

Actividad 3

Tema 3

TRABAJO REALIZADO POR: CARMEN M^a SÁNCHEZ CAMPOY

PROFESORES: RAMÓN GUTIÉRREZ SÁNCHEZ
MARIA DOLORES RUIZ MEDINA

CURSO: DISEÑO ESTADÍSTICO EXPERIMENTAL Y CONTROL DE CALIDAD.
APLICACIONES EN BIOCIENCIAS E INGENIERÍA

- MASTER ESTADÍSTICA APLICADA -

A1. CUESTIONES TEÓRICAS

Resolver tres actividades teóricas.

1.- Deducir la expresión de los estimadores mínimo-cuadráticos de los parámetros del modelo de regresión lineal simple.

Definimos las siguientes variables:

X: variable de regresión o explicativa, continua y controlable por el experimentador. En el diseño del experimento se determinan sus valores.

Y: variable respuesta, para la que se supone una relación lineal entre Y y la variable explicativa X.

El modelo que define la observación de la variable respuesta Y viene dado por:

$$Y = a_0 + a_1 X + \varepsilon$$

representando ε , la componente de error aleatoria, se supone que ε es una variable aleatoria con media cero y varianza σ^2 y que el conjunto de componentes aleatorias de error no están correlacionadas.

Tomando n pares de datos $(x_1, y_1), \dots, (x_n, y_n)$, presentamos la siguiente **demostración**:

El proceso para la obtención por mínimos cuadrados de los estimadores a_0 y a_1 tiene por objetivo minimizar la suma de los cuadrados de los residuos, que denotamos por L . Partiendo de dicha función su expresión viene dada por:

$$L = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - a_0 - a_1 x_i)^2$$

Para minimizar L , derivamos parcialmente respecto de a_0 y a_1 :

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial a_0} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) \\ \frac{\partial L}{\partial a_1} = -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) x_i \end{array} \right.$$

Los estimadores mínimo-cuadráticos se obtienen igualando las anteriores derivadas a cero:

$$\left\{ \begin{array}{l} -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) = 0 \\ -2 \sum_{i=1}^n (y_i - a_0 - a_1 x_i) x_i = 0 \end{array} \right.$$

Operando se tiene:

$$\left\{ \begin{array}{l} \sum_{i=1}^n y_i = n a_0 + a_1 \sum_{i=1}^n x_i \\ \sum_{i=1}^n y_i x_i = a_0 \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 \end{array} \right.$$

Para resolver este sistema de ecuaciones, realizamos los siguientes pasos:

Dividimos la primera ecuación por n: $\bar{Y} = a_0 + a_1 \bar{X}$

Despejando: $a_0 = \bar{Y} - a_1 \bar{X}$

Sustituyendo a_0 en la segunda ecuación:

$$\begin{aligned} \sum_{i=1}^n y_i x_i &= (\bar{Y} - a_1 \bar{X}) \sum_{i=1}^n x_i + a_1 \sum_{i=1}^n x_i^2 \\ \sum_{i=1}^n y_i x_i - \bar{Y} \sum_{i=1}^n x_i &= a_1 \left(\sum_{i=1}^n x_i^2 - \bar{X} \sum_{i=1}^n x_i \right) \quad (*) \end{aligned}$$

Por otra parte:

$$\begin{aligned} \bullet \sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X}) &= \sum_{i=1}^n (y_i x_i - \bar{X} y_i - \bar{Y} x_i + \bar{Y} \bar{X}) = \\ &= \sum_{i=1}^n y_i x_i - \bar{X} \sum_{i=1}^n y_i - \bar{Y} \sum_{i=1}^n x_i + n \bar{Y} \bar{X} = \\ &= \sum_{i=1}^n y_i x_i - n \bar{X} \bar{Y} - \bar{Y} \sum_{i=1}^n x_i + n \bar{Y} \bar{X} = \sum_{i=1}^n y_i x_i - \bar{Y} \sum_{i=1}^n x_i \\ \bullet \sum_{i=1}^n (x_i - \bar{X})^2 &= \sum_{i=1}^n (x_i^2 - 2 \bar{X} x_i + \bar{X}^2) = \sum_{i=1}^n x_i^2 - 2 \bar{X} \sum_{i=1}^n x_i + n \bar{X}^2 = \\ &= \sum_{i=1}^n x_i^2 - 2 \bar{X} \sum_{i=1}^n x_i + \bar{X} \sum_{i=1}^n x_i = \sum_{i=1}^n x_i^2 - \bar{X} \sum_{i=1}^n x_i \end{aligned}$$

Teniendo en cuenta estas igualdades obtenidas y sustituyéndolas en la ecuación (*), tenemos que:

$$\sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X}) = a_1 \sum_{i=1}^n (x_i - \bar{X})^2$$

De donde deducimos que el estimador de a_1 viene dado por el cociente:

$$a_1 = \frac{\sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^n (x_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}}$$

Siendo:

$$S_{XY} = \sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X}) = \sum_{i=1}^n y_i x_i - n \bar{X} \bar{Y}$$

$$S_{XX} = \sum_{i=1}^n (x_i - \bar{X})^2 = \sum_{i=1}^n x_i^2 - n \bar{X}^2$$

Basta sustituir a_1 en la expresión despejada de la primera ecuación para obtener:

$$a_0 = \bar{Y} - a_1 \bar{X}$$

Luego, hemos deducido que los estimadores a_0 y a_1 son:

$$\begin{cases} a_1 = \frac{S_{XY}}{S_{XX}} \\ a_0 = \bar{Y} - a_1 \bar{X} \end{cases}$$

2.- Deducir la expresión

$$SS_E = S_{YY} - a_1 S_{XY}$$

de la suma de cuadrados de los residuos.

Tenemos las siguientes notaciones:

$$\left\{ \begin{array}{l} Y = a_0 + a_1 X \\ SS_E : \text{Suma de cuadrados de los residuos: } SS_E = \sum_{i=1}^n \varepsilon_i^2 \\ S_{YY} = \sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n y_i^2 - n \bar{Y}^2 \\ S_{XX} = \sum_{i=1}^n (x_i - \bar{X})^2 = \sum_{i=1}^n x_i^2 - n \bar{X}^2 \\ S_{XY} = \sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X}) = \sum_{i=1}^n y_i x_i - n \bar{X} \bar{Y} \end{array} \right.$$

Para poder llegar a la expresión deseada, comenzamos con la siguiente igualdad:

$$y_i = y_i + \varepsilon_i$$

Restamos \bar{Y} a ambos lados: $y_i - \bar{Y} = y_i - \bar{Y} + \varepsilon_i$

Si elevamos al cuadrado ambos miembros se obtiene que:

$$(y_i - \bar{Y})^2 = (y_i - \bar{Y} + \varepsilon_i)^2$$

Es decir: $(y_i - \bar{Y})^2 = (y_i - \bar{Y})^2 + \varepsilon_i^2 + 2(y_i - \bar{Y})\varepsilon_i$

Sumando ambos miembros de la expresión de $i = 1$ hasta n , se tiene

$$\sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n \varepsilon_i^2 + 2 \sum_{i=1}^n (y_i - \bar{Y})\varepsilon_i$$

Ahora bien, el último término de la expresión anterior es cero, hacemos la demostración en el siguiente marco:

$$\sum_{i=1}^n (y_i - \bar{Y})\varepsilon_i = \sum_{i=1}^n y_i \varepsilon_i - \bar{Y} \sum_{i=1}^n \varepsilon_i \text{ y sabemos que:}$$

- La suma de los residuos mínimo-cuadráticos es igual a cero:

$$\begin{aligned} \sum_{i=1}^n \varepsilon_i &= \sum_{i=1}^n (y_i - \bar{y}_i) = \sum_{i=1}^n y_i - \sum_{i=1}^n \bar{y}_i = \sum_{i=1}^n y_i - \sum_{i=1}^n (a_0 + a_1 x_i) = \\ &= \sum_{i=1}^n y_i - n a_0 - a_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i - n (\bar{Y} - a_1 \bar{X}) - a_1 \sum_{i=1}^n x_i = \\ &= \sum_{i=1}^n y_i - \left(\sum_{i=1}^n y_i - a_1 \sum_{i=1}^n x_i \right) - a_1 \sum_{i=1}^n x_i = 0 \end{aligned}$$

- La suma de los productos cruzados entre los valores ajustados y los residuos es igual a 0:

$$\sum_{i=1}^n y_i \varepsilon_i = \sum_{i=1}^n (a_0 + a_1 x_i) \varepsilon_i = a_0 \sum_{i=1}^n \varepsilon_i + a_1 \sum_{i=1}^n x_i \varepsilon_i = 0$$

Puesto que:

$$\varepsilon \text{ es una variable aleatoria con media cero luego: } n \bar{\varepsilon} = \sum_{i=1}^n \varepsilon_i = 0$$

$$\sum_{i=1}^n x_i \varepsilon_i = \sum_{i=1}^n (y_i - a_0 - a_1 x_i) x_i = 0 \quad \text{por la segunda ecuación del sistemas de ecuaciones obtenido en la estimación por mínimos cuadrados.}$$

Luego:

$$\sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n \varepsilon_i^2$$

Por tanto, hemos llegado a que:

$$S_{YY} = S_{YY} + SS_E$$

Como $Y = a_0 + a_1X$, equivale a un cambio de escala y origen de la variable X, por las propiedades de la varianza ante estos cambios, se tiene que:

$$\frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{n} = a_1^2 \frac{\sum_{i=1}^n (x_i - \bar{X})^2}{n} \Rightarrow \sum_{i=1}^n (y_i - \bar{Y})^2 = a_1^2 \sum_{i=1}^n (x_i - \bar{X})^2$$

Luego:

$$S_{YY} = a_1^2 S_{XX} = a_1 a_1 S_{XX} = a_1 \frac{S_{XY}}{S_{XX}} S_{XX} = a_1 S_{XY}$$

Así llegamos a la igualdad deseada:

$$S_{YY} = a_1 S_{XY} + SS_E \Rightarrow SS_E = S_{YY} - a_1 S_{XY}$$

3.- Explicar brevemente la interpretación de los valores del coeficiente de determinación.

