Nonparametric Tests For Bias In Estimates And Forecasts

Charles, Coleman

US Census Bureau

1999

American Statistical Association

1999
Proceedings

of the

Business and Economic Statistics Section

Papers presented at the
Annual Meeting of the American Statistical Association,
Baltimore, Maryland, August 8–12, 1999,
under the sponsorship of the
Business and Economic Statistics Section

American Statistical Association ■ 1429 Duke Street, Alexandria, VA 22314

NONPARAMETRIC TESTS FOR BIAS IN ESTIMATES AND FORECASTS

Charles D. Coleman, U.S. Bureau of the Census
Washington, DC 20233-8800 (ccoleman@census.gov)

Key Words: MALPE, non-i.i.d., significance tests, nonidentical distributions, %POS.

Abstract:

Nonparametric tests can be made for bias in estimate and forecast errors without assuming identical and independent distributions. Tests are created for bias in the median and the mean. The test for median bias is a form of the familiar Sign Test for the median. For mean bias, an asymptotically normal test statistic is derived from the mean algebraic percentage error. These statistics are then applied to cross-sectional and time series contexts.

1 Introduction

The proportion of positive (or negative) errors (%POS) and the mean algebraic percentage error (MALPE) are common metrics for measuring bias in estimates and forecasts. Commonly, a set of estimates or forecasts is made, then the "truth" is revealed and the original estimates or forecasts are evaluated. These estimates and forecasts come from random distributions of unknown forms. The variables that can be estimated or forecast include population, employment and income, by geographic area or time, to give but a few examples. The errors are simply the differences between the estimates or forecasts and the true values. A set of estimates or forecasts is mean unbiased if the expected (i.e., average) value of each error is zero. Likewise, they are median unbiased if the median of each error's distribution is zero. Since the errors cannot usually be assumed to be independent and identically distributed or even normally distributed, parametric (i.e., using distributional assumptions) tests for the equality of their means to zero cannot be used.

This paper begins by assuming only independent errors. This assumption is literally untrue when estimates are constrained to sum up to a control total. However, in the absence of a suitable theory for dependence, this assumption must be made. In any case, the effect of the violation of independence may be insubstantial. For %POS, the median of the each error's distribution is assumed equal to zero under the null hypothesis of no median bias. This makes the sign of each error Bernoulli and their sum Binomial. Applying Liapounov's Central Limit Theorem creates an asymptotically normal test statistic for mean bias using MALPE. If the underlying distributions are symmetric, then their means and medians coincide, so a test for one is also a test for the other.

One should interpret the results of these statistical tests with caution. It is assumed that the "true" values are indeed the truth. However, it is quite possible that they themselves are estimates of the truth. The "true" values may thus themselves be biased. To provide a concrete example, many believe that every U.S. decennial Census since 1790 has had some undercount. A finding of positive bias in population estimates for a Census year may indicate that the estimates are correctly accounting for the persons missed by the decennial Census.[1] Nonetheless, we will be referring to true values in the discussion below.

Section 2 of this paper formally defines median bias and constructs tests using %POS. Section 3 does the same for mean bias and MALPE. Section 4 applies the bias tests to a cross-sectional example, a comparison of state population forecasts. The example is for purposes of illustration only: it is not meant to make definitive statistical statements about the underlying data. Section 5 discusses applications to time series. Section 6 concludes this paper.

2 Proportion of Positive Errors (%POS) and Testing for Median Bias

The median of a distribution is the point which divides the values of a random variable into two sets

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

[1] I would like to thank Prithwis Das Gupta for pointing this out.

