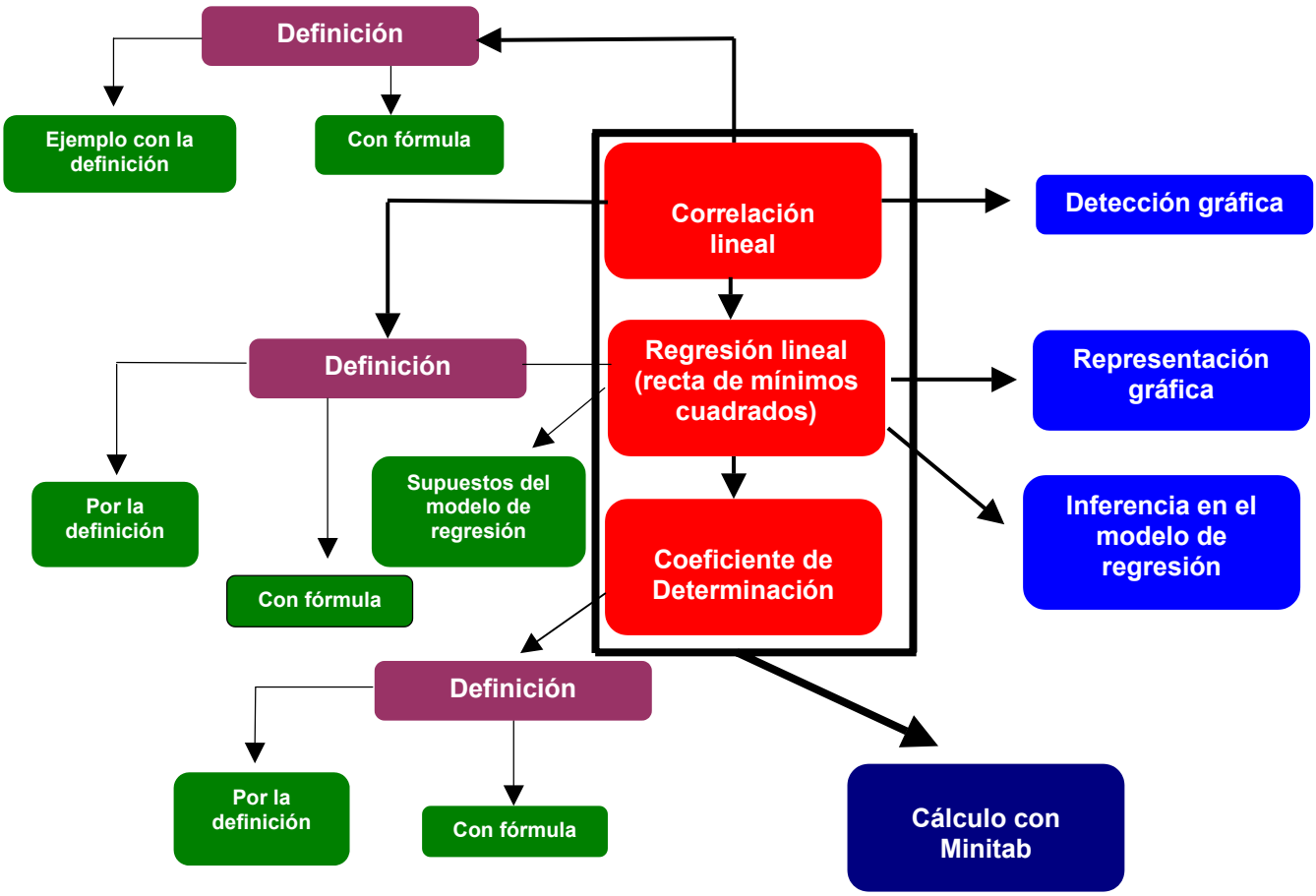
CORRELACIÓN LINEAL Y ANÁLISIS DE REGRESIÓN

Autores: Alicia Vila (avilag@uoc.edu), Máximo Sedano (msedanoh@uoc.edu), Ana López (alopezrat@uoc.edu), Ángel A. Juan (ajuanp@uoc.edu),

MAPA CONCEPTUAL



Financiado por la Secretaría de Estado de Educación y Universidades (MECD)

INTRODUCCIÓN

El objetivo de este math-block es analizar el grado de la relación existente entre variables utilizando modelos matemáticos y representaciones gráficas. Así pues, para representar la relación entre dos o más variables desarrollaremos una ecuación que permitirá estimar una variable en función de la otra.

Por ejemplo, ¿en qué medida, un aumento de los gastos en publicidad hace aumentar las ventas de un determinado producto?, ¿cómo representamos que la bajada de temperaturas implica un aumento del consumo de la calefacción?,...

A continuación, estudiaremos dicho grado de relación entre dos variables en lo que llamaremos *análisis de correlación*. Para representar esta relación utilizaremos una representación gráfica llamada *diagrama de dispersión* y, finalmente, estudiaremos un modelo matemático para estimar el valor de una variable basándonos en el valor de otra, en lo que llamaremos *análisis de regresión*.

OBJETIVOS

- Aprender a calcular la correlación entre dos variables
- Saber dibujar un diagrama de dispersión
- Representar la recta que define la relación lineal entre dos variables
- Saber estimar la recta de regresión por el método de mínimos cuadrados e interpretar su ajuste.
- Realizar inferencia sobre los parámetros de la recta de regresión
- Construir e interpretar intervalos de confianza e intervalos de predicción para la variable dependiente
- Realizar una prueba de hipótesis para determinar si el coeficiente de correlación es distinto de cero

CONOCIMIENTOS PREVIOS

Es recomendable haber leído, previamente, los *math-blocks* "Estimación puntual e intervalos de confianza" y "Contraste de hipótesis para dos poblaciones", así como los ejercicios asociados resueltos con Minitab.

