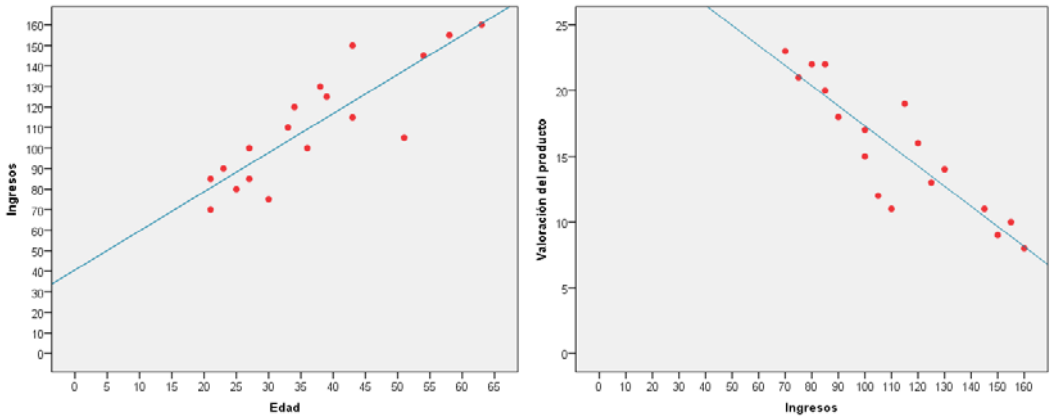
estrechamente relacionada con el cálculo de una medida de la relación entre variables cuantitativas: el coeficiente de correlación de Pearson. Como veremos más tarde un buen ajuste de una recta de regresión se basa en la existencia de una alta correlación entre las variables consideradas. Pero no habrá que olvidar que no es lo mismo; una cosa es la correlación (la interdependencia) y otra la dependencia, que puede ser, en particular, causal. Antes de profundizar en el análisis de regresión será necesario precisar el concepto y la forma de calcular la correlación entre las variables.

2. Concepto, medida y representación de la correlación

El concepto de correlación es una de las ideas básicas del análisis estadístico de datos a la hora de establecer relaciones entre dos variables cuantitativas. La correlación es una medida de asociación entre variables cuantitativas que expresa la fuerza o la intensidad de la relación entre dos variables así como la dirección de la misma, siempre y cuando esta relación sea de tipo lineal.

El llamado coeficiente de correlación producto-momento de Pearson (Pearson, 1896) nos muestra la variación concomitante entre las observaciones de dos variables. La relación entre ambas se establece a partir de la observación y del cálculo de las puntuaciones o pares de valores para las dos variables. Gráficamente esta relación se observa con un gráfico de dispersión. Consideremos, por ejemplo, la relación de antes entre los ingresos y la edad, y también consideramos una escala de valoración de un producto de consumo relacionada con los ingresos. Los gráficos de dispersión que se obtienen son los recogidos en el Gráfico III.9.3.

Gráfico III.9.3. Diagramas de dispersión con las nubes de puntos



(a) Nube de Ingresos y Edad
Relación lineal creciente

(b) Nube de Valoración e Ingresos
Relación lineal decreciente

La representación gráfica de las puntuaciones para todos los individuos da lugar a una nube de puntos que muestra en ambos casos la existencia de una relación lineal de asociación entre las variables. Pero nos encontramos ante dos situaciones claramente diferentes. Por un lado, observamos el gráfico (a) que a medida que aumenta la edad también aumentan los ingresos, la variación de ambas variables se encamina en el mismo sentido, se dice que la relación o la correlación es positiva: a más edad más

ingresos. Por otro lado, el gráfico (b) también muestra la existencia de relación pero en este caso en sentido negativo, las dos variables varían en sentido inverso, a medida que aumentan los ingresos la valoración del producto decrece, varían de forma concomitante en sentido contrario, la correlación es pues negativa, cuanto más rico menos se valora. Ambas situaciones indican visualmente la existencia de correlación que se expresará de forma precisa y cuantificada a través del cálculo del coeficiente de correlación de Pearson r. Veremos seguidamente que este coeficiente es un valor que varía entre 0 y 1, en función de que la correlación sea menor o mayor, y tendrá un signo positivo o negativo que expresará la direccionalidad de la correlación. La mayor o menor intensidad de la relación se traduce gráficamente en la mayor o menor dispersión de los puntos entorno a la recta de regresión, cuanto más próximos estén los puntos de la recta de regresión y, por tanto, más se parezcan a una línea recta mayor será la fuerza o intensidad de la relación.

El coeficiente de correlación entre y y x, r_{yx}, se define como el cociente entre la covarianza de las dos variables s_{yx} y el producto de las desviaciones, s_y y s_x, es decir:

$$r_{yx} = \frac{s_{yx}}{s_y \cdot s_x} \qquad \textit{Ecuación 5}$$

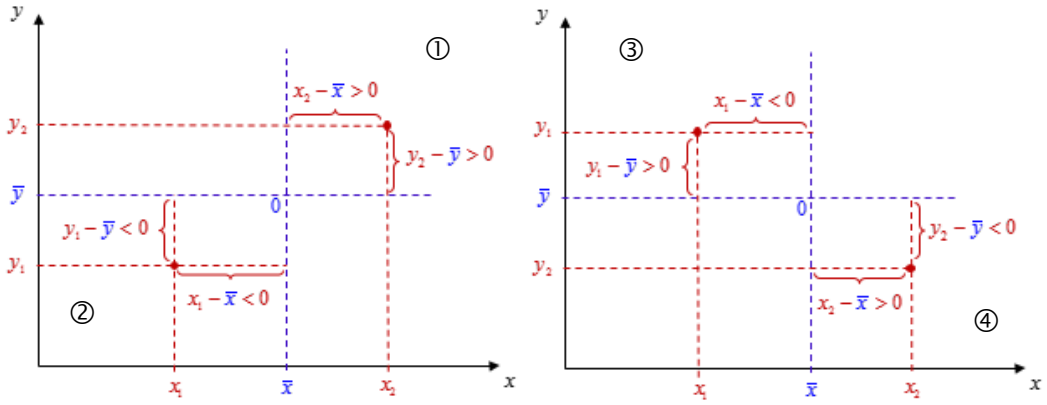
La covarianza s_{yx} expresa la cantidad de varianza común entre las dos variables, y su fórmula es:

$$\text{cov}(y,x) = s_{yx} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n-1} = \frac{SPD_{yx}}{n-1} \qquad \textit{Ecuación 6}$$

El Gráfico III.9.4 representa el concepto de covariación. Cuando (a) la relación es positiva entre los puntos las diferencias entre la puntuación de cada variable y su media dan lugar a valores positivos que se multiplican entre sí, o bien a valores negativos que se multiplican entre sí, en ambos casos dando lugar a un resultado positivo que expresa la existencia de variaciones conjuntas positivas o relaciones lineales crecientes y por tanto una correlación positiva. Cuando (b) la relación es negativa entre los puntos las diferencias entre la puntuación de cada variable y su media dan lugar siempre a valores positivos y negativos que al multiplicarse entre sí generan un resultado negativo poniendo de manifiesto la variación en sentido contrario de sus valores o una relación lineal decreciente y por tanto de una correlación negativa.

El valor de la covarianza es mayor cuanto más más intensa es la relación lineal entre las variables y el signo expresa la direccionalidad de la relación: en sentido positivo o negativo. Un valor cero indica el valor mínimo y la ausencia de covariación. No obstante, el valor máximo manifiesta un problema relevante pues dependen del grado de dispersión de las variables y, en consecuencia, impide su interpretación como medida relativa de la intensidad de la relación entre variables y muestras distintas. Para solucionarlo se considera valorar el valor de la covarianza con respecto al valor máximo que puede alcanzar en cada caso. El valor máximo de la covarianza entre dos variables se sabe que es igual al producto de las desviaciones típicas de ambas variables, por lo que dividiendo la covarianza en el máximo valor que puede alcanzar podemos relativizarla. Esto es lo se hace cuando se calcula el coeficiente de correlación.

Gráfico III.9.4. Representación de la covariación



(a) Relación positiva *(b) Relación negativa*

Cuadrante	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
①	+	+	+
②	−	−	+
③	+	−	−
④	−	+	−

Por tanto, la correlación se interpreta como el grado en que la covarianza alcanza su máximo, o también como la proporción de la varianza compartida por ambas variables sobre el total de la covarianza[4].

A partir de la fórmula del coeficiente de correlación (*Ecuación 5*), si consideramos las expresiones de la varianza de y y de x:

$$var(y) = s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{SCD_y}{n-1} \quad \text{y}$$

$$var(x) = s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{SCD_x}{n-1}$$

la expresión del coeficiente de correlación queda:

$$r_{yx} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{SPD_{yx}}{\sqrt{SCD_y \cdot SCD_x}} \qquad \textit{Ecuación 7}$$

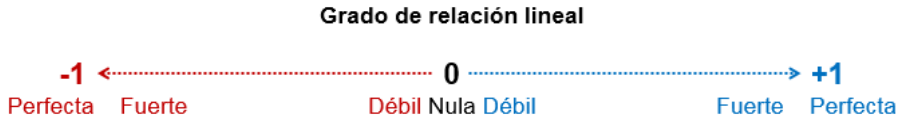
[4] Se puede comprobar además que el resultado es equivalente a calcular la covarianza en las dos variables después de tipificarlas o estandarizarlas. En ese caso la dispersión es siempre la misma independientemente de los datos y del grado de dispersión original de las variables, pues las variables tipificadas siempre tienen una desviación típica igual a 1.

donde SPD es la abreviación de suma de productos de las diferencias, en este caso entre y y x, y SCD de suma de cuadrados de las diferencias de cada variable, expresiones que representan los contenidos del numerador y del denominador del coeficiente de correlación[5].

El coeficiente de correlación da valores que nos indican la intensidad y el sentido de la asociación entre las dos variables consideradas. Estos valores oscilan entre −1 y 1, valores extremos que indican la máxima correlación (correlación perfecta entre las variables), negativa o positiva, mientras que el 0 indica la ausencia de correlación (independencia entre las variables). Como medida de la relación entre dos variables es simétrica, es decir, la correlación entre y y x es la misma que entre x e y.

La evaluación de la importancia del coeficiente de correlación obtenido no es automática. Según el estudio, la naturaleza de los datos, las experiencias anteriores, etc., pueden determinar la relevancia de un coeficiente.

De forma general nos movemos entre grados de correlación débil y fuerte:



Que podrían ser valorados, de forma orientativa y sin tener en cuenta el contexto en el cual se aplican, de la forma siguiente:

$\|r\|=0$	Correlación lineal nula, independencia lineal
$0 <\|r\| \leq 0{,}2$	Correlación lineal muy débil
$0{,}2 <\|r\| \leq 0{,}5$	Correlación lineal débil
$0{,}5 <\|r\| \leq 0{,}7$	Correlación lineal media
$0{,}7 <\|r\| \leq 0{,}9$	Correlación lineal fuerte
$0{,}9 <\|r\| < 1$	Correlación lineal muy fuerte
$\|r\|=1$	Correlación lineal perfecta entre las variables

Para ejemplificar el cálculo y la interpretación del coeficiente de correlación consideraremos el caso de los gráficos anteriores donde hemos comentado las relaciones entre la valoración de un producto, los ingresos y la edad. Será también el mismo ejemplo que utilizaremos más tarde para los cálculos de la regresión. Los datos originales se reproducen en la Tabla III.9.1, con un total de 18 casos.

[5] En términos operativos la fórmula del coeficiente de correlación también se puede calcular con las tres siguientes expresiones:

$$r_{yx}=\frac{n\cdot\sum_{i=1}^{n}y_i\cdot x_i-\left(\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i\right)}{\sqrt{\left[n\cdot\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right]\cdot\left[n\cdot\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right]}}$$

$$r_{yx}=\frac{\sum_{i=1}^{n}y_i\cdot x_i-\frac{\sum_{i=1}^{n}y_i\sum_{i=1}^{n}x_i}{n}}{\sqrt{\left[\sum_{i=1}^{n}y_i^2-\frac{\left(\sum_{i=1}^{n}y_i\right)^2}{n}\right]\cdot\left[\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right]}}$$

$$r_{yx}=\frac{\frac{\sum_{i=1}^{n}y_i\cdot x_i}{n}-\bar{x}\cdot\bar{y}}{\sqrt{\left(\frac{\sum_{i=1}^{n}y_i^2}{n}-\bar{y}^2\right)\cdot\left(\frac{\sum_{i=1}^{n}x_i^2}{n}-\bar{x}^2\right)}}$$

Tabla III.9.1. Matriz de datos de valoración del producto, ingresos y edad

Caso	*Valoración del producto (V)*	*Ingresos (I)*	*Edad (E)*
1	18	90	23
2	11	110	33
3	23	70	21
4	13	125	39
5	17	100	27
6	19	115	43
7	22	85	21
8	20	85	27
9	9	150	43
10	14	130	38
11	21	75	30
12	11	145	54
13	8	160	63
14	10	155	58
15	22	80	25
16	12	105	51
17	15	100	36
18	16	120	34

En la Tabla III.9.2 se han añadido una serie de columnas que facilitan la realización de todos los cálculos manuales necesarios de la fórmula del coeficiente de correlación y también de la regresión.

Tabla III.9.2. Cálculos con valoración, ingresos y edad

	V	I	E	V^2	I^2	E^2	$V \times I$	$V \times E$	$I \times E$
1	18	90	23	324	810	529	1620	2070	414
2	11	110	33	121	1210	1089	1210	3630	363
3	23	70	21	529	490	441	1610	1470	483
4	13	125	39	169	1562	1521	1625	4875	507
5	17	100	27	289	1000	729	1700	2700	459
6	19	115	43	361	1322	1849	2185	4945	817
7	22	85	21	484	722	441	1870	1785	462
8	20	85	27	400	722	729	1700	2295	540
9	9	150	43	81	2250	1849	1350	6450	387
10	14	130	38	196	1690	1444	1820	4940	532
11	21	75	30	441	562	900	1575	2250	630
12	11	145	54	121	2102	2916	1595	7830	594
13	8	160	63	64	2560	3969	1280	10080	504
14	10	155	58	100	2402	3364	1550	8990	580
15	22	80	25	484	640	625	1760	2000	550
16	12	105	51	144	1102	2601	1260	5355	612
17	15	100	36	225	1000	1296	1500	3600	540
18	16	120	34	256	1440	1156	1920	4080	544
$\sum y_i$	281	2000	666						
$\sum y_i^2$				4789	23590	27448			
$\sum y_i \cdot x_i$							29130	79345	9518
$(\sum y_i)^2$	78961	4000000	443556						
$\bar{y}$	15,6	111,1	37						
s	4,9	28,4	12,8						
s^2	23,7	804,6	165,1						

La correlación entre los ingresos y la edad se obtiene de la siguiente forma:

$$r_{V,E} = \frac{\sum_{i=1}^{n} y_i \cdot x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sqrt{\left[\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right] \cdot \left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}} = \frac{79345 - \frac{2000 \times 666}{18}}{\left(235900 - \frac{4000000}{18}\right) \times \left(27448 - \frac{443556}{18}\right)} = 0,863$$

El resto de coeficientes se pueden obtener fácilmente a partir de los datos que hemos reproducido.

Con la ayuda del software estadístico hemos realizado estos cálculos para todas las correlaciones entre las tres variables, obteniendo los resultados de la Tabla III.9.3.

Tabla III.9.3. Correlación entre Valoración del producto, Edad e Ingresos

		Valoración	Edad	Ingresos
Correlación de Pearson	Valoración	1	-0,827	-,892
	Edad	-0,827	1	,863
	Ingresos	-0,892	0,863	1
Significación bilateral	Valoración		0,000	0,000
	Edad	0,000		0,000
	Ingresos	0,000	0,000	
Suma de cuadrados y productos cruzados	Valoración	402,278	-879,000	-2.092,222
	Edad	-879,000	2.806,000	5.345,000
	Ingresos	-2.092,222	5.345,000	13.677,778
Covarianza	Valoración	23,663	-51,706	-123,072
	Edad	-51,706	165,059	314,412
	Ingresos	-123,072	314,412	804,575
Casos	Valoración	18	18	18
	Edad	18	18	18
	Ingresos	18	18	18

La correlación entre ingresos y edad es positiva y alta, mientras que las correlaciones donde interviene la valoración del producto dan valores altos pero negativos: cuanta más edad y más ingresos se tienen menos se valora el producto de consumo.

Con valores altos de correlación es de esperar que estos sean significativos desde un punto de vista estadístico, sobre todo si la muestra es de un tamaño suficiente. Cuando se obtiene un coeficiente de correlación se puede realizar una prueba estadística para valorar la hipótesis nula de si el coeficiente poblacional, ρ_{xy}, es significativamente distinto de cero o, por el contrario, las diferencias en relación al cero se deben al azar y se puede concluir que no existe correlación.

A partir de una muestra aleatoria de valores (x,y) de n casos independientes entre sí, de una población normal o con muestras suficientemente grandes (a partir de 30 casos), la prueba estadística bilateral se plantea en los términos siguientes:

1. Formulación de las hipótesis
   H_0: El coeficiente de correlación es cero, $\rho_{xy}=0$
   H_A: El coeficiente de correlación no es cero, $\rho_{xy}\neq 0$

2. Cálculo del valor del estadístico muestral
   El estadístico de contraste sigue una distribución teórica de la t de Student y se calcula como:

$$t = \frac{r_{yx}}{\sqrt{\frac{1-r_{yx}^2}{n-2}}}$$

3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico a partir del valor concreto t_o del estadístico t.[6]

4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, con $n-2$ grados de libertad, la decisión se formaliza de la siguiente manera:
   Si $Pr(t_o) \geq \alpha$ aceptamos la hipótesis nula, la correlación lineal es cero.
   Si $Pr(t_o) < \alpha$ rechazamos la hipótesis nula, la correlación lineal no es cero.

Se trata de una prueba estadística que genera resultados significativos en cuanto tenemos correlaciones que se alejan de cero aunque sean débiles.

En el caso del ejemplo podemos ver como todas las correlaciones son significativas, la tabla anterior nos muestra en todos los casos una probabilidad de 0,000 de que no haya correlación. En general, si la probabilidad asociada es inferior al valor 0,05 se puede rechazar la hipótesis nula que afirma que el coeficiente de correlación es cero.

Es necesario destacar varios comentarios en relación a la interpretación y utilización del coeficiente de correlación:

- En primer lugar hay que hacer notar que no siempre toda correlación significativa expresa una dependencia entre las variables si no es sustentada por una interpretación teórica-conceptual que establezca la pertenencia de esta relación, debemos confirmar por tanto que esta correlación no sea espuria, que su vínculo no esté mediado por terceras variable que son las que lo motivan[7].

---

[6] Con tamaños muestrales a partir de 50 casos se puede emplear también la distribución normal con el estadístico siguiente: $t = \frac{|r_{yx}|}{\sqrt{n}}$.

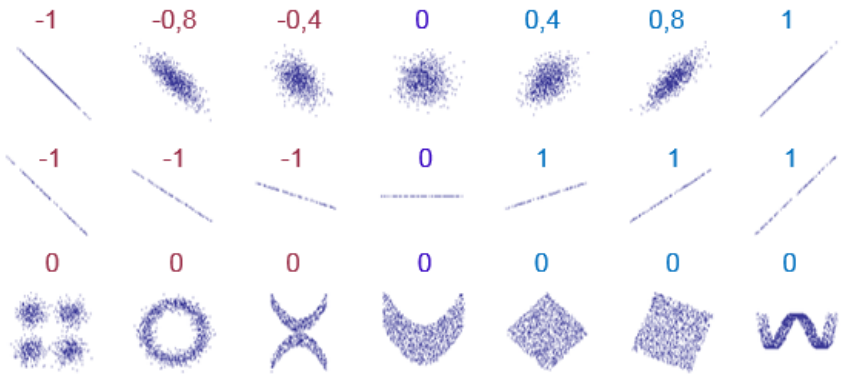
[7] Como señaló Neyman (1952) que de la correlación existente entre el número de cigüeñas y el número de niños y niñas nacidos no se derivan que los niños/as los traigan las cigüeñas. También se puede observar una correlación en los ingresos y la altura de las personas ¿a qué se debe? a que los ingresos son mayores entre los varones que entre

- De la obtención de un coeficiente de correlación elevado no siempre se debe concluir necesariamente la existencia de asociación lineal entre las variables. De hecho nos podemos encontrar con correlaciones importantes con variaciones de tipo exponencial.

- Por otra parte, un coeficiente casi nulo puede mostrar la no existencia de correlación de tipo lineal, pero las dos variables pueden estar fuertemente asociadas a partir de relaciones no lineales, por ejemplo, con forma de U. Por este motivo es importante la inspección visual previa de las representaciones gráficas. El Gráfico III.9.5 muestra diversas situaciones de nubes de puntos asociadas a diversos valores de correlación y el Gráfico III.9.6 diversas formas de gráficos de dispersión que se asocian a distintas funciones matemáticas. La correlación es importante en la medida en que la nube de puntos se acerca a una línea recta y deja de ser dispersa, esto es, deja de ser gruesa y los puntos se alinean. Formas diversas que evidencian un patrón de relación muestran una correlación que es nula desde el punto de vista lineal pero adoptan otras formas que se asimilan a otras forma funcionales o funciones matemáticas (exponencial, logarítmica, parabólica, o inversa).

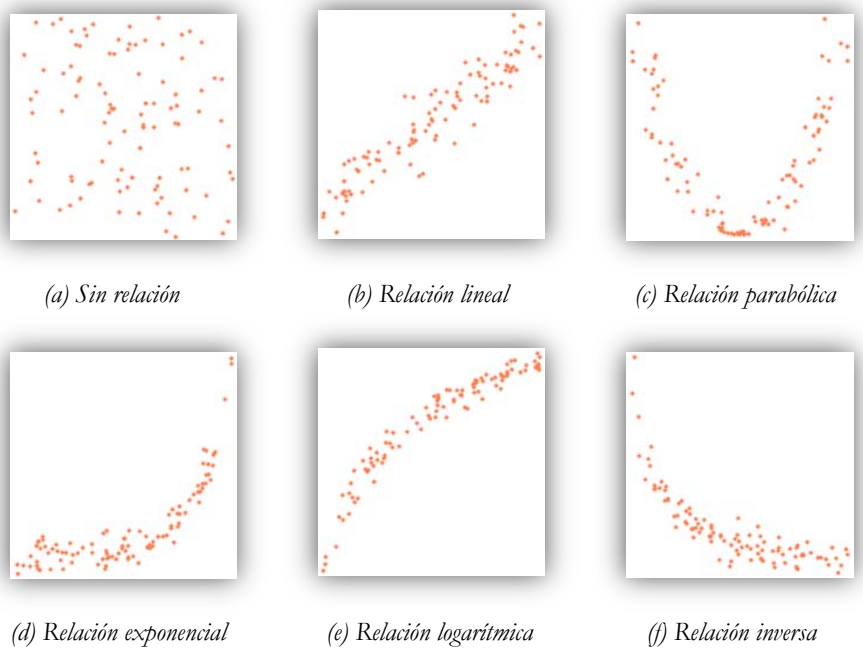
- A la hora de comparar coeficientes de correlación es necesario ser cautos. En general se puede afirmar la mayor o menor fuerza de la relación entre variables pero no que es tantas veces más o menos fuerte. Para ello es mejor utilizar el cuadrado del coeficiente de correlación, R^2, el llamado coeficiente de determinación, que se interpreta como la proporción de la variabilidad de una variable y explicada por otra x, es la variabilidad conjunta de las dos variables, o también la reducción proporcional del error cometido al predecir los valores de y a partir de la ecuación de regresión, mientras que $1-R^2_{yx}$ es la variabilidad de y no explicada por otra x (a la expresión se le llama coeficiente de alienación). Así por ejemplo, un r=0,4 indica que la variable x explica un 16% de la variabilidad de y, mientras que un r=0,8, explica un 64%, es decir, la relación es más del doble.

Gráfico III.9.5. Diagramas de dispersión y coeficientes de correlación



las mujeres y, como los primeros son más altos, ingresan más frente a las mujeres, las que son menos altas. La aparente correlación inicial es falsa pues está modulada por el sexo de las personas.

Gráfico III.9.6. Diagramas de dispersión y funciones de relación



(a) Sin relación *(b) Relación lineal* *(c) Relación parabólica*

(d) Relación exponencial *(e) Relación logarítmica* *(f) Relación inversa*

- Cuando decimos que la correlación es de 0,8 y afirmamos que el 64% de la variabilidad de *y* es explicada por *x*, también estamos afirmando que el 64% del variabilidad de *x* es debida a *y*, en otras palabras, que estamos ante una medida simétrica y que correlación es diferente de causación, con el coeficiente estimamos covariaciones, tan sólo determinamos si dos variables están relacionadas linealmente.

- El valor del coeficiente de correlación se puede ver afectado por la presencia de valores anómalos o influyentes.

- Finalmente, hay que tener presente que la significatividad del coeficiente de correlación aumenta cuando lo hace el número de casos, siempre que realicemos comparaciones entre coeficientes obtenidos de muestras con diferente tamaño se debe indicar el número de casos.

En la Tabla III.9.4 y en el Gráfico III.9.7 se presenta la información del análisis de correlación entre las variables Lifeexpectancy (*Life expectancy at birth*), GNIpercapita2011 (*Gross national income, GNI per capita 2011 PPP $*) y Schooling (*Mean years of schooling*) extraídas de la bases de de datos que publica Naciones Unidas en relación al estudio sobre el Índice de Desarrollo Humano (IDH).

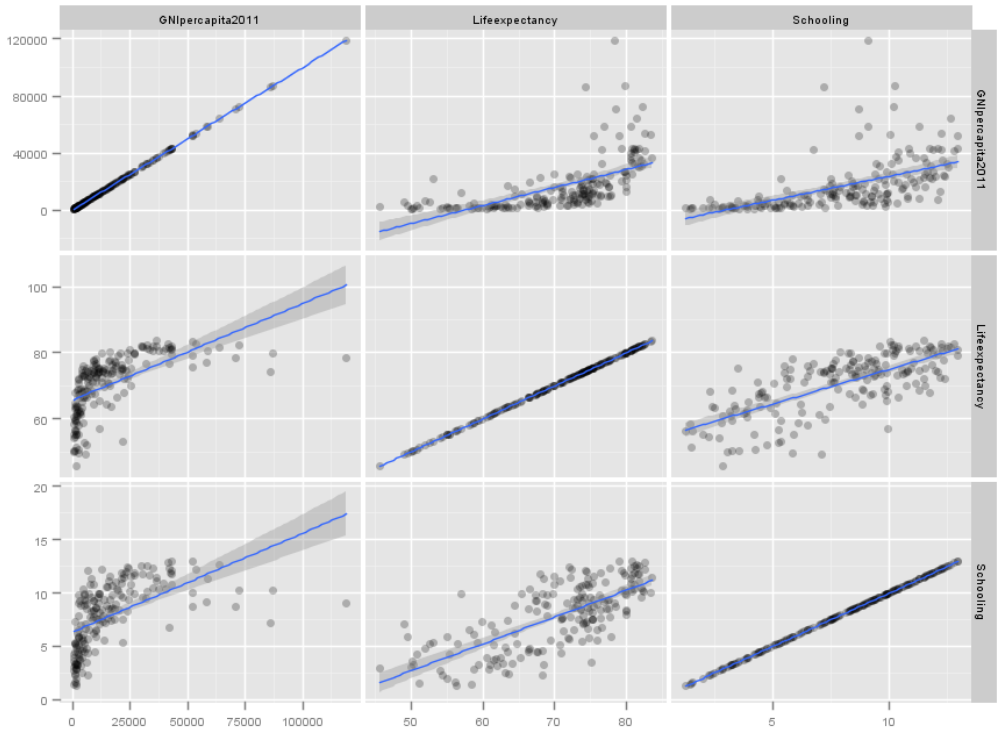
Se observa la relación lineal positiva y destacable que se da entre la esperanza de vida y la media de años de escolarización de los países (0,729), así como una relación más moderada con la renta per cápita (0,608). Si nos fijamos en los gráficos, la relación entre la esperanza de vida y la escolarización dibuja una forma lineal, pero en el caso de los gráficos donde interviene la variable de renta per cápita la forma de la nube de puntos es curvada mostrando la existencia de una relación entre las variables pero que

se acerca más a la forma exponencial o logarítmica. En consecuencia, cabe esperar que la intensidad de la relación sea mayor que la que expresa una medida del grado de relación lineal.

Tabla III.9.4. Correlación entre la Esperanza de vida, la Renta per cápita y los Años de escolarización de los países del mundo

		Lifeexpectancy	Schooling	GNIpercapita2011
Correlación de Pearson	Lifeexpectancy	1	0,729	0,608
	Schooling	0,729	1	0,560
	GNIpercapita2011	0,608	0,560	1
Significación bilateral	Lifeexpectancy		0,000	0,000
	Schooling	0,000		0,000
	GNIpercapita2011	0,000	0,000	
Suma de cuadrados y productos cruzados	Lifeexpectancy	14.994	3.663	18.644.175
	Schooling	3.663	1.740	5.876.421
	GNIpercapita2011	18.644.175	5.876.421	63.869.667.100
Covarianza	Lifeexpectancy	78,915	19,691	99.171
	Schooling	19,691	9,356	31.594
	GNIpercapita2011	99.171	31.594	337.934.747
Casos	Lifeexpectancy	191	187	189
	Schooling	187	187	187
	GNIpercapita2011	189	187	190

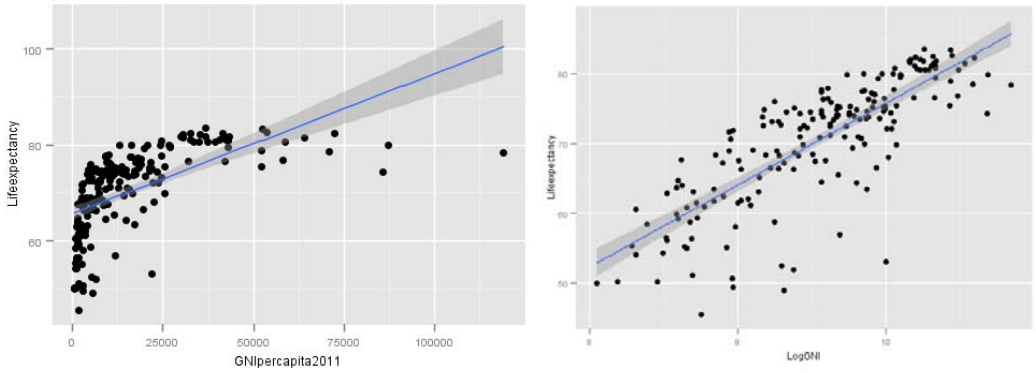
Gráfico III.9.7. Diagrama de dispersión matricial entre la Esperanza de vida, la Renta per cápita y los Años de escolarización de los países del mundo



Si consideramos, en particular, la relación entre la esperanza de vida y la renta per cápita, evidenciamos una forma semejante a la imagen (e) del Gráfico III.9.6, es decir, donde se manifiesta una relación de tipo logarítmica. En estos casos si se transforma la variable original, en nuestro caso la renta per cápita, por el logaritmo de ésta (variable LogGNI) se obtiene una representación que expresa linealmente la relación existente entre las dos variables (Gráfico III.9.8).

Si con esta transformación volvemos a calcular los coeficientes de correlación entre las variables obtenemos otras medidas más elevadas que no se expresaban con anterioridad (Tabla III.9.5). En concreto, la correlación entre entre la esperanza de vida y la renta per cápita ahora es del 0,802, una correlación incluso más alta que la que manifiesta con los años de escolarización.

Gráfico III.9.8. Diagramas de dispersión original y transformado entre la Esperanza de vida y la Renta per cápita



(a) Gráfico original *(b) Gráfico transformado*

Tabla III.9.5. Correlación entre la Esperanza de vida, los Años de escolarización y la Renta per cápita transformada

		Lifeexpectancy	Schooling	GNIpercapita2011
Correlación de Pearson	Lifeexpectancy	1	0,729	0,802
	Schooling	0,729	1	0,782
	GNIpercapita2011	0,802	0,782	1

► **Ejercicio 1.**

A partir de la variable **y** (por ejemplo *las notas de una asignatura*) elegimos los valores de las variables *v, w, x, z* (por ejemplo *las horas de estudio*) que se corresponden con las situaciones siguientes:

a) *y,x* tienen correlación perfecta positiva
b) *y,v* tienen correlación perfecta negativa
c) *y,w* tienen correlación nula
d) *y,z* tienen una correlación alta positiva

Los datos pueden ser los siguientes:

y	*x*	*v*	*w*	*z*
6	3	2	2	6
6	3	2	4	5
4	2	3	2	4
4	2	3	4	3

Realizamos el cálculo de la correlación entre *y* y *x* con la ayuda de la tabla siguiente:

	y_i	$(y_i-\bar{y})$	$(y_i-\bar{y})^2$	x_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})\cdot(y_i-\bar{y})$
	6	1	1	3	0,5	0,25	0,5
	6	1	1	3	0,5	0,25	0,5
	4	-1	1	2	-0,5	0,25	0,5
	4	-1	1	2	-0,5	0,25	0,5
Suma	20		4	10		1	2
Media	5		1,333	2,5		0,333	0,667

De donde el coeficiente de correlación es:

$$r_{yx}=\frac{s_{yx}}{s_y\cdot s_x}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\frac{SPD_{yx}}{\sqrt{SCD_y\cdot SCD_x}}=\frac{2}{\sqrt{4\times 1}}=1$$

El resto de las correlaciones se pueden calcular siguiendo el mismo procedimiento comprobando los datos con los que aparecen en la tabla siguiente:

		Y	X	V	W	Z
Correlación de Pearson	Y	1	1,000	-1,000	,000	,894
	X	1,000	1	-1,000	,000	,894
	V	-1,000	-1,000	1	,000	-,894
	W	,000	,000	,000	1	-,447
	Z	,894	,894	-,894	-,447	1
Suma de cuadrados y productos cruzados	Y	4,000	2,000	-2,000	,000	4,000
	X	2,000	1,000	-1,000	,000	2,000
	V	-2,000	-1,000	1,000	,000	-2,000
	W	,000	,000	,000	4,000	-2,000
	Z	4,000	2,000	-2,000	-2,000	5,000
Covarianza	Y	1,333	,667	-,667	,000	1,333
	X	,667	,333	-,333	,000	,667
	V	-,667	-,333	,333	,000	-,667
	W	,000	,000	,000	1,333	-,667
	Z	1,333	,667	-,667	-,667	1,667

3. El análisis de regresión simple

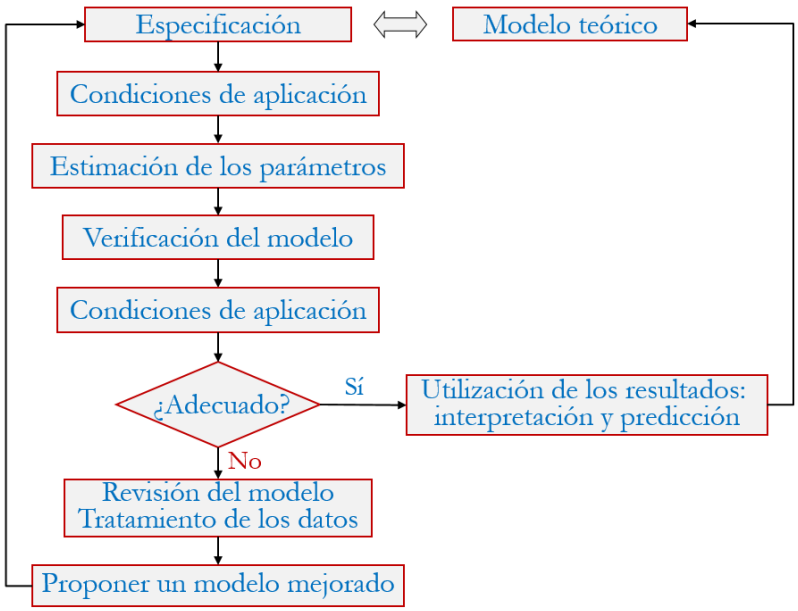
Por análisis de regresión simple entenderemos el análisis que tiene en cuenta la relación de una variable dependiente con un solo factor causal o variable independiente. En este contexto destacamos en primer lugar que el análisis de correlación que hemos visto en el apartado anterior está estrechamente vinculado con la regresión pues: se asumen las mismas condiciones de nivel de medida de las variables, se establecen relaciones lineales y el coeficiente de determinación es un indicador de la bondad de ajuste de la recta de regresión a la nube de puntos.

