3.Análisis de Regresión

Análisis de caso: Las ventas de Make-Up S.L.
La empresa Make-Up se dedica a la comercialización de productos de maquillaje por venta directa domiciliaria. Los datos recogen los resultados de 34 vendedores y vendedoras. Se pretende estudiar la relación que puede existir entre las ventas, por una parte, y la edad, la experiencia profesional (variable exp, registrada en meses de trabajo en la empresa) y el descuento (variable dto) que aplican, por otra. También analizaremos si la relación entre estas variables cambia en función del sexo o de la zona geográfica de los vendedores.

1. Aspectos básicos de regresión lineal simple: La influencia de la edad en las ventas.

Primero vamos a introducir la regresión lineal simple minimocuadrática. En el primer análisis la variable dependiente será ventas, y la independiente la edad del vendedor.

El output se presenta espaciado para poder introducir cuadros explicativos.
En este primer ejemplo de Regresión Lineal Simple los cuadros comentan los aspectos básicos del output. En los sucesivos ejemplos comentaremos aspectos técnicos más sofisticados.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.

/regress
 depend= ventas.
 indep = edad.

#/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# manes(sexo) = varon, mujer.

/end
0 0 21 10 1  430
1 0 37 18 9 2640
0 0 43 23 2 2340
1 0 54 44 6 5790
0 0 18  6 2 1000
1 1 23 38 5 2600
0 1 35 56 3 7658
1 1 47 21 9 3300
0 1 54 33 2 3700
1 1 19 17 7 1600
0 0 22 13 2  990
1 0 34 19 5 1610
0 0 23 27 0 1940
1 0 35 24 3 3790
0 0 19 12 2 1000
1 1 29 43 5 3800
0 1 46 56 6 8858
1 1 47 19 0 4940
0 1 54 53 1 4700
1 1 26 13 9 1240
0 0 28 12 8 1590
1 0 32 15 7 1640
0 0 36 33 9 2340
1 0 35 44 8 6340
0 0 28 16 9 1050
1 1 50 38 1 4810
0 1 55 16 9 4658
1 1 43 30 4 5300
0 1 22 33 9 1390
1 1 21  9 5  200
0 0 18  1 9  100
1 1 19  1 1  150
0 0 26  1 4  200
1 0 45  1 3 1700
/end
```

Aquí se define el análisis de regresión. La variable dependiente es ventas (registrada en miles) y la independiente es edad (registrada en años cumplidos). Ambas son variables cuantitativas.

Preparamos el INPUT con un párrafo /GROUP para realizar después análisis por grupos con la variable zona y, en otro análisis, con la variable sexo. Pero ahora no deseamos todavía análisis por grupos, por eso mantenemos todo el párrafo desactivado línea a línea mediante el signo # al comienzo de cada línea.

OUTPUT SELECCIONADO

```
BMDP1R - LINEAR REGRESSION BY GROUPS
```

El programa 1R se caracteriza porque puede hacer regresiones lineales por grupos y compararlas, pero también puede hacer regresión lineal simple y múltiple sin establecer grupos

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = edad.
#/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# manes(sexo) = varon, mujer.
/end
```

El output repite el archivo de input, para facilitar su identificación posterior

```
  REGRESSION INTERCEPT. . . . . .NON-ZERO
  GROUPING VARIABLE . . . . . .
  REGRESSION WEIGHT VARIABLE  . . . . . .
  PRINT COVARIANCE MATRIX . . . . . .      NO
  PRINT CORRELATION MATRIX. . . . . .      NO
  PRINT CORRELATION OF REGRESSION COEFFICIENTS. .  NO
  PRINT RESIDUALS . . . . . .      NO
  PRINT NORMAL PROBABILITY PLOT . . . . . .      NO
  PRINT DETRENDED NORMAL PROBABILITY PLOT . . . .      NO
  PRINT TRANSFORMATION PLOT . . . . . .      NO
```

A continuación ofrece siempre un diagnóstico de lo que hemos solicitado y de las condiciones en que efectuara el análisis.

```
NUMBER OF CASES READ. . . . . .      34
```

Número de casos leídos. 1R opera solo con casos completos: Es decir, sin faltantes en ninguna de las variables, o en su caso en ninguna de las especificadas en la instrucción USE si la hay

VARIABLE	MEAN	STANDARD DEVIATION	COEFFICIENT OF VARIATION	MINIMUM	MAXIMUM
1 sexo	0.5000	0.5075	1.01504	0.00000	1.00000
2 zona	0.4706	0.5066	1.07661	0.00000	1.00000
3 edad	33.6471	12.2571	0.36428	18.00000	55.00000
4 exp	23.3824	16.0133	0.68484	1.00000	56.00000
5 dto	4.8529	3.1540	0.64991	0.00000	9.00000
6 ventas	2805.7056	2239.6968	0.79827	100.00000	8858.00000

Estadísticos descriptivos de cada variable. Permiten detectar errores de introducción y obtener una visión primera de como se han comportado las variables que trataremos como dependiente e independiente. Aquí podemos ver que el rango de la independiente EDAD está entre 18 y 55 años: Entre esos límites será legítimo considerar el comportamiento de la ecuación de regresión que produzca el análisis, supuesto que esta ajuste razonablemente. Fuera de los mismos habrá que ver los pronósticos con prudencia. También es de interés observar cual es la media de la variable dependiente VENTAS (en este caso 2805'7) y considerar entre que valores venden los 34 vendedores de la empresa y con que desviación típica. Para calcular estos estadísticos el programa solo utiliza los casos completos (sin ningún faltante o fuera de rango) para todas las variables (o para todas las que se hayan especificado en una instrucción USE).

```
DEPENDENT VARIABLE. . . . . .          6 ventas
TOLERANCE . . . . . .  0.0100
```

La variable dependiente VENTAS es la número 6 de la lista de variables.

```
ALL DATA CONSIDERED AS A SINGLE GROUP
```

Todos los datos se analizan como un solo grupo.

La R múltiple expresa la correlación lineal múltiple entre la variable dependiente y las variables independientes. Esto equivale a decir la correlación de Pearson entre la variable dependiente Y y los valores predichos por la ecuación de regresión para la variable dependiente (Y').
Si se trata de regresión lineal simple, como en este caso, esta correlación equivale al coeficiente de correlación de Pearson entre la variable dependiente y la independiente.

```
MULTIPLE R              0.6936          STD. ERROR OF EST.      1638.4393
```

```
MULTIPLE R-SQUARE       0.4811
```

R múltiple al cuadrado ó Correlación múltiple al cuadrado. Es el coeficiente de determinación múltiple. Expresa la proporción de la varianza de la variable dependiente de la que dan cuenta las variables independientes tomadas conjuntamente. Si se trata de regresión lineal simple equivale al coeficiente de correlación de Pearson entre dependiente e independiente elevado al cuadrado. En este caso la edad explica el 48'11% de la varianza de las ventas. En el ejemplo, en buena medida las ventas son explicables por la edad del vendedor.

ANALYSIS OF VARIANCE

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	79632504.0000	1	79632504.0000	29.664	0.0000
RESIDUAL	85903464.0000	32	2684483.2500		

Esta parte del output contiene la ecuación de regresión propiamente dicha. En la primera columna INTERCEPT se refiere a la constante del modelo de regresión, y, a continuación, se presentan las variables independientes, en este caso solo una por tratarse de regresión simple.

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		-1458.61035					
edad	3	126.7367	23.2695	0.69	5.45	0.00	1.0000

```
La ecuación de regresión lineal simple:                    Y'  =     a       +    b      *   X
La ecuación de regresión en puntuaciones directas:    Ventas' = -1458'61035 +  126'7367 * edad.
Tomando las dos variables en diferenciales:           Ventas' =                126'7367 * edad
Tomando las dos variables en típicas:                 Ventas' =                  0'69   * edad
Obsérvese que la intercept vale 0 tanto en diferenciales como en típicas. En el caso de la
regresión lineal simple el coeficiente de regresión estandarizado (STD. REG. COEFF., es decir,
el valor del parámetro b con la ecuación escrita para puntuaciones típicas) es el coeficiente
de correlación de Pearson entre las dos variables.
En el caso de las puntuaciones directas y de las diferenciales b es también el coeficiente de
correlación de Pearson, pero multiplicado por el cociente entre la desviación típica de Y y la
de X. Estas ecuaciones permiten pronosticar las Ventas de un vendedor a partir de su Edad.
```

```
END OF INSTRUCTIONS
PROGRAM TERMINATED
```

El coeficiente a, también llamado intercept, es el punto en que la recta de regresión corta el eje de la variable dependiente Y (en este caso el eje de Ventas). Es el valor que pronosticamos para la dependiente Y cuando las independientes X valen 0. En este caso para una persona de edad 0 años pronosticaríamos unas ventas de -1458'61. En este caso ese pronostico no tiene interés, recuérdese que la ecuación opera entre 18 y 55 años, como hemos visto en la descripción de edad.

El coeficiente b, también llamado inclinación, representa la inclinación de la recta. Lo que crece la variable dependiente cuando la independiente crece una unidad (manteniendo constantes todas las demás independientes en el caso de que las hubiera). En este caso cada año más que tiene un vendedor supone una esperanza de ventas de 126'73 más que el año anterior.

2. Contrastes Estadísticos en la Regresión Lineal Simple: La relación entre experiencia y ventas.

En este segundo análisis vamos a efectuar también una regresión lineal simple tomando como la variable dependiente ventas, y como independiente la experiencia (llamada exp) del vendedor medida en meses.

En este segundo ejemplo de Regresión Lineal Simple los cuadros comentan los aspectos del output referidos al contraste estadístico de la ecuación.
De nuevo el output se presenta más espaciado de lo normal para poder introducir cuadros explicativos, pero ahora y en adelante se suprimen aspectos no esenciales o que permanecen constantes a través de estos ejemplos, como los datos.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = exp.
/end
0 0 21 10 1  430
...
datos de los restantes 32 casos, como en el ejemplo primero
...
1 0 45  1 3 1700
/end
```

La variable independiente es ahora exp (experiencia). Se dice que la variable ventas regresa sobre la variable experiencia.

OUTPUT SELECCIONADO

```
NUMBER OF CASES READ. . . . . .      34

ALL DATA CONSIDERED AS A SINGLE GROUP
```

Error Típico de Estimación (ETE) de la Variable Dependiente a partir de las Independientes. Es un indicador del error de estimación que comete la ecuación globalmente. Es igual a

$$ETE = (\Sigma(Y-Y')^2 / (N - K))^{1/2}$$

donde N es el número de casos y K el número de variables independientes. Cuanto mayor es el Error Típico de Estimación peor pronostica la ecuación los valores reales de la variable dependiente Y.

```
MULTIPLE R              0.8299          STD. ERROR OF EST.       1268.9082

MULTIPLE R-SQUARE       0.6887

ANALYSIS OF VARIANCE
```

El apartado de Análisis de Varianza asociado a la Regresión presenta una evaluación del ajuste del conjunto de la ecuación.

La Suma de Cuadrados de la Regresión (SCR) es:

$$SCR = \sum (Y' - \overline{Y})^2$$

Los grados de libertad (GL_{SCR}) de la SCR son igual al número K de coeficientes en la ecuación menos 1. En la ecuación de regresión simple hay dos coeficientes (a, b) y por tanto DF = 1

La Media Cuadrática de la Regresión (MCR) es la Suma de Cuadrados de la Regresión (SCR) dividida por sus grados de libertad:
MCR = SCR / (K - 1)

La razón F es un test estadístico igual al cociente entre MCR y MCE :
F = MCR / MCE
con K - 1 grados de libertad en el numerador y N - K grados de libertad en el

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	114011880.0000	1	114011880.0000	70.809	0.0000
RESIDUAL	51524092.0000	32	1610127.8800		

La Suma de Cuadrados del Error o Residual (SCE) es

$$SCE = \sum (Y - Y')^2$$

Los grados de libertad (GL_{SCE}) de la SCE son igual al número de casos N menos el número de coeficientes K en la ecuación . En la ecuación de regresión simple hay dos coeficientes en la ecuación (a, b) y por tanto 34-2=32

La Media Cuadrática del Error o residual (MCE) es la Suma de Cuadrados del Error (SCE) dividida por sus grados de libertad:
MCE = SCE / (N - K)

Probabilidad de F para esos grados de libertad bajo la Hipótesis Nula. Es un test de la Hipótesis Nula de NO-Relación entre la dependiente Y y todas las independientes X en la ecuación. P es la probabilidad de encontrar por azar una relación tan fuerte como esta (ó más) en estas circunstancias.
P se interpreta como el Nivel de Significación. Si este valor es menor o igual a 0'05 se dice que la ecuación presenta una relación significativa. En

Relaciones en la Tabla de Análisis de Varianza: Descomposición de la Suma de Cuadrados Total.

La suma de SCR y la SCE se denomina Suma de Cuadrados Total (SCT) y es igual a $\sum (Y - \overline{Y})^2$.

Por tanto:

SCR + SCE = SCT

Es decir:

$$\sum (Y' - \overline{Y})^2 + \sum (Y - Y')^2 = \sum (Y - \overline{Y})^2$$

Ecuación de Regresión de la Variable Ventas sobre la Variable

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		91.60449					
exp	4	116.0748	13.7941	0.83	8.41	0.00	1.0000

Error Típico del Coeficiente b:

$$ET_B = \frac{ETE}{\sqrt{\sum (X - \overline{X})^2}}$$

Prueba T de que la inclinación (coeficiente) b es distinta de 0:

$$T = \frac{b}{ET_b} = \frac{116{,}0748}{13{,}7941} = 8{,}4148$$

Probabilidad asociada a ese valor T bajo la Hipótesis Nula. Expresa el Nivel de Significación del coeficiente b. Si este valor es menor o igual a 0,05 rechazamos la hipótesis nula de que el valor del coeficiente es 0. Decimos que el valor de b es significativamente distinto de 0 para esta variable

3. Aspectos adicionales de la regresión Lineal: Análisis de la relación entre descuento y ventas.

En este tercer ejemplo de regresión lineal comentaremos algunos aspectos adicionales del análisis y mostraremos un caso en que la relación no es significativa.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = dto.
/end
...
datos como en el ejemplo
del punto 1
...
/end
```

Ecuación de Regresión Lineal Simple de la variable Ventas (dependiente) sobre la variable descuento (independiente).

Si hubieran datos faltantes o fuera de rango en cualquier otra variable (por ejemplo en edad o en zona) sería mejor definir un USE en el párrafo /VAR restringiendo únicamente a las variables dependiente e independientes mencionadas en la ecuación para evitar perder casos con datos completos.

OUTPUT SELECCIONADO

```
NUMBER OF CASES READ. . . . . .      34

  VARIABLE          MEAN    STANDARD    COEFFICIENT      MINIMUM       MAXIMUM
                          DEVIATION   OF VARIATION
-----------  ------------ ----------- -------------- ------------- -------------
  5 dto            4.8529      3.1540        0.64991       0.00000       9.00000
  6 ventas      2805.7056   2239.6968        0.79827     100.00000    8858.00000
```

El descuento medio es 4,85, y el rango de la variable va desde descuento 0 hasta descuento 9. La desviación típica es 3,15.

