Divulgación

Comparación de subconjuntos de regresiones cuando algunas de las variables explicativas provenientes de bases de datos biológicos están observadas parcialmente.

RECIBIDO: 30/05/2008
ACEPTADO: 19/03/2009

Badler, C. E. • Puigsubirá, C. R. • Vitelleschi, M. S.

Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística. Facultad de Ciencias Económicas y Estadística. Universidad Nacional de Rosario.

Bvrd. Oroño 1261. (2000) Rosario, Santa Fe, Argentina.
Teléfono: (0341) 480 2793 int. 151.
Email: cpuigsu@fcecon.unr.edu.ar

RESUMEN: El coeficiente de correlación múltiple es usualmente utilizado para comparar conjuntos de variables explicativas con respecto a cuán bien ellas predicen los valores futuros de la variable respuesta. Si la base de datos contiene variables observadas parcialmente, Donald Rubin propuso una adaptación del coeficiente de correlación múltiple con el fin de no descartar información. Para el cálculo de dicho coeficiente se requieren las estimaciones máximo verosímiles de ciertos parámetros, las cuales se realizan a través del operador "Sweep". La metodología es aplicada a un conjunto de datos provenientes de las historias clínicas perinatales de niños nacidos en el Hospital Roque Sáenz Peña de la ciudad de Rosario durante el año 2002.

PALABRAS CLAVE: Información faltante, Selección de variables, Coeficiente de correlación múltiple, Operador "Sweep".

SUMMARY: *Comparison of regression subsets when some predictor variables of biological data are partially observed.* The multiple correlation coefficient is often used to compare sets of independent variables regarding how well they predict the future values of a dependent variable. If the data set contains partially observed variables, Donald Rubin proposed an adaptation of the multiple correlation coefficient in order not to discard information. For the calculation of this coefficient the maximum likelihood estimates of certain parameters are required, which are carried out through the "Sweep" operator. The methodology is applied to a data set taken from perinatal clinical histories of children born in the Roque Sáenz Peña Hospital of Rosario city, during the year 2002.

KEYWORDS: Missing data, Selection of variables, Multiple correlation coefficient, "Sweep" operator.

Introducción

En las bases de datos provenientes del área biológica, es frecuente que algunas variables no registren información para ciertas unidades observacionales. Esto sucede tanto en bases de datos provenientes de relevamientos censales como muestrales. La falta de información puede ser originada por múltiples causas y en general las variables "sensibles" o difíciles de recabar son las más afectadas; lo cual no sólo repercute en la calidad de la información disponible, sino también en los resultados de los análisis realizados a partir de ellos, alterando las conclusiones.

Actualmente, el problema de la falta de información es considerado fundamental para la inferencia dada su permanente presencia, hecho que continúa motivando la aparición de diferentes tratamientos para su solución.

Los softwares de análisis estadístico no siempre contemplan la posibilidad de aplicar sus procedimientos incorporando las unidades con pérdida en algunas variables, sino que suprimen dichas unidades afectando el análisis y por ende los resultados.

Cuando el objetivo del investigador es construir un modelo de regresión múltiple a partir de una base de datos con una gran cantidad de variables, previo a la construcción del mismo, debe realizar una selección de variables para obtener el "mejor" subconjunto de las mismas.

El coeficiente de correlación múltiple es una de las posibles medidas para comparar subconjuntos de variables explicativas y decidir cuál de ellos es el "mejor" para predecir valores de la variable respuesta.

Cuando algunas de las variables regresoras están observadas parcialmente, para no descartar información, Donald Rubin (1) propone una adaptación del coeficiente de correlaciones múltiple para esta situación. En este trabajo se presenta dicha adaptación y la metodología es aplicada a un conjunto de datos provenientes de las historias clínicas perinatales de niños nacidos en el Hospital Roque Sáenz Peña de la ciudad de Rosario durante el año 2002.

