Práctica 5

Vamos a analizar con STATGRAPHICS el problema de estimación, descomposición de la variabilidad, contraste de regresión, intervalos de confianza para los parámetros, y predicciones, sobre un modelo de regresión lineal múltiple. Para ello, consideraremos las observaciones sobre 8 provincias españolas de un vector tridimensional (X_1, X_2, Y), donde la variable Y está dada por un indicador global del consumo, y las variables X_1 y X_2 representan, respectivamente, el número de automóviles y teléfonos por cada 1000 habitantes. Dichas observaciones, almacenadas en el fichero **P5.SF3**, se recogen en la siguiente tabla:

Ind. Global	Automóviles	Teléfonos
64	58	111
78	84	131
83	78	158
88	81	147
89	82	121
99	102	165
101	85	174
102	102	169

Vamos a trabajar con el modelo de regresión de Y sobre X_1 y X_2.

Estimación y contrastes de nulidad sobre los parámetros del modelo. Tabla ANOVA y contraste de regresión.

Debemos ejecutar la opción **RELATE/MULTIPLE REGRESSION**, y entonces establecer la variable Y como dependiente, y las variables X_1 y X_2 como independientes. Al hacer click sobre el botón OK se obtiene la siguiente salida:

```
Multiple Regression Analysis
-----------------------------------------------------------------------------
Dependent variable: Y
-----------------------------------------------------------------------------
                            Standard          T
Parameter      Estimate         Error       Statistic    P-Value
-----------------------------------------------------------------------------
CONSTANT        9.05385        14.9631      0.605079      0.5715
```

```
X1              0.520279       0.231662       2.24585      0.0747
X2              0.239746       0.137752       1.74042      0.1423
------------------------------------------------------------------------------

                           Analysis of Variance
------------------------------------------------------------------------------
Source     Sum of Squares   Df   Mean Square    F-Ratio   P-Value
------------------------------------------------------------------------------
Model          1011.65       2     505.827       14.34     0.0085
Residual       176.345       5     35.2691
------------------------------------------------------------------------------
Total (Corr.)   1188.0       7

R-squared = 85.1561 percent

R-squared (adjusted for d.f.) = 79.2185 percent

Standard Error of Est. = 5.93878

Mean absolute error = 3.96034

Durbin-Watson statistic = 1.94516


The StatAdvisor
---------------
   The output shows the results of fitting a multiple linear
regression model to describe the relationship between Y and 2
independent variables.  The equation of the fitted model is

Y = 9.05385 + 0.520279*X1 + 0.239746*X2

Since the P-value in the ANOVA table is less than 0.01, there is a
statistically significant relationship between the variables at
the 99% confidence level.

     The R-Squared statistic indicates that the model as fitted
explains 85.1561% of the variability in Y.  The adjusted
R-squared statistic, which is more suitable for comparing models
with different numbers of independent variables, is 79.2185%.
The standard error of the estimate shows the standard deviation of
the residuals to be 5.93878.  This value can be used to construct
prediction limits for new observations by selecting the Reports
option from the text menu. The mean absolute error (MAE) of
3.96034 is the average value of the residuals.  The Durbin-Watson
(DW) statistic tests the residuals to determine if there is any
significant correlation based on the order in which they occur in
```

```
your data file.  Since the DW value is greater than 1.4, there is
probably not any serious autocorrelation in the residuals.

   In determining whether the model can be simplified, notice that the
highest P-value on the independent variables is 0.1423, belonging
to X2.  Since the P-value is greater or equal to 0.10, that
term is not statistically significant at the 90% or higher confidence
level.  Consequently, you should consider removing X2 from the
model.
```

The StatAdvisor da una ligera explicación de los resultados de mayor relevancia. La explicación detallada debe ser establecida por el alumno. Toda la información obtenida se interpreta como sigue:

Modelo estimado. Contrastes $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$ y $H_0 : \beta_2 = 0$.

