Estadística II
Curso 2010/11

Guión de la Práctica 2
El modelo de regresión lineal y su tratamiento en Statgraphics

1. Contenidos de la práctica

2. Introducción

En esta práctica veremos una introducción al empleo de Statgraphics para el tratamiento de modelos de regresión lineal. Para motivar las distintas tareas que vamos a realizar, nos vamos a basar en la resolución de uno de los ejercicios del Tema 4 (el problema 4):

> "*Una gasolinera ha recogido información acerca de su recaudación diaria durante una semana, así como del número de clientes que acudieron a la misma en cada día:*
>
> | | | | | | | | |
> |---|---|---|---|---|---|---|---|
> | *Recaudación* | *1,5* | *10* | *8* | *3* | *5* | *15* | *2* |
> | *Número de clientes* | *3* | *6* | *5* | *3,5* | *4* | *8* | *3,2* |
>
> *a) Realizar un ajuste lineal que exprese la recaudación en función del número de clientes.*
> *b) Determinar cuál sería la recaudación media prevista para aquellos días en los que lleguen a la gasolinera 720 clientes (7,2). Obtener un intervalo de confianza al 95 % para dicha predicción.*
> *c) Determinar cuál sería la recaudación prevista para un día en el que lleguen a la gasolinera 720 clientes (7,2). Obtener un intervalo de confianza al 95 % para dicha predicción."*

El primer paso que debemos dar consiste en introducir en Statgraphics los datos anteriores. Una manera posible (ver el guión de la práctica 1) de hacer esto es entrar en Statgraphics y seleccionar "Cancel" en el "StatWizard" para salir del mismo. A continuación introducimos manualmente en dos columnas de una ventana en blanco los datos anteriores. En lo que sigue asumiremos que en "Col_1" se han introducido los datos de recaudación y en "Col_2" se han introducido los datos de número de clientes.

3. El modelo de regresión lineal simple

3.1 Estimación de los parámetros del modelo

Comenzamos calculando los valores de los parámetros del modelo de regresión lineal simple correspondiente a los valores de la muestra que hemos introducido. Para ello

seleccionamos el menú "Relate", y en el mismo escogemos "Simple Regression ..." e introducimos en el cuadro de diálogo "Col_1" para la variable Y y "Col_2" para la variable X, ya que parece razonable considerar el número de clientes como la variable independiente en este caso.

Relate → Simple Regression …
Y: Col_1 , X: Col_2

Tras pulsar OK, Statgraphics genera dos ventanas y en la ventana izquierda (la ventana de texto) obtenemos la salida siguiente:

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------------------------
Dependent variable: Col_1
Independent variable: Col_2
-------------------------------------------------------------------------------
                            Standard          T
Parameter       Estimate        Error       Statistic        P-Value
-------------------------------------------------------------------------------
Intercept       -6,31226       0,547767       -11,5236         0,0001
Slope             2,7121       0,110348        24,5778         0,0000
-------------------------------------------------------------------------------


                         Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio       P-Value
-------------------------------------------------------------------------------
Model                      145,156      1      145,156     604,07        0,0000
Residual                   1,20148      5     0,240296
-------------------------------------------------------------------------------
Total (Corr.)              146,357      6

Correlation Coefficient = 0,995887
R-squared = 99,1791 percent
Standard Error of Est. = 0,4902
```

La ventana derecha muestra un gráfico de dispersión con el modelo ajustado.

En la salida anterior podemos encontrar los valores de los diferentes parámetros del modelo (indicados en rojo). En particular, obtenemos las siguientes estimaciones