Material

Se trabaja con una base de datos correspondiente a 179 niños nacidos en el Hospital Roque Sáenz Peña de la ciudad de Rosario en el año 2002. Las variables consideradas son:

- Peso del recién nacido en gramos (Y).
- Edad gestacional en semanas cumplidas hasta el parto (X_1).
- Percentil del peso por edad gestacional (X_2).
- Perímetro cefálico en milímetros (X_3).
- Edad de la embarazada en años cumplidos (X_4).

Metodología

La selección de un subconjunto S de p variables explicativas observadas completamente, se puede realizar utilizando el coeficiente de correlación múltiple (R_S^2), ya que mide la habilidad de dicho subconjunto para predecir valores futuros de la variable respuesta. La determinación del mismo se realiza a través del mayor valor del coeficiente de correlación múltilple. Si existen dos subconjuntos de variables explicativas con igual valor de R_S^2, se seleccionará aquel cuyas variables sean menos costosas de registrarlas. Mientras que, si una nueva variable es agregada al subconjunto S y el coeficiente de correlación múltiple es ligeramente mayor, dicha variable podría no ser considerada ya que no aporta mayor información que la obtenida.

Una vez determinado el subconjunto S de variables explicativas, se selecciona aleatoriamente una unidad y el error esperado de predecir el valor de Y para esa unidad es:

$$(1-R_S^2)\,\sigma^2$$

Siendo:

R_S^2 : coeficiente de correlación múltilple poblacional,

σ^2: variancia poblacional de Y.

Se considera, ahora, que algunos valores de las variables explicativas están perdidos y esto es causado por un mecanismo probabilístico (2). Es decir, los datos perdidos están siempre perdidos al azar, los datos observados están siempre observados al azar y esto ocurre cualquiera sea la variable a registrar.

Primero se supone que se registran las p variables explicativas sobre las unidades elegidas aleatoriamente. Sea π_T la probabilidad que sólo las variables X en el conjunto T están observadas, es decir, si T={1,2}, π_T es la probabilidad que las variables explicativas 1 y 2 estén observadas y las variables explicativas 3,….,p estén perdidas. La suma de los a través de los 2^p esquemas posibles de observaciones perdidas en las variables X es igual a uno ($\sum_T \pi_T = 1$). Si las p variables explicativas están observadas para una unidad determinada, se las utiliza a todas para predecir Y; si todas las variables X están observadas menos una, se usan todas menos esa variable X para predecir Y; y así sucesivamente. Si se registran las p variables explicativas, el cuadrado medio error de la predicción para una unidad elegida aleatoriamente es:

$$\sum_T \pi_T (1-R_T^2)\,\sigma^2$$

Se supone ahora que se ha seleccionado el subconjunto S de variables X y se considera una unidad elegida aleatoriamente sobre la cual se trata de registrar todas las variables en S y no las restantes variables. Además, se supone que el conjunto T de variables X habría sido observado para esa unidad que se han tratado de registrar las p variables explicativas. Luego se han registrado valores para el conjunto $S \cap T$ de variables X, y de esta forma se puede utilizar este conjunto para predecir Y.

De manera tal que se ha elegido el subconjunto S de variables explicativas para predecir Y y el cuadrado medio error de la predicción para una unidad elegida aleatoriamente es:

$$\sum_T \pi_T (1-R_{S\cap T}^2)\,\sigma^2 \quad (1)$$

La cual se puede escribir de la forma:

$$(1-Q_S^2)\,\sigma^2$$

siendo:

$$Q_S^2 = \sum_T \pi_T\, R_{S\cap T}^2 \quad (2)$$

Puesto que Q_S^2 es el porcentaje de variación de Y que puede ser predicho por las variables X en S, es apropiado usarlo como una medida de la habilidad del subconjunto S para predecir valores futuros de Y. Se deduce de la ecuación (2) que $0 \le Q_S^2 \le R_S^2$ y $Q_S^2 = R_S^2$ solamente cuando las variables en S están siempre observadas.

El estimador máximo verosímil de Q_S^2 es:

$$\hat{Q}_S^{\,2} = \sum_T \hat{\pi}_T\, \hat{R}_{S\cap T}^{\,2}$$

siendo:

$\hat{\pi}_T$: el estimador máximo verosímil de π_T

$\hat{R}_{S\cap T}^{\,2}$: es el estimador máximo verosímil de $R_{S\cap T}^2$.

