Regresión lineal

Francisco Montes
Departament d'Estadística i I. O.
Universitat de València
http://www.uv.es/~montes

Recta de regresión

Latitud y temperatura

	Latitud	Temperatura		Latitud	Temperatura
Mobile, Ala	30	61	Honolulu, Hawaii	21	79
Montgomery, Ala	32	59	Boise, Idaho	43	36
Juneau, Alaska	58	30	San Juan, Puerto Rico	18	81
Phoenix, Ariz	33	64	Louisville, Ky	38	44
Litle Rock, Ark	34	51	New Orleans, La	29	64
Los Angeles, Cal	34	65	Portland, Maine	43	32
San Francisco, Cal	37	55	Baltimore, Md	39	44
Denver, Col	39	42	Boston, Mass	42	37
New Haven, Conn	41	37	Detroit, Mich	42	33
Wilmington, Del	39	41	Sault Ste Marie, Mich	46	23
Washington, DC	38	44	Minn St Paul, Minn	44	22
Jacksonville, Fla	38	67	St Louis, Missouri	38	40
Key West, Fla	24	74	Charlestown, SC	32	61
Miami, Fla	25	76	Houston, Tx	29	64
Atlanta, Ga	33	52			

Latitud y media de la temperatura máxima en enero desde 1931 a 1960 en 29 ciudades de los EE. UU.

Altura y peso

Altura	Peso	Altura	Peso	Altura	Peso	Altura	Peso
190	80	149	67	162	80	162	39
155	56	190	93	183	88	162	65
167	41	162	58	162	65	160	68
171	49	181	78	173	78	162	63
182	89	166	69	147	60	200	86
173	71	160	52	189	85	202	96
151	53	165	58	185	56	182	84
172	71	182	86	159	58	150	45
175	89	151	48	150	55	168	58
189	93	192	109				

Alturas y pesos de una muestra de 38 individuos

Gráficos de dispersión

La experiencia demuestra que, en general, las personas altas tienen mayor peso y que a mayor latitud (más al Norte) menor es la temperatura. Una adecuada representación gráfica de los pares de observaciones anteriores puede corrobar nuestra conjetura.

Un gráfico de dispersión es la representación más apropiada a este tipo de datos. Veamos los gráficos correspondientes a ambos conjuntos de datos.

Gráfico de dispersión de temperatura y latitud

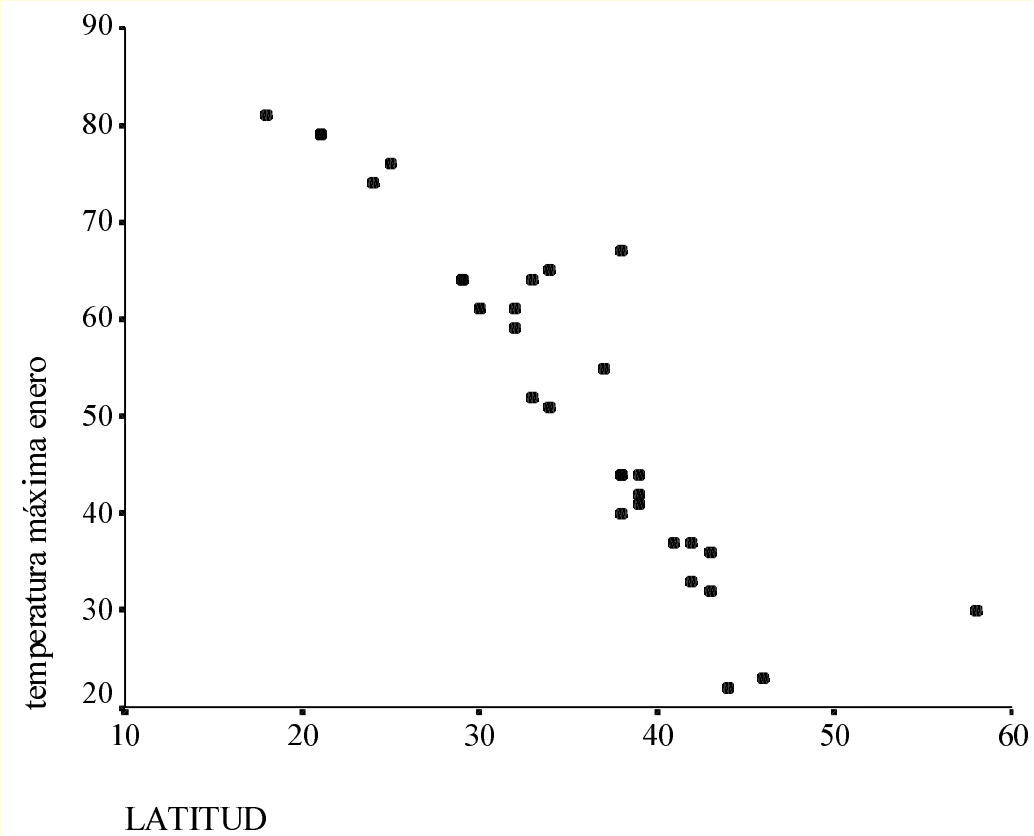
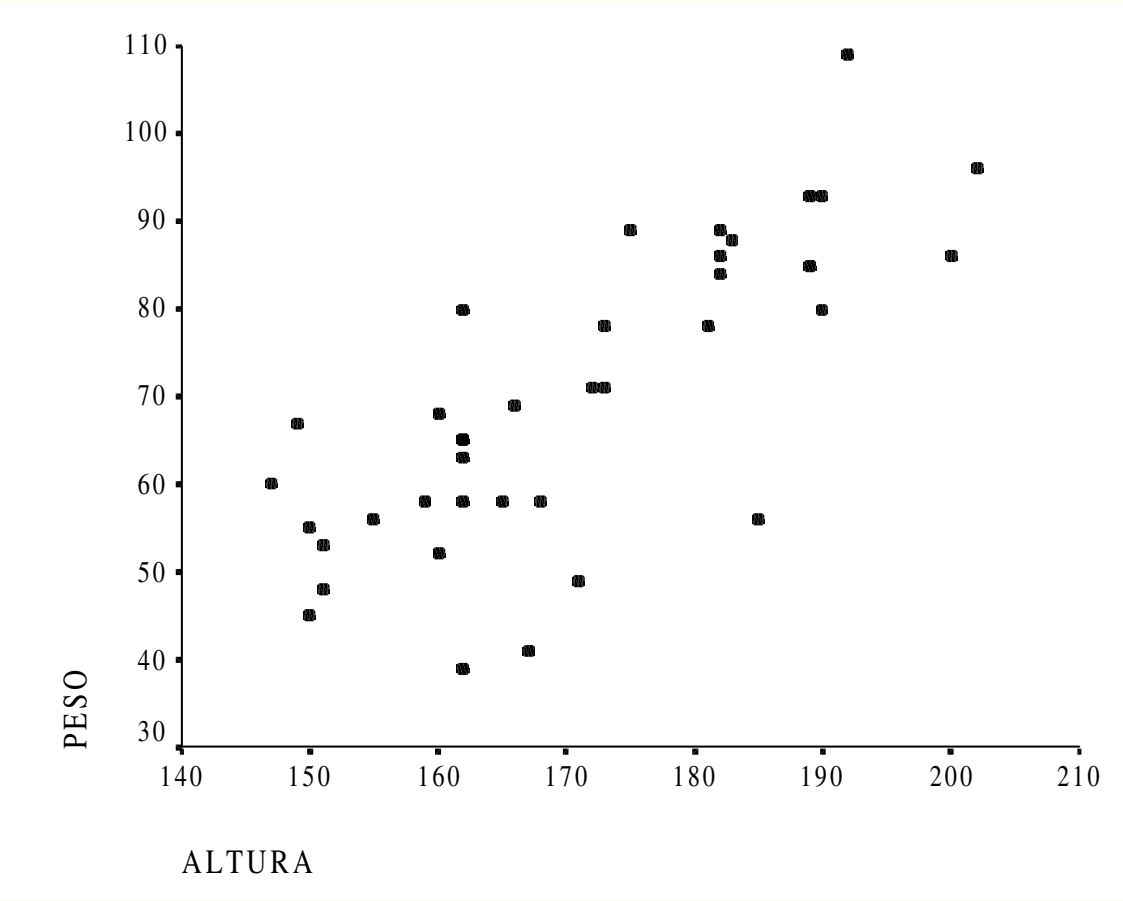