La primera tabla muestra los coeficientes de regresión estimados, en la columna **Estimate**; la estimación de la desviación típica de los coeficientes de regresión, en la columna **Standard Error**; los valores del estadístico t_{exp} para contrastar si los verdaderos coeficientes de regresión son nulos, en la columna **T Statistic**; y los P-valores asociados a los contrastes anteriores, esto es, la probabilidad de encontrar un valor de una T_{n-3} mayor o igual en valor absoluto que $|t_{exp}|$, en la columna **P-value**. Resulta

$$\hat{\beta}_0 = \widehat{CONSTANT} = 9.05385\ , \quad \widehat{SE(\beta_0)} = 14.9631\ , \quad T_{exp} = 0.605079\ para\ \ H_0:\ \beta_0 = 0$$
$$\hat{\beta}_1 = \widehat{X1} = 0.520279\ , \quad \widehat{SE(\beta_1)} = 0.231662\ , \quad T_{exp} = 2.24585\ para\ \ H_0:\ \beta_1 = 0$$
$$\hat{\beta}_2 = \widehat{X2} = 0.239746 \quad \widehat{SE(\beta_2)} = 0.137752\ , \quad T_{exp} = 1.74042\ para\ \ H_0:\ \beta_2 = 0$$

Entonces, el plano de regresión estimado está dada por

$$\widehat{Y} = 9.05385 + 0.520279 * X_1 + 0.239746 * X_2\ .$$

Así, cuando las variables X_1 y X_2 tomen el valor 0, predeciremos el valor medio de la variable Y por $\widehat{Y} = 9.05385$. Además, por cada unidad que aumentemos la variable X_1, manteniendo constante X_2, la variable Y aumentará por término medio en 0.520279 unidades; y, por cada unidad que aumentemos la variable X_2, manteniendo constante X_1, la variable Y aumentará por término medio en 0.239746 unidades.

Fijado un nivel de significación $\alpha = 0.05$ se concluye, dado que todos los P-valores son mayores que α, que hay evidencia empírica para admitir que cualquiera de los coeficientes de regresión puedan ser nulos. Notemos que dichos contrastes se hacen por separado, y que esto no significa que simultáneamente dos cualesquiera de los parámetros, o los tres, sean nulos. De hecho, para admitir o rechazar que Y se relaciona de forma lineal con X_1 y X_2 debemos resolver el contraste de regresión $H_0 : R^2 = 0$ mediante la tabla ANOVA, o equivalentemente, resolver el contraste $H_0 : \beta_1 = 0, \beta_2 = 0$.

Valores del P-valor muy próximos a α con un número reducido de datos, como ocurre en este caso con el P-valor para β_1 que resulta ser 0.0747 y $\alpha = 0.05$, requieren aumentar el tamaño muestral para confirmar si se acepta o rechaza la hipótesis nula.

Tabla ANOVA. Contraste $H_0 : R^2 = 0$.

La lectura de la tabla ANOVA nos indica que la variabilidad total, **VT = Total Sum of Squares = 1188**, se descompone en variabilidad explicada, **VE = Model Sum of Squares = 1011.65** y no explicada o residual, **VNE = Residual Sum of Squares = 176.345**. El estadístico de contraste $\mathbf{F}_{exp}$ toma el valor 14.34. Nuevamente el **P-value** asociado a este contraste representa la probabilidad de encontrar un valor de una $F_{1,n-3}$ mayor o igual que F_{exp}. Puesto que el P-valor asociado es $0.0085 < \alpha = 0.05$, entonces debemos rechazar $H_0 : R^2 = 0$, esto es, no hay evidencia empírica para rechazar una relación lineal entre las variables.

La tabla ANOVA también proporciona el valor de la varianza residual, dada por

$$\widehat{\sigma}^2 = Residual\ Mean\ Square = 35.2691\ .$$

Su raiz cuadrada, esto es, la estimación por mínimos cuadrados de la desviación típica σ de los errores, resulta ser

$$\widehat{\sigma} = Standard\ Error\ of\ Est. = 5.93878\ .$$

El valor del coeficiente de determinación $R^2 =$ **R-squared =** 85.1561 indica que el 85.1561 % de la variabilidad total de la variable dependiente es explicada por el plano de regresión.

Intervalos de confianza para los parámetros del modelo.