Verificamos que la variable dependiente es Ventas, la número 6.

```
DEPENDENT VARIABLE. . . . . .      6 ventas

TOLERANCE . . . . . .  0.0100
```

Una variable independiente NO podrá entrar a formar parte de la ecuación si su Correlación Múltiple al Cuadrado con las demás independientes en la ecuación excede el límite (1 - Tolerancia), ó bien, si al entrar en la ecuación provoca que la Correlación Múltiple al Cuadrado de otra independiente con el resto de independientes supere este límite.
La Tolerancia preasignada es 0'01 (la que muestra aquí el output), pero puede variarse a cualquier valor mayor que 0,001 y cualquier valor por debajo de 1,00. Generalmente se acepta el valor dado por defecto. Para variarlo a 0'002 por ejemplo, en el párrafo regress escribiríamos TOL=0.002.
El fenómeno de que los predictores (variables independientes) estén correlacionados o muy correlacionados entre sí se denomina colinealidad o multicolinealidad. Si la correlación es muy alta se produce una perdida de precisión en los cálculos de la ecuación debido a la inclusión de variables cuva información es redundante entre sí.

```
ALL DATA CONSIDERED AS A SINGLE GROUP

MULTIPLE R              0.0336        STD. ERROR OF EST.      2273.1401
MULTIPLE R-SQUARE       0.0011
```

La variable ventas no es en absoluto explicable por la variable descuento.: solo un 0'11%
La ecuación no es significativa

La parte residual (MCE) es muy grande en comparación con la parte de Y explicada por la Regresión (MCR), consecuentemente F es muy bajo

```
ANALYSIS OF VARIANCE
               SUM OF SQUARES      DF    MEAN SQUARE     F RATIO     P(TAIL)
  REGRESSION      186648.0000       1     186648.0000       0.036      0.8505
  RESIDUAL     165349328.0000      32    5167166.5000

                                 STD.    STD. REG
  VARIABLE     COEFFICIENT      ERROR     COEFF          T       P(2 TAIL)  TOLERANCE

INTERCEPT          2921.42383
dto         5      -23.8450    125.4626      -0.03     -0.19       0.85      1.0000
```

El error típico de b es muy grande en relación con b, consecuentemente t es bajo

La relación entre descuento y ventas no es significativa. En el caso de la regresión lineal simple los tests F y T muestran exactamente la misma cosa, de hecho en este caso $F = T^2$ y la probabilidad es necesariamente la misma, como puede verse en el

Como la Tolerancia de una variable es igual a uno menos su correlación múltiple al cuadrado con las demás variables independientes, en el caso de una sola independiente la tolerancia vale

```
END OF INSTRUCTIONS
PROGRAM TERMINATED
```

4. Regresión lineal múltiple: Explicando las ventas a partir de la edad y el descuento simultáneamente.

Un análisis de regresión múltiple no es simplemente la "suma" de los análisis simples y desde luego no puede sustituirse por estos últimos. Por ejemplo una variable puede

aparecer como un predictor claro en una relación de regresión simple y sin embargo desvanecerse su efecto al colocarla en una ecuación de regresión múltiple junto a otra con la que comparte varianza. Sin embargo, si una variable independiente no presenta una relación con la dependiente en un análisis de regresión lineal simple difícilmente mostrará una relación en otro múltiple.
Hemos introducido primero las ecuaciones de regresión simple por razones didácticas. Dada la información que hemos acumulado en el análisis del caso en esas ecuaciones al plantear ahora una ecuación de regresión lineal múltiple podríamos prescindir de la variable descuento como predictor, no obstante, vamos a dejarla para que pueda apreciarse en el análisis el comportamiento de una variable independiente que no es un predictor adecuado de la variable dependiente que se pretende explicar.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = edad, exp, dto.

#/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# manes(sexo) = varon, mujer.
/end
...
aquí los datos, como en el apartado 1
...
/end
```

Ecuación de Regresión Lineal Múltiple de la variable (dependiente) Ventas sobre las variables (independientes) edad, experiencia (llamada exp) y descuento (llamada dto).

El análisis por grupos sigue inhibido por # al principio de cada linea del párrafo /GROUP

OUTPUT SELECCIONADO

```
BMDP1R - LINEAR REGRESSION BY GROUPS
```

Diagnóstico de las instrucciones, interpretación de las instrucciones, número de casos leídos y estadísticos descriptivos como en ejemplos anteriores.

```
  DEPENDENT VARIABLE. . . . . .        6 ventas
  TOLERANCE . . . . . .  0.0100
ALL DATA CONSIDERED AS A SINGLE GROUP
```

El Error Típico de Estimación de Ventas a partir de las 3 independientes es de 1052'9691. Algo mejor que el mejor obtenido con ecuaciones de regresión lineal simple (fue de 1268'9 con experiencia como independiente)

La ecuación en su conjunto presenta una razón F significativa al nivel 0,0005. Ahora hay que calcular 4 coeficientes en la ecuación (uno por cada variable independiente más la intercept), lo que se refleja en los grados de libertad.

La Correlación Múltiple de Ventas con los 3 predictores tomados conjuntamente es 0'8939, y, elevando al cuadrado, el coeficiente de determinación múltiple 0'7991. Estos 3 predictores tomados conjuntamente dan cuenta del 79'91% de la varianza de la variable dependiente Ventas.

```
MULTIPLE R              0.8939
MULTIPLE R-SQUARE       0.7991

STD. ERROR OF EST.      1052.9691

ANALYSIS OF VARIANCE
                SUM OF SQUARES      DF     MEAN SQUARE        F RATIO     P(TAIL)
  REGRESSION   132273656.0000        3   44091220.0000         39.767      0.0000
  RESIDUAL      33262316.0000       30    1108743.8800
```

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		-1636.32617					
edad	3	69.4320	17.1237	0.38	4.05	0.00	0.7627
exp	4	90.2443	13.0971	0.65	6.89	0.00	0.7638
dto	5	-0.8812	58.1768	0.00	-0.02	0.99	0.9979

La ecuación en puntuaciones directas es: Ventas' = -1636'32617 + 69'4320 * Edad + 90'2443 * exp + (-0'8812) * dto.
La tolerancia de edad y exp ha bajado algo, acusando que estas variables tienen una correlación no nula. La tolerancia de la variable dto es prácticamente 1, mostrando que no está relacionada en absoluto con los demás predictores.
La b de la variable dto no difiere significativamente de 0 (P=0'99); su error típico es altísimo (58'1768) comparado con su coeficiente b (-0'8812) y realmente puede decirse que no aporta nada a la ecuación de predicción (b estandarizada = 0'00).
Los coeficientes de regresión b NO-estandarizados (para la ecuación en directas o en diferenciales) NO pueden compararse entre sí debido a que dependen de la escala en que estén medidas las variables y de su dispersión. Sin embargo, los coeficientes de regresión estandarizados (STD.REG.COEFF.) actúan todos sobre las variables puestas en típicas (todas con media 0 y d.t. 1) y por tanto sí son comparables entre sí, ofreciendo una imagen de cual es la aportación relativa de cada variable a la predicción.

5. Análisis de Regresión por grupos: ¿Existen diferencias significativas en el modo en que la edad y la experiencia permiten estimar las ventas según la zona?.

La posibilidad de calcular ecuaciones de regresión por separado para diferentes grupos y compararlas entre sí para establecer si la relación entre dependiente e independientes

es significativamente distinta entre grupos, es una característica muy interesante y útil del programa 1R (que no es usual en otros paquetes estadísticos).
Aquí vamos a interesarnos por las diferencias en la recta de regresión entre vendedores de la zona sur y de la zona norte (en un análisis posterior evaluaremos la ecuación en varones frente a la ecuación en mujeres). Debe observarse claramente que el test de igualdad de las líneas de regresión NO compara si los dos grupos difieren significativamente en la variable dependiente o en las independientes (p.e. NO dice si hay diferencias significativas en cuanto venden los del norte frente a los del sur). Para evaluar esta cuestión se utiliza otro test (una prueba t de contraste entre medias o un análisis de varianza -distinto del que hemos visto asociado a la regresión-). El test de igualdad de las líneas de regresión entre grupos pone a prueba si las líneas de regresión (es decir, la intercept y los coeficientes b de regresión) entre los grupos son iguales, si la dependiente mantiene la misma relación con el conjunto de las independientes en los diferentes grupos (dos o más) en que se plantee el análisis.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = edad, exp, dto.
/group
 var=zona. codes(zona)= 0, 1.
 names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# names(sexo) = varon, mujer.
/end
...
aquí los datos como en apartado 1
...
/end
```

Hemos activado (quitando # del principio de linea) las instrucciones del párrafo /GROUP que definen la variable zona como agrupadora. Mantenemos desactivadas las instrucciones de ese mismo párrafo que después utilizaremos para comparar las regresiones de varones versus mujeres. Activar y desactivar instrucciones mediante # tiene la ventaja de facilitar la reutilización de los archivos de instrucciones para diferentes análisis sin tener que reescribir las líneas de instrucciones.

OUTPUT SELECCIONADO

```
BMDP Program Output File: C:\E5\1R_7.OUT
BMDP1R - LINEAR REGRESSION BY GROUPS

/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = edad, exp, dto.
/group
 var=zona. codes(zona)= 0, 1.
 names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# names(sexo) = varon, mujer.
/end

  REGRESSION INTERCEPT. . . . . .NON-ZERO

  GROUPING VARIABLE . . . . . .zona

  REGRESSION WEIGHT VARIABLE  . . . . . .
  PRINT COVARIANCE MATRIX . . . . . NO
  PRINT CORRELATION MATRIX. . . . . NO
  PRINT CORRELATION OF REGRESSION COEFFICIENTS. . NO
  PRINT RESIDUALS . . . . . NO
  PRINT NORMAL PROBABILITY PLOT . . . . . NO
  PRINT DETRENDED NORMAL PROBABILITY PLOT . . . . NO
  PRINT TRANSFORMATION PLOT . . . . . NO

NUMBER OF CASES READ. . . . . 34
```

Los outputs siempre mencionan al principio en que archivo se guarda ese output.

En el diagnóstico se especifica que la intercept de la Regresión No se ha forzado a ser igual a 0 (esta es una opción de my raro uso en Ciencias Sociales) que implica que la escala de las variables no es arbitraria y se sabe que para el valor 0 de todas las independientes la dependiente ha de valer también 0. Para forzar la intercept a ser 0 en el párrafo /REGRESS escribiríamos

El diagnóstico de las instrucciones muestra ahora que vamos a utilizar la variable Zona como variable agrupadora.

```
  VARIABLE          MEAN    STANDARD    COEFFICIENT     MINIMUM      MAXIMUM
                            DEVIATION   OF VARIATION
----------  ------------ ----------- -------------- ------------ ------------
  1 sexo          0.5000      0.5075         1.01504      0.00000      1.00000
  3 edad         33.6471     12.2571         0.36428     18.00000     55.00000
  4 exp          23.3824     16.0133         0.68484      1.00000     56.00000
  5 dto           4.8529      3.1540         0.64991      0.00000      9.00000
  6 ventas     2805.7056   2239.6968         0.79827    100.00000   8858.00000
```

Presenta los estadísticos descriptivos para los 34 casos tomados conjuntamente, como en ejemplos anteriores.

Primero mostrará los resultados del análisis de regresión para todos los datos considerados como un solo grupo (igual que en ejemplos anteriores).

```
  DEPENDENT VARIABLE. . . . . .          6 ventas
  TOLERANCE . . . . . .  0.0100
ALL DATA CONSIDERED AS A SINGLE GROUP

MULTIPLE R              0.8939          STD. ERROR OF EST.      1052.9691
MULTIPLE R-SQUARE       0.7991

ANALYSIS OF VARIANCE
               SUM OF SQUARES      DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION   132273656.0000       3   44091220.0000        39.767     0.0000
  RESIDUAL      33262316.0000      30    1108743.8800

                                 STD.   STD. REG
  VARIABLE     COEFFICIENT      ERROR     COEFF      T     P(2 TAIL) TOLERANCE

INTERCEPT          -1636.32617
edad        3         69.4320    17.1237        0.38     4.05      0.00      0.7627
exp         4         90.2443    13.0971        0.65     6.89      0.00      0.7638
dto         5         -0.8812    58.1768        0.00    -0.02      0.99      0.9979

NUMBER OF INTEGER WORDS USED IN PRECEDING    PROBLEM     764

BMDP1R - LINEAR REGRESSION BY GROUPS

REGRESSION FOR GROUP  1  norte

NUMBER OF CASES READ. . . . . .       34
   CASES WITH GROUPING VALUES NOT USED. . . .       16
      REMAINING NUMBER OF CASES . . . . . .       18
```

Ahora reanalizará para cada grupo. Aquí comienzan los resultados para el GRUPO NORTE de la variable ZONA. En ese grupo hay 18 de los 34 casos.

```
  VARIABLE          MEAN    STANDARD    COEFFICIENT     MINIMUM      MAXIMUM
                            DEVIATION   OF VARIATION
----------  ------------ ----------- -------------- ------------ ------------
  1 sexo          0.3889      0.5016         1.28991      0.00000      1.00000
  3 edad         30.7778     10.1202         0.32881     18.00000     54.00000
  4 exp          17.7222     13.0691         0.73744      1.00000     44.00000
  5 dto           4.9444      3.1895         0.64508      0.00000      9.00000
  6 ventas     2027.2222   1728.8582         0.85282    100.00000   6340.00000
  REGRESSION TITLE IS

  DEPENDENT VARIABLE. . . . . .          6 ventas
  TOLERANCE . . . . . .  0.0100

MULTIPLE R              0.9110          STD. ERROR OF EST.       785.5007
MULTIPLE R-SQUARE       0.8300

ANALYSIS OF VARIANCE
               SUM OF SQUARES      DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION   42174000.0000       3   14058000.0000        22.784     0.0000
  RESIDUAL      8638159.0000      14     617011.3750

                                 STD.   STD. REG
  VARIABLE     COEFFICIENT      ERROR     COEFF      T     P(2 TAIL) TOLERANCE

INTERCEPT          -1206.21924
edad        3         49.4172    22.7264        0.29     2.17      0.05      0.6861
exp         4         94.8190    17.5164        0.72     5.41      0.00      0.6926
dto         5          6.4893    61.5821        0.01     0.11      0.92      0.9408

NUMBER OF INTEGER WORDS USED IN PRECEDING    PROBLEM     764
```

Los resultados parecen muy semejantes a los ya comentados para el grupo total.

```
BMDP1R - LINEAR REGRESSION BY GROUPS

REGRESSION FOR GROUP  2  sur

NUMBER OF CASES READ. . . . . .      34
   CASES WITH GROUPING VALUES NOT USED. . . .      18
      REMAINING NUMBER OF CASES . . . . . .      16
```

Aquí comienzan los análisis para el GRUPO 2, vendedores de la ZONA SUR.
Hay 16 casos de los 34 en este grupo.

```
  VARIABLE          MEAN    STANDARD   COEFFICIENT     MINIMUM       MAXIMUM
                            DEVIATION  OF VARIATION
----------  ----------- ----------- -------------- ------------ ------------
  1 sexo          0.6250      0.5000        0.80000      0.00000      1.00000
  3 edad         36.8750     13.9086        0.37718     19.00000     55.00000
  4 exp          29.7500     16.9961        0.57130      1.00000     56.00000
  5 dto           4.7500      3.2146        0.67675      0.00000      9.00000
  6 ventas     3681.4998   2470.3967        0.67103    150.00000   8858.00000
  REGRESSION TITLE IS

  DEPENDENT VARIABLE. . . . . .       6 ventas
  TOLERANCE . . . . . .  0.0100

MULTIPLE R             0.8608          STD. ERROR OF EST.      1405.6779
MULTIPLE R-SQUARE      0.7410

ANALYSIS OF VARIANCE
              SUM OF SQUARES     DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION    67831744.0000      3   22610582.0000        11.443     0.0008
  RESIDUAL      23711164.0000     12    1975930.3800

                               STD.    STD. REG
  VARIABLE     COEFFICIENT      ERROR     COEFF      T     P(2 TAIL)  TOLERANCE

INTERCEPT         -1859.28198
edad        3        78.7953    28.5988        0.44      2.76      0.02      0.8326
exp         4        86.8559    23.0961        0.60      3.76      0.00      0.8549
dto         5        10.7878   117.5623        0.01      0.09      0.93      0.9224
```

Aparentemente los resultados también son semejantes a los del grupo total.

A continuación se presenta el test de igualdad de la líneas de regresión. Este test se efectúa mediante una razón F basada en el principio de que si las intercepts **a** y los coeficientes **b** de regresión son iguales (es decir, si las ecuaciones de regresión son iguales) la Suma de Cuadrados Total de los Errores o residuales sobre los grupos, será igual a la Suma de Cuadrados de los Errores o residuales para el grupo total (sin efectuar grupos, el análisis antes

```
ANALYSIS OF VARIANCE OF REGRESSION COEFFICIENTS OVER GROUPS
           REDUCTION OF RESIDUALS DUE TO GROUPING

                        SUM OF SQUARES  DF  MEAN SQUARE  F RATIO    P(TAIL)
                        ---------------------------------------------------
 REGRESSION OVER GROUPS    912992.000   4   228248.000     0.183    0.94491
 RESIDUAL WITHIN GROUPS 32349324.000   26  1244204.750

 A SIGNIFICANT F RATIO INDICATES THAT THE SLOPES OR INTERCEPTS DIFFER
 BEYOND CHANCE BETWEEN THE GROUPS.


PROGRAM TERMINATED
```

Como dice el mensaje incluido en el output, una razón F significativa indica que las inclinaciones (coeficientes b) o las intercepts (ó ambas) difieren entre grupos más allá de lo esperable por azar en estas circunstancias de tamaño de muestra, número de grupos y número de variables independientes. El valor P final permite tomar una decisión estadística de rechazar o no rechazar la Hipótesis Nula de que las ecuaciones de regresión no difieren entre sí. Si ese valor es menor o igual que el punto de corte convencional 0,05 decimos que hay diferencias significativas en la regresión debidas al agrupamiento. En ese caso, por tanto, convendría utilizar las ecuaciones de regresión separadas para cada grupo, pues las ecuaciones difieren entre si significativamente. Si, como en este caso, la probabilidad es mayor que 0,05 y por tanto, no hay diferencias significativas, puede usarse la ecuación de regresión obtenida para todos los datos considerados como un solo grupo.

