


El Análisis de la Regresión a través de SPSS

M. Dolores Martínez Miranda

Profesora del Dpto. Estadística e I.O.

Universidad de Granada

Referencias bibliográficas

1. Hair, J.F., Anderson, R.E., Tatham, R.L. y Black, W.C. (1999) **Análisis Multivariante** (5ª edición). Ed. Prentice Hall.
 2. Pérez, C. (2001) **Técnicas estadísticas con SPSS**. Ed. Prentice Hall.
-



INTRODUCCIÓN

- El Análisis de Regresión tiene como objetivo estudiar la relación entre variables.
 - Permite expresar dicha relación en términos de una ecuación que conecta una variable de respuesta Y , con una o más variables explicativas X_1, X_2, \dots, X_k .
 - Finalidad:
 - Determinación explícita del funcional que relaciona las variables. (Predicción)
 - Comprensión por parte del analista de las interrelaciones entre las variables que intervienen en el análisis.
-



PLANTEAMIENTO GENERAL

■ Notación:

Y variable de respuesta (dependiente, endógena, explicada)

X_1, X_2, \dots, X_k variables explicativas (independientes, exógenas, regresores)

■ Modelo general de Regresión:

$$Y = m(X_1, X_2, \dots, X_k) + \varepsilon$$

m función de regresión

ε residuos del modelo (errores de observación, inadecuación del modelo)



Variantes del Análisis de Regresión en SPSS

- Según el número de v. explicativas: Simple o Múltiple.
- Supuestos sobre la función de regresión

Regresión lineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

Estimación curvilínea (Potencial, exponencial, hiperbólica, etc.)

$$Y = \exp(a + b X) \quad \xrightarrow{\text{Linealización}} \quad \ln Y = a + b X$$

Regresión no lineal (Algoritmos de estimación iterativos)



Variantes del Análisis de Regresión en SPSS

➤ Tipo de datos

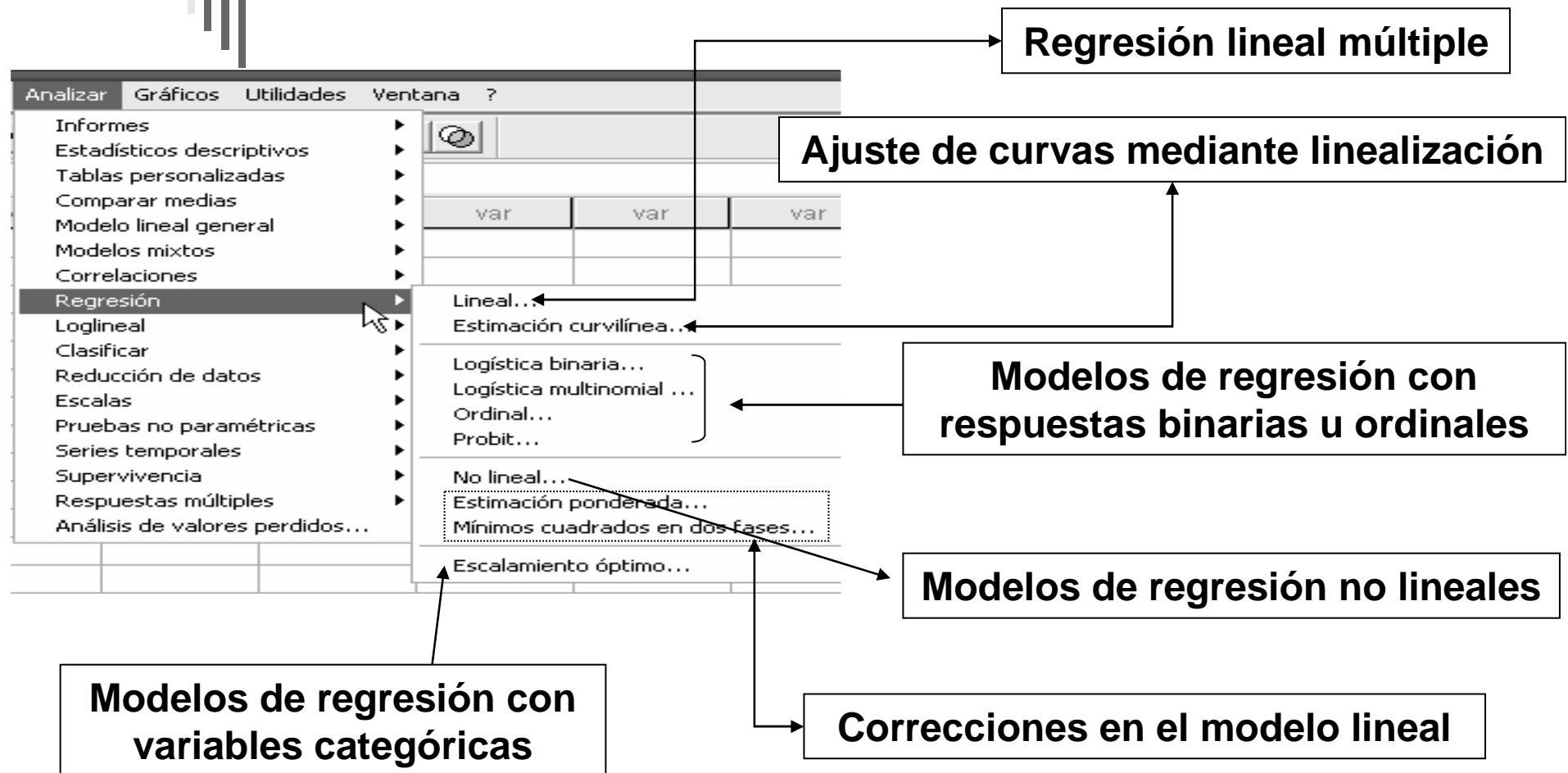
Regresión logística, Modelos Probit (La variable de respuesta es binaria)