CONCEPTOS FUNDAMENTALES

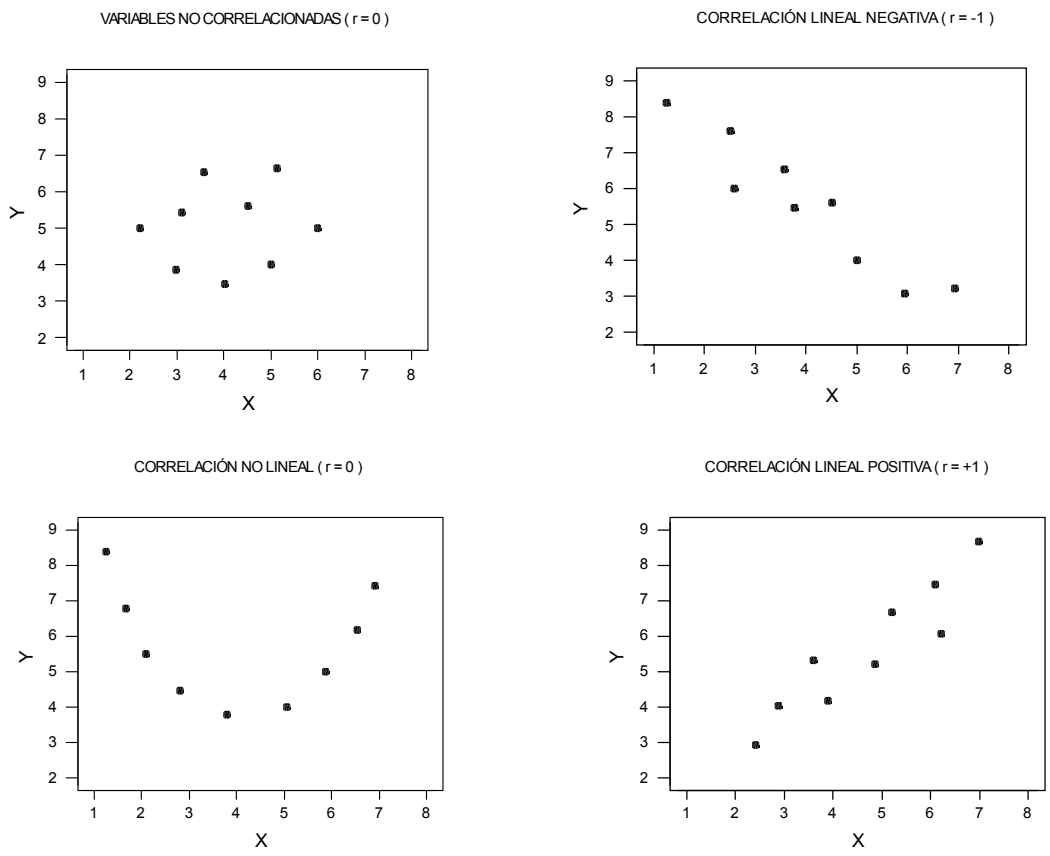
❑ Definición de Correlación Lineal

En ocasiones nos puede interesar estudiar si existe o no algún tipo de relación entre dos variables aleatorias. Así, por ejemplo, podemos preguntarnos si hay alguna relación entre las notas de la asignatura Estadística I y las de Matemáticas I. Una primera aproximación al problema consistiría en dibujar en el plano R^2 un punto por cada alumno: la primera coordenada de cada punto sería su nota en estadística, mientras que la segunda sería su nota en matemáticas. Así, obtendríamos una nube de puntos la cual podría indicarnos visualmente la existencia o no de algún tipo de relación (lineal, parabólica, exponencial, etc.) entre ambas notas.

Otro ejemplo, consistiría en analizar la facturación de una empresa en un periodo de tiempo dado y de cómo influyen los gastos de promoción y publicidad en dicha facturación. Si consideramos un periodo de tiempo de 10 años, una posible representación sería situar un punto por cada año de forma que la primera coordenada de cada punto sería la cantidad en euros invertidos en publicidad, mientras que la segunda sería la cantidad en euros obtenidos de su facturación. De esta manera, obtendríamos una nube de puntos que nos indicaría el tipo de relación existente entre ambas variables.

En particular, nos interesa cuantificar la intensidad de la relación **lineal** entre dos variables. El parámetro que nos da tal cuantificación es el **coeficiente de correlación lineal de Pearson r**, cuyo valor oscila entre –1 y +1 :

$$-1 \le r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum_{t=1}^{n}(X_t - \overline{X}) * (Y_t - \overline{Y})}{\sqrt{\sum_{t=1}^{n}(X_t - \overline{X})^2} * \sqrt{\sum_{t=1}^{n}(Y_t - \overline{Y})^2}} \le +1$$

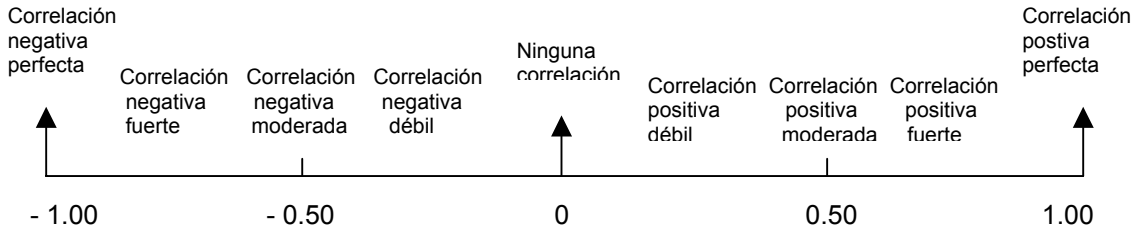


Como se observa en los diagramas anteriores, el valor de *r* se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de *X* significan mayores valores de *Y*), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa.

Es importante notar que la existencia de correlación entre variables no implica causalidad.

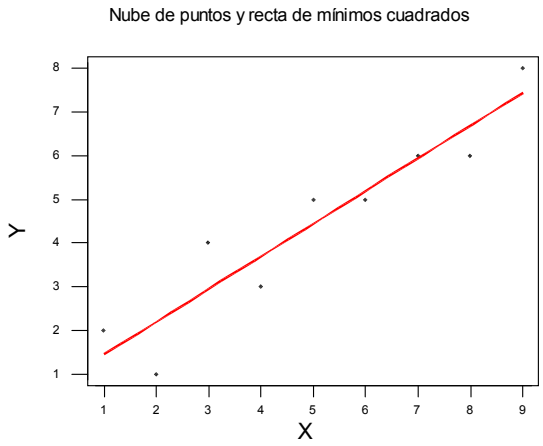
¡Atención!: si no hay correlación de ningún tipo entre dos v.a., entonces tampoco habrá correlación lineal, por lo que r = 0. Sin embargo, el que ocurra r = 0 sólo nos dice que no hay correlación lineal, pero puede que la haya de otro tipo.

El siguiente diagrama resume el análisis del coeficiente de correlación entre dos variables:



Definición y características del concepto de Regresión Lineal

En aquellos casos en que el coeficiente de regresión lineal sea "cercano" a +1 o a –1, tiene sentido considerar la ecuación de la recta que "mejor se ajuste" a la nube de puntos (recta de mínimos cuadrados). Uno de los principales usos de dicha recta será el de predecir o estimar los valores de Y que obtendríamos para distintos valores de X. Estos conceptos quedarán representados en lo que llamamos **diagrama de dispersión**:



La ecuación de la **recta de mínimos cuadrados** (en forma punto-pendiente) es la siguiente:

$$y - \bar{y} = \frac{Cov(X,Y)}{s_x^{\,2}}(x - \bar{x})$$

Veamos con detalle estos conceptos mediante un ejemplo:

Si queremos estudiar la relación existente entre ambas variables, siguiendo con el ejemplo anterior referente a la relación entre las ventas de una empresa (V_t) y sus gastos en publicidad (GP_t), lo que podemos hacer es representar gráficamente el modelo matemático lineal que podemos considerar para analizar dicha relación.

$$V_t = \beta_1 + \beta_2 * GP_t + u_t$$

Supongamos que disponemos de los siguientes datos:

Año	Ventas en millones de euros.	Gastos en publicidad en millones de euros.
1998	200	30
1999	400	50
2000	800	50
2001	1.200	60
2002	900	60

A partir de este modelo matemático lineal, vamos a analizar la relación entre ambas variables, la variable ventas (V_t) que es la variable dependiente del modelo y la variable que vamos a analizar y los gastos en publicidad (GP_t) que es la variable independiente o la variable explicativa que vamos a utilizar para estudiar las ventas.

