Práctica 9

REGRESION LINEAL Y CORRELACIÓN

Objetivos:

En esta práctica utilizaremos el paquete SPSS para estudiar la regresión lineal entre dos variables y la regresión lineal múltiple entre una variable dependiente y varias independientes obteniendo la estimación de los parámetros del modelo y realizando los contrastes estadísticos oportunos para verificar la validez del modelo construido.

Índice:

1. REGRESIÓN SIMPLE: RECTA DE REGRESIÓN Y COEFICIENTE DE CORRELACION

En la regresión simple se ajusta una recta de regresión de la forma

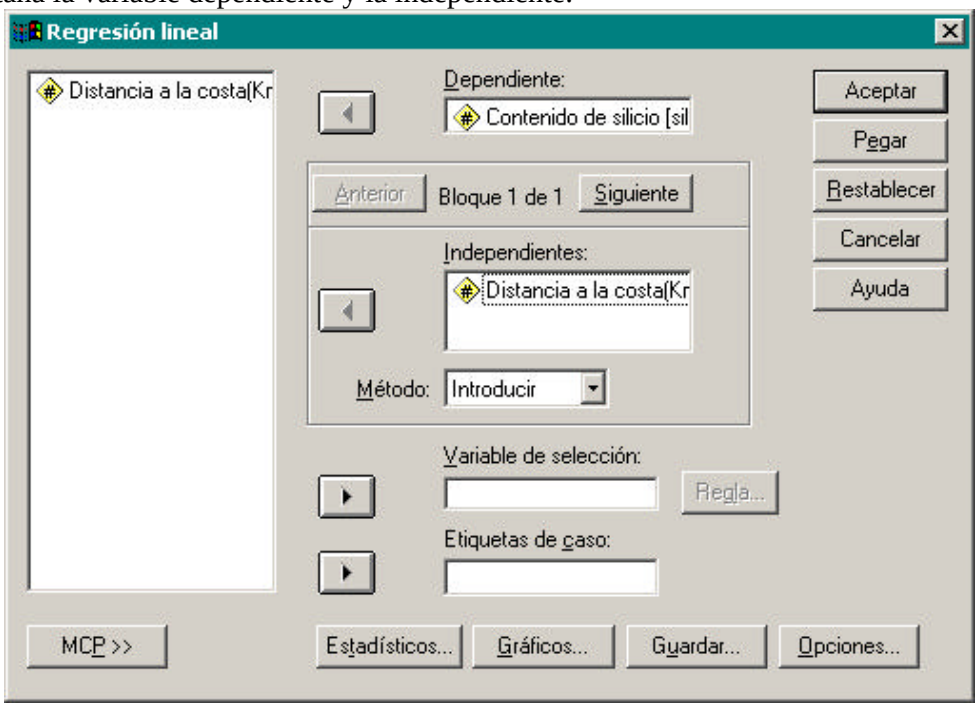
$$Y = a + b\,X$$

para, en el caso de que el modelo lineal sea el adecuado, estimar valores de la variable dependiente, Y, para ciertos valores de la variable independiente, X.

Ejercicio 1:
En el fichero SILICIO están los datos del contenido de silicio en muestras de agua de mar recogida a ciertas distancias prefijadas de la costa. Se trata de estudiar la relación lineal entre las dos variables, y predecir el contenido de silicio del agua en función de la distancia a la costa.

1.1.- Preparación del análisis

Para realizar el estudio de la relación lineal entre ambas variables seleccionamos, en el menú principal **Analizar / Regresión / Lineal...**. En primer lugar debemos poner en la ventana la variable dependiente y la independiente.



Puedes obtener información adicional de los resultados del análisis con las siguientes opciones de los botones de la ventana:

Regresión lineal: Guardar...

Puedes guardar los valores pronosticados, los residuos y otros estadísticos útiles para los diagnósticos. Cada selección añade una o más variables nuevas a tu archivo de datos activo.

- Valores pronosticados. Son los valores que el modelo de regresión pronostica para cada caso.

- Intervalos de pronóstico. Los límites superior e inferior para los intervalos de predicción individual y promedio .
- Residuos. El valor actual de la variable dependiente menos el valor pronosticado por la ecuación de regresión.

Regresión lineal: Gráficos...