of equal probability. That is, a random observation drawn from this distribution has equal probabilities of being less than the median or of being greater than the median. We can test the null hypothesis that the distributions generating the estimates x_i are generated by distributions $F_i(x_i)$ with medians equal to the true values t_i. Another way of stating this is $\mathbf{M}(x_i) = t_i$ for all observations i, where $\mathbf{M}$ is the median operator. The error of each observation is defined as $e_i = x_i - t_i$. A set of errors is said to be median unbiased if $\mathbf{M}(e_i) = 0$ for all i. Assuming the independence of the e_i implies that the sign of each e_i (sgn e_i) has a Bernoulli distribution with mean $p = .5$ under the null hypothesis. That is, the probability that any e_i is positive is the same as the probability that it is negative, and this probability is $.5 = 1/2$. This can be used to construct a version of the familiar Sign Test for the median.[2] The proportion of n positive (or negative) errors (i.e., the sum of the n Bernoulli variables divided by n) is thus Binomial with parameters $(.5, n)$ divided by n and the probability that $\%\text{POS} = y/n$ is $\binom{n}{y}\frac{1}{2^n}$. This distribution can be used to construct exact critical values r_α such that, under the null assumption of no bias in the median, %POS is less than r_α with probability α.[3] When n grows large, the Binomial distribution can be approximated by the normal distribution with mean .5 and standard deviation $1/(2\sqrt{n})$. As a general rule, the normal approximation is used when $n \geq 20$.

3 Mean Algebraic Percentage Error and Testing for Mean Bias

A set of estimates is unbiased in the mean if the expected (or mean) values of the errors all equal zero. Equivalently, $\mathbf{E}(e_i) = 0$ for all observations i, where $\mathbf{E}$ is the expected value operator. This forms our null hypothesis. The mean algebraic percentage error (MALPE) can be used as the basis of a test for mean bias since the distribution of a function of MALPE is shown to be asymptotically normal.

MALPE is defined as:

$$\text{MALPE} = n^{-1}\sum_{i=1}^{n}\frac{x_i - t_i}{t_i} = n^{-1}\sum_{i=1}^{n}\frac{e_i}{t_i}$$

Let the relative (or algebraic percentage) error be defined by $\tilde{e}_i = e_i/t_i$. Since $\mathbf{E}(e_i) = 0$ for all i under the null hypothesis, $\mathbf{E}(\tilde{e}_i) = 0$ for all i. If we further assume that (White, 1984:112):

Assumption 1: $\mathbf{E}|\tilde{e}_i|^{2+d} < D < \infty$ for all i and some d and $D > 0$.

Then, under the null hypothesis, $z = (\sqrt{n}\bar{\sigma}_n)^{-1}\sum_{i=1}^{n}\tilde{e}_i$ converges asymptotically to a standard normal distribution by the Liapounov Central Limit Theorem, where $\bar{\sigma}_n = n^{-1}\sum_{i=1}^{n}\sigma_i$ is the average standard deviation of the $\tilde{e}_i$.[4] Note that Assumption 1 is satisfied whenever the $\tilde{e}_i$ have finite, bounded supports.[5] To apply this result, we need an asymptotically consistent estimate of $\bar{\sigma}_n$. Assumption 2 provides a necessary condition for s_n, the sample standard deviation, to be asymptotically consistent. To wit, the errors must be asymptotically homoscedastic:

Assumption 2: $\lim_{n\to\infty}\bar{\sigma}_n = \bar{\sigma}$.

It is trivial to show that $\mathbf{E}s_n^2 = n^{-1}\sum_{i=1}^{n}\sigma_i^2$. However, $\mathbf{E}s_n \geq \bar{\sigma}_n$ for all finite n, with equality holding iff $\sigma_i \equiv \bar{\sigma}$ (Mitrinović and Vasić, 1970:85).[6] When Assumption 2 holds, $\lim_{n\to\infty}s_n^2 = (\bar{\sigma})^2$. Therefore, s_n converges to $\bar{\sigma}$. We can compute s_n by

$$\begin{aligned} s_n &= \sqrt{(n-1)^{-1}\sum_{i=1}^{n}(\tilde{e}_i - \text{MALPE})^2} \\ &= \sqrt{(n-1)^{-1}\sum_{i=1}^{n}(\tilde{e}_i^2 - \text{MALPE}^2)} \\ &= \sqrt{(n-1)^{-1}(\text{RMSPE}^2 - \text{MALPE}^2)} \end{aligned}$$

where RMSPE is the root mean squared percentage error $\left(n^{-1}\sum_{i=1}^{n}\tilde{e}_i^2\right)^{1/2}$, another commonly reported statistic. This finding of the asymptotic normality of MALPE explains the empirical findings of normality by Smith and Sincich (1988).

