La práctica del análisis de correspondencias

MICHAEL GREENACRE

Catedrático de Estadística en la Universidad Pompeu Fabra

Separata del capítulo 14

Relaciones de transición y regresión

Primera edición: julio 2008
ISBN: 978-84-96515-71-0

Traducción: Jordi Comas Angelet
Revisión: Carles M. Cuadras Avellana

www.fbbva.es

CAPÍTULO 14

Relaciones de transición y regresión

En AC realizamos mapas en los que representamos conjuntamente filas y columnas como puntos cuya interpretación depende de las escalas escogidas para filas y columnas. Hemos visto cómo geométricamente las posiciones de las filas dependen de las posiciones de las columnas, y viceversa. En este capítulo nos centraremos en las relaciones matemáticas existentes entre las filas y las columnas; las ecuaciones de transición. Además, dado que el análisis de regresión es un método estadístico bien conocido, mostraremos cómo las coordenadas de las filas y las coordenadas de las columnas se pueden relacionar con los datos originales a través de modelos de regresión lineal. En realidad podríamos omitir este capítulo sin perder el hilo sobre la interpretación geométrica del AC.

Contenido

Las coordenadas en el primer eje del ejemplo sobre la financiación de la investigación científica

En este capítulo estamos interesados en las relaciones existentes entre las coordenadas principales y las coordenadas estándares de filas y de columnas, que emanan del AC, así como en su relación con los datos originales. Empecemos por fijarnos en las relaciones existentes en los ejes principales. En la tabla de la imagen 14.1 mostramos todos los resultados del primer eje principal del ejemplo

Imagen 14.1:
Coordenadas principales y coordenadas estándares de las disciplinas científicas y de las categorías de financiación en el primer eje principal del AC (datos originales en la imagen 10.1)

Disciplina científica	*Coordenada principal*	*Coordenada estándar*	Categoría de financiación	*Coordenada principal*	*Coordenada estándar*
Geología	0,076	0,386	*A*	0,478	2,417
Bioquímica	0,180	0,910	*B*	0,127	0,643
Química	0,038	0,190	*C*	0,083	0,417
Zoología	–0,327	–1,655	*D*	–0,390	–1,974
Física	0,316	1,595	*E*	–0,032	–0,161
Ingeniería	–0,117	–0,594			
Microbiología	0,013	0,065			
Botánica	–0,179	–0,904			
Estadística	0,125	0,630			
Matemáticas	0,107	0,540			

sobre la financiación de la investigación científica. Este eje tiene una inercia de $\lambda_1 = 0{,}03912$, siendo $\sqrt{\lambda_1} = 0{,}1978$. En el capítulo 8 vimos que este último valor es el factor de escala que relaciona coordenadas principales con coordenadas estándares. También vimos que lo podemos interpretar como un coeficiente de correlación entre las coordenadas de las filas y columnas en la primera dimensión. Dado que la correlación está relacionada con la regresión, en primer lugar veremos la regresión de las coordenadas de las filas sobre las de las columnas y viceversa.

Regresión entre coordenadas

En la imagen 8.5, utilizando los datos sobre la encuesta de la salud, mostramos el diagrama de dispersión de las coordenadas de las filas sobre las coordenadas de las columnas en el primer eje principal, de todos los individuos de la tabla de contingencia. En la imagen 14.2 mostramos el mismo tipo de representación gráfica para las coordenadas estándares de los datos sobre financiación de la investigación científica. En esta última imagen aparecen 50 cuadrados que corresponden a las 50 celdas de la tabla de contingencia de la imagen 10.1. En el diagrama, los cuadrados se sitúan en los correspondientes valores de filas y columnas de la tabla y tienen un área proporcional al número de individuos (científicos) de las celdas. Sabemos que la correlación, calculada para los 796 individuos representados por los 50 puntos de este diagrama, es de 0,1978. Ahora nos interesamos en la regresión de las disciplinas científicas sobre las categorías de financiación, y de las categorías de financiación sobre las disciplinas científicas. Para llevar a cabo este análisis de regresión haremos un listado en el que asignaremos, según los valores de la tabla de la imagen 14.1, a cada uno de los 796 científicos sus correspondientes pares de valores. Por ejemplo, las coordenadas estándares de un geólogo en la categoría *A*, son 0,386 (para la variable *y*) y 2,417 (variable *x*). Dado que sólo existen 50 pares distintos, una alternativa sería hacer un listado con únicamente los 50 pares en los que, junto con los valores de las coordenadas, aparezcan sus frecuencias y luego llevar a cabo una regresión ponderada tomando como pesos