Los gráficos pueden ayudar a validar los supuestos de normalidad, linealidad e igualdad de las varianzas. También son útiles para detectar valores atípicos, observaciones poco usuales y casos de influencia. Tras guardarlos como nuevas variables, dispondrás en el Editor de datos de los valores pronosticados, los residuos y otros valores diagnósticos, con los cuales podrás crear gráficos respecto a las variables independientes. Se encuentran disponibles los siguientes gráficos:

- Diagramas de dispersión. debemos representar los residuos tipificados frente a los valores pronosticados tipificados para contrastar la linealidad y la igualdad de las varianzas.
- Gráficos de residuos tipificados. Puedes obtener histogramas de los residuos tipificados y gráficos de probabilidad normal que comparen la distribución de los residuos tipificados con una distribución normal.

1.2.- Interpretación de los resultados

En el resultado del análisis de la regresión se presenta la siguiente información:

Resumen del modelo

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,936[a]	,876	,871	.3936

a. Variables predictoras: (Constante), Distancia a la costa(Km)

b. Variable dependiente: Contenido de silicio

- R: Coeficiente de correlación (en valor absoluto) entre los valores observados y pronosticados en la variable dependiente. Su valor tiene un rango de 0 a 1. Un valor pequeño indica que hay poca o ninguna relación lineal entre la variable dependiente y la variable independiente.
- R cuadrado: Medida de la bondad de ajuste de un modelo lineal. Recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable dependiente explicada por el modelo de regresión. Sus valores van desde 0 a 1. Los valores pequeños indican que el modelo no se ajusta bien a los datos.
- R cuadrado corregida: La R cuadrado muestral tiende a estimar de manera demasiado optimista el ajuste del modelo en la población. Habitualmente el modelo no se ajusta a la población tan bien como se ajusta a la muestra de la que se ha derivado. La R cuadrado corregida intenta corregir la R cuadrado para reflejar más estrechamente la bondad de ajuste en la población.
- Error típico de la estimación: Medida de cuánto puede variar el valor de un estadístico de contraste de muestra en muestra. Es la desviación típica de la distribución muestral de un estadístico. En este caso nos da la desviación típica residual.

ANOVA

Presenta el análisis de varianza del modelo de regresión lineal. Contrasta si el modelo de regresión lineal es una buena explicación de la variabilidad de la variable dependiente.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	24,118	1	24,118	155,656	,000[a]
	Residual	3,409	22	,155		
	Total	27,526	23			

a. Variables predictoras: (Constante), Distancia a la costa(Km)

b. Variable dependiente: Contenido de silicio

Coeficientes

Para cada uno de los parámetros de la recta de regresión se presenta su estimación, el error estándar y el resultado del contraste de hipótesis en el que la hipótesis nula es que el correspondiente parámetro vale cero.

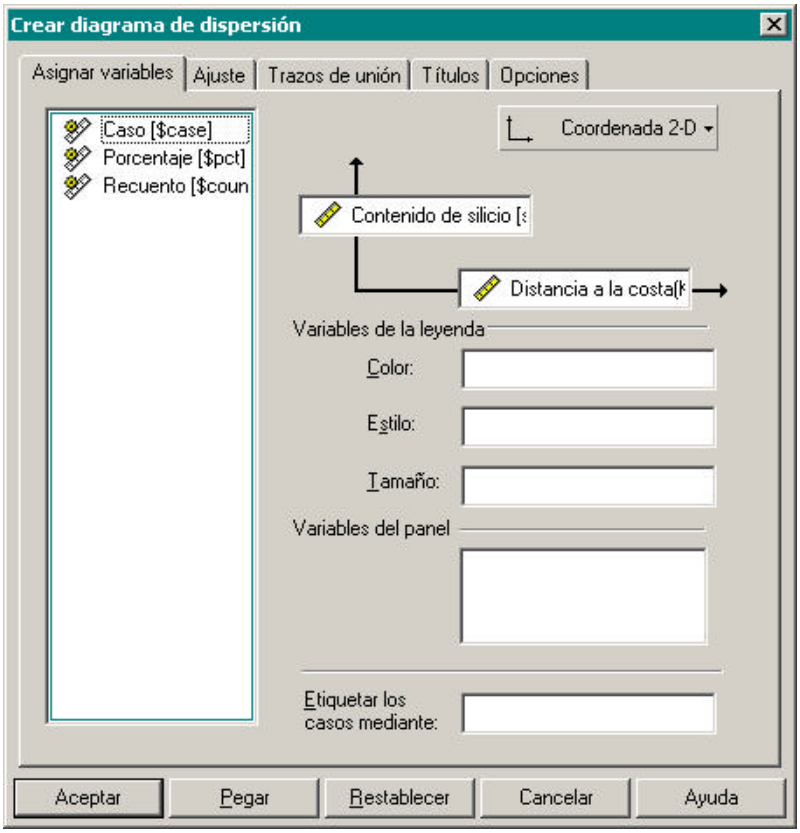
Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	5,945	,162		36,660	,000
	Distancia a la costa(Km)	-6,06E-02	,005	-,936	-12,476	,000

