Rounding errors and index numbers

Thomas A. Gittings

Each month, different agencies of the government release the most recent percentage changes in a variety of indexes. For example, the Consumer Price Index (CPI) and the Producer Price Index (PPI), released by the Bureau of Labor Statistics of the Department of Labor, are two important price indexes. The Industrial Production Index reported by the Federal Reserve Board is an important quantity index.

When the most recent percentage change in these index numbers is released, it is given major coverage by wire services and news agencies, both here and abroad. Unexpected changes can have immediate effects on various markets and can be important factors influencing policy making processes. The pattern and timing of these changes need to be accurately preserved in historical data bases so that information is not lost or distorted.

Rounding of historical index numbers can distort the pattern of monthly growth rates, especially when indexes are revised or "rebased." This can create serious econometric problems when working with monthly index number series such as the CPI. These problems include biased estimates and spurious regressions and are especially severe for periods when the growth rates have a low mean and/or variance and the index numbers have been subjected to large base adjustments.[1]

The purpose of this article is to show how index numbers are being systematically distorted by significant rounding errors in the revised or rebased data. After a brief discussion of why index numbers are especially vulnerable to rounding errors, the CPI for urban wage earners and clerical workers (CPI-W) is used to illustrate how large these errors can be and how they now distort the pattern of monthly changes in the 1960s. Next, a Monte Carlo study is presented to show some of the econometric problems associated with rounding errors of this magnitude. Finally, I recommend steps to better understand and/or alleviate this problem.

Why rounding poses a problem

The rounding-error problem of index numbers is not a theoretical issue, rather, it is an issue of how index numbers are handled in practice. Consider the case of a typical price index: the Consumer Price Index for all urban workers (CPI-W). In order to calculate the CPI-W, each month the Department of Commerce samples a broad range of prices for items that a typical urban consumer would purchase. These items include apples and oranges, cars and gasoline, medical care services and baseballs, and so on.

These actual prices are then weighted by their relative importance as determined periodically by detailed surveys of consumption expenditure patterns. By construction the weights are adjusted so that the total index averages 100.0 for some period of time. Each month the new index numbers are calculated

Thomas A. Gittings is a senior economist at the Federal Reserve Bank of Chicago. The author thanks Kenneth Kuttner and Steve Strongin for their advice and help.

and released to the public after being rounded to the nearest tenth of a point. This first type of rounding error is noticeable in the data when monthly changes are calculated, but is relatively unimportant in practice.

The rounding-error problem that is unique to index numbers begins when the index is adjusted or "rebased" so that it averages 100.0 for a new period of time. Because most prices have increased substantially over the past thirty years, the result of rebasing is that prices from earlier years are scaled down so as to provide a continuous series.

The new historical numbers are then rounded to the nearest tenth of a point before being released. Herein lies the source of the major rounding-error problem with index numbers. To illustrate the magnitude of this second type of rounding error, we can examine how it has affected the monthly changes for the CPI-W.

TABLE 1

CPI-W: levels, changes, and percentage changes

		1957-59=100.0			1982-84 = 100.0		
Year	Month	Level	Change	Percent change	Level	Change	Percent change
60	8	103.2			29.8		
60	9	103.3	0.1	0.0969	29.8	0.0	0.0000
60	10	103.7	0.4	0.3872	29.9	0.1	0.3356
60	11	103.8	0.1	0.0964	30.0	0.1	0.3344
60	12	103.9	0.1	0.0963	30.0	0.0	0.0000