```
Intercepto                          -6,31226
Pendiente                            2,7121
Varianza de los errores              0,240296
```

El modelo de regresión estimado será

$$y = -6,31226 + 2,7121\ x$$

Observamos que en la salida anterior aparece también el valor de R^2, el coeficiente de determinación. Este valor nos indica la variación en la recaudación que se puede explicar si se conoce el número de clientes. En este caso el valor es igual al 99,18% (valor en azul en la salida), esto es, es un valor muy elevado y el ajuste del modelo a los datos es muy bueno, como se puede verificar en el gráfico de dispersión (ventana derecha).

3.2 Intervalos de confianza y contrastes para los parámetros del modelo

Un paso importante a dar una vez estimado el modelo es contrastar el cumplimiento de las hipótesis de dicho modelo de regresión lineal. En esta práctica vamos a dejar

este paso para más adelante (Diagnosis del modelo), y seguiremos con el orden de las de las materias visto en clase, que se corresponde con el orden de preguntas en el ejercicio. Supondremos por el momento que dichas hipótesis se cumplen razonablemente.

Supongamos que queremos contrastar al 5% de significación si la pendiente del modelo es diferente de cero. Para ello podemos calcular el p-valor asociado al estadístico t para la pendiente, visto en clase. Dicho valor aparece en la tabla de resultados generada por Statgraphics, y está indicado a continuación en rojo.

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Col_1
Independent variable: Col_2
------------------------------------------------------------------------------
                            Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept       -6,31226       0,547767       -11,5236         0,0001
Slope             2,7121       0,110348        24,5778         0,0000
------------------------------------------------------------------------------


                          Analysis of Variance
------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                      145,156      1      145,156     604,07       0,0000
Residual                   1,20148      5     0,240296
------------------------------------------------------------------------------
Total (Corr.)              146,357      6

Correlation Coefficient = 0,995887
R-squared = 99,1791 percent
Standard Error of Est. = 0,4902
```

Este p-valor es el correspondiente a un contraste bilateral con hipótesis nula que establece que la pendiente sea igual a cero. Por tanto, en este caso dado que el p-valor vale 0,0000, rechazamos dicha hipótesis nula al 5%, y concluimos que tenemos suficiente evidencia para considerar que la pendiente es distinta de cero.

Supongamos ahora que estamos interesados en obtener un intervalo de confianza para esta pendiente del modelo de regresión (para un nivel de confianza del 95%, por ejemplo). Este intervalo se puede calcular aplicando las fórmulas vistas en clase, dado que disponemos de la estimación del parámetro y la varianza residual, y solo tendríamos que calcular la varianza de la variable independiente y el cuantil de la distribución t. Esto lo haríamos siguiendo los procedimientos vistos en la primera práctica.

Una manera más sencilla de calcular dichos intervalos en Statgraphics consiste en realizar una regresión múltiple, esto es, seleccionar "Multiple Regression ..." en el menú "Relate",

Relate → Multiple Regression …
Dependent Variable: Col_1
Independent Variables: Col_2

Tras pulsar OK obtenemos dos ventanas y la siguiente salida en la ventana izquierda

```
Multiple Regression Analysis
------------------------------------------------------------------------------
Dependent variable: Col_1
```

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-6,31226	0,547767	-11,5236	0,0001
Col_2	2,7121	0,110348	24,5778	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	145,156	1	145,156	604,07	0,0000
Residual	1,20148	5	0,240296		
Total (Corr.)	146,357	6			

R-squared = 99,1791 percent
R-squared (adjusted for d.f.) = 99,0149 percent
Standard Error of Est. = 0,4902
Mean absolute error = 0,358632
Durbin-Watson statistic = 2,19937

con los mismos valores de los parámetros que en el caso anterior. Haciendo esto tenemos la ventaja de que al pulsar el icono "Tabular Options" podemos marcar la casilla "Confidence Intervals" en el cuadro de diálogo

(I) Tabular options → Confidence Intervals → OK

para obtener la salida

95,0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	-6,31226	0,547767	-7,72034	-4,90417
Col_2	2,7121	0,110348	2,42845	2,99576

y el intervalo de confianza al 95% para la pendiente corresponde a los valores indicados en rojo.

