Regiones de predicción mejoradas en modelos de regresión simétricos

Fernando Lucambio

Departamento de Estatística
Universidade Federal do Paraná
Curitiba, Paraná, Brasil

Enero de 2012

Resumo

Recientemente diversos trabajos tratan sobre intervalos de predicción en modelos paramétricos, el objetivo de eses trabajos és el mejoramiento del limite de predicción para nuevas observaciones Z, o sea, encontrar funciones $c_\alpha(y)$ tales que $P\{Z < c_\alpha(Y); \vartheta\} = \alpha$, de manera exata o aproximada. Una de las ideas en ese sentido és presentada en Vidoni (1998) en la qual se utiliza la fórmula p^*, posteriormente extendida a los modelos lineares generalizados. Esa propuesta solamente funciona en el caso de dimensión uno de Z, por otro lado, Barndorff-Nielsen & Cox (1996) utilizaron teoría asintótica con el mismo objetivo y obtuvieron un resultado posteriormente extendido por Corcuera & Giummolè (2006) al caso multidimensional. En este trabajo utilizaremos el resultado en Corcuera & Giummolè (2006) para encontrar regiones de predicción mejoradas en modelos de regresión simétricos y mostraremos, a través de ejemplos, que la solución propuesta és una mejoria con relación al procedimiento padrón.

Palabras clave: probabilidad de cobertura, regiones de predicción, modelos de regresión simétricos.

1 Introdução

En este trabajo será considerado el problema de predicción, por regiones de confianza, de un vector aleatório $Z = (Z_1, \ldots, Z_m)$ con base en una muestra $y = (y_1, \ldots, y_n)$ del vector aleatório $Y = (Y_1, \ldots, Y_n)$. Asumiremos que la función de densidad de Y y la densidad condicional de Z dado $Y = y$, $g(z; \vartheta|y)$, son conocidas a menos un vector de parámetros $\vartheta \in \Theta \subset \mathbb{R}^d$. La predicción de Z será a través de regiones $R_\alpha(Y) \subset \mathbb{R}^m$ tales que

$$P\{Z \in R_\alpha(Y)\} = \alpha,$$

para todo $\vartheta \in \Theta$ y para $\alpha \in (0,1)$, fijo. Esta probabilidad és conocida como probabilidad de cobertura y és calculada con relación a la función de densidad conjunta de Z y Y.

La forma mas simple de realizar predicciones de Z és a través de la conocida función de densidad estimada predictiva $g(z;\widetilde{\vartheta}|y)$, donde $\widetilde{\vartheta}$ és un estimador de ϑ asintoticamente eficiente, usualmente siendo el estimador de máxima verosimilitud $\widehat{\vartheta}$. Las regiones de predicción obtenidas de esta maneira son imprecisas, con error de cobertura de orden $O(n^{-1})$, o sea, $P\{Z \in R_\alpha(Y)\} = \alpha + O(n^{-1})$. Este és un resultado conocido cuando Z és de dimensión uno (Barndorff-Nielsen & Cox, 1994).

En este sentido Barndorff-Nielsen & Cox (1996) y Vidoni (1998) sugieren maneras de corregir los quantiles de la densidad estimada predictiva obtiendo limites de predicción con error de cobertura de orden $o(n^{-1})$. Estas propuestas pueden ser aplicadas unicamente en el caso uni-dimensional, pués no consideran la dependencia induzida entre las componentes de Z al estimar ϑ con los mismos datos.

De manera general el problema de predicción por regiones de confianza para nuevas observaciones se reduce a encontrar funciones $c_1(\widehat{\vartheta})$, $c_2(Z^{(1)};\widehat{\vartheta})$, ..., $c_m(Z^{(m-1)};\widehat{\vartheta})$ tales que

$$P\{Z_1 < c_1(\widehat{\vartheta}), Z_2 < c_2(Z^{(1)};\widehat{\vartheta}), \ldots, Z_m < c_m(Z^{(m-1)};\widehat{\vartheta})\} = \alpha + o(n^{-1}),$$

donde la igualdad se cumple para todo $\vartheta \in \Theta$ con $Z^{(i)} = (Z_1, \ldots, Z_i)$, las primeras i componentes de Z e $i = 1, \ldots, m-1$. El sistema de funciones $c_1(\widehat{\vartheta})$, $c_2(Z^{(1)};\widehat{\vartheta})$, ..., $c_m(Z^{(m-1)};\widehat{\vartheta})$ que cumplen la igualdad anterior será llamado de limites simultaneos de predicción para Z.

