Volume Title: Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance

Volume Author/Editor: Jacob A. Mincer, editor

Volume Publisher: NBER

Volume ISBN: 0-870-14202-X

Volume URL: http://www.nber.org/books/minc69-1

Publication Date: 1969

Chapter Title: Data Errors and Forecasting Accuracy

Chapter Author: Rosanne Cole

Chapter URL: http://www.nber.org/chapters/c1215

Chapter pages in book: (p. 47 - 82)

TWO

Data Errors and Forecasting Accuracy

ROSANNE COLE

INTRODUCTION

A basic requirement for successful economic forecasting is accurate data. Though it is widely recognized that most economic statistics contain measurement errors, relatively little effort has been made to determine how much of the error in forecasts might be attributed to errors in the underlying data. An analysis of the importance of this source of forecast error, however, is indispensable for a proper evaluation of forecasting accuracy. For example, conclusions about the quality of a set of forecasts (and hence the model used to generate them) would vary according to whether data errors were found to be a major or a negligible component of forecast error. Moreover, an analysis of data errors can provide an indication of the potential for improving forecasts by improving the accuracy of the underlying data.

The major difficulty confronting an empirical analysis of the effects of data errors on forecasting accuracy is that very little is known about the errors in many economic series. One type of information is avail-

NOTE: The reader who has traveled the preceding chapter will recognize throughout this report my indebtedness to Jacob Mincer and Victor Zarnowitz. This is my opportunity to do so. I should also like to thank Phillip Cagan, John Kendrick, and Julius Shiskin for their helpful comments on an earlier version of the paper.

able, however, and it approximates the data problems of a forecasting situation rather well. The data underlying many forecasts are preliminary estimates. Revised estimates based on more complete information are published at a later date. Since the revised estimates are presumably more accurate, the consequences of using preliminary rather than revised data can be viewed as an illustration of the effects of data errors.[1]

Another difficulty arises from the fact that the majority of published forecasts are not scientific: neither the forecasting models nor the data they use are explicitly specified. Such forecasts are nevertheless of considerable and practical interest to a broad class of business and economic analysts. Since they cannot be replicated, the effect of data errors on their accuracy cannot be assessed directly. However, by relating their errors to errors in data available at the time the forecasts were made, it is possible to infer indirectly the element of forecast error attributable to data errors. More specifically, forecasts may be assumed to rely partly on extrapolations of recent levels of the series to be predicted. Any shortcomings in these data would thus be transferred to the forecasts and become a source of error.

This chapter shows the effect of using preliminary rather than revised data on the accuracy of three types of short-term forecasts of GNP and its major components: (1) forecasts which consist only of extrapolations (naive models); (2) business forecasts which may rely partly on extrapolations and partly on other information; and (3) an analytical model of consumption in which the as yet unknown value of an exogenous variable is obtained by extrapolation. The main emphasis is thus on errors in the data underlying the extrapolative component of a forecast. The effects of errors in other data on which business forecasts may rely are not explicitly considered.

The first section of this study contains an analysis of the ways in which data errors would impair forecast accuracy and a brief review of the characteristics of the errors (as indicated by subsequent revisions) in preliminary GNP data. These errors are shown to be a potential source of forecast bias as well as inefficiency. Though the

[1] A recent study of this nature was made by Denton and Kuiper [2]. They constructed a small econometric forecasting model based on the Canadian national accounts, generated forecasts, and observed directly the effects of using preliminary rather than revised data on the parameters of the model and on the accuracy of the simulated forecasts.

analysis is formulated in terms of errors in the data underlying the extrapolative component of forecasts, it is general enough to apply to errors in any data on which forecasts may draw.

In the empirical analysis that follows, estimates are made of the extent to which the use of preliminary rather than revised GNP data reduced the over-all accuracy of naive projections and business forecasts. The importance of data errors as a source of bias and inefficiency in business forecasts is then assessed. Finally, the effect of such errors on the parameter estimates and predictive accuracy of a quarterly consumption function is shown. The findings are summarized in the last section.

I. THE DATA ERROR COMPONENT OF FORECASTS, REALIZATIONS, AND FORECAST ERRORS

An observed forecast error may contain data errors of two kinds: (1) measurement errors in the data used to construct the forecast and (2) measurement error in the realized value. Data errors of the first kind will be a component of the true forecast error. Data errors of the second kind will cause the observed forecast error to differ from the true error. In order to illustrate their different effects, we shall first assume that realizations are measured without error and consider only the consequences of errors in the data underlying forecasts. This assumption will then be relaxed and the effect of errors in realizations shown.

ERRORS IN THE UNDERLYING DATA

Let A be the series forecast. A forecast P_t, made in period $t-1$, of the value of A in period t can be considered as consisting partly of a projection of past values of the series as in

$$P_t = \gamma_1 A^\circ_{t-1} + \gamma_2 A^\circ_{t-2} + \gamma_3 A^\circ_{t-3} + \cdots + h_t, \tag{1}$$

where A° denotes the series of estimates available to the forecaster at the time the forecast is made; γ_i is the weight assigned to A°_{t-i}, the value of the series in period $t-i$; and h_t is an autonomous component

summarizing all other information on which the forecast may draw. The linear formulation is assumed for the sake of simplicity.

Most forecasters use the series of best available estimates. This series (A°) consists of mixed data: The values for the most recent periods are provisional estimates while those for periods further into the past have been revised at least once. An estimate from this series of the value of A in period $t - i$ differs from the final series by the error ϵ. In symbols,

$$A^\circ_{t-i} = A_{t-i} + \epsilon_{t-i}. \tag{2}$$

As a consequence of using preliminary (A°) rather than final data (A), errors in A° are incorporated into the forecast. Using equation (2), the forecast can be rewritten as the sum of P'_t, the forecast that could have been made if final data were available to the forecaster, and an element of data error,

$$P_t = \Sigma\gamma_i A_{t-i} + h_t + \Sigma\gamma_i\epsilon_{t-i} = P'_t + \Sigma\gamma_i\epsilon_{t-i}, \tag{1'}$$

where $P'_t = \Sigma\gamma_i A_{t-i} + h_t$.

The forecast error (u) is defined

$$u_t = P_t - A_t = (P'_t - A_t) + \Sigma\gamma_i\epsilon_{t-i} = u'_t + \Sigma\gamma_i\epsilon_{t-i}, \tag{3}$$

where u' is the error of the forecast (P') based on final data. Errors in the preliminary data (A°) are thus a component of both the forecast and its error.

Let us first consider the case in which the preliminary data are unbiased [i.e., $E(A^\circ) = E(A)$, such that $E(\epsilon) = 0$, where E denotes expected value]. How would such data errors affect forecasting accuracy? The expected value of the forecast error,

$$E(u) = E(P) - E(A) = E(u') + \Sigma\gamma_i E(\epsilon), \tag{4}$$

is a measure of the bias. Forecasts are unbiased if $E(P) = E(A)$ and in that case $E(u) = 0$. It is clear from (4) that if the preliminary data were unbiased [$E(\epsilon) = 0$], their error would not be a source of bias in P.

Whether the preliminary data are biased or not, their errors are likely to reduce forecasting efficiency. Provided the two components of the forecast error in (3) are uncorrelated, the variance of the forecast error is

$$\sigma^2(u) = \sigma^2(u') + \sigma^2(\Sigma\gamma_i\epsilon_{t-i}), \tag{5}$$

and $\sigma^2(u)$ must exceed $\sigma^2(u')$. The observed forecast (P) is therefore less efficient than the forecast (P') that could have been made with final data.

A particular aspect of inefficiency is the presence of a "slope error." This obtains if a linear correction of the forecast would reduce the variance of its error.[2] Such a correction is given by a least squares regression of A on P:

$$A_t = \alpha + \beta P_t + v_t. \tag{6}$$

The corrected forecast is $\alpha + \beta P_t$ and the variance of its error is the residual variance, $\sigma^2(v) = (1 - \rho_{AP}^2)\sigma^2(A)$, where ρ_{AP}^2 denotes the coefficient of determination. The variance of the forecast error can then be expressed as the sum of a potentially reducible, or systematic, component and the residual variance,

$$\sigma^2(u) = (1 - \beta)^2\sigma^2(P) + (1 - \rho_{AP}^2)\sigma^2(A). \tag{7}$$

It is clear that unless $\beta = 1$, P would be inefficient because $\sigma^2(u)$ would exceed $(1 - \rho_{AP}^2)\sigma^2(A)$.

It can be readily shown that the forecast would be efficient (i.e., $\beta = 1$) only if it is uncorrelated with its error. The correction factor β is, by definition,

$$\beta = \frac{\text{Cov } (A,P)}{\sigma^2(P)}. \tag{8}$$

It follows from the identity $A \equiv P - u$ that

$$1 - \beta = \frac{\text{Cov } (u,P)}{\sigma^2(P)}, \tag{9}$$

and Cov $(u,P) = 0$ implies $\rho_{uP} = 0$.

We have seen, however, that P and u share a common error $(\Sigma\gamma_i\epsilon_{t-i})$, which creates a positive correlation between them. As a result, $1 - \beta$ would not equal zero even if P' and u' were uncorrelated, since, using (1′) and (3),

$$1 - \beta = \frac{\text{Cov } (u', P') + \sigma^2(\Sigma\gamma_i\epsilon_{t-i})}{\sigma^2(P)}. \tag{9′}$$

[2] Jacob Mincer and Victor Zarnowitz in Chapter 1 of this volume propose this component of forecast (in)efficiency in their decomposition of mean square errors.

