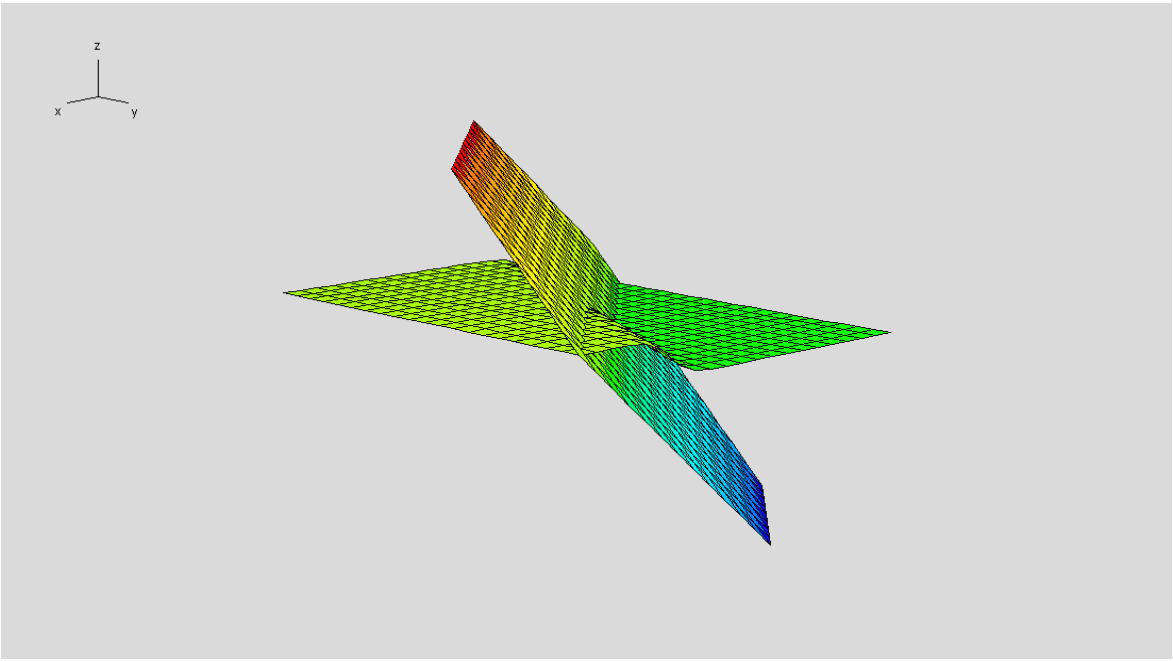
Regresión lineal múltiple



J. M. Rojo Abuín

Instituto de Economía y Geografía

Madrid, II-2007

Índice

I. Introducción

En el capitulo anterior se ha estudiado el modelo de regresión lineal simple, donde se analizaba la influencia de una variable explicativa X en los valores que toma otra variable denominada dependiente (Y).

En la regresión lineal múltiple vamos a utilizar más de una variable explicativa; esto nos va a ofrecer la ventaja de utilizar más información en la construcción del modelo y, consecuentemente, realizar estimaciones más precisas.

Al tener más de una variable explicativa (no se debe de emplear el término independiente) surgirán algunas diferencias con el modelo de regresión lineal simple.

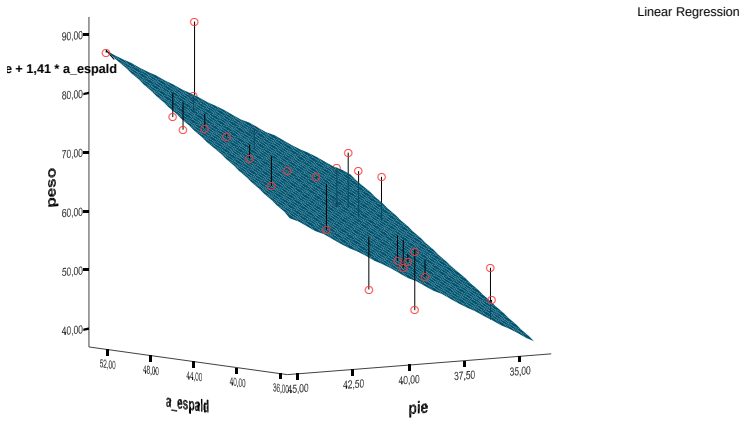
Una cuestión de gran interés será responder a la siguiente pregunta: de un vasto conjunto de variables explicativas: $\mathbf{x_1, x_2, \ldots, xk}$, cuáles son las que más influyen en la variable dependiente Y.

En definitiva, y al igual que en regresión lineal simple, vamos a considerar que los valores de la variable dependiente Y han sido generados por una combinación lineal de los valores de una o más variables explicativas y un término aleatorio:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_k \cdot x_k + u$$

Los coeficientes son elegidos de forma que la suma de cuadrados entre los valores observados y los pronosticados sea mínima, es decir, que se va a minimizar la varianza residual.

Esta ecuación recibe el nombre de hiperplano, pues cuando tenemos dos variables explicativas, en vez de recta de regresión tenemos un plano:



Con tres variables explicativas tendríamos un espacio de tres dimensiones, y así sucesivamente.

Vamos a ir introduciendo los elementos de este análisis a través de un sencillo ejemplo.

Consideramos una muestra de personas como la que sigue a continuación:

Registro	sexo	estatura	l_roxto	pie	l_brazo	a_espald	d_cráneo	peso
		X_1	X_6	X_2	X_3	X_4	X_5	Y
1	mujer	158	39	36	68	43	55	43
2	mujer	152	38	34	66	40	55	45
3	mujer	168	43	39	72.5	41	54.5	48
4	mujer	159	40	36	68.5	42	57	49
5	mujer	158	41	36	68.5	44	57	50
6	mujer	164	40	36	71	44.5	54	51
7	mujer	156	41	36	67	36	56	52
8	mujer	167	44	37	73	41.5	58	52

En base a estos datos, vamos a construir un modelo para predecir el **peso** de una persona (Y). Esto equivale a estudiar la relación existente entre este conjunto de variables $x_1,...,x_5$ y la variable peso (Y).