Regresión ordinal (La variable de respuesta es de tipo ordinal)

Escalamiento óptimo o regresión categórica (Las variables explicativas y/o explicada, pueden ser nominales)

➤ Situaciones especiales en la estimación del modelo lineal: Mínimos cuadrados en dos fases (correlación entre residuos y v. explicativas), estimación ponderada (situación de heterocedasticidad)

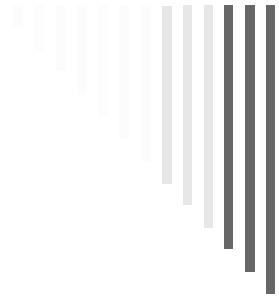
Submenú REGRESIÓN





Contenidos: Aplicaciones con SPSS

- Regresión lineal (múltiple)
 - Estimación ponderada
 - Mínimos cuadrados en dos fases
 - Escalamiento óptimo
 - Regresión curvilínea
 - Regresión no lineal
-



Regresión lineal múltiple

-Modelo teórico-

Modelo lineal $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon \quad (1)$

Parámetros

β_j magnitud del efecto que X_j tienen sobre Y (incremento en la media de Y cuando X_j aumenta una unidad)

β_0 término constante (promedio de Y cuando las v. explicativas valen 0)

ε residuos (perturbaciones aleatorias, error del modelo)

Datos (observaciones, muestra) $\{ (Y_i, X_{1i}, \dots, X_{ki}) : i = 1, \dots, n \}$

PROBLEMA Suponiendo que la relación entre las variables es como en (1), estimar los coeficientes (β_j) utilizando la información proporcionada por la muestra

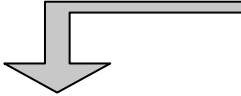
Regresión lineal múltiple

-Modelo teórico-

Expresión matricial

$$Y = X \beta + \varepsilon$$

HIPÓTESIS


$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{21} & \cdots & X_{k1} \\ X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1n} & X_{2n} & \cdots & X_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

- ε_j son v.v.a.a. con media 0 e independientes de las X_j
- **Homocedasticidad:** ε_j tienen varianzas iguales (σ^2)
- **No autocorrelación:** ε_j son incorreladas entre sí
- ε_j son normales e independientes (Inferencia sobre el modelo)
- **No multicolinealidad:** Las columnas de X son linealmente independientes ($\text{rango}(X) = k+1$)

Estimación del modelo

Problema de mínimos cuadrados

\hat{Y}_i valor predicho

Solución
($n > k+1$)

Minimizar
 $\beta_0, \beta_1, \dots, \beta_k$

$$\sum_{i=1}^n \{ Y_i - (\beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}) \}^2$$

Suma residual de cuadrados

Residuo estimado : $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$

$$\hat{\beta} = (\hat{\beta}_0 \quad \hat{\beta}_1 \quad \dots \quad \hat{\beta}_k)^T = (X^T X)^{-1} X^T Y$$

Estimación de los coeficientes



Ejemplo con SPSS (Coches.sav)

Objetivo: Ajustar un modelo lineal que permita predecir el consumo en función de motor, cv, peso y acel

Variable dependiente

CONSUMO	Consumo (l/100Km)
---------	-------------------

Variables independientes

MOTOR	Cilindrada en cc
-------	------------------

CV	Potencia (CV)
----	---------------

PESO	Peso total (kg)
------	-----------------

ACEL	Aceleración 0 a 100 km/h (segundos)
------	-------------------------------------



Analizar Regresión Lineal

Regresión lineal

Dependiente: ☐ Consumo [l/100Km] [c]

Independientes: ☐ Cilindrada en cc [m] ☐ Potencia [CV] [cv]

Método:

Variable de selección:

Etiquetas de caso:

Botones: Aceptar, Pegar, Restablecer, Cancelar, Ayuda

Botones inferiores: MCP >>, Estadísticos..., Gráficos..., Guardar..., Opciones...

Coeficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	,432	1,166		,370	,711
	Cilindrada en cc	3,093E-04	,000	,134	1,612	,108
	Potencia (CV)	4,386E-02	,008	,424	5,582	,000
	Peso total (kg)	4,948E-03	,001	,355	4,404	,000
	Aceleración 0 a 100 km/h (segundos)	2,504E-02	,059	,018	,424	,672

a. Variable dependiente: Consumo [l/100Km]

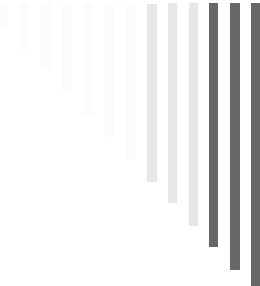
$$\hat{\beta} = (\hat{\beta}_0 \quad \hat{\beta}_2 \quad \dots \quad \hat{\beta}_k)^T$$

$$\text{Consumo} = 0.432 + 3.093\text{E-}04 \text{ Motor} + 4.386\text{E-}02 \text{ CV} + 4.948\text{E-}03 \text{ Peso} + 2.504\text{E-}02 \text{ Acel}$$

**Coeficientes tipificados
Comparación de los efectos**

$$b_j = \frac{\sigma_{X_j}}{\sigma_Y} \hat{\beta}_j$$

Significación individual de las variables y de la constante (Inferencia)



Inferencia sobre el modelo

Significación individual de las variables

Utilidad: Verificar si cada variable aporta información significativa al análisis

Nota: Depende de las interrelaciones entre las variables, **no es concluyente**

Contraste de hipótesis (X_j)

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Aceptar H_0 significa que la variable
“no aporta información significativa”
en el análisis de regresión realizado

Resolución

$$T = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \xrightarrow{\text{Bajo } H_0} t_{n-k-1}$$

Inferencia sobre el modelo

Significación de la constante

Utilidad: Verificar si la v.dependiente tiene media 0 cuando las v.explicativas se anulan

Contraste de hipótesis

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

Aceptar H_0 significa que
“no es conveniente incluir un término constante”
en el análisis de regresión realizado

Resolución

$$T = \frac{\hat{\beta}_0}{SE(\hat{\beta}_0)} \xrightarrow{\text{Bajo } H_0} t_{n-2}$$

Ejemplo (Coches.sav)

Interpretación del p-valor

(en un contraste al nivel de significación α)

Si $p\text{-valor} < \alpha$ entonces se rechaza la hipótesis nula

Coeficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	,432	1,166		,370	,711
	Cilindrada en cc	3,093E-04	,000	,134	1,612	,108
	Potencia (CV)	4,386E-02	,008	,424	5,582	,000
	Peso total (kg)	4,948E-03	,001	,355	4,404	,000
	Aceleración 0 a 100 km/h (segundos)	2,504E-02	,059	,018	,424	,672

a. Variable dependiente: Consumo (l/100Km)

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

Al 5% se puede **no incluir constante** en el modelo

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Al nivel de significación del 5%:

- Motor (0.108) y Acel (0.672) “no son significativas”
- CV (0.000) y Peso (0.000) “sí son significativas”

Inferencia sobre el modelo

Bondad de ajuste

Descomposición de la variabilidad

$$\underbrace{\sum_{i=1}^n (Y_i - \bar{y})^2}_{VT} = \underbrace{\sum_{i=1}^n (\hat{Y}_i - \bar{y})^2}_{VE} + \underbrace{\sum_{i=1}^n \hat{\varepsilon}_i^2}_{VNE}$$

Coeficiente de determinación \longrightarrow R: Coeficiente correlación lineal múltiple

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

Indica la mayor correlación entre Y y las
c.l. de las v. explicativas

Inconveniente: Sobrevalora la bondad del ajuste

Coeficiente de determinación corregido

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} R^2$$

Ejemplo (Coches.sav)

Resumen del modelo^b

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,869 ^a	,755	,752	1,970

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg), Potencia (CV), Cilindrada en cc

b. Variable dependiente: Consumo (l/100Km)

$$R^2 = 0.755$$

Consumo queda explicada en un 75.5% por las variables explicativas según el modelo lineal considerado

$$R^2 \text{ corregido} = 0.752 \quad (\text{siempre algo menor que } R^2)$$



Inferencia sobre el modelo

Contraste de regresión (ANOVA)

Utilidad: Verificar que (de forma conjunta) las v.explicativas aportan información en la explicación de la variable de respuesta

Contraste:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$$H_1 : \text{Algún } \beta_j \neq 0$$



$$H_0 : R = 0$$

$$H_1 : R \neq 0$$

Aceptar H_0 significa que
“las v.explicativas no están relacionadas linealmente con Y”

Resolución (ANOVA)

$$F = \frac{VE / k}{VNE / (n - k - 1)} \xrightarrow{\text{Bajo } H_0} F_{k, n-k-1}$$

Ejemplo (Coches.sav)

Contraste de regresión

ANOVA^b

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	4626,220	4	1156,555	297,956	,000 ^a
	Residual	1502,188	387	3,882		
	Total	6128,408	391			

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg), Potencia (CV), Cilindrada en cc

b. Variable dependiente: Consumo (l/100Km)