El coeficiente de determinación se define como la proporción de la varianza total explicada por la regresión. Su expresión viene dada por:

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{Y})^2}{\sum_{i=1}^n (y_i - \bar{Y})^2} = \frac{SS_R}{S_{YY}}$$

De forma equivalente, y en aplicación de la igualdad:

$$\sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 + \sum_{i=1}^n \varepsilon_i^2$$

obtenida en el apartado anterior de esta actividad, podemos expresar el coeficiente de determinación como, como uno menos la proporción no explicada por la regresión, es decir:

$$R^2 = 1 - \frac{\sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n (y_i - \bar{Y})^2} = 1 - \frac{SS_E}{S_{YY}}$$

El criterio mínimo-cuadrático equivale a maximizar R^2 .

Dicho coeficiente toma valores en el intervalo (0, 1) y se interpreta como la proporción de variabilidad de los datos explicada por el modelo de regresión. Por este motivo, se

suele utilizar, como un indicador de la adecuación del modelo de regresión (medida relativa del grado de asociación lineal entre X e Y), mide la correlación entre el valor observado y el valor predicho o ajustado con la regresión.

$$0 \leq R^2 \leq 1$$

- Si $R^2 = 1 \Rightarrow \sum_{i=1}^n (y_i - \bar{Y})^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 \quad y \quad \sum_{i=1}^n \varepsilon_i^2 = 0$

Lo que implica un ajuste perfecto, Y depende funcionalmente de X, la varianza de los residuos se hace cero y la varianza de los valores observados y la variable respuesta coincide.

- Si $R^2 < 1 \Rightarrow \sum_{i=1}^n (y_i - \bar{Y})^2 \neq 0 \quad y \quad \sum_{i=1}^n \varepsilon_i^2 \neq 0$

Se tiene que:

$$\begin{cases} \sum_{i=1}^n (y_i - \bar{Y})^2 = R^2 \sum_{i=1}^n (y_i - \bar{Y})^2 \\ \sum_{i=1}^n \varepsilon_i^2 = (1 - R^2) \sum_{i=1}^n (y_i - \bar{Y})^2 \end{cases}$$

Un valor de R^2 cercano a 0 implica baja capacidad explicativa de la recta, por otro lado, un valor próximo a 1, equivale a alta capacidad explicativa de la recta.

- Si $R^2 = 0 \Rightarrow \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 \quad y \quad \sum_{i=1}^n (y_i - \bar{Y})^2 = 0$

El modelo no explica nada de Y a partir de X.

En resumen:

El coeficiente de determinación toma valores entre 0 y 1, tomando el valor 0 cuando el modelo no explica nada de Y a partir de X, es decir el ajuste es el peor posible, y tomando el valor 1 cuando todos los residuos son nulos, es decir el ajuste es perfecto. Para valores intermedios, según estén más próximos a 0 o 1, nos indicarán un peor o mejor ajuste respectivamente, por poner datos numéricos algunos autores, consideran un buen ajuste para valores de R^2 mayores de 0.75, es decir cuando al menos el 75% de la varianza total quede explicada por la regresión.

Para terminar damos otras fórmulas para el coeficiente de determinación:

$$R^2 = \frac{SS_R}{S_{YY}} = \frac{S_{XY}^2}{S_{YY} S_{XX}} = a_1 a_1' = r^2$$

A2. TRABAJO

Elaborar un resumen sobre los contrastes de ajuste en el modelo de regresión lineal. Indicando algunos casos particulares interesantes para el análisis de la adecuación del modelo

Los estimadores a_0 y a_1 dependen de la muestra seleccionada, por lo tanto son variables aleatorias y presentarán una distribución de probabilidad. Estas distribuciones de probabilidad de los estimadores pueden utilizarse para construir intervalos de confianza o contrastes sobre los parámetros del modelo de regresión.

Suponiendo que los residuos se distribuyen normalmente, realizamos un resumen de los contrastes de ajuste sobre el modelo de regresión lineal simple:

1.- Ajuste de la pendiente de la recta, contrastes para el parámetro a_1 :

En términos generales planteamos los siguientes contrastes para a_1 :

- **Unilateral a la izquierda** (contraste de una cola):

$$\begin{cases} H_0 : a_1 = a \\ H_1 : a_1 < a \end{cases}$$

- El estadístico pivot para este contraste es:

$$t_0 = \frac{a_1 - a}{\sqrt{\frac{MS_E}{S_{XX}}}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con $n-2$ grados de libertad.

- La hipótesis nula se rechaza cuando:

$$t_0 < t_{\alpha, n-2}$$

siendo $t_{\alpha, n-2}$ el percentil $1-\alpha$ de la distribución t-Student con $n-2$ grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

Rechazamos H_0 si : p-valor < α

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = P(t_{n-2} < t_0)$$

- **Unilateral a la derecha** (contraste de una cola):

$$\begin{cases} H_0 : a_1 = a \\ H_1 : a_1 > a \end{cases}$$

- El estadístico pivot para este contraste es:

$$t_0 = \frac{a_1 - a}{\sqrt{\frac{MS_E}{S_{XX}}}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$t_0 > t_{1-\alpha, n-2}$$

siendo $t_{1-\alpha, n-2}$ el percentil α de la distribución t-Student con n-2 grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

Rechazamos H_0 si : p-valor < α

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = P(t_{n-2} > t_0)$$

• Bilateral (contraste de dos colas):

$$\begin{cases} H_0 : a_1 = a \\ H_1 : a_1 \neq a \end{cases}$$

- El estadístico pivot para este contraste es:

$$t_0 = \frac{a_1 - a}{\sqrt{\frac{MS_E}{S_{XX}}}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$|t_0| > t_{\alpha/2, n-2}$$

siendo $t_{\alpha/2, n-2}$ el percentil $1 - \alpha / 2$ de la distribución t-Student con n-2 grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

Rechazamos H_0 si : p-valor < α

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = 2P(t_{n-2} > |t_0|)$$

Caso especial

Se puede considerar el contraste de ausencia de una relación lineal entre X e Y; o bien, la ausencia de una relación causal entre dichas variables, en términos del primer contraste de ajuste sobre la pendiente. Es decir,

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

- El estadístico pivote para este contraste es:

$$t_0 = \frac{a_1}{\sqrt{\frac{MS_E}{S_{XX}}}}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$|t_0| > t_{\alpha/2, n-2}$$

es decir, $t_0 > t_{\alpha/2, n-2}$ o $t_0 < -t_{\alpha/2, n-2}$

siendo $t_{\alpha/2, n-2}$ el percentil $1 - \alpha / 2$ de la distribución t-Student con n-2 grados de libertad.

Por lo tanto, si el estadístico de prueba cae en la región crítica, se rechaza la hipótesis nula y se dice que el estadístico hallado es estadísticamente significativo con un nivel de confianza del $100(1 - \alpha)\%$.

Ajuste de la pendiente de la recta, contrastes para el parámetro a_0 :

En términos generales planteamos los siguientes contrastes para a_0 :

• Unilateral a la izquierda (contraste de una cola):

$$\begin{cases} H_0 : a_0 = a \\ H_1 : a_0 < a \end{cases}$$

- El estadístico pivote para este contraste es:

$$t_0 = \frac{a_0 - a}{\sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$t_0 < t_{\alpha, n-2}$$

siendo $t_{\alpha, n-2}$ el percentil $1 - \alpha$ de la distribución t-Student con n-2 grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

Rechazamos H_0 si : p-valor < α

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = P(t_{n-2} < t_0)$$

- **Unilateral a la derecha** (contraste de una cola):

$$\begin{cases} H_0 : a_0 = a \\ H_1 : a_0 > a \end{cases}$$

- El estadístico pivoté para este contraste es:

$$t_0 = \frac{a_0 - a}{\sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$t_0 > t_{1-\alpha, n-2}$$

siendo $t_{1-\alpha, n-2}$ el percentil α de la distribución t-Student con n-2 grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

$$\text{Rechazamos } H_0 \text{ si : p-valor} < \alpha$$

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = P(t_{n-2} > t_0)$$

- **Bilateral** (contraste de dos colas):

$$\begin{cases} H_0 : a_0 = a \\ H_1 : a_0 \neq a \end{cases}$$

- El estadístico pivoté para este contraste es:

$$t_0 = \frac{a_0 - a}{\sqrt{MS_E \left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}} \right)}} \quad \text{donde: } MS_E = \frac{S_{YY} - a_1 S_{XY}}{n-2}$$

Bajo la hipótesis nula, dicho estadístico sigue una distribución t-Student con n-2 grados de libertad.

- La hipótesis nula se rechaza cuando:

$$|t_0| > t_{\alpha/2, n-2}$$

siendo $t_{\alpha/2, n-2}$ el percentil $1 - \alpha / 2$ de la distribución t-Student con n-2 grados de libertad.

Otra forma ver si rechazamos o no la hipótesis nula es con el p-valor:

Rechazamos H_0 si : p-valor < α

Calculándose el p-valor en este caso de la forma siguiente:

$$\text{p-valor} = 2P(t_{n-2} > |t_0|)$$

A3. ANÁLISIS DE DATOS

Para realizar los ejercicios voy a utilizar el software SPSS.