Pero correlación y regresión persiguen objetivos diferentes. En el primer caso se trata de medir la relación lineal, en el segundo caso, se trata de describir la naturaleza de la relación, cuantificarla, explicarla y realizar predicciones. Por otra parte hay que tener presente que la correlación es una medida simétrica de la relación entre dos variables, la regresión sin embargo implica la distinción de cuál es la variable dependiente y cuál la independiente, no es lo mismo la regresión de y sobre x, que de x sobre y.

A continuación desarrollaremos el modelo del análisis de regresión en cuatro apartados: la especificación del modelo lineal de la regresión; las condiciones de aplicación; la estimación de los parámetros de este modelo y su significación; la varianza explicada por la recta de regresión y la verificación de la bondad del ajuste de la regresión; y finalmente otros análisis adicionales.

Estos diferentes aspectos hacen referencia a las distintas tareas implicadas en un proceso de análisis de regresión y que podemos esquematizar como aparece en el Gráfico III.9.9.

Gráfico III.9.9. Esquema del proceso de un análisis de regresión



En él se destaca en primer lugar la importancia de la especificación del modelo de regresión a partir de una selección de las variables que debe obedecer a criterios

teóricos que fundamenten la elección y la posterior interpretación de los resultados. Bajo ese modelo, se trata de comprobar que se dan algunas de las condiciones de aplicación de un análisis de regresión para proceder a estimar los parámetros del modelo, esto es, los coeficientes de la ecuación de regresión. Seguidamente se establece su verificación, es decir, se determina la capacidad explicativa del modelo y se comprueban otras condiciones de aplicación a partir de los resultados generados en la regresión.

Si el modelo se considera adecuado se procede a la interpretación y al razonamiento predictivo que proporciona el modelo, contrastando las hipótesis que se hayan formulado y extrayendo las consiguientes conclusiones teóricas del análisis. En caso contrario se tratará de revisarlo tanto en los aspectos de tratamiento de los datos como de especificación de un nuevo modelo que finalmente sea adecuado y pueda interpretarse.

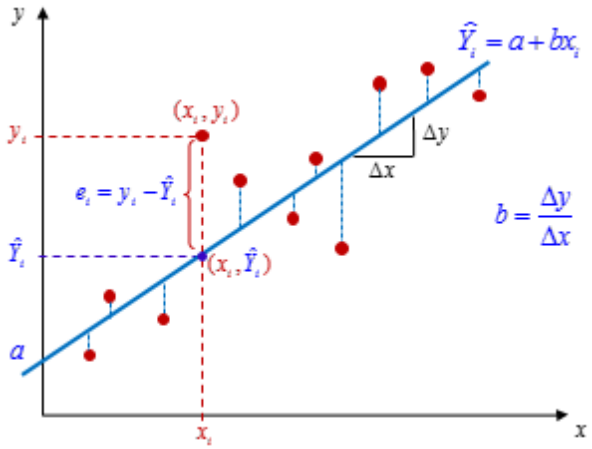
3.1. Especificación del modelo

El modelo de regresión es el modelo lineal. Si consideramos una población sobre la que se han observado dos variables, una dependiente, y (por ejemplo, la escala de valoración de un producto) y otra independiente predictora de la dependiente, x (por ejemplo, los ingresos), observadas para cada individuo i ($i=1,...,n$) de una muestra de la población, la especificación del modelo matemático clásico para predecir el valor de y_i a partir de los valores x_i es la *Ecuación 3* y la *Ecuación 4* que vimos anteriormente:

$$y_i = a + bx_i + e_i = \hat{Y}_i + e_i$$

donde $\hat{Y}_i$ son los valores de y_i dados por la recta de regresión para cada x_i, siendo a y b los parámetros del modelo y e_i el término de error o residuo, es decir, la diferencia entre el valor observado y_i y el dado por la recta de regresión $\hat{Y}_i$. En el Gráfico III.9.10 se representa esta información del modelo.

Gráfico III.9.10. Modelo lineal de regresión simple

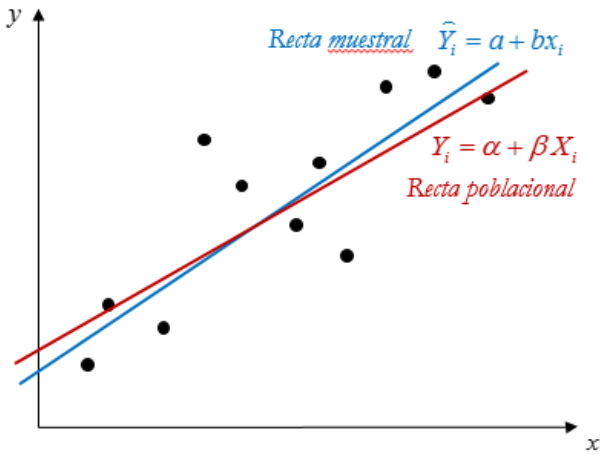


El parámetro a es la constante, el valor que adopta y_i cuando x_i es igual a cero. En el gráfico corresponde al punto donde la recta corta el eje vertical, el eje de ordenadas, denominado ordenada en el origen o intersección. Por su parte, el parámetro b corresponde a la inclinación o pendiente de la recta, es el llamado coeficiente de regresión e indica cuánto se varía y_i cuando x_i varía. En particular, cuantifica cuánto se incrementa o reduce y_i por cada variación de una unidad de x_i.

La ecuación $\widehat{Y}_i = a + bx_i$ sin el término de error es la ecuación de regresión que representa una recta. Si consideramos la expresión, para cada individuo i, el modelo lineal de regresión se puede expresar como: $y_i = \widehat{Y}_i + e_i$, donde $\widehat{Y}_i$ representa la media (o valor esperado) de la variable y para el individuo i dado por la recta de regresión, es decir, el valor que se espera en la población para todos los individuos con un valor como el del individuo i, siendo e_i el error que se comete al representar al individuo por ese valor medio de la recta.

Los valores de a y b se refieren a los valores observados y, por tanto, son el resultado de medidas y de muestras observadas, por lo que los valores para toda la población no coincidirán con estas estimaciones muestrales. Cuando trabajamos con muestras de la población hacemos estimaciones de los parámetros verdaderos: a y b serán una estimación más o menos buena de los valores correspondientes al total de la población α y β. Así pues obtenemos de una recta de regresión muestral que estima la recta de regresión poblacional:

Gráfico III.9.11. Recta de regresión muestral y poblacional



3.2. Condiciones de aplicación

La utilización del modelo de regresión precisa que se cumplan las siguientes condiciones o hipótesis del modelo lineal, en particular, para poder extrapolar los resultados de una muestra al conjunto de la población[8]. Son condiciones que aplicaremos tanto a los análisis de regresión simple como múltiple. Siguiendo a

[8] Habitualmente trabajamos con muestras estadísticas donde este razonamiento es necesario, pero si los datos son poblacionales la inferencia no es relevante y la descripción del modelo de regresión resultante se realiza al margen de las condiciones inferenciales, no obstante, otras condiciones como la linealidad deberán verificarse.

Domènech y Riba (1985) enumeraremos estas condiciones generales y más adelante detallaremos cómo tratar de forma específica la verificación de su cumplimiento o el tratamiento necesario para resolver el problema de su incumplimiento.

a) Hipótesis básicas (modelo descriptivo):

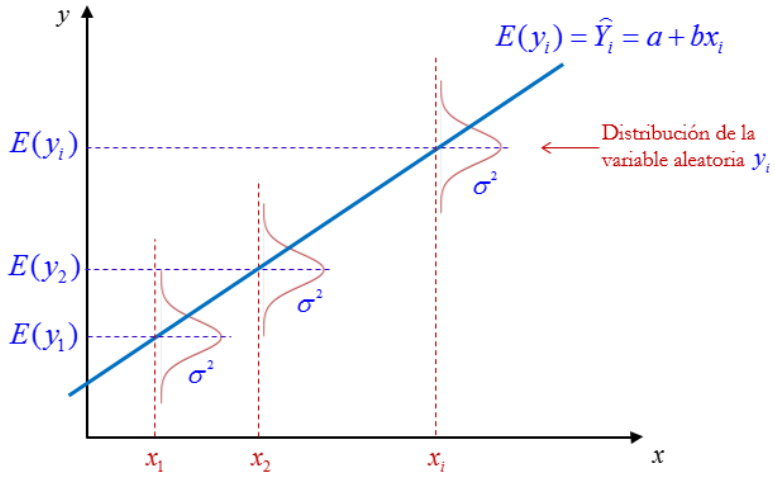
1. Las observaciones se corresponden con una muestra aleatoria de la población.
2. Las observaciones son una muestra aleatoria suficiente de la población. No existe una regla establecida, Harrell (2001) propone un mínimo de 10 o 20 veces el número de variables independientes, otros autores sugieren 50 o 100 casos más el número de variables independientes del modelo. En todo caso cuanto mayor sea la muestra mejor, pero depende de cada contexto de investigación.
3. El modelo está correctamente especificado: no se han dejado de considerar variables relevantes y no se han incluido variables independientes irrelevantes.
4. La relación de la variable dependiente con las independientes es lineal.
5. Las variables independientes se miden sin error, se corresponden con valores exactos.
6. Las medias o los valores esperados de las observaciones, $E(y_i)$, están situadas sobre una recta de regresión verdadera: $E(y_i) = \hat{Y}_i = a + bx_i$, es decir, en términos poblacionales que las medias de Y para cada valor de X están en la recta de regresión.

b) Hipótesis adicionales (modelo inferencial) que aseguran que a y b son las mejores estimaciones de los parámetros poblacionales y se pueden aplicar las pruebas de hipótesis.

7. Normalidad: se asume que la variable dependiente, para cualquier valor de la variable independiente, sigue una distribución normal con el valor de la media que corresponde a la recta poblacional.
8. Homoscedasticidad: las varianzas de la variable dependiente son iguales para los diferentes valores de la variable independiente:
$$var(Y_1) = var(Y_2) = \ldots = var(Y_i) = \sigma^2.$$
9. Ausencia de autocorrelación: las variables aleatorias Y_i y Y_j, ligadas a cualquier par de valores X_i y X_j de la variable independiente, son estocásticamente independientes: $cov(Y_i, Y_j)=0$.
10. Ausencia de (multi)colinealidad cuando el modelo es de regresión múltiple, esto es, cuando existe una combinación lineal entre las variables independientes, y evitar así la inflación de los errores típicos de las estimaciones.

Los supuestos 6, 7, 8 y 9 se pueden enunciar en términos de los errores: los errores ε_i son variables aleatorias independientes distribuidas según la ley normal con media 0 y varianza σ^2, es decir, $N(0, \sigma^2)$, y $cov(\varepsilon_i, \varepsilon_j)=0$. El Gráfico III.9.12 ilustra esta afirmación. A cada valor de x_i le corresponde una población de valores de y_i que se asume que se distribuye normalmente y que está centrada en el valor esperado de y_i, lugar por donde pasa la recta de regresión poblacional. Cada una de estas distribuciones se asume que tiene igual varianza.

Gráfico III.9.12. Condiciones de aplicación de la regresión. Distribución de los errores



3.3. Estimación y significación de los parámetros del modelo

Con el modelo especificado se trata de calcular los parámetros del mismo a y b que son los que definen la recta de regresión. Para ello se precisa un método de ajuste de la nube de puntos a la recta. El método que se emplea es aquel que hace mínimo el error que se comete al predecir los valores de la variable dependiente, es decir, e_i, la diferencia entre el valor de la variable y_i, para cada individuo, y el valor $\hat{Y}_i$ de esta variable obtenido con la recta de regresión. Por tanto, el que minimiza las distancias verticales de todos los puntos a la recta (ver Gráfico III.9.10). El ajuste se hace por el método de mínimos cuadrados ordinarios (MCO), método matemático que consiste en calcular los coeficientes a y b de forma que la suma de los cuadrados de los errores sea mínima:

$$Min \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \qquad \textit{Ecuación 8}$$

Estos valores se encuentran aplicando el teorema fundamental del cálculo infinitesimal: son las soluciones del sistema de ecuaciones que resulta de igualar a cero las derivadas parciales de la función con respecto a los parámetros a y b. Realizando estos cálculos se encuentra que los valores que minimizan la función anterior tienen la expresión:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} - b \cdot \frac{\sum_{i=1}^{n} x_i}{n} = \bar{y} - b\bar{x} \qquad \textit{Ecuación 9}$$

$$b = \frac{\sum_{i=1}^{n} y_i \cdot x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{SPD_{yx}}{SCD_x} \qquad \textit{Ecuación 10}$$

De la *Ecuación 9* se deriva que la recta de regresión pasa por el punto $(\overline{x},\overline{y})$. Con estos valores la recta de regresión se puede expresar como:

$$\hat{Y}_i = a + bx_i = \overline{y} - b\cdot(x_i - \overline{x}) \qquad \textit{Ecuación 11}$$

Si aplicamos estas fórmulas al caso anterior del análisis de la relación entre la valoración del producto de consumo considerada como variable dependiente para ser explicada en función de los ingresos obtenemos los siguientes resultados:

$$b = \frac{\sum_{i=1}^{n} y_i \cdot x_i - \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{29130 - \frac{281\times 2000}{18}}{235900 - \frac{(2000)^2}{18}} = -0{,}153$$

con el valor de b calculamos el de a:

$$a = \overline{y} - b\overline{x} = 15{,}6 - (-0{,}153)\times 111{,}1 = 32{,}607$$

Por lo tanto, la ecuación de regresión que describe la nube de puntos es $\hat{Y}_i = 32{,}61 - 0{,}15x_i$. En la Tabla III.9.6 se presentan estos resultados, como suelen aparecer con el software estadístico. Junto a la estimación de los coeficientes de regresión que describen la relación entre las variables aparece la información de las pruebas de significación de los coeficientes y los intervalos de confianza.

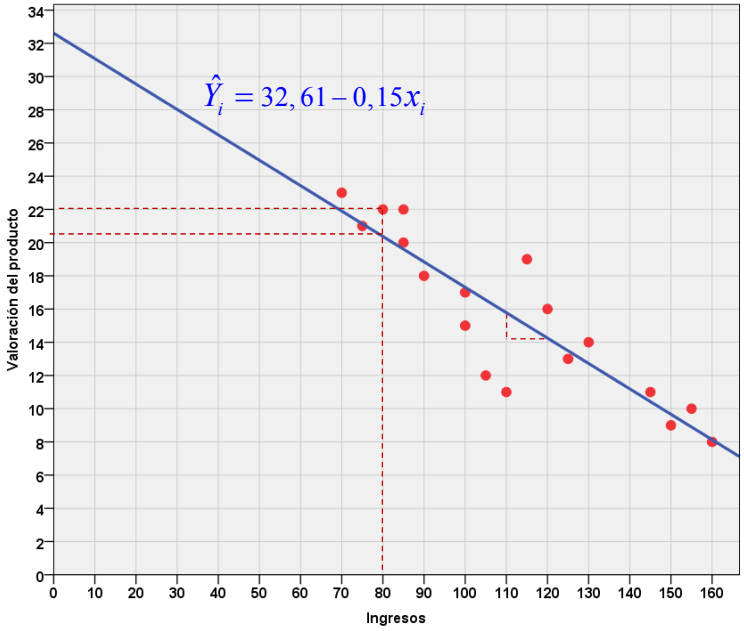
Tabla III.9.6. Tabla de coeficientes de regresión entre Valoración e Ingresos

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	32,607	2,219		14,693	0,000	27,903	37,312
Ingresos	−0,153	0,019	−0,892	−7,891	0,000	−0,194	−0,112

A continuación veremos cómo se interpretan estos valores con la representación gráfica de la recta de regresión (Gráfico III.9.13). El coeficiente de regresión b, en este caso −0,153, representa la inclinación de la recta, la pendiente de la recta de regresión respecto del eje de abscisas, indicando el número de unidades que se espera que varíe la valoración cuando los ingresos varían en una unidad. Así, una unidad más de ingresos (1 euro más) supone 0,153 de reducción en la escala de valoración. De la misma forma podemos considerar otra cantidad de unidades, por ejemplo, 10 euros. Si miramos en el gráfico el paso de 110€ a 120€ implica bajar la valoración en 1,53 puntos. Por su parte, la constante es el valor de la ordenada en el origen, donde corta la recta de regresión del eje de ordenadas, es el valor esperado correspondiente a no tener ingresos, 32,61 en este caso. No obstante este valor no tiene mayor relevancia pues en nuestros datos no hemos observado a nadie con cero ingresos. En un análisis

de regresión solamente se interpreta el tramo de la recta que corresponde a valores observados de x.

Gráfico III.9.13. Gráfico de dispersión de la regresión entre Valoración e Ingresos



El valor que se obtiene de la recta de regresión para cada valor x_i representa la media prevista de los valores de la valoración de y_i para los individuos que tienen unos ingresos de una cantidad determinada. En este sentido podemos predecir cuál será el grado en que una persona es más o menos favorable en relación al producto. Por ejemplo, la valoración media que cabe esperar de una persona con unos ingresos de 80 euros será de 20,6:

$$\hat{Y}_i = 32,61 - 0,15 \times 80 = 20,6$$

En los datos del ejemplo tenemos un caso observado con unos ingresos de 80 euros que realiza una valoración de 22 puntos en la escala. En este caso, y en todos los casos con unos ingresos de 80, se espera que en promedio valoren el producto en 20,6. Por tanto, en este caso se genera un residuo, un valor por encima del esperado por la recta de regresión, de 22-20,6=1,4.
Hay que tener en cuenta, como hemos comentado, que los valores que se pueden predecir, los que tienen valor estadístico, son aquellos que se obtienen a partir de valores de la variable independiente que estén en un rango de variación limitado por los valores observados en la muestra.

Por otra parte, un aspecto que también hay que destacar es que las variables x e y de la recta de regresión se han considerado en su propia unidad de medida. Cuando introduzcamos una segunda variable independiente en un modelo de regresión múltiple será necesario hacer comparables las variaciones de cada variable independiente en una misma unidad de medida por lo que habrá que estandarizarlas.

Lo veremos más adelante. En la Tabla III.9.6 podemos ver que los coeficientes aparecen primero sin estandarizar y luego estandarizados o tipificados. Cuando se estandarizan los coeficientes, llamados beta, se expresan en unidades de desviación típica y el coeficiente de la constante es cero. Por tanto, la ecuación de regresión estandarizada es $\hat{Y}_i^s = -0,892x_i^s$. La variable independiente de ingresos tiene una desviación típica de 28,365 euros mientras que la de la dependiente valoración es 4,865 puntos en la escala. Así, la ecuación estandarizada se interpreta de la forma siguiente: por cada unidad de desviación típica que varíen los ingresos, es decir, por cada 28,365 euros más, la valoración cabe esperar que se vea reducida en 0,892 unidades de desviación, es decir, en $\hat{Y}_i^s = -0,892 \times 4,865 = -4,340$ puntos. Es el mismo valor que se obtendría en la ecuación de regresión no estandarizada: $\hat{Y}_i = 32,61 - 0,15 \times 28,365 = -4,340$. En regresión simple el coeficiente de regresión estandarizado coincide con el coeficiente de correlación entre las dos variables.

Una vez se han obtenido valores a y b de la recta de regresión se plantea la cuestión de su significación estadística. La recta de regresión que se obtiene a partir de los datos muestrales es una estimación de la verdadera recta de regresión poblacional como vimos. Por lo tanto, los parámetros a y b son estimaciones de los verdaderos parámetros poblacionales α y β con un error determinado. La prueba estadística de significación contrasta la hipótesis nula de que el parámetro poblacional β sea cero, y como alternativa que sea diferente de cero. Es decir, se trata de determinar si la inclinación de la recta es o no distinta de cero, si es distinta de una recta plana que implicaría constancia en la variación de y_i. La prueba estadística se plantea en los siguientes términos:

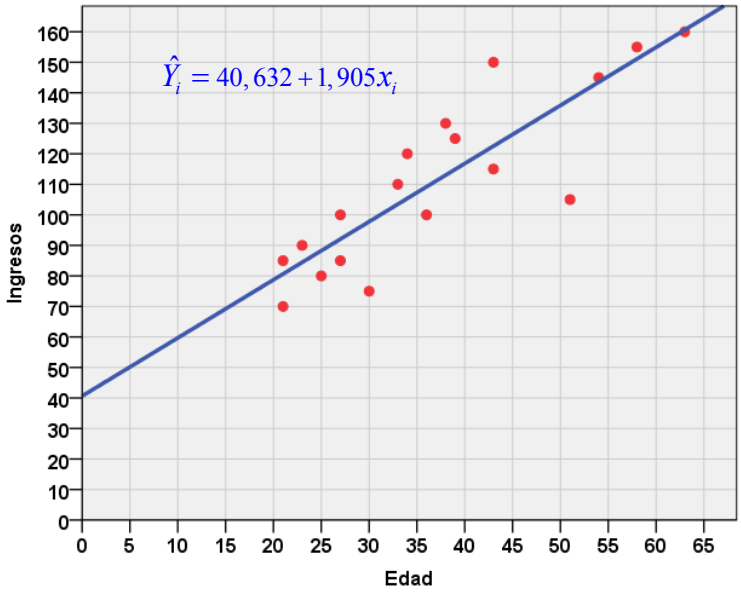
1. Formulación de las hipótesis
   H_0: El parámetro poblacional es cero, $\beta=0$
   H_A: El parámetro poblacional no es cero, $\beta\neq0$
2. Cálculo del valor del estadístico muestral
   El estadístico tiene la expresión: $t = \dfrac{b}{s_b}$
   con un error típico de $s_b = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{(n-2)\cdot\sum_{i=1}^{n}(x_i - \bar{x})^2}$
   y se distribuye como una t-Student con $v=n-2$ grados de libertad.
3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico a partir del valor concreto t_o del estadístico t.
4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, con $n-2$ grados de libertad, la decisión se formaliza de la siguiente manera:
   Si $Pr(t_o) \geq \alpha$ aceptamos la hipótesis nula, el coeficiente de regresión es cero.
   Si $Pr(t_o) < \alpha$ rechazamos la hipótesis nula, el coeficiente de regresión no es cero.

El intervalo de confianza se obtiene mediante la expresión: $b \pm t \cdot s_b$.

Con la ayuda del software estadístico los resultados de la estimación de los parámetros de la recta de regresión que se obtienen son los que se reproducen en la Tabla III.9.6, con la prueba de significación correspondiente y el intervalo de confianza. Como podemos ver el coeficiente es significativo y podemos rechazar la hipótesis nula de que sean nulos.

Si consideramos ahora el caso de la relación entre los ingresos y la edad con los datos del ejemplo presentado nos encontramos con una relación de tipo lineal positiva (Gráfico III.9.14).

Gráfico III.9.14. Gráfico de dispersión de la regresión entre Ingresos y Edad



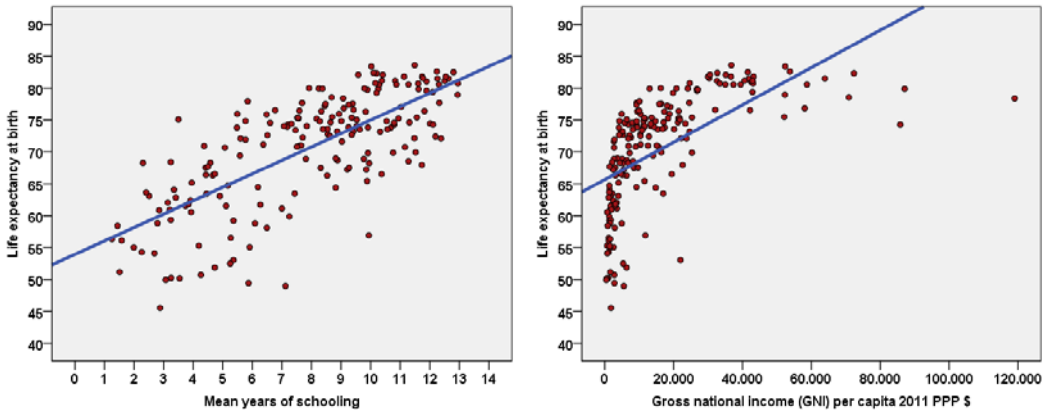
La ecuación de regresión de la Tabla III.9.7 muestra la significación del coeficiente de regresión. Su valor de 1,905 nos indica que por cada año más de edad los ingresos diarios se incrementan en casi 2 euros.

Tabla III.9.7. Tabla de coeficientes de regresión entre Ingresos y Edad

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	40,632	10,897		3,729	0,002	17,530	63,733
Edad	1,905	0,279	0,863	6,826	0,000	1,313	2,496

Tomemos ahora el ejemplo de los datos sobre el Índice de Desarrollo Humano de los países y efectuemos las regresiones simples entre la esperanza de vida al nacer (Lifeexpectancy) como variable dependiente y las variables media de años de escolarización (Schooling) y renta per cápita (GNIpercapita2011) como independientes, se obtienen los resultados que se muestran en el Gráfico III.9.15, la Tabla III.9.8 y la Tabla III.9.9.

Gráfico III.9.15. Gráficos de dispersión de la regresión entre la Esperanza de vida al nacer y los Años de escolarización y la Renta per cápita



En los dos casos nos encontramos unos coeficientes de regresión significativos que muestran una relación positiva con la esperanza de vida, es decir, que a medida que aumentan los valores de escolarización y de renta de un país aumenta la esperanza de vida.

Tabla III.9.8. Tabla de coeficientes de regresión entre la Esperanza de vida al nacer y los Años de escolarización

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	53,952	1,232		43,802	0,000	51,522	56,382
Schooling	2,105	0,145	0,729	14,472	0,000	1,818	2,392

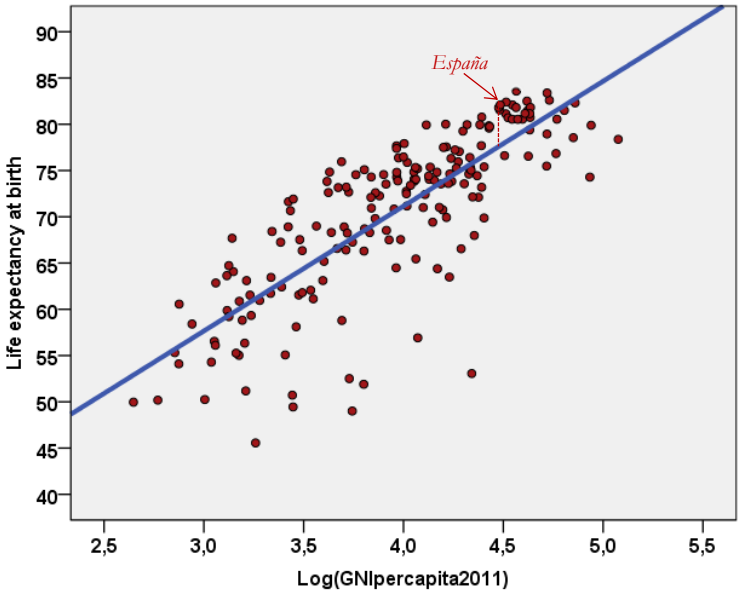
En el primer caso, por cada año de más de escolarización que tenga un país se espera que la esperanza de vida aumente en 2,105 años. En el segundo, por cada dólar más de producto interior bruto per cápita del país la esperanza de vida se espera que aumente en 0,00029 años, o lo que es lo mismo, por cada 10.000 dólares de incremento de la renta per cápita la esperanza de vida aumentará en 2,9 años.

Tabla III.9.9. Tabla de coeficientes de regresión entre la Esperanza de vida al nacer y la Renta per cápita

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	65,670	0,691		95,096	0,000	64,308	67,033
GNIpercapita2011	0,00029	0,00003	0,608	10,470	0,000	0,00024	0,00035

No obstante, en el caso de la renta, como comentamos anteriormente y se muestra en el Gráfico III.9.15, la relación entre las variables es deficiente desde el punto de vista lineal. Si transformamos la variable calculando el logaritmo en base 10 de la variable GNIpercapita2011, y la llamamos LogGNI, la nube de puntos que se obtiene es la del Gráfico III.9.16.

Gráfico III.9.16. Gráfico de dispersión de la regresión entre la Esperanza de vida al nacer y el logaritmo decimal de la Renta per cápita



La ecuación de regresión estimada da lugar a un coeficiente de regresión de 13,499 (Tabla III.9.10), un valor que se interpreta de la forma siguiente: por cada incremento de una unidad en el logaritmo de la renta per cápita, la esperanza de vida aumenta en 13,499 años. Así, si consideramos un valor de la variable independiente de 1 (lo que corresponde a una renta de 10 dólares, si calculamos el antilogaritmo, es decir, si operamos 10^1) la esperanza de vida se verá aumentada en 13,499. Si, por ejemplo, el valor fuera de 4 (lo que corresponde a una renta de 10.000 dólares, si hacemos el antilogaritmo, es decir, 10^4), el aumento sería de: $13{,}499 \times 4 = 53{,}996$ años. En consecuencia, los países con una renta per cápita de 10.000 dólares esperan alcanzar, en promedio, una esperanza de vida de: $17{,}161 + 13{,}499 \times 4 = 71{,}157$ años.

Tabla III.9.10. Tabla de coeficientes de regresión entre la Esperanza de vida al nacer y el logaritmo de la Renta per cápita

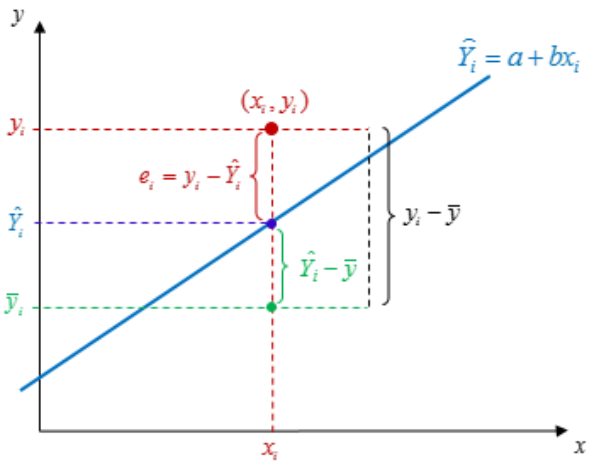
Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	17,161	2,935		5,847	0,000	11,371	22,951
LogGNI	13,499	0,736	0,802	18,335	0,000	12,047	14,951

Si miramos los datos para el caso de España, la renta per cápita es de 30.561,47 dólares, mientras que su esperanza de vida es de 82,10 años. Si calculamos el valor de la recta para el caso de España, esto es, el valor pronosticado por el modelo de regresión, obtenemos este resultado: $17{,}161 + 13{,}499 \times \log_{10}(30561{,}47) = 77{,}706$ años, un valor esperado que infraestima el valor observado de 82,10, generando un residuo positivo de $82{,}10 - 77{,}706 = 4{,}393$ años.

3.4. Verificación del modelo: la bondad de ajuste

Después de estimar los parámetros que describen la relación lineal entre las dos variables según el modelo de regresión, y ver su significación, se trata de verificarlo, de medir la calidad del ajuste, lo que significa también medir su poder de predicción. Este ejercicio de verificación es equivalente a ver si las observaciones están más o menos agrupadas en torno a la recta de regresión que se ha calculado. Si es así, predeciremos el valor de la variable dependiente con un error bajo, si las observaciones se dispersan, se alejan de la recta, la predicción será débil y el error cometido alto. Por tanto, en realidad, aunque sean mínimos los valores de los términos de error e_i, la parte no explicada por la recta de regresión, no significa que se haya suprimido el error; una parte de la varianza de y será explicada por la variable x pero otra quedará inexplicada. Se trata de determinar qué proporción representa la parte explicada. En el Gráfico III.9.17 se representa esta idea.

Gráfico III.9.17. Verificación del modelo de regresión



Si tuviéramos que realizar la predicción del valor $\hat{Y}_i$ sin tener la información de la variable independiente x_i, una posible predicción sería dar el valor de la media de las y_i para todos los valores posibles de x_i. En esta situación se ajustaría una recta de regresión plana que pasaría por la media de y, eso significaría que ambas variables serían linealmente independientes. Si lo hacemos así cometemos un error elevado e igual a $(y_i - \bar{y})$, la distancia en negro en el gráfico. Si, por otro lado, utilizamos la información de la recta de regresión el error será $(y_i - \hat{Y}_i)$, la distancia en rojo, menor que la cantidad anterior porque el modelo ha explicado una parte de la desviación: $(\hat{Y}_i - \bar{y})$, la distancia en verde. Por lo tanto, para un individuo i, la desviación total respecto a la media de la muestra se descompone en dos partes: la desviación explicada por el modelo de regresión y la desviación no explicada:

$$(y_i - \bar{y}) = (\hat{Y}_i - \bar{y}) + (y_i - \hat{Y}_i)$$

Esta distinción nos remite al planteamiento que hemos visto en el análisis de varianza en el capítulo anterior. En este caso a partir de las componentes de la variable dependiente en el contexto de un análisis de regresión.

A partir de la expresión anterior, si elevamos al cuadrado y tomamos sumatorios, el criterio mínimo cuadrático garantiza que:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{Variación Total}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{y})^2}_{\text{Variación Explicada}} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{Y}_i)^2}_{\text{Variación No Explicada}} \qquad \textit{Ecuación 12}$$

$$SCD_{Total} = SCD_{Regresión} + SCD_{Residual}$$

Es decir, que se puede descomponer la variación total (o la suma de cuadrados de las diferencias total) de la variable dependiente en una parte explicada y una parte no explicada. La variación total se identifica también como una medida del error total. Por su parte, la variación explicada es el error sistemático, predicho, explicado por la recta de regresión. Finalmente el error no predicho, no explicado, es el error residual o aleatorio.