Now you see it, now you don't

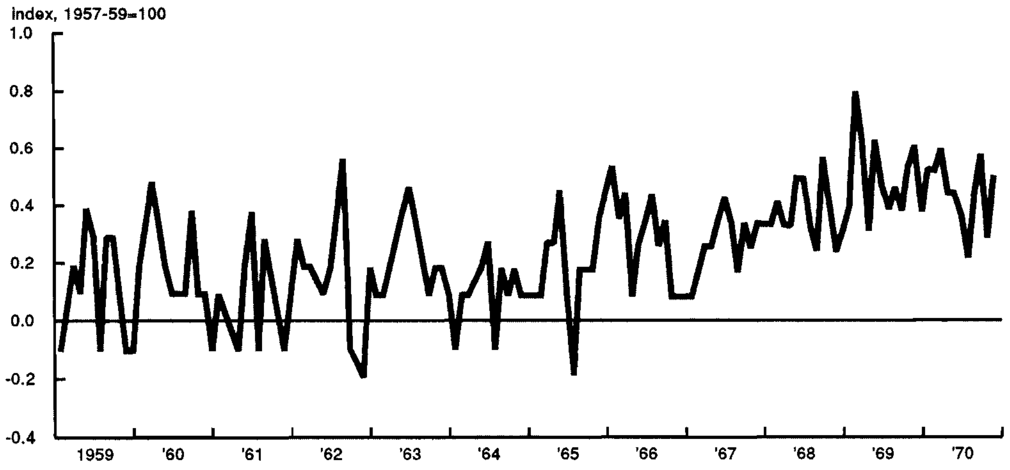
During the 1960s the CPI-W was benchmarked to a base period of 1957-59. Between January 1959 and December 1970, the CPI-W increased from 99.7 to 138.5. Figure 1 plots the monthly percentage changes as they were originally released. This Figure shows that the monthly percentage changes were quite volatile and that there was a pronounced increase in the average rate of inflation beginning in 1968.

The most recent release of the CPI-W series has been benchmarked to a base period of 1982-84. Because of the high rates of inflation during the 1970s, the adjustment or scaling factor for the 1960s is about 0.3. Table 1 lists the original CPI-W numbers and the current rebased numbers for the last five months of 1960. For each of these series, the month to month changes are tabulated and the percent changes are calculated to four decimal places.

Figure 2 plots the monthly percentage changes that are calculated from the currently released numbers for the period 1959-70. The pattern of monthly percentage changes is now dominated by rounding errors. The timing and magnitudes of monthly percentage changes

FIGURE 1

CPI-W monthly growth rates



have been significantly distorted, although it still is possible to discern the increase in inflation starting in 1968 and to calculate accurately average growth rates for each year.

Another way to see the effects of rounding is to plot the scatter diagram of the monthly percentage changes for 1959 through 1970. Figure 3 is this scatter diagram where the original growth rates are plotted along the horizontal axis and the revised growth rates are plotted along the vertical axis.

In the absence of rounding errors, there would be a perfect correlation between these two series, and the scatter diagram would be a straight line from the lower left corner to the upper right corner. The pronounced clustering is due to the rounding errors associated with the rebasing, and the "rays" pointing towards the origin are due to the rounding of the original data before the growth rates are calculated.

Table 2 tabulates the monthly changes for the two bases. With the low inflation of the early 1960s and the lower base for the revised numbers, it not surprising that almost half (62/144) of the monthly changes in the CPI-W for 1959 through 1970 are now reported as unchanged or zero.

Monte Carlo simulations

In order to identify what characteristics of the CPI-W data for the 1960s make this series so sensitive to rounding errors, a set of Monte Carlo simulations was run. In a Monte Carlo study, one assumes an underlying model, a set of parameter values, and a distribution for random errors. A number of samples are then generated using different random numbers, and the parameters are estimated by various techniques. The distributions of these estimates can then be compared with the true parameter values to study the bias and fit of the estimation technique with and without rounding errors in the data.

Inflation is typically modeled as a first-order, autoregressive process or a moving average process. Both of these inflation models were simulated so as to demonstrate some of the different effects of rounding errors. Model 1 assumes that inflation is a first-order autoregressive process where the error terms are normally distributed with mean μ and standard deviation σ:

1) $Dp(t) = \beta * Dp(t-1) + \varepsilon(t)$,

where *Dp(t)* is the rate of inflation in period *t* and $\varepsilon(t)$ is the normally distributed error term with mean μ and standard deviation σ. The rate of inflation is equal to the difference in the natural logarithms of the price level in periods *t* and *t*–1.