6. Análisis de regresión por grupos: Comparando los coeficientes de las ecuaciones de regresión para hombres y mujeres.

El análisis siguiente utiliza sexo como variable agrupadora.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
/regress
 depend= ventas.
 indep = edad, exp, dto.
/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
 var=sexo. codes(sexo)= 0, 1.
 names(sexo) = varon, mujer.
/end
...
aquí los mismos datos que en los análisis anteriores
...
/end
```

La misma ecuación de regresión, pero ahora se inhiben las líneas del párrafo group dedicadas a la variable zona y se liberan las líneas relativas al agrupamiento según la variable sexo.

OUTPUT SELECCIONADO

```
BMDP1R - LINEAR REGRESSION BY GROUPS

NUMBER OF CASES READ. . . . . .      34

  DEPENDENT VARIABLE. . . . . .       6 ventas
  TOLERANCE . . . . . .  0.0100
ALL DATA CONSIDERED AS A SINGLE GROUP

MULTIPLE R             0.8939          STD. ERROR OF EST.      1052.9691
MULTIPLE R-SQUARE      0.7991

ANALYSIS OF VARIANCE
              SUM OF SQUARES     DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION  132273656.0000      3  44091220.0000        39.767     0.0000
  RESIDUAL     33262316.0000     30   1108743.8800

                               STD.   STD. REG
  VARIABLE    COEFFICIENT      ERROR     COEFF      T     P(2 TAIL) TOLERANCE

INTERCEPT         -1636.32617
edad       3       69.4320    17.1237       0.38     4.05      0.00      0.7627
exp        4       90.2443    13.0971       0.65     6.89      0.00      0.7638
dto        5       -0.8812    58.1768       0.00    -0.02      0.99      0.9979

REGRESSION FOR GROUP  1  varon

NUMBER OF CASES READ. . . . . .      34
   CASES WITH GROUPING VALUES NOT USED. . . .      17
      REMAINING NUMBER OF CASES . . . . . .      17


  VARIABLE          MEAN   STANDARD   COEFFICIENT     MINIMUM      MAXIMUM
                           DEVIATION  OF VARIATION
-----------  ----------- ----------- -------------- ------------ ------------
  2 zona          0.3529      0.4926        1.39568      0.00000      1.00000
  3 edad         32.2353     13.3909        0.41541     18.00000     55.00000
  4 exp          23.5882     18.1213        0.76824      1.00000     56.00000
  5 dto           4.5882      3.4832        0.75915      0.00000      9.00000
  6 ventas     2584.9409   2555.4590        0.98859    100.00000   8858.00000
  REGRESSION TITLE IS

  DEPENDENT VARIABLE. . . . . .       6 ventas
  TOLERANCE . . . . . .  0.0100

MULTIPLE R             0.8863          STD. ERROR OF EST.      1312.9266
MULTIPLE R-SQUARE      0.7855

ANALYSIS OF VARIANCE
              SUM OF SQUARES     DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION   82076832.0000      3  27358944.0000        15.872     0.0001
  RESIDUAL     22409094.0000     13   1723776.5000


                               STD.   STD. REG
  VARIABLE    COEFFICIENT      ERROR     COEFF      T     P(2 TAIL) TOLERANCE
```

El análisis para todos los datos considerados como un solo grupo será siempre el mismo (salvo que hubieran ligeras variaciones debidas a la exclusión de casos por valores faltantes si se variara el USE)

Análisis para los 17 casos varones incluidos en los datos.

```
INTERCEPT          -1576.14502
edad        3        51.6421     30.7096        0.27      1.68      0.12      0.6371
exp         4        98.7669     22.8204        0.70      4.33      0.00      0.6300
dto         5        36.3216     95.2653        0.05      0.38      0.71      0.9785
```

Una ligera variación, para los varones la variable edad entra en la ecuación sin alcanzar significación estadística.

```
REGRESSION FOR GROUP  2  mujer

NUMBER OF CASES READ. . . . . .          34
   CASES WITH GROUPING VALUES NOT USED. . . .          17
      REMAINING NUMBER OF CASES . . . . . .          17
```

Análisis para las 17 mujeres.

```
  VARIABLE           MEAN    STANDARD   COEFFICIENT      MINIMUM       MAXIMUM
                             DEVIATION  OF VARIATION
-----------  ------------ ----------- -------------- ------------ ------------
  2 zona           0.5882      0.5073        0.86241      0.00000      1.00000
  3 edad          35.0588     11.2387        0.32057     19.00000     54.00000
  4 exp           23.1765     14.1564        0.61081      1.00000     44.00000
  5 dto            5.1176      2.8697        0.56075      0.00000      9.00000
  6 ventas      3026.4707   1926.6711        0.63661    150.00000   6340.00000
  REGRESSION TITLE IS

  DEPENDENT VARIABLE. . . . . .          6 ventas
  TOLERANCE . . . . . .  0.0100

MULTIPLE R             0.9230          STD. ERROR OF EST.       822.5944
MULTIPLE R-SQUARE      0.8519

ANALYSIS OF VARIANCE
               SUM OF SQUARES     DF    MEAN SQUARE       F RATIO    P(TAIL)
  REGRESSION   50596388.0000      3   16865462.0000        24.925     0.0000
  RESIDUAL      8796599.0000     13     676661.4380

                              STD.   STD. REG
  VARIABLE    COEFFICIENT     ERROR    COEFF      T     P(2 TAIL) TOLERANCE

INTERCEPT          -1582.43652
edad        3        79.8168     19.8249        0.47      4.03      0.00      0.8519
exp         4        89.0494     15.6260        0.65      5.70      0.00      0.8643
dto         5       -49.4809     73.7303       -0.07     -0.67      0.51      0.9447

 ANALYSIS OF VARIANCE OF REGRESSION COEFFICIENTS OVER GROUPS
           REDUCTION OF RESIDUALS DUE TO GROUPING

                      SUM OF SQUARES  DF  MEAN SQUARE  F RATIO    P(TAIL)
                      ---------------------------------------------------
 REGRESSION OVER GROUPS  2056624.000   4   514156.000    0.428    0.78676
 RESIDUAL WITHIN GROUPS 31205692.000  26  1200218.880

 A SIGNIFICANT F RATIO INDICATES THAT THE SLOPES OR INTERCEPTS DIFFER
 BEYOND CHANCE BETWEEN THE GROUPS.
```

Las ecuaciones de regresión no difieren significativamente, por tanto, podemos aceptar la ecuación general sin agrupamientos como una solución general independientemente del sexo de los casos.

```
PROGRAM TERMINATED
```

7. Valores estimados, residuales, correlaciones, covarianzas y plots.

Una ecuación de regresión permite analizar la relación entre una variable dependiente y otras tomadas como independientes, y contrastar si estas contribuyen o no al pronóstico de aquella, como hemos visto. Además, una ecuación de regresión permite estimar o pronosticar valores en la variable dependiente (en nuestro ejemplo en ventas) conocidos los valores en las variables independientes. En ese ejemplo veremos como podemos acceder a esos valores estimados para cada caso, y obtener sus residuales (sus diferencias con los valores reales conocidos de la variable dependiente). Además veremos otras posibilidades del programa 1R obteniendo correlaciones, covarianzas y algunos gráficos.

INSTRUCCIONES. PROGRAMA 1R.

```
/input var=6. format=free.
/var
 names= sexo, zona, edad, exp, dto, ventas.
 use = edad, exp, ventas.
/regress
 depend= ventas.
 indep = edad, exp.
#/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# names(sexo) = varon, mujer.
/print
 data.
 correlations.
 covariances.
 rreg.
/plot
 residuals.
 norm.
 dnorm.
 variable=edad.
/end
0 0 21 10 1  430
1 0 37 18 9 2640
0 0 43 23 2 2340
1 0 54 44 6 5790
0 0 18  6 2 1000
1 1 23 38 5 2600
0 1 35 56 3 7658
1 1 47 21 9 3300
0 1 54 33 2 3700
1 1 19 17 7 1600
0 0 22 13 2  990
1 0 34 19 5 1610
0 0 23 27 0 1940
1 0 35 24 3 3790
0 0 19 12 2 1000
1 1 29 43 5 3800
0 1 46 56 6 8858
1 1 47 19 0 4940
0 1 54 53 1 4700
1 1 26 13 9 1240
0 0 28 12 8 1590
1 0 32 15 7 1640
0 0 36 33 9 2340
1 0 35 44 8 6340
0 0 28 16 9 1050
1 1 50 38 1 4810
0 1 55 16 9 4658
1 1 43 30 4 5300
0 1 22 33 9 1390
1 1 21  9 5  200
0 0 18  1 9  100
1 1 19  1 1  150
0 0 26  1 4  200
1 0 45  1 3 1700
/end
```

Se introduce un USE restringiendo la atención del programa a la variable dependiente y a las independientes que se van a utilizar.
Se ha eliminado la variable dto como independiente después de llegar a la conclusión que no afecta a las ventas en el grupo total (ni el los grupos debidos a zona o a sexo).
Se inhiben todas las instrucciones relativas a agrupamientos.

La instrucción DATA del párrafo /PRINT solicita que se presente un listado de los datos con los valores estimados y con los residuales para cada caso.
La instrucción CORRELATIONS solicita que se presente la matriz de correlaciones.
La instrucción COVARIANCES provocará la presentación de la matriz de covarianzas.

El párrafo /PLOT solicita algunos gráficos relacionados básicamente con los residuales. Estos gráficos ayudan a diagnosticar el comportamiento de la regresión.
RESIDUALS solicita plots de los residuales y de los residuales al cuadrado frente a los valores predichos, lo que permite apreciar si hay alguna tendencia entre estas variables.
NORM solicita un plot de probabilidad normal de los residuales.
DNORM solicita un plot de probabilidad de los residuales eliminada la tendencia.
VARIABLE=EDAD solicita dos plots más. El primero es un plot de los valores observados y estimados frente a la variable EDAD. El segundo, de los residuales frente a EDAD.

Datos:
Los mismos que hemos venido analizando en todos los análisis del caso.

OUTPUT SELECCIONADO

```
/input var=6. format=free.
/var names= sexo, zona, edad, exp, dto, ventas.
 use = edad, exp, ventas.
/regress
 depend= ventas.
 indep = edad, exp.
#/group
# var=zona. codes(zona)= 0, 1.
# names(zona) = norte, sur.
# var=sexo. codes(sexo)= 0, 1.
# names(sexo) = varon, mujer.
/print
 data.
 correlations.
 covariances.
 rreg.
/plot
 residuals.
 norm.
 dnorm.
 variable=edad.
/end
```

Instrucciones y Diagnósticos del input.

```
  REGRESSION INTERCEPT. . . . . .NON-ZERO
  GROUPING VARIABLE . . . . . .
  REGRESSION WEIGHT VARIABLE  . . . . . .
  PRINT COVARIANCE MATRIX . . . . . YES
  PRINT CORRELATION MATRIX. . . . . YES
  PRINT CORRELATION OF REGRESSION COEFFICIENTS. . YES
  PRINT RESIDUALS . . . . . YES
  PRINT NORMAL PROBABILITY PLOT . . . . . YES
  PRINT DETRENDED NORMAL PROBABILITY PLOT . . . . YES
  PRINT TRANSFORMATION PLOT . . . . . NO

NUMBER OF CASES READ. . . . . 34
```

Estadísticos descriptivos (solo considera las variables incluidas en el USE).

```
  VARIABLE          MEAN    STANDARD    COEFFICIENT     MINIMUM      MAXIMUM
                            DEVIATION   OF VARIATION
-----------  ----------- ----------- -------------- ------------ ------------
  3 edad         33.6471     12.2571        0.36428     18.00000     55.00000
  4 exp          23.3824     16.0133        0.68484      1.00000     56.00000
  6 ventas     2805.7056   2239.6968        0.79827    100.00000   8858.00000
```

```
 COVARIANCE MATRIX

                  edad         exp          ventas
                     3           4               6

edad        3     150.2353
exp         4      95.3815    256.4251
ventas      6   19040.3184  29764.4824  5.01624E+6
```

Matriz de Varianzas - Covarianzas.
En la diagonal se encuentran las varianzas de las variables.
Fuera de la diagonal las covarianzas entre las variables que encabezan la fila y columna respectivas.
El valor de la varianza de Ventas es tan grande que se presenta una aproximación en notación científica:

```
 CORRELATION MATRIX

                  edad         exp          ventas
                     3           4               6

edad        3      1.0000
exp         4      0.4860      1.0000
ventas      6      0.6936      0.8299      1.0000
```

Matriz de Correlaciones de Pearson.
En la diagonal de esta matriz siempre hay unos porque toda variable correlaciona 1 consigo misma. Solo se muestra una parte de la matriz porque la otra parte sería simétrica (la correlación de Y con X es la misma que la de X con Y).
Podemos apreciar que Ventas correlaciona notoriamente con Experiencia y con Edad.

```
 DEPENDENT VARIABLE. . . . . .  6 ventas
  TOLERANCE . . . . . .  0.0100
```

ALL DATA CONSIDERED AS A SINGLE GROUP

Utilizando como independientes solo edad y exp se obtiene un modelo predictivo más simple (que utilizando también dto,) sin apenas perdida de poder explicativo. El modelo explica ahora el 79'91% de la varianza de Ventas con un ETE de 1035'8505.

```
MULTIPLE R              0.8939          STD. ERROR OF EST.      1035.8505
MULTIPLE R-SQUARE       0.7991
```

ANALYSIS OF VARIANCE

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO	P(TAIL)
REGRESSION	132273400.0000	2	66136700.0000	61.638	0.0000
RESIDUAL	33262570.0000	31	1072986.1300		

VARIABLE		COEFFICIENT	STD. ERROR	STD. REG COEFF	T	P(2 TAIL)	TOLERANCE
INTERCEPT		-1640.94824					
edad	3	69.4421	16.8326	0.38	4.13	0.00	0.7638
exp	4	90.2447	12.8842	0.65	7.00	0.00	0.7638

Ambas variables independientes entran en la ecuación como significativas y (consecuentemente) la ecuación también aparece como significativa.

CORRELATION MATRIX OF REGRESSION COEFFICIENTS

		edad 3	exp 4
edad	3	1.0000	
exp	4	-0.4860	1.0000

Matriz de correlaciones de los coeficientes de regresión.

Listado de Residuales, Valores Pronosticados (o Estimados) y datos en las variables