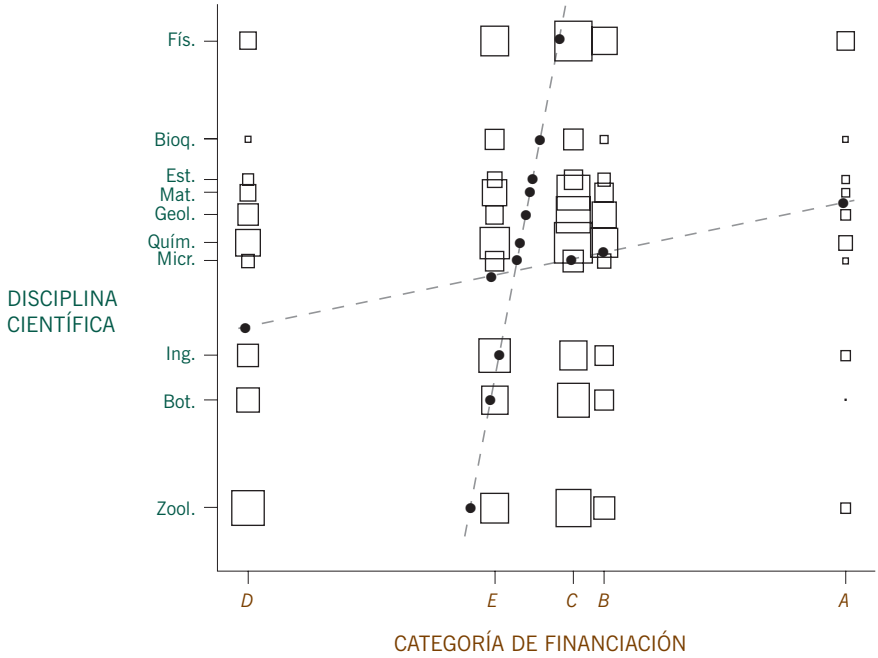

Imagen 14.2:
Diagrama de dispersión de las coordenadas estándares de las filas sobre las coordenadas estándares de las columnas en la primera dimensión del AC (imagen 14.1). Los cuadrados se sitúan en cada combinación de valores, con áreas proporcionales al número de individuos. Las dos rectas de regresión, de filas sobre columnas y de columnas sobre filas, tienen pendientes de 0,1978 y de 5,056, siendo cada una de ellas el valor inverso de la otra. Los puntos • indican medias condicionales (medias ponderadas), es decir, las coordenadas principales

las frecuencias (lo veremos en el apéndice de cálculo, B). Un resultado bien conocido de la regresión lineal simple es que la pendiente es igual a la correlación multiplicada por el cociente de la desviación típica de la variable *y* con la de la variable *x*. Las varianzas de las filas y de las columnas en coordenadas estándares son iguales a 1; por tanto, la pendiente de la recta de regresión de *y* sobre *x* será igual al coeficiente de correlación, concretamente igual a 0,1978 (imagen 14.2). De forma simétrica, la regresión de *x* sobre *y* tendrá también una pendiente de 0,1978, pero en este caso *x* estaría situada en el eje vertical y la *y* en el horizontal; sin embargo, dado que en el diagrama de la imagen 14.2 hemos situado *y* en el eje vertical, la pendiente entre *x* e *y* será 1/0,1978 = 5,056.

