Índice general

Capítulo 6

Regresión Múltiple

El modelo de regresión múltiple es la extensión a k variables explicativas del modelo de regresión simple estudiado en el capítulo anterior. En general, una variable de interés y depende de varias variables $x_1, \ldots, x_k$ y no sólo de una única variable de predicción x. Por ejemplo, para estudiar la variación del precio de una vivienda, parece razonable considerar más de una variable explicativa, como pueden ser el precio del suelo, la superficie del piso, número de cuartos de baño, edad de la vivienda, etc. Además de las variables observables, la variable de interés puede depender de otras desconocidas para el investigador. Un modelo de regresión representa el efecto de estas variables en lo que se conoce como error aleatorio o pertubación.

Si suponemos un modelo de regresión teórico en el que las variables se pueden relacionar mediante una función de tipo lineal, éste puede escribirse

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon \tag{6.1}$$

donde β_0, $\beta_1, \ldots,$ β_k son los parámetros desconocidos que vamos a estimar y ε es el error aleatorio o perturbación. y es la variable de interés que queremos predecir, también llamada variable respuesta o variable dependiente. Las variables $x_1, \ldots, x_k$ se llaman variables independientes, explicativas o de predicción. El error aleatorio ε representa el efecto de todas las variables que pueden afectar a la variable dependiente y no están incluidas en el modelo (6.9).

Algunos ejemplos de modelos de regresión mútiple pueden ser:

- El consumo de combustible de un vehículo, cuya variación puede ser explicada por la velocidad media del mismo y por el tipo de carretera. Podemos incluir en el término de error variables como el efecto del conductor, las condiciones meteorológicas, etc.
- El presupuesto de una universidad, cuya variación puede ser explicada por el número de alumnos. También podríamos considerar en el modelo variables como el número de profesores, el número de laboratorios, superficie disponible de instalaciones, personal de administración, etc.

Si se desea explicar los valores de una variable aleatoria y mediante k variables $x_1, \ldots, x_k$ que a su vez toman n valores, tenemos entonces

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \tag{6.2}$$

Las perturbaciones deben verificar las siguientes hipótesis:

- Su esperanza es cero
- Su varianza es constante

- Son independientes entre sí
- Su distribución es normal

Los parámetros desconocidos son estimados por mínimos cuadrados, resultando la ecuación estimada de regresión dada por

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_k x_k \tag{6.3}$$

donde cada coeficiente $\widehat{\beta}_i$ representa el efecto sobre la respuesta cuando la variable aumenta en una unidad y las demás variables permanecen constantes. Puede interpretarse como el efecto diferencial de esta variable sobre la variable respuesta cuando controlamos los efectos de las otras variables. $\widehat{\beta}_0$ es el valor de la respuesta ajustada cuando todas las variables explicativas toman el valor cero.

6.1. Descomposición de la variabilidad y contrastes de hipótesis

La variabilidad de la respuesta puede descomponerse de igual forma que en regresión simple

$$\sum (y_i - \overline{y}) = \sum (\widehat{y}_i - \overline{y})^2 + \sum (y_i - \widehat{y}_i)^2. \tag{6.4}$$

Esta descomposición la notamos por

SCT=SCReg+ SCE

donde SCT es la suma de cuadrados total y representa la variablidad total, $SCReg$ es la suma de cuadrados de la regresión y representa la variabilidad explicada por el modelo de regresión. SCE es la suma de cuadrados residual y representa la variabilidad que queda sin explicar. Esta descomposición suele escribirse en la siguiente tabla

TABLA ANOVA

Fuente	Suma de cuadrados	g.l.	Varianza	Contraste
Regresión	$SCReg = \sum (\widehat{y}_i - \overline{y})^2$	k	$\widehat{s}_e^2 = \frac{\sum (\widehat{y}_i - \overline{y})^2}{k}$	$F_{exp} = \frac{\widehat{s}_e^2}{\widehat{s}_R^2}$
Error	$SCE = \sum e_i^2$	$n-k-1$	$\widehat{s}_R^2 = \frac{\sum e_i^2}{n-k-1}$	
Total	$SCT = \sum (y_i - \overline{y})$	$n-1$	$\widehat{s}_y^2 = \frac{\sum (y_i - \overline{y})}{n}$	

donde $e_i = y_i - \widehat{y}_i$ son los residuos, $\widehat{s}_R^2$ es la varianza residual y $\widehat{s}_y^2$ es la varianza de y.

El valor del estadístico $F_{exp} = \frac{\widehat{s}_e^2}{\widehat{s}_R^2}$ permite resolver el contraste de regresión, dado por

$$\begin{cases} H_0: & \beta_1 = \ldots = \beta_k = 0 \\ H_1: & \text{Algún} \quad \beta_i \neq 0 \quad \text{para} \quad i = 1, \ldots, k \end{cases}$$

Fijado un nivel de significación α, se rechaza H_0 si $F_{exp} > F_{\alpha,k,n-k-1}$. En la práctica Statgraphics proporciona el $p - valor$ o nivel mínimo de significación para el rechazo de H_0, que permite resolver el contraste de hipótesis fijado un nivel de significación.