En primer lugar tenemos que la variable dependiente es el peso; y las variables que vamos a utilizar para predecir el peso reciben el nombre de variables independientes o explicativas.

En la práctica deberemos de elegir cuidadosamente qué variables vamos a considerar como explicativas. **Algunos criterios que deben de cumplir serán los siguientes**:

- *Tener sentido numérico.*
- *No deberá de haber variables repetidas o redundantes*
- *Las variables introducidas en el modelo deberán de tener una cierta justificación teórica.*
- *La relación entre variables explicativas en el modelo y casos debe de ser como mínimo de 1 a 10.*
- *La relación de las variables explicativas con la variable dependiente debe de ser lineal, es decir, proporcional.*

II. El Modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple es idéntico al modelo de regresión lineal simple, con la única diferencia de que aparecen más variables explicativas:

Modelo de regresión simple:

$$y = b_0 + b_1 \cdot x + u$$

Modelo de regresión múltiple:

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + ... + b_k \cdot x_k + u$$

Siguiendo con nuestro ejemplo, si consideramos el peso como variable dependiente y como posibles variables explicativas:

- *estatura*
- *pie*
- *l_brazo*
- *a_espald*
- *d_craneo*

El modelo que deseamos construir es:

$$\boxed{peso = b_0 + b_1 \cdot estatura + b_2 \cdot pie + b_3 \cdot l\_brazo + b_4 \cdot a\_espald + b_5 \cdot d\_craneo}$$

Al igual que en regresión lineal simple, los coeficientes **b** van a indicar el incremento en el peso por el incremento unitario de la correspondiente variable explicativa. Por lo tanto, estos coeficientes van a tener las correspondientes unidades de medida.

III. Hipótesis

Para realizar un análisis de regresión lineal múltiple se hacen las siguientes consideraciones sobre los datos:

a) Linealidad: los valores de la variable dependiente están generados por el siguiente modelo lineal:

$$Y = X * B + U$$

b) Homocedasticidad: todas las perturbaciones tienen las misma varianza:

$$V(u_i) = \sigma^2$$

c) Independencia: las perturbaciones aleatorias son independientes entre sí:

$$E(u_i \cdot u_j) = 0, \forall i \neq j$$

d) Normalidad: la distribución de la perturbación aleatoria tiene distribución normal:

$$U \approx N(0, \sigma^2)$$

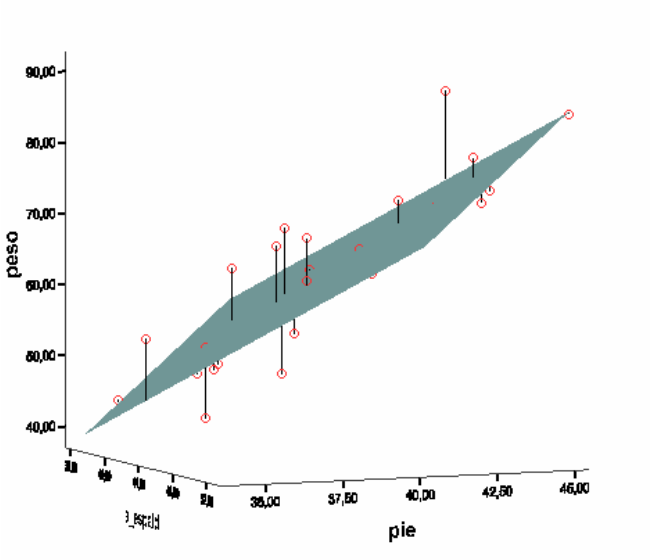
e) Las variables explicativas $\mathbf{X_k}$ se obtienen sin errores de medida.

Si admitimos que los datos presentan estas hipótesis entonces el teorema de Gauss-Markov establece que el método de estimación de mínimos cuadrados va a producir estimadores óptimos, en el sentido que los parámetros estimados van a estar centrados y van a ser de mínima varianza.

IV. Estimación de los parámetros por mínimos cuadrados

Vamos a calcular un hiperplano de regresión de forma que se minimice la varianza residual:

$$Min\sum(y_j - \hat{y}_j)^2$$



Donde:

$$\hat{y}_j = b_0 + b_1 * x_{1,1} + b_2 * x_{2,j} + ...b_k * x_{k,j}$$

Utilizando notación matricial:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{bmatrix} = \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ . \\ . \\ y_n - \hat{y}_n \end{bmatrix} = y - \hat{y}$$

Y teniendo en cuenta la definición de $\hat{y}$:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ u_n \end{bmatrix} = \begin{bmatrix} y_1 - b_0 - b_1 * x_{1,1} - b_2 * x_{2,1} - b_3 * x_{3,1} - ... - b_k * x_{k,1} \\ y_2 - b_0 - b_1 * x_{1,2} - b_2 * x_{2,2} - b_3 * x_{3,2} - ... - b_k * x_{k,2} \\ . \\ . \\ y_n - b_0 - b_1 * x_{1,n} - b_2 * x_{2,n} - b_3 * x_{3,n} - ... - b_k * x_{k,n} \end{bmatrix} = y - \hat{y}$$

Por lo tanto:

$$u = \begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_{1,1} & . & . & x_{k,1} \\ 1 & x_{1,2} & . & . & x_{k,2} \\ & & . & & \\ & & . & & \\ 1 & x_{1,n} & & & x_{k,n} \end{bmatrix} * \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_k \end{bmatrix} = y - X * b$$

