UNIVERSIDADE DE SÃO PAULO
CAMPUS DE SÃO CARLOS
INSTITUTO DE CIÊNCIAS MATEMÁTICAS DE SÃO CARLOS
I.C.M.S.C.

Ratio predictor under error-in variables structural superpopulation models

Josemar Rodrigues

nº 45

Notas do ICMSC - USP

ISSN 0103-2577

Ratio predictor under error-in variables structural superpopulation models

Josemar Rodrigues

nº 45

São Carlos - SP
1989

RATIO PREDICTOR UNDER ERROR-IN-VARIABLES STRUCTURAL SUPERPOPULATION MODELS

by

JOSEMAR RODRIGUES

Instituto de Ciências Matemáticas de São Carlos

Universidade de São Paulo

Caixa Postal 668, CEP-13.560, São Carlos, S.P, Brasil

SUMMARY

In this paper we introduce the conditional linear regression superpopulation model containing measurement error and investigate the effects of measurement error on the ratio predictor of a finite population mean. The ratio predictors corrected by atenuation are shown to be asymptotically normal, under the conditional superpopulation model. Also, it is presented the necessary and sufficient conditions for protecting the ratio predictor against measurement errors. Some results of regression analysis for finite population are derived.

Key words: Ratio predictor; conditional structural superpopulation model: measurement error.

1. INTRODUCTION

Let $P = \{1, \ldots, N\}$ denote a finite population of N units, and suppose there is a pair of real numbers (x_t, Y_t) with $x_t > 0$ associated with the t-th unit for $t = 1, \ldots, N$. It is assumed that the finite population $y_1, \ldots, y_N$ is a realization of the random vector Y' which is related to the konwn vector $x' = (x_1, \ldots, x_N)$ through a linear regression model:

$$Y = x\beta + \varepsilon, \tag{1}$$

where ε given x is$N(0, \Sigma_x)$, $\Sigma_x = \sigma^2 \text{diag}\{x_1, \ldots, x_N\}$ and β and σ^2 are unknown. The objective is to predict the population mean $\overline{Y} = \frac{T}{N}$, $T = 1'_N Y$, from a sample s of n units where 1_N denotes a column vector of ones. We use the notation r for the remainder of the population and partition Y and x in the following fashion:

$$Y' = (Y'_s, Y'_r) \qquad \text{and} \qquad x' = (x'_s, x'_r), \tag{2}.$$

It is well known that the ratio predictor (Royall,1971)

$$\overline{Y}_R = \overline{x}\frac{\overline{Y}_s}{\overline{x}_s}, \tag{3}$$

where $\overline{x}, \overline{x}_s$ and $\overline{Y}_s$ are population and sample means, respectively, has minimum mean square error among all predictors that are unbiased under the model (1) . Now, let us suppose that one is unable to observe x_t directly. Instead of observing x_t one observes the sum

$$X_t = x_t + u_t, \tag{4}$$

where u_t is a $(0, \sigma_{uu})$ random variable (see Fuller, 1987, for a good discussion of this model). In this paper we investigate the effect of measurement error on the ratio

$$\overline{Y}_{Re} = \overline{X}\frac{\overline{Y}_s}{\overline{X}_s}, \tag{5}$$

computed from the observed variables, in the model (1) and (4), under the assumption that

$$(\varepsilon', x', u') \sim N((0', \mu_x 1'_N, 0'); \Sigma), \tag{6}$$

where $\Sigma = \text{diag}(\Sigma_\varepsilon, \sigma_{xx} I, \sigma_{uu} I)$, I is the identity matrix and Σ is an unknown diagonal matrix. In this paper we refer to (1),(4) and (6) as the structural superpopulation model. Because (Y_t, X_t) is distributed as a bivariate normal the conditional distribution of Y given $X' = (X_1, X_2, \ldots, X_N)$, which we refer as the conditional structural superpopulation model, is given by:

$$Y = Z\beta + \varepsilon \quad \text{where} \quad E[\varepsilon \mid X] = 0, Var[\varepsilon \mid X] = H_\varepsilon, \tag{7},$$

H_ε is an unknown diagonal matrix, $Z_t = k_{xx} X_t + (1 - K_{xx})\mu_x$ and

$$K_{xx} = \frac{\sigma_{xx}}{\sigma_{xx} + \sigma_{uu}}. \tag{8}$$

The factor K_{xx} is known in the literature as the reliability ratio (see Fuller,1987, for more details). In this paper K_{xx} and μ_x are supposed to be known.