Se puede demostrar además que el cálculo de la suma de cuadrados de la regresión se puede expresar como:

$$SCD_{\text{Regresión}} = b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{Ecuación 13}$$

de donde:

$$SCD_{\text{Residual}} = SCD_{\text{Total}} - b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad \textit{Ecuación 14}$$

Esta descomposición de la suma de cuadrados de las diferencias nos permite plantear dos tipos de cálculos y de interpretaciones destinadas a la verificación del modelo de

regresión: el coeficiente de determinación, que nos indicará el poder de predicción del modelo, y un análisis de varianza, que nos proporciona una prueba de significación global de la regresión, la bondad de ajuste del modelo.

3.4.1. *El coeficiente de determinación*

El llamado coeficiente de determinación nos permite medir el poder de predicción del modelo de regresión, cuál es la parte relativa explicada por la recta de regresión, es decir, su bondad de ajuste. Como habíamos señalado anteriormente al hablar de la correlación, este coeficiente mide la proporción de la varianza total explicada por el modelo de regresión lineal y se calcula de la siguiente forma:

$$R^2_{yx} = \frac{SCD_{Total} - SCD_{Residual}}{SCD_{Total}} = \frac{SCD_{Regresión}}{SCD_{Total}} \qquad \textit{Ecuación 15}$$

Se puede demostrar que R^2_{yx} es el cuadrado del coeficiente de correlación:

$$R^2_{yx} = \frac{SPD^2_{yx}}{SCD_y \cdot SCD_x} = \left(\frac{SPD_{yx}}{\sqrt{SCD_y \cdot SCD_x}} \right)^2 = r^2_{yx} \quad \Rightarrow \quad r_{yx} = \sqrt{R^2_{yx}}$$

Igualmente se cumple la siguiente relación entre el coeficiente de la regresión y el coeficiente de correlación:

$$b = r_{yx} \cdot \frac{S_y}{S_x} \qquad \textit{Ecuación 16}$$

El coeficiente de determinación ajustado RA^2_{yx} es un cálculo que corrige el valor inicial teniendo en cuenta el número de variables independientes del modelo (p) y el tamaño de la muestra (n) y refleja más fielmente la variación explicada por la variable independiente. El resultado del ajuste es siempre una pequeña reducción del valor del R^2_{yx}. Su expresión es:

$$RA^2_{yx} = R^2_{yx} - \frac{p \cdot (1 - R^2_{yx})}{n - p - 1} \qquad \textit{Ecuación 17}$$

En el ejemplo de la relación entre la valoración del producto y los ingresos habíamos visto la existencia de una fuerte asociación al calcular el coeficiente de correlación entre las variables, un valor de 0,892. El coeficiente de determinación, el cuadrado del coeficiente de correlación, es de 0,796, que se interpreta en el sentido siguiente: a partir de los ingresos se puede predecir el 80% de la varianza de la variable dependiente de valoración, el 20% restante, al margen de errores de medición, es la parte no explicada debida a otros factores, variables que no se han tenido en cuenta en este modelo. De la misma forma se puede afirmar que la recta de regresión de y sobre x permite reducir los errores de predicción en un 80%, siendo la cantidad de varianza compartida por ambas variables.

El software estadístico nos proporciona un resultado como el de la Tabla III.9.11.

Tabla III.9.11. Capacidad explicativa del modelo de regresión entre Valoración e Ingresos

R	R cuadrado	R cuadrado corregido	Error típico de la estimación
0,892	0,796	0,783	2,267

3.4.2. *Prueba de significación de la regresión*

La descomposición de la suma de cuadrados nos permite también aplicar la prueba estadística del análisis de varianza para validar el modelo y verificar si la varianza explicada es superior a la varianza no explicada o residual. Cuando disponemos tan sólo de una variable independiente el análisis de varianza que veremos a continuación nos permite contrastar la hipótesis nula según la cual el coeficiente de determinación poblacional es igual a cero: $\rho^2_{yx} = 0$, lo que es lo mismo que poner a prueba la hipótesis nula de si el coeficiente de regresión es o no significativamente distinto de cero.

El análisis de varianza se plantea en los términos que se especifican en la Tabla III.9.12. Como hemos comentado anteriormente, la variabilidad total observada en la variable dependiente se divide en dos partes, son dos fuentes de variación: la atribuible a la regresión y la residual. Las sumas de cuadrados correspondientes se dividen por los grados de libertad y se obtienen las varianzas o medias cuadráticas. Si los supuestos de la regresión se cumplen, el cociente entre la media cuadrática de la regresión y la media cuadrática se distribuye como un estadístico F de Fisher-Snedecor con $\nu_1=1$ y $\nu_2=n-2$ grados de libertad, y nos sirve para contrastar hasta qué punto el modelo de regresión se ajusta a los datos observados.

Tabla III.9.12. Análisis de varianza del modelo de regresión simple

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Varianzas o medias cuadráticas*	*F*
Regresión	$SCD_{Reg} = b^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$	1	$V_{Reg} = SCD_{reg} / 1$	
Residual o Error	$SCD_{Res} = SCD_{Tot} - SCD_{Reg}$	$n-2$	$V_{Res} = SCD_{Res} / n-2$	$F = \frac{V_{Reg}}{V_{Res}}$
Total	$SCD_{Tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	$n-1$	$V_{Tot} = SCD_{Tot} / n-1$	

La prueba estadística se plantea así:

1. Formulación de las hipótesis
   H_0: El modelo de regresión no se ajusta, $\rho^2_{yx} = 0$.
   H_A: El modelo de regresión se ajusta, $\rho^2_{yx} \neq 0$.
2. Cálculo del valor del estadístico muestral
   El estadístico F se calcula según la Tabla III.9.12: $F = \frac{V_{Reg}}{V_{Res}}$ y se distribuye como una F de Fisher-Snedecor.
3. Determinación de la significación
   Se estima la probabilidad asociada al estadístico a partir del valor concreto F_o del estadístico F.
4. Decisión sobre la significación del estadístico
   Tomando el valor de significación $\alpha=0,05$, con $\nu_1=1$ y $\nu_2=n-2$ grados de libertad, la decisión se formaliza de la siguiente manera:
   Si $Pr(F_o) \geq \alpha$ aceptamos la hipótesis nula, el modelo no ajusta.
   Si $Pr(F_o) < \alpha$ rechazamos la hipótesis nula, el modelo ajusta.

Los cálculos del análisis de la varianza a partir de los resultados que presenta el software estadístico se presentan en la Tabla III.9.13.

Se concluye la significatividad de la prueba estadística de la F, rechazamos la hipótesis nula según la cual el coeficiente de determinación es nulo, o lo que es lo mismo, que el coeficiente b de la regresión sea igual a cero. El modelo es válido, tiene un poder explicativo y predictivo que hemos evaluado con un coeficiente de determinación de $R^2_{yx} = 0,8$.

Tabla III.9.13. Análisis de varianza del modelo de regresión entre Valoración e Ingresos

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	F	Sig.
Regresión	320,037	1	320,037	62,263	0,000
Residual	82,241	16	5,140		
Total	402,278	17			

Las pruebas estadísticas que hemos presentado poseen una propiedad de interés: la raíz cuadrada del valor de F es el valor del estadístico t. Se verifica que el cuadrado del valor de t con k grados de libertad es un valor F con 1 y k grados de libertad: $t^2=F$. Por tanto, se concluye que podemos utilizar la t o la F para probar si el coeficiente b de la regresión es o no igual a cero.

Si consideramos ahora el caso de la relación entre los ingresos y la edad, esta regresión alcanza una capacidad explicativa del 73% según nos indica el valor del coeficiente de

determinación corregido o R^2 ajustado (Tabla III.9.14), un resultado significativo estadísticamente según el resultado de la prueba de la tabla ANOVA (Tabla III.9.15).

Tabla III.9.14. Capacidad explicativa del modelo de regresión entre Ingresos y Edad

R	R cuadrado	R cuadrado corregido	Error típico de la estimación
0,863	0,744	0,728	14,783

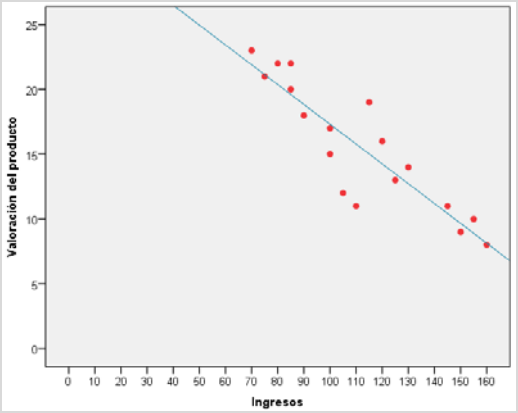
Tabla III.9.15. Análisis de varianza del modelo de regresión entre Ingresos y Edad

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	F	Sig.
Regresión	320,037	1	320,037	62,263	0,000
Residual	82,241	16	5,140		
Total	402,278	17			

En los ejemplos sobre el Índice de Desarrollo Humano que analizamos anteriormente, la regresión entre la esperanza de vida y los años de escolarización da lugar a un coeficiente de determinación significativo de 0,531. Cuando relacionamos la esperanza de vida con los datos originales de la renta per cápita el R^2 que se obtiene es de 0,370, un valor que no refleja la intensidad de la verdadera relación existente entre ambas variables. Cuando transformamos la variable de renta per cápita en su logaritmo decimal el valor del R^2 se convierte en 0,643, un valor mayor que en el caso de la escolarización.

► Ejercicio 2. Propuesto

Analizar la relación entre las variables Valoración del producto e Ingresos mediante un análisis de regresión a partir de los resultados siguientes.



Coeficientes[a]

Modelo		Coeficientes no estandarizados B	Coeficientes no estandarizados Error estándar	Coeficientes estandarizados Beta	t	Sig.	95,0% intervalo de confianza para B Límite inferior	95,0% intervalo de confianza para B Límite superior
1	(Constante)	27,202	2,076		13,101	0,000	22,800	31,603
	E Edad	-0,313	0,053	-0,827	-5,892	0,000	-0,426	-0,201

a. Variable dependiente: V Valoración del producto

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	0,827[a]	0,684	0,665	2,817

a. Predictores: (Constante), E Edad

ANOVA[a]

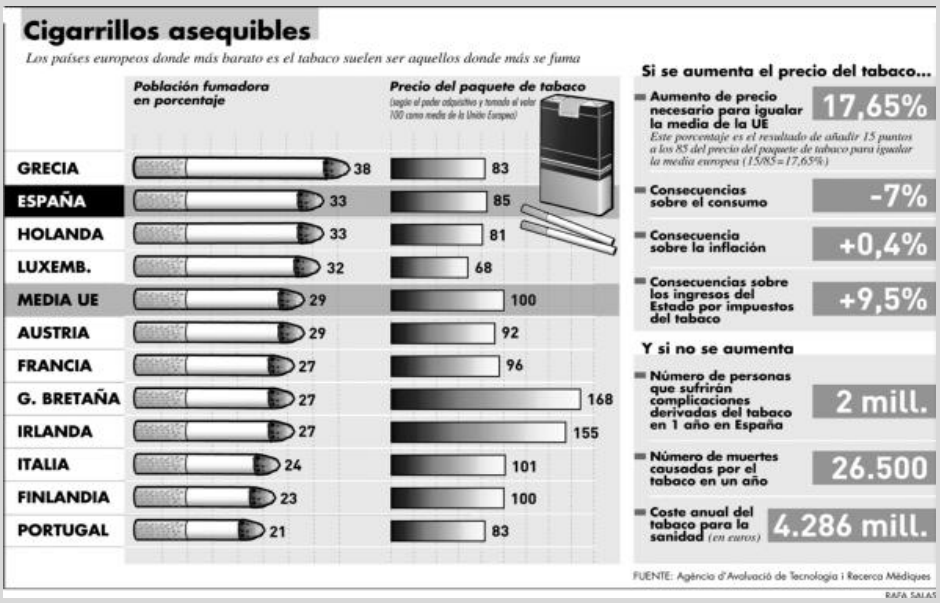
Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	275,353	1	275,353	34,711	0,000[b]
	Residuo	126,925	16	7,933		
	Total	402,278	17			

a. Variable dependiente: V Valoración del producto

b. Predictores: (Constante), E Edad

► Ejercicio 3. Propuesto

En un periódico apareció publicada la noticia siguiente con el titular: "*Los países europeos donde más barato es el tabaco suelen ser aquellos donde más se fuma*".



Con los datos que aparecen en la noticia y mediante un análisis de regresión analizar si podemos llegar a esa conclusión.

► **Ejercicio 4. Propuesto**
Imagina los datos de cuatro situaciones de relación entre dos variables cuantitativas (las notas obtenidas en un examen y las horas de estudio dedicadas), con 10 individuos, que correspondan a diferentes configuraciones de distribución de puntos en el plano en un gráfico de dispersión:
- Desde la completa indeterminación.
- Pasando por un nivel moderado de determinación.
- Por un alto nivel de correlación.
- Y hasta la determinación completa en la relación entre las variables.

Adicionalmente, calcula a mano y/o con la ayuda del software estadístico, la correlación, la recta de regresión y el R^2.

4. El análisis de regresión múltiple

El análisis de regresión múltiple se plantea cuando el modelo considera dos o más variables independientes. Ello nos permitirá tener en cuenta modelos más ricos y completos para dar cuenta de fenómenos sociales que se caracterizan por la multidimensionalidad y, en este caso, por la existencia de diversos factores explicativos que dan cuenta de un fenómeno. Al tener en cuenta múltiples variables independientes de forma simultánea, además, estamos controlando el efecto neto de cada una de ellas en relación a las demás y podremos establecer la la importancia diferenciada de cada una de ellas.

Así, en relación al análisis de regresión simple, la introducción de una o más variables independientes adicionales introduce diversas cuestiones de interés: ¿cómo medir de forma diferenciada el efecto de cada variable independiente?, ¿qué sucede si las variables independientes está altamente relacionadas entre sí, si existe interacción entre ellas?, cuando el número de variables es más numeroso ¿existe un procedimiento para seleccionar las variables del modelo?

Con la regresión múltiple no se trata de ajustar una recta de regresión sino un hiperplano de regresión. La ecuación de regresión múltiple tiene de forma genérica la expresión:

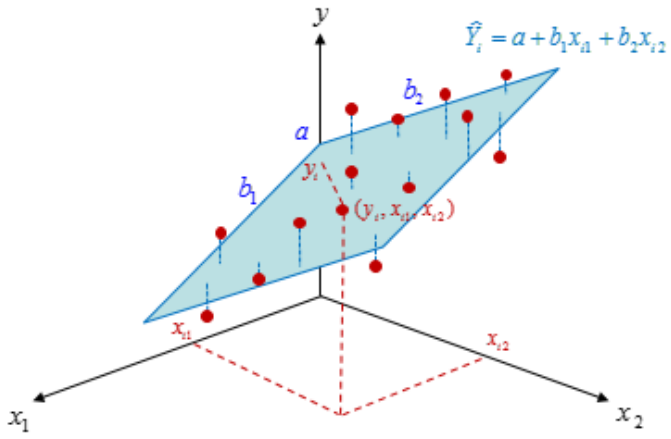
$$y_i = a + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} + e_i \qquad \textit{Ecuación 18}$$

donde los distintos coeficientes de regresión parciales b_j, con $j=1...p$, se interpretan como el efecto específico de cada variable independiente. De forma gráfica se corresponden con las distintas pendientes del hiperplano que se ajusta a la nube de puntos.

En el caso de dos variables independientes podemos representar gráficamente el plano de regresión que se ajusta (Gráfico III.9.18) y donde b_1 y b_2 determinan su inclinación, siendo la constante a el valor de y_i cuando x_{i1} y x_{i2} son cero, es decir, donde el plano de regresión corta el eje de la variable dependiente y que tan solo interpretamos en el caso en que tenga sentido considerar los valores cero de las variables independientes. Con un número mayor de variables independientes no será posible visualizar su

relación mediante un gráfico de dispersión. En cualquier caso, será posible obtener los valores de esa ecuación aplicando también el método de mínimos cuadrado ordinarios, aquél que minimiza la distancia vertical de cada punto al hiperplano, esto es, con dos variables, el plano que mejor se ajusta a la forma de la nube de puntos, el que consigue en definitiva obtener una combinación de las variables independientes que mejor explica la variabilidad de la variable dependiente.

Gráfico III.9.18. Modelo lineal de regresión múltiple



En la *Ecuación 17* la variable dependiente *y* se explica como resultado de una combinación lineal de las variables independientes: mediante la suma ponderada por el coeficiente de regresión que afecta (multiplica) a cada variable y establece la importancia relativa de cada una de ellas, es decir, expresan el cambio que experimenta la variable dependiente debido al cambio en una unidad de cada variable independiente cuando el resto permanecen constantes.

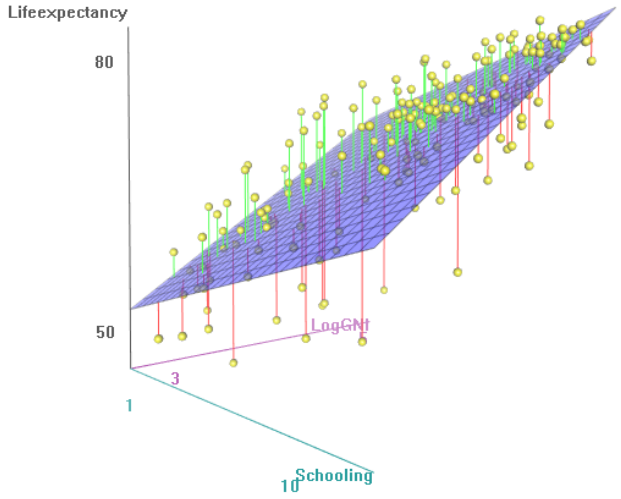
En la Tabla III.9.16 se presentan los coeficientes obtenidos en el ejemplo de la regresión de la esperanza de vida en función de la media de los años de escolarización y la renta per cápita, transformada en su logaritmo decimal.

Tabla III.9.16. Tabla de coeficientes de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
	Coeficiente	Error típico	Beta			Inferior	Superior
(Constante)	24,989	3,477		7,187	0,000	18,130	31,848
LogGNI	9,994	1,145	0,594	8,729	0,000	7,735	12,253
Schooling	0,764	0,197	0,264	3,885	0,000	0,376	1,151

Esta ecuación de regresión ajusta un plano a la nube de puntos donde cada coeficiente marca la pendiente de las dos dimensiones del plano (Gráfico III.9.19).

Gráfico III.9.19. Gráfico de dispersión de la Esperanza de vida según los años de escolarización y la renta per cápita transformada



En la Tabla III.9.16 se muestran, en primer lugar, los coeficientes no estandarizados. Podemos observar que el coeficiente no estandarizado de LogGNI (9,994) ha cambiado respecto del modelo de regresión simple (13,499). El efecto de una variable independiente puede ser muy distinto cuando se analiza en una regresión simple respecto a cuando se analiza junto a otra u otras variables en una regresión múltiple. Puede ser que se atenúe o aumente el efecto de una variable o incluso que desaparezca como resultado de la interacción o control con la(s) otra(s) variable(s) independiente(s). En nuestro ejemplo la diferencia se debe al efecto compartido con la variable de Schooling, dado que ambas variables están correlacionadas (su coeficiente de correlación es de 0,782). La reducción del coeficiente de regresión por la introducción de una segunda variable independiente permite diferenciar la parte específica del efecto de cada variable de la parte compartida con otras, facilitando así la regresión múltiple alcanzar una mayor validez del modelo. La diferencia entre 13,499 y 9,994 es el efecto compartido de la variable de renta con la de escolarización, siendo ahora 9,994 el efecto neto, es decir, el efecto sobre la esperanza de vida controlando por escolarización.

En el análisis de regresión con una variable independiente comentamos que la interpretación de los coeficientes la realizábamos con la unidad de medida propia de las variables. Cuando introducimos una segunda o sucesivas variables independientes en un modelo de regresión múltiple el efecto de cada factor se expresa en la escala propia, por lo que si son diferentes unidades los coeficientes no son comparables y no podemos utilizarlos como indicadores de la importancia relativa de cada variable independiente. Para disponer de una misma unidad de medida comparable se procede a estandarizar o tipificar las variables independientes con puntuaciones típicas *z*, es decir, a cada valor de la variable se le resta la media (se centra) y se divide por la desviación típica (se cambia la escala).

Cuando centramos una variable estamos considerando un nuevo centro, un nuevo valor 0 de origen. Este nuevo centro de referencia de las puntuaciones de una variable es la media de ésta. La variable original se transforma en una variable centrada cuando

se resta a cada puntuación, cada valor del individuo y el valor de la media. En el caso de las variables x e y:

$$y^c = y_i - \overline{y} \quad \text{y} \quad x^c = x_i - \overline{x}$$

En el caso particular de considerar el valor $x_i = \overline{x}$ y $y_i = \overline{y}$ entonces las medias de las variables centradas son cero: $\overline{x}^c = 0$ y $\overline{y}^c = 0$. El nuevo valor de a en la recta de regresión con las variables centradas será también nulo: $a = 0$.

Si además de centrar las variables dividimos por la desviación suprimiremos el efecto de las unidades de medida:

$$y^s = \frac{y_i - \overline{y}}{s_y} \quad \text{y} \quad x_{ij}^s = \frac{x_{ij} - \overline{x}_j}{s_{x_j}}$$

de nuevo $a = 0$ y la ecuación de regresión múltiple estandarizada queda:

$$\widehat{Y}_i^s = b_1^s \cdot x_{i1}^s + b_2^s \cdot x_{i2}^s + \cdots + b_p^s \cdot x_{ip}^s \qquad \textit{Ecuación 19}$$

donde b_j^s es el coeficiente de regresión con variables estandarizadas y tiene un significado equivalente a b, pero siendo ahora las unidades de cada variable las unidades de desviación. Por tanto, nos indica el número de desviaciones en que aumenta o disminuye y cuando lo hace x_j en una unidad de desviación, permaneciendo el resto constante, y permitiendo así realizar comparaciones relativas entre varias variables.

En el ejemplo sobre la esperanza de vida podemos constatar cómo el factor económico, la renta per cápita, tiene una importancia relativa mayor que el factor educativo, la escolarización, es decir, 0,594 es mayor que 0,264 (valor Beta en la Tabla III.9.16). De forma equivalente al coeficiente no estandarizado, si comparamos la regresión simple con la variable LogGNI con la regresión múltiple donde se añade la variable Schooling, observamos una reducción del efecto de la renta per cápita al controlar su importancia por la introducción de esta segunda variable independiente: el coeficiente estandarizado pasa de 0,802 a 0,594 por efecto de la correlación compartida con la variable Schooling.

La valoración estadística de los coeficientes de regresión se realiza, como en regresión simple, con una prueba donde se calcula el estadístico t que sigue en este caso una distribución de t de Student con $n-p-1$ grados de libertad. Para cada coeficiente el contraste se realiza mediante de la hipótesis nula de si el mismo es nulo en términos poblacionales:

H_0: $\beta_j = 0$

H_A: $\beta_j \neq 0$

El estadístico t se obtiene dividiendo el coeficiente obtenido en la muestra por su error típico de cada coeficiente: $t = \frac{b_j}{s_{b_j}}$, y aceptamos que es significativo si la probabilidad asociada al valor obtenido según la distribución de la t de student es inferior a 0,05. Su

intervalo de confianza se obtiene mediante la expresión: $b_j \pm t \cdot s_{b_j}$ [9]. En caso contrario, la no significación, implica eliminar la variable de nuestro modelo.

4.1. La colinealidad

Cuando dos o más variables explican la variabilidad de la variable dependiente se plantea la cuestión de hasta qué punto la relación entre las dos variables independientes determina o condiciona el modelo explicativo. En regresión múltiple no se modeliza necesariamente la interacción, se trata de analizar el efecto independiente de cada variable, si bien es posible incluir la interacción entre las variables independientes para reflejar mejor la aditividad del modelo mediante la inclusión del producto de las variables que interactúan. Por ello es deseable elegir factores explicativos que no estén altamente correlacionados entre sí pues ello afectará a nuestras estimaciones de los parámetros. Si la multicolinealidad fuera perfecta, y una de las variables pudiera expresarse como una una combinación lineal de las demás, entonces se podría alcanzar una solución de la ecuación de regresión.

La existencia de correlación entre las variables independientes se denomina colinealidad (entre dos variables) o multicolinealidad (entre más de dos). En general nos vamos a encontrar con variables independientes que siempre mantendrán un cierto grado de correlación, las variables originales difícilmente serán perfectamente independientes, por lo que la colinealidad será una cuestión de grado. Niveles altos de colinealidad serán perjudiciales para los resultados de la regresión, no en términos descriptivos pues la ecuación de regresión no se modifica, aunque sí en términos inferenciales, pues la presencia de colinealidad afecta a los errores típicos de los coeficientes de regresión haciéndolos mayores e inestables, lo que provoca que aumenten los intervalos de confianza y aumente por tanto la aceptación de coeficientes significativos que no lo serían.

La existencia de correlación entre las variables es un indicador claro del posible problema de la colinealidad. Pero hay que determinar su importancia. A través de diversos estadísticos se puede diagnosticar la colinealidad y el grado de tolerancia para cada una de las variables independientes. La tolerancia es un estadístico que determina en qué medida están relacionadas las variables independientes. La tolerancia de una variable es la proporción de su varianza no explicada por las otras variables independientes de la ecuación. Una variable con una tolerancia muy baja contribuye con poca información a un modelo, es colineal, y puede causar problemas de cálculo. Se calcula como 1 menos el coeficiente de determinación R_j^2 cuando se realiza la regresión de la variable independiente, es decir, cuando es pronosticada por las demás variables independientes incluidas en el análisis. Un valor próximo a 1 del índice de tolerancia para la variable independiente señala que la variable no está correlacionada con el resto de variables independientes. Por el contrario, si el valor se aproxima al

[9] Siendo $s_{b_j} = \dfrac{SCD_{Res} \big/ n-p-1}{(n-1)s_j^2(1-R_j^2)}$, donde s_j^2 es la varianza de la variable j y R_j^2 es el coeficiente de determinación obtenido en la regresión entre la variable independiente j y el resto de variables independientes.

cero sí que estaría relacionada con las demás. Se suele considerar el valor de 0,1 como la referencia a partir de la cual cabe considerar que con valores inferiores nos encontramos en una situación problemática de colinealidad.

De forma complementaria, el factor de inflación de la varianza (*FIV*) es el recíproco de la tolerancia: $1/1-R_j^2$. Cuando el *FIV* crece también lo hace la varianza del coeficiente de regresión, provocando que el estimador sea inestable. Valores de *FIV* grandes son un indicador de la existencia de colinealidad. La regla empírica de Kleinbaum (Kleinbaum y Kupper, 1978) señala que valores del *FIV* superiores a 10 implican problemas reales de colinealidad (0,1 por tanto para el valor de tolerancia).

En el ejemplo sobre el IDH se obtienen los valores de tolerancia y FIV de la Tabla III.9.17. Observamos en este caso que los valores se alejan de un valor inferior a 0,1 en el estadístico de tolerancia y 10 en el factor de inflación.

Tabla III.9.17. Tabla de coeficientes de regresión y diagnóstico de colinealidad de la Esperanza de vida según los años de escolarización y la renta per cápita

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Colinealidad	
	Coeficiente	Error típico	Beta			Tolerancia	FIV
(Constante)	24,989	3,477		7,187	0,000		
LogGNI	9,994	1,145	0,594	8,729	0,000	0,389	2,569
Schooling	0,764	0,197	0,264	3,885	0,000	0,389	2,569

Una alternativa para efectuar el diagnóstico de colinealidad es realizar un análisis de componentes principales entre las variables independientes[10]. Valores cercanos a cero de los valores propios o autovalores indican problemas de colinealidad. Y el número de valores propios nulos señala el número de variables que son combinación lineal. Para determinar cuándo un autovalor pequeño está suficientemente próximo a cero se utiliza su valor relativo respecto del mayor, estimando para cada autovalor el índice de condición como la raíz cuadrada del cociente entre el mayor de ellos y dicho autovalor. Y se llama número de condición al mayor de los índices de condición (Tabla III.9.18).

Tabla III.9.18. Diagnóstico de colinealidad. Índice de condición de la regresión de la Esperanza de vida según los años de escolarización y la renta per cápita

Dimensión	Autovalor	Índice de condición	Proporciones de varianza		
			(Constante)	LogGNI	Schooling
1	2,924	1,000	0,00	0,00	0,01
2	0,072	6,388	0,05	0,00	0,44
3	0,004	26,557	0,95	1,00	0,55

[10] Esta técnica se presenta en el Capítulo III.11 sobre el análisis factorial.

Para Belsey (1991) los índices de condición se pueden valorar de la forma siguiente:

Entre 5 y 10	Colinealidad débil
Entre 30 y 100	Colinealidad de moderada a fuerte
A partir de 1000	Colinealidad severa

En la regresión que analizamos el índice de condición de 26,557 nos indica un nivel moderado.

Determinada la presencia y el número de colinealidades, hay que saber entre qué variables se da. Si dos o más variables tienen una proporción de varianza alta en una componente indica que estas variables están implicadas en la colinealidad y, por tanto, la estimación de sus coeficientes está degradada por la misma. Belsey (1991) propone utilizar conjuntamente los índices de condición y la proporción de descomposición de varianza para realizar el diagnóstico de colinealidad, utilizando como umbral de proporción alta 0,5. Así los índices de condición altos (mayores que 30) indican el número de colinealidades y la magnitud de los mismos mide su importancia relativa. Si una componente tiene un índice de condición mayor que 30 y dos o más variables tienen una proporción de varianza alta en el mismo, estas variables son colineales.

¿Cómo solucionar un problema de colinealidad? Lo más sencillo, y drástico, sería eliminar la variable que manifiesta la colinealidad, pero ello evidentemente puede entrar en contradicción con una selección de esta variable basada en criterios teóricos que justificaría su relevancia e inclusión en el modelo de regresión. De forma alternativa se puede proceder a combinar las variables colineales en una sola o realizar esta síntesis mediante un análisis de componentes principales. Alternativamente, en un procedimiento de introducción por pasos de las variables en el modelo de regresión, cuando la incorporación de una nueva variable no reporta un incremento de la varianza explicada y no aporta un elemento explicativo propio de la variabilidad de la variable dependiente.

4.2. Bondad de ajuste del modelo

Una vez establecida la ecuación de regresión se trata de determinar la bondad de ajuste del modelo, de forma similar a como la vimos en regresión simple. De hecho, el establecimiento de la ecuación de regresión y los análisis que comporta supone observar recurrentemente los efectos que tiene en términos del coeficiente de determinación, como vimos también regresión simple. En regresión múltiple, para cuantificar la relación entre la variable dependiente con las independientes en términos de capacidad explicativa se emplea el coeficiente de correlación múltiple. Con dos variables explicativas lo denotamos como r_{y,x_1,x_2} y, en general, con p variables como $r_{y,x_1,\ldots,x_p}$, y expresa el grado de relación entre la variable dependiente y la combinación de las variables independientes, es decir, entre la y observada y la $\hat{y}$ pronosticada por la ecuación de regresión. El cuadrado del coeficiente de correlación múltiple da lugar al coeficiente de determinación $R^2_{y,x_1,\ldots,x_p}$, de la misma forma que en regresión simple, y se interpreta como la proporción de varianza explicada por el modelo. Para ajustar

esta estimación de la bondad de ajuste, que tiende a elevarse por efecto del número de variables y del número de casos, se calcula el $R^2_{y,x_1,x_2,\ldots}$ corregido o ajustado.

La significación estadística de la relación entre la variable dependiente y las independientes se realiza mediante el contraste de la hipótesis nula de si el coeficiente de determinación poblacional es nulo o no:

H_0: $\rho^2_{y,x_1,x_2,\ldots} = 0$

H_A: $\rho^2_{y,x_1,x_2,\ldots} \neq 0$

Este contraste implica igualmente plantearse si los p diferentes coeficientes de regresión son, simultáneamente, iguales a cero. Por lo tanto, valorar globalmente la relevancia del modelo con las variables independientes consideradas y, si la prueba es significativa, ello implica que por lo menos uno de los coeficientes de regresión es diferente de cero y contribuye a mejorar el ajuste en relación a no considerar ninguna variable. Como en la regresión simple se obtiene un estadístico F de Fisher-Snedecor ahora con $\nu_1=p$ y $\nu_2=n-p-1$ grados de libertad resultado de la descomposición de la varianza (Tabla III.9.19).

Tabla III.9.19. Análisis de varianza del modelo de regresión múltiple

Fuente de variación	*Suma de cuadrados*	*Grados de libertad*	*Varianzas o medias cuadráticas*	*F*
Regresión	SCD_{Reg}	p	$V_{Reg} = \frac{SCD_{reg}}{p}$	
Residual o Error	$SCD_{Res} = SCD_{Tot} - SCD_{Reg}$	$n-p-1$	$V_{Res} = \frac{SCD_{Res}}{n-p-1}$	$F = \frac{V_{Reg}}{V_{Res}}$
Total	$SCD_{Tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	$V_{Tot} = \frac{SCD_{Tot}}{n-1}$	

La prueba estadística se plantea de la misma forma que vimos en el caso de la regresión simple. En este caso, si concluimos la significación del modelo, debemos precisar qué coeficientes de regresión en concreto son significativos. Su valoración estadística se realiza de nuevo con una prueba donde se calcula el estadístico t que sigue una distribución de t de Student con $n-p-1$ grados de libertad.

En el análisis de la esperanza de vida de los países del mundo se obtienen los datos de la Tabla III.9.20 y de la Tabla III.9.21. En relación al modelo de regresión simple donde se consideraba solamente la variable de renta per cápita (LogGNI), los resultados que se obtienen permiten concluir que la capacidad explicativa mejora ligeramente como resultado de incorporar la variable de escolarización (Schooling) para alcanzar un coeficiente de determinación significativo de 0,668 (0,665 corregido por el número de variables y de casos) frente al 0,643 anterior.

Tabla III.9.20. Capacidad explicativa del modelo de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita

R	R cuadrado	R cuadrado corregido	Error típico de la estimación
0,818	0,668	0,665	5,11570

Tabla III.9.21. Análisis de varianza del modelo de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita

Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	F	Sig.
Regresión	9.702,693	2	4.851,346	185,376	0,000
Residual	4.815,346	184	26,170		
Total	14.518,038	186			

4.3. La importancia relativa: correlación parcial y semiparcial

La importancia relativa de las variables puede establecerse, además de por el valor del coeficiente de regresión estandarizado como vimos anteriormente, por la contribución de cada variable al ajuste, es decir, en la explicación de la varianza de la variable dependiente. Para ello se emplea el cuadrado del coeficiente de correlación semiparcial. La correlación semiparcial es un medida diferente de la correlación parcial. Veamos estos conceptos.

