THE USE AND ABUSE OF REAL-TIME DATA IN ECONOMIC FORECASTING

Evan F. Koenig, Sheila Dolmas, and Jeremy Piger

Abstract—We distinguish between three different strategies for estimating forecasting equations with real-time data and argue that the most popular approach should generally be avoided. The point is illustrated with a model that uses current-quarter monthly industrial production, employment, and retail sales data to predict real GDP growth. When the model is estimated using either of our two alternative methods, its out-of-sample forecasting performance is superior to that obtained using conventional estimation and compares favorably with that of the Blue Chip consensus.

I. Introduction

Many economic time series are subject to revision. Revisions to measures of real economic activity—such as employment, sales, and production—are sometimes large, and may occur years after official figures are first released. Nevertheless, analysts typically use only data of the most recent vintage when estimating and evaluating their forecasting models. Current-vintage data are often used even in ex post recursive forecasting exercises that are classified as out-of-sample. The use of current-vintage data can lead an analyst to include variables in his forecasting model that, in real time, have little marginal predictive power (Diebold & Rudebusch, 1991; Swanson, 1996). It can exaggerate the forecasting performance of a model relative to alternative models and relative to predictions that were actually available at the time (Fair and Shiller, 1990; Orphanides, 1999).

In those studies where the pitfalls of relying on current-vintage data have been taken seriously, a common response has been to use *end-of-sample*-vintage data for estimation and evaluation purposes instead of current-vintage data.[1] As the sample period over which the forecasting equation is estimated is extended, the vintage of the data used to estimate the equation is updated. Thus, rather than forecast, say, 1990:Q1–1997:Q4 GDP growth using recursive regressions all estimated with current-vintage data, one uses the prediction of an equation estimated with 1990:Q1-vintage data to forecast GDP growth in 1990:Q1, the prediction of an equation estimated with 1990:Q2-vintage data to forecast GDP growth in 1990:Q2, and so forth. This procedure mimics the actual practice of many professional forecasters and provides a level playing field for comparing their performance with that of the model. However, there is reason to suspect that this conventional approach to real-time estimation and forecasting is often suboptimal. A fortiori, the predictions made by professional forecasters may also often be suboptimal.[2]

Rather than use end-of-sample-vintage data to estimate their forecasting equations, we argue that analysts should generally use data of as many different vintages as there are dates in their samples. More specifically, at every date within a sample, right-side variables ought to be the most up-to-date estimates available *at that time.* We call these *real-time-vintage data.* For example, when the left-side variable is 1990:Q1 GDP growth, all right-side variables should be measured as they appeared in 1990:Q1. Only 1990:Q1-vintage data should be used in forecasting 1990:Q1 GDP growth, regardless of whether or not the sample period extends beyond 1990:Q1. Thus, we argue that when a data point is added to the end of the sample period, the data that appear on the right-hand side of the equation earlier on in the sample ought *not* to be updated. We also point out that, under reasonable conditions, first available official estimates should be used for the left-side variable when estimating the forecasting equation. First available estimates ought to be used even if one is ultimately interested in predicting final revised data.

The intuition underlying our arguments is simple. The empirical relationship between GDP growth and early estimates of (say) employment growth will typically differ from that between GDP growth and estimates of employment growth available several years after the fact. It is the former relationship that is of interest to the forecaster. The trouble with the conventional approach to real-time estimation is that the data on the right-hand side of the forecasting equation range from extensively revised (early in the sample) to nearly unrevised (at end of sample). What we call real-time-vintage estimation avoids this problem by including at each point in the sample only right-side data that would have been available to a forecaster at that point.

The argument for using first-release left-side data is based on the assumption that the government's initial release is an efficient estimate of subsequent releases, meaning that

Received for publication January 27, 2000. Revision accepted for publication July 15, 2002.

* Federal Reserve Bank of Dallas, unaffiliated, and Federal Reserve Bank of St. Louis, respectively.

This paper had its origins in a GDP forecasting project undertaken jointly with Ken Emery. Helpful comments and suggestions were offered by Jim Stock, an anonymous referee, Nathan Balke, Todd Clark, Dean Croushore, Jon Faust, Preston Miller, John Robertson, Ellis Tallman, Jonathan Wright, and attendees of the November 1999 meeting of the Federal Reserve System Committee on Macroeconomics. Dean Croushore and Ellis Tallman were kind enough to share their data, and Dong Fu provided research assistance. Early drafts of the paper were completed when all three authors were affiliated with the Federal Reserve Bank of Dallas. The opinions expressed do not necessarily represent the views of the Federal Reserve Bank of Dallas, the Federal Reserve Bank of St. Louis, or the Federal Reserve System.

[1] See, for example, Fitzgerald and Miller (1989), Trehan (1989, 1992), Diebold and Rudebusch (1991), Miller and Chin (1996), and Robertson and Tallman (1998). Our "end-of-sample-vintage data" are what Swanson (1996) calls "partially revised" or "real-time" data.

[2] Ehrbeck and Waldmann (1996) and Laster, Bennett, and Geoum (1999) also argue that professional forecasts are typically suboptimal. Their arguments depend on strategic behavior by forecasters, and are completely different from the arguments advanced here.

revisions to the initial release are unpredictable using data available at the time it is issued. If this assumption holds, unbiased estimates of the parameters linking true GDP growth to the real-time-vintage right-side variables will be obtained regardless of the vintage of the data on the left-hand side of the estimated equation. Less obviously, the parameter estimates obtained using initial-release left-side data will be more accurate, in finite samples, than those obtained using revised data. Any revised GDP release will incorporate information that cannot, under the efficiency assumption, be predicted using data available at the time of the initial release. To the econometrician estimating the forecasting equation, this additional information is extraneous noise. Even if the government's initial estimates are not fully efficient, they may be good enough that our proposed approach will perform well in practice.

It might appear that collecting the data required to implement our version of real-time forecasting would be prohibitively difficult. However, for most variables of interest it is easier to obtain short data series of many vintages than it is to reconstruct long series of a few vintages.

The specific application we consider is forecasting same-quarter real GDP growth using monthly data on employment, industrial production, and retail sales.[3] Economists devote substantial effort to forecasting GDP growth, and their prognostications receive much press attention. Despite this effort and scrutiny, GDP forecasts are not very accurate. For example, since 1990, the root-mean-square error of the Blue Chip consensus GDP growth forecast has been 1.5 percentage points, based on forecasts published three weeks prior to the release of the official advance GDP estimate. The corresponding 95% confidence interval is 5.8 percentage points wide. Consensus forecasts are known to be more accurate than those of most individuals (Graham, 1996; McNees, 1987).

Despite the limited set of monthly indicators included in our model and our relatively short sample period, we are able to achieve out-of-sample forecasting performance that is as good as, or better than, that of the Blue Chip consensus. An important contributor to our model's strong performance is the fact that we estimate it with real-time rather than end-of-sample-vintage data. When our model is estimated conventionally, its forecasting performance suffers.

II. The Forecasting Problem

Consider the problem of forecasting a single variable, y, using time series observations on a $1 \times k$ vector, $\boldsymbol{x}$, of other variables (which might possibly include lagged values of y). Official estimates of both $\boldsymbol{x}$ and y are subject to revision. The initial official estimate of y is based on at least as much information as is available to the econometrician. We adopt the following notation:

$y(t) \equiv$ the "true" period-t value of y, $0 \le t$,
$\boldsymbol{x}(t) \equiv$ the "true" period-t value of $\boldsymbol{x}$, $0 \le t$,
$y(t)_s \equiv$ the official estimate of $y(t)$ released at time s, where $s \ge t$,
$\boldsymbol{x}(t)_s \equiv$ the official estimate of $\boldsymbol{x}(t)$ available at time $s \ge t$.