Por lo tanto la varianza residual se puede expresar de la siguiente forma:

$$n * \sigma^2 = u' * u = (y - X * b)' * (y - X * b)$$

Es decir:

$$\Phi(b) = \sum (y_j - \hat{y}_j)^2 = u' * u$$

Por tanto, la varianza residual es una función del vector de parámetros **b** y la condición para que tenga un mínimo será:

$$\frac{\partial \phi(b)}{\partial b} = 0$$

Antes de derivar vamos a simplificar la expresión de la varianza residual:

$$n*\sigma^2 = u'*u = (y - x*b)'*(y - x*b) = y'*y - y'*x*b - b'*x'*y + b'*x'*x*b$$

Por lo tanto:

$$\Phi(b) = \sum (y_j - \hat{y}_j)^2 = u'*u = y'*y - y'*x*b - b'*x'*y + b'*x'*x*b$$

$$\frac{\partial \phi(b)}{\partial b} = \frac{\partial (y - X*b)'*(y - X*b)}{\partial b} = -2*X'*Y + 2*X'*X*B$$

Igualando a cero y despejando:

$$X'*Y = X'*X*B$$

y si $X'*X$ es matriz no singular y por lo tanto tiene inversa, tenemos:

$$X'*Y = X'*X*B$$

Multiplicando por $(X'*X)^{-1}$

$$(X'*X)^{-1}X'*Y = (X'*X)^{-1}X'*X*B$$

$$(X'*X)^{-1}X'*Y = I*B$$

$$\boxed{B = (X'*X)^{-1}*X'*Y}$$

Ésta es la expresión del estimador de parámetros **B** .

Además

$$X' * Y = X' * X * B$$

$$X' * Y - X' * X * B = 0$$

$$X' * (Y - X * B) = 0$$

$$\boxed{X' * U = 0}$$

Es decir, los residuos obtenidos del modelo estimado por mínimos cuadrados no van a estar correlacionados con las variables explicativas.

Nota

Es importante observar que si las variables explicativas X están muy correlacionadas entre si, la matriz $(X' * X)$ va a tener el determinante con valor cero o muy cercano a cero.

Si hay al menos una variable que puede ser expresada como combinación lineal del resto (ingresos mensuales, ingresos anuales) el determinante de esta matriz es cero y dicha matriz será singular y por lo tanto no tendrá inversa.

Si no hay variables que sean combinación lineal de las demás, pero están fuertemente correlacionadas, el determinante no será cero pero tendrá un valor muy próximo a cero; este caso va a producir una inestabilidad en la solución del estimador, en general, se va a producir un aumento en su varianza.

En estos casos se impone la utilización de un método de selección de variables explicativas.

A los problemas provocados por la fuerte correlación entre las variables explicativas se les llama **multicolinealidad**.

V. Varianza residual

Al igual que en el caso de regresión lineal simple, vamos a descomponer la variabilidad de la variable dependiente Y en dos componentes o fuentes de variabilidad: una componente va a representar la variabilidad explicada por el modelo de regresión y la otra componente va a representar la variabilidad no explicada por el modelo y, por tanto, atribuida a factores aleatorios.

Consideramos la variabilidad de la variable dependiente como:

$$n*\sigma^2 = \sum (y_i - \bar{Y})^2$$

Es decir, la variabilidad de Y es la suma cuadrática de los valores que toma la variable respecto a la media de la variable.

Sumando y restando el valor pronosticado por el modelo de regresión obtenemos la siguiente expresión:

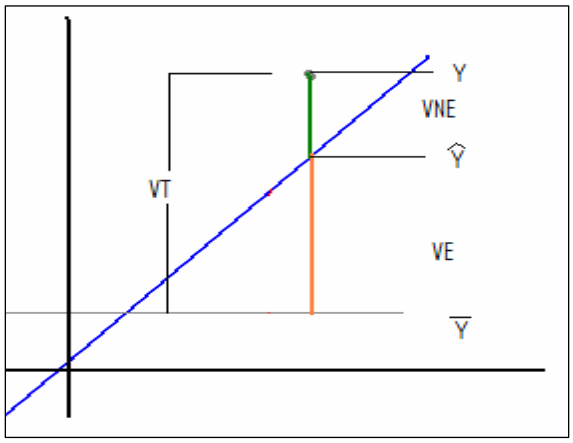
$$\sum (y_i - y)^2 = \sum (\hat{y}_i - y)^2 + \sum (y_i - \hat{y}_i)^2$$

Es decir, que la suma de cuadrados de la variable Y respecto a su media se puede descomponer en términos de la varianza residual. De esta expresión se deduce que “la distancia de Y a su media se descompone como la distancia de Y a su estimación más la distancia de su estimación a la media”.

Teniendo en cuenta que el último término representa la varianza no explicada, tenemos:

$$VT = VE + VNE$$

Gráficamente es fácil ver la relación:



Dividiendo la variabilidad total entre sus grados de libertad obtenemos la varianza de la variable dependiente Y :

$$S_Y^2 = \frac{VT}{n-1}$$

Dividiendo la variabilidad no explicada entre sus grados de libertad obtenemos la varianza residual de la variable dependiente Y :

$$S_R^2 = \frac{VNE}{n-(k+1)}$$

Tabla resumen

	Suma de cuadrados	Grados de libertad	
VT	$\sum(y-\bar{y})^2$	n-1	$S_Y^2 = \frac{VT}{n-1}$
VE	$\sum(\hat{y}-\bar{y})^2$	k-1	
VNE	$\sum (y-\hat{y})^2$	n-k-1	$S_R^2 = \frac{VNE}{n-k-1}$

VI. Contraste de regresión

Como estamos sacando conclusiones de una muestra de un conjunto mucho más amplio de datos, a veces este conjunto será infinito, es obvio que distintas muestras van a dar distintos valores de los parámetros.