2-THE EFFECTS OF MEASUREMENT ERRORS ON THE RATIO PREDICTOR.

The following results give conditions for protecting the ratio predictor (5) against measurement errors.

LEMMA 1- Under the conditional structural superpopulation model defined by (7) we have that

$$E[(\overline{Y}_{Re} - \overline{Y}) \mid X] = 0 \quad \text{if and only if} \quad \overline{X}_s = \overline{X}, \tag{9}$$

that is, the sample is balanced on X (Pereira and Rodrigues,1983).

Proof: The result is straightforward from (7).

LEMMA 2- Under the conditional structural superpopulation model the ratio predictor (5) is optimal in Royall's sense (1970) if and only if

$$\begin{cases} \text{(i)- The sample s is balanced on X} \\ \text{(ii)- } Var\{Y \mid X\}1_N = \delta Z \text{ for some constant } \delta. \end{cases} \tag{10}$$

Proof: The result follows trivially from Tam's Theorem (1986). It is interesting to note that

$$h_{te} = Var\{\varepsilon_t \mid X\} = (1-\rho_t^2)Var\{Y_t\},$$

where ρ_t is the correlation coefficient so from (10) $Var\{Y_t\}$ must depend on Z_t. Now, we are going to study the effect of the measurement error on the ratio $\frac{\overline{Y}_s}{\overline{X}_s}$ under the conditional structural superpopulation model. It follows from (7) that

$$\begin{aligned} E\{\frac{\overline{Y}_s}{\overline{X}_s} \mid X\} &= \beta\frac{[K_{xx}\overline{X}_s + (1-K_{xx})\mu_x]}{\overline{X}_s} \\ &= \beta[K_{xx} + (1-K_{xx})\frac{\mu_x}{\overline{X}_s}] = \beta f_{xx}. \end{aligned} \tag{11}$$

We conclude from (11) that the ratio $\frac{\overline{Y}_s}{\overline{X}_s}$ is biased with respect to β. One way to examine the effect of measurement error in X is to say that the regression coefficient estimator has been biased by the measurement error or by the finite reliability ratio f_{xx}. An unbiased estimator of the regression coefficient β relating Y_t to the true value x_t of model (4) is given by

$$\hat{\beta} = f_{xx}^{-1}\frac{\overline{Y}_s}{\overline{X}_s}. \tag{12}$$

Let

$$\begin{aligned} &\overline{Y}_{R1} = f_{xx}^{-1}\overline{Y}_{Re} \quad \text{and} \\ &\overline{Y}_{R2} = \overline{Z}\hat{\beta} = K_{xx}f_{xx}^{-1}\overline{Y}_{Re} + (1-K_{xx})f_{xx}^{-1}\mu_x\frac{\overline{Y}_s}{\overline{X}_s}. \end{aligned} \tag{13}$$

be the ratio predictors of $\overline{Y}$ corrected by the factor f_{xx}. The next theorems give a version of Theorem 6.4 (Cochran, 1962) containing measurement error.

THEOREM-1 The necessary and sufficient condition for $\overline{Y}_{R2}$ to be optimal in Royall's sense (1970) under the conditional structural superpopulation model defined by (7) is

that

$$Var\{Y \mid X\} = \delta Z, \tag{14}$$

for some constant δ.

PROOF:

The proof is similar to the one in LEMMA 2.

THEOREM 2.

The necessary and sufficienty condition for $\overline{Y}_{R1}$ to be optimal in Royall's sense (1970) under the conditional structure model (7) is that

$$\begin{aligned} &(i) - K_x = 1, \quad or \quad \mu_x = \overline{X} \\ &(ii) - Var(Y \mid X) = \delta Z, \quad \text{for some constant } \delta. \end{aligned} \tag{15}$$

PROOF:

Similar to the proof of Theorem 1.

3-REGRESSION ANALYSIS FOR A FINITE POPULATION.

In this section, we assume the following orthogonal regression model derived from (7) : (with $H_\epsilon = \sigma^2 Q, \quad Q = diag(q_1, \ldots, q_N)$: known matrix)

$$\begin{aligned} Y &= ZB_N + \varepsilon \quad \text{where} \\ B_N &= \{\sum_{i=1}^{N} Z_i^2 q_i^{-1}\}^{-1} \sum_{i=1}^{N} Z_i q_i^{-1} Y_i, \\ E[\varepsilon \mid X] &= 0 \quad \text{and} \\ Var[\varepsilon \mid X] &= Q - Z(Z'Q^{-1}Z)^{-1}Z'. \end{aligned} \tag{16}$$

As in Hartley (1975), Fuller (1975) and Shah et al. (1977) our purpose is to make inferences about the linear function B_N.