For concreteness, the reader may want to think of $y(t)_s$ as an official estimate of GDP growth. The econometrician is trying to forecast GDP growth, $y(T)$, in quarter $T > 0$. No official estimate of quarter-T GDP growth has yet been released [$y(T)_T$ is, as yet, unavailable]. However, an initial estimate of (say) quarter-T employment growth is available and is included in the vector of variables observed by the econometrician [$\boldsymbol{x}(T)_T$].[4] Also available and included in $\boldsymbol{x}(T)_T$ might be partially revised official estimates of lagged employment growth rates.

Since our goal is to forecast $y(T)$ using $\boldsymbol{x}(T)_T$, it seems natural to posit that $y(t)$ and $\boldsymbol{x}(t)_t$ are linearly related:

$$y(t) = \boldsymbol{x}(t)_t\boldsymbol{\alpha} + w(t), \tag{1}$$

where $w(t)$ captures both information available to the government at time t, but not to the econometrician, and information that is not available, even to the government, until after time t. From the perspective of the econometrician, $w(t)$ is a mean-zero disturbance orthogonal to $\boldsymbol{x}(t)_t$. While heteroskedasticity cannot be ruled out, there are no a priori reasons to expect it.

Given an estimate of $\boldsymbol{\alpha}$, the econometrician forecasts $y(T)$ in the obvious way, by setting

$$\hat{y}(T) = \boldsymbol{x}(T)_T\hat{\boldsymbol{\alpha}}, \tag{2}$$

where hats indicate forecasts or estimates. But how is $\hat{\boldsymbol{\alpha}}$ to be obtained?

III. Alternative Estimation Strategies

In practice, $y(t)$ is often observed with a substantial lag—if it is observed at all. Consequently, direct estimation of equation (1) will usually either be impossible or be feasible only over a truncated sample period.[5] We consider three different ways around this problem.

Our preferred strategy (henceforth, *strategy 1*) is to estimate the forecasting equation with first-release data on its left-hand side and real-time-vintage data (data, at each point within the sample, that are the latest available at the time) on the right. In the notation introduced above, strategy-1

[3] Closely related work includes Fitzgerald and Miller (1989), Trehan (1989, 1992), Braun (1990), Ingenito and Trehan (1996), and Miller and Chin (1996). Zadrozny (1990) and Rathjens and Robins (1993) use monthly data to improve forecasts of *next* quarter's output growth.

[4] Jobs data are released almost a month before the first official estimate of GDP growth.

[5] Suppose that the truth is revealed after $S > 0$ periods, so that $y(t)_{t+S} = y(t)$. Then equation (1) can be estimated over a sample period that runs from $t = 0$ to $t = T - S - 1$. We argue that even when it is possible, direct estimation of equation (1) is likely to be suboptimal.

coefficient estimates are obtained by applying least squares to the equation

$$y(t)_t = \boldsymbol{x}(t)_t\boldsymbol{\alpha} + \omega(t) \tag{3}$$

for $0 \leq t \leq T - 1$.[6] There is no a priori reason to expect that $\omega(t)$ will be heteroskedastic. Serial correlation can usually be eliminated by expanding the vector $\boldsymbol{x}(t)_t$ to include additional lags.[7]

Strategy 2 is to estimate an equation of the form

$$y(t)_{T-1} = \boldsymbol{x}(t)_t\boldsymbol{\alpha} + \omega'(t) \tag{3$'$}$$

for $0 \leq t \leq T - 1$. Intuitively, by using an end-of-sample-vintage estimate of $y(t)$ on the left-hand side of the equation and real-time-vintage data on the right, the econometrician comes as close to directly estimating equation (1) as possible.[8]

Real-world professional forecasters generally use end-of-sample-vintage data on both the left-hand and right-hand sides of their forecasting equations. Economists who undertake conventional real-time forecasting exercises imitate this process, updating the vintage of the data used for estimating their models as they gradually extend their samples. In our notation, the conventional approach to real-time forecasting amounts to estimating

$$y(t)_{T-1} = \boldsymbol{x}(t)_T\boldsymbol{\alpha} + \omega''(t) \tag{3$''$}$$

for $0 \leq t \leq T - 1$. We call this approach *strategy 3.*

IV. Strategy-3 Bias

Strategy 3 is consistent only under exceptional circumstances. The exact formula for its bias depends on the nature of the revisions to the right-side variables. At one extreme, $\boldsymbol{\xi}(t)_T \equiv \boldsymbol{x}(t)_T - \boldsymbol{x}(t)_t$ might be pure *noise,* uncorrelated with $\boldsymbol{x}(t)_T$ and $y(t)_{T-1}$. At the other extreme, $\boldsymbol{\xi}(t)_T$ might be pure *news,* uncorrelated with all variables in the government's information set at time t. We consider both possibilities in turn. Throughout we assume that the left-side-variable error, $y(t) - y(t)_{T-1}$, is uncorrelated with $\boldsymbol{x}(t)_t$. This assumption is standard in textbook treatments of the errors-in-variables problem. In the case where deviations of $\boldsymbol{x}(t)_T$ from $\boldsymbol{x}(t)_t$ are pure news, assuming that deviations of $y(t)$ from $y(t)_{T-1}$ are also unpredictable using time-t information seems natural.

We begin by taking the probability limit of the coefficient estimate, $\hat{\boldsymbol{\alpha}}''$, obtained by applying least squares to equation (3$''$). If $\boldsymbol{\xi}(t)_T \equiv \boldsymbol{x}(t)_T - \boldsymbol{x}(t)_t$ is noise, uncorrelated with $\boldsymbol{x}(t)_T$ and $y(t)_{T-1}$, standard calculations yield

$$\text{plim } \hat{\boldsymbol{\alpha}}'' = \boldsymbol{\alpha} + \Sigma_{X_TX_T}^{-1}\Sigma_{\xi\xi}\boldsymbol{\alpha}, \tag{4}$$

where $\Sigma_{X_TX_T} \equiv \text{plim}[X_T'X_T/T]$, $\Sigma_{\xi\xi} \equiv \text{plim}[\Xi'\Xi/T]$, and X_T and Ξ are $T \times k$ matrices whose t^{th} rows are $\boldsymbol{x}(t)_T$ and $\boldsymbol{\xi}(t)_T$, respectively. Thus, the least-squares coefficient estimates are inconsistent.

Equation 4 is the counterpart of the textbook formula for coefficient bias when right-side variables are subject to measurement error. [The textbook formula can be obtained from equation (4) by treating $\boldsymbol{\alpha}''$ as the true coefficient vector and $\boldsymbol{\alpha}$ as the limiting value of a least squares estimate of $\boldsymbol{\alpha}''$, and then solving for $\boldsymbol{\alpha}$ as a function of $\boldsymbol{\alpha}''$.] The intuition is straightforward. Textbook treatments of the errors-in-variables problem take it for granted that the analyst wants to find the relationship between the dependent variable and error-free values of the right-side variables. However, real-world forecasters have no choice but to substitute noisy, preliminary data into their forecasting equations. So, contrary to the textbook assumption, it is the coefficient vector $\boldsymbol{\alpha}$ from equation (1) that matters to the forecaster, rather than the coefficient vector that relates $y(t)$ to $\boldsymbol{x}(t)$.[9]

Now consider the case where revisions to the right-side variables, $\boldsymbol{\xi}(t)_T$, are news, uncorrelated with all variables in the government's time-t information set. In this case, applying least squares to equation (3$''$) and taking the probability limit of the resultant coefficient estimate yields

$$\text{plim } \hat{\boldsymbol{\alpha}}'' = \boldsymbol{\alpha} + (\Sigma_{XX} + \Sigma_{\xi\xi})^{-1}[\Sigma_{\xi\nu} - \Sigma_{\xi\xi}\boldsymbol{\alpha}], \tag{5}$$

where $\Sigma_{XX} \equiv \text{plim}(X'X/T)$, $\Sigma_{\xi\nu} \equiv \text{plim}(\Xi'\boldsymbol{N}/T)$, X is a $T \times k$ matrix whose t^{th} row is $\boldsymbol{x}(t)_t$, and $\boldsymbol{N}$ is a $T \times 1$ vector whose t^{th} element is $\nu(t)_{T-1} \equiv y(t)_{T-1} - y(t)_t$. It is theoretically possible for the factor in square brackets in equation (5) to equal zero [such will be the case if $\nu(t)_{T-1}$ is related to $\boldsymbol{\xi}(t)_T$ in the same way that $y(t)_{T-1}$ is related to $\boldsymbol{x}(t)_t$], but there are no a priori grounds for believing that this condition will hold. Hence, strategy 3 will typically yield an inconsistent estimate of $\boldsymbol{\alpha}$.