Si se quisiera cambiar el nivel de confianza, se pulsaría el botón derecho del ratón en la ventana izquierda y seleccionaríamos "Pane Options".

3.3 Estimación para respuestas promedio y predicciones

Describimos a continuación la manera de responder a las preguntas b) y c) del ejercicio introducido al comienzo de la práctica. En particular, queremos calcular una estimación puntual y un intervalo de confianza para la recaudación media prevista para aquellos días en los que lleguen a la gasolinera 720 clientes (x_0 = 7,2), y para la recaudación prevista para un día en el que lleguen a la gasolinera 720 clientes.

Queremos por tanto llevar a cabo inferencia sobre respuestas promedio y predicciones. Podemos emplear los estadísticos vistos en clase, al final del Tema 4. Alternativamente, y directamente desde Statgraphics, con los datos introducidos obtenemos el modelo de regresión mediante

Relate → Simple Regression …
Y: Col_1 , X: Col_2

Una vez generadas las ventanas de resultados, pulsamos el icono "Tabular Options" y marcamos la casilla "Forecasts" para obtener las predicciones y los intervalos de confianza deseados. Para modificar el valor de la variable independiente a 7,2, pulsamos con el botón derecho en la ventana izquierda, seleccionamos "Pane Options" e introducimos en el campo de la ventana emergente "Forecast at X:" el valor deseado.

(I) Tabular Options → Forecasts
(BD) Pane Options → Forecast at X: → 7,2

El resultado es el siguiente:

Predicted Values

		95,00%		95,00%	
	Predicted	Prediction Limits		Confidence Limits	
X	Y	Lower	Upper	Lower	Upper
7,2	13,2149	11,6887	14,741	12,3539	14,0759

El valor predicho correspondiente a las dos preguntas es el mismo, 13,2149. Sin embargo, obtenemos dos intervalos: uno para la predicción puntual, correspondiente a los valores

[11,6887 , 14,741]

y otro para la estimación de la respuesta promedio, dado por

[12,3539 , 14,0759]

Estos valores proporcionan las respuestas a los apartados c) y b) respectivamente (ya que por defecto están calculados para un nivel de confianza del 95%).

3.4 Diagnosis del modelo

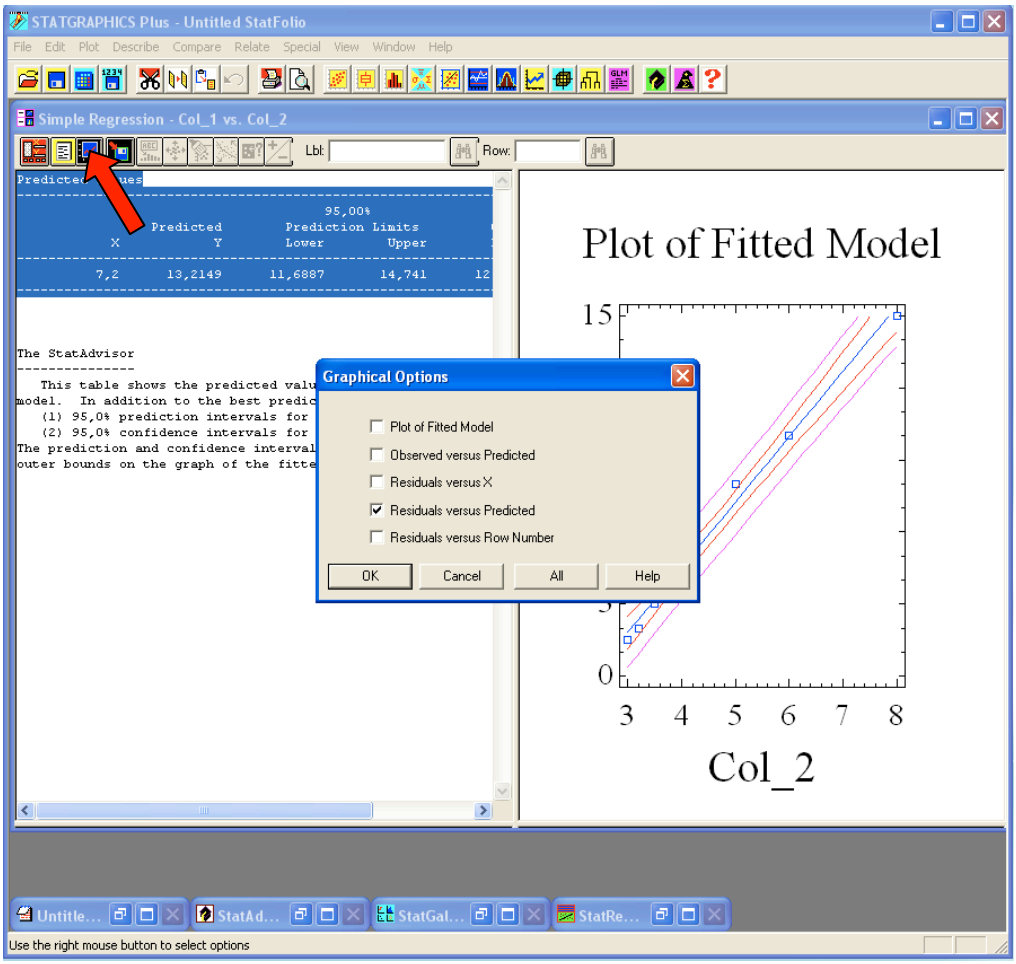
En este apartado retomamos la valoración del cumplimiento de las hipótesis del modelo, que dejamos pendiente anteriormente. Una manera de llevar a cabo este análisis, al menos de forma gráfica, consiste en generar un diagrama de los residuos, y verificar en dicho diagrama las propiedades de los mismos.