```
LIST OF PREDICTED VALUES, RESIDUALS, AND VARIABLES
```

```
*** N O T E *** A NEGATIVE CASE NUMBER DENOTES A CASE WITH MISSING VALUES.  THE
                NUMBER OF STANDARD DEVIATIONS FROM THE MEAN IS DENOTED BY
                ASTERISKS (UP TO 3) TO THE RIGHT OF EACH RESIDUAL OR VARIABLE.
                MISSING VALUES AND VALUES OUT OF RANGE ARE DENOTED BY VALUES
                GREATER THAN OR EQUAL TO  0.3245E+33 IN ABSOLUTE VALUE.
```

CASE LABEL	NO.	RESIDUAL	PREDICTED VALUE	VARIABLES 3 edad		4 exp		6 ventas	
	1	-289.8	719.8	21.0	*	10.0		430.	*
	2	87.19	2553.	37.0		18.0		0.264E+04	
	3	-1081.	* 3421.	43.0		23.0		0.234E+04	
	4	-289.7	6080.	54.0	*	44.0	*	0.579E+04	*
	5	849.5	150.5	18.0	*	6.00	*	0.100E+04	
	6	-785.5	3386.	23.0		38.0		0.260E+04	
	7	1815.	* 5843.	35.0		56.0	**	0.766E+04	**
	8	-218.0	3518.	47.0	*	21.0		0.330E+04	
	9	-1387.	* 5087.	54.0	*	33.0		0.370E+04	
	10	387.4	1213.	19.0	*	17.0		0.160E+04	
	11	-69.96	1060.	22.0		13.0		990.	
	12	-824.7	2435.	34.0		19.0		0.161E+04	
	13	-452.8	2393.	23.0		27.0		0.194E+04	
	14	834.6	2955.	35.0		24.0		0.379E+04	
	15	238.6	761.4	19.0	*	12.0		0.100E+04	
	16	-453.4	4253.	29.0		43.0	*	0.380E+04	
	17	2251.	** 6607.	46.0	*	56.0	**	0.886E+04	**
	18	1603.	* 3337.	47.0	*	19.0		0.494E+04	
	19	-2192.	** 6892.	54.0	*	53.0	*	0.470E+04	
	20	-97.73	1338.	26.0		13.0		0.124E+04	
	21	203.6	1386.	28.0		12.0		0.159E+04	
	22	-294.9	1935.	32.0		15.0		0.164E+04	
	23	-1497.	* 3837.	36.0		33.0		0.234E+04	
	24	1580.	* 4760.	35.0		44.0	*	0.634E+04	*
	25	-697.3	1747.	28.0		16.0		0.105E+04	
	26	-450.5	5260.	50.0	*	38.0		0.481E+04	
	27	1036.	3622.	55.0	*	16.0		0.466E+04	
	28	1248.	* 4052.	43.0		30.0		0.530E+04	*
	29	-1475.	* 2865.	22.0		33.0		0.139E+04	
	30	-429.5	629.5	21.0	*	9.00		200.	*
	31	400.7	-300.7	18.0	*	1.00	*	100.	*
	32	381.3	-231.3	19.0	*	1.00	*	150.	*
	33	-54.79	254.8	26.0		1.00	*	200.	*
	34	125.8	1574.	45.0		1.00	*	0.170E+04	

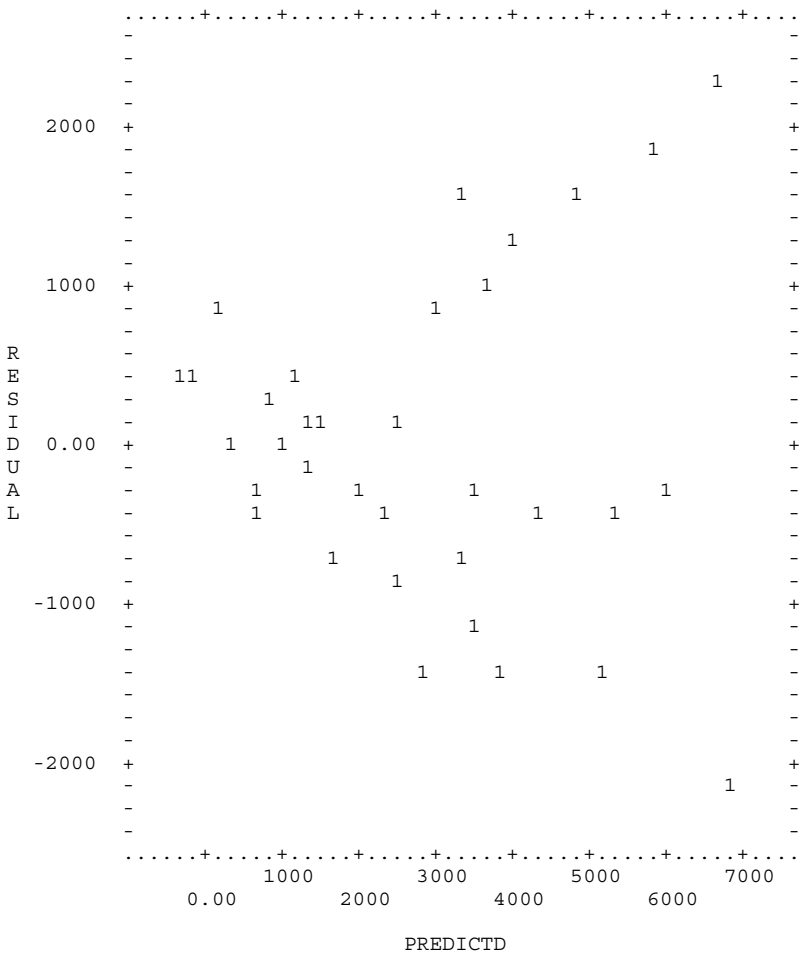
Correlación serial de los residuales.

```
SERIAL CORRELATION OF RESIDUALS = -0.1984
```

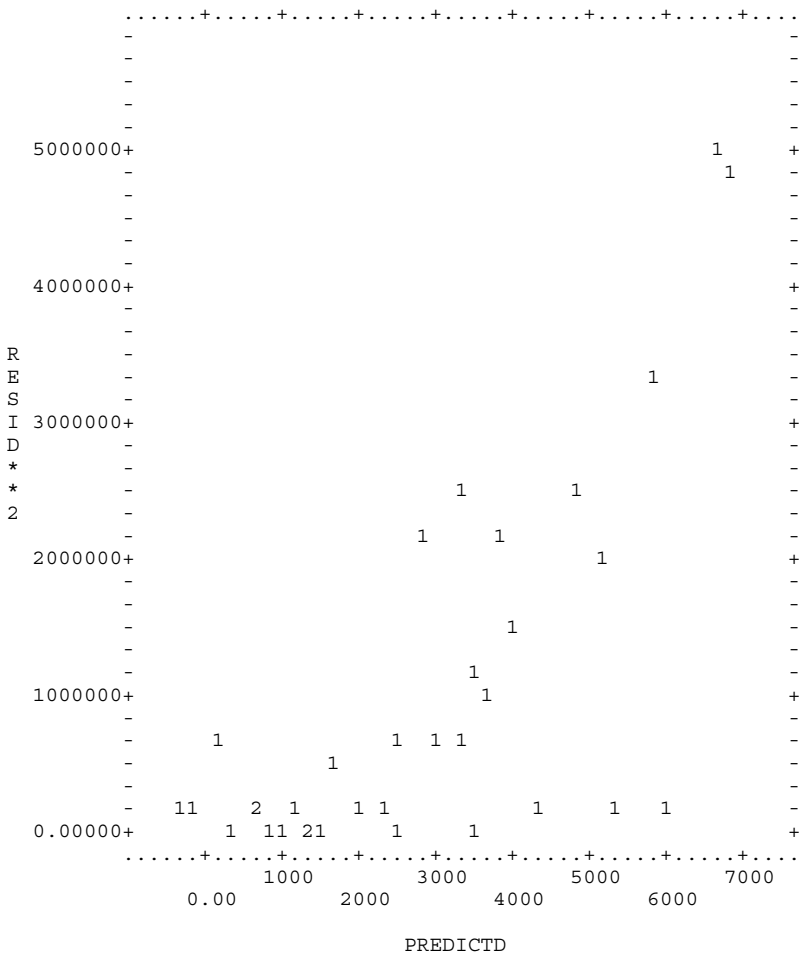
```
DURBIN-WATSON STATISTICS =  2.3932 BASED ON    34 RESIDUALS
```

Estadístico de Durbin-Watson.

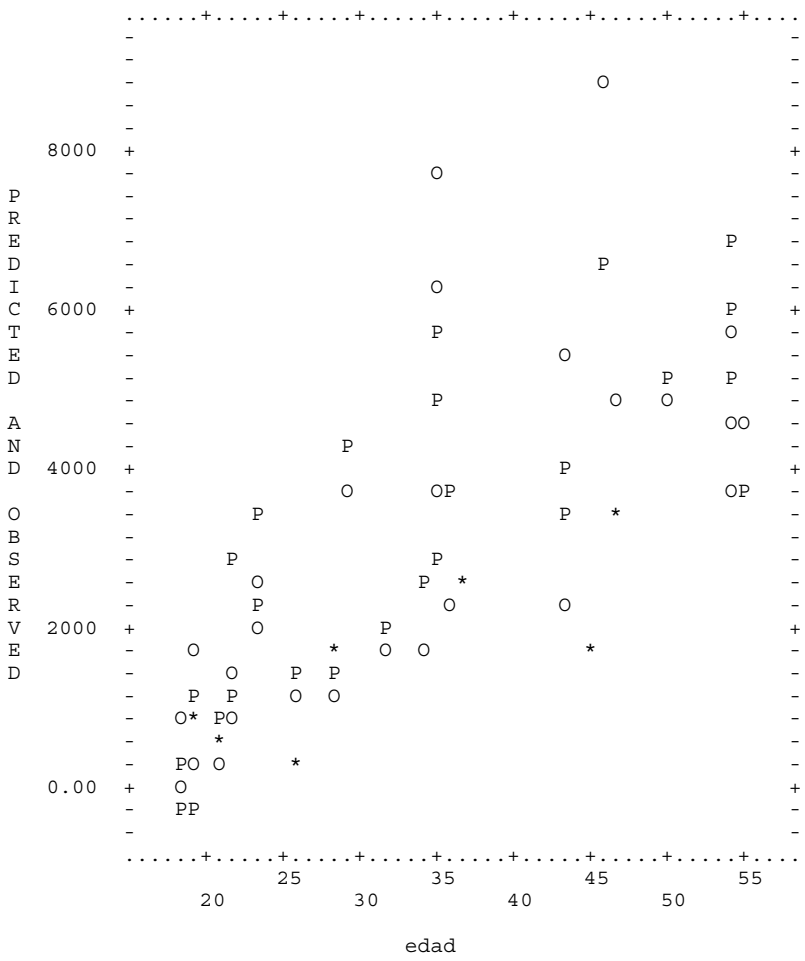
Plot de Valores Estimados (Abcisas) versus Residuales (ordenadas).



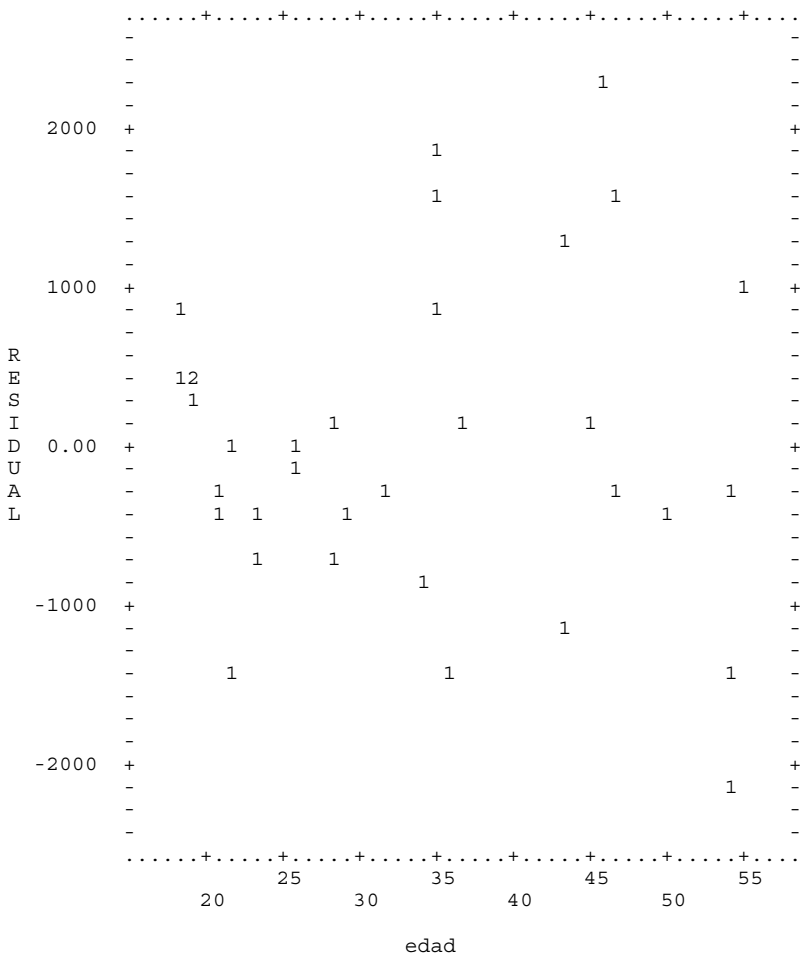
Plot de Valores Estimados (Abcisas) versus Residuales elevados al cuadrado (ordenadas).



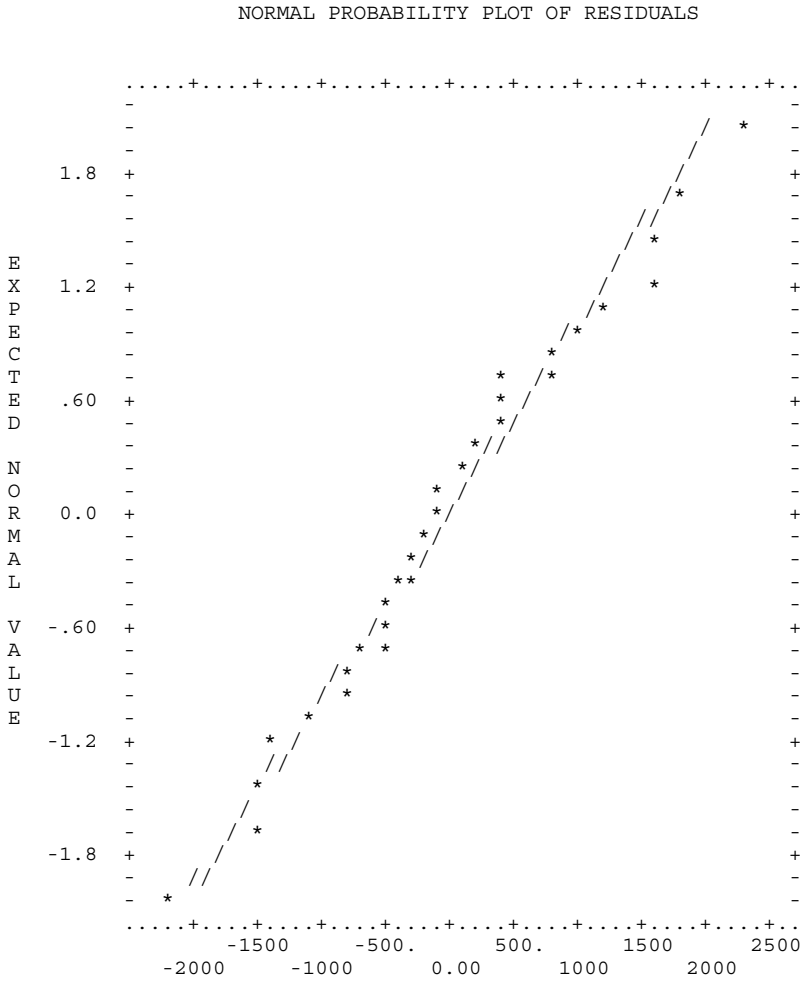
Plot de Valores de la Variable Edad (Abcisas) versus Valores Observados (O) y Estimados (P)de la variable dependiente Ventas (en ordenadas).



Plot de Valores de la variable independiente Edad (Abcisas) versus Residuales (ordenadas).

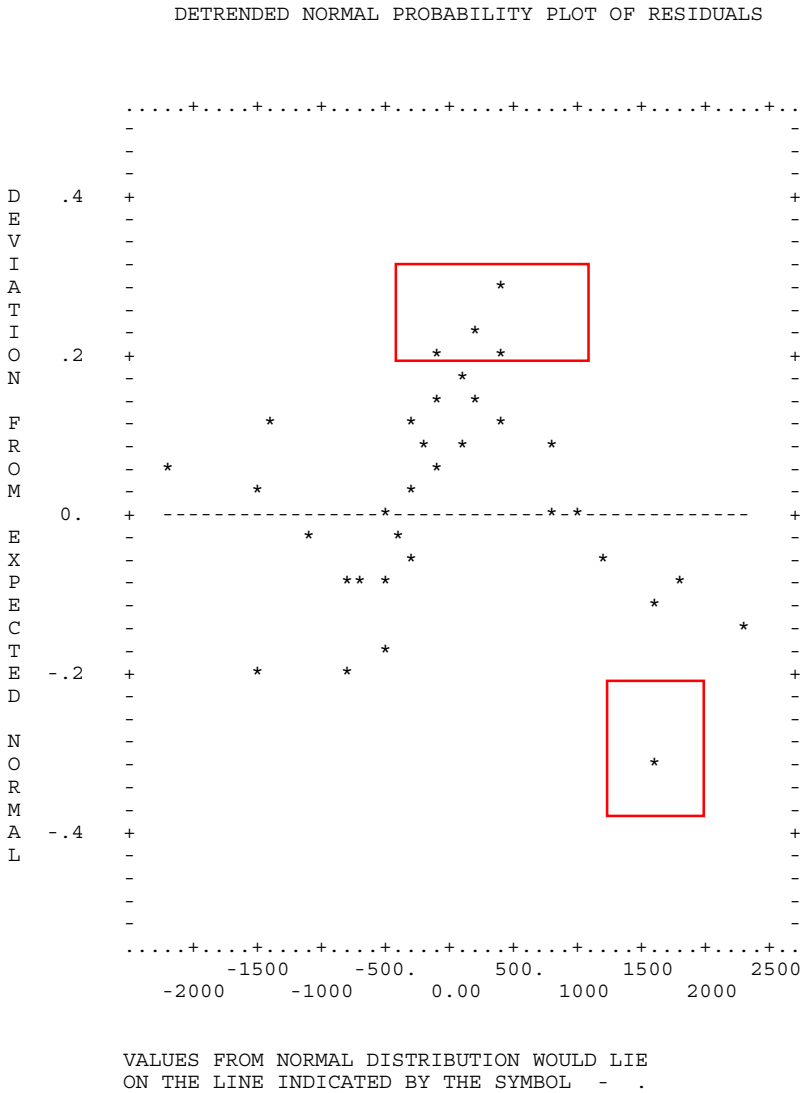


Plot de Probabilidad Normal de los Residuales.
Valores de los residuales (Abcisas) versus Valor esperado de los residuales si tuvieran una distribución normal perfecta (ordenadas).



Si los residuales presentaran una distribución normal deberían ubicarse sobre la linea marcada con el signo / . En el ejemplo los residuales * siguen razonablemente esa linea, con solo algún caso un poco más alejado.

Plot de Probabilidad Normal de los Residuales eliminada la tendencia.
Valores de los residuales (Abcisas) versus desviación de una distribución normal perfecta (ordenadas).



Este plot permite diagnosticar la normalidad de los residuales (como el anterior) de un modo más claro. Los valores de una distribución normal se ubicarían sobre la linea central. En este caso los residuales marcados en cuadrados aparecen alejados de esa linea a más de dos desviaciones típicas.

PROGRAM TERMINATED

8. Análisis de Regresión Stepwise.

El programa 2R ofrece un amplio abanico de posibilidades de análisis utilizando regresión paso a paso. En este tipo de regresión las variables independientes entran de una en una (o también en grupos) en la ecuación, pudiendo controlar el criterio para su inclusión. El programa 2R permite además diagnósticos adicionales muy sofisticados del modelo de regresión.
Para que el análisis paso a paso tenga más interes hemos incluido dos variables nuevas. La variable formación recoge el número de semanas de formación en ventas que ha recibido de la empresa cada vendedor. La variable interes es una puntuación entre 0 y 10 otorgada a cada vendedor por su supervisor de zona para expresar el grado en que considera al vendedor implicado en su trabajo y comprometido con el esfuerzo comercial de la empresa.

INSTRUCCIONES. PROGRAMA 2R.