El estimador de $R_{S\cap T}^2$ se puede calcular a partir de la matriz de covariancias estimada ($\hat{\sum}$), la cual fue obtenida mediante el operor “SWEEP”.

El operador "Sweep"

El operador "Sweep" (3,4) provee una forma simple y conveniente de realizar los cálculos para obtener las estimaciones máximo verosímiles de ciertos parámetros cuando existe falta de información en la base de datos.

La aplicación del mismo requiere que las variables que componen la base de datos, puedan ser arregladas en un esquema de pérdida monótono y que las mismas se distribuyan conjuntamente normal.

Sea **G** una matriz simétrica de dimensión pxp. Para cualquier k ε {1,...,p} el operador "Sweep" (1 - 2) en posición k, SWP[k], produce otra matriz **H** simétrica pxp, cuyos elementos son de la forma:

$$h_{kk} = -1/g_{kk}$$
$$h_{jk} = h_{kj} = g_{jk}/g_{kk} \qquad k \neq j$$
$$h_{jl} = g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad k \neq j, k \neq l$$

Aplicar el operador Sweep recorriendo todas las posiciones k de la matriz **G**, es equivalente al cálculo de $-\mathbf{G}^{-1}$. Esta inversa existe sí y sólo sí ninguno de los barridos involucra la división por 0. Es decir:

$$SWP[1,...,p]\mathbf{G} = SW\,P[1]...\,SW\,P[p]\mathbf{G} = -\mathbf{G}^{-1}$$

Cuando se realizan barridos en varias posiciones no es necesario llevarlo a cabo en un orden particular, dado que el operador "Sweep" es conmutativo:

$$SWP[j,k]\mathbf{G} = SWP[k,j]\mathbf{G}$$

para cualquier $j \neq k$ con j, k ε {1,...,p}.

Se define el operador "Sweep" inverso en posición k, y se lo denota con **H** = RSW[k]. Sus componentes son de la forma:

$$h_{kk} = -1/g_{kk}$$
$$h_{jk} = h_{kj} = -g_{jk}/g_{kk} \qquad k \neq j$$
$$h_{jl} = g_{jl} - g_{jk}g_{kl}/g_{kk} \qquad k \neq j, l \neq j$$

El operador "Sweep" inverso es también conmutativo y además, es el inverso del operador "Sweep", es decir:

$$RSW[k]SWP[k]\mathbf{G} = \mathbf{G},$$

para cualquier k ε {1,...,p}.

Se presenta el caso cuando el operador "Sweep" y el "Sweep" inverso pueden ser aplicados para encontrar las estimaciones máximo verosímiles del vector de promedios y la matriz de covariancias de una distribución normal multivariada a partir de un conjunto de datos incompletos en el que las variables han sido convenientemente agrupadas en bloques con igual número de observaciones dentro de cada uno de ellos, para obtener un esquema de pérdidas monótono. Por simplicidad se consideran tres bloques de variables (Z_1, Z_2 y Z_3). La extensión a más de tres bloques es inmediata. Los pasos a seguir son:

Paso 1: encontrar los estimadores máximo verosímiles $\hat{\mu}_1$ y $\hat{\Sigma}_{11}$ del vector de promedios y de la matriz de covariancias Σ_{11} del primer bloque de variables, las cuales están completamente observadas. Estas estimaciones son simplemente el vector de promedios y la matriz de covariancias de Z_1, calculados a partir de todas las observaciones muestrales.

Paso 2: encontrar los estimadores máximo verosímiles $\hat{\beta}_{20.1}$, $\hat{\beta}_{21.1}$ y $\hat{\Sigma}_{22.1}$ de los interceptos, los coeficientes de regresión y la matriz de covariancias residual de la regresión de Z_2 en Z_1. Estas pueden ser encontradas barriendo las variables Z_1 fuera de la matriz de covariancias ampliada de Z_1 y Z_2 basadas en las observaciones con Z_1 y Z_2 ambas observadas.