Al 5% se rechaza H_0
(las variables explicativas influyen de forma conjunta y lineal sobre Y)

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Media cuadrática	F exp.
Modelo	VE	k	VE / k	$\frac{VE / k}{VNE / (n-k-1)}$
Residual	VNE	n-k-1	VNE / (n-k-1)	
Total	VT	n-1		

Predicción

Bandas de confianza

Predicciones para Y
(dentro del rango de predicción)

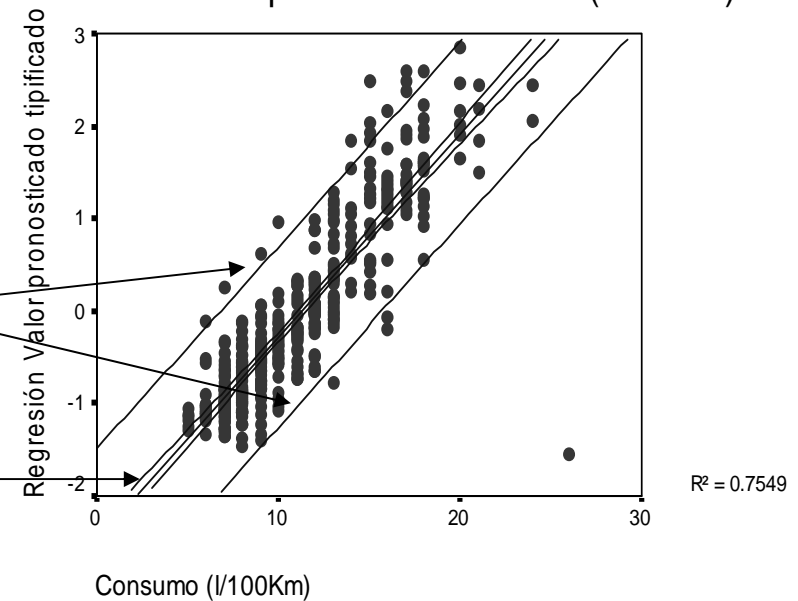
$$\hat{Y}(x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$$

I.C. para los valores predichos de Y

I.C. para la media de Y

Gráfico de dispersión

Variable dependiente: Consumo (l/100Km)

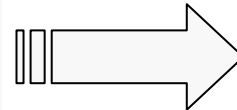




El análisis de los residuos

Objetivo: Verificar que no se violan las hipótesis sobre las que se estima el modelo y se realiza la inferencia

1. Normalidad de los residuos
2. No autocorrelación
3. Homocedasticidad
4. Falta de linealidad
5. No multicolinealidad



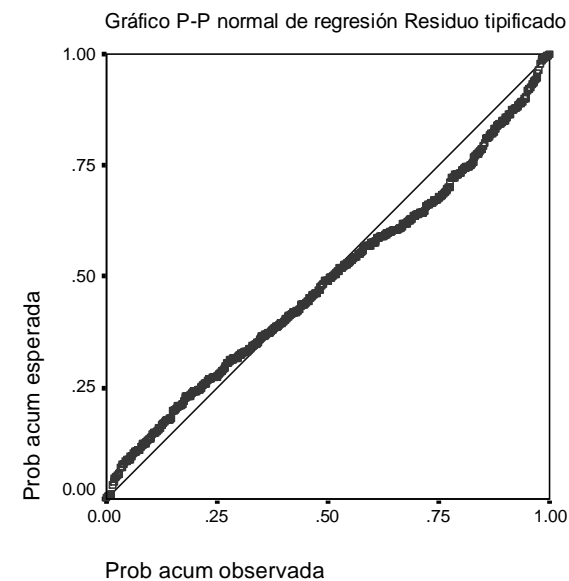
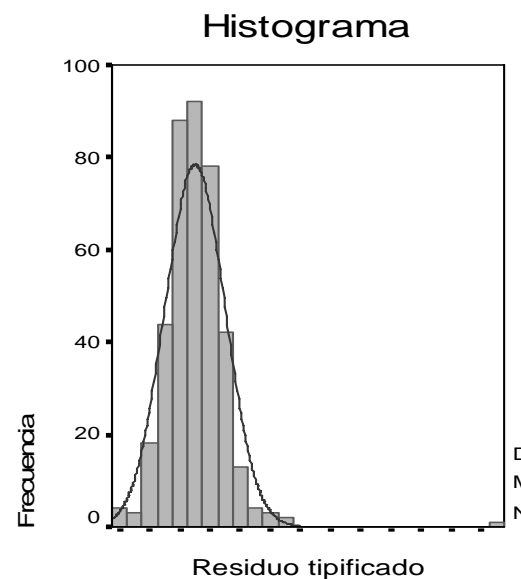
Posibles correcciones:

- Detección de atípicos y puntos influyentes
 - Transformaciones
 - Variables ficticias
 - Ajustes polinomiales
 - Términos de interacción
-