Para obtener los diagramas de residuos, generamos en primer lugar el modelo de regresión,

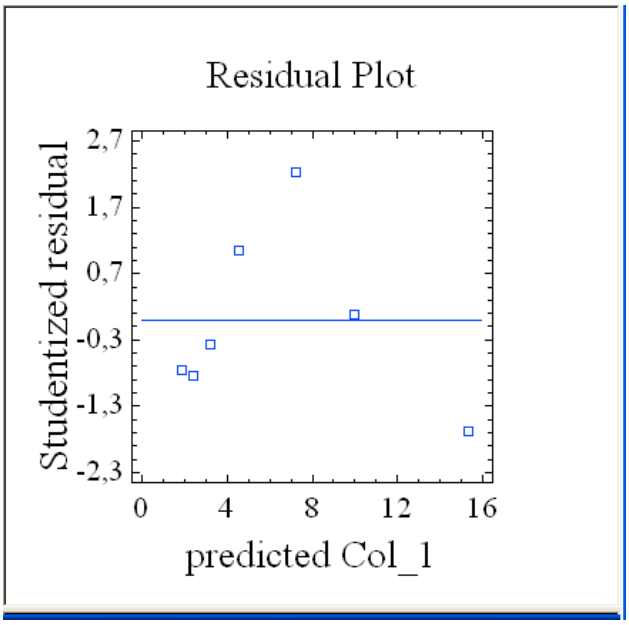
Relate → Simple Regression …
Y: Col_1 , X: Col_2

y a continuación seleccionamos el botón "Graphical Options" (el tercer botón en la parte superior), y marcamos la caja "Residuals versus Predicted", como se indica en la imagen siguiente.

(I) Graphical Options → Residuals versus Predicted → OK



El gráfico resultante (que aparece en la ventana derecha generada por Statgraphics) presenta los valores de los residuos "estudentizados" (esto es, divididos por la estimación de su desviación típica, s_R) frente a los valores predichos para la variable dependiente (la recaudación) y se incluye a continuación.