En este modelo queremos comprobar qué influencia tienen los gastos de publicidad sobre el volumen de facturación o las ventas de la empresa.

Para poder cuantificar dicha relación, debemos también representar la recta de regresión que subyace en el modelo matemático que relaciona ambas variables.

Para cuantificar la relación entre ambas variables y tener un aproximación de la magnitud de la influencia de los gastos en publicidad sobre las ventas de la empresa debemos estimar el modelo por mínimos cuadrados ordinarios (M.C.O.) donde se minimiza la suma de los cuadrados de los residuos.

La recta en rojo (que aparece a continuación en el gráfico), es la que mejor se ajusta a la nube de puntos que tenemos. Dicho de otra forma, es la recta que hace que el error de estimación, definido como la distancia entre el valor observado y el valor estimado de la variable endógena (en el gráfico, es la distancia vertical señalada por la flecha en rojo), sea la mínima para cada una de las observaciones (recta de mínimos cuadrados), esta recta será la que utilizaremos para predecir o estimar los valores de Y que obtendremos para distintos valores de X.

La diferencia entre un valor observado y el valor estimado lo denominaremos **residuo**.

$$\textit{Residuo} = Y_t - \hat{Y}_t$$

Nuestro problema consiste en minimizar la suma de los cuadrados de los residuos de los cuadrados de los residuos, $\sum_{t=1}^{n} \hat{u}_t^2$. De este problema de optimización se deduce la expresión de mínimos cuadrados ordinarios del MRLM:

$$\text{Criterio MCO:} \qquad Min \sum_{t=1}^{n} \hat{u}_t^2$$

Como ya hemos citado anteriormente, la ecuación de la **recta de mínimos cuadrados** (en forma punto-pendiente) es la siguiente:

$$Y - \overline{Y} = \frac{Cov(X,Y)}{s_x^{\ 2}}(X - \overline{X}) = \frac{\sum_{t=1}^{n}(X_t - \overline{X}) * (Y_t - \overline{Y})}{\sum_{t=1}^{n}(X_t - \overline{X})^2}(X - \overline{X})$$

$\beta_2 = \dfrac{\sum_{t=1998}^{2002}(GP_t - \overline{GP})(V_t - \overline{V})}{\sum_{t=1998}^{2002}(GP_t - \overline{GP})^2} = \dfrac{17.000}{600} = 28{,}3$, ésta sería la estimación de la pendiente de la recta por mínimos cuadrados.

Por otro lado, $\beta_1 = \overline{V} - \beta_2 \overline{GP} = 700 - 28{,}333 * 50 = -716{,}6$, y ésta sería la estimación de la ordenada de la recta de regresión ó el punto de corte de la recta con los ejes.

Por tanto, $\mathrm{Y} = -716{,}6 + 28{,}3\mathrm{X}$

La representación gráfica de los datos anteriores es la que sigue:

Del diagrama anterior, cabe observar que no todos los puntos están en la línea de regresión. Si todos lo estuvieran y, además, si el número de observaciones fuera suficientemente grande, no habría ningún error de estimación. En ese caso, no habría ninguna diferencia entre el valor observado y el valor de predicción.

Como imaginamos, en los casos reales, las predicciones perfectas son prácticamente imposibles y lo que necesitamos es una medida que describa cómo de precisa es la predicción de Y en función de X o, inversamente, qué inexacta puede ser la estimación.

A esta medida se le llama **error estándar de estimación** y se denota $\mathbf{S_{yx}}$. El error estándar de estimación, es el mismo concepto que la desviación estándar, aunque ésta mide la dispersión alrededor de la media y el error estándar mide la dispersión alrededor de la línea de regresión.

Interpretación de los coeficientes estimados

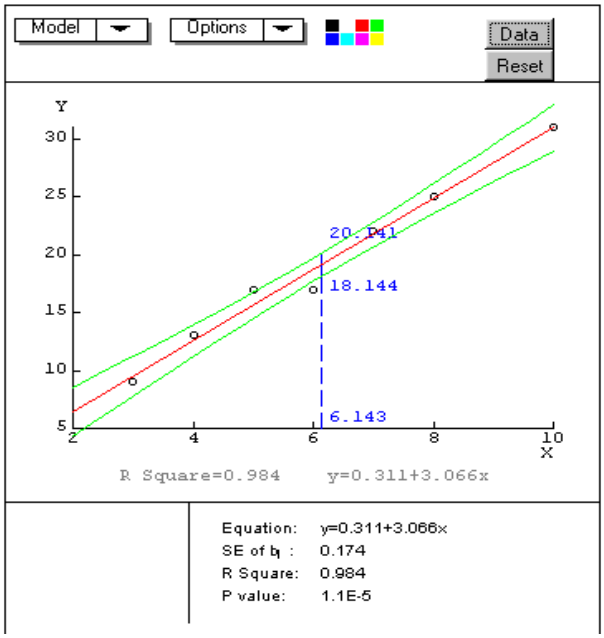
Según la recta de mínimos cuadrados, al incrementarse en un millón de euros los gastos en publicidad, la cantidad de facturación obtenida se incrementará en 28,3 millones de euros. Y cuando no se haga ningún esfuerzo publicitario, las ventas según la recta serán negativas. Esto se puede entender como que no se vende nada o que si no se hace ningún esfuerzo publicitario se obtienen unas ventas negativas, en el sentido de que hay otros gastos a la hora de vender que provocan que al final haya ventas negativas.

La correlación entre ambas variables es muy alta, ya que el coeficiente de correlación r = 0.87 está muy próximo a 1.

