Research Program on Forecasting

How Biased Are U.S. Government Forecasts of the Federal Debt?

Neil R. Ericsson
Division of International Finance
Board of Governors of the Federal Reserve System
Washington, DC 20551 USA
ericsson@frb.gov
and
Department of Economics
The George Washington University
Washington, DC 20052 USA
ericsson@gwu.edu

RPF Working Paper No. 2017-001
http://www2.gwu.edu/~forcpgm/2017-001.pdf

January 6, 2017

RESEARCH PROGRAM ON FORECASTING
Center of Economic Research
Department of Economics
The George Washington University
Washington, DC 20052
http://www2.gwu.edu/~forcpgm

HOW BIASED ARE U.S. GOVERNMENT FORECASTS OF THE FEDERAL DEBT?

Neil R. Ericsson*

January 6, 2017

Abstract: Government debt and forecasts thereof attracted considerable attention during the recent financial crisis. The current paper analyzes potential biases in different U.S. government agencies' one-year-ahead forecasts of U.S. gross federal debt over 1984–2012. Standard tests typically fail to detect biases in these forecasts. However, impulse indicator saturation (IIS) detects economically large and highly significant time-varying biases, particularly at turning points in the business cycle. These biases do not appear to be politically related. IIS defines a generic procedure for examining forecast properties; it explains why standard tests fail to detect bias; and it provides a mechanism for potentially improving forecasts.

Keywords: Autometrics, bias, debt, federal government, forecasts, impulse indicator saturation, heteroscedasticity, projections, United States.

JEL classifications: H68, C53.

*Forthcoming in the *International Journal of Forecasting* as the articles "How Biased Are U.S. Government Forecasts of the Federal Debt?" (Sections 1–7 below) and "Interpreting Estimates of Forecast Bias" (Appendix A below). Appendix B below lists the data and forecasts analyzed. An earlier version of this paper was titled "Detecting and Quantifying Biases in Government Forecasts of the U.S. Gross Federal Debt". The author is a staff economist in the Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551 USA, and a Research Professor of Economics, Department of Economics, The George Washington University, Washington, DC 20052 USA. He may be reached on the Internet at ericsson@frb.gov and ericsson@gwu.edu. The views in this paper are solely the responsibility of the author and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System. The author is grateful to Danny Bachman, Russell Davidson, Ed Gamber, David Hendry, Stedman Hood, Rob Hyndman, Søren Johansen, Fred Joutz, Andrew Kane, Kajal Lahiri, Jeffrey Liebner, Prakash Loungani, Aaron Markiewitz, Jaime Marquez, Andrew Martinez, Toshihiko Mukoyama, Bent Nielsen, Felix Pretis, John Rogers, Tara Sinclair, Herman Stekler, Ben Taylor, Christopher Williams, and two anonymous referees for helpful discussions and comments; and, in addition, to Stedman Hood for invaluable research assistance, and to Andrew Martinez for providing the data and forecasts analyzed and for stimulating my interest in this topic. Numerical results were obtained using Microsoft's 32-bit Excel 2013 and Doornik and Hendry's (2013) PcGive Version 14.1, Autometrics Version 1.5g, and Ox Professional Version 7.10 in 64-bit OxMetrics Version 7.10.

1 Introduction

Government debt attracted considerable attention during the recent financial crisis and Great Recession. In the United States, federal debt limits, sequestration, and the federal government shut-down have posed substantial economic, political, and policy challenges; see *The Economist* (November 20, 2010), Podkul (2011), Bernanke (2011, 2013), Chokshi (2013), and Yellen (2014, pp. 20–21) *inter alia*. In Europe, government debt and fiscal policy are central to current discussions about the euro-area crisis. Because future outcomes of government debt are unknown, forecasts of that debt may matter in government policy, so it is of interest to ascertain how good those forecasts are, and how they might be improved. A central focus in forecast evaluation is forecast bias, especially because forecast biases are systematic, and because ignored forecast biases may have substantive adverse consequences for policy.

Building on Martinez (2011, 2015), the current paper analyzes potential biases in different U.S. government agencies' one-year-ahead forecasts of the U.S. gross federal debt over 1984–2012. Standard tests typically do not detect biases in these forecasts. However, a recently developed technique—impulse indicator saturation—detects economically large and highly statistically significant time-varying biases in the forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. Biases differ according to the agency making the forecasts as well as over time. Biases are typically associated with turning points in the business cycle and (to a lesser degree) economic expansions, and thus are highly nonlinear and dynamic. That said, the forecast biases do not appear to be politically related. Impulse indicator saturation defines a generic procedure for examining forecast properties; it explains why standard tests fail to detect forecast bias; and it provides a mechanism for potentially improving the forecasts.

This paper is organized as follows. Section 2 describes the data and the forecasts being analyzed. Section 3 discusses different approaches to testing for potential forecast bias and proposes impulse indicator saturation as a generic test of forecast bias. Section 4 describes indicator saturation techniques, including impulse indicator saturation and several of its extensions. Section 5 presents evidence on forecast bias, using the methods detailed in Sections 3 and 4. Section 6 re-examines the forecast biases in light of business-cycle turning points. Section 7 concludes.

2 The Data and the Forecasts

This section describes the data on the United States gross federal debt and the three different one-year-ahead forecasts of that debt that are analyzed herein. The forecasts are denoted by their sources:

- CBO (Congressional Budget Office) in its *Budget and Economic Outlook*,
- OMB (Office of Management and Budget) in its *Budget of the U.S. Government*, and

- APB (*Analysis of the President's Budget*).

The Congressional Budget Office and the Office of Management and Budget are different agencies within the U.S. federal government. The *Analysis of the President's Budget* is produced by the Congressional Budget Office, but the forecast in the *Analysis of the President's Budget* is referred to as the "APB forecast" in order to distinguish it from the "CBO forecast", which appears in the CBO's *Budget and Economic Outlook*. The agencies' publications detail how debt is forecast and the assumptions made in generating those forecasts. Significantly, the CBO forecast assumes that current law remains unchanged, whereas the OMB and APB forecasts assume that the president's proposed budget is implemented. The assumptions underlying the forecasts, the complex process involved in generating the forecasts, and the goals and objectives of that process are of considerable interest in their own right and merit detailed examination. However, in the spirit of Stekler (1972), Chong and Hendry (1986), and Fildes and Stekler (2002) *inter alia*, the current paper focuses on the properties of the forecasts themselves. The data on the debt are published by the Financial Management Service at the U.S. Department of the Treasury in the *Treasury Bulletin*.

The data on debt are annual (end of fiscal year) over 1984–2012 (29 observations) and are for total gross federal debt outstanding held by the public and the government. The CBO, OMB, and APB forecasts typically are published in late January, early February, and early March respectively, where those months directly precede the end of the fiscal year (September 30); see Martinez (2011, Table 2; 2015) for details. For convenience, these forecasts are called "one-year-ahead", even though the actual horizon is somewhat less than one year, differs for the three forecasts, and varies somewhat from one year to the next. Debt and its forecasts are in billions of U.S. dollars (nominal), and the analysis below is of the logs of debt and of its forecasts.

Figure 1 plots actual U.S. gross federal debt and its forecasts by the CBO, OMB, and APB (in logs, denoted by lowercase). Actual and forecast values appear close, reflecting in part the scale of the graph: debt increases by approximately an order of magnitude over the sample. Figure 2 plots the forecast errors for the log of U.S. gross federal debt. The forecast errors for all three forecasts are often small—under 2% in absolute value—but sometimes they are much larger, and with the magnitude and even the sign differing across agency as well as by forecast date. Forecast errors are often persistent, suggestive of systematic biases in the forecasts. For comparison, the growth rate of debt is 8.3% on average, and its standard deviation is 4.1%.

The presence of forecast bias has both economic significance and statistical significance. That said, the particular sense in which forecast bias is significant depends in part on whether an agency's forecasts are interpreted as "forecasts" or as "projections", where "projections" are in the sense of being policy simulations conditional upon a certain set of assumptions. If the agency's forecasts are interpreted *qua* forecasts, then forecast bias implies potential room for improvement in terms of standard

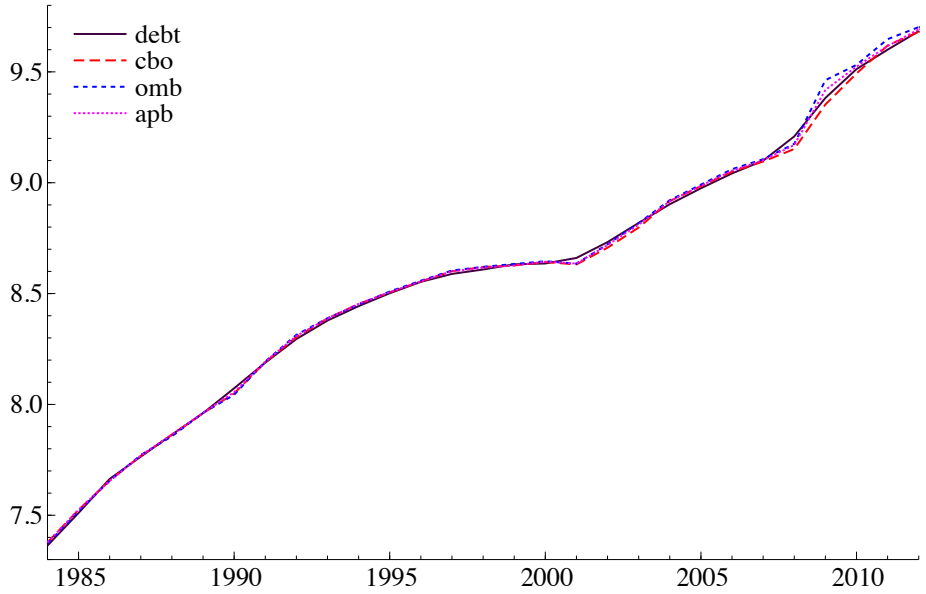

Figure 1: Actual U.S. gross federal debt and its forecasts by the CBO, OMB, and APB (in logs).

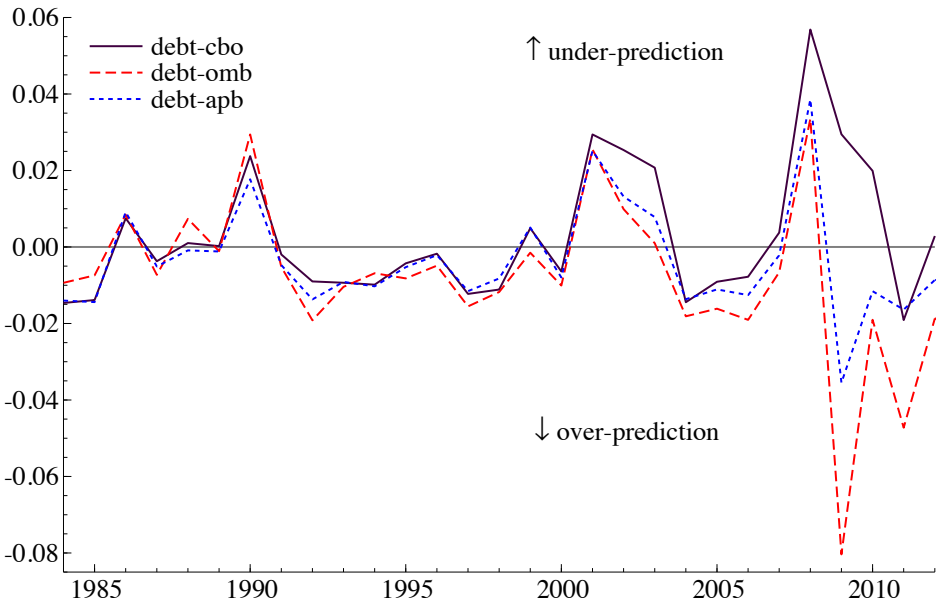

Figure 2: Forecast errors for the log of U.S. gross federal debt.

performance measures such as the root mean squared error. If the forecasts are interpreted *qua* projections, then forecast bias implies a limited usefulness of the forecasts as representing interesting hypothetical paths for economic policy. With that in mind, the agencies' forecasts are always referred to as "forecasts" below, while recognizing that some of these forecasts may be more usefully viewed as projections. This broader usage of the term "forecast" is also in line with Clements and Hendry (2002b, p. 2): "A forecast is any statement about the future". For some previous analyses of these and other governmental and institutional forecasts, see Corder (2005), Engstrom and Kernell (1999), Frankel (2011), Joutz and Stekler (2000), Nunes (2013), Sinclair, Joutz, and Stekler (2010), Romer and Romer (2008), and Tsuchiya (2013). Finally, many prior studies have compared forecasts whose assumptions differ from each other. Hence, the differing assumptions of the CBO, OMB, and APB forecasts are not grounds *per se* for not comparing the forecasts.