Para estudiar el cumplimiento de las hipótesis del modelo podemos comprobar si en este gráfico los valores de los residuos no muestran estructura. Si se cumplen dichas hipótesis, los valores de estos residuos deben corresponder a valores independientes obtenidos de una variable aleatoria normal con media cero y desviación típica igual a uno.

En este caso por ejemplo podemos observar una cierta estructura no lineal en los mismos, y podríamos dudar del cumplimiento de la condición de linealidad en la relación entre las variables.

4. Relaciones no lineales y linealización

En algunos casos los datos de que disponemos parecen seguir una relación no lineal, pero podemos aplicar una transformación no lineal que haga que los datos transformados se aproximen mejor a través de una relación lineal.

Para ilustrar el uso de estas transformaciones en Statgraphics empleamos algunos de los datos del fichero "undata.sf3" (datos de la UNECE Statistical Database). Almacenamos dichos datos en una carpeta adecuada, y a continuación los leemos en Statgraphics empleando

File → Open→ Open Data File …

Vamos a estudiar la relación entre las variables correspondientes al GDP per capita y la mortalidad infantil (Infant_Mortality_Rate). Para ello seleccionamos

Relate → Simple Regression …
Y: Infant_Mortality_Rate , X: GDP_per_Capita

Los resultados obtenidos se pueden ver en la salida siguiente:

```
Regression Analysis - Linear model: Y = a + b*X
-------------------------------------------------------------------------------
Dependent variable: Infant_Mortality_Rate
Independent variable: GDP_per_Capita
-------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate         Error       Statistic        P-Value
-------------------------------------------------------------------------------
Intercept        11,3761       0,941033         12,089         0,0000
Slope       -0,000184079   0,0000312106       -5,89798         0,0000
-------------------------------------------------------------------------------


                          Analysis of Variance
-------------------------------------------------------------------------------
Source             Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------
Model                     386,703      1      386,703      34,79       0,0000
Residual                   489,13     44      11,1166
-------------------------------------------------------------------------------
Total (Corr.)             875,833     45

Correlation Coefficient = -0,664474
R-squared = 44,1526 percent
Standard Error of Est. = 3,33415
```

En particular, el valor de R^2 es 0.44, que no es demasiado elevado.