Si $p - valor < \alpha$ entonces **No rechazamos** H_0

Si $p - valor \geq \alpha$ entonces **Rechazamos** H_0

Si estamos interesados en estudiar el efecto individual de una variable explicativa sobre la variable respuesta se considera el siguiente contraste

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

En este caso el estadístico de contraste sigue una F con 1 y $n-k-1$ grados de libertad. Este contraste es equivalente al contraste de regresión con una única variable explicativa, estudiado en el capítulo anterior. El rechazo de la hipótesis nula implica admitir la validez de la variable explicativa x_i para predecir la variable de interés y.

6.2. Coeficiente de determinación

Para construir una medida descriptiva del ajuste global de un modelo de regresión se emplea el coeficiente de determinación, dado por

$$R^2 = \frac{SCReg}{SCT}. \tag{6.5}$$

R^2 representa la proporción de variación de y explicada por el modelo de regresión. Por construcción, es evidente que $0 \leq R^2 \leq 1$. Si $R^2 = 1$ entonces $SCReg = SCT$, por lo que toda la variación de y es explicada por el modelo de regresión. Si $R^2 = 0$ entonces $SCT = SCE$, por lo que toda la variación de y queda sin explicar. En general, cuanto más próximo esté a 1, mayor es la variación de y explicada por el modelo de regresión.

Sin embargo, en regresión múltiple, el coeficiente de determinación presenta el inconveniente de que su valor aumenta al añadir nuevas variables al modelo de regresión, independientemente de que éstas contribuyan de forma significativa a la explicación de la variable respuesta. Para evitar un aumento injustificado de este coeficiente se introduce el coeficiente de determinación corregido, que notamos por $\overline{R}^2$ y que se obtiene a partir de R^2 de la forma

$$\overline{R}^2 = \frac{n-1}{n-k-1} R^2. \tag{6.6}$$

Este coeficiente no aumenta su valor cuando se añaden nuevas variables, sino que en caso de añadir variables superfluas al modelo, el valor de $\overline{R}^2$ disminuirá considerablemente respecto al valor del coeficiente $\overline{R}^2$.

Ejercicio

Una empresa fabricante de cereales para el desayuno desea conocer la ecuación que permita predecir las ventas (en miles de euros) en función de los gastos en publicidad infantil en televisión (en miles de euros), la inversión en publicidad en radio (en miles de euros) y la inversión en publicidad en periódicos (en miles de euros). Se realiza un estudio en el que se reúnen los datos mensuales correspondientes a los últimos 20 meses. Estos datos se pueden obtener en el fichero ventas.sf3.

Se pide:

a) Representar el gráfico de dispersión matricial

b) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación

c) Realiza el contraste global de significación del modelo regresión

d) ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales

e) ¿Cuáles serán las ventas estimadas para una inversión en publicidad en televisión, en radio y en periódicos de 1.77, 0.83 y 0.7 miles de euros, respectivamente?

f) Obtener e interpretar el valor de la suma de cuadrados residual

g) Coeficiente de determinación y de determinación corregido

SOLUCIÓN:

a) Representar el gráfico de dispersión matricial.

En el estudio de un modelo de regresión lineal múltiple el gráfico de dispersión matricial es el primer gráfico que se debe observar. Proporciona una primera idea de la existencia de relación lineal o de otro tipo entre la respuesta y las variables explicativas y también da una idea de posibles relaciones lineales entre las variables de predicción, lo que crea problemas de multicolinealidad. Para realizar este gráfico con Statgraphics se selecciona en los menús **Gráficos → Gráficos de dispersión → Gráfico de Matriz...**, se introducen las variables (ver Figuras 6.1 y 6.2) y pulsando **Aceptar** obtenemos el gráfico de dispersión matricial.

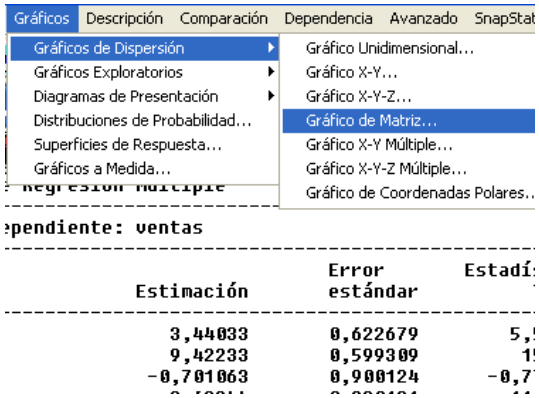


Figura 6.1: Elección en los menús de Statgraphics para representar el gráfico de dispersión matricial

b) Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación

Notamos ventas, pubtv, pubradio y pubper las variables que intervienen en el ejercicio. La variable ventas es la variable dependiente, mientras que pubtv, pubradio y pubper son las variables explicativas.