```
/input var=8. format=free.
/var
 names= sexo, zona, edad, exp, dto, ventas, formacion, interes.
 use = edad to interes.
/regress
 depend= ventas.
/print
 data.
 correlation.
 partial.
/plot
 residuals.
 variable=exp.
 normal.
 dnormal.
/end
0 0 21 10 1  430 4 0
1 0 37 18 9 2640 2 2
0 0 43 23 2 2340 2 1
1 0 54 44 6 5790 5 4
0 0 18  6 2 1000 1 2
1 1 23 38 5 2600 2 5
0 1 35 56 3 7658 6 6
1 1 47 21 9 3300 2 7
0 1 54 33 2 3700 3 8
1 1 19 17 7 1600 2 9
0 0 22 13 2  990 0 2
1 0 34 19 5 1610 2 3
0 0 23 27 0 1940 2 4
1 0 35 24 3 3790 4 5
0 0 19 12 2 1000 0 4
1 1 29 43 5 3800 5 8
0 1 46 56 6 8858 5 9
1 1 47 19 0 4940 5 10
0 1 54 53 1 4700 4 10
1 1 26 13 9 1240 1 5
0 0 28 12 8 1590 1 1
1 0 32 15 7 1640 0 0
0 0 36 33 9 2340 2 3
1 0 35 44 8 6340 5 5
0 0 28 16 9 1050 0 6
1 1 50 38 1 4810 5 9
0 1 55 16 9 4658 5 7
1 1 43 30 4 5300 5 2
0 1 22 33 9 1390 1 4
1 1 21  9 5  200 0 5
0 0 18  1 9  100 0 7
1 1 19  1 1  150 0 9
0 0 26  1 4  200 0 5
1 0 45  1 3 1700 0 5
/end
```

En 2R al especificar en /REGRESS solo la variable dependiente se entiende que todas las demás del USE se toman como independientes.

En el USE utilizamos la expresión TO para expresar que la lista va desde edad hasta interes, sin necesidad de escribir los nombres de las variables intermedias de la lista una a una..

Las instrucciones de /PRINT y de /PLOT utilizadas aquí nos resultan familiares del programa 1R. La instrucción PARTIAL solicita correlaciones parciales.

Los datos son los mismos del caso analizados con anterioridad con el añadido de dos nuevas columnas (las dos últimas) que son respectivamente formacion (número de semanas de formación comercial recibidas de la empresa) e interes (una calificación de 0 a 10 otorgada a cada vendedor por su supervisor, donde 10 es el extremo positivo).

OUTPUT SELECCIONADO

```
BMDP2R - STEPWISE REGRESSION

 CASE     3         4         5         6         7         8
  NO. edad      exp       dto      ventas    formacio interes
----- -------- --------- -------- --------- --------- ---------
    1    21.00     10.00     1.00    430.00      4.00      0.00
    2    37.00     18.00     9.00   2640.00      2.00      2.00
    3    43.00     23.00     2.00   2340.00      2.00      1.00
    4    54.00     44.00     6.00   5790.00      5.00      4.00
    5    18.00      6.00     2.00   1000.00      1.00      2.00
    6    23.00     38.00     5.00   2600.00      2.00      5.00
    7    35.00     56.00     3.00   7658.00      6.00      6.00
    8    47.00     21.00     9.00   3300.00      2.00      7.00
    9    54.00     33.00     2.00   3700.00      3.00      8.00
   10    19.00     17.00     7.00   1600.00      2.00      9.00

NUMBER OF CASES READ. . . . . .      34
```

Por defecto presenta los datos de los 10 primeros casos. Para evitarlo incluir en /PRINT la instrucción: CASE=0.

```
DESCRIPTIVE STATISTICS OF DATA
----------- ---------- -- ----

 VARIABLE    TOTAL          STANDARD  SKEW-              SMALLEST      LARGEST
 NO. NAME     FREQ.  MEAN     DEV.    NESS  KURTOSIS  VALUE  Z-SCR  VALUE  Z-SCR

   3 edad        34  33.647  12.257   0.338  -1.314  18.000 -1.28  55.000  1.74
   4 exp         34  23.382  16.013   0.480  -0.807  1.0000 -1.40  56.000  2.04
   5 dto         34  4.8529  3.1540   0.058  -1.524  0.0000 -1.54  9.0000  1.31
   6 ventas      34  2805.7  2239.7   0.863  -0.050  100.00 -1.21  8858.0  2.70
   7 formacio    34  2.3824  2.0303   0.250  -1.490  0.0000 -1.17  6.0000  1.78
   8 interes     34  5.0588  2.8914   0.038  -1.066  0.0000 -1.75  10.000  1.71

*** N O T E *** KURTOSIS VALUES GREATER THAN ZERO INDICATE A DISTRIBUTION WITH
                HEAVIER TAILS THAN NORMAL DISTRIBUTION.

CORRELATION MATRIX
----------- ------

                edad      exp       dto      ventas   formacio interes
                   3        4         5         6         7         8

edad         3   1.0000
exp          4   0.4860   1.0000
dto          5  -0.0453  -0.0235   1.0000
ventas       6   0.6936   0.8299  -0.0336   1.0000
formacio     7   0.6083   0.7410  -0.1613   0.8657   1.0000
interes      8   0.3067   0.2953  -0.1120   0.3636   0.3109   1.0000
```

Matriz de correlaciones: formacion aparece muy correlacionada con ventas, y también con exp

```
REGRESSION INTERCEPT. . . . . .NON ZERO
REGRESSION WEIGHT VARIABLE. . . . . .
PRINT COVARIANCE MATRIX . . . . . NO
PRINT CORRELATION MATRIX. . . . . YES
PRINT ANOVA AT EACH STEP. . . . . YES
PRINT STEP OUTPUT . . . . . YES
PRINT REGRESSION COEFFICIENT SUMMARY TABLE. . . YES
PRINT PARTIAL CORRELATION SUMMARY TABLE . . . . YES
PRINT F-RATIO SUMMARY TABLE . . . . . NO
PRINT SUMMARY TABLE . . . . . YES
PRINT DATA OR DIAGNOSTICS . . . . . YES
PRINT CORRELATION OF REGRESSION COEFFICIENTS. . NO
PRINT NORMAL PROBABILITY PLOT . . . . . YES
PRINT DETRENDED NORMAL PROBABILITY PLOT . . . . YES
PRINT PLOTS FOR XVAR AND YVAR . . . . . NO
PRINT PLOTS AND DATA. . . . . NO
PRINT PLOTS WITH STATISTICS . . . . . NO
PRINT DIAGNOSTIC PLOT(S). . . . . NO
PRINT CASE-BY-STATISTIC PLOTS . . . . . NO
PRINT ADDED VARIABLE MINIPLOTS. . . . . NO
```

Diagnósticos. Los análisis no solicitados enumeran otras posibilidades del programa

```
STEPPING ALGORITHM. . . . . .F
MAXIMUM NUMBER OF STEPS . . . . . 16
DEPENDENT VARIABLE. . . . . 6 ventas
MINIMUM ACCEPTABLE F-TO-ENTER . . . . . 4.000,   4.000
MAXIMUM ACCEPTABLE F-TO-REMOVE. . . . . 3.900,   3.900
MINIMUM ACCEPTABLE TOLERANCE. . . . . 0.01000
SUBSCRIPTS OF THE INDEPENDENT VARIABLES . . . .   3    4    5    7    8
```

Items relativos al control de la entrada de variables en la ecuación.

Paso número 0

En el Paso 0 la columna PARTIAL CORR. contiene la correlación de Pearson entre la dependiente y cada independiente, todas las tolerancias son 1 (porque todavía no hay ninguna independiente en la ecuación) y el test F es la prueba estadística para una correlación de Pearson simple.

```
STEP NO.    0
--------------

STD. ERROR OF EST.  2239.6968

ANALYSIS OF VARIANCE
                    SUM OF SQUARES     DF    MEAN SQUARE
         RESIDUAL    0.16553597E+09     33   5016242.

       VARIABLES IN EQUATION                    VARIABLES NOT IN EQUATION
-------------------------------------------  ------------------------------
                     STD.ERR        F                    PARTIAL       F
VARIABLE    COEFF.   OF COEFF  TOL. REMOVE(L) VARIABLE   CORR.   TOL.  ENTER(L)
-------------------------------------------  ------------------------------
(CONSTANT2805.7056)
                                             edad      0.6936 1.0000   29.66(1)
                                             exp       0.8299 1.0000   70.81(1)
                                             dto      -0.0336 1.0000    0.04(1)
                                             formacio 0.8657 1.0000   95.71(1)
                                             interes  0.3636 1.0000    4.88(1)

ENTER VARIABLE TO MOVE NEXT :
!V to View Output; ENTER to accept: formacio--->
```

En este punto al ejecutar el trabajo interjectivamente el programa se para y pregunta qué variable deseamos incluir en la ecuación. Si dejamos que el programa haga la elección este introduce la variable formacion que es la que ha presentado una correlación mayor en el paso anterior y la mayor F para entrar.

Paso nª 1. Entra la independiente

```
STEP NO.    1
--------------
VARIABLE ENTERED    7 formacio

MULTIPLE R              0.8657
MULTIPLE R-SQUARE       0.7494
ADJUSTED R-SQUARE       0.7416

STD. ERROR OF EST.  1138.5028
```

El ETE se ha reducido considerablemente al introducir el variable formacion (respecto al que teníamos en el paso 0 solo con la constante). La R múltiple es en este paso la correlación de Pearson entre ventas y formacion (todavía no es propiamente múltiple porque hay una sola independiente). El coeficiente de determinación múltiple (que en este caso también es en realidad simple) indica que formacion explica el 74,94% de la varianza de ventas. este coeficiente se dice que capitaliza el azar, pues esta optimado para los valores de la muestra; si volviéramos a calcular R^2 en otra muestra de la misma población, para este mismo modelo, sería más razonable esperar la R^2 ajustada 0,7416.

$$R^2_{AJUS} = R^2 - \frac{(1-R^2)(k-1)}{N-k}$$

Donde N es el número de casos y K el número de coeficientes a estimar en la ecuación. Una fórmula equivalente para la R^2 ajustada es:

$$R^2_{AJUS} = 1 - \frac{(1-R^2)(N-1)}{N-k}$$

```
ANALYSIS OF VARIANCE
                    SUM OF SQUARES     DF    MEAN SQUARE      F RATIO
         REGRESSION  0.12405794E+09      1  0.1240579E+09        95.71
         RESIDUAL     41478036.         32   1296189.
```

```
       VARIABLES IN EQUATION                          VARIABLES NOT IN EQUATION
-------------------------------------------  -------------------------------
                      STD.ERR        F                      PARTIAL        F
VARIABLE    COEFF.    OF COEFF   TOL. REMOVE(L)  VARIABLE   CORR.    TOL.   ENTER(L)
-------------------------------------------  -------------------------------
(CONSTANT 530.5977)
formacio   954.9835  97.6152 1.0000   95.71(1) edad      0.4202 0.6299     6.65(1)
                                                exp       0.5605 0.4509    14.20(1)
                                                dto       0.2147 0.9740     1.50(1)
                                                interes   0.1985 0.9033     1.27(1)
```

Coeficiente b para cada variable independiente. La constante ha cambiado al incluir formación en la ecuación.

Error típico del coeficiente b para cada independiente

Tolerancia es igual a 1 menos la correlación múltiple al cuadrado de esa independiente con todas las demás. Como solo hay una independiente aquí vale 1. Cuanto mayor la colinealidad menor el valor de tolerancia.

Correlación parcial. Correlación de cada variable independiente con la variable dependiente una vez sustraído el efecto de las variables que ya han entrado en la ecuación.

F to enter. F para entrar a la ecuación. Contrasta el coeficiente de regresión que podría aportar en el paso siguiente cada independiente. Indica cuanto mejoraría la predicción si esa i dependiente entrara en el paso siguiente.

F to remove: F para salir de la ecuación. Contrasta los coeficientes b para determinar la relativa importancia de las variables que ya están en la ecuación. Es igual al (cociente entre b y error típico) elevado al cuadrado. Si este valor es menor que el límite aceptable (ver arriba en valores que controlan la entrada de variables) y menor que el valor de otras variables entradas a la ecuación, la variable se sacará de la ecuación en el próximo paso

```
 ENTER VARIABLE TO MOVE NEXT :
!V to View Output; ENTER to accept: exp      --->
```

El programa propone exp para entrar a continuación debido a su mayor F to enter.

```
STEP NO.    2
---------------
VARIABLE ENTERED    4 exp

MULTIPLE R              0.9100
MULTIPLE R-SQUARE       0.8282
ADJUSTED R-SQUARE       0.8171

STD. ERROR OF EST.    957.9185

ANALYSIS OF VARIANCE
                     SUM OF SQUARES      DF    MEAN SQUARE       F RATIO
         REGRESSION  0.13709013E+09       2  0.6854506E+08         74.70
         RESIDUAL     28445844.          31   917607.9
```

En el paso 2 tenemos ya los 2 mejores predictores y la capacidad explicativa de la ecuación ha mejorado sensiblemente.

```
       VARIABLES IN EQUATION                          VARIABLES NOT IN EQUATION
-------------------------------------------  -------------------------------
                      STD.ERR        F                      PARTIAL        F
VARIABLE    COEFF.    OF COEFF   TOL. REMOVE(L)  VARIABLE   CORR.    TOL.   ENTER(L)
-------------------------------------------  -------------------------------
(CONSTANT -22.1953)
exp         58.4435   15.5080 0.4509   14.20(1) edad      0.4638 0.6272     8.22(1)
formacio   613.4081  122.3142 0.4509   25.15(1) dto       0.1629 0.9535     0.82(1)
                                                interes   0.1717 0.8940     0.91(1)
```

Ecuación de regresión lineal múltiple en el paso 2.

Tolerancia es igual a 1 menos la correlación múltiple al cuadrado de esa independiente con todas las demás ya incluidas en la ecuación. Cuanto mayor la colinealidad menor el valor de tolerancia.. Si es menor que el valor 0'01 la variable no podrá entrar en la ecuación

```
ENTER VARIABLE TO MOVE NEXT :
!V to View Output; ENTER to accept: edad     --->
```

Debido a su F to Enter el programa propone incluir a continuación la variable Edad.

