PRÁCTICA 3. REGRESIÓN LINEAL SIMPLE CON SPSS

3.1. Gráfico de dispersión

3.2. Ajuste de un modelo de regresión lineal simple

3.3. Porcentaje de variabilidad explicado

3.4 ¿Es adecuado este modelo para ajustar los datos?

3.5 Estudio de las hipótesis requeridas para utilizar este modelo

3.6 Inferencias para los parámetros

3.7 Inferencias para las predicciones

3.8 Ejercicio

Ejemplo: Con los datos del archivo miel.sav vamos a estudiar un modelo de regresión lineal simple que permita predecir la *acidez total* a partir de la acidez *libre*.

3.1. Gráfico de dispersión

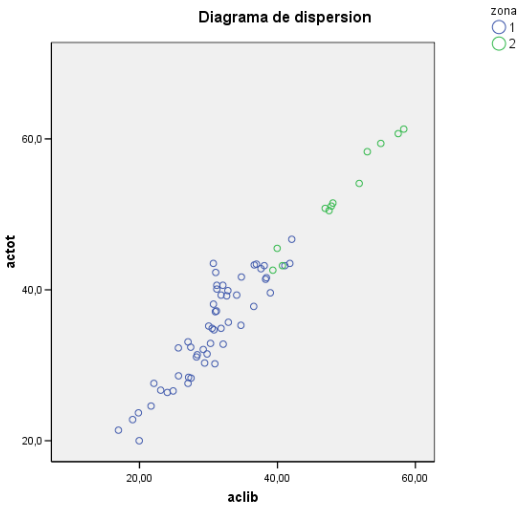
Dibujamos un gráfico de dispersión de la acidez total (AcTot) frente a la acidez libre (AcLib).

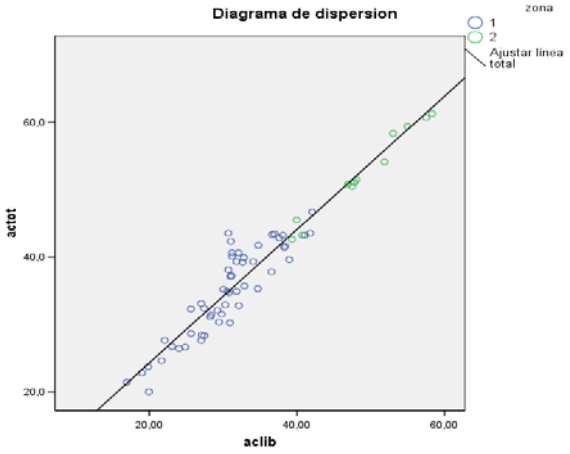
Gráficos/Dispersión-puntos

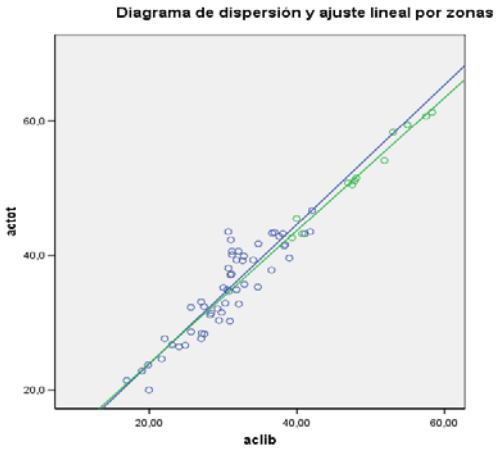
Eje de las X: *Acidez libre*

Eje de las Y: *Acided total.*

Establecer marcas por: *zona*







3.2. Ajuste de un modelo de regresión lineal simple

Ajustamos un modelo de regresión lineal simple para explicar la acidez total (AcTot) frente a la acidez libre (AcLib).

Analizar/Regresión/ Lineal

Variable dependiente: *Acidez total*.

Método: *introducir*

Obtenemos la tabla siguiente tabla de coeficientes:

Model		Coeficientes no estandarizados	
		B	Error típ.
	Constante	4.469	1.260
	aclib	.990	.036

Que nos lleva a que la recta ajustada es la siguiente:

$$\hat{Y} = 4.469 + 0.990X$$

3.3. Porcentaje de variabilidad explicado

Veamos qué porcentaje de variabilidad de la acidez total queda explicada por la acidez libre.