3 Approaches to Detecting Forecast Bias

This section considers different approaches for assessing potential forecast bias, starting with the standard test of forecast bias by Mincer and Zarnowitz (1969). This section then discusses how Chong and Hendry's (1986) forecast-encompassing test is interpretable as a test of time-varying forecast bias. Finally, this section proposes using impulse indicator saturation as a generic test of arbitrarily time-varying forecast bias. This generic test generalizes the Mincer–Zarnowitz test, which is a test of a constant (i.e., time-invariant) forecast bias.

Mincer and Zarnowitz (1969, pp. 8–11) suggest testing for forecast bias by regressing the forecast error on an intercept and testing whether the intercept is statistically significant. That is, for a variable y_t at time t and its forecast $\hat{y}_t$, estimate the equation:

$$(y_t - \hat{y}_t) \;=\; a \;+\; e_t \qquad t = 1, \ldots, T, \tag{1}$$

where a is the intercept, e_t is the error term at time t, and T is the number of observations. A test of $a = 0$ is interpretable as a test that the forecast $\hat{y}_t$ is unbiased for the variable y_t. For one-step ahead forecasts, the error e_t may be serially uncorrelated, in which case a t- or F-statistic for $a = 0$ may be appropriate. For multi-step ahead forecasts, e_t generally will be serially correlated; hence, inference about the intercept a may require some accounting for that autocorrelation.

Mincer and Zarnowitz (1969, p. 11) also propose a variant of equation (1) in which the coefficient on $\hat{y}_t$ is estimated rather than imposed. That variant is:

$$y_t \;=\; a_0 \;+ a_1 \hat{y}_t + \; e_t \qquad t = 1, \ldots, T, \tag{2}$$

where a_0 is the intercept, and a_1 is the coefficient on $\hat{y}_t$. Mincer and Zarnowitz (1969) interpret a test that $a_1 = 1$ as a test of the efficiency of the forecast $\hat{y}_t$ for

the outcome y_t. The joint hypothesis $\{a_0 = 0, a_1 = 1\}$ is of interest to test as well. Subtracting $\hat{y}_t$ from both sides, equation (2) may be conveniently rewritten as:

$$(y_t - \hat{y}_t) \;=\; a_0 \;+ a_1^* \hat{y}_t + \; e_t \qquad t = 1, \ldots, T, \tag{3}$$

where $a_1^* = a_1 - 1$. Hence, the hypothesis $\{a_0 = 0, a_1^* = 0\}$ in equation (3) is equivalent to $\{a_0 = 0, a_1 = 1\}$ in equation (2).

Below, "Mincer–Zarnowitz A" denotes the regression-based test of $a = 0$ in equation (1), whereas "Mincer–Zarnowitz B" denotes the regression-based test of $\{a_0 = 0, a_1^* = 0\}$ in equation (3). While equations (2) and (3) are equivalent, equation (3) is reported below because it parallels the structure of equation (1), with $y_t - \hat{y}_t$ as the dependent variable. Mincer–Zarnowitz A (i.e., testing $a = 0$ in equation (1)) is itself equivalent to testing $a_0 = 0$ in equation (3), subject to the restriction that $a_1^* = 0$. See Holden and Peel (1990) and Stekler (2002) for expositions on these tests as tests of unbiasedness and efficiency, and Sinclair, Stekler, and Carnow (2012) for a recent discussion.

Chong and Hendry (1986) propose another test about forecast errors, namely, a test of whether one model's forecasts provide information about another model's forecast errors. If one model's forecasts do provide information about another model's forecast errors, then those forecast errors are in part predictable. If not, then the latter model "forecast-encompasses" the first model. As Ericsson (1992) discusses, a necessary condition for forecast encompassing is having the smallest mean squared forecast error (MSFE). Granger (1989) and Diebold and Mariano (1995) propose tests of whether one model's MSFE is less than another model's MSFE.

Chong and Hendry (1986) and subsequent authors implement many versions of the forecast-encompassing test. One appealing version is based on the regression:

$$\begin{aligned} (y_t - \hat{y}_t) \;&=\; b_0 \;+\; b_1 \cdot (\tilde{y}_t - \hat{y}_t) \;+\; e_t \\ &=\; a_t \;+\; e_t \qquad\qquad\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{4}$$

where $\hat{y}_t$ is the forecast of y_t by model 1 (say), $\tilde{y}_t$ is the forecast of y_t by model 2, and b_0 and b_1 are regression coefficients. A test of $b_1 = 0$ is interpretable as a test of whether discrepancies between the two models' forecasts are helpful in explaining model 1's forecast errors. The joint hypothesis $\{b_0 = 0, b_1 = 0\}$ is also of interest to test. Equation (4) can be extended to compare several forecasts at once, in which case the right-hand side of equation (4) includes the differential of each alternative model's forecast relative to model 1's forecast; see Ericsson and Marquez (1993).

Tests of forecast encompassing are interpretable as tests of time-varying forecast bias, as the second line in equation (4) indicates. The subscript t on the intercept a_t emphasizes the time dependence of the potential bias, which here is parameterized as $b_0 + b_1 \cdot (\tilde{y}_t - \hat{y}_t)$. The forecast-encompassing test thus focuses on a specific time-varying form of potential forecast bias.

The time dependence of the forecast bias could be completely general, as follows:

$$\begin{aligned}(y_t - \hat{y}_t) &= \textstyle\sum_{i=1}^{T} c_i I_{it} \; + \; e_t \\ &= a_t \; + \; e_t \qquad\qquad\qquad t = 1, \ldots, T, \end{aligned} \tag{5}$$

where the impulse indicator I_{it} is a dummy variable that is unity for $t = i$ and zero otherwise, and c_i is the corresponding coefficient for I_{it}. Because the $\{c_i\}$ may have any values whatsoever, the intercept a_t in equation (5) may vary arbitrarily over time. In this context, a test that all coefficients c_i are equal to zero is a generic test of forecast unbiasedness. Because equation (5) includes T coefficients, equation (5) cannot be estimated unrestrictedly. However, the question being asked can be answered by using impulse indicator saturation, as is discussed in the following section.

4 Indicator Saturation Techniques

Impulse indicator saturation (IIS) is a general procedure for model evaluation, and in particular for testing parameter constancy. As this section shows, IIS also can be used to test for time-varying forecast bias. Doing so provides a new application of impulse indicator saturation—as a generic test of forecast bias—noting that IIS has previously been employed for model evaluation, model design, and robust estimation. Section 4.1 discusses IIS and its extensions as a procedure for testing parameter constancy. Section 4.2 re-interprets existing tests of forecast bias as special cases of IIS and shows how IIS can be used to detect arbitrarily time-varying forecast bias. Sections 5 and 6 then apply IIS and its extensions to analyze potential bias in forecasts of the U.S. gross federal debt.

4.1 Impulse Indicator Saturation and Extensions

This subsection summarizes how impulse indicator saturation provides a general procedure for analyzing a model's constancy. Specifically, IIS is a generic test for an unknown number of breaks, occurring at unknown times anywhere in the sample, with unknown duration, magnitude, and functional form. IIS is a powerful empirical tool for both evaluating and improving existing empirical models. Hendry (1999) proposed IIS as a procedure for testing parameter constancy. See Hendry, Johansen, and Santos (2008), Doornik (2009a), Johansen and Nielsen (2009, 2013), Hendry and Santos (2010), Ericsson (2011a, 2011b, 2012, 2016), Ericsson and Reisman (2012), Bergamelli and Urga (2014), Hendry and Pretis (2013), Hendry and Doornik (2014), Castle, Doornik, Hendry, and Pretis (2015), and Marczak and Proietti (2016) for further discussion and recent developments.

Impulse indicator saturation uses the zero–one impulse indicator dummies $\{I_{it}\}$ to analyze properties of a model. For a sample of T observations, there are T such dummies, so the unrestricted inclusion of all T dummies in an estimated model (thereby

"saturating" the sample) is infeasible. However, blocks of dummies *can* be included, and that insight provides the basis for IIS. To motivate how IIS is implemented in practice, this subsection employs a bare-bones version of IIS in two simple Monte Carlo examples.

Example 1. This example illustrates the behavior of IIS when the model is correctly specified. Suppose that the data generation process (DGP) for the variable w_t is:

$$w_t = \mu_0 + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2), \qquad t = 1, \ldots, T, \tag{6}$$

where w_t is normally and independently distributed with mean μ_0 and variance σ^2. Furthermore, suppose that the model estimated is a regression of w_t on an intercept, i.e., the model is correctly specified. Figure 3a plots Monte Carlo data from the DGP in equation (6) with $\mu_0 = 20$, $\sigma^2 = 1$, and $T = 100$. Figure 3b plots the estimated model's residuals, scaled by that model's residual standard error.

The bare-bones version of IIS is as follows.

1. Estimate the model, including impulse indicator dummies for the first half of the sample, as represented by Figure 4a. That estimation is equivalent to estimating the model over the second half of the sample, ignoring the first half. Drop all statistically insignificant impulse indicator dummies and retain the statistically significant ones (Figure 4b).
2. Repeat this process, but start by including impulse indicator dummies for the *second* half of the sample (Figure 4d), and retain the significant ones (Figure 4e).
3. Re-estimate the original model, including all dummies retained in the two block searches (Figure 4g), and select the statistically significant dummies from that combined set (Figure 4h).

Hendry, Johansen, and Santos (2008) and Johansen and Nielsen (2009) have shown that, under the null hypothesis of correct specification, the expected number of impulse indicator dummies retained is roughly αT, where α is the target size. In Figure 4h, five dummies are retained; $\alpha = 5\%$; and $\alpha T = (5\% \cdot 100) = 5$, an exact match.

Example 2. This example illustrates the behavior of IIS when there is an unmodeled break. Suppose that the DGP for the variable w_t is:

$$w_t = \mu_0 + \mu_1 S_{64t} + \varepsilon_t \quad \varepsilon_t \sim NID(0, \sigma^2), \qquad t = 1, \ldots, T, \tag{7}$$

where S_{64t} is a one-off step dummy that is equal to 0 ($t = 1, \ldots, 63$) or 1 ($t = 64, \ldots, 100$), and μ_1 is its coefficient in the DGP. The model estimated is a regression of w_t on an intercept alone, ignoring the break induced by the step dummy S_{64t}. As in Example 1, w_t is normally and independently distributed with a nonzero mean. However, that mean alters at $t = 64$. The model ignores that change in mean (aka

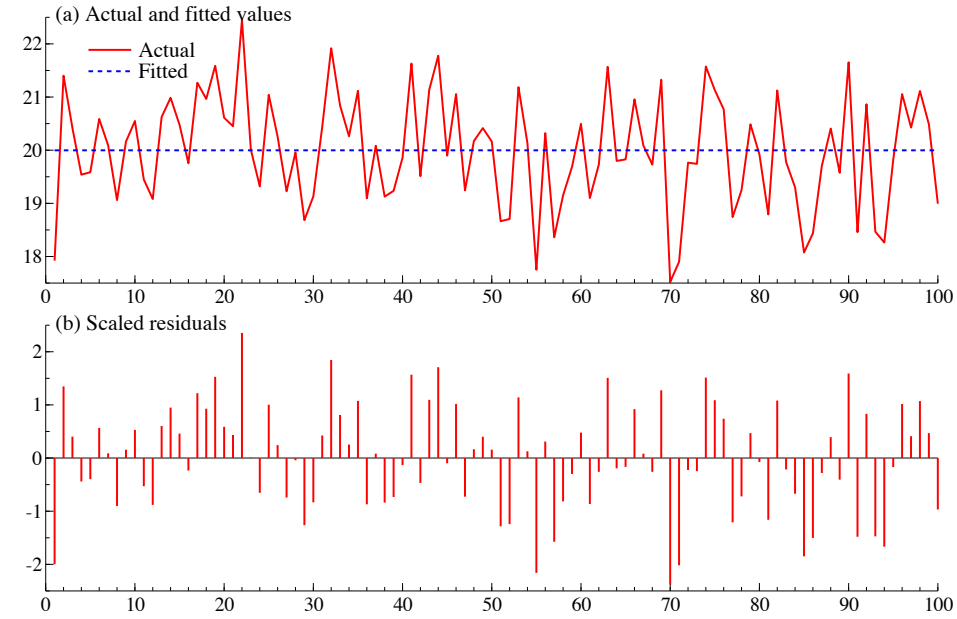

Figure 3: Actual and fitted values and the corresponding scaled residuals for the estimated model when the DGP does not have a break.

