Estadística

INGENIERÍA MULTIMEDIA

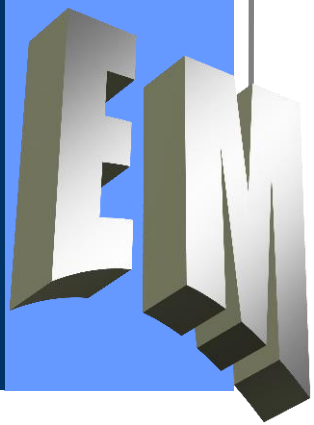
Violeta Migallón

Dept. de Ciència de la Computació i Intel·ligència artificial
Dpto. de Ciencia de la Computación e Inteligencia artificial

Introducción

Punto 1

El problema de la dependencia entre variables consiste en el estudio de la relación existente entre dos o más características de los elementos de una población. Aquí nos vamos a dedicar sólo al estudio de la relación entre dos características de una población que puedan ser representadas mediante variables medibles. La relación entre variables cualitativas o categorizadas se tratará en otros temas.



Introducción

Punto 1

- Dependencia funcional: Decimos que una variable Y depende funcionalmente de otra X cuando podemos establecer una aplicación f que nos transforme los elementos de X en elementos de Y, es decir Y = f(X).

Ejemplo: Relación existente entre el espacio, e, recorrido por un móvil y su velocidad v para un tiempo dado t_0:

$$e = v \cdot t_0$$

- Dependencia estadística: Si la relación entre ambas variables no es exacta sino que hay un componente aleatorio que no es posible controlar.

Ejemplo: Relación existente entre estatura y peso, consumo y renta, ….

Introducción

Punto 1

El análisis **de regresión** consiste en ajustar una función f que represente la relación estadística entre dos variables. Una de las dos variables se considera la variable independiente, X, y la otra, la variable dependiente Y, de forma que mediante la función ajustada, para cada valor de X se obtenga un valor f(X) que pueda servir como predicción del valor de la variable Y.

Según la forma que adopte el diagrama de dispersión de la muestra bidimensional se elegirá el tipo de curva que se ajuste al diagrama.

Diagrama de dispersión o nube de puntos

Punto 1

Punto 2

Un **diagrama de dispersión** o nube de puntos es un tipo de gráfico que utiliza las coordenadas cartesianas para mostrar los valores de dos variables cuantitativas X e Y para un conjunto de datos. Los datos se muestran como un conjunto de puntos, cada uno con el valor de una variable que determina la posición en el eje horizontal y el valor de la otra variable determinado por la posición en el eje vertical.

Ejemplo

Diagrama de dispersión o nube de puntos

Punto 1

Punto 2

X	Y
1.0	1.5
3.0	2.7
6.1	4.32
4.3	1.0
6.0	3.2

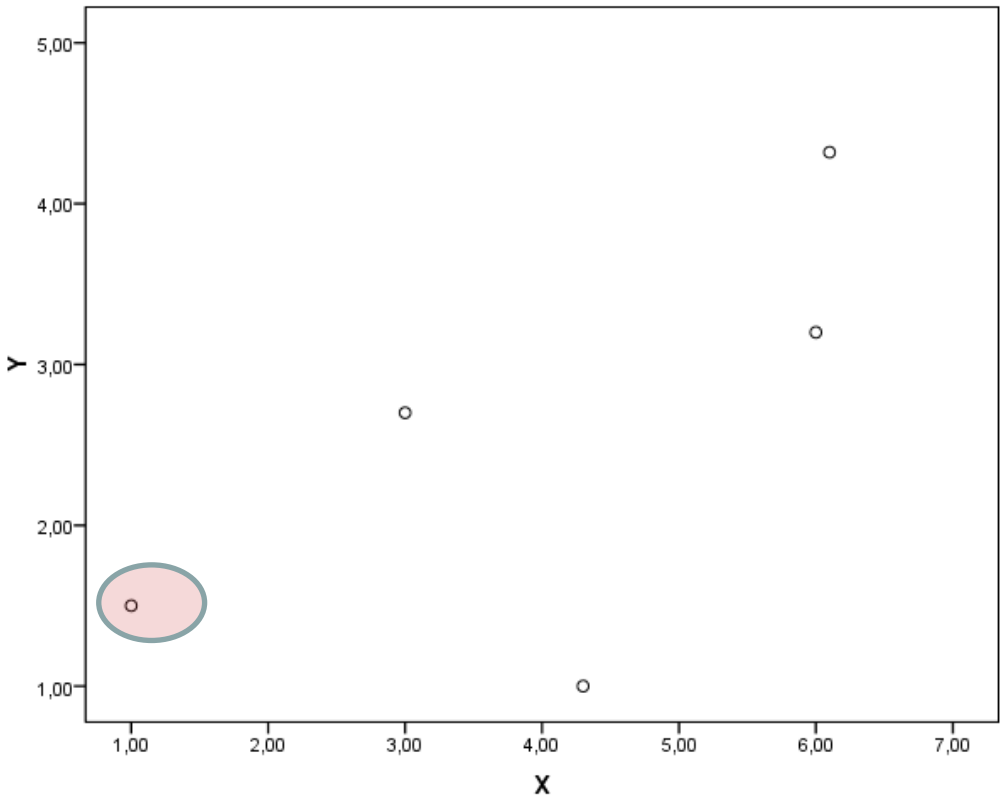
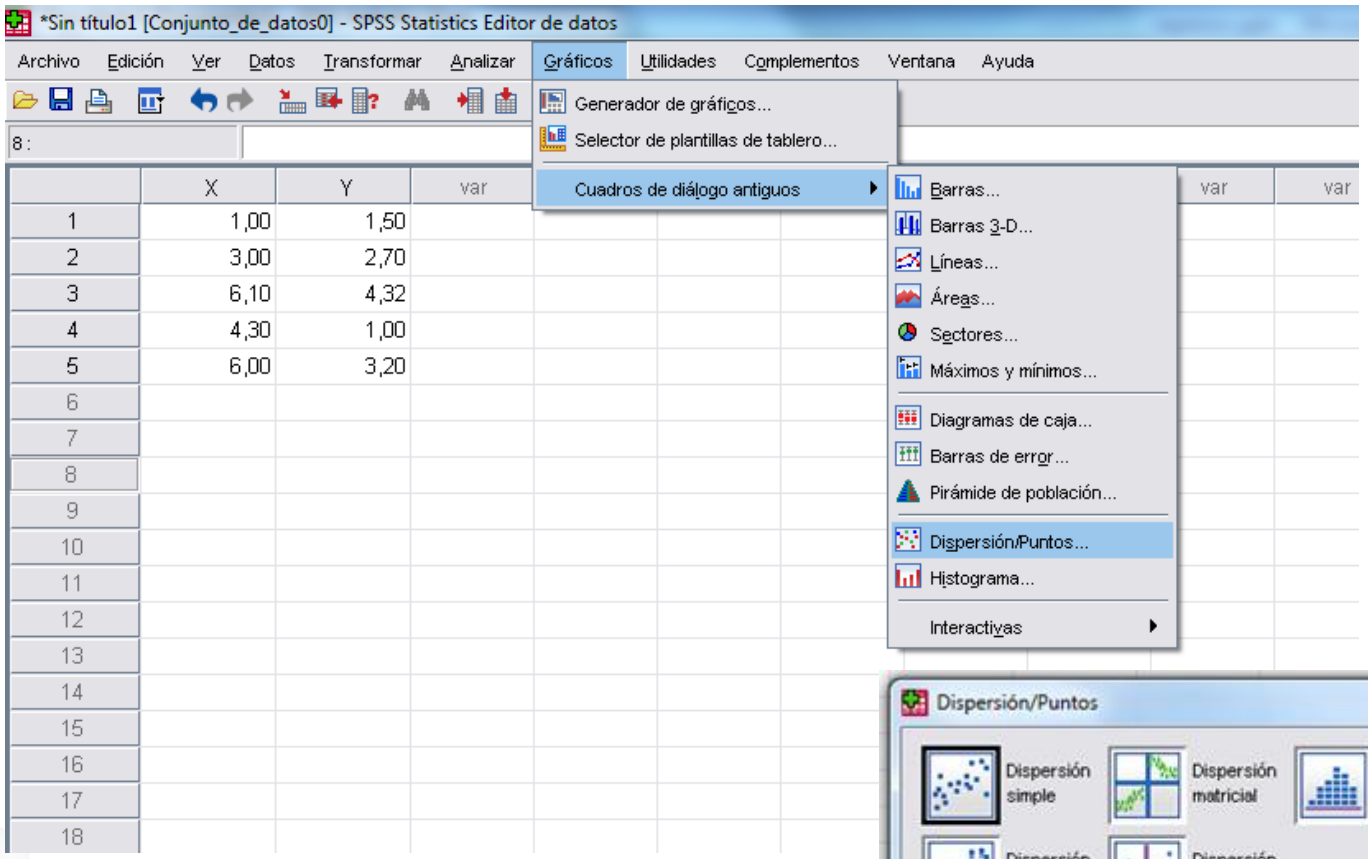