Random data errors could therefore augment the slope component of $\sigma^2(u)$. They would also increase the residual variance component. If ϵ_{t-i} were uncorrelated with A and P', and if β' denotes the regression coefficient, and $\rho^2_{AP'}$, the coefficient of determination, in the regression of A on P', the slope component would be [3]

$$(10) \qquad (1-\beta)^2\sigma^2(P) = (1-\beta')^2\sigma^2(P') + (1-\beta'\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i}),$$

and the residual variance would be [4]

$$(11) \qquad (1-\rho^2_{AP})\sigma^2(A) = (1-\rho^2_{AP'})\sigma^2(A) + \beta'\beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i}).$$

The extent to which data errors increase the mean square error and each of its three components can now be seen. The mean square error of P' is defined as $E(P'-A)^2$ and equals

$$(12) \qquad M' = [E(P'-A)]^2 + (1-\beta')^2\sigma^2(P') + (1-\rho^2_{AP'})\sigma^2(A).$$

Because of data errors, the mean square error of the observed forecast P would be

$$(13) \quad M = [E(P'-A+\Sigma\gamma_i\epsilon_{t-i})]^2 + [(1-\beta')^2\sigma^2(P') + (1-\beta'\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i})] + [(1-\rho^2_{AP'})\sigma^2(A) + \beta'\beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i})].$$

Let P' be unbiased and efficient. Then $E(P'-A)=0$ and $\beta'=1$. The first two components of M', the mean and slope components, would then vanish and $M' = (1-\rho^2_{AP'})\sigma^2(A)$. However, the mean component of M would be

$$[E(\Sigma\gamma_i\epsilon_{t-i})]^2;$$

the slope component would be

$$(1-\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i});$$

and the residual component would be

$$(1-\rho^2_{AP'})\sigma^2(A) + \beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i}).$$

[3] Since $\beta = \beta' \dfrac{\sigma^2(P')}{\sigma^2(P)}$ and $\sigma^2(P) = \sigma^2(P') + \sigma^2(\Sigma\gamma_i\epsilon_{t-i})$, $(1-\beta)^2\sigma^2(P) = \sigma^2(P') + \sigma^2(\Sigma\gamma_i\epsilon_{t-i}) - 2\beta'\sigma^2(P') + \beta'^2\sigma^2(P') \dfrac{\sigma^2(P')}{\sigma^2(P)}$. Add and subtract $\beta'^2\sigma^2(P')$ and rearrange terms to obtain equation (10).

[4] The proof is similar to that given in footnote 3.

Thus, errors in the underlying data could convert an unbiased, efficient forecast into a biased, inefficient one. The size of the bias would be $E(\Sigma\gamma_i\epsilon_{t-i})$.[5] The variance of the forecast error would be augmented by $\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$ in such a way that a slope error of $(1-\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$ would be created and the residual variance would be augmented by $\beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$.

On the assumption that P' is unbiased and efficient, M would reduce to

$$M = (1-\rho^2_{AP'})\sigma^2(A) + E(\Sigma\gamma_i\epsilon_{t-i})^2. \tag{13'}$$

Let the relative magnitude of errors in the preliminary data be $k^2 = \frac{E(\epsilon)^2}{\sigma^2(A)}$. The data error component of M can then be expressed

$$E(\Sigma\gamma_i\epsilon_{t-i})^2 = \Sigma\gamma_i^2 k^2\sigma^2(A), \tag{14}$$

provided that $E(\epsilon_{t-i})^2 = E(\epsilon)^2$ for all i.

The relative mean square error, $RM' = M/M'$, shows the extent to which data errors increase the pure forecast error. Substituting (14) into (13'),

$$RM' = \frac{(1-\rho^2_{AP'})\sigma^2(A) + \Sigma\gamma_i^2 k^2\sigma^2(A)}{(1-\rho^2_{AP'})\sigma^2(A)} = 1 + \frac{\Sigma\gamma_i^2 k^2}{1-\rho^2_{AP'}}. \tag{15}$$

Thus, given the relative size of errors in the preliminary data (k^2), the increase in forecast error will be greater, the greater the weights assigned to these data ($\Sigma\gamma_i^2$) and the better the forecast (i.e., the greater $\rho^2_{AP'}$).

ERRORS IN REALIZATIONS

Thus far it has been assumed that realizations are measured without error, and we have seen only the effect of errors in the data underlying forecasts. In practice, however, realizations are also likely to contain measurement errors which obscure true forecast errors. In keeping with the preceding example of measurement errors, let us now assume that realizations consist of preliminary (A°) rather than revised (A) data. The observed forecast error (u°) is then defined as $P - A^\circ$ and equals

[5] If the preliminary data are unbiased [i.e., $E(\epsilon_{t-i}) = 0$ for all i]. the size of the bias would of course be zero.

(16)
$$u_t^\circ = P_t - A_t^\circ = (P_t' + \Sigma\gamma_i\epsilon_{t-i}) - (A_t + \epsilon_t) = u_t - \epsilon_t = u_t' + \Sigma\gamma_i\epsilon_{t-i} - \epsilon_t.$$

With the aid of (16), three types of forecast error can be distinguished and their relation to each other shown: the observed forecast error (u°); the true forecast error (u); and the pure forecast error (u'). If there were no data errors, the three forecast errors would be identical. Errors in the data used to construct forecasts ($\Sigma\gamma_i\epsilon_{t-i}$) augment the pure forecast error (u') and become a component of the true forecase error (u). Errors in realizations data (ϵ_t) cause the observed error (u°) to differ from the true error (u).

The observed error thus consists of a pure forecast error and two components of data error. The expected value of u° is

(17)
$$E(u^\circ) = E(u - \epsilon) = E(u') + (\Sigma\gamma_i - 1)E(\epsilon),$$

provided $E(\epsilon_{t-i}) = E(\epsilon)$ for all i; and the variance is

(18)
$$\sigma^2(u^\circ) = \sigma^2(u') + \sigma^2(\Sigma\gamma_i\epsilon_{t-i}) + \sigma^2(\epsilon_t) - 2\ \mathrm{Cov}\ (\Sigma\gamma_i\epsilon_{t-i},\epsilon_t),$$

provided u' and ϵ are uncorrelated.

If the two sets of data errors are independent of each other; that is, if ϵ is serially independent, the covariance term in (18) would vanish. The variance of the forecast error would then be augmented by the errors in the realizations data, $\sigma^2(\epsilon_t)$, as well as by errors in the data underlying the forecast, $\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$.

However, $\sigma^2(\epsilon_t)$ would not be distributed among the slope and residual components in the same way as $\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$. If β° and $\rho^2_{A^\circ P}$ denote the coefficients of regression and determination, respectively, the observed mean square error (M°) equals

(19)
$$M^\circ = [E(P - A^\circ)]^2 + (1 - \beta^\circ)\sigma^2(P) + (1 - \rho^2_{A^\circ P})\sigma^2(A^\circ).$$

The mean component of M° would equal

$$[E(P - A^\circ)]^2 = [E(u - \epsilon)]^2 = [E(u') + (\Sigma\gamma_i - 1)E(\epsilon)]^2;$$

the slope component would be

$$(1 - \beta^\circ)^2\sigma^2(P) = (1 - \beta)^2\sigma^2(P);$$

and the residual component would be

$$(1 - \rho^2_{A^\circ P})\sigma^2(A^\circ) = (1 - \rho^2_{AP})\sigma^2(A) + \sigma^2(\epsilon),$$

provided Cov $(\epsilon_t,\epsilon_{t-i})$ and Cov (A,ϵ) are zero. If P' is unbiased and efficient, M° would become, using (10) and (11),

$$(19') \qquad M^\circ = [(\Sigma\gamma_i - 1)E(\epsilon)]^2 + [(1-\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i})] + [(1-\rho^2_{AP'})\sigma^2(A) + \beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i}) + \sigma^2(\epsilon)].$$

Data errors would thus affect observed forecasting accuracy in the following ways: Given that errors in the data used to construct the forecast are independent of errors in the realized values (i.e., ϵ is serially independent), then, if the preliminary data are biased, the bias in realizations data would tend to offset the bias induced by errors in the data underlying the forecast. Indeed, in the special case in which $\Sigma\gamma_i = 1$, the biases would be exactly offsetting. Both sets of data errors would reduce forecasting efficiency. Errors in the underlying data would increase the variance of the forecast error by $\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$. They would create a slope error of $(1-\beta)\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$ and increase the residual variance by $\beta\sigma^2(\Sigma\gamma_i\epsilon_{t-i})$. Errors in realizations would have no effect on slope error; they would reduce the forecast's efficiency by augmenting the residual variance by $\sigma^2(\epsilon_t)$.

There would be a smaller reduction in forecasting efficiency if the two data errors were related. If ϵ were serially correlated, the reduction in efficiency arising from errors in the underlying data would be attenuated by errors in the realizations data. The extent to which the two data errors are offsetting depends, as (18) shows, on the weights assigned to the underlying data $(\Sigma\gamma_i)$ and the strength of the serial correlation in ϵ (ρ_ϵ). At the one extreme, in the special case in which $\Sigma\gamma_i = 1$ and $\rho_{\epsilon_t\epsilon_{t-i}} = 1$ for all i, errors in realizations would exactly offset the reduction in efficiency caused by errors in the underlying data. Then $\sigma^2(u^\circ) = \sigma^2(u') < \sigma^2(u)$, and the observed forecast error would be an unbiased, efficient estimate of the pure forecast error. At the other extreme, $\rho_{\epsilon_t\epsilon_{t-i}} = 0$ for all i, and, as we have seen, errors in realizations would augment the reduction in efficiency and $\sigma^2(u^\circ) > \sigma^2(u) > \sigma^2(u')$. In practice, the effect of errors in the realizations data is likely to fall somewhere in between.

ERRORS IN PRELIMINARY GNP DATA

Table 2-1 shows summary statistics of the errors (as measured by revisions) in preliminary estimates of GNP and its components. We would expect the preliminary data to be unbiased and their errors to be serially independent if they were generated by a probability sampling process. This is not the case, however, for the national accounts estimates. As Table 2-1 shows, and, as is well documented elsewhere,[6] the preliminary product (or expenditures) data have a negative bias: the preliminary estimates underestimate revised levels of GNP and most of its components. These data could thus be a source of negative bias in forecasts which rely on them. Since the biases in the detailed variables do not offset one another, we could expect the bias induced by data errors to be larger in forecasts of aggregates than in forecasts of detailed variables.

Errors in aggregate variables, however, would be likely to increase the mean square error (from M' to M) in forecasts of these variables less than that of the detailed variables. This is suggested by the k-ratios in column 4 of Table 2-1. The relative size of the data errors (k) tends to be larger in details than in aggregates. Thus, other things being constant (i.e., $\Sigma\gamma_i^2$ and $\rho_{AP'}^2$), data errors would be expected to cause the greatest increase in the over-all error of forecasts of net exports, expenditures on consumer and producer durables, new construction, and change in business inventories; they should have the smallest effect on the accuracy of forecasts of GNP, personal consumption, and government expenditures on goods and services.

Errors in the preliminary data for some GNP components show strong, positive serial correlation (column 5). Therefore, the errors would tend to be offsetting when these data are used as realizations as well as inputs to forecasts. Indeed, the possibility for offsetting the errors in the data underlying forecasts by the errors in realizations data has led several investigators to choose preliminary rather than revised GNP data as the set of realized values.[7] In effect, they are using the observed forecast error (u°) to approximate the pure forecast error (u'). The accuracy of this approximation depends, as shown elsewhere,

[6] See, for example, [7], [1], and Rosanne Cole, "Errors in Provisional Estimates of Gross National Product," NBER, forthcoming.

[7] See, for example, [5], [6], and Chapter 1 of this volume.