Ajustamos un modelo de regresión que responde a una expresión del tipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{6.7}$$

donde y representa las ventas de cereales (en miles de euros), x_1 es la inversión en publicidad en televisión (en miles de euros), x_2 representa el gasto en publicidad por radio (en miles de euros) y x_3 es la publicidad en periódicos (en miles de euros).

Los párametros desconocidos β_0, β_1, β_2, y β_3 son estimados por mínimos cuadrados. Para obtener dicha estimación con Statgraphics se selecciona **Dependencia → Regresión Múltiple...** (Ver Figura 6.5)

Se introducen las tres variables explicativas en el campo Variables Independientes y la variable ventas en el campo Variable Dependiente, como muestra la Figura 6.6.

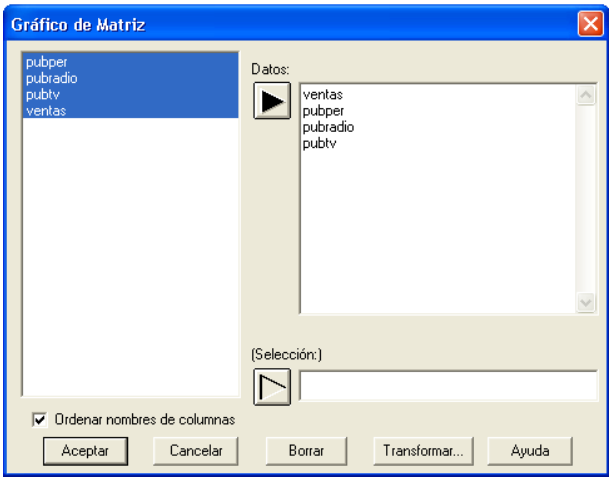


Figura 6.2: Se introducen las variables en el campo Datos

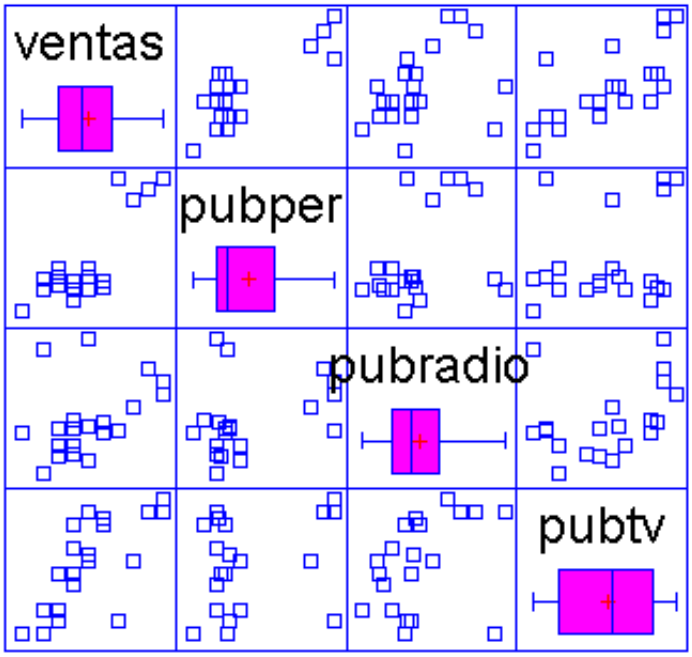


Figura 6.3: Gráfico de dispersión matricial

	ventas	pubtv	pubradio	pubper
1	10	1	0,5	0,4
2	12	1,2	0,57	0,4
3	11	1,3	0,56	0,42
4	13	1,4	0,55	0,5
5	12	1,5	0,6	0,4
6	14	1,7	0,65	0,44
7	14	1,75	0,69	0,41
8	12	1,3	0,67	0,44
9	13	1,45	0,68	0,46
10	11	0,9	0,67	0,46
11	10	0,8	0,97	0,45
12	15	0,9	0,66	0,9
13	8,5	0,8	0,65	0,3
14	11	1	0,6	0,5

Figura 6.4: []

Figura 6.5: Elección en los menús de Statgraphics para realizar un ajuste de un modelo de regresión lineal múltiple

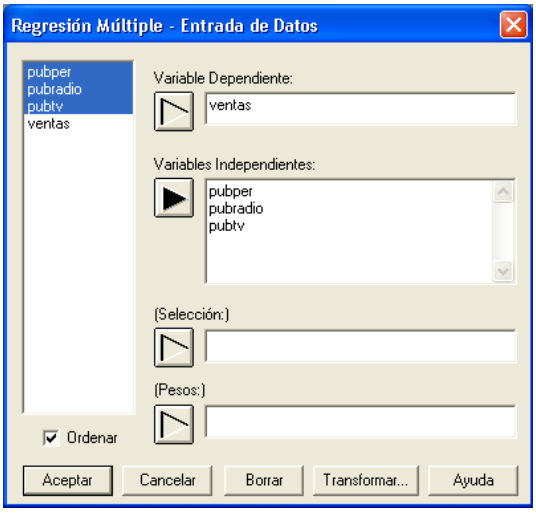


Figura 6.6: Introducimos las variables en los campos correspondientes

Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.7: Salida de Statgraphics donde aparecen las estimaciones de los coeficientes

Se pulsa Aceptar y se obtiene como resultado la salida del programa que recoge la Figura 6.7.