Gráfico de dispersión de altura y peso



Relación funcional entre dos variables

La relación funcional que liga a ambas variables puede ser de muy diversos tipos. El objetivo del gráfico de dispersión es precisamente ayudarnos a elegir el más apropiado a la distribución de puntos que estamos observando.

Tanto para el par latitud-temperatura, como para el par altura-peso, el gráfico sugiere la existencia de una relación lineal, la más sencilla posible entre dos variables.

$$y = ax + b$$

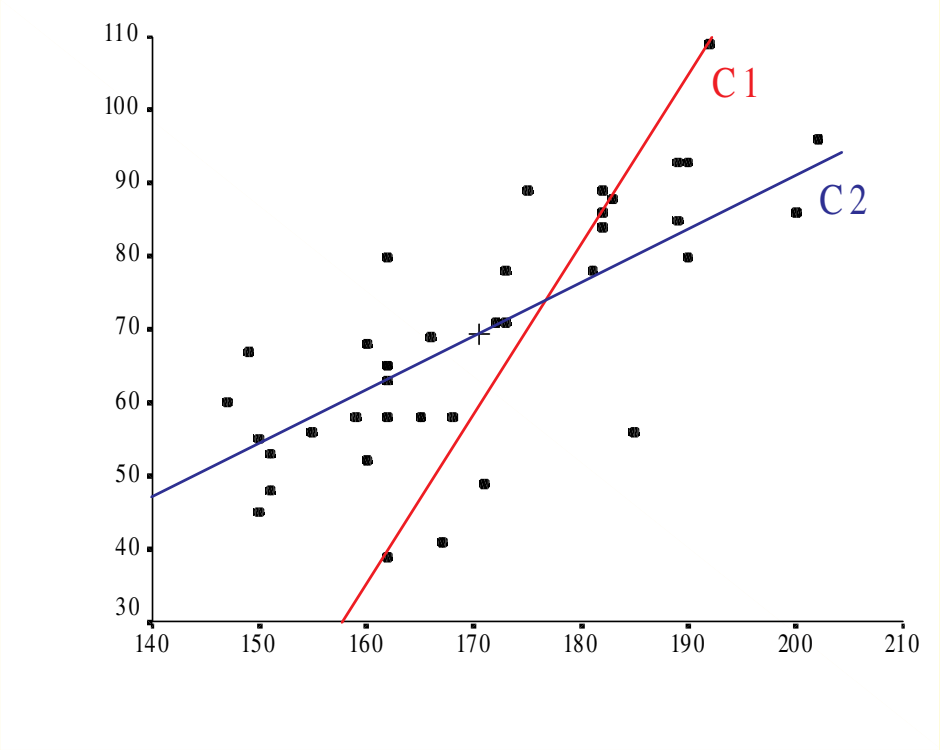
Elección de la recta

Si hemos decidido que una recta puede describir adecuadamente la relación entre y y x. Se trata ahora de elegir aquella que mejor la describa (mejor se ajuste a las observaciones).

$$\text{elegir recta} \Rightarrow \text{obtener valores para } a \text{ y } b$$

Habrá que precisar que entendemos por *mejor se ajuste*. Lo que implica fijar los criterios con los que poder elegir los mejores valores para a y para b.

Un par de criterios de elección



C1) Puntos extremos

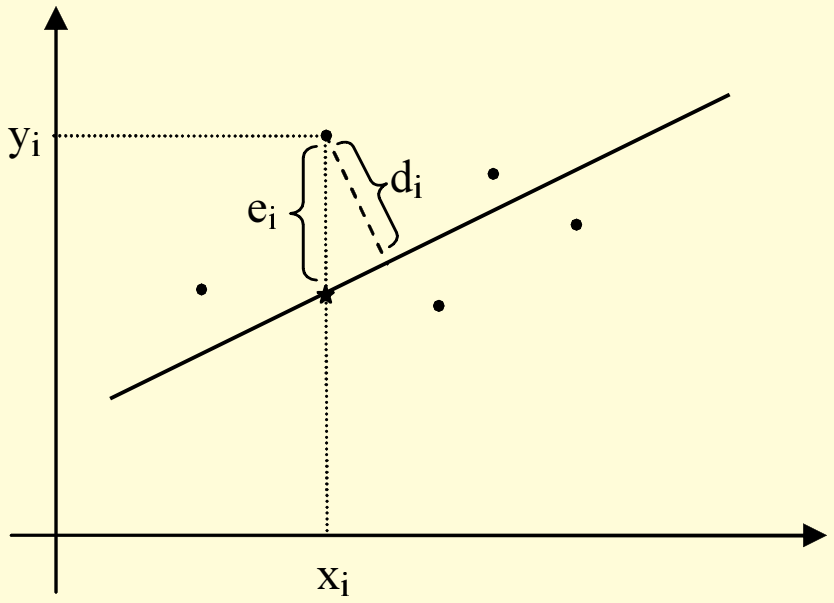
C2) Igual reparto

Pero, ¿son buenos?

La recta **C1** no parece producir un buen ajuste, mientras que la recta **C2** goza de mejor calidad, pero tiene el inconveniente de ser un método gráfico poco eficiente e impreciso porque la recta a determinar no es única.

Necesitamos criterios que determinen unívocamente la recta mediante un método sencillo de obtención.