1.1. Normalidad de los residuos

Herramientas disponibles en SPSS

Gráficos: Histograma, gráfico probabilístico normal

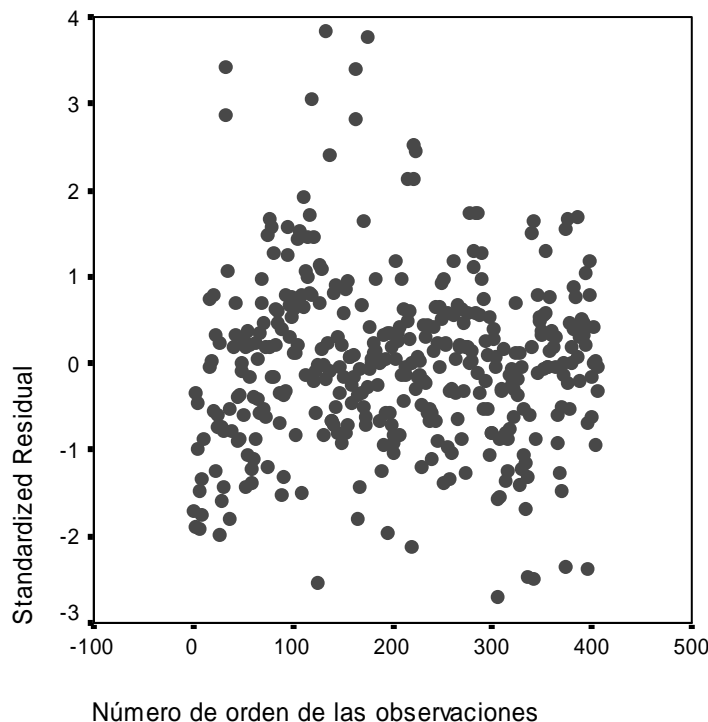


Contrastes: Kolmogorov-Smirnov, Shapiro-Wilks,...

1.2. No autocorrelación

Hace referencia a los efectos de la inercia de una observación a otra que pueda indicar la no independencia entre los residuos.
Se trata de buscar modelos o pautas en los gráficos residuales frente al número de caso (incluso con cada variable independiente).

Herramientas disponibles en SPSS: Gráficos residuales y el estadístico de Durbin-Watson



de la	Durbin-W
ción	atson
1,970	1,228
m/h (segundos),	

H_0 : No hay autocorrelación

- Si $d < 1.18$ rechazar,
- Si $d > 1.4$ no rechazar.
- Si $1.18 < d < 1.4$ no es concluyente

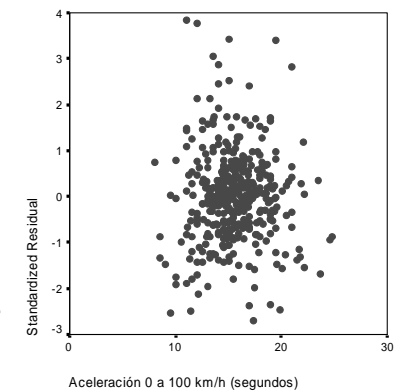
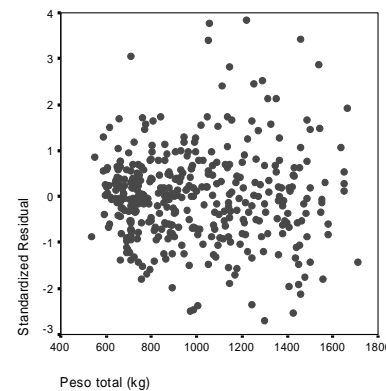
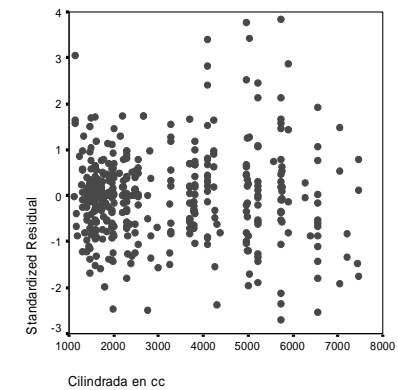
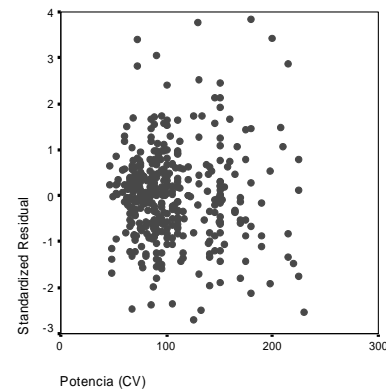
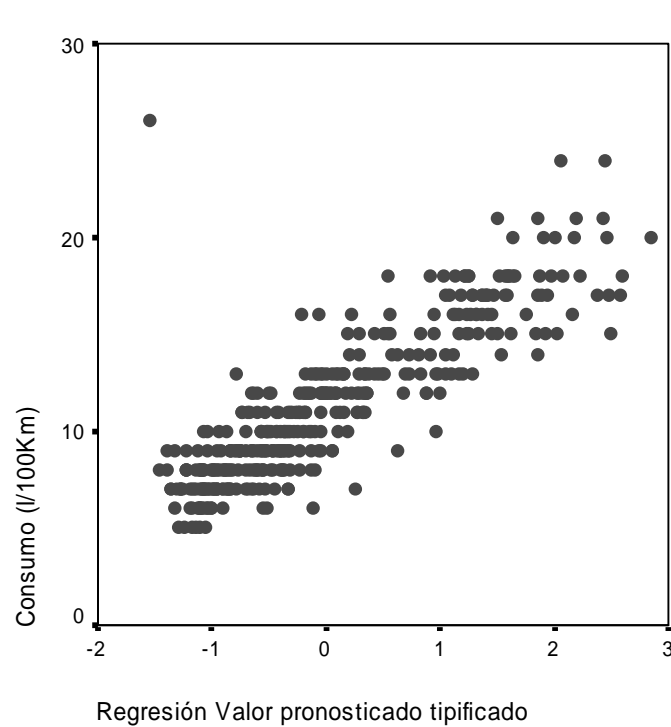
Posibles soluciones:

- Transformaciones
- Añadir variables

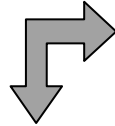
1.3. Homocedasticidad

Hace referencia a la constancia de los residuos para los valores que van tomando las variables independientes.

Herramientas disponibles en SPSS: Gráficos residuales

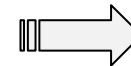
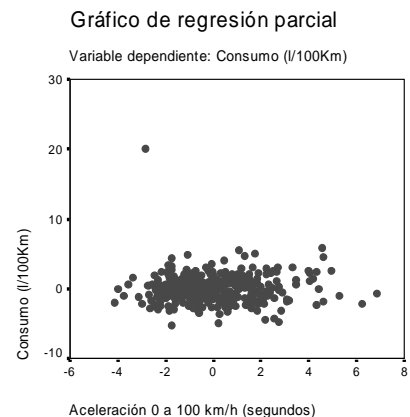
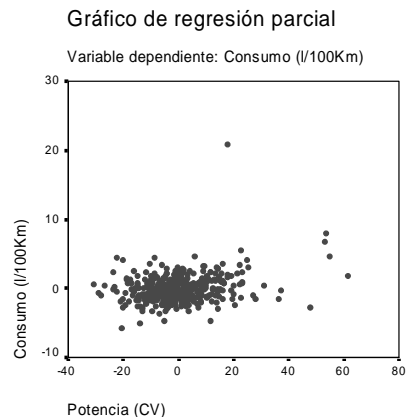
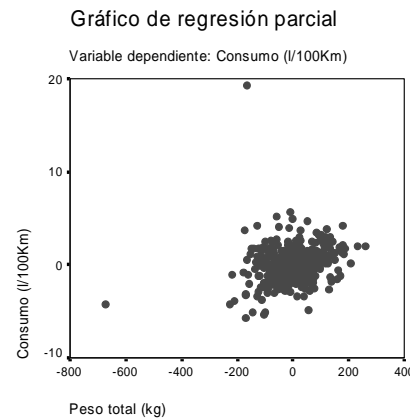
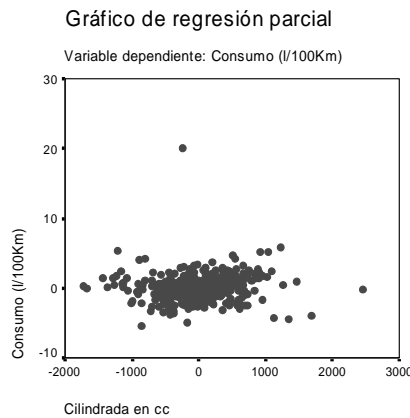


1.4. Falta de linealidad



Hace referencia a las posibles desviaciones de los datos desde el modelo lineal que se está ajustando.

Herramientas disponibles en SPSS: Gráficos de regresión parcial y gráficos residuales



Son diagramas de dispersión de los residuos de cada v. independiente y los residuos de la v. dependiente cuando se regresan ambas por separado sobre las restantes v. independientes.

1.5. No multicolinealidad

Colinealidad es la asociación, medida como correlación, entre dos variables explicativas (el término **multicolinealidad** se utiliza para tres o más variables explicativas).

Impacto de la multicolinealidad

Reducción del poder explicativo de cualquier v. explicativa individual en la medida en que está correlada con las otras v. explicativas presentes en el modelo.

Herramientas disponibles en SPSS: Índices de condicionamiento, FIV

Diagnósticos de colinealidad

Modelo	Dimensión	Autovalor	Índice de condición	Proporciones de la varianza				
				(Constante)	Cilindrada en cc	Potencia (CV)	Peso total (kg)	Aceleración 0 a 100 km/h (segundos)
1	1	4,729	1,000	,00	,00	,00	,00	,00
	2	,238	4,454	,00	,03	,00	,00	,02
	3	2,268E-02	14,440	,03	,22	,29	,01	,06
	4	6,265E-03	27,474	,20	,75	,02	,70	,00
	5	3,612E-03	36,185	,76	,01	,69	,29	,92

1. Identificar los índices que estén por encima del umbral: 30
2. Para los índices identificados, identificar las variables con proporciones de varianza por encima del 90%: Habrá multicolinealidad si ocurre con dos o más coeficientes.