En la Sección 2 presentamos la teoria para encontrar el sistema de limites simultaneos de predicción mejorados. La Sección 3 és dedicada a la obtención de estes limites en modelos de regresión simétricos y en la Sección 4 presentamos diversos resultados de simulación que corroboran la calidad del sistema de limites simultaneos mejorados en estos modelos de regresión.

2 Región de predicción mejorada

El resultado en Corcuera & Giummolè (2006) extiende la idea de Barndorff-Nielsen & Cox (1996) para el caso Z m-dimensional con componentes absolutamente contínuas, independientes o no, también asume independencia entre Z e Y. Aqui hacemos un resumén de este resultado.

Utilizaremos la notación para sumas de Einstein e asi escribir de manera simples los resultados, por tanto entenderemos que cuando aparezcan índices repetidos como subescritos y sobrescritos en una expresión significa que estamos sumando en estes índices.

Definimos

$$B^i_{rs}(z;\vartheta) = \frac{1}{2}\left(\frac{g^i_r(z;\vartheta)}{g^i(z;\vartheta)}G^i_s(z;\vartheta) + \frac{g^i_s(z;\vartheta)}{g^i(z;\vartheta)}G^i_r(z;\vartheta) - G^i_{rs}(z;\vartheta)\right),$$

donde los subíndices r y s indican derivación con relación a las correspondientes componentes del vector de parámetros, $g^i(z;\vartheta)$ és la función de densidad marginal de Z_i y $G^i(z;\vartheta)$ denota la función de distribuición acumulada de Z_i, $i = 1,\ldots,m$. Observemos que nuestra situación estamos asumiendo independencia entre las componentes do vetor Z e entre Z e Y, podemos entonces escribir la función de densidad condicional de Z dado $Y = y$ como $g(z;\vartheta) = \prod_{i=1}^{m} g^i(z_i;\vartheta)$.

Sean

$$b^r(\vartheta) = \mathrm{E}\{(\widehat{\vartheta} - \vartheta)^r\}$$

y

$$i^{rs}(\vartheta) = \mathrm{E}\{(\widehat{\vartheta} - \vartheta)^r(\widehat{\vartheta} - \vartheta)^s\},$$

el vício de la esperanza y de la variancia de $\widehat{\vartheta}$, respectivamente. Los limites simultaneos de predicción mejorados, según fueron obtenidos en Corcuera & Giummolè (2006) son

$$c_1(\widehat{\vartheta}) = q^1_{\alpha_1}(\widehat{\vartheta}) - \frac{h^1(q^1_{\alpha_1}(\widehat{\vartheta});\widehat{\vartheta})}{g^1(q^1_{\alpha_1}(\widehat{\vartheta});\widehat{\vartheta})} \tag{1}$$

y, para $i = 2,\ldots,m$,

$$c_i(z^{(i-1)};\widehat{\vartheta}) = q^i_{\alpha_i}(\widehat{\vartheta}) - \frac{h^i(q^i_{\alpha_i}(\widehat{\vartheta});\widehat{\vartheta})}{g^i(q^i_{\alpha_i}(\widehat{\vartheta});\widehat{\vartheta})} - \sum_{j<i} \frac{h^{ij}(q^i_{\alpha_i}(\widehat{\vartheta}), z_j;\widehat{\vartheta})}{g^i(q^i_{\alpha_i}(\widehat{\vartheta});\widehat{\vartheta})}, \tag{2}$$

donde $q^i_{\alpha_i}$ és el cuantil α_i de $g^i(z;\vartheta)$. Las funciones $h^i(z_i;\vartheta)$ y $h^{ij}(z_i,z_j;\vartheta)$ son de orden $O(n^{-1})$, las quales pueden ser escritas para $i = 1,\ldots,m$ e $j < i$, como

$$h^i(z;\vartheta) = -b^r(\vartheta)G^i_r(z;\vartheta) + i^{rs}(\vartheta)B^i_{rs}(z;\vartheta) \tag{3}$$

y

$$h^{ij}(z_i,z_j;\vartheta) = i^{rs}(\vartheta)G^i_r(z_i;\vartheta)\frac{g^j_s(z_j;\vartheta)}{g^j(z_j;\vartheta)}. \tag{4}$$