Debemos ejecutar el icono **Tabular options** que aparece debajo de la barra del título de la ventana **Multiple Regression**. Aparece así, un menú con casillas de verificación. Por defecto está activada la casilla correspondiente a la opción **Analysis Summary**, que produce la salida descrita en la sección anterior. Con el ratón debemos activar además la casilla de verificación correspondiente a la opción **Confidence Intervals**. Al hacer click sobre el botón *OK* se obtiene la siguiente salida:

95.0% confidence intervals for coefficient estimates

Parameter	Estimate	Standard Error	Lower Limit	Upper Limit
CONSTANT	9.05385	14.9631	-29.4101	47.5178
X1	0.520279	0.231662	-0.0752289	1.11579

```
X2            0.239746     0.137752     -0.114357       0.593849
-------------------------------------------------------------------------------


The StatAdvisor
---------------
   This table shows 95.0% confidence intervals for the coefficients in
the model.  Confidence intervals show how precisely the
coefficients can be estimated given the amount of available data
and the noise which is present.
```

La tabla anterior muestra de nuevo en sus dos primeras columnas los coeficientes de regresión estimados, en la columna **Estimate**, y la estimación de la desviación típica de los coeficientes de regresión, en la columna **Standard Error**. En la tercera columna aparecen los intervalos de confianza al 95 % para cada parámetro. Obsérvese como cada uno de ellos contiene al valor 0, por lo que puede admitirse que cualquiera de los parámetros pueda ser nulo. No obstante para admitir o rechazar valores de los parámetros muy próximos a los extremos de cada intervalo deberíamos aumentar el tamaño muestral, que es muy pequeño.

Podemos cambiar el nivel de confianza haciendo click con el botón derecho del ratón y seleccionando la opción **Pane Options...**.

Predicciones.

Para llevar a cabo predicciones con el modelo estimado, debemos añadir al final del fichero de datos un caso por cada valor a predecir con los valores de las variables independientes, dejando vacía la celda correspondiente a la variable dependiente. Por ejemplo, si queremos predecir el valor del Indicador de Consumo para dos provincias, la primera con 75 automóviles y 150 teléfonos por cada 1000 habitantes, y la segunda con 90 automóviles y 120 teléfonos por cada 1000 habitantes; debemos añadir dichos valores como observaciones de las variables correspondientes en las filas 9 y 10 del fichero de datos.

Entonces, volveremos a la ventana **Multiple Regression** y ejecutaremos el icono **Tabular options** que aparece debajo de la barra del título. A continuación, debemos activar con el ratón la casilla de verificación correspondiente a la opción **Reports**. Al hacer click sobre el botón OK se obtiene la siguiente salida:

```
Regression Results for Y
-------------------------------------------------------------------------
         Fitted    Stnd. Error     Lower 95.0% CL    Upper 95.0% CL
Row      Value     for Forecast    for Forecast      forForecast
-------------------------------------------------------------------------
   9     84.0367     6.74094           66.7085         101.365
  10     84.6485     7.93164           64.2595         105.038
-------------------------------------------------------------------------
```

```
---------------------------------
Lower 95.0% CL     Upper 95.0% CL
for Mean           for Mean
---------------------------------
  75.8385             92.2349
  71.1335             98.1635
---------------------------------

The StatAdvisor
---------------
   This table contains information about Y generated using
the fitted model.  The table includes:
   (1) the predicted value of Y using the fitted model
   (2) the standard error for each predicted value
   (3) 95.0% prediction limits for new observations
   (4) 95.0% confidence limits for the mean response
Each item corresponds to the values of the independent variables
in a specific row of your data file.  To generate forecasts for
additional combinations of the variables, add additional rows to
the bottom of your data file.  In each new row, enter values for
the independent variables but leave the cell for the dependent
variable empty.  When you return to this pane, forecasts will be
added to the table for the new rows, but the model will be
unaffected.
```

La tabla anterior muestra para cada fila de observaciones de las variables independientes, y de izquierda a derecha, las siguientes columnas: número de fila, valores predichos de la variable dependiente según el plano de regresión estimado, intervalos de confianza para la variable independiente e intervalos de confianza para la media de la variable independiente.

Modelo sin término constante.

Al realizar los contrastes de nulidad sobre los parámetros del modelo pudimos comprobar que al nivel de significación $\alpha = 0.05$ había evidencia para admitir que el plano de regresión pasara por el origen, esto es, que $\beta_0 = 0$. Es posible obtener con Statgraphics el modelo reestimado bajo esta hipótesis, es decir, el modelo sin término constante. Para ello debemos situarnos sobre la ventana **Multiple Regression Analysis** y hacer click con el botón derecho del ratón. Entonces seleccionaremos **Analysis Options...** y desactivaremos la casilla de verificación **Constant in Model**. Automáticamente todas las ventanas de regresión actualizarán sus salidas para adaptarse a un modelo sin término constante.