La relación entre perfiles y vértices

En el capítulo 3 vimos que la posición de los perfiles fila resulta de calcular las medias ponderadas de los vértices de las columnas, siendo los pesos los valores de los perfiles fila. Para los perfiles columna y los vértices fila se cumple la misma relación. Estas relaciones, basadas en el cálculo de medias ponderadas, también se cumplen para las proyecciones de filas y columnas sobre cualquier subespacio. En concreto, tal como vimos en el capítulo 8, se cumplen para las proyecciones de las coordenadas en los ejes principales. Es decir, en un eje principal *k*, las posiciones de las coordenadas principales de las filas son medias ponderadas de las coordenadas estándares de las columnas, y viceversa. En la imagen 14.2, ilustramos esta relación en el primer eje principal; hemos calculado las posiciones de

las filas —que mostramos como puntos, a partir de medias ponderadas de las coordenadas estándares de las columnas, y viceversa—. Podemos ver que las mencionadas coordenadas principales se sitúan en dos rectas.

En regresión, las coordenadas principales son medias condicionadas

Una regresión es un modelo de medias condicionales de la variable respuesta con relación a la variable explicativa. Los puntos negros de la imagen 14.2 no son más que medias condicionales de *y* sobre *x* (cinco medias en la recta de pendiente 0,1978) y de *x* sobre *y* (diez medias en la recta de pendiente 5,056). Estas medias condicionales son las coordenadas principales que, como podemos ver, en el diagrama definen dos funciones de regresión. Así, por ejemplo, la primera coordenada principal de Física, que mostramos como un punto negro en la parte superior del diagrama, es la media condicionada de las categorías de financiación expresadas en coordenadas estándares en el primer eje principal y ponderadas con las respectivas frecuencias de la matriz de datos de la imagen 10.1 (en este diagrama las hemos simbolizado por los cuadrados situados en la misma vertical que la de coordenada estándar de Física y, horizontalmente, según las coordenadas estándares de las columnas). De manera similar, hemos representado como un punto negro a la derecha la primera coordenada principal de *A*, que es la media condicionada de las categorías científicas expresadas en coordenadas estándares en el primer eje principal y ponderadas con las respectivas frecuencias de la matriz de datos de la imagen 10.1, que hemos simbolizado con los cuadrados situados en la misma horizontal que la coordenada estándar de *A* y, verticalmente, según las coordenadas estándares de las disciplinas científicas. Por tanto, en la imagen 14.2 mostramos simultáneamente coordenadas principales y coordenadas estándares. Las coordenadas principales de las filas son las coordenadas de los diez puntos situados sobre la recta de regresión en la escala de las coordenadas estándares de las columnas (escala horizontal) y viceversa.

Regresiones lineales simultáneas

El hecho de que en el AC las regresiones de *y* sobre *x* (filas sobre columnas) y de *x* sobre *y* (columnas sobre filas) sean rectas dio lugar a que inicialmente se presentara el AC como un sistema de *regresiones lineales simultáneas*. Si la correlación entre filas y columnas es alta, entonces las dos rectas de regresión serán muy parecidas y, en consecuencia, las coordenadas principales se hallarán más separadas; es decir, la inercia será mayor (recordemos que la inercia principal es igual al cuadrado de la correlación). Es decir, el AC se podría definir como un método que trata de buscar rectas de regresión simultáneas (como las mostradas en la imagen 14.2) que formen el menor ángulo posible entre ellas, lo que es equivalente a maximizar la correlación entre filas y columnas.

