PESQUIMAT Revista de la Fac.CC.MM.de la
Universidad Nacional Mayor de San Marcos
Vol.III, $N°2$, Noviembre 2000, pag.34-43. Lima-Perú

INFLUENCIA DE LOS DATOS DISCORDANTES SOBRE LA MATRIZ DE CORRELACIÓN Y CORRELACIÓN PARCIAL EN UN CONJUNTO DE DATOS MULTIVARIADO

Seier, E.[1] ;Cambillo,E.[1] ;Cárdenas,A.[1] ;Adriazola,Y.[1]

ABSTRACT. *Una parte importante de la utilidad del análisis de influencia es determinar la debilidad de algunos estimadores tradicionales frente a la presencia de datos discordantes tales como los $\hat{B}_j$ (coeficientes de regresión) y los r_i (residuales) en un análisis de regresión; la $\overline{X}$ (media muestral) en un análisis inferencial, etc. los cuales han sido desarrollados ampliamente. Pero como muy poco se ha difundido sobre la correlación parcial, tratamos de presentar los efectos que se producen en dichos estimadores ante la presencia de datos discordantes o influyentes con una aplicación a un conjunto de datos socio-económicos del Perú.*

1. INTRODUCCIÓN

El análisis de influencia está desarrollado en la literatura especializada para algunos aspectos del análisis de regresión, pero hemos detectado algunos estadísticos como las correlaciones, correlaciones parciales entre otras, en las que se ha desarrollado muy poco el análisis de influencia.

De ahí que surgió la siguiente pregunta, ¿Qué efectos se producen en la matriz de correlación y correlación parcial en el conjunto de datos ante la presencia de datos discordantes?

La respuesta a esta pregunta es muy importante porque existen técnicas estadísticas que requieren que las variables estén correlacionadas y no vaya a suceder que ante la presencia de datos discordantes o influyentes se esté enmascarando una alta o pobre correlación y como

[1]Universidad Nacional Mayor de San Marcos. Laboratorio de Series de Tiempo.

consecuencia de ello se esté subestimando ó sobrestimando β en el modelo de regresión.

Como teóricamente resulta complicado establecer los efectos que se producen en la matriz de correlación y correlación parcial ante la presencia de datos discordantes o influyentes, se tomó un conjunto de datos socio-económico del Perú para analizar los posibles cambios que se pueden presentar debido a que en [2] se probó que LIMA y MADRE DE DIOS son datos discordantes.

JUSTIFICACIÓN

Denotemos por $X_{n\times p}$ la matriz de datos $S = (s_{ij})$ la matriz de covarianza muestral y $R = (r_{ij})$ la matriz de correlación muestral con $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, p$.

R es obtenida de S escalando filas y columnas tal que todos los elementos de la diagonal sea igual a la unidad.

La varianza y correlación han sido bien estudiados en la mayoría de los libros de estadística, lo que no es muy conocido es la interpretación de los elementos de la inversa de las matrices de covarianza S^{-1} y correlacin R^{-1}.

Si un elemento de S^{-1} es cero, entonces las variables correspondientes son independientes condicionalmente dado el resto de las variables. Los recíprocos de los elementos de la diagonal de S^{-1} vienen hacer las varianzas residuales habiendo predicho las otras variables restantes y nos permite presentar la proporción de varianza de cada variable que permanece no explicada.

$$\frac{Var(X_i/x_i, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)}{Var(X_i)} \qquad (1)$$

El recíproco de los elementos de la diagonal de R^{-1} coincide con la proporción de varianza de cada variable X_i que permanece no explicada. Los elementos de la diagonal de R^{-1} nos permite encontrar los coeficientes de determinación ya que cada elemento está relacionado con (1).

Es decir cada elemento de la diagonal de R^{-1}, según algunos autores (ver [1]) se les conoce como factor de inflación de varianza (VIF_j) ya que al ser R^2 próximo a uno, la cantidad $(1/1 - R^2)^{-1}$ se hace cada vez más grande.

Si hacemos cambio de escala en R^2 hasta obtener uno en la diagonal, se obtiene una matriz llamada de correlación inversa. Nuestro interés además de las interpretaciones antes mencionadas, está en la interpretación de los elementos de dicha matriz.