[6] It is the requirement that right-side data at each point within the sample incorporate then-available revisions that distinguishes our preferred strategy from estimation that uses exclusively first-release data (Swanson, 1996; Swanson & White, 1996, 1997). Because it discards potentially useful information, first-release estimation can be expected to yield poorer in-sample fit and out-of-sample forecasting precision than strategy 1. As shown in the appendix, restricting $\boldsymbol{x}(t)_t$ to first-release data also induces serial correlation in equation (3) when it would otherwise be absent.

[7] An exception is when changes in the government's methodology for calculating $y(t)$ are so great as to shift $\boldsymbol{\alpha}$. Ideally, the analyst would apply the government's latest methodology retroactively, using real-time vintage source data to obtain methodologically consistent series for $y(t)_t$ and $\boldsymbol{x}(t)_t$. A more practical alternative is to test for structural breaks coincident with major methodological revisions and—if necessary—introduce one or more dummy variables on the right-hand side of equation (3).

[8] We assume that the release of $y(t)_T$ is delayed relative to the release of $\boldsymbol{x}(t)_T$ just as the release of $y(T)_T$ is delayed relative to the release of $\boldsymbol{x}(T)_T$. Hence, $y(t)_{T-1}$ is the most up-to-date available estimate of $y(t)$.

[9] Howrey (1978) suggests a three-step forecasting procedure in the revisions-as-noise case. First, latest available data are used to estimate a relationship between the left-side and right-side variables that approximates the relationship between $y(t)$ and $\boldsymbol{x}(t)$. Second, real-time-vintage right-side data, $\boldsymbol{x}(t)_t$, are regressed on latest available data to approximate the relationship between $\boldsymbol{x}(t)_t$ and $\boldsymbol{x}(t)$. Finally, the analyst uses the Kalman filter to find $E_T(\boldsymbol{x}(T))$, which he substitutes into the equation estimated in step 1 to obtain a forecast of $y(T)$.

As an extreme example, consider the special case in which $\boldsymbol{\alpha} = \mathbf{0}$, so that $\boldsymbol{x}(T)_T$ is of absolutely no use in forecasting $y(T)$. Equation (5) says that estimating a relationship between y and $\boldsymbol{x}$ using end-of-sample data (strategy 3) will, nevertheless, yield a nonzero estimate of $\boldsymbol{\alpha}$ insofar as *revisions* to y are correlated with *revisions* to the elements of $\boldsymbol{x}$. Hence, one of the complaints that has been directed against forecasting analyses that use current-vintage data—that such analyses can lead the econometrician to rely on indicators that, in real time, have little marginal predictive power—applies also to any real-time analysis that uses end-of-sample-vintage data. Intuitively, if sample periods extend back very far at all, both current-vintage and strategy-3 estimations are dominated by heavily revised data—data that may contain more information on how revisions to the forecasted variable are related to revisions to the right-side variables than on how early estimates of the right-side variables are related to the forecasted variable.

V. Comparing Strategies 1 and 2

A. *Bias*

The equation (3) and equation (1) error terms are related by $\omega(t) = w(t) - \nu(t)$, where $\nu(t) \equiv y(t) - y(t)_t$. It immediately follows that if $\boldsymbol{x}(t)_t$ and $\nu(t)$ are uncorrelated, then ordinary least squares applied to equation (3) (strategy 1) will yield an unbiased estimate of the coefficient vector $\boldsymbol{\alpha}$. The zero-correlation condition will be satisfied if the government uses the information in $\boldsymbol{x}(t)_t$ efficiently when constructing its initial official estimate of $y(t)$.

Similarly, the equation (3′) and equation (1) error terms are related by $\omega'(t) = w(t) - [y(t) - y(t)_{T-1}]$. Hence, least squares applied to equation (3′) (strategy 2) will yield an unbiased estimate of $\boldsymbol{\alpha}$ provided that $\boldsymbol{x}(t)_t$ and $y(t) - y(t)_{T-1}$ are uncorrelated, as will certainly be the case if the government uses the information in $\boldsymbol{x}(t)_t$ efficiently when constructing $y(t)_{T-1}$. Because it applies to the government's end-of-sample estimate of $y(t)$ rather than its initial estimate, this condition is slightly weaker than that needed to guarantee the unbiasedness of strategy-1 estimation.

Although empirical tests of the efficiency of government statistical releases are not uniformly supportive (Runkle, 1998; Croushore and Stark, 1999), the fact that attempts to second-guess government estimates at the time of their release are rare suggests that inefficiencies in the government's estimation process are generally small. Hence, prescriptions based on the efficiency hypothesis may be a good guide to econometric practice.

B. *Coefficient Precision and Forecast Accuracy*

The strategy-2 and strategy-1 error terms are related by $\omega'(t) = \omega(t) + \nu(t)_{T-1}$, where $\nu(t)_{T-1} \equiv y(t)_{T-1} - y(t)_t$ is the difference between the government's end-of-sample and initial estimates of $y(t)$. Because $\omega(t) = y(t)_t - \boldsymbol{x}(t)_t\boldsymbol{\alpha}$ and because both $y(t)_t$ and $\boldsymbol{x}(t)_t$ are in the government's period-t information set, $\omega(t)$ will be uncorrelated with $\nu(t)_{T-1}$ provided that $y(t)_t$ is an efficient estimate of $y(t)_{T-1}$. Hence, the strategy-2 error term will have a greater variance than the strategy-1 error term: $\text{var}[\omega'(t)] = \text{var}[\omega(t)] + \text{var}[\nu(t)_{T-1}] \geq \text{var}[\omega(t)]$. Strict inequality will hold for at least one $t < T - 1$ unless the government's initial estimate of $y(t)$ is never revised. More generally, the variance of the strategy-2 error term can be expected to be decreasing in t, but is always at least as great as the variance of the strategy-1 error term.[10]

The intuition behind these results is straightforward. The strategy-1 error term is the difference between the government's period-t estimate of $y(t)$ and the econometrician's estimate of $y(t)$. As such, it has a nonzero variance only to the extent that the government has more information in period t than does the econometrician. The strategy-2 error term, in contrast, reflects both the period-t government-private information gap and additions to the government's information set over time. If the government uses information efficiently, the government-private-information gap will be uncorrelated with additions to the government's information set. Hence the strategy-1 error term must have lower variance than the strategy-2 error term.[11]

That the strategy-2 error term has larger variance than the strategy-1 error term means that strategy 1 will yield more-precise finite-sample parameter estimates.[12] More-precise parameter estimates translate into superior forecasting performance. To see that this is so, note that the forecast error under either strategy has two components: the government's error in forecasting $y(T)$ based on information available to it at time T, and the econometrician's error in forecasting the government's initial estimate of $y(T)$. Under our efficiency assumption, the two components are uncorrelated. Hence,

$$\text{var}[y(T) - \hat{y}(T)] = \sigma_\nu^2 + \sigma_\omega^2 + \text{var}[\boldsymbol{x}(T)_T(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})]. \quad (6)$$

The final term in equation (6) is the variance due to coefficient uncertainty. For any fixed $\boldsymbol{x}(T)_T$, this variance goes to zero as the sample size increases. For any given sample size, it is increasing in the distance between $\boldsymbol{x}(T)_T$ and the sample mean of the $\boldsymbol{x}(t)_t$ $(t = 0, 1, \ldots, T - 1)$. In the special case in which $\boldsymbol{x}(T)_T$ and the sample mean of

[10] Thus, the error term in equation (3′) can be expected to be heteroskedastic. Likely serial correlation in $\nu(t)_{T-1}$ means that $\omega'(t)$ will also typically be serially correlated.