Ecuaciones de transición entre filas y columnas

Utilizando la notación que vimos anteriormente (págs. 51 y 139), y recordando que las coordenadas principales correspondan a perfiles y que las coordenadas estándares correspondan a vértices, podemos expresar la relación entre filas y columnas basada en el cálculo de medias ponderadas (o medias condicionales), de la manera siguiente:

perfil fila ← vértices de las columnas: $$f_{ik} = \sum_j \left(\frac{p_{ij}}{r_i} \right) \gamma_{jk} \qquad (14.1)$$

perfil columna ← vértices de las filas: $$g_{jk} = \sum_i \left(\frac{p_{ij}}{c_j} \right) \phi_{ik} \qquad (14.2)$$

(← simboliza «obtenido de», por ejemplo, «perfil fila ← vértices de las columnas» indica que hemos obtenido las coordenadas principales de una fila a partir las coordenadas estándares de todas las columnas utilizando la relación mostrada en la ecuación). Utilizamos la notación f y g para las coordenadas principales de las filas y de las columnas, respectivamente; y γ y ϕ para las coordenadas estándares de las filas y de las columnas, respectivamente. Para las filas utiizamos el subíndice i, para las columnas el subíndice j, y para dimensiones el subíndice k. Entre paréntesis mostramos los pesos, que son los perfiles de las filas en (14.1) y los de las columnas en (14.2). Llamamos *ecuaciones de transición* a las medias ponderadas de las expresiones (14.1) y (14.2). Recordemos que las relaciones existentes entre las coordenadas principales y las coordenadas estándares son:

perfil fila ← vértice fila: $$f_{ik} = \sqrt{\lambda_k}\, \phi_{ik} \qquad (14.3)$$

perfil columna ← vértice columna: $$g_{jk} = \sqrt{\lambda_k}\, \gamma_{jk} \qquad (14.4)$$

donde λ_k es la inercia principal (valor propio) del k-ésimo eje. Luego, las ecuaciones de transición entre las coordenadas principales de las filas y las coordenaas principales de las columnas, las podemos expresar de la siguiente manera:

perfil fila ← perfiles columna: $$f_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{r_i} \right) g_{jk} \qquad (14.5)$$

perfil columna ← perfiles fila: $$g_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_{ij}}{c_j} \right) f_{ik} \qquad (14.6)$$

y, de manera similar, podemos expresar las ecuaciones de transición entre las coordenadas estándares de filas y las coordenadas estándares de columnas como:

vértice fila ← vértices columna: $$\phi_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_j \left(\frac{p_{ij}}{r_i} \right) \gamma_{jk} \qquad (14.7)$$

vértice columna ← vértices fila: $$\gamma_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_i \left(\frac{p_{ij}}{c_j} \right) \phi_{ik} \qquad (14.8)$$

Regresión entre coordenadas usando ecuaciones de transición

Podemos utilizar cualquiera de las ecuaciones de transición anteriores, para «estimar» por regresión las coordenadas a partir de los perfiles (variables explicativas). Por ejemplo, supongamos que queremos utilizar la ecuación (14.1) para obtener las coordenadas estándares de las columnas. Las variables respuesta serán las primeras coordenadas principales de las filas (primera columna de la imagen 14.1), y las variables explicativas los diez perfiles de las filas de la matriz 10 × 5 de la imagen 10.1. El análisis de la regresión da los siguientes coeficientes de regresión:

Fuente de variación	*Coeficiente*
Ordenada en el origen	0,000
A	2,417
B	0,643
C	0,417
D	–1,974
E	–0,161

$R^2 = 1{,}000$

La varianza explicada es del 100% y los coeficientes de regresión son las coordenadas estándares de las columnas en el primer eje (última columna de la imagen 14.1).