Un caso de especial interés es asignar una medida de probabilidad a la siguiente afirmación o hipótesis:

$$H_0 \equiv b_1 = b_2 = ... = b_k = 0$$

La afirmación contraria sería:

$$H_1 \equiv \exists b_j \neq 0$$

Nota

La hipótesis nula es que todos los coeficientes menos b_0 son nulos y la hipótesis alternativa o complementaria es que existe al menos uno que es distinto de 0, puede haber varios que sean nulos, pero al menos existe uno distinto de cero.

Se denomina contraste de regresión al estudio de la posibilidad de que el modelo de regresión sea nulo, es decir, los valores de las variables explicativas X no van a influir en la variable Peso.

Construcción del contraste

Si los residuos siguen una distribución normal y $b_1 = b_2 = ... = b_k = 0$, tenemos que:

$$\frac{VT}{\sigma^2} \approx \chi^2_{n-1}$$

$$\frac{VE}{\sigma^2} \approx \chi^2_1$$

$$\frac{VNE}{\sigma^2} \approx \chi^2_{n-(k+1)}$$

Por tanto:

$$\frac{VE \big/ 1}{VNE \big/ n-(k+1)} = \frac{VE}{S_R^2} \approx F_{1,n-(k+1)}$$

Es decir, el cociente entre la varianza explicada y la varianza no explicada será aproximadamente 1. Además, al seguir una distribución F, podemos asignar una medida de probabilidad (p-value) a la hipótesis de que la varianza explicada es igual a la varianza no explicada.

En caso contrario la varianza no explicada será muy inferior a la varianza explicada y, por lo tanto, este cociente tendrá un valor muy superior a 1.

Nota

En general si el p-value es menor de 0.05 se acepta que el modelo de regresión es significativo; en caso contrario no podemos hablar de regresión, pues el modelo sería nulo.

Si aceptamos que el modelo de regresión es significativo, es habitual mostrar el p-value; por ejemplo:

Encontramos que este modelo de regresión es estadísticamente significativo con un p-value de 0.0003

VII. Coeficiente de determinación R^2

Vamos a construir un coeficiente (estadístico) que mida la bondad del ajuste del modelo. Si bien la varianza residual (S_R^2) nos indica cómo están de cerca las estimaciones respecto de los puntos, esta varianza está influida por la varianza de la variable dependiente, la cual, a su vez, está influida por su unidad de medida. Por lo tanto, una medida adecuada es la proporción de la varianza explicada (VE) entre la varianza total (VT); de este modo, definimos el coeficiente de determinación R^2:

$$R^2 = \frac{VE}{VT} = \frac{VT - VNE}{VT} = 1 - \frac{VNE}{VT}$$

Por ser cociente de sumas de cuadrados, este coeficiente será siempre positivo.

Si todos los puntos están sobre la recta de regresión, la varianza no explicada será 0, y por lo tanto:

$$R^2 = \frac{VE}{VT} = 1 - \frac{0}{VT} = 1$$

Este coeficiente es muy importante pues determina qué porcentaje (en tantos por uno) de la varianza de la variable dependiente es explicado por el modelo de regresión.

En general, se pueden clasificar los valores de R^2 de la siguiente manera:

Menor de 0.3	0.3 a 0.4	0.4 a 0.5	0.5 a 0.85	Mayor de 0.85
Muy malo	Malo	Regular	Bueno	Sospechoso

Además, a diferencia de la varianza residual, este coeficiente es adimensional; esto quiere decir que no está afectado por transformaciones lineales de las variables; por ello, si cambiamos las unidades de medida, el coeficiente de determinación permanecerá invariante.

VIII. Diagnosis y validación de un modelo de regresión lineal múltiple

VIII.1. Multicolinealidad

Si las variables explicativas se pueden expresar como una combinación lineal:

$$\alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_k x_k + \alpha_0 = 0$$

Se dice que tenemos un problema de multicolinealidad.

En general, este problema va a afectar incrementando la varianza de los estimadores.

Este problema se detecta fácilmente:

- Solicitando el determinante de la matriz de varianzas-covarianzas, que estará cercano a cero.
- Calculando el cociente entre el primer y último autovalor de la matriz de varianzas-covarianzas que será mayor de 50.
- Calculando para cada variable el coeficiente de determinación (R^2) de dicha variable con el resto.

La solución es eliminar del modelo aquellas variables explicativas que dependen unas de otras. En general, los métodos de selección de variables solucionan automáticamente este problema.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3485,401	6	580,900	14,986	,000[a]
	Residual	775,265	20	38,763		
	Total	4260,667	26			

a. Predictors: (Constant), l_roxto Longitud de rodilla a tobillo, d_cráneo, a_espald, l_ brazo, pie, estatura

b. Dependent Variable: peso

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-133,261	43,985		-3,030	,007		
	estatura	-,354	,445	-,283	-,796	,435	,072	13,882
	pie	2,187	1,248	,489	1,752	,095	,117	8,574
	l_brazo	,821	,621	,317	1,323	,201	,159	6,307
	a_espald	1,067	,660	,335	1,616	,122	,212	4,724
	d_cráneo	1,093	,922	,157	1,186	,250	,517	1,933
	l_roxto Longitud de rodilla a tobillo	-,003	,841	-,001	-,004	,997	,212	4,724

a. Dependent Variable: peso

En esta tabla se muestra el valor de los estimadores del hiperplano de regresión.

La columna denominada tolerancia es:

$$1 - R^2$$

Donde la variable correspondiente entra como variable dependiente y el resto de las variables explicativas actúan como regresoras.

A la vista de estos resultados, la variable estatura esta provocando problemas de multicolinealidad.

Es interesante observar que si bien el contraste de regresión es significativo, ninguna de las variables explicativas lo es.