Se encuentra que los elementos fuera de la diagonal de la matriz correlación inversa son los negativos de los coeficientes de correlación parcial entre X_i y X_j dado $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$.

Los resultados de estas interpretaciones se pueden resumir en las siguientes proposiciones:

Proposición 1.- Cada elemento de la diagonal de la inversa de la matriz varianza es el recíproco de la varianza parcial.

Proposición 2.- Cada elemento fuera de la diagonal de la inversa de la matriz de varianza, escalada a tener en la diagonal la unidad, es el negativo de la correlación parcial entre las 2 correspondientes variables dado el resto de las variables.

2. DATOS UTILIZADOS

APLICACIÓN

En base al conjunto de datos PERÚ.DAT [2] correspondiente a las variables socio-económicas del Perú (Censo 1983) se mostrará el efecto que tienen los datos discordantes en las correlaciones simples y correlaciones parciales respectivamente.

Se probó en [3] a través del análisis se regresión que los datos correspondientes al departamento de Madre de Dios y Lima son discordantes y potencialmente influyentes. Las siguientes matrices corresponden a la matriz de covarianzas y matriz de correlación y sus matrices inversas respectivas de las variables MIG, PEAG, VSC, IFP, ANALF.

MATRIZ DE COVARIANZA

	MIG	PEAG	VSC	IFP	ANALF
MIG	11.481				
PEAG	-34.427	346.302		SIM	
VSC	27.508	-227.242	224.672		
IFP	56.459	-468.751	231.831	935.582	
ANALF	-22.206	211.443	-124.354	-340.828	181.462

MATRIZ DE CORRELACIÓN

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.546	1		SIM	
VSC	0.541	-0.815	1		
IFP	0.548	-0.824	0.507	1	
ANALF	-0.489	0.851	-0.621	-0.832	1

INVERSA DE LA MATRIZ DE COVARIANZA

	MIG	PEAG	VSC	IFP	ANALF
MIG	0.147916				
PEAG	-0.010481	0.03064		SIM	
VSC	-0.019349	0.01912	0.020236		
IFP	-0.010039	0.00806	0.005719	0.005874	
ANALF	-0.001802	-0.001802	-0.00036	0.004327	0.02358

INVERSA DE LA MATRIZ DE CORRELACIÓN

	MIG	PEAG	VSC	IFP	ANALF
MIG	1.695				
PEAG	-0.647	10.773		SIM	
VSC	-0.973	5.341	4.532		
IFP	-1.030	4.485	2.574	5.468	
ANALF	-0.080	-2.4433	-0.069	1.824	4.515

MIG : Tasa de migración.
PEAG : Población agropecuaria.
VSC : Vivienda sin construir.
IFP : Ingreso familiar promedio.
ANALF : Tasa de analfabetismo.

Puede de observarse que las correlaciones varían de -0.489 hasta 0.851, el valor -0.4893 corresponde al índice de correlación entre las variables ANALF y MIG indicando que existe una relación inversa entre tasa de migración y la tasa de analfabetismo y el valor 0.851 está en correspondencia con las variables ANALF y PEAG, indicando que existe una relación directa entre el analfabetismo y la población agraria.

De acuerdo a lo presentado en la sección justificación del presente artículo observamos que todos los elementos de la inversa de la matriz de covarianza se aproximan a cero, lo cual indica que las variables correspondientes son independientes condicionalmente dado el resto de variables.

VARIANZAS RESIDUALES

MIG	PEAG	VSC	IFP	ANALF
6.7712	32.1399	49.6697	171.0364	40.1929

Puede observarse que la variable MIG tiene la varianza residual más pequeña (6.7712) y la varianza IFP la más grande (171.0864).

La proporción de varianza que permanece no explicada se determina dividiendo la varianza de la variable condicionada a las restantes con la varianza de la variable, por ejemplo:

$$\frac{Var\ \text{MIG / (rest. de variables)}}{Var\ \text{MIG}} = \frac{Var\ \text{MIG(Residual de MIG)}}{Var\ \text{MIG}} = \frac{6.771}{11.481} = 0.59$$

Este valor indica que la migración, los servicios básicos (agua, desagüe, luz), el ingreso y el analfabetismo están muy relacionados con el hecho de que la población sea urbana o rural.

Se presenta a continuación la proporción de varianza no explicada para cada una de las variables consideradas en el estudio.