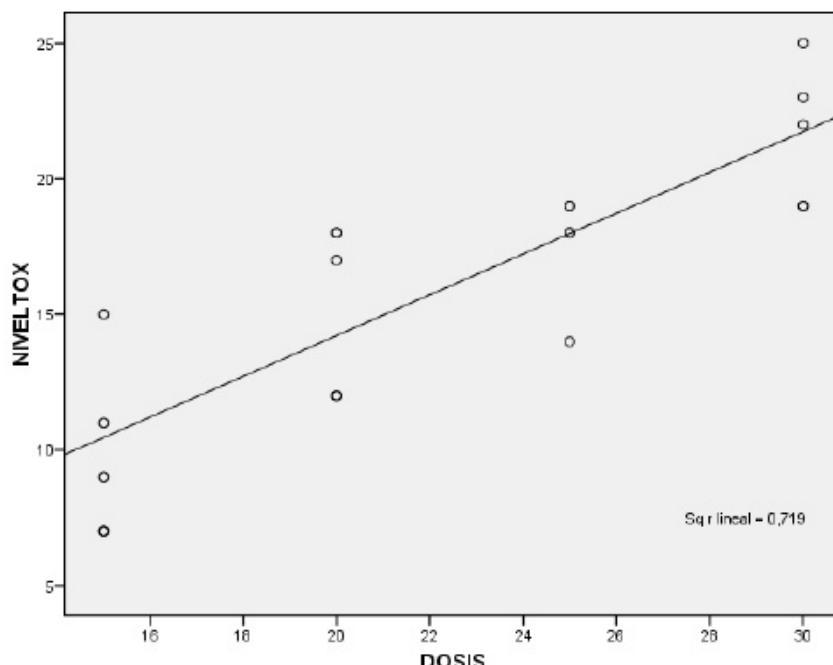
1. Se realiza un experimento para estudiar la relación entre los niveles de paracetamol ingeridos y los niveles de toxicidad hepática. En la tabla siguiente se reflejan los niveles de toxicidad observados en cinco individuos a los que se le administraron diferentes dosis de paracetamol diarias.

| Dosis | Niveles de toxicidad hepática observados | | | | |
|-------|--|----|----|----|----|
| 15 | 7 | 7 | 15 | 11 | 9 |
| 20 | 12 | 17 | 12 | 18 | 18 |
| 25 | 14 | 18 | 18 | 19 | 19 |
| 30 | 19 | 25 | 22 | 19 | 23 |

Contrastar la idoneidad del modelo de regresión lineal para un tamaño $\alpha = 0,05$ del test.

Llamamos a las variables DOSIS (Variable Independiente) y NIVELTOX (Variable Dependiente) que vienen recogidas en archivo ejercicio1.sav de la carpeta de datos.

Empezamos el problema mediante la presentación del diagrama de dispersión entre ambas variables, y la representación de la recta de regresión aproximada. Los diagramas de dispersión ofrecen una idea bastante aproximada sobre el tipo de relación existente entre dos variables, además, también puede utilizarse como una forma de cuantificar el grado de relación lineal existente entre dos variables, basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.



El gráfico muestra una posible adecuación del modelo lineal y la tendencia creciente del mismo.

Para obtener la recta de regresión mínima cuadrática de NIVELTOX sobre DOSIS , representada en la nube de puntos,

$$Y = a_0 + a_1 X$$

Para ello utilizamos la opción Analizar/Regresión/Lineales... que proporciona el SPSS, obtenemos los siguientes resultados:

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación |
|--------|-------------------|------------|----------------------|-----------------------------|
| 1 | ,848 ^a | ,719 | ,703 | 2,772 |

a. Variables predictoras: (Constante), DOSIS

b. Variable dependiente: NIVELTOX

En la tabla Resumen del modelo, se muestran los resultados del ajuste del modelo de regresión. El valor del coeficiente de determinación, R cuadrado, mide la bondad del ajuste de la recta de regresión a la nube de puntos, valores pequeños de R cuadrado indican que el modelo no se ajusta bien a los datos.

R cuadrado toma un valor de 0.719 que nos indica que el 71.9% de la variabilidad de NIVELTOX, es explicada por la relación lineal con DOSIS.

El valor R (0.848) representa el valor absoluto del Coeficiente de Correlación, es decir es un valor entre 0 y 1. Valores próximos a 1 indican una fuerte relación entre las variables. La última columna nos muestra el Error típico de la estimación (raíz cuadrada de la varianza residual) con un valor igual a 2,772.

En cuadro siguiente se tiene la tabla ANOVA:

ANOVA^b

| Modelo | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|-------------|-------------------|----|------------------|--------|-------------------|
| 1 Regresión | 353,440 | 1 | 353,440 | 45,981 | ,000 ^a |
| Residual | 138,360 | 18 | 7,687 | | |
| Total | 491,800 | 19 | | | |

a. Variables predictoras: (Constante), DOSIS

b. Variable dependiente: NIVELTOX

En la Tabla ANOVA, se muestra la descomposición de la Variabilidad Total (491,8) en la Variabilidad debida a la Regresión (353,44) y la Variabilidad Residual (138,36), es decir, en Variabilidad explicada por el modelo de regresión y la Variabilidad no explicada. La Tabla de Análisis de la Varianza (Tabla ANOVA) se construye a partir de esta descomposición y proporciona el valor del estadístico F que permite contrastar la hipótesis nula de que la pendiente de la recta de regresión es igual a cero contra la alternativa de que la pendiente es distinta de cero, es decir:

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

donde H_0 se conoce, en general, como hipótesis de no linealidad entre X e Y.

La Tabla ANOVA muestra el valor del estadístico de contraste, $F = 45.981$, que se define como el cociente entre el Cuadrado medio debido a la regresión (353.44) y el Cuadrado medio residual (7.687), por tanto cuanto mayor sea su valor, mejor será la predicción mediante el modelo lineal. El p-valor asociado a F, en la columna Sig, es cero en su redondeo, menor que el nivel de significación $\alpha = 0.05$, lo que conduce a rechazar la hipótesis nula, es decir existe una relación lineal significativa entre las variables del problema.

"**Esto indica que es válido el modelo de regresión considerado, en este caso el modelo lineal simple.**"

La siguiente tabla muestra las estimaciones de los parámetros del modelo de regresión lineal simple:

Coeficientes^a

| Modelo | Coeficientes no estandarizados | | Coeficientes estandarizados | t | Sig. | Intervalo de confianza para B al 95% | |
|---------------|--------------------------------|------------|-----------------------------|-------|------|--------------------------------------|-----------------|
| | B | Error típ. | | | | Límite inferior | Límite superior |
| 1 (Constante) | -,820 | 2,571 | ,848 | -,319 | ,753 | -6,222 | 4,582 |
| | .752 | ,111 | | 6,781 | ,000 | ,519 | ,985 |

a. Variable dependiente: NIVELTOX

El modelo presenta los siguientes parámetros: como ordenada en el origen, $a_0 = -0.82$ y la pendiente $a_1 = 0.752$.

Por tanto, la ecuación de la recta estimada o ajustada es: $y = -0.82 + 0.752x$. Así mismo, en esta tabla se presentan los resultados de los dos contrastes individuales de la significación de cada uno de estos parámetros:

$$\begin{cases} H_0 : a_0 = 0 \\ H_1 : a_0 \neq 0 \end{cases} \quad \begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

El primero de estos contrastes carece de interés en la mayoría de los casos ya que raramente el punto de corte de la recta de regresión con el eje de ordenadas (ordenada en el origen) será el punto (0,0). Además dicho punto de corte carece de significado casi siempre.

El segundo contraste, el contraste de la pendiente de la recta, es una alternativa equivalente al contraste que acabamos de comentar en la Tabla ANOVA. El estadístico de contraste que aparece en la columna t vale 6.781 tiene un p-valor asociado, columna Sig, menor que 0.001, menor que el nivel de significación $\alpha = 0.05$ que conduce al rechazo de la hipótesis nula y podemos afirmar que existe una relación lineal significativa entre Y y X.

En la última columna de la tabla se muestran los intervalos de confianza para a_0 y a_1 , al 95%. El intervalo para a_0 es (-6.222, 4.582), puesto que el cero pertenece al intervalo, se aceptaría la hipótesis nula y concluir que si la DOSIS es cero el NIVELTOX también lo es, por tanto al nivel de confianza del 95% el parámetro a_0 podría considerarse igual a cero.

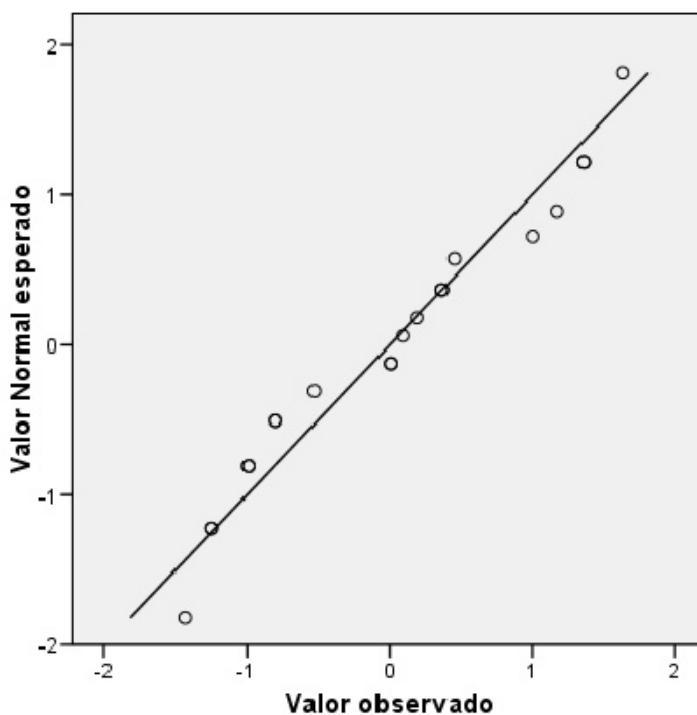
VALIDACIÓN Y DIAGNOSIS DEL MODELO

En este apartado vamos a comprobar que se verifican los supuestos del modelo de regresión lineal (normalidad, homocedasticidad (igualdad de varianzas) y linealidad) estos supuestos resultan necesarios para validar la inferencia respecto a los parámetros. Utilizaremos el análisis de los residuos para realizar los contrastes a posteriori de dichas hipótesis del modelo.