Posibles soluciones:

- ACP y utilizar las componentes principales como regresores.
- A la vista de las correlaciones eliminar variables “redundantes”.



Datos anómalos

Medidas de influencia

Objetivo: Detectar datos anómalos y datos influyentes

Datos anómalos (atípicos)

Individuos cuyo residuo tipificado es superior a 3 (en valor absoluto)

Datos influyentes

Individuos cuya omisión produce cambios notables en los resultados del análisis

Herramientas estadísticas (medidas de influencia)

- Identificación de puntos de apalancamiento (observaciones aisladas del resto sobre una o más v.independientes)
 - Observaciones influyentes: influencias sobre coeficientes individuales, medidas globales de influencia.
-



Medidas para identificar puntos de apalancamiento:

Leverage o **medida de influencia**: Límite: $2(k+1) / n$ (Si $n > 50$, $3(k+1) / n$)

Distancia de **Mahalanobis**: Considera la distancia de cada observación desde los valores medios de las v.independientes. Existen tablas para contrastar, pero en general se procede a identificar valores considerablemente altos respecto al resto.

Medidas para identificar observaciones influyentes:

- Influencias sobre coeficientes individuales:

DFBETA Mide el efecto del dato i-ésimo ejerce sobre β_j . Límites para la versión estandarizada: $\pm 2 n^{-1/2}$ (si $n < 50$ usar los límites de la normal)

- Medidas globales de influencia:

DFITTS Mide el efecto del dato i-ésimo ejerce en su propia predicción. Límites para la versión estandarizada: $\pm 2 [(k+2) / (n-k-2)]^{1/2}$

COVRATIO Representa el grado al que una observación tiene impacto sobre los errores estándar de los coeficientes. Límites: $1 \pm 3(k+1) / n$

Distancia de **Cook**: Localizar valores que exceden a $4 / (n-k-1)$

Detección de residuos atípicos: Los valores tipificados deben estar entre -3 y 3

Diagnósticos por caso^a

Número de caso	Residuo tip.	Consumo (l/100Km)	Valor pronosticado	Residuo bruto
35	10,176	26	5,95	20,05

a. Variable dependiente: Consumo (l/100Km)

Detección de puntos influyentes: Dist. De Mahalanobis, Cook, valor de influencia

Estadísticos sobre los residuos^a

	Mínimo	Máximo	Media	Desviación típ.	N
Valor pronosticado	5,95	21,05	11,27	3,440	392
Valor pronosticado tip.	-1,545	2,843	,000	1,000	392
Error típico del valor pronosticado	,107	,831	,210	,075	392
Valor pronosticado corregido	4,57	21,08	11,26	3,447	392
Residuo bruto	-5,16	20,05	,00	1,960	392
Residuo tip.	-2,618	10,176	,000	,995	392
Residuo estud.	-2,641	10,520	,001	1,011	392
Residuo eliminado	-5,25	21,43	,00	2,024	392
Residuo eliminado estud.	-2,661	12,433	,006	1,067	392
Dist. de Mahalanobis	,166	68,628	3,990	4,866	392
Distancia de Cook	,000	1,520	,007	,077	392
Valor de influencia centrado	,000	,176	,010	,012	392

a. Variable dependiente: Consumo (l/100Km)

El rango de valores para la distancia de Mahalanobis es elevado

Hay valores de la distancia de Cook superiores a $4 / (n-k-1) = 0.010$

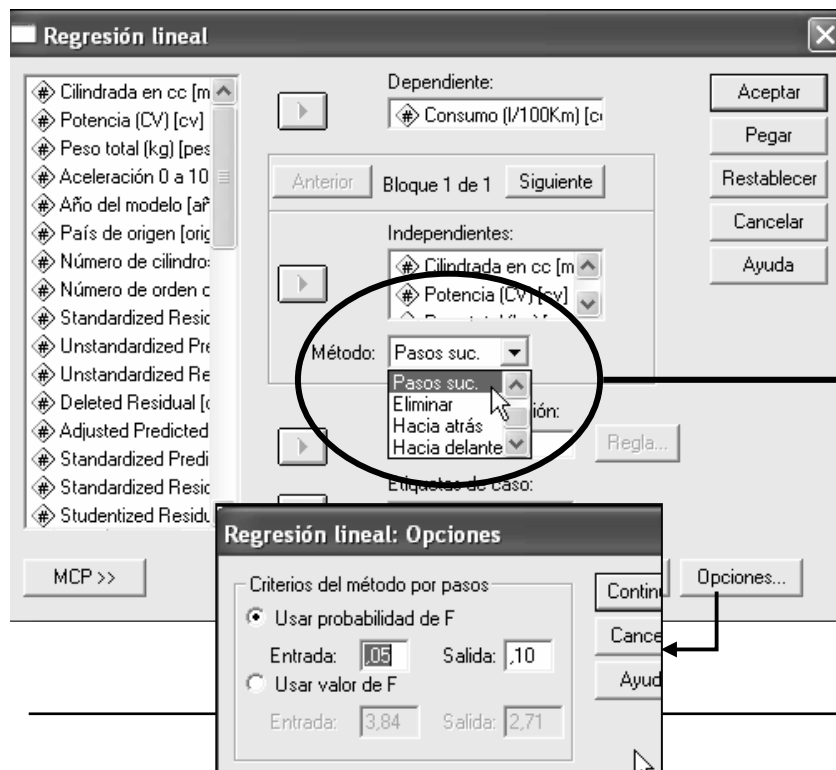
Medida de influencia. Límite ($k=4$): 0.038