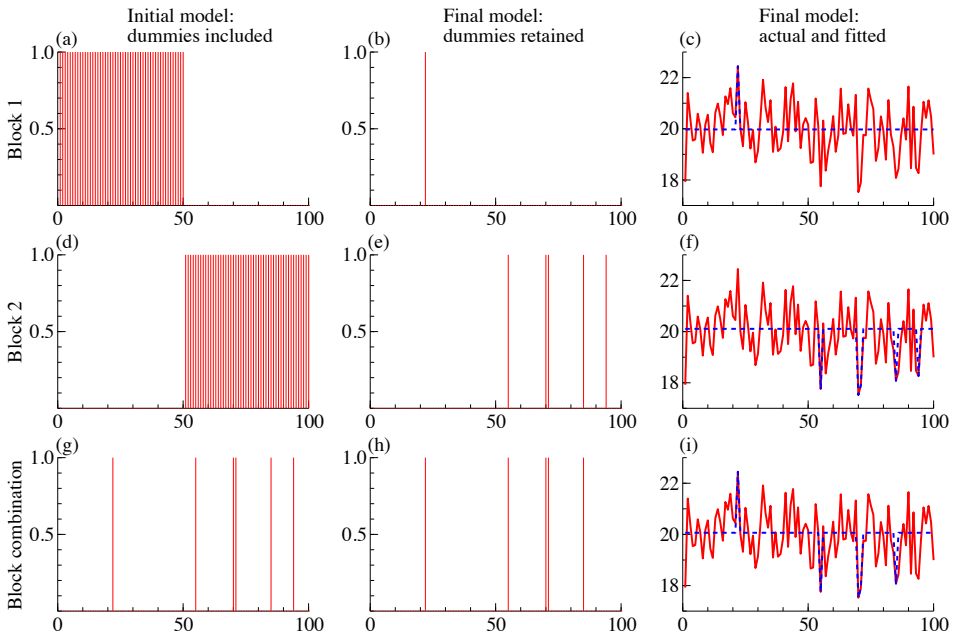

Figure 4: A characterization of bare-bones impulse indicator saturation with a target size of 5% when the DGP does not have a break.

a "location shift") and hence is mis-specified. Figure 5a plots Monte Carlo data from the DGP in equation (7) with $\mu_0 = 20$, $\mu_1 = -10$, $\sigma^2 = 1$, and $T = 100$. Figure 5b plots the estimated model's residuals. Interestingly, no residuals lie outside the estimated 95% confidence region, even though the break is -10σ. The model has no "outliers".

Figure 6 plots the corresponding graphs for the bare-bones implementation of IIS described in Example 1, as applied to the Monte Carlo data in Example 2. As the penultimate graph (Figure 6h) shows, the procedure has high power to detect the break, even although the nature of the break is not utilized in the procedure itself.

In practice, IIS as an algorithm may be more complicated than this bare-bones version, which employs two equally sized blocks, selects dummies by t-tests, and is non-iterative. In Doornik and Hendry's (2013) Autometrics econometrics software, IIS utilizes many possibly unequally sized blocks, rather than just two blocks; the partitioning of the sample into blocks may vary over iterations of searches; dummy selection includes F-tests against a general model; and residual diagnostics help guide model selection. Notably, the specific algorithm for IIS can make or break IIS's usefulness; cf. Doornik (2009a), Castle, Fawcett, and Hendry (2010), and Hendry and Doornik (2014). IIS is a statistically valid procedure for integrated, cointegrated data; see Johansen and Nielsen (2009). IIS can serve as a diagnostic statistic, and it can aid in model development, as discussed in Ericsson (2011a).

Many existing procedures can be interpreted as special cases of IIS in that they represent particular algorithmic implementations of IIS. Such special cases include recursive estimation, rolling regression, the Chow (1960) predictive failure statistic (including the 1-step, breakpoint, and forecast versions implemented in OxMetrics), the Andrews (1993) unknown breakpoint test, the Bai and Perron (1998) multiple breakpoint test, tests of extended constancy in Ericsson, Hendry, and Prestwich (1998, pp. 305ff), tests of nonlinearity, intercept correction (in forecasting), and robust estimation. IIS thus provides a general and generic procedure for analyzing a model's constancy. Algorithmically, IIS also solves the problem of having more potential regressors than observations by testing and selecting over blocks of variables.

Table 1 summarizes IIS and two extensions of IIS, drawing on expositions and developments in Ericsson (2011b, 2012) and Ericsson and Reisman (2012). Throughout, T is the sample size, t is the index for time, i and j are the indexes for indicators, k is the index for economic variables (denoted x_{kt}), and K is the total number of potential regressors considered. A few remarks may be helpful for interpreting the entries in Table 1.

Impulse indicator saturation. This is the standard IIS procedure proposed by Hendry (1999), with selection among the T zero–one impulse indicators $\{I_{it}\}$.

Super saturation. Super saturation searches across all possible one-off step functions $\{S_{it}\}$, in addition to $\{I_{it}\}$. Step functions are of economic interest because they

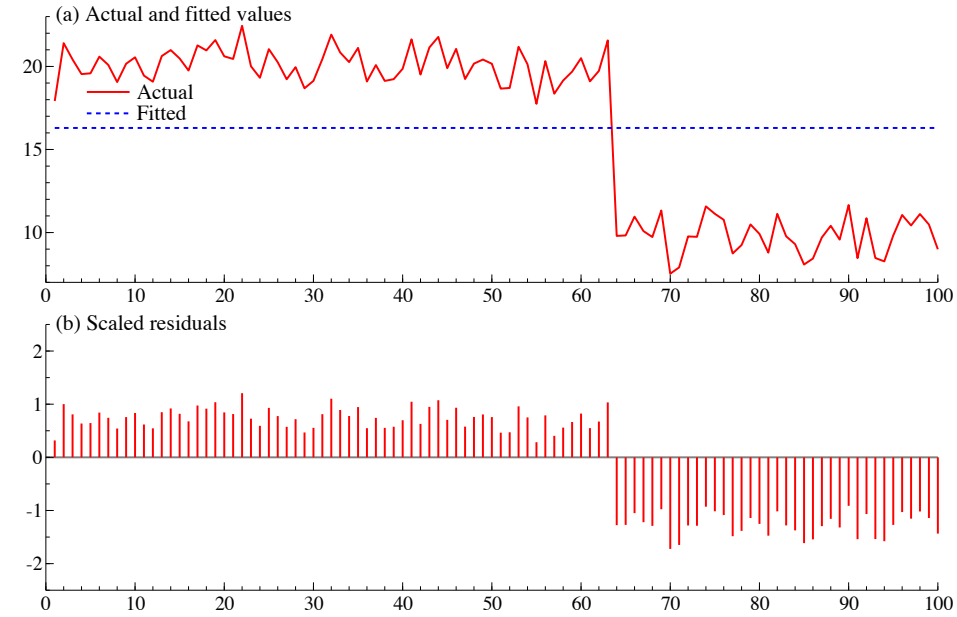

Figure 5: Actual and fitted values and the corresponding scaled residuals for the estimated model when the DGP has a break and the model ignores that break.

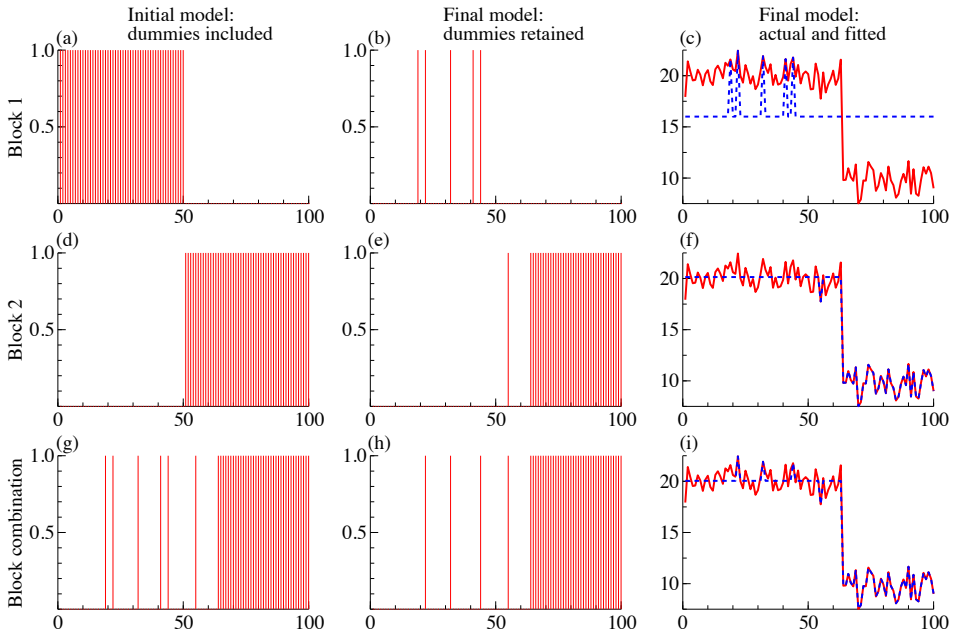

Figure 6: A characterization of bare-bones impulse indicator saturation with a target size of 5% when the DGP has a break and the model ignores that break.

Table 1: Impulse indicator saturation and two extensions, as characterized by the variables involved.

Name	Description	Variables	Definition
Impulse indicator saturation	Zero–one dummies	$\{I_{it}\}$	$I_{it} = 1$ for $t = i$, zero otherwise
Super saturation	Step functions	$\{I_{it}, S_{it}\}$	$S_{it} = 1$ for $t \geq i$, zero otherwise
Ultra saturation	Broken linear trends	$\{I_{it}, S_{it}, T_{it}\}$	$T_{it} = t - i + 1$ for $t \geq i$, zero otherwise

may capture permanent or long-lasting changes that are not otherwise incorporated into a specific empirical model. A step function is a partial sum of impulse indicators. Equivalently, a step function is a parsimonious representation of a sequential subset of impulse indicators that have equal coefficients. Castle, Doornik, Hendry, and Pretis (2015) investigate the statistical properties of a closely related saturation estimator—step indicator saturation (SIS)—which searches among only the step indicator variables $\{S_{it}\}$. Autometrics now includes IIS, SIS, super saturation (IIS+SIS), and zero-sum pairwise IIS (mentioned below); see Doornik and Hendry (2013).

Ultra saturation. Ultra saturation (earlier, sometimes called "super duper" saturation) searches across $\{I_{it}, S_{it}, T_{it}\}$, where the $\{T_{it}\}$ are broken linear trends. Broken linear trends may be of economic interest. Mathematically, the $\{T_{it}\}$ are partial sums of the partial sums of impulse indicators. Broken quadratic trends, broken cubic trends, and higher-order broken trends are also feasible.

Table 1 is by no means an exhaustive list of extensions to IIS. Other extensions include sequential ($j = 1$) and non-sequential ($j > 1$) pairwise impulse indicator saturation for an indicator P_{it}, defined as $I_{it} + I_{i+j,t}$; zero-sum pairwise IIS for an indicator Z_{it}, defined as ΔI_{it}; many many variables for a set of K potential regressors $\{x_{kt},\ k = 1, \ldots, K\}$ for $K > T$; factors; principal components; and multiplicative indicator saturation for the set of $S_{it}x_{kt}$. See Ericsson (2011b, 2012) and Castle, Clements, and Hendry (2013) for details, discussion, and examples in the literature. Also, the saturation procedure chosen may itself be a combination of extensions; and that choice may affect the power of the procedure to detect specific alternatives. For instance, in Example 2 above, the 37 impulse indicators $\{I_{it},\ i = 64, \ldots, 100\}$ are not a particularly parsimonious way of expressing the step shift that occurs two thirds of the way through the sample, whereas the single one-off step dummy S_{64t} is.

