Práctica 4

Vamos a analizar con STATGRAPHICS el problema de estimación, contrastes de regresión, contraste de linealidad, y predicciones, sobre un modelo de regresión lineal simple. Para ello utilizaremos los datos de la Práctica 3.

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

Debemos ejecutar la opción **RELATE/SIMPLE REGRESSION**, y entonces establecer la variable dependiente como Y, y la independiente como X. Al hacer click sobre el botón OK se obtiene la siguiente salida:

```
Regression Analysis - Linear model: Y = a + b*X
------------------------------------------------------------------------------
Dependent variable: Y
Independent variable: X
------------------------------------------------------------------------------
                              Standard          T
Parameter       Estimate        Error       Statistic        P-Value
------------------------------------------------------------------------------
Intercept        5.12182      0.164997        31.0419         0.0000
Slope           -2.06948     0.0671017       -30.8409         0.0000
------------------------------------------------------------------------------


                           Analysis of Variance
------------------------------------------------------------------------------
Source          Sum of Squares     Df  Mean Square    F-Ratio      P-Value
------------------------------------------------------------------------------
Model                  253.586      1      253.586     951.16       0.0000
Residual               12.7971     48     0.266607
------------------------------------------------------------------------------
Total (Corr.)          266.383     49

Correlation Coefficient = -0.975684
```

```
R-squared = 95.196 percent
Standard Error of Est. = 0.51634


The StatAdvisor
---------------
   The output shows the results of fitting a linear model to describe
the relationship between Y and X.  The equation of the fitted model is

   Y = 5.12182 - 2.06948*X

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

   The R-Squared statistic indicates that the model as fitted explains
95.196% of the variability in Y.  The correlation coefficient equals
-0.975684, indicating a relatively strong relationship between the
variables.  The standard error of the estimate shows the standard
deviation of the residuals to be 0.51634.  This value can be used to
construct prediction limits for new observations by selecting the
Forecasts option from the text menu.
```

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 $H_0 : \beta_0 = 0$ y $H_0 : \beta_1 = 0$.

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-2} mayor o igual en valor absoluto que $|t_{exp}|$, en la columna **P-value**. Resulta

$$\hat{\beta}_0 = \widehat{Intercept} = 5.12182\ , \quad \widehat{SE(\hat{\beta}_0)} = 0.164997\ , \quad T_{exp} = 31.0419\ para\ \ H_0 :\ \beta_0 = 0$$
$$\hat{\beta}_1 = \widehat{Slope} = -2.06948\ , \quad \widehat{SE(\hat{\beta}_1)} = 0.0671017\ , \quad T_{exp} = -30.8409\ para\ \ H_0 :\ \beta_1 = 0$$

Entonces, la recta de regresión estimada está dada por

$$\widehat{Y} = 5.12182 - 2.06984\ X\ .$$

Así, cuando la variable X toma el valor 0, vamos a predecir el valor medio de la variable Y por $\widehat{Y} = 5.12182$. Y, por cada unidad que aumente la variable X, la variable Y disminuirá por término medio en 2.06984 unidades.

Fijado un nivel de significación $\alpha = 0.05$ se concluye, dado que los dos P-valores son menores que 10^{-5}, y por tanto menores que α, que no hay evidencia empírica para admitir que los valores de la pendiente y la ordenada sean nulos. Notemos que ambos contrastes se hacen por separado y que la conclusión obtenida para la pendiente de la recta nos lleva a afirmar la existencia de relación lineal entre ambas variables, hecho que será confirmado de nuevo mediante la tabla ANOVA.

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

En la tabla ANOVA se tiene que la variabilidad total, **VT = Total Sum of Squares = 266.383**, se descompone en variabilidad explicada, **VE = Model Sum of Squares = 253.586** y no explicada o residual, **VNE = Residual Sum of Squares =12.7971**. El estadístico de contraste $\mathbf{F}_{exp}$ toma el valor 951.16. Nuevamente el **P-value** asociado a este contraste representa la probabilidad de encontrar un valor de una $F_{1,n-2}$ mayor o igual que F_{exp}. Puesto que el P-valor asociado es menor que 10^{-5}, entonces dicho P-valor es menor que $\alpha = 0.05$, y debemos rechazar $H_0 : R^2 = 0$, esto es, no hay evidencia empírica para rechazar una relación lineal entre las variables. Observemos además cómo el valor F_{exp} es el cuadrado de t_{exp} para el contraste de la pendiente de la recta, como ya quedó demostrado de forma teórica.