TABLE 2-1. Errors in Preliminary Estimates of Annual Levels of Gross National Product and Its Components, 1953–63 [a] (*dollars in billions*)

Variable	Mean Error $\bar{\epsilon}$ (1)	Standard Deviation of Error S_ϵ (2)	Root Mean Square Error $\sqrt{M_\epsilon}$ (3)	k-Ratio $\sqrt{\frac{M_\epsilon}{\sigma^2(A)}}$ (4)	Serial Correlation Coefficient $r_{\epsilon_t\epsilon_{t-1}}$ (5)
Gross national product	−11.0	4.5	12.8	.166	.570
Personal consumption expenditures	−6.0	1.6	6.2	.125	.841
Durables	−2.9	1.6	3.2	.508	.976
Nondurables	1.8	1.9	2.5	.141	.948
Services	−4.9	1.4	5.1	.198	.457
Gross private domestic investment	−4.2	2.7	4.9	.441	.429
Producers durable equipment	−0.5	2.9	2.8	.642	.464
New construction	−2.7	2.6	3.6	.764	.976
Change in business inventories	−1.0	1.6	1.8	.625	.176
Gov't expend. on goods and services	1.0	1.7	1.9	.114	.574
Federal	0.6	1.6	1.6	.239	.679
State and local	0.3	0.8	0.8	.072	.568
Net exports	−1.8	0.7	1.9	.905	.958

[a] Errors are computed as $\epsilon = A^\circ - A$, where A° denotes provisional estimates and A, the 1965 statistically revised estimates. Provisional estimates of the value of GNP and its components during a given year are from the next year's February issue of the *Survey of Current Business* (*SCB*). The 1965 statistically revised estimates are from the August 1965 *SCB*. The figures published are the result of both statistical and definitional revisions. The major definitional change was to exclude interest paid by consumers from the estimates (see the report article, "National Income and Product Accounts," *SCB*, August 1965, Tables 2 and 3). This item was added to the published figures (expenditures on consumer services, and hence to the aggregates, personal consumption expenditures and gross national product), to obtain estimates of only the statistically revised data. This procedure does not entirely eliminate the definitional changes, and the resulting series (A), therefore, includes some minor definitional changes in federal government expenditures and net exports.

on the strength of the serial correlation in ϵ and on the importance of the extrapolative component in the forecast.

In the following sections, estimates are made of the extent to which errors in the preliminary data augmented the pure forecast error of three types of forecasts of GNP and its components. Though u' (and hence the effect of data errors) can be directly observed for two of the three, it must be estimated indirectly for the third, business forecasts. Two estimates of u' are made for business forecasts. Regression analysis is used to decompose u into its two components: u' and $\Sigma\gamma_i\epsilon_{t-i}$. The results are then compared with those obtained when u° is used as an estimate of u'.

II. EFFECT OF DATA ERRORS ON THE ACCURACY OF NAIVE PROJECTIONS

Naive models are a class of forecasts constructed from past values of the target series. They are widely used not so often as forecasts per se but as yardsticks for appraising the performance of more sophisticated forecasts.

Table 2-2 shows the root mean square errors of naive projections of annual levels in GNP and its components for the period 1953–63. Errors of projections constructed with preliminary and with revised (1965) data are compared for three types of naive projections, denoted N1, N2, and N3. The projections based on preliminary data, P_t, are specified

$$\text{N1:} \quad P_t = A^{\circ}_{t-1}$$

$$\text{N2:} \quad P_t = A^{\circ}_{t-1} + (A^{\circ}_{t-1} - A^{\circ}_{t-2})$$

$$\text{N3:} \quad P_t = A^{\circ}_{t-1} + (A^{\circ}_{t-1} - A^{\circ}_{t-n})/n,$$

where n is the number of observations in the series. The projections based on revised data, P'_t, are the same except that A_{t-i} replaces A°_{t-i} in each case.

These models are thus special cases of the forecast described in equation (1) above, in which the autonomous component of the forecast (h_t) is zero, and the weights (γ_i) assigned to past values of the series are set arbitrarily. In the case of N1, γ_1 equals unity and all other γ coefficients are zero. The model N2 projects the last known level plus the last known change in the series ($\gamma_1 = 2$, $\gamma_2 = -1$, and $\gamma_i = 0$ for $i > 2$); N3 projects the last known level plus the average change [$\gamma_1 = (n+1)/n$, $\gamma_n = -1/n$, and $\gamma_i = 0$ for $1 < i < n$].

Table 2-2 shows that preliminary data errors increase the root mean square error in naive projections of GNP by 12 to nearly 40 per cent (line 1, columns 7–9). Of the four major sectors, errors in the early estimates of government expenditures affect forecast accuracy the least. The greatest reductions in accuracy are produced by errors in the detailed components of personal consumption expenditures and of gross private domestic investment data.

TABLE 2-2. Effect of Data Errors on the Accuracy of Three Naive Model Projections of Annual Levels of Gross National Product and Its Components, 1953–63

Line	Variable	Root Mean Square Error of: [a]						Relative Root Mean Square Error [b]		
		Naive Model N1		Naive Model N2		Naive Model N3				
		$\sqrt{M}$	$\sqrt{M'}$	$\sqrt{M}$	$\sqrt{M'}$	$\sqrt{M}$	$\sqrt{M'}$	$\sqrt{RM'_1}$	$\sqrt{RM'_2}$	$\sqrt{RM'_3}$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	Gross national product	28.4	22.9	20.6	18.4	16.4	11.9	1.240	1.120	1.378
2	Personal consumption expenditures	13.0	12.3	11.6	10.1	7.9	6.2	1.057	1.148	1.274
3	Durables	6.3	4.0	8.8	6.0	5.2	3.8	1.575	1.467	1.368
4	Nondurables	3.4	4.6	6.5	8.9	3.9	1.7	.739	.730	2.294
5	Services	5.9	6.4	4.5	2.4	2.9	1.0	.922	1.875	2.900
6	Gross private domestic investment	11.2	7.6	17.2	15.4	9.9	7.4	1.474	1.117	1.338
7	Producers durable equipment	4.4	2.7	5.4	6.9	4.7	2.6	1.630	.783	1.808
8	New construction	5.4	2.6	5.5	2.9	3.9	2.3	2.077	1.896	1.696
9	Change in business inventories	5.4	3.8	21.3	17.9	5.7	4.0	1.421	1.190	1.425
10	Gov't. expend. on goods and services	5.6	5.7	4.8	5.3	6.0	5.5	.982	.906	1.091
11	Federal	4.0	4.0	3.9	5.5	5.6	5.7	1.000	.709	.982
12	State and local	2.1	3.0	1.7	1.8	1.0	0.7	.700	.944	1.428
13	Net exports	2.5	1.9	2.8	2.9	2.7	2.4	1.316	.966	1.125

[a] In billion dollars. See text for a description of the naive models N1, N2, and N3; see Table 2-1, note a, for sources of data.

[b] Relative root mean square error is defined $\sqrt{RM'_i} = \sqrt{\frac{M_i}{M'_i}}$, where i denotes the naive model.

The data errors do not always increase forecast error. For some variables (line 4 or line 10, for example), the bias arising from the data offsets the bias in projections.[8] On the whole, however, Table 2-2 shows that data errors reduce the accuracy of the naive models, particularly that of the simple trend projection, N3. The root mean square errors of this projection were increased by an average of 55 per cent.

III. EFFECT OF DATA ERRORS ON THE ACCURACY OF BUSINESS FORECASTS

FORECASTS OF ANNUAL LEVELS

Naive projections can be considered scientific forecasts in the sense that the models generating the predictions are specified and the projections can be replicated. The exact weights (γ_i) that naive models assign to past values of the series are known. It was, therefore, possible to compute directly the element of forecast error that can be traced to data errors.

The effect of data errors on the accuracy of business forecasts, however, must be determined indirectly. The models underlying the forecasts in the Zarnowitz sample are not explicitly specified and it is necessary to infer their dependence on preliminary GNP statistics. Some of the forecasts may rely primarily on extrapolations; others may use them hardly at all. Mincer and Zarnowitz found that the patterns of observed forecast errors are consistent with a hypothesis that forecasters tend to be selective and use extrapolations when they provide the greatest advantage: Forecasts of fairly smooth and strongly serially correlated series rely more on extrapolations than do forecasts of somewhat volatile series. For example, forecasts of consumption expenditures tend to rely more on extrapolations than do investment forecasts.

It might be tempting, therefore, to predict that errors in the preliminary statistics would reduce the accuracy of consumption more than that of investment forecasts. Other things being equal, this prediction would be correct. But the "other things" are, according to (15),

[8] Note from equations (12) and (13) above that $M' > M$ if $[E(u') + \Sigma\gamma_i E(\epsilon)]^2 < [E(u')]^2$, or, in other words, if $E(u')$ and $\Sigma\gamma_i E(\epsilon)$ are of opposite sign.

the relative magnitudes of the data errors (k^2) and the efficiency of the pure forecast ($\rho^2_{AP'}$). Even though consumption forecasts may rely more on preliminary statistics than do forecasts of more volatile GNP components, the errors in the aggregate consumption data tend to be smaller (as shown by the k-ratios in Table 2-1). The effect on the forecast error depends not only on the importance of extrapolations to the forecast but on the size of the data errors as well.

In order to determine the effect of data errors on the accuracy of business forecasts, let us assume that these forecasts contain an extrapolative component and that it is a linear combination of past values of the target series as expressed in equation (1) above. The forecast error (u) would then consist of the pure forecast error (u') and the error ($\Sigma\gamma_i\epsilon_{t-i}$) induced by errors in the preliminary data, as shown in equation (3) above.

Provided u' and ϵ_{t-i} are uncorrelated, a least squares regression of u_t on past data errors decomposes u into its two components:

$$u_t = \gamma_1\epsilon_{t-1} + \gamma_2\epsilon_{t-2} + \gamma_3\epsilon_{t-3} + \cdots + u'_t, \quad \text{or} \tag{20}$$

$$u_t = \gamma_0 + \gamma_1\epsilon_{t-1} + \gamma_2\epsilon_{t-2} + \gamma_3\epsilon_{t-3} + \cdots + w_t,$$

where $u'_t = \gamma_0 + w_t$. If the forecast contains an extrapolative component (i.e., if the weights γ_i are not zero), then data errors would be a source of error in the forecast and they would account for part of its variability. The regression intercept (γ_0) provides an estimate of the mean error, $\overline{u'}$, and the residual variance [$\sigma^2(w)$], an estimate of $\sigma^2(u')$. Regression (20) can thus be used to determine whether or not data errors are a component of business forecast errors. If they are, they can be a source of bias and will reduce the forecast's efficiency.

