Badler, Clara Elisabeth
Puigsubirá, Cristina Raquel[1]
Vitelleschi, María Susana[1]
Instituto de Investigaciones Teóricas y Aplicadas de la Escuela de Estadística (IITAE)

COMPARACIÓN DE SUBCONJUNTOS DE REGRESIONES CUANDO ALGUNAS DE LAS VARIABLES EXPLICATIVAS ESTÁN OBSERVADAS PARCIALMENTE

INTRODUCIÓN

En la mayoría de los problemas de regresión, el objetivo es determinar la habilidad de un conjunto de variables explicativas (X) para predecir una variable respuesta (Y). Si se tienen p variables explicativas, las mismas podrían usarse para predecir valores futuros de Y. Sin embargo, a veces resulta costoso registrarlas a todas y además algunas de ellas podrían estar observadas parcialmente. Cuando ocurre alguna de estas situaciones, el investigador deberá decidir cuál será el "mejor" subconjunto de variables explicativas.

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

En este trabajo se presenta una adaptación del coeficiente de correlación múltiple propuesto por Donald Rubin (1976), el cual es apropiado cuando existe información faltante en algunas de las variables X. La metodología es aplicada a un conjunto de datos provenientes de las historias clínicas perinatales de niños nacidos en el Hospital Sáenz Peña de Rosario durante el año 2002.

METODOLOGÍA

La selección de un subconjunto S de las 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 correla-

[1] Docente-Investigador e Investigador del Consejo de Investigaciones de la Universidad Nacional de Rosario

ción 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, podría no ser considerada esa nueva variable, 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 estén perdidos y esto es causado por un mecanismo probabilístico. 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 π_T a través de las 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 esta 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:

$$\sum_T \pi_T (1 - R^2_{S \cap T})\, \sigma^2 \qquad (1).$$

La cual se puede escribir de la forma:

$$(1 - Q^2_S)\, \sigma^2$$

siendo:

$$Q^2_S = \sum_T \pi_T \, R^2_{S \cap T} \; . \qquad (2)$$

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

El estimador máximo verosímil de Q^2_S es:

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

siendo:

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

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

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

OPERADOR "SWEEP"

El operador "Sweep" 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:

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

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]**G** = SWP[k,j]**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]**G** = **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 μ_1 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:

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

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:

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

Esta ecuación define la transformación de $\hat{\phi}$ a $\hat{\theta}$.

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):

n	y	x_1	x_4	x_3	x_2
1					
2					
.					
105					
.					
129					
130					
.					
179					

Figura 1: Esquema de pérdida monótono

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:

$$\hat{\mu} = \begin{bmatrix} 3113.29 & 38.47 & 21.99 & 343.19 & 39.41 \end{bmatrix}$$

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

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

Se presentan en la Tabla 1 las estimaciones máximo verosímiles de los $\hat{R}^2$ para el conjunto de datos originales B1 y la de los $\hat{R}_S^2$ y $\hat{Q}_S^2$ para el conjunto con informnación faltante B2. 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	$\frac{24}{179}$	$\frac{49}{179}$ 0.4373+ $\frac{1}{179}$ 0.4687+ $\frac{24}{179}$ 0.7189+ $\frac{105}{179}$ 0.7189=0.6404
234	0.8357	0.9307	0	$\frac{1}{179}$ 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 $\hat{Q}^2_S$ son, substancialmente, menores que los correspondientes a $\hat{R}^2_S$. Esto indica que la habilidad de predecir un nuevo valor de Y usando un subconjunto de variables explicativas puede ser sustancialmente menor que si no existieran datos perdidos;

- la variable perímetro cefálico (x_3) es el mejor predictor simple de valores futuros de Y puesto que está altamente correlacionada con ella. Dicha variable está observada en el 72% de las unidades. Cabe destacar que a pesar que $\hat{R}_2^2$ 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 Y ya que está levemente correlacionada con ella. Dicha variable está observada en el 73% del total de unidades;
- las variables edad gestacional en semanas cumplidas (x_1) y percentil del peso por edad gestacional (x_2) son los mejores pares de predictores de valores futuros de Y;
- variables edad gestacional en semanas cumplidas (x_1), percentil del peso por edad gestacional (x_2) y perímetro cefálico (x_3), es la mejor terna de los predictores de valores futuros de 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 conjuntos 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 solo cuan correlacionadas están las variables X con 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 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

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.

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

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

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