Normalidad

Podemos comprobarla de forma gráfica o analíticamente, gráficamente podemos estudiar el gráfico probabilístico normal, Para obtener dicho gráfico seleccionamos Analizar/Estadísticos descriptivos/Gráficos Q-Q..., obtenemos lo siguiente:

Gráfico Q-Q Normal de Standardized Residual



El Gráfico representa las funciones de distribución teórica y empírica de los residuos tipificados. Desviaciones de los puntos del gráfico respecto de la diagonal indican alteraciones de la normalidad. Observamos la ubicación de los puntos del gráfico, estos puntos se aproximan razonablemente bien a la diagonal lo que confirma la hipótesis de normalidad. Lo conformamos de forma analística mediante el contraste de Kolmogorov-Smirnov:

Prueba de Kolmogorov-Smirnov para una muestra

| | | Standardized Residual |
|--|--|-----------------------|
| N | | 20 |
| Parámetros normales^{a,b} | | |
| Media | | ,0000000 |
| Desviación típica | | ,97332853 |
| Diferencias más extremas | | |
| Absoluta | | ,145 |
| Positiva | | ,145 |
| Negativa | | -,103 |
| Z de Kolmogorov-Smirnov | | ,647 |
| Sig. asintót. (bilateral) | | ,797 |

a. La distribución de contraste es la Normal.

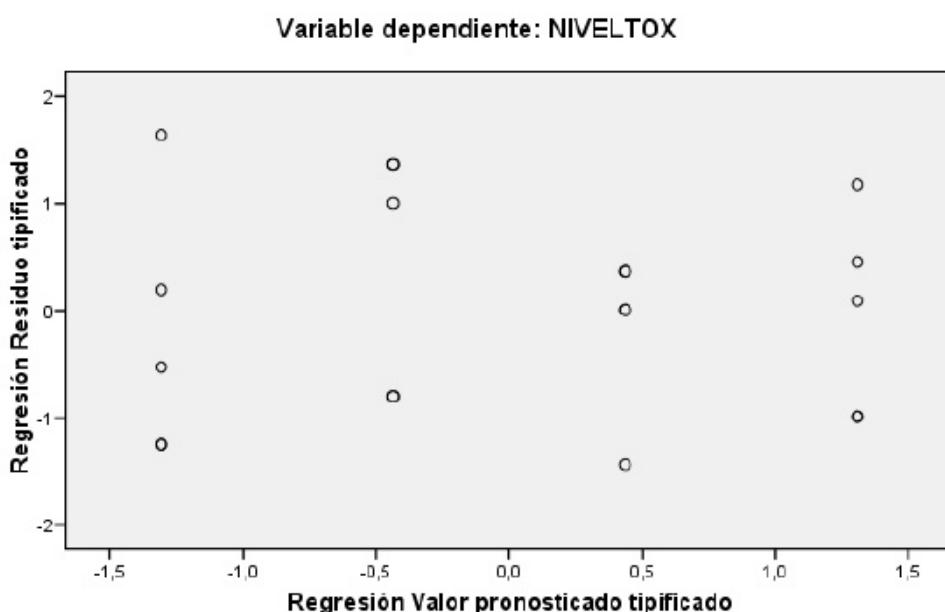
b. Se han calculado a partir de los datos.

Esta tabla muestra la mayor diferencia entre los resultados esperados en caso de que los residuos surgieran de una distribución normal y los valores observados. Se distingue entre la mayor diferencia en valor absoluto, la mayor diferencia positiva y la

mayor diferencia negativa. Se muestra el valor del estadístico Z (0.647) y el valor del p-valor asociado (0.797). Por lo tanto no se puede rechazar la hipótesis de normalidad de los residuos.

Homocedasticidad

Comprobamos la hipótesis de homogeneidad de las varianzas gráficamente representando los residuos tipificados frente a los tiempos de incubación estimados tipificados. El análisis de este gráfico puede revelar una posible violación de la hipótesis de homocedasticidad, por ejemplo si detectamos que el tamaño de los residuos aumenta o disminuye de forma sistemática para algunos valores ajustados de la variable NIVELTOX , si observamos que el gráfico muestra forma de embudo... Si por el contrario dicho gráfico no muestra patrón alguno, entonces no podemos rechazar la hipótesis de igualdad de varianzas.



No apreciamos tendencia clara en este gráfico, los residuos no presentan estructura definida respecto de los valores predichos por el modelo por lo que no debemos rechazar la hipótesis de homocedasticidad.

Este mismo gráfico resulta muy útil para detectar indicios de falta de adecuación del modelo propuesto a los datos, posibles desviaciones de la hipótesis de linealidad. Si observamos trayectorias de comportamiento no aleatorio esto es indicio de que el modelo propuesto no describe adecuadamente los datos.

Independencia de los residuos

La hipótesis de independencia de los residuos la realizaremos mediante el contraste de Durbin-Watson. Para ello se selecciona Analizar/Regresión/Lineal...

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación | Durbin-Watson |
|--------|-------------------|------------|----------------------|-----------------------------|---------------|
| 1 | ,848 ^a | ,719 | ,703 | 2,772 | 2,399 |

a. Variables predictoras: (Constante), DOSIS

b. Variable dependiente: NIVELTOX

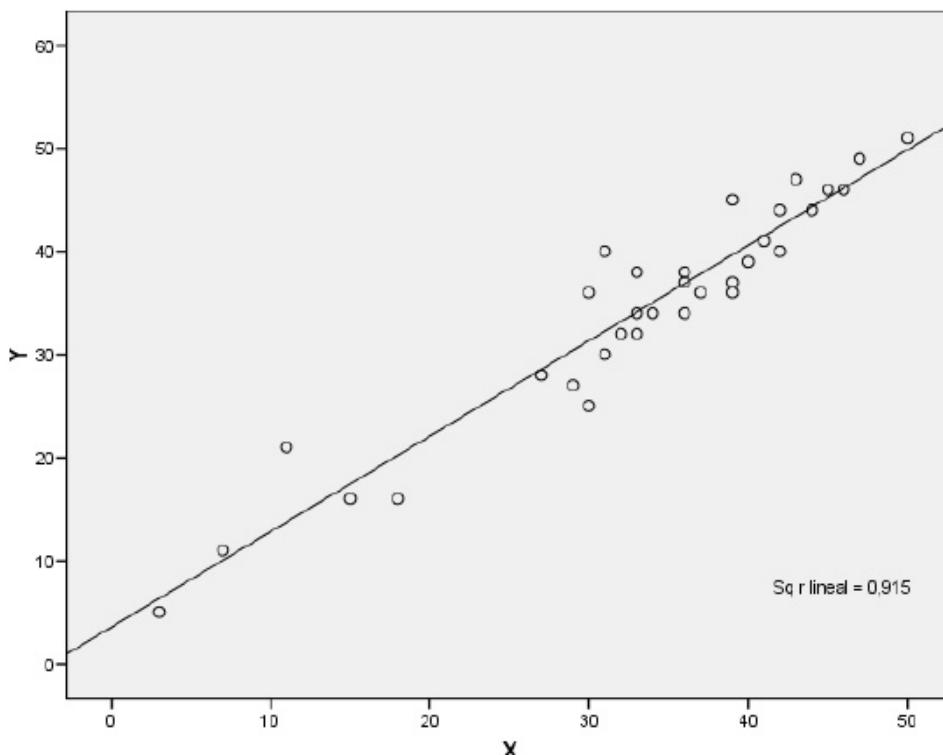
SPSS proporciona el valor del estadístico de *Durbin-Watson* pero no muestra el p-valor asociado por lo que hay que utilizar las tablas correspondientes. El estadístico de *Durbin-Watson* mide el grado de autocorrelación entre el residuo correspondiente a cada observación y la anterior. Si su valor está próximo a 2, entonces los residuos están incorrelados, si se aproxima a 4, estarán negativamente autocorrelados y si su valor está cercano a 0 estarán positivamente autocorrelados. En nuestro caso, toma el valor 2.399 próximo a 2 lo que indica la incorrelación de los residuos.

2. Uno de los problemas con los que se encuentran las industrias químicas es el tratamiento de las aguas residuales. Dichas aguas son químicamente complejas puesto que se caracterizan por altos valores de gases tóxicos, sólidos volátiles y otras sustancias nocivas. Los datos siguientes se obtuvieron a partir de 33 muestras de aguas residuales tratadas en el Instituto Politécnico de Virginia. En dichos datos se muestran los porcentajes de reducción de gases tóxicos (Y) durante el proceso de depuración para distintos valores de reducción en los sólidos volátiles (X). El objetivo es determinar un modelo que permita predecir la disminución porcentual en la concentración de gases tóxicos conocida la disminución de sólidos volátiles durante la depuración de las aguas residuales.

| X | Y | X | Y | X | Y | X | Y |
|----|----|----|----|----|----|----|----|
| 3 | 5 | 31 | 30 | 37 | 36 | 42 | 44 |
| 7 | 11 | 31 | 40 | 33 | 38 | 43 | 47 |
| 11 | 21 | 32 | 32 | 39 | 37 | 44 | 44 |
| 15 | 16 | 33 | 34 | 39 | 36 | 45 | 46 |
| 18 | 16 | 33 | 32 | 39 | 45 | 46 | 46 |
| 27 | 28 | 34 | 34 | 40 | 39 | 47 | 49 |
| 29 | 27 | 36 | 37 | 41 | 41 | 50 | 51 |
| 30 | 25 | 36 | 38 | 42 | 40 | | |
| 30 | 36 | 36 | 34 | | | | |

Realizar un análisis de regresión simple sobre los datos, incluyendo el contraste de linealidad. Estudiar posibles datos atípicos, afectan al ajuste realizado? Afectan a alguna de las hipótesis del modelo? Es admisible que la variabilidad de los residuos aumente o disminuya con la propia variable explicativa? Justifica estas cuestiones.