Model 2 assumes that the change in the rate of inflation is a first-order moving average process with normally distributed error terms:

2) $D^2p(t) = \delta * MA(1) + \varepsilon(t)$,

where $D^2p(t)$ is the change in inflation and δ is

FIGURE 2

CPI-W monthly growth rates

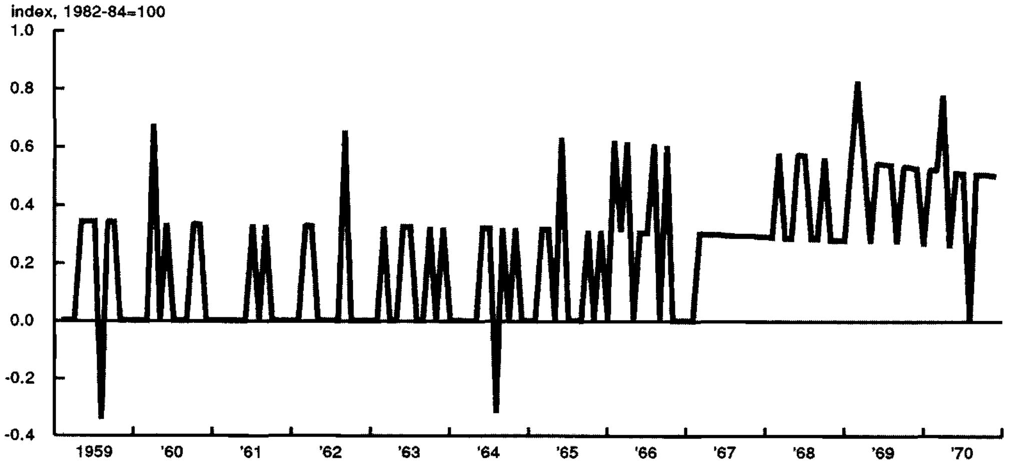
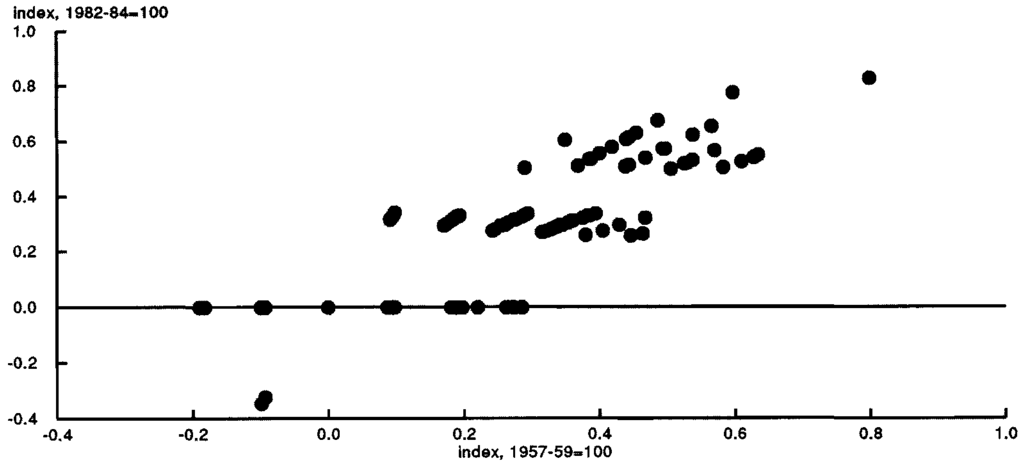


FIGURE 3

CPI-W scatter diagram

(Monthly percentage changes)



the estimated coefficient for the first-order moving average process. The change in inflation is calculated as the inflation in period t minus the inflation in period t–1. Using monthly CPI-W data, Pearce (1979) found that this equation provided stable estimates for sample periods from 1947 to 1976.