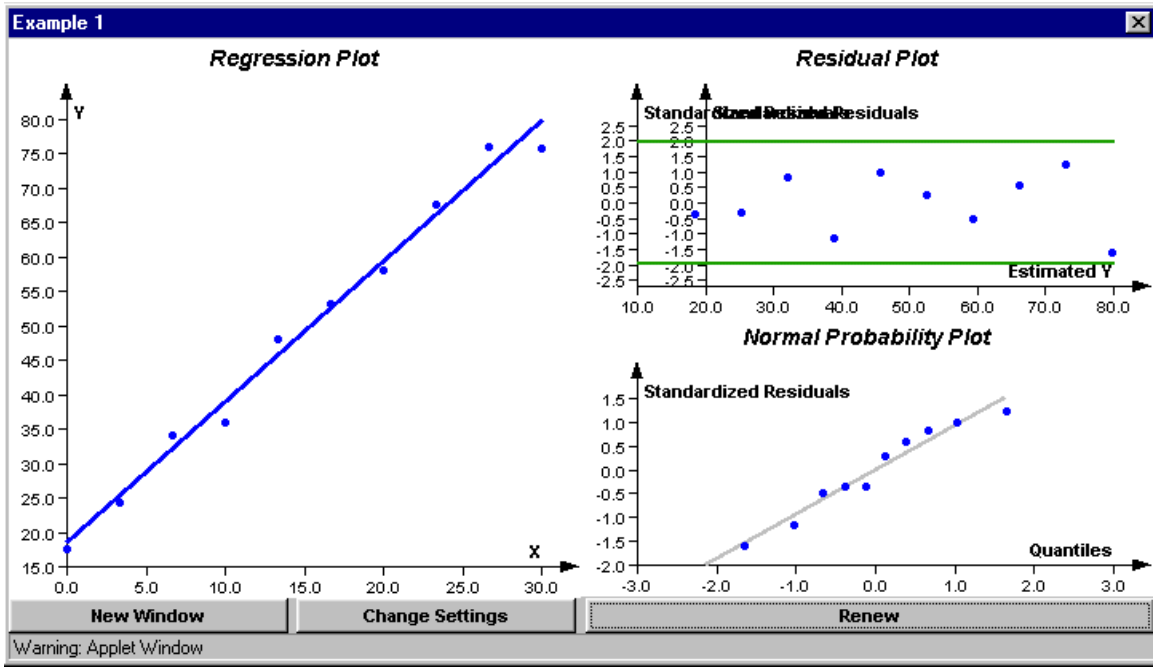
$$r = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum_{t=1998}^{2002}(GP_t - \overline{GP}) * (V_t - \overline{V})}{\sqrt{\sum_{t=1998}^{2002}(GP_t - \overline{GP})^2} * \sqrt{\sum_{t=1998}^{2002}(V_t - \overline{V})^2}} = 0{,}868$$

Para profundizar más en los conceptos vistos hasta el momento o para entender gráficamente como funcionan, a continuación citamos algunos enlaces web interesantes:

En el enlace: http://www.stat.wvu.edu/SRS/Modules/Applets/Regression/regression.html encontraremos un applet en el que modificando los datos de la variable X e Y podemos construir la recta de regresión. El gráfico resultante será similar al siguiente:

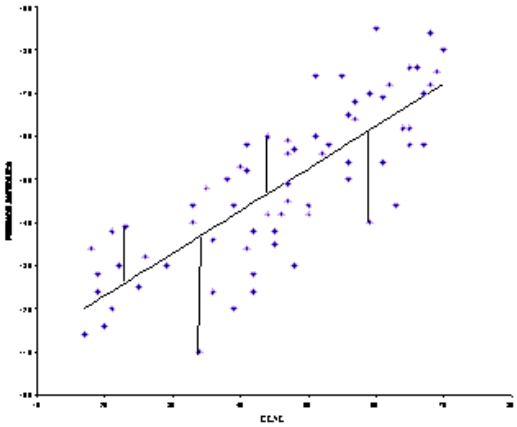


Un applet de similares características lo encontraremos en:
http://www.kuleuven.ac.be/ucs/java/version2.0/Applet010.html



Los conceptos de regresión lineal y correlación entre variables se aplican a innumerables aspectos de la vida real, tanto en el ámbito social, como científico,... En el siguiente enlace: http://www.fisterra.com/material/investiga/regre_lineal_simple/regre_lineal_simple.htm#1 encontramos un claro ejemplo de cómo utilizar estos conceptos para ver la relación entre la Tensión arterial sistólica y la edad, a partir de una muestra de 69 pacientes.

Figura 1. Relación entre la Edad y Presión Sistólica. Recta de Regresión y diferencias entre los valores observados y ajustados



Supuestos del modelo de regresión lineal

En el caso en que nuestras observaciones sean una muestra aleatoria proveniente de una población, estaremos interesados en realizar inferencias sobre la misma. A fin de que estas inferencias sean "estadísticamente razonables", se han de cumplir las siguientes condiciones:

1. En la población, la relación entre las variables *X* e *Y* debe ser aproximadamente lineal, i.e.: $y = \beta_1 + \beta_2 x + \varepsilon$, siendo ε la v.a. que representa los **residuos** (diferencias entre el valor estimado por el modelo y el verdadero valor de *Y*).

2. Los residuos se distribuyen según una Normal de media 0, i.e., $\varepsilon \approx N(0, \sigma^2)$.

3. Los residuos son independientes unos de otros.

4. Los residuos tienen varianza σ^2 constante.

Afortunadamente, el modelo de regresión lineal es bastante "robusto", lo que significa que no es necesario que las condiciones anteriores se cumplan con exactitud (en particular las tres últimas).

Definición del Coeficiente de Determinación

Denominamos **coeficiente de determinación R^2** como el coeficiente que nos indica el porcentaje del ajuste que se ha conseguido con el modelo lineal, es decir el porcentaje de la variación de Y(ventas) que se explica a través del modelo lineal que se ha estimado, es decir a través del comportamiento de X (publicidad) . A mayor porcentaje mejor es nuestro modelo para predecir el comportamiento de la variable Y

También se puede entender este coeficiente de determinación como el porcentaje de varianza explicada por la recta de regresión y su valor siempre estará entre 0 y 1 y siempre es igual al cuadrado del coeficiente de correlación (r).

$$R^2 = r^2$$

Es una medida de la proximidad o de ajuste de la recta de regresión a la nube de puntos. También se le denomina *bondad del ajuste*.

$1 - R^2$ nos indica qué porcentaje de las variaciones no se explica a través del modelo de regresión, es como si fuera la varianza inexplicada que es la varianza de los residuos.

En nuestro ejemplo, el coeficiente de determinación nos da bajo, el 75,3%, por lo que sólo conseguimos explicar el 75,3 % de las variaciones de las ventas a través del ajuste por medio de los gastos en publicidad.

Inferencia en el modelo de regresión

Una vez que hemos calculado la recta de regresión y el ajuste que hemos conseguido con el modelo de regresión lineal, el siguiente paso consiste en analizar si la regresión en efecto es válida y la podemos utilizar para predecir. Para ello debemos contrastar si la correlación entre ambas variables es distinta de cero o si el modelo de regresión es válido en el sentido de contrastar si el análisis de nuestra variable endógena (Y). es válido a través de la influencia de la variable explicativa (X).