Más criterios de elección



C3) Mínima distancia, d_i

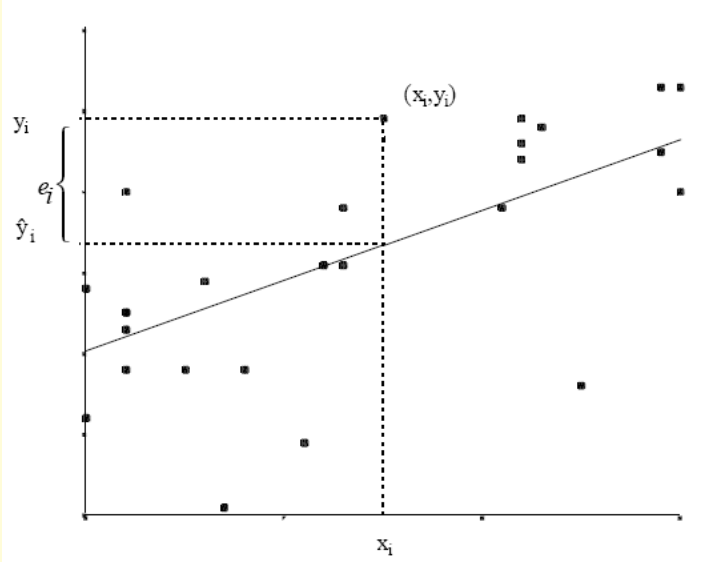
C4) Mínimos cuadrados, e_i^2

Recta de regresión mínimo-cuadrática

Los nuevos criterios cumplen con lo deseado: unicidad y facilidad de obtención. En ambos casos la solución pasa por minimizar una función, entonces, ¿cuál elegir? Si queremos efectuar predicciones con ella, el mejor criterio es el que minimiza el error de predicción, es decir, el de mínimos cuadrados.

La recta así obtenida se denomina recta de regresión mínimo-cuadrática.

Mínimos cuadrados



- Pareja observada: (x_i, y_i)
- Predicción: $\hat{y}_i = ax_i + b$
- Error: $e_i = y_i - \hat{y}_i$

Los valores de a y b deben ser elegido de manera que la llamada suma de cuadrados de los errores sea mínima

Es decir,

$$SC_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (ax_i - b))^2.$$

Valores de a y de b

Con la anterior condición, los valores para a y b vienen dados por las expresiones,

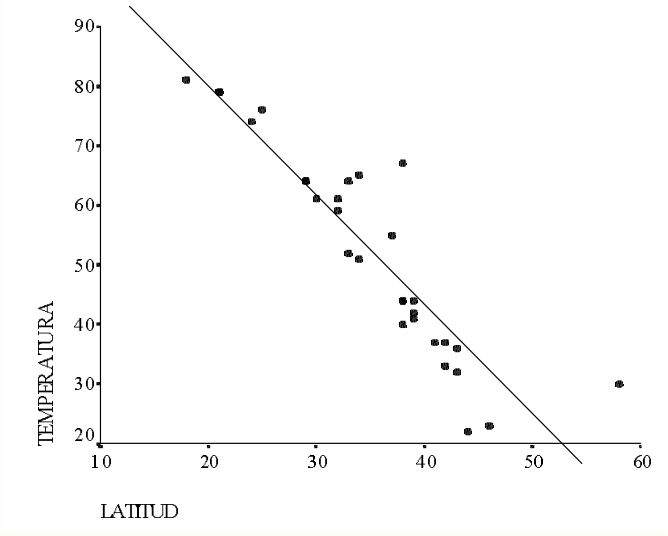
$$a = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathsf{cov}(x,y)}{\mathsf{var}(x)} \qquad b = \overline{y} - a\overline{x}.$$

Obsérvese que la recta de regresión así obtenida pasa por el centro de gravedad de las observaciones: $(\overline{x}, \overline{y})$

Rectas ajustadas

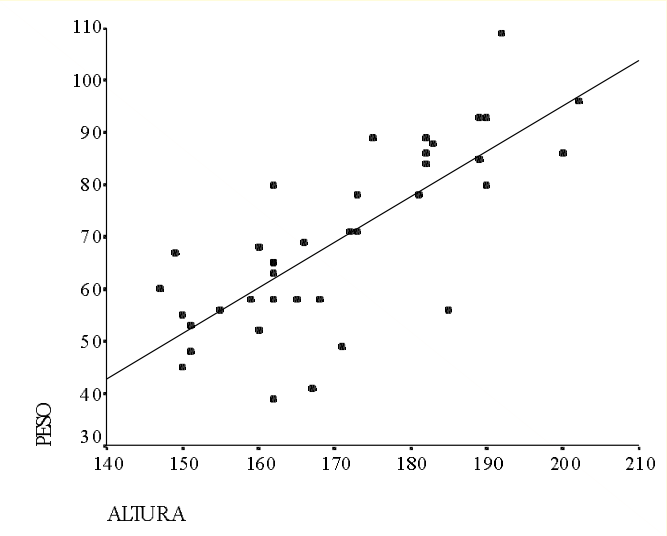
latitud-temperatura

$$y = -1.83x + 116.75$$



altura-peso

$$y = 0.87x - 79.32$$



La bondad del ajuste

La bondad del ajuste se mide mediante la varianza de los errores, s_e^2,

$$s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \overline{e})^2}{n-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-1} = \frac{SC_e}{n-1}$$

porque $\overline{e} = 0$. Pero el valor de s_e^2 depende de la magnitud de las y_i y tiene su dimensión^2, lo que hace difícil comparar las bondades de distintas rectas. Por ejemplo

$$s_e^2(\text{alt-peso}) = 119,49 \quad \text{y} \quad s_e^2(\text{lat-temp}) = 52,15,$$

¿cuál de las dos rectas se ajusta mejor?

El coeficiente de correlación, $\mathbf{r_{xy}}$

El llamado coeficiente de correlación entre x e y, r_{xy}, mide en una escala 0-1. Sus propiedades:

1. $|r_{xy}| \leq 1$
2. si $r > 0 \Longrightarrow a > 0$ y la recta es creciente
3. si $r < 0 \Longrightarrow a < 0$ y la recta es decreciente
4. si $r = 1$, relación lineal perfecta
5. si $r = 0$, ausencia total de relación lineal

Ahora podemos comparar

$$r_{altura-peso} = 0,76 \quad \text{y} \quad r_{lat-temp} = -0,90$$

El coeficiente de determinación,

Otra forma de medir la bondad del ajuste es mediante el concepto de reducción de la varianza.

