Descripción simultánea de varias variables con STATGRAPHICS CENTURION

Ficheros empleados: Cardata.sf3

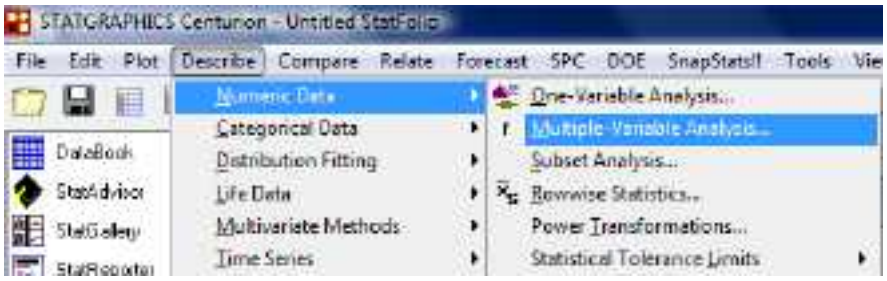
1. Introducción

El fichero Cardata.sf3 tiene datos de una muestra de vehículos. Entre las variables del fichero se encuentra el precio (price) de los vehículos. Queremos saber qué variables ayudan a explicar, o predecir, que un coche valga más o menos. Para ello construiremos un modelo de regresión múltiple que explique el precio de los coches. Las variables cuantitativas de este fichero que pueden ser de interés para explicar el precio de los coches son:

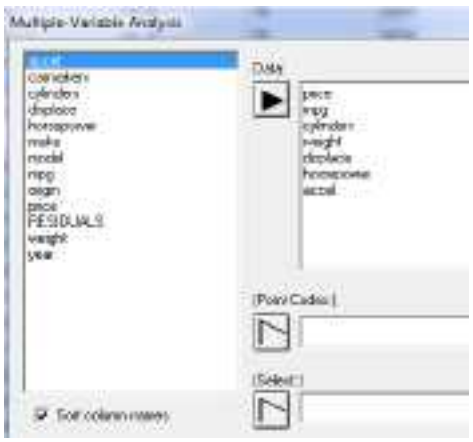
- mpg: millas recorridas por galón de combustible consumido.
- cylinders: número de cilindros del motor.
- weight: peso del vehículo (libras).
- displace: capacidad de los cilindros (pulgadas cúbicas).
- horsepower: potencia del motor.
- accel: tiempo que tarda en alcanzar la velocidad de 60 millas por hora

2. Gráficos XY

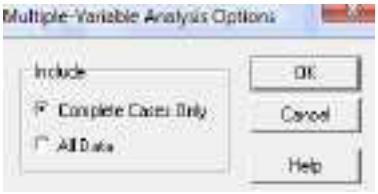
Una regresión múltiple recoge la relación lineal que hay entre una variable Xi y otra Y eliminando la influencia de otras variables explicativas. Por tanto, esa relación no es la misma que visualizamos en un gráfico XY. No obstante, es útil hacer un primer ejercicio de visualización de los datos dibujando el gráfico XY de cada regresor con Y. En estos gráficos podemos ya anticipar si puede haber relaciones más o menos fuertes, valores atípicos, no linealidades, etc. Una opción sencilla para hacer este tipo de gráficos con muchas variables es en:



y allí introducimos las variables:

