Práctica 5

Análisis de regresión

Contenido

En esta práctica vamos a usar algunas de las prestaciones disponibles en el paquete SPSS para el análisis de regresión. SPSS lleva a cabo tanto la regresión lineal simple como la múltiple, pero como esta última no la hemos estudiado el presente curso, tendremos que ignorar muchas de las opciones en los menús y cuadros de diálogo que aparecerán durante el análisis. Asimismo, SPSS puede realizar análisis de regresión no lineal, para el que emplea técnicas más sofisticadas, que tampoco consideraremos en esta práctica.

1 Objetivos

1. Trazar diagramas de dispersión para deducir visualmente la posible forma de la ecuación de regresión a partir de la observación de la nube de puntos.
2. Calcular las estimaciones de los coeficientes de regresión lineal simple, incluyendo la regresión curvilínea.
3. Calcular los valores pronosticados por el modelo de regresión, de la variable de respuesta y los residuos.

2 Observando diagramas de dispersión

Suele iniciarse un análisis estadístico llevando a cabo un *análisis exploratorio*, es decir, clasificando u ordenando los datos, elaborando tablas de frecuencias, trazando gráficos y calculando estadísticos descriptivos simples como porcentajes, medias y otros, de manera que podamos hacernos una idea inicial del problema, y adquirir indicios que nos orienten en las técnicas estadísticas que deben emplearse.

El paquete SPSS realiza estas tareas exploratorias pulsando en la barra de menú

Analizar ⟶ Estadísticos descriptivos

apareciendo el siguiente menú desplegable

Frecuencias...
Descriptivos...
Explorar...
Tablas de contingencia...
Razón...

del que podemos utilizar una o varias de sus opciones para el análisis exploratorio. De ellas, la opción Explorar... proporciona estadísticos y gráficos diversos.

Para el análisis de regresión lo primero que debemos conocer es la forma que presenta el conjunto de los datos cuando se representan en un diagrama, que en el argot del análisis de regresión se conoce como *nube de puntos*. La observación de la misma, nos sugerirá ideas acerca de la curva de ajuste más adecuada.

En SPSS, este tipo de representación se llama *gráfico de dispersión*. En la práctica 3 de este curso hemos hecho uso del mismo, y ahora en los ejercicios que siguen volveremos a hacerlo.

Ejercicio 1

Haz un diagrama de dispersión de los datos de la siguiente tabla

presión (lb × pulg^2)	30	31	32	33	34	35	36
nº de millas (× 1000)	29.5	32.1	36.3	38.2	37.7	33.6	26.8

que representan la duración en millas de neumáticos inflados a diferentes presiones.

SOLUCIÓN: Primero escribimos los datos en el *Editor de datos* usando dos variables que llamaremos **presión** y **millas** y a continuación, para trazar un diagrama de dispersión de estos datos, en la barra de menú pulsamos

Gráficos ⟶ Dispersión/Puntos...

con lo que aparece un cuadro de diálogo en el que tomamos la opción *Dispersión simple*; la pulsación del botón Definir despliega otro cuadro donde elegimos las variables que deseamos, seleccionándolas de la lista de la izquierda, y trasladándolas a los campos situados a la derecha mediante el botón ▶ (tómese **presión** para el *Eje x* y **millas** para el *Eje Y*). No usaremos los otros campos o botones, cuya descripción puede encontrarse al pulsar en cada uno de ellos con el botón derecho del ratón, o bien pulsando el botón Ayuda de este cuadro de diálogo. Por último pulsamos Aceptar. Como consecuencia de estas acciones, en el *Visor de resultados* aparecerá una gráfica en la que los puntos correspondientes a los datos se distribuyen de forma que asemejan una parábola con el vértice hacia arriba. Tal imagen nos sugiere que una ecuación de regresión adecuada podría ser una de segundo grado.

Veamos un segundo ejemplo en la misma línea.

Ejercicio 2

Considera la tabla

x	-0.20	0.20	0.40	0.60	0.70	0.80
y	0.96	1.40	1.56	1.74	1.92	2.04

y traza un diagrama de dispersión con esos datos.