Regressions of forecast errors on past data errors were computed for sixteen forecasts from the Zarnowitz sample.[9] Table 2-3 shows error statistics for these forecasts and the corresponding regression

[9] The forecasts are from the eight different sets of business forecasts, denoted by eight capital letters, A–H. They are a subset of the records of several hundred forecasts which were assembled for the NBER study of short-term economic forecasting. The particular variables predicted differ from one forecast set to another; however, all eight sets include forecasts of GNP and two of them (B and F) include forecasts of the major GNP components. For purposes of illustration, only the eight GNP forecasts and eight GNP component forecasts are used here. Though a somewhat different subset of forecasts is used by Mincer and Zarnowitz in Chapter 1 of this volume, we both include the GNP forecasts of sets E, F, and G and the personal consumption forecasts of set F.

TABLE 2-3. Estimates of the Effect of Data Errors on the Accuracy of Forecasts of Annual Levels of Gross National Product and Its Major Components, 1953–63

Line	Code of Forecast [a] and Period Covered	Forecast Error (billion dollars)			Regression Estimates of Pure Forecast Error [b] (billion dollars)				Tests of Bias and Efficiency [c]				
									t-test for				Adjusted
		$\bar{u}$	S_u	$\sqrt{M}$	$\bar{u}'$	$S_{u'}$	$\sqrt{M'}$	$\sqrt{RM'}$	$E(u)=0$	$E(u')=0$	$r_{P'u}$	$r_{P'u'}$	$R^2_{u\cdot\epsilon_{t-1},\cdots,\epsilon_{t-k}}$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Gross National Product												
1	Set A, 1954–63	−19.1	11.6	22.1	−23.0	8.8	24.5	.902	5.20 *	2.80 *	.473	.417	.432
2	Set B, 1953–63	−13.8	11.2	17.5	−7.3	6.5	9.7	1.804	4.09 *	1.42	.061	−.091	.659 *
3	Set C, 1958–63	−14.5	11.0	17.7	14.5	9.5	16.9	1.047	3.23 *	0.52	.179	.073	.260
4	Set D, 1956–63	−18.5	9.9	20.7	3.5	6.5	7.0	2.957	5.27 *	0.19	−.012	−.305	.569 *
5	Set E, 1953–63	−21.8	13.5	25.3	−13.8	7.7	15.6	1.622	5.35 *	2.28 *	.240	.210	.677 *
6	Set F, 1953–63	−15.8	8.8	18.0	−6.6	7.6	9.8	1.837	5.95 *	1.00	.315	.257	.264
7	Set G, 1953–63	−9.2	7.9	11.0	−8.2	7.4	10.8	1.018	3.88 *	1.30	.684 *	.778 *	.122
8	Set H, 1954–63	−19.7	10.5	22.1	−24.6	6.1	25.3	.873	5.95 *	4.59 *	.432	.408	.668 *
	Personal Consumption Expenditures												
9	Set B, 1953–63	−7.3	6.4	9.5	2.1	4.7	4.9	1.939	3.82 *	0.62	.338	−.045	.460 *
10	Set F, 1953–63	−9.1	5.0	10.2	4.1	4.5	5.9	1.729	6.09 *	0.50	.550 *	.582 *	.181
	Gross Private Domestic Investment												
11	Set B, 1953–63	−6.9	6.7	9.4	−8.2	3.6	8.9	1.056	3.40 *	3.65 *	.264	−.288	.718 *
12	Set F, 1953–63	−5.5	5.0	7.3	−10.8	3.6	11.3	.646	3.61 *	2.70 *	.372	.473	.494 *
	Gov't. Expend. on Goods and Services												
13	Set B, 1953–63	2.2	3.2	3.8	1.4	3.2	3.3	1.152	2.35 *	1.13	.015	.203	−.007
14	Set F, 1953–63	1.2	2.5	2.7	1.3	2.6	2.8	.964	1.56	1.27	−.156	.171	−.105
	Net Exports												
15	Set B, 1953–63	−1.9	2.2	2.9	−0.7	2.2	2.3	1.261	2.82 *	0.54	.497	.386	−.001
16	Set F, 1953–63	−2.5	1.6	2.9	−1.4	1.7	2.1	1.381	5.01 *	0.70	−.015	.116	−.070

Note: * denotes significance at the 10 per cent level.

[a] For a description of the forecasts, see Zarnowitz [6] and footnote 9 above.

[b] Error statistics are from regression (20) in text:

$$u_t = \gamma_0 + \gamma_1\epsilon_{t-1} + \gamma_2\epsilon_{t-2} + \gamma_3\epsilon_{t-3} + \gamma_4\epsilon_{t-4} + w_t.$$

$\bar{u}'$ is estimated as the regression intercept γ_0; $S_{u'}$, as the adjusted standard error of estimate, S_w; and $\sqrt{M'}$ is computed as $\sqrt{\gamma_0^2 + (n-1)/n\, S_w^2}$, where n is the number of observations. A step-wise regression was used and the number of lags (up to 4) that produced the maximum adjusted R^2 was used. The statistics for: lines 1, 5, 7, 8, 10, and 12 are based on 4 lags; lines 2–4 are based on 3 lags; line 11 is based on 2 lags; and lines 6, 9, 13–16 are based on 1 lag.

[c] The correlation coefficient $r_{P'u'}$ is estimated as $r_{P'u|\epsilon_{t-1},\cdots,\epsilon_{t-k}}$, where k is the number of lags used in the regressions above.

estimates of $\overline{u'}$, $\sigma^2(u')$, and M'. In addition, the table gives the results of tests for forecast bias and efficiency. The samples are small and sampling variation alone could produce a relation between forecast errors and data errors, as well as nonzero mean errors or nonzero correlations between forecasts and their errors. Tests of significance are therefore indicated.

There is some relation between forecast errors and data errors for each of the twelve sets of GNP, consumption, and investment forecasts, but judging by the coefficients of determination ($R^2_{u \cdot \epsilon_{t-1}, \ldots, \epsilon_{t-k}}$ in column 12), it is significant (at the 10 per cent level) in only seven cases. In these seven sets, however, data errors account for about 65 per cent of $\sigma^2(u)$. No relation is indicated between the errors in forecasts and errors in the preliminary data for government expenditures and net exports (column 12, lines 13–16).[10]

We have seen that though data errors would reduce the efficiency of P by augmenting the random component of $\sigma^2(u)$, they could also affect the systematic component. This is because data errors would create a positive correlation between P and u. Thus, in general, we would expect the correlation coefficient r_{Pu} to be greater than $r_{P'u'}$. If data errors were the only source of inefficiency, $r_{P'u'}$ would be zero. The partial correlation coefficient, $r_{P_t u_t \cdot \epsilon_{t-1}, \ldots, \epsilon_{t-k}}$, which holds the effect of data errors constant, provides an estimate of $r_{P'u'}$.

The two correlation coefficients, r_{Pu} and $r_{P'u'}$, are given in columns 10 and 11 of Table 2-3. A comparison shows that r_{Pu} exceeds $r_{P'u'}$ for six of the seven forecasts in which data errors comprise a significant element of forecast error. However, the correlations that data errors create between P and u are not strong, and we could conclude that data errors reduced the efficiency of P primarily by increasing the random component of $\sigma^2(u)$ rather than by increasing its slope component.

The hypothesis that each of the seven forecasts affected by data errors is unbiased would be rejected at the 10 per cent level in every case (column 8). This hypothesis for the forecasts net of data errors (P') would be rejected for four sets (column 9). Data errors, therefore, could be considered the only source of bias in three of the seven sets.

[10] Data errors are measured by revisions. Definitional revisions, however, have not been excluded from the revisions of government expenditures and net exports (see Table 2-1, note a). It is possible, therefore, that the definitional revisions obscure any relation that may exist between forecast errors and statistical data errors for these variables.

To sum up, the results in Table 2-3 suggest that at least seven of the sixteen forecast sets relied on extrapolations.[11] Errors in the data used to construct the extrapolations reduced the efficiency of these seven forecasts. The reduction was considerable: It is estimated that data errors account for 50 to 70 per cent of $\sigma^2(u)$. All seven forecasts are biased, and data errors could be considered the *primary* source of the bias in three.

FORECASTS OF ANNUAL CHANGES

Errors in forecasts of changes would be exactly the same as errors in forecasts of levels, except for the presence of data errors. The error of a forecast of the change in series A from year $t-1$ to year t is defined

$$(21) \quad u_{\Delta t} = (P_t - A^\circ_{t-1}) - (A_t - A_{t-1}) = (P_t - A_t) - (A^\circ_{t-1} - A_{t-1})$$
$$= u_t - \epsilon_{t-1} = u'_t + \Sigma\gamma_i\epsilon_{t-i} - \epsilon_{t-1},$$

where A°_{t-1} is the forecast base.[12] If there were no data errors, the error of both the change and level forecast would be u'_t.

Provided as before that u' and ϵ_{t-i} are uncorrelated, a least squares regression of the observed forecast error (u_Δ) on past data errors (ϵ_{t-i}) decomposes u_Δ into its two components:

$$(22) \quad u_{\Delta t} = u_t - \epsilon_{t-1} = \gamma_0 + (\gamma_1 - 1)\epsilon_{t-1} + \gamma_2\epsilon_{t-2} + \gamma_3\epsilon_{t-3} + \cdots + w_t,$$

where $u'_t = \gamma_0 + w_t$.

Since (22) is simply a linear transformation of (20), the regression equation for levels, all of the coefficients (except the coefficient of

[11] These results suggest less widespread use of extrapolations than that found by Mincer and Zarnowitz. This is because the method that they used to decompose forecasts into extrapolative and "autonomous" components, a regression of P_t on past values of A°, estimates the autonomous component as a residual. Thus, as they point out, the autonomous component would include only that element of the forecast that is statistically independent of past values of A°. If the forecast in fact relied on variable B, which is correlated with A°_{t-i}, its importance to the forecast would be attributed to A°_{t-i}. It is unlikely, however, that measurement errors in B would also be correlated with the measurement errors in A°_{t-i}. Therefore, a regression of the forecast *error* on the *errors* in past values of A° would be unlikely to associate B with the extrapolative component of the forecast.

[12] For the sake of simplicity, it is assumed that the forecaster's estimate of the current level of the series, the forecast base, equals the first official estimate, A°_{t-1}. This is a reasonable assumption: the forecaster's estimate of the base year is an average of the official estimates of the first three quarters and the forecaster's estimate of the fourth quarter.

