Anales de la Real Academia de Doctores de España
Volumen 13, pp. 147-157, 2009

ACP DEL TRIPLETE ESTADÍSTICO

$$\left[\left(I_{nxn} - \frac{1_n\, 1_n^T}{n}\right) X_{nxp}, Q_{pxp}, D_p\right]$$

ANTES DE UNA REGRESIÓN LINEAL MÚLTIPLE EN UNA SITUACIÓN LÍMITE

Dr. Fco. Javier Díaz-Llanos Sáinz-Calleja [*]
Dr. Juan Cayón Peña
Dra. M.ª del Carmen Cermeño Carrasco
[*] *Académico de Número de la Real Academia de Doctores de España*

RESUMEN

El objetivo de este artículo es el de advertir a los investigadores científicos —no acreditados en estadística— que, cuando tengan intención de hacer una **predicción** de una variable aleatoria —cuantitativa— observada frente a p variables —no aleatorias cuantitativas— que estén altamente correlacionadas, se les aconseja que apliquen un ACP del triplete estadístico antes de la regresión lineal múltiple.

PALABRAS CLAVES

Regresión lineal múltiple, métricas Q y D_p, triplete estadístico, operador WD_p de Yves Escoufier, análisis en componentes principales (ACP).

INTRODUCCIÓN

Nos parece lamentable que, a principios del año 2009 aún tan sólo se contemplen en los paquetes de programas comercializados de Análisis Estadístico Multidimensional, el **ACP** del **triplete estadístico:**

$$\left[\left(I_{nxn} - \frac{1_n\, 1_n^T}{n}\right) X_{nxp}, Q_{pxp}, D_p\right]$$

tan sólo con dos de las múltiples opciones para la **métrica Q,** manteniendo constante la **métrica D_p.**

Ni que decir tiene, tanto los investigadores científicos —no acreditados en Estadística— como algunos que sí lo están, han estado aplicando sistemáticamente el **ACP,** tal como se contempla en los paquetes de programas comercializados. Más adelante, indicaremos, no sólo cuáles son estas dos opciones, sino también, consejos para construir las **métricas** contenidas en el **triplete estadístico,** que se adapten mejor al tema de investigación de cada uno de los investigadores científicos.

No obstante, hemos de advertir a los investigadores científicos que, estas dos opciones que mencionaremos más adelante no son las únicas en un **ACP** y es más, en ciertas ocasiones, la aplicación de una de ellas puede conllevar que las **componentes principales** que se obtengan en sus análisis, no reflejen la realidad del problema concreto que estén investigando.

Una vez definido el **triplete estadístico,** estamos en condiciones de construir el **operador WD_p,** de Yves Escoufier, que más se adecúe para alcanzar los objetivos fijados *a priori* por los investigadores.

A partir de este operador, construiremos las **componentes principales,** que serán, las variables no aleatorias explicativas, contenidas en el modelo de regresión lineal múltiple.

Actuando de esta manera, podremos estimar —sin ningún problema— los **coeficientes de regresión.**

Finalmente, procederemos a deshacer la transformación lineal realizada y obtendremos la línea de regresión la cual nos permitirá, predecir una variable aleatoria observada, a partir de p variables no aleatorias.

Dado que, el objetivo de este artículo no es el de explicar el análisis de regresión lineal múltiple sobre las componentes principales para aquellos investigadores que deseen iniciarse en el análisis en **componentes principales** de manera formalmente correcta, les aconsejamos especialmente que consulten el libro de F. Cailliez y J. P. Pages (1). En dicho libro, también se contempla el **operador WD_p** de Yves Escoufier.

MATERIAL

Antes de mostrar con detalle los puntos más relevantes de la metodología a seguir (que se detallará en el apartado de **método),** nos parece conveniente explicar los siguientes puntos:

1. **El significado de la nomenclatura contenida en el triplete estadístico.**

2. **El grado de fiabilidad de la matriz *X* de dimensiones (n, p) que contiene las p variables originales no aleatorias explicativas.**

3. **Qué dimensión debe de tener la matriz *X* para que los análisis sean interpretables.**

4. Cuáles deben ser los tres requisitos básicos para aplicar con éxito el método de mínimos cuadrados ordinarios.

5. Cuál debe ser el criterio que tendrán que tener en cuenta los investigadores para construir las métricas Q y D_p:

5.1. Similitud entre el nivel de significación y las métricas Q y D_p.

5.2. Consejos prácticos dirigidos a los investigadores científicos para que construyan adecuadamente sus métricas Q y D_p.