En diversos trabajos, por ejemplo, Vidoni (2001) e Vidoni (2003) son obtenidas expresiones del intervalo de predicción en modelos de regresión limitándose a la predicción de uma única variáble. El objetivo de este trabajo és extender los resultados presentados en esta sección a los modelos de regresión simétricos cuando nos interesa hacer predicciones por regiones de confianza.

3 Modelos de regresión simétricos

En artículo reciente Cordeiro *et al.* (2000) generalizaron trabajos anteriores y desenvolvieron una nueva clase de modelos llamados de modelos de regresión no lineares simétricos. És conocido que el modelo de regresión normal no siempre és un buen modelo para representar datos conteniendo observaciones extremas o periféricas. Para superar este problema, nuevos modelos estadísticos han sido propuestos. Distribuciones

simétricas estan apareciendo con frequencia cresciente en la literatura estadística para modelar diversos tipos de datos que contienen observaciones mas distantes de lo que seria esperado tomando como base la distribución normal.

Esta nueva clase de modelos incluye distribuciones tales como normal, t-Student, potencia exponencial, logísticas I e II e los modelos normais contaminados, permitiendo aplicar una amplia variedad de modelos para varios tipos de datos.

Decimos que la variáble aleatória Y tiene distribución absolutamente contínua simétrica con parámetro de locación $\mu \in R$ y parámetro de dispersión $\phi > 0$, si la función de densidad és

$$f(y;\mu,\phi) = \frac{1}{\phi} g\left\{ \left(\frac{y-\mu}{\phi} \right)^2 \right\}, \quad y \in R, \tag{5}$$

para alguna función g (independiente de y, μ e ϕ), tal que $g(u) > 0$, para $u > 0$ e $\int_0^\infty u^{-1/2} g(u) du = 1$, donde y és una realización de Y. Escribimos $Y \sim S(\mu, \phi)$.

La família de densidades simétricas de locación y escala (5) retiene la estructura de la distribución normal eliminando la forma específica de la densidad normal. Esta família incluye densidades simétricas que tienen extremos menores o mas altos que los de la normal. Una relación detallada de estas distribuciones e áreas de aplicación puede ser encontrada en Chmielewski (1981).

Para introduzir una estructura de regresión en la clase de modelos (5), definimos los modelos de regresión linear simétricos como

$$Y_i = \mu_i(\beta) + \epsilon_i, \tag{6}$$

donde $\mu_i(\beta) = x_i^\top \beta$, $\beta = (\beta_1, \ldots, \beta_p)^\top$, x_i és un vector de dimensión $p \times 1$ con los valores de las variábles explicativas e $\epsilon_i \sim S(0, \phi)$, $i = 1, \ldots, n$. Caso existan, tenemos que $\mathrm{E}\{Y_i\} = \mu_i$ e $\mathrm{Var}\{Y_i\} = \xi\phi$, siendo $\xi > 0$ una constante que puede ser obtenida para cada situación particular, por ejemplo, para la distribución normal $\xi = 1$ y en el caso de la distribución t-Student con ν grados de libertad tenemos que $\xi = \nu/(\nu-2)$ si $\nu > 2$. Observamos que el argumento de la función $g(\cdot)$ en (5) depende de y, β e ϕ, luego podemos escribir $g(u^2) = g(y, \beta, \phi)$, donde $u = u(y, \beta, \phi)$ e $u(y, \beta, \phi) = (y - x^\top\beta)/\phi$. En el artículo de Galea *et al.* (2003) son descritas medidas de diagnóstico para los modelos definidos en (6).

