Capítulo 6: Regresión múltiple

Utilizamos regresión múltiple cuando estudiamos la posible relación entre varias variables *independientes* (*predictoras* o *explicativas*) y otra variable dependiente (*criterio*, *explicada, respuesta*).

Por ejemplo, podemos estar interesados en estudiar la inteligencia humana (IQ como variable respuesta), y es posible que consideremos que puede estar relacionado con otras variables como el tamaño del cerebro (explicativa). Es posible que el tamaño de la persona y su sexo también deban ser tenidos en cuenta. Podríamos añadirlas al estudio como variables independientes. Un modelo de regresión podría ofrecer una respuesta como:

IQ = 80 + 0.02 Volumen cerebro + 0.15 Tamaño -0.8 Sexo,

donde la variable sexo es una variable dicotómica o indicadora, codificada como 0 para las mujeres y 1 para los hombres. Para interpretar un modelo así hay que ser muy cautelosos. Las modelos de regresión nos informan de la presencia de relaciones, pero no del mecanismo causal. Por ejemplo muchos conductores asocian que cuanta más policía local hay dirigiendo el tráfico, mayores son los atascos y concluyen erróneamente que es la policía la causa de los atascos. Olvidan terceras variables que no han sido tenidas en cuenta como las averías previas en los semáforos o la ocurrencia de accidentes.

Otra fuente de problemas de interpretación es la relación entre variables independientes (colinealidad). Por ejemplo el sexo puede parecer influir en la inteligencia mirando inocentemente la ecuación, pero hay que considerar que las mujeres son habitualmente más pequeñas que los hombres. Si observamos los signos, apreciamos que compensa el efecto de una con la otra.

La técnica de regresión múltiple se usa frecuentemente en investigación. Se aplica al caso en que la variable respuesta es de tipo numérico. Cuando la respuesta es de tipo dicotómico (muere/vive, enferma/no enferma), usamos otra técnica denominada regresión logística y que tratamos en un capítulo posterior.

6.1 Aplicaciones de la regresión múltiple

Es cierto que la regresión múltiple se utiliza para la predicción de respuestas a partir de variables explicativas. Pero no es ésta realmente la aplicación que se le suele dar en investigación. Los usos que con mayor frecuencia encontraremos en las publicaciones son los siguientes:

- **Identificación de variables explicativas.** Nos ayuda a crear un modelo donde se seleccionen las variables que puedan influir en la respuesta, descartando aquellas que no aporten información.
- **Detección de interacciones** entre variables independientes que afectan a la variable respuesta. Un ejemplo de interacción clásico es el de estudiar la respuesta de un paciente al alcohol y a un barbitúrico, y observar que cuando se ingieren ambos, el efecto es mucho mayor del esperado como suma de los dos.

- **Identificación de variables confusoras.** Es un problema difícil el de su detección, pero de interés en investigación no experimental, ya que el investigador frecuentemente no tiene control sobre las variables independientes.

6.2 Requisitos y limitaciones

Hay ciertos requerimientos necesarios para poder utilizar la técnica de regresión múltiple:

- **Linealidad**: Se supone que la variable respuesta depende linealmente de las variables explicativas. Si la respuesta no aparenta ser lineal, debemos introducir en el modelo componentes no lineales (como incluir transformaciones no lineales de las variables independientes en el modelo). Otro tipo de respuesta no lineal es la interacción. Para ello se ha de incluir en el modelo términos de interacción, que equivalen a introducir nuevas variables explicativas que en realidad son el producto de dos o más de las independientes.
- **Normalidad y equidistribución de los residuos**: Se llaman residuos las diferencias entre los valores calculados por el modelo y los realmente observados en la variable dependiente. Para tener un buen modelo de regresión no es suficiente con que los residuos sean pequeños. La validez del modelo requiere que los mismos se distribuyan de modo normal y con la misma dispersión para (siéntese antes de leer el resto de la frase) ¡cada combinación de valores de las variables independientes!
  Por supuesto, esta condición en la práctica es inverificable, puesto que para cada combinación de variables independientes tendremos normalmente ninguna o una respuesta. Lo que se suele hacer es examinar una serie de gráficos de residuos que nos hagan *sospechar.* Por ejemplo si los residuos aumentan al aumentar la respuesta, o vemos que aparecen tendencias,… Es decir, hay una serie de reglas heurísticas que nos ayudan a decidir si aceptar o no el modelo de regresión, pero no están basadas en contrastes de hipótesis como hemos usado hasta ahora. Es la experiencia del investigador observando residuos la que le decide a usarlo o no.
- **Número de variables independientes**: Podemos estar tentados en incluir en el modelo cualquier cosa que tengamos en una base de datos, con la esperanza de que cuantas más variables incluyamos, más posibilidades hay de que "*suene la flauta*". Si nos aborda esta tentación, hemos de recordar que corremos el riesgo de cometer error de tipo I. Otra razón es que si esperamos ajustar unas pocas observaciones usando muchas variables, muy probablemente consigamos una aproximación muy artificial, y además muy sensible a los valores observados. La inclusión de una nueva observación puede cambiar completamente el valor de los coeficientes del modelo. Esto se traducirá al realizar el contraste como justo todo lo contrario de lo que deseábamos: ¡Todas las variables independientes del modelo serán consideradas no significativas!
  Una regla que se suele recomendar es la de incluir al menos 20 observaciones por cada variable independiente que estimemos *a priori* interesantes en el modelo. Números inferiores nos llevarán posiblemente a no poder obtener conclusiones y errores de tipo II.
- **Colinealidad**: Si dos variables independientes están estrechamente relacionadas (consumo de refrescos y temperatura ambiente por ejemplo) y ambas son incluidas en un modelo, muy posiblemente ninguna de las dos sea considerada significativa, aunque si hubiésemos incluido sólo una de ellas, sí. Hay diferentes técnicas para detectar la colinealidad pero que requieren profundizar en

documentos más sofisticados. Aquí vamos a indicar una técnica muy simple: examinar los coeficientes del modelo para ver si se vuelven inestables al introducir una nueva variable. Si es así posiblemente hay colinealidad entre la nueva variable y las anteriores.

- **Observaciones anómalas**: Está muy relacionada con la cuestión de los residuos, pero merece destacarlo aparte. Debemos poner especial cuidado en identificarlas (y descartarlas si procede), pues tienen gran influencia en el resultado. A veces, son sólo errores en la entrada de datos, pero de gran consecuencia en el análisis. Hay técnicas de regresión robustas que permiten minimizar su efecto.

6.3 Variables numéricas e indicadoras (dummy)

Un modelo de regresión lineal tiene el aspecto:

$$Y = b_0 + b_1 X_1 + \cdots + b_n X_n$$

- Y es la variable dependiente
- Los términos X_i representan las variables independientes o explicativas
- Los coeficientes del modelo b_i son calculados por el programa estadístico, de mode que se minimicen los residuos.
  Esencialmente cuando obtengamos para los coeficientes valores "compatibles" con cero (no significativos), la variable asociada se elimina del modelo, y en otro caso se considera a la variable asociada de interés. Esta regla no hay que aplicarla ciegamente. Si por ejemplo la variable con coeficiente no significativo se observa que es confusora, debemos considerarla como parte del modelo, bien explícitamente o estratificando la muestra según los diferentes valores de la misma.

Está claro que para ajustar el modelo la variable respuesta debe ser numérica. Sin embargo, aunque pueda parecer extraño no tienen por qué serlo las variables explicativas. Aunque requiere un artificio, podemos utilizar predictores categóricos mediante la introducción de variables indicadoras (también denominadas mudas o *dummy*)

Si una variable es dicotómica, puede ser codificada como 0 ó 1. Así si estudiamos la explicación del peso de una persona como función de su altura y su sexo, un modelo como:

$Peso = -100 + 1 \cdot Altura - 5 \cdot Sexo$

donde se ha codificado con Sexo=0 a los hombres y Sexo=1 a las mujeres, puede ser interpretado como que las mujeres, a igualdad de altura, pesan de media 5 Kg menos que los hombres. El coeficiente 1 de la altura, se interpreta como que por cada diferencia de altura de un centímetro en personas que tienen el resto de variables independientes iguales (mismo sexo), el peso aumenta, por término medio, en un kg.