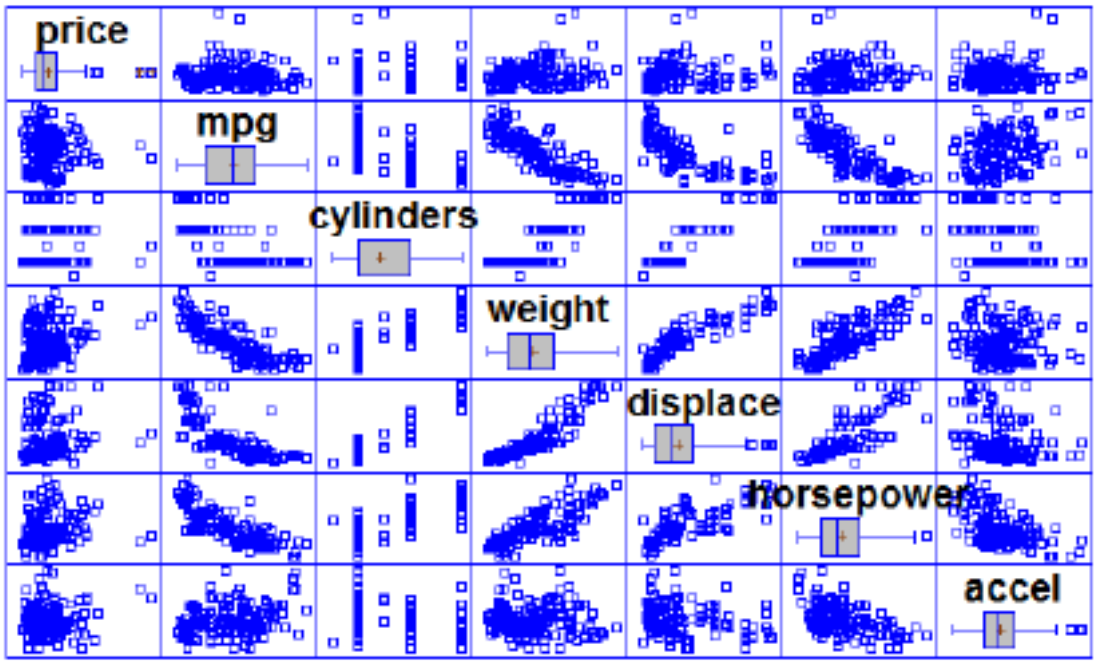


Aparece entonces la siguiente ventana,

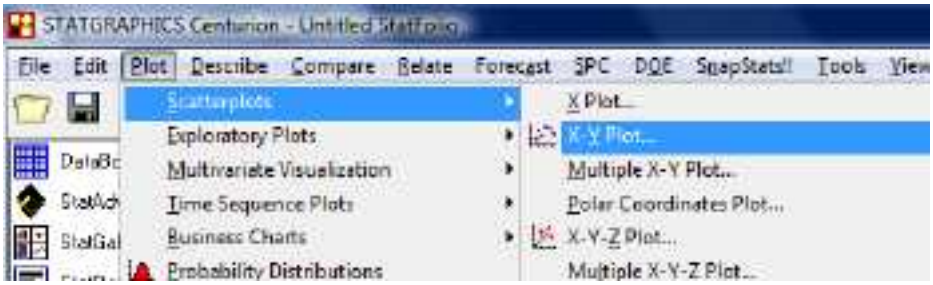


Si seleccionamos la primera opción, se eliminarán aquellas filas que tengan algún valor ausente en alguna de las variables. Si se selecciona la segunda opción, se usarán todos los datos posibles. Por ejemplo, si tenemos tres variables V1, V2 y V3, y en V1 falta el dato de la fila 12, bajo la opción primera se elimina esa fila a todos los efectos. Con la segunda opción, la fila 12 se utiliza en aquellos en aquellos cálculos en los que no intervenga V1; por ejemplo, para calcular la correlación entre V2 y V3.

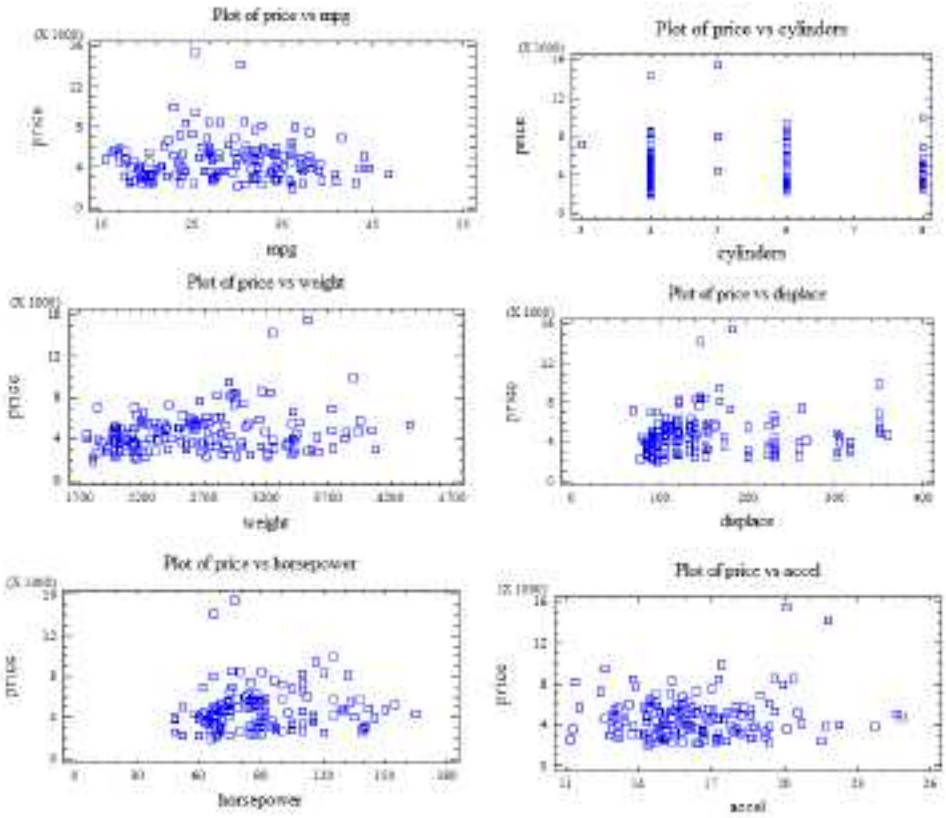
Obtenemos entonces la siguiente matriz de gráficos:



De esta matriz, lo que más nos interesa es la primera fila, que tiene en el eje Y la variable price, y en el X cada una del las restantes variables. Un aspecto destacable de estos gráficos es la presencia de dos puntos atípicos. Para ver mejor estas relaciones, haremos gráficos XY de cada par de variables.