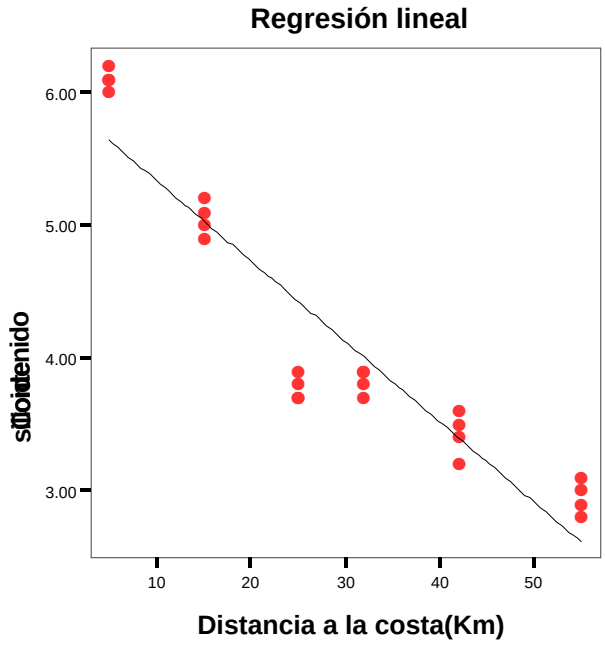
a. Variable dependiente: Contenido de silicio

1.3.- Obtención de la recta de regresión

Se puede realizar en el menú **Gráficos / Interactivos / Diagrama de dispersión...**
Se arrastra la variable dependiente al eje Y y la independiente al eje X.

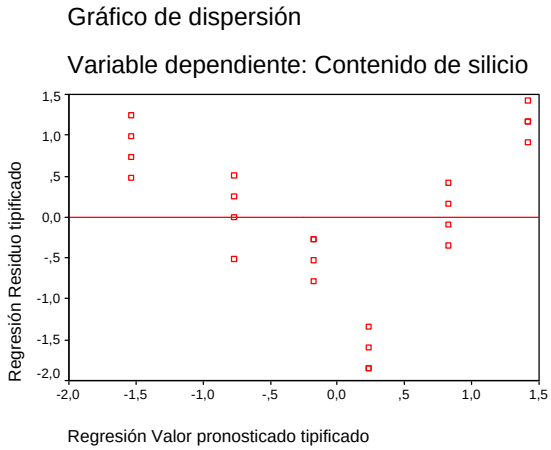


En **Ajuste** se elige como método **Regresión**.

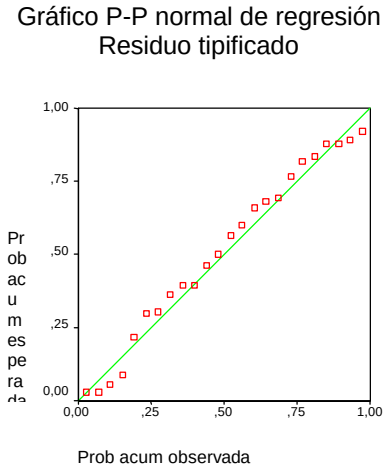


1.4.- Gráficos

La gráfica presenta el diagrama de dispersión de los residuos tipificados en relación a los valores tipificados, que nos da una indicación de si se cumplen las hipótesis de linealidad e igualdad de varianzas.



Otra gráfica interesante es la de probabilidad normal:



.5.- Predicción

Si se desea hacer estimaciones de la variable dependiente para ciertos valores de la independiente se han de añadir estos valores a la tabla de datos y repetir el análisis de regresión, marcando en la opción guardar los resultados que se quieren obtener (valores pronosticados y/o intervalos).

Como ejemplo, puedes obtener un intervalo de estimación para la media del contenido de silicio en lugares situados a 12 y a 40 Km. de la costa. ¿Podrías estimar la cantidad de silicio en un punto situado a 70 Km. de la costa?