Si creemos que la dieta puede influir en la respuesta, y tenemos 3 dietas posibles (es decir, hay un factor llamado dieta, que es variable categórica con tres modalidades), como por ejemplo *dieta normal*, *alta en proteínas* y *vegetariana*, podemos introducirlo usando dos variables indicadoras creadas por nosotros, *indProteína*, *indVegetal*, de manera que recodifiquemos el factor "*Dieta*" usando las nuevas variables como sigue:

	indProteína	**indVegetal**

Dieta		
Normal (grupo control)	0	0
Alta en proteínas	1	0
Vegetariana	0	1

Estéticamente no hemos ganado mucho, pero observe que ahora es fácil interpretar un modelo como sigue:

$$Peso = -100 + 1 \cdot Altura - 5 \cdot Sexo + 4 \cdot ind\,\text{Pr}\,oteína - 6 \cdot indVegetal$$

- Por cada centímetro de altura que un individuo supere a otro, se espera un aumento de peso de 1kg (a igualdad del resto de variables).
- Las mujeres pesan de media 5 kg menos que los hombres (a igualdad del resto de variables).
- Si un individuo sigue una dieta alta en proteínas pesa 4 kg. más de media que un individuo control (dieta normal, indProteína=indVegetal=0) cuando todas las demás variables coinciden.
- Si un individuo sigue una dieta vegetariana pesa 6 kg. menos de media que un individuo cuando todas las demás variables coinciden.

Observe que el modelo de regresión múltiple generaliza a otras técnicas estadísticas que conocemos a estas alturas como el modelo t-student para 2 muestras independientes o ANOVA de un factor. Un contraste de dos medias independientes puede resolverse con una regresión de la variable respuesta en función de una variable indicadora que identifica sendas muestras. Un modelo ANOVA de un factor se puede expresar usando variables indicadoras suficientes para codificar el grupo. Para poder interpretar cómodamente los resultados es importante que los grupos sean equilibrados (cada muestra debe tener un número similar de elementos).

6.4 Interpretación de los resultados

En la sección anterior hemos interpretado sólo una parte del modelo, pero hay muchos términos que es necesario conocer para poder contrastar hipótesis. Es más, son tantos que hay que tener cuidado con no cometer errores de tipo I debido a que por puro azar, obtengamos resultados significativos donde no debería haberlos (por la misma razón que si compramos muchos billetes en un tómbola puede tocarnos un premio, sin que seamos personas especialmente afortunadas).

- **La significación del modelo de regresión**: La hipótesis nula es que la variable respuesta no está influenciada por las variables independientes. Dicho de otro modo, la variabilidad observada en las respuestas son causadas por el azar, sin influencia de las variables independientes. La hipótesis alternativa es que hay algún tipo de influencia. La significación del contraste se calcula haciendo un análisis de la varianza.
- **Los coeficientes**: Los programas estadísticos ofrecen una estimación de los mismos, junto a un error típico de la estimación, un valor de significación, o mejor aún, un intervalo de confianza. Una aplicación interesante del mismo es la siguiente: Si la significación es pequeña, el intervalo no contiene el valor cero. Esto lo consideramos como una indicación de que esa variable es interesante en el modelo. Si contiene al cero (no significativa), posiblemente sea preferible eliminarla del modelo para simplificar. Pero, ¡atención!, si al hacerlo otros

coeficientes cambian muy claramente, incluso pasando de positivos a negativos siendo de nuevo significativos, estamos posiblemente ante una variable confusora. Encontrar este tipo de variables es uno de los objetivos en regresión, y debemos conservarlas para cualquier interpretación aunque sus coeficientes no sean significativos.