En la Figura 6.7 aparecen los parámetros estimados de regresión $\widehat{\beta}_0 = 3{,}44$, $\widehat{\beta}_1 = 3{,}682$, $\widehat{\beta}_2 = -0{,}701$ y $\widehat{\beta}_3 = 9{,}422$ y la ecuación estimada de regresión, que está dada por:

$$\widehat{y} = 3{,}44 + 3{,}682x_1 - 0{,}701x_2 + 9{,}422x_3. \tag{6.8}$$

Las ventas estimadas son iguales a 3440 euros si no se produce inversión en publicidad ni en televisión, ni en radio, ni en periódicos.

Por cada mil euros invertidos en publicidad en televisión las ventas esperadas aumentan en 3682 euros, supuesto que permanecen constantes las otras variables.

Por cada mil euros invertidos en publicidad en radio las ventas estimadas disminuyen 701 euros, suponiendo que se mantienen constantes las otras variables independientes.

Por cada mil euros invertidos en publicidad en periódicos se produce un incremento en las ventas esperadas de 9422 euros, supuestas constantes las restantes variables predictivas.

c) Contrasta la significación global del modelo regresión

El contraste de significación del modelo de regresión permite verificar si ninguna variable explicatica es válida para la predicción de la variable de interés. Este contraste puede escribirse por:

$$\begin{cases} H_0: & \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1: & \text{Algún} \quad \beta_i \neq 0 \quad \text{para} \quad i = 1, 2, 3 \end{cases}$$

El p-valor asociado al contraste (Figura 6.8) es menor que $\alpha = 0,05$, por lo que rechazamos la hipótesis nula. Esto implica que al menos una de las variables independientes contribuye de forma significativa a la explicación de la variable respuesta.

d) ¿Puede eliminarse alguna variable del modelo? Realiza los contrastes de significación individuales

Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.8: Salida de Statgraphics donde aparece el $p-valor$ asociado al contraste global de regresión

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{6.9}$$

Realizamos tres contrastes de hipótesis, uno para cada coeficiente que acompaña a cada variable explicativa $i = 1, 2, 3$.

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

Para la variable pubradio, $p - valor = 0{,}4474 > \alpha = 0{,}05$ (Figura 6.9), por lo que no rechazamos la hipótesis nula de significación de la variable pubradio. Esta variable no es válida para predecir las ventas de cereales. Se puede considerar eliminar dicha variable del modelo.

e) ¿Cuáles serán las ventas estimadas para una inversión en publicidad en televisión, en radio y en periódicos de 1.77, 0.83 y 0.7 miles de euros, respectivamente?

Introducimos 1.77 en la columna pubtv, 0.83 en la de pubradio y 0.7 en la columna de pubper, como se aprecia en la Figura 6.10. Pulsamos el botón **Opciones tabulares** y elegimos la opción **Informes**. Statgraphics calcula automáticamente el valor de las ventas estimadas para la inversión en publicidad realizada. En la Figura 6.11 se muestra que el valor de las ventas estimadas es igual a 15972 euros.

f) Obtener e interpretar el valor de la suma de cuadrados residual

El valor de la suma de cuadrados residual puede obtenerse de la tabla ANOVA (Ver Figura 6.12), de donde $SCE = 3,86706$. Este valor es considerablemente menor que el correspondiente a la suma de cuadrados de regresión, $SCReg = 131,37$. De la variación de las ventas $SCT = 135,237$, únicamente queda sin explicar la cantidad $SCE = 3,86706$, mientras que $SCReg = 131,37$ son explicadas por el modelo de regresión.

g) Coeficiente de determinación y de determinación corregido

Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.9: Salida de Statgraphics donde aparece los $p-valores$ asociados a los contrastes de regresión individuales

12	15	0,9	0,66	0,9
13	8,5	0,8	0,65	0,3
14	11	1	0,6	0,5
15	12	1,7	0,7	0,35
16	13	1,8	1,01	0,4
17	16	1,4	0,75	0,8
18	18	1,9	0,8	0,9
19	18	1,8	0,85	0,9
20	17	1,8	0,9	0,85
21		1,77	0,83	0,7

Figura 6.10: Datos para realizar la predicción de ventas para ciertas inversiones en publicidad

Resultados de la Regresión para ventas

Fila	Ajustado Valor	Error Est. para la Predicción	Inf. 95,0% CL la Predicción	Sup. 95,0% CL la Predicción	Inf. 95,0% CL para la Media	Sup. 95,0% CL para la Media
21	15,972	0,523879	14,8614	17,0826	15,5883	16,3557

Figura 6.11: Salida de Stagraphics donde aparece la predicción de ventas para ciertas inversiones en publicidad

El coeficiente de determinación corregido es igual a 0,96604 y representa la proporción en la variación de las ventas que son explicadas por el modelo de regresión. Este coeficiente tomar un valor muy satisfactorio.

Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.12: Valor de la suma de cuadrados residual

6.3. Hipótesis del modelo

6.3.1. Normalidad de los residuos

La normalidad de los residuos puede contrastarse gráficamente mediante el gráfico probabilístico normal. En dicho gráfico la diagonal representa la ubicación teórica de los residuos en el caso de que éstos sigan una distribución normal. Desviaciones de estos puntos respecto de la diagonal indican alteraciones de la normalidad de los residuos.

Para obtener un gráfico de normalidad de los residuos es necesario guardar previamente los residuos pulsando el botón **Guardar** y seleccionar la casilla **Residuos** de la ventana **Opciones Guardar Resultados**, como se muestra en la Figura 6.14. Los residuos son guardados en el Editor de datos con el nombre RESIDUALS. Se selecciona **Gráficos → Gráficos exploratorios → Gráfico Probabilístico...**
Se introduce la variable RESIDUALS en la casilla Datos, se pulsa **Aceptar** y se obtiene el gráfico probabilístico normal.
Podemos observar que los puntos del gráfico se aproximan razonablemente bien a la diagonal, por lo que se debe aceptar la hipótesis de normalidad de los residuos.

6.3.2. Homocedasticidad

Para estudiar si se cumple la hipótesis de homocedasticidad, se pulsa Opciones Gráficas y se selecciona el gráfico **Residuos frente a Predicho**(Ver Figura 6.3.2), donde se representan los residuos estudentizados frente a los valores estimados. El análisis de este gráfico puede revelar una posible violación de la hipótesis de homocedasticidad, por ejemplo si detectamos que el tamaño de los residuos aumenta o disminuye de forma sistemática a medida que aumenta $\widehat{y}$. Si dicho gráfico no muestra patrón alguno

Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.13: Valor del coeficiente de determinación corregido

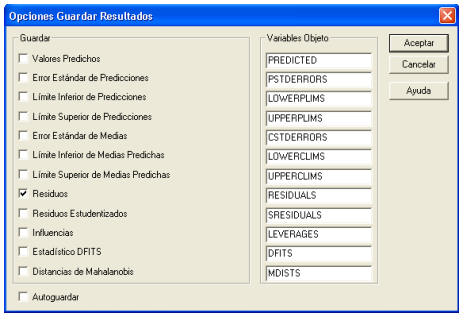


Figura 6.14: [Opciones Guardar Residuos]

entonces podemos aceptar que se cumple la hipótesis de varianza de los residuos constante. En la Figura 6.18 se muestra un ejemplo de un ajuste de regresión en el que existe heterocedasticidad.
En la Figura 6.17 no se aprecia tendencia clara en este gráfico. Los residuos no presentan estructura definida respecto de los valores ajustados $\widehat{y}_i$, por lo que podemos aceptar la hipótesis de homocedasticidad.

Además de contrastar la homocedasticidad, este gráfico sirve para detectar indicios de falta de adecuación del modelo propuesto a los datos. Si observamos trayectorias de comportamiento no aleatorio, esto es indicio de que el modelo propuesto no describe adecuadamente los datos.

6.3.3. Independencia de los residuos

La falta de independencia, se produce fundamentalmente cuando se trabaja con variables aleatorias que se observan a lo largo del tiempo, esto es, cuando se trabaja con series temporales. Por ello, una primera medida para tratar de evitar la dependencia de las observaciones consiste en aleatorizar la recogida muestral. El que no se cumpla la hipótesis de independencia afecta gravemente a los resultados del modelo de regresión puesto que se obtienen estimadores de los parámetros y predicciones ineficientes y los intervalos de confianza y contrastes que se deducen de la tabla ANOVA no son válidos. Para contrastar si existe dependencia entre los residuos podemos emplear el contraste de Durbin-Watson. En la Figura 6.19 se muestra el valor del estadístico de Durbin-Watson, así como el p-valor asociado a dicho contraste. A un nivel de significación del 5 %, $p - valor = 0{,}2479 > \alpha = 0{,}05$, por lo que se acepta la hipótesis nula de independencia de los residuos.

6.4. Regresión con variables cualitativas

Ejemplo

La variable de interés de este ejemplo problema es salari07, que representa el salario al mes en el año 2007 de un grupo de trabajadores de una empresa. Se pretende explicar dicha variable a partir de la variable cuantitativa edad y a partir de la variable cualitativa sexo. El conjunto de datos se puede conseguir del fichero miempresa.sf3. Se pide:

a) Recodifica la variable sexo en una variable A que valga 1 para la modalidad hombre y 0 para la modalidad mujer

b) Realiza un ajuste lineal múltiple. Interpreta los valores de los coeficientes estimados de regresión