ϵ_{t-1}), their standard errors, and the residual variance will be the same as in (20). The only difference would occur in the coefficient of determination. Generally one might expect this coefficient to be larger for levels than for changes because its denominator, the variance of the forecast error, would be expected to be smaller for levels than for changes. In other words, base errors (ϵ_{t-1}) and level errors (u_t) might be expected to be uncorrelated and, if so, $\sigma^2(u_\Delta) = \sigma^2(u) + \sigma^2(\epsilon)$.

However, it is clear from (21) that the presence of data errors in the forecast would create a positive correlation between u_t and ϵ_{t-1}, and as a result,

$$\sigma^2(u_\Delta) \gtreqless \sigma^2(u) \quad \text{as} \quad r_{u_t\epsilon_{t-1}} \lesseqgtr \frac{1}{2}\frac{\sigma(\epsilon)}{\sigma(u)}. \tag{23}$$

This can be expressed differently: Since

$$\sigma^2(u_\Delta) = \sigma^2(u) + \sigma^2(\epsilon_{t-1}) - 2\mathrm{Cov}(u_t,\epsilon_{t-1}) \tag{24}$$

$$= \sigma^2(u) + (1 - 2\gamma_1 - 2\sum_{i=2} \gamma_i \rho_{\epsilon_{t-i},\epsilon_{t-1}})\sigma^2(\epsilon),$$

provided $\mathrm{Cov}\,(u'_t,\epsilon_{t-i}) = 0$ and $\sigma^2(\epsilon_{t-i}) = \sigma^2(\epsilon)$ for all i, then even if there were no serial correlation in ϵ (i.e., $\rho_{\epsilon_{t-i},\epsilon_{t-1}} = 0$), as long as γ_1, the weight that the forecast assigns to the forecast base (A°_{t-1}, the last known value of the series), exceeds $\frac{1}{2}$, data errors would be a smaller component of predicted change than of predicted level errors and $\sigma^2(u_\Delta)$ would be smaller than $\sigma^2(u)$.

Table 2-4 shows error statistics for forecasts of annual change in GNP and its major components similar to those given in Table 2-3 for level forecasts. A comparison shows that data errors were indeed a smaller component of predicted change errors than of level errors (column 7 and 12 in Table 2-4 compared with columns 7 and 12 in Table 2-3).

Comparison of columns 8 through 11 in Table 2-4 with the corresponding columns in Table 2-3 shows a striking difference between the characteristics of change and level forecast errors. Though both show systematic, or potentially reducible, error, it takes a different form: Change forecasts show much less bias than level forecasts, but they tend to be inefficient (in the sense that the predicted changes are correlated with their errors) whereas level forecasts do not. It might be

TABLE 2-4. Estimates of the Effect of Data Errors on the Accuracy of Forecasts of Annual Changes in Gross National Product and Its Major Components, 1953–63

Line	Code of Forecast [a] and Period Covered	Forecast Error (billion dollars)			Regression Estimates of Pure Forecast Error [b] (billion dollars)				Tests of Bias and Efficiency [c]				
									t-test for				
		$\bar{u}$	S_u	$\sqrt{M}$	$\bar{u}'$	$S_{u'}$	$\sqrt{M'}$	$\sqrt{RM'}$	$E(u)=0$	$E(u')=0$	r_{Pu}	$r_{P'u'}$	Adjusted $R^2_{u\cdot e_{t-1},\cdots,e_{t-k}}$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
						Gross National Product							
1	Set A, 1954–63	−7.8	11.7	13.6	−23.0	8.8	24.5	.555	2.11 *	2.80 *	.730 *	.736 *	.441
2	Set B, 1953–63	−4.2	8.8	9.8	−7.3	6.5	9.7	1.010	1.76	1.42	.436	.246	.452 *
3	Set C, 1958–63	−2.7	10.1	9.6	14.5	9.5	16.9	.568	0.66	0.52	.722 *	.628 *	.115
4	Set D, 1956–63	−6.2	8.1	9.7	3.5	6.5	7.0	1.386	2.16 *	0.19	.487	.446	.357
5	Set E, 1953–63	−11.4	10.2	15.0	−13.8	7.7	15.6	.962	3.71 *	2.28 *	.483	.262	.435
6	Set F, 1953–63	−5.2	8.4	9.5	−6.6	7.6	9.8	.969	2.05 *	1.00	.726 *	.782 *	.185
7	Set G, 1953–63	0.1	8.9	9.0	−8.2	7.4	10.8	.833	0.04	1.30	.684 *	.602 *	.237
8	Set H, 1954–63	−8.9	10.1	13.1	−24.6	6.1	25.3	.518	2.79 *	4.59 *	.672 *	.607 *	.642 *
						Personal Consumption Expenditures							
9	Set B, 1953–63	−2.4	4.9	5.3	2.1	4.7	4.9	1.082	1.62	0.62	.548 *	.585 *	.102
10	Set F, 1953–63	−3.6	4.7	5.8	4.1	4.5	5.9	.983	2.54 *	0.50	.850 *	.858 *	.101
						Gross Private Domestic Investment							
11	Set B, 1953–63	−2.2	6.6	6.6	−8.2	3.6	8.9	.742	1.11	3.65 *	.405	.371	.707 *
12	Set F, 1953–63	−1.4	5.1	5.1	−10.8	3.6	11.3	.451	0.91	2.70 *	.358	.556	.513 *
						Gov't. Expend. on Goods and Services							
13	Set B, 1953–63	0.7	3.1	3.1	1.4	3.2	3.3	.939	0.75	1.13	.274	−.377	−.036
14	Set F, 1953–63	0.3	2.2	2.1	1.3	2.6	2.8	.750	0.45	1.27	−.322	−.356	−.062
						Net Exports							
15	Set B, 1953–63	−0.2	1.7	1.6	−0.7	2.2	2.3	.696	0.39	0.54	−.532 *	−.624 *	.013
16	Set F, 1953–63	−0.4	1.6	1.6	−1.4	1.7	2.1	.762	0.83	0.70	−.236	−.268	.081

Note: * denotes significance at the 10 per cent level. See notes to Table 2-3.

supposed that data errors are the primary source of the correlations between P and u, but the correlations remain when data errors are held constant (columns 10 and 11 of Table 2-4).

As Mincer and Zarnowitz showed in the preceding chapter, however, the criteria for efficient change forecasts are more stringent than those for levels. Consider the regression

$$A_t - A_{t-1} = \alpha_\Delta + \beta_\Delta(P_t - A^\circ_{t-1}) + v'_t. \tag{25}$$

A change forecast is efficient if $\beta_\Delta = 1$. The coefficient is, by definition,

$$\beta_\Delta = \frac{\text{Cov}\,(A_t - A_{t-1},\, P_t - A^\circ_{t-1})}{\sigma^2(P_t - A^\circ_{t-1})}. \tag{26}$$

Using the identity $(A_t - A_{t-1}) \equiv (P_t - A^\circ_{t-1}) - u_{\Delta t}$, $\beta_\Delta = 1$ only if Cov $(P_t - A^\circ_{t-1}, u_{\Delta t}) = 0$. Since

$$1 - \beta_\Delta = \frac{\text{Cov}\,(P_t - A^\circ_{t-1}, u_{\Delta t})}{\sigma^2(P_t - A^\circ_{t-1})} = \tag{27}$$

$$\frac{(1 - \Sigma\gamma_i\rho_{\epsilon_{t-i},\epsilon_{t-1}})\sigma^2(\epsilon) + \text{Cov}\,(u'_t, P'_t) - \text{Cov}\,(u'_t, A_{t-1})}{\sigma^2(P_t - A^\circ_{t-1})},$$

a change forecast would be efficient if there were no data errors and if both $r_{u'_tP'_t}$ and $r_{u'_tA_{t-1}}$ are zero. The last requirement means that the forecast should utilize the extrapolative potential of the series. Otherwise, the forecast error could be reduced by taking account of the last known value of the series. The estimate of this value that would be available for the forecast is the preliminary estimate A°_{t-1}. Thus, in some respects the forecaster is in a box: Full use of A°_{t-1} would transfer its error to the forecast, but failure to do so would also result in an inefficient forecast.

The data in Table 2-4 suggest that forecasting efficiency was impaired much less by data errors than by failure to use the forecasting potential of the series. Eight of the forecasts of change are inefficient before as well as after the effect of data errors is taken into account (column 10 compared with column 11). Since six of these forecasts were not considered inefficient level forecasts (i.e., $r_{P'u'}$ did not differ significantly from zero in Table 2-3, column 10), we could conclude that the change forecasts were inefficient because they did not make effective use of A°_{t-1} (i.e., $r_{u'_tA_{t-1}} \neq 0$).

Failure of forecasts to use A°_{t-1} effectively could cause the variance of predicted changes to exceed that of actual changes. Efficient use of A°_{t-1} requires the forecast to take account of the serial correlation in A, such that $r_{P'_tA_{t-1}} = r_{A_tA_{t-1}}$. The variance of the predicted change is $\sigma^2(P_t - A^{\circ}_{t-1}) = \sigma^2(P'_t - A_{t-1}) + (\Sigma\gamma_i^2 + 1 - 2\Sigma\gamma_i\rho_{\epsilon_{t-i},\epsilon_{t-1}})\sigma^2(\epsilon)$. Since $\sigma^2(P'_t - A_{t-1}) = \sigma^2(P'_t) + \sigma^2(A_{t-1}) - 2r_{P'_tA_{t-1}}\sigma(P'_t)\sigma(A_{t-1})$,

$$(28)\quad \sigma^2(P_t - A^{\circ}_{t-1}) = \sigma^2(P'_t) + \sigma^2(A_{t-1}) - 2r_{P'_tA_{t-1}}\sigma(P'_t)\sigma(A_{t-1}) + (\Sigma\gamma_i^2 + 1 - 2\Sigma\gamma_i\rho_{\epsilon_{t-i},\epsilon_{t-1}})\sigma^2(\epsilon).$$

The variance of the actual change is

$$(29)\quad \sigma^2(A_t - A_{t-1}) = \sigma^2(A_t) + \sigma^2(A_{t-1}) - 2r_{A_tA_{t-1}}\sigma(A_t)\sigma(A_{t-1}).$$

Thus $\sigma^2(P_t - A^{\circ}_{t-1})$ could exceed $\sigma^2(A_t - A_{t-1})$ because of data errors, because $\sigma^2(P')$ exceeds $\sigma^2(A)$, or because the forecast did not make sufficient use of the serial correlation in A (i.e., $r_{P'_tA_{t-1}} < r_{A_tA_{t-1}}$).