Los 6 gráficos XY que buscamos son los siguientes:



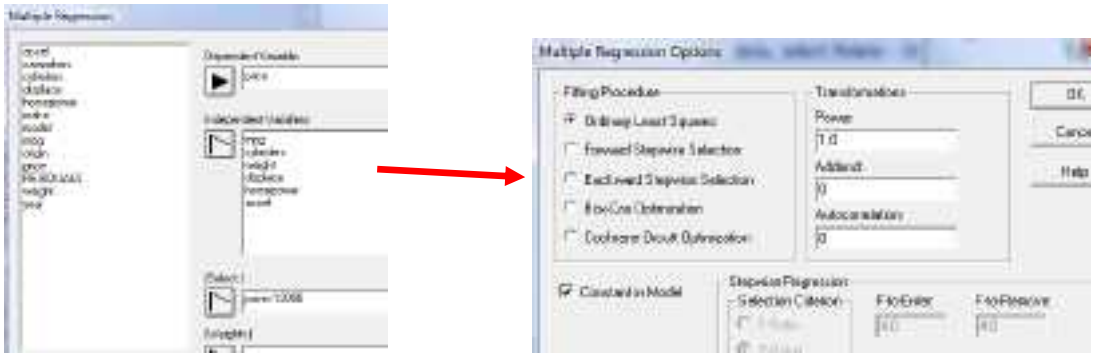
Lo más destacable de estos gráficos es:

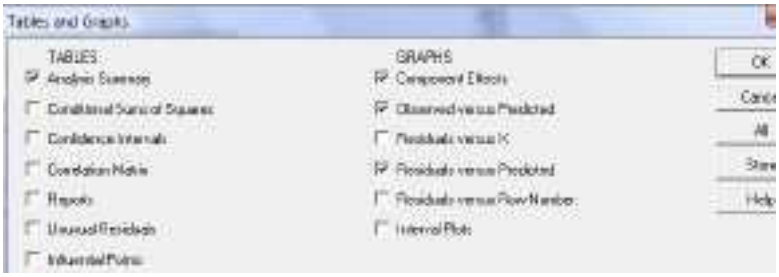
- Hay dos puntos anómalos. Se trata de dos coches de lujo cuyo perfil se aleja del resto de los coches y que muy posiblemente distorsione los resultados. Lo más conveniente sería eliminarlos.
- La influencia del resto de las variables aparece muy difusa. No hay ninguna variable que tenga una relación que sea fuerte. No parece que la regresión vaya a tener un R^2 muy elevado.

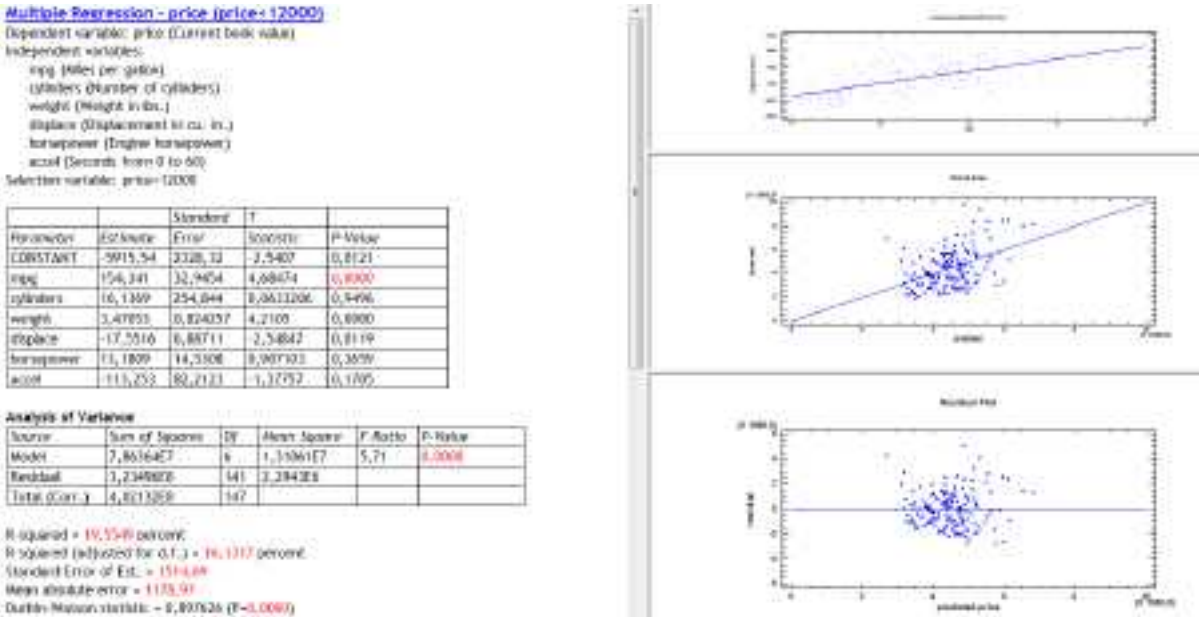
3. Regresión múltiple inicial

Como las variables estarán relacionadas, no es igual el resultado de una regresión simple que una regresión múltiple. Por tanto, lo más aconsejable es empezar con un modelo que contenga todas las variables (eliminando los dos datos atípicos). Posteriormente eliminaremos las variables que no resulten significativas. La eliminación de las variables significativas hay que hacerla eliminándolas una a una. Como el valor de los parámetros depende de qué variables están incluidas en el modelo, al eliminar una de ellas las estimaciones cambiarán. Por tanto, alguna variable que aparecía como no significativa puede pasar a serlo.