La correlación parcial entre dos variables x_j y $x_{j'}$, en el contexto de p variables, donde $p*$ serían todas las variables restantes que no son x_j y $x_{j'}$, expresada como $r_{x_j,x_{j'}|p^*}$, mide el grado de correlación neta entre dos variables, es decir, la correlación que queda cuando se elimina, se controla, el efecto de las demás variables en estas dos. Su fórmula, para el caso de tres variables, y considerando la correlación parcial entre las dos primeras variables controlando por la tercera, es la siguiente[11]:

$$r_{x_1,x_2|x_3} = \frac{r_{x_1,x_2} - r_{x_1,x_3} r_{x_2,x_3}}{\sqrt{(1-r^2_{x_1,x_3})(1-r^2_{x_2,x_3})}} \qquad \textit{Ecuación 20}$$

Esta correlación se denomina de primer orden pues se controla el efecto de una variable sola, x_3, sobre x_1 y x_2. Si se toman cuatro variables habría que considerar correlaciones de segundo orden, se calcula la correlación entre x_1 y x_2 eliminando el

[11] La significación del coeficiente de correlación parcial, controlando por p variables, se establece a partir del estadístico: $t = \frac{r_{x_1,x_2|p} - \sqrt{n-p-2}}{\sqrt{(1-r_{x_1,x_2|p})}}$ que sigue una distribución de una t de student con $n-p-2$ grados de libertad.

efecto de x_3 y x_4. Y así sucesivamente con más variables. Cuando no se controla el efecto de otras variables entonces la correlación se denomina de orden cero.

Para obtener el coeficiente de correlación parcial es preciso realizar la regresión de cada una de las variables independientes que se correlacionan, x_j y $x_{j'}$, con todas las otras. De estas dos regresiones se obtienen unos residuos cuya correlación es la medida de correlación parcial, que como veremos es la medida relativa de la correlación parcial. Esta medida nos indica qué parte de la varianza total es explicada por la introducción de cada nueva variable, es decir, en relación a la varianza que está o queda por explicar, qué reducción del error se consigue por la contribución de la variable, sin tener en cuenta el efecto de la(s) variable(s) que está(n) incluida(s) en la ecuación. Esta es una información de interés en un análisis de regresión para evaluar la selección de las variables en un procedimiento de selección por pasos.

Por su parte, la correlación semiparcial, expresada como $r_{y,x_j(x_{j'}|p)}$, en un conjunto de p variables, mide la correlación lineal entre dos variables, x_j y $x_{j'}$, cuando se elimina el efecto de las demás variables, solamente sobre una de las dos, y se obtiene correlacionando la primera variable x_j con los residuos derivados de la regresión de $x_{j'}$ con las demás variables. Mediante la correlación semiparcial podemos evaluar así la contribución neta de cada variable independiente para explicar la dependiente.

Por ejemplo, si consideramos la relación entre la valoración de un producto (variable 1, dependiente), los ingresos (variable 2, independiente) y la edad (variable 3, independiente), la correlación parcial entre valoración e ingresos, $r_{1,2|3}$, mide su grado de relación lineal cuando se elimina el efecto sobre ambas variables de la edad. La correlación semiparcial entre valoración e ingresos $r_{1(2|3)}$ mide la relación lineal entre ambas cuando se elimina el efecto de la edad sobre los ingresos.

A partir de la correlación semiparcial y del coeficiente de determinación, y considerando una variable dependiente y y dos independientes x_1 y x_2, se llega a la siguiente relación fundamental: el coeficiente de determinación R^2_{y,x_1,x_2}, es decir, la capacidad explicativa conjunta del modelo de regresión, se puede descomponer en la suma de dos partes:

$$R^2_{y,x_1,x_2} = R^2_{y,x_1} + R^2_{y,x_2(x_1)} \qquad \textit{Ecuación 21}$$

a) La contribución de la primera variable x_1, que se expresa en el cuadrado del coeficiente de correlación simple entre y y x_1, es decir, R^2_{y,x_1}, y
b) La contribución de segunda variable x_2 cuando se extrae el efecto de x_1 sobre y, que se expresa en el cuadrado de la correlación semiparcial de la segunda variable x_2, $R^2_{y,x_2(x_1)}$, esto es, la correlación entre y y x_2 cuando se elimina el efecto de x_1 sobre y. Este valor se interpreta como el efecto adicional de x_2 cuando se añade a

la ecuación de regresión donde está x_1, por tanto, la varianza explicada aportada por x_2.

De manera genérica, considerando sucesivas variables independientes, la descomposición de la varianza total dada por la aportación neta de cada nueva variable se expresa como:

$$R^2_{y,x_1,\ldots,x_p} = R^2_{y,x_1} + R^2_{y,x_2(x_1)} + R^2_{y,x_3(x_1,x_2)} + \cdots + R^2_{y,x_p(x_1,x_2,\cdots,x_{p-1})} \qquad \textit{Ecuación 22}$$

Con los datos de nuestro ejemplo del IDH obtenemos la información de las correlaciones tal y como se presenta en la Tabla III.9.22.

Tabla III.9.22. Tabla de coeficientes de regresión y coeficientes de correlación simple, parcial y semiparcial de la Esperanza de vida según los años de escolarización y la renta per cápita

Modelo	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
	Coeficiente	Error típico	Beta			Orden cero	Parcial	Parte
(Constante)	24,989	3,477		7,187	0,000			
LogGNI	9,994	1,145	0,594	8,729	0,000	0,801	0,541	0,371
Schooling	0,764	0,197	0,264	3,885	0,000	0,729	0,275	0,165

La correlación simple o de orden cero entre la esperanza de vida cada variable independiente aparece en la primera columna de correlaciones. El valor de 0,801 de la variable LogGNI indica que si consideramos solamente esta variable explicaríamos el 64,2% de la varianza de la variable dependiente Lifeexpectancy, valor que se obtiene elevando al cuadrado el coeficiente de correlación. Este valor es del 53,1% en el caso de la variable Schooling. Estos resultados los conocíamos ya de las regresiones simples que realizamos anteriormente. Y conjuntamente, ambas variables explican el 66,8% de la varianza como hemos visto con el coeficiente de determinación que se obtiene en la regresión múltiple. Por tanto, a partir de la regresión simple con LogGNI, cuando se añade Schooling se aumenta la capacidad explicativa del modelo en un 2,7%. Del mismo modo, a partir de la regresión simple con Schooling, cuando se añade LogGNI se aumenta la capacidad explicativa del modelo en un 13,7%. Como ambas variables están relacionadas entre sí, su correlación es de 0,782, ambas explican una parte que también es explicada por la otra variable, por lo que teniéndolas en cuenta a las dos simultáneamente la varianza explicada es solamente del 66,8%. Estos resultados se expresan en términos de la correlación semiparcial.

La correlación semiparcial aparece en la última columna con el nombre Parte, que si elevamos al cuadrado nos informa sobre la mejora que supone la variable en cuestión al incorporarse a la ecuación de regresión donde está incluida la otra variable. El cuadrado del coeficiente de coeficiente de correlación semiparcial entre y y x_2 cuando se extrae el efecto de x_1, en nuestro caso, en particular, entre Lifeexpectancy y Schooling, es la diferencia entre el coeficiente de determinación que resulta de incluir las dos variables en la ecuación $R^2_{y,x_1,x_2} = 0,668$, y el coeficiente de determinación de

la variable LogGNI solamente: $R^2_{y,x_1} = 0,641$, una diferencia pues de 0,027, es decir, de un 2,7%. Esto es, con la ecuación de regresión donde está la variable LogGNI, la incorporación de la variable Schooling contribuye a mejorar el ajuste del modelo en un 2,7%, el resultado de elevar al cuadrado el valor de la correlación semiparcial 0,165 y multiplicar por cien.

De la misma forma, sobre la ecuación de regresión donde está la variable Schooling, la incorporación de la variable LogGNI contribuye a mejorar el ajuste del modelo en un 13,7%, resultado de elevar al cuadrado 0,371 y multiplicar por cien. La suma de estos valores no alcanza el valor de coeficiente de determinación R^2, de un 66,8% de varianza explicada, pues entre ellas se da un importante grado de correlación y de una superposición, por tanto, en términos de la varianza explicada de la variable dependiente.

La correlación parcial nos expresa en términos relativos las contribuciones que acabamos de comentar. Es decir, nos informa de cómo contribuye cada variable en la explicación de la dependiente en relación a la parte no explicada por las variables que ya están en la ecuación. Tomando el coeficiente de determinación que resulta de incluir las dos variables en la ecuación $R^2_{y,x_1,x_2} = 0,668$, si le restamos el coeficiente de determinación de la variable LogGNI: $R^2_{y,x_1} = 0,641$, la diferencia es de 0,027, un 2,7%, que dividida por la varianza no explicada $1 - R^2_{y,x_1} = 1 - 0,641 = 0,359$ da como resultado la siguiente proporción:

$$R^2_{y,x_2|x_1} = \frac{R^2_{y,x_1,x_2} - R^2_{y,x_1}}{1 - R^2_{y,x_1}} = \frac{0,668 - 0,641}{1 - 0,641} = \frac{0,027}{0,359} = 0,076$$

Es decir, que teniendo en cuenta que la variable LogGNI explica 64,1%, queda por explicar un 35,9%, y del total de esta parte sin explicar un 7,6% viene explicado por la variable Schooling. Este valor es el cuadrado del coeficiente de correlación parcial que aparece en la tabla, el cuadrado de 0,275.

De forma equivalente, teniendo en cuenta que la variable Schooling explica 53,1% de Lifeexpectancy, queda por explicar un 46,9%, y del total de esta parte sin explicar un 29,3% viene explicado por la variable LogGNI, valor que corresponde al cuadrado del coeficiente de correlación parcial de esta variable.

$$R^2_{y,x_1|x_2} = \frac{R^2_{y,x_1,x_2} - R^2_{y,x_2}}{1 - R^2_{y,x_2}} = \frac{0,668 - 0,531}{1 - 0,531} = \frac{0,137}{0,469} = 0,293$$

De forma general, en el contexto de p variables independientes, donde $p*$ serían todas las variables que no son x_j, esta fórmula se puede expresar como:

$$R^2_{y,x_j|p^*} = \frac{R^2_{y,p} - R^2_{y,p^*}}{1 - R^2_{y,p^*}} \qquad \textit{Ecuación 23}$$

4.4. El análisis de regresión por pasos

Cuando se realiza un análisis de regresión, a partir de un modelo que define el conjunto de variables independientes que intervendrán en la explicación de la variable dependiente, es interesante proceder a un análisis secuencial y progresivo de introducción por etapas de las variables independientes para observar las implicaciones que supone la incorporación de cada nueva variable y establecer también un criterio de jerarquía en la selección de las variables.

Sobre un conjunto de variables que son pertinentes para dar cuenta de la variable dependiente se busca encontrar una selección de variables que expliquen la mayor parte de la varianza explicada. Sabiendo que la incorporación de nuevas variables siempre puede ayudar a mejorar el coeficiente de determinación no se trata de realizar un ejercicio empiricista al margen de la necesaria reflexión teórico-conceptual de nuestro modelo[12], se debe procurar equilibrar la selección de variables que contribuyan a mejorar la bondad de ajuste del modelo con un principo de parsimonia que exige la práctica del conocimiento científico. Por tanto, se trata de seguir una dinámica de selección que comporte una contribución sustantiva y estadísticamente signficativa.

Para ilustrar esta manera de proceder emplearemos el sencillo análisis de los datos del IDH con el método de inclusión de variables Hacia adelante. Con este método se selecciona en primer lugar la variable que tiene un mayor coeficiente de correlación con la variable dependiente. A continuación se calculan los coeficientes de correlación parciales de las variables no incluidas en la ecuación y la variable dependiente, excluyendo, por tanto, el efecto de la variable ya incluida en la ecuación. La variable con mayor coeficiente de correlación parcial se elige para ser incluida en la ecuación, y se reitera el proceso con el resto de variables independientes. En ese proceso se emplea el estadístico F para evaluar la signficación del cambio en el R^2[13].

Con las variables LogGNI y Schooling explicando Lifeexpectancy, el procedimiento de selección hacia adelante elige en primer lugar la variable LogGNI en una primera etapa configurando el Modelo 1 con una sola variables independiente, la que tiene una mayor correlación con Lifeexpectancy. En la siguiente etapa elige la variable Schooling, configurando el Modelo 2 con ambas variables. En la Tabla III.9.23 vemos estos dos modelos con los valores de R^2 asociados a cada uno. Junto a estos valores se incluyen los llamados estadísticos del cambio. El Cambio en R cuadrado muestra cómo aumenta el coeficiente de determinación en cada etapa: por la introducción de la primera variable LogGNI (0,641) y la segunda variable Schooling (0,027). Estos valores son, como vimos en el apartado anterior, los valores del cuadrado del coeficiente de correlación semiparcial. El estadístico Cambio en F junto con su significación contrasta la hipótesis nula de que el cambio en el R^2 es cero, ello implica que solamente se incorporarán al modelo variables que supongan una contribución significativa ($sig.<0,05$) al incremento del R^2.

[12] El coeficiente de determinación R^2 siempre aumenta como resultado de la introducción de nuevas variables, incluso puede llegar a ser 1 si el número de variables es igual al número de casos.
[13] Alternativamente se pueden emplear métodos como el de introducción Hacia atrás o Por pasos en SPSS.

Tabla III.9.23. Capacidad explicativa del modelo de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita. Procedimiento de introducción hacia adelante

Modelo	R	R cuadrado	R cuadrado corregido	Error típico de la estimación	Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. cambio en F
1	0,801	0,641	0,639	5,30698	0,641	330,482	1	185	0,000
2	0,818	0,668	0,665	5,11570	0,027	15,093	1	184	0,000

En el procedimiento de introducción secuencial se obtiene la evaluación de la significación del coeficiente de determinación de cada modelo (Tabla III.9.24).

Tabla III.9.24. Análisis de varianza del modelo de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita. Procedimiento de introducción hacia adelante

Modelo	Fuente de variación	Suma de cuadrados	Grados de libertad	Media cuadrática	F	Sig.
1	Regresión	9.307,693	1	9.307,693	330,482	0,000
	Residual	5.210,345	185	28,164		
	Total	14.518,038	186			
2	Regresión	9.702,693	2	4.851,346	185,376	0,000
	Residual	4.815,346	184	26,170		
	Total	14.518,038	186			

También se obtiene una tabla de coeficientes de la ecuación de regresión para cada modelo (Tabla III.9.25). Finalmente se evalúa la exclusión de las variables del modelo (Tabla III.9.26): si la significación asociada a cada variable bajo la hipótesis nula de que los coeficientes de regresión estadarizados (En beta) son cero es mayor o igual a 0,05 se excluye, en caso contrario se mantienen.

Tabla III.9.25. Tabla de coeficientes de regresión de la Esperanza de vida según los años de escolarización y la renta per cápita. Procedimiento de introducción hacia adelante

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza (95%)	
		Coeficiente	Error típico	Beta			Inferior	Superior
1	(Constante)	17,263	2,958		5,835	0,000	11,426	23,100
	LogGNI	13,471	0,741	0,801	18,179	0,000	12,009	14,932
2	(Constante)	24,989	3,477		7,187	0,000	18,130	31,848
	LogGNI	9,994	1,145	0,594	8,729	0,000	7,735	12,253
	Schooling	0,764	0,197	0,264	3,885	0,000	0,376	1,151

Tabla III.9.26. Variables excluidas en la regresión de la Esperanza de vida según los años de escolarización y la renta per cápita. Procedimiento de introducción hacia adelante

Modelo		En beta	t	Sig.	Correlación parcial	Estadísticas de colinealidad		
						Tolerancia	VIF	Tolerancia mínima
1	Schooling	0,264	3,885	0,000	0,275	0,389	2,569	0,389

► **Ejercicio 5. Propuesto**
Con los datos del ejemplo de valoración del producto en función de los ingresos y la edad visto en el apartado de regresión simple realiza un análisis de regresión múltiple e interpreta los resultados: ecuación de regresión, bondad de ajuste del modelo, colinealidad e importancia relativa.

5. El análisis de regresión con variables cualitativas

En un análisis de regresión se pueden emplear variables cualitativas siempre que se traten de forma adecuada. Para ello debemos utilizar variables cualitativas dicotómicas codificadas con 0 y 1 (llamadas también binarias, ficticias o dummies).

Si la variable cualitativa tiene dos valores éstos se codifican sencillamente en 0 y 1. Si la variable tiene más de dos valores entonces creamos tantas variables dicotómicas como valores tenga, y las utilizamos todas menos una. Por ejemplo la variable Nivel de estudios con tres valores: Primarios, Secundarios y Superiores se codificaría así:

	Primarios	*Secundarios*	*Superiores*
	x_1	x_2	x_3
Primarios	1	0	0
Secundarios	0	1	0
Superiores	0	0	0

La última categoría actúa de base de referencia y no se considera en el modelo, pues daría combinación lineal perfecta (colinealidad) y no se podría calcular el modelo de regresión.

Si realizamos un análisis de regresión simple con una variable independiente cualitativa codificada de esta forma obtendríamos los mismos resultados del contraste de dos medias (prueba de la t) o en una ANOVA como se muestra en la Tabla III.9.27. En la relación entre el índice socioeconómico y el sexo (donde las mujeres son codificadas

con 1 y los varones con 0) el valor de la constante en la ecuación de regresión se corresponde con la media de los varones en la variable dependiente (49,934), es decir, cuando la variable independiente toma el valor cero. El valor del coeficiente de regresión −4,857 es el valor de la variable dependiente cuando la independiente varía en una unidad, cuando pasa de 0 a 1, es decir, cuando se pasa de ser varón a ser mujer, por tanto, la diferencia entre la media de varones y mujeres, una reducción en la media del índice socieconómico de −4,857.

Tabla III.9.27. Resultados de una ANOVA y de una Regresión con variables cualitativas

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
									95% Intervalo de confianza para la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
indsocec Índice socioeconómico del encuestado	Se han asumido varianzas iguales	3,935	,047	4,877	1417	,000	4,8567	,9957	2,9034	6,8100
	No se han asumido varianzas iguales			4,831	1281,46	,000	4,8567	1,0054	2,8843	6,8290

Estadísticos de grupo

	sexo Sexo del entrevistado	N	Media	Desviación típ.	Error típ. de la media
indsocec Índice socioeconómico del encuestado	0 Hombre	622	49,934	19,4156	,7785
	1 Mujer	797	45,077	17,9592	,6361

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	49,934	,746		66,913	,000
	sexo Sexo del entrevistado	-4,857	,996	-,128	-4,877	,000

a. Variable dependiente: indsocec Índice socioeconómico del encuestado

0 Hombre $\hat{Y}_i = 49{,}934 - 4{,}857 \times 0 = 49{,}934$

1 Mujer $\hat{Y}_i = 49{,}934 - 4{,}857 \times 1 = 45{,}077$

ANOVA

indsocec Índice socioeconómico del encuestado

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	8240,303	1	8240,303	23,789	,000
Intra-grupos	490833,811	1417	346,389		
Total	499074,114	1418			

Medidas de asociación

	Eta	Eta cuadrado
indsocec Índice socioeconómico del encuestado * sexo Sexo del entrevistado	,128	,017

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	8240,303	1	8240,303	23,789	,000[a]
	Residual	490833,811	1417	346,389		
	Total	499074,114	1418			

a. Variables predictoras: (Constante), sexo Sexo del entrevistado
b. Variable dependiente: indsocec Índice socioeconómico del encuestado

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,128[a]	,017	,016	18,6115

a. Variables predictoras: (Constante), sexo Sexo del entrevistado

Para ilustrar el uso de la regresión múltiple con variables cualitativas se puede consultar el análisis realizado por Fachelli y Planas (2016) sobre equidad en el acceso y en la inserción profesional de los graduados universitarios[14].

[14] El informe "Equitat en l'accés i en la inserció professional dels graduats i graduades universitaris" se puede obtener en esta página: http://www.aqu.cat/aqu/publicacions/insercio_laboral.html y el capítulo concreto al que hacemos referencia en este apartado es: Fachelli, S. y Planas J. (2016) "Capítulo 1: Evolución de la inserción profesional de los universitarios: de la expansión a la crisis duradera", Barcelona: AQU, y se puede encontrar aquí http://www.aqu.cat/doc/doc_10339347_1.pdf

6. Análisis adicionales en regresión

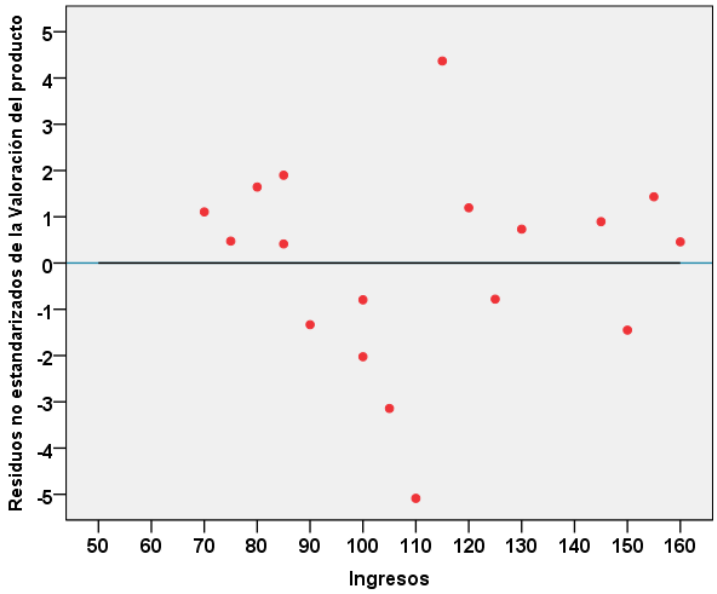
Los análisis que hemos reproducido hasta aquí, tanto de regresión simple como de regresión múltiple, se deben ampliar con otros adicionales destinados a verificar las condiciones de aplicación del modelo de regresión lineal y a estudiar la presencia de casos atípico e influyentes.

6.1. Linealidad

La linealidad es una condición básica del modelo clásico de regresión. En regresión simple se puede comprobar con el gráfico de dispersión y el cálculo del coeficiente de correlación como vimos anteriormente. Adicionalmente se puede representar un gráfico de dispersión con los residuos en el eje vertical y la variable independiente en el eje horizontal[15]. Si la pauta de relación entre *y* y *x* es de tipo lineal la nube de puntos que se configura no debe mostrar un patrón observable sino que los puntos deben aparecer distribuidos aleatoriamente a lo largo del espacio del gráfico en torno al valor cero de los residuos.

En el Gráfico III.9.20 se representa este gráfico para el caso de la regresión de la Valoración sobre los Ingresos y donde se puede comprobar la disposición aleatoria de los casos alrededor de la línea del cero.

Gráfico III.9.20. Diagrama de dispersión entre los residuos no estandarizados de la Valoración y las variables independientes Ingresos y Edad



En el caso de la regresión múltiple se pueden utilizar los gráficos de dispersión con cada variable independiente, si bien solamente muestran relaciones individuales y no la del conjunto. Se pueden obtener gráficos tridimensionales con dos variables

[15] También se pueden representar los valores pronosticados por la ecuación de regresión en el eje horizontal en vez de la variables independiente, y estos valores pueden estandarizarse también.

independientes, pero con más de dos variables no es posible este tipo de representación. Alternativamente se pueden emplear diagramas de dispersión parciales donde se consideran los residuos de los valores predichos de la variable dependiente a partir del conjunto de las variables independientes excepto una dada x_j, junto con los residuos que resultan de pronosticar la variable x_j en cuestión con el resto de variables independientes. Así se obtiene una representación donde se muestra la correlación parcial entre y y x_j. La pendiente de la recta de regresión que se ajusta se corresponde con el coeficiente de regresión de dicha variable, b_j, y de existir una relación lineal debe dibujarse una nube de puntos con una disposición alineada, en caso contrario, se concluye la ausencia de linealidad.

Cuando no se da una situación de relación de tipo lineal los residuos son considerables y el ajuste de la recta de regresión es deficiente. Por eso hemos comentado que la inspección visual, cuando es posible, es de gran ayuda. Pero disponer de una configuración de puntos con una forma no lineal no significa necesariamente que no podamos hacer un análisis de regresión. Si tenemos, por ejemplo, una configuración de la nube de puntos que evidencia una relación exponencial se puede ajustar una recta de regresión siempre que la variable independiente se transforme a través de una función matemática, por ejemplo calculando el logaritmo. En la Tabla III.9.28 se recogen distintas formas de curvas con la ecuación de la función matemática.

Tabla III.9.28. Familia de curvas y ecuación

Tipo de curva	Ecuación
Lineal	$y = a + bx$
Logarítmica	$y = a + b\log(x)$
Parabólica	$y = a + bx + cx^2$
Polinómica de grado n	$y = a_o + a_1 x + \cdots + a_n x^n$
Exponencial	$y = a \cdot e^{bx}$
Potencial	$y = a \cdot x^b$
Inversa	$y = a + \frac{b}{x}$
Curva S	$y = e^{a+\frac{b}{x}}$

6.2. Independencia

En un modelo de regresión lineal se establece la hipótesis de que los errores, la diferencia entre el valor observado de la variable dependiente y el pronosticado, se distribuyen con media cero y son independientes entre sí, es decir, que no están autocorrelacionados, lo que implicaría el seguimiento de patrones de comportamientos crecientes o decrecientes. Se trata de una situación que puede darse en el caso de datos longitudinales donde se dispone de una serie de datos secuenciales que pueden expresar comportamientos tendenciales en los residuos como cuando tenemos series temporales. En esos casos las varianzas de los coeficientes de regresión son menores

aumentando las situaciones donde se concluye la significación de los coeficientes. Si los datos son transversales, recogidos en un solo momento en el tiempo, no se plantea este problema.

Para detectar la presencia de la autocorrelación se pueden representar en un gráfico de dispersión los casos en el orden secuencial que tuvieran, en el eje horizontal, y los residuos en el eje vertical. Si no se detecta ningún patrón de comportamiento sino que los puntos se distribuyen de forma aleatoria concluimos la ausencia de autocorrelación. Si se da autocorrelación positiva los puntos del gráfico se alinean en tramos crecientes o decrecientes. Si se da autocorrelación negativa los residuos positivos y negativo se suceden de forma alternativa.

El grado de autocorrelación se puede evaluar mediante el estadístico de Durbin-Watson:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \qquad \textit{Ecuación 24}$$

que toma valores entre 0 y 4. Cuando los valores son próximos a 2 y, en general, comprendios entre 1,5 y 2,5, se considera que los residuos son independientes. Los inferiores indicarían autocorrelación positiva y los superiores autocorrelación negativa.

6.3. Normalidad

El supuesto de normalidad es una condición relevante en el ejercicio de inferencia del modelo de regresión y para poder extraer conclusiones adecuadas sobre la significación de los coeficientes de regresión a la hora de estimar los valores de la variables dependiente. Los errores que cometemos en ese ejercicio de estimación a partir de cada valor de las variables independientes se exige que sigan una distribución normal. El contraste sobre la forma de la distribución normal, así como el histograma de los residuos y el gráfico de probabilidad normal (gráfico Q-Q) permiten detectar el posible incumplimiento del supuesto de normalidad.

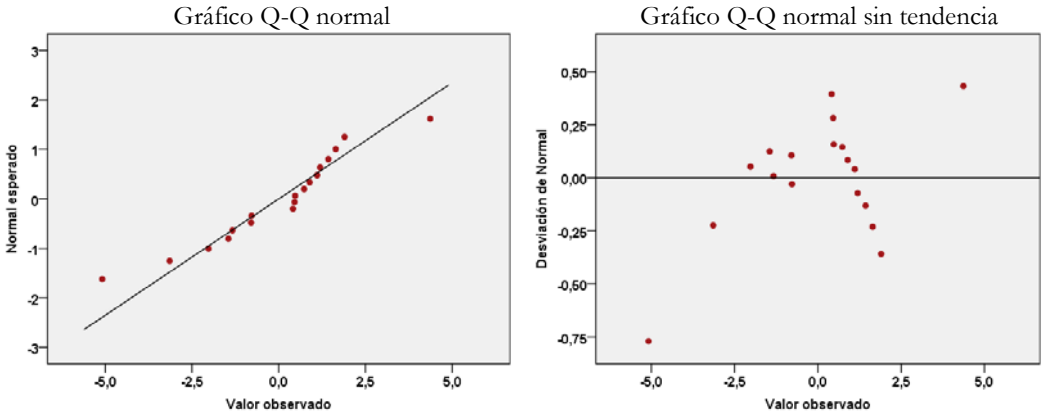
Tabla III.9.29. Prueba de normalidad de los residuos de la regresión. Valoración del producto según Edad e Ingresos

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Residuos	0,188	18	0,093	0,953	18	0,469

a. Corrección de significación de Lilliefors

En este caso la pruebas estadísticas arrojan una probabilidad superior a 0,05 que permiten mantener la hipótesis nula de normalidad. Los gráficos muestran esta conclusión por el alineamiento a lo largo de la recta en el gráfico Q-Q normal así como una distribución aleatoria de los puntos en el gráfico Q-Q normal sin tendencia.

Gráfico III.9.21. Gráficos de probabilidad normal Q-Q de los residuos. Valoración del producto según Edad e Ingresos



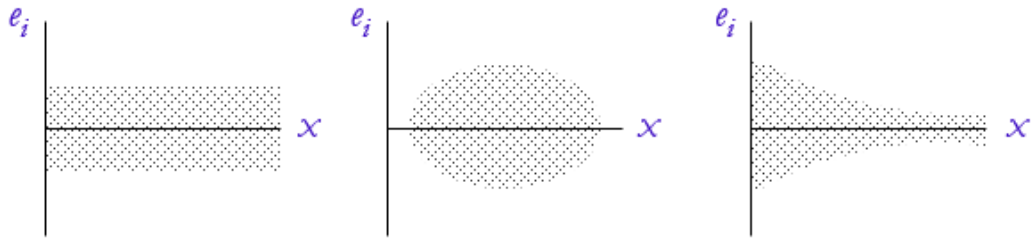
Ciertas desviaciones de la normalidad no son problemáticas si la muestra es suficientemente grande, especialmente en el contexto de estudios por encuesta con muestreos de 1000 o 2000 personas entrevistadas.

6.4. Homoscedasticidad

En regresión lineal es relevante también el supuesto según el cual la varianza de los errores debe ser la misma para cada valor de la independiente (homoscedasticidad). Si la varianza de los errores no es constante las estimaciones de los coeficientes de regresión se hacen imprecisas, no se obtienen varianzas mínimas (bajas) cuando se aplica la regresión por mínimos cuadrados ordinarios (MCO), el procedimiento habitual. Para detectar la heteroscedasticidad suele construir un gráfico de dispersión entre los residuos y los valores de la variable dependiente estimados por la regresión, o bien con la variable independiente, como el comentado anteriormente del Gráfico III.9.20.

En el Gráfico III.9.22 se esquematizan las representaciones de diferentes situaciones en relación al comportamiento de la varianza de los residuos. Cuando no se da la heteroscedasticidad obtenemos un gráfico de la forma de (a), con muestras grandes, o de la forma de (b), con muestras pequeñas. Si se da la heteroscedasticidad la forma gráfica es la de (c).

Gráfico III.9.22. Representación gráfica de la varianza de los errores



(a) Homoscedasticidad (muestra grande) *(b) Homoscedasticidad (muestra pequeña)* *(c) Heteroscedasticidad*

En el caso de que se dé heteroscedasticidad, una alternativa es utilizar estimaciones a través del llamado método de Mínimo Cuadrados Ponderados (MCP), dando menos peso a los valores de la variable independiente que tienen mayor variabilidad. También es posible transformar la variable dependiente haciendo la raíz cuadrada y aplicando la regresión por MCO.

6.5. Casos atípicos e influyentes

Como venimos destacando, un análisis de regresión se completa con un análisis de verificación de las condiciones de aplicación para que las conclusiones derivadas de las estimación de la ecuación de regresión sean válidas. Estos análisis se completan además con el estudio de los casos atípicos e influyentes. En este sentido, la ecuación de regresión final que se obtiene en el análisis es un proceso que implica habitualmente tareas progresivas de comprobaciones, correcciones y generación de diversas regresiones hasta que se validan los resultados del análisis.

En estos tratamientos de los datos la información de los residuos que se obtienen en cada regresión es fundamental como vimos. El análisis de residuos permite también detectar la existencia de casos atípicos o extremos (*outliers*), es decir, observaciones que tienen un comportamiento que difiere notablemente del conjunto de casos y que impiden un mejor ajuste de la recta de regresión. Los casos atípicos pueden observarse tanto en la variable dependiente como en las independientes, y pueden constatarse de forma individual o conjunta.

Los residuos, la diferencia entre el valor observado y el valor estimado por la regresión ($e_i = y_i - \hat{Y}_i$), reflejan el grado de error que cometemos al expesar los datos observados en la muestra por los de la recta de regresión. De ello se deriva que los casos más extremos son los que tendrán un valor absoluto mayor. Estos valores se suelen estandarizar, dividiendo por la desviación típica de los residuos, con lo que tendremos un estadístico de puntuaciones típicas que sigue una distribución normal. Los residuos estandarizados o tipificados e_i^z se expresan como:

$$e_i^z = \frac{y_i - \hat{Y}_i}{\sqrt{V_{Res}}} \qquad \textit{Ecuación 25}$$

donde V_{Res} es la varianza o media cuadrática de los residuos o errores del modelo, y tienen una distribución con media 0 y desviación típica 1.

Los residuos estandarizados o tipificados con valores superiores a 3 unidades de desviación corresponden a casos muy atípicos, muy alejados de la recta o el hiperplano de regresión y, en consecuencia, no son bien pronosticados por la ecuación de regresión. Una vez se han detectado y valorado, estos casos se pueden excluir del análisis consiguiendo así una mejora en la bondad de ajuste del modelo. Este aspecto tiene su importancia de cara a la predicción cuando se consideran diferentes valores de la variable independiente. La recta de regresión que se obtiene muestra una relación lineal para el rango de valores observados, los cuales no siempre están alineados a lo largo de la recta de regresión, en algunos casos o valores extremos la relación lineal se impone sin estar allí.

Es posible también considerar los residuos studentizados, estos es, a partir de tipificar los residuos en relación al propio error típico de cada puntuación y no en relación al promedio que representa V_{Res}, siguiendo una distribución de t-student con $n-p-1$ grados de libertad. La expresión es:

$$e_i^t = \frac{y_i - \widehat{Y}_i}{s_{e_i}} \qquad \textit{Ecuación 26}$$

donde s_{e_i} es el error típico de los residuos. Con estos residuos podemos evaluar si un caso concreto se aleja de forma significativa del valor pronosticado por la recta de regresión. Si la muestra es suficientemente grande, el criterio anterior de considerar valores superiores a 3 como casos extremos se aplica igualmente.