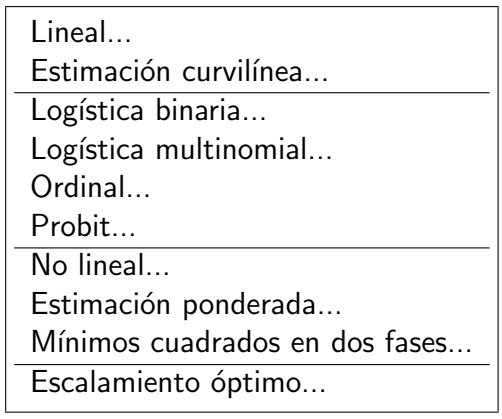
SOLUCIÓN: Como antes, introducimos los datos en el *Editor de datos*, usando para ello dos variables a las que podemos llamar **x** e **y**. Procediendo como en el Ejercicio 1 obtenemos un diagrama de dispersión que evidencia ahora una relación notoriamente lineal entre las dos variables. Para el análisis de regresión tomaríamos pues un modelo lineal.

3 Modelo de regresión lineal

El paquete SPSS realiza diferentes modalidades del análisis de regresión, como puede observarse al pulsar en la barra de menú

Analizar ⟶ Regresión

en cuyo caso aparece un menú desplegable con diversas opciones:



En esta práctica nos limitaremos a las dos primeras opciones de este menú. Para una descripción detallada de todas las opciones, puede consultarse la obra *Técnicas estadísticas con SPSS* citada en la bibliografía, o la ayuda de SPSS que puede obtenerse pulsando en la barra de menú

? ⟶ Temas

y en la ventana de ayuda que aparece se selecciona la pestaña Contenido. En la lista de temas, elegimos Análisis estadístico, y dentro de él, Regresión, apareciendo entonces una lista con las diferentes opciones, de modo que al pulsar cualquiera de ellas, aparece una descripción de la misma.

En el siguiente ejercicio vamos a llevar a cabo un análisis de regresión lineal, para el que vamos a usar los datos del Ejercicio 2. Usaremos muchas, aunque no todas, las prestaciones de SPSS en este tipo de análisis.

Ejercicio 3

Usando los datos del Ejercicio 2, realiza un análisis de regresión lineal.

SOLUCIÓN: Comenzamos pulsando en la barra de menú

Analizar ⟶ Regresión ⟶ Lineal...

El cuadro de diálogo Regresión lineal que aparece nos permite introducir toda la información que necesita SPSS para llevar a cabo el análisis:

- Dependiente. El nombre que le hayamos asignado a la variable de respuesta en el *Editor de datos* (en nuestro caso **y**). Se selecciona de la lista de la izquierda y se traslada al campo correspondiente pulsando el botón ▶ situado entre ambos.
- Independientes. Los nombres de las variables de regresión en el *Editor de datos* elegidos de la lista de la izquierda y trasladados al campo correspondiente como se acaba de explicar. Nos limitaremos al análisis de regresión lineal simple, por lo que sólo usaremos una variable de regresión (la variable **x**).
- Método. Indica la manera cómo SPSS va a incorporar las variables de regresión al modelo. Como estamos en el caso de regresión lineal simple, sólo tenemos una variable, por eso elegiremos la opción *Introducir* (las demás opciones se usan en la regresión lineal múltiple cuando hay más de una variable de regresión).
- Variable de selección. El nombre de una variable del *Editor de datos* que se usa para seleccionar un subconjunto de casos (de valores) de las variables de regresión y de respuesta. No haremos uso de esta opción.
- Etiquetas de caso. El nombre de una variable cuyos valores se usarán como etiquetas para identificar los puntos en un diagrama de dispersión. Esta variable suele ser de tipo cadena. Para hacer uso de esta opción, pulsa la pestaña Vista de variables en la parte inferior de la ventana de SPSS y situado allí, define una variable de tipo *cadena* llamada **rótulos**. Pasa a la Vista de datos pulsando la pestaña correspondiente en la parte inferior de la ventana e introduce en la variable que acabas de definir, los valores:

  punto 1 punto 2 punto 3 punto 4 punto 5 punto 6

El botón Estadísticos... situado en la parte inferior del cuadro de diálogo, permite seleccionar los estadísticos que deseamos que SPSS calcule como parte del análisis de regresión. Al pulsarlo se abre un cuadro en el que podemos elegir entre diversas opciones no excluyentes. Para este ejercicio marca las opciones que siguen:

- *Estimaciones*. Calcula estimaciones b_0 y b_1 de los coeficientes de regresión, usando para ello los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de mínimos cuadrados de los mismos. Además

calcula estimaciones de las desviaciones típicas de estos estimadores usando los estimadores

$$DT(\hat{\beta}_0)_{\text{estimada}} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}\,, \qquad DT(\hat{\beta}_1)_{\text{estimada}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \qquad (*)$$

(Observa que son las mismas expresiones que hemos estudiado en la asignatura). Recuerda que la desviación típica de un estadístico se conoce como el *error típico*, de modo que las anteriores expresiones constituyen estimaciones de los errores típicos de los estadísticos $\hat{\beta}_0$ y $\hat{\beta}_1$.

- *Intervalos de confianza.* Calcula los límites de confianza del 95 % para los coeficientes de regresión β_0 y β_1.

- *Ajuste del modelo.* Calcula los coeficientes de correlación y regresión, así como una estimación de la desviación típica de la población, usando para ello las expresiones

$$R = \sqrt{R^2}\,, \qquad R^2 = \frac{SSR}{S_{yy}}\,, \qquad \hat{\sigma} = \sqrt{\frac{SSE}{n-2}}\,.$$

  Además, lleva a cabo un análisis de varianza (ANOVA) (no lo hemos estudiado en este tema) para medir el ajuste de la ecuación de regresión a los datos.

- *Descriptivos.* Calcula la media y la desviación típica muestrales y el número de casos para cada una de las variables. Además calcula los coeficientes de correlación de Pearson para cada par de variables.

Una vez elegidas las opciones deseadas, el botón Continuar cierra este cuadro, regresando al cuadro anterior, en el que podemos pulsar el botón Aceptar para que SPSS lleve a cabo el anĺisis de regresión lineal simple.

En el *Visor de resultados* aparecen diversos cuadros con los cálculos efectuados:

Estadísticos descriptivos

	Media	Desviación típ.	N
y	1.60333333	0.39159503	6
x	0.41666667	0.37103459	6

En este cuadro figuran la media y la desviación típica muestral de las variables **x** e **y**, así como el número de datos.

Correlaciones

		y	x
Correlación de Pearson	y	1	0.99613224
	x	0.99613224	1
Sig. (unilateral)	y	.	1.1205E-05
	x	1.1205E-05	.
N	y	6	6
	x	6	6

En el cuadro Correlaciones aparecen los *coeficientes de correlación de Pearson* (la raíz cuadrada de los *coeficientes de determinación* R^2), para cada par de variables: x − y, x − x, y − x e y − y. Observa que el coeficiente correlación de una variable consigo misma es 1. Además están los valores P (que SPSS llama Sig.) del contraste de la hipótesis de que dichos coeficientes son cero frente a la hipótesis alternativa de que no lo son. Al ser tan pequeños esos valores, indican que la correlación es significativa.

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	...
		B	Error típ.	Beta	
1	(Constante)	1.16527845	0.02489801		
	x	1.05133172	0.04636788	0.99613224	

a Variable dependiente: y

...			Intervalo de confianza para B al 95%	
	t	Sig.	Límite inferior	Límite superior
	46.8020679	1.2467E-06	1.09615049	1.23440641
	22.6737065	2.241E-05	0.92259384	1.18006959

