VI Congreso Galego de Estatística e Investigación de Operacións
Vigo 5–7 de Novembro de 2003

ESTIMACIÓN DE LA FUNCIÓN DE DISTRIBUCIÓN EN POBLACIONES FINITAS. UNA APLICACIÓN A DATOS REALES.

Lombardía M.J.[1], González-Manteiga W.[1], Prada-Sánchez, J.M.[1]

[1]Departamento de Estatística e Investigación Operativa
Universidade de Santiago de Compostela

RESUMO

Consideremos una población finita **P**, entendida como una realización de un modelo de superpoblación, y supongamos como modelo de superpoblación un modelo de regresión lineal con errores heterocedásticos. En este trabajo se estudia el estimador de Chambers-Dunstan de la función de distribución de las poblaciones finitas, generadas a partir del modelo de superpoblación indicado.

Palabras e frases chave: Modelo de superpoblación, Información auxiliar, Parámetro ventana, Estimación tipo núcleo.
Clasificación AMS: 62D05.

1. INTRODUCCIÓN

En poblaciones finitas, el estimador de Chambers y Dunstan (1986) es de gran utilidad en el estudio de la función de distribución de una variable, cuando ésta se relaciona con una variable auxiliar a través de un modelo de regresión lineal homocedástico. En este trabajo proponemos una extensión del estimador de Chambers y Dunstan (1986) a modelos de regresión lineal heterocedásticos con varianza del modelo desconocida, estimando la función varianza no paramétricamente, de acuerdo a lo indicado en Carroll (1982). Se presentan dos estimadores de la función de distribución. El primero se construye a partir de los estimadores de mínimos cuadrados ponderados de los parámetros del modelo y el segundo utiliza los estimadores de mínimos cuadrados de modelos homocedásticos; en ambos casos se estima de forma no paramétrica la varianza. La muestra que se toma de la población es sin reemplazamiento.

El error de predicción de estos estimadores es asintóticamente normal e insesgado y su varianza asintótica generaliza la obtenida por Chambers et al. (1992) para modelos de regresión lineal homocedásticos como modelo de superpoblación.

A continuación, en la Sección 2 se presentan los estimadores de la función de distribución basados en el modelo de regresión lineal heterocedástico, siguiendo la idea de Chambers y Dunstan (1986). En esta sección también se estudia su comportamiento asintótico. Finalmente, en la Sección 3 analizamos el comportamiento de los estimadores propuestos sobre una población finita de plantaciones de caña de azúcar (Chambers y Dunstan, 1986).

2. EL ESTIMADOR PARAMÉTRICO DE CHAMBERS-DUNSTAN

Sea P el conjunto de enteros $\{1, ..., N\}$, S un subconjunto de n-elementos de P y $P - S$ el complementario de S en P. Consideremos una población finita $\mathbf{P} = \{(Y_k, x_k)\}_{k \in P}$, donde los valores x_k de la variable auxiliar X son conocidos para todos los elementos de la población y la variable Y se relaciona con X mediante el modelo

$$\xi : Y_k = \alpha + \beta x_k + \eta_k = \alpha + \beta x_k + \sigma(x_k)\varepsilon_k, \tag{1}$$

siendo los ε_k variables aleatorias independientes, idénticamente distribuidas, de media cero y varianza unidad, y los parámetros del modelo (α, β) y la varianza $\sigma^2(x_k)$ cantidades desconocidas que se estimarán a partir de los datos muestrales. Las variables Y_k sólo son conocidas para los elementos de la muestra $\mathbf{S} = \{(Y_i, x_i)\}_{i \in S}$, tomada aleatoriamente y sin reemplazamiento de $\mathbf{P}$.

Los estimadores de α y β más utilizados en la literatura son los mínimos cuadráticos, estos estimadores son insesgados pero no de mínima varianza cuando se trata de modelos de regresión lineal heterocedásticos. Una modificación adecuada consiste en tomar como estimadores los que minimizan la expresión

$$\sum_{i \in S} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2(x_i)}, \tag{2}$$

que denotaremos por

$$\begin{aligned} \hat{\alpha}_T &= \bar{y}_w - \hat{\beta}_T \bar{x}_w, \\ \hat{\beta}_T &= \frac{\sum_{i \in S} w_i Y_i x_i - \bar{y}_w \bar{x}_w}{\sum_{i \in S} w_i (x_i - \bar{x}_w)^2} \end{aligned} \tag{3}$$