THEOREM 3.

Under the orthogonal model (15), B_n is the optimal predictor for B_N, that is,
(i)-$E[B_n - B_N \mid X] = 0$
(ii)-$Var[B_n - B_N \mid X]$ is minimum for all unbiased predictor of B_N, (17)

PROOF

The statement of our theorem follows in a straighforward manner from Rao (1971) by verifying that

$$\{Z_s'[Var(\varepsilon_s \mid X)]^{-1} Z_s\}^{-1} Z_s'[Var(\varepsilon_s \mid X)]^{-1} Y_s = B_n. \tag{18}$$

It is interesting to observe, from Theorem 3, that $\hat{\beta}$ is optimal for $\frac{t_N Y}{t_N Z}$ if we take $Var[Y \mid X] = \delta Z$. The next result gives the limiting distribution of B_n.

THEOREM 4.

Let $\{z_n : n = 1, \ldots\}$ be a sequence of finite populations, where z_n is a random sample of size N_n, $N_n > N_{n-1}$,selected from an infinite population .Let a simple random nonreplacement sample of size n be selected from the t-th finite population ,$n = 1, \ldots$. Let $f_n = \frac{n}{N_n}$,$\{Z_t^2 q_t^{-1}\}$ i.i.d.r.v. and

$$\lim f_n = f, \qquad 0 < f < 1.$$

Then, under the conditional structural superpopulation model (7) (with $H_e = \delta Q$)

$$n^{\frac{1}{2}}[B_n - B_N] \Longrightarrow N\{0; \delta(1-f)E^{-1}(Z_t^2 q_t^{-1})\}. \tag{19}$$

PROOF:

We may write

$$\begin{aligned} B_n - \beta &= E^{-1}[Z_t^2 q_t^{-1}] \sum_{t=1}^{n} \frac{Z_t q_t^{-1} \varepsilon_t}{n} + o_p(n^{-1}), \\ B_N - \beta &= E^{-1}[Z_t^2 q_t^{-1}] \sum_{t=1}^{N} \frac{Z_t q_t^{-1} \varepsilon_t}{N} + o_p(n^{-1}), \\ \varepsilon_t &= Y_t - \beta Z_t. \end{aligned} \tag{20}$$

The result follows by the Central Limit Theorem .

COROLLARY 1.

Assuming $q_t = Z_t$, under the conditions of Theorem 4, we have that

$$n^{\frac{1}{2}}[\hat{\beta} - \frac{1'_N Y}{1'_N Z}] \Longrightarrow N\{0; \delta(1-f)\}. \tag{21}$$

PROOF

It trivially follows from Theorem 4.

COROLLARY 2.

Assuming $q_t = Z_t$, under the conditions of Theorem 4, we have that

$$n^{\frac{1}{2}}[\overline{Y}_{R2} - \overline{Y}] \Longrightarrow N\{0; \delta(1-f)\}. \tag{22}$$

PROOF:

We may write

$$\overline{Y}_{R2} - \overline{Y} = \hat{\beta} - \frac{\overline{Y}}{\overline{Z}} + o_p(n^{-1}), \tag{23}$$

so , our result follows from Theorem 4. It is considerable interest to introduce a consistent estimator for δ given by

$$(\frac{1}{n-2}) \sum_{t=1}^{n} \frac{\varepsilon_{c,t}^2}{Z_t}, \tag{24}$$

where $\varepsilon_{c,t} = Y_t - \hat{\beta} Z_t$.