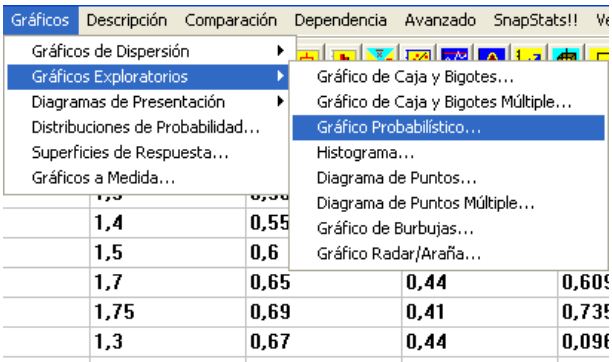


Figura 6.15: Elección en los menús de Statgraphics para representar el gráfico probabilístico normal

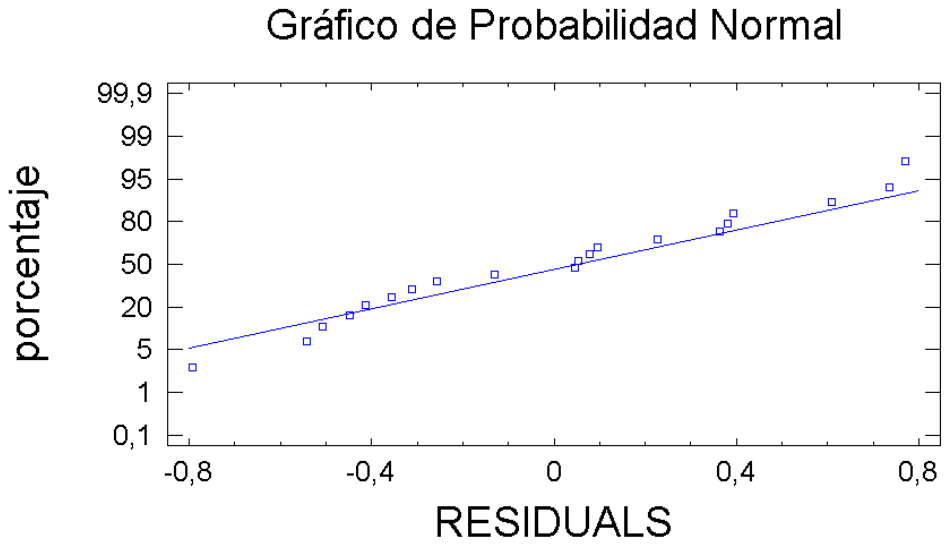
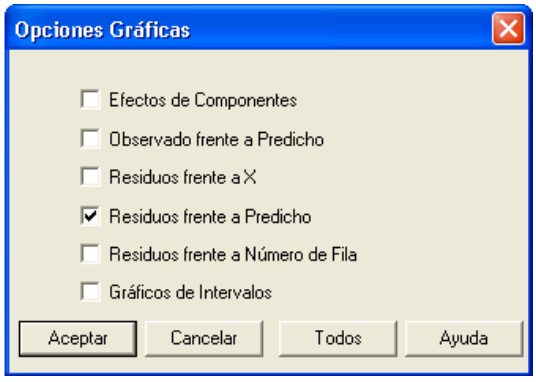