Ahora estamos en condiciones de desarrollar e investigar la calidad de las regiones de predicción en los modelos de regresión simétricos.

3.1 Regiones de predicción mejoradas en modelos de regresión simétricos

Utilizando las expresiones (1), (2), (3) e (4) vamos encontrar el sistema de limites de predicción mejorados en los modelos de regresión simétricos observando que $\vartheta = (\beta, \phi)$, o sea, que el vector de parámetros completo és de dimensión $p+1$. Tambien

devemos observar que el vector de respuestas en estos modelos és independiente, luego nuestras expresiones deven ser bastante simplificadas.

Con este objetivo vemos que el vício de la esperanza de orden $O(n^{-1})$ del vector $\widehat{\beta}$ és nulo, resultado que puede ser demostrado utilizando la fórmula de Cox & Snell (1968), entonces

$$h^i(z;\vartheta) = -b^{\phi}(\vartheta)G^i_{\phi}(z;\vartheta) + i^{rs}(\vartheta)B^i_{rs}(z;\vartheta),$$

para $i = 1, \ldots, m$, el vicio de $\widehat{\phi}$ de orden $O(n^{-1})$ fue obtenido en Cordeiro *et al.* (2000) como

$$b^{\phi}(\vartheta) = \frac{\phi}{2n(\alpha_{2,2}-1)}\left\{p\left(\frac{\alpha_{3,1}}{\alpha_{2,0}}+2\right)+\frac{\alpha_{3,3}+2\alpha_{2,2}}{(\alpha_{2,2}-1)}\right\},$$

donde $\alpha_{r,s} = \mathrm{E}\{t^{(r)}(U)U^s\}$, para $r, s = 0, 1, 2, 3$, $t(u) = \log g(u^2)$, $t^{(r)} = \mathrm{d}^r t(u)/\mathrm{d}u^r$ e $U \sim S(0,1)$.

En este último artículo los autores también mostraron que la matriz de información de Fisher de $\widehat{\vartheta}$ és

$$K(\widehat{\vartheta}) = \begin{pmatrix} -\dfrac{\alpha_{2,0}}{\phi^2}X^{\top}X & 0 \\ 0 & \dfrac{n}{\phi^2}(1-\alpha_{2,2}) \end{pmatrix},$$

también están disponibles expresiones de $\alpha_{r,s}$ em diversas situaciones particulares de modelos de regresión simétricos.

A seguir reservaremos los índices r e s para indicar las componentes del vector de parámetros de regresión β, asi $\vartheta_r = \beta_r$, $r = 1, \ldots, p$ e $\vartheta_{\phi} = \phi$. Consideramos que el vector de nuevas observaciones tiene función de densidad de la forma (5), que $u_i = (z_i - \mu_i)/\phi$ y que el predictor linear sea $\mu_i = \omega_i^{\top}\beta$, onde $\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{ip})^{\top}$ és un vector conocido para cada nueva observación, $i = 1, \ldots, m$.

Entonces, para cada componente del vector Z, obtenemos que

$$\frac{g_r(z;\beta,\phi)}{g(z;\beta,\phi)} = -2u\omega_r W_g(u^2)/\phi$$

e

$$\frac{g_{\phi}(z;\beta,\phi)}{g(z;\beta,\phi)} = -\left\{1+2u^2W_g(u^2)\right\}/\phi\cdot$$

Tambien encontramos las expressiones de $G_r(z;\beta,\phi) = -\omega_r g(u^2)/\phi$ e $G_{\phi}(z;\beta,\phi) = -ug(u^2)/\phi$. Las segundas derivadas tienen formas mas complicadas, asi

$$G_{rs}(z;\beta,\phi) = 2u\omega_r\omega_s W_g(u^2)g(u^2)/\phi^2,$$

$$G_{r\phi}(z;\beta,\phi) = \omega_r g(u^2)\left\{1+2u^2W_g(u^2)\right\}/\phi^2,$$

$$G_{\phi s}(z;\beta,\phi) = 2u^2\omega_s W_g(u^2)g(u^2)/\phi^2$$

y que

$$G_{\phi\phi}(z;\beta,\phi) = ug(u^2)\left\{3 + 2u^2 W_g(u^2)\right\}/\phi^2\cdot$$