Recordatorio del modelo bilineal del AC

Realizamos un análisis de regresión más interesante y más relevante, cuando predecimos los datos a partir de las coordenadas, como vimos de forma resumida en el capítulo 13 al tratar sobre el modelo del AC. Vamos a repetir aquí tres versiones del mencionado modelo; la «versión simétrica», utilizando sólo coordenadas estándares [véase (13.4)], y las dos versiones asimétricas con filas o columnas respectivamente en coordenadas principales:

$$\frac{p_{ij}}{r_i c_j} = 1 + \sum_{k=1}^{K^*} \sqrt{\lambda_k}\,\phi_{ik}\gamma_{jk} + e_{ij} \tag{14.9}$$

$$\left(\frac{p_{ij}}{r_i}\right) / c_j = 1 + \sum_{k=1}^{K^*} f_{ik}\gamma_{jk} + e_{ij} \tag{14.10}$$

$$\left(\frac{p_{ij}}{c_j}\right) / r_i = 1 + \sum_{k=1}^{K^*} \phi_{ik} g_{jk} + e_{ij} \tag{14.11}$$

Este modelo se llama *bilineal* porque es lineal con relación a los productos de dos parámetros. Tendremos que fijar los valores de las coordenadas estándares de las filas o de las columnas, para obtener las coordenadas principales mediante análisis de regresión múltiple.

Regresión ponderada

En los lados izquierdos de las ecuaciones (14.9), (14.10) y (14.11) aparecen los cocientes de contingencia que definimos en el capítulo 13, escritos de tres maneras equivalentes. Tomando como ejemplo (14.10), y suponiendo que conocemos

las coordenadas estándares γ_{jk} de las columnas, tenemos a la derecha un modelo de regresión que predice los valores de las filas de la izquierda. Supongamos que estamos interesados en la primera fila (Geología) y que queremos llevar a cabo una regresión para $K^* = 2$. Para ajustar el modelo de AC, tenemos que minimizar una suma ponderada de residuos en la que ponderamos las categorías (columnas) con sus masas. Otra forma de verlo es decir que en (14.10) normalizamos las «variables explicativas» γ_{jk} con las masas de las columnas de la siguiente manera: $\sum_j c_j \gamma_{jk}^2 = 1$. Además, como las variables explicativas son ortogonales, cuando ponderamos con las masas de las columnas: $\sum_j c_j \gamma_{jk} \gamma_{j'k} = 0$ si $j \neq j'$. Para llevar a cabo la regresión, acomodamos el vector respuesta como un vector de 5×1 con los cocientes de contingencia de Geología, y las variables explicativas la disponemos como una matriz de 5×2 con las coordenadas estándares de las columnas en los dos primeros ejes principales. Llevamos a cabo el análisis de la regresión ponderada, ponderando con las masas de las columnas c_j. Los datos (los cocientes de contingencia de Geología (fila 1) indicados como $p_{1j} / (r_1 c_j)$, las coordenadas estándares de las columnas en las dimensiones 1 y 2, indicadas como γ_1 y γ_2, y los pesos c_j) son los siguientes:

Categoría	Geología	γ_1	γ_2	*Masa*
A	0,9063	2,4175	–0,4147	0,0389
B	1,3901	0,6434	–0,9948	0,1608
C	1,1781	0,4171	–0,2858	0,3894
D	1,0163	–1,9741	–0,7991	0,1621
E	0,4730	–0,1613	1,6762	0,2487

Los resultados de la regresión son:

Fuente	*Coeficiente*	*Coeficiente estandarizado*
Ordenada en el origen	1,000	—
f_{11}	0,076	0,234
f_{12}	–0,303	–0,928

$R^2 = 0{,}916$

Los coeficientes son las coordenadas principales f_{11} y f_{12} de Geología (el primer valor lo encontramos en la imagen 14.1). La varianza explicada (R^2) es la calidad de la representación de Geología en el mapa bidimensional (imagen 11.8).