```
STEP NO.     3
---------------
VARIABLE ENTERED     3 edad

MULTIPLE R              0.9301
MULTIPLE R-SQUARE       0.8651
ADJUSTED R-SQUARE       0.8516

STD. ERROR OF EST.    862.7079

ANALYSIS OF VARIANCE
                    SUM OF SQUARES      DF     MEAN SQUARE       F RATIO
         REGRESSION  0.14320802E+09       3  0.4773600E+08         64.14
         RESIDUAL     22327948.          30   744264.9

       VARIABLES IN EQUATION                        VARIABLES NOT IN EQUATION
-------------------------------------------  --------------------------------
                     STD.ERR        F                        PARTIAL       F
VARIABLE    COEFF.   OF COEFF  TOL. REMOVE(L)  VARIABLE  CORR.    TOL.  ENTER(L)
-------------------------------------------  --------------------------------
(CONSTANT*********)
edad        44.3573  15.4714 0.6272    8.22(1) dto      0.1534 0.9502    0.70(1)
exp         55.7960  13.9971 0.4489   15.89(1) interes  0.1165 0.8738    0.40(1)
formacio   465.9724 121.5688 0.3702   14.69(1)

 ENTER VARIABLE TO MOVE NEXT :
!V to View Output; ENTER to accept: NONE     --->

***** F LEVELS(  4.000,  3.900) OR TOLERANCE INSUFFICIENT FOR FURTHER STEPPING

         STEPWISE REGRESSION COEFFICIENTS

  VARIABLES  0 Y-INTCPT  3 edad      4 exp       5 dto       7 formacio
STEP
   0        2805.7056*   126.7367    116.0748    -23.8450    954.9835
   1         530.5977*    48.4260     58.4435     77.3366    954.9835*
   2         -22.1953*    44.3573     58.4435*    49.0957    613.4081*
   3       -1101.5417*    44.3573*    55.7960*    41.0355    465.9724*

         STEPWISE REGRESSION COEFFICIENTS

  VARIABLES  8 interes
STEP
   0          281.6584
   1           80.9795
   2           58.3219
   3           35.4603

*** N O T E *** 1) REGRESSION COEFFICIENTS FOR VARIABLES IN THE EQUATION ARE
                   INDICATED BY AN ASTERISK.
                2) THE REMAINING COEFFICIENTS ARE THOSE WHICH WOULD BE
                   OBTAINED IF THAT VARIABLE WERE TO ENTER IN THE NEXT STEP.
```

En el paso 3, al incluir edad como tercera variable independiente el ajuste de la ecuación mejora todavía un poco más, alcanzando una R múltiple cuadrado ajustada de 0'85 y reduciendo el ETE hasta 862`7079.

Ecuación de regresión lineal con 3 independientes. La intercept no la imprime aquí posiblemente por cuestión de formato, pero la muestra abajo en el resumen para cada paso.

Las F to Enter de dto e interes están por debajo del límite mínimo para aceptar una variable en la ecuación. Consecuentemente el programa propone no introducir ninguna variable más.

(Sigue aquí dado que no cabe todo en una misma linea)

Resumen de los coeficientes de la ecuación en cada paso. Obsérvese que solo aquellos indicados con un asterisco son los que corresponden a variables en la ecuación en ese paso. Los demás expresan cual sería el coeficiente de esa variable de haber entrado en el SIGUIENTE paso. Por ejemplo, en el paso 3 las variables dto e interes están fuera, pero, si se hubiesen aceptado cada una de ellas separadamente en un hipotético paso 4 hubieran presentado los coeficientes b 41´0355 y 35'4603 respectivamente.

Resumen de las correlaciones parciales paso a paso: Correlaciones de cada indep con la dependiente eliminado el efecto lineal de todas las dependientes ya incluidas en la

PARTIAL CORRELATIONS

VARIABLES	3 edad	4 exp	5 dto	7 formacio	8 interes
STEP					
0	0.6936	0.8299	-0.0336	0.8657	0.3636
1	0.4202	0.5605	0.2147	0.8657*	0.1985
2	0.4638	0.5605*	0.1629	0.6693*	0.1717
3	0.4638*	0.5884*	0.1534	0.5734*	0.1165

Tabla resumen, de pasos, variables entradas en cada paso , (variables que salen en cada paso, aquí ninguna), valores de R y R cuadrado, cambio en R de un paso al siguiente, F para entrar en cada paso , (F para salir en cada paso, aquí no se ha producido ninguna salida) y número de variables en la ecuación.

SUMMARY TABLE

STEP NO.	VARIABLE ENTERED	VARIABLE REMOVED	MULTIPLE R	RSQ	CHANGE IN RSQ	F TO ENTER	F TO REMOVE	NO.OF VAR. INCLUDED
1	7 formacio		0.8657	0.7494	0.7494	95.71		1
2	4 exp		0.9100	0.8282	0.0787	14.20		2
3	3 edad		0.9301	0.8651	0.0370	8.22		3

SERIAL CORRELATION -0.2519
DURBIN-WATSON STATISTIC 2.2851 BASED ON 34 CASES

Listado de casos con sus valores estimados (PREDICTED) y sus residuales. Cada asterisco indica una desviación típica de distancia a la media (en este caso dos son particularmente grandes).
A continuación una columna de ponderaciones (aquí todas igual a 1) y los datos de las variables

LIST OF PREDICTED VALUES, RESIDUALS, AND VARIABLES

- CASES WITH MISSING VALUES ARE MARKED WITH A MINUS SIGN BETWEEN THE CASE NUMBER AND CASE LABEL.
- ASTERISKS (UP TO 3) TO THE RIGHT OF A RESIDUAL INDICATE THAT THE RESIDUAL DEVIATES FROM THE MEAN BY MORE THAN THAT NUMBER OF STANDARD DEVIATIONS.
- MISSING VALUES AND VALUES OUT OF RANGE ARE DENOTED BY VALUES GREATER THAN OR EQUAL TO 3.24519E+32 IN ABSOLUTE VALUE.

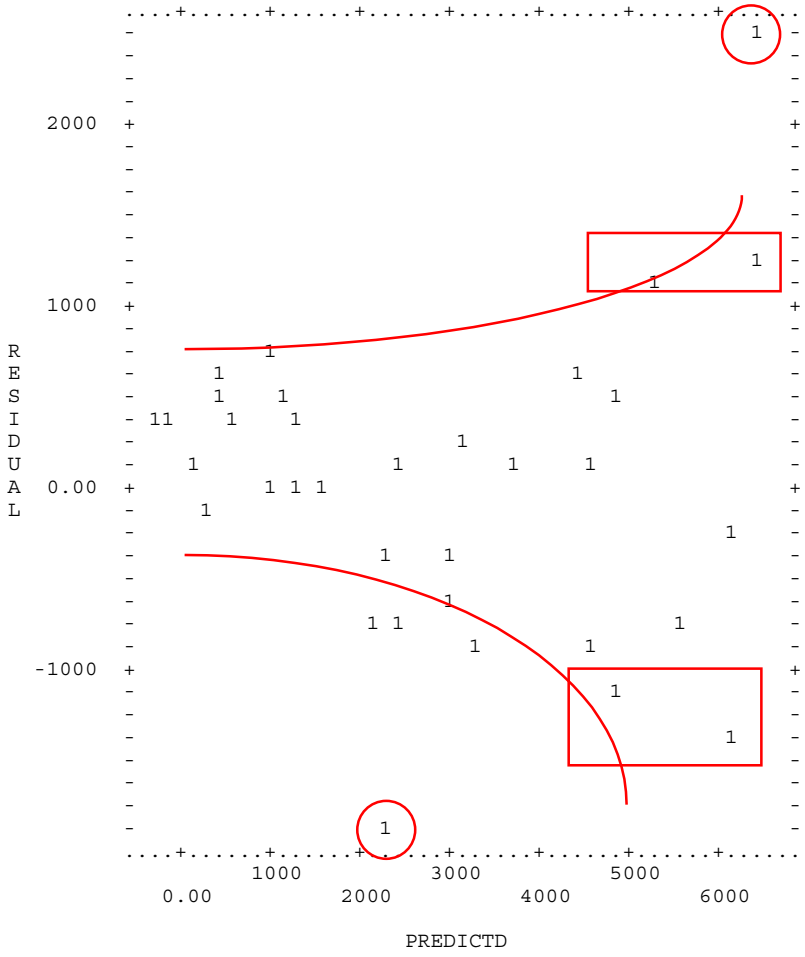
CASE NO.	LABEL	PREDICTED	RESIDUAL	WEIGHT	6 ventas	3 edad	4 exp
1		2251.8123	-1821.8123**	1.000	430.0000	21.0000	10.0000
2		2475.9529	164.0471	1.000	2640.0000	37.0000	18.0000
3		3021.0769	-681.0769	1.000	2340.0000	43.0000	23.0000
4		6078.6416	-288.6416	1.000	5790.0000	54.0000	44.0000
5		497.6389	502.3611	1.000	1000.0000	18.0000	6.0000
6		2970.8706	-370.8706	1.000	2600.0000	23.0000	38.0000
7		6371.3770	1286.6230*	1.000	7658.0000	35.0000	56.0000
8		3086.9143	213.0857	1.000	3300.0000	47.0000	21.0000
9		4532.9404	-832.9404	1.000	3700.0000	54.0000	33.0000
10		1621.7249	-21.7249	1.000	1600.0000	19.0000	17.0000
11		599.6680	390.3320	1.000	990.0000	22.0000	13.0000
12		2398.6770	-788.6770	1.000	1610.0000	34.0000	19.0000
13		2357.1145	-417.1145	1.000	1940.0000	23.0000	27.0000
14		3653.9592	136.0408	1.000	3790.0000	35.0000	24.0000
15		410.7999	589.2001	1.000	1000.0000	19.0000	12.0000
16		4913.9121	-1113.9121*	1.000	3800.0000	29.0000	43.0000
17		6393.3350	2464.6650**	1.000	8858.0000	46.0000	56.0000
18		4373.2402	566.7598	1.000	4940.0000	47.0000	19.0000
19		6114.8330	-1414.8330*	1.000	4700.0000	54.0000	53.0000
20		1243.0698	-3.0698	1.000	1240.0000	26.0000	13.0000
21		1275.9885	314.0115	1.000	1590.0000	28.0000	12.0000
22		1154.8333	485.1667	1.000	1640.0000	32.0000	15.0000
23		3268.5359	-928.5359*	1.000	2340.0000	36.0000	33.0000
24		5235.8525	1104.1475*	1.000	6340.0000	35.0000	44.0000
25		1033.2000	16.8000	1.000	1050.0000	28.0000	16.0000
26		5566.4365	-756.4365	1.000	4810.0000	50.0000	38.0000
27		4560.7100	97.2900	1.000	4658.0000	55.0000	16.0000
28		4809.5664	490.4336	1.000	5300.0000	43.0000	30.0000
29		2181.5608	-791.5608	1.000	1390.0000	22.0000	33.0000
30		332.1265	-132.1265	1.000	200.0000	21.0000	9.0000
31		-247.3137	347.3137	1.000	100.0000	18.0000	1.0000
32		-202.9563	352.9563	1.000	150.0000	19.0000	1.0000
33		107.5450	92.4550	1.000	200.0000	26.0000	1.0000
34		950.3345	749.6655	1.000	1700.0000	45.0000	1.0000

LIST OF VARIABLES (CONTINUED)

CASE NO.	LABEL	5 dto	7 formacio	8 interes
1		1.0000	4.0000	0.0000
2		9.0000	2.0000	2.0000
3		2.0000	2.0000	1.0000
4		6.0000	5.0000	4.0000
5		2.0000	1.0000	2.0000
6		5.0000	2.0000	5.0000
7		3.0000	6.0000	6.0000
8		9.0000	2.0000	7.0000
9		2.0000	3.0000	8.0000
10		7.0000	2.0000	9.0000
11		2.0000	0.0000	2.0000
12		5.0000	2.0000	3.0000
13		0.0000	2.0000	4.0000
14		3.0000	4.0000	5.0000
15		2.0000	0.0000	4.0000
16		5.0000	5.0000	8.0000
17		6.0000	5.0000	9.0000
18		0.0000	5.0000	10.0000
19		1.0000	4.0000	10.0000
20		9.0000	1.0000	5.0000
21		8.0000	1.0000	1.0000
22		7.0000	0.0000	0.0000
23		9.0000	2.0000	3.0000
24		8.0000	5.0000	5.0000
25		9.0000	0.0000	6.0000
26		1.0000	5.0000	9.0000
27		9.0000	5.0000	7.0000
28		4.0000	5.0000	2.0000
29		9.0000	1.0000	4.0000
30		5.0000	0.0000	5.0000
31		9.0000	0.0000	7.0000
32		1.0000	0.0000	9.0000
33		4.0000	0.0000	5.0000
34		3.0000	0.0000	5.0000

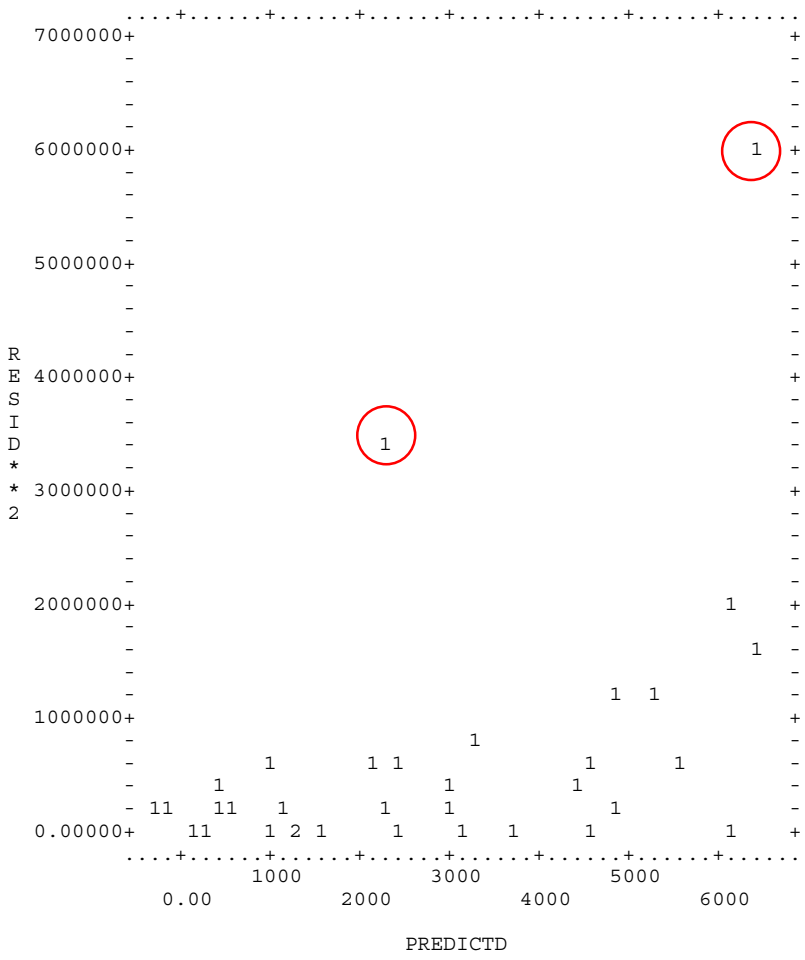
Continua el listado de datos caso a caso para cada variable.

Plot de Valores Estimados (Abcisas) versus Residuales (ordenadas).



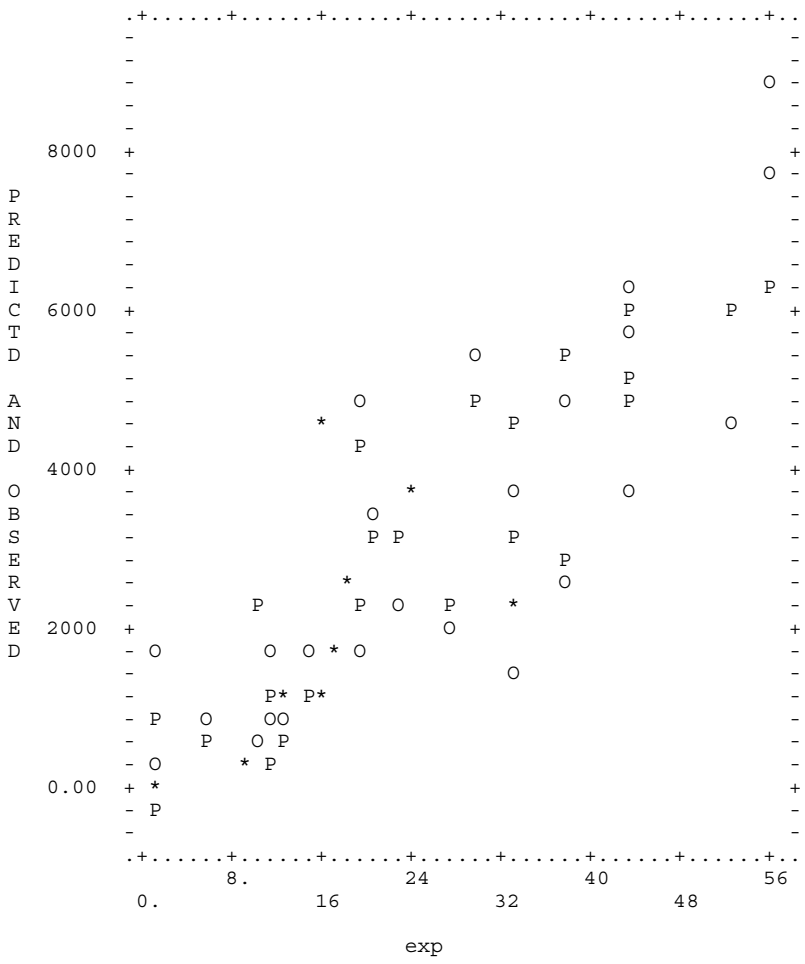
Se señalan algunos residuales pronunciadamente altos. La dispersión de los residuales parece aumentar según aumentan los valores pronosticados sugiriendo (como se ha marcado con las líneas rojas trazadas sobre el gráfico indicativamente) que pueden tener un comportamiento

Plot de Valores Estimados (Abcisas) versus Residuales elevados al cuadrado (ordenadas).

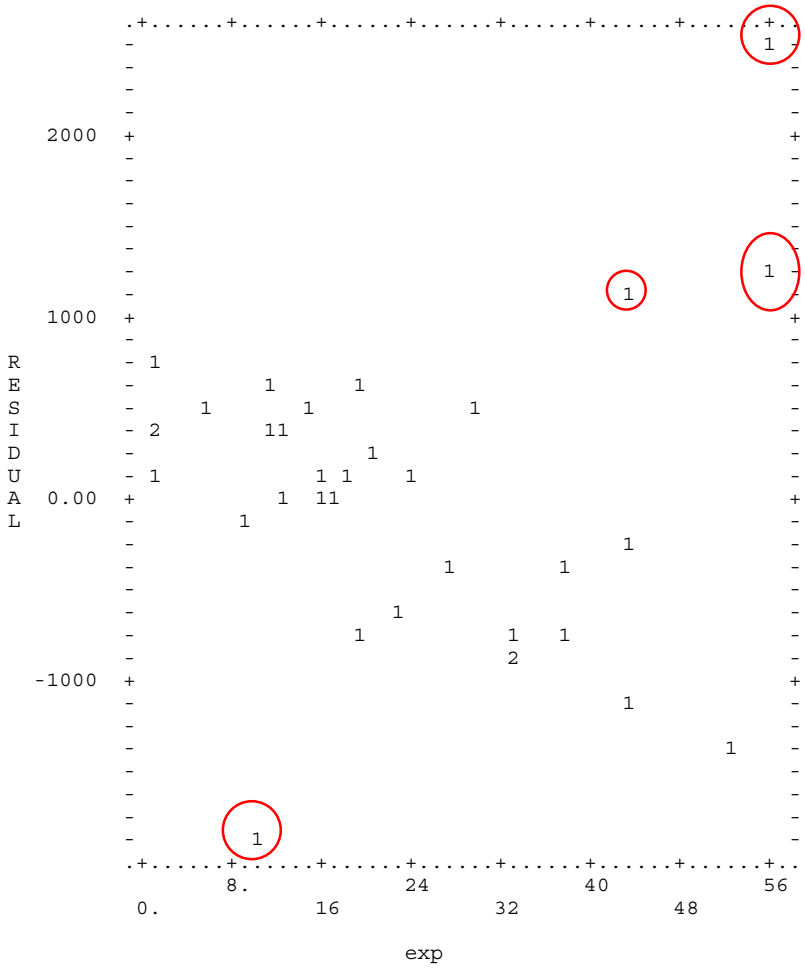


Al elevar al cuadrado los residuales estos pierden el signo y presentan solo su magnitud, acentuada por el cuadrado, de modo que los grandes residuales tienen a parecer en este gráfico proporcionalmente más grandes todavía.
El gráfico muestra 2 casos con residuales substancialmente grandes que habría que identificar y analizar cuidadosamente. Para empezar hay que descartar siempre que no se trate de un mero error de datos al tomar la información o al transcribirla. Conviene tratar de averiguar que pasa en estos casos para que se comporten de modo ajeno al modelo general.
Excluidos estos dos casos los residuales parecen ser algo

Plot de Valores de la Variable Experiencia (Abcisas) versus Valores Observados (O) y Estimados (P)de la variable dependiente Ventas (en ordenadas).

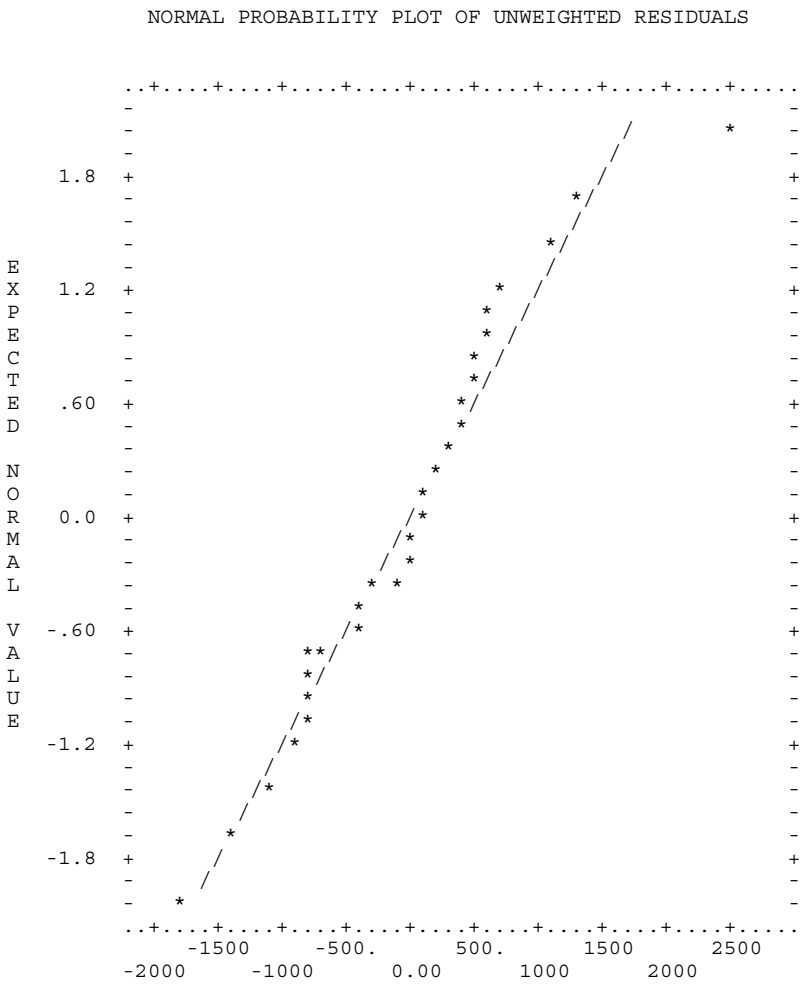


Plot de Valores de la variable independiente Experiencia (Abcisas) versus Residuales (ordenadas).



Se marcan algunos casos de residuales con valores grandes en términos absolutos que posiblemente tienen una influencia marcada en la solución de regresión. Por otra parte, excluidos esos residuales mayores, parece apreciarse una tendencia lineal en los demás residuales.:
La ecuación infraestimaría para valores bajos de experiencia y sobrestimaría para valores altos de experiencia.
Podría convenir identificar estos casos especiales con altos residuales, analizarlos más de cerca (2R ofrece muy potentes elementos diagnósticos para este fin) y quizás recalcular el modelo excluyéndolos o atenuando su impacto en la ecuación.

Plot de Probabilidad Normal de los Residuales.
Valores de los residuales (Abcisas) versus Valor esperado de los residuales si tuvieran una distribución normal perfecta (ordenadas).



Si los residuales presentaran una distribución normal deberían ubicarse sobre la linea marcada con el signo / .
En este caso los residuales * se alejan de esa linea manifiestamente en algunos puntos.

Plot de Probabilidad Normal de los Residuales eliminada la tendencia.
Valores de los residuales (Abcisas) versus desviación de una distribución normal perfecta (ordenadas).