2. USO DE TRANSFORMACIONES DE LOS DATOS. EFECTO DE VALORES ATÍPICOS (OUTLIERS) EN LA REGRESIÓN Y CORRELACIÓN

Ejercicio 2
En el fichero CEREBROS están los datos del peso del cerebro y del cuerpo de diferentes razas de animales identificados mediante una etiqueta. Se desea estudiar la relación que existe entre estas medidas.

2.1.- Estudia la relación lineal entre el peso del cerebro y el peso del cuerpo. Comenta los resultados tanto de los contrastes de hipótesis que se pueden realizar como del análisis de los residuos.

2.2.- Crea dos nuevas variables que sean el logaritmo del peso del cerebro y el del peso del cuerpo. Estudia la relación que existe entre estas dos nuevas variables. Obtén la gráfica de la recta de regresión ¿observas algún dato anómalo?. Identifica cuales son.

2.3.- Haz una selección de casos que no contenga los datos anómalos observados anteriormente y repite el proceso del paso 2.2 comentando las diferencias que encuentres.

3.- REGRESIÓN LINEAL MÚLTIPLE

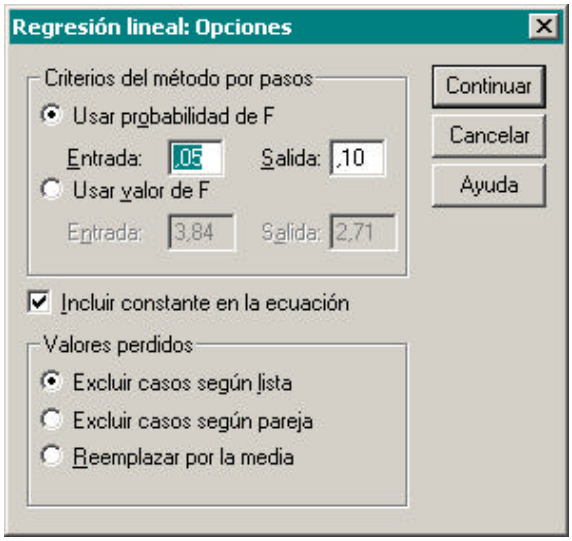
El proceso es idéntico al anterior excepto que en las variables independientes pondremos todas las que deseamos estudiar y elegiremos el método deseado para introducirlas en el modelo.

Métodos de selección de variables en el análisis de regresión lineal

La selección del método permite especificar cómo se introducen las variables independientes en el análisis. Utilizando distintos métodos se pueden construir diversos modelos de regresión a partir del mismo conjunto de variables.

Para introducir las variables del bloque en un solo paso seleccione **Introducir**. La selección de variables **Hacia adelante** introduce las variables del bloque una a una basándose en los criterios de entrada. La eliminación de variables **Hacia atrás** introduce todas las variables del bloque en un único paso y después las elimina una a una basándose en los criterios de salida. La entrada y salida de variables mediante **Pasos sucesivos** examina las variables del bloque en cada paso para introducirlas o excluirlas. Se trata de un procedimiento hacia adelante por pasos.

Los criterios de entrada y salida de las variables se especifican en **Opciones** (el criterio de entrada debe ser menor que el criterio de salida).



Ejercicio 3

En la industria de la explotación forestal es necesario hacer estimaciones del total de tableros que pueden obtenerse de un árbol. Existe una gran variedad de reglas para este propósito, típicamente relacionan la productividad, Y, con la longitud y diámetro del tronco. Intuitivamente podríamos esperar que estas reglas tengan en cuenta productos de estas variables ya que la fórmula del volumen de un cilindro depende del cuadrado del radio y de la longitud.

La unidad de volumen para los tableros es el "board foot". Por definición un "board foot" es el volumen de un tablero de 1 pulgada de espesor por 12 pulgadas de ancho por 12 pulgadas de largo. En el fichero TABLEROS están los datos de una muestra de 24 árboles de los que se registraron el "board foot" (BDFT), Y, la longitud (LONG),X_1, y el diámetro al final del árbol (DIAM), X_2.

Considera el modelo lineal de segundo orden

$$Y=a + b_1 x_1+ b_2 x_2+ b_3 x_1^2+ b_4 x_2^2+ b_5 x_1 x_2$$

1.- Estima los parámetros de este modelo. ¿Es adecuado el modelo lineal elegido?
2.- Elimina la variable que no consideres significativa y vuelve a realizar el análisis.
3.- Con todas las variables, haz un análisis hacia delante, hacia atrás y por pasos sucesivos. Comenta los diferentes resultados.