Al mismo tiempo, hay que mirar a los gráficos de residuos, por si se detectase alguna no linealidad que aconsejase realizar alguna transformación. La regresión con todas las variables se hace de la siguiente manera::



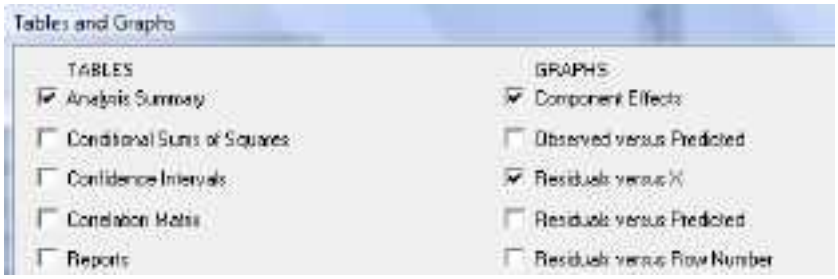


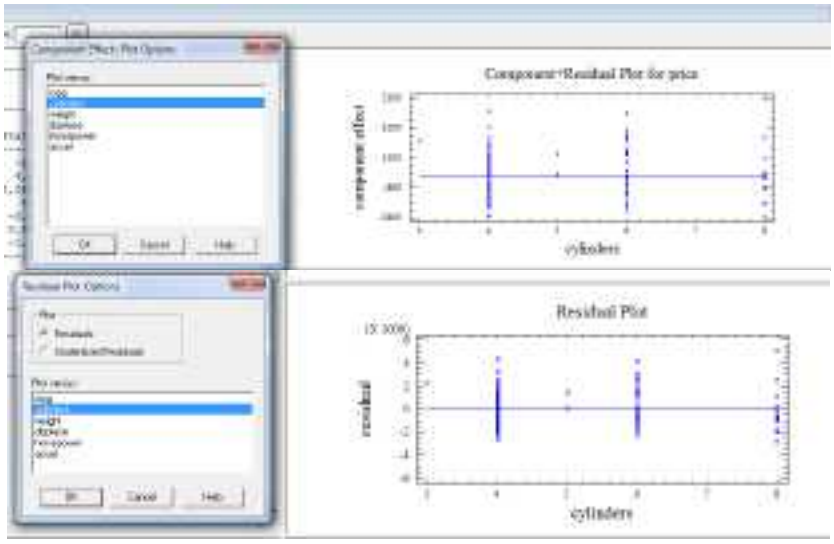


Lo primero que hay que mirar son los residuos (gráfico de residuos frente a valores previstos) , por si hay algún tipo de patrón que invalide la regresión. La interpretación de los residuos es igual que en regresión simple. En este caso los residuos no muestran nada relevante (en los gráficos de residuos, tenemos la opción de representar los “residuos” o los llamados “residuos estudentizados”. Elegiremos los primeros, pues son los únicos que hemos visto en clase). Los resultados de la regresión son por tanto válidos. Vemos que hay variables que no son significativas. Por lo que tendremos que eliminarlas.

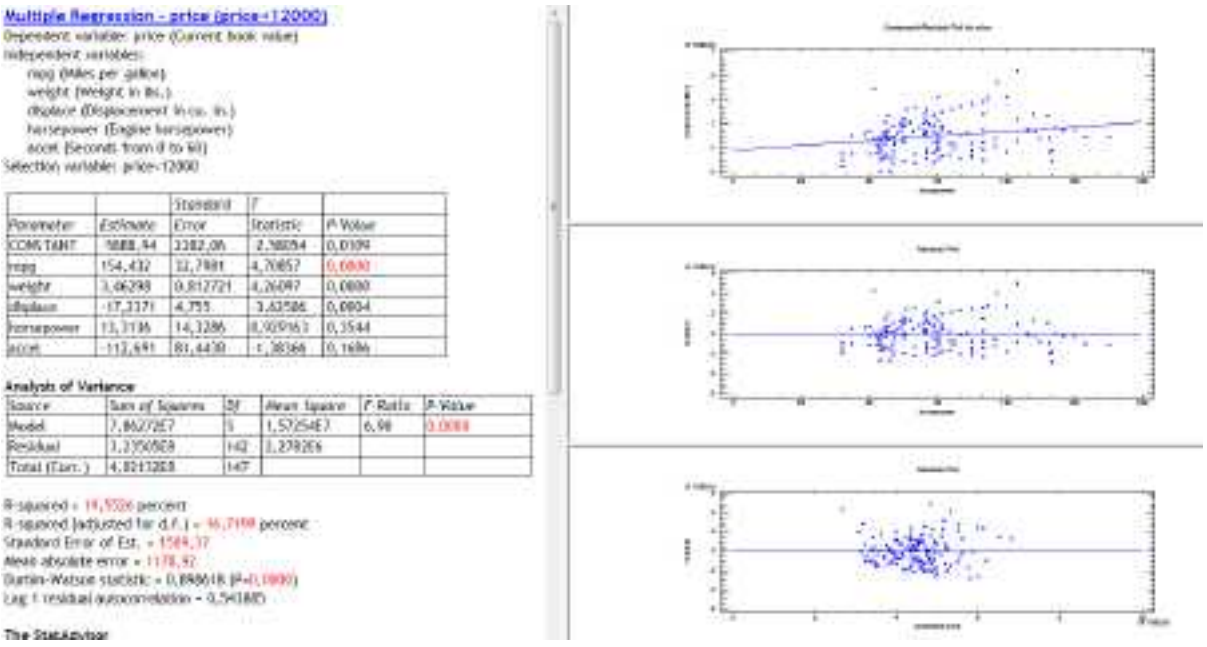
4. Eliminación de variables no significativas. Modelo final.