VIII.2. Análisis de residuos

Definimos como residuo del i-esimo caso a:

$$u_i = y_i - \hat{y}_i$$

Los residuos son variables aleatorias que siguen (¿?) una distribución normal. Los residuos tienen unidades de medida y, por tanto no se puede determinar si es grande o pequeño a simple vista.

Para solventar este problema se define el **residuo estandarizado** como:

$$Zu_i = \frac{u_i}{\hat{S}_R} * \frac{1}{\sqrt{1-h_{ii}}}$$

Se considera que un residuo tiene un valor alto, y por lo tanto puede influir negativamente en el análisis, si su residuo estandarizado es mayor de 3 en valor absoluto.

$$\lfloor Zu_i \rfloor \geq 3$$

Para evitar la dependencia entre numerador y denominador de la expresión anterior, también se utilizan los **residuos estudentizados.**

$$SZu_i = \frac{u_i}{\hat{S}(i)_R} * \frac{1}{\sqrt{1-h_{ii}}}$$

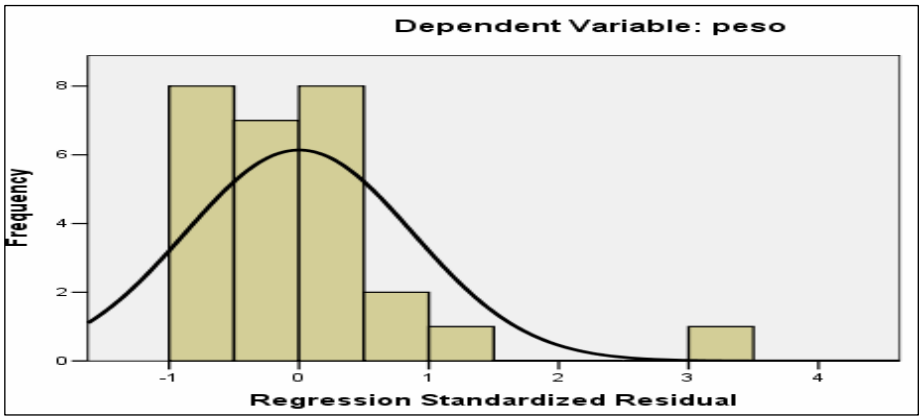
Donde $\hat{S}(i)_R$ es la varianza residual calculada sin considerar el i-esimo caso.

El análisis descriptivo y el histograma de los residuos nos indicarán si existen casos que no se adapten bien al modelo lineal.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	23,9527	138,1509	71,2963	25,44848	27
Residual	-31,69022	117,84905	,00000	29,60339	27
Std. Predicted Value	-1,860	2,627	,000	1,000	27
Std. Residual	-,939	3,492	,000	,877	27

a. Dependent Variable: peso

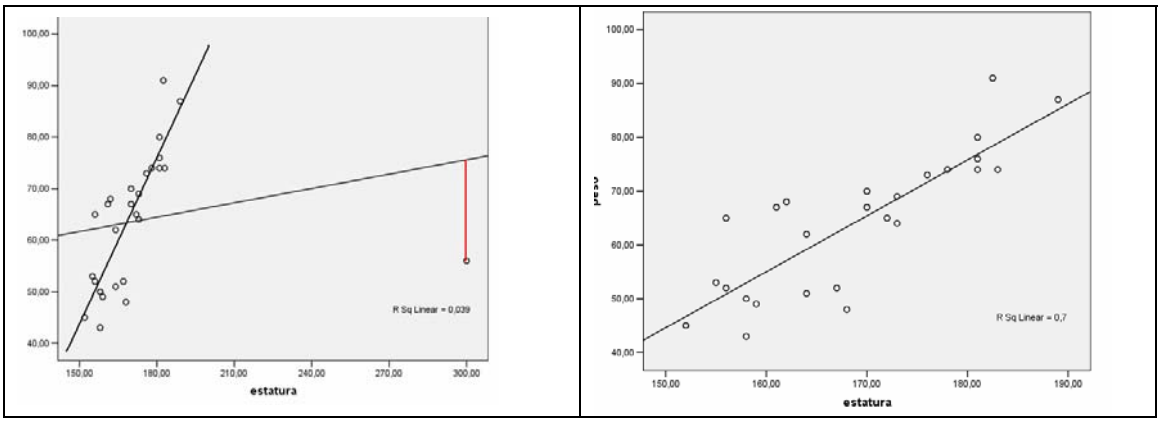


Podemos observar que hay un caso que tiene un residuo anormal, pues su valor tipificado es 3.49.

VIII.3. Valores de influencia (leverage)

Se considera que una observación es influyente a priori si su inclusión en el análisis modifica sustancialmente el sentido del mismo.

Una observación puede ser influyente si es un outlayer respecto a alguna de las variables explicativas:



Para detectar estos problemas se utiliza la **medida de Leverage**:

$$l(i) = \frac{1}{n}(1 + \frac{(x_i - \bar{x})^2}{s_x^2})$$

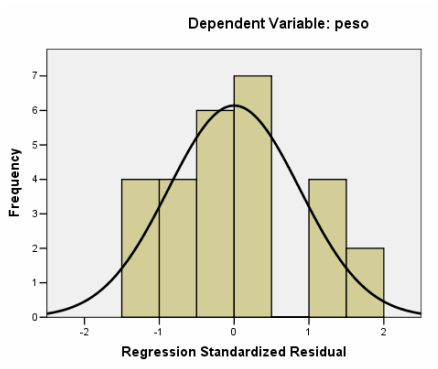
Este estadístico mide la distancia de un punto a la media de la distribución.

Valores cercanos a **2/n** indican casos que pueden influir negativamente en la estimación del modelo introduciendo un fuerte sesgo en el valor de los estimadores.

VIII.4. Contrastando las hipótesis básicas

Normalidad de los residuos.

Para verificar esta hipótesis se suele utilizar el histograma de los residuos y en caso necesario el test de Kolgomorov Smirnov.