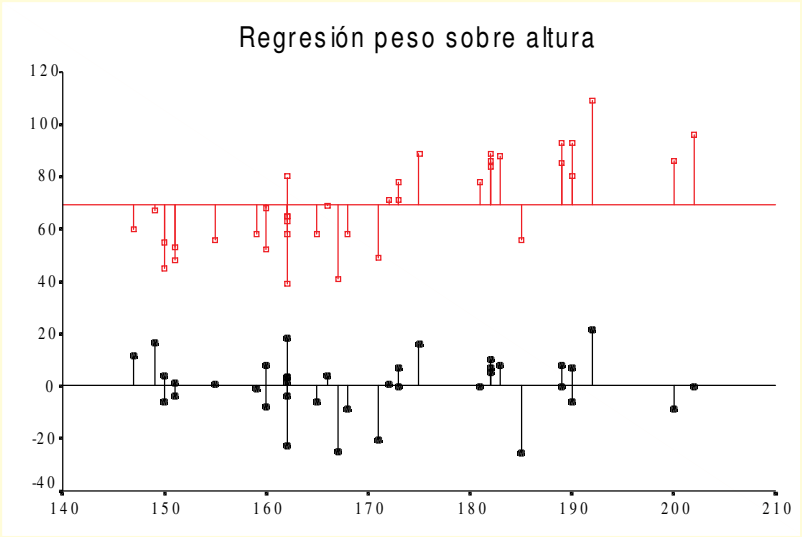
abscisa	**y-estimada**	**error**	**error total**
x_i	$\overline{y}$	$y_i - \overline{y}$	s_y^2
x_i	$\hat{y}_i$	$y_i - \hat{y}_i$	s_e^2

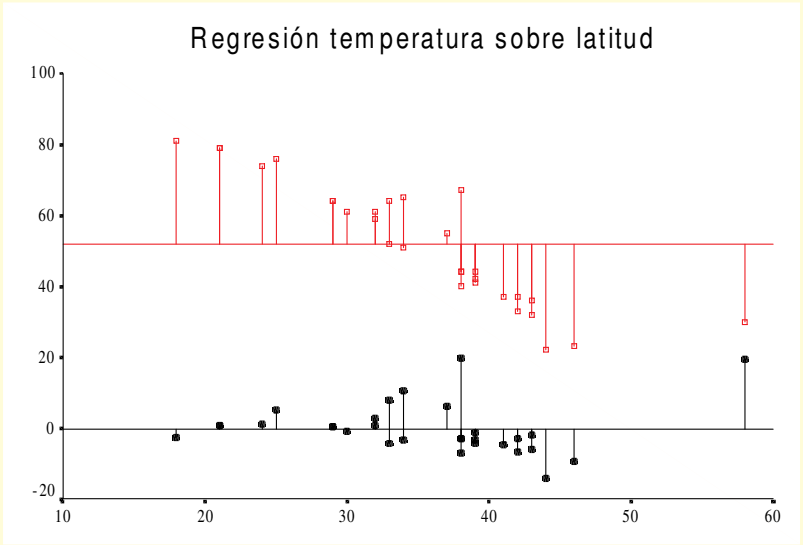
La reducción del error o reducción de la varianza vale,

$$\frac{s_y^2 - s_e^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2} = r_{xy}^2$$

y suele expresarse en porcentaje.

Reducción de la varianza





	r_{xy}	r_{xy}^2	red. varianza
altura-peso	0.7644	0.5843	58.43 %
temperatura-latitud	-0.9051	0.8192	81.92 %

Hasta ahora, ...

todo cuanto hemos dicho cae en el campo del Análisis Matemático. El cálculo de a y b es un sencillo problema de mínimos.

¿Cuando aparecen, si han de hacerlo, la Estadística y la Probabilidad? Cuando nuestras observaciones son una muestra aleatoria de tamaño n,

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

de las variables $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_y, \sigma_y^2)$ con un coeficiente de correlación entre ambas, ρ_{xy}.

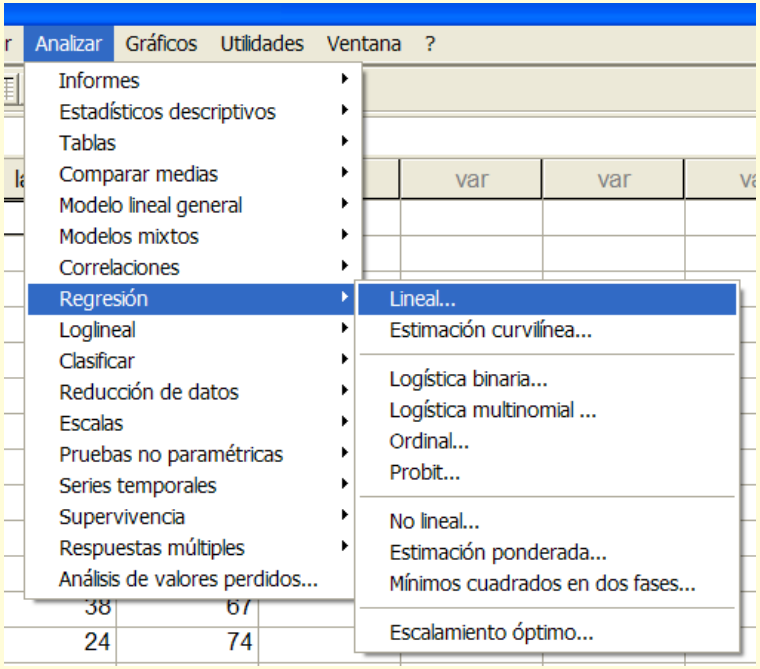
Recta de regresión aleatoria

La nueva situación implica que la recta de regresión que obtengamos es una estimación de la recta que describe la relación entre x e y. Es, ella también, aleatoria. Más concretamente, lo son sus parámetros a y b.

Este hecho nos obliga a conjeturar sobre el verdadero valor (valor poblacional) de los parámetros y a efectuar los contrastes de hipótesis adecuados.

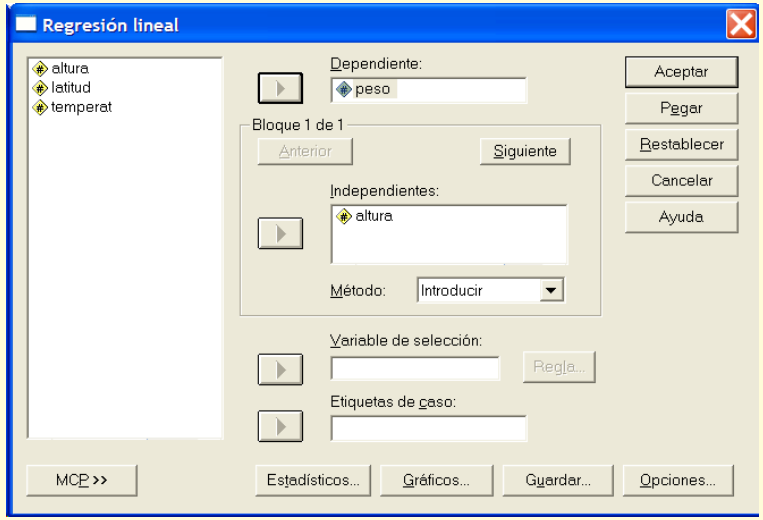
Veamos estos aspectos a través de la muestra de *alturas* y *pesos*, aplicando el software SPSS.

Paso 1



Leídos los datos, desde la opción `Analizar` de la barra del menú accedemos al procedimiento `Regresión Lineal`, tal como muestra la imagen. Se trata, como ya sabemos de un menú interactivo, por el que nos movemos con ayuda del ratón.

Paso 2



Se despliega la ventana de diálogo que mostramos en la que nos limitamos a indicar cuál es la variable *dependiente* (peso) y cuál la *independiente* (altura). El procedimiento se ejecutará al pulsar `Aceptar`.