It should be noted that this argument works the same way using the average error, $\bar{e} = n^{-1}\sum_{i=1}^{n}e_i$, and the root mean squared error, $\left(n^{-1}\sum_{i=1}^{n}e_i^2\right)^{1/2}$, provided that analogous versions of Assumptions 1 and 2 are made, most notably that the e_i are asymptotically homoscedastic. However, there is generally little reason to believe this to be true of cross-sectional estimates. Other contexts, such as simulation, may satisfy this assumption.

[2] Conover, 1999:157–164. One should note that the other assumptions besides independence are satisfied: ordinal data and internal consistency.

[3] These values can be looked up from tables of the Binomial distribution.

[4] Note that if the $\tilde{e}_i$ are normally distributed, then MALPE, being an average of normal distributions, is always normally distributed.

[5] To see this, choose some number $N > \max_i|\tilde{e}_i|$. Then, $|\tilde{e}_i|^{2+d} < N^{2+d}$ for all i and for all $d > 0$. We can thus choose $D = N^{2+d}$ in Assumption 1.

[6] It should be noted that this bias in finite samples biases z towards zero, thereby favoring the null hypothesis. This bias increases in the absence of Assumption 2.

4 An Application to Cross-Sectional Data

Smith and Sincich (1992) have published average (i.e., pooled) %POS, MALPE and RMSPE for projections of the populations of all 50 states for four projection horizon lengths.[7] This Section applies the methods of Sections 2 and 3 to their data to determine if significant bias exists. It should be noted that independence is not necessarily satisfied in all cases. However, the purpose of this Section is illustration, not statistical proof. All results should be interpreted with this in mind.

4.1 %POS

Table 1 summarizes %POS for data pooled at various horizons, as reported by Smith and Sincich (1992), along with their significance levels. n is also included, being computed from data in Smith and Sincich (1992), Exhibit 1.[8] The significance levels are obtained by the normal approximation.

Table 1 shows 10 and 1 cases significant at 1% and 5%, respectively. In addition, 2 cases are significant at 10% and possibly at 5%. Of the mathematical techniques, EXPO and ARIMA show highly significant bias at all horizons.[9] All of the nonmathematical techniques (CB, NPA, OBERS) show significant bias at at least one horizon. CB, NPA and OBERS are significant at 3, 1 and 2 of their 4, 3 and 3 horizons, respectively.

4.2 MALPE

Table 2 presents MALPE and RMSPE, respectively, from Smith and Sincich (1992). In addition, it presents values of $z = \sqrt{n-1}\text{MALPE}/\sqrt{\text{RMSPE}^2 - \text{MALPE}^2}$,[10] which are asymptotically standard normal under the null hypothesis, and indicates significant values using two-tailed tests.

Table 2 shows 1 and 9 cases which are significant at the 1% and 5% levels or less, respectively. One additional case may be significant at 10% (LINE at 20 years). The ambiguity is due to rounding error. Thus, there is significant evidence of mean bias in these samples. Of the mathematical techniques, only SHIFT and SHARE show no significant bias at any horizon. The absolute value of LINE's z increases in the horizon, possibly achieving significance at 20 years. EXPO's z is practically constant for all horizons. ARIMA's z is about the same for the 10, 15 and 20 year horizons. The nonmathematical techniques all show bias. OBERS is biased at every horizon. CB is biased at three of four horizons. NPA shows bias at only one time horizon, at 1%, which may be spurious.[11]

5 Comparison of %POS and MALPE

MALPE shows greater evidence of bias than %POS. In all cases in which %POS reports bias, MALPE reports bias at the 1% level, while %POS reports higher significance levels in several cases. MALPE additionally reports bias in OBERS at the 15-year horizon and possible bias in LINE at the 20-year horizon. While the evidence of bias is generally overlapping, it is curious that evidence of median bias is weaker than for mean bias. A priori, one would think that the data generating processes are asymmetric, with their means different from their medians. However, the median is only sensitive to the direction of errors, unlike MALPE which sensitive to both their direction and magnitude, reducing the power of tests based on it, compared to tests based on the mean.[12] CB further indicates that sample size is of little relevance to both tests: both the 10 and 15 year horizons have the same n (592), but only the former is significant.

