

ESTIMADORES INDIRECTOS. UN ESTUDIO DE SIMULACION.

Pérez Villalta, R., Ortega Irizo, J., Arias Martín, C.

Dpto. Economía Aplicada I. Universidad de Sevilla.

1. Introducción.

Es conocido que los estimadores de razón y de regresión ofrecen una metodología para aprovechar la información disponible sobre la variable relacionada con la característica objeto de estudio. Sea Y la característica objeto de estudio, y sea X la variable relacionada con Y , y supongamos que se conoce el total poblacional de la variable X , es decir, t_x . En adelante, también vamos a suponer que la característica de interés es cuantitativa.

Entonces, si se considera una población de tamaño N , donde u_i $i=1,2,\dots,N$ denota la unidad i -ésima, y_i representa la medida de la variable Y mientras que x_i denota la medida de la variable X , ambas sobre la misma unidad u_i , y se utiliza un diseño aleatorio simple con reemplazamiento de tamaño n , el estimador insesgado de t_y es precisamente

$$\hat{t}_y = \frac{N}{n} \sum_{i=1}^n y_i e_i$$

donde $e_i \sim B(n, p_i)$ con $p_i = 1/N$, $i=1,2,\dots,N$, siendo su estimación, $N\bar{y}$, donde \bar{y} denota la media muestral de la variable Y .

El objetivo de los métodos indirectos de estimación es aprovechar la información que la variable X pueda aportar sobre la variable Y para conseguir estimaciones más precisas, es decir, con menor error cuadrático medio que las proporcionadas al utilizar directamente el diseño aleatorio simple con reemplazamiento.

Un planteamiento general del problema, puede ser el siguiente. Dado que se conoce el valor del total de la variable X , al considerar un determinado diseño muestral, la variable aleatoria

$$\hat{x}_x = t_x - \hat{t}_x$$

cuyos valores son los errores en las estimaciones del total de la variable X, se encuentra perfectamente especificada. Sin embargo, no ocurre lo mismo cuando se considera la variable Y, pues la variable aleatoria

$$\hat{x}_Y = t_Y - \hat{t}_Y$$

cuyos valores son precisamente los errores en las estimaciones del total de la variable Y, depende de t_Y , que es la característica que se trata de estimar. Los métodos indirectos de estimación tratan de mejorar el estimador \hat{t}_Y mediante la construcción de otro estimador a partir de éste. Un primer camino sería considerar el estimador:

$$\hat{t}_Y^* = \hat{t}_Y + \hat{x}_Y = t_Y$$

que evidentemente es inviable pues coincide con la característica que se trata de estimar. Sin embargo, si las variables X e Y están relacionadas, una hipótesis razonable es que también lo estén las variables aleatorias \hat{x}_X y \hat{x}_Y , y entonces, podemos definir, siguiendo a Azorín y Sánchez-Crespo (1986),

$$\hat{t}_Y^g = \hat{t}_Y + g\hat{x}_X = \hat{t}_Y + g(t_X - \hat{t}_X)$$

donde g se debe especificar de forma que el estimador \hat{t}_Y^g sea preferible al estimador \hat{t}_Y .

Evidentemente, si $g=0$, no se produce ninguna mejora en el estimador \hat{t}_Y^0 con respecto a \hat{t}_Y , pues ambos coinciden. El mejor de los supuestos posibles sería que

$$\hat{t}_Y^g = t_Y$$

y entonces, al ser

$$\hat{t}_Y^g = \hat{t}_Y + g(t_X - \hat{t}_X) = t_Y$$

se tiene que

$$g = \frac{t_Y - \hat{t}_Y}{t_X - \hat{t}_X} = \frac{\hat{x}_Y}{\hat{x}_X}$$

2. Una construcción de estimadores indirectos.

Dado que g no puede depender de la característica que se trata de estimar, la solución propuesta en el epígrafe anterior no es viable. Sin embargo, el problema puede estudiarse si realizamos alguna hipótesis adicional sobre el cociente de las variables aleatorias \hat{x}_Y y \hat{x}_X . De esta forma, si en primer lugar se supone que

$$\frac{\hat{x}_Y}{\hat{x}_X}$$

es aproximadamente constante, esto es, \hat{x}_Y es aproximadamente proporcional a \hat{x}_X , entonces podemos expresar

$$g = \frac{\hat{x}_Y}{\hat{x}_X} = b$$

y así:

$$\hat{t}_Y^b = \hat{t}_Y + b(t_X - \hat{t}_X)$$

resultando que \hat{t}_Y^b es el estimador de regresión del total de la variable Y.