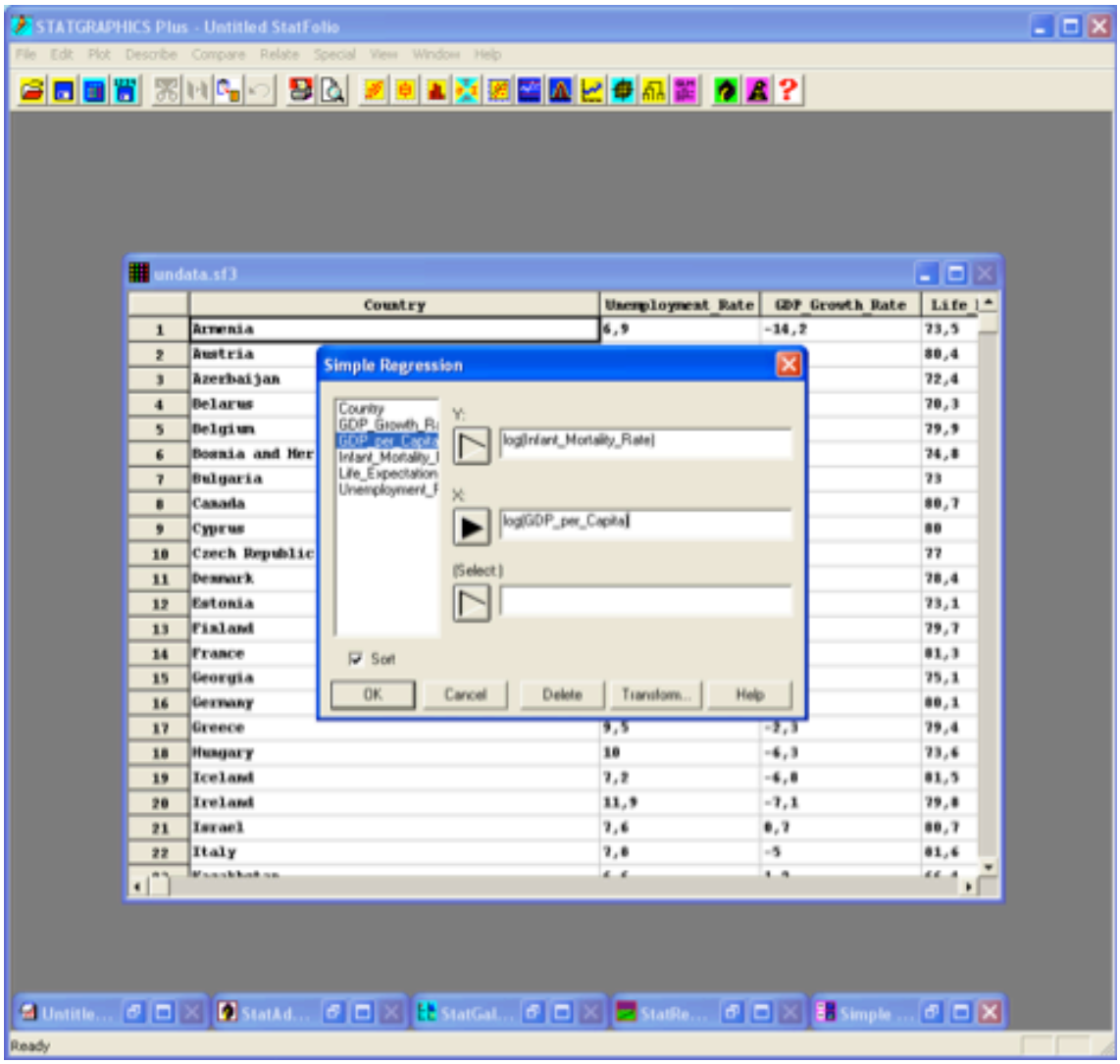
Vamos a repetir este ejercicio, pero ahora llevaremos a cabo la regresión empleando los logaritmos de las mismas variables, intentando atenuar el efecto no lineal asociado a los valores mayores de ambas variables. Para ello, seleccionamos

Relate → Simple Regression …

y al introducir la definición de las variables escribimos directamente la forma de la transformación no lineal que queremos llevar a cabo. En este caso, en la ventana correspondiente de Statgraphics escribimos "log(Infant_Mortality_Rate)" para Y y "log(GDP_per_Capita)" para X.

Y: log(Infant_Mortality_Rate) , X: log(GDP_per_Capita)

La siguiente captura de pantalla muestra como debe quedar la ventana de selección de variables:



Una vez pulsado "OK", obtenemos como resultado

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: log(Infant_Mortality_Rate)
Independent variable: log(GDP_per_Capita)
```

```
-------------------------------------------------------------------------------
                                 Standard          T
Parameter       Estimate          Error       Statistic        P-Value
-------------------------------------------------------------------------------
Intercept        8,26339       0,743168         11,1191         0,0000
Slope          -0,659579      0,0746038         -8,8411         0,0000
-------------------------------------------------------------------------------


                           Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
-------------------------------------------------------------------------------
Model                      10,3666      1      10,3666      78,17       0,0000
Residual                   5,83546     44     0,132624
-------------------------------------------------------------------------------
Total (Corr.)               16,202     45