6 Time Series Applications

Both %POS and MALPE can be used to detect median or mean bias in time series, respectively. These tests can be made with an even weaker null hypothesis than independence: either jointly uncorrelated in median errors in the case of %POS or a mix-

[7] In the case of ARIMA, only 48 states are used: Alaska and Hawaii are excluded.

[8] Each technique generates 50 observations for each launch year, except for launch year 1955 and ARIMA, which both generate 48 observations. Technique CB, except for launch year 1980, consists of multiple methods, which are pooled.

[9] See Smith and Sincich (1992) for the technique definitions.

[10] This calculation can be seen by noting that $z = \left(\sqrt{n}\bar{\sigma}_n\right)^{-1}\sum_{i=1}^n \tilde{e}_i$ is asymptotically standard normal under the null hypothesis and substituting: $\left(\sqrt{n}\bar{\sigma}_n\right)^{-1}\sum_{i=1}^n \tilde{e}_i = \sqrt{n}\text{MALPE}/\bar{\sigma}_n \approx \sqrt{n}\text{MALPE}/s_n = \sqrt{n-1}\text{MALPE}/\sqrt{\text{RMSPE}^2 - \text{MALPE}^2}$.

[11] Note that three of the biased techniques (OBERS at the 10 and 20 year horizons, ARIMA and NPA) are negatively biased. If one believes that the Census has a sizable undercount, the evidence of bias for these techniques is further strengthened.

[12] Remember, we have found the asymptotic distribution of z by assuming only a moment condition, so the usual caveats about requiring parametric assumptions do not apply.

Table 1: %POS (percent)

Technique	Length of Projection Period (years)			
	5	10	15	20
LINE	51.3 (248)	46.7 (248)	47.5 (200)	44.7 (150)
EXPO	59.0[c] (248)	60.4[c] (248)	61.5[c] (200)	60.7[c] (150)
ARIMA	40.6[c] (240)	40.0[c] (240)	36.0[c] (192)	34.7[c] (144)
SHIFT	54.0 (248)	51.6 (248)	51.5 (200)	46.7 (150)
SHARE	54.0 (248)	51.6 (248)	54.0 (200)	49.3 (150)
CB	44.4[c] (642)	46.0[a] (592)	50.3 (592)	55.7[a] (300)
NPA	33.0[b] (150)	46.0 (100)	49.0 (100)	—
OBERS	64.0[c] (150)	34.0[c] (100)	43.0 (100)	—

[a]Significant at 10%, possibly significant at 5%. Rounding error prevents exact determination.
[b]Significant at 5%.
[c]Significant at 1%.

Note: %POS data are from Smith and Sincich (1992), Exhibit 2. n is calculated from Smith and Sincich (1992), Exhibit 1 and reported in parentheses.

ing condition in the case of MALPE.[13,14] In effect, these tests are for the joint null hypothesis of no bias and either no median correlations or a mixing condition.[15] %POS can be used for sequential testing of a data generation process. A run of five consecutive errors of the same sign has probability $1/2^5 = 1/32 \approx .031$ under the null hypothesis. This is significant at the 5% level. If this occurs at the beginning of a data generation process, it provides evidence that the process is median biased. More complicated patterns require the use of the Binomial distribution for testing.

MALPE is a more difficult case. Its distribution has only been obtained asymptotically, rendering it of little use in short time series. Finding its small sample properties by simulation is difficult, since the data generation processes are generally poorly understood probabilistically.

[13]One should note that, while times series themselves typically contain dependence (e.g., tomorrow's population depends on today's population), the errors in estimating these series need not be dependent.

[14]Consider a sequence of errors $(e_1, e_2, \ldots e_t)$. These are jointly uncorrelated in median if $\mathbf{M}(e_{t+1})$ does not depend on the preceding error sequence for all t. To give a concrete example, assume that the e_t come from symmetric distributions, the variance of each distribution depending on the preceding errors, but with identical medians (and means). Then, this sequence is dependent but jointly uncorrelated in median.