```
          DETRENDED NORMAL PROBABILITY PLOT OF UNWEIGHTED RESIDUALS

                  ..+....+....+....+....+....+....+....+....+....+....+.....
                  -                                                       -
                  -                                                       -
                  -                                    *                  -
D    .4           +                                    *                  +
E                 -                                                       -
V                 -                                                       -
I                 -                                                       -
A                 -                                   **                  -
T                 -                                                       -
I                 -                                                       -
O    .2           +                                                       +
N                 -                 *                 *                 -
                  -                  *                     * *            -
F                 -                                  *                    -
R                 -                                   *                   -
O                 -                 *                 *                   -
M                 -    *                            * *                   -
     0.           +    ----------------------*-*-----------------------    +
E                 -         *           *                                 -
X                 -                  *     *     *                        -
P                 -                              *                        -
E                 -                         *    *                        -
C                 -              *                 *                      -
T                 -                 *              *                      -
E   -.2           +                 *                                     +
D                 -                              *                        -
                  -                               *                       -
N                 -                                                       -
O                 -                                                       -
R                 -                                                       -
M                 -                                                       -
A   -.4           +                                                       +
L                 -                                                       -
                  -                                                       -
                  -                                                       -
                  -                                                       -
                  ..+....+....+....+....+....+....+....+....+....L.....
                       -1500      -500.       500.       1500       2500
                  -2000      -1000       0.00       1000       2000

                  VALUES FROM NORMAL DISTRIBUTION WOULD LIE
                  ON THE LINE INDICATED BY THE SYMBOL  -   .
```

Valores de residuales desviados respecto a los esperables en una distribución normal.

```
PROGRAM TERMINATED
```

9. Explorando los principales subconjuntos de variables independientes.

El programa 9R permite obtener todos los subconjuntos de variables independientes disponibles para el pronóstico lineal de una variable dependiente dada, lo que facilita extraordinariamente un trabajo exploratorio cuando el número de posibles predictores es grande.

En contrapartida debe tenerse en cuenta que este procedimiento (como también sucede con el método stepwise) capitaliza el azar a favor de los resultados, por lo que es poco riguroso cuando se trata de contrastar un modelo establecido. Esto supone que los hallazgos empíricos fundados en estas técnicas exigen estudios de replicación (es decir, repetir el estudio en otra muestra bajo las mismas condiciones -replicación directa- o variando estas sistemáticamente -replicación sistemática-).

INSTRUCCIONES. PROGRAMA 9R.

```
/input var=8. format=free.
/var
 names= sexo, zona, edad, exp, dto, ventas, formacion, interes.
 use = edad to interes.
/regress
 depend= ventas.
 indep=edad, exp, dto, formacion, interes.
/end
aquí van los mismos datos
que en el análisis anterior
/end
```

Como puede apreciarse no se han introducido instrucciones especiales, para un output básico se utilizan las mismas que otros programas de regresión. Eso sí, a 9R hay que especificarle cuales son las

OUTPUT SELECCIONADO

```
BMDP9R - ALL POSSIBLE SUBSETS REGRESSION

/input var=8. format=free.
/var
 names= sexo, zona, edad, exp, dto, ventas, formacion, interes.
 use = edad to interes.
/regress
 depend= ventas.
 indep=edad, exp, dto, formacion, interes.
/end

DATA AFTER TRANSFORMATIONS
--------------------------
 CASE    3        4        5        7        8        6
  NO. edad      exp      dto      formacio interes  ventas
----- -------- -------- -------- -------- -------- --------
    1    21.00    10.00     1.00     4.00     0.00   430.00
    2    37.00    18.00     9.00     2.00     2.00  2640.00
    3    43.00    23.00     2.00     2.00     1.00  2340.00
    4    54.00    44.00     6.00     5.00     4.00  5790.00
    5    18.00     6.00     2.00     1.00     2.00  1000.00
    6    23.00    38.00     5.00     2.00     5.00  2600.00
    7    35.00    56.00     3.00     6.00     6.00  7658.00
    8    47.00    21.00     9.00     2.00     7.00  3300.00
    9    54.00    33.00     2.00     3.00     8.00  3700.00
   10    19.00    17.00     7.00     2.00     9.00  1600.00

NUMBER OF CASES READ. . . . . .      34


SUMMARY STATISTICS FOR EACH VARIABLE
------------------------------------
                          STANDARD    COEFFICIENT   SMALLEST      LARGEST
  VARIABLE     MEAN       DEVIATION   OF VARIATION   VALUE         VALUE
------------ ------------ ------------ ------------ ------------ -----------

  3 edad         33.64706     12.25705      0.364283     18.00000     55.00000
  4 exp          23.38235     16.01328      0.684845      1.00000     56.00000
  5 dto           4.85294      3.15395      0.649905      0.00000      9.00000
  7 formacio      2.38235      2.03030      0.852223      0.00000      6.00000
  8 interes       5.05882      2.89138      0.571552      0.00000     10.00000
  6 ventas     2805.70588   2239.69685      0.798265    100.00000   8858.00000
```

Instrucciones repetidas, y lista de los 10 primeros casos presentada por defecto.

Tabla de estadísticos descriptivos de las variables.

```
CORRELATIONS
------------


                edad      exp       dto      formacio interes  ventas
                   3        4         5         7        8         6

edad         3    1.000
exp          4    0.486     1.000
dto          5   -0.045    -0.023     1.000
formacio     7    0.608     0.741    -0.161     1.000
interes      8    0.307     0.295    -0.112     0.311     1.000
ventas       6    0.694     0.830    -0.034     0.866     0.364     1.000
```

Matriz de coeficientes de Correlación de Pearson entre las variables. Coloca la variable dependiente

```
FIRST DIGITS OF CORRELATIONS
----------------------------

   3 edad      *
   4 exp       4*
   7 formacio 67*
   6 ventas   688*
   8 interes  3233*
   5 dto        1 1*
```

Este diagrama es un intento de representar gráficamente la matriz de correlaciones presentando el primer dígito (cuidado al interpretarla porque las variables no están en el mismo orden que en la matriz anterior).