Si en segundo lugar suponemos que, aproximadamente,

$$\hat{x}_Y = \hat{x}_X$$

entonces,

$$g = \frac{\hat{x}_Y}{\hat{x}_X} = 1$$

y así:

$$t_Y^d = t_Y + (t_X - \hat{t}_X)$$

resultando que t_Y^d es el estimador de diferencia del total de la variable Y.

Por otra parte, si en lugar de considerar los errores absolutos de las estimaciones del total de la variable Y y de la variable X, se consideran los errores relativizados, esto es,

$$\frac{\hat{x}_Y}{\hat{t}_Y} \text{ y } \frac{\hat{x}_X}{\hat{t}_X}$$

con lo que se podría resolver además el problema de las unidades de medida en las variables Y y X, si los hubiera, entonces, para el mejor de los casos posibles, esto es, $t_Y^g = t_Y$, al ser

$$g = \frac{\hat{x}_Y}{\hat{x}_X}$$

se tiene que

$$g \frac{\hat{t}_X}{\hat{t}_Y} = \frac{\frac{\hat{x}_Y}{\hat{t}_Y}}{\frac{\hat{x}_X}{\hat{t}_X}} = \frac{\frac{t_Y - \hat{t}_Y}{\hat{t}_Y}}{\frac{t_X - \hat{t}_X}{\hat{t}_X}}$$

solución que no es viable pues g depende nuevamente de la característica que trata de estimar. Ahora bien, si realizamos alguna hipótesis adicional sobre el cociente de las variables

aleatorias \hat{x}_Y / \hat{t}_Y y \hat{x}_X / \hat{t}_X , se podrá especificar el correspondiente estimador del total de la variable Y. Así, en primer lugar, bajo el supuesto de que

$$\frac{\frac{\hat{x}_Y}{\hat{t}_Y}}{\frac{\hat{x}_X}{\hat{t}_X}}$$

sea aproximadamente constante, esto es, \hat{x}_Y / \hat{t}_Y sea aproximadamente proporcional a \hat{x}_X / \hat{t}_X , entonces,

$$g \frac{\hat{t}_x}{\hat{t}_y} = \frac{\frac{\hat{x}_y}{\hat{t}_y}}{\frac{\hat{x}_x}{\hat{t}_x}} = 1$$

y por lo tanto, como

$$g = 1 \frac{\hat{t}_y}{\hat{t}_x}$$

se tiene que

$$\hat{t}_y^l = \hat{t}_y + 1 \frac{\hat{t}_y}{\hat{t}_x} (t_x - \hat{t}_x) = (1-1)\hat{t}_y + 1 \frac{\hat{t}_y}{\hat{t}_x} t_x$$

estimador que denominamos estimador de razón generalizado.

Si en segundo lugar, se considera que aproximadamente

$$\frac{\hat{x}_y}{\hat{t}_y} = \frac{\hat{x}_x}{\hat{t}_x}$$

al ser aproximadamente $\lambda=1$, se tiene que:

$$\hat{t}_y^R = \frac{\hat{t}_y}{\hat{t}_x} t_x$$

siendo \hat{t}_y^R el estimador de razón del total de la variable Y.