Supongamos por un lado que el coeficiente de correlación lineal *r*, está próximo a +1 o a –1, y por tanto parece indicar la existencia de una correlación lineal entre los valores de la muestra. Pero este valor del coeficiente de correlación lineal muestral entre ambas variables no garantiza que también estén correlacionadas en la población.

Para poder contrastar esta suposición, una vez que hemos estimado la recta de regresión y hemos obtenido las estimaciones de los parámetros del modelo; $V_t = \beta_1 + \beta_2 * GP_t + u_t$ como $\hat{V}_t = \hat{\beta}_1 + \hat{\beta}_2 * GP_t$.

Ahora lo que debemos es comprobar si esta estimación de este modelo es válida en el sentido de si es significativa de forma que la variable Publicidad (X) es relevante para explicar (Y) que son las ventas. Entonces debemos contrastar si la **pendiente de la recta de regresión poblacional** β_2 es significativamente distinta de cero, de ahí tendríamos que, en efecto, existe una correlación lineal entre ambas variables poblacionales.

Los dos contrastes siguientes son equivalentes porque si el coeficiente de correlación , r, es cero también lo será la estimación de la pendiente, $\hat{\beta}_2$ puesto que: $\hat{\beta}_2 = r * \frac{S_Y}{S_X}$

$$(1) \quad \begin{Bmatrix} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{Bmatrix} \qquad \text{y} \qquad (2) \quad \begin{Bmatrix} H_0 : \beta_2 = 0 \\ H_1 : \beta_2 \neq 0 \end{Bmatrix}$$

donde ρ es el coeficiente de correlación entre ambas variables.

El estadístico (t-Student) que se utiliza para realizar el test (2) es el siguiente:

$$t = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} \approx t\left(n-2, \alpha/2\right), \quad \text{donde} \quad S_{\hat{\beta}_1} = \sqrt{\frac{\sum Y^2 - \hat{\beta}_1 \sum Y - \hat{\beta}_2 \sum XY}{(n-2)\left[\sum X^2 - \frac{\left(\sum X\right)^2}{n}\right]}}$$

donde *t(n-2,α/2)* es el valor asociado a una t-Student con *n-2* grados de libertad que deja a su derecha un área de $\alpha/2$ (o, equivalentemente, deje a su izquierda un área de *1 - α/2*).

OJO!: si en vez de realizar el contraste bilateral (2) deseamos hacer un contraste unilateral (en el cual la hipótesis alternativa sería $H_1 : \beta_2 > 0$ ó $H_1 : \beta_2 < 0$), deberemos sustituir en la fórmula anterior $\alpha/2$ por α (ya que ahora trabajaremos con una única cola de la distribución).

Finalmente, también podemos obtener el intervalo de confianza para α_1 a nivel de confianza *(1-α)* utilizando la expresión:

$$\hat{\beta}_2 \pm t\left(n-2, \alpha/2\right) * s_{\hat{\beta}_2}$$

Siguiendo con el ejemplo anterior, el estadístico de contraste nos queda:

$$t = \frac{\hat{\beta}_2 - \beta_2}{S_{\hat{\beta}_2}} = \frac{28,\widehat{3} - 0}{9,38} = 3,02$$

Si calculamos el p-valor de t = 3,02 con tres grados de libertad, vamos a la tabla t-student y debemos calcular el área que hay por encima de t = 3,02 y el área por debajo de t= -3,02, si miramos en la tabla , el valor de t más cercano es t = 3,1824 que le corresponde un área de 0,025, por lo que a t>=3,02 le corresponderá un área menor, por lo que el p-valor será algo menor del 0,05=2*0,025.

Por lo que, si el nivel de significación es del 5%, como el p-valor es menor que 0,05, rechazaremos la hipótesis nula a un nivel de significación del 5%,. Esto indica que existen evidencias estadísticas de que la variable gastos en publicidad es una variable relevante o que influye sobre las ventas.

Es interesante notar que todo lo que hemos realizado sobre el coeficiente β_2 es también aplicable al coeficiente β_1.

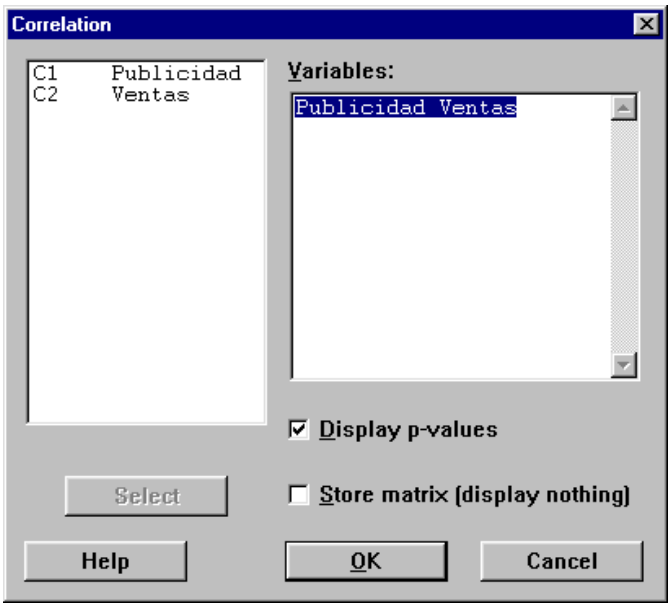
CASOS PRÁCTICOS CON SOFTWARE

1. En la siguiente tabla, se muestran los datos registrados de las ventas en millones de euros y de los gastos incurridos en publicidad, también en millones de euros, por una empresa industrial que fabrica sillas abatibles para oficina:

Gtos de publicidad (millones euros) (X)	*Volumen de ventas (millones euros) (Y)*
14,2226	95,065
13,9336	97,281
15,5040	103,159
16,3105	107,607
17,4936	113,860
19,8906	121,153
21,4803	129,102
20,4046	132,340
21,4776	138,663
22,6821	142,856
20,9722	143,120
23,3538	147,928
26,1040	155,955
29,1101	164,946
27,2418	163,921
23,0096	163,426
27,6116	172,485
32,1111	180,519
36,1788	190,509
37,5671	196,497
33,5069	196,024
36,6088	200,832
31,1554	196,769
32,7752	205,341
41,1886	220,230
39,9715	228,703
39,6866	236,500
40,2991	244,560
40,9538	254,771
41,9323	263,683
39,8393	268,304

a) Calcular el coeficiente de correlación lineal entre las variables ventas y gastos de publicidad.