Empezamos el problema mediante la presentación del diagrama de dispersión entre ambas variables, y la representación de la recta de regresión aproximada. Los diagramas de dispersión ofrecen una idea bastante aproximada sobre el tipo de relación existente entre dos variables, además, también puede utilizarse como una forma de cuantificar el grado de relación lineal existente entre dos variables, basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.



El gráfico muestra una posible adecuación del modelo lineal y la tendencia creciente del mismo.

Para obtener la recta de regresión mínima cuadrática de Y sobre X , representada en la nube de puntos,

$$Y = a_0 + a_1 X$$

Para ello utilizamos la opción Analizar/Regresión/Lineales... que proporciona el SPSS, obtenemos los siguientes resultados:

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación | Durbin-Watson |
|--------|-------------------|------------|----------------------|-----------------------------|---------------|
| 1 | ,957 ^a | ,915 | ,912 | 3,254 | 2,395 |

a. Variables predictoras: (Constante), X

b. Variable dependiente: Y

En la tabla Resumen del modelo, se muestran los resultados del ajuste del modelo de regresión. El valor del coeficiente de determinación, R cuadrado, mide la bondad del ajuste de la recta de regresión a la nube de puntos, valores pequeños de R cuadrado indican que el modelo no se ajusta bien a los datos.

R cuadrado toma un valor de 0.915 que nos indica que el 91.5% de la variabilidad de Y, es explicada por la relación lineal con X.

El valor R (0.957) representa el valor absoluto del Coeficiente de Correlación, es decir es un valor entre 0 y 1. Valores próximos a 1 indican una fuerte relación entre las variables. La penúltima columna nos muestra el Error típico de la estimación (raíz cuadrada de la varianza residual) con un valor igual a 3.254.

En cuadro siguiente se tiene la tabla ANOVA:

| ANOVA ^b | | | | | |
|--------------------|-------------------|----|------------------|---------|-------------------|
| Modelo | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
| 1 Regresión | 3543,657 | 1 | 3543,657 | 334,693 | ,000 ^a |
| Residual | 328,222 | 31 | 10,588 | | |
| Total | 3871,879 | 32 | | | |

a. Variables predictoras: (Constante), X

b. Variable dependiente: Y

En la Tabla ANOVA, se muestra la descomposición de la Variabilidad Total (3871.879) en la Variabilidad debida a la Regresión (3543.657) y la Variabilidad Residual (328.222), es decir, en Variabilidad explicada por el modelo de regresión y la Variabilidad no explicada. La Tabla de Análisis de la Varianza (Tabla ANOVA) se construye a partir de esta descomposición y proporciona el valor del estadístico F que permite contrastar la hipótesis nula de que la pendiente de la recta de regresión es igual a cero contra la alternativa de que la pendiente es distinta de cero, es decir:

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

donde H_0 se conoce, en general, como hipótesis de no linealidad entre X e Y.

La Tabla ANOVA muestra el valor del estadístico de contraste, $F = 334.693$, que se define como el cociente entre el Cuadrado medio debido a la regresión (3543.657) y el Cuadrado medio residual (10.588), por tanto cuanto mayor sea su valor, mejor será la predicción mediante el modelo lineal. El p-valor asociado a F, en la columna Sig, es cero en su redondeo, menor que el nivel de significación $\alpha = 0.05$, lo que conduce a rechazar la hipótesis nula, es decir existe una relación lineal significativa entre las variables del problema.

"Esto indica que es válido el modelo de regresión considerado, en este caso el modelo lineal simple."

La siguiente tabla muestra las estimaciones de los parámetros del modelo de regresión lineal simple:

Coeficientes^a

| Modelo | Coeficientes no estandarizados | | Coeficientes estandarizados | t | Sig. | Intervalo de confianza para B al 95% | |
|---------------|--------------------------------|------------|-----------------------------|--------|------|--------------------------------------|-----------------|
| | B | Error típ. | | | | Límite inferior | Límite superior |
| 1 (Constante) | 3,549 | 1,779 | | 1,995 | ,055 | -,078 | 7,177 |
| X | ,926 | ,051 | ,957 | 18,295 | ,000 | ,823 | 1,029 |

a. Variable dependiente: Y

El modelo presenta los siguientes parámetros: como ordenada en el origen, $a_0 = 3.549$ y la pendiente $a_1 = 0.926$.

Por tanto, la ecuación de la recta estimada o ajustada es: $y = 3.549 + 0.926x$. Así mismo, en esta tabla se presentan los resultados de los dos contrastes individuales de la significación de cada uno de estos parámetros:

$$\begin{cases} H_0 : a_0 = 0 \\ H_1 : a_0 \neq 0 \end{cases} \quad \begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

El primero de estos contrastes carece de interés en la mayoría de los casos ya que raramente el punto de corte de la recta de regresión con el eje de ordenadas (ordenada en el origen) será el punto (0,0). Además dicho punto de corte carece de significado casi siempre.

El segundo contraste, el contraste de la pendiente de la recta, es una alternativa equivalente al contraste que acabamos de comentar en la Tabla ANOVA. El estadístico de contraste que aparece en la columna t vale 18.295 tiene un p-valor asociado, columna Sig, menor que 0.001, menor que el nivel de significación $\alpha = 0.05$ que conduce al rechazo de la hipótesis nula y podemos afirmar que existe una relación lineal significativa entre Y y X.

En la última columna de la tabla se muestran los intervalos de confianza para a_0 y a_1 , al 95%. El intervalo para a_0 es (-0.078, 7.177), puesto que el cero pertenece al intervalo, se aceptaría la hipótesis nula y concluir que si la variable X es cero la variable Y también lo es, por tanto al nivel de confianza del 95% el parámetro a_0 podría considerarse igual a cero.

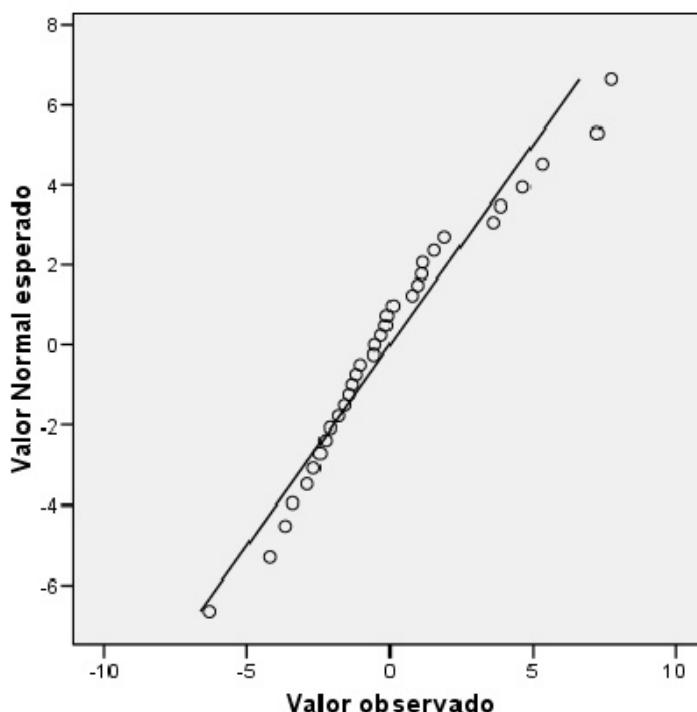
VALIDACIÓN Y DIAGNOSIS DEL MODELO

En este apartado vamos a comprobar que se verifican los supuestos del modelo de regresión lineal (normalidad, homocedasticidad (igualdad de varianzas) y linealidad) estos supuestos resultan necesarios para validar la inferencia respecto a los parámetros. Utilizaremos el análisis de los residuos para realizar los contrastes a posteriori de dichas hipótesis del modelo.

Normalidad

Podemos comprobarla de forma gráfica o analíticamente, gráficamente podemos estudiar el gráfico probabilístico normal. Para obtener dicho gráfico seleccionamos Analizar/Estadísticos descriptivos/Gráficos Q-Q..., obtenemos lo siguiente:

Gráfico Q-Q Normal de Unstandardized Residual



El Gráfico representa las funciones de distribución teórica y empírica de los residuos tipificados. Desviaciones de los puntos del gráfico respecto de la diagonal indican alteraciones de la normalidad. Observamos la ubicación de los puntos del gráfico, estos puntos se aproximan razonablemente bien a la diagonal lo que confirma la hipótesis de normalidad. Lo conformamos de forma analística mediante el contraste de Kolmogorov-Smirnov:

Prueba de Kolmogorov-Smirnov para una muestra

| | | Unstandardiz ed Residual |
|------------------------------------|-------------------|-----------------------------|
| N | | 33 |
| Parámetros normales ^{a,b} | Media | ,0000000 |
| | Desviación típica | 3,20264426 |
| Diferencias más extre mas | Absoluta | ,123 |
| | Positiva | ,123 |
| | Negativa | -,065 |
| Z de Kolmogorov-Smirnov | | ,706 |
| Sig. asintót. (bilateral) | | ,701 |

a. La distribución de contraste es la Normal.