Diagrama de dispersión o nube de puntos con SPSS

- Punto 1
- Punto 2



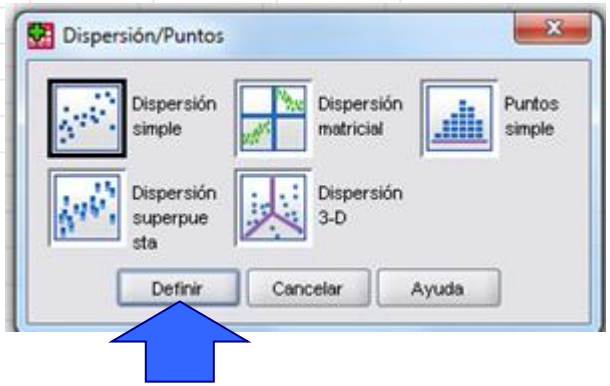


Diagrama de dispersión o nube de puntos con SPSS

Punto 1

Punto 2

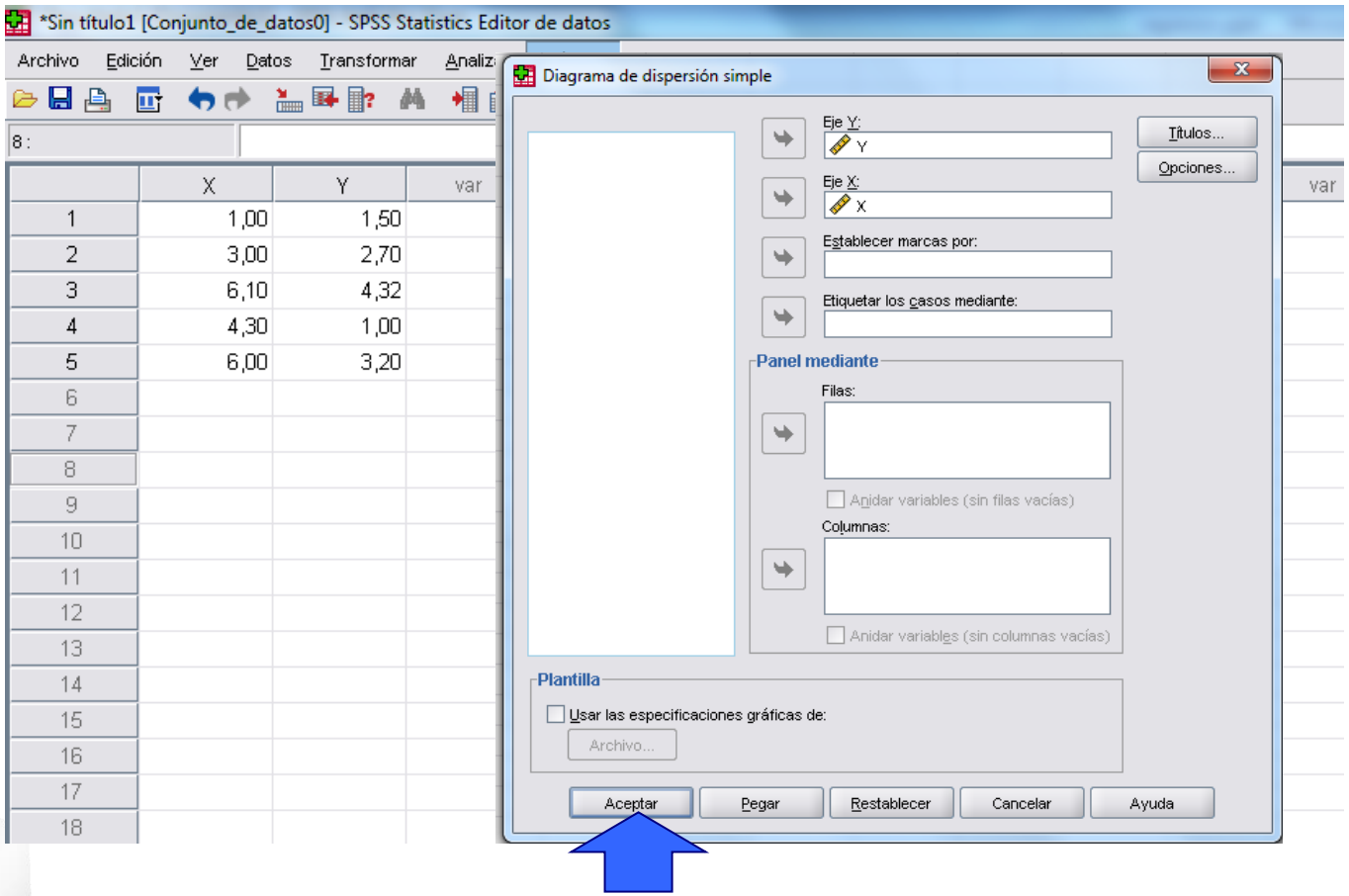


Diagrama de dispersión o nube de puntos con SPSS

Punto 1

Punto 2

X	Y
1.0	1.5
3.0	2.7
6.1	4.32
4.3	1.0
6.0	3.2

SPSS