Resultados

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,764[a]	,584	,573	11,231

a. Variables predictoras: (Constante), ALTURA

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	6382,743	1	6382,743	50,605	,000[a]
	Residual	4540,652	36	126,129		
	Total	10923,395	37			

a. Variables predictoras: (Constante), ALTURA

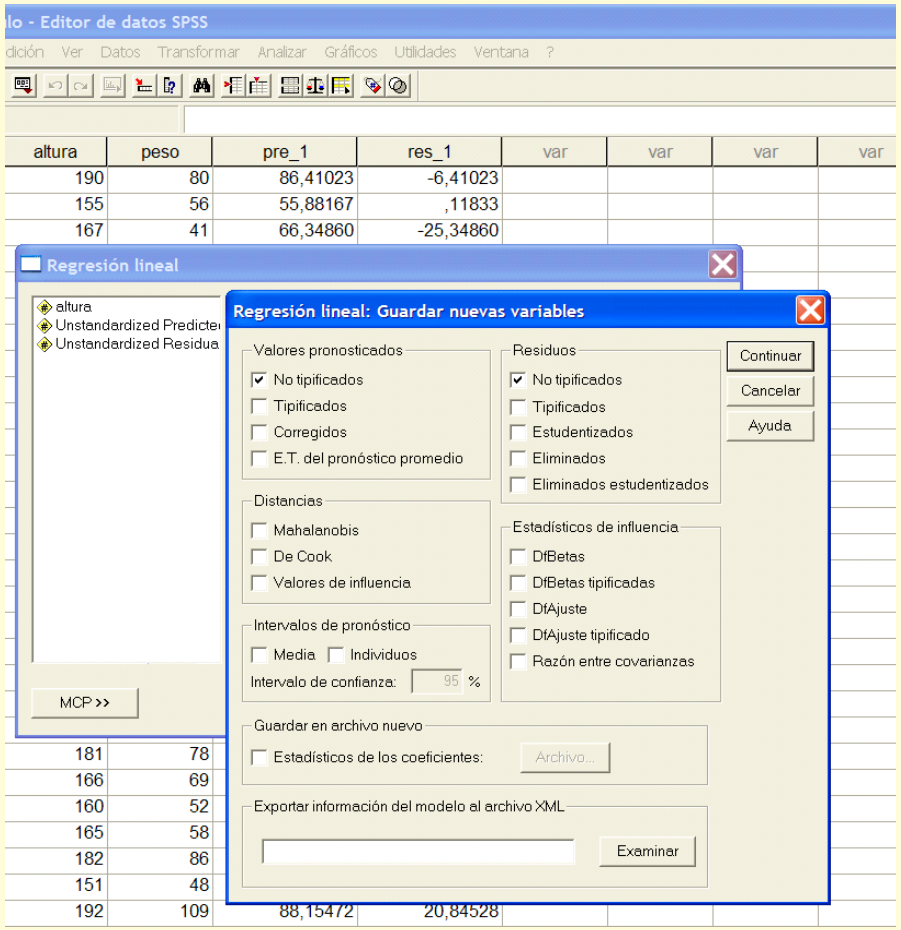
b. Variable dependiente: PESO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-79,316	20,991		-3,778	,001
	ALTURA	,872	,123	,764	7,114	,000

a. Variable dependiente: PESO

Predicción y residuos



Para interpretar las tablas anteriores es conveniente guardar las predicciones $(\hat{y}_i)$ y sus residuos $(e_i = y_i - \hat{y}_i)$. Para ello elegimos la opción `Guardar` en el cuadro de diálogo.

El programa añade dos nuevas variables, `pred_1` y `res_1`, con los valores obtenidos.

Interpretación de los resultados: Resumen del modelo

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,764[a]	,584	,573	11,231

a. Variables predictoras: (Constante), ALTURA

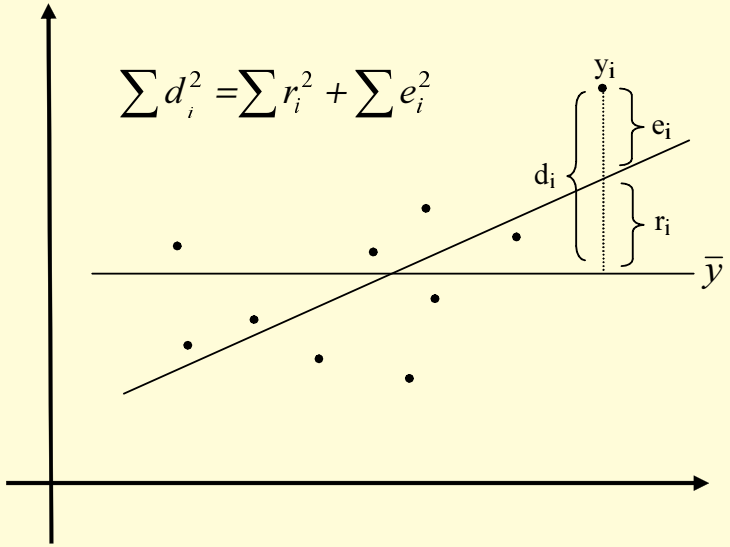
La tabla resume la bondad del modelo mediante

$$R^2 = 1 - \frac{s_e^2}{s_y^2} \qquad \text{con } s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-1}$$

$$R^2_{corr} = 1 - \frac{(s_e^2)^*}{s_y^2} \qquad \text{con } (s_e^2)^* = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

$$se_p = \sqrt{(s_e^2)^*}$$

Interpretación de los resultados: Tabla ANOVA



ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	6382,743	1	6382,743	50,605	,000[a]
	Residual	4540,652	36	126,129		
	Total	10923,395	37			

a. Variables predictoras: (Constante), ALTURA

b. Variable dependiente: PESO

La tabla ANOVA contrasta si la reducción de la varianza es significativa. En este caso lo es pues el valor del estadístico

$$F = 50,605 \text{ con un } p_{\text{valor}} < 0{,}001$$

.

Interpretación de los resultados: Coeficientes

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error típ.	Beta	t	Sig.
1	(Constante)	-79,316	20,991		-3,778	,001
	ALTURA	,872	,123	,764	7,114	,000

a. Variable dependiente: PESO

La tabla muestra los valores estimados de los coeficientes de la recta y contrasta la hipótesis de si pueden ser nulos. Es de particular interés el del coeficiente de la altura, pues si aceptamos que vale 0 no hay recta de regresión.