Figura 6.16: Gráfico probabilístico normal de los residuos



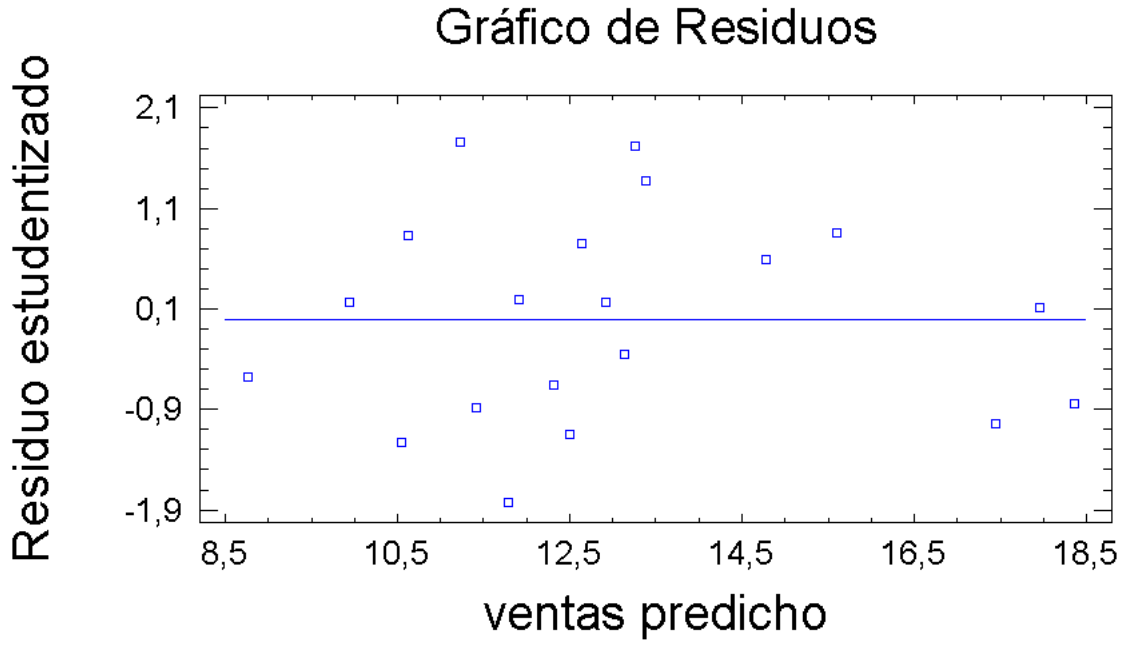


Figura 6.17: Gráfico de los residuos frente a las predicciones de las ventas

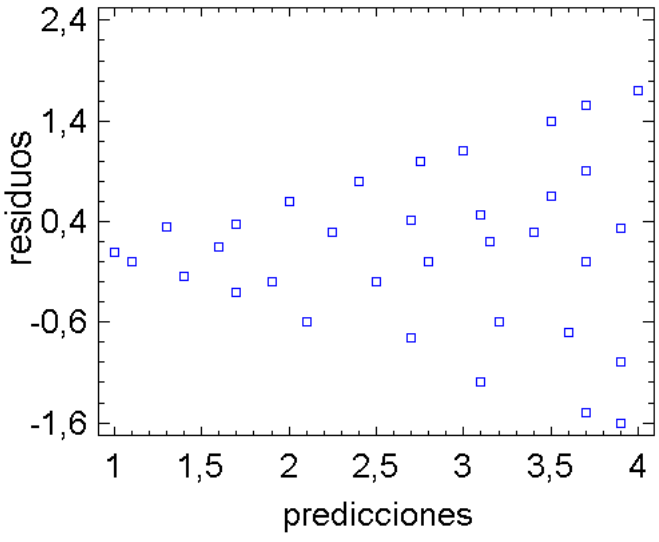
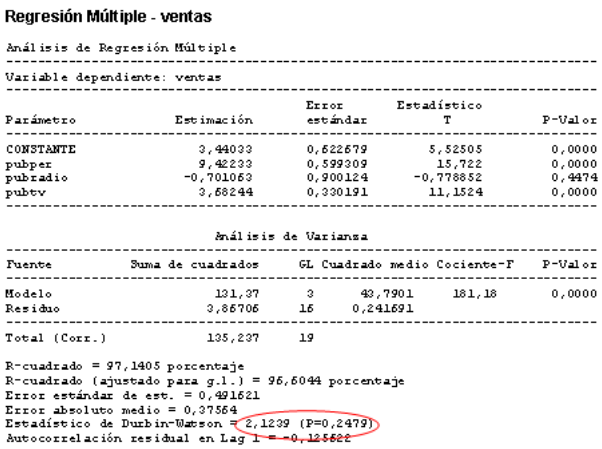


Figura 6.18: Ejemplo de gráfico de residuos frente a predicciones en el que existe heterocedasticidad



Regresión Múltiple - ventas

Análisis de Regresión Múltiple

Variable dependiente: ventas

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	3,44033	0,622679	5,52505	0,0000
pubper	9,42233	0,599309	15,722	0,0000
pubradio	-0,701063	0,900124	-0,778852	0,4474
pubtv	3,68244	0,330191	11,1524	0,0000

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	131,37	3	43,7901	181,18	0,0000
Residuo	3,86706	16	0,241691		
Total (Corr.)	135,237	19			

R-cuadrado = 97,1405 porcentaje
R-cuadrado (ajustado para g.l.) = 96,6044 porcentaje
Error estándar de est. = 0,491621
Error absoluto medio = 0,37564
Estadístico de Durbin-Watson = 2,1239 (P=0,2479)
Autocorrelación residual en Lag 1 = -0,125622

Figura 6.19: Valor del estadístico de Durbin-Watson y $p-valor$ asociado al contraste de Durbin-Watson

c) ¿Es significativa la contribución de la variable A a la explicación del salario de los trabajadores en el año 2007?

SOLUCIÓN:

a) **Recodifica la variable sexo en una variable A que valga 1 para la modalidad hombre y 0 para la modalidad mujer**

Para constuir un modelo de regresión múltiple que incorpore la información de tipo cualitativo es necesario recodificar la variable sexo en otra variable, que notamos por A y que se define por

$$A = \begin{cases} 1 & \text{si} \quad \text{sexo} = h \\ 0 & \text{si} \quad \text{sexo} = m \end{cases}$$

Para conservar en el editor de datos la variable de partida sexo es conveniente copiarla en una columna en blanco. Llamamos esta nueva variable como $sexo1$. Para recodificarla se pulsa sobre el nombre de la variable con el botón derecho del ratón y se elige **Generar datos**. En la ventana que aparece se escribe **RECODE(sexo1)** para recodificar esta variable en valores 1 y 2 para hombre y mujer respectivamente. Es necesario comprobar que la variable sexo1 está definida como tipo **Numérica**, puesto que de lo contario Statgraphics no permite continuar el ejercicio.