Table 2-5 compares the variance of level and change forecasts, both adjusted and unadjusted for data errors, with the variance of actual levels and changes. Regressions of forecast errors on past data errors were used to obtain estimates of the variance of forecasts net of data errors. That is, $\sigma^2(P')$ was estimated as

$$S^2_{P'} = S^2_P - R^2_{u\cdot\epsilon_{t-1},\ldots,\epsilon_{t-k}}S^2_u,$$

where S^2_P is the variance of the forecast (level or change); S^2_u, the variance of its error (level or change, respectively); and $R^2_{u\cdot\epsilon_{t-1},\ldots,\epsilon_{t-k}}$, the coefficient of determination in the regression of u on ϵ_{t-i}.

The variance of GNP, consumption, and investment forecasts exceeds that of the actual values, especially for the changes. Data errors were only in part responsible. Although they increased the variance of predicted levels by about 1 per cent and that of predicted changes by 5 to 25 per cent, the variance of the adjusted predictions in most cases exceeds that of actuals. Comparison of the two correlation coefficients [13] shows that $r_{P'_tA_{t-1}}$ is generally less than $r_{A_tA_{t-1}}$. Thus failure to exploit the extrapolative potential of the series is the main reason why the change forecasts are inefficient and why the variance of forecasts exceeds the variance of realizations.

[13] The partial correlation coefficient, $r_{P_tA^{\circ}_{t-1}\cdot\epsilon_{t-1},\ldots,\epsilon_{t-k}}$, which holds the effect of data errors constant, was used to estimate $r_{P'_tA_{t-1}}$.

TABLE 2-5. Estimates of the Effect of Data Errors on the Variability of Forecasts of Annual Levels and Changes in Gross National Product and Its Major Components, 1953–63 (*dollars in billions*)

Line	Code of Forecast and Period Covered	Level Forecasts			Change Forecasts				
		S_P (1)	$S_{P'}$ (2)	S_A (3)	S_P (4)	$S_{P'}$ (5)	S_A (6)	$r_{P_tA_{t-1}}$ (7)	$r_{A_tA_{t-1}}$ (8)
	Gross National Product								
1	Set A, 1954–63	78.1	77.4	73.3	18.7	17.0	12.9	.976	.986
2	Set B, 1953–63	77.0	76.4	77.1	13.2	11.9	12.3	.986	.988
3	Set C, 1958–63	53.4	53.2	52.6	18.3	17.9	13.0	.926	.973
4	Set D, 1956–63	62.4	62.0	60.9	13.7	12.8	11.1	.980	.985
5	Set E, 1953–63	79.3	78.5	77.1	13.4	11.0	12.3	.989	.989
6	Set F, 1953–63	79.5	79.3	77.1	17.0	16.6	12.3	.979	.989
7	Set G, 1953–63	82.3	82.2	77.1	16.6	15.8	12.3	.989	.989
8	Set H, 1954–63	77.2	76.8	73.3	17.4	15.4	12.9	.979	.986
	Personal Consumption Expenditures								
9	Set B, 1953–63	48.9	48.7	49.5	5.4	5.1	4.9	.997	.996
10	Set F, 1953–63	52.1	42.0	49.5	8.3	8.1	4.9	.990	.996
	Gross Private Domestic Investment								
11	Set B, 1953–63	12.5	11.2	11.1	7.1	4.4	7.4	.804	.764
12	Set F, 1953–63	12.0	11.4	11.1	7.6	6.6	7.4	.748	.764
	Gov't. Expend. on Goods and Services								
13	Set B, 1953–63	16.4	16.4	16.7	4.0	4.0	4.8	.983	.963
14	Set F, 1953–63	16.1	16.1	16.7	3.6	3.6	4.8	.981	.963
	Net Exports								
15	Set B, 1953–63	2.0	2.0	2.1	0.7	0.7	2.2	.932	.465
16	Set F, 1953–63	1.4	1.4	2.1	1.1	1.1	2.2	.816	.465

Note: See notes to Table 2-3.

Data errors then were not the only source of systematic error in the business forecasts examined here: Though they could be considered the primary source of bias in three of the sixteen forecast sets, in no case were they the only source of the "slope component" of forecast errors. Data errors mainly impaired forecasting accuracy by augmenting the random component of the variance of the forecast error.

ALTERNATIVE ESTIMATES

Since the method of least squares necessarily yields an estimate of the maximum amount of the variation in u that is statistically related to data errors, and hence a minimum estimate of $\sigma^2(u')$, the regression estimates may overstate the effect of data errors on forecasting efficiency. This would be the case if u' were correlated with ϵ_{t-i}. A comparison of the regression estimates with alternative estimates of u' is therefore worthwhile.

An obvious alternative and one that is commonly used is the observed forecast error u°. It was argued earlier (Section I) that if ϵ is serially correlated, the error in realizations data (ϵ_t) would tend to offset the data errors ($\Sigma\gamma_i\epsilon_{t-i}$) transmitted to the forecast through its extrapolative component. Indeed, if $\Sigma\gamma_i = 1$, $E(u^\circ) = E(u')$. The extent to which ϵ_t would offset the loss in forecast efficiency depends on the strength of the serial correlation in ϵ. If the correlation were perfect (and positive), $\sigma^2(u^\circ) = \sigma^2(u')$, since

$$(30)\quad \sigma^2(u^\circ) = \sigma^2(u'_t + \Sigma\gamma_i\epsilon_{t-i} - \epsilon_t)$$
$$= \sigma^2(u'_t) + \sigma^2(\Sigma\gamma_i\epsilon_{t-i}) + \sigma^2(\epsilon_t) - 2\ \mathrm{Cov}\ (\Sigma\gamma_i\epsilon_{t-i}, \epsilon_t)$$
$$+ 2\ \mathrm{Cov}\ (u'_t, \Sigma\gamma_i\epsilon_{t-i}) - 2\ \mathrm{Cov}\ (u'_t, \epsilon_t).$$

In general, $\Sigma\gamma_i \neq 1$ and u° is not a very satisfactory estimate of u'. For example, if the bias in A° were in the same (opposite) direction as the bias in P', $E(u^\circ)$ would understate (overstate) $E(u')$ for those forecasts in which $\Sigma\gamma_i < 1$. Moreover, ϵ is not perfectly serially correlated and $\sigma^2(u^\circ)$ would generally exceed $\sigma^2(u')$. Indeed, $\sigma^2(u^\circ)$ would exceed $\sigma^2(u)$ for those forecasts which do not rely on extrapolations at all ($\gamma_i = 0$). Since $\sigma^2(u^\circ)$ could overestimate, and the regression estimates could underestimate, the two would bracket $\sigma^2(u')$. The alternative estimates based on u° are given in Table 2-6.

As expected, S_{u° exceeds the regression estimate of $S_{u'}$ for all of

TABLE 2-6. Alternative Estimates of the Effect of Data Errors on the Accuracy of Forecasts of Annual Levels of Gross National Product and Its Major Components, 1953–63 (*dollars in billions*)

Line	Code of Forecast and Period Covered	Alternative Estimates of Pure Forecast Error				
		$\bar{u}^\circ$ (1)	S_{u° (2)	$\sqrt{M^\circ}$ (3)	$\sqrt{RM^\circ}$ (4)	t-test for $E(u^\circ) = 0$ (5)
		Gross National Product				
1	Set A, 1954–63	−7.1	13.5	12.5	1.768	1.68
2	Set B, 1953–63	−2.8	10.8	10.7	1.636	0.87
3	Set C, 1958–63	−3.2	11.6	11.0	1.609	0.67
4	Set D, 1956–63	−6.9	9.7	11.4	1.816	2.01 *
5	Set E, 1953–63	−10.8	13.4	16.7	1.515	2.68 *
6	Set F, 1953–63	−4.8	7.8	8.8	2.045	2.07 *
7	Set G, 1953–63	1.8	8.1	7.9	1.392	0.73
8	Set H, 1954–63	−7.8	9.7	12.0	1.842	2.52 *
		Personal Consumption Expenditures				
9	Set B, 1953–63	−1.3	5.6	5.7	1.667	0.76
10	Set F, 1953–63	−3.0	4.8	5.5	1.855	2.06 *
		Gross Private Domestic Investment				
11	Set B, 1953–63	−2.7	6.9	7.1	1.323	1.32
12	Set F, 1953–63	−1.3	4.4	4.4	1.659	1.00
		Gov't. Expend. on Goods and Services				
13	Set B, 1953–63	1.3	2.1	2.3	1.652	2.01 *
14	Set F, 1953–63	0.2	1.6	1.6	1.688	0.43
		Net Exports				
15	Set B, 1953–63	−0.1	1.8	1.7	1.706	0.22
16	Set F, 1953–63	−0.7	1.5	1.6	1.812	1.53

Note: * denotes significance at the 10 per cent level. See Table 2-3, note a, for source.

the forecasts of GNP, consumption, and investment expenditures (compare column 2, Table 2-6, with column 5, Table 2-3). The opposite relation, however, holds for forecasts of government expenditures and net exports and it is probably due to the fact that the errors in both the preliminary data and the forecasts primarily reflect definitional revisions (see footnote 10 above).[14] In most cases in which the regressions indicated no significant relation between past data errors and forecast errors, S_{u° exceeds S_u.

The estimates based on u° suggest that P' is biased downward. However, u° may underestimate the size of the bias in P' and therefore overestimate the importance of data errors as a source of bias in P. The hypothesis that P is unbiased would be rejected at the 10 per cent level for fifteen of the sixteen forecasts shown in Table 2-3 (column 8). This hypothesis for P' would be rejected for only six of the forecasts in Table 2-6 (column 5). Thus, when u° is used to estimate u', data errors would be considered the only source of bias in nine of the forecasts examined. The regression estimates, however, indicated that data errors were solely responsible for the bias in only three sets.

When there is a significant relation between forecast errors and past data errors, the regression estimates attribute a somewhat larger fraction of forecast error to data errors than that suggested by the alternatives based on u° (column 7, Table 2-3 compared with column 4, Table 2-6). These differences, however, are relatively small. The greatest differences occur for the nine forecasts that the regression estimates indicate were unaffected by data error. The alternative estimates suggest that data errors reduced the accuracy of these forecasts by an average of 70 per cent! This huge overestimation occurs because bias is a very large component of the over-all forecast error (M) and u° overstates the bias arising from data errors.[15] Thus, even though

[14] More explicitly, the definitional revisions would create a positive correlation between u_t and ϵ_t and therefore reduce the variance of u°. Since the regressions for these variables indicated γ_i not different from zero, $u_t = u'_t$ and

$$\sigma^2(u^\circ) = \sigma^2(u) + \sigma^2(\epsilon) - 2 \text{ Cov } (u, \epsilon).$$

[15] This is readily shown. The regressions for these nine forecasts did not show a relation between forecast errors (u_t) and past data errors (ϵ_{t-i}) strong enough to reject the null hypothesis $\gamma_i = 0$. Thus equation (16) would become

$$(16') \qquad u_t^\circ = P_t - A_t^\circ = u_t - \epsilon_t = u'_t - \epsilon_t,$$

and it follows that $E(u) = E(u')$, but $E(u^\circ) = E(u') - E(\epsilon)$.

the regression estimates may overstate the effect of data errors, they are preferable to the alternatives. The regression estimates permit a test for the presence of data errors; the alternative estimates indiscriminately adjust for data errors, whether they were incorporated into the forecast or not.