Paso 3: encontrar los estimadores máximo verosímiles $\hat{\beta}_{30.12}$, $\hat{\beta}_{31.12}$, $\hat{\beta}_{32.12}$ y $\hat{\Sigma}_{33.12}$ de los interceptos, los coeficientes de regresión y la matriz de covariancias residual de la regresión de Z_3 en Z_1 y Z_2. Estas pueden ser obtenidas mediante el barrido de las variables Z_1 y Z_2 fuera de la matriz de covariancias ampliadas

de Z_1, Z_2 y Z_3 basadas en las observaciones completas de Z_1, Z_2 y las observadas de Z_3.

Paso 4: calcular la matriz **A** de la forma:

$$\mathbf{A} = \mathrm{SWP}[\mathbf{1}]\begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

donde SWP[1] indica el barrido a través del conjunto de variables Z_1.

Paso 5: calcular la matriz:

$$\mathbf{B} = \mathrm{SWP}[\mathbf{2}]\begin{bmatrix} a_{11} & a_{12} & \hat{\beta}_{20.1}^T \\ a_{21} & a_{22} & \hat{\beta}_{21.1}^T \\ \hat{\beta}_{20.1} & \hat{\beta}_{21.1} & \hat{\Sigma}_{22.1} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

donde SWP[2] indica el barrido a través del conjunto de variables Z_2.

Paso 6: finalmente, la estimación máximo verosímil de la matriz de covariancias ampliada de Z_1, Z_2 y Z_3 está dada por:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \mathrm{RSW}[1,2]\begin{bmatrix} c_{11} & c_{12} & c_{13} & \hat{\beta}_{20.1}^T \\ c_{21} & c_{22} & c_{23} & \hat{\beta}_{31.12}^T \\ c_{31} & c_{23} & c_{33} & \hat{\beta}_{32.12}^T \\ \hat{\beta}_{20.1} & \hat{\beta}_{31.12} & \hat{\beta}_{32.12} & \hat{\Sigma}_{33.12} \end{bmatrix}$$

Esta matriz contiene las estimaciones máximo verosímiles del vector de los promedios y la matriz de covariancias de Z_1, Z_2 y Z_3 (4).

Los pasos 4 a 6 pueden ser representados concisamente por la ecuación:

$$\begin{bmatrix} -1 & \hat{\mu}^T \\ \hat{\mu} & \hat{\Sigma} \end{bmatrix} = \mathrm{RSW}[1,2]\left[\ \mathrm{SWP}[2]\ \begin{bmatrix} \mathrm{SWP}[1]\begin{bmatrix} -1 & \hat{\mu}_1^T \\ \hat{\mu}_1 & \hat{\Sigma}_{11} \end{bmatrix} & \hat{\beta}_{20.1}^T & \hat{\beta}_{30.12}^T \\ & \hat{\beta}_{21.1}^T & \hat{\beta}_{31.12}^T \\ \hat{\beta}_{20.1}\ \hat{\beta}_{21.1} & \hat{\Sigma}_{22.1} & \hat{\beta}_{32.12}^T \\ \hat{\beta}_{30.12}\ \hat{\beta}_{31.12}\ \hat{\beta}_{32.12} & & \hat{\Sigma}_{33.12} \end{bmatrix}\right]$$

Esta ecuación define la transformación para obtener las estimaciones máximo verosímiles de los parámetros de la distribución normal multivariada.

Aplicación

Los datos que se analizan fueron extraídos de las historias clínicas perinatales de 179 niños nacidos en el Hospital Roque Sáenz Peña de Rosario en el año 2002. Se trabaja con las variables:

- Peso del recién nacido en gramos (Y).
- Edad gestacional en semanas cumplidas hasta el parto (X_1).
- Percentil del peso por edad gestacional (X_2).
- Perímetro cefálico en milímetros (X_3).
- Edad de la embarazada en años cumplidos (X_4).