In each of the Monte Carlo simulations the following steps were followed. First, one of the underlying models was selected and values were assigned for each of the parameters in the model. In each simulation the initial value of the price level was set equal to 100.0. Next a series of 143 normally distributed error terms were generated. These random changes together with the specified model were then used to calculate a time series for the level of a price index.

From this series of 144 total observations, three additional series were calculated to study the effects of rounding errors. The first rounded series consists of the original data rounded to one digit after the decimal point. This series corresponds to the rounded monthly CPI numbers as they are first released by the Commerce Department, that is, without rebasing. To show the effects of rebasing on rounding errors, the original data were multiplied by two scaling factors, 0.5 and 0.3, in order to obtain two additional series. These two series were then rounded to one decimal point. The 0.5 scaling factor would be used when rebasing a price index series after prices have doubled. The 0.3 factor is the approxi-

TABLE 2

Monthly changes in CPI-W

(1959-70)

	1957-59 = 100.0		1982-84=100.0	
Change	Frequency	Percent	Frequency	Percent
–0.2	2	1.4		
–0.1	11	7.6	2	1.4
0.0	20	13.9	62	43.1
0.1	23	16.0	51	35.4
0.2	19	13.2	27	18.8
0.3	15	10.4	2	1.4
0.4	20	13.9		
0.5	15	10.4		
0.6	8	5.6		
0.7	5	3.5		
0.8	5	4.5		
0.9	0	0.0		
1.0	1	0.7		
Total	**144**	**100.0**	**144**	**100.0**

mate scaling factor that is now being used on CPI-W data from the 1960s.

This process of generating four time series of 144 observations was repeated 400 times using different sequences of random numbers. The model parameters were estimated by ordinary least squares (OLS) for each time series to determine the effects of rounding errors. The statistical results are presented in the following sections for each of the underlying models. In the following discussion of the results of the simulation, Series 1 is the unrounded data, Series 2 is the rounded data, Series 3 is the data multiplied by the 0.5 scaling factor and then rounded, and Series 4 is the data multiplied by the 0.3 scaling factor and then rounded.

Model 1

A Monte Carlo simulation of Model 1 has three coefficients (β, μ, and σ) that must be specified. These values were selected so as to approximate the time series properties of the CPI-W since the late 1950s. Typically, the value of β could be expected to fall between 0 and 1. In each of the simulations reported in this article, the value of β, the lagged inflation coefficient, was 0.5. The results were similar when alternative values between 0 and 1 were tried. For brevity, these simulations are not reported.

Three values for μ, the mean of the error term, were selected so that the average rate of inflation, that is, the average monthly change in the logarithm of the price index, would be 0.000, 0.002, or 0.004. Because the model includes a lagged dependent variable, the equation used to calculate μ was

3) $\mu = (1-\beta) * \pi$,

where π is the desired average rate of inflation. The two positive values for π were chosen because they approximate the average monthly inflation rate in the CPI-W over the 1960s and 1980s, respectively.

Two values for σ (0.001 and 0.002) were used so that θ, the expected standard deviation of the dependent variable, would be 0.002 or 0.004. These values were also chosen to approximate the standard deviation of inflation during the 1960s and the 1980s. The equation used to calculate σ was

4) $\sigma = \theta * \sqrt{1-\beta^2}$,

where θ is the desired standard deviation of inflation.

The equation estimated was

5) $Dp(t) = \alpha + \beta * Dp(t-1)$,

where α is an intercept term and is included because the errors in the underlying model can have a nonzero mean. Some of the estimated coefficients are listed in Tables 3, 4, and 5, where the coefficients, t-statistics, and correlation coefficients are the average values of the 400 simulations.