1. EL SIGNIFICADO DE LA NOMENCLATURA CONTENIDA EN EL TRIPLETE ESTADÍSTICO

Para no expertos en el tema, consideramos —indudablemente— imprescindible explicar el significado de la misma:

1_n^T : *vector transpuesto de* 1_n.

X_{nxp} : *tabla de datos originales donde se encuentran las p variables no aleatorias explicativas.*

$\left(I_{nxn} - \frac{1_n\, 1_n^T}{n}\right) X_{nxp}$: *tabla de datos originales centrada por columnas.*

Q_{pxp} : *es la métrica de dimensiones (p, p) que nos va a permitir el cálculo de las distancias entre los individuos. Es una matriz simétrica definida positiva.*

En el caso hipotético que las variables sean independientes y además tengan el mismo peso :

$$Q_{pxp} = I_{pxp}$$

I_{pxp} : *matriz identidad de orden p.*

D_p : *es la métrica de dimensiones (n, n) que nos permite el cálculo de las distancias entre las variables. Es una matriz diagonal y se define :*

$$D_p = diag\,(p_i) \qquad \sum_{i=1}^{i=n} p_i = 1$$

p_i : *son los pesos de los individuos.*

En el caso hipotético que todos los individuos tengan el mismo peso :

$$D_p = \frac{1}{n} I_{nxn}$$

2. GRADO DE FIABILIDAD DE LA MATRIZ X DE DIMENSIONES (n,p) QUE CONTIENE LAS p VARIABLES ORIGINALES NO ALEATORIAS EXPLICATIVAS

Las variables contenidas en la matriz X, deben tener un grado de fiabilidad aceptable. Es obvio que, si las mediciones de las variables se han realizado de forma incorrecta, no tiene sentido la aplicación de ningún tipo de análisis.

3. EN CUANTO A LAS DIMENSIONES QUE DEBE TENER LA MATRIZ X PARA QUE LOS ANÁLISIS SEAN INTERPRETABLES

El profesor Thiery Foucart, indica claramente en su libro (2) que, para que un **ACP** se pueda interpretar, la tabla de datos debe de contener más de quince individuos y más de cuatro variables cuantitativas.

4. ¿CUÁLES DEBEN SER LOS TRES REQUISITOS BÁSICOS PARA APLICAR CON ÉXITO EL MÉTODO DE MÍNIMOS CUADRADOS ORDINARIOS?

1.º Datos ausentes

Tanto la variable aleatoria cuantitativa observada a explicar como las p variables no aleatorias cuantitativas explicativas deben de contener todas sus observaciones. Pues, de no ser así, no es posible la aplicación del método de mínimos cuadrados ordinarios.

2.º p > n

Es decir, el número de variables no aleatorias cuantitativas explicativas, debe ser mayor que el número de individuos.

Esta condición es obvia ya que, de no ser así, sería totalmente imposible encontrar los estimadores de los coeficientes de regresión que, más adelante, nos permitirán encontrar sus estimaciones, dado que, no podríamos invertir la matriz

$$\boldsymbol{X}^T I_{nxn} \boldsymbol{X}$$

para el cálculo de los estimadores de los coeficientes de regresión:

$$\hat{\beta}_{p(MO)} = (\mathbf{X}^T I_{nxn} \mathbf{X})^{-1} \mathbf{X}^T y^o$$

y^o: es la variable aleatoria cuantitativa observada centrada

$$\left(I_{nxn} - \frac{1_n 1_n^T}{n} \right) X_{xxp} = \mathbf{X}$$

3.º Se eligirán aquellas variables no aleatorias explicativas cuya relación entre ellas sea la menor posible

En este sentido, recordaremos que, las variables sujetas a estudio, se elegirán guardando entre ellas el mayor grado de independencia posible. Si no se conoce demasiado bien la naturaleza de dichas variables, se aconseja la realización de un test de multicoliniaridad. Entre los dos test de multicolinialidad que conocemos: test de Klein y test de Farrar y Glauber, recomendamos el segundo (3, 4) puesto que, proporciona resultados más fidedignos que el de Klein (4).