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

$$\widehat{\sigma}^2 = Residual\ Mean\ Square = 0.266607\ .$$

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

$$\widehat{\sigma} = Standard\ Error\ of\ Est. = 0.51634\ .$$

El valor del coeficiente de determinación $R^2 =$ **R-squared = 0.95196** indica que el 95.196 % de la variabilidad total de la variable dependiente es explicada por la recta de regresión. Su raíz cuadrada, con el signo de la covarianza, define el coeficiente de correlación lineal entre las variables X e Y, $R =$ **Correlation Coefficient = -0.975684**. En este caso, su signo negativo indica que la recta de regresión es decreciente, o equivalentemente, la relación lineal entre las variables es de tipo inversa, esto es, al aumentar X disminuye Y, y recíprocamente.

Contraste de linealidad o de falta de ajuste.

Debemos ejecutar el icono **Tabular options** que aparece debajo de la barra del título de la ventana **Simple 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 **Lack-of-fit-Test**. Al hacer click sobre el botón OK se obtiene la siguiente salida:

Analysis of Variance with Lack-of-Fit

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	253.586	1	253.586	951.16	0.0000
Residual	12.7971	48	0.266607		
Lack-of-Fit	2.68407	13	0.206467	0.71	0.7368
Pure Error	10.113	35	0.288944		
Total (Corr.)	266.383	49			

The StatAdvisor

The lack of fit test is designed to determine whether the selected model is adequate to describe the observed data, or whether a more complicated model should be used. The test is performed by comparing the variability of the current model residuals to the variability between observations at replicate values of the independent variable X. Since the P-value for lack-of-fit in the ANOVA table is greater or equal to 0.10, the model appears to be adequate for the observed data.

La tabla anterior es la tabla ANOVA de la regresión incluyendo la descomposición de la variabilidad no explicada. Dicha descomposición ha sido posible gracias a que se dispone de varias observaciones de la variable respuesta para idénticos valores de la variable independiente. En concreto hay $d = 15$ valores distintos de X.

El término de falta de ajuste está dado por

$$\sum_{i=1}^{d} n_i(\bar{y}_i - \widehat{y}_i)^2 = \textbf{Lack-of-Fit Sum of Squares} = 2.68407 \ .$$

Y el término de errores puros está dado por

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \textbf{Pure Error Sum of Squares} = 10.113 \ .$$

La suma de ambos términos coincide, por supuesto, con $VNE = 12.7971$. Y el estadístico F para el contraste de linealidad toma el valor $F_{exp} = \textbf{F-ratio} = 0.71$. El **P-value** asociado es la probabilidad de encontrar un valor de una $F_{d-2,n-d}$ mayor o igual que F_{exp}, y resulta ser 0.7368. Y, como $\alpha = 0.05$ es menor que dicho P-valor, debemos admitir la hipótesis de linealidad. Así, la variabilidad no explicada se debe a la variabilidad inherente a los datos, y no a que la medias de las distribuciones condicionadas de Y a cada valor de X no estén sobre una recta.

Predicciones.

Nuevamente debemos ejecutar el icono **Tabular options** que aparece debajo de la barra del título de la ventana **Simple Regression** y activar con el ratón la casilla de verificación correspondiente a la opción **Forecasts**. Al hacer click sobre el botón OK se obtiene la siguiente salida:

Predicted Values

		95.00% Prediction Limits		95.00% Confidence Limits	
X	Predicted Y	Lower	Upper	Lower	Upper
0.5	4.08708	3.01364	5.16052	3.81418	4.35997
4.5	-4.19083	-5.28409	-3.09756	-4.53351	-3.84815

```
The StatAdvisor
---------------
   This table shows the predicted values for Y using the fitted model.
In addition to the best predictions, the table shows:
   (1) 95.0% prediction intervals for new observations
   (2) 95.0% confidence intervals for the mean of many observations
The prediction and confidence intervals correspond to the inner and
outer bounds on the graph of the fitted model.
```

La tabla anterior muestra de izquierda a derecha las siguientes columnas: valores de la variable independiente, valores predichos de la variable dependiente según la recta de regresión estimada, intervalos de confianza para la variable e intervalos de confianza para la media de la variable.

Por defecto, STATGRAPHICS presenta únicamente la tabla antes descrita para dos valores de la variable independiente, a saber, el menor y el mayor de todos. Para indicar sobre qué valores deseamos realizar las predicciones debemos hacer click con el botón derecho del ratón, con lo que desplegaremos un menu emergente. La opción **Pane Options...** nos da acceso a la ventana **Forecats Options** en la que podemos indicar los valores deseados de la variable independiente, así como el nivel de confianza para los intervalos de confianza.

Si indicamos un valor de la variable independiente igual a 0, el intervalo de confianza para la media de la variable independiente coincidirá con el intervalo de confianza para el parámetro β_0. En este caso la tabla de predicción está dada por

Predicted Values

		95.00%		95.00%	
	Predicted	Prediction Limits		Confidence Limits	
X	Y	Lower	Upper	Lower	Upper
0.0	5.12182	4.03193	6.21171	4.79007	5.45357

Entonces, el intervalo de confianza al 95 % para β_0 está dado por $(4.79007, 5.45357)$.