con pesos $w_i = \sigma^{-2}(x_i)$ y $\bar{y}_w = \left[\sum_{i \in S} w_i\right]^{-1} \sum_{i \in S} w_i Y_i$, $\bar{x}_w = \left[\sum_{i \in S} w_i\right]^{-1} \sum_{i \in S} w_i x_i$ los promedios ponderados en los elementos de la muestra relativos a las variables Y y X. En base a las propiedades de consistencia de la estimación de la función de regresión, estimaremos la función varianza por (Carroll, 1982)

$$\hat{\sigma}_h^2(x) = \frac{\sum_{i \in S} k\left\{(x - x_i)/h\right\} \left(Y_i - \hat{\alpha}_L - \hat{\beta}_L x_i\right)^2}{\sum_{i \in S} k\left\{(x - x_i)/h\right\}}, \tag{4}$$

y denotaremos por $\hat{\alpha}_E$ y $\hat{\beta}_E$ a los estimadores de α y β, obtenidos al reemplazar $\sigma^2(x)$ por $\hat{\sigma}_h^2(x)$ en las correspondientes expresiones de $\hat{\alpha}_T$ y $\hat{\beta}_T$.

Nuestro objetivo es estudiar la función de distribución en la población finita $\mathbf{P}$,

$$F_N(t) = N^{-1} \sum_{k \in P} I(Y_k \leq t) = (n/N) F_n(t) + (1 - n/N) F_r(t) \tag{5}$$

donde

$$\begin{aligned} F_n(t) &= n^{-1} \sum_{i \in S} I(Y_i \leq t) \\ F_r(t) &= (N - n)^{-1} \sum_{j \in P - S} I(Y_j \leq t). \end{aligned}$$

Como $F_n(t)$ es conocido, el problema va a ser estimar $F_r(t)$. Chambers y Dunstan (1986) construyen un estimador para este término a partir de los residuos del modelo de superpoblación, considerando un modelo de regresión lineal heterocedástico con la varianza del modelo conocida. Sin embargo, en la práctica son pocas las ocasiones en que esta varianza se conoce, por lo que trabajaremos con los residuos $r_i = Y_i - \hat{\alpha}_E - \hat{\beta}_E x_i$. Por lo tanto, el estimador de Chambers y Dunstan respecto del modelo (ξ) es

$$\begin{aligned}\hat{F}(t) &= N^{-1}\left\{nF_n(t) + (N-n)\hat{F}_r(t)\right\} \\ &= N^{-1}\left\{\sum_{i\in S} I\left(Y_i \leq t\right) + \sum_{j\in P-S} \hat{G}\left(\frac{t-\hat{\alpha}_E - \hat{\beta}_E x_j}{\hat{\sigma}_h(x_j)}\right)\right\},\end{aligned} \tag{6}$$

con $\hat{G}(u) = n^{-1}\sum_{i\in S} I(\frac{r_i}{\hat{\sigma}_h(x_i)} \leq u)$ la distribución empírica de los errores normalizados, con función de distribución G.

Trabajando con los estimadores de los parámetros del modelo de superpoblación $\hat{\alpha}_L$ y $\hat{\beta}_L$, los cuales son insesgados pero no de mínima varianza, el estimador de Chambers y Dunstan resultante es

$$\hat{F}_L(t) = N^{-1}\left\{\sum_{i\in S} I\left(Y_i \leq t\right) + \sum_{j\in P-S} \hat{G}_L\left(\frac{t-\hat{\alpha}_L - \hat{\beta}_L x_j}{\hat{\sigma}_h(x_j)}\right)\right\}, \tag{7}$$

con $\hat{G}_L(u) = n^{-1}\sum_{i\in S} I(\frac{\tilde{r}_i}{\hat{\sigma}_h(x_i)} \leq u)$ y $\tilde{r}_i = Y_i - \hat{\alpha}_L - \hat{\beta}_L x_i$.

Comportamiento asintótico

Notación

Indicamos por $a \wedge b$ el mínimo de a y b. Sea $g = G'$ la función de densidad del error normalizado y $Var\{Y_k\} = \sigma^2(x_k)$ la función varianza del error del modelo (ξ). Los puntos del diseño $\{x_k\}$ tienen función de densidad d con soporte compacto Γ y media y varianza asintóticas $\mu_x = \int x d(x)dx$ y $\tau_x^2 = \int (x-\mu_x)^2 d(x)dx$, respectivamente. Indicamos por W los pesos $\int \sigma^{-2}(x)d(x)dx$, definiendo la media ponderada y varianza ponderada asintóticas de los puntos del diseño por $\mu_{wx} = W^{-1}\int x\sigma^{-2}(x)d(x)dx$ y $\tau_{wx}^2 = W^{-1}\int (x-\mu_{wx})^2\sigma^{-2}(x)d(x)dx$, respectivamente.