Correlation Coefficient = -0,799895
R-squared = 63,9831 percent
Standard Error of Est. = 0,364176
```

y ahora el valor de R^2 pasa a ser 0.64, mejorando el valor anterior.

5. Regresión lineal múltiple

Comentamos muy brevemente en este apartado como ajustar modelos de regresión lineal múltiple utilizando Statgraphics. Lo haremos utilizando datos del fichero "cardata.sf3", disponibles en la página Web de prácticas de la asignatura.

Leemos dichos datos (después de haberlos descargado a una carpeta adecuada) seleccionando

File → Open→ Open Data File …

e indicando la carpeta y el nombre del archivo deseado ("cardata.sf3").

Supongamos que deseamos calcular los parámetros de un modelo de regresión para explicar los valores de la variable "accel" (aceleración) en función de la variable "horsepower" (potencia). Para ello seleccionamos

Relate → Multiple Regression …
Dependent Variable: accel
Independent Variables: horsepower

y obtenemos como resultado la siguiente salida

```
Multiple Regression Analysis
-------------------------------------------------------------------------------
Dependent variable: accel
-------------------------------------------------------------------------------
                                    Standard          T
Parameter                Estimate       Error       Statistic        P-Value
-------------------------------------------------------------------------------
CONSTANT                  20,7949    0,679216         30,6161         0,0000
horsepower             -0,0509812  0,00736125        -6,92562         0,0000
-------------------------------------------------------------------------------

                           Analysis of Variance
-------------------------------------------------------------------------------
Source              Sum of Squares     Df  Mean Square    F-Ratio      P-Value
```

Model	232,566	1	232,566	47,96	0,0000
Residual	722,463	149	4,84874		
Total (Corr.)	955,029	150			

R-squared = 24,3517 percent
R-squared (adjusted for d.f.) = 23,844 percent
Standard Error of Est. = 2,20199
Mean absolute error = 1,73477
Durbin-Watson statistic = 1,52416

y el modelo

$$accel = 20,7949 - 0,0509812 \times horsepower$$

donde ambos coeficientes son significativos (ver sus p-valores) y el valor de R^2 es igual a 23,84% (un valor relativamente bajo).

Si ahora añadimos una nueva variable explicativa, y construimos el modelo para "accel" en función de "horsepower" y la nueva variable "weight" (peso)

Relate → Multiple Regression …
Dependent Variable: accel
Independent Variables: horsepower , weight

obtenemos el resultado siguiente

Multiple Regression Analysis

Dependent variable: accel

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	16,9714	0,594887	28,5287	0,0000
horsepower	-0,13636	0,00912925	-14,9366	0,0000
weight	0,00426685	0,000369447	11,5493	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	575,037	2	287,518	111,98	0,0000
Residual	379,992	148	2,56751		
Total (Corr.)	955,029	150			

R-squared = 60,2114 percent
R-squared (adjusted for d.f.) = 59,6738 percent
Standard Error of Est. = 1,60235
Mean absolute error = 1,19181
Durbin-Watson statistic = 1,95717

El modelo es ahora

$$accel = 16,9714 - 0,13363 \times horsepower + 0,00426685 \times weight$$

donde los tres coeficientes son significativos y el valor de R^2 es igual a 59,67% (bastante más elevado que para el modelo anterior).

6. *Ejercicio de aplicación*

En la página Web de la asignatura y para la práctica 1 tienes disponible un conjunto de datos denominado Datos_1, "cardata.sf3". Sobre este conjunto de datos se pide que realices las tareas siguientes:

I. *Calcula los coeficientes del modelo de regresión lineal simple para las variables "weight" (variable dependiente) en función de "horsepower" (variable independiente) (Resultados. Intercepto: 896,111, Pendiente: 20,0097, R^2: 65,57%)*
II. *Determina si el valor de la pendiente es significativo (Respuesta: sí ¿Por qué?)*
III. *Calcula un intervalo de confianza al 98% para el valor de la pendiente (Respuesta: [17,2164 ; 22,803])*
IV. *Da una estimación para el valor del peso del vehículo correspondiente a una potencia de 130 hp e indica un intervalo de confianza para dicha predicción al 99% (Respuestas: 3497,37 y [2558,57 ; 4436,17])*
V. *Genera el gráfico de los residuos frente a los valores predichos por el modelo y comenta sobre el cumplimiento de las condiciones del mismo.*