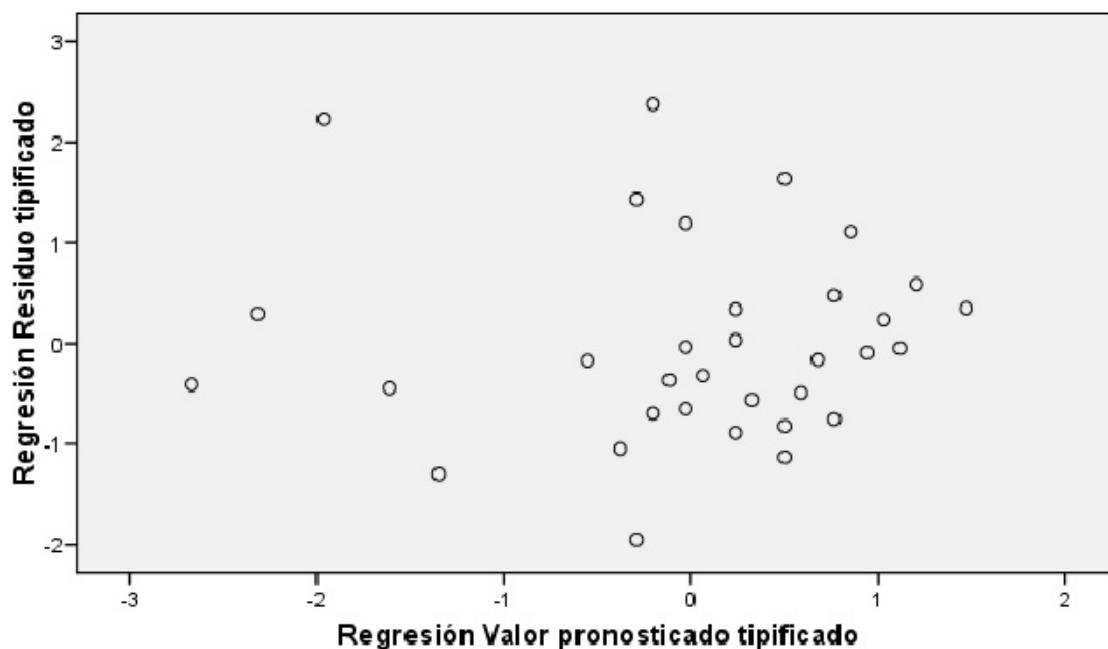
b. Se han calculado a partir de los datos.

Esta tabla muestra la mayor diferencia entre los resultados esperados en caso de que los residuos surgieran de una distribución normal y los valores observados. Se distingue entre la mayor diferencia en valor absoluto, la mayor diferencia positiva y la mayor diferencia negativa. Se muestra el valor del estadístico Z (0.706) y el valor del p-valor asociado (0.701). Por lo tanto no se puede rechazar la hipótesis de normalidad de los residuos.

Homocedasticidad

Comprobamos la hipótesis de homogeneidad de las varianzas gráficamente representando los residuos tipificados frente a los tiempos de incubación estimados tipificados. El análisis de este gráfico puede revelar una posible violación de la hipótesis de homocedasticidad, por ejemplo si detectamos que el tamaño de los residuos aumenta o disminuye de forma sistemática para algunos valores ajustados de la variable Y, si observamos que el gráfico muestra forma de embudo... Si por el contrario dicho gráfico no muestra patrón alguno, entonces no podemos rechazar la hipótesis de igualdad de varianzas.

Variable dependiente: Y



No apreciamos tendencia clara en este gráfico, los residuos no presentan estructura definida respecto de los valores predichos por el modelo por lo que no debemos rechazar la hipótesis de homocedasticidad.

Este mismo gráfico resulta muy útil para detectar indicios de falta de adecuación del modelo propuesto a los datos, posibles desviaciones de la hipótesis de linealidad. Si observamos trayectorias de comportamiento no aleatorio esto es indicio de que el modelo propuesto no describe adecuadamente los datos.

Independencia de los residuos

La hipótesis de independencia de los residuos la realizaremos mediante el contraste de Durbin-Watson. Para ello se selecciona Analizar/Regresión/Lineal...

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación | Durbin-Watson |
|--------|-------------------|------------|----------------------|-----------------------------|---------------|
| 1 | .957 ^a | .915 | .912 | 3,254 | 2,395 |

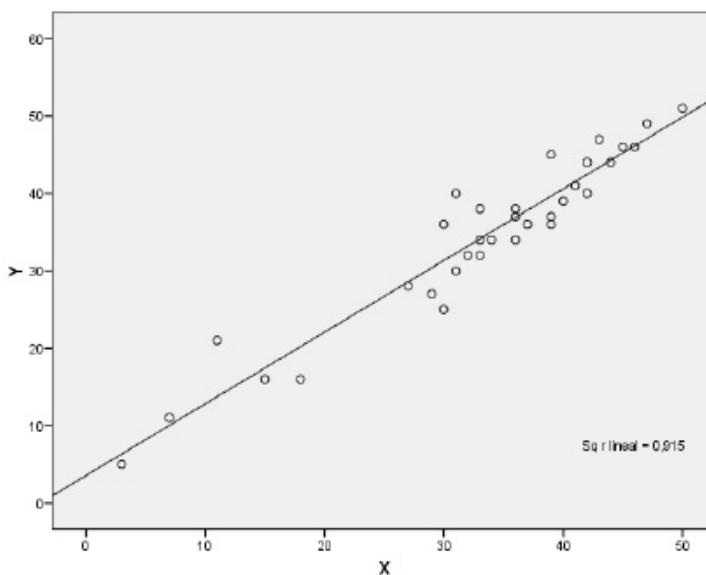
a. Variables predictoras: (Constante), X

b. Variable dependiente: Y

SPSS proporciona el valor del estadístico de *Durbin-Watson* pero no muestra el p-valor asociado por lo que hay que utilizar las tablas correspondientes. El estadístico de *Durbin-Watson* mide el grado de autocorrelación entre el residuo correspondiente a cada observación y la anterior. Si su valor está próximo a 2, entonces los residuos están incorrelados, si se aproxima a 4, estarán negativamente autocorrelados y si su valor está cercano a 0 estarán positivamente autocorrelados. En nuestro caso, toma el valor 2.395 próximo a 2 lo que indica la incorrelación de los residuos.

Estudiar posibles datos atípicos, afectan al ajuste realizado? ¿Afectan a alguna de las hipótesis del modelo?

En la nube de puntos podemos ver de forma gráfica si existen o no datos atípicos o anómalos que puedan influir en el estudio regresión lineal, para nuestro caso observamos datos separados de la recta de regresión generada, aunque no se observa datos con gran relevancia. Para hacer un estudio de si hay o no datos atípicos podemos analizar los residuos.



Los residuos son muy importantes en el análisis de regresión. En primer lugar, nos informan sobre el grado de exactitud de los pronósticos: cuanto más pequeño es el error típico de los residuos, mejores son los pronósticos, o lo que es lo mismo, mejor

se ajusta la recta de regresión a la nube de puntos. En segundo lugar, el análisis de las características de los casos con residuos grandes (grandes en valor absoluto) puede ayudarnos a detectar casos atípicos y, consecuentemente, a perfeccionar la ecuación de regresión a través de un estudio detallado de los mismos. El SPSS nos ofrece la opción "Diagnósticos por caso" del cuadro de diálogo Regresión lineal: Estadísticos, ofrece un listado de todos los residuos o, alternativamente (y esto es más interesante), un listado de los residuos que se alejan de cero (el valor esperado de los residuos) en más de un determinado número de desviaciones típicas. Es fácil, por tanto, identificar los casos que poseen residuos grandes.

Hemos elegido el valor de 2, puesto que no hay residuos que se alejen más de 3, que es el valor que viene por defecto. El resultado que proporciona el SPSS es de dos valores atípicos encontrados:

Diagnósticos por caso^a

| Número de caso | Residuo tip. | Y | Valor pronosticado | Residuo bruto |
|----------------|--------------|----|--------------------|---------------|
| 3 | 2,232 | 21 | 13,74 | 7,263 |
| 11 | 2,378 | 40 | 32,26 | 7,739 |

a. Variable dependiente: Y

Los datos atípicos pueden afectar al modelo estimado de regresión, así como a las hipótesis de normalidad y homocedasticidad cuando estos sean relevantes por lo que merecen un estudio en profundidad, cuando se tienen identificados los datos atípicos podemos:

- Eliminar los puntos si realmente no presentan ningún interés.
- Crear una variable ficticia que trate de medir el efecto del punto sobre el modelo y que lo caracterice como punto especial proveniente de otra población.

Es admisible que la variabilidad de los residuos aumente o disminuya con la propia variable explicativa?

Puesto que siempre se cumple la igualdad:

$$SS_E = S_{YY} - S_{\hat{Y}\hat{Y}}$$

y la variable Y depende de la variable X , es lógico que la variabilidad de los residuos aumente o disminuya con la propia variable explicativa.

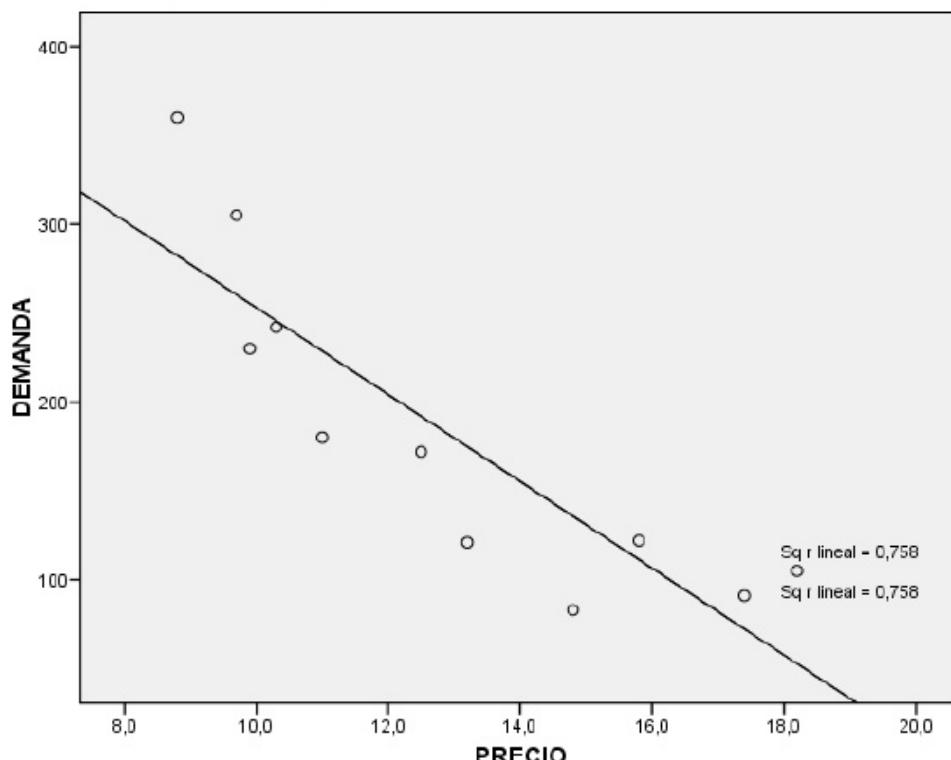
3. Un comerciante est interesado en conocer el comportamiento de la demanda de un cierto producto en funcin de los precios sobre los que oscila su venta, con el objeto de introducir dicho producto en un nuevo pas. Para ello observ, en distintos pases, el precio de venta al publico y la demanda en unidades de dicho producto. Los datos fueron los siguientes:

| | | | | | | | | | | | |
|---------|-----|-----|-----|------|-----|------|------|------|------|------|------|
| Demanda | 360 | 305 | 230 | 242 | 180 | 172 | 121 | 83 | 122 | 91 | 105 |
| Precio | 8.8 | 9.7 | 9.9 | 10.3 | 11 | 12.5 | 13.2 | 14.8 | 15.8 | 17.4 | 18.2 |

Se pretende estudiar la evolucin de la demanda en funcin de los precios. Realizar la regresin simple asociada a este experimento (contrastos de linealidad, sobre los parmetros, datos atpicos...).