- **La bondad del ajuste**: Hay un término denominado *R cuadrado*, que se interpreta del siguiente modo. La variable respuesta presenta cierta variabilidad (incertidumbre), pero cuando se conoce el valor de las variables independientes, dicha incertidumbre disminuye. El término *R cuadrado* es una cantidad que puede interpretarse como un factor (porcentaje) de reducción de la incertidumbre cuando son conocidas las variables independientes. Cuanto más se acerque a uno, más poder explicativo tendrá el modelo. Pero esto esconde una trampa. Cada vez que introducimos una nueva variable independiente en el modelo, *R cuadrado* no puede hacer otra cosa que aumentar. Si introducimos un número artificialmente grande de ellas, podremos llegar a acercarla a uno tanto como queramos.
  Los programas estadísticos nos muestran un término *R cuadrado corregida*, que puede interpretarse como una corrección de honestidad. Nos castigará disminuyendo cuando introduzcamos variables innecesarias. Si al ir complicando el modelo este término aumenta una cantidad "razonable", podemos considerarlo posiblemente una variable de interés, pero si disminuye, deberíamos pensar dos veces si nos merece la pena la complejidad del modelo para tan poco beneficio.
- **La matriz de correlaciones**: Nos ayudan a identificar correlaciones lineales entre pares de variables. Encontrar correlaciones lineales entre la variable dependiente y cualquiera de las independientes es siempre de interés. Pero es una mala señal la correlación entre variables independientes. Alguna de las dos debería salir del modelo.
  La matriz de correlaciones está formada por todos los **coeficientes de correlación lineal de Pearson** para cada par de variables. Los mismos son cantidades que pueden tomar valores comprendidos entre -1 y +1. Cuanto más extremo sea el coeficiente, mejor asociación lineal existe entre el par de variables. Cuando es cercano a cero, no. El signo positivo del coeficiente nos indica que la asociación es directa (cuando una variable crece la otra también). Un valor negativo indica que la relación es inversa (cuando una crece, la otra decrece).

6.5 Variables confusoras

Dos variables o más variables están confundidas cuando sus efectos sobre la variable dependiente no pueden ser separados. Dicho de otra forma, una variable es confusora cuando estando relacionada con alguna variable independiente, a su vez afecta a la dependiente.

Como ilustración consideremos un estudio en la que se consideran familias monoparentales y se relaciona el tiempo que un progenitor dedica diariamente a su hijo, y el momento en el que éste empieza a hablar. No es difícil imaginar razones por la que el sexo del progenitor sea variable confusora. Si el estudio no es experimental (sería difícil hacerlo compatible con la ética), no conseguiremos muchas madres que dediquen

muy poco tiempo a los hijos, ni muchos padres que le dediquen muchísimo. Además seguramente el sexo del progenitor influirá en la respuesta.

Cuando se identifica una variable que está confundida con alguna de las variables independientes significativa, es necesario dejarla formar parte del modelo, tenga o no mucha significación. *Las variables confusoras no pueden ser ignoradas*.

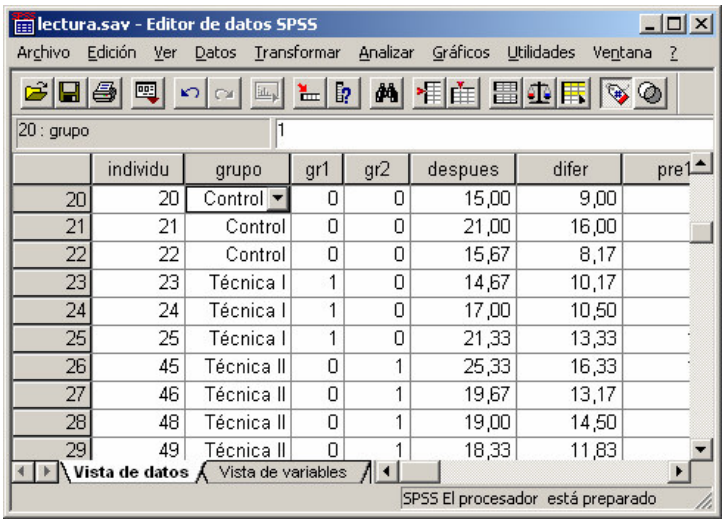
Puede ayudarnos a identificar una variable confusora el encontrarnos con un modelo en el que una variable independiente parece tener cierta influencia significativa del signo que sea en la variable respuesta, pero al incluir una nueva variable previamente ignorada (variable confusora) se observa que la primera tiene una influencia claramente diferente (incluso con el signo cambiado y aún significativa). El simple hecho de que lo anteriormente mencionado ocurra, no es la prueba de que ambas variables estén confundidas, pero nos invita a reflexionar. Es de mayor utilidad el estudio de los residuos, pero aquí no entraremos en ello.