Así pues, y a **modo de resumen** se tiene que:

a) Si $\hat{x}_y / \hat{x}_x = b$,

$$\hat{t}_y^b = \hat{t}_y + b(t_x - \hat{t}_x)$$

b) Si $\hat{x}_y = \hat{x}_x$, entonces

$$\hat{t}_y^d = \hat{t}_y + (t_x - \hat{t}_x)$$

c) Si $(\hat{x}_y / \hat{t}_y) / (\hat{x}_x / \hat{t}_x) = \lambda$, se tiene que

$$\hat{t}_y^l = (1-1)\hat{t}_y + 1 \frac{\hat{t}_y}{\hat{t}_x} t_x$$

d) Si $(\hat{\mathbf{x}}_Y/\hat{\mathbf{t}}_Y) = (\hat{\mathbf{x}}_X/\hat{\mathbf{t}}_X)$, entonces

$$\hat{\mathbf{t}}_Y^R = \frac{\hat{\mathbf{t}}_Y}{\hat{\mathbf{t}}_X} \mathbf{t}_X$$

Si en lugar de estimar el total de la variable Y se pretende estimar su media, los resultados son análogos.

Es fácil comprobar que si se supone la relación funcional $Y=kX$, entonces se tiene que $(\hat{\mathbf{x}}_Y/\hat{\mathbf{t}}_Y) = (\hat{\mathbf{x}}_X/\hat{\mathbf{t}}_X)$, y por lo tanto,

$$\hat{\mathbf{t}}_Y^R = \frac{\hat{\mathbf{t}}_Y}{\hat{\mathbf{t}}_X} \mathbf{t}_X$$

3. Simulación.

Vamos a comprobar mediante simulación que el estimador de razón no es estable para cambios en el modelo de la forma $Y=kX^\lambda$, con λ próximo a la unidad. Sin embargo, el estimador de razón generalizado presenta un comportamiento mucho más adecuado en esta situación.

Para ello, se ha diseñado un programa en TurboC, versión 2.0, que, en primer lugar, genera valores de una variable aleatoria distribuida según un modelo $N(\mu, \sigma^2)$ de la siguiente forma. A partir del método congruencial que el programa incorpora para generar valores de una distribución Uniforme discreta, los convertimos posteriormente en valores de una distribución Uniforme de parámetros 0 y 1; generando posteriormente, de forma independiente, y según el mismo procedimiento, valores de otra distribución Uniforme de parámetros 0 y 1. A continuación, dado que mediante el empleo de dos variables U y V, independientes, distribuidas según un modelo $U(0,1)$, se tiene que

$$Z = \sqrt{-2\text{Ln}(U)} \cos(2\pi V)$$

se distribuye según un modelo $N(0,1)$, basta considerar $X=\mu+\sigma Z$ para obtener los valores de la variable aleatoria de interés, es decir, valores de una variable aleatoria distribuida según un modelo $N(\mu,\sigma^2)$.

A partir de estos valores normales x , se obtienen los valores $y=kx^\lambda$, donde se ha observado que los valores de k no tienen especial relevancia sobre los aspectos de los que trata este trabajo, por lo que únicamente se van a ofrecer resultados para $k=2$. Se ha estudiado el valor de λ , desde 0.80 a 1.20, con un paso de 0.001, aunque a título de ejemplo únicamente vamos a ofrecer los resultados con un paso de 0.02. Fijados los parámetros de la variable aleatoria X , se generan 1000 valores de dicha variable. Posteriormente, para cada valor de λ , y $k=2$, se obtienen mediante la relación funcional $Y=2X^\lambda$, los 1000 valores correspondientes de la variable Y , obteniendo de esta forma 1000 valores (x,y) , que constituyen la población objeto de estudio.

Se toman de esta población 1000 muestras de tamaño 100, obteniéndose para cada una de dichas muestras la estimación de λ , mediante las técnicas clásicas del ajuste de una función potencial, lo que obliga a que los valores de la población sean positivos, y las estimaciones a que dan lugar tanto el estimador de razón como el estimador de razón generalizado.

Posteriormente, se ha calculado la media de cada una de estas estimaciones,

$$\bar{t}_R = \frac{\sum_{i=1}^{1000} t_{iR}}{1000}$$

donde t_{iR} es la estimación del total empleando el estimador de razón para la i -ésima muestra, $i=1,\dots,1000$; así como una aproximación del error cuadrático medio,

$$ECM(\bar{t}_R) = \frac{\sum_{i=1}^{1000} (t_{iR} - t_Y)^2}{1000}$$

obteniéndose para el estimador de razón generalizado las mismas

características, \bar{t}_I y $ECM(\bar{t}_I)$.