El resumen de las estimaciones acerca de los coeficientes de regresión figura en el cuadro Coeficientes. Bajo el epígrafe Coeficientes no estandarizados, en la columna encabezada por la letra B están las estimaciones b_0 y b_1 de los coeficientes de regresión β_0 y β_1, y junto a ella bajo el encabezamiento Error típ., los errores típicos de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, que vienen dados por las expresiones (*) de más arriba.
Bajo el epígrafe t aparecen los valores de los estadísticos

$$T = \frac{\hat{\beta}_0}{\hat{\sigma}\sqrt{\dfrac{1}{n} + \dfrac{\overline{x}^2}{S_{xx}}}} \qquad \text{y} \qquad T = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}},$$

que son los estadísticos de prueba para los contrastes de hipótesis

$$\begin{cases} H_0 : \beta_0 = 0; \\ H_1 : \beta_0 \neq 0; \end{cases} \qquad \text{y} \qquad \begin{cases} H_0 : \beta_1 = 0; \\ H_1 : \beta_1 \neq 0. \end{cases}$$

Bajo el epígrafe Sig. están los correspondientes valores P. Observa que los dos son muy pequeños, lo que nos permite en ambos casos rechazar las hipótesis nulas, y en particular el segundo de ellos, nos lleva a concluir que *la regresión es significativa.*
Por último bajo el epígrafe Intervalo de confianza para B al 95% aparecen los extremos de los intervalos de confianza para los parámetros de regresión β_0 y β_1.

Vamos a continuar explorando las posibilidades de SPSS en el análisis de regresión, para lo que volvemos a la barra de menús y pulsamos

Analizar ⟶ Regresión ⟶ Lineal...

De nuevo en el cuadro de diálogo Regresión lineal pulsa otra vez el botón Estadísticos... y desmarca todas las opciones que se marcaron antes. Ahora podemos efectuar un análisis gráfico del ajuste y de los residuos pulsando el botón Gráficos... . Al hacerlo se abre una

caja de diálogo en la que podemos seleccionar las variables que se van a representar en distintos diagramas de dispersión. Algunas de estas variables (que no figuran en el *Editor de datos*, ya que son resultados del análisis de regresión) son

DEPENDNT: Los valores de la variable de respuesta que figuran en la tabla de datos (es decir, la variable que hemos llamado y).

*ZPRED: Los valores calculados con la ecuación de ajuste de la variable de respuesta, tipificados, es decir restándoles su media aritmética y dividiéndolos por su cuasidesviación típica.

*ZRESID: Los residuos, es decir las diferencias entre los valores calculados (con la ecuación de ajuste) y los valores observados (los de la tabla de datos) de la variable de respuesta, también tipificados como la variable anterior.

Selecciona DEPENDNT y *ZPRED, y sitúalas en los campos X e Y, después pulsa el botón Siguiente y coloca DEPENDNT y *ZRESID en ese orden en los mismos campos. Con ello has indicado a SPSS que realice dos gráficos de dispersión. Pulsa Continuar para volver al cuadro anterior y pulsando ahora Aceptar, SPSS efectúa los cálculos.

Los resultados aparecen, como es habitual, en el *Visor de resultados*. Aunque hemos desactivado todas las opciones que antes marcamos en el cuadro de diálogo Estadísticos, SPSS, de una manera algo reiterativa vuelve a mostrar los cuadros Variables introducidas/eliminadas, Resumen del modelo, ANOVA y Coeficientes (aunque este último sin los intervalos de confianza). Ignora estos cuadros que ya se han comentado, y observa los gráficos. Interprétalos. ¿Sabrías explicar por qué en el primero de ellos los puntos están bastante alineados y en el segundo no?

Proseguimos el análisis de regresión, para lo que volvemos a la barra de menús y pulsamos

Analizar ⟶ Regresión ⟶ Lineal...