4.2 Re-interpretation and Generalization

This subsection discusses how IIS and its extensions provide a conceptual framework for re-interpreting existing tests of forecast bias. Equally, saturation procedures generalize those existing tests to allow for arbitrarily time-varying forecast bias.

For instance, the Mincer–Zarnowitz A test (based on equation (1)) is a special case of super saturation in which only the step dummy S_{1t} (equivalent to the intercept) is included. The Mincer–Zarnowitz A test is also interpretable as the IIS test based on equation (5), but where $c_1 = c_2 = \ldots = c_T$ is *imposed*, and the hypothesis $c_1 = 0$ is tested. The Mincer–Zarnowitz B test (based on equation (3)) is a special case of multiplicative indicator saturation in which the dependent variable is the forecast error, the x's are the intercept and the forecast, and the only multiplicative indicators considered are those multiplied by the step indicator S_{1t}. Multiplicative indicator saturation also includes the forecast encompassing test and standard tests of strong efficiency as special cases; cf. Holden and Peel (1990) and Stekler (2002).

As equation (5) entails, saturation-based tests generalize the Mincer–Zarnowitz tests to allow for time-varying forecast bias. This observation and the observations above highlight the strength of the Mincer–Zarnowitz tests (that they focus on detecting a constant nonzero forecast bias) and also their weakness (that they assume that the forecast bias *is* constant over time). These characteristics of the Mincer–Zarnowitz tests bear directly on the empirical results in the next two sections.

Certain challenges arise when interpreting a saturation-based test as a test of forecast *bias*. Specifically, saturation-based tests can detect not only time-varying forecast bias but also other forms of mis-specification, as reflected by discrepancies between the actual data and their assumed distribution as implied by the model. Such mis-specifications include outliers due to heteroscedasticity (as from a change in the forecast error variance) and thick tails (thick, relative to the assumed distribution). IIS's ability to detect many forms of mis-specification is thus a caveat for the *interpretation* of IIS results *per se*. Two items can help resolve this interpretational challenge: the retained dummies themselves, and outside information.

First, the structure of the retained dummies may have implications for their interpretation. For instance, for mis-specification due to heteroscedasticity or thick tails, retained impulses typically would not be sequential or—even if they were—would not be of the same sign and of similar magnitude. Because (e.g.) step indicators characterize sequential, same-signed, same-magnitude features, any retained step indicators from super saturation would be unlikely to arise from heteroscedasticity or thick tails. Hence, the interpretational caveat may not be germane to extended forms of IIS such as super saturation. In that light, saturation procedures can serve as tools for characterizing time-varying forecast bias *qua* bias, rather than as some unknown form of mis-specification. Saturation procedures thus provide a generic approach to estimating time-varying forecast bias, albeit a generic approach that is atheoretical,

economically speaking.

Second, outside information—such as from economic, institutional, and historical knowledge—may assist in interpreting saturation-based results. For instance, Section 6 integrates saturation procedures with an economically based interpretation of the estimated biases in light of the dates of business-cycle turning points. Features of the government's budget may imply systematic forecast errors (i.e., biases) at business-cycle turning points. Such an economic interpretation holds, even although impulse (rather than step) dummies statistically characterize the time-varying forecast bias.

As a more general observation, different types of indicators are adept at characterizing different sorts of bias: impulse dummies $\{I_{it}\}$ for date-specific anomalies, step dummies $\{S_{it}\}$ for level shifts, and broken trends $\{T_{it}\}$ for evolving developments. Transformations of the variable being forecast also may affect the interpretation of the retained indicators. For instance, an impulse dummy for a growth rate implies a level shift in the (log) level of the variable.

Saturation-based tests of forecast bias can serve both as diagnostic tools to detect what is wrong with the forecasts, and as developmental tools to suggest how the forecasts can be improved. Clearly, "rejection of the null doesn't imply the alternative". However, for time series data, the date-specific nature of saturation procedures can aid in identifying important sources of forecast error. Use of these tests in forecast development is consistent with a progressive modeling approach; see White (1990) and Hendry and Doornik (2014).

5 Evidence on Biases in the Forecasts of Debt

This section examines the CBO, OMB, and APB forecasts of U.S. gross federal debt for potential bias over 1984–2012. Standard (Mincer–Zarnowitz) tests of forecast bias typically fail to detect economically and statistically important biases. By contrast, saturation-based tests detect large time-varying biases in the CBO, OMB, and APB forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. Forecast biases for a given year differ numerically across the CBO, OMB, and APB, albeit with some similarities.

Table 2 reports the Mincer–Zarnowitz regressions in equations (1) and (3) for the CBO, OMB, and APB forecasts, with columns alternating between the "A" and "B" versions of the Mincer–Zarnowitz regression. Here and in subsequent tables, estimated standard errors appear in parentheses $(\cdot)$ under regression coefficients, t-ratios appear in curly brackets $\{\cdot\}$, p-values appear in square brackets $[\cdot]$, and $\hat{\sigma}$ denotes the residual standard error. For the Mincer–Zarnowitz test statistic in Table 2, and for other test statistics here and below, the entries within a given block of numbers are the F-statistic for testing the null hypothesis against the designated main-

Table 2: Coefficients, estimated standard errors, t-ratios, and summary statistics for Mincer–Zarnowitz A and B regressions of the CBO, OMB, and APB forecast errors.

Regressor or statistic	CBO	CBO	OMB	OMB	APB	APB
Intercept	0.27 (0.33) {0.82}	−6.07 (4.60) {−1.32}	−0.79 (0.40) {−1.99}	11.56 (5.15) {2.24}	−0.36 (0.27) {−1.34}	1.89 (3.82) {0.50}
Forecast $\hat{y}_t$	–	0.0074 (0.0054) {1.38}	–	−0.0144 (0.0060) {−2.40}	–	−0.0026 (0.0044) {−0.59}
$\hat{\sigma}$	1.757%	1.729%	2.136%	1.974%	1.432%	1.449%
RMSE of the forecast	1.746%	1.746%	2.243%	2.243%	1.452%	1.452%
Mincer–Zarnowitz test statistic	0.67 [0.421] $F(1, 28)$	1.30 [0.290] $F(2, 27)$	3.96 [0.056] $F(1, 28)$	5.21* [0.012] $F(2, 27)$	1.80 [0.191] $F(1, 28)$	1.05 [0.363] $F(2, 27)$
Normality statistic	10.6** [0.005] $\chi^2(2)$	5.40 [0.067] $\chi^2(2)$	13.5** [0.001] $\chi^2(2)$	17.2** [0.000] $\chi^2(2)$	6.00* [0.050] $\chi^2(2)$	6.79* [0.034] $\chi^2(2)$
Variance instability statistic	0.38	0.48*	0.43	0.39	0.37	0.30

tained hypothesis, the tail probability associated with that value of the F-statistic, the degrees of freedom for the F-statistic (in parentheses), and (for saturation-based statistics) the retained dummy variables. Superscript asterisks * and ** denote rejections of the null hypothesis at the 5% and 1% levels respectively, and the null hypothesis typically includes setting the coefficient on the intercept to zero. Doornik and Hendry (2013) provide a description of the residual diagnostic statistics. For the saturation-based statistics reported below, K is the number of *potential* regressors for selection, and the target size is chosen much smaller than $1/K$ in order to help ensure that few if any indicators are retained fortuitously.

The results in Table 2 provide little evidence of forecast bias for any of the forecasts. From the first column for the CBO, the estimate of the forecast bias a in equation (1) is 0.27, which is statistically insignificantly different from zero, with an F-statistic of 0.67. From the second column for the CBO, the estimates of a_0 and

a_1^* in equation (3) are -6.07 and 0.0074, which are individually insignificant with t-statistics of -1.32 and 1.38, and jointly insignificant with an F-statistic of 1.30. The Mincer–Zarnowitz statistics for OMB and APB are likewise insignificant, except that the Mincer–Zarnowitz B statistic for OMB is significant at around the 1% level. Thus, the Mincer–Zarnowitz A test fails to detect bias in all three forecasts, and the Mincer–Zarnowitz B test fails to detect bias in two of three forecasts. Standard tests thus provide little evidence of forecast bias.

Table 3 reports forecast-encompassing statistics and saturation-based test statistics of forecast bias for the CBO, OMB, and APB forecasts. Table 3 also includes the Mincer–Zarnowitz statistics for comparison. The forecast-encompassing statistic detects bias for all three forecasts; cf. Martinez (2011, 2015). Likewise, IIS and its extensions *always* detect bias, and they do so for historically and economically consequential years. The dates of several retained impulse and step dummies are indicative of the following important events that potentially affected the actual federal debt after its forecasts were made.

1990: Iraq invasion of Kuwait on August 2, 1990; July 1990–March 1991 recession.

2001: March–November 2001 recession; September 11, 2001.

2008, 2009: December 2007–June 2009 recession.

Recessions are dated per the National Bureau of Economic Research (2012). Business-cycle turning points are prominent among the events listed. The four years listed also highlight the difficulties in forecasting the debt, especially in light of unanticipated events that affect both government expenditures and government revenues; cf. Alexander and Stekler (1959) and Stekler (1967).

The saturation-based tests in Table 3 focus on the statistical significance of the biases for each set of forecasts. The corresponding regressions permit assessing the extent and economic and numerical importance of the bias for each set of forecasts. Figure 2 plots the CBO, OMB, and APB forecast errors; and Figure 7 plots the estimates of forecast bias obtained from ultra saturation. (Figure 8 provides an alternative calculation of the forecast biases, as discussed in Section 6 below.)

The forecast biases vary markedly over time, and they exhibit some similarities across agencies. For the CBO forecasts, the bias is approximately 2.5% for 1990 and 2001–2003, 5% for 2008, for the most part declining thereafter, and –0.5% (and statistically detectably so) for all other years. For the OMB forecasts, the bias is approximately –8% for 2009 and –0.5% for all other years. For the APB forecasts, the bias is approximately 4% for 2008 and –0.5% for all other years. As a reference, the residual standard errors for the regressions with ultra saturation are 0.68%, 1.65%, and 1.20% respectively. In several instances, forecast biases exceed 2% in absolute value. These biases are economically large, especially considering that debt is a stock (not a flow), and that the forecasts are made less than nine months prior to the end of the fiscal year.

Table 3: Statistics for testing for bias in the CBO, OMB, and APB forecasts.