Para la variable aleatoria X se han generado distintas distribuciones normales, de entre las que ofrecemos las de parámetros $\mu=100$ y $\sigma^2=10$, cuadro 1, y $\mu=100$ y $\sigma^2=500$, cuadro 2.

En ambos cuadros se observa que conforme λ se aproxima a la unidad, van disminuyendo los valores de error cuadrático medio de los dos estimadores considerados, aumentando conforme λ se aleja de dicho valor, aunque, en términos absolutos, la diferencia entre ambos es notable, poniendo de manifiesto la gran sensibilidad que muestra el estimador de razón ante pequeños cambios en el modelo. Podemos observar también, que la diferencia entre los valores medios de las estimaciones de los dos estimadores considerados es pequeña, lo que nos indica que la desviación respecto al verdadero valor del parámetro es del mismo orden en ambos estimadores.

Bibliografía.

Azorín, F. (1969). Curso de muestreo y aplicaciones. Aguilar.

Azorín, F. y Sánchez-Crespo, J.L. (1986). Métodos y aplicaciones del muestreo. Alianza Editorial.

Fernández, F.R. y Mayor, J.A. (1994). Muestreo en poblaciones finitas: curso básico. P.P.U.

Kennedy, W.J. y Ghentle, J.E. (1980). Statistical computing. Marcel Dekker.

Cuadro 1

λ	\bar{t}_R	\bar{t}_I	$ECM(\bar{t}_R)$	$ECM(\bar{t}_I)$
0.80	79984.608	79686.888	2833.497	0.940
0.82	87374.904	87376.608	2462.631	0.979
0.84	95808.264	95809.112	2228.163	0.811
0.86	105056.864	105056.112	2120.134	0.843
0.88	115195.528	115194.896	1858.801	0.779
0.90	126313.560	126312.192	1517.373	0.702

0.92	138505.536	138503.808	1181.912	0.602
0.94	151870.352	151871.680	888.853	0.409
0.96	166527.520	166528.048	435.452	0.262
0.98	182600.928	182601.872	131.429	0.071
1.00	200223.984	200223.984	0.004	0.004
1.02	219548.240	219549.968	189.815	0.126
1.04	240735.488	240736.400	954.060	0.526
1.06	263970.768	263970.208	2546.668	1.549
1.08	289452.448	289450.240	5515.863	3.148
1.10	317380.640	317385.696	10598.597	5.854
1.12	348015.968	348019.712	17232.014	10.743
1.14	381608.608	381605.888	29895.712	19.409
1.16	418449.376	418438.016	46596.040	33.499
1.18	458822.688	458824.288	64497.784	50.117
1.20	503123.584	503109.920	101234.192	81.303

Cuadro 2

λ	\bar{t}_R	\bar{t}_I	ECM(\bar{t}_R)	ECM(\bar{t}_I)
0.80	79525.944	79532.152	108714.800	2742.737
0.82	87252.352	87239.360	113793.576	2519.287
0.84	95701.056	95697.168	100632.104	2734.666
0.86	104977.384	104971.880	97641.440	2411.735
0.88	115144.344	115150.648	86196.712	2389.509
0.90	126323.704	126320.232	68328.664	2082.663
0.92	138581.136	138570.400	52898.828	1623.781
0.94	152025.344	152015.152	35060.600	990.727
0.96	166773.584	166766.560	20072.040	631.506
0.98	182952.640	182953.584	5514.003	201.601
1.00	200715.936	200715.936	0.002	0.002
1.02	220207.360	220205.120	7532.143	280.360
1.04	241591.312	241592.080	42379.232	1592.648
1.06	265068.656	265061.680	104431.416	3968.389
1.08	290824.832	290810.304	221932.416	9658.094
1.10	319101.664	319071.328	427813.280	18204.670
1.12	350093.312	350091.136	803114.432	31042.760
1.14	384141.792	384131.584	1248121.60	57561.972
1.16	421528.736	421466.816	1926114.81	86821.92
1.18	462347.168	462502.464	2741999.36	130562.480
1.20	507444.544	507477.760	4856664.06	216715.744