Seleccionamos *Stat* > *Basic Statistics* > *Correlation* :

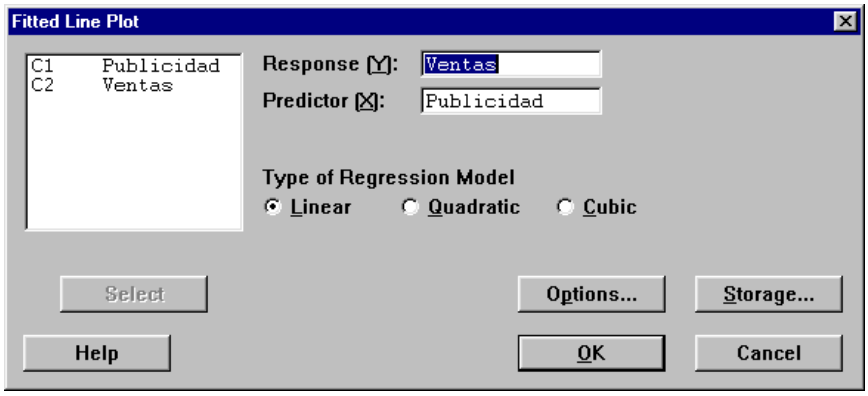


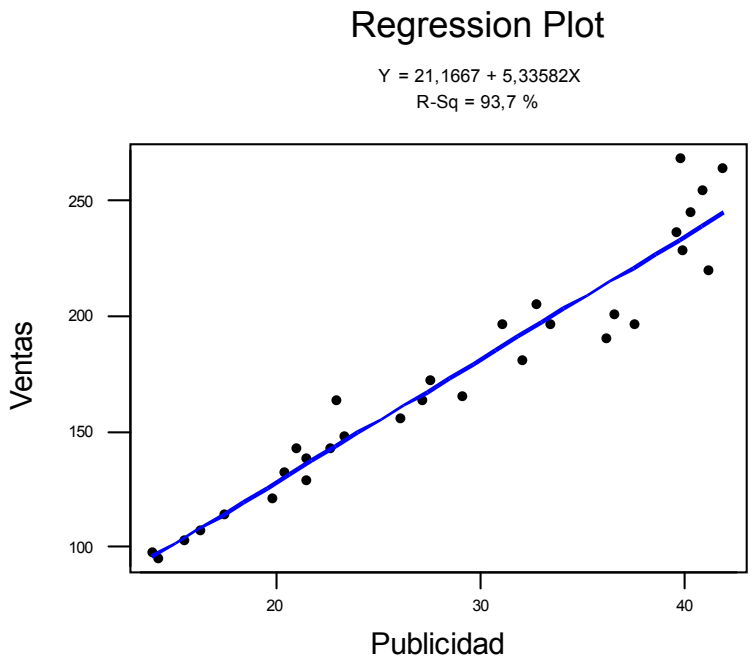

```
Correlations (Pearson)

Correlation of Publicidad y ventas = 0.973, P-Value = 0.000
```

El valor obtenido para el coeficiente de correlación es de 0,973, lo cual hace suponer que, en principio, la correlación entre ambas variables es muy alta por lo que se puede prever que en la regresión obtendremos un buen ajuste.

b) Representar la nube de puntos (gráfico x-y) ventas vs. publicidad, junto con la recta de regresión asociada. ¿Piensas que el modelo obtenido sirve para explicar las ventas obtenidas por esta empresa en los últimos treinta años en función de lo que se ha gastado en publicidad?

Seleccionamos *Stat* > *Regression* > *Fitted Line Plot* :






```
Regression

The regression equation is
y = 21,2 + 5,34 x

Predictor        Coef       StDev          T        P
Constant       21,167       7,687       2,75    0,010
x              5,3358      0,2568      20,78    0,000

S = 12,94       R-Sq = 93,7%     R-Sq(adj) = 93,5%
```

Como se aprecia en el gráfico, el modelo lineal simple ajusta con mínimos errores la evolución de las ventas en función de los gastos en publicidad. De hecho, si nos fijamos en el valor del coeficiente de determinación R-sq, veremos que este modelo explica casi el 94% del comportamiento de las ventas a través de la evolución, por lo que es un buen ajuste y por tanto, los residuos son mínimos.

c) ¿Presenta la muestra suficiente evidencia, a un nivel de significación de 0,05, como para rechazar la hipótesis nula sobre la pendiente (H_0: pendiente de la recta es cero)?

En el output anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es casi cero. Como este valor es menor que α = 0,05, debemos rechazar la hipótesis nula, i.e., concluiremos que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo tiene sentido).

2. La información estadística obtenida de una muestra de tamaño 12 sobre la relación existente entre la inversión hecha y el rendimiento obtenido en miles de euros para explotaciones agropecuarias se muestra la tabla siguiente:

Inv	11	14	16	15	16	18	20	31	14	20	19	11
Rend.	2	3	5	6	5	3	7	10	6	10	5	6

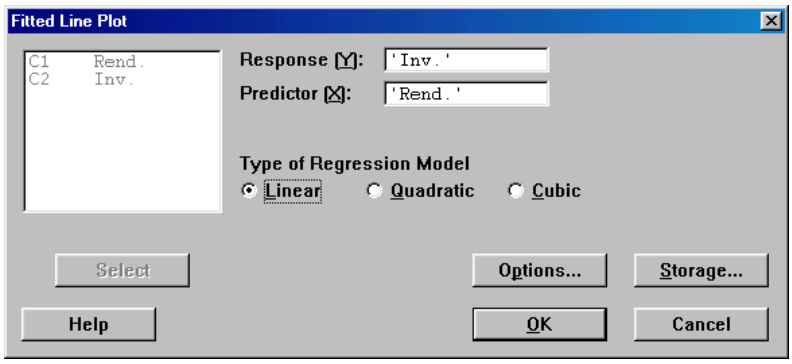
a) Calcula el coeficiente de correlación lineal, así como la recta de regresión. Calcula además, la previsión de inversión que se obtendrá con un rendimiento de 8000 €

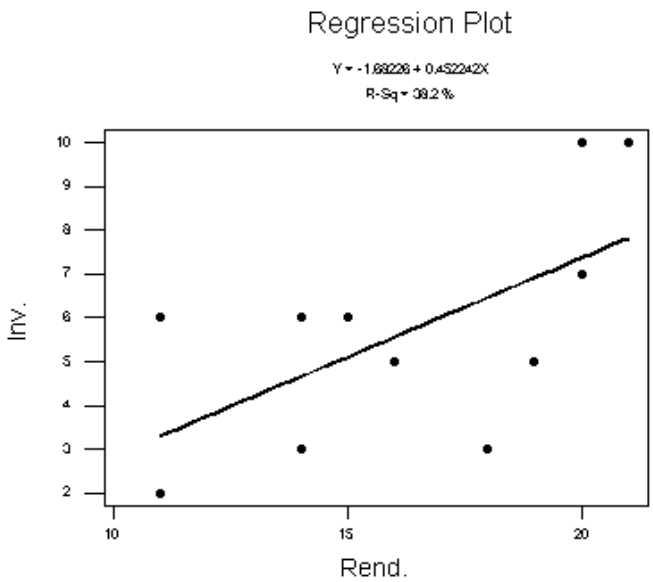
Seleccionamos *Stat* > *Basic Statistics* > *Correlation* y obtenemos:

```
Correlations (Pearson)

Correlation of Rend. and Inv. = 0.618, P-Value = 0.032
```

Como el coeficiente de correlación lineal es 0.618 no podemos deducir que exista una relación fuerte ni débil, tendríamos que realizar un contraste de hipótesis para saberlo con claridad.