Statistic or regressor (target size)	K	CBO	OMB	APB
Mincer–Zarnowitz A	1	0.67 [0.421] $F(1, 28)$	3.96 [0.056] $F(1, 28)$	1.80 [0.191] $F(1, 28)$
Mincer–Zarnowitz B	2	1.30 [0.290] $F(2, 27)$	5.21* [0.012] $F(2, 27)$	1.05 [0.363] $F(2, 27)$
Forecast-encompassing	3	8.38** [0.000] $F(3, 26)$	19.44** [0.000] $F(3, 26)$	3.12* [0.043] $F(3, 26)$
Impulse indicator saturation (1%)	29	18.50** [0.000] $F(8, 21)$ I_{1990}, $I_{2001}, I_{2002}, I_{2003}$, $I_{2008}, I_{2009}, I_{2010}$	28.04** [0.000] $F(6, 23)$ I_{1990}, I_{2001}, $I_{2008}, I_{2009}, I_{2011}$	14.40** [0.000] $F(5, 24)$ I_{1990}, I_{2001}, I_{2008}, I_{2009}
Super saturation (0.5%)	56	17.66** [0.000] $F(8, 21)$ $I_{2008}, S_{1990}, S_{1991}$, S_{2001}, S_{2004}, S_{2008}, S_{2011}	15.44** [0.000] $F(4, 25)$ I_{2008}, S_{2008}, S_{2010}	7.63** [0.002] $F(2, 27)$ I_{2008}
Ultra saturation (0.3%)	84	24.16** [0.000] $F(7, 22)$ I_{1990}, I_{2011}, S_{2001}, S_{2004}, T_{2008}, T_{2009}	13.33** [0.000] $F(2, 27)$ I_{2009}	7.63** [0.002] $F(2, 27)$ I_{2008}
Intercept (OLS)	1	0.27 (0.33) {0.82} [0.421]	−0.79 (0.40) {−1.99} [0.056]	−0.36 (0.27) {−1.34} [0.191]
Intercept (IIS at 1%)	29	−0.58 (0.15) {−3.78} [0.001]	−0.79 (0.18) {−4.43} [0.000]	−0.60 (0.16) {−3.74} [0.001]

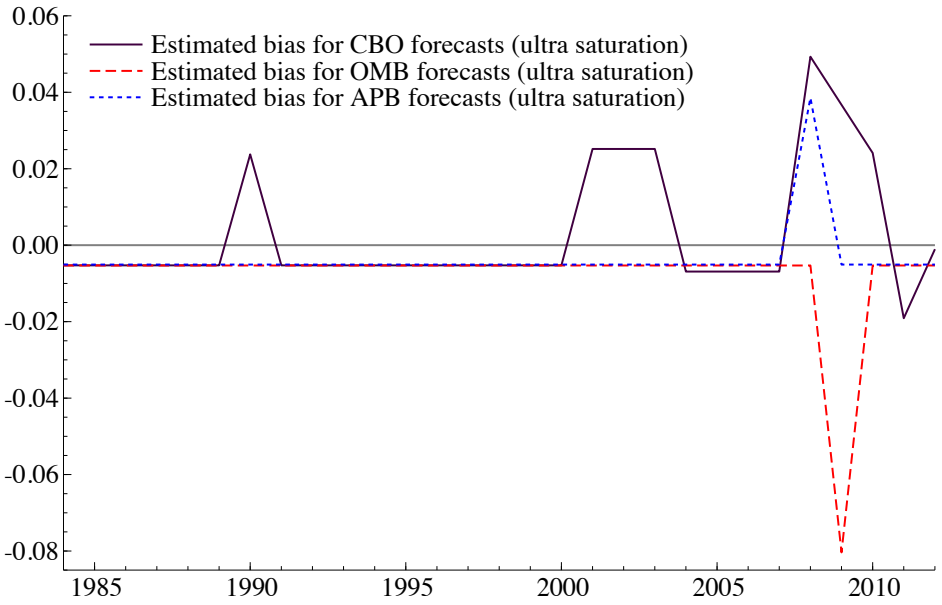

Figure 7: Estimates of forecast bias for the log of U.S. gross federal debt using ultra saturation.

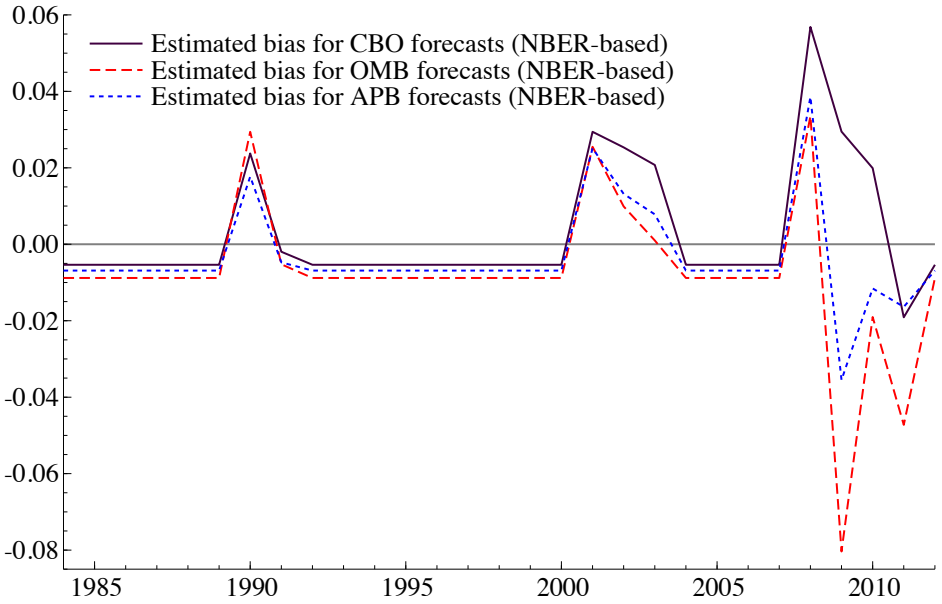

Figure 8: Estimates of forecast bias for the log of U.S. gross federal debt using a standardized set of NBER-dated and other impulse dummies.

As Figure 7 shows, forecast biases are sometimes positive and othertimes negative. The Mincer–Zarnowitz tests have particular difficulty in detecting such biases because the Mincer–Zarnowitz tests average all biases (both negative and positive) over time, and because the Mincer–Zarnowitz tests assign any time variation in bias to the residual rather than to the bias itself. As an extreme hypothetical example, the Mincer–Zarnowitz A test has no power whatsoever to detect a forecast bias that is $+\$10^{100}$ for the first half of the sample and $-\$10^{100}$ for the second half of the sample, even though this bias would be obvious from (e.g.) plotting the forecast errors.

Mincer–Zarnowitz tests also can lack power to detect forecast bias if forecast errors have thick tails or are heteroscedastic. Indeed, for every Mincer–Zarnowitz regression in Table 2, residual diagnostic statistics reject either normality or homoscedasticity. As follows from Johansen and Nielsen (2009), IIS can provide robust inference about the intercept in such a situation. While heteroscedasticity-consistent standard errors may provide consistent inference, they fail to improve efficiency of coefficient estimates, whereas robust estimation techniques such as IIS can. Those differences are highlighted in the bottom two rows of Table 3, which compare the estimated intercepts in the (OLS) Mincer–Zarnowitz A regressions with the estimated intercepts using IIS. The intercepts in the standard Mincer–Zarnowitz A regressions are statistically insignificant, whereas the intercepts estimated using IIS are highly significant. Even when IIS is viewed purely as a robust estimation procedure, empirical inferences about bias alter dramatically for the CBO, OMB, and APB forecasts. Bias is present in all three forecasts, and the standard Mincer–Zarnowitz tests typically fail to detect that bias. Section 6 goes further by re-interpreting the selected indicators themselves as resulting from an economically based time-varying forecast bias.

6 An Economic Interpretation of the Forecast Biases

This section examines the forecast biases in light of the business cycle. Section 6.1 re-interprets the estimated biases in light of the dates for the peaks and troughs of the business cycle, as determined by the National Bureau of Economic Research (NBER). This re-interpretation leads to a standardized reformulation of the estimated forecast biases in terms of business-cycle turning points, augmented by a few additional adjustments. Thus, this approach draws on Sinclair, Joutz, and Stekler (2010), who analyze the Fed's Greenbook forecasts similarly; and on Hendry (1999), who re-interprets IIS-detected outliers in an economic and institutional framework. See also Dyckman and Stekler (1966) and Stekler (1972, 2003). Section 6.2 evaluates these new models of forecast bias, including with tests for biases associated with political factors. Section 6.3 discusses some implications of forecast bias for forecasting.

Table 4: NBER reference dates and announcement dates for 1984–2012.

Event	Reference date of the event	Announcement date of the reference date	Length of determination (in months)	Impulse indicator
Peak	July 1990	April 25, 1991	9	I^P_{1990}
Trough	March 1991	December 22, 1992	21	I^T_{1991}
Peak	March 2001	November 26, 2001	8	I^P_{2001}
Trough	November 2001	July 17, 2003	20	I^T_{2002}
Peak	December 2007	December 1, 2008	11	I^P_{2008}
Trough	June 2009	September 20, 2010	15	I^T_{2009}

Notes. The length of determination is the time elapsed from the end of the month of the reference date to the announcement date of the reference date, rounded to the nearest end of month. The date of the impulse indicator is the calendar year in which the fiscal year ends for the fiscal year that spans the reference date. A superscript P or T on an impulse indicator denotes peak or trough; and that superscript emphasizes the event associated with the superscripted indicator. Source for events and dates: National Bureau of Economic Research (2012).

6.1 Forecast Biases and Turning Points

The previous section noted that several of the years associated with forecast bias are years in which major events—such as business-cycle turning points—occurred after the forecasts were made. As a way of capturing these phenomena in an economically interpretable manner, this subsection re-analyzes the forecast errors, specifically accounting for the effects of business-cycle turning points with impulse indicators. Additionally, IIS and its extensions are re-calculated, conditional on including the impulse indicators for these turning-point events.

The analysis below allows the forecast bias to alter for years in which an NBER-dated peak or trough occurs *after* the publication of the forecast but *before* the end of the fiscal year. In practice, impulse indicators are constructed for these turning-point events, and these dummy variables are included in regressions such as those for calculating the Mincer–Zarnowitz tests. From an economic and budgetary perspective, turning-point events could generate systematic biases in forecasts of debt because the advent of a recession or an expansion is likely to affect both sides of the federal government's balance sheet. For instance, the onset of a recession could lead to higher-than-anticipated outlays (as through higher unemployment compensation) *and* lower-than-anticipated revenues (as through lower individual and corporate income taxes).

Table 4 reports the NBER's turning-point events ("peak" or "trough") within the sample, the date of the event (the "reference date" in the NBER's terminology), the

Table 5: Statistics for testing for additional time-varying forecast bias in regressions of the CBO, OMB, and APB forecast errors on a standardized set of NBER-dated impulse indicators.

Statistic (target size)	K	CBO	OMB	APB
IIS (1%)	23	11.69** [0.000] $F(2, 20)$ I_{2003}, I_{2010}	20.99** [0.000] $F(1, 21)$ I_{2011}	No impulses selected

date on which the NBER announced the determination of that event, and the length of time taken to determine that an event had occurred; see the National Bureau of Economic Research (2012) for details. The corresponding impulse dummies are denoted I^P_{1990}, I^T_{1991}, I^P_{2001}, I^T_{2002}, I^P_{2008}, and I^T_{2009}, where a superscript P or T denotes that the event was a peak or trough, and the subscript indicates the year of the event (i.e., i in the notation above for the subscript on an indicator dummy).

The turning-point dummies appear necessary to capture the time variation in the forecast bias, but they do not appear sufficient. When the turning-point dummies are added to (e.g.) the Mincer–Zarnowitz A regression in equation (1), those dummies do capture economically and statistically important time dependence of the forecast bias. However, there is also evidence of time-varying bias, additional to what is associated with those turning points. Specifically, when IIS is applied to the version of equation (1) that is augmented by the turning-point dummies, IIS detects three additional years (2003, 2010, 2011) with bias for the CBO and OMB forecasts. Those three years immediately follow troughs, suggesting a potential explanation.

Table 5 reports the additional impulse dummies detected and the corresponding test statistics. The additional dummies detected differ across agencies: I_{2003} and I_{2010} for the CBO, I_{2011} for the OMB, and none for the APB. For a given forecast, the indicators selected are the same across saturation procedures, whether IIS at 1% ($K = 23$), super saturation at 0.5% ($K = 50$), or ultra saturation at 0.3% ($K = 78$).

To provide a unified and encompassing approach, the agencies' forecast errors are re-analyzed in regressions that include an intercept, all turning-point dummies, and all three of the additional dummies from Table 5. Table 6 reports these regressions, and Figure 8 (above) graphs the corresponding estimated forecast biases. This unified approach is also in line with the methodology in Hendry and Johansen (2015), who advocate (and provide the statistical underpinnings for) empirical analysis that embodies the available economic theory, while allowing model selection to detect additional phenomena that are also incorporated into the empirical model.

Table 6: Coefficients, estimated standard errors, and summary statistics for regressions of the CBO, OMB, and APB forecast errors on a standardized set of NBER-dated and other impulse indicators.