Junto al cálculo de estos estadísticos que nos informan de la presencia de estos casos se pueden obtener representaciones gráficas de la nube de puntos que ofrecen una forma visual de localizarlos. No obstante, cuando dispongamos de más de tres variables en el análisis de regresión la información del estadístico será la única referencia informativa y no dispondremos de una representación gráfica. El diagrama de dispersión que se emplea para mostrar los casos atípicos sitúa en el eje horizontal los valores pronosticados y los residuos en el eje vertical. Cuando se observa este tipo de representación no debe aparecer ningún tipo de estructura en la secuencia de errores (sobre todo temporal); parcialmente puede aparecer pero no debe ser constante.

Los casos atípicos también se valoran en relación a las variables independientes y es posible cuantificar estas situaciones mediante una medida de influencia h_i, donde se valora el grado de alejamiento de cada caso respecto del centro definidos por el conjunto de variables independientes y, por tanto, de influencia de cada caso en la forma lineal de la regresión. Estos valores se puede calcular también a partir de la medida de influencia centrada. Cuanto mayor sea el valor obtenido mayor será su influencia. Una regla práctica (Pardo y San Martín, 2015: 455) es considerar los valores inferiores a 0,2 como poco problemáticos, entre 0,2 y 0,5 como arriesgados y los superiores a 0,5 son los que deberían revisarse.

Un caso atípico difiere de un caso influyente. Un caso puede comportarse de forma atípica pero no alterar de forma significativa los resultados del análisis, es decir, si no modifican de forma importante la ecuación de regresión y las pendientes del hiperplano de regresión. En regresiones con una o dos variables independientes puede analizarse este efecto gráficamente, pero, en general, es necesario analizar el comportamiento en los coeficientes de regresión como resultado de eliminar el caso en cuestión de los datos analizados, y ver también el efecto en los valores pronosticados y los residuos.

El efecto en los coeficientes de regresión de los casos influyentes se puede valorar a partir del cálculo de la diferencia entre los coeficientes de regresión tipificados, los dfbetas (Belsey, Kuh y Welsch, 1980). Se obtiene un valor para caso como resultado de la diferencia en cada coeficiente de regresión incluyendo y excluyendo el caso evaluado, que es además tipificado. Los valores por encima de $2/\sqrt{n}$ se cosidera que deben ser valorados como casos influyentes.

Alternativamente se puede emplear otro estadístico denominado Distancia de Cook, D_i (Cook, 1979) que expresa el cambio experimentado por los coeficientes de regresión de forma conjunta al eliminar cada caso. Este cálculo se comporta siguiendo una distribución F com $p+1$ y $n-p-1$ grados de libertad, y se considera que valores superiores a 1 deben ser revisados.

El efecto de los casos influyentes se puede ver reflejado también en los valores pronosticados calculando la diferencia tipificada para cada caso, los dffits (Belsey, Kuh y Welsch, 1980). Los valores superiores a $2\sqrt{(p+1)/n}$ deben ser revisados.

Analizando los cambios en los valores de los residuos se pueden comparar los residuos studentizados originales con los obtenidos como resultado de eliminar el caso en cuestión. Si el caso es influyente el resultado de la diferencia será un valor diferente de cero. Los casos que se corresponden a un residuo eliminado studentizado superior a 3 deberán ser revisados para evaluar las particularidades de su comportamiento y la eventual eliminación en aras del mejor ajuste que configuran el resto de los casos que son los típicos.

7. El análisis de regresión con SPSS

Para ejemplificar el análisis de regresión con el software SPSS trabajaremos con la matriz de datos sobre el Índice de Desarrollo Humano IDH2014.sav y reproduciremos en ejercicio de análisis que hemos ido explicando a lo largo del capítulo. Se trata de una matriz de datos de países con la realizamos el análisis con tres variables: la variable dependiente que queremos explicar es la esperanza de vida al nacer (variable Lifeexpectancy) en función de la riqueza del país, medido a través de los ingresos nacional brutos per cápita de 2011 a precio de paridad de compra (variable GNIpercapita2011 -Gross National Income (GNI) per capita 2011 PPP $-), y el nivel educativo, medido a través de la media de años de escolarización (variable Schooling). Presentaremos primero un análisis de regresión simple y lo extenderemos a la regresión múltiple. El archivo de sintaxis ARL-Esperanza.sps reproduce los análisis que aquí se presentan.

Utilizaremos tres procedimientos del SPSS destinados a reproducir los resultados del análisis de la relación entre variables con el modelo de regresión:

1) Gráficos, a través del menú o de los comandos **GRAPH**, **GGRAPH** o **STATS REGRESS PLOT**. Con ellos obtendremos gráficos de dispersión de la relación entre las variables cuantitativas que se analizan, así como otras variables que se obtienen de la aplicación de la técnica del análisis de regresión lineal.
2) Correlaciones bivariadas a través del menú Analizar o del comando **CORRELATION**, para calcular y contrastar la significación de la correlación entre pares de variables cuantitativas.
3) Regresión lineal del menú Analizar, comando **REGRESSION**, que realiza el análisis estadístico de la relación de dependencia lineal entre una variable cuantitativa y una o más variables cuantitativas independientes.

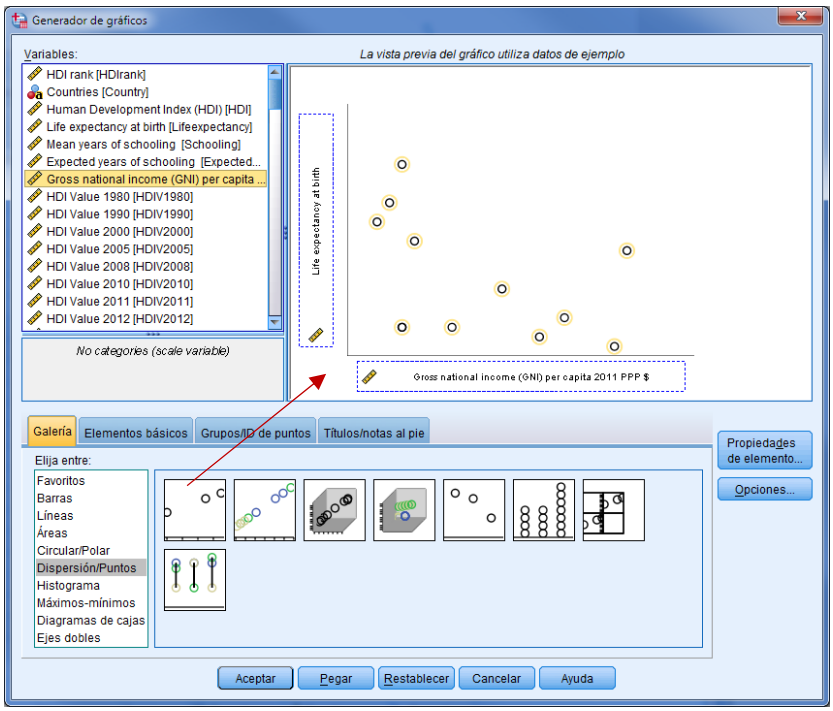
7.1. El análisis de regresión simple

En primer lugar analizaremos la correlación y representaremos los gráficos de dispersión con las variables consideradas. Se trata de mostrar la existencia de una relación lineal por lo que la inspección visual inicial permite ver esta característica del vínculo entre ellas y ayuda a interpretar mejor el posterior cálculo del coeficiente de correlación lineal.

7.1.1. *Gráficos de dispersión*

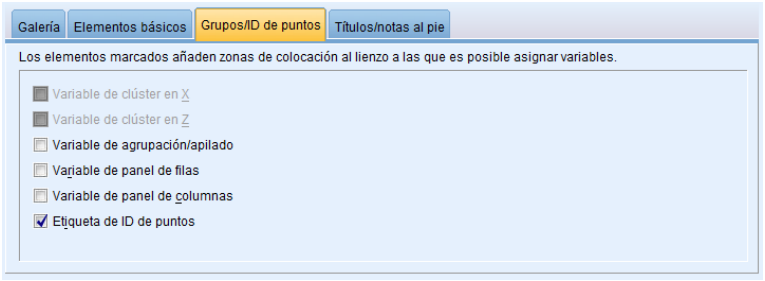
Los diagramas de dispersión representan gráficamente la relación entre dos o tres variables cuantitativas. Esta representación permite observar la naturaleza de la relación y poder constatar si ésta es de tipo lineal, curvilínea, aleatoria, etc., y que los cálculos estadísticos no revelarían, o también detectar valores extremos atípicos que podrían limitar el análisis.

Para obtener un gráfico de dispersión iremos al menú Gráficos / Generador de gráficos y en el cuadro de diálogo principal, en la pestaña Galería, elegiremos Dispersión / Puntos, y entre las diversas opciones haremos doble clic sobre el gráfico de Dispersión Simple.

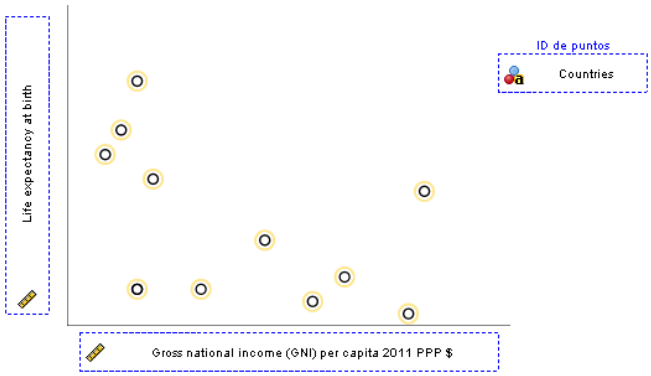


Seguidamente traspasamos, seleccionando y arrastrando, la variable Lifeexpectancy en el cuadro del Eje Y, el vertical, que consideramos como convención el eje de la variable dependiente, y la variable GNIpercapita2011 en el Eje X, el horizontal que consideramos como eje de la variable independiente.

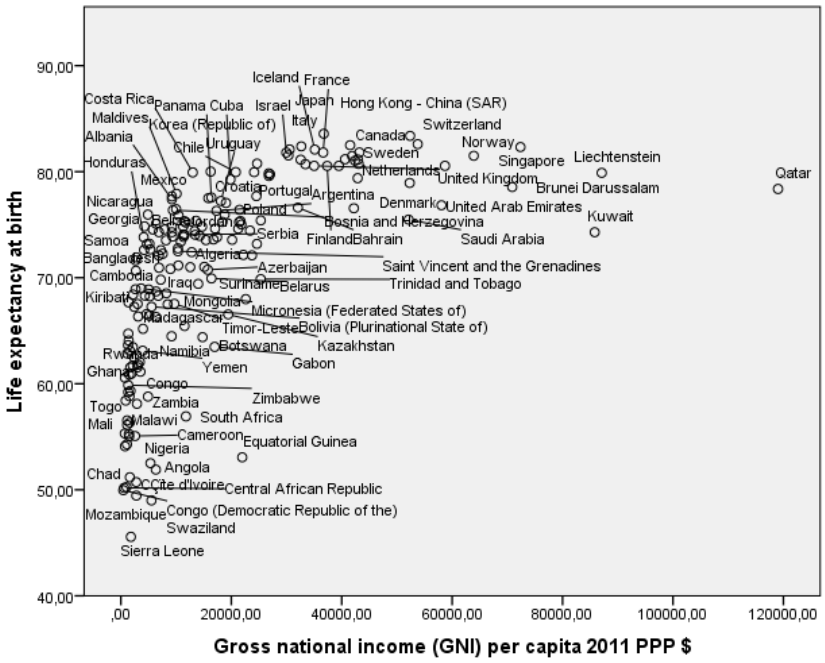
Adicionalmente, podemos etiquetar los puntos de gráfico con la identificación del nombre del país (variable Countries). Para ello, elegimos Etiqueta de ID de puntos en la pestaña Grupos/ID de puntos:



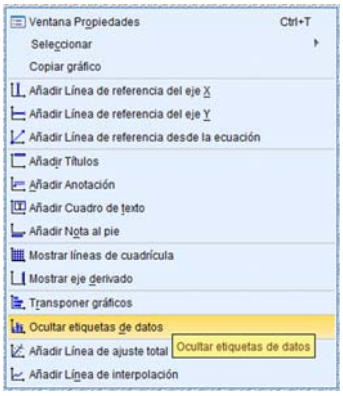
Sobre el gráfico aparecer un recuadro donde traspasaremos la variable Countries:



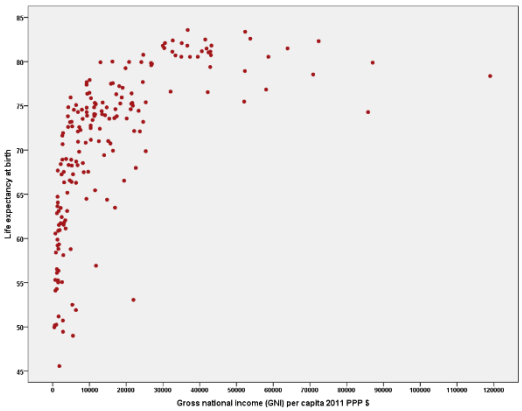
Después de Aceptar obtenemos esta representación:



Para editarlo clicamos doblemente sobre él. Primeramente modificaremos el gráfico de dispersión para presentar más claramente la forma de la nube de puntos sin etiquetas. Le daremos al botón derecho del ratón y sobre el menú contextual elegiremos Ocultar etiquetas de datos:



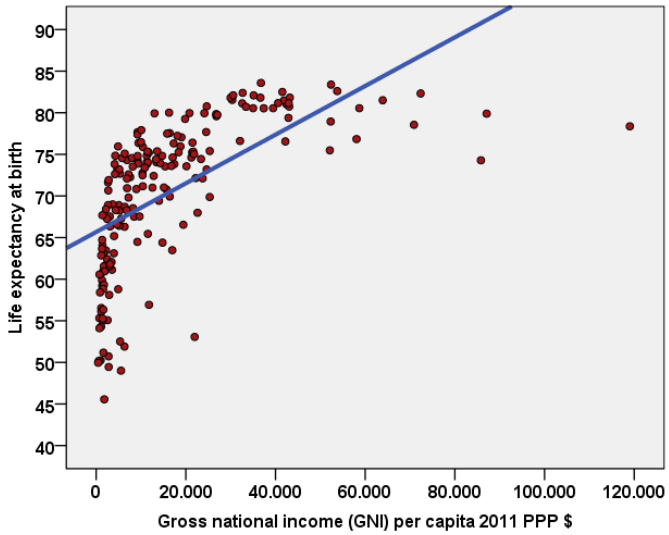
También cambiaremos las marcas de los puntos y ajustaremos los valores de la escala de los ejes quitando decimales y ampliando el número de marcas. La nueva representación es como la siguiente:



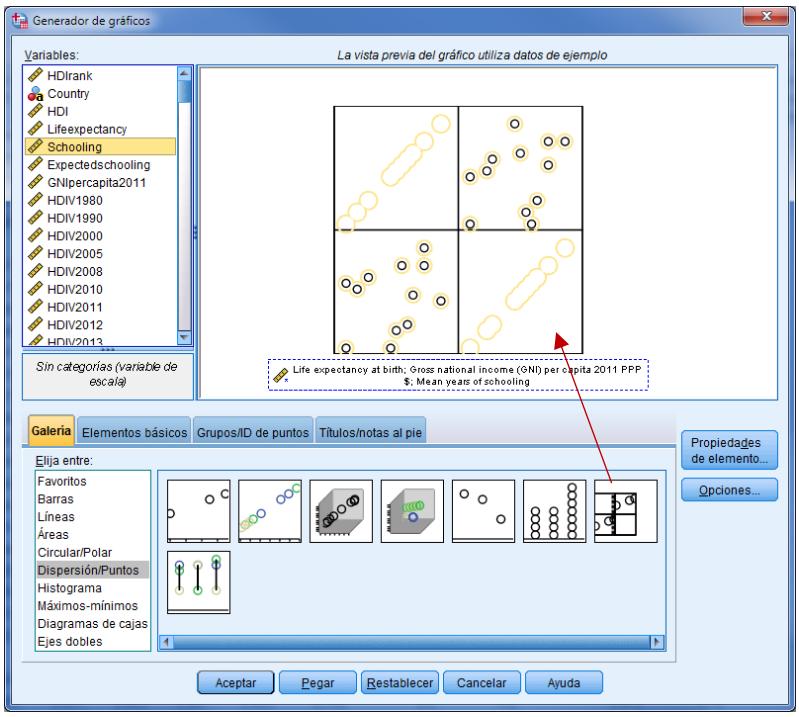
Con la herramienta del menú Elementos denominada Modo de etiqueta de datos hemos etiquetado una selección de países:

Con la herramienta Ir a caso del menú Editar se puede ver en la matriz de datos el caso que se seleccione en el gráfico.

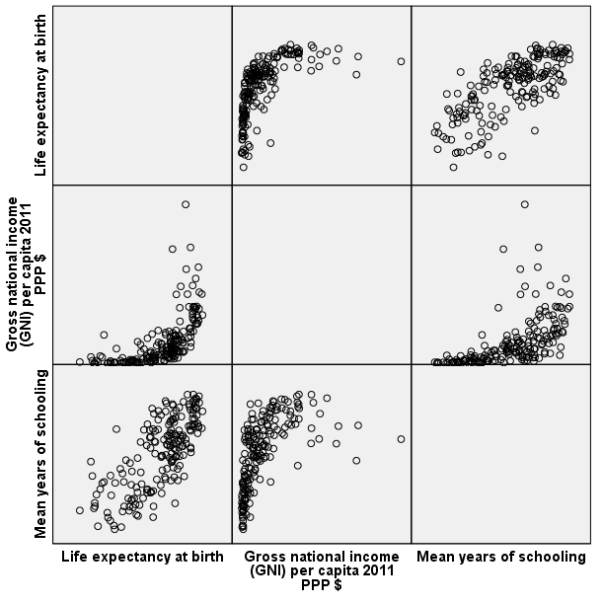
Una posibilidad de las representaciones de diagramas de dispersión es la opción de ajustar una línea a la nube de puntos, es decir, ajustar la recta de regresión. Esta opción se configura y realiza fuera de este procedimiento, en el editor de gráficos en el visor de resultados. Una vez editado el gráfico hay que acceder a las opciones, o a través del menú contextual, y marcar Añadir línea de ajuste total.



Con las tres variables que consideramos de la matriz de datos del IDH podemos realizar un diagrama de dispersión matricial para disponer de una representación simultánea de todos los gráficos de dispersión por parejas de variables. Iremos al menú Gráficos / Generador de gráficos y en el cuadro de diálogo principal, en la pestaña Galería, elegiremos Dispersión / Puntos, y entre las diversas opciones haremos doble clic sobre el gráfico de Dispersión Matricial. En el recuadro ¿Matriz de dispersión? Incluiremos las tres variables Lifeexpectancy, GNIpercapita2011 y Schooling.



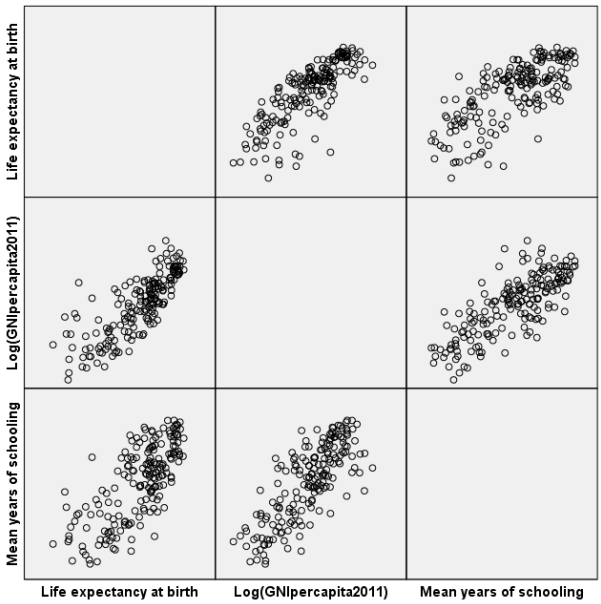
El resultado es este gráfico:



Donde se puede observar la relación no lineal que mantienen las variables variables Lifeexpectancy, y Schooling con GNIpercapita2011. Para transformar la variable GNIpercapita2011 por su logaritmo en base 10 generaremos una nueva variable de nombre LogGNI, a través del menú Transformar / Calcular variable o a través de la sintaxis con la instrucción siguiente:

```
COMPUTE LogGNI=LG10(GNIpercapita2011).
```

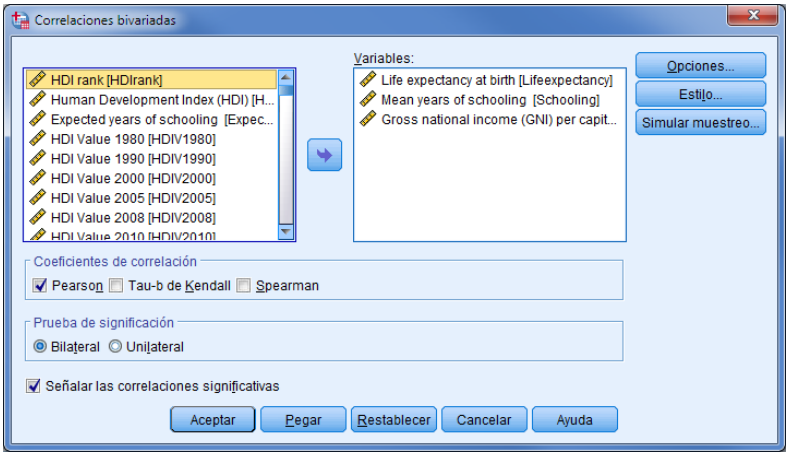
Si obtenemos de nuevo el gráfico de dispersión matricial el resultado es el siguiente:



Donde constatamos ahora la disposición lineal de las diferentes nubes de puntos.

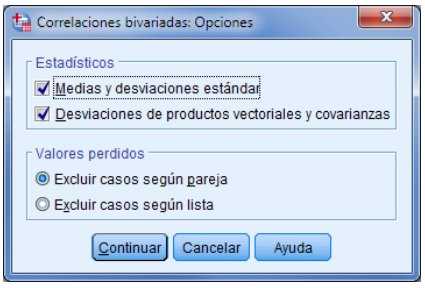
7.1.2. *Análisis de correlación*

El procedimiento **CORRELATIONS**, menú Analizar / Correlacionar / Bivariadas, calcula el coeficiente de correlación de Pearson, la rho de Spearman y la tau-b de Kendall, contrastando también su significación. El cuadro de diálogo inicial es el siguiente:



En este caso hemos traspasado las tres variables que analizamos Lifeexpectancy, GNIpercapita2011 y Schooling al recuadro Variables para calcular las correlaciones de Pearson bivariables entre todas ellas. Por defecto se hacen las pruebas de significación en dos colas (bilateral) y se marcarán las correlaciones que sean significativas.

Para completar las especificaciones de este procedimiento el botón de Opciones nos ofrece la posibilidad de pedir varios estadísticos: la media y la desviación típica de cada variable, y los productos cruzados de las desviaciones y las covarianzas para cada pareja de variables. También controla el tratamiento de los valores perdidos.



Los resultados que se obtienen son los siguientes:

Estadísticos descriptivos

	Media	Desviación estándar	N
Life expectancy at birth	70,4266	8,88343	191
Mean years of schooling	7,9008	3,05877	187
Gross national income (GNI) per capita 2011 PPP $	16.486,5138	18.383,00157	190

Correlaciones

		Life expectancy at birth	Mean years of schooling	Gross national income (GNI) per capita 2011 PPP $
Life expectancy at birth	Correlación de Pearson	1	0,729	0,608
	Sig. (bilateral)		0,000	0,000
	Suma de cuadrados y productos vectoriales	14.993,925	3.662,611	18644174,63
	Covarianza	78,915	19,691	99.171,142
	N	191	187	189
Mean years of schooling	Correlación de Pearson	0,729	1	0,560
	Sig. (bilateral)	0,000		0,000
	Suma de cuadrados y productos vectoriales	3.662,611	1.740,233	5876420,898
	Covarianza	19,691	9,356	31.593,661
	N	187	187	187
Gross national income (GNI) per capita 2011 PPP $	Correlación de Pearson	0,608	0,560	1
	Sig. (bilateral)	0,000	0,000	
	Suma de cuadrados y productos vectoriales	18644174,63	5876420,898	6,387E+10
	Covarianza	99.171,142	31.593,661	337934746,6
	N	189	187	190

Los coeficientes obtenidos son moderadamente altos y positivos. En el caso de las relaciones donde interviene la riqueza del país, no obstante, la tendencia no lineal observada gráficamente el coeficiente de correlación lineal subvalora el grado de relación existente y nos llevó a transformarla en su logaritmo decimal. Con la variable transformada obtenemos los resultados siguientes que muestran la mejora de los coeficientes de correlación.

Estadísticos descriptivos

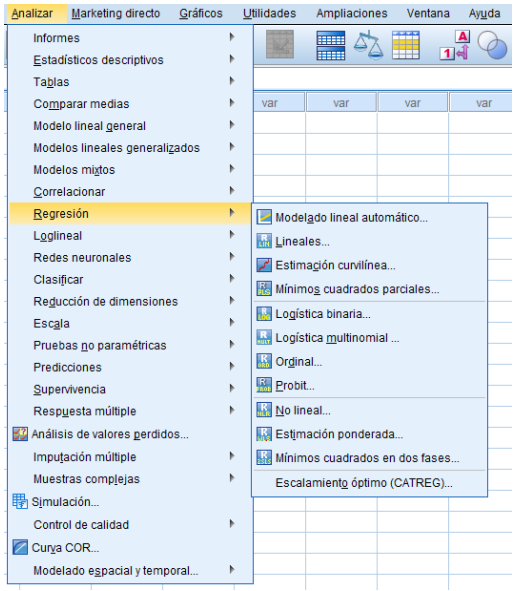
	Media	Desviación estándar	N
Life expectancy at birth	70,4266	8,88343	191
Mean years of schooling	7,9008	3,05877	187
Log(GNIpercapita2011)	3,9508	0,52499	190

Correlaciones

		Life expectancy at birth	Mean years of schooling	Log (GNIpercapita 2011)
Life expectancy at birth	Correlación de Pearson	1	0,729	0,802
	Sig. (bilateral)		0,000	0,000
	Suma de cuadrados y productos vectoriales	14.993,925	3.662,611	702,409
	Covarianza	78,915	19,691	3,736
	N	191	187	189
Mean years of schooling	Correlación de Pearson	0,729	1	0,782
	Sig. (bilateral)	0,000		0,000
	Suma de cuadrados y productos vectoriales	3.662,611	1.740,233	233,499
	Covarianza	19,691	9,356	1,255
	N	187	187	187
Log(GNIpercapita2011)	Correlación de Pearson	0,802	0,782	1
	Sig. (bilateral)	0,000	0,000	
	Suma de cuadrados y productos vectoriales	702,409	233,499	52,091
	Covarianza	3,736	1,255	0,276
	N	189	187	190

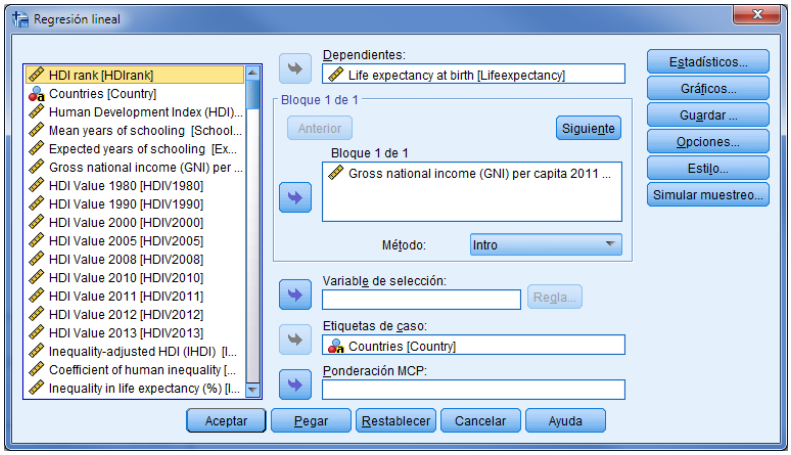
7.1.3. *Análisis de regresión*

El análisis de regresión lineal es el procedimiento destinado a estimar los coeficientes de la ecuación lineal, con una variable independiente (análisis de regresión simple) o más de una variable independiente (análisis de regresión múltiple), que mejor prediga los valores de la variable dependiente. Se corresponde con el comando **REGRESSION** y se accede por el menú a través Analizar / Regresión / Lineales.



Podemos ver que existe todo un conjunto de procedimientos que realizan también regresiones en condiciones o métodos distintos. La regresión lineal clásica es la que presentamos aquí.

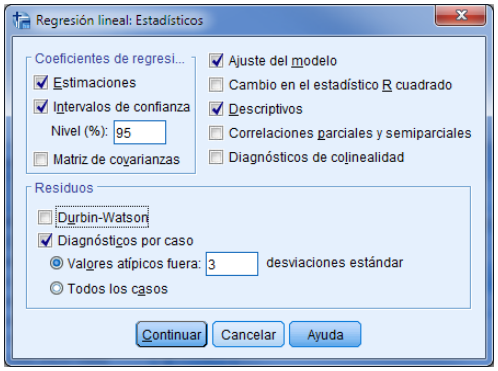
El cuadro de diálogo principal del procedimiento requiere que se especifique la variable dependiente, en nuestro caso elegiremos la variable Lifeexpectancy, y consideraremos la variable independiente GNIpercapita2011 incialmente. Pero como hemos comentado se trata de una variable que manifiesta una relación con la variable dependiente que viola la condición de aplicación de la linealidad. Por ello repetiremos este análisis, con las mismas especificaciones, para la variable transformada con el logaritmo decimal LogGNI. El cuadro de diálogo inicial es el siguiente, donde hemos introducido las variables para el análisis de regresión simple:



A partir de esta selección básica de variables tenemos la opción de considerar una submuestra de casos que se seleccionarían a través de una variable que debería incluirse en el cuadro Variable de selección. Igualmente tenemos la opción de elegir una variable para identificar los casos en el recuadro Etiquetas de caso, hemos utilizado la variable del nombre del país Country. También se debe seleccionar el método de selección de las variables; en este caso utilizaremos por defecto la opción Intro que implica forzar la introducción simultánea de todas las variables independientes seleccionadas. Por su parte el recuadro de MCP permite obtener un modelo de Mínimos Cuadrados Ponderados en condiciones de heteroscedasticidad, donde se pretende que las observaciones con varianzas grandes tengan un menor impacto en el análisis frente a las observaciones asociadas a varianzas pequeñas. Esta opción no la utilizaremos.

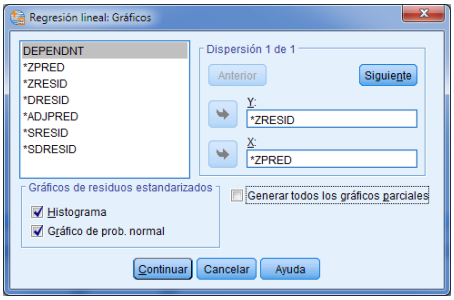
A continuación se trata de realizar las especificaciones necesarias a partir de los botones Estadísticos, Gráficos, Guardar y Opciones que presentamos a continuación.

La especificación de los Estadísticos se hace a partir del cuadro de diálogo siguiente:



En este caso hemos elegido la opción de obtener la tabla de las Estimaciones de los coeficientes de regresión del modelo con su significación así como los Intervalos de confianza, considerado un 95% de nivel de confianza. Por otra parte, la opción Ajuste del modelo nos proporciona dos tablas, una de resumen del modelo donde se calcula el coeficiente de determinación R^2 para medir la capacidad explicativa y el poder de predicción del modelo, y otra tabla que reproduce un análisis de varianza para contrastar si el coeficiente de determinación poblacional es significativamente diferente de cero. Marcaremos también la opción de Diagnóstico por caso para que se listen los casos atípicos, aquellos que se alejan de la recta de regresión 3 unidades de desviación. Si existen se puede optar por eliminarlos del análisis con el objetivo de mejorar el ajuste de la recta de regresión.

El procedimiento de la regresión nos facilita la obtención de diversas representaciones gráficas:

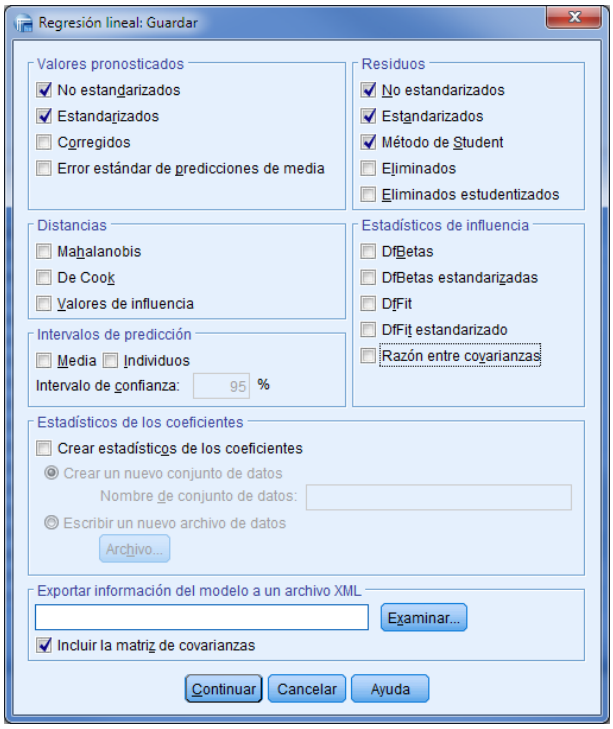


Se puede considerar la variable dependiente (DEPENDNT) y las variables pronosticadas y residuales siguientes: valores pronosticados tipificados (*ZPRED), residuos tipificados (*ZRESID), residuos eliminados (*DRESID), valores pronosticados corregidos (*ADJPRED), residuos estudentizados (*SRESID) y residuos estudentizados eliminados (*SDRESID).

Estos gráficos nos ayudan a validar los supuestos de normalidad, linealidad e igualdad de las varianzas. También son de utilidad para detectar valores atípicos, observaciones poco habituales y casos de influencia. En particular si representamos los residuos tipificados con los valores pronosticados tipificados podemos contrastar la linealidad de la relación y el supuesto de igualdad de las varianzas.

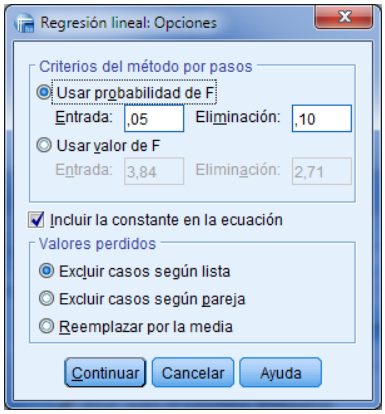
El cuadro de diálogo nos ofrece la opción de generar Gráficos de residuos estandarizados: gráficos de residuos tipificados (histogramas de los residuos tipificados y gráficos de probabilidad normal) y también optar por gráficos parciales cuando hay más de una variable independiente.