Este contraste y el de tabla ANOVA son equivalentes, pues

$$F_{1,\nu} = t_\nu^2 \quad \text{y, en efecto,} \quad (7,114)^2 = 50,605$$

Estudio de los residuos

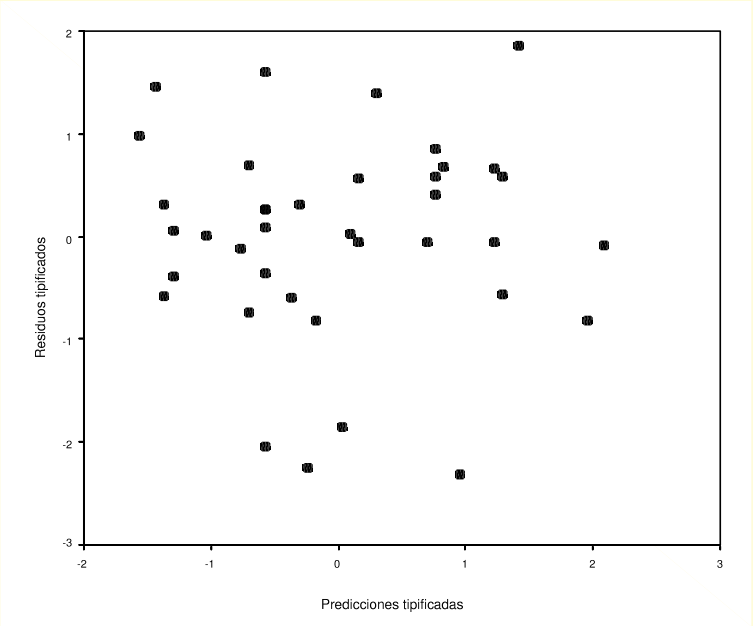
En el modelo de regresión ajustado se supone que

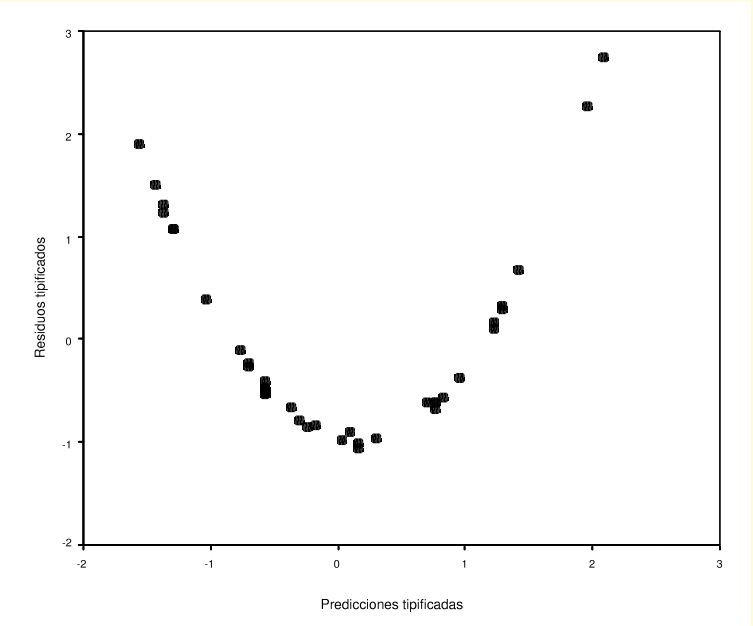
$$Y_i = aX_i + b + E_i, \tag{1}$$

donde los E_i son los residuos o errores, variables aleatorias *i.i.d.* con distribución común $N(0, \sigma^2)$.

Los errores observados, $e_i = y_i - \hat{y}_i$, son estimaciones de aquellas variables y por tanto su análisis es una forma de comprobar que el modelo (1) es correcto y que las condiciones previas de normalidad e independencia se satisfacen.

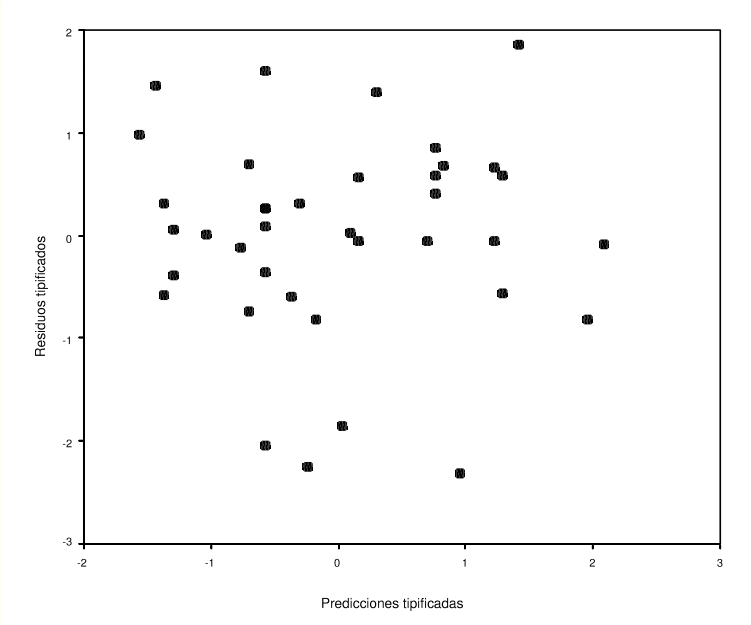
Estudio de los residuos: linealidad

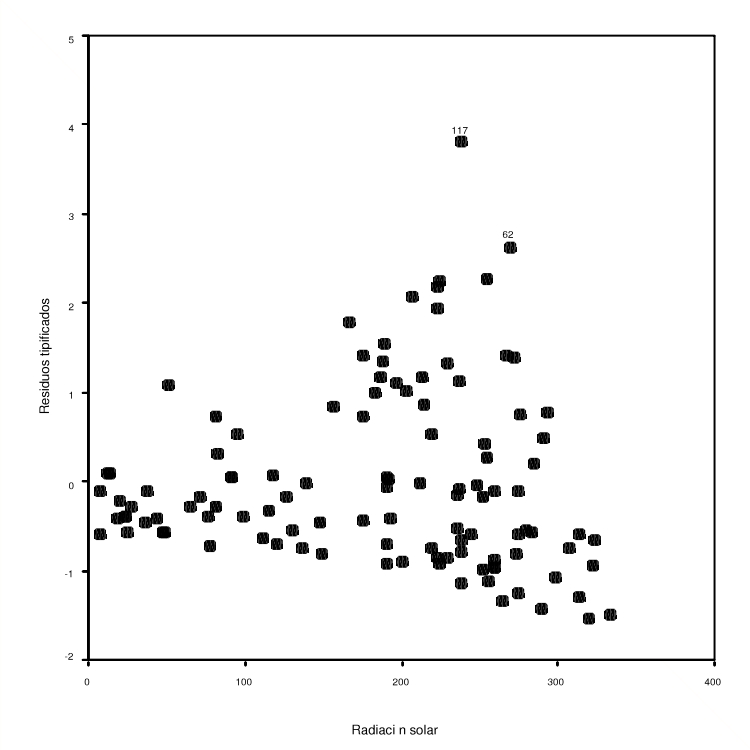




La presencia de un patrón para los residuos es síntoma de no linealidad en la relación entre X e Y.