Expressões para $\alpha_{2,0}$, $\alpha_{2,1}$, $\alpha_{3,1}$ e $\alpha_{3,3}$ son encontradas em Cordeiro *et al.* (2000) para algunas situaciones particulares de modelos de regresión simétricos y en el artículo de Cysneiros & Paula (2005) podemos encontrar expressiones de $W_g(u^2) = g'(u)/g(u^2)$, tambien para algunas situaciones particulares, donde $g'(u) = \mathrm{d}g(u)/\mathrm{d}u$. Devemos observar que por la particular forma de la función de densidad en los modelos de regresión simétricos $f_r/f = g_r/g$ asi como en las otras expresiones.

4 Ejemplos de modelos de regresión simétricos

Aqui mostraremos la calidad en las predicciones por regiones de confianza obtenida cuando usamos los limites simultaneos de predicción mejorados. En algunos modelos especificos comparamos la probabilidad de cobertura cuando utilizamos los límites de predicción mejorados y aquella obtenida por la forma tradicional. Realizaremos simulaciones Monte Carlo y en cada caso calculamos ambas regiones de predicción.

4.1 Modelo de regresión t-Student

Consideremos que $Y_1,\ldots,Y_n$ e $Z_1,\ldots,Z_m$ son variábles aleatórias independientes con distribución t-Student de parámetros β, ϕ y ν grados de libertad e funcion de densidad

$$g(u) = \frac{\nu^{\nu/2}}{\mathrm{B}(1/2,\nu/2)}(\nu+u^2)^{-(\nu+1)/2},$$

donde $\mathrm{B}(\cdot,\cdot)$ és la función Beta.

La forma mas simples de obter los limites simultanios de predicción és definirlos como los quantis de la función de densidade estimada $g(z;\widehat{\vartheta})$, la qual és obtenida substituyendo θ por su estimador consistente $\widehat{\theta}$, de esta forma los limites simultanios de predicción son

$$c_i(\widehat{\theta}) = G^{-1}(\alpha_i;\widehat{\theta})$$

o sea, $c_i(\widehat{\theta}) = q_{\alpha_i}(\widehat{\theta})$ donde $q_{\alpha_i}(\widehat{\theta})$ és el quantil α_i de $g(z;\widehat{\theta})$, $i = 1,\ldots,m$ e $\alpha = \prod_{i=1}^m \alpha_i$.

Para obtener los limites simultanios de predicción mejorados son necesarias las expressiones de la sección anterior y también $\alpha_{2,0} = -1/2$, $\alpha_{2,2} = 1/2$, $\alpha_{3,1} = 1/3$, $\alpha_{3,3} = -1/2$,

$$W_g(u^2) = -\frac{\nu+1}{2(\nu+u^2)}, \qquad b^{\phi}(\vartheta) = \phi(1-2p/3)/n$$

y también

$$G_r(z;\vartheta) = -\omega_r g(u^2)/\phi, \qquad G_{\phi}(z;\vartheta) = -ug(u^2)/\phi\cdot$$

En la Tabla 1 mostramos los resultados de un estudio de simulación realizado con el objetivo de mostrar la mejoria obtenida en la probabilidad de cobertura cuando

Tabela 1: Valores médios de la probabilidad de cobertura, en 20.000 réplicas Monte Carlo, para diferentes tamaños de muestra n e diferentes número de perámetros de regresión p y $m = 1$, en el modelo de regresión t-Student. Las estimativas de la probabilidad de cobertura fueron obtenidas utilizando los métodos (a) tradicional y (b) mejorado.