En el cuadro de diálogo Regresión lineal pulsa otra vez el botón Gráficos... y desactiva los gráficos que se activaron antes, devolviendo las variables de los campos de la derecha al campo de la izquierda, con objeto de que en el próximo análisis, SPSS no vuelva a presentar los gráficos en el *Visor de resultados*. Cierra este cuadro de diálogo y de nuevo en el cuadro Regresión lineal pulsa el botón Guardar... . Éste nos permite guardar los resultados numéricos del análisis (por ejemplo los valores tipificados de las variables de regresión, los intervalos de confianza para la respuesta media, los residuos...) que deseemos, bien como nuevas variables en el archivo de datos actual, bien en un nuevo archivo. De entre las muchas opciones que nos ofrece, vamos a elegir:

- En el campo Valores pronosticados, la opción No tipificados. Se trata de los valores de la variable de respuesta para cada valor de la variable de regresión, que se obtienen al usar la ecuación de regresión ajustada (es decir, aquella cuyos coeficientes son las estimaciones de los coeficientes de regresión).

- En el campo Residuos, de nuevo la opción No tipificados. Los residuos son las diferencias entre los valores de la variable de respuesta medidos u observados, y los pronosticados por el modelo ajustado.

- En el campo Intervalos de pronóstico marcamos (no son excluyentes) ambas opciones Media e Individuos. El nivel de confianza podemos dejarlo tal cual o elegir otro. La primera opción calcula los límites de confianza para la respuesta media correspondientes a cada valor de la variable de regresión que figure en la tabla de datos, y la segunda los límites de los intervalos de predicción correspondientes a esos mismos valores.

No usaremos ninguna otra de las muchas opciones disponibles. El botón Continuar cierra la caja de diálogo aceptando las elecciones que hayamos hecho y volviendo al cuadro Regresión lineal. Al pulsar Aceptar, SPSS efectúa los cálculos pertinentes, mostrando en el *Visor de resultados*, de nuevo algunos cuadros ya comentados antes, más otro cuadro de título Estadísticos sobre los residuos.

Observemos ahora las nuevas variables que han aparecido en el Editor de datos, (las que hemos pedido a SPSS que guarde):

PRE_1: Los valores estimados de la recta de regresión, es decir, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

RES_1: Los residuos, es decir $Y|x - \hat{y}$.

LMCI_1 y UMCI_1: Los extremos inferior y superior del intervalo de confianza para la respuesta media $\mu_{Y|x}$.

LICI_1 y UICI_1: Los extremos inferior y superior del intervalo de predicción al 95% para la variable de respuesta $Y|x$.

Explica por qué SPSS calcula tantos intervalos como valores tiene la variable x.

Una vez que estas nuevas variables están disponibles en el Editor de datos, podemos almacenarlas en un fichero, o usarlas para trazar gráficas.

---

4 Modelo de regresión curvilíneo

De los diversos modelos de regresión que proporciona SPSS, nos limitamos al caso lineal ya estudiado y al curvilíneo. Los ejercicios que siguen, exploran parcialmente las potencialidades de SPSS en estos últimos modelos

Ejercicio 4

Con los datos de la tabla del Ejercicio 1 lleva a cabo un análisis de regresión usando un modelo de regresión curvilínea.

SOLUCIÓN: En el examen exploratorio de los datos observamos que una parábola podría ser una curva que se ajustara bien a los datos, así pues elegiremos como curva de regresión, una de segundo grado. El procedimiento se inicia pulsando en la barra de menú

Analizar ⟶ Regresión ⟶ Estimación curvilínea...

El cuadro de diálogo Estimación curvilínea... que aparece nos permite introducir toda la información que necesita SPSS para llevar a cabo el análisis:

- Dependiente. El nombre que le hayamos asignado a la variable de respuesta en el *Editor de datos* (en nuestro caso **millas**). Se selecciona de la lista de la izquierda y se traslada al campo correspondiente como se ha explicado antes.

- Independientes. Los nombres de las variables de regresión en el *Editor de datos* elegidos de la lista de la izquierda y trasladados al campo correspondiente. Nos limitaremos a un análisis de regresión simple, por lo que sólo usaremos una variable de regresión (la variable **presión**). También puede elegirse como variable independiente el *tiempo*, lo cual SPSS efectúa automáticamente sin que sea necesario tener una variable en el *Editor de datos*.

- Etiquetas de caso. Igual que en la regresión lineal, podemos, como allí, introducir en el *Editor de datos* una variable de tipo cadena de nombre **rótulos** cuyos valores sean

  punto 1 punto 2 punto 3 punto 4 punto 5 punto 6 punto 7

  y usarla para este propósito.