Denotamos por:

$$\begin{aligned}
I_1 &= \int\int \left[G\left(\frac{t-\alpha-\beta u}{\sigma(u)} \wedge \frac{t-\alpha-\beta v}{\sigma(v)}\right) - G\left(\frac{t-\alpha-\beta u}{\sigma(u)}\right) G\left(\frac{t-\alpha-\beta v}{\sigma(v)}\right) d(u)d(v)dudv\right], \\
I_2 &= \int\int g\left(\frac{t-\alpha-\beta v}{\sigma(v)}\right)\left(\frac{1}{\sigma(u)} - \frac{1}{\sigma(v)}\right) d(u)d(v)dudv, \\
I_3 &= \int\int g\left(\frac{t-\alpha-\beta v}{\sigma(v)}\right)\left(\frac{u}{\sigma(u)} - \frac{v}{\sigma(v)}\right) d(u)d(v)dudv, \\
I_4 &= \int\int_{-\infty}^{\frac{t-\alpha-\beta u}{\sigma(u)}} z g(z) d(u)dudz, \\
I_5 &= \int \left[G\left(\frac{t-\alpha-\beta v}{\sigma(v)}\right) - G\left(\frac{t-\alpha-\beta v}{\sigma(v)}\right)^2\right] d(v)dv.
\end{aligned}$$

Resultados

Chambers y Dunstan (1986) probaron, bajo ciertas condiciones de regularidad, la normalidad asintótica del error de predicción, considerando como modelo de superpoblación un modelo de regresión lineal heterocedástico pasando por el origen y con varianza conocida.

Ahora, estudiaremos el comportamiento asintótico del error de predicción $\{\hat{F}(t) - F_N(t)\}$ del estimador de Chambers y Dunstan construido a partir de un modelo de regresión lineal heterocedástico con varianza desconocida (1) y estimada no paramétricamente, de acuerdo a lo indicado en el apartado anterior.

Teorema 1 *Bajo ciertas condiciones de regularidad, se verifica*

$$\sqrt{n}\left\{\hat{F}(t) - F_N(t)\right\} \to_d N(0, V), \tag{8}$$

donde

$$\begin{aligned} V = \{&(1-f)^2 \left[I_1 + I_2^2 W^{-1}\left(1 + \tau_{wx}^{-2}\mu_{wx}^2\right) + I_3^2 W^{-1}\tau_{wx}^{-2} - I_2 I_3 W^{-1}\tau_{wx}^{-2}\mu_{wx}\right. \\ &+ I_2 I_4 \int \left(1 - \tau_{wx}^{-2}\mu_{wx}(x - \mu_{wx})\right)\sigma^{-1}(x)d(x)dx + I_3 I_4 \tau_{wx}^{-2}\left.\int (x - \mu_{wx})\sigma^{-1}(x)d(x)dx\right] \\ &+ f(1-f)I_5\}^{1/2}. \end{aligned} \tag{9}$$

Respecto al comportamiento asintótico del error de predicción del estimador de Chambers-Dunstan, considerando los estimadores de mínimos cuadrados de los parámetros del modelo de superpoblación (ξ), $\{\hat{F}_L(t) - F_N(t)\}$, se tiene el siguiente resultado:

Teorema 2 *Bajo ciertas condiciones de regularidad, se verifica*

$$\sqrt{n}\left\{\hat{F}_L(t) - F_N(t)\right\} \to_d N(0, V_L), \tag{10}$$

donde

$$\begin{aligned} V_L = \Big\{&(1-f)^2 \left[I_1 + I_2^2 \int \left(1 - \tau_x^{-2}\mu_x(x - \mu_x)\right)^2 \sigma^2(x)d(x)dx\right. \\ &+ I_3^2 \int (\tau_x^2)^{-2}(x - \mu_x)^2\sigma^2(x)d(x)dx \\ &+ I_2 I_3 \int \left(1 - \tau_x^{-2}\mu_x(x - \mu_x)\right)\tau_x^{-2}(x - \mu_x)\sigma^2(x)d(x)dx \\ &+ I_2 I_4 \int \left(1 - \tau_x^{-2}\mu_x(x - \mu_x)\right)\sigma(x)d(x)dx + I_3 I_4 \tau_x^{-2}\left.\int (x - \mu_x)\sigma(x)d(x)dx\right] \\ &+ f(1-f)I_5\Big\}^{1/2}. \end{aligned} \tag{11}$$

3. ESTUDIO DE SIMULACIÓN: POBLACIÓN DE PLANTACIONES DE CAÑA DE AZÚCAR

Por similitud con las situaciones reales trabajaremos con una población finita particular. A continuación comparamos las distintas versiones del estimador de Chambers-Dunstan para estimar la función de distribución de una población finita que consiste en 338 plantaciones de caña de azúcar. Los datos proceden de una encuesta realizada en 1982 sobre la industria azucarera de Queensland. Se consideran tres variables de interés: Y_1, cosecha total de caña; Y_2, ingresos brutos de la caña de azúcar; Y_3, gasto total de la plantación. La variable auxiliar X es el área asignada para la plantación de caña. Vamos a tomar muestras de tamaño $n = 85$, obtenidas por muestreo aleatorio sin reemplazamiento. Este ejemplo práctico fue estudiado por Chambers y Dunstan (1986).