[11] Consider, for example, the extreme case in which $y(t) = x(t) + z(t)$ where $x(t)$ and $z(t)$ are both white noise. Suppose that $x(t)$ is observed by both the government and the private sector at time t [so that $x(t)_t = x(t)$], but $z(t)$ is not observed until later. In the notation used above, $\alpha = 1$, $y(t)_t = x(t)\alpha$, $\omega(t) \equiv 0$, and $w(t) = \nu(t) = z(t)$. Clearly, in any finite sample the analyst will do better to "estimate" the equation $y(t)_t = x(t)\alpha$ (which fits the data perfectly) than to estimate the equation $y(t) = x(t)\alpha + w(t)$, although both strategies yield unbiased estimates of α.

[12] The dropoff in precision will be particularly great if, in estimating equation (3′), heteroskedasticity and serial correlation are not properly taken into account. See footnote 10.

the $\boldsymbol{x}(t)_t$ coincide, the variance due to coefficient uncertainty is $\sigma_\omega^2/\underline{T}$ under strategy 1 and $(\sigma_\omega^2 + \overline{\sigma_\nu^2})/T$ under strategy 2, where $\overline{\sigma_\nu^2}$ is the average variance of the revisions to the left-side variable. The relative inefficiency of strategy 2 is clear.

C. More on the Performance Advantage of Strategy 1 over Strategy 2

We can say more about the gains from strategy-1 estimation in the special case where the econometrician forecasts $y(T)$ using a single right-side variable and a constant, so that equation (1) takes the form

$$y(t) = \alpha_0 + \alpha_1 x(t)_t + w(t), \tag{7}$$

and equation (3) becomes

$$y(t)_t = \alpha_0 + \alpha_1 x(t)_t + \omega(t), \tag{8}$$

where $\omega(t) = w(t) - \nu(t) = w(t) - [y(t) - y(t)_t]$. For example, $y(t)$ might be quarterly real GDP growth and $x(t)$ same-quarter employment growth. We assume that the government's initial $y(t)$ estimate is efficient, so that $\nu(t)$ is uncorrelated with all variables in the government's period-t information set.

To further simplify the analysis, suppose that $y(t)$ becomes available within one period of the initial release, so that direct estimation of equation (7) is possible—an extreme version of strategy-2 estimation. A real-world example would be the analyst who wants to forecast the so-called final estimate of real GDP growth, which is published two months after GDP figures first become available. The question is whether the analyst ought to estimate his forecasting equation with final GDP data on its left-hand side [equation (7)], or with first-release data [equation (8)].

With a sample extending from $t = 0$ to $t = T - 1$, the variance of the period-T forecast error is

$$\text{var}[y(T) - \hat{y}(T)] = \sigma_w^2(1 + K) = (\sigma_\omega^2 + \sigma_\nu^2)(1 + K) \tag{9}$$

using strategy 2 [equation (7)], and

$$\text{var}[y(T) - \hat{y}(T)_T] = \sigma_\nu^2 + \sigma_\omega^2(1 + K) \tag{10}$$

using strategy 1 [equation (8)], where $K \equiv [1 + F(T + 1)/(T - 1)]/T$,

$$F \equiv \frac{[\boldsymbol{x}(T)_T - \bar{\boldsymbol{x}}]^2/(T + 1)}{\{(1/T)\sum_0^{T-1}[\boldsymbol{x}(t)_t - \bar{\boldsymbol{x}}]^2\}/(T - 1)}, \tag{11}$$

and $\bar{x}$ is the sample mean of the $x(t)_t$. Note that strategy 1 unambiguously outperforms strategy 2.

Comparing equations (9) and (10), the ratio of the strategy-1 forecast-error variance to the strategy-2 forecast-error variance is

$$\begin{aligned}\frac{\text{var}[y(T) - \hat{y}(T)_T]}{\text{var}[y(T) - \hat{y}(T)]} &= 1 - \left(\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}\right)\left(\frac{K}{1 + K}\right) \\ &= 1 - \left[\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}\right] \\ &\quad \times \left[\frac{1 + \dfrac{T + 1}{T - 1}F}{(T + 1) + \dfrac{T + 1}{T - 1}F}\right].\end{aligned} \tag{12}$$

The first factor in brackets on the right-hand side of this equation is the fraction of σ_w^2 accounted for by revisions to y. The more important are these revisions, the bigger the payoff from stripping them out of the equation before estimation. The second factor in brackets measures the fraction of the total strategy-2 forecast-error variance accounted for by parameter uncertainty. This fraction is larger the smaller is the sample size and the more unusual is $x(T)_T$ relative to in-sample values of $x(t)_t$.

To reach conclusions about the ex ante expected performance advantage of strategy 1 relative to strategy 2, we must assume something about the distribution from which the $x(t)_t$ are drawn. Suppose, for example, that the $x(t)_t$ ($t = 0, 1, \ldots, T$) are independent draws from a normal distribution. Then $F \sim F_{1,T-1}$ and $(T - 1)/(T - 1 + F)$ has a beta distribution with parameters $(T - 1)/2$ and $\frac{1}{2}$ (Hogg and Craig, 1970). It follows that the mean of the forecast-variance ratio is

$$E\left(\frac{\text{var}[y(T) - \hat{y}(T)_T]}{\text{var}[y(T) - \hat{y}(T)]}\right) = 1 - \left(\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}\right)\left(\frac{2}{T + 1}\right). \tag{13}$$

Consistent with results for the case in which the $x(t)_t$ are known, equation (13) says that the expected performance gain from strategy 1 is greater the more important are revisions to official estimates of $y(t)$ and the smaller is the sample size over which the forecasting equation will be estimated.

In the case where there are $k > 1$ mutually independent right-side variables, Jensen's inequality can be used to bound the expected loss in forecast efficiency from using strategy 2:

$$\begin{aligned}E\left(\frac{\text{var}[y(T) - \hat{y}(T)_T]}{\text{var}[y(T) - \hat{y}(T)]}\right) &\geq \left(\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}\right)\left(\frac{T}{T + 1}\right) \\ &\quad \times \left(\frac{T - 1}{T - 1 + E(F)}\right) \\ &\quad + \frac{\sigma_\omega^2}{\sigma_\nu^2 + \sigma_\omega^2},\end{aligned} \tag{14}$$

where F is now the sum of k random variables, each defined as in equation (11). If the components of $x(t)_t$ are i.i.d. over time and normal, then each term in the sum has an $F_{1,T-1}$ distribution, and

$$E\left(\frac{\text{var}[y(T) - \hat{y}(T)_T]}{\text{var}[y(T) - \hat{y}(T)]}\right) \geq 1 - \left[\frac{\sigma_\nu^2}{\sigma_\nu^2 + \sigma_\omega^2}\right] \times \left[\frac{(T-3) + (T+1)k}{(T+1)(T-3+k)}\right]. \tag{15}$$

The expressions in square brackets on right-hand side of equation (15) are increasing in k, so the right-hand side as a whole is decreasing in k.

VI. An Example: Forecasting Current-Quarter GDP Growth

A. *The Model*

To illustrate the importance of estimating forecasting models with real-time-vintage data, we predict current-quarter real GDP growth using monthly measures of real economic activity. Following Trehan (1992), our set of monthly indicator variables includes nonfarm employment, real retail sales (nominal sales deflated by the consumer price index), and industrial production. These variables are all important and closely watched direct measures of current real economic activity. Nonfarm employment and industrial production are among only four variables included in the Conference Board's composite coincident index, and real retail sales serve as a timely proxy for a third component of that index (real manufacturing and trade sales).[13]

We obtain our forecasting models by regressing quarter-to-quarter changes in real GDP on a constant and five month-to-month changes in each of our three coincident indicators:

$$y(t) = \alpha_0 + \sum_{j=0}^{4} \beta_j e(t, 3-j) + \sum_{j=0}^{4} \gamma_j ip(t, 3-j) + \sum_{j=0}^{4} \epsilon_j rs(t, 3-j) + w(t), \tag{16}$$

where $y(t)$ denotes the annualized quarterly percentage change in real GDP in quarter t; where $e(t, s)$, $ip(t, s)$, and $rs(t, s)$ are the annualized monthly percentage changes in nonfarm employment, industrial production, and real retail sales, respectively, in month s of quarter t; and where $w(t)$ is an error term.[14] The lag length is chosen entirely on a priori grounds.[15]

The forecasting models differ only in the data vintages they use for estimating equation (16). All estimations start in the first quarter of 1980 and end on or before the third quarter of 1997. Forecasts cover the period from 1990:Q1 through 1997:Q4.