A continuación, sobre una columna en blanco, se selecciona **Generar datos** con el botón derecho del ratón. En el campo **Expresiones** de la ventana resultante se escribe sexo1=1 como se muestra en la Figura 6.20. Pulsamos **Aceptar** para generar una nueva variable que toma valor 1 si el trabajador es hombre y toma el valor 0 si el trabajador es una mujer. Finalmente, se renombra esta variable como A.

b) **Realiza un ajuste lineal múltiple. Interpreta los valores de los coeficientes estimados de regresión**

La variable A se introduce en el modelo de igual forma que el resto de variables, analizando si la contribución de esta variable a la explicación de la variable salario es signifitiva del mismo modo que cualquier otra variable explicativa. Se selecciona en los menús de Statgraphics la opción

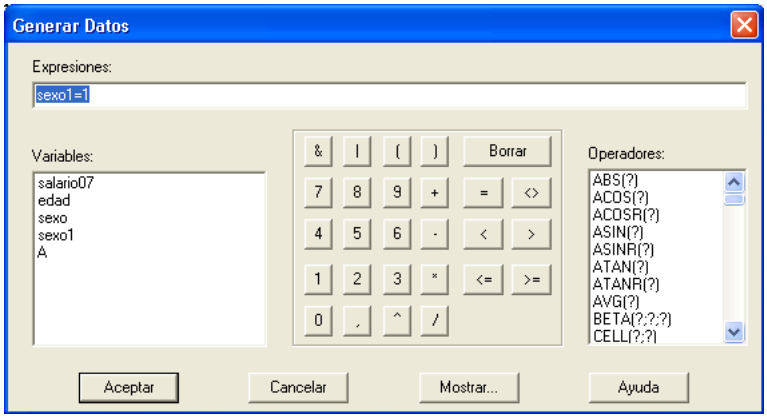


Figura 6.20: Generar Datos

Regresión Múltiple - salario07

Análisis de Regresión Múltiple

Variable dependiente: salario07

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	1,4652	0,175689	8,33972	0,0000
edad	0,0105206	0,00518785	2,02792	0,0498
A	0,0264294	0,0845919	0,312435	0,7565

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	0,3119	2	0,15595	2,21	0,1240
Residuo	2,61122	37	0,0705735		
Total (Corr.)	2,92312	39			

R-cuadrado = 10,6701 porcentaje
R-cuadrado (ajustado para g.l.) = 5,84146 porcentaje
Error estándar de est. = 0,265657
Error absoluto medio = 0,208791
Estadístico de Durbin-Watson = 0,568288 (P=0,0000)
Autocorrelación residual en Lag 1 = 0,711798

Figura 6.21: Salida de Statgraphics donde aparecen los coeficientes estimados del ajuste realizado

Dependencia → Regresión Múltiple e introducimos las variables en sus respectivos campos. Se pulsa **Aceptar** y se obtiene la salida de Statgraphics que aparece en la Figura 6.21.

El modelo ajustado es

$$\widehat{y} = 1{,}4652 + 0{,}01052x + 0{,}02643A, \tag{6.10}$$

donde x es la edad y A es la variable ficticia.

La pendiente 0.01052 representa el aumento que se produce en el salario estimado cuando se incrementa en un año la edad del trabajador. Por otra parte, 0.02643 es el aumento que se produce en el valor esperado del salario entre los individuos que presentan la modalidad hombre y los que presentan la modalidad mujer.

c) **¿Es significativa la contribución de la variable A a la explicación del salario de los trabajadores en el año 2007?**

Statgraphics proporciona los $p - valores$ (figura 6.22) asociados a los contrastes individuales de significación, como se muestra en la Figura anterior. La variable edad tiene una contribución significativa a la explicación de la variable respuesta a un nivel de significación del 5 %. En cambio, $p-valor = 0{,}7565 > \alpha = 0{,}05$, por lo que la variable A no es válida para la predicción de la variable de interés salario07.

Regresión Múltiple - salario07

Análisis de Regresión Múltiple

Variable dependiente: salario07

Parámetro	Estimación	Error estándar	Estadístico T	P-Valor
CONSTANTE	1,4652	0,175589	8,33972	0,0000
edad	0,0105206	0,00518785	2,02792	0,0498
A	0,0264294	0,0845919	0,312435	0,7565

Análisis de Varianza

Fuente	Suma de cuadrados	GL	Cuadrado medio	Cociente-F	P-Valor
Modelo	0,3119	2	0,15595	2,21	0,1240
Residuo	2,61122	37	0,0705735		
Total (Corr.)	2,92312	39			

R-cuadrado = 10,6701 porcentaje
R-cuadrado (ajustado para g.l.) = 5,84146 porcentaje
Error estándar de est. = 0,265657
Error absoluto medio = 0,208791
Estadístico de Durbin-Watson = 0,568288 (P=0,0000)
Autocorrelación residual en Lag 1 = 0,711798

Figura 6.22: Indicación de los p-valores