		α								
n, p		0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
15, 2	(a)	0.138	0.241	0.331	0.416	0.499	0.582	0.667	0.757	0.860
	(b)	0.105	0.205	0.303	0.401	0.499	0.597	0.694	0.793	0.894
20, 3	(a)	0.137	0.240	0.331	0.416	0.499	0.582	0.667	0.759	0.863
	(b)	0.105	0.206	0.305	0.402	0.499	0.596	0.694	0.793	0.894
25, 4	(a)	0.155	0.262	0.349	0.427	0.501	0.575	0.652	0.739	0.846
	(b)	0.116	0.221	0.319	0.411	0.501	0.591	0.683	0.780	0.885

utilizamos los resultados presentados en las Secciones 2, 3 e 3.1. Fueron realizadas, en cada caso, 20.000 réplicas Monte Carlo del modelo de regresión t-Student con número de parámetros de regresión p iguales a 2, 3 y 4, tamaños de muestra n, 15, 20 e 25, $\phi = 2$, grados de libertad $\nu = 3$ e solamente consideramos el vector Z de nuevas observaciones univariado. Para realizar este trabajo se utilizo el programa Ox versión 6.21 (Windows/U) (C) J.A. Doornik, 1994-2011 (www.doornik.com).

Podemos observar que la probablidad de cobertura fue siempre subestimada si utilizamos los limites simultanios de predicción simples y la predicción pretendida és menor que 0.5, en 0.5 existe coincidencia entre ambas formas de obter los limites dimultanios de predicción y si la predicción pretendida és mayor que 0.5 la probabilidad de cobertura és subrestimada. Si utilizamos los limites simultaneos de predicción mejorados los resultados obtenidos fueron impresionantemente satisfactórios. En todos los casos el error de estimación de la probabilidad de cobertura mejorada fué menor de 0.06.

5 Considerações finais

Extendemos aqui los resultados de mejoramiento de regiones de predicción multivariadas para nuevas observaciones a los modelos lineares simétricos. Encontramos expresiones gerenrales que, para cada modelo particular, son facilmente programadas. En el modelo de regresión t-Student mostramos que la probabilidad de cobertura és mejorada con los resultados presentados aqui en situaciones de tamaño de muestra relativamente pequeño. Más estudios de simulación son necesários en los quales consideraremos el vector de nuevas observaciones multivariado.

6 Agradecimiento

Agradezco el apoyo de la Coordenação de Aperfeiçoamento de Pessoal de Nível Superior CAPES para la presentación de este trabajo en este evento.

Referências

Barndorff-Nielsen, O.E. & Cox, D.R. (1989). *Asymptotics techniques for use in statistics*. Chapman and Hall, London.

Barndorff-Nielsen, O.E. & Cox, D.R. (1994). *Inference and Asymptotics*. Chapman and Hall, Oxford.

Barndorff-Nielsen, O.E. & Cox, D.R. (1996). Prediction and asymptotics. *Bernoulli*, **2**, 319–340.

Chmielewski, M.A. (1981). Elliptically symmetric distributions: a review and bibliography. *International Statistical Review*, **49**(1), 67–74.

Corcuera, J.M. & Giummolè, F. (2006). Multivariate prediction. *Bernoulli*, **12**(1), 157–168.

Cordeiro, G.M., Ferrari, S.L.P., Uribe-Opazo, M.A. & Vasconcellos, K.L.P. (2000). Corrected maximum-likelihood estimation in a class of symmetric nonlinear regression models. *Statistics & Probability Letters*, (46), 317–328.

Cox, D.R & Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, **30**, 248–278.

Cysneiros, F.J.A. & Paula, G.A. (2005). Restricted methods in symmetrical linear regression models. *Computational Statistics & Data Analysis*, **49**, 689–708.

Galea, M., Paula, G.A. & Uribe-Opazo, M. (2003). On influence diagnostic in symmetrical linear regression models. *Statistical Papers*, **44**, 23–45.

Vidoni, P. (1998). A note on modified estimative prediction limits and distributions. *Biometrika*, **85**, 949–953.

Vidoni, P. (2001). Improved prediction limits for continuous and discrete observations in generalized linear models. *Biometrika*, **88**, 881–887.

Vidoni, P. (2003). Prediction and calibration in generalized linear models. *Annals of the Institute of Statistical Mathematics*, **55**(1), 169–185.