B. *The Data*

In our end-of-sample-vintage data set, right-side variables are measured as of the end of the sample period (quarter T), which ranges from 1989:Q4 to 1997:Q4.[16] Data toward the beginning of a given end-of-sample-vintage data set have undergone extensive revision. Data at the end of the data set are "first release." In contrast, our real-time-vintage data set consists of a sequence of 5-month "snapshot" histories of our right-side variables—one snapshot for each quarter from 1980:Q1 through 1997:Q4. Regardless of when the estimation period ends, only data that would have been available at the close of 1980:Q1 are used to forecast 1980:Q1 GDP growth; only data that would have been available at the close of 1980:Q2 are used to forecast 1980:Q2 GDP growth; and so on. There is exactly one vintage of data for each point in the estimation period. Formally, $e(t, 3-j)$, $ip(t, 3-j)$, and $rs(t, 3-j)$ ($j = 0, \ldots, 4$) are all measured at the close of quarter t, for $t =$ 1980:Q1–1997:Q4.[17] For each t, some data are first-release ($j = 0$) and some are partially revised ($j > 0$).

Depending on the estimation strategy, we use either end-of-sample-vintage or first-release measures of real GDP growth as the left-side variable in equation (16).[18] The

[13] The final component of the coincident index—real personal income—is released substantially later than the others. For a discussion of the various statistical releases, see Rogers (1998).

[14] If $x(t, s)$ is a monthly variable, then $x(t, 0) \equiv x(t-1, 3)$ and $x(t, -1) \equiv x(t-1, 2)$. Restricting the information set to one or two months of current-quarter data results in poorer forecasting performance, but does not change the relative ranking of the estimation strategies.

[15] Let $Y(t)$ denote the logarithm of quarterly aggregate output, and suppose there is a monthly coincident indicator, $X(t, s)$, such that $Y(t) = [X(t, 3) + X(t, 2) + X(t, 1)]/3$ for all t. Then

$$\begin{aligned} Y(t) - Y(t-1) = \{&[X(t, 3) - X(t, 2)] + 2[X(t, 2) - X(t, 1)] \\ &+ 3[X(t, 1) - X(t-1, 3)] \\ &+ 2[X(t-1, 1) - X(t-1, 3)] \\ &+ [X(t-1, 2) - X(t-1, 1)]\}/3. \end{aligned}$$

Thus, the quarter-to-quarter percentage change in output is a weighted average of five month-to-month percentage changes in the coincident indicator. In our GDP model, the exact pattern of weights suggested by this example was rejected in formal statistical tests, so we left the coefficient weights attached to the right-side variables unrestricted in our regressions.

[16] Ellis Tallman graciously provided industrial-production data. Other data were culled from a variety of official government sources.

[17] Data are actually measured as of the middle of the month following the close of the quarter. The initial official GDP estimate is released about two weeks later.

[18] End-of-sample-vintage "GDP" data come from the Philadelphia Fed's on-line, real-time database. For vintages through 1991:Q3, this database actually contains GNP rather than GDP data. Similarly, Blue Chip survey participants forecast GNP through 1991:Q3 and GDP thereafter, reflecting the fact that the Commerce Department emphasized GNP until November,

switch from fixed- to chain-weight GDP is treated just like any other data revision.

C. *Efficiency Tests*

Recall that our preferred estimation technique (strategy 1) puts first-release data on the left-hand side of the forecasting equation for estimation purposes and real-time-vintage data on the right. Strategy 1 yields unbiased coefficient estimates if revisions to official estimates of the forecasted variable $[y(t)]$ are uncorrelated with the vector of right-hand-side variables, $\boldsymbol{x}(t)_t$. In the present context, the question is whether GDP revisions are correlated with real-time-vintage growth in employment, industrial production, and retail sales. Results from a regression of GDP revisions on the set of monthly coincident indicators—presented on line 1 of table 1—provide no evidence that correlation is a problem. The F-statistic associated with the regression has a P-value of 0.448.

The main competition for our preferred estimation technique is strategy 2, which puts real-time-vintage data on the right-hand side of the forecasting equation, but end-of-sample data on the left. For strategy 1 to yield more precise finite-sample parameter estimates than strategy 2 it is sufficient that revisions to official $y(t)$ estimates be uncorrelated with the initial releases $[y(t)_t]$ and with the vector of right-side variables $[\boldsymbol{x}(t)_t]$. In the present context, strategy 1 will yield more accurate GDP forecasts than strategy 2 provided GDP revisions are not predictable using first-release GDP data and real-time-vintage growth in jobs, industrial production, and retail sales. Results from the relevant regression—displayed on line 2 of table 1—are ambiguous: the hypothesis that GDP revisions are unpredictable is rejected at the 10% significance level, but not the 5% level.

D. *In-Sample Performance*

Table 2 presents in-sample results from the application of strategies 1, 2, and 3 to equation (16). Q-test statistics indicate that serial correlation is not a problem, and Goldfeld-Quandt tests signal that heteroskedasticity is of importance only for strategy 2.[19] Nevertheless, all standard errors are calculated using the Newey-West estimator. Monthly changes in employment, industrial production, and retail sales are highly statistically significant in every regression.

As expected, strategy 1 produces a tighter in-sample fit than strategy 2. Indeed, the strategy-1 error term appears to have about half the variance of the strategy-2 error term. Given this difference in in-sample fit, equation (15) suggests that strategy 1's mean squared forecast error can be expected to be as much as 9% below that of strategy 2. As we shall see, strategy 1's performance advantage over strategy 2 is actually considerably greater.

TABLE 1.—TESTING THE EFFICIENCY OF THE GOVERNMENT'S INITIAL GDP ESTIMATES

Information Set	Adjusted R^2	F-Statistic	Constant Term
$x(t)_t$	0.005	1.022	−0.062
		(P = 0.448)	(P = 0.828)
$y(t)_t$ and $x(t)_t$	0.156	1.806	0.461
		(P = 0.055)	(P = 0.140)

Summary statistics obtained by regressing GDP revisions on alternative information sets, 1980:Q1–1997:Q3.

TABLE 2.—SUMMARY OF IN-SAMPLE ESTIMATION RESULTS, 1980:Q1–1997:Q3

	Value		
Statistic	Strategy 1	Strategy 2	Strategy 3
Employment:			
Joint significance	0.001	0.000	0.020
Sum of coefficients	0.293	0.488	0.476
Industrial production:			
Joint significance	0.000	0.000	0.000
Sum of coefficients	0.244	0.228	0.262
Real retail sales:			
Joint significance	0.000	0.000	0.000
Sum of coefficients	0.139	0.119	0.191
Overall:			
Adjusted R^2	0.858	0.794	0.723
Standard error of est.	1.085	1.512	1.755
Significance of GQ statistic	0.097	0.030	0.189
Significance of Q statistic	0.324	0.787	0.723

E. *Recursive Forecast Comparisons*

To rank our alternative estimation strategies, we constructed a series of recursive forecasts running from 1990:Q1 through 1997:Q4. The forecasts were then compared with *actual* GDP growth, that is, GDP growth as measured in January 1999.[20] Presumably, current-vintage GDP growth statistics are the best available estimates of what happened to real economic activity over the period in question.