No obstante, si las variables analizadas son de origen biológico o socioeconómico, es muy posible que, dicho test dé positivo y, por tal razón, no quede más alternativa que realizar previamente un **ACP** del **triplete estadístico:**

$$\left[\left(I_{nxn} - \frac{1_n\, 1_n^T}{n}\right) X_{nxp}, Q_{pxp}, D_p\right]$$

Hacer lo que hemos propuesto es imprescindible ya que, si no se verifica la hipótesis estructural H6 mencionada por el profesor Régis Bourbonnais en su libro (4), es totalmente imposible encontrar los estimadores de los coeficientes de regresión, ya que

$$\det\ (\mathbf{X}^T\, I_{nxn}\, \mathbf{X}) = 0$$

Se sabe —por experiencia— que los procesos biológicos, tienen una similitud con los socioeconómicos y a su vez, que tanto las variables biológicas como las socioeconómicas están —casi— igualmente correlacionadas entre ellas.

En este sentido, a los investigadores que deseen constatar la veracidad de la correlación existente entre las variables socioeconómicas —sobre todo en un país en situación de inflación— les aconsejamos que acudan al libro de Karl A. Fox, Jati K. Sengupta y Erik Thorbecke (5). En dicho libro, se expone —de forma detallada— tal tipo de correlación.

5. ¿CUÁL DEBE SER EL CRITERIO QUE TENDRÁN QUE TENER EN CUENTA LOS INVESTIGADORES PARA CONSTRUIR LAS MÉTRICAS Q Y D_p?

5.1. Similitud entre el nivel de significación y las métricas Q y D_p

Así como, en Inferencia Estadística uno de los puntos claves para el investigador científico es el de fijar —debidamente— el nivel de significación; es decir, el umbral crítico de decisión a partir del cual aceptamos o rechazamos la hipótesis nula en un test de hipótesis, en el **ACP** los puntos claves son las **métricas**

$$Q_{nxn} \text{ y } D_p \text{ contenidas en el triplete estadístico.}$$

5.2. Consejos prácticos dirigidos a los investigadores científicos para que construyan adecuadamente sus métricas

Tanto para la construcción de la **métrica Q** como de la **métrica D_{pi},** tan sólo es imprescindible que, se conozcan perfectamente los **datos empíricos** sujetos a análisis.

5.2.1. *Métrica Q*

Conociendo tan sólo los **datos empíricos,** se está en condiciones de saber si las variables no aleatorias cuantitativas que se han elegido para el estudio de los objetivos, **son independientes** y si además, **tienen el mismo peso.**

La condición de **independencia** para las variables no aleatorias cuantitativas lleva consigo que la matriz

$$Q_{pxp} \text{ sea una matriz diagonal.}$$

Si además, todas las variables aportan el mismo peso, los elementos de dicha diagonal serán todos unos y, por tanto, la **métrica** adopta la siguiente forma:

$$Q_{pxp} = I_{pxp}$$

$$I_{pxp}: \text{ matriz identidad de orden } p$$

5.2.2. *Métrica D_p*

Partiendo de la base de que se conocen los **datos empíricos,** se está en condiciones de saber cuál es el peso de cada individuo.

La condición de que todos los individuos aporten el mismo peso lleva consigo que:

$$D_p = \frac{1}{n} I_{nxn}$$

$$I_{nxn}: \text{ matriz identidad de orden } n$$

Observación de interés: Aunque es obvio que ninguna de estas dos situaciones que hemos contemplado se dan en la práctica, los paquetes de programas comercializados de Análisis Estadístico Multidimensional, **sí** la incluyen para la aplicación de un **ACP** como una de las dos opciones aludidas en la **introducción.**

5.2.3. *¿Cuáles son las dos métricas para Q manteniendo constante la métrica D_p contenidas en los paquetes de programas comercializados de Análisis Estadístico Multidimensional para la aplicación de un ACP?*

5.2.3.1.