IV. EFFECT OF DATA ERRORS ON THE ACCURACY OF AN ANALYTICAL MODEL OF CONSUMPTION

The use of preliminary rather than final data affects not only the values of the variables underlying a forecast, it affects the estimates of the parameters of relationships among these variables as well. Thus far the indirect effects on forecasting accuracy of errors in the data used to estimate the parameters of the forecast model have not been considered. There are no indirect effects on naive models—their parameters are not estimated but set arbitrarily—and they could not be determined for business forecasts because the forecasting models are not explicitly specified. The backbone of many GNP models, however, is a consumption function of one kind or another, and one is therefore used in this section to illustrate the total effect (indirect as well as direct) on predictive accuracy of errors in the underlying data.

EFFECT ON PARAMETER ESTIMATES

The consumption function chosen is one of the quarterly models first estimated by Zellner [8] and reestimated with revised data by Griliches *et al.* [3]. This function is

$$C_t = \alpha + \beta Y_t + \gamma C_{t-1} + v_t, \tag{31}$$

where C denotes personal consumption expenditures; Y, personal disposable income; and v, the residual.

The preliminary consumption (C°) and income (Y°) estimates are written

$$C^\circ = C + \epsilon(C) \quad \text{and} \quad Y^\circ = Y + \epsilon(Y), \tag{32}$$

where $\epsilon(C)$ and $\epsilon(Y)$ are errors in measuring C and Y, respectively. If preliminary data are used, equation (31) becomes

$$C_t^\circ = \alpha + \beta Y_t^\circ + \gamma C_{t-1}^\circ + v_t^\circ, \tag{31'}$$

where $$v_t^\circ = v_t + \epsilon(C)_t - \beta\epsilon(Y)_t - \gamma\epsilon(C)_{t-1}.$$

It is well known that the method of least squares applied to (31′) would yield biased estimates of the coefficients. The magnitude and direction of the bias would depend on the correlation between the explanatory variables ($r_{Y_tC_{t-1}}$) and on the relative magnitude of the data errors ($\lambda_C = \sigma^2[\epsilon(C)]/\sigma^2(C)$ and $\lambda_Y = \sigma^2[\epsilon(Y)]/\sigma^2(Y)$) as well as their intercorrelations.[16]

The following tabulation, where C and Y denote 1965 data, shows the relevant statistics for the sample periods used by Zellner and by Griliches *et al.*:

Data Used	Period Covered [a]	Error Statistics				
		λ_C	λ_Y	$r_{\epsilon(C)_t,t-1}$	$r_{\epsilon(C)_t\epsilon(Y)_t}$	$r_{\epsilon(Y)_t\epsilon(C)_{t-1}}$
Zellner						
Available in July 1955	1947-I–55-I	.024	.006	.960	.365	.342
Griliches et al.						
Available in Aug. 1961	1947-I–55-I	.013	.004	.890	.013	−.048
	1947-I–60-IV	.003	.002	.832	−.129	−.202

		Correlations Among Dependent and Independent Variables		
		$r_{C_tY_t}$	$r_{C_tC_{t-1}}$	$r_{Y_tC_{t-1}}$
1965 Revised Data				
Available in Aug. 1965	1947-I–55-I	.984	.994 [b]	.983 [b]
			.989 [c]	.981 [c]
	1947-I–60-IV	.996	.998 [b]	.995 [b]
			.997 [c]	.995 [c]

[a] Excluding 1950–III and 1951–I. [b] Based on Zellner method of excluding observations (see text below). [c] Based on Griliches *et al.* method of excluding observations.

Though there is strong, positive serial correlation in the consumption data errors, the relative magnitude of these errors, as well as that of the income errors, is small. Thus, in the absence of intercorrelation (i.e., $r_{Y_tC_{t-1}} = 0$), the data errors would have only negligible effects on the estimated coefficients. However, $r_{Y_tC_{t-1}}$ is close to unity, indicating

[16] For a detailed treatment of the effects of errors in the variables, see Johnston [4, Chapter 6].

strong multicollinearity. As a consequence, the effects of the errors would be substantially magnified. The strong, positive serial correlation in $\epsilon(C)$ would tend to bias c, the estimated coefficient of lagged consumption, upward. Because of the multicollinearity, serial correlation in $\epsilon(C)$ would also affect b, the estimated coefficient of current income, and tend to bias it downward.

On the whole these expectations are borne out in Table 2-7, which compares the coefficients obtained by Zellner with those obtained from revised data. The coefficients based on preliminary data tend to underestimate β and overestimate γ.

Zellner found that b was not significantly different from zero (line 1, column 2). Griliches *et al.* reestimated the coefficients from revised data. Their estimates differed substantially from Zellner's and reversed the conclusion for b (line 5, column 2). But as they point out [3, p. 494, note 6]:

> There is one minor difference between Zellner's and our way of computing the same equation. When Zellner leaves out an observation, e.g., 1951-I, in the next period the value of lagged consumption is taken to be that of two periods ago, whereas when we delete an "observation," we do not change the independent variables, and in 1951-II, C_{t-1} equals the actual C of 1951-I, even though this value itself does not appear in the series for the dependent variable. Whichever procedure is right depends on one's interpretation of why 1951-I is "out of line" and should be excluded in the first place.

As Table 2-7 shows, this small difference has a large effect on the coefficients. The Zellner method results in much lower estimates of the coefficient of Y_t and much higher estimates of the coefficient of C_{t-1} than those obtained using the method of [3]. Hence a simple comparison of the original Zellner coefficients (line 1) with the Griliches *et al.* coefficients (line 5) grossly overstates the effect of data errors. Indeed, if Griliches *et al.* had used the same method as Zellner, they too would have found the coefficient of current income lacking statistical significance (line 2, column 2), and if Zellner had used the Griliches *et al.* method, he would have found a statistically significant b (line 4, column 2). Estimates based on the 1965 data yield similar results (lines 3 and 6, column 2).

None of the coefficients estimated from data covering the longer period lacks statistical significance (lines 7–10). Moreover, it is worth noting that sets of considerably different coefficients are associated

TABLE 2-7 Coefficients of Zellner's Quarterly Consumption Function: Original Compared With Those Computed From Revised Data

Line	Method of Excluding Observations [a]	Constant Term (1)	Coefficients of Y_t (2)	Coefficients of C_{t-1} (3)	R^2 (4)	*LR MPC* [b] (5)
	Period Covered: 1947-I–1955-I, Excluding 1950-III and 1951-I					
	Zellner Method					
1	Zellner (July 1955 data)	0.1	.128 (.093)	.870 (.127)	.978	.98
2	Griliches *et al.* (August 1961 data)	0.6	.071 (.099)	.928 (.129)	.984	.99
3	1965 revised data	0.3	.168 (.093)	.827 (.106)	.989	.97
	Griliches *et al.* Method					
4	Zellner data	13.1	.296 (.115)	.618 (.158)	.959	.77
5	Griliches *et al.* data	15.4	.335 (.130)	.574 (.169)	.971	.79
6	1965 revised data	3.1	.318 (.111)	.650 (.126)	.982	.91
	Period Covered: 1947-I–1960-IV, Excluding 1950-III and 1951-I					
	Zellner Method					
7	Griliches *et al.* data	2.7	.185 (.068)	.796 (.078)	.996	.91
8	1965 revised data	0.4	.258 (.069)	.728 (.073)	.997	.95
	Griliches *et al.* Method					
9	Griliches *et al.* data	3.1	.300 (.085)	.670 (.097)	.994	.91
10	1965 revised data	−0.4	.330 (.081)	.652 (.086)	.996	.95

[a] The Zellner data are in 1947–49 dollars, the Griliches *et al.* data are in 1954 dollars, and the 1965 revised data are in 1958 dollars.

[b] *LR MPC* is the long-run marginal propensity to consume, computed by dividing the coefficient of current income by 1 minus the coefficient of lagged consumption.

with the same long-run marginal propensity to consume (line 7 compared with line 9 and line 8 compared with line 10).

To sum up, errors in the underlying data bias estimates of the coefficients of the consumption function in (31): The coefficient of current income was biased downward by about 25 per cent and the coefficient of lagged consumption was biased upward by about 7 per cent when the Zellner method of excluding observations is used. The biases are much smaller when the Griliches *et al.* method is used. The effect of data errors on the coefficients, however, was much weaker than the effect of a small difference in the method of excluding observations.

EFFECT ON PREDICTIVE ACCURACY

Strictly considered, the consumption function in (31) is not a forecasting model because it requires knowledge of the value of personal disposable income during the prediction period $t+1$. It could be used as one if a prediction of income $\hat{Y}_{t+1}$ were somehow obtained. Since we are interested in the effect of errors in the preliminary data, an obvious choice is a simple extrapolation of these data,

(33)
$$\hat{Y}^{\circ}_{t+1} = d_0 + d_1 Y^{\circ}_t + d_2 Y^{\circ}_{t-1} + \cdots = d_0 + \Sigma d_i Y_{t-i+1} + \Sigma d_i \epsilon(Y)_{t-i+1}.$$

The forecast, made in period t, of consumption in period $t+1$ ($\hat{C}^{\circ}_{t+1}$) would then be

(34)
$$\hat{C}^{\circ}_{t+1} = a + b\hat{Y}^{\circ}_{t+1} + cC^{\circ}_t,$$

where a, b, and c are the coefficients estimated from preliminary data.

If (31) were correctly specified, and if the true values of the variables are denoted C and Y, the value of consumption in period $t+1$ is

(35)
$$C_{t+1} = \alpha + \beta Y_{t+1} + \gamma C_t + v_{t+1}.$$

The error of the forecast is then defined

(36)
$$\hat{C}^{\circ}_{t+1} - C_{t+1} = (a-\alpha) + (b-\beta)Y_{t+1} + (c-\gamma)C_t + b(\hat{Y}^{\circ}_{t+1} - Y_{t+1}) + c(C^{\circ}_t - C_t) - v_{t+1}.$$

In addition to the error (v) arising because the forecast model (31) is a stochastic rather than an exact relation, (36) shows that error in $\hat{C}^{\circ}_{t+1}$

could also arise from: (1) biased parameter estimates, (2) error in extrapolating income, and (3) error in the preliminary consumption data.