Estas variables y otras derivadas del análisis de regresión se pueden Guardar y se pueden reproducir nuevos gráficos y análisis. Para guardar las variables en el archivo de datos y hacer uso posteriormente para completar el análisis disponemos de varias opciones que se pueden ver en el siguiente cuadro de diálogo:



Hemos optado por guardar los Valores pronosticados por la recta de regresión, estandarizados y no, así como los Residuos: absolutos, estandarizados y los estudentizados. Adicionalmente, para realizar un estudio de los casos atípicos influyente se pueden elegir los Estadísticos de influencia y e Distancias.

Varias especificaciones adicionales se encuentran dentro de Opciones:



Se definen los criterios de probabilidad o valor crítico del método por pasos que se aplican a los métodos de selección de variables definido en el cuadro de diálogo principal, la inclusión de la constante en la ecuación y el tratamiento de valores perdidos. Dejaremos las opciones por defecto que marca el procedimiento estadístico del SPSS.

Tras la ejecución del procedimiento obtenemos los resultados que se presentan a continuación.

Estadísticos descriptivos

	Media	Desviación estándar	N
Life expectancy at birth	70,5102	8,85942	189
Gross national income (GNI) per capita 2011 PPP $	16.546,4927	18.413,17780	189

Correlaciones

		Life expectancy at birth	Gross national income (GNI) per capita 2011 PPP $
Correlación de Pearson	Life expectancy at birth	1,000	0,608
	Gross national income (GNI) per capita 2011 PPP $	0,608	1,000
Sig. (unilateral)	Life expectancy at birth	.	0,000
	Gross national income (GNI) per capita 2011 PPP $	0,000	.
N	Life expectancy at birth	189	189
	Gross national income (GNI) per capita 2011 PPP $	189	189

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	Gross national income (GNI) per capita 2011 PPP $[b]	.	Entrar

a. Variable dependiente: Life expectancy at birth

b. Todas las variables solicitadas introducidas.

Como podremos comprobar con relación a los resultados de la regresión con la variable de renta transformada, la bondad de ajuste del modelo que expresa el R^2 se verá notablemente incrementada en relación al valor obtenido de 0,370.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	0,608[a]	0,370	0,366	7,05310

a. Predictores: (Constante), Gross national income (GNI) per capita 2011 PPP $

b. Variable dependiente: Life expectancy at birth

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	5.453,446	1	5.453,446	109,625	0,000[b]
	Residuo	9.302,536	187	49,746		
	Total	14.755,982	188			

a. Variable dependiente: Life expectancy at birth

b. Predictores: (Constante), Gross national income (GNI) per capita 2011 PPP $

La recta de regresión que se obtiene es *Lifeexpectancy* = 65,67 + 0,0002925 *GNIpercapita2011* indicando que ante una variación de un dólar la esperanza de vida aumenta en 0,000293 años, o lo que es lo mismo, por cada 10.000$ el aumento es de casi 3 años[16].

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B	
		B	Error estándar	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	65,670	0,691		95,096	0,000	64,308	67,033
	Gross national income (GNI) per capita 2011 PPP $	0,000293	0,000028	0,608	10,470	0,000	0,000237	0,000348

a. Variable dependiente: Life expectancy at birth

En la regresión se puede observar la existencia de un caso extremo, Qatar, con un residuo estandarizado superior a 3 unidades de desviación por debajo de la recta de regresión.

Diagnósticos por casos[a]

Número del caso	Countries	Residuo estándar	Life expectancy at birth	Valor pronosticado	Residuo
31	Qatar	-3,136	78,37	100,4865	-22,11651

a. Variable dependiente: Life expectancy at birth

Observando los valores de mínimo y máximo de la tabla Estadística de residuos se pueden ver, en particular, si existen residuos estandarizados o estudentizados que superan las tres unidades de desviación, como nos encontramos en esta regresión. En el caso de los valores influyentes, se darían casos problemáticos si hubiera valores superiores a 0,5. No es el caso como se puede verificar en la línea de Variable de influencia centrado. Sí se observan valores por encima de 1 en el caso de la Distancia de Cook. Como resultado de la transformación y del tratamiento de los datos que operaremos seguidamente estos valores serán tenidos en cuenta y corregidos sus efectos.

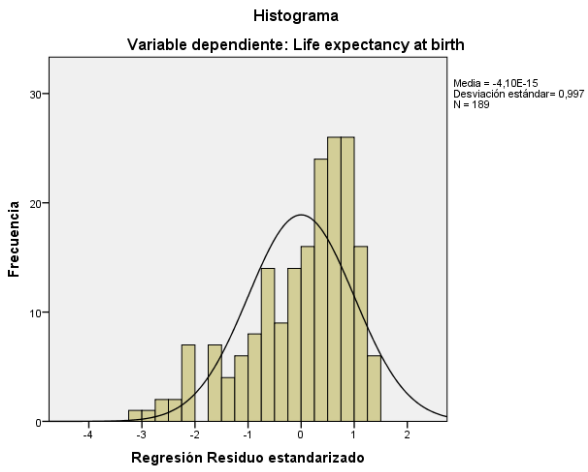
[16] El coeficiente de la constante de hecho no es interpretable: si un país tuviera 0$ de renta la esperanza de vida sería de 65,67 años, pero no se observan países en esa situación.

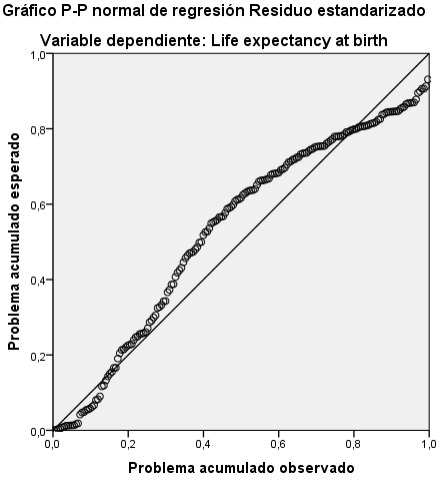
Estadísticas de residuos[a]

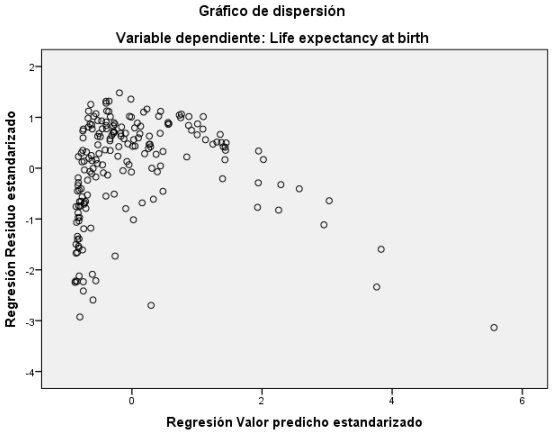
	Mínimo	Máximo	Media	Desviación estándar	N
Valor pronosticado	65,8002	100,4865	70,5102	5,38588	189
Valor pronosticado estándar	-0,875	5,566	0,000	1,000	189
Error estándar de valor pronosticado	0,513	2,909	0,670	0,279	189
Valor pronosticado corregido	65,9398	105,0184	70,5527	5,57865	189
Residuo	-22,11651	10,45372	0,00000	7,03431	189
Residuo estándar	-3,136	1,482	0,000	0,997	189
Residuo estudentizado	-3,442	1,486	-0,003	1,008	189
Residuo eliminado	-26,64845	10,51139	-0,04253	7,19785	189
Residuo estudentizado suprimido	-3,547	1,491	-0,006	1,016	189
Distancia de Mahal.	0,000	30,977	0,995	2,896	189
Distancia de Cook	0,000	1,214	0,012	0,090	189
Valor de influencia centrado	0,000	0,165	0,005	0,015	189

a. Variable dependiente: Life expectancy at birth

Los gráficos siguientes (Histograma y Gráfico P-P) nos permiten observar la distribución de los residuos estandarizados que muestran un alejamiento respecto de la normalidad. Por su parte, el gráfico de dispersión entre el residuo estandarizado y el valor pronosticado estandarizado no presenta una distribución aleatoria de los puntos a lo largo del gráfico, lo que nos indica un alejamiento de las condiciones deseables de distribución de los residuos para cumplir con las condiciones de linealidad y homoscedasticidad.







Para mejorar estos resultados ejecutaremos las mismas especificaciones del análisis de regresión que acabamos de realizar pero con la variable transformada LogGNI. Los

resultados siguientes son diferentes y los distintos cálculos y resultados se verán modificados y mejorados.

Estadísticos descriptivos

	Media	Desviación estándar	N
Life expectancy at birth	70,5102	8,85942	189
Log(GNIpercapita2011)	3,9521	0,52610	189

Correlaciones

		Life expectancy at birth	Log (GNIpercapita 2011)
Correlación de Pearson	Life expectancy at birth	1,000	0,802
	Log(GNIpercapita2011)	0,802	1,000
Sig. (unilateral)	Life expectancy at birth	.	0,000
	Log(GNIpercapita2011)	0,000	.
N	Life expectancy at birth	189	189
	Log(GNIpercapita2011)	189	189

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	Log (GNIpercapita2011)[b]	.	Entrar

a. Variable dependiente: Life expectancy at birth

b. Todas las variables solicitadas introducidas.

En la nueva regresión con la variable transformada se alcanza un coeficiente de determinación de 0,643, lo que supone un 27,35 más de varianza explicada de la variable dependiente.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	0,802[a]	0,643	0,641	5,31073

a. Predictores: (Constante), Log(GNIpercapita2011)

b. Variable dependiente: Life expectancy at birth

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	9.481,863	1	9.481,863	336,190	0,000[b]
	Residuo	5.274,119	187	28,204		
	Total	14.755,982	188			

a. Variable dependiente: Life expectancy at birth

b. Predictores: (Constante), Log(GNIpercapita2011)

La ecuación de regresión estimada da lugar a un coeficiente de regresión de 13,499 que nos indica que por cada incremento de una unidad en el logaritmo de la renta per cápita, la esperanza de vida aumenta en 13,499 años, es decir, lo que corresponde a una renta de 10 dólares, si calculamos el antilogaritmo, es decir, si operamos 10^1, la esperanza de vida se verá aumentada en 13,499.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B	
		B	Error estándar	Beta	t	Sig.	Límite inferior	Límite superior
1	(Constante)	17,161	2,935		5,847	0,000	11,371	22,951
	Log(GNIpercapita2011)	13,499	0,736	0,802	18,335	0,000	12,047	14,951

a. Variable dependiente: Life expectancy at birth

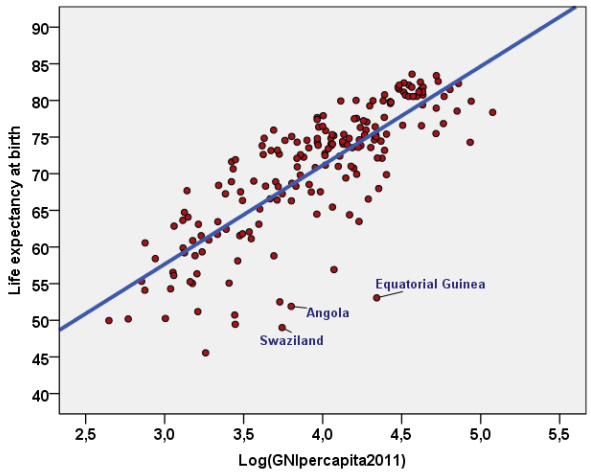
Con esta ecuación de regresión observamos que en relación a la recta ajustada tres países se encuentran a una distancia importante superior a las 3 unidades de desviación de residuo estandarizado:

Diagnósticos por casos[a]

Número del caso	Countries	Residuo estándar	Life expectancy at birth	Valor pronosticado	Residuo
144	Equatorial Guinea	-4,277	53,06	75,7720	-22,71202
148	Swaziland	-3,519	49,00	67,6909	-18,69089
149	Angola	-3,120	51,90	68,4696	-16,56963

a. Variable dependiente: Life expectancy at birth

Son los casos que se identifican en el gráfico de dispersión siguiente:

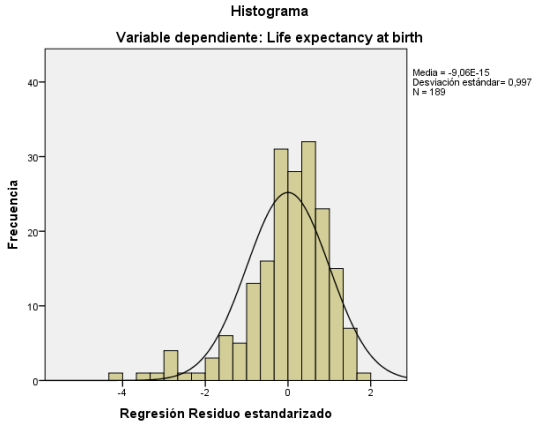


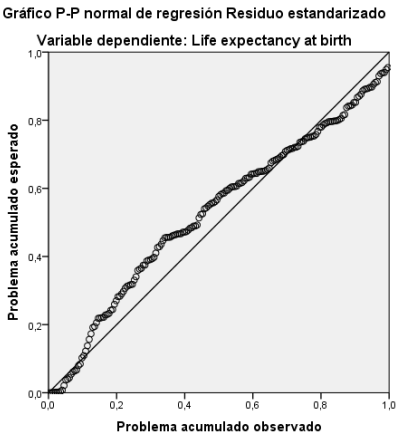
Estadísticas de residuos[a]

	Mínimo	Máximo	Media	Desviación estándar	N
Valor pronosticado	52,8974	85,6773	70,5102	7,10179	189
Valor pronosticado estándar	-2,480	2,136	0,000	1,000	189
Error estándar de valor pronosticado	0,386	1,035	0,528	0,140	189
Valor pronosticado corregido	53,0135	85,8999	70,5144	7,10278	189
Residuo	-22,71202	8,98412	0,00000	5,29659	189
Residuo estándar	-4,277	1,692	0,000	0,997	189
Residuo estudentizado	-4,294	1,697	0,000	1,002	189
Residuo eliminado	-22,90005	9,04395	-0,00417	5,34823	189
Residuo estudentizado suprimido	-4,511	1,706	-0,004	1,015	189
Distancia de Mahal.	0,000	6,151	0,995	1,136	189
Distancia de Cook	0,000	0,076	0,005	0,010	189
Valor de influencia centrado	0,000	0,033	0,005	0,006	189

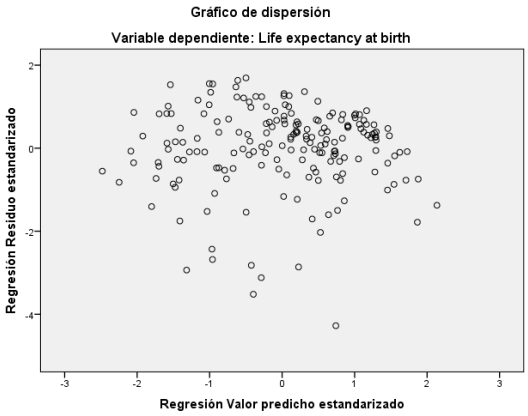
a. Variable dependiente: Life expectancy at birth

En relación al caso anterior vemos como los residuos mejoran en su aproximación al comportamiento normal:





Y se verifica una distribución aleatoria de los puntos en el gráfico de dispersión entre los residuos y los pronóstico tipificados:



Podemos seguir mejorando el ajuste del modelo si eliminamos los tres casos extremos que hemos encontrado. Si ejecutamos de nuevo el procedimiento sin esos tres países, el coeficiente de determinación sube de 0,643 a 0,698:

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	0,835[a]	0,698	0,696	4,72976

a. Predictores: (Constante), Log(GNIpercapita2011)

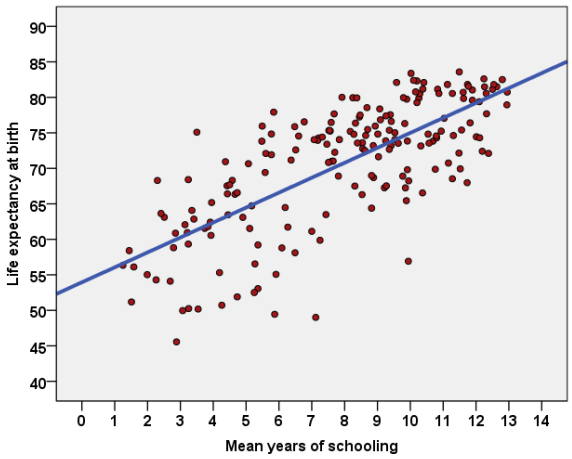
b. Variable dependiente: Life expectancy at birth

7.2. El análisis de regresión múltiple

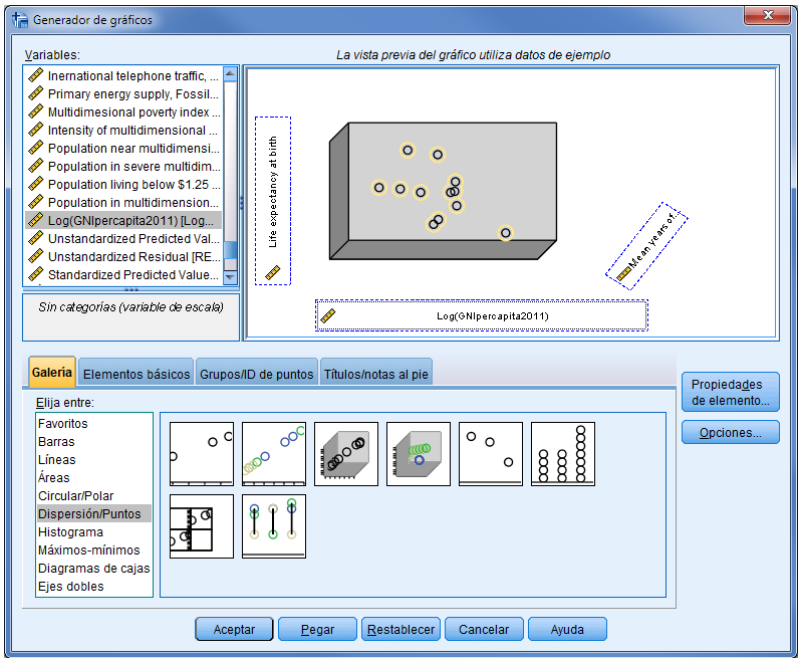
El modelo de regresión simple que acabamos de obtener mejoró como resultado de constatar que la relación funcional entre las variables era logarítmica, la función inversa de la exponencial, y su transformación nos proporcionó una mejora en el coeficiente de determinación muy notable. También mejoró como resultado de la supresión de tres casos extremos. Aun así queda una parte no explicada que genera errores de

predicción. La existencia de los residuos se debe a distintas razones que expresa nuestro modelo. Junto a la adecuada relación funcional puede suceder que tengamos errores de medición de nuestras variables. También es posible que nuestro modelo sea incompleto y que hayamos omitido variables relevantes que actúan como variables independientes explicativas que deberíamos incorporar al modelo. La mejora en la bondad de ajuste de la regresión se consigue con la especificación de más variables independientes relevantes como factores explicativos de la variable dependiente.

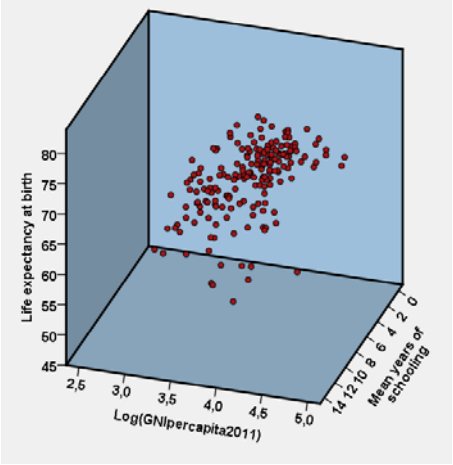
Vamos a considerar un modelo ampliado de regresión múltiple incorporando una segunda variable independiente para analizar el efecto del factor educativo, la variable Schooling, la media de años de escolarización. El gráfico de dispersión de la esperanza de vida con la nueva variable evidencia una clara tendencia lineal:



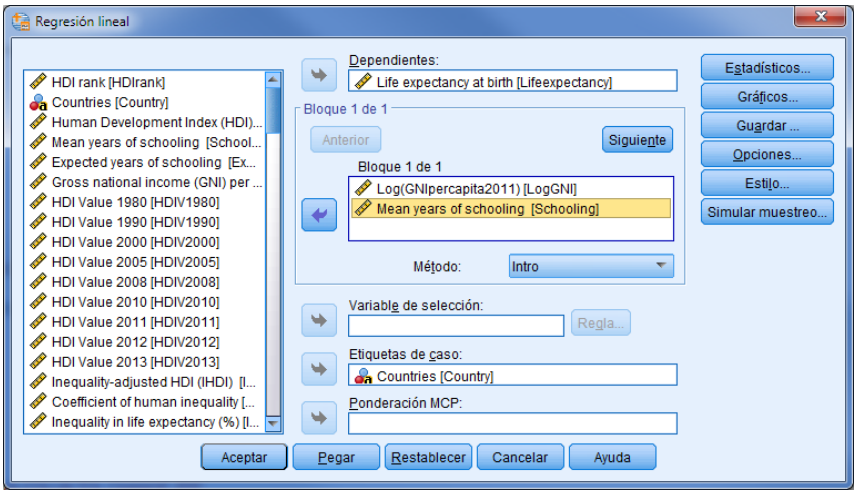
A través del menú Gráficos / Gráficos interactivos podemos representar conjuntamente las tres variables en un gráfico de tres dimensiones, seleccionando un gráfico de Dispersión 3D Simple:



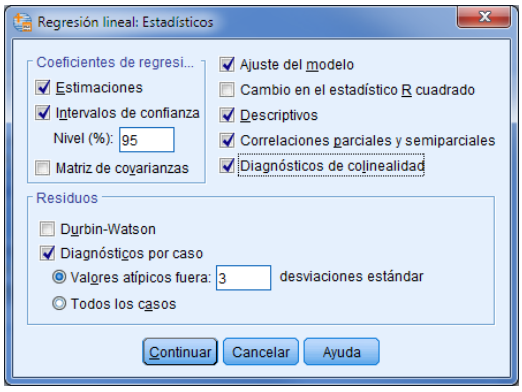
Después de editarlo se obtiene un gráfico como este:



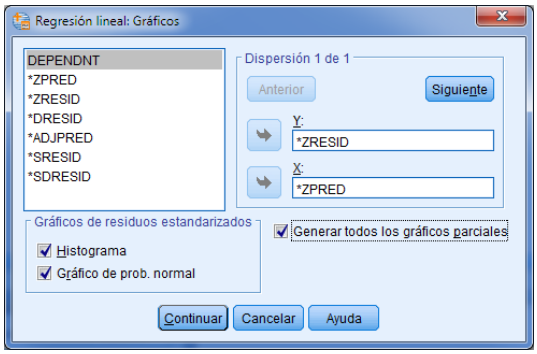
Los resultados del análisis de regresión que se presentan seguidamente incluyen a todos los casos, no se excluyen los tres países que en el anterior análisis se posicionaron como extremos. Se han sido obtenido con las mismas especificaciones que vimos con anterioridad introduciendo la variable Schooling como segunda variable independiente:



Además, en Estadísticos, se han marcado las opciones que nos permite visualizar las Correlaciones parciales y semiparciales así como los Diagnósticos de colinealidad:



Y en Gráficos se ha marcado Generar todos los gráficos parciales con cada variable independiente:



Como resultado se obtienen las siguientes tablas y gráficos.

Estadísticos descriptivos

	Media	Desviación estándar	N
Life expectancy at birth	70,5805	8,83482	187
Log(GNIpercapita2011)	3,9581	0,52514	187
Mean years of schooling	7,9008	3,05877	187

Correlaciones

		Life expectancy at birth	Log (GNIpercapita 2011)	Mean years of schooling
Correlación de Pearson	Life expectancy at birth	1,000	0,801	0,729
	Log(GNIpercapita2011)	0,801	1,000	0,782
	Mean years of schooling	0,729	0,782	1,000
Sig. (unilateral)	Life expectancy at birth	.	0,000	0,000
	Log(GNIpercapita2011)	0,000	.	0,000
	Mean years of schooling	0,000	0,000	.
N	Life expectancy at birth	187	187	187
	Log(GNIpercapita2011)	187	187	187
	Mean years of schooling	187	187	187

Variables entradas/eliminadas[a]

Modelo	Variables entradas	Variables eliminadas	Método
1	Mean years of schooling , Log (GNIpercapita2011)[b]	.	Entrar

a. Variable dependiente: Life expectancy at birth

b. Todas las variables solicitadas introducidas.

Alcanzamos un coeficiente de determinación de 0,668 con la introducción de la de una segunda variable independiente, es decir, conjuntamente, ambas variables explican el 66,8% de la varianza con una mejora. En relación a la regresión simple con LogGNI cuando se añade Schooling se aumenta la capacidad explicativa del modelo moderadamente en un 2,7%, dada la correlación existente entre las dos variables independientes. Este valor es el que se puede derivar de la información sobre las correlaciones que se presenta en la tabla de coeficientes de la regresión. La diferencia entre el coeficiente de determinación que resulta de incluir las dos variables en la ecuación, 0,668, y el coeficiente de determinación de la variable LogGNI solamente, 0,641, es una diferencia, 0,027, que equivale al cuadrado del coeficiente de coeficiente de correlación semiparcial de la variable Schooling con la variable dependiente, 0,165. Esto es, con la ecuación de regresión donde está la variable LogGNI, la incorporación

de la variable Schooling contribuye a mejorar el ajuste del modelo en un 2,7%, el resultado de elevar al cuadrado el valor de la correlación semiparcial 0,165 y multiplicar por cien.

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	0,818[a]	0,668	0,665	5,11570

a. Predictores: (Constante), Mean years of schooling , Log (GNIpercapita2011)

b. Variable dependiente: Life expectancy at birth

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	9.702,693	2	4.851,346	185,376	0,000[b]
	Residuo	4.815,346	184	26,170		
	Total	14.518,038	186			

a. Variable dependiente: Life expectancy at birth

b. Predictores: (Constante), Mean years of schooling , Log(GNIpercapita2011)

En la ecuación de regresión vemos cómo el efecto económico es más importante que el educativo. Un incremento de una unidad de desviación de la renta per cápita determina un aumento de la esperanza de vida de 0,59 mientras que con una unidad de desviación de la escolarización, el incremento es de 0,26. Si comparamos el coeficiente de regresión obtenido de la variable de renta con el de la regresión simple vemos que se ha reducido, esta diferencia se debe al efecto compartido con la variable de Schooling, dado que ambas variables están correlacionadas.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados			95,0% intervalo de confianza para B		Correlaciones			Estadísticas de colinealidad	
		B	Error estándar	Beta	t	Sig.	Límite inferior	Límite superior	Orden cero	Parcial	Parte	Tolerancia	VIF
1	(Constante)	24,989	3,477		7,187	0,000	18,130	31,848					
	Log(GNIpercapita2011)	9,994	1,145	0,594	8,729	0,000	7,735	12,253	0,801	0,541	0,371	0,389	2,569
	Mean years of schooling	0,764	0,197	0,264	3,885	0,000	0,376	1,151	0,729	0,275	0,165	0,389	2,569

a. Variable dependiente: Life expectancy at birth

Los diagnósticos de colinealidad nos indican la existencia de una colinealidad moderada. Hemos observado una correlación alta entre las variables independientes, de 0,78, indicativa de un cierto grado de colinealidad. La (multi)colinealidad no afecta a la predicción de los valores de la variable dependiente, y el R^2 puede ser elevado, pero sí a las pruebas estadísticas ya que cuando más importante sea la colinealidad mayor será el error típico de los coeficientes de regresión, aumentando la probabilidad de no significación de estos coeficientes (aumentan los intervalos) a pesar de que las variables independientes correspondientes determinen la variable dependiente. Para determinar su importancia hemos solicitado el estadístico de Tolerancia y su inverso el VIF. La regla empírica de Kleinbaum señala que valores del VIF superiores a 10 implican problemas reales de colinealidad. Por tanto, no estaríamos en tal caso. El índice de condición inferior a 30 también indicaría este hecho.

Diagnósticos de colinealidad[a]

Modelo	Dimensión	Autovalor	Índice de condición	Proporciones de varianza (Constante)	Log (GNIpercapita 2011)	Mean years of schooling
1	1	2,924	1,000	0,00	0,00	0,01
	2	0,072	6,388	0,05	0,00	0,44
	3	0,004	26,557	0,95	1,00	0,55

a. Variable dependiente: Life expectancy at birth

La tabla de diagnósticos de los residuos para detectar casos extremos nos muestra que los casos 119 (South Africa), 144 (Equatorial Guinea) y 148 (Swaziland) son los que más se alejan de la recta de regresión. Los valores estandarizados o estudentizados superiores a ±3 que pueden afectar a los resultados de la regresión y reducir su capacidad explicativa. Para obtener una mejora en el ajuste del modelo proceder a eliminar estos casos extremos, si lo hiciéramos en una nueva regresión el coeficiente de determinación alcanzaría el valor del 72%.

Diagnósticos por casos[a]

Número del caso	Countries	Residuo estándar	Life expectancy at birth	Valor pronosticado	Residuo
119	South Africa	-3,196	56,92	73,2708	-16,35079
144	Equatorial Guinea	-3,795	53,06	72,4760	-19,41602
148	Swaziland	-3,682	49,00	67,8371	-18,83707

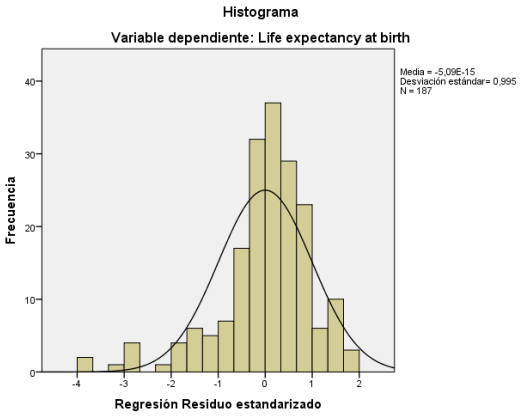
a. Variable dependiente: Life expectancy at birth

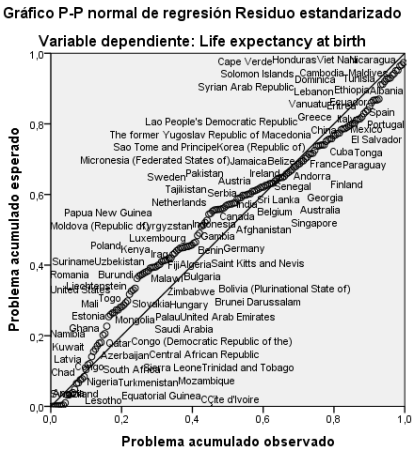
Estadísticas de residuos[a]

	Mínimo	Máximo	Media	Desviación estándar	N
Valor pronosticado	53,7916	82,6621	70,5805	7,22254	187
Valor pronosticado estándar	-2,325	1,673	0,000	1,000	187
Error estándar de valor pronosticado	0,377	1,282	0,624	0,174	187
Valor pronosticado corregido	53,9544	82,8789	70,5881	7,22636	187
Residuo	-19,41602	9,89447	0,00000	5,08812	187
Residuo estándar	-3,795	1,934	0,000	0,995	187
Residuo estudentizado	-3,864	1,943	-0,001	1,002	187
Residuo eliminado	-20,12786	9,98271	-0,00761	5,16642	187
Residuo estudentizado suprimido	-4,020	1,958	-0,004	1,015	187
Distancia de Mahal.	0,018	10,682	1,989	1,756	187
Distancia de Cook	0,000	0,182	0,005	0,015	187
Valor de influencia centrado	0,000	0,057	0,011	0,009	187

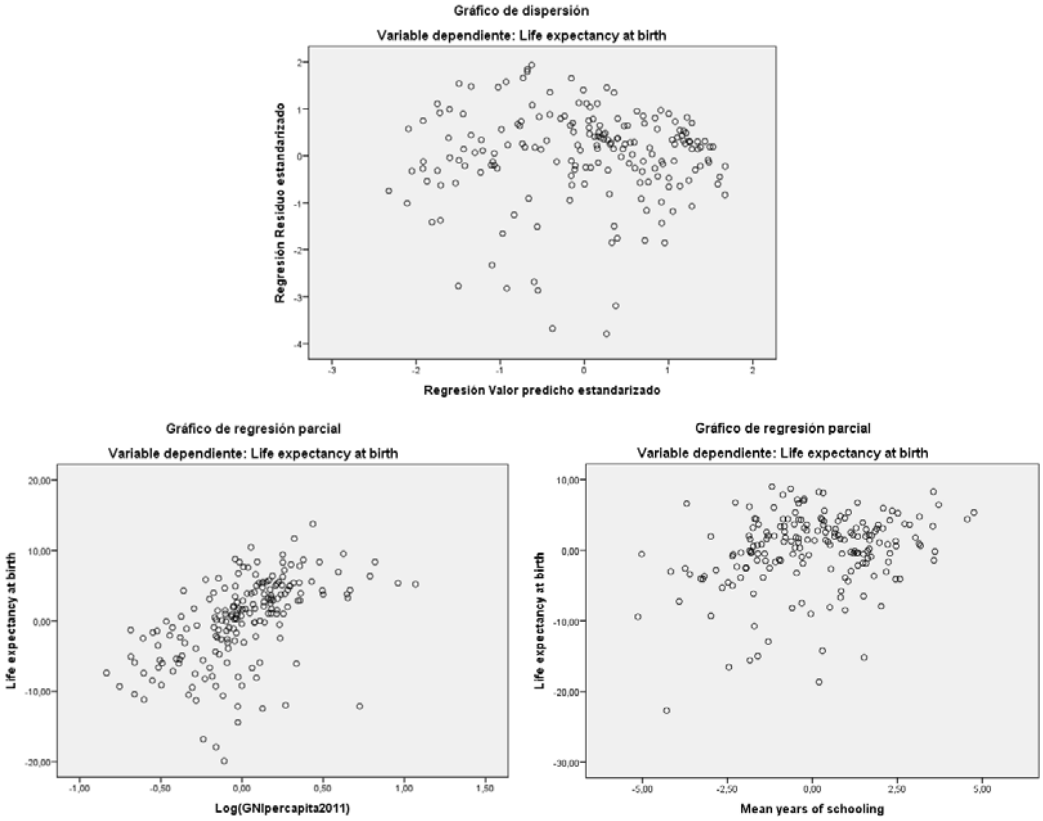
a. Variable dependiente: Life expectancy at birth

En los gráficos adjuntos se observan como anteriormente algunas pequeñas desviaciones de la normalidad.