Calculemos ahora la recta de regresión. Para ello, seleccionaremos *Stat* > *Regression* > *Fitted Line Plot*:





A partir de este gráfico, observamos que no existe ninguna correlación entre las dos variables.

Para conocer más detalles, seleccionamos *Stat* > *Regression* > *Regression*:

```
Regression Analysis

The regression equation is
Inv. = - 1.68 + 0.452 Rend.

Predictor        Coef       StDev          T        P
Constant       -1.682       3.015      -0.56    0.589
Rend.          0.4522      0.1819       2.49    0.032

S = 2.060       R-Sq = 38.2%     R-Sq(adj) = 32.0%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      26.230      26.230      6.18    0.032
Residual Error    10      42.437       4.244
Total             11      68.667
```

Así pues, la recta de regresión será:

Inv=-1.68 + 0.452*Rend

Por tanto, para obtener un rendimiento de 8000 €, tendríamos que hacer una inversión de...

Inv = -1.68 + 0.452*8000 = 3614.32 €

b) ¿Presenta la muestra suficiente evidencia, a un nivel de significación de 0,05, como para rechazar la hipótesis nula sobre la pendiente (H_0: pendiente de la recta es cero)?

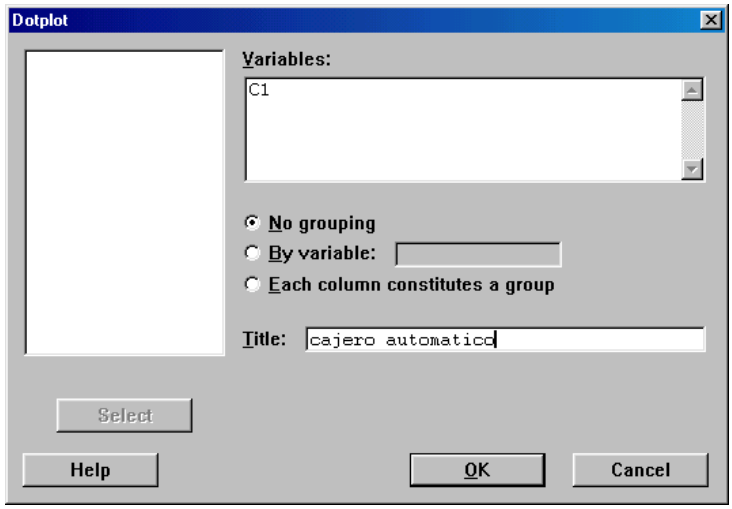
En el output anterior podemos ver que el p-valor asociado al contraste de hipótesis anterior es 0,032. Como este valor es menor que α = 0,05, debemos rechazar la hipótesis nula, i.e., concluiremos que la pendiente de la recta es distinta de cero o, lo que es lo mismo, que el coeficiente de correlación poblacional es no nulo (es decir, que ambas variables están correlacionadas y que, por tanto, el modelo tiene sentido).

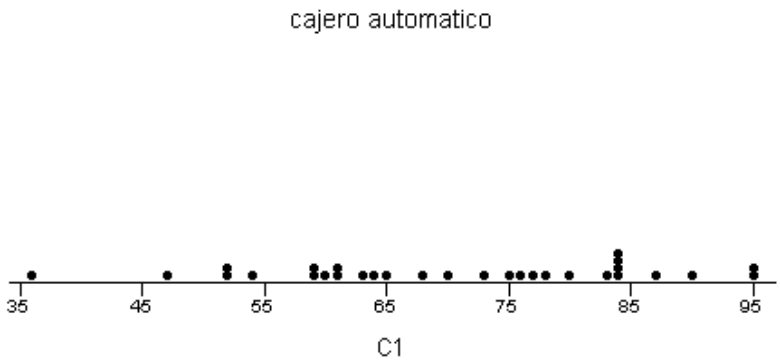
3. La entidad bancaria City Banking está estudiando el número de veces por día que se usa el cajero automático localizado en un barrio de una determina ciudad española del sur. Los siguientes datos son las veces por día que fue usado el cajero en los últimos 30 días:

83	64	84	76	84	54	75	59	70	61
63	80	84	73	68	52	65	90	52	77
95	36	78	61	59	84	95	47	87	60

a) Realiza un dotplot de los valores anteriores y comenta los resultados.

Para dibujar el dotplot, seleccionamos *Graph* > *Dotplot*:

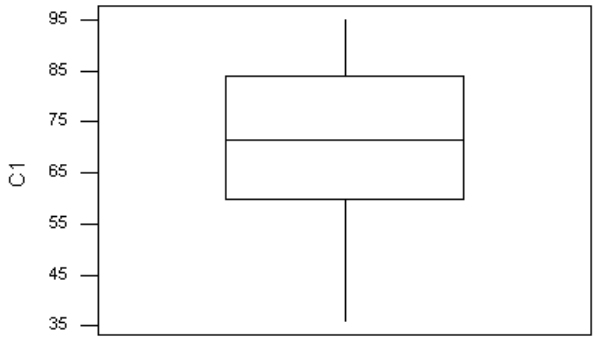




Del gráfico anterior podríamos concluir que el valor que más se repite es 84 y, además, podemos apreciar que los datos no parecen seguir una distribución normal.

b) Dibujar un diagrama de cajas (boxplot) asociado a los datos anteriores, así como también los estadísticos descriptivos correspondientes.

Para realizar el diagrama de cajas, seleccionamos *Graph* > *Boxplot*, y en el eje de las Y, insertamos cada una de las columnas:



Del anterior gráfico se desprende que el valor máximo es 95 y el mínimo 36. Así mismo, el valor de la mediana estará aproximadamente entre 70 y 75. Los cuartiles primero y tercero serán 60 y 85 aproximadamente.

Verifiquemos estos resultados anteriores calculando los estadísticos descriptivos. Seleccionamos *Stat* > *Basic Statistics* > *Display Basic Statistics*:

Descriptive Statistics

Variable	N	**Mean**	**Median**	TrMean	StDev	SE Mean
C1	30	**70.53**	**71.50**	70.88	14.82	2.71

Variable	**Minimum**	**Maximum**	**Q1**	**Q3**
C1	**36.00**	**95.00**	**59.75**	**84.00**

Por tanto, como vemos en este resultado, los valores correspondientes a la media, mediana, máximo, mínimo y cuartiles coinciden con los comentados a partir del diagrama de cajas.