- Incluir constante en la ecuación. Para que la ecuación de regresión tenga término independiente.

- Representar los modelos. Al elegir esta opción, SPSS traza en un diagrama los datos observados (los de nuestra tabla) unidos por una línea quebrada, y la curva de regresión ajustada.

- Modelos. Para elegir el tipo de curva de regresión que deseamos. Puede elegirse más de uno, y si hemos pedido que represente los modelos, dibuja las curvas sobre el mismo diagrama. Los más habituales son:

  1. Regresión lineal: $y = b_o + b_1 t$.
  2. Regresión cuadrática: $y = b_o + b_1 t + b_2 t^2$.
  3. Regresión potencial: $y = b_o t^{b_1}$.
  4. Regresión exponencial: $y = b_o e^{b_1 t}$.
  5. Regresión compuesta: $y = b_o b_1^t$.

- Mostrar tabla ANOVA. Produce una tabla de análisis de varianza del modelo.

En este cuadro de diálogo hay también algunos botones. En la parte inferior, el botón Guardar permite guardar en unas variables nuevas que SPSS define en el *Editor de datos*, algunos de los resultados del análisis de regresión. Al pulsarlo, se abre una caja de diálogo en la que podemos elegir los *Valores pronosticados*, los *Residuos* y los *Intervalos de pronóstico* (es decir, los límites de los intervalos de predicción), así como el nivel de confianza que deseamos para estos intervalos. A diferencia del caso lineal, SPSS no da opción a guardar esta información en un fichero.

De los demás botones, sólo mencionaremos Continuar, Cancelar y Ayuda cuya utilidad la indica su nombre.

Por último, realiza los ejercicios que siguen.

Ejercicio 5

Para la siguiente tabla de datos, traza un diagrama de dispersión que te oriente acerca de la ecuación de regresión más adecuada, y una vez elegida, procede al análisis de regresión. (Sugerencia: usa un modelo cúbico, aunque puedes ensayar además con otro a criterio tuyo después de haber observado el diagrama de dispersión).

x	1.0	1.5	2.0	2.5	3.0	3.5	4.2
y	1.46	2.14	2.63	2.99	3.32	3.69	4.44

SOLUCIÓN: Procede como en el Ejercicio 4.

Ejercicio 6

Para la siguiente tabla de datos, traza un diagrama de dispersión que te oriente acerca de la ecuación de regresión más adecuada, y una vez elegida, procede al análisis de regresión. (Sugerencia: usa un modelo potencial, aunque puedes ensayar además con otro a criterio tuyo después de haber observado el diagrama de dispersión).

x	2	3	6	10	20	30
y	126.1	141.5	170.1	193.2	228.0	250.3

SOLUCIÓN: Como en los dos ejercicios anteriores.

Ejercicio 7

Para estudiar un caso real, abre el fichero `Coches.sav` que se encuentra en el directorio raíz de SPSS y estudia la relación que existe entre las variables *consumo* y *peso*. Para ello, analiza el comportamiento de dichas variables calculando los estadísticos más usuales y los histogramas. Decide si la regresión lineal es aceptable y, aplicando *regresión curvilínea*, compara los resultados para los modelos *lineal*, *logarítmico* y *exponencial*.

5 Bibliografía

Manual de SPSS de la Universidad de Cádiz.
http://www2.uca.es/serv/ai/formacion/spss/Inicio.pdf

Cuaderno de prácticas de SPSS de la asignatura Análisis de datos en Psicología I. Universidad Autónoma de Madrid.
http://www.uam.es/personal_pdi/psicologia/carmenx/MaterialD.html

Manual de SPSS de la Universidad de Cádiz.
http://www2.uca.es/serv/ai/formacion/spss/Inicio.pdf

Pérez, César, *Técnicas Estadísticas con SPSS*. Prentice Hall

Portilla, M. et al. *Manual práctico del paquete estadístico SPSS 9 para Windows*. Universidad Pública de Navarra.