The errors obey a mixing condition if they are asymptotically independent. (White, 1984:44–46) The Central Limit Theorems for these processes require Assumption 2. (White, 1984:124)

[15]These tests should be used in addition to the standard tests for dependence, such as those for serial correlation and trending. See, for example, Krishnaiah and Sen (1984).

Table 2: MALPE, RMSPE and z

Technique	Statistic	Length of Projection Period (years)			
		5	10	15	20
LINE	MALPE	0.1	−0.5	−1.1	−1.9
	RMSPE	5.1	8.2	10.8	14.6
	z	0.31	−0.96	−1.44	-1.64^a
EXPO	MALPE	1.2	2.4	4.3	−6.0
	RMSPE	6.3	11.7	20.2	33.0
	z	3.05^c	3.29^c	3.07^c	2.97^c
ARIMA	MALPE	−1.1	−2.8	−4.4	−6.0
	RMSPE	4.6	8.2	11.7	14.8
	z	-3.81^c	-5.62^c	-5.61^c	-5.30^c
SHIFT	MALPE	0.4	0.2	−0.2	−0.8
	RMSPE	5.5	9.3	13.2	18.7
	z	1.15	0.34	0.25	0.32
SHARE	MALPE	0.4	0.2	0.2	0.4
	RMSPE	5.2	8.4	11.3	15.2
	z	1.21	0.37	0.25	0.32
CB	MALPE	−0.7	−1.1	−0.4	2.4
	RMSPE	5.0	8.2	10.7	15.1
	z	-3.58^c	-3.29^c	−0.91	2.78^c
NPA	MALPE	−2.4	−0.9	−0.6	—
	RMSPE	5.3	8.5	10.3	
	z	-6.20^c	−1.06	−0.58	
OBERS	MALPE	1.7	−3.6	−2.6	—
	RMSPE	5.8	8.8	11.6	
	z	3.74^c	-4.46^c	-2.29^b	

[a]Possibly significant at 10%.
[b]Significant at 5%.
[c]Significant at 1%.

Note: MALPE and RMSPE data are from Smith and Sincich (1992), Exhibit 2. z is calculated with the n reported in Table 1.

7 Conclusion

This paper has created nonparametric tests for bias in estimates and forecasts without assuming identically and independently distributed random variables. It has developed a test for median bias that can be used on both cross-sectional and time series data. Alternatively, MALPE is the basis of an asymptotically normal test for mean bias, with unclear small sample properties. The assumption of independence is very strong and, often, unrealistic. Evidence of significant bias, under the assumption of independence, in published datasets has been found using two different tests.[16] The empirical finding that MALPE is normally distributed now has a theoretical basis. MALPE's small sample behavior using nonnormal distributions is open to research,

[16]It should be noted, again, that independence has not been established for any of the data sets. It is not clear whether there is any substantial effect on the tests performed herein as a result.

most likely, by simulation studies. The behaviors of both tests under different forms of dependence is also open to research.

8 References

Conover, W.J. (1999), *Practical Nonparametric Statistics*, third edition, John Wiley & Sons, New York.

Krishnaiah, P.R. and P.K. Sen, eds. (1984), *Nonparametric Methods: Handbook of Statistics*, Volume 4, Elsevier, New York.

Mitrinović, D.S. and P.M. Vasić (1970), *Analytic Inequalities*, Springer-Verlag, New York.

Smith, Stanley K. and Terry Sincich (1992), "Evaluating the Forecast Accuracy and Bias of Alternative Population Projections for States," *International Journal of Forecasting* **8**, 495–508.

Smith, Stanley K. and Terry Sincich (1988), "Stability over Time in the Distribution of Population Forecast Errors," *Demography* **25**, 464–474.

White, Halbert (1984) *Asymptotic Theory for Econometricians*, Academic Press, Orlando, Florida.

9 Acknowledgements

I would like to thank Muhammad Habib and Stan Smith for helpful conversations and Prithwis Das Gupta, Raymond J. Soukop, Jenny Thompson and William R. Bell for peer review.