One-Sample Kolmogorov-Smirnov Test

		ZRE_1 Standardized Residual
N		27
Normal Parameters[a,b]	Mean	,0000000
	Std. Deviation	,87705802
Most Extreme Differences	Absolute	,117
	Positive	,117
	Negative	-,105
Kolmogorov-Smirnov Z		,609
Asymp. Sig. (2-tailed)		,852

a. Test distribution is Normal.

b. Calculated from data.

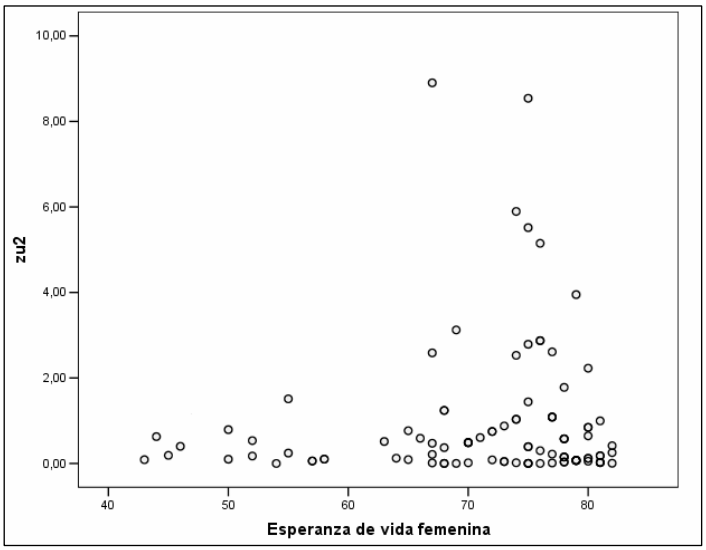
En este caso no se detecta falta de normalidad, el pvalue del test KS es de 0.852, por lo tanto se concluye que:

No se encuentran diferencias estadísticamente significativas para rechazar la hipótesis de normalidad.

VIII.5. Homocedasticidad

La hipótesis de homocedasticidad establece que la variabilidad de los residuos es independiente de las variables explicativas.

En general, la variabilidad de los residuos estará en función de las variables explicativas, pero como las variables explicativas están fuertemente correlacionadas con la variable dependiente, bastara con examinar el gráfico de valores pronosticados versus residuos al cuadrado.



Este es un claro ejemplo de falta de homocedasticidad.

Existe una familia de transformaciones denominada **Box-CCOS** que se realizan sobre la variable dependiente encaminadas a conseguir homocedasticidad. La transformación más habitual para conseguir homocedasticidad es:

$$Y' = \log(Y)$$

En cualquier caso, es conveniente examinar detenidamente las implicaciones de realizar este tipo de transformaciones, pues en muchas ocasiones es peor el remedio que la enfermedad, ya que la variable dependiente puede llegar a perder el sentido.

VIII.6. Errores que deben de evitarse

Errores que son fáciles pasar por alto al realizar un modelo de regresión lineal múltiple son los siguientes:

- No controlar el factor tamaño.
- Si hay un factor de ponderación, no tenerlo en cuenta.
- Al calcular los grados de libertad en los contrastes de hipótesis.
- No incluir una variable relevante en el modelo.
- Incluir una variable irrelevante.
- Especificar una relación lineal que no lo es.

IX. Selección de las variables regresoras

Los procedimientos para seleccionar las variables regresoras son los siguientes:

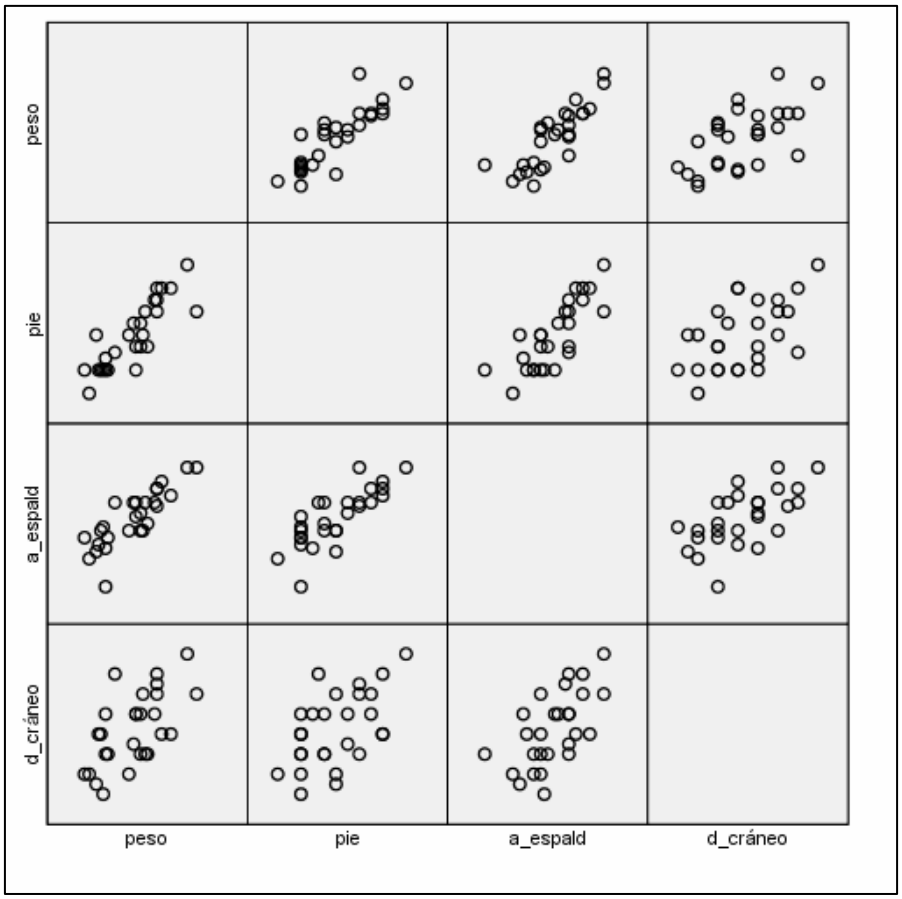
- Eliminación progresiva.
- Introducción progresiva.
- Regresión paso a paso (Stepwise Regression**)**.