Results from these recursive forecasting exercises are displayed in table 3. In addition to results for strategies 1–3, table 3 shows how well equation (16) appears to perform when estimated and evaluated naively, using current-vintage data throughout. This last approach yields potentially misleading results, since both the data used for estimation and the data substituted into the estimated equation's right-hand side to obtain a "forecast" would not actually have been available to an analyst in real time. Nevertheless, the naive

1991. So as not to unfairly disadvantage strategies 2 and 3 and the Blue Chip survey relative to strategy 1, we compare all forecasts with GNP data from 1990:Q1 through 1991:Q3, and with GDP data thereafter. (We could have avoided the GNP-GDP problem by starting our forecasting exercise in 1991:Q4, but this would have excluded the 1990–1991 recession from the analysis.) To streamline the main text, we refer only to "GDP growth," even though our data sets are actually GNP-GDP hybrids.

[19] An estimation using all first-release data showed strong serial correlation. See footnote 6.

[20] As indicated above (footnote 18), 1990:Q1–1991:Q3 forecasts are actually compared with 1999-vintage GNP growth, and 1991:Q4–1997:Q4 forecasts with 1999-vintage GDP growth.

TABLE 3.—SUMMARY STATISTICS FOR RECURSIVE FORECASTING EXERCISE, 1990:Q1–1997:Q4

Estimation Strategy	Mean Error	Mean Abs. Error	RMS Error
Strategy 1	0.21	1.09	1.36
Strategy 2	0.20	1.30	1.53
Strategy 3	0.05	1.45	1.68
Naive	−0.57	1.44	1.76
Autoregression	0.16	1.62	2.00
Blue Chip consensus	0.35	1.18	1.46

approach to forecast evaluation is frequently used in practice.

The table confirms that conventional real-time estimation (strategy 3) performs poorly in comparison with strategies 1 and 2. Moreover, strategy 1's root-mean-square forecast error is about 11% below that of strategy 2—implying a 22%-lower mean squared error.[21]

Is real-time analysis worth the extra bother? If the only real-time approach being considered is strategy 3, table 3 suggests that the answer is "probably not." Thus, naive estimation and evaluation predicts the real-time forecasting performance of strategy 3 fairly well. In contrast, the naive approach gives a markedly too pessimistic view of the real-time performance of strategy 1.

The final two rows of table 3 compare the recursive forecasting performance of our monthly-indicators model to two benchmark alternatives—a purely autoregressive model (estimated using strategy 1) and the Blue Chip consensus forecast.[22] The autoregression performs poorly relative to equation (16) regardless of how the equation is estimated. The comparison with the Blue Chip consensus forecast is more interesting. It illustrates that a simple model, correctly estimated, can sometimes match the real-time performance of experienced, professional forecasters using conventional techniques. Thus, our strategy-1 forecasts do somewhat better than the Blue Chip forecasts over the period from 1990:Q1 to 1997:Q4, and strategy-2 forecasts do nearly as well. When estimated using strategy 3 or with current-vintage data, our model performs comparatively poorly.

Table 4 presents results of a formal comparison of how well various versions of the monthly indicators model perform relative to the Blue Chip consensus. In the left half of the table, recursive forecast errors from the indicators model are regressed on the Blue Chip forecast to see whether the Blue Chip forecast has marginal predictive power. In the right half of the table, similarly, we test whether the different versions of the indicators model have predictive power beyond the Blue Chip forecast. Although such encompassing tests are known to be problematic when predictions depend on estimated parameters, they are asymptotically valid when the forecasts under consideration are recursive and the models being compared are non-nested—conditions that are met here.[23]

Results displayed in the left half of table 4 indicate that the Blue Chip forecast contains no information beyond that captured by the monthly indicators model, regardless of how the model is estimated. In contrast, results vary considerably across estimation strategies when—as in the right half of table 4—the roles of the Blue Chip and model forecasts are reversed. In particular, an analyst using strategy-3 estimation would conclude from these results that the indicators model is of little help in predicting GDP growth given the Blue Chip survey. In actuality, the failure to improve on the Blue Chip survey is the fault of the data used to estimate the model rather than of the model per se. Thus, the strategy-1 and strategy-2 forecasts are statistically significant at the 10% and 5% levels, respectively, when used to predict Blue Chip survey errors.

F. *Why Does Strategy 1 Perform So Well?*

As reported above, strategy 1's mean squared forecast error is 22% below that of strategy 2. This difference is very much larger than the 9% improvement predicted by equation (15). Might it simply be the result of a lucky draw? To see, we conducted a series of 10,000 strategy-1 and strategy-2 estimation and recursive-forecasting exercises using simulated GDP data.[24] The average ratio of strategy-1 to strategy-2 squared errors in the Monte Carlo experiment was 0.80—insignificantly different from the observed ratio.[25] We conclude that our forecasting results are not atypical.

[21] Using first-release data raises the Strategy-1 RMSE to 1.58 percentage points. See footnote 6.

[22] The AR lag length is selected recursively by the Bayesian information criterion. Results change little if the lag length is held fixed, or if the model is estimated using strategy 2 or strategy 3.

[23] See West and McCracken (1998). Their result assumes that certain technical conditions are also met. Among them is the requirement that the estimation methodology yield unbiased parameter estimates. The unbiasedness condition is typically violated by strategy 3, but an analyst unfamiliar with our arguments would not be aware of this problem.

[24] For each estimation-forecasting exercise, artificial first-release GDP data were generated using a full-sample estimate of equation (3). Revisions to each quarter's GDP growth from one vintage to the next were assumed to be normally distributed white noise, with a variance matching that observed in the real-world data. Right-side data were real-time-vintage actual observations—the same data used in the empirical portion of the paper.

[25] The Monte Carlo forecast-variance ratio was 0.78 or below 42% of the time.

TABLE 4.—ENCOMPASSING TESTS FOR RECURSIVE FORECASTS OF CURRENT-VINTAGE GDP GROWTH, 1990:Q1–1997:Q4

	$y(T)$ − (strategy i) = θ × (Blue Chip)		$y(T)$ − (Blue Chip) = θ × (strategy i)	
Strategy	θ	θ = 0?	θ	θ = 0?
1	0.037 (0.097)	P = 0.703	0.171 (0.086)	P = 0.056
2	0.017 (0.110)	P = 0.880	0.168 (0.080)	P = 0.045
3	−0.050 (0.120)	P = 0.683	0.126 (0.080)	P = 0.126

Evidence on the likely source of strategy 1's unexpectedly strong performance is provided in table 5, which compares strategies 1 and 2 when the set of right-side variables is restricted to monthly changes in a single indicator variable. When the right-side variables are changes in employment or retail sales, the observed strategy-1 advantage is only slightly larger than that predicted by equation (15). However, when industrial production data are used to forecast GDP growth, strategy 1 performs markedly better than expected. Multicollinearity in the industrial production data may be the culprit: Strategy-2 regressions (especially those ending in the early 1990s) do a notably poor job of pinning down the coefficients attached to individual monthly changes in industrial production, even though their overall fit is very good.

We conclude that equation (15) is probably best regarded as a qualitative rather than a quantitative guide to strategy 1's performance advantage. Correlation between right-side variables—neglected in the equation—can be important in practice.

G. *Forecasting Advance and Final GDP*

The performance statistics presented up to this point have assumed that the analyst wants the best possible prediction of "true" GDP growth. However, early GDP releases may affect the decisions of households and businesses (and hence the future course of the economy) more than statistics released years after the fact. Accordingly, we briefly consider the performance of alternative estimation strategies in forecasting the first (*advance*) and third (*final*) GDP estimates, which become available one month and three months after the end of the quarter, respectively. Results are presented in table 6, which is similar in format to table 3.

Two main conclusions emerge from table 6. First, early GDP releases are easier to predict than are late releases. Thus, the root-mean-square errors recorded in table 6 are uniformly lower than the corresponding errors recorded in table 3. Within table 6, root-mean-square errors are lower when forecasting the advance GDP release than when forecasting the final release. These findings are consistent with in-sample estimation results reported in table 2, and are exactly what one would expect to see if revisions to the government's GDP estimates are unpredictable. The second main conclusion to emerge from table 6 is that the relative ranking of the alternative forecasts is largely unaffected by whether the goal is to predict an early or a late GDP release. Strategy 1 continues to produce root-mean-square errors substantially below those of strategy 3, and its performance advantage relative to the Blue Chip consensus forecast increases, if anything.[26] Performance differences between strategies 1 and 2 are small when predicting third-release GDP, because the differences between the left-side data used for estimation are also fairly small.