Ejemplo: Vamos a retomar un ejemplo que desarrollamos en el capítulo donde tratábamos la técnica ANOVA, para usar con él un modelo de regresión múltiple. Recordamos que se realizó un experimento para comparar tres métodos de aprendizaje de lectura. Se asignó aleatoriamente los estudiantes a cada uno de los tres métodos. Cada método fue probado con 22 estudiantes. Se evaluó mediante diferentes pruebas la capacidad de comprensión de los estudiantes, antes y después de recibir la instrucción. Por tanto tenemos 3 variables numéricas que son la capacidad al inicio del experimento, al final, y la que resulta más interesante, la diferencia. Se encontró evidencia estadísticamente significativa a favor de que las medias en los tres grupos son diferentes, pero no pudimos concluir mucho más.
Nos interesa ahora reformular la cuestión usando análisis de regresión. En este caso proponemos estudiar el siguiente modelo:

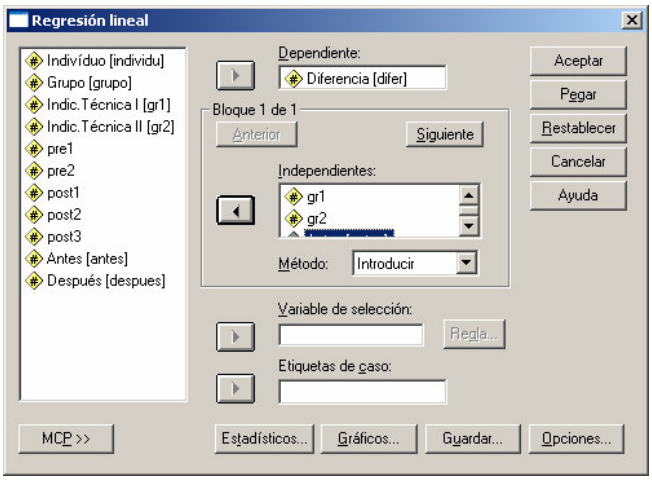
- Variable dependiente: La diferencia entre la capacidad “después” y “antes”.
- Variables explicativas:
    - La capacidad al inicio del experimento. Posiblemente los estudiantes con mejor capacidad inicial sacaron menos provecho que el resto.
    - La técnica utilizada. Como es una variable categórica que se utiliza para identificar la muestra y tiene tres categorías podemos codificarla usando dos variables indicadoras:

Grupo	Indic. Técnica I	Indic. Técnica II
Control	0	0
Técnica I	1	0
Técnica II	0	1

Las variables indicadoras se aprecian en la captura de pantalla se la ventana de datos de SPSS como variables *gr1* y *gr2*.



Para utilizar un modelo de regresión múltiple, elegimos la opción del menú "Analizar – Regresión - Lineal…", y situamos en su lugar a las variables independientes y la dependiente.



Observamos en el resultado del análisis de la varianza que el modelo resulta significativo (p aproximadamente cero). Por tanto rechazamos la hipótesis nula de que la variabilidad observada en la variable respuesta sea explicable por el azar, y admitimos que hay algún tipo de asociación entre la variable dependiente y las independientes.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	237,770	3	79,257	12,248	,000[a]
	Residual	401,192	62	6,471		
	Total	638,962	65			

a. Variables predictoras: (Constante), Antes, Indic.Técnica I, Indic.Técnica II

b. Variable dependiente: Diferencia

En la siguiente tabla tenemos información sobre los coeficientes. El resultado se lee del siguiente modo:

- **El modelo ajustado** de regresión lineal múltiple es:

$$\text{Diferencia} = 13.557 + 3.406 \cdot \text{Técnica I} + 2.827 \cdot \text{Técnica II} - 0.467 \cdot \text{Antes}$$

- **Término constante**: vale 13.557; Se puede interpretar como la diferencia en puntuación obtenida para un individuo del grupo de control, que al empezar tuviese una puntuación 0. Si contrastamos si el mismo vale 0, se rechaza la hipótesis (p aproximadamente 0).
- **Término para la variable indicadora de la técnica I**: Es significativo (p aproximadamente 0), es decir, se rechaza que sea nulo. Vale 3.4 con un intervalo de confianza al 95% que va desde 1.87 a 4.95; El valor del coeficiente nos indica cuánto esperamos que aumente de la variable respuesta, en caso de que a un estudiante decidiésemos aplicarle la técnica I frente a la técnica del grupo de control.
- **Término para la variable indicadora de la técnica II**: Es significativo (p=0.001), es decir, se rechaza que sea nulo (para el nivel de significación habitual de 0.05); Vale 2.8 con un intervalo de confianza al 95% que va desde 1.27 a 4.38; El valor del coeficiente nos indica cuánto esperamos que aumente la variable respuesta, en caso de que a un estudiante decidiésemos aplicarle la técnica II frente a la técnica del grupo de control.
- **Término para la variable "Antes"**: Es significativo (p=0.003); es decir, se rechaza que sea nulo (para el nivel de significación habitual de 0.05); Vale -0.47 con un intervalo de confianza al 95% que va desde -0.76 a -0.17; El valor negativo del coeficiente nos indica que sea cual sea la técnica de aprendizaje que se aplique a un estudiante, se espera que disminuya en -0.46 unidades por cada punto obtenido en la prueba inicial de capacidad (con un margen de error, como se aprecia en el intervalo de confianza).