Este último método es una combinación de los procedimientos anteriores. Parte del modelo sin ninguna variable regresora y en cada etapa se introduce la más significativa, pero en cada etapa examina si todas las variables introducidas en el modelo deben de permanecer. Termina el algoritmo cuando ninguna variable entra o sale del modelo.

X. Ejemplo 1

Statistics

		estatura	peso	pie	l_brazo	a_espald	d_cráneo	l_roxto Longitud de rodilla a tobillo
N	Valid	27	27	27	27	27	27	27
	Missing	0	0	0	0	0	0	0
Mean		168,7963	63,8889	38,9815	73,4815	45,8519	57,2407	43,0926
Median		168,0000	65,0000	39,0000	73,0000	46,0000	57,0000	43,0000
Std. Deviation		10,22089	12,80124	2,86384	4,93707	4,02113	1,84167	3,15630
Skewness		,173	,187	,303	,427	-,249	,178	,632
Std. Error of Skewness		,448	,448	,448	,448	,448	,448	,448
Kurtosis		-1,016	-,658	-,855	-,605	,075	-,740	1,044
Std. Error of Kurtosis		,872	,872	,872	,872	,872	,872	,872
Minimum		152,00	43,00	34,00	66,00	36,00	54,00	38,00
Maximum		189,00	91,00	45,00	83,00	53,00	61,00	52,00



Correlations

		peso	estatura	pie	l_brazo	a_espald	d_cráneo	l_roxto Longitud de rodilla a tobillo
peso	Pearson Correlation	1	,832**	,850**	,819**	,839**	,619**	,718**
	Sig. (2-tailed)		,000	,000	,000	,000	,001	,000
	N	27	27	27	27	27	27	27
estatura	Pearson Correlation	,832**	1	,927**	,909**	,841**	,588**	,842**
	Sig. (2-tailed)	,000		,000	,000	,000	,001	,000
	N	27	27	27	27	27	27	27
pie	Pearson Correlation	,850**	,927**	1	,851**	,798**	,548**	,851**
	Sig. (2-tailed)	,000	,000		,000	,000	,003	,000
	N	27	27	27	27	27	27	27
l_brazo	Pearson Correlation	,819**	,909**	,851**	1	,801**	,476*	,765**
	Sig. (2-tailed)	,000	,000	,000		,000	,012	,000
	N	27	27	27	27	27	27	27
a_espald	Pearson Correlation	,839**	,841**	,798**	,801**	1	,627**	,631**
	Sig. (2-tailed)	,000	,000	,000	,000		,000	,000
	N	27	27	27	27	27	27	27
d_cráneo	Pearson Correlation	,619**	,588**	,548**	,476*	,627**	1	,555**
	Sig. (2-tailed)	,001	,001	,003	,012	,000		,003
	N	27	27	27	27	27	27	27
l_roxto Longitud de rodilla a tobillo	Pearson Correlation	,718**	,842**	,851**	,765**	,631**	,555**	1
	Sig. (2-tailed)	,000	,000	,000	,000	,000	,003	
	N	27	27	27	27	27	27	27

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,904[a]	,818	,763	6,22602	2,274

a. Predictors: (Constant), l_roxto Longitud de rodilla a tobillo, d_cráneo, a_espald, l_brazo, pie, estatura

b. Dependent Variable: peso

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3485,401	6	580,900	14,986	,000[a]
	Residual	775,265	20	38,763		
	Total	4260,667	26			

a. Predictors: (Constant), l_roxto Longitud de rodilla a tobillo, d_cráneo, a_espald, l_brazo, pie, estatura

b. Dependent Variable: peso

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-133,261	43,985		-3,030	,007		
	estatura	-,354	,445	-,283	-,796	,435	,072	13,882
	pie	2,187	1,248	,489	1,752	,095	,117	8,574
	l_brazo	,821	,621	,317	1,323	,201	,159	6,307
	a_espald	1,067	,660	,335	1,616	,122	,212	4,724
	d_cráneo	1,093	,922	,157	1,186	,250	,517	1,933
	l_roxto Longitud de rodilla a tobillo	-,003	,841	-,001	-,004	,997	,212	4,724

a. Dependent Variable: peso

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	44,1230	88,5975	63,8889	11,57816	27
Residual	-8,21203	11,34415	,00000	5,46058	27
Std. Predicted Value	-1,707	2,134	,000	1,000	27
Std. Residual	-1,319	1,822	,000	,877	27

a. Dependent Variable: peso

El mismo análisis pero utilizando un algoritmo de selección de variables.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,850[a]	,722	,711	6,88269	
2	,891[b]	,794	,777	6,05049	2,120

a. Predictors: (Constant), pie

b. Predictors: (Constant), pie, a_espald

c. Dependent Variable: peso

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3076,382	1	3076,382	64,942	,000[a]
	Residual	1184,285	25	47,371		
	Total	4260,667	26			
2	Regression	3382,065	2	1691,032	46,192	,000[b]
	Residual	878,602	24	36,608		
	Total	4260,667	26			

a. Predictors: (Constant), pie

b. Predictors: (Constant), pie, a_espald

c. Dependent Variable: peso

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-84,173	18,421		-4,569	,000		
	pie	3,798	,471	,850	8,059	,000	1,000	1,000
2	(Constant)	-87,250	16,228		-5,376	,000		
	pie	2,213	,687	,495	3,219	,004	,363	2,753
	a_espald	1,415	,490	,444	2,890	,008	,363	2,753