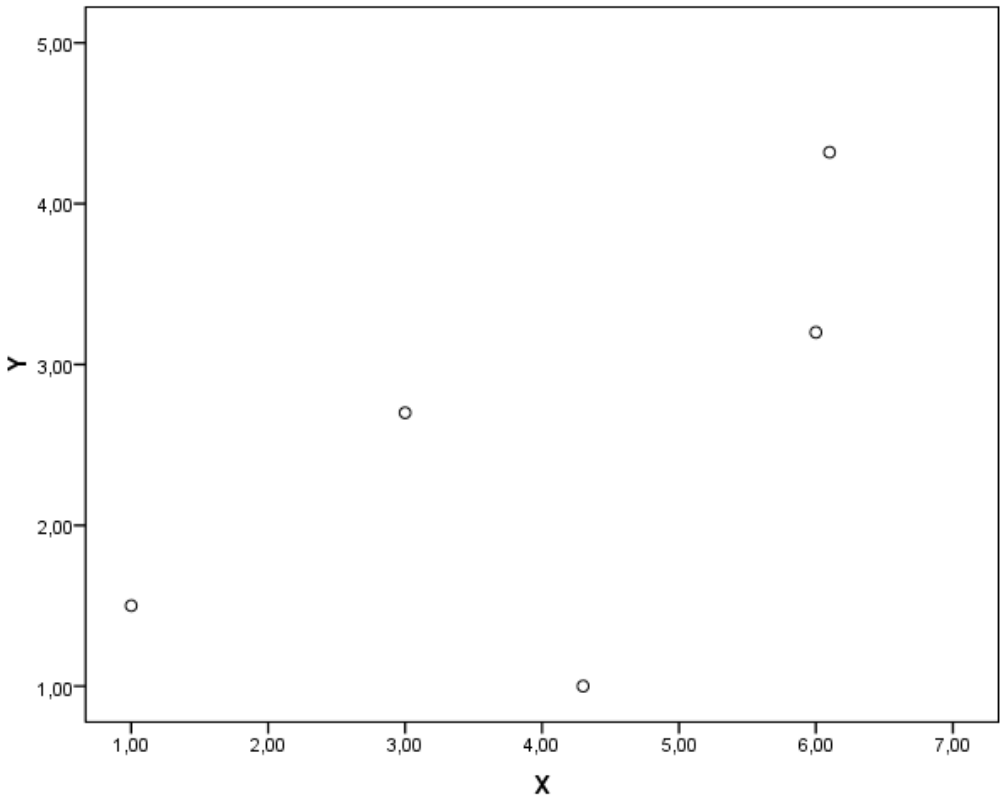


Diagrama de dispersión o nube de puntos

Punto 1

Punto 2

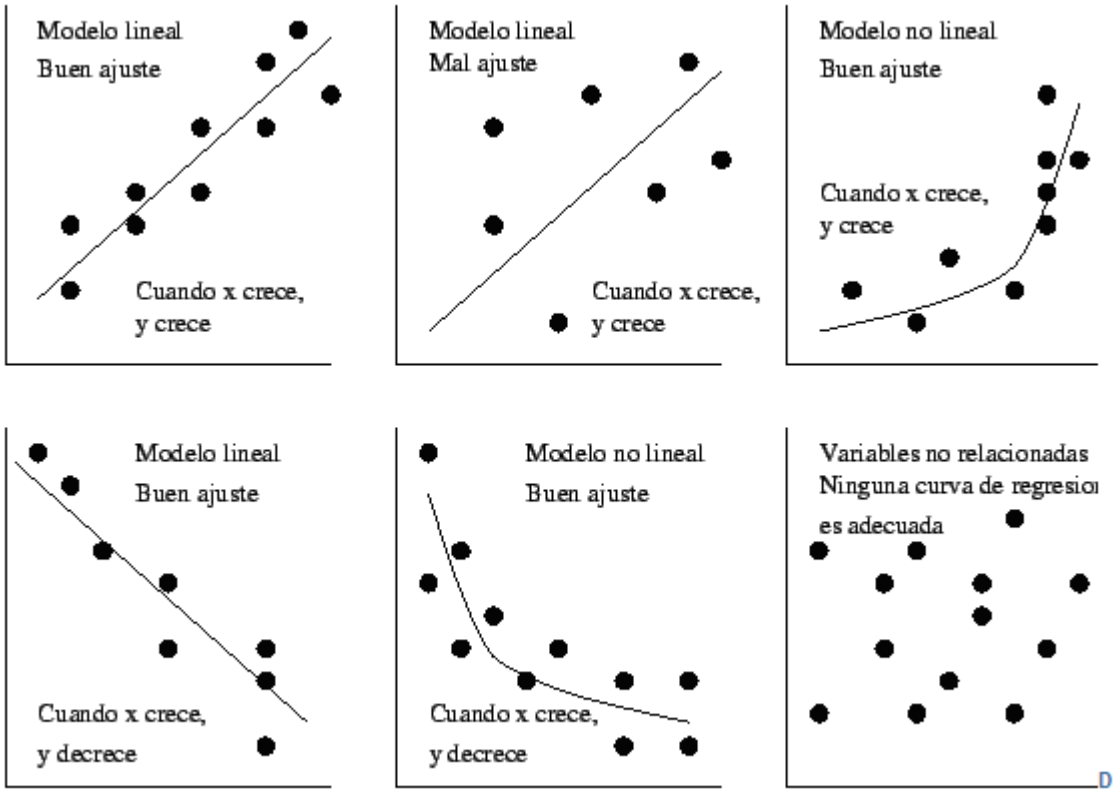


Figura 1. Diferentes nubes de puntos y modelos de regresión para ellas

Nos restringiremos al ajuste lineal

Medidas de relación lineal: Covarianza y correlación

- Consideremos una muestra bidimensional de tamaño n:

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

Para medir la idoneidad del ajuste lineal podemos usar:

Covarianza:

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

o equivalentemente

$$\mathrm{Cov}(X,Y) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - n\,\bar{x}\,\bar{y}\right]$$

Medidas de relación lineal: Covarianza y correlación

Punto 1

Punto 2

Punto 3

La covarianza tiene dos inconvenientes:

- Los puntos atípicos influyen considerablemente en el resultado
- Depende de las unidades de medida de las variables

Para evitar estos inconvenientes utilizamos el **coeficiente de correlación de Pearson:**

$$R = \frac{\mathrm{Cov}(X,Y)}{S_x S_y}$$

Medidas de relación lineal: Covarianza y correlación

Punto 1

Punto 2

Punto 3

Coeficiente de correlación de Pearson:

$$R = \frac{\mathrm{Cov}(X,Y)}{S_x S_y}$$

- R es una medida de la precisión de la dependencia lineal entre X e Y.
- $|R| \leq 1$.
- Si R está próximo a 1 indica una gran dependencia lineal creciente.
- Si R está próximo a -1 indica una gran dependencia lineal decreciente.
- Si R está próximo a 0 la dependencia lineal es muy pobre. Si R=0 no hay dependencia lineal y se dice que las variables son incorreladas.
- $100R^2$ representa el porcentaje de la varianza común entre las dos variables. R^2 se denomina **coeficiente de determinación**.

Recta de regresión de Y sobre X

Punto 1

Punto 2

Punto 3

Sean dos variables medibles X e Y entre las que suponemos que existe dependencia estadística lineal. Buscamos la recta de regresión:

Y=aX+b

Recta que mejor representa la relación de dependencia lineal entre dichas variables.

Método de los mínimos cuadrados:

Recta que minimiza la suma de las diferencias cuadráticas entre los valores observados y los ajustados.

Y=aX+b

$$a = \frac{\mathrm{Cov}(X,Y)}{S_x^2}$$

$$b = \bar{y} - a\bar{x}$$

Recta de regresión de Y sobre X

- Con ayuda de la recta de regresión se puede predecir el valor que tomará una variable para un valor determinado de la otra.
- La predicción de Y para $X = x_0$ será simplemente el valor obtenido en la recta de regresión de Y sobre X al sustituir el valor de X por x_0.
- La fiabilidad de esta predicción será mayor cuanto mayor sea la correlación entre las variables

$$a = \frac{\mathrm{Cov}(X,Y)}{S_x^2}$$

$y_0 = ax_0 + b$

$$b = \bar{y} - a\bar{x}$$

Regresión lineal simple con SPSS

Ejercicio: Se ha solicitado a un grupo de estudiantes de Ingeniería Multimedia información sobre el número de horas que han dedicado al estudio de un examen y la calificación del mismo. Los datos se han incluido en la siguiente tabla.

X: horas de estudio	20	16	34	10	23
Y: calificación	6.5	6	8	4	7

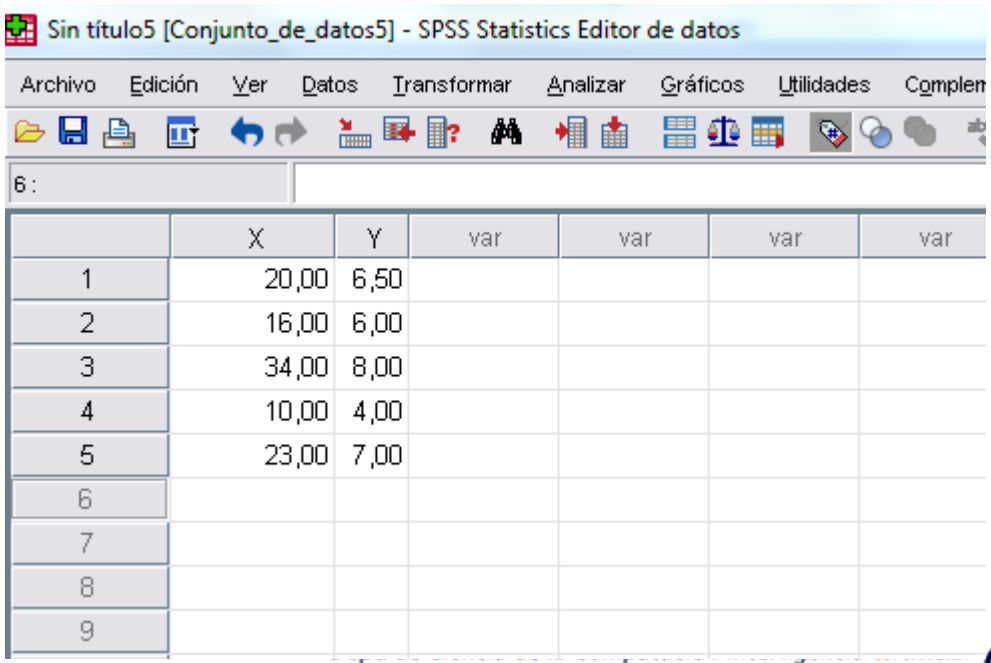
- Calcula la covarianza, el coeficiente de correlación y el coeficiente de determinación.
- Determina la ecuación de la recta de regresión de Y sobre X.
- Si una persona ha estudiado 15 horas , ¿cuánto cabe esperar que haya sacado en el examen?

Regresión lineal simple con SPSS

Ejercicio: Calcula la covarianza y el coeficiente de correlación y el de determinación.