Para la aplicación de la metodología presentada, dado que se parte de una base de datos sin información faltante (B1), se generan pérdidas completamente al azar en las variables percentil del peso, perímetro cefálico y edad de la embarazada, aproximadamente, en un 42, 28 y 27 por ciento, respectivamente.

Las variables en la base de datos incompleta (B2) se reordenaron, convenientemente, de manera de obtener un esquema de pérdida monótono (Figura 1).

Figura 1: Esquema de pérdida monótono

N	Y	X_1	X_4	X_3	X_2
1	■	■	■	■	■
2	■	■	■	■	■
.	■	■	■	■	■
105	■	■	■	■	■
.	■	■	■	■	
129	■	■	■	■	
130	■	■	■		
.	■	■			
179	■	■			

Mediante el operador "Sweep" se obtuvieron las estimaciones máximo verosímiles del vector de promedios y la matriz de covariancias del conjunto de datos B2, resultando:

$$\hat{\mu} = [3113.29 \quad 38.47 \quad 21.99 \quad 343.19 \quad 39.41]$$

$$\hat{\Sigma} = \begin{bmatrix} 324489.69 & 812.48 & 930.61 & 7693.62 & 15479.57 \\ 812.48 & 4.65 & 2.17 & 24.73 & 15.47 \\ 930.61 & 2.17 & 30.98 & 21.78 & 43.68 \\ 7693.62 & 24.73 & 21.78 & 289.94 & 293.15 \\ 15479.57 & 15.47 & 43.68 & 293.15 & 955.45 \end{bmatrix}$$

A partir $\hat{\Sigma}$ se calcularon las estimaciones máximo verosímiles de los coeficientes de correlación múltiples, $\hat{R}^2_S$, del conjunto de datos B2.

Las estimaciones máximo verosímiles de los $\hat{R}^2$ para el conjunto de datos originales B1 y la de los $\hat{R}^2_S$ y $\hat{Q}^2_S$ para el conjunto con información faltante B2, se presentan en la Tabla 1. Dichos coeficientes fueron calculados para todas las regresiones posibles (2^4) con las 4 variables explicativas disponibles.

Tabla 1: Coeficientes de correlación múltiple de los conjuntos de datos B1 y B2 para todas las regresiones posibles.

S Conjunto de Predictores	$\hat{R}^2$ Datos Originales	$\hat{R}^2_S$ Datos incompletos	$\hat{\pi}_S$	$\hat{Q}^2_S$
0	0	0	0	0
1	0.4373	0.4373	$\frac{49}{179}$	$\frac{49}{179}0.4373+\frac{1}{179}0.4373+\frac{24}{179}0.4373+\frac{105}{179}0.4373=0.4373$
2	0.5648	0.7728	0	$\frac{105}{179}0.7728=0.4533$
3	0.6313	0.7017	0	$\frac{24}{179}0.70175+\frac{105}{179}0.70175=0.5007$
4	0.0626	0.0861	0	$\frac{1}{179}0.0861+\frac{24}{179}0.0861+\frac{105}{179}0.0861=0.0120$
12	0.9291	0.9938	0	$\frac{49}{179}0.4373+\frac{1}{179}0.4373+\frac{24}{179}0.4373+\frac{105}{179}0.9938=0.7637$
13	0.6672	0.7104	0	$\frac{49}{179}0.4373+\frac{24}{179}0.7104+\frac{105}{179}0.7104=0.6317$
14	0.4692	0.4687	$\frac{1}{179}$	$\frac{49}{179}0.4373+\frac{1}{179}0.4687+\frac{24}{179}0.4687+\frac{105}{179}0.4687=0.4601$
23	0.8340	0.9300	0	$\frac{24}{179}0.7017+\frac{105}{179}09300=0.6396$
24	0.5711	0.7781	0	$\frac{24}{179}0.0861+\frac{105}{179}0.7781=0.4680$
34	0.6457	0.7104	0	$\frac{1}{179}0.0861+\frac{24}{179}0.7104+\frac{105}{179}0.7104=0.5124$
123	0.9495	0.9998	0	$\frac{49}{179}0.4373+\frac{1}{179}0.4373+\frac{24}{179}0.7104+\frac{105}{179}0.9998=0.8039$
124	0.9296	0.9940	0	$\frac{49}{179}0.4373+\frac{1}{179}0.4687+\frac{24}{179}0.4687+\frac{105}{179}0.994=0.6462$
134	0.6814	0.7189	0	$\frac{49}{179}0.4373+\frac{1}{179}0.4687+\frac{24}{179}0.7189+\frac{105}{179}0.7189=0.6404$
234	0.8357	0.9307	$\frac{24}{179}$	$\frac{1}{179}0.0861+\frac{24}{179}0.7104+\frac{105}{179}0.9307=0.6416$
1234	0.9499	0.999	$\frac{105}{179}$	$\frac{49}{179}0.4373+\frac{1}{179}0.4687+\frac{24}{179}0.7189+\frac{105}{179}0.999=0.8052$