a. Dependent Variable: peso

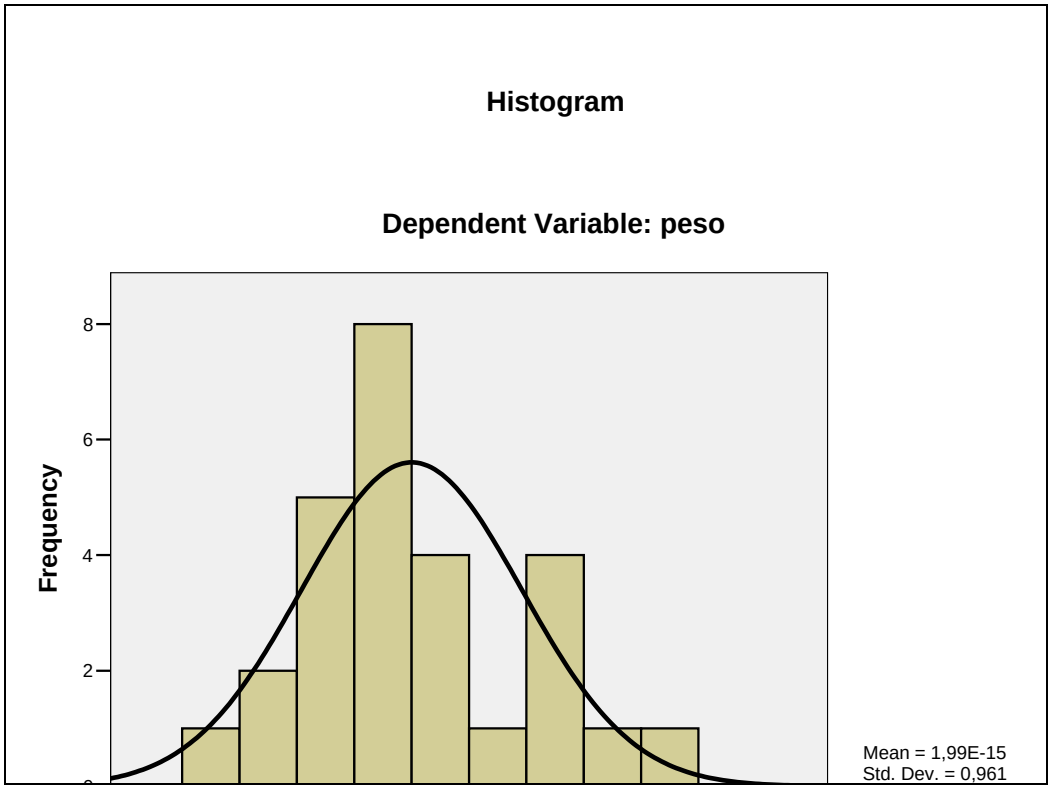
Collinearity Diagnostics[a]

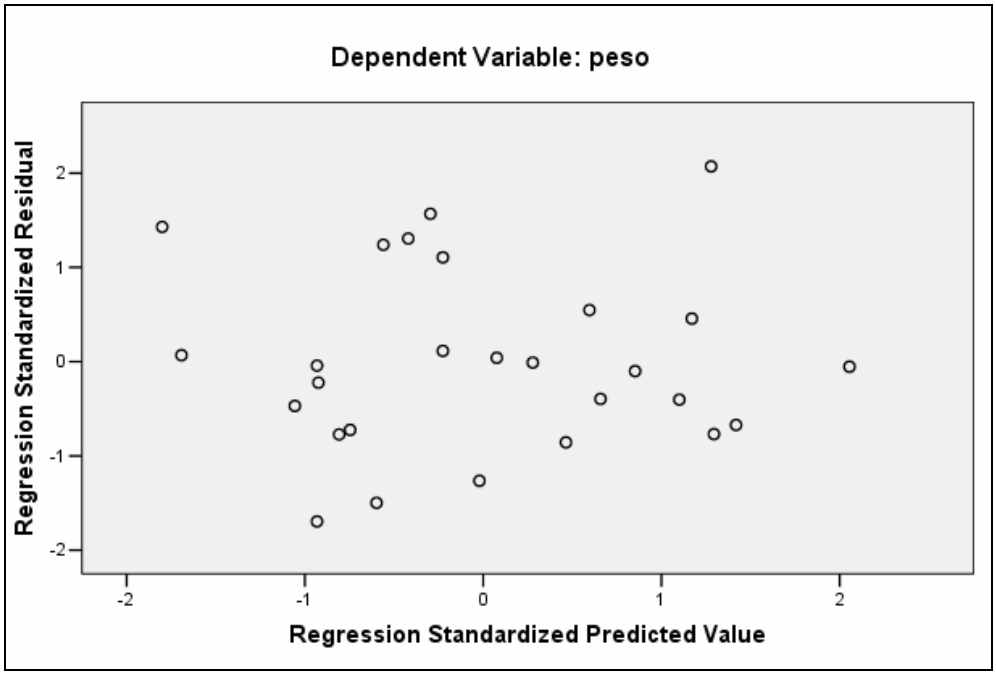
		Condition	Variance Proportions		

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	43,3520	87,3214	63,8889	11,40524	27
Residual	-10,25595	12,53056	,00000	5,81312	27
Std. Predicted Value	-1,801	2,055	,000	1,000	27
Std. Residual	-1,695	2,071	,000	,961	27

a. Dependent Variable: peso


Histogram
Dependent Variable: peso
Frequency
8
6
4
2
Mean = 1,99E-15
Std. Dev. = 0,961


Dependent Variable: peso
Regression Standardized Residual
2
1
0
-1
-2
-2
-1
0
1
2
Regression Standardized Predicted Value