X: horas de estudio	20	16	34	10	23
Y: calificación	6.5	6	8	4	7

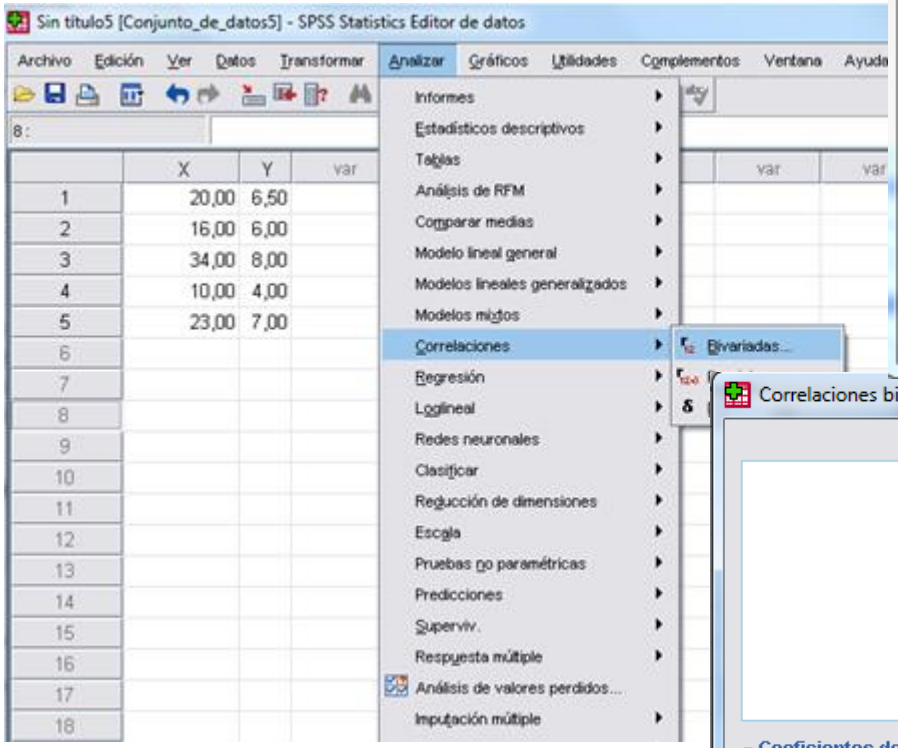
Se introducen los datos en el SPSS

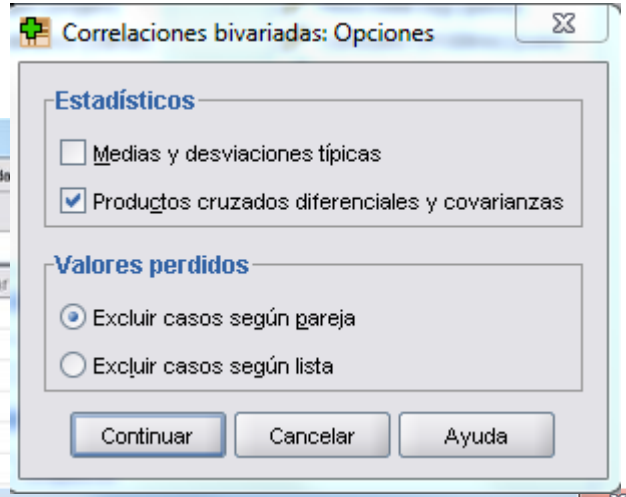


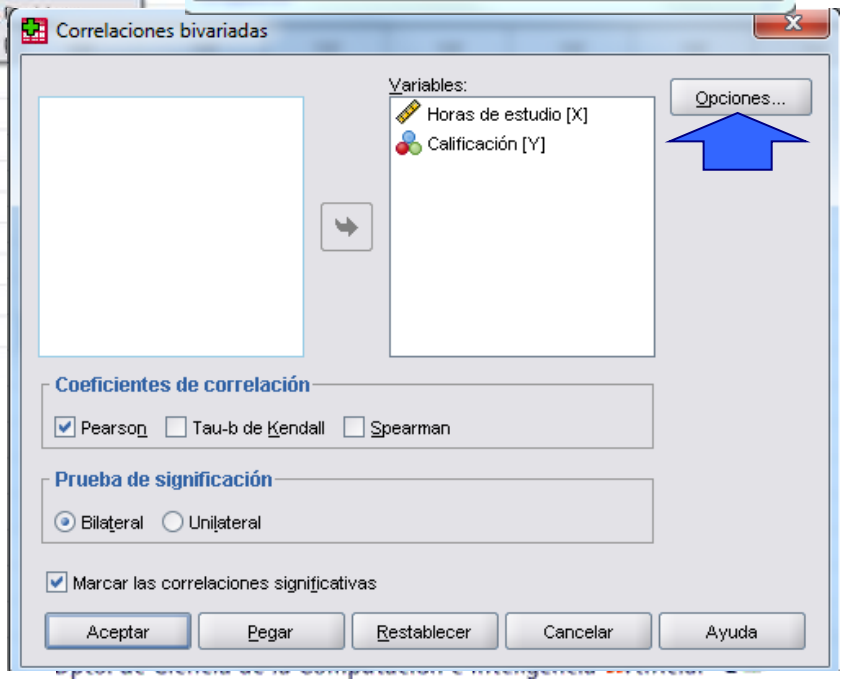
Regresión lineal simple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5

Ejercicio:







Regresión lineal simple con SPSS

Punto 1

Punto 2

Punto 3

Punto 4

Punto 5

Ejercicio:

Correlaciones

		Horas de estudio	Calificación
Horas de estudio (X)	Correlación de Pearson	1	,945*
	Sig. (bilateral)		,015
	Suma de cuadrados y productos cruzados	319,200	50,100
	Covarianza	79,800	12,525
	N	5	5
Calificación (Y)	Correlación de Pearson	,945*	1
	Sig. (bilateral)	,015	
	Suma de cuadrados y productos cruzados	50,100	8,800
	Covarianza	12,525	2,200
	N	5	5

*. La correlación es significante al nivel 0,05 (bilateral).

- Covarianza: Cov(X,Y)=12.525
- Correlación de Pearson:

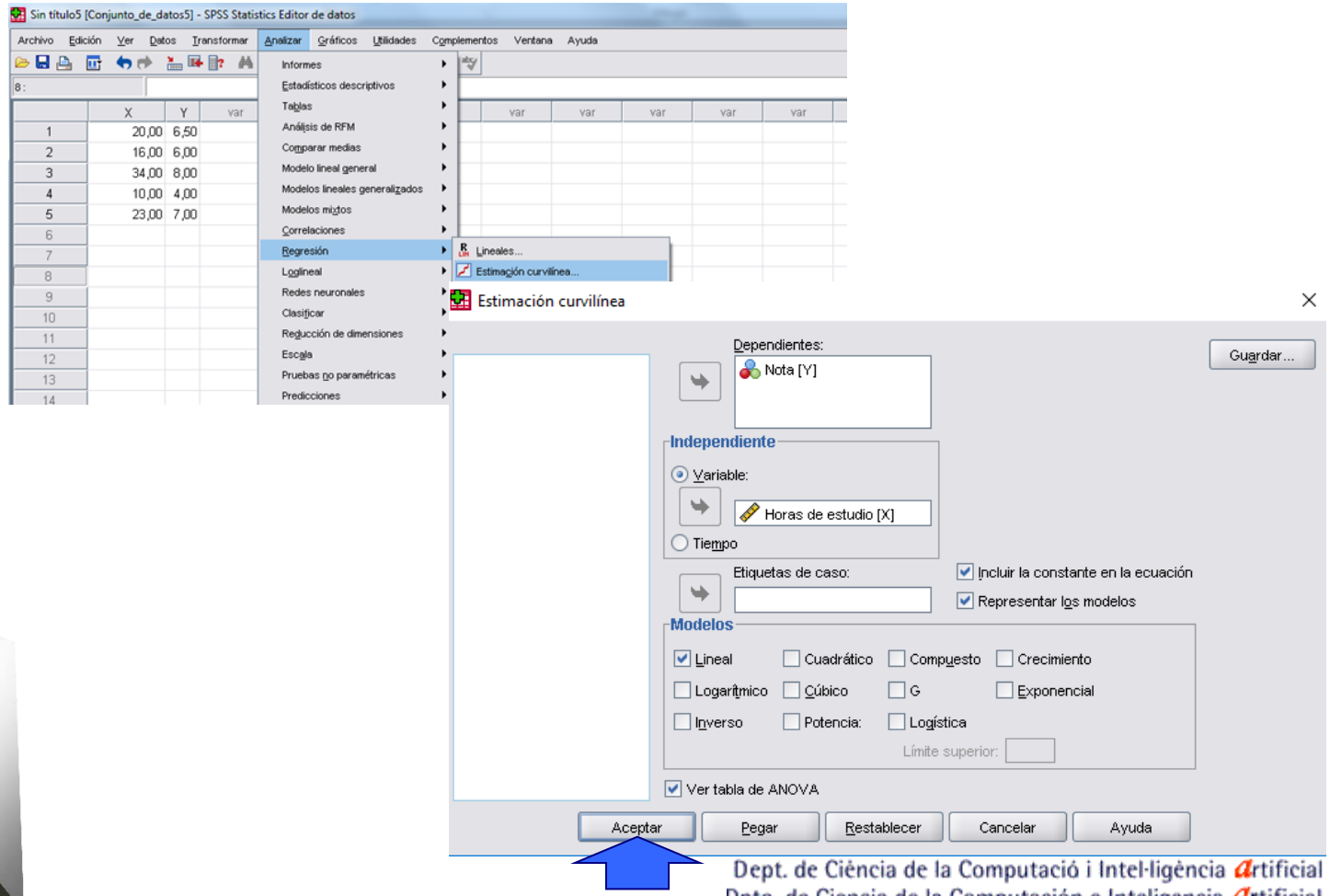
R=0.945 →Fuerte dependencia lineal creciente, el ajuste lineal es apropiado.

- Coeficiente de determinación: R^2=0.893 →La variable horas de estudio explica el 89.3% de la variabilidad de la variable calificación.

Regresión lineal simple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5

Ejercicio: Determina la ecuación de la recta de regresión de Y sobre X.



Regresión lineal simple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5

Ejercicio:

Resumen del modelo y estimaciones de los parámetros

Variable dependiente:Calificación

Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,894	25,188	1	3	,015	3,067	,157

La variable independiente esHoras de estudio.

Coeficientes

	Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
	B	Error típico	Beta		
Horas de estudio	,157	,031	,945	5,019	,015
(Constante)	3,067	,691		4,438	,021

Sig.<0.05 o equivalentemente |t|>2, entonces la variable es significativa en el modelo

Recta de regresión

Y=0.157X+3.067

NOTA: Según este modelo, manteniendo el resto constante un aumento de una hora de estudio produciría un aumento de 0.157 en la calificación ya que la variable horas de estudio es significativa en el modelo.

Regresión lineal simple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5

Ejercicio: Si una persona ha estudiado 15 horas , ¿cuánto cabe esperar que haya sacado en el examen? 5.422

Resumen del modelo y estimaciones de los parámetros

Variable dependiente:Calificación

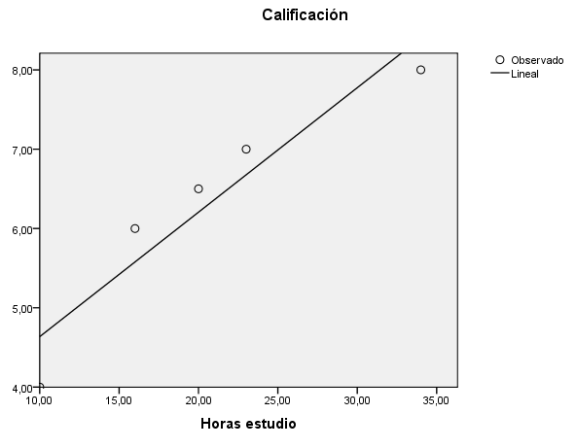
Ecuación	Resumen del modelo					Estimaciones de los parámetros	
	R cuadrado	F	gl1	gl2	Sig.	Constante	b1
Lineal	,894	25,188	1	3	,015	3,067	,157

La variable independiente esHoras de estudio.

Recta de regresión

$Y=0.157X+3.067$

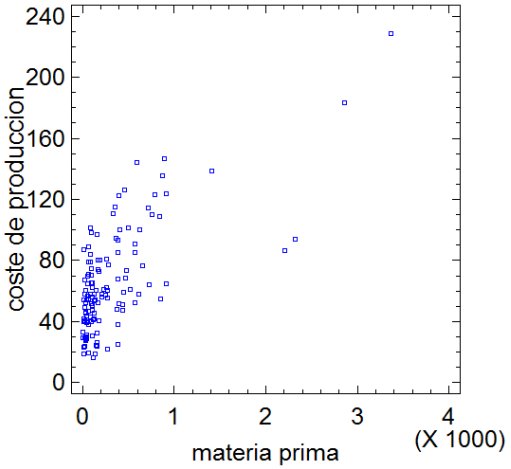
$Y=0.157\cdot 15+3.067=5.422$

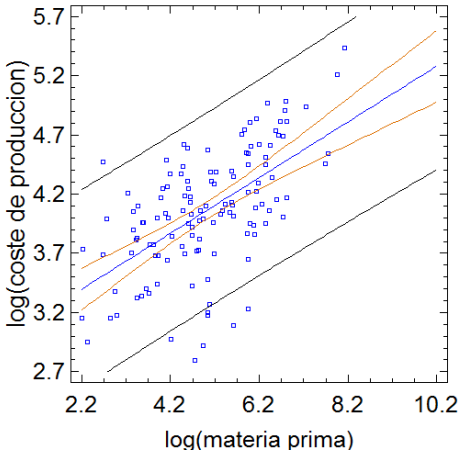


Regresión lineal simple con SPSS

Validación y diagnosis del modelo: Para que las conclusiones de nuestro modelo, en problemas reales donde se desea hacer pronósticos, sean las correctas, los datos deben satisfacer las siguientes hipótesis:

1. **Linealidad:** se puede analizar a priori con el gráfico de dispersión.