La variable menos significativa es el número de cilindros. Antes de eliminarla veamos si la razón de su escasa importancia es porque la verdadera relación sea no lineal. Para analizarlo hacemos el gráfico de residuos frente a mpg así como el gráfico de componentes de esta variable.





Ambos gráficos muestran que realmente esta variable no tiene ninguna contribución marginal al modelo. El nuevo modelo sin esta variable es:



Multiple Regression - price (price<12000)

Dependent variable: price (Current book value)

Independent variables:

mpg (Miles per gallon)

weight (Weight in lbs.)

displace (Displacement in cu. in.)

horsepower (Engine horsepower)

accel (Seconds from 0 to 60)

Selection variable: price<12000

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	5888.44	2382.08	2.48054	0.0139
mpg	154.432	32.7881	4.70857	0.0000
weight	3.06298	0.812721	4.26047	0.0000
displace	-17.2371	4.755	-3.62586	0.0004
horsepower	13.3136	14.3286	0.929163	0.3544
accel	-112.691	81.4430	-1.38366	0.1686

Analysis of Variance

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-Value
Model	7.8627267	5	1.57254E7	6.98	0.0000
Residual	1.2350503	142	2.229226		
Total (Corr.)	4.8213203	147			

R-squared = 19.5226 percent

R-squared (adjusted for d.f.) = 16.7149 percent

Standard Error of Est. = 1584.37

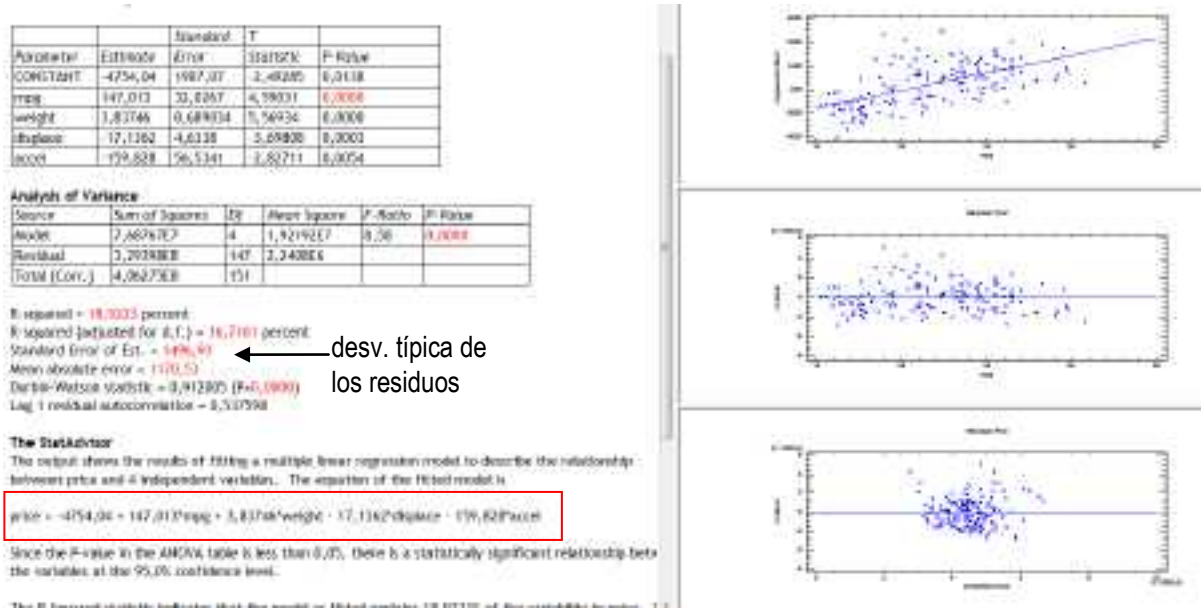
Mean absolute error = 1178.42

Durbin-Watson statistic = 0.886618 (P=0.0000)

Lag 1 residual autocorrelation = 0.540885

The StatAdvisor

Ahora eliminamos la variable horsepower, que es la variable que en este modelo es menos significativa. Los gráficos de residuos y el de componentes no muestran nada anómalo. Esta variable parece que no tiene ninguna aportación adicional para explicar el precio que no esté recogida en el resto de variables. El nuevo modelo es:



Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-4754,04	1987,07	-2,39205	0,0118
mpg	147,013	32,8267	4,47931	0,0000
weight	3,83746	0,689034	5,56934	0,0000
displace	-17,1362	4,6138	-3,71408	0,0003
accel	-159,828	56,5341	-2,82711	0,0054

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	2,46767E7	4	1,92192E7	8,58	0,0000
Residual	3,29348E8	147	2,24048E6		
Total (Corr.)	4,06273E8	151			

R-squared = 18,5333 percent
R-squared (adjusted for d.f.) = 16,3701 percent
Standard Error of Est. = 1496,93
Mean absolute error = 1120,52
Durbin-Watson statistic = 0,912805 (P=0,0000)
Lag 1 residual autocorrelation = 0,535598

desv. típica de los residuos

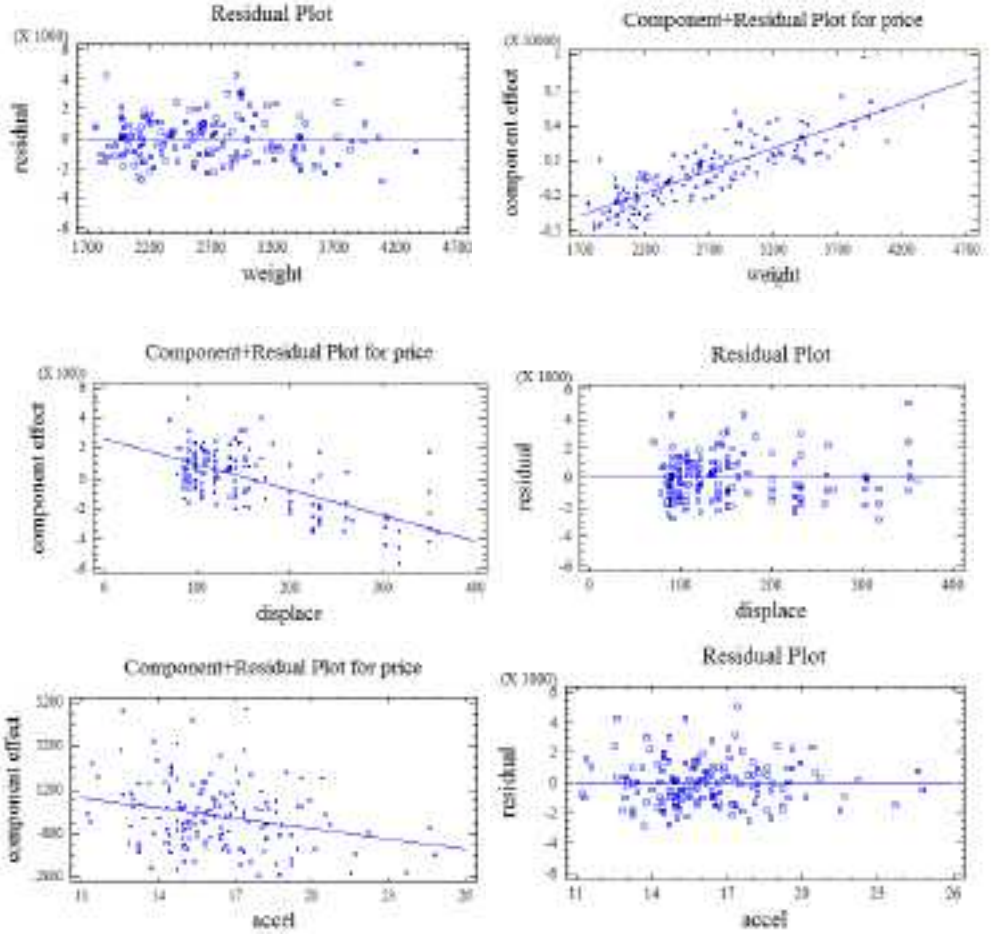
The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between price and 4 independent variables. The equation of the fitted model is

price = -4754,04 + 147,013*mpg + 3,83746*weight - 17,1362*displace - 159,828*accel

Since the P-value in the ANOVA table is less than 0,05, there is a statistically significant relationship bet the variables at the 95,0% confidence level.

Este modelo tiene ya todas las variables significativas (vemos que accel antes no era significativa y ahora sí). Explica sólo el 18.9% de la variabilidad de la variable precio. Los residuos frente a valores previstos no muestran ningún patrón destacable. Vemos que el gráfico de residuos y el de componentes para la variable mpg tampoco muestran falta de linealidad. Tenemos que analizar el resto de las variables:



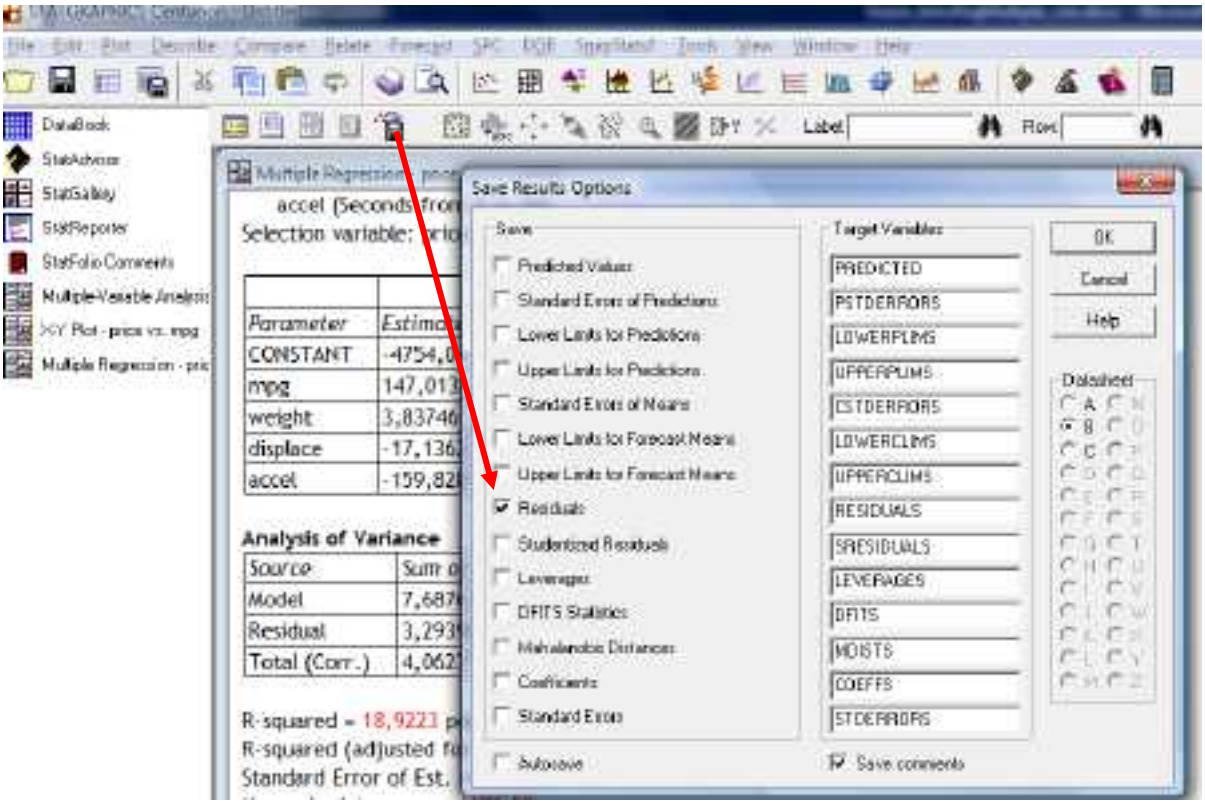
Estas figuras no muestran ningún problema evidente. La relación lineal parece adecuada. El modelo que explica el precio de los automóviles es:

```
price = -4754,04 + 147,013*mpg + 3,83746*weight - 17,1362*displace -159,828*accel
```

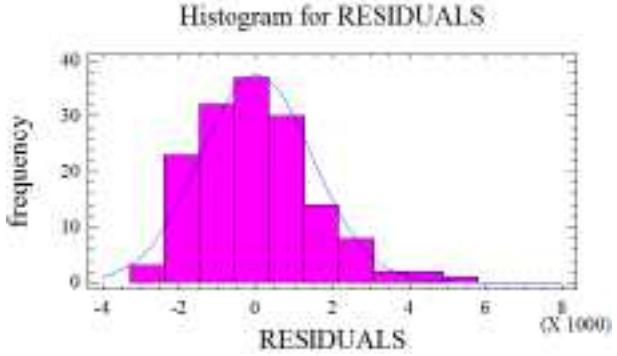
donde habría que añadir el término de error, que es una variable aleatoria de varianza estimada (varianza residual)=2.24. Del modelo estimado se concluye:

- A igualdad de consumo, peso, cilindrada y tiempo de aceleración, el número de cilindros y la potencia no ayudan a explicar el precio.
- A igual de de peso, cilindrada y tiempo de aceleración, los coches que menos consumen (más mpg) son más caros.
- A igualdad de consumo, cilindrada, y tiempo de aceleración, los coches más pesados, son más caros.
- A igualdad de consumo, peso y tiempo de aceleración, cuanto más grande son los cilindros, el coche es más barato. Este resultado no es muy intuitivo. Debe ser que entre dos coches con un motor con igual rendimiento, el que tenga ese rendimiento con un cilindro más pequeño es porque el motor es más eficiente: más válvulas, turboalimentado, inyectores electrónicos, etc.
- A igualdad de consumo, peso y cilindrada, a mayor tiempo de aceleración más barato es el coche.

Para terminar de hacer la diagnosis, analizaremos la normalidad de los residuos. Para ello primero salvamos sus valores:



De esta forma crearemos una nueva variable con los residuos, y podremos hacer análisis estadísticos con ella. El histograma de los residuos es:



Este histograma muestra que los residuos tienen una ligera asimetría, pero que se aproximan bastante a la normal. El p-valor del test de la chi-cuadrado es >0.10. Por tanto el modelo parece adecuado.