Figure 4 plots the Monte Carlo results for Model 1 with π (the average inflation rate) equal to 0.002 and θ (the standard deviation of inflation) equal to 0.002. The horizontal axis plots the three values of the base scaling factor, and the vertical axis plots the estimated values of β. The values for the estimated model parameters are listed in Table 3.

The simulation results for one value of π are presented in Figure 4 in order to show the confidence interval for each estimated β. These confidence intervals are indicated by the vertical bars around each of the labeled points in the Figure. In the 400 simulations, 68 percent of the estimated β coefficients were in the range delimited by these vertical bars.

Figure 4 illustrates what happens to the sample bias when the original time series of price changes is rebased and rounded. The point labeled "ORG" in Figure 4 is the OLS estimate of β for the unrounded data (Series 1). As shown in Table 3, the estimated β for Series 1 is 0.481, which is close to the actual β

TABLE 3

Model 1 (β=0.5, π=0.002, θ=0.002)

Model	est. α	t-stat	est. β	t-stat	R^2
Series 1	0.00103	4.97	0.481	6.59	0.237
Series 2	0.00110	5.13	0.450	6.06	0.209
Series 3	0.00126	5.54	0.366	4.72	0.141
Series 4	0.00155	6.16	0.220	2.71	0.057

TABLE 4					
Model 1 (β=0.5, π=0.004, θ=0.002)					
Model	est. α	t-stat	est. β	t-stat	R^2
Series 1	0.00207	6.33	0.481	6.62	0.239
Series 2	0.00216	6.50	0.458	6.21	0.217
Series 3	0.00243	7.01	0.390	5.10	0.160
Series 4	0.00292	7.90	0.267	3.34	0.080

TABLE 5					
Model 1 (β=0.5, π=0.002, θ=0.004)					
Model	est. α	t-stat	est. β	t-stat	R^2
Series 1	0.00103	3.12	0.481	6.58	0.237
Series 2	0.00105	3.14	0.475	6.45	0.230
Series 3	0.00110	3.21	0.449	6.03	0.208
Series 4	0.00119	3.33	0.400	5.24	0.167

of 0.50. The sample bias of OLS when there is a lagged dependent variable and an intercept is consistent with the estimated bias calculated by Marriott and Pope (1954). The point labeled "RND" in Figure 4 is the estimated β for Series 2, in which the original data has been rounded to one decimal point. The estimated β for Series 2 is 0.450, according to Table 3. Thus, the sample bias for the rounded data is larger than the sample bias for the unrounded data. Notice also that the R^2 for Series 2 is 0.209, compared to 0.237 for Series 1. Thus, the degree of fit of the model decreases somewhat when the data is rounded.

The points labeled "A" and "B" in Figure 4 are the estimated βs, respectively, for Series 3 and 4, the rebased series. The estimated β for Series 3 is 0.366 and the estimated β for Series 4 is 0.220. Clearly, the sample bias increases markedly when the data is rebased by a factor of 0.5, and even more when the data is rebased by a factor of 0.3. Thus, as the rebasing factor decreases (and the difference between the adjusted data and the original data increases) the sample bias increases. Also, as shown in Table 3, the R^2 or degree of fit decreases when the rebasing factor decreases. Figure 4 and Table 3 show that the consequences of rebasing and rounding data for a first-order, lagged dependent variable model are increasingly biased estimates and decreasing fit.

Figure 5 plots the estimated coefficients for the lagged inflation term using three alternative average rates of inflation. The middle line in Figure 5 is the same line shown in Figure 4. The top line in Figure 5 represents the simulation in which the average inflation rate is 0.004, and the bottom line represents the simulation in which the average inflation rate is 0.000. Figure 5 shows that sample bias is increased when the average inflation rate is lower because the "noise" introduced by rebasing and rounding is imposed on a time series that has less "signal".