2. **Homocedasticidad:** la variabilidad de los residuos (valores reales – valores estimados con la recta) debe ser constante. Se puede analizar también a priori mediante el gráfico de dispersión. Gráficamente, la nube de puntos de los datos debe tener una anchura más o menos constante a lo largo de la recta de regresión.

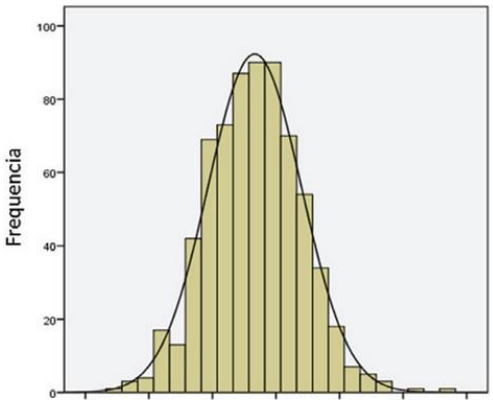
.

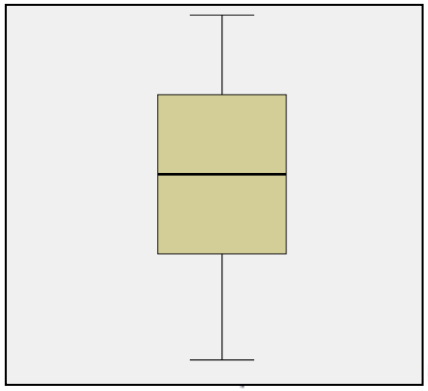
Regresión lineal simple con SPSS

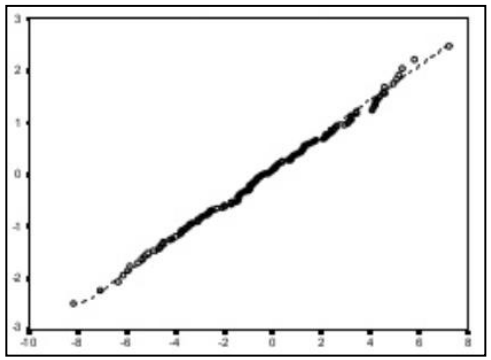
Punto 1
Punto 2
Punto 3
Punto 4
Punto 5

3. **Independencia**: Los residuos deben ser independientes, es decir no disminuyen o aumentan siguiendo una pauta discernible. Una primera medida para tratar de evitar la dependencia consiste en utilizar muestras aleatorias.
La falta de independencia, se produce fundamentalmente cuando se trabaja con series temporales por lo que no vamos a utilizar este tipo de variables.

4. **Normalidad:** Los residuos deben seguir una distribución Normal. Este tipo de distribuciones las veremos en el tema 5 y en el tema 7 estudiaremos como contrastar la normalidad.
Por el momento se puede hacer un análisis gráfico de los residuos (que se habrán guardado previamente al hacer la regresión). Este análisis se puede hacer desde la opción Analizar→Estadísticos Descriptivos →Explorar, mediante histogramas (debe asemejarse a una campana de Gauss), gráficos caja o con el gráfico probabilístico normal de los residuos. No obstante, para muestras grandes el supuesto de distribución normal no es crucial.







Regresión lineal simple con SPSS

Cuando las hipótesis del modelo no se cumplen es necesario transformar los datos, de manera que los datos transformados cumplan las hipótesis. El uso de logaritmos es especialmente útil cuando hay falta de linealidad o hay heterocedasticidad.

Interpretación de los coeficientes: Se analizan una vez estudiada la significatividad de los mismos.

Y=aX+b → Un incremento de *X* en una unidad incrementaría Y en *a* unidades.
ln(Y)=aX+b → Un incremento de *X* en una unidad incrementaría *Y* el 100*a* %.
ln(Y)=aln(X)+b → Un incremento de *X* del 1% incrementaría *Y* el *a* %.
Y=aln(X)+b → Un incremento de *X* del 1% incrementaría Y en $a/100$ *unidades.*

Regresión lineal multiple con SPSS

Se trata conocer el valor de una variable respuesta Y a partir de más de una variable explicativa.