La primera opción es:

$$Q_{pxp} = I_{pxp} \qquad D_p = I_{pxp}$$

5.2.3.2.

La segunda opción es:

$$Q_{pxp} = diag\left(\frac{1}{s_j^2}\right) \quad j = 1,...,p \qquad D_p = \frac{1}{n} I_{nxn}$$

s_j^2 *: varianzas de cada una de las p variables explicativas*

MÉTODO

La metodología que vamos a mostrar tan sólo está recomendada cuando deseemos predecir una variable aleatoria cuantitativa, observada en función de p variables no aleatorias cuantitativas que den positivo el test de Farrar y Glauber.

Las etapas a seguir son las que mostramos a continuación:

1.ª Etapa

Partimos del término i-ésimo del modelo de regresión lineal múltiple:

$$y_i = \beta_o + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i$$

Dicho modelo puesto en forma matricial adopta la forma siguiente:

$$y_{nx1} = \left(1_n \mid X_{nxp}\right)_{n,p+1} \beta_{p+1} + \varepsilon_{nx1}$$

2.ª Etapa

Procedemos a centrar por columnas tanto la variable aleatoria cuantitativa observada como las p variables no aleatorias explicativas.

Como resultado de esta operación, el término i-ésimo del modelo de regresión lineal múltiple es el siguiente:

$$y_1^o = \beta_1 \chi_{i1} + ... + \beta_p \chi_{ip} + \varepsilon_i$$
$$y_i^o = yi - \bar{y} \quad i = 1,...,n$$
$$\boldsymbol{X}_{ij} = x_{ij} - \bar{x}_j \quad j = 1,...,p$$
$$\varepsilon_i = \varepsilon_i - \bar{\varepsilon}$$

Puesto en forma matricial es:

$$y^o_{nx1} = \mathbf{X}_{nxp}\ \beta^o{}_{px1} + \varepsilon_{nx1}$$

ε_{nx1} : *es un vector aleatorio centrado.*

3.ª Etapa

Aplicamos el test de Farrar y Glauber para constatar la existencia de multicolinealidad y observamos que el test da positivo.

4.ª Etapa

Aplicamos un **ACP** del **triplete estadístico**

$$(\boldsymbol{X}_{nxp},\ Q_{pxp},\ D_p)$$

Para conseguir este objetivo, lo primero que hemos de llevar a cabo es la construcción del operador

$$WD_p \textit{ de Yves Escoufier}$$

Dicho operador adopta la forma siguiente:

$$WD_p = \boldsymbol{X}_{nxp}\ Q_{pxp}\ \boldsymbol{X}^T_{pxn}\ Dp$$

Las **componentes principales** se obtienen diagonalizando este operador.

5.ª Etapa

Aplicaremos el **método de mínimos cuadrados ordinarios** sustituyendo las p variable no aleatorias cuantitativas por las p **componentes principales.**

6.ª Etapa

Desharemos la transformación lineal realizada con el fin de obtener la línea de regresión que contenga las variables no aleatorias cuantitativas originales.

De estas **seis etapas,** tan sólo vamos a desarrollar la **cuarta** ya que, las restantes, son muy conocidas.

4.ª Etapa

Para desarrollar dicha **etapa,** vamos a estructurarla en **dos sub-etapas:**

1.ª *Sub-etapa*

En esta parte mostramos las **métricas Q y D_p, el triplete estadístico,** el operador WD_p y cómo se obtienen las **componentes principales,** en la primera opción que se contempla en los paquetes de programas comercializados de Análisis Estadístico Multidimensional para el **ACP.**

a) Métricas:

$$Q_{pxp} = I_{pxp} \qquad D_p = \frac{1}{n} I_{nxn}$$

b) Triplete estadístico:

$$\left(\mathbf{X}_{nxp}, I_{pxp}, \frac{1}{n} I_{nxn} \right)$$

c) Operador WD_p:

$$WD_p = \mathbf{X}_{nxp}\, Q_{pxp}\, \mathbf{X}^T_{pxn} \frac{1}{n} I_{nxn}$$

d) Componentes principales:

Las **componentes principales** se obtienen directamente diagonalizando el **operador**

$$WD_p$$

2.ª *Sub-etapa*

En la segunda **sub-etapa,** mostraremos las **métricas**

$$Q_{pxp} \; y \; D_p$$

, el triplete estadístico, el operador

$$WD_p$$

y, cómo se obtienen las **componentes principales** en la segunda opción que se contempla en los paquetes de programas comercializados de Análisis Estadístico Multidimensional para el **ACP.**

a) Métricas:

$$Q_{pxp} = diag\left(\frac{1}{s_j^2}\right) \qquad j = 1,...,p \qquad D_p = \frac{1}{n} I_{nxn}$$

b) Triplete estadístico:

$$\left[\mathbf{X}_{n,p}, diag\left(\frac{1}{s_j^2}\right), \frac{1}{n} I_{nxn}\right]$$

c) Operador WD_p:

$$WD_p = \mathbf{X}_{nxp}\ diag\left(\frac{1}{s_j^2}\right)\mathbf{X}_{pxn}^T \frac{1}{n} I_{nxn}$$

d) Componentes principales:

Las **componentes principales** se obtienen directamente diagonalizando el **operador**

$$WD_P$$

Un aspecto de interés sobre el ACP de un triplete estadístico

La aplicación de un **ACP** del **triplete estadístico**

$$\left(\boldsymbol{X}_{n,p}, Q_{pxp}, D_p\right)$$

conduce al mismo operador

$$WD_p$$

que la aplicación de un **ACP** del **triplete estadístico**

$$\left(\mathbf{X}_{nxp}\ Q_{pxp}^{\frac{1}{2}}\ , I_{pxp}, D_p\right)$$

Búsqueda de WD_p en la primera opción

La construcción de dicho operador es inmediata:

$$WD_p = \mathbf{X}_{nxp}\, Q_{pxp}\, \mathbf{X}^T_{pxn}\, D_p$$

Búsqueda de WD_p en la segunda opción:

La construcción de dicho operador es casi inmediata:

$$WD_p = \mathbf{X}_{nxp}\, Q^{\frac{1}{2}}_{pxp}\, I_{pxp}\, Q^{\frac{1}{2}}_{pxp}\, \mathbf{X}^T_{pxn}\, D_p =$$
$$= \mathbf{X}_{nxp}\, Q_{pxp}\, \mathbf{X}^T_{pxn}\, D_p$$
$$WD_p = \mathbf{X}_{nxp}\, Q_{pxp}\, \mathbf{X}^T_{pxn}\, D_p$$

De lo que se desprende que el operador es el mismo en las dos situaciones.

CONCLUSIÓN

De la lectura de este artículo se concluye que, es imprescindible incorporar al programa del **ACP** —contenido en los paquetes de programas comercializados de Análisis Estadístico Multidimensional— una opción general que nos permita acceder a las **métricas Q** y **$\mathbf{D_p}$,** que más de adecúen a los **datos empíricos.**

Es obvio que este tipo de implementación deberá ser realizada por informáticos cualificados para dicha labor, de no ser así, los resultados de los análisis obtenidos, contendrían un alto grado de errores que harían inviable su credibilidad.

BIBLIOGRAFÍA

(1) Cailliez. F., Pages, J. P. (1976): *Introduction à l'analyse des données.* SMASH.

(2) Foucart, Th. (1997): *L'analyse des données. Méthodes et études de cas.* Presses Universitaires de Rennes.

(3) Farrar, D. E., Glauber R. R. (1967): «Multicolinearity in regression analysis». *Review of Economics and Statistics,* vol. 49.

(4) Bourbonnais, R. (1998): *Manuel et Exercices corrigés. Économetrie.* Deuxième édition. Dunod.

(5) Fox, K. A.; Sengupta, J. K., Thorbecke, E. (1979): *La teoría de la política económica cuantitativa,* oikos-tau, s. a-ediciones.