Table 4 lists the estimated parameters for the case where the average inflation rate is 0.004. Comparing Table 4 with Table 3, we see similar patterns in the estimates for the coefficients, t-statistics, and R^2. The estimated intercept terms for the αs are approximately twice as large in Table 4 because the average inflation rate has been doubled in these simulations. The bias estimates for β are slightly smaller when the average inflation rate is higher.

FIGURE 4

Model 1: Rebasing bias Confidence intervals

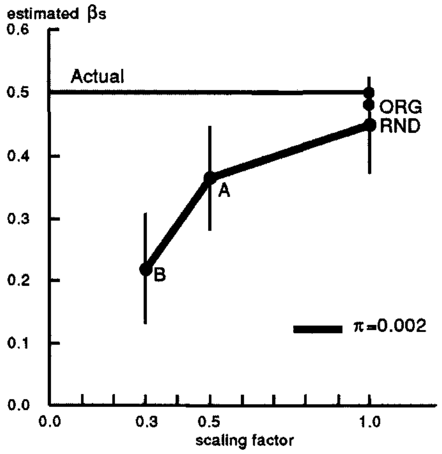


FIGURE 5

Model 1: Rebasing bias
Different rates of inflation

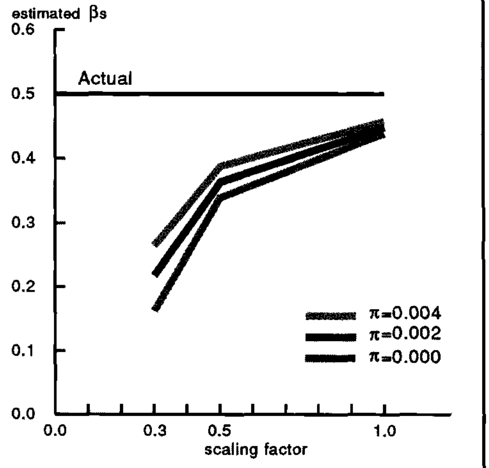


FIGURE 6

Model 1: Rebasing bias
Higher standard deviation of inflation

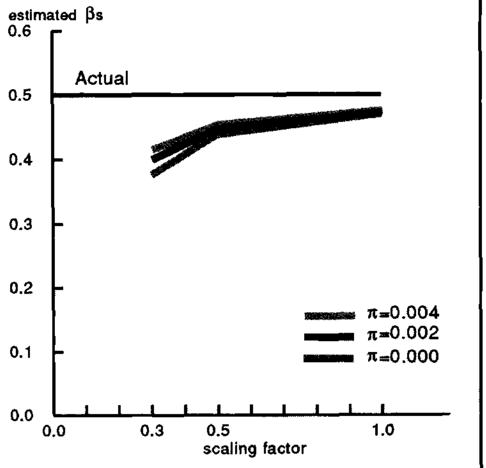


Figure 6 plots the estimated coefficients for the same average inflation rate as in Figure 5, but with the standard deviation of inflation doubled to 0.004. Comparing Figure 5 and Figure 6, we see that the sample biases in Figure 6 are consistently lower because the higher standard deviation means that there is a higher signal to noise ratio when the data is rebased and rounded. The estimated parameters for one of these simulations are listed in Table 5. Comparing Table 5 and Table 3, we see that the sample biases for α and β are consistently smaller when the standard deviation of inflation is higher.

In summary, we can draw the following conclusions from Model 1:

1) The bias of OLS estimates is increased when the data is rounded.

2) This bias is greater when the mean of the error term is closer to zero and/or the standard deviation is lower.

3) The bias increases as the scaling factor decreases.

4) The correlation coefficient or degree of fit decreases as the scaling factor decreases.

The Monte Carlo simulations using Model 1 show that rounding errors are a systematic source of noise that can be imposed on a data series. The effect of rounding errors and rebasing is to systematically remove information and bias the estimates away from the actual coefficients. The smaller the signal in the data, that is, the lower the mean and standard deviation for the error process, the greater is the effect on estimations.