$$Y=a_0+a_1X_1+a_2X_2+\ldots+a_nX_n$$

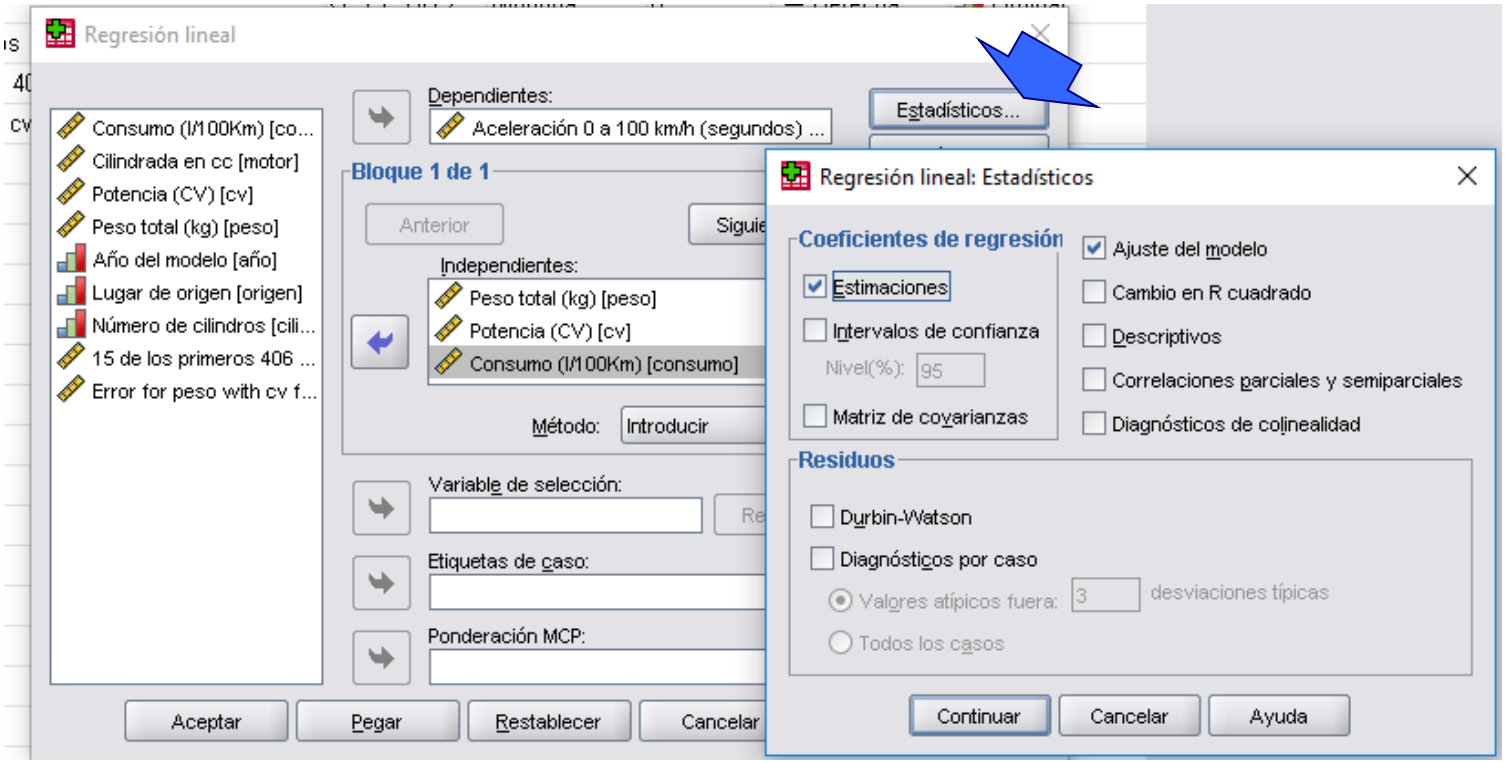
Requiere el cumplimiento de hipótesis análogas a las del modelo de regresión lineal simple: Linealidad, homocedasticidad, independencia y normalidad.

Ejercicio: Se desea realizar un modelo de regresión múltiple para ver si se puede explicar la variable aceleración del fichero car.sav en función de las variables: peso, potencia y consumo.

Regresión lineal multiple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5
- Punto 6

Desde Analizar →Regresión → lineales, podemos hacer tanto regresiones lineales simples como múltiples.



Regresión lineal multiple con SPSS

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.789[a]	.622	.619	1.71464

a. Predictores: (Constante), Consumo (l/100Km), Potencia (CV), Peso total (kg)

Al aumentar el numero de variables en el modelo, R^2 puede aumentar aunque las variables no sean significativas. R^2 ajustado nos explicará de forma más correcta la variabilidad de la variable dependiente que es explicada en el modelo de regresión múltiple.

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	18.451	.317		58.192	.000
	Peso total (kg)	.007	.001	.723	10.808	.000
	Potencia (CV)	-.096	.005	-1.323	-19.789	.000
	Consumo (l/100Km)	.008	.044	.011	.175	.861

a. Variable dependiente: Aceleración 0 a 100 km/h (segundos)

La variable consumo no es significativa (sig.>0.05), por tanto la quitamos del modelo y volvemos a realizar la regresión lineal

Regresión lineal multiple con SPSS

- Punto 1
- Punto 2
- Punto 3
- Punto 4
- Punto 5
- Punto 6

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.792[a]	.628	.626	1.72654

a. Predictores: (Constante), Potencia (CV), Peso total (kg)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	18.513	.317		58.334	.000
	Peso total (kg)	.007	.001	.720	12.027	.000
	Potencia (CV)	-.097	.004	-1.320	-22.047	.000

a. Variable dependiente: Aceleración 0 a 100 km/h (segundos)

Todas las variables son significativas (sig.<0.05), por tanto el modelo de regresión lineal múltiple obtenido sería:

Aceleración=18.513+0.007*Peso total-0.097*Potencia

Explica cómo afecta al tiempo de aceleración un incremento de una unidad en el peso si el resto se mantiene constante ¿y un aumento de una unidad en la potencia?

Multicolinealidad entre variables

Si hay variables explicativas con mucha relación entre ellas, el modelo no va a poder distinguir qué parte de *Y* es explicada por cada variable. En este caso hay variables independientes que contienen prácticamente la misma información y no tendrá sentido incluirlas todas en el modelo.

El problema de la multicolinealidad aparece frecuentemente

Detección de la colinealidad:

Cuando en la regresión simple las variables son significativas y al introducir nuevas variables, dejan de ser significativas hay colinealidad. Dichas variables (que dejan de ser significativas) son las que se eliminarán del modelo.

Variables dummies o ficticias

Una variable ficticia (o dummy) es una variable artificial que suele tomar valores 0 y 1.
Nos es útil cuando disponemos de información cualitativa acerca de un conjunto de elementos o personas (sexo, raza,…).
Ejercicio: Consideremos el modelo de regresión múltiple anterior. Queremos explicar la diferencia de aceleración de los coches según sean de EEUU o no.

Se crea una variable eeuu que tenga un 1 para los coches de EEUU y 0 en el resto de casos, por ejemplo con recodificar variables, e introducimos la nueva variable en el análisis de regresión:

Variables dummies o ficticias

Ejercicio:

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.795[a]	.632	.629	1.71918

a. Predictores: (Constante), eeuu, Potencia (CV), Peso total (kg)

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Error estándar	Beta	t	Sig.
1	(Constante)	18.317	.329		55.602	.000
	Peso total (kg)	.008	.001	.776	11.884	.000
	Potencia (CV)	-.097	.004	-1.329	-22.232	.000
	eeuu	-.469	.223	-.081	-2.100	.036

a. Variable dependiente: Aceleración 0 a 100 km/h (segundos)

Aceleración=18.317+0.008*Peso total-0.097*Potencia-0.469*eeuu

Variables dummies o ficticias

Ejercicio: Ahora el modelo ha quedado:

Aceleración=18.317+0.008*Peso total-0.097*Potencia-0.469*eeuu

Como la variable eeuu es significativa, obtenemos que un coche de EEUU con el mismo peso y potencia tardará 0.469 segundos menos en pasar de 0 a 100 km/h que otro que no sea de EEUU