EJERCICIOS

1.- En un estudio, por medio de detectores radioactivos, de la capacidad corporal para absorber hierro y plomo, participaron diez sujetos. A cada uno se le da una dosis oral idéntica de hierro (sulfato ferroso) y de plomo (cloruro de plomo-203). Después de doce días se mide la cantidad de cada componente retenida en el sistema corporal y, a partir de éstas, se determinan los porcentajes absorbidos por el cuerpo. Los datos obtenidos fueron:

Hierro (%)	17	22	35	43	80	85	91	92	96	100
Plomo (%)	8	17	18	25	58	59	41	30	43	58

a) Dibuja la nube de puntos. Basándose en ella, ¿se puede esperar que el coeficiente de correlación esté próximo a 1, -1 ó 0?.
b) Halla e interpreta el coeficiente de determinación.
c) Comprueba la idoneidad del modelo de regresión lineal. Si éste es apropiado, estima la recta de regresión y utilízala para predecir el porcentaje de hierro absorbido por un individuo cuyo sistema corporal absorbe el 15% del plomo ingerido.

2.- Se han obtenido importantes ventajas de enseñar a los diabéticos a medir su propia glucosa en sangre. Se investiga una nueva técnica menos costosa que el procedimiento habitual. La técnica utiliza una varilla. La varilla desarrolla dos colores simultáneamente y estos colores son comparados a ojo con una tarjeta que da el nivel de glucosa. Si se puede probar que el procedimiento es preciso, se generalizará su uso. Se obtuvieron los datos de X, nivel de glucosa en sangre medido por un paciente utilizando la varilla e Y, nivel de glucosa en sangre del paciente medido en el laboratorio (medidos en milimoles/litro). Los datos se encuentran el fichero DIABETES.
a) Dibuja la nube de puntos. Basándose en ella, ¿crees que hay una fuerte correlación positiva entre el nivel de glucosa en sangre establecido por el paciente y el que se ha medido en el laboratorio?
b) Halla el coeficiente de correlación y el de determinación.
c) Halla estimaciones puntuales para la pendiente y la ordenada en el origen de la recta de regresión.
d) Hallar una estimación del nivel de glucosa establecido en el laboratorio de un paciente que lo sitúa en 4.0 mmol/litro. Hallar un intervalo de confianza al 90% para este valor.

3.- Se realiza un estudio de fotoperiodo en aves acuáticas. Se pretende establecer una ecuación mediante la cual pueda predecirse el tiempo de reproducción, Y, en base al conocimiento del fotoperiodo (número de horas de luz por día) bajo el que se inició la reproducción, X. Se obtuvieron datos del comportamiento de 11 *Aythya* (patos buceadores). Los resultados fueron los siguientes:

Tiempo reproducción	110	54	98	50	67	58	52	50	43	15	28
Fotoperiodo	12.8	13.9	14.1	14.7	15.0	15.1	16.0	16.5	16.6	17.2	17.9

Halla la recta de regresión correspondiente. Calcula una predicción del tiempo de reproducción para un fotoperiodo de 14.5 horas. ¿Tendría sentido realizar en este caso una predicción para un fotoperiodo de 24 horas?.

4.- En un estudio sobre seguridad vial se consideró la relación entre la velocidad del vehículo y la distancia de frenado, para poder predecir la distancia de frenado conociendo la velocidad. Entre otros objetivos se trata de decidir si la relación entre ambas variables es lineal o más bien si la velocidad está relacionada linealmente con la raíz cuadrada de la distancia de frenado, tal como sugiere una ley física relativa a la disipación de fuerzas de inercia. En nuestro caso, se trata de decidir utilizando únicamente procedimientos estadísticos.

Los datos del fichero FRENADO muestran velocidades en millas por hora (una milla = 1609.34 metros) y distancias de frenado en pies (un pie= 30.48 centímetros).

Ajusta a los datos una recta de regresión de la distancia (en metros) sobre la velocidad (en km/hora) y otra de la raíz de la distancia sobre la velocidad. Dibuja ambas, realiza los correspondientes análisis y decide el modelo que mejor refleje la relación entre ambas variables.

5.- En el fichero SO2 están los datos de la polución del aire de 41 ciudades americanas. La variable dependiente es la media de la concentración de dióxido de sulfuro, expresada en microgramos por metro cúbico. Las variables de predicción son seis de tipo ecológico. Encuentra un modelo de regresión múltiple adecuado al problema.