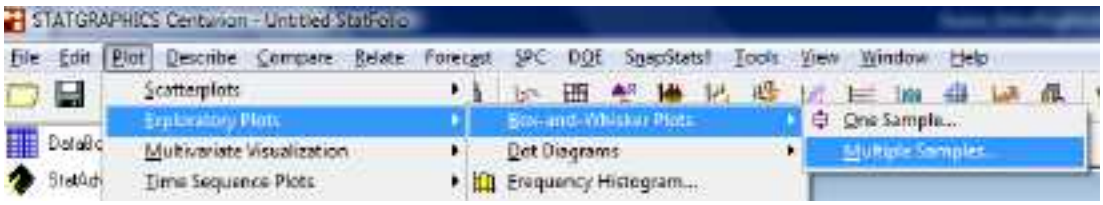
5. Regresión con variables binarias

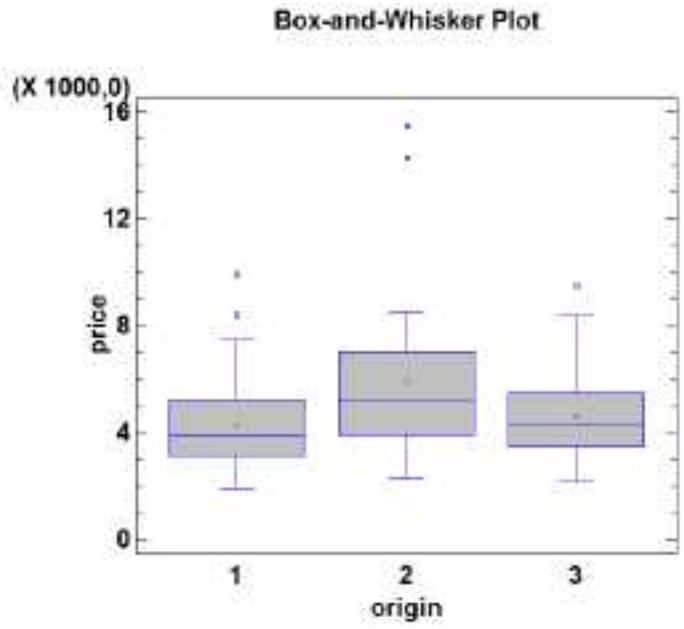
Una de las variables que se tienen en el fichero es el origen del coche (origin) y sus valores son:

- America=1,
- Europa=2,
- Japón=3.

Podemos analizar entonces si estos vehículos tienen un precio medio diferente según origen, y que no tenga que ver con las variables anteriores. Es decir, si dos coches iguales cuestan distinto por ser un coche americano, europeo o japonés.

En primer lugar hacemos un análisis descriptivo mediante box-plots:

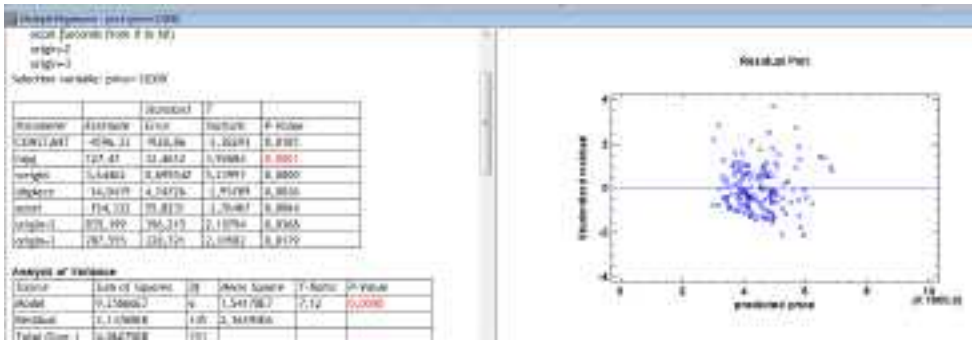




Los datos muestran un menor precio en los coches americanos, luego en los japoneses y finalmente los más caros son los europeos. Queremos saber si esas diferencias encontradas en la muestra son significativas, es decir, si son extrapolables a la población o son sólo producidas por el azar del muestreo. En caso afirmativo, querremos saber si las diferencias vienen explicadas ya por las variables de la regresión.

Tomaremos como referencia el precio de los coches americanos, que tienen una media muestral menor, e introduciremos dos variables binarias, una para los coches europeos (origin=2) y otra para los japoneses (origin=3), y miraremos si sus coeficientes son significativos.

El modelo que resulta es:



Vemos que las variables binarias son significativas: el precio de los coches es diferente según su origen y esas diferencias no parecen venir explicadas sólo por sus características mecánicas como peso, consumo, etc. Los coches más baratos, para un conjunto de características dadas, son los americanos, después los japoneses (en media cuestan 787.5 dólares más que los americanos a igualdad del resto de factores) y después los europeos (cuestan en media 835.199 dólares más que los americanos, a igualdad del resto de factores). Los residuos siguen siendo mostrando un patrón sin estructura.