En la regresión ponderada, las correlaciones recuperan las contribuciones relativas

Dado que en la regresión ponderada las variables explicativas están estandarizadas y son ortogonales, los coeficientes de regresión estandarizados serán también las correlaciones parciales entre la variable respuesta y las variables explicativas. La matriz de correlaciones de las tres variables es la siguiente (recordemos que en los cálculos hemos tenido en cuenta los pesos):

Variables	Geología	γ_1	γ_2
Geología	1,000	0,234	–0,928
γ_1	0,234	1,000	0,000
γ_2	–0,928	0,000	1,000

Como esperábamos, las dos variables explicativas no están correlacionadas. Las correlaciones entre Geología y las dos variables explicativas son exactamente los coeficientes de regresión estandarizados. Los cuadrados de estas correlaciones, $0{,}234^2 = 0{,}055$ y $(-0{,}928)^2 = 0{,}861$, son los cosenos al cuadrado (contribuciones relativas) que vimos en la imagen 11.6. Los resultados que acabamos de ver, ilustran la propiedad de la regresión que establece que si las variables explicativas no están correlacionadas, la varianza explicada R^2 es igual a la suma de los cuadrados de las correlaciones parciales.

Cálculo recíproco de medias y mínimos cuadrados alternados

Las ecuaciones de transición (14.1) y (14.2) son la base de un conocido algoritmo para hallar la solución del AC, llamado *cálculo recíproco de medias*. Empezamos el algoritmo con unas coordenadas estándares de las columnas —que hemos centrado y normalizado con medias y sumas de cuadrados ponderadas—. Aplicando la fórmula (14.1) de cálculo de medias ponderadas, calculamos valores para las coordenadas de las filas, a continuación aplicando la fórmula (14.2) a los valores anteriores de las filas, calculamos nuevos valores para las coordenadas de las columnas. Seguidamente estandarizamos estos valores y repetimos el proceso desde el inicio hasta la convergencia de los resultados, es decir, hasta obtener las coordenadas principales en el primer eje principal (es necesario estandarizar los valores de las coordenadas de las columnas que obtenemos en cada iteración, en caso contrario en los sucesivos cálculos de medias llegaríamos al valor cero). Hallar el segundo conjunto de coordenadas es más complicado, ya que tenemos que asegurar la ortogonalidad con las primeras coordenadas, no obstante la idea es la misma. Anteriormente, hemos visto que el paso de coordenadas columna a coordenadas fila, y de coordenadas fila a coordenadas columna lo podemos hacer mediante regresión. Por este motivo, este algoritmo se conoce también como *mínimos cuadrados alternados*, o regresiones alternadas. De todas formas, numéricamente es mejor llevar a cabo los cálculos utilizando la DVS (veánse los apéndices A y B), pero conocer estos algoritmos alternativos nos ayuda a profundizar en la comprensión del AC.

RESUMEN: Relaciones de transición y regresión

1. Cualesquiera que sean los valores asignados a las categorías de filas y columnas, podemos calcular las medias condicionales (es decir, las regresiones) de las filas con relación a las columnas o de las columnas con relación a las filas.
2. Realizado el AC, las coordenadas estándares de filas y de columnas cumplen las siguientes propiedades:

— la regresión de filas sobre columnas, y viceversa, son lineales (de aquí el nombre de *regresiones lineales simultáneas):*
— se minimiza el ángulo entre las dos regresiones;
— las medias condicionales que se hallan en las dos rectas de regresión son las coordenadas principales.

3. Llamamos *ecuaciones de transición* a las medias ponderadas entre coordenadas de filas y columnas, ponderadas con los elementos de los perfiles (de filas o columnas según el caso). Llamamos *cálculo recíproco* de medias a un algoritmo que permite hallar la solución del AC mediante la aplicación sucesiva de un par de ecuaciones de transición.

4. Podemos definir el AC como un *modelo de regresión bilineal,* ya que podemos recuperar los datos originales a partir de un modelo lineal de productos de coordenadas de filas y columnas. Este modelo se convierte en lineal si contemplamos como fijos los valores de las coordenadas de filas o columnas, lo que conduce a un algoritmo para hallar la solución del AC llamado *regresión de mínimos cuadrados alternada* (que, en realidad, es idéntico al algoritmo del cálculo recíproco de medias).