If the measurement error variance σ_{uu} is known and k_x is unknown, a natural estimator for k_x, is $\hat{k}_x = 1 - \frac{\sigma_{uu}}{S_X^2}$ where $S_X^2 = \frac{\sum_{t=1}^{n}(X_t - \overline{X})^2}{n-1}$. Let :

$$\hat{f}_{xx} = \hat{k}_x + (1 - \hat{k}_x)\frac{\mu_x}{\overline{X}_s}, \tag{25}$$

In this case Corollary 1 is still true when replacing k_x by $\hat{k}_x$, f_{xx} by $\hat{f}_{xx}$ and taking $Z_t = \hat{k}_x X_t + (1 - \hat{k}_x)\mu_x$ and $\hat{\beta}_1 = \hat{f}_{xx}^{-1}\frac{\overline{Y}_s}{\overline{X}_s}$. In many pratical situations f_{xx} can be estimated by $\hat{f}_{x1}$ which is independent of the data $(Y_t, X_t), t \in s$. In the sequel an interesting result follows:

THEOREM 5

Let the structural superpopulation model holds and let $\hat{f}_{x1}$ be an estimator of f_{xx} and

$$\sqrt{n}(\hat{f}_{x1} - f_{xx}) \Longrightarrow N\{0; \sigma_{ff}\}.$$

Let $\hat{\beta}_{11} = \hat{f}_{x1}^{-1} \frac{\overline{Y}_s}{\overline{X}_s}$, then

$$n^{\frac{1}{2}}(\hat{\beta}_{11} - \beta) \Longrightarrow N\{0; f_{xx}^{-2}[\mu_x^{-2}\sigma_r + \beta^2\sigma_{ff}]\}, \tag{26}$$

where

$$\sigma_r = \lim_{n\to\infty} \frac{\sum_{t=1}^{n}[Var(Y_t - \hat{f}_{x1}\beta X_t)]}{n}$$

PROOF:

The result follows by noting that

$$\sqrt{n}(\hat{\beta}_{11} - \beta) = \hat{f}_{x1}^{-1}\sqrt{n}\frac{\sum_{t=1}^{n} r_t}{\sum_{t=1}^{n} X_t} + \beta\hat{f}_{x1}^{-1}\sqrt{n}(\hat{f}_{x1} - f_{xx}) \tag{27}$$

where $r_t = Y_t - \hat{f}_{x1}\beta X_t$.

Using data from Table 6.1 (Cochran, p.156, 1962), our Table 1. shows very clearly how error in measurement affects the ratio estimator $\overline{Y}_{R_r}$ and how robust is $\overline{Y}_{R1}$. The Table 1. shows that as k_x decreases we have serious effect on $\overline{Y}_{R_r}$.Fortunately , we can correct our ratio estimator by utilizing the reliability factor to obtain robust predictor for the population total.

TABELA 1.: Effect of Measurement Error ($\beta = 1, \delta = 1$).

k_x	f_{xx}	$\overline{Y}_{R1}$	$\overline{Y}_{R1} - 29.34$	MSE:$\overline{Y}_{Re}$	MSE:$\overline{Y}_{R1}$
1	1	28.39	0.95	2.10	2.10
0.99	0.99	28.61	0.73	2.92	2.13
0.98	0.98	28.84	0.50	5.39	2.15
0.97	0.97	29.07	0.27	9.51	2.17
0.96	0.96	29.30	0.04	15.29	2.19
0.95	0.96	29.54	-0.119	22.72	2.21
0.94	0.95	29.78	-0.43	31.81	2.24
0.93	0.94	30.0	-0.68	42.55	2.26
0.92	0.93	30.28	-0.93	54.94	2.28
0.91	0.92	30.53	-1.44	68.99	2.31
0.9	0.92	30.7	-1.41	84.69	2.33
0.8	0.84	33.62	-4.27	332.72	2.60

REFERENCES

Cochran,G.C.(1962). Sampling Techniques, J.Wiley.

Fuller,A.W. (1975). Regression analysis for sample survey, Sankhya, Series C,V.37.Part 3,pp.117-132.

Fuller,A.W. (1987). Measurement error models. J.Wiley.

Hartley, H.O. and Sielken,R.L. (1975). A superpopulation viewpoint for finite population sampling. Biometrics,31,411-422 .

Pereira,C.A.B. and Rodrigues,J. (1983). Robust linear prediction in finite populations,International Statistical Review,51,293-300.

Rao,C.R.(1973) Linear statistical inference and its applications, 2nd. ed.New York, Wiley.

Royall,R.M. (1970). On finite population sampling theory under certain linear regressions models. Biometrika,337-387.

Shah,B.V.,Holt,M.M. and Folsom, R.E.(1977). Inference about regression models for sample survey data. Bull. Internat. Statist. Inst. 47(3),43-57.

Tam,S.M.(1986). Characterization of best model-based predictors in survey sampling. Biometrika, 73,1,pp. 232-5.