These results, with hindsight, are somewhat intuitive. Consider an idealized case somewhat similar to the early 1960s. Suppose that each month the absolute (not percentage) change in the index is approximately 0.1. When the data is rebased with a scaling factor of 0.3, the three month pattern will become two months with no change followed by one month with an absolute change of about 0.1. Over the course of a year this will yield essentially the same average rate of change as the original.

Using the original data in this idealized case, Model 1 without an intercept term would have an estimated coefficient of approximately 1.0 and would be very significant. After the 0.3 rebasing, the new three month pattern would generate an estimated coefficient close to 0.0 which would not be statistically significant. This is an extreme case of the bias found in these simulations using Model 1.

Model 2

In the Monte Carlo simulations using Model 2, in which the change in inflation is assumed to be a first-order moving average process, three values for the coefficient of the moving average process are presented to show how the size of the bias due to rounding is dependent upon the size of this coefficient.

Typically, the value of the moving average coefficient could be expected to fall between 0 and −1. Therefore, the alternative values used for the moving average coefficient were −0.25, −0.50, and −0.75. In these simulations the mean of the error term was assumed to be zero because there is no trend in the change in inflation during the 1960s.

The standard deviation of the error term was selected so that the expected standard deviation of the change in inflation was 0.002. This value approximates the standard deviation of the change in CPI-W inflation during the 1960s. The equation to calculate σ was

$$6)\quad \sigma = \psi / \sqrt{1 + \delta^2},$$

where ψ is the desired standard deviation of the change in inflation. Simulations using alternative standard deviations are not reported because the results are similar to the estimates for this simulation.

Figure 7 plots the estimated values for the moving average coefficient and the corresponding confidence intervals. The average values for δ, as well as the t-statistics and R^2s, for Series 1-4 are presented in Table 6. The point labeled "ORG" in Figure 7 is the estimated moving average coefficient for Series 1, the unrounded series. As shown in Figure 7, this estimate is very close to the actual value of the coefficient. According to Table 6, the estimated value is −0.497, compared to the actual value of −0.5. Thus, OLS gives an unbiased estimate of the moving average coefficient before rounding and rebasing.

The points labeled "RND", "A", and "B" in Figure 7 show the effects of rounding and rebasing on estimates of δ. As in Model 1, this effect is to increase the bias or difference between the estimated coefficient and the actual coefficient. However, note that as shown in Table 6, the t-statistics for the moving average coefficient and the R^2 for the model *increase* for Series 2, 3, and 4. Thus the effect of rounding and rebasing in Model 2 are to bias the results to accepting a spurious fit.

In Figure 8 the three lines correspond to the three alternative values for the coefficient of the moving average process. Tables 7 and 8 list the estimated parameters for the two additional values of the moving average coefficient. For each of the three lines shown in Figure 8, the points labeled "ORG" are the estimated coefficients before rounding and are very close to the actual values used in the different simulations. In each case, the effect of rounding and rebasing is to estimate a larger negative coefficient for the moving average term and to increase the corresponding t-statistic. The increases in the coefficient and t-statistic are essentially proportional so that the standard error of the estimate is essentially unchanged.

As shown in Figure 8, the bias towards a spurious fit resulting from rounding and rebasing is most noticeable for $\delta = -0.25$ and least noticeable for $\delta = -0.75$. In each case the effect is to bias the estimated moving average coefficient towards −1.0.