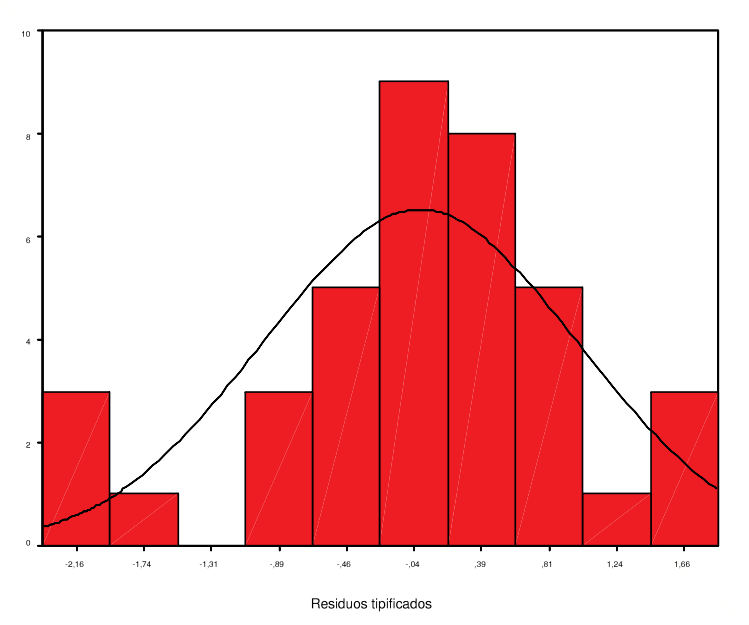
Estudio de los residuos: homcedasticidad





El aumento de la amplitud de la varianza con X es síntoma de que la varianza de Y no permanece constante a lo largo de X. La gráfica permite localizar los outliers o valores atípicos.

Estudio de los residuos: prueba de normalidad

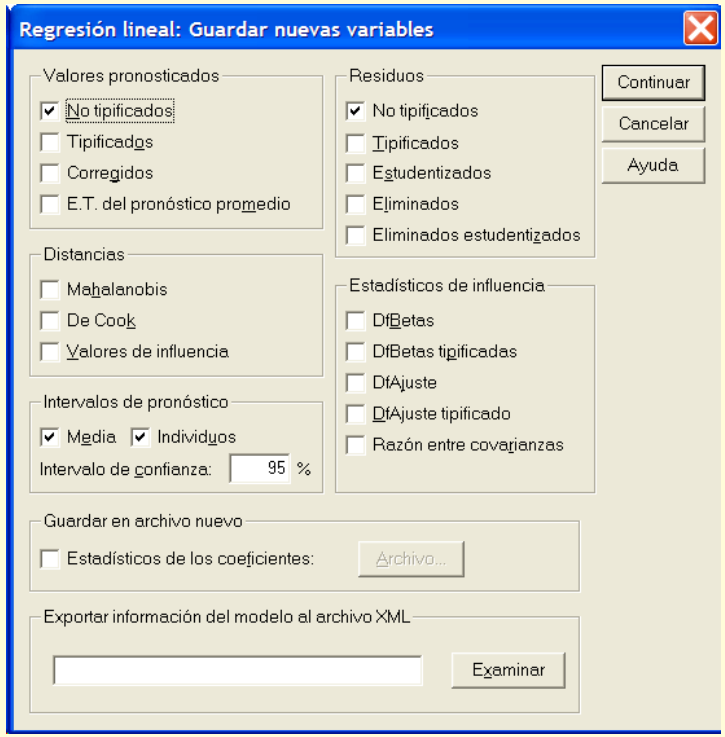


Prueba de K-S para los residuos altura-peso

		Residuos tipificados
N		38
Parámetros normales	Media	,0000000
	Des. típica	,98639392
Diferencias más extremas	Absoluta	,135
	Positiva	,084
	Negativa	-,135
Z de Kolmogorov-Smirnov		,832
Sig. asintót. (bilateral)		,493

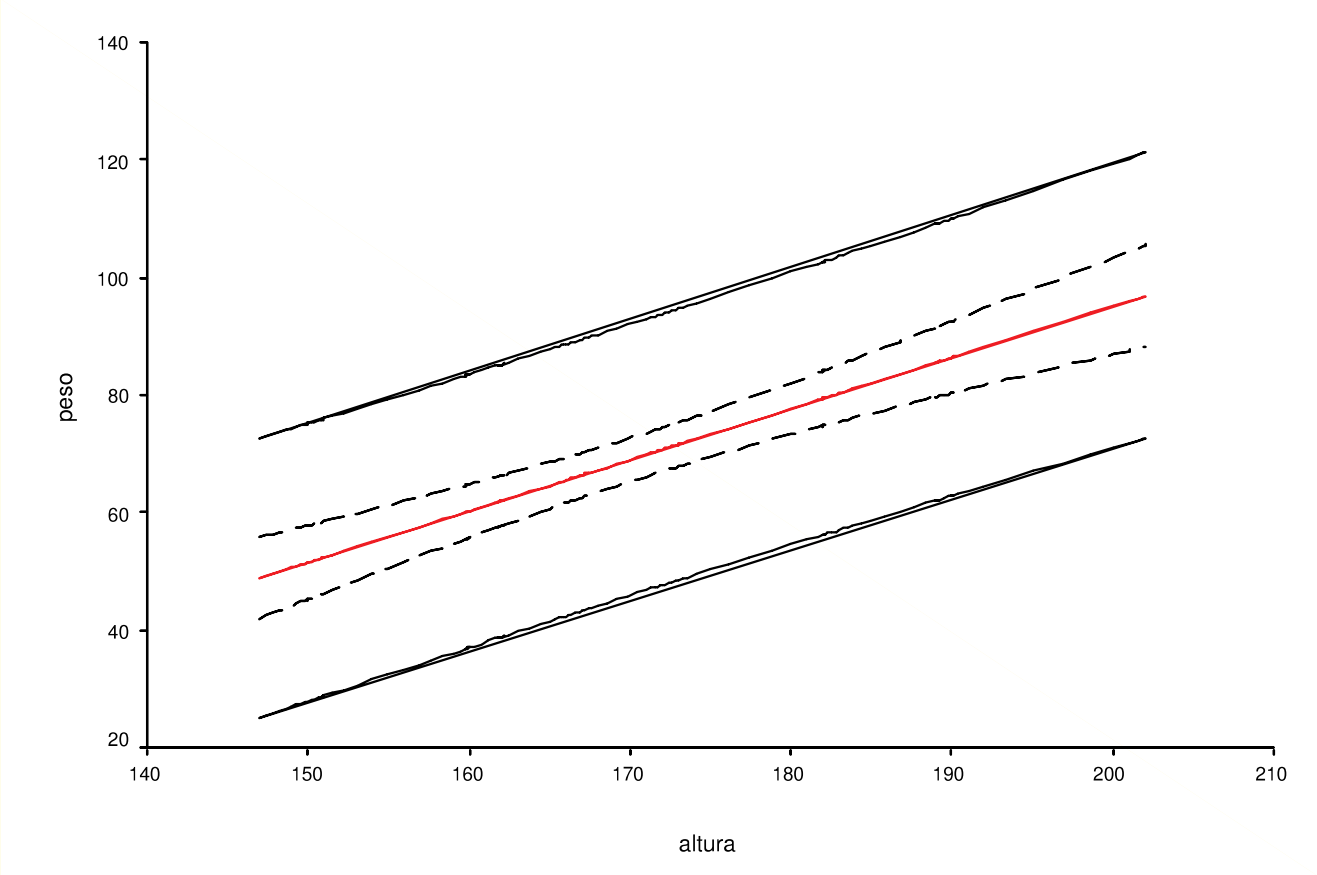
Un histograma y un contraste de normalidad son aconsejables.