Regressor or statistic	CBO	OMB	APB
Intercept	−0.54	−0.88	−0.69
	(0.15)	(0.18)	(0.15)
I^P_{1990}	2.91	3.82	2.46
	(0.71)	(0.82)	(0.67)
I^T_{1991}	0.34	0.36	0.21
	(0.71)	(0.82)	(0.67)
I^P_{2001}	3.48	3.43	3.20
	(0.71)	(0.82)	(0.67)
I^T_{2002}	3.07	1.87	2.01
	(0.71)	(0.82)	(0.67)
I^P_{2008}	6.22	4.24	4.54
	(0.71)	(0.82)	(0.67)
I^T_{2009}	3.48	−7.16	−2.86
	(0.71)	(0.82)	(0.67)
I_{2003}	2.61	0.98	1.47
	(0.71)	(0.82)	(0.67)
I_{2010}	2.53	−1.02	−0.47
	(0.71)	(0.82)	(0.67)
I_{2011}	−1.37	−3.84	−0.95
	(0.71)	(0.82)	(0.67)
$\hat{\sigma}$	0.690%	0.798%	0.654%
RMSE of the forecast	1.746%	2.243%	1.452%
AR(2) LM statistic	0.47	1.45	0.10
	[0.634]	[0.261]	[0.910]
	$F(2, 17)$	$F(2, 17)$	$F(2, 17)$
ARCH(1) LM statistic	0.18	0.36	0.00
	[0.674]	[0.554]	[0.951]
	$F(1, 27)$	$F(1, 27)$	$F(1, 27)$
Normality statistic	1.08	4.63	6.09*
	[0.582]	[0.099]	[0.048]
	$\chi^2(2)$	$\chi^2(2)$	$\chi^2(2)$
Ramsey (1969)	0.00	0.00	0.00
RESET statistic	[0.999]	[0.996]	[0.999]
	$F(2, 17)$	$F(2, 17)$	$F(2, 17)$

The estimated biases in Figure 8 are thus interpretable economically, and they arise primarily from turning points in the business cycle. The three business-cycle peaks are all associated with substantial under-prediction of the debt: approximately 2%–3% in 1990 and 2001, and 3%–6% in 2008. Debt tends to be slightly over-predicted (by roughly 0.5%–1%) during 1984–1989, 1992–2000, 2004–2007, and 2012, which correspond to expansionary periods: see the intercepts in Table 6. Numerically and economically, the estimated biases are very similar across forecasts through 2008, but differ markedly thereafter. Statistically, the estimated biases in Figure 8 are substantial, noting the large difference between the residual standard error ($\hat{\sigma}$) of a given regression in Table 6 and the root mean squared error (RMSE) of the corresponding forecast. Interestingly, these economically based estimated forecast biases are generally similar to the "atheoretically based" estimated biases in Figure 7, which are derived from ultra saturation alone.

The estimated biases in Figure 8 also can be assessed statistically through residual diagnostics of the corresponding estimated equations. The standard diagnostics reported in Table 6 do not detect any substantial evidence of mis-specification. In particular, the Ramsey (1969) RESET test does not detect any nonlinearity, additional to that found by IIS. Conversely, the saturation-based tests for time-varying bias in Section 5 are very much in the spirit of Ramsey's RESET test for nonlinear mis-specification.

6.2 Assessment of the Economic Interpretation

This subsection assesses the economic interpretation of the models of forecast bias in Table 6 by testing various hypotheses about these models. Table 7 examines hypotheses that restrict the parameters estimated in Table 6—hypotheses of unbiasedness, the degree of bias induced by turning points, and biases across different forecasts. Table 8 examines hypotheses that focus on the potential importance of information *excluded* from the regressions in Table 6: assumed efficiency (in Mincer and Zarnowitz's sense), alternative forecasts, the phase of the NBER business cycle, the White House administration, the political party in the White House, and dates of presidential elections. Empirically, the magnitude of the forecast bias varies across business cycles; and the forecast bias does not appear politically related. The remainder of this subsection considers the results in Tables 7 and 8 in detail.

Table 7 examines restrictions on the parameters estimated in Table 6. Hypothesis (i) in Table 7 restricts all coefficients (including the intercept) to equal zero. This is denoted the Mincer–Zarnowitz A* test because it generalizes the Mincer–Zarnowitz A test by allowing for time-varying forecast bias. If all coefficients are zero, then the forecasts are unbiased. Unbiasedness is strongly rejected for all agencies' forecasts, contrasting with non-rejection by the Mincer–Zarnowitz A tests in Table 3.

Table 7: Tests of coefficient restrictions in regressions of the CBO, OMB, and APB forecast errors on a standardized set of NBER-dated and other impulse indicators.

Hypothesis or statistic		CBO	OMB	APB
(i)	Mincer–Zarnowitz A*	16.70** [0.000] $F(10, 19)$	20.99** [0.000] $F(10, 19)$	12.38** [0.000] $F(10, 19)$
(ii)	Mincer–Zarnowitz A**	12.08** [0.002] $F(1, 19)$	24.40** [0.000] $F(1, 19)$	22.15** [0.000] $F(1, 19)$
(iii)	Equal coefficients (by event)	6.36** [0.002] $F(4, 19)$	18.47** [0.000] $F(4, 19)$	8.39** [0.000] $F(4, 19)$
(iv)	Equal magnitude, opposite-signed coefficients	25.59** [0.000] $F(5, 19)$	16.51** [0.000] $F(5, 19)$	12.19** [0.000] $F(5, 19)$
(v)	Equality of biases across forecasts	86.23** [0.000] $\chi^2(10)$ (CBO = OMB)	63.77** [0.000] $\chi^2(10)$ (OMB = APB)	100.31** [0.000] $\chi^2(10)$ (APB = CBO)

Hypothesis (ii) restricts just the intercept in Table 6 to equal zero. This also is a variant of the hypothesis underlying the Mincer–Zarnowitz A test, so it is denoted Mincer–Zarnowitz A**. This hypothesis is rejected for all forecasts. Rejection implies a bias for all years *without* an impulse indicator in the regression, i.e., for the years 1984–1989, 1992–2000, 2004–2007, and 2012, all of which are during expansions. The estimated biases for those years are –0.54%, –0.88%, and –0.69% for the CBO, OMB, and APB respectively. That is, during these expansionary years, forecasts tend to over-predict the debt by about two-thirds of a percent.

Hypotheses (iii) and (iv) restrict the bias associated with turning points: either so that that bias is equal across dates for a specific event (peak or trough), or so that that bias is of equal magnitude across all events and opposite-signed for peaks and troughs. These hypotheses thus examine whether all peaks have the same bias (and likewise, all troughs), and additionally whether the nature of the event (peak or trough) affects only the sign of the bias. Hypotheses (iii) and (iv) are rejected for all agencies' forecasts. Not all peaks—nor all troughs—are equal in their effect on bias.

Hypothesis (v) imposes equality of the bias across different forecasts, e.g., testing

whether the CBO and OMB forecast biases are equal. Hypothesis (v) is strongly rejected, whether comparing CBO and OMB forecast biases, OMB and APB forecast biases, or APB and CBO forecast biases. Furthermore, the hypothesis of equality across CBO, OMB, and APB forecast biases is rejected, with the likelihood ratio statistic being $\chi^2(20) = 149.9^{**}$ [0.000]. Thus, in Figure 8, the time-varying forecast biases for the CBO, OMB, and APB are all significantly different from each other: the CBO, OMB, and APB forecasts do not share the same bias.

Table 8 focuses on the potential importance of information *excluded* from the regressions in Table 6. While IIS directly applied to Table 6's regressions would implicitly test the hypotheses listed in Table 8, explicit tests of these hypotheses may have more power than IIS. Hypothesis #1 in Table 8 imposes efficiency in the sense of Mincer and Zarnowitz, generalizing on the hypothesis $a_1 = 1$ in equation (2) and hence denoted Mincer–Zarnowitz B*. This hypothesis is examined by testing for the significance of the forecast itself, if the forecast is added to a regression in Table 6. This test is not rejected for the CBO or the APB, but it is rejected for the OMB. That said, the implied estimate of a_1 for the OMB is 0.9922, which is very close to unity numerically. Hypothesis #2 (forecast encompassing) considers whether alternative forecasts help explain a given forecast's forecast error. Only for the OMB do the other agencies' forecasts aid in explaining the forecast error, and then, only marginally so. Hypothesis #3 considers whether the phase of the NBER business cycle (expansion or contraction) matters for the forecast bias, above and beyond the presence of turning points. The phase does not matter for the CBO or APB but does matter (marginally) for the OMB.

The remaining hypotheses in Table 8 examine whether various political factors bias the forecasts. These hypotheses are very much in the spirit of Faust and Irons (1999), who test for presidential-cycle effects in U.S. macro-economic data. These hypotheses about political factors are of interest for *all* of the forecasts, even though the Congressional Budget Office produces "nonpartisan analysis for the U.S. Congress" (CBO website). In particular, *outcomes* of debt might be influenced by political factors, in which case the forecast errors could be, too. That is, a forecast could be biased because it failed to account for political factors that affected the actual outcome.

Hypotheses #4 and #5 consider the administration in the White House and the political party of the administration, where the "administration" is defined by the four-year presidential term. Neither the administration nor its political party appear to affect the forecast bias of any of the agencies. Hypotheses #6–#8 consider the presidential elections themselves, as measured by the year of the election, or by the political party of the president elected in that year. Furthermore, because the forecasts are made early in the calendar year and the presidential elections are held shortly after the end of the fiscal year, these hypotheses are also tested for the year *after* the presidential election (Hypotheses #9–#11). As the statistics for Hypotheses #6–#11 indicate, presidential elections do not appear to affect the forecast bias of any of the

Table 8: Diagnostic statistics for regressions of the CBO, OMB, and APB forecast errors on a standardized set of NBER-dated and other impulse indicators.

Hypothesis or statistic	CBO	OMB	APB
1. Mincer–Zarnowitz B*	0.25 [0.620] $F(1, 18)$	9.35** [0.007] $F(1, 18)$	0.41 [0.532] $F(1, 18)$
2. Forecast encompassing	1.17 [0.334] $F(2, 17)$	4.45* [0.028] $F(2, 17)$	0.20 [0.820] $F(2, 17)$
3. Phase of the NBER business cycle	0.69 [0.574] $F(3, 16)$	4.83* [0.014] $F(3, 16)$	0.55 [0.656] $F(3, 16)$
4. White House administration	0.79 [0.606] $F(7, 12)$	1.72 [0.194] $F(7, 12)$	0.49 [0.823] $F(7, 12)$
5. Political party of the administration	0.00 [0.948] $F(1, 18)$	0.22 [0.642] $F(1, 18)$	0.07 [0.797] $F(1, 18)$
6. Presidential election year	0.88 [0.550] $F(7, 12)$	1.58 [0.233] $F(7, 12)$	0.63 [0.725] $F(7, 12)$
7. Year that a Democratic president was elected	0.54 [0.471] $F(1, 18)$	1.70 [0.209] $F(1, 18)$	0.13 [0.727] $F(1, 18)$
8. Year that a Republican president was elected	1.11 [0.305] $F(1, 18)$	0.12 [0.732] $F(1, 18)$	0.57 [0.458] $F(1, 18)$
9. Year after a presidential election	0.81 [0.565] $F(5, 14)$	0.44 [0.811] $F(5, 14)$	0.61 [0.696] $F(5, 14)$
10. Year after a Democratic president was elected	1.42 [0.249] $F(1, 18)$	0.59 [0.453] $F(1, 18)$	0.60 [0.448] $F(1, 18)$
11. Year after a Republican president was elected	0.35 [0.560] $F(1, 18)$	0.02 [0.892] $F(1, 18)$	0.32 [0.578] $F(1, 18)$

agencies, regardless of the particular measure used for the presidential election year. Notably, the results on Hypotheses #4–#11 pertain to forecast *errors* and are mute about whether politics affects the government debt and its forecasts.

In summary, debt forecasts by the CBO, OMB, and APB exhibit time-varying biases that are primarily associated with turning points in the business cycle. The biases are not the same across the agencies making the forecasts, nor are they the same for peaks (or troughs) across different business cycles. Biases appear little affected by other factors. In particular, the biases do not appear to be politically related.

6.3 Remarks and Implications

This subsection discusses some potential implications of forecast bias. As background, this subsection first discusses forecast bias as a conditional expectation and then examines the importance of the information set on which that expectation is taken.