TABLE 5.—COMPARING THE FORECAST PERFORMANCE OF STRATEGIES 1 AND 2

Right-Side Variable	RMS Error: Strategy 1	RMS Error: Strategy 2	(MSE Strategy 1)/(MSE Strategy 2): Observed	(MSE Strategy 1)/(MSE Strategy 2): Equation (15)
Employment	1.56	1.59	0.96	0.98
Industrial production	1.64	1.85	0.78	0.96
Real retail sales	1.62	1.67	0.93	0.97

Observed and predicted forecast variances when the set of right-side variables is restricted.

TABLE 6.—SUMMARY STATISTICS FOR RECURSIVE FORECASTING EXERCISES, 1990:Q1–1997:Q4

Estimation Strategy	Mean Error	Mean Abs. Error	RMS Error
Advance GDP:			
Strategy 1[a]	0.11	0.67	0.80
Strategy 3	−0.05	0.99	1.15
Blue Chip	0.25	0.82	0.98
Final GDP:			
Strategy 1	0.20	0.88	1.06
Strategy 2[b]	0.25	0.88	1.09
Strategy 3	0.04	1.19	1.34
Blue Chip	0.34	1.02	1.24

[a] Identical to strategy 2.
[b] Final (third release) GDP growth used as the left-side variable when estimating the forecasting equation.

VII. Concluding Remarks

In most economic forecasting applications, the data that are substituted into the right-hand side of the forecasting equation to obtain an actual out-of-sample forecast have undergone few, if any, revisions. We have argued that this fact should be taken into account when the forecasting equation is first estimated. In particular, at each date within his sample the econometrician estimating a forecasting equation ought to use only right-side data that would have been available at the time. We call these real-time-vintage data. Real-time-vintage data sets are more complete than first-release data sets in that at each within-sample date they include revisions that would have been known at that date. For typical lag specifications (extending back a year or less), real-time-vintage data are readily available in back issues of government publications.

Most analysts do not use real-time vintage data to estimate their forecasting models. Instead, they use the most up-to-date numbers available at the time of the estimation. Economists often label as "real time" forecasting exercises in which this practice is reproduced after the fact. In these exercises, the economist gradually extends the period over which a forecasting equation is estimated, each time using data as they would have appeared at the close of the sample period. The problem with using end-of-sample-vintage data in this way is that correlation between revisions to the right-side variables and revisions to the left-side variable can make it appear that a forecasting relationship exists when, in fact, early vintages of the right-hand-side

[26] Encompassing tests similar to those reported in table 4 indicate that strategy-1 forecasts help explain Blue Chip forecast errors, but Blue Chip forecasts do not explain strategy-1 errors.

variables—the only vintages actually relevant to constructing current forecasts—have little or no marginal predictive power. In other words, the linkages between the right-side and left-side variables near the start of the sample period (where both are heavily revised) may be quite different from the linkages at the end of the sample period (where the available data have undergone little, if any, revision). It is only the latter linkages that are relevant for constructing an accurate current forecast.

A more subtle question than whether right-side variables in a forecasting equation ought to be of real-time vintage or end-of-sample vintage is whether the *left*-side variable ought to be of first-release or end-of-sample vintage. We have argued that as long as the government's first release fully exploits available information, superior forecasting performance can be expected if first-release data are used on the left-hand side of the equation during estimation. Intuitively, if the government's initial release is efficient, using it as the dependent variable strips unpredictable noise out of the equation and yields more precise coefficient estimates. First-release data are to be preferred for estimation even if the analyst is ultimately interested in predicting revised data.

In the application considered here—forecasting current-quarter GDP using monthly jobs, industrial production, and retail sales data—we find that our theoretical findings are borne out. A substantial improvement in out-of-sample performance is achieved if the forecasting equation is estimated with real-time-vintage data on its right-hand side, rather than end-of-sample-vintage data. There is a further improvement if first-release GDP growth is used as the left-side variable. Properly estimated, our simple model is competitive with the Blue Chip consensus GDP forecast.

REFERENCES

Blue Chip Economic Indicators, Capitol Publications, Alexandria, VA (various issues).

Braun, Steven N., "Estimation of Current-Quarter Gross National Product by Pooling Preliminary Labor-Market Data," *Journal of Business and Economic Statistics* 8 (1990), 293–304.

Croushore, Dean, and Tom Stark, "A Real-Time Data Set for Macroeconomists," Federal Reserve Bank of Philadelphia working paper no. 99-4 (1999).

Diebold, Francis X., and Glenn D. Rudebusch, "Forecasting Output with the Composite Leading Index: A Real-Time Analysis," *Journal of the American Statistical Association* 86 (1991), 603–610.

Ehrbeck, Tilman, and Robert Waldmann, "Why Are Professional Forecasters Biased? Agency versus Behavioral Explanations," *Quarterly Journal of Economics* 111 (1996), 21–40.

Fair, Ray C., and Robert J. Shiller, "Comparing Information in Forecasts from Econometric Models," *American Economic Review* 80, 375–389 (1990).

Fitzgerald, Terry J., and Preston J. Miller, "A Simple Way to Estimate Current-Quarter GNP," Federal Reserve Bank of Minneapolis *Quarterly Review* (Fall 1989), 27–31.

Graham, John R., "Is a Group of Economists Better than One? Than None?" *Journal of Business* 69 (1996), 193–232.

Hogg, Robert V., and Allen T. Craig, *Introduction to Mathematical Statistics* (New York: Macmillan, 1970).

Howrey, E. Phillip, "The Use of Preliminary Data in Econometric Forecasting," *Review of Economics and Statistics* 60 (1978), 193–201.

Ingenito, Robert, and Bharat Trehan, "Using Monthly Data to Predict Quarterly Output," Federal Reserve Bank of San Francisco *Economic Review,* No. 3 (1996), 3–11.

Laster, David, Paul Bennett, and In Sun Geoum, "Rational Bias in Economic Forecasts," *Quarterly Journal of Economics* 114 (1999), 293–318.

McNees, Stephen K., "Consensus Forecasts: Tyranny of the Majority," *New England Economic Review* (November-December 1987), 15–21.

Miller, Preston J., and Daniel M. Chin, "Using Monthly Data to Improve Quarterly Model Forecasts," Federal Reserve Bank of Minneapolis *Quarterly Review* (Spring 1996), 16–33.

Orphanides, Athanasios, "The Quest for Prosperity without Inflation," Board of Governors of the Federal Reserve System working paper (1999).

Rathjens, Peter, and Russell P. Robins, "Forecasting Quarterly Data Using Monthly Information," *Journal of Forecasting* 12 (1993), 321–330.

Robertson, John C., and Ellis W. Tallman, "Data Vintages and Measuring Forecast Model Performance," Federal Reserve Bank of Atlanta *Economic Review* (Fourth Quarter 1998), 4–20.

Rogers, R. Mark, "A Primer on Short-Term Linkages between Key Economic Data Series," Federal Reserve Bank of Atlanta *Economic Review* (Second Quarter 1998), 40–54.

Runkle, David E., "Revisionist History: How Data Revisions Distort Economic Policy Research," Federal Reserve Bank of Minneapolis *Quarterly Review* (Fall 1998), 3–12.

Swanson, Norman R., "Forecasting Using First Available versus Fully Revised Economic Time Series Data," Pennsylvania State University working paper no. 4-96-7 (1996).

Swanson, Norman R., and Halbert White, "Forecasting Economic Time Series Using Flexible versus Fixed Specification and Linear vs. Nonlinear Econometric Models," Pennsylvania State University manuscript (1996).