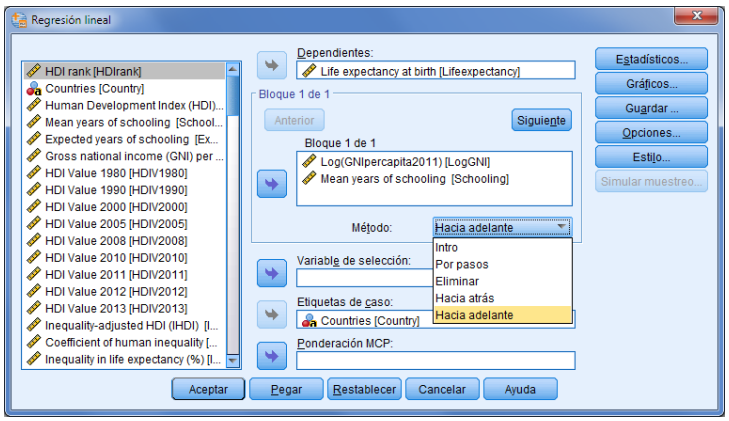


Y de nuevo se verifican las condiciones de linealidad y homoscedasticidad al no observarse una representación gráfica ningún patrón definido, se observa como la varianza de los residuos experimenta algunos comportamientos más o menos variables a lo largo de los valores pronosticados de la variable.

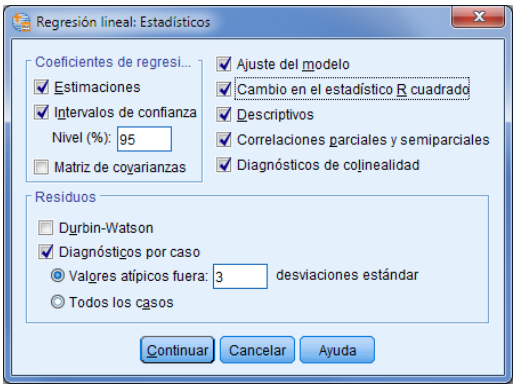


El análisis de regresión realizado se puede completar con un ejercicio de introducción secuencial de variables que nos permitiría, en cada paso, ver la regresión y los estadísticos que se generan por la introducción en cada etapa de una nueva variable independiente, según un criterio estadístico de selección que implica la mayor mejora en la capacidad explicativa del modelo.

Para ello se elegiría, por ejemplo, el método de introducción Hacia delante, entre otros que dispone el SPSS:



Y marcaríamos adicionalmente la opción Cambios en el estadístico R cuadrado para ver los efectos y la significación de la variación del coeficiente de determinación.

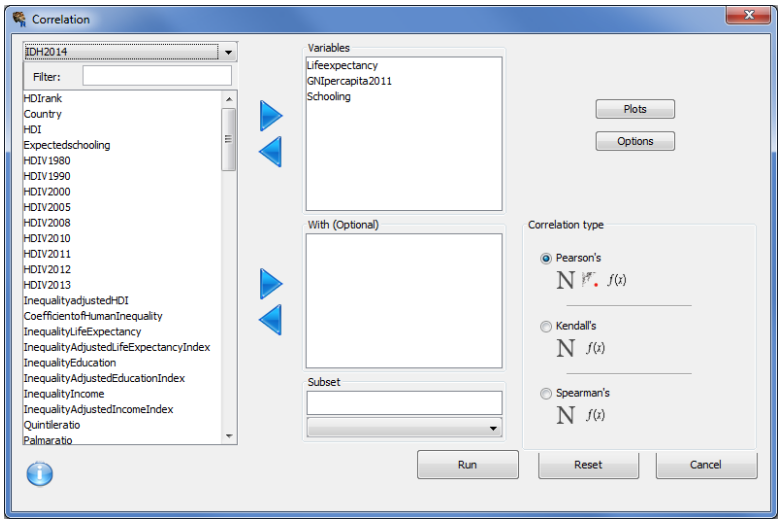


8. El análisis de regresión con R

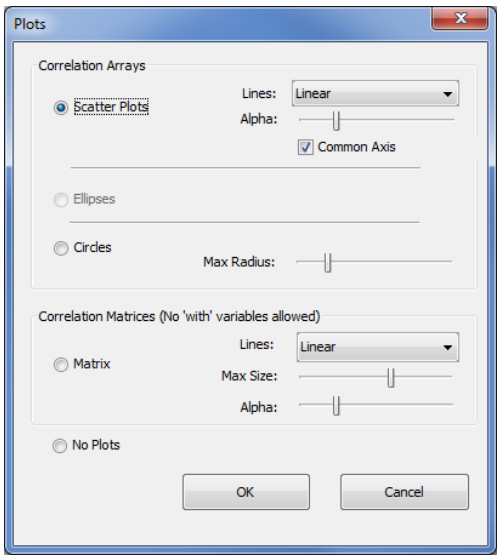
Para ejemplificar el análisis de regresión con el software R trabajaremos con la matriz de datos sobre el índice de desarrollo humano IDH2014.rda. Se trata de una matriz de datos de países con la que presentamos un ejercicio de análisis con tres variables: la variable dependiente que queremos explicar es la esperanza de vida al nacer (variable Lifeexpectancy) en función de la riqueza del país medido a través de los ingresos nacional brutos per cápita de 2011 a precio de paridad de compra (variable GNIpercapita2011 -Gross National Income (GNI) per capita 2011 PPP $-) y el nivel educativo medido a través de la media de años de escolarización (variable Schooling).

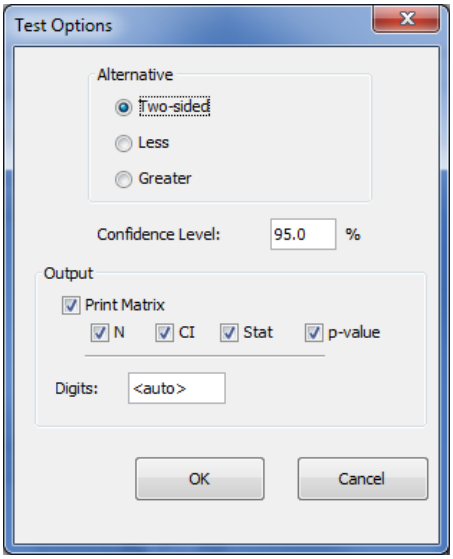
Para realizar un análisis de regresión en R se puede utilizar Deducer, en el menú Analysis se dispone de dos procedimientos: Linear Model y Generalized Linear Model. Además podemos optar por diversos comandos o funciones de R, en particular `lm` y `glm` del paquete stats.

En primer lugar analizaremos la correlación y representaremos los gráficos de dispersión con las variables consideradas. A través del menú de Deducer Analysis/Correlation se pueden analizar los coeficientes y los gráficos.



En Plots marcaremos la opción Scatter Plots para obtener los gráficos de dispersión conjuntos, mientras que mantendremos las especificaciones por defecto que se pueden elegir a través del botón de Options.





Los resultados son los siguientes:

Correlation

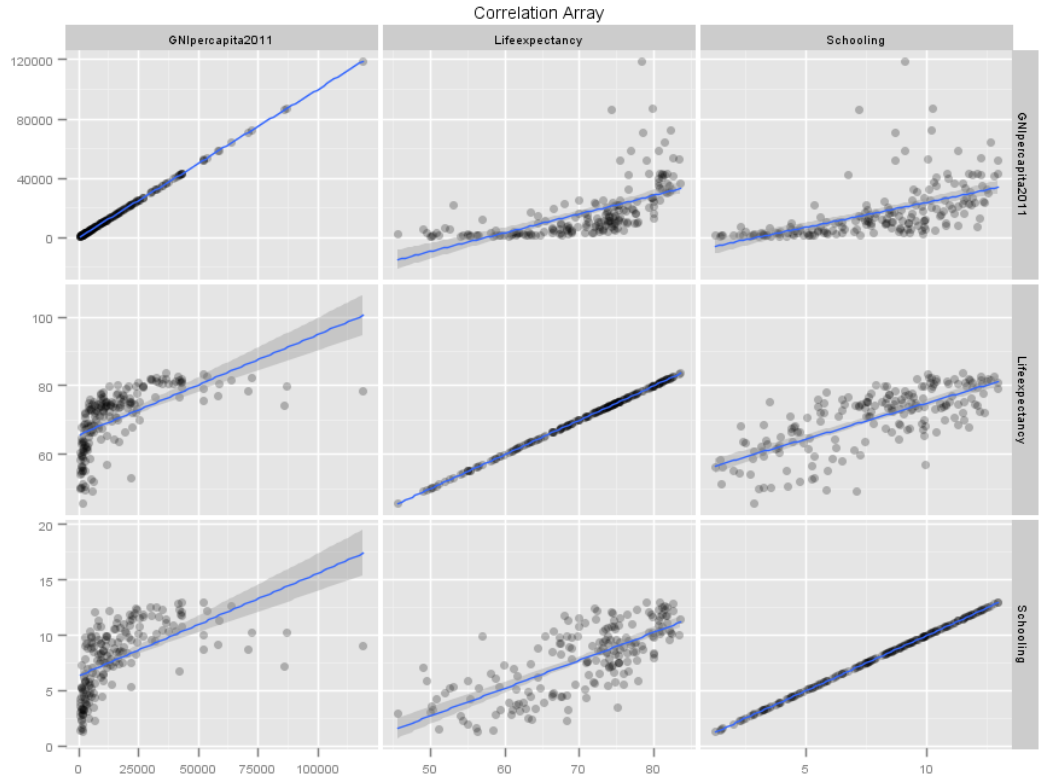
Pearson's product-moment correlation

		Lifeexpectancy	*GNIpercapita2011*	*Schooling*
Lifeexpectancy	cor		0.608	0.729
	95% CI		[0.509, 0.691]	[0.654, 0.790]
	N		189	187
	t (df)		10.470 (187)	14.472 (185)
	*p-value**		<0.001	<0.001
GNIpercapita2011	cor	0.608		0.560
	95% CI	[0.509, 0.691]		[0.452, 0.651]
	N	189		187
	t (df)	10.470 (187)		9.185 (185)
	*p-value**	<0.001		<0.001
Schooling	cor	0.729	0.560	
	95% CI	[0.654, 0.790]	[0.452, 0.651]	
	N	187	187	
	t (df)	14.472 (185)	9.185 (185)	
	*p-value**	<0.001	<0.001	

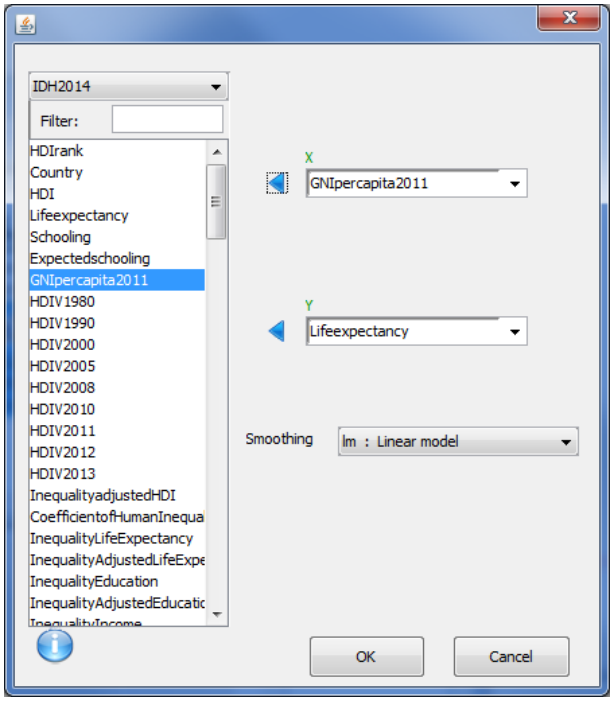
Notes:

H0: correlation = 0
*HA: two.sided

Los coeficientes obtenidos son moderadamente altos y positivos. En el caso de las relaciones donde interviene la riqueza del país, no obstante, se dibuja una nube de puntos con una tendencia no lineal, la forma es logarítmica o exponencial, según el caso, por lo que el coeficiente de correlación lineal subvalora el grado de relación existente.



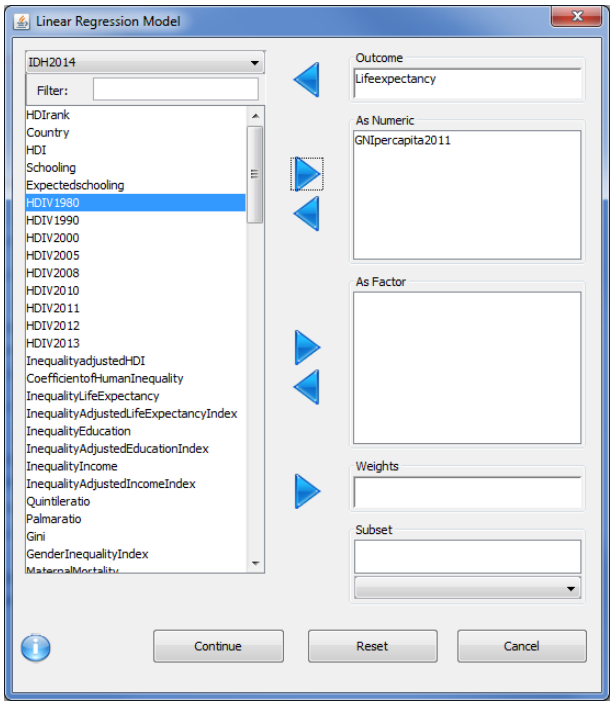
Estos gráficos se pueden obtener de forma individual especificando dos variables o bien a través de Plot Builder eligiendo el template scatter smooth, y especificando en el desplegable smooting la opción Linear model:



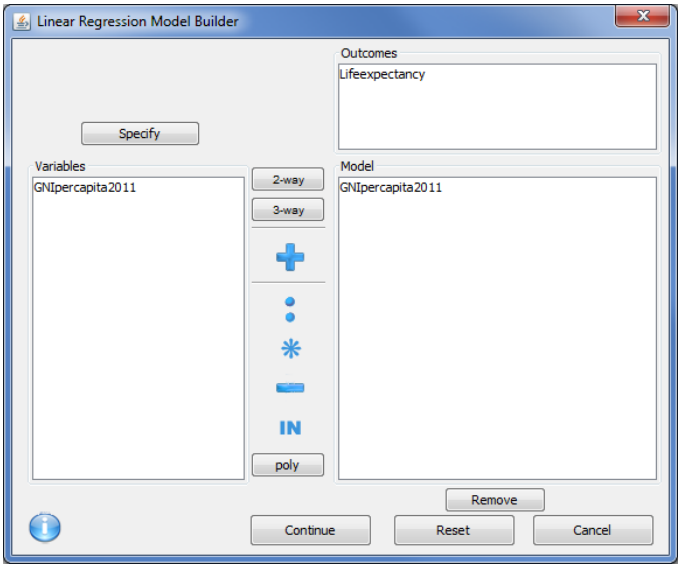
Realizaremos en primer lugar el análisis de regresión simple entre la esperanza de vida y la riqueza del país para constatar el ajuste lineal que se obtiene. Dada la forma de la

nube de puntos el ajuste lineal no será muy bueno por lo que será necesario transformar los datos para obtener un buen ajuste lineal.

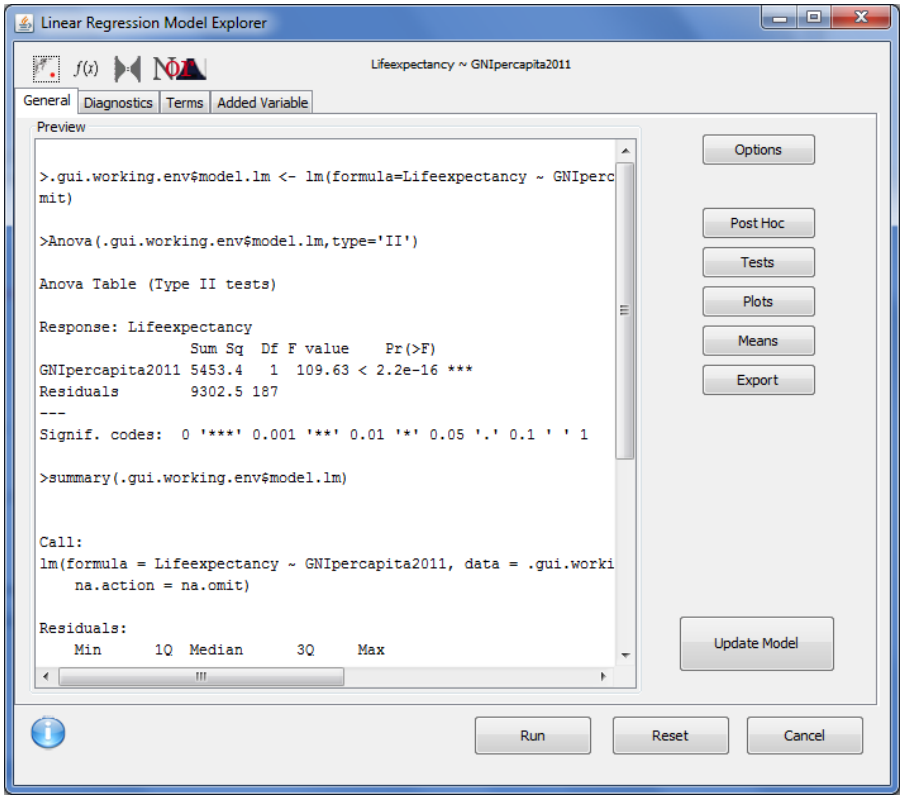
Si obtenemos la regresión a través de Linear Model del menú Analysis en Deducer:



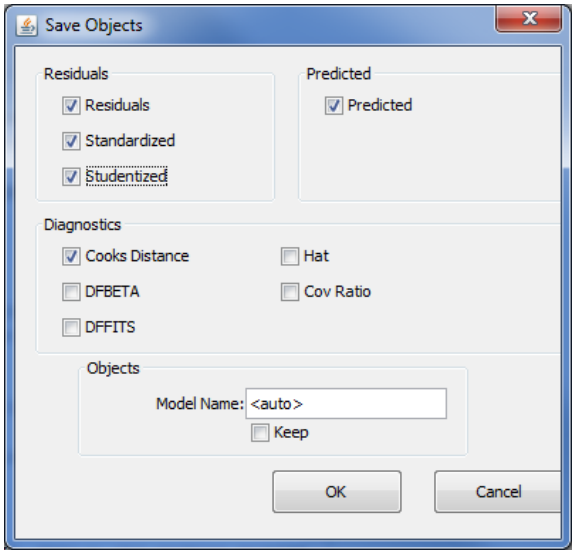
Tras especificar las variables, la dependiente en Outcome y la independiente en As Numeric aparece la especificación del modelo:



Las distintas especificaciones de ejecución del modelo aparecen en el cuadro de diálogo siguiente:



Podemos solicitar que nos genere diversas variables que se obtienen con el análisis de regresión a través del botón Export:

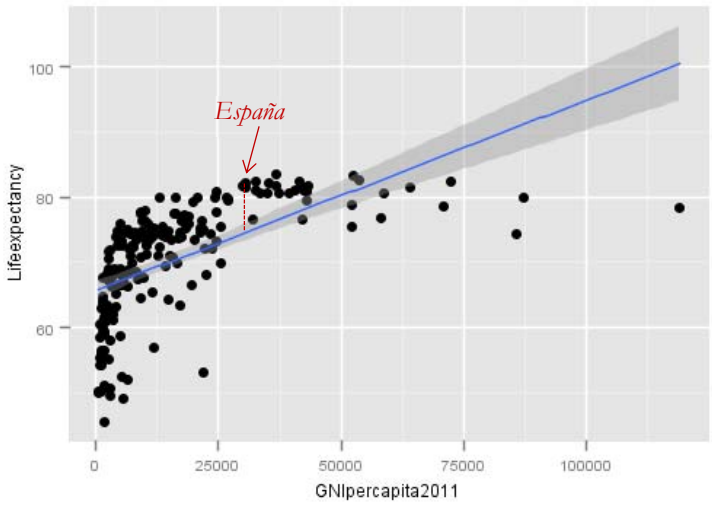


El resto de las especificaciones las dejamos por defecto. Podemos analizar los gráficos que aparecen en las pestañas Diagnostics, Terms y Added Variable.

Una vez introducidas las variables y ejecutado el procedimiento con el resto de las especificaciones por defecto se obtienen los resultados siguientes:

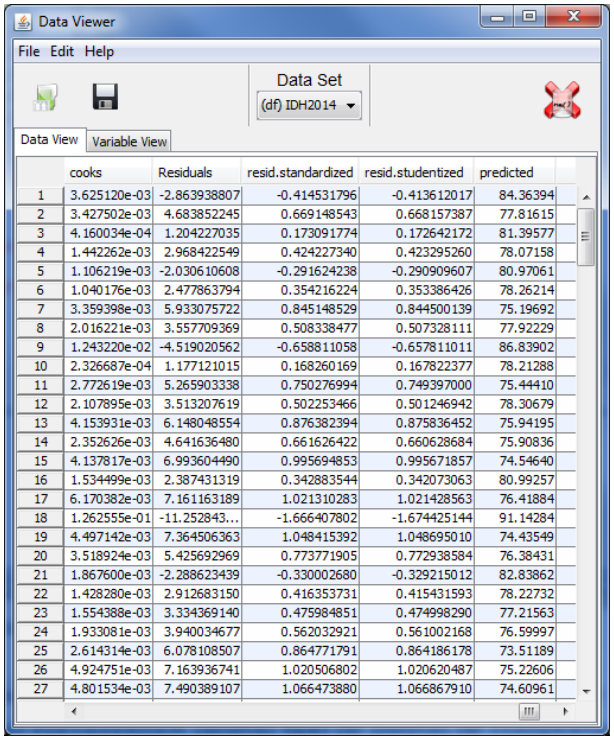
```
> model.lm <- lm(formula=Lifeexpectancy ~ GNIpercapita2011,data=IDH2014,na.action=na.omit)
> Anova(model.lm,type='II')
Anova Table (Type II tests)

Response: Lifeexpectancy
                Sum Sq  Df F value    Pr(>F)
GNIpercapita2011 5453.4   1  109.63 < 2.2e-16 ***
Residuals        9302.5 187
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model.lm)

Call:
lm(formula = Lifeexpectancy ~ GNIpercapita2011, data = IDH2014,
    na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-22.116  -4.306   2.098   5.426  10.454

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.567e+01  6.906e-01   95.10   <2e-16 ***
GNIpercapita2011 2.925e-04  2.794e-05   10.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.053 on 187 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.3696,     Adjusted R-squared: 0.3662
F-statistic: 109.6 on 1 and 187 DF,  p-value: < 2.2e-16
```

La recta de regresión es *Lifeexpectancy* = 65,67 + 0,0002925 *GNIpercapita2011* indicando que ante una variación de un dólar la esperanza de vida aumenta en 0,0002925 años, o lo que es lo mismo, por cada 10.000$ el aumento es de 3 años[17].



En el caso de España, que en el ranquin de países ocupa el lugar 27, con una renta per cápita en 2011 de 30.561,47 dólares tenía una esperanza de vida de 82,1. Según la recta de regresión el valor predicho de España es de 74,6, un error (un residuo) de 7,5 años que infraestima la esperanza de vida.

[17] El coeficiente de la constante de hecho no es interpretable: si un país tuviera 0$ de renta la esperanza de vida sería de 65,67 años, pero no se observan países en esa situación.

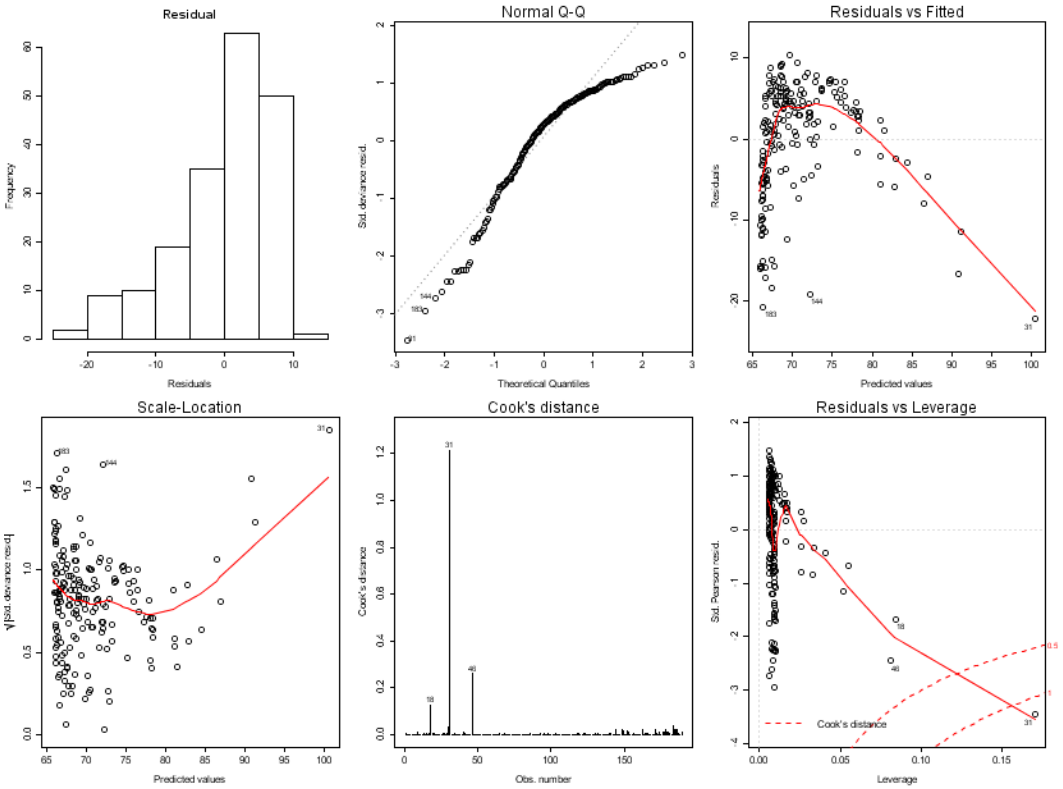


	cooks	Residuals	resid.standardized	resid.studentized	predicted
1	3.625120e-03	-2.863938807	-0.414531796	-0.413612017	84.36394
2	3.427502e-03	4.683852245	0.669148543	0.668157387	77.81615
3	4.160034e-04	1.204227035	0.173091774	0.172642172	81.39577
4	1.442262e-03	2.968422549	0.424227340	0.423295260	78.07158
5	1.106219e-03	-2.030610608	-0.291624238	-0.290909607	80.97061
6	1.040176e-03	2.477863794	0.354216224	0.353386426	78.26214
7	3.359398e-03	5.933075722	0.845148529	0.844500139	75.19692
8	2.016221e-03	3.557709369	0.508338477	0.507328111	77.92229
9	1.243220e-02	-4.519020562	-0.658811058	-0.657811011	86.83902
10	2.326687e-04	1.177121015	0.168260169	0.167822377	78.21288
11	2.772619e-03	5.265903338	0.750276994	0.749397000	75.44410
12	2.107895e-03	3.513207619	0.502253466	0.501246942	78.30679
13	4.153931e-03	6.148048554	0.876382394	0.875836452	75.94195
14	2.352626e-03	4.641636480	0.661626422	0.660628684	75.90836
15	4.137817e-03	6.993604490	0.995694853	0.995671857	74.54640
16	1.534499e-03	2.387431319	0.342883544	0.342073063	80.99257
17	6.170382e-03	7.161163189	1.021310283	1.021428563	76.41884
18	1.262555e-01	-11.252843...	-1.666407802	-1.674425144	91.14284
19	4.497142e-03	7.364506363	1.048415392	1.048695010	74.43549
20	3.518924e-03	5.425692969	0.773771905	0.772938584	76.38431
21	1.867600e-03	-2.288623439	-0.330002680	-0.329215012	82.83862
22	1.428280e-03	2.912683150	0.416353731	0.415431593	78.22732
23	1.554388e-03	3.334369140	0.475984851	0.474998290	77.21563
24	1.933081e-03	3.940034677	0.562032921	0.561002168	76.59997
25	2.614314e-03	6.078108507	0.864771791	0.864186178	73.51189
26	4.924751e-03	7.163936741	1.020506802	1.020620487	75.22606
27	4.801534e-03	7.490389107	1.066473880	1.066867910	74.60961

El coeficiente de determinación R^2 es de 0,37, es decir, que el modelo explica una 37% de la variabilidad de la variable dependiente, siendo la parte residual o no explicada del 63%. Las pruebas estadísticas del modelo, tanto la ANOVA como la del coeficiente de regresión muestran que son significativas.

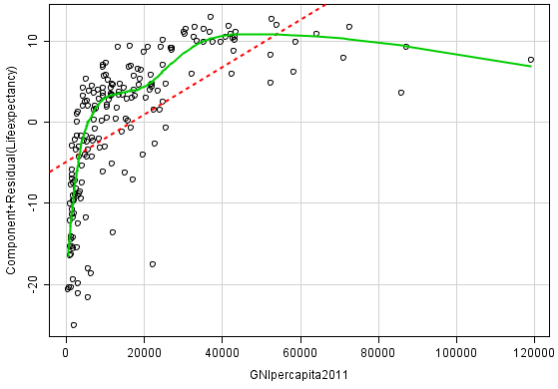
En los gráficos adjuntos se pueden analizar algunas de las diferentes condiciones de aplicación del modelo de regresión. El diagnóstico de esas condiciones se realiza a partir de un análisis de los residuos y en donde se puede constatar: su independencia, la normalidad y la presencia de casos anómalos.

El diagrama de dispersión que relaciona residuos y los valores permite validar el modelo de regresión si no encontramos ningún tipo de patrón de comportamiento. El análisis realizado presenta seis gráficos. El denominado Residual es el histograma de los residuos y donde puede observarse su comportamiento no simétrico que evidencia el alejamiento de la normalidad. Del mismo modo el gráfico Normal Q-Q dibuja un alejamiento de la línea diagonal que representa la normalidad, entre los valores bajos y sobre todo entre los altos. El diagrama de dispersión Residuals vs Fitted nos permite comprobar la constancia de la varianza si los puntos se distribuyen alrededor de una línea recta, situación que no se produce en este caso, la dispersión varía notablemente según los valores predichos. El gráfico Scale-Location evalúa la existencia de un patrón en los residuos, como este caso, donde los valores bajos de la esperanza de vida predichos muestran una distribución de los residuos muy diferente que con valores altos.

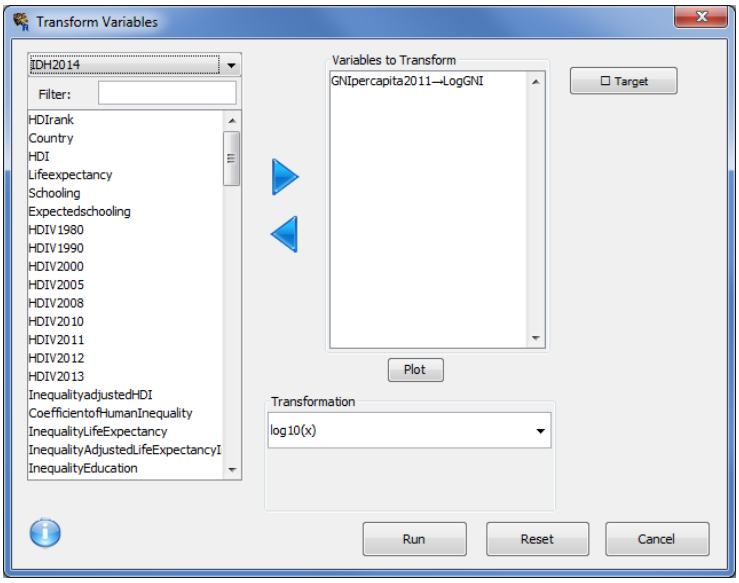


La representación de los valores de la distancia de Cook, gráfico Cook's distance, permite detectar la existencia de valores atípicos o extremos que se alejan excesivamente de la recta de regresión, son los que aparecen alejados y numerados en el gráfico, pudiendo provocar un destacable desajuste de la recta que se obtiene, reduciendo la capacidad explicativa del modelo, o incluso, aunque no necesariamente, alterar la propia recta de regresión. Este aspecto, la capacidad de influencia de un punto para alterar la recta se puede observar a través del diagrama de Residuals vs Leverage que detecta la presencia de valores influyentes que pueden alterar de forma destacada la ecuación de regresión, es decir, su posición en el espacio. En este caso vemos numerados los mismos casos extremos del gráfico anterior con un alejamiento destacado e influyente del caso 31.

Vemos por tanto que el ajuste de la recta de regresión presenta diversos problemas derivados principalmente de la forma no lineal de la relación y la disposición logarítmica de la nube de puntos:



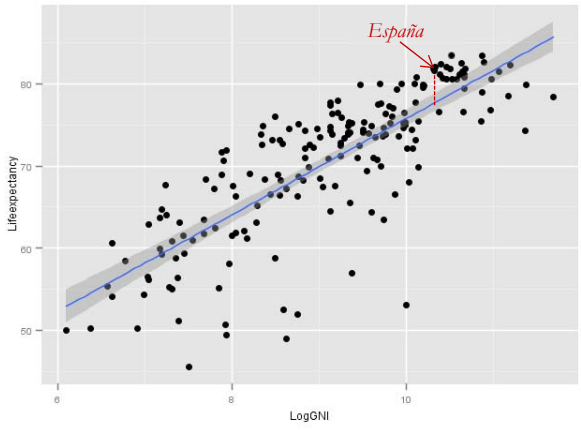
Para resolver esta situación operamos la transformación de los datos a través de Data/Transform para calcular el logaritmo en base 10 de la variable independiente creando la nueva variable LogGNI. Para ello elegiremos en el desplegable de Transformation la opción Custom/Enter Function y escribiremos log10(x):



Las instrucciones de Deducer que se ejecutan son las siguientes:

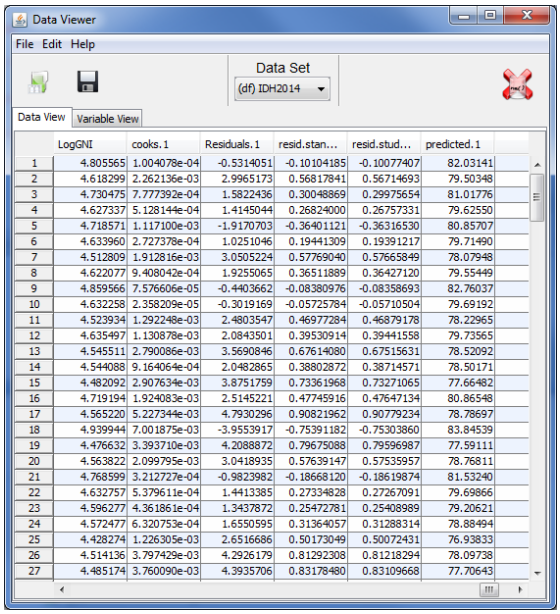
```
> trans.function <- function(x) log10(x)
> IDH2014[['LogGNI']] <- trans.function(IDH2014[['GNIpercapita2011']])
```

Con esta transformación volvemos a realizar la regresión. Vemos en primer lugar el cambio producido en el diagrama de dispersión entre Lifeexpectancy y GNIpercapita2011 al aplicar el logaritmo en la variable independiente con la nueva forma alineada de la nube de puntos.