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Intervalo de confianza para B al 95%	
		B	Error típ.	Beta			Límite inferior	Límite superior
1	(Constante)	**13,557**	1,292		10,494	,000	10,975	16,140
	Indic.Técnica I	**3,406**	,770	,516	4,422	,000	1,866	4,945
	Indic.Técnica II	**2,827**	,777	,428	3,637	,001	1,273	4,380
	Antes	**-,467**	,149	-,321	-3,144	,003	-,765	-,170

a. Variable dependiente: Diferencia

Para medir la bondad del ajuste tenemos el término R cuadrado y R cuadrado corregida. Que R cuadrado sea igual a 0.372, se puede interpretar de la siguiente forma: Elegido un individuo al azar, del que no sabemos nada, tenemos una cierta incertidumbre (varianza) de cuál será el valor de la variable respuesta. Si disponemos de información adicional sobre las variables independientes (el grupo al que será asignado y su capacidad inicial), gracias al modelo lineal de regresión, podemos hacer una predicción donde la incertidumbre (varianza) está disminuida en un 37.2% con respecto a la original.

Resumen del modelo[b]

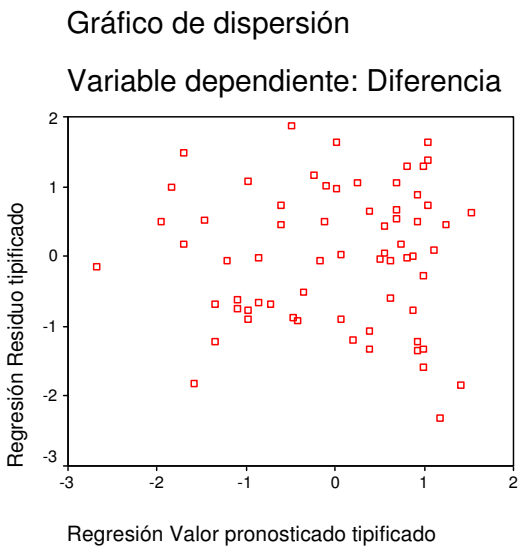
Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,610[a]	,372	,342	2,54378

a. Variables predictoras: (Constante), Antes, Indic.Técnica I, Indic.Técnica II

b. Variable dependiente: Diferencia

Estamos también interesados en saber si se dan las condiciones de validez para los contrastes que se han realizado. Como mencionamos anteriormente, es difícil de contrastar, y simplemente se suele inspeccionar los residuos para ver si detectamos algún alejamiento de las mismas. Para ello pulsamos el botón "Graficos…" de la ventana donde especificamos el modelo de regresión lineal. El análisis de residuos puede ser muy extenso. A continuación tiene la combinación que hemos elegido, junto con algunos comentarios de parte de los resultados (la salida es muy extensa y no la analizamos completa).

En el diagrama de dispersión del valor pronóstico tipificado (ZPRED), frente al residuo tipificado (ZRESID), observamos que los puntos se distribuyen en una banda horizontal con respecto al eje de abcisas. No se aprecia ninguna tendencia especial. No encontramos nada que nos haga sospechar de la falta de adecuación del modelo de regresión.



En el gráfico de probabilidad normal que sigue a continuación, si los puntos se alejan mucho de la línea diagonal, se lee como falta de normalidad. Se aprecia un cierto alejamiento de la normalidad, pero no excesivo. La práctica en examinar este tipo de gráficos ayuda a diferenciar lo que es normal y lo que no.