Intervalos de confianza para las predicciones



El cuadro de diálogo de la opción Guardar permite obtener como nuevas variables los extremos de los intervalos de confianza para las medias de las predicciones (`lmci_#`, `umci_#`) y para las predicciones individuales (`lici_#`, `uici_#`).

Gráfico de las bandas de confianza para las predicciones



Bandas de confianza del 95 % para las predicciones de la media (linea discontinua) y para las predicciones puntuales (linea continua)

Regresión Múltiple

Calidad del aire

ozono	viento	temp	mes	día	rad_solar
41	7,4	67	5	1	120
36	8	72	5	2	39
12	12,6	74	5	3	182
18	11,5	62	5	4	269
	14,3	56	5	5	265
28	14,9	66	5	6	286
23	8,6	65	5	7	14
19	13,8	59	5	8	127
8	20,1	61	5	9	259
	8,6	69	5	10	50
7	6,9	74	5	11	80
....	....	....	....	....	....
....	....	....	....	....	....

Datos de calidad del aire en 153 estaciones de NY

Regresión lineal múltiple

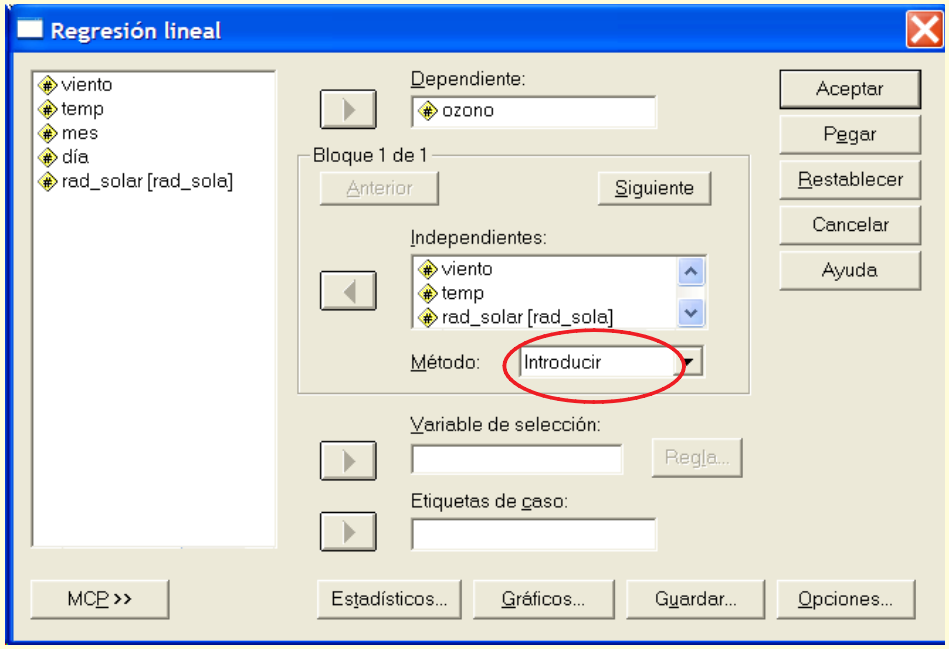
Se trata ahora de describir linealmente la relación entre una variable dependiente Y y un conjunto de p variables independientes X_j, $j = 1, \ldots, p$. El modelo tiene la expresión

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + E_i, \qquad (2)$$

donde, como en el caso de la recta de regresión, E_i es un error aleatorio $N(0, \sigma^2)$.

El problema comparte solución con la regresión simple (univariante) pero posee aspectos propios que conviene señalar.

Calidad del aire: ajuste del ozono



Las instrucciones de la ventana estiman el modelo de dependencia lienal del *ozono* respecto del *viento*, *temperatura* y *radiación solar*.

¿Por qué hemos destacado en rojo el Método? Veamos primero el resultado del ajuste y volvamos luego sobre ello.

Calidad del aire: resultados del ajuste del ozono

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,757	,573	,561	21,849

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	71676,213	3	23892,071	50,048	,000
	Residual	53466,848	112	477,383		
	Total	125143,060	115			

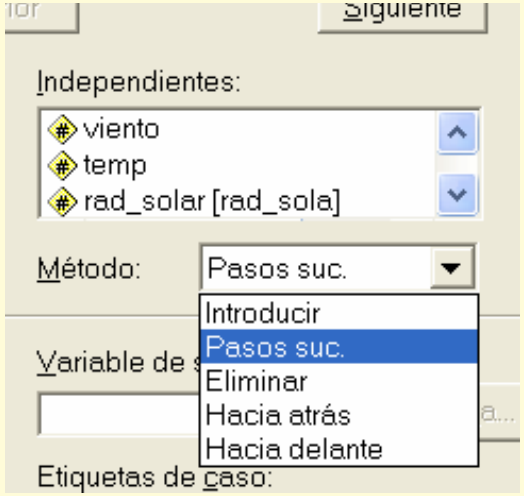
Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-72,075	23,593		-3,055	,003
	VIENTO	-3,141	,668	-,340	-4,700	,000
	TEMP	1,815	,251	,522	7,229	,000
	rad_solar	,024	,024	,064	1,030	,305

a. Variable dependiente: OZONO

Calidad del aire: ajuste del ozono paso a paso

Si observamos la última tabla, el coeficiente de la *radiación solar* no es significativo. Es decir, podemos prescindir de él y efectuar un nuevo ajuste del modelo sólo para las otras dos variables independientes.



Existe una utilidad que nos permite elegir directamente la mejor combinación de variables sin necesidad de ir repitiendo el proceso. Se trata de la opción `Método` que antes destacábamos en rojo.

Calidad del aire: resultado del ajuste paso a paso

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,698[a]	,488	,483	23,714
2	,754[b]	,569	,561	21,855

a. Variables predictoras: (Constante), TEMP

b. Variables predictoras: (Constante), TEMP, VIENTO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	-146,995	18,287		-8,038	,000
	TEMP	2,429	,233	,698	10,418	,000
2	(Constante)	-71,033	23,578		-3,013	,003
	TEMP	1,840	,250	,529	7,362	,000
	VIENTO	-3,055	,663	-,331	-4,607	,000

a. Variable dependiente: OZONO

Calidad del aire: comparación de métodos

Al comparar los resultados de ambos métodos vemos que la calidad del ajuste R^2 es prácticamente la misma con la *radiación solar* que sin ella y el modelo final es más sencillo (principio de parsimonia)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,757	,573	,561	21,849

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,698[a]	,488	,483	23,714
2	,754[b]	,569	,561	21,855

a. Variables predictoras: (Constante), TEMP

b. Variables predictoras: (Constante), TEMP, VIENTO