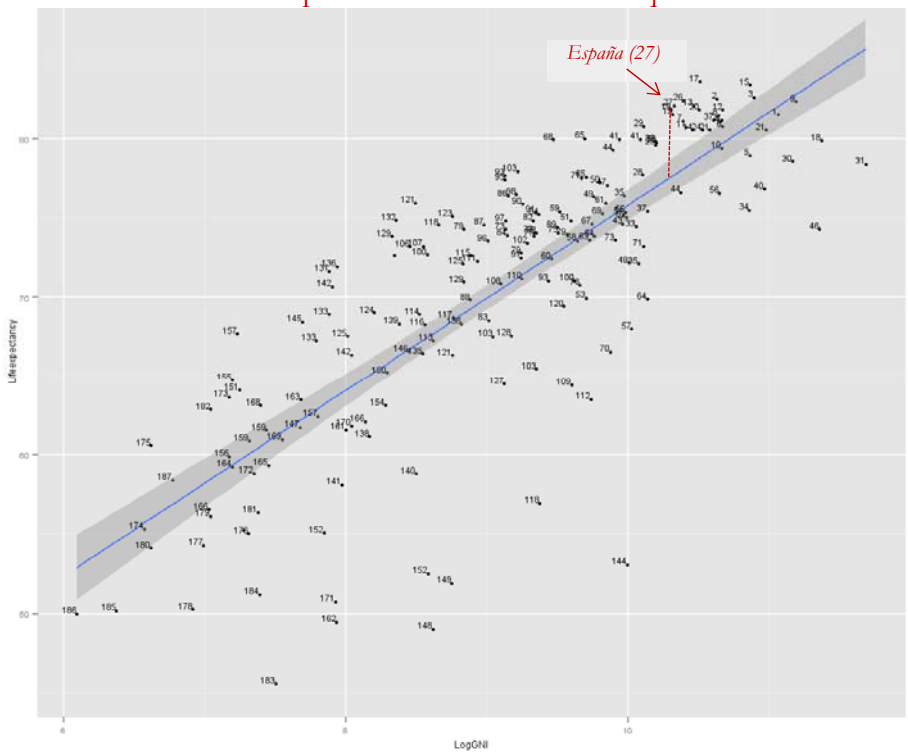
La nueva recta de regresión es ahora *Lifeexpectancy* = 17,161 + 5,863 *LogGNI*, es decir, que ante una variación de una unidad del logaritmo de GNI, es decir, cada 10$ (la inversa del logaritmo en base 10: si $log_{10}(x)=1$ entonces $x=10^1$), la esperanza de vida aumenta en 5,863 años. Y también, por cada 1.000$, cuyo logaritmo es 3, el aumento es de 5,863 × 3 = 17,589 años.

En esta ocasión el valor predicho para España es de 77,71, un valor que se ajusta más al valor observado. Se comprueba que este resultado se obtiene mediante el cálculo 17,161 + 5,863 × log(30.561,47), generando un residuo de 4,4 años, la diferencia entre el valor observado 82,1 y el predicho 77,7.



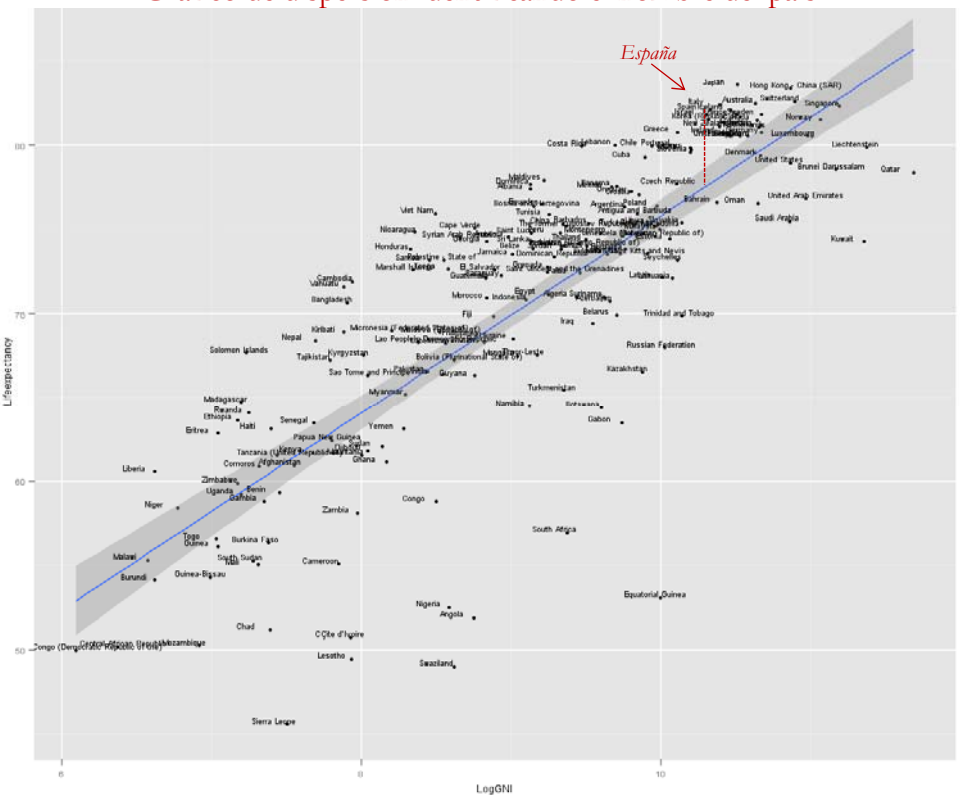
	LogGNI	cooks.1	Residuals.1	resid.stan...	resid.stud...	predicted.1
1	4.805565	1.004078e-04	-0.5314051	-0.10104185	-0.10077407	82.03141
2	4.618299	2.262136e-03	2.9965173	0.56817841	0.5671469	79.50348
3	4.730475	7.777392e-04	1.5822436	0.30048869	0.29975654	81.01776
4	4.627337	5.128144e-04	1.4145044	0.26824000	0.26757331	79.62550
5	4.718571	1.117100e-03	-1.9170703	-0.36401121	-0.36316530	80.85707
6	4.633960	2.727378e-04	1.0251046	0.19441309	0.19391217	79.71490
7	4.512809	1.912816e-03	3.0505224	0.57769040	0.57665849	78.07948
8	4.622077	9.408042e-04	1.9255065	0.36511889	0.36427120	79.55449
9	4.859566	7.576606e-05	-0.4403662	-0.08380976	-0.08358693	82.76037
10	4.632258	2.358209e-05	-0.3019169	-0.05725784	-0.05710504	79.69192
11	4.523934	1.292248e-03	2.4803547	0.46977284	0.46879178	78.22965
12	4.635497	1.130878e-03	2.0843501	0.39530914	0.39441558	79.73565
13	4.545511	2.790086e-03	3.5690846	0.67614080	0.67515631	78.52092
14	4.544088	9.164064e-04	2.0482865	0.38802872	0.38714571	78.50171
15	4.482092	2.907634e-03	3.8751759	0.73361968	0.73271065	77.66482
16	4.719194	1.924083e-03	2.5145221	0.47745916	0.47647134	80.86548
17	4.565220	5.227344e-03	4.7930296	0.90821962	0.90779234	78.78697
18	4.939944	7.001875e-03	-3.9553917	-0.75391182	-0.75303860	83.84539
19	4.476632	3.393710e-03	4.2088872	0.79675088	0.79596987	77.59111
20	4.563822	2.099795e-03	3.0418935	0.57639147	0.57535957	78.76811
21	4.768599	3.212727e-04	-0.9823982	-0.18668120	-0.18619874	81.53240
22	4.632757	5.379611e-04	1.4413385	0.27334828	0.27267091	79.69866
23	4.596277	4.361861e-04	1.3437872	0.25472781	0.25408989	79.20621
24	4.572477	6.320753e-04	1.6550595	0.31364057	0.31288314	78.88494
25	4.428274	1.226305e-03	2.6516686	0.50173049	0.50072431	76.93833
26	4.514136	3.797429e-03	4.2926179	0.81292308	0.81218294	78.09738
27	4.485174	3.760090e-03	4.3935706	0.83178480	0.83109668	77.70643

Gráfico de dispersión identificando el ranquin IDH[18]



[18] La identificación se obtiene en Plot Builder añadiendo Text desde la pestaña de Geometric Elements.y eligiendo la variable de identificación: HDIrank o Country.

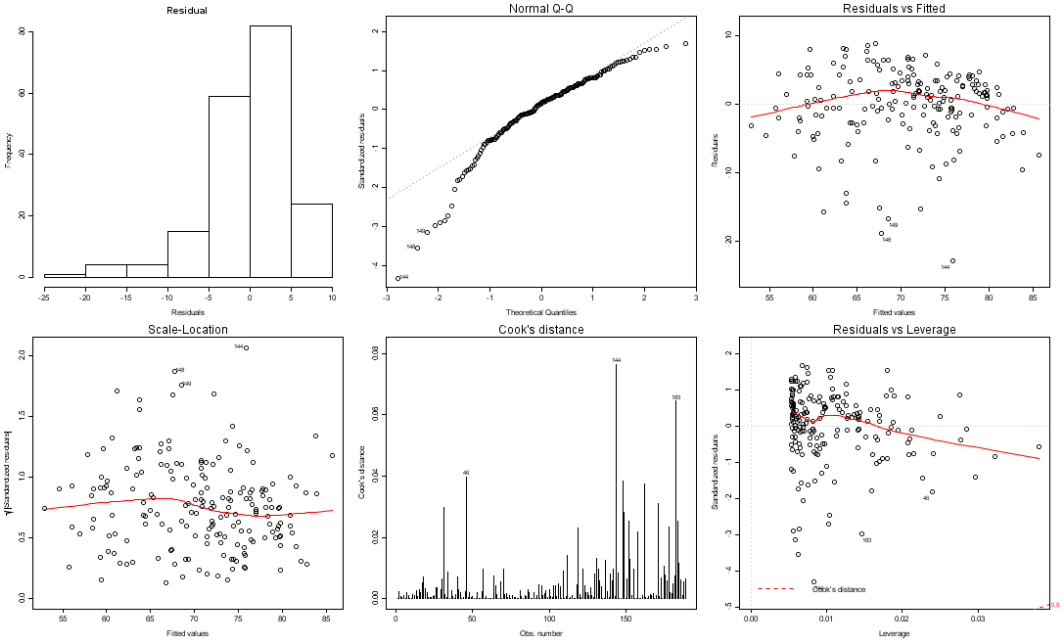
Gráfico de dispersión identificando el nombre del país

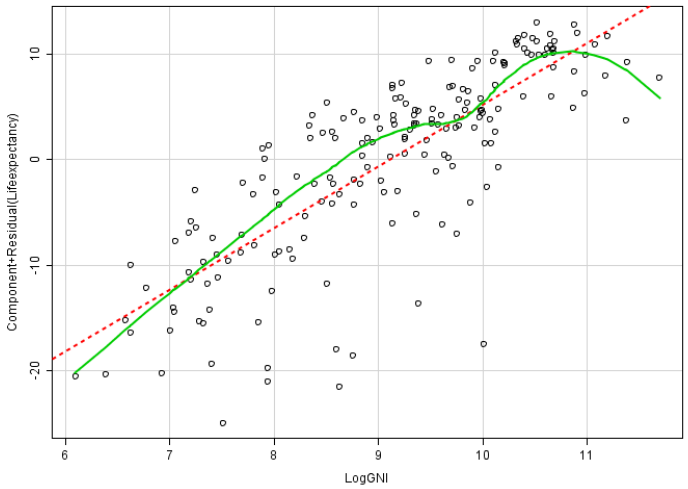



```
> model.lm <- lm(formula=Lifeexpectancy ~ LogGNI,data=IDH2014,na.action=na.omit)
> Anova(model.lm,type='II')
Anova Table (Type II tests)

Response: Lifeexpectancy
          Sum Sq  Df F value    Pr(>F)
LogGNI    9481.9   1  336.19 < 2.2e-16 ***
Residuals 5274.1 187
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model.lm)

Call:
lm(formula = Lifeexpectancy ~ LogGNI, data = IDH2014, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-22.712  -2.326   1.025   3.389   8.984

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.1608     2.9352   5.847  2.2e-08 ***
LogGNI        5.8626     0.3197  18.335  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.311 on 187 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared: 0.6426,     Adjusted R-squared: 0.6407
F-statistic: 336.2 on 1 and 187 DF,  p-value: < 2.2e-16
```

La mejora en la predicción de la recta de regresión obedece al mejor ajuste obtenido. La bondad de dicho ajuste se expresa en un R^2 que ahora es del 64%, siendo la parte residual del 36%. La mejora respecto al modelo anterior ha sido del 27%.

Los gráficos que analizan las condiciones de aplicación siguen mostrando la ausencia de normalidad, que no es problemática dado el número suficiente de casos, y la presencia de casos extremos según muestra la distancia de Cook. No obstante mejoró la disposición de los valores residuales respecto de los ajustados generando una línea que se aproxima a una recta que permite concluir la ausencia de patrones diferenciados y la idoneidad del análisis de regresión.

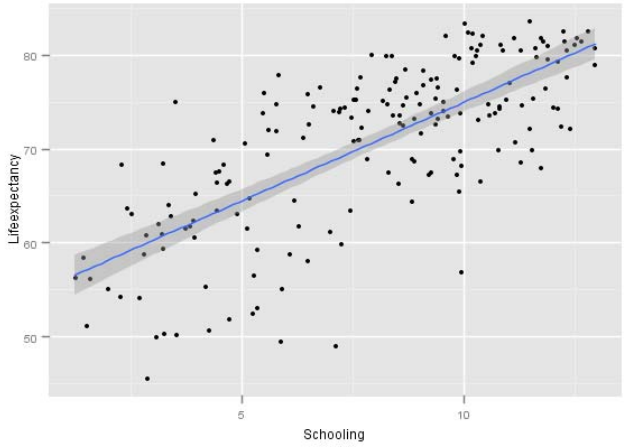




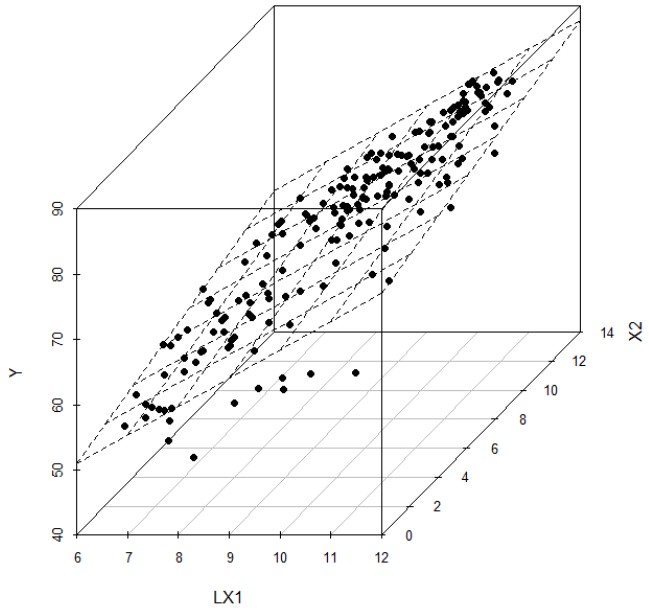
El modelo de regresión simple que acabamos de obtener mejoró como resultado de constatar que la relación funcional entre las variables era logarítmica, la función inversa de la exponencial, y su transformación nos proporcionó una mejora en el coeficiente de determinación muy notable. Aun así queda una parte no explicada que genera

errores de predicción. La existencia de los residuos se debe a distintas razones que expresa nuestro modelo. Junto a la adecuada relación funcional puede suceder que tengamos errores de medición de nuestras variables. Además cabe contemplar los efectos de la presencia de ciertos casos extremos e influyentes que alteran tanto la bondad del ajuste como la ecuación de regresión. Y también es posible que nuestro modelo se incompleto y que hayamos omitido variables relevantes que actúan como variables independientes explicativas que deberíamos incorporar al modelo.

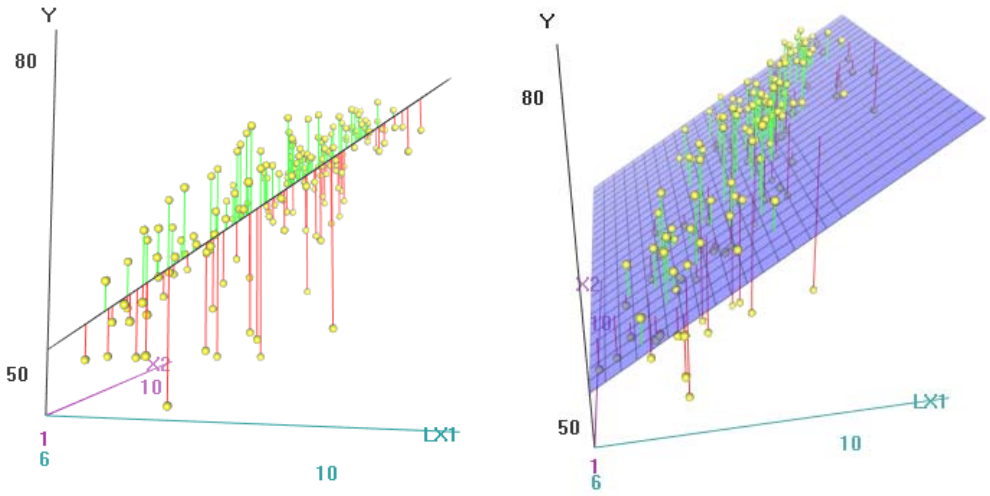
Trataremos estos dos últimos aspectos. Empezaremos considerando un modelo ampliado de regresión múltiple incorporando una segunda variable independiente para contemplar el efecto del factor educativo, la variable Schooling, la media de años de escolarización. El gráfico de dispersión de la esperanza de vida con la nueva variable evidencia una cierta tendencia lineal como vimos anteriormente.



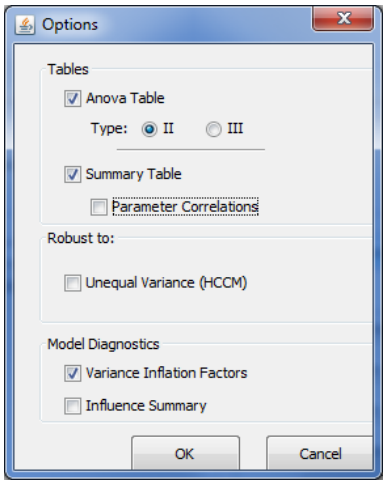
Si representamos conjuntamente las tres variables en un gráfico de tres dimensiones ajustando el plano de regresión podemos obtener diferentes representaciones[19]:



[19] Ver el programa de instrucciones AR-Esperanza.R.



En el cuadro de diálogo de Linear Model de Deducer colocaremos las segunda variable numérica en el recuadro de la primera ventana y en la segunda ventana la añadiremos con el símbolo de + en el recuadro Model. Con la introducción de una tercera variable debemos comprobar la existencia de un problema de colinealidad. Para obtener un diagnóstico clicaremos en el botón de Options y marcaremos la especificación de Variance Inflation Factors (VIF):



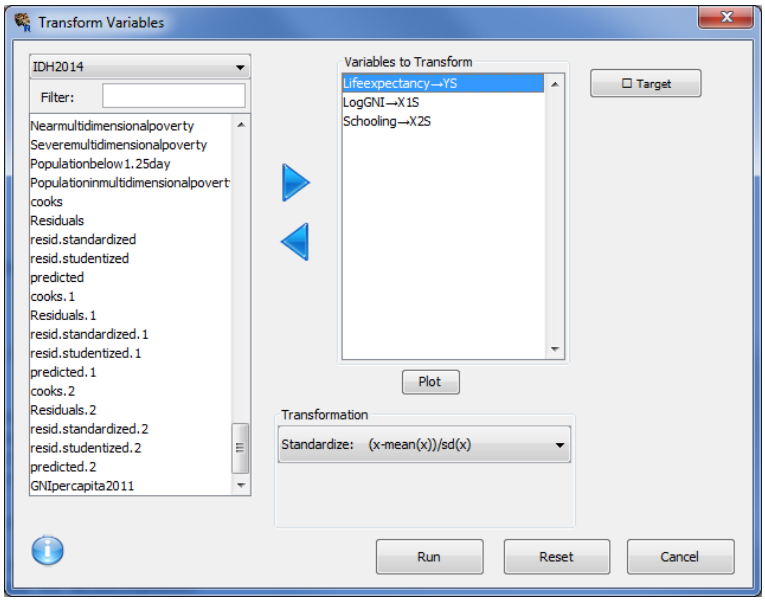
Los resultados que se obtienen permiten concluir, por un lado, que la capacidad explicativa mejora ligeramente al alcanzar el coeficiente de determinación el valor 0,67 (0,66 corregido por el número de variables y de casos) frente al 0,64 anterior. Por otro, que ambas variables contribuyen positivamente.

En un modelo de regresión múltiple cada coeficiente (parcial) se interpreta como el efecto de la variable independiente en cuestión cuando el resto permanecen constantes. Para hacer comparables los efectos y determinar la importancia relativa de cada uno de ellos tenemos que considerar los coeficientes estandarizados, así no tiene en cuenta la unidad de medida particular de cada una al expresarse en unidades de desviación típica. Los resultados obtenidos con R no presentan los datos estandarizados por lo que será necesario ejecutar la regresión con las variables estandarizadas. Adicionalmente podemos valorar la importancia relativa como porcentaje de la

varianza explicada. Ahora lo veremos. No obstante podemos observar que el coeficiente no estandarizado de LogGNI (9,99) se ha modificado respecto del modelo de regresión simple (5,86). Esta diferencia se debe al efecto compartido con la variable de Schooling, dado que ambas variables están correlacionadas (el problema de la colinealidad).

```
> model.lm <- lm(formula=Lifeexpectancy ~ LogGNI + Schooling,data=IDH2014,na.action=na.omit)
> Anova(model.lm,type='II')
Anova Table (Type II tests)

Response: Lifeexpectancy
           Sum Sq  Df F value    Pr(>F)
LogGNI     1994.1   1  76.197 1.533e-15 ***
Schooling   395.0   1  15.093 0.0001426 ***
Residuals  4815.3 184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model.lm)

Call:
lm(formula = Lifeexpectancy ~ LogGNI + Schooling, data = IDH2014,
    na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4160  -1.6920   0.9123   3.1460   9.8945

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.9889     3.4767   7.187 1.60e-11 ***
LogGNI        9.9943     1.1449   8.729 1.53e-15 ***
Schooling     0.7637     0.1966   3.885 0.000143 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.116 on 184 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.6683,     Adjusted R-squared: 0.6647
F-statistic: 185.4 on 2 and 184 DF,  p-value: < 2.2e-16
```

A través del menú Data / Transform procedemos a estandarizar las tres variables y les asignamos los nombres YS, X1S y X2S.



Ejecutando de nuevo la regresión obtenemos estos nuevos resultados:

```
> model.lm <- lm(formula=YS ~ X1S + X2S,data=IDH2014,na.action=na.omit)
> Anova(model.lm,type='II')
Anova Table (Type II tests)

Response: YS
          Sum Sq  Df F value    Pr(>F)
X1S       25.269   1  76.197 1.533e-15 ***
X2S        5.005   1  15.093 0.0001426 ***
Residuals 61.019 184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(model.lm)

Call:
lm(formula = YS ~ X1S + X2S, data = IDH2014, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1856 -0.1905  0.1027  0.3541  1.1138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.009169   0.042122   0.218 0.827931
X1S         0.590638   0.067663   8.729 1.53e-15 ***
X2S         0.262949   0.067683   3.885 0.000143 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5759 on 184 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.6683,     Adjusted R-squared: 0.6647
F-statistic: 185.4 on 2 and 184 DF,  p-value: < 2.2e-16
```

Vemos cómo el efecto económico es más importante que el educativo. Un incremento de una unidad de desviación de la renta per cápita, aquí X1S, determina un aumento de la esperanza de vida de 0,59 mientras que con una unidad de desviación de la escolarización, aquí X2S, el incremento es de 0,26.

Con la función `calc.relimp` del paquete relaimpo podemos valorar el porcentaje de varianza explicada por cada variable en relación al 66,8% obtenido del R^2 anteriormente[20].

```
Response variable: YS
Total response variance: 0.9890842
Analysis based on 187 observations

2 Regressors:
LX1S X2S
Proportion of variance explained by model: 66.83%
Metrics are not normalized (rela=FALSE).

Relative importance metrics:

           lmg
LX1S 0.3892332
X2S  0.2790866

Average coefficients for different model sizes:

            1X       2Xs
LX1S 0.7960814 0.5906381
X2S  0.7246860 0.2629487
```

[20] Ver el programa de instrucciones de R: ARE-Esperanza.R.

La variable de renta representa un 39% (el 58% del total) y la variable educativa un 28% (el 42% del total).

Hemos observado una correlación alta entre las variables independientes, de 0,78, indicativa de un cierto grado de colinealidad. La (multi)colinealidad no afecta a la predicción de los valores de la variable dependiente, y el R^2 puede ser elevado, pero sí a las pruebas estadísticas ya que cuando más importante sea la colinealidad mayor será el error típico de los coeficientes de regresión, aumentando la probabilidad de no significación de estos coeficientes (aumentan los intervalos) a pesar de que las variables independientes correspondientes determinen la variable dependiente. Para determinar su importancia hemos solicitado el estadístico VIF para su diagnóstico:

```
> vif(model.lm)
   LogGNI Schooling
 2.569332  2.569332
```

La regla empírica de Kleinbaum señala que valores del VIF superiores a 10 implican problemas reales de colinealidad. Por tanto, no estaríamos en tal caso.

Por último podríamos proceder a eliminar los casos extremos e influyentes para ver si se obtiene una mejora en el ajuste del modelo. Si miramos los residuos estudentizados podemos detectar valores superiores a ±3 que pueden afectar a los resultados de la regresión y reducir su capacidad explicativa. Los casos 144 (Equatorial Guinea) y 148 (Swaziland) son los que más se alejan de la recta de regresión, si los eliminamos del análisis el coeficiente de determinación alcanza el 70%.

9. Bibliografía

Achen, C. H. (1982). *Interpreting and Using Regression*. Beverly Hills: Sage Publications.

Allison, P. D. (1984). *Event History Analysis: Regression for Longitudinal Event Data*. Beverly Hills: Sage Publications.

Belsey, D. A. (1991). *Conditioning Diagnostics: Collinearity and weak data in regression*. New York: John Wiley and Sons.

Belsey, D. A.; Kuh, E.; Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influencial Data and Sources of Collinearity*. New York: Wiley.

Berry, W. D. (1993). *Understanding Regression Assumptions*. Newbury Park, California: Sage Publications.

Berry, W. D.; Feldman, S. (1984). *Multiple Regression in Practice*. Beverly Hills: Sage Publications.

Bisquerra, R. (1987). *Introducción a la estadística aplicada a la investigación educativa. Un enfoque informático con los paquetes BMDP y SPSSX*. Barcelona: Promociones y Publicaciones Universitarias. Cap. 9.

Bisquerra, R. (1989). *Introducción conceptual al análisis multivariable*. Barcelona: Promociones y Publicaciones Universitarias. Cap. 8.

Blalock, H. M. Jr. (1978). *Estadística Social*. 2a. edició. México: Fondo de Cultura Económica. Cap. 13, 17 a 20.

Bosque, J.; Moreno, A. (1994). *Práctica de análisis exploratorio y multivariante de datos*. Barcelona: Oikos-Tau. Cap. 2 y 3.

Bryman, A.; Cramer, D. (1990). *Quantitative Data Analysis fpr Social Scientist*. London: Routledge.

Castro, C. (2010). Análisis de regresión para sociólogos con aplicaciones en R. Santiago de Chile. http://www.bubok.es/libros/175431/Analisis-de-regresion-para-sociologos

Chatterjee, S.; Price, B. (1977). *Regression Analysis by Example*. New York: John Wiley.

Cohen, J.; Cohen, P.; West, S. G.; Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Cook, R. D. (1979). Influential Observations in Regression Analysis. *Journal of the American Statistical Association*, 74, 169-174.

Daudin, J. J. (1980). Régression qualitative: choix de l'espace prédicteur. En *Data analysis and Informatics*, editado por E. Diday et al. Amsterdam: North-Holland, 324-345.

Doebreske, J.-J.; Tassi, Ph. (1990). *Histoire de la Statistique*. Paris: Presses Universitaires de France.

Domenech, J. M.; Riba, M. D. (1983). *Introducción al modelo lineal: regresión múltiple*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona.

Domenech, J. M.; Riba, M. D. (1981). *Una síntesis de los métodos estadísticos bivariantes*. Barcelona: Herder.

Domenech, J. M.; Riba, M. D. (1985). *Métodos estadísticos. Modelo lineal de regresión*. Barcelona: Herder.

Draper, N. R.; Smith, H. (1998). *Applied Regression Analysis*. New York: Wiley. Third Edition

Edwards, A. L. (1984). *An introduction to linear regression and correlation*. New York: Freeman.

Etxeberría, J. (1999). *Regresión Múltiple*. Madrid. La Muralla.

Escribà, A. (2006). Estructura familiar, estatus ocupacional y movilidad social intrageneracional en España. *Revista Internacional de Sociología*, LXIV, 45, septiembre-diciembre, 145-170.

Fachelli, S.; Planas J. (2016). Evolución de la inserción profesional de los universitarios: de la expansión a la crisis duradera". En AQU Catalunya, *Equitat en l'accés i en la inserció professional dels graduats i graduades universitaris*. Barcelona: AQU. http://www.aqu.cat/doc/doc_10339347_1.pdf

Faraway, J. J. (2002). *Practical Regression and Anova using R*. CRAN R Project. https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf

Faraway, J. J. (2015). *Linear Models with R*. 2nd edition. Boca Raton, FL: Taylor and Francis.

Farré, M. (2006). La recta de mínims quadrats. Departament de Matemàtiques, Universitat Autònoma de Barcelona. *MATerials MATemàtics*, 10. http://www.mat.uab.cat/matmat

Fox, J. (1991). *Regression Diagnostics*. Newbury Park, California: Sage Publications.

Fox, J. (2002). *An R and S Plus Companion to Applied Regression*. Thousand Oaks, California: Sage Publications.

Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models*. Newbury Park, California: Sage Publications. Third Edition. http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/index.html

García Ferrando, M. (1987) *Socioestadística. Introducción a la estadística en sociología*. 2a edición amp. Madrid: Alianza. Alianza Universidad Textos, 96. Cap. 9 y 14.

Glass, G. V.; Stanley, J. C. (1970). *Métodos Estadísticos Aplicados a las Ciencias Sociales*. Méxixo: Prentice-Hall.

Goodman, L. A. (1972). A modified multiple regression approach to the analysis of dichotomous variables. *American Sociological Review, 37*, 28-46.

Guillén, M. F. (1992). *Análisis de regresión múltiple*. Madrid: Centro de Investigaciones Sociológicas.

Harrell, F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, California: Sage Publications.

Jaccard, J. J.; Turrisi, R. (2003). *Interaction Effects in Multiple Regression*. Thousand Oaks, California: Sage Publications.

Jovell, A. J. (1995). *Análisis de regresión logística*. Madrid: Centro de Investigaciones Sociológicas.

Kerlinger, F. N.; Pedhazur, E. M. (1973). *Multiple regression analysis in behavioral research*. New York: Holt, Rinehart & Winston.

Kleinbaum, D. G.; Kupper, L. L. (1978). *Applied Regression Analysis and Other Multivariate Methods*. North Scituate (Massachusets): Duxburg Press.

Kutner, M. H. et al. (2005). *Applied Linear Statistical Models*. Boston: McGraw-Hill Irvin. 5ª edición.

Lebart, L.; Morineau, A.; Fénelon, J.-P. (1985). *Tratamiento estadístico de datos. Métodos y Programas*. Barcelona: Marcombo.

Lévy, J.-P., Varela, J.. (2003). *Análisis Multivariable para las Ciencias Sociales*. Madrid: Pearson Prentice Hall. Cap. 7.

Lewis-Beck, M. S. (1980). *Applied Regression: An Introduction*. Beverly Hills: Sage Publications.

López González, E. (1998). Tratamiento de la colinealidad en regresión múltiple. *Psicothema,* 10, 2, 491-507.

Mendenhall, W.; Sincich, T. (2003). *A Second Course in Statistics : Regression Analysis*. Upper Saddle River, New Jersey: Pearson Education.

Mosteler, F.; Tukey, J. W. (1977). *Data Analysis and Regression*. New York: Addison-Wesley.

Myers, R. H. (1986). *Classical and modern regression with applications*. Boston (Massachussets): Duxbury Press.

Newbold, P.; Bos, T. (1985). *Stochastic Parameter Regression Models*. Beverly Hills: Sage Publications.

Ostrom, C. W. Jr. (1978). *Time-series analysis: regression techniques*. Beverly Hills: Sage Publications.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253-318.

Palacio, J. I.; Simón, H. J. (2006). Segregación laboral y diferencias salariales por razón de sexo en España. *Estadística Española*, 48, 163, 493-524.

Pardo, A.; Ruíz, M. A. (2001). Análisis de regresión lineal. El procedimiento Regresión Lineal. En A. Pardo y M. A. Ruíz, *SPSS 10.0. Guía para el análisis de datos*. Madrid: Hispanoportuguesa SPSS.

Pardo, A.; Ruíz, M. A. (2015). *Análisis de datos en ciencias sociales y de la salud*. Madrid: Síntesis.

Peró, M. et al. (2012). *Estadística aplicada a las ciencias sociales mediante R y R-Commander*. Madrid: Ibergarceta Publicaciones.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research. Explanation and prediction*. New York: Holt, Rinehart and Winston.

Pérez López, C. (2004). Técnicas de Análisis Multivariante de Datos: aplicaciones con SPSS. Madrid: Pearson.

Riera, P. (2012). La abstención diferencial en la España de las autonomías. Pautas significativas y mecanismos explicativos. *Revista Internacional de Sociología*, 70, 3, 615–642. doi:10.3989/ris.2010.10.07

Sánchez Carrión, J. J. (1999). *Manual de análisis de datos*. Madrid: Alianza. Alianza Universidad Textos, 150. Cap. 8 y 9.

Schroeder, L. D.; Sloquist, D. L.; Sthepan, P. E. (1986). *Understanding Regression Analysis: An introductory Guide*. Bevely Hills: Sage Publications.

Sierra Bravo, R. (1994). *Análisis estadístico multivariable. Teoría y ejercicios*. Madrid: Paraninfo. Cap. 1y 2.

Tabachnick, B. G.; Fidell, L. S. (1989). *Using Multivariate Statistics*. 2a. edició. New York: Harper Collins.

Tusell, F. (2011). *Análisis de Regresión. Introducción Teórica y Práctica basada en R*. www.et.bs.ehu.es/~etptupaf/nuevo/ficheros/estad3/nreg1.pdf

Vallès, J. M.; Liñeira, R. (2014). Abstención diferencial en Cataluña y en la Comunidad de Madrid: explicación sociopolítica de un fenómeno urbano. *Revista Española de Investigaciones Sociológicas,* 146, 69-92. http://dx.doi.org/10.5477/cis/reis.146.69

Wonnacott, Th. H.; Wonnacott, R. J. (1981). *Fundamentos de Estadística para Administración y Economía*. México: Limusa. Cap. 11 a 14.