Posible solución: Eliminar observaciones

Selección de un subconjunto óptimo de variables independientes

Objetivo: Seleccionar aquellas variables que sin ser redundantes proporcionen la mejor explicación de la v. dependiente.

Métodos secuenciales en SPSS: Hacia atrás, Hacia delante, Pasos sucesivos



En términos muy muy generales...

...Evalúan estadísticos F que controlan la entrada y salida de variables, además de las correlaciones parciales de la v. dependiente con cada regresor.

Método forward (hacia adelante)

Inicialmente no hay regresores, se van introduciendo uno a uno aquellos que tienen alta correlación parcial con la v. dependiente y que son significativos (valor F-entrar).

Variables introducidas/eliminadas

Modelo	Variables introducidas	Variables eliminadas	Método
1	Peso total (kg)	,	Hacia adelante (criterio: Prob. de F para entrar <= ,050)
2	Potencia (CV)	,	Hacia adelante (criterio: Prob. de F para entrar <= ,050)

a. Variable dependiente: Consumo (l/100Km)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,837 ^a	,700	,699	2,172	,700	909,085	1	390	,000
2	,868 ^b	,753	,752	1,972	,053	84,214	1	389	,000

a. Variables predictoras: (Constante), Peso total (kg)

b. Variables predictoras: (Constante), Peso total (kg), Potencia (CV)

Método backward (hacia atrás)

Inicialmente se incluyen todos las v. independientes, se van eliminando una a una las que van resultando significativas (valor F-salir).

Variables introducidas/eliminadas^b

Modelo	Variables introducidas	Variables eliminadas	Método
1	Aceleración 0 a 100 km/h (segundos), Peso total (kg), ^a Potencia (CV), Cilindrada en cc	,	Introducir
2	,	Aceleración 0 a 100 km/h (segundos)	Hacia atrás (criterio: Prob. de F para eliminar \geq ,100).
3	,	Cilindrada en cc	Hacia atrás (criterio: Prob. de F para eliminar \geq ,100).

a. Todas las variables solicitadas introducidas

b. Variable dependiente: Consumo (l/100Km)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,869 ^a	,755	,752	1,970	,755	297,956	4	387	,000
2	,869 ^b	,755	,753	1,968	,000	,180	1	389	,672
3	,868 ^c	,753	,752	1,972	-,002	2,456	1	390	,118

a. Variables predictoras: (Constante), Aceleración 0 a 100 km/h (segundos), Peso total (kg), Potencia (CV), Cilindrada en cc

b. Variables predictoras: (Constante), Peso total (kg), Potencia (CV), Cilindrada en cc

c. Variables predictoras: (Constante), Peso total (kg), Potencia (CV)

Método Stepwise (pasos sucesivos)

Combina los dos métodos anteriores definiendo un procedimiento en el que las variables independientes entran o salen del modelo dependiendo de su significación (valores F-entrar y F-salir).

Variables introducidas/eliminadas^a

Modelo	Variables introducidas	Variables eliminadas	Método
1	Peso total (kg)	,	Por pasos (criterio: Prob. de F para entrar \leq ,050, Prob. de F para salir \geq ,100).
2	Potencia (CV)	,	Por pasos (criterio: Prob. de F para entrar \leq ,050, Prob. de F para salir \geq ,100).

a. Variable dependiente: Consumo (l/100Km)

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Estadísticos de cambio				
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
1	,837 ^a	,700	,699	2,172	,700	909,085	1	390	,000
2	,868 ^b	,753	,752	1,972	,053	84,214	1	389	,000

a. Variables predictoras: (Constante), Peso total (kg)

b. Variables predictoras: (Constante), Peso total (kg), Potencia (CV)



Resumen

Pasos a seguir en un análisis de regresión

Paso 1. Objetivos del análisis

Paso 2. Diseño de la investigación mediante regresión múltiple

Paso 3. Supuestos del análisis

Paso 4. Estimación del modelo de regresión y valoración global del ajuste

Paso 5. Interpretación y validación de los resultados.