In order to compare the results for Model 2 with the results of Model 1, recall the idealized case presented at the end of the discussion of Model 1, in which the average change

FIGURE 7

Model 2: Rebasing bias
Confidence intervals

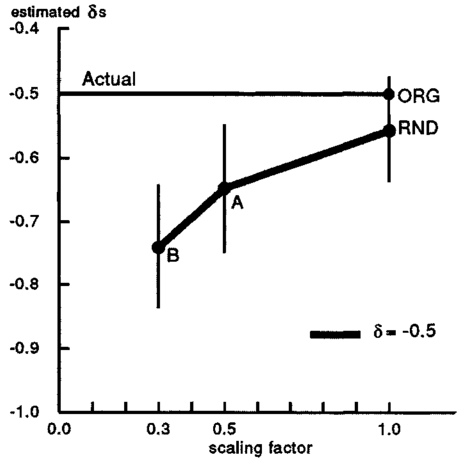


TABLE 6

Model 2 (δ=−0.50)

Model	est. δ	t-stat	R^2
Series 1	−0.497	−5.86	0.202
Series 2	−0.555	−6.56	0.243
Series 3	−0.646	−7.63	0.312
Series 4	−0.740	−8.73	0.389

TABLE 7			
Model 2 (δ=–0.25)			
Model	est. δ	t-stat	R^2
Series 1	–0.250	–2.95	0.062
Series 2	–0.355	–4.18	0.123
Series 3	–0.476	–5.60	0.205
Series 4	–0.600	–7.06	0.299

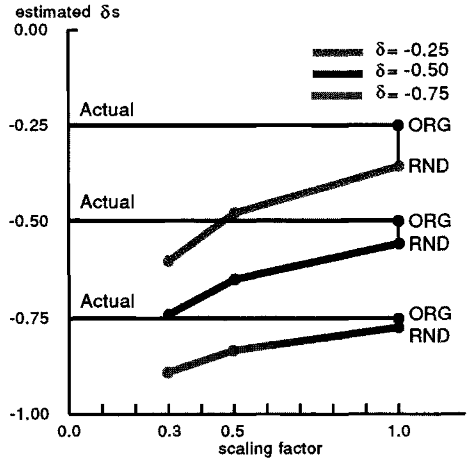
TABLE 8			
Model 2 (δ=–0.75)			
Model	est. δ	t-stat	R^2
Series 1	–0.739	–8.74	0.369
Series 2	–0.771	–9.11	0.385
Series 3	–0.832	–9.83	0.439
Series 4	–0.890	–10.53	0.501

in the price index for each month is approximately 0.1. In Model 1, the result of the 0.3 scaling of the series was to generate a statistically insignificant coefficient close to 0, as compared to a significant coefficient of approximately 1.0 for the original data when there is not an intercept term.

In the case of Model 2, rebasing this series for a factor of 0.3 and then differencing to get the change in inflation would result in a three month pattern of approximately 0, 0.1, and –0.1. If a series with this type of pattern is estimated as a first-order moving average process, the expected coefficient would be close to –1.0 and would be highly significant, as compared to a coefficient of 0 for the original data. This is an extreme case of the bias found in Model 2.

FIGURE 8

Model 2: Rebasing bias
Different MA coefficients



What should be done?

The distortions in timing and magnitude and biases in estimations that are caused by rounding errors in index numbers can be a serious problem for some time periods and economic time series. Of course the obvious solution would be for the reporting agency to provide more digits when they release the historical revisions. In many cases this would amount to a simple change in the way they format the computer tapes and/or diskettes that they sell to the public.

Until this additional data is made available, the users of historical data should be careful to check if rounding errors are affecting their results. While smoothing filters and/or the use of more time aggregated data can alleviate these high frequency distortions and biases, these techniques may not always be appropriate. In the end, there is no substitute for better data.

FOOTNOTES

[1]Technically, this problem is not purely the result of rebasing, but also of publishing the numbers to a fixed accuracy, so that there are fewer significant digits the farther back you go in a series with an upward historical trend.

REFERENCES

Marriott, F. H. C., and J. A. Pope, "Bias in the estimation of autocorrelation," *Biometrika*, 41, 1954, pp. 390-402.

Pearce, Douglas K., "Comparing survey and rational measures of expected inflation," *Journal of Money, Credit, and Banking*, 11, November 1979, pp. 447-56.