Llamamos a las variables PRECIO (Variable Independiente) y DEMANDA (Variable Dependiente) que vienen recogidas en archivo ejercicio3.sav de la carpeta de datos.

Empezamos el problema mediante la presentación del diagrama de dispersión entre ambas variables, y la representación de la recta de regresión aproximada. Los diagramas de dispersión ofrecen una idea bastante aproximada sobre el tipo de relación existente entre dos variables, además, también puede utilizarse como una forma de cuantificar el grado de relación lineal existente entre dos variables, basta con observar el grado en el que la nube de puntos se ajusta a una línea recta.



El gráfico muestra una posible adecuación del modelo lineal y la tendencia decreciente del mismo.

Para obtener la recta de regresión mínima cuadrática de Y sobre X , representada en la nube de puntos,

$$Y = a_0 + a_1 X$$

Para ello utilizamos la opción Analizar/Regresión/Lineales... que proporciona el SPSS, obtenemos los siguientes resultados:

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error típ. de la estimación | Durbin-Watson |
|--------|-------------------|------------|----------------------|-----------------------------|---------------|
| 1 | ,871 ^a | ,758 | ,732 | 47,528 | ,771 |

a. Variables predictoras: (Constante), PRECIO

b. Variable dependiente: DEMANDA

En la tabla Resumen del modelo, se muestran los resultados del ajuste del modelo de regresión. El valor del coeficiente de determinación, R cuadrado, mide la bondad del ajuste de la recta de regresión a la nube de puntos, valores pequeños de R cuadrado indican que el modelo no se ajusta bien a los datos.

R cuadrado toma un valor de 0.758 que nos indica que el 75.8% de la variabilidad de DEMANDA, es explicada por la relación lineal con PRECIO.

El valor R (0.871) representa el valor absoluto del Coeficiente de Correlación, es decir es un valor entre 0 y 1. Valores próximos a 1 indican una fuerte relación entre las variables. La penúltima columna nos muestra el Error típico de la estimación (raíz cuadrada de la varianza residual) con un valor igual a 47,528.

En cuadro siguiente se tiene la tabla ANOVA:

ANOVA^b

| Modelo | Suma de cuadrados | gl | Media cuadrática | F | Sig. |
|-------------|-------------------|----|------------------|--------|-------------------|
| 1 Regresión | 63815,230 | 1 | 63815,230 | 28,250 | ,000 ^a |
| Residual | 20330,406 | 9 | 2258,934 | | |
| Total | 84145,636 | 10 | | | |

a. Variables predictoras: (Constante), PRECIO

b. Variable dependiente: DEMANDA

En la Tabla ANOVA, se muestra la descomposición de la Variabilidad Total (84145.636) en la Variabilidad debida a la Regresión (63815.23) y la Variabilidad Residual (20330.406), es decir, en Variabilidad explicada por el modelo de regresión y la Variabilidad no explicada. La Tabla de Análisis de la Varianza (Tabla ANOVA) se construye a partir de esta descomposición y proporciona el valor del estadístico F que permite contrastar la hipótesis nula de que la pendiente de la recta de regresión es igual a cero contra la alternativa de que la pendiente es distinta de cero, es decir:

$$\begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

donde H_0 se conoce, en general, como hipótesis de no linealidad entre X e Y.

La Tabla ANOVA muestra el valor del estadístico de contraste, $F = 28.25$, que se define como el cociente entre el Cuadrado medio debido a la regresión (63815.23) y el Cuadrado medio residual (2258.934), por tanto cuanto mayor sea su valor, mejor será la predicción mediante el modelo lineal. El p-valor asociado a F , en la columna Sig, es cero en su redondeo, menor que el nivel de significación $\alpha = 0.05$, lo que conduce a rechazar la hipótesis nula, es decir existe una relación lineal significativa entre las variables del problema.

"Esto indica que es válido el modelo de regresión considerado, en este caso el modelo lineal simple."

La siguiente tabla muestra las estimaciones de los parámetros del modelo de regresión lineal simple:

Coeficientes^a

| Modelo | Coeficientes no estandarizados | | Coeficientes estandarizados Betas | t | Sig. | Intervalo de confianza para B al 95% | |
|---------------|--------------------------------|------------|--------------------------------------|--------|------|--------------------------------------|-----------------|
| | B | Error típ. | | | | Límite inferior | Límite superior |
| 1 (Constante) | 497,156 | 60,852 | | 8,170 | ,000 | 359,499 | 634,813 |
| PRECIO | -24,419 | 4,594 | -,871 | -5,315 | ,000 | -34,812 | -14,026 |

a. Variable dependiente: DEMANDA

El modelo presenta los siguientes parámetros: como ordenada en el origen, $a_0 = 497.156$ y la pendiente $a_1 = -24.419$.

Por tanto, la ecuación de la recta estimada o ajustada es: $y = 497.156 - 24.419x$. Así mismo, en esta tabla se presentan los resultados de los dos contrastes individuales de la significación de cada uno de estos parámetros:

$$\begin{cases} H_0 : a_0 = 0 \\ H_1 : a_0 \neq 0 \end{cases} \quad \begin{cases} H_0 : a_1 = 0 \\ H_1 : a_1 \neq 0 \end{cases}$$

El primero de estos contrastes carece de interés en la mayoría de los casos ya que raramente el punto de corte de la recta de regresión con el eje de ordenadas (ordenada en el origen) será el punto (0,0). Además dicho punto de corte carece de significado casi siempre.

El segundo contraste, el contraste de la pendiente de la recta, es una alternativa equivalente al contraste que acabamos de comentar en la Tabla ANOVA. El estadístico de contraste que aparece en la columna t vale 8.17 tiene un p-valor asociado, columna Sig, menor que 0.001, menor que el nivel de significación $\alpha = 0.05$ que conduce al rechazo de la hipótesis nula y podemos afirmar que existe una relación lineal significativa entre Y y X.

En la última columna de la tabla se muestran los intervalos de confianza para a_0 y a_1 , al 95%. El intervalo para a_0 es (359.499 , 634.813), puesto que el cero no pertenece al intervalo, se rechazaría la hipótesis nula.

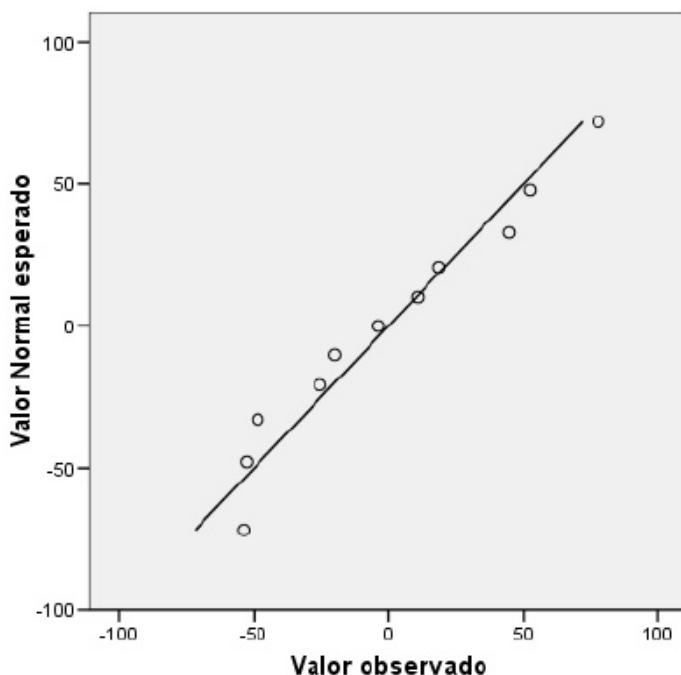
VALIDACIÓN Y DIAGNOSIS DEL MODELO

En este apartado vamos a comprobar que se verifican los supuestos del modelo de regresión lineal (normalidad, homocedasticidad (igualdad de varianzas) y linealidad) estos supuestos resultan necesarios para validar la inferencia respecto a los parámetros. Utilizaremos el análisis de los residuos para realizar los contrastes a posteriori de dichas hipótesis del modelo.