A partir de estos resultados se puede destacar que:

• la mayoría de los valores de $\hat{Q}_S^2$ son menores que los correspondientes valores de $\hat{R}_S^2$. Esto indica que la habilidad de predecir un nuevo valor de la variable Y usando un subconjunto de variables explicativas puede ser substancialmente menor que si no existieran datos perdidos;

• el perímetro cefálico (X_3) es el mejor predictor simple de valores futuros de la variable Y puesto que está altamente correlacionada con ella. Dicha variable está observada en el 72% de las unidades. Cabe destacar que a pesar que $\hat{R}_S^2$ es más grande que $\hat{R}_3^2$, esto sucede porque la variable X_2 tiene más pérdidas que la X_3. Mientras que, la variable edad de la embarazada (X_4) es el peor predictor simple de valores futuros de la variable Y ya que está levemente correlacionada con ella. Dicha variable está observada en el 73% del total de unidades;

• la edad gestacional en semanas cumplidas (X1) y el percentil del peso por edad gestacional (X2) son los mejores pares de predictores de valores futuros de la variable Y;

• la edad gestacional en semanas cumplidas (X_1), el percentil del peso por edad gestacional (X_2) y el perímetro cefálico (X_3), es la mejor terna de los predictores de valores futuros de la variable Y.

Discusión

El método de máxima verosimilitud resulta una opción válida para la estimación de los parámetros, cuando se dispone de un conjunto de datos que presenta pérdidas parciales en algunas variables. El mismo permite incorporar toda la información disponible de las variables observadas incompletamente y su implementación se ve facilitada a través de la utilización del operador “Sweep”.

Para comparar subconjuntos de variables explicativas y decidir cuál de ellos es el “mejor” para predecir valores de la variable respuesta, una de las posibles medidas a utilizar es el coeficiente de correlación múltiple. Si los datos están observados parcialmente, la comparación a menudo debería reflejar no sólo cuan correlacionadas están las variables X con la Y, sino también cuan probablemente ellas estén observadas. Así una variable X que está altamente correlacionada con Y pero está observada parcialmente no es útil para predecir valores futuros de Y como una variable X menos correlacionada pero observada totalmente. Se presenta una aplicación de la generalización del coeficiente de correlación múltiple, el cual es apropiado cuando existen valores perdidos y coincide con el coeficiente de correlación múltiple cuando las variables están observadas completamente.

Referencias bibliográficas

1. Rubin, D. B. (1976). “Comparing regressions when some predictor values are missing”. Technometrics, vol. 18, Nº 2, pp. 201-205.

2. Little, R. J. (1992). ”Regression with missing X´s: a review”. Journal of the American Statistical Association, vol. 87, pp. 1227-1237.

3. Little, R. J. and D. B. Rubin. (1987). “Statistical Analysis with Missing Data”. John Wiley & Sons. New York.

4. Badler, C; Alsina, S.; Beltrán, C.; Bussi, J.; Puigsubirá, C.; Vitelleschi, M. (2001).”¿Eliminar o utilizar?.Estimación máximo verosímil ante la presencia de información faltante”. Revista de la Sociedad Argentina de Estadística. Vol. 5, Nº 1-2, pp.17-32.