Dentro del procedimiento anterior el SPSS nos da la siguiente tabla de resumen:

Resumen del modelo

Model	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
	.960(a)	.922	.921	2.6762

a Variables predictoras: (Constante), aclib

Vemos que la acidez libre **nos explica el 92,2% de la variabilidad de la acidez total**.

3.4 ¿Es adecuado este modelo para ajustar los datos?

El modelo es adecuado si podemos afirmar que existe dependencia lineal entre las dos variables.

De la tabla de los coeficientes vemos que b= .990 es distinto de cero por lo que parece que hay una relación lineal. Además es positivo por lo que la relación es directa. Sin embargo no sabemos si ese coeficiente es significativamente distinto de cero. Para comprobarlo debemos realizar el **contraste de regresión**.

$$\begin{aligned} H_0 &: \quad \beta = 0 \\ H_1 &: \quad \beta \neq 0 \end{aligned}$$

En este caso este contraste equivale a contrastar que el coeficiente de correlación poblacional es significativamente distinto de cero.

Dentro del procedimiento anterior el SPSS nos da la siguiente tabla :

ANOVA

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	**Regresión**	5442.731	1	5442.731	**759.928**	**.000**
	Residual	458.379	64	7.162		
	Total	5901.110	65			

El valor de F =759,928 tiene un P_valor igual a 0 < 0,05 por lo que rechazamos la hipótesis nula y concluimos que la dependencia lineal es estadísticamente significativa. Por lo tanto **el modelo es adecuado**.

3.5 Estudio de las hipótesis requeridas para utilizar este modelo

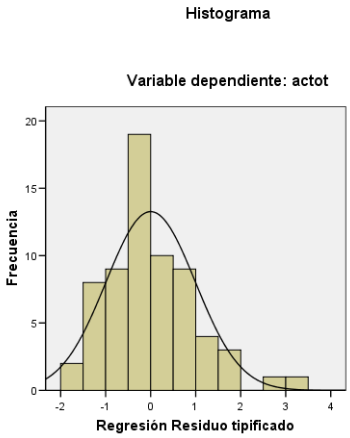
Veamos si se cumplen las hipótesis requeridas, es decir: normalidad, independencia y homocedasticidad.

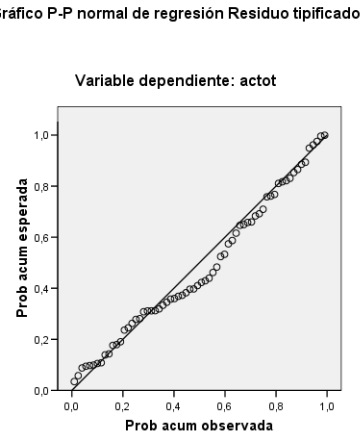
Normalidad:

Analizar/Regresión/ Lineal

Gráficos/Gráficos residuos tipificados/Histograma y gráfico de probabilidad normal

Obtenemos las dos gráficas siguientes las cuales parecen indicar normalidad.





Para el contraste de *normalidad*:

Analizar / Estadísticos descriptivos / Explorar

Gráficos: > Gráficos con pruebas de normalidad para: *los residuo tipificados* (que debemos tener guardados con la opción GUARDAR)

Obtenemos la siguiente tabla:

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Standardized Residual	**,107**	66	**,061**	**,959**	66	**,027**

a. Corrección de la significación de Lilliefors

Con K-S aceptamos normalidad, con S-W no. Nos quedamos con que son normales pues n> 30.

Independencia:

Para comprobar la *independencia* obtenemos el estadístico de Durbin-Watson:

Analizar / Regresión / Lineal

Estadísticos: Durbin-Watson

Resumen del modelo(b)

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación	Durbin-Watson
1	.960(a)	.922	.921	2.6762	1.624

a Variables predictoras: (Constante), aclib

D-W =1.624, está entre 1.5 y 2.5 y entonces concluimos que los residuos son incorrelados.