```
                                SUBSETS WITH    1 VARIABLES
                                ----------------------------
            ADJUSTED
 R-SQUARED  R-SQUARED         CP

  0.749432   0.741601      24.22 formacio

  0.688743   0.679017      37.36 exp

  0.481059   0.464842      82.30 edad

  0.132214   0.105096     157.79 interes

  0.001127  -0.030087     186.16 dto
```

Subconjuntos con 1 predictor. El programa muestra ordenadas las cinco mejores soluciones de regresión con un solo predictor en cada una de ellas.
La mejor solución con un solo predictor es la que ofrece la variable formación, seguida por experiencia y luego por edad. Las otras dos son ya muy malas.

¿Cómo comparar entre sí diversas soluciones de regresión lineales para los mismos datos?

Por defecto el criterio de selección de subconjuntos del 9R es el estadístico Cp de Mallows. Cp es esencialmente un cociente entre un índice del residual que produce una solución determinada (modelo restringido) y un índice del residual que produciría una solución basada en todos los predictores (modelo completo). Esto ofrece un estadístico útil para comparar soluciones de regresión entre si. Cuanto mejor es una solución menor residual produce y por tanto menor es su Cp. Exactamente Cp es la Suma de Cuadrados del Error o residual del modelo restringido dividida por la media cuadrática del error del modelo completo, menos (N menos 2 * (k menos 1)) donde N y K son como de costumbre el número de casos y el número de coeficientes en la ecuación del modelo restringido. En forma de fórmula:

$$C_p de Mallows = \frac{SCE_{RESTR.}}{MCE_{COMPL.}} - (N - 2(k-1))$$

El programa 9R también puede seleccionar los subconjuntos basándose en que presenten el mayor coeficiente de correlación múltiple al cuadrado (R^2). En ese caso en el párrafo /REGRESS hay que añadir:
METHOD = RSQ.
En muchas ocasiones los dos criterios de selección producen una selección de mejores subconjuntos semejante o muy semejante entre sí.

SUBSETS WITH 2 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP			
0.828159	0.817072	9.19	VARIABLE	COEFFICIENT	T-STATISTIC
			4 exp	58.4435	3.77
			7 formacio	613.408	5.02
			INTERCEPT	-22.1945	
0.799061	0.786097	15.48	edad	exp	
0.793673	0.780362	16.65	edad	formacio	
0.760983	0.745563	23.72	dto	formacio	
0.759304	0.743775	24.09	formacio	interes	
0.704132	0.685044	36.03	exp	interes	
0.688943	0.668875	39.31	exp	dto	
0.506187	0.474328	78.86	edad	interes	
0.481063	0.447584	84.30	edad	dto	
0.132266	0.076283	159.78	dto	interes	

Subconjuntos con dos i independientes. El mejor empíricamente es el primero, con exp y formacio como independientes, pero las 3 ó 4 soluciones siguientes también son buenas empíricamente y podrían ser elegidas por otras razones (razones de teoría , razones prácticas de accesibilidad de los datos de los pedictores, etc.). Las últimas de la lista son definitivamente mucho peores.
De la mejor solución el programa ofrece los coeficientes de regresión b y la intercept y el estadístico t para cada coeficiente b (que es el coeficiente b dividido por su error típico).
De R cuadrado y de R cuadrado ajustada vale lo explicado anteriormente para el 2R.

SUBSETS WITH 3 VARIABLES

R-SQUARED	ADJUSTED R-SQUARED	CP				
0.865117	0.851629	3.19	VARIABLE	COEFFICIENT	T-STATISTIC	
			3 edad	44.3574	2.87	
			4 exp	55.7960	3.99	
			7 formacio	465.972	3.83	
			INTERCEPT	-1101.54		
0.833227	0.816549	10.09	exp	formacio	interes	
0.832716	0.815988	10.20	exp	dto	formacio	
0.802697	0.782967	16.70	edad	exp	interes	
0.802415	0.782657	16.76	edad	dto	formacio	
0.799063	0.778969	17.48	edad	exp	dto	
0.798216	0.778038	17.67	edad	formacio	interes	
0.772364	0.749600	23.26	dto	formacio	interes	
0.704132	0.674546	38.03	exp	dto	interes	
0.506391	0.457030	80.82	edad	dto	interes	

Subconjuntos de 3 variables independientes. El mejor añade ahora la variable edad a la mejor solución anterior de 2 independientes.
Obsérvese que, en general, al aumentar el número de predictores tiende a aumentar el valor de R cuadrado. Con una sola independiente la mejor solución obtenía un 0'74, con 2 independientes se obtiene un 0'82 como máximo, y con 3 se alcanza un 0'86. Como puede sospecharse, en general, el crecimiento de R cuadrado al aumentar el número de variables independientes en la ecuación es negativamente acelerado. La cuestión es encontrar un modelo que explique satisfactoriamente siendo lo más

```
                                        SUBSETS WITH    4 VARIABLES
                                        ---------------------------
             ADJUSTED
 R-SQUARED  R-SQUARED        CP

  0.868290   0.850123       4.50  VARIABLE      COEFFICIENT  T-STATISTIC
                                   3 edad            43.5936      2.80
                                   4 exp             54.1192      3.81
                                   5 dto             41.0355      0.84
                                   7 formacio        488.861      3.90
                                    INTERCEPT       -1290.31

  0.866948   0.848596       4.79  VARIABLE      COEFFICIENT  T-STATISTIC
                                   3 edad            42.8570      2.71
                                   4 exp             54.9637      3.87
                                   7 formacio        460.645      3.74
                                   8 interes         35.4604      0.63
                                    INTERCEPT       -1198.29

  0.838643   0.816386      10.92 exp       dto       formacio  interes

  0.808023   0.781543      17.54 edad      dto       formacio  interes

  0.802723   0.775512      18.69 edad      exp       dto       interes

                                        SUBSETS WITH    5 VARIABLES
                                        ---------------------------
             ADJUSTED
 R-SQUARED  R-SQUARED        CP

  0.870611   0.847506       6.00  VARIABLE      COEFFICIENT  T-STATISTIC
                                   3 edad            41.8368      2.63
                                   4 exp             53.0457      3.68
                                   5 dto             44.2776      0.89
                                   7 formacio        484.645      3.83
                                   8 interes         40.0932      0.71
                                    INTERCEPT       -1414.61

STATISTICS FOR 'BEST' SUBSET
----------------------------
MALLOWS' CP                          3.19
SQUARED MULTIPLE CORRELATION      0.86512
MULTIPLE CORRELATION              0.93012
ADJUSTED SQUARED MULT. CORR.      0.85163
RESIDUAL MEAN SQUARE         744266.607168
STANDARD ERROR OF EST.          862.708877
F-STATISTIC                      64.14
NUMERATOR DEGREES OF FREEDOM         3
DENOMINATOR DEGREES OF FREEDOM      30
SIGNIFICANCE (TAIL PROB.)         0.0000

*** N O T E *** THE ABOVE F-STATISTIC AND ASSOCIATED SIGNIFICANCE TEND TO BE
                LIBERAL WHENEVER A SUBSET OF VARIABLES IS SELECTED BY THE CP OR
                ADJUSTED R-SQUARED CRITERIA.


----------------------------------------

                                                                    CONTRI-
  VARIABLE      REGRESSION      STANDARD  STAND.     T-   2TAIL     TOL-  BUTION
NO.   NAME     COEFFICIENT        ERROR   COEF.   STAT.    SIG.   ERANCE TO R-SQ

   INTERCEPT      -1101.54       460.566  -0.492  -2.39   0.023
  3 edad           44.3573       15.4714   0.243   2.87   0.008 0.627170 0.03696
  4 exp            55.7960       13.9971   0.399   3.99   0.000 0.448928 0.07144
  7 formacio       465.972       121.569   0.422   3.83   0.001 0.370212 0.06606

PROGRAM TERMINATED
```

Los mejores subconjuntos de 4 variables no mejoran la R cuadrado ajustada del mejor subconjunto de 3, y, además producen un estadístico CP peor (es decir más alto). Es claro que la inclusión de una cuarta variable independiente en el modelo está injustificada empíricamente. El comentario anterior puede extenderse, lógicamente, a los modelos con 5 variables

Por último el programa presenta el mejor subconjunto de predicción (considerando los de todos los tamaños analizados) y ofrece una detallada descripción del mismo con estadísticos que ya hemos explicado en detalle en análisis anteriores.
Dado que la ecuación es el resultado de un proceso sistemático de búsqueda los tests estadísticos de contraste deben considerarse con mucha prudencia. Un contraste del modelo exigiría un estudio de replicación.

Por último se presenta la ecuación para el mejor modelo:
Ventas' = -1101'54 + 44'3573*Edad + 55'796*Exp + 465'972*formacio

10. Correlaciones parciales y regresión multivariada.

INSTRUCCIONES. PROGRAMA 6R.

```
/input var=8. format=free.
/var
 names= sexo, zona, edad, exp, dto,
 ventas, formacion, interes.
 use = edad to interes.
/regress
 depend= ventas, interes.
 indep=edad, exp, dto, formacion.
/end
 (aquí los mismos datos
  del análisis anterior)
/end
```

El programa 6R sirve para dos propósitos:
1) Obtener correlaciones parciales: es decir obtener las correlaciones entre un conjunto de variables después de eliminar de cada una de ellas los efectos lineales de un segundo conjunto de variables.
2) Obtener regresiones lineales (simples o múltiples) multivariadas. Se denomina regresión multivariada a aquella en la que se desea pronosticar un conjunto (dos o más) de variables dependientes con el mismo conjunto (una o más) de variables independientes. (En realidad se trata de realizar varios análisis de regresión separados en los que para diferentes dependientes se utilizan las mismas independientes).
En las instrucciones adjuntas se solicita que tanto ventas, (un indicador objetivo de desempeño,) como interes, (un indicador subjetivo de desempeño,) sean pronosticadas por las independientes edad, exp, dto y formacio.

OUTPUT SELECCIONADO

```
BMDP6R - PARTIAL CORRELATION AND MULTIVARIATE REGRESSION
```

Aquí el output presenta las instrucciones, los datos de los 10 primeros casos y los estadísticos descriptivos habituales.

```
NUMBER OF CASES READ. . . . . .      34
```

```
CORRELATIONS

                edad       exp       dto      formacio ventas    interes
                   3         4         5         7         6         8

edad        3    1.000
exp         4    0.486     1.000
dto         5   -0.045    -0.023     1.000
formacio    7    0.608     0.741    -0.161     1.000
ventas      6    0.694     0.830    -0.034     0.866     1.000
interes     8    0.307     0.295    -0.112     0.311     0.364     1.000
```

Matriz de correlaciones (coeficientes de correlación de Pearson)

```
FIRST DIGITS OF CORRELATIONS
----------------------------

   3 edad      *
   4 exp       4*
   5 dto         *
   7 formacio 671*
   6 ventas    68 8*
   8 interes   32133*
```

Representación semi-gráfica de la matriz de correlaciones mostrando el primer dígito de las mismas.

El cuadro siguiente es autoexplicativo. Presenta "la correlación múltiple al cuadrado de cada variable independiente con todas las demás variables independientes y los tests de significación de la regresión múltiple". Los grados de libertad de esos test F son (k - 1) en el numerador (es decir 3) y (N -k) en el denominador (es decir 30). (N= Número de casos; k=Nº de variables independientes en una ecuación de regresión más 1).
(Ver interpretación de la linea referida a formación como ejemplo de lectura de estos resultados en el cuadro explicativo inferior)

```
SQUARED MULTIPLE CORRELATION OF EACH INDEPENDENT
VARIABLE WITH ALL OTHER INDEPENDENT VARIABLES
AND TESTS OF SIGNIFICANCE OF MULTIPLE REGRESSION
DEGREES OF FREEDOM FOR F-STATISTICS ARE     3 AND     30

                SQUARED
  VARIABLE      MULTIPLE       MULTIPLE       F-STATISTIC   SIGNIFICANCE
NO.     NAME    CORRELATION    CORRELATION                  (P LESS THAN)

  3 edad          0.37499        0.61236          6.00          0.00250
  4 exp           0.56002        0.74834         12.73          0.00002
  5 dto           0.04978        0.22311          0.52          0.66920
  7 formacio      0.64749        0.80467         18.37          0.00000
```

El cuadro presenta un panorama detallado de las relaciones que cada variable mencionada en las instrucciones como independiente mantiene con todas las demás allí mencionadas

Si efectuáramos una regresión lineal múltiple con formacio como dependiente y edad, exp, y dto como independientes, la correlación múltiple entre la primera y las segundas sería 0'80467 y la correlación múltiple al cuadrado sería 0'64749. El análisis de varianza asociado a esa regresión (que fue explicado detalladamente con anterioridad) presentaría una razón F igual 18'37, con 3 g.l en el numerador y 30 en el denominador, cuyo nivel de significación sería 0'00001 (interpretación

El siguiente cuadro de resultados es de estructura y lectura semejante al anterior, pero ahora analiza las relaciones de cada variable mencionada en las instrucciones como dependiente con todas las variables mencionadas en las instrucciones como independientes. La nota explicativa dice " Correlación Múltiple al Cuadrado de cada variable dependiente con las variables independientes, y test de significación de la regresión múltiple. Los grados de libertad de los estadísticos F son 4 (en el numerador k-1) y 29 (en el denominador N-k).

```
SQUARED MULTIPLE CORRELATION OF EACH DEPENDENT VARIABLE
WITH THE INDEPENDENT VARIABLES
AND TESTS OF SIGNIFICANCE OF MULTIPLE REGRESSION
DEGREES OF FREEDOM FOR F-STATISTICS ARE     4 AND    29

                SQUARED
  VARIABLE      MULTIPLE       MULTIPLE       F-STATISTIC   SIGNIFICANCE
NO.     NAME    CORRELATION    CORRELATION                  (P LESS THAN)

  6 ventas       0.86829        0.93182        47.80          0.00000
  8 interes      0.13360        0.36551         1.12          0.36716
```

```
PARTIAL CORRELATIONS OF DEPENDENT VARIABLES AFTER REMOVING
LINEAR EFFECTS OF INDEPENDENT VARIABLES

                ventas    interes
                     6          8

ventas      6     1.000
interes     8     0.133     1.000
```

Correlaciones parciales entre las variables dependientes después de descontar de cada una de ellas los efectos lineales del conjunto de las variables independientes.

(Equivale a la correlación entre los residuales que quedarían de cada una de estas dependientes en las ecuaciones de regresión lineal múltiple en las que todas las independientes entraran como

La correlación entre ventas e interes "limpias del efecto lineal" sobre cada una de ellas de edad, experiencia, descuento y formación es de 0'133. Elevando al cuadrado ese valor tenemos que solo el 1'77% de la variabilidad de interes (excluido de esta variable el efecto de las independientes) puede atribuirse a las ventas (excluido de esta variable el efecto de las independientes).

Estas variables correlacionaban 0'364, de modo que el 13'25% de la variabilidad de las calificaciones en interes podía atribuirse a las ventas. Ahora se ve que, si se descuenta de cada una de estas variables el efecto de otras que las explican, la relación entre ellas "por sí" es

```
PROGRAM TERMINATED
```

Correlaciones y Covarianzas.

INSTRUCCIONES:

```
# Correlaciones y Covarianzas. Programa 8D.

/input
       case=15. Var=3. format=free.
/var
       names=descuent, ventas, benefic.
/print
       cova. #Solicita las covarianzas entre las variables.
             #El 8D calcula por defecto las correlaciones de Pearson.
/end
5 1 10
3 1 12
3 2 18
5 5 50
5 5 47
5 6 56
6 7 60
6 8 70
6 7 55
4 3 23
4 3 25
5 1  5
4 4 49
3 2 23
5 3 40
/end
```

RESULTADOS:

```
BMDP8D - CORRELATIONS WITH MISSING DATA
NUMBER OF CASES READ. . . . . .      15
```

PAIRWISE FREQUENCIES WITH BOTH VARIABLES PRESENT
(Tabla de Frecuencias conjuntas de las variables, es relevante solo si hay datos faltantes)

		descuent 1	ventas 2	benefic 3
descuent	1	15		
ventas	2	15	15	
benefic	3	15	15	15

Hay 15 casos con datos para ambas variables ventas y benefic

CORPAIR COVARIANCE MATRIX
(Matriz de varianzas- covarianzas)

		descuent 1	ventas 2	benefic 3
descuent	1	1.114		
ventas	2	1.800	5.552	
benefic	3	14.229	46.814	427.886

Varianza de la variable descuent

Covarianza entre las variables ventas y benefic

CORPAIR CORRELATION MATRIX
(Matriz de Correlaciones)

		descuent 1	ventas 2	benefic 3
descuent	1	1.0000		
ventas	2	0.7237	1.0000	
benefic	3	0.6516	0.9605	1.0000

Coefiente de correlación de Pearson entre las variables descuent y benefic

Regresión Lineal Simple.

INSTRUCCIONES:

```
#Regresion Lineal Simple (Síntesis de aspectos básicos). Programa 1R.

/input
       case=15. Var=3. format=free.
/var
       names=descuent, ventas, benefic.
/regress
       depend=benefic.  #Variable a utilizar como dependiente.
       indep=ventas.    #Variable a utilizar como independiente.
/end
5 1 10
3 1 12
3 2 18
5 5 50
5 5 47
5 6 56
6 7 60
6 8 70
6 7 55
4 3 23
4 3 25
5 1  5
4 4 49
3 2 23
5 3 40
/end
```

En regresión lineal simple R expresa el coeficiente de correlación de Pearson entre la variable dependiente e independiente, en este caso benefic y ventas respectivam.

RESULTADOS:

```
BMDP Program Output File: D:\EBMDP\E205_1.OUT
BMDP1R - LINEAR REGRESSION BY GROUPS
NUMBER OF CASES READ. . . . . .      15
```

Coefiente de determinación, es decir, Coef. de Corr. de Pearson elevado al cuadrado. El 92´25% de la varianza de los "beneficios" se explica por las "ventas".

```
MULTIPLE R            0.9605          STD. ERROR OF EST.          5.9773
MULTIPLE R-SQUARE     0.9225

ANALYSIS OF VARIANCE
                SUM OF SQUARES      DF     MEAN SQUARE       F RATIO     P(TAIL)
  REGRESSION        5525.9316       1       5525.9316       154.665      0.0000
  RESIDUAL           464.4682      13         35.7283

                                    STD.    STD. REG
  VARIABLE     COEFFICIENT          ERROR     COEFF      T      P(2 TAIL)  TOLERANCE

INTERCEPT            3.59862
ventas      2        8.4314        0.6780      0.96     12.44      0.00      1.0000
```

Ecuación de Regresión lineal simple:

Y' = a + b · X

Beneficios = 3'59862 + 8'4314 x Ventas

Expresa la Ecuación de Regresión Simple en puntuaciones directas.

En puntuaciones diferenciales el valor de "b" es el mismo y "a" siempre vale 0.

En la ecuación de Regresión lineal simple en puntuaciones típicas (es decir, estandarizada) "a" siempre vale 0 y "b" vale 0'96.