Normalidad

Podemos comprobarla de forma gráfica o analíticamente, gráficamente podemos estudiar el gráfico probabilístico normal, Para obtener dicho gráfico seleccionamos Analizar/Estadísticos descriptivos/Gráficos Q-Q..., obtenemos lo siguiente:

Gráfico Q-Q Normal de Unstandardized Residual



El Gráfico representa las funciones de distribución teórica y empírica de los residuos tipificados. Desviaciones de los puntos del gráfico respecto de la diagonal indican alteraciones de la normalidad. Observamos la ubicación de los puntos del gráfico, estos puntos se aproximan razonablemente bien a la diagonal lo que confirma la hipótesis de normalidad. Lo conformamos de forma analística mediante el contraste de Kolmogorov-Smirnov:

Prueba de Kolmogorov-Smirnov para una muestra

| | | Unstandardized Residual |
|------------------------------------|-------------------|-------------------------|
| N | | 11 |
| Parámetros normales ^{a,b} | Media | ,0000000 |
| | Desviación típica | 45,08925150 |
| Diferencias más extremas | Absoluta | ,132 |
| | Positiva | ,132 |
| | Negativa | -,116 |
| Z de Kolmogorov-Smirnov | | ,438 |
| Sig. asintót. (bilateral) | | ,991 |

a. La distribución de contraste es la Normal.

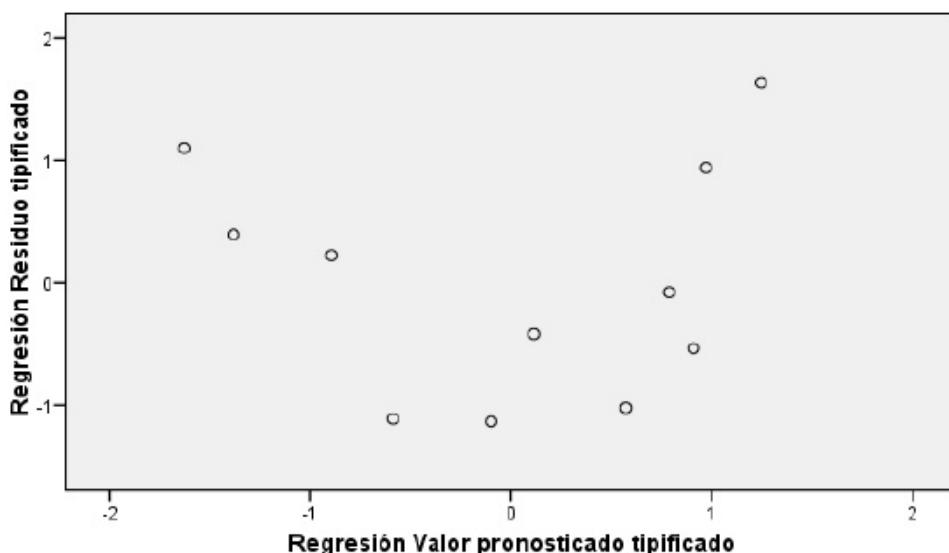
b. Se han calculado a partir de los datos.

Esta tabla muestra la mayor diferencia entre los resultados esperados en caso de que los residuos surgieran de una distribución normal y los valores observados. Se distingue entre la mayor diferencia en valor absoluto, la mayor diferencia positiva y la mayor diferencia negativa. Se muestra el valor del estadístico Z (0.438) y el valor del p-valor asociado (0.991). Por lo tanto no se puede rechazar la hipótesis de normalidad de los residuos.

Homocedasticidad

Comprobamos la hipótesis de homogeneidad de las varianzas gráficamente representando los residuos tipificados frente a los tiempos de incubación estimados tipificados. El análisis de este gráfico puede revelar una posible violación de la hipótesis de homocedasticidad, por ejemplo si detectamos que el tamaño de los residuos aumenta o disminuye de forma sistemática para algunos valores ajustados de la variable Y, si observamos que el gráfico muestra forma de embudo... Si por el contrario dicho gráfico no muestra patrón alguno, entonces no podemos rechazar la hipótesis de igualdad de varianzas.

Variable dependiente: DEMANDA



No apreciamos tendencia clara en este gráfico, los residuos no presentan estructura definida respecto de los valores predichos por el modelo por lo que no debemos rechazar la hipótesis de homocedasticidad.

Este mismo gráfico resulta muy útil para detectar indicios de falta de adecuación del modelo propuesto a los datos, posibles desviaciones de la hipótesis de linealidad. Si observamos trayectorias de comportamiento no aleatorio esto es indicio de que el modelo propuesto no describe adecuadamente los datos.

Independencia de los residuos

La hipótesis de independencia de los residuos la realizaremos mediante el contraste de Durbin-Watson. Para ello se selecciona Analizar/Regresión/Lineal...

Resumen del modelo^b

| Modelo | R | R cuadrado | R cuadrado corregida | Error t _p de la estimación | Durbin-Watson |
|--------|-------------------|------------|----------------------|---------------------------------------|---------------|
| 1 | ,871 ^a | ,758 | ,732 | 47,528 | ,771 |

a. Variables predictoras: (Constante), PRECIO

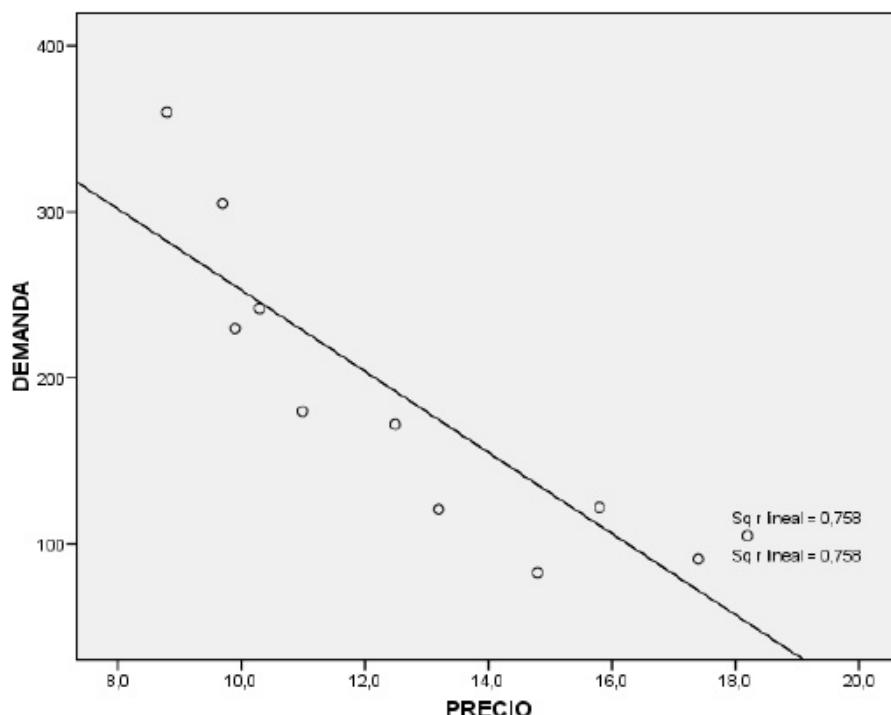
b. Variable dependiente: DEMANDA

SPSS proporciona el valor del estadístico de *Durbin-Watson* pero no muestra el p-valor asociado por lo que hay que utilizar las tablas correspondientes. El estadístico de *Durbin-Watson* mide el grado de autocorrelación entre el residuo correspondiente a cada observación y la anterior. Si su valor está próximo a 2, entonces los residuos están incorrelados, si se aproxima a 4, estarán negativamente autocorrelados y si su valor está cercano a 0 estarán positivamente autocorrelados. En nuestro caso, toma el valor 0.771 próximo a 0 lo que indica la correlación de los residuos.

Par solucionar la falta independencia al existir una correlación entre los residuos, podemos plantear una transformación de los valores o el añadir más datos al problema.

DATOS ATÍPICOS

En la nube de puntos podemos ver de forma gráfica si existen o no datos atípicos que puedan influir en el estudio regresión lineal, para nuestro caso observamos datos separados de la recta de regresión generada, que de entrada nos hace una idea de que puedan existir datos anómalos. Para hacer un estudio de si hay o no datos atípicos podemos analizar los residuos.



Los residuos son muy importantes en el análisis de regresión. En primer lugar, nos informan sobre el grado de exactitud de los pronósticos: cuanto más pequeño es el error típico de los residuos, mejores son los pronósticos, o lo que es lo mismo, mejor se ajusta la recta de regresión a la nube de puntos. En segundo lugar, el análisis de las características de los casos con residuos grandes (grandes en valor absoluto) puede ayudarnos a detectar casos atípicos y, consecuentemente, a perfeccionar la ecuación de regresión a través de un estudio detallado de los mismos. El SPSS nos ofrece la opción "Diagnósticos por caso" del cuadro de diálogo Regresión lineal: Estadísticos, ofrece un listado de todos los residuos o, alternativamente (y esto es más interesante), un listado de los residuos que se alejan de cero (el valor esperado de los residuos) en más de un determinado número de desviaciones típicas. Es fácil, por tanto, identificar los casos que poseen residuos grandes.

Hemos elegido el valor de 1.5, puesto que no hay residuos que se alejen más de 2. El resultado que proporciona el SPSS es de un valor atípico encontrado:

Diagnósticos por caso^a

| Número de caso | Residuo tip. | DEMANDA | Valor pronosticado | Residuo bruto |
|----------------|--------------|---------|--------------------|---------------|
| 1 | 1,635 | 360 | 282,27 | 77,730 |

a. Variable dependiente: DEMANDA

Los datos atípicos pueden afectar al modelo estimado de regresión, así como a las hipótesis de normalidad y homocedasticidad cuando estos sean relevantes por lo que merecen un estudio en profundidad, cuando se tienen identificados los datos atípicos podemos:

- Eliminar los puntos si realmente no presentan ningún interés.

- Crear una variable ficticia que trate de medir el efecto del punto sobre el modelo y que lo caracterice como punto especial proveniente de otra población.

EL RESTO DE EJERCICIOS SON DE REGRESIÓN LINEAL MÚLTIPLE Y VIENEN REPETIDOS EN LA ACTIVIDAD 4 Y HE DECIDIDO HACERLOS EN ESA ACTIVIDAD.