If Y were a linear autoregressive series, Y_{t+1} would be

$$Y_{t+1} = \delta_0 + \Sigma\delta_i Y_{t-i+1} + w_t, \tag{37}$$

and the error of the income extrapolation would be

$$\hat{Y}^{\circ}_{t+1} - Y_{t+1} = (d_0 - \delta_0) + \Sigma(d_i - \delta_i)Y_{t-i+1} + \Sigma d_i \epsilon(Y)_{t-i+1} - w_{t+1}. \tag{38}$$

Thus the error in extrapolating income would be partly induced by errors in the preliminary income data. These errors would affect the forecast directly as well as indirectly through their effects on the estimated parameters of the extrapolation model.

Now if C and Y and their respective errors, $\epsilon(Y)$ and $\epsilon(C)$, were stationary series, their means would be independent of t. Then

$$a - \alpha = \overline{\epsilon(C)} - (b - \beta)\bar{Y} - (c - \gamma)\bar{C} - b\overline{\epsilon(Y)} - c\overline{\epsilon(C)}, \tag{39}$$

and

$$d_0 - \delta_0 = \overline{\epsilon(Y)} - \Sigma(d_i - \delta_i)\bar{Y} - \Sigma d_i \overline{\epsilon(Y)}, \tag{40}$$

where the bar denotes mean value. Using (38), (39), and (40) to rewrite (36), the forecast error would become

$$\begin{aligned} \hat{C}^{\circ}_{t+1} - C_{t+1} = {} & \overline{\epsilon(C)} + b\overline{\epsilon(Y)} + (b - \beta)(Y_{t+1} - \bar{Y}) + (c - \gamma)(C_t - \bar{C}) \\ & + b\Sigma(d_i - \delta_i)(Y_{t-i+1} - \bar{Y}) + c[\epsilon(C)_t - \overline{\epsilon(C)}] \\ & + b\Sigma d_i[\epsilon(Y)_{t-i+1} - \overline{\epsilon(Y)}] - v_{t+1} - bw_{t+1}. \end{aligned} \tag{41}$$

Though errors in the independent variables would bias the parameter estimates [i.e.. $E(b - \beta)$, $E(c - \gamma)$, and $E(d_i - \delta_i)$ would not equal zero], it is well known that biased parameter estimates would not bias the forecast if C, Y, and their errors were stationary series. Under stationarity assumptions,

$$E(Y_{t-i+1} - \bar{Y}),\ E(C_t - \bar{C}),\ E[\epsilon(C)_t - \overline{\epsilon(C)}],\ \text{and}\ E[\epsilon(Y)_{t-i+1} - \overline{\epsilon(Y)}]$$

would all be zero, and hence the bias in b, c, and d_i would create no bias in $\hat{C}^{\circ}_{t+1}$. This is not to say that $\hat{C}^{\circ}_{t+1}$ would be unbiased, however. The expected value of the forecast error would be

$$E(\hat{C}^{\circ}_{t+1} - C_{t+1}) = \overline{\epsilon(C)} + b\overline{\epsilon(Y)}, \tag{42}$$

assuming $E(v)$ and $E(w)$ are zero. Thus the forecast would be unbiased only if the preliminary data were unbiased. Since these data have a negative bias, we would expect that consumption forecasts would also have a negative bias, and Table 2-8 shows that they do.

Table 2-8 illustrates the direct as well as indirect effects of using preliminary rather than revised data on the accuracy of consumption forecasts. The table shows error statistics for forecasts constructed in three ways: (1) by inserting variables based on preliminary data ($\hat{Y}^{\circ}_{t+1}$ and C°_{t}) into the equation estimated from preliminary data; (2) by inserting variables based on 1965 revised data ($\hat{Y}^{65}_{t+1}$ and C^{65}_{t}) into the preliminary equation; and (3) by inserting the revised data variables into the equation estimated from revised data. The effect of errors in the variables used to construct the forecast (the direct effect) is shown by comparing the errors in forecasts of type (1) with those in forecasts of type (2). The effect of data errors on the parameter estimates (the indirect effect) is shown by comparing the errors in type (2) forecasts with those in type (3) forecasts. The total effect of data errors is seen by comparing the errors in type (1) with those in type (3) forecasts.

The use of preliminary rather than revised data resulted in a *doubling* of the forecast error (line 1 compared with line 3, 4 with 6, and 7 and 9 with 11). Though the direct effect is clearly more important and accounts for most of the increase in error, the indirect effect is by no means negligible.[17]

V. SUMMARY

According to our analysis, the use of preliminary rather than revised GNP data impaired forecasting accuracy and by a substantial amount: The accuracy of naive model projections of GNP and its components

[17] Denton and Kuiper [2] found somewhat similar results for the Canadian data: The direct effects were much larger than the indirect effects of errors in the preliminary data. This is not to say that the parameter estimates were unaffected. Indeed, they found that the choice of data had a stronger effect on the estimates of the parameters of their small econometric model than that resulting from the choice of estimating procedures (direct least squares or two-stage least squares).

TABLE 2-8. Effect of Data Errors on the Predictive Accuracy of Zellner's Quarterly Consumption Function, 1961-I–1964-IV [a]

Line	Prediction Equation: $\hat{C}_{t+1} = a + b\hat{Y}_{t+1} + cC_t$ Coefficients	Variables Based on Preliminary or Revised Data	Prediction Errors (billion 1958 dollars) $\bar{E}$ (1)	S_E (2)	$\bar{M}$ (3)
	1947-I–1960-IV				
	Zellner Method				
	Preliminary Data Coefficients				
1	$a = 2.7, b = .185, c = .796$	Preliminary	−7.2	3.8	8.1
2		Revised	−4.0	2.4	4.6
	1965 Revised Data Coefficients				
3	$a = 0.4, b = .258, c = .728$	Revised	−3.0	2.3	3.7
	Griliches et al. Method				
	Preliminary Data Coefficients				
4	$a = 3.1, b = .300, c = .670$	Preliminary	−7.3	3.8	8.2
5		Revised	−5.0	2.4	5.5
	1965 Revised Data Coefficients				
6	$a = -0.4, b = .330, c = .652$	Revised	−3.6	2.3	4.2
	1947-I–1955-I				
	Zellner Method				
	Zellner Coefficients				
7	$a = 0.1, b = .128, c = .870$	Preliminary	−5.5	3.7	6.6
8		Revised	−4.0	2.4	3.0
	Griliches *et al.* Coefficients				
9	$a = 0.6, b = .071, c = .928$	Preliminary	−6.3	3.8	7.3
10		Revised	−2.3	2.4	3.3
	1965 Revised Data Coefficients				
11	$a = 0.3, b = .168, c = .827$	Revised	−1.7	2.4	2.9

[a] The sample period excludes 1950-III and 1951-I. The coefficients are from Table 2-7. The predictions in lines 1, 4, 7, and 9 are based on C_t°, $\hat{Y}_{t+1}^{\circ}$; the remainder are based on C_t^{65}, $\hat{Y}_t^{65}$, and the actual value is C_{t+1}^{65}; where C_{t+1}^{65} and C_t^{65} are estimated 1965 statistically revised estimates, C_t° denotes preliminary estimates, and $\hat{Y}_{t+1}^{\circ}$ and $\hat{Y}_{t+1}^{65}$ denote extrapolations based on preliminary and on 1965 statistically revised estimates, respectively.

The coefficients used to obtain $\hat{Y}_{t+1}^{\circ}$ were estimated from the data used by Griliches. Those used to obtain $\hat{Y}_{t+1}^{65}$, were estimated from the 1965 revised data. In both cases the regression was of the form

$$Y_t = d_0 + d_1 Y_{t-1} + \cdots + d_6 Y_{t-6} + v_t$$

and the sample period was 1948-II–1960-IV. Extrapolations, $\hat{Y}_t = d_0 + d_1 Y_t + \cdots + d_6 Y_{t-5}$, were then generated for the 1961-I–1964-IV period, $\hat{Y}_{t+1}^{\circ}$ used preliminary data in the equation estimated from the Griliches data and $\hat{Y}_{t+1}^{65}$ used 1965 data in the equation estimated from revised data.

was reduced by about 30 per cent, while that of business forecasts was reduced by nearly 40 per cent.

Data errors were not the major source of systematic error in the business forecasts examined here. Though they could be considered the primary source of the bias in three of the sixteen forecast sets, in no instance did they materially contribute to the slope component of inefficiency. This does not mean that forecast efficiency was unaffected. Indeed, the reduction in efficiency was considerable. It is estimated that data errors accounted for 50 to 70 per cent of the variance of the error in seven of the sixteen forecasts.

Data errors affect not only the variables underlying a forecast (the direct effect); they affect the estimates of the parameters of the relationships among these variables as well (the indirect effect). A well-known quarterly consumption function was used to illustrate the indirect as well as direct effects of data errors. Consumption forecasts were generated from preliminary and from 1965 revised data. The use of preliminary rather than revised data led to a *doubling* of the forecast errors. The direct effect accounted for 70 per cent of the increase; the remaining 30 per cent was due to the indirect effect of data errors on the parameter estimates.

These results suggest that there is considerable scope for improving forecasting accuracy by improving the accuracy of preliminary data.

REFERENCES

[1] De Janosi, Peter E., "A Note on Provisional Estimates of the Gross National Product and Its Major Components," *Journal of Business*, October 1962.

[2] Denton, Frank T. and Kuiper, John, "The Effect of Measurement Errors on Parameter Estimates and Forecasts: A Case Study Based on the Canadian Preliminary National Accounts," *Review of Economics and Statistics*, May 1965.

[3] Griliches, Zvi, Maddala, G. S., Lucas, R., and Wallace, N., "Notes on Estimated Aggregate Quarterly Consumption Functions," *Econometrica*, July 1962.

[4] Johnston, J., *Econometric Methods*, New York, 1963.

[5] Suits, Daniel, "Forecasting and Analysis with an Econometric Model," *American Economic Review*, March 1962.

[6] Zarnowitz, Victor, *An Appraisal of Short-Term Economic Forecasts*, Occasional Paper 104, NBER, New York, 1967.

[7] Zellner, Arnold, "A Statistical Analysis of Provisional Estimates of Gross National Product and Its Components, of Selected National Income Components, and of Personal Savings," *American Statistical Association Journal*, March 1958.

[8] ———, "The Short-Run Consumption Function," *Econometrica*, October 1957.