Presentamos la aproximación de Monte Carlo del error cuadrático medio del error de predicción en los cuartiles de la variable respuesta, para la distribución empírica $(F_n(t) - F_N(t))$, para el estimador de Chambers-Dunstan $\hat{F}(t)$, $(\hat{F}(t) - F_N(t))$ y $\hat{F}_L(t)$, $(\hat{F}_L(t) - F_N(t))$ estudiados en la sección anterior, y además, para el estimador de Chambers-Dunstan

	Y=COSECHA			Y=INGRESOS			Y=GASTO		
$F_N(t)$	0.249	0.500	0.749	0.246	0.497	0.746	0.246	0.497	0.746
MSE_n	16.21	21.47	15.40	15.13	20.37	14.78	15.83	22.34	16.15
MSE_{cd}	4.610	14.59	6.359	6.035	12.35	7.360	4.218	18.64	10.21
MSE_L	4.554	11.95	5.043	6.486	9.661	6.190	4.169	14.63	7.165
MSE	4.479	11.96	5.188	5.861	10.41	6.017	3.689	15.62	6.679

Tabla I. $MSE \times 10^{-4}$, *aproximado por Monte Carlo de* $(F_n(t) - F_N(t))$, $(\hat{F}_{cd}(t) - F_N(t))$, $(\hat{F}_L(t) - F_N(t))$ *y* $(\hat{F}(t) - F_N(t))$.

considerando como modelo de trabajo un modelo de regresión lineal homocedástico $(\hat{F}_{cd}(t) - F_N(t))$ (Chambers et al., 1992). Ahora la población finita sobre la que se realiza el estudio es fija, por lo que nuestro interés se centrará en el estudio del error de predicción para esta población finita particular: $E\{(\hat{F}(t) - F_N(t))^2/\mathbf{P}\}$.

Para la aproximación de Monte Carlo se obtuvieron $I = 1000$ muestras iniciales de la población de 338 plantaciones de caña. En este estudio no sólo se tiene en cuenta el modelo que genera la población sino también el diseño con el que se toma la muestra. Entonces, la aproximación de Monte Carlo del error cuadrático medio del error de predicción del estimador de Chambers-Dunstan en esta población finita es,

$$\begin{aligned} MSE_{cd}(t) &\approx \frac{1}{I}\sum_{i=1}^{I}\left(\hat{F}_{cd}^i(t) - F_N(t)\right)^2, \\ MSE(t) &\approx \frac{1}{I}\sum_{i=1}^{I}\left(\hat{F}^i(t) - F_N(t)\right)^2, \\ MSE_L(t) &\approx \frac{1}{I}\sum_{i=1}^{I}\left(\hat{F}_L^i(t) - F_N(t)\right)^2, \end{aligned} \tag{12}$$

si se toma $\hat{F}_{cd}(t)$, $\hat{F}(t)$ o $\hat{F}_L(t)$ como estimador de $F_N(t)$, respectivamente, en cada muestra $\mathbf{S}^i$ $(i = 1, ..., I)$.

La Tabla I muestra los resultados de la aproximación de Monte Carlo del error cuadrático medio del error de predicción. Nótese que el modelo de superpoblación que genera la población finita presenta una heterocedasticidad moderada, $\sigma(x) = \sqrt{x}$ (Chambers y Dunstan, 1986), lo que favorece la similitud de resultados entre los tres estimadores de Chambers-Dunstan.

4. AGRADECEMENTOS

Este trabajo ha sido financiado por el Ministerio de Educación y Cultura, mediante el proyecto BFM2002-03213 (70% FEDER).

5. REFERENCIAS

R.J. Carroll, Adapting for heteroscddaticity in linear models, *The Annals of Statistics*, **10**, 4, (1982), pp.1224-1233.

R.L. Chambers, A.H. Dorfman and P. Hall, Properties of estimators of the finite population distribution function, *Biometrika*, **79**, 3, (1992), pp.577-582.

R.L. Chambers and R. Dunstan, Estimating distribution functions from survey data, *Biometrika*, **73**, 3, (1986), pp.597-604.