Forecast bias is defined as the expectation of the deviation between the actual outcome and the forecast itself. Either implicitly or explicitly, this expectation is conditional on an information set, such as past data; and the choice of that information set can affect the forecast bias. For instance, a forecast error may be unanticipated, unpredictable, and non-systematic conditional on one information set—but anticipated, predictable, and systematic conditional on another information set.

To illustrate, consider a simple (albeit slightly modified) example from Granger's (1983) paper "Forecasting White Noise". Define the forecast error u_t as $(y_t - \hat{y}_t)$, and assume that u_t is white noise and has an unconditional zero mean. Hence, from the properties of white noise, the expectation of the forecast error conditional on its own lag is zero:

$$\mathcal{E}[(y_t - \hat{y}_t) \,|\, (y_{t-1} - \hat{y}_{t-1})] \;=\; \mathcal{E}[u_t \,|\, u_{t-1}] \;=\; 0, \tag{8}$$

where $\mathcal{E}[\cdot]$ is the expectations operator. In fact, conditional on the lagged forecast error u_{t-1}, the current forecast error u_t is unpredictable, and the forecast $\hat{y}_t$ is unbiased for y_t. That said, whether the forecast is *really* unbiased—and whether the forecast error is *really* unpredictable—depends on the information set being conditioned upon.

To see the importance of the information set chosen, suppose that the white-noise forecast error u_t is made up of two white-noise processes g and η:

$$u_t \;=\; g_{t-1} + \eta_t. \tag{9}$$

Conditional on g_{t-1}, rather than on u_{t-1}, the forecast error is both biased and predictable:

$$\mathcal{E}[u_t \,|\, g_{t-1}] \;=\; \mathcal{E}[(g_{t-1} + \eta_t) \,|\, g_{t-1}] \;=\; g_{t-1}. \tag{10}$$

If g_{t-1} is large relative to η_t, the forecast error may appear to be an outlier conditional on the lagged forecast error u_{t-1}, whereas the forecast error simply may be biased

conditional on g_{t-1}. As equations (8), (9), and (10) highlight, the choice of conditioning set can matter; see Clements and Hendry (1999, Chapter 1.4.1) and Hendry and Mizon (2014) for discussion.

Some additional observations are germane. First, the forecast bias in equation (10) is *systematic* in that it depends directly on g_{t-1}. Second, that forecast bias is *not persistent*, noting that g_t is white noise. Third, the conditioning information sets in both equations (8) and (10) include only *lagged* information.

Several different information sets are relevant for analyzing the forecasts of debt, including:

(a) knowledge about the economy, as available at the time that the forecast is made;

(b) knowledge about the economy, as available on September 30; and

(c) the actual state of the economy on and before September 30.

Information set (a) is relevant for formulating the forecasts themselves, whereas information sets (a), (b), and (c) are all valuable for casting light on the sources of forecast error. In particular, the NBER dates for business-cycle turning points can be viewed as information in (c) and hence as valid information for *ex post* analysis of the forecast errors. Those dates provide the basis for the statistical and economic interpretation of the estimated forecast biases in Section 6.2. Information about *upcoming* turning points may be present in (a), but not fully utilized in formulating the forecasts, thereby leading to forecast biases relative to (a).

The presence of forecast bias implies the potential for improved forecasts. The feasibility of improvement may depend on the information in (a)–(c), as the large biases in 2001 and 1990 illustrate. For 2001, the NBER-dated peak of the business cycle is March. Because the 2001 forecasts were released on January 31 (CBO), February 28 (OMB), and May 1 (APB), exploitable information about the 2001 recession may have been available when the forecasts were being prepared. For 1990, however, the NBER-dated peak of the business cycle is July, which is much later in the forecast period than March. Hence, at the time of forecasting in 1990, evidence about the upcoming recession may have been more limited than in 2001. Additionally, Iraq's invasion of Kuwait begins on August 2, 1990; and that event and its timing would have been difficult to predict when the debt forecasts were being prepared in early 1990. These two examples highlight the importance of developing robust and accurate forecasts, and some of the difficulties in doing so.

As Fildes and Stekler (2002) and others have documented, turning points have been difficult to forecast. The large forecast biases for debt appear to reflect that challenge. From an institutional perspective, it may be useful to isolate the causes of the forecast errors according to the various assumptions made about fiscal policy, outlays and revenues, and the path of the economy in terms of variables such as output, inflation, and interest rates. Such an analysis could lead to improved forecasts, or at least provide a deeper understanding of the sources of forecast error.

7 Conclusions

Government debt and its forecasts feature prominently in current economic and political discussions. The properties of these forecasts are thus of interest, and it matters how these properties are assessed. Mincer–Zarnowitz tests typically fail to detect biases in the CBO, OMB, and APB one-year-ahead forecasts of U.S. gross federal debt over 1984–2012. By contrast, more general tests based on impulse indicator saturation detect economically large, systematic, and statistically highly significant time-varying biases in the CBO, OMB, and APB forecasts, particularly for 1990, 1991, 2001–2003, and 2008–2011. These biases differ according to the agency making the forecasts, and these biases are closely linked to turning points in the business cycle and (to a lesser degree) economic expansions. However, these biases do not appear to be politically related. The IIS approach also explains *why* Mincer–Zarnowitz tests may fail to detect bias. The Mincer–Zarnowitz tests average over the biases for *all* observations, but those biases may be positive for some observations and negative for others, thereby reducing the tests' power.

Impulse indicator saturation defines a generic procedure for examining forecast properties and, in particular, for detecting and quantifying forecast bias. Forecast bias can be systematic yet time-varying; it can be difficult to detect in a timely fashion; and it may have substantive implications for policy analysis. IIS and its extensions can help address these issues by characterizing systematic properties in the forecast errors. The IIS approach also links directly to existing techniques for robustifying forecasts, noting that intercept correction is a variant of super saturation; see Clements and Hendry (1996, 1999, 2002a), Hendry (2006), Castle, Fawcett, and Hendry (2010), and Castle, Clements, and Hendry (2015).

The IIS approach has many potential applications, beyond its initial roles in model evaluation and robust estimation. Ericsson (2012) considers its uses for detecting crises, jumps, and changes in regime. IIS also provides a framework for creating near real-time early-warning and rapid-detection devices, such as of financial market anomalies; cf. Vere-Jones (1995) on forecasting earthquakes and earthquake risk, and Goldstein, Kaminsky, and Reinhart (2000) on early warning systems for emerging market economies. Relatedly, the model selection approach in IIS is applicable to nowcasting with a large set of potential explanatory variables, such as those generated from Google Trends; see Doornik (2009b), Choi and Varian (2012), and Castle, Hendry, and Kitov (2016). Finally, IIS generalizes to systems, and so is consonant with the approach proposed in Sinclair, Stekler, and Carnow (2012) for evaluating economic forecasts.

Appendix A. Interpreting Estimates of Forecast Bias

This appendix resolves differences in results and interpretation between Ericsson's (2017) and Gamber and Liebner's (2017) assessments of forecasts of U.S. gross federal debt. As Gamber and Liebner (2017) discuss, heteroscedasticity could explain the empirical results in Ericsson (2017). However, the combined evidence in Ericsson (2017) and Gamber and Liebner (2017) supports the interpretation that these forecasts have significant time-varying biases. Both Ericsson (2017) and Gamber and Liebner (2017) advocate using impulse indicator saturation in empirical modeling.

A.1 Introduction

Using impulse indicator saturation (IIS), Ericsson (2017) tests for and detects economically large and statistically highly significant time-varying biases in forecasts of U.S. gross federal debt over 1984–2012, particularly at turning points in the business cycle. Gamber and Liebner (2017) discuss Ericsson (2017), obtaining different empirical results and offering a different interpretation. This appendix resolves those differences through a re-examination of IIS.

Gamber and Liebner (2017) examine Ericsson's (2017) choice of IIS's significance level and interpretation of the estimated bias, concluding that the empirical basis for *time-varying* bias *per se* is weaker than claimed, and that the outliers detected by IIS could easily arise from heteroscedasticity rather than from time-varying bias. Because IIS does have power to detect heteroscedasticity, heteroscedasticity could explain the IIS results in Ericsson (2017). However, as Sections A.2 and A.3 below show, time-varying bias is more consistent with the combined evidence in Ericsson (2017) and Gamber and Liebner (2017). Section A.4 comments further on modeling with IIS.

A.2 Analysis of Alternative Model Specifications

Ericsson (2017) and Gamber and Liebner (2017) assess forecasts of U.S. federal debt, focusing on the economic and statistical bases for the selected impulse indicators from IIS. Although Ericsson (2017) and Gamber and Liebner (2017) evaluate the same set of forecasts, they obtain different empirical results and offer different interpretations of those results. Section A.3 below resolves the differences in interpretation through a re-examination of IIS. The current section resolves the differences in the empirical results themselves—both qualitatively and quantitatively—through an encompassing approach by examining alternative model specifications.

In particular, encompassing analysis of an analytical example demonstrates how certain model specifications reduce the power of tests to detect impulse indicators,

where that power depends directly on t-ratios for the indicators. The encompassing analysis implies that some relevant indicators may nonetheless appear unimportant in certain models, simply because those models omit relevant variables, thereby increasing the residual standard error and hence reducing the t-ratios. The current section first presents the analytical example and then applies it to the disparate empirical results with IIS.

This type of assessment is sometimes called "mis-specification analysis" because some models analyzed omit certain relevant variables and hence are mis-specified, relative to the data generation process; see Sargan (1988, Chapter 8). Mizon and Richard (1986) propose a constructive utilization of mis-specification analysis—known as the encompassing approach—in which a given model (Model M0, below) is shown to explain or "encompass" properties of the other models (Models M1 and M2, below). In the current section, model properties include t-ratios, residual variances, and the selection of impulse dummies. See Davidson, Hendry, Srba, and Yeo (1978), Mizon and Richard (1986), and Bontemps and Mizon (2008) for further discussion.

Analytical example. To put the encompassing analysis in context, suppose that both blocks of observations for bare-bones IIS include impulse dummies that have nonzero coefficients in the data generation process (DGP). In bare-bones IIS, estimation of coefficients for dummies that saturate a given block then implies omission of the other block's relevant dummies in the corresponding model. These omitted dummies typically result in reduced power to detect the significance of included dummies. An analytical example illustrates.[1]

In a notation similar to that in Ericsson (2017, Example 2), let the DGP for the variable w_t be as follows.

$$\text{DGP: } w_t = \delta_0 + \delta_1 I_{1t} + \delta_2 I_{2t} + \varepsilon_t, \qquad \varepsilon_t \sim NID(0, \sigma^2), \qquad t = 1, \ldots, T. \tag{A1}$$

That is, w_t is normally and independently distributed with a constant mean δ_0 and constant variance σ^2 over T observations, except that w_t's mean is $\delta_0 + \delta_1$ in period $t = t_1$ (when the impulse indicator I_{1t} is nonzero) and $\delta_0 + \delta_2$ in period $t = t_2$ (when $I_{2t} \neq 0$). For expository purposes, assume that δ_1 and δ_2 are both strictly positive, and that t_1 and t_2 are in the first and second blocks of observations respectively.

Consider three models, denoted M0, M1, and M2. Model M0 is specified as the DGP (A1) itself.

$$\text{Model M0: } w_t = \delta_0 + \delta_1 I_{1t} + \delta_2 I_{2t} + \varepsilon_t. \tag{A2}$$

Models M1 and M2 entail omitted variables. Model M1 includes I_{1t} but omits I_{2t}.

$$\text{Model M1: } w_t = \delta_0 + \delta_1 I_{1t} + v_{1t}. \tag{A3}$$

[1] This analysis and its empirical application below ignore changes in the estimated coefficients that arise from the omitted impulse indicators. However, because impulse indicators are orthogonal, those changes should not be an important consideration here.