MIG	PEAG	VSC	IFP	ANALF
0.59	0.09	0.18	0.18	0.22

En relación a la inversa de la matriz de correlación, ésta nos permite encontrar los coeficientes de determinación o la proporción de variación para cada una de las variables.

Proporción de variación de cada una de las variables que son explicadas por los restantes.

MIG	PEAG	VSC	IFP	ANALF
0.4101	0.77797	0.8180	0.7778	0.9072

MATRIZ DE LA CORRELACIÓN INVERSA

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.152	1		SIM	
VSC	-0.360	0.764	1		
IFP	-0.338	0.584	0.517	1	
ANALF	-0.029	-0.350	-0.015	0.367	-1

MATRIZ DE CORRELACIÓN PARCIAL

	MIG	PEAG	VSC	IFP	ANALF
MIG	-1				
PEAG	0.152	-1		SIM	
VSC	0.360	-0.764	-1		
IFP	0.338	-0.584	-0.517	-1	
ANALF	0.029	0.350	0.015	-0.367	-1

Los elementos fuera de la diagonal de la matriz de correlación inversa corresponden a los negativos de los coeficientes de correlación parcial entre las correspondientes pares de variables condicionada a las restantes.

Por ejemplo el valor -0.764 indica que existe una relación inversa entre las variables PEAG y VSC, condicionada al conocimiento de las variables restantes.

Puede también afirmarse que las variables ANALF y VSC no están correlacionadas dadas las restantes.

Comparaciones entre las matrices de correlación para el conjunto de datos completos PERU.DAT que contiene datos discordantes y/o influyentes y para el mismo conjunto de datos eliminando dichos puntos discordantes y/o influyentes.

- PERU.16 (eliminando Madre de Dios)
- PERU.14 (eliminando Lima)
- PERU.1416 (eliminando Lima y Madre de Dios)

PERÚ

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.546	1		SIM	
VSC	0.541	-0.815	1		
IFP	0.545	-0.824	0.507	1	
ANALF	-0.489	0.851	-0.621	0.832	1

PERÚ16

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.552	1		SIM	
VSC	0.556	-0.840	1		
IFP	0.642	-0.870	0.752	1	
ANALF	-0.491	0.846	-0.672	0.906	1

PERÚ14

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.348	1		SIM	
VSC	0.136	-0.760	1		
IFP	0.526	-0.796	0.401	1	
ANALF	-0.523	0.846	-0.582	0.816	1

PERÚ1416

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	-0.321	1		SIM	
VSC	0.168	-0.856	1		
IFP	0.560	-0.834	0.673	1	
ANALF	-0.508	0.840	-0.641	0.908	1

Observando estas cuatro matrices de correlación se encuentra que en las matrices PERÚ y PERÚ16, la correlación entre VSC e IFP varía notoriamente de 0.507 a 0.752. En las matrices PERÚ y PERÚ1416 hay un ligero incremento entre la correlación de VSC e IFP, en lo que se refiere a las matrices PERÚ y PERÚ14 hay una ligera disminución entre VSC e IFP de 0.507 a 0.401.

Comparaciones entre las matrices de correlación parcial para el conjunto de datos completos PERÚ.DAT que contiene datos discordantes y/o influyentes y para el mismo conjunto de datos eliminando dichos puntos discordantes y/o influyentes.

PARCIPERÚ

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	0.152	1		SIM	
VSC	0.351	-0.764	1		
IFP	0.338	-0.584	-0.517	1	
ANALF	0.029	0.350	0.015	-0.367	1

PAR16

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	0.066	1		SIM	
VSC	0.141	-0.764	1		
IFP	0.453	-0.182	0.708	1	
ANALF	0.226	0.408	0.308	-0.670	1

PAR14

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	0.212	1		SIM	
VSC	-0.001	-0.707	1		
IFP	0.259	-0.509	-0.483	1	
ANALF	-0.327	0.380	-0.066	-0.295	1

PAR1416

	MIG	PEAG	VSC	IFP	ANALF
MIG	1				
PEAG	0.139	1		SIM	
VSC	-0.142	-0.706	1		
IFP	0.380	-0.199	0.118	1	
ANALF	-0.091	0.423	0.232	-0.599	1