Homocedasticidad

Para comprobar la *homocedasticidad* dibujaremos un diagrama de dispersión de las estimaciones (valores predichos por el modelo) tipificadas (ZPRED) frente a los residuos tipificados (ZRESID) (Aquí entiende los ZRESID aunque no los hayamos guardado con anterioridad)

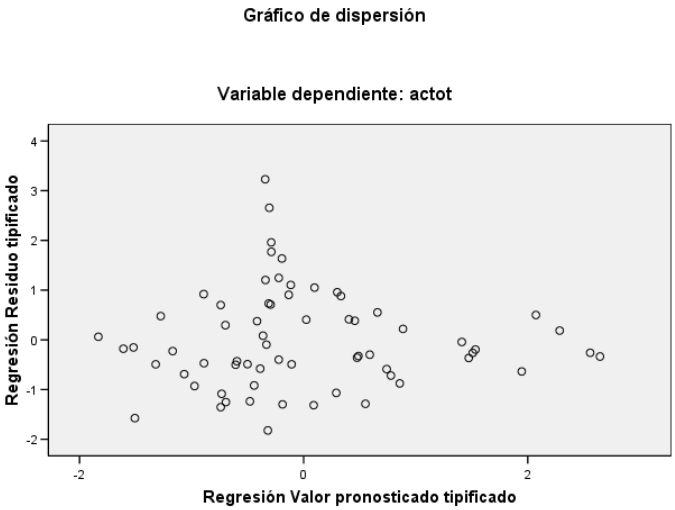
Analizar/Regresión/ Lineal

Gráficos:

Eje Y: *residuos tipificados (ZRESID)*

Eje X: *valores predichos por el modelo tipificados (ZPRED)*

El gráfico no presenta ningún patrón. **Aceptamos igualdad de varianzas**.



3.6 Inferencias para los parámetros

Analizar/Regresión/ Lineal

Estadísticos: Intervalos de confianza

Coeficientes[a]

Modelo	Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0%	
	B	Error típ.	Beta			Límite inferior	Límite superior
Constante	4,469	1,260		3,547	,001	1,952	6,987
aclib	,990	,036	,960	27,567	,000	,918	1,062

a. Variable dependiente: actot

Los dos coeficientes son significatívamente distintos de cero pues para los dos casos el P_valor es < 0,05.

Las dos últimas columnas nos dan los correspondientes intervalos de confianza.

3.7 Inferencias para la predicción

a) ¿Cuál sería el valor medio de la Acidez Total predicho para una Acidez Libre igual a 30,03? Dar el intervalo de confianza para esa predicción y el error que se comete en la estimación.

En este caso, Aclibre= 30,03 es uno de los datos de la muestra. Entonces tanto el valor medio predicho como el intervalo de confianza pueden ser obtenidos directamente en el proceso de ajuste con el SPSS (En **GUARDAR** activar: *valores pronosticados no tipificados*; *residuos tipificados; intervalos de pronóstico para la media y para los individuos*. **Todos estos resultados aparecen en vista de datos)**.

El valor **medio predicho es 34,19** (ver valores pronosticados).

El intervalo de confianza es (33,48 34,91) (ver intervalos de pronóstico para la media).

El error que se comete en la estimación es (34,91-33,48)/2 = **0,715**

b) ¿Cuál sería el valor predicho para la acidez total si consideramos la acidez libre media?

La recta pasa por siempre por el centro de gravedad de los datos.

Estadísticos descriptivos

	N	Media
actot	66	**37,998**
aclib	66	**33,8727**
N válido (según lista)	66	

Entonces:

$$\hat{Y} = \overline{Y} = 37,998$$

3.8 Ejercicio

Con los datos de la miel el investigador tiene interés en establecer una relación lineal entre la variable azuhum y la humedad.

a) **¿Tiene sentido plantearse un modelo de regresión? ¿Por qué?**

b) **Obtener la recta de regresión tomando como variable dependiente azuhum y como independiente la humedad.**

c) **¿Cuál sería el valor medio de azuhum predicho para un grado de humedad de 17,8? Dar el intervalo de confianza para esa predicción y el error que se comete en la estimación.**

d) **¿Qué valor tomaría la variable azuhum si considerásemos el grado medio de humedad?**

e) **Hacer un estudio de los residuos.**