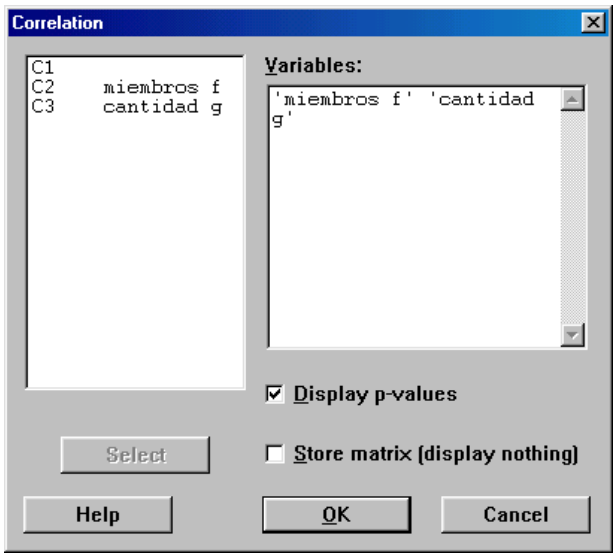
b) Además, se quiere también estudiar cuál es la relación entre la cantidad gastada semanalmente en comida (en euros) y el número de miembros de una familia. Para ello, cogemos una muestra de 10 familias del barrio obteniendo los siguientes resultados:

Miembros familia	3	6	5	6	3	4	4	5	3	6
Cantidad gastada	99	104	151	129	142	74	91	119	91	142

Determina el coeficiente de correlación entre las dos variables. Calcula y representa también la recta de regresión.

¿Qué cantidad gastada en comida cabría esperar si el número de miembros de una familia aumenta a 8?

Para calcular el coeficiente de correlación, seleccionamos *Stat* > *Basic Statistics* > *Correlation*:

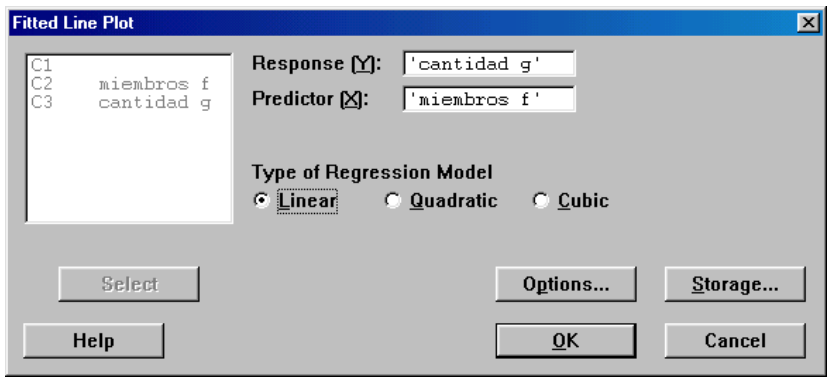


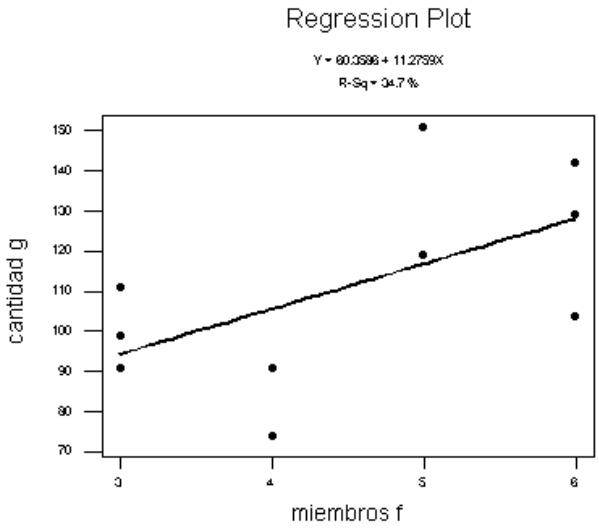
Correlations (Pearson)

Correlation of miembros f and cantidad g = 0.589, P-Value = 0.073

Como vemos, el coeficiente de correlación es de 0.589, lo cual indica que existe cierta correlación entre el número de miembros de una familia y la cantidad gastada semanalmente.

Para representar la recta de regresión, utilizamos la opción *Stat* > *Regresión* > *Fitted Line Plot* :





A partir de este gráfico observamos que sorprendentemente, parece no existir apenas correlación entre el número de miembros de una familia y la cantidad gastada en alimentos semanalmente.

```
The regression equation is
y = 60.4 + 11.3 x

Predictor        Coef       StDev          T        P
Constant        60.36       25.47       2.37    0.045
x              11.276       5.467       2.06    0.073

S = 20.82       R-Sq = 34.7%     R-Sq(adj) = 26.6%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      1843.6      1843.6      4.25    0.073
Residual Error     8      3467.3       433.4
Total              9      5310.9
```

Por tanto, la recta de regresión es:

cantidad_g = 60.4 + 11.3(miembros_f)

Así pues, la cantidad que esperamos gastar en una familia de 8 miembros será:

Cantidad_g = 60.4 + 11.3 * 8 = **150.8**

BIBLIOGRAFÍA

[1] D.A. Lind, R.D. Mason, W.G. Marchal (2001): "Estadística para Administración y Economía". Ed. Irwin McGraw-Hill.F.

[2] Kvanli, A. "Introduction to Business Statistics" South-Western

[3] R. Johnson (1996): "Elementary Statistics". Ed. Duxbury

[4] Richard I. Levin & David S. Rubin (1996): "Estadística para Administradores". Ed. Prentice Hall.

[5] E. Farber (1995): "A Guide to Minitab". Ed. McGraw-Hill.

ENLACES

- http://www.unalmed.edu.co/~estadist/regression/regresion.htm : Características y applet de Regresión lineal.
- http://kitchen.stat.vt.edu/~sundar/java/applets/ : Applets de Java de Estadística
- http://huizen.dds.nl/~berrie/ : Colección de enlaces a applets de Java de Estadística
- http://e-stadistica.bio.ucm.es/mod_regresion/regresion_applet.html : Características y applets de regresión lineal simple
- http://www.stat.wvu.edu/SRS/Modules/Applets/Regression/regression.html : Applet de Java para calcular la recta de regresión
- http://www2.egr.uh.edu/%7Eemw30693/applet.htm : Applet de Java para calcular la recta de regresión
- http://www.ruf.rice.edu/%7Elane/stat_sim/reg_by_eye/index.html : Ejemplo de recta de regresión y correlación lineal
- http://www.kuleuven.ac.be/ucs/java/version2.0/Applet010.html : Applet para calcular la recta de regresión
- http://www.kuleuven.ac.be/ucs/java/index.htm : Colección de applets para mostrar conceptos de estadística.
- http://ima.udg.es/Docencia/02-03/3105100015/Dossier_Rev.pdf : Ejercicios resueltos con Minitab de la Universitat de Girona.