Analizando las cuatro matrices de correlación parcial se encuentran cambios sustanciales entre dichas correlaciones. Por ejemplo, en la matriz PARCIPERÚ y PAR16 la correlación parcial entre VSC e IFP cambia de -0.571 a 0.708 debido a que el dato 16 correspondiente a Madre de Dios no ha sido considerado (es discordante e influyente). Algunas

otras correlaciones cambian moderadamente manteniendo su signo. En cambio la correlación parcial entre VSC y IFP que se observa en las matrices PARCIPERÚ y PAR14 mantienen su propio signo y no sufren un cambio fuerte. Además se observa que las variables MIG y ANALF son incorrelacionadas pero al eliminar el dato 14, ésta correlación es moderada y cambia de signo.

Respecto a la matriz de correlación PAR1416, la correlación parcial entre VSC e IFP (eliminando ambos datos 14 y 16) cambia de -0.517 a 0.118 como era de esperar en la forma similar cuando se eliminó solamente el dato 16 que es influyente.

Examinando las 4 matrices de correlación observamos que la correlación parcial entre PEAG y VSC se mantienen casi igual en valor y signo en cada uno de los casos, es decir eliminando el dato 16, eliminando el dato 14 y eliminando ambos datos 14 y 16. Por lo tanto la correlación parcial entre PEAG y VSC no se ve afectada por la presencia del dato influyente ni por la presencia de las otras variables en estudio.

Comparando la matriz de correlación y la matriz de correlación parcial para los datos de PERÚ observamos que la correlación entre VSC e IFP es de 0.507, y la correlación parcial es de -0.517, es decir sufre un cambio de signo, a más IFP menos VSC, lo cual indica que la relación se mantiene en forma negativa aunque estén presentes las otras variables.

También observamos que mientras la correlación entre ANALF y VSC es igual a -0.621, la correlación parcial entre ellas es 0.015, es decir son incorrelacionadas ante la presencia de las otras variables.

Siguiendo la comparación para los datos PERÚ16 se observa un fuerte cambio entre ambas correlaciones simples y parciales para las variables IFP y VSC. Es decir al eliminar el dato 16 correspondiente a Madre de Dios la correlación simple que es fuerte entre VSC e IFP se desvanece cuando se encuentran presentes las variables PEAG, ANALF, MIG.

Esto nos hace pensar que la presencia de un dato influyente afecta la correlación parcial entre las variables, ya que la correlación simple entre VSC y PEAG y la correlación parcial entre VSC y PEAG se mantienen altas e inversas.

3. CONCLUSIONES

- Para el caso multivariado los r_{ij}, resultan afectados ante la presencia de datos influyentes, ya que estas estadísticas cambian de

valor siendo el caso mas notorio el de las variables VSC e IFP cuya correlación varia de 0,507 a 0,752 cuando se elimina el dato influyente correspondiente al departamento de Madre de Dios.

- Para el caso multivariado los r_{ij}, 1, 2, 3, ... , k también resultan afectados ante la presencia de datos influyentes, debido a que no sólo cambian de valor sino en signo; siendo el caso mas notorio el de las variables VSC e IFP cuya correlación parcial cambió de $-0,571$ a 0,708 cuando se eliminó el dato influyente correspondiente a Madres de Dios, y cambia a 0,118 cuando se eliminaron los datos influyentes correspondientes a los departamentos Madre de Dios y Lima.
- Resulta complicado tratar de medir los cambios que sufren tanto las correlaciones como las correlaciones parciales en un conjunto de datos multivariado, porque además de la posición del dato influyente existen diferentes clases de relaciones entre dos variables estando presente las restantes.

4. BIBLIOGRAFIA

[1] Chatterjee, S. and B. Price *Regression Analysis by Example*, Jhon Wiley and Sons; New York.(1977).

[2] Seiers, E.; Cárdenas, A.; Cambillo, E.; Adriazola, Y.; Medina, F. *Análisis de Influencia.* Laboratorio de Series de Tiempo.(1994).

[3] Seiers, Edith *Análisis de Influencia en regresión para algunas variables Socioeconómicos en el Perú.* Doc.Téc. Laboratorio de Series De Tiempo. FCM UNMSM. (1991).

[4] Whittaker Joe *Graphical models in Applied Multivariate Statistics.*(1989).