——— "A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks," *Review of Economics and Statistics* 79 (1997), 540–550.

Trehan, Bharat, "Forecasting Growth in Current Quarter Real GNP," Federal Reserve Bank of San Francisco *Economic Review* (Winter 1997), 39–52.

——— "Predicting Contemporaneous Output," Federal Reserve Bank of San Francisco *Economic Review*, no. 2 (1992), 3–11.

West, Kenneth D., and Michael W. McCracken, "Regression-Based Tests of Predictive Ability," *International Economic Review* 39 (1998), 817–840.

Zadrozny, Peter A., "Forecasting U.S. GNP at Monthly Intervals with an Estimated Bivariate Time Series Model," Federal Reserve Bank of Atlanta *Economic Review* (November-December 1990), 2–15.

APPENDIX

The Data Requirements of the Alternative Estimation Strategies—An Example

Suppose that we want to forecast GDP growth using current and two lags of jobs growth. Available GDP data extend from $t = 0$ to $t = T - 1$. As shown in table A1, we can arrange the data in a triangular array with

TABLE A1.—AVAILABLE LEFT-SIDE-VARIABLE DATA LAID OUT IN AN ARRAY

$\mathbf{y(0)_0}$	$y(0)_1$	$y(0)_2$	·	·	·	$y(0)_{T-2}$	$y(0)_{T-1}$		
	$\mathbf{y(1)_1}$	$y(1)_2$	·	·	·	$y(1)_{T-2}$	$y(1)_{T-1}$		
		$\mathbf{y(2)_2}$	·	·	·	$y(2)_{T-2}$	$y(2)_{T-1}$	More recent periods ↓	
						·	·		
						·	·		
						·	·		
						$\mathbf{y(T-2)_{T-2}}$	$y(T-2)_{T-1}$		
							$\mathbf{y(T-1)_{T-1}}$		

More recent vintages →

A typical entry, $y(t)_s$, is the official estimate of $y(t)$ released at time $s \geq t$. Strategy 1 uses only entries in boldfaced type (along the main diagonal) for estimation of the forecasting equation. Strategies 2 and 3 use only entries from the rightmost column.

TABLE A2.—AVAILABLE RIGHT-SIDE-VARIABLE DATA LAID OUT IN AN ARRAY

$\mathbf{e(-2)_0}$	$e(-2)_1$	$e(-2)_2$	· · ·	$e(-2)_{T-2}$	$e(-2)_{T-1}$	$e(-2)_T$	
$\mathbf{e(-1)_0}$	$\mathbf{e(-1)_1}$	$e(-1)_2$	· · ·	$e(-1)_{T-2}$	$e(-1)_{T-1}$	$e(-1)_T$	
$\mathbf{e(0)_0}$	$\mathbf{e(0)_1}$	$\mathbf{e(0)_2}$	· · ·	$e(0)_{T-2}$	$e(0)_{T-1}$	$e(0)_T$	
	$\mathbf{e(1)_1}$	$\mathbf{e(1)_2}$	· · ·	$e(1)_{T-2}$	$e(1)_{T-1}$	$e(1)_T$	
		$\mathbf{e(2)_2}$	· · ·	$e(2)_{T-2}$	$e(2)_{T-1}$	$e(2)_T$	More recent periods ↓
				·	·	·	
				·	·	·	
				$\mathbf{e(T-4)_{T-2}}$	·	·	
				$\mathbf{e(T-3)_{T-2}}$	$\mathbf{e(T-3)_{T-1}}$	·	
				$\mathbf{e(T-2)_{T-2}}$	$\mathbf{e(T-2)_{T-1}}$	$e(T-2)_T$	
					$\mathbf{e(T-1)_{T-1}}$	$e(T-1)_T$	
						$e(T)_T$	

More recent vintages →

In the notation of the main text, $\boldsymbol{x}(t)_s = [e(t)_s \;\; e(t-1)_s \;\; e(t-2)_s]$. A typical entry of the array, $e(t)_s$, is the official estimate of $e(t)$ released at time $s \geq t$. Strategies 1 and 2 use only the entries in boldface type for estimation. The bottom three entries of the rightmost column are then substituted into the estimated equation to obtain a forecasted value of $y(T)$. Strategy 3 uses only entries from the rightmost column for estimation and forecasting. First-release estimation uses only entries from the main diagonal.

T rows and T columns. In the upper left-hand corner is the period-0 official estimate of period-0 GDP growth $[y(0)_0]$. In the upper right-hand corner is the period-$T-1$ official estimate of period-0 GDP growth $[y(0)_{T-1}]$. And in the lower right-hand corner is the period-$T-1$ official estimate of period-$T-1$ GDP growth $[y(T-1)_{T-1}]$. More generally, increasingly up-to-date estimates of a particular quarter's GDP growth appear as one moves from left to right along a given row. Moving from top to bottom along a column, the *vintage* of the data stays constant, but one sees GDP growth in ever more recent periods (Diebold and Rudebusch, 1991). Strategy-1 estimation uses data from the main diagonal. Strategies 2 and 3 use data from the rightmost column.

Under strategy 1, the econometrician simply adds a new GDP growth observation to the end of his data set as T increases. In table A1, all that's needed is one new diagonal element at the lower right of the data array. Under strategies 2 and 3, the entire data set is discarded and replaced with a new set of GDP-growth observations of vintage T—an entire new column of data must be added to the data array. Suppose that the econometrician wishes to conduct an ex post, real-time recursive-forecast exercise. Under strategy 1, the econometrician need collect only *one* series of GDP growth estimates—a series consisting entirely of initial releases. Under strategies 2 and 3, the econometrician must collect a data set of vintage $T-1$ that covers GDP growth over the entire interval from $t = 0$ to $t = T-1$, a data set of vintage T that extends from $t = 0$ to $t = T$, and so forth.

The jobs-growth data in table A2 are organized much like the GDP-growth data in table A1, except that there are two additional rows at the top of the array (to accommodate lags of jobs growth) and one additional column (reflecting the availability of vintage-T jobs data). As before, strategy 3 uses only entries from the extreme right-hand column of the array. If the sample period is extended by one quarter, the old data set must be discarded and replaced with a new set of jobs-growth observations of vintage $T+1$. To conduct a real-time recursive-forecasting exercise, the econometrician must collect a sequence of long data sets, each of a different vintage.

Forecasting strategies 1 and 2 use jobs-growth data from the bottom three elements of each column, reflecting the fact that the jobs-growth terms that appear on the right-hand side of the forecasting equation at a given date are all of the same vintage. Extending the sample period by one quarter simply requires adding three new entries at the lower right of the jobs-growth array—an entirely new observation of period-$T+1$ jobs growth, and newly revised estimates of period-T and period-$T-1$ jobs growth. Conducting a real-time recursive-forecast exercise requires that the econometrician collect many three-element jobs-growth snapshots, each of a different vintage.

Restricting oneself to first-release right-side data amounts to taking data only from the main diagonal of the array, rather than from the main and two adjacent diagonals. To see how serial correlation can arise under this approach, consider the difference between the right-side variables under the two strategies. Using strategy 1, the right-side variables at time t are $e(t)_t$, $e(t-1)_t$, and $e(t-2)_t$. With first-release data, the right-side variables are $e(t)_t$, $e(t-1)_{t-1}$, and $e(t-2)_{t-2}$. The differences between the right-side variables under the two strategies are thus 0, $e(t-1)_t - e(t-1)_{t-1}$, and $e(t-2)_t - e(t-2)_{t-2} = [e(t-2)_t - e(t-2)_{t-1}] + \mathbf{[}\boldsymbol{e(t-2)_{t-1} - e(t-2)_{t-2}}\mathbf{]}$. At time $t-1$, the corresponding differences are 0, $\boldsymbol{e(t-2)_{t-1} - e(t-2)_{t-2}}$, and $e(t-3)_{t-1} - e(t-3)_{t-3}$. The overlap (the expression in boldface type), means that the information that is left out as one goes from strategy-1 to first-release estimation can be expected to be correlated over time.