Model M2 includes I_{2t} but omits I_{1t}.

$$\text{Model M2: } w_t = \delta_0 + \delta_2 I_{2t} + v_{2t}. \tag{A4}$$

For Model M1, the error v_{1t} is $(\delta_2 I_{2t} + \varepsilon_t)$, so Model M1's mean squared error σ_1^2 is:

$$\sigma_1^2 = (\sigma^2 + \delta_2^2/T), \tag{A5}$$

which is larger than σ^2, the error variance for Model M0. Likewise, for Model M2, the error v_{2t} is $(\delta_1 I_{1t} + \varepsilon_t)$, and the mean squared error σ_2^2 is:

$$\sigma_2^2 = (\sigma^2 + \delta_1^2/T), \tag{A6}$$

which also is larger than σ^2.

One possible consequence of model specifications such as M1 and M2 is to shrink t-ratios on included variables. As equations (A5) and (A6) imply, the estimated residual variance in a model with an omitted relevant variable is typically larger than the estimated residual variance in the DGP. Hence, the estimated standard error on the coefficient of a variable included in that model is larger than the corresponding coefficient's estimated standard error in the DGP. That shrinks the coefficient's t-ratio in the model with the omitted variable.

For example, the t-ratio for I_{1t} in Model M1 uses $\hat{\sigma}_1$ in the coefficient's estimated standard error, rather than $\hat{\sigma}$, which would be used for its t-ratio in Model M0. Thus, I_{1t} might be significant in Model M0 but appear insignificant in Model M1, simply because Model M1 excludes I_{2t} and so $\hat{\sigma}_1 > \hat{\sigma}$. Likewise, the t-ratio for I_{2t} in Model M2 uses $\hat{\sigma}_2$ in the coefficient's estimated standard error, rather than $\hat{\sigma}$. Hence, I_{2t} might be significant in Model M0 but appear insignificant in Model M2 because Model M2 excludes I_{1t} and so $\hat{\sigma}_2 > \hat{\sigma}$. As Hendry and Doornik (2014, p. 243) summarize, "[w]hen there is more than a single break, a failure to detect one [break] increases the residual variance and so lowers the probability of detecting any others."

Empirical application. Gamber and Liebner (2017) discuss t-ratios, significance levels, and empirical power for IIS, illustrating with the CBO forecasts. To interpret these empirical results in an encompassing framework, consider a baseline specification that includes all seven impulse indicators selected in Ericsson (2017). The observed t-ratios on retained impulses in Gamber and Liebner's models are closely matched by t-ratios as numerically solved from an encompassing analysis that starts with that baseline seven-indicator model. This comparison appears in Table A1. Moreover, the retention (or not) of individual impulse indicators in Gamber and Liebner (2017) is consistent with the losses in power implied by the encompassing analysis.

Key empirical results can be summarized, as follows. Using the "bare-bones" implementation of IIS, Gamber and Liebner (2017, Section 3) detect the following impulse indicators in the second subsample (1998–2012):

Table A1: Actual and solved t-ratios and residual standard errors for regressions of the CBO forecast errors on various impulse indicators.

Regressor or statistic	Block analyzed, significance level or target size, and result and column					
	Bare-bones IIS				Autometrics IIS	
	2nd block (1%)	2nd block (1%, 1%)	2nd block (5%)	1st block (–)	Multi-block (1%)	Estimated coefficient $\hat{\delta}_i$
	(a) col. #1	(b) col. #2	(c) col. #3	(d) col. #4	(e) col. #5a	(e) col. #5b
I_{1990}				1.2 $\langle 1.5 \rangle$	4.0^{***}	2.96
I_{2001}	2.4^{*} $\langle 2.7^{*} \rangle$		3.5^{**} $\langle 3.8^{**} \rangle$		4.8^{***}	3.52
I_{2002}			3.1^{**} $\langle 3.4^{**} \rangle$		4.2^{***}	3.12
I_{2003}			2.6^{*} $\langle 2.9^{**} \rangle$		3.6^{**}	2.66
I_{2008}	4.6^{***} $\langle 4.8^{***} \rangle$	3.8^{***} $\langle 3.9^{***} \rangle$	6.4^{***} $\langle 6.7^{***} \rangle$		8.5^{***}	6.27
I_{2009}	2.5^{*} $\langle 2.7^{*} \rangle$		3.6^{**} $\langle 3.8^{**} \rangle$		4.8^{***}	3.53
I_{2010}			2.6^{*} $\langle 2.8^{*} \rangle$		3.5^{**}	2.57
$\hat{\sigma}$	1.24 $\langle 1.28 \rangle$	1.44 $\langle 1.58 \rangle$	0.94 $\langle 0.91 \rangle$	1.74 $\langle 1.88 \rangle$	0.72	—
Calculated rescaling factor	$\langle 0.57 \rangle$	$\langle 0.46 \rangle$	$\langle 0.80 \rangle$	$\langle 0.38 \rangle$	–	—

Notes. Column headers indicate the version of IIS employed, the block(s) analyzed, the significance level (for bare-bones IIS) or target size (for Autometrics IIS), associated result (a)–(e), and the column number. Unbracketed numerical values are observed empirical t-ratios, $\hat{\sigma}$, and (for Column #5b) estimated coefficients from the designated regressions. Values in angled brackets $\langle \cdot \rangle$ are as solved from the encompassing analysis. Superscript asterisks *, **, and *** denote rejections of the null hypothesis at the 5%, 1%, and 0.1% levels respectively; and the null hypothesis is that the coefficient on the corresponding impulse indicator is zero. All actual and solved values are reported to just one or two decimals for readability, but solved quantities are calculated from *unrounded* actual values. All regressions include an intercept; $\hat{\sigma}$ is in percent; and the sample period is 1984–2012. In Column #2, selection at the 1% significance level is repeated.

(a) I_{2001}, I_{2008}, and I_{2009} (at a 1% significance level);

(b) I_{2008} only (at a 1% significance level, but re-selected from (a)); and

(c) I_{2001}, I_{2002}, I_{2003}, I_{2008}, I_{2009}, and I_{2010} (at a 5% significance level).

For the first subsample (1984–1997), Gamber and Liebner find that:

(d) I_{1990} is not significant, nor is any other impulse indicator.

Columns ##1–4 in Table A1 report the t-ratios from (a)–(d). Using IIS in Autometrics, Ericsson (2017, Table 3) detects seven impulse indicators:

(e) I_{1990}, I_{2001}, I_{2002}, I_{2003}, I_{2008}, I_{2009}, and I_{2010} (at a 1% target size).

Column #5a in Table A1 reports the t-ratios in that specification.

The results in (a)–(e) present a puzzle. From (a)–(d) combined, Gamber and Liebner (2017) find that only I_{2008} is significant at the 1% level. By contrast, all seven impulses in (e) are significant at not only the 1% level but at the 0.5% level; and all but I_{2003} and I_{2010} are significant at the 0.1% level.

These apparently contradictory results can be reconciled by an encompassing analysis that treats (e) as Model M0 (the DGP), (a)–(c) as versions of model M1, and (d) as model M2. In this context, specifications (e), (a)–(c), and (d) generalize equations (A2), (A3), and (A4) to (potentially) include multiple indicators in each subsample.

The encompassing analysis begins with $\hat{\sigma}$. Note that $\hat{\sigma}$ in Column #5a is 0.72, which is $\hat{\sigma}$ for the assumed DGP. In Columns ##1–4, the values of $\hat{\sigma}$ are much larger, as would be expected with omitted relevant indicators. Directly under those four values of $\hat{\sigma}$, the values in angled brackets $\langle\cdot\rangle$ report the corresponding residual standard errors, as solved numerically from the analytical example above. These solved values are calculated from formulas (A5) and (A6), generalized for multiple impulses, and using the values of $\hat{\sigma}$ and $\hat{\delta}_i$ for the model in Column #5. The solved values for $\hat{\sigma}$ are very close to the actual values for $\hat{\sigma}$, indicating how well the analytical example helps explain (and encompass) Gamber and Liebner's empirical results.

Similarly, the values in angled brackets $\langle\cdot\rangle$ under actual t-ratios report the t-ratios as solved from the encompassing analysis. To obtain a "solved" t-ratio, the actual t-ratio is rescaled by the ratio of Column #5's $\hat{\sigma}$ to the solved value of the residual standard error. The values of the solved t-ratios also are very close to their actual values. The last line in Table A1 reports the calculated rescaling factor, which highlights the considerable anticipated loss of information from the omitted impulse indicators in (a)–(d).

To illustrate concretely how these encompassing calculations proceeded, consider the solved values for Column #3. From equation (A6), the solved value of $\hat{\sigma}$ is the square root of $(0.72^2 + (2.96^2/29))$, or 0.91. The solved t-ratio on (e.g.) I_{2001} is $4.8 \cdot (0.72/0.91)$, or 3.8. These solved values for $\hat{\sigma}$ and the t-ratio are very close to the actual values of 0.94 and 3.5.

A.3 The Power of Impulse Indicator Saturation

Gamber and Liebner (2017) observe that IIS has power to detect heteroscedasticity in the disturbances as well as nonconstancy in the forecast bias. Gamber and Liebner then conduct Monte Carlo simulations, which suggest that heteroscedasticity is a likely interpretation of the empirical results from IIS in Ericsson (2017). Paralleling Gamber and Liebner's Monte Carlo simulations, a direct analytical solution shows that heteroscedasticity can give rise to IIS detecting multiple impulse dummies. However, the number of impulse dummies actually detected by IIS for the government debt forecast errors would likely require substantially more heteroscedasticity than assumed. This section summarizes the statistical framework for Gamber and Liebner's Monte Carlo simulations, derives an alternative analytical solution, summarizes implications for the empirical results, and reconsiders the potential role of heteroscedasticity.

To show that pure heteroscedasticity might explain the empirical results from IIS, Gamber and Liebner (2017) adopt the following DGP for w_t:

$$w_t \sim NID(0, \sigma_a^2), \qquad t = 1, \ldots, T_a; \text{ and} \tag{A7}$$

$$w_t \sim NID(0, \sigma_b^2), \qquad t = (T_a + 1), \ldots, T. \tag{A8}$$

Based on the empirical setting for debt forecasts as analyzed with bare-bones IIS, Gamber and Liebner choose equations (A7)–(A8) with subsamples of length $T_a = 14$ and $T_b = 15$ where $T_b \equiv (T - T_a)$, and subsample standard deviations of $\sigma_a = 1.007\%$ and $\sigma_b = 2.122\%$. Gamber and Liebner generate 10^4 replications of Monte Carlo data with these properties, apply bare-bones IIS to each replication, and count the number of dummies retained across replications. Table A2's column labeled "Monte Carlo (5%)" reports Gamber and Liebner's (2017, Table 1) estimated probabilities for retaining different numbers of impulse indicator dummies when selecting them at a 5% significance level on individual t-ratios in bare-bones IIS. These estimated probabilities imply a nearly one-in-three chance of detecting six or more impulse indicators, six being the number of indicators detected in (c) above. The average number of indicators detected in the Monte Carlo simulation is 4.4.

The statistical problem posed by Gamber and Liebner can also be solved analytically, noting the following features. First, the t-ratios on the impulse indicators in bare-bones IIS have t-distributions, once the t-ratios are rescaled by σ_a/σ_b or σ_b/σ_a, as appropriate. Second, the probability of retaining a specific number of dummies can be derived from a generalization of the binomial distribution; see Stuart and Ord (1987, Chapter 5). Solving that probability obtains the values in Table A2's column "Binomial solution (5%)", which closely matches the previous column, "Monte Carlo (5%)".

As Section A.2 discusses, the empirically relevant target size is 1% (not 5%), and it is of interest to calculate the probability of retaining at least seven dummies (rather

Table A2: Calculated probabilities for retaining different numbers of impulse indicator dummies under an assumption of heteroscedasticity, at 5% and 1% target sizes.

Number of retained dummies	Monte Carlo (5%)	Binomial solution (5%)	Binomial solution (1%)	Binomial solution (1%) $[\sigma_b = 2.842]$
0	1.9	0.3	5.4	0.4
1	6.4	2.0	17.5	2.8
2	13.3	6.8	26.2	8.6
3	16.5	14.0	24.3	16.4
4	17.4	20.2	15.6	21.6
5	15.1	21.4	7.4	20.9
6	11.9	17.1	2.6	15.3
7	8.1	10.6	0.7	8.6
8	5.0	5.1	0.2	3.8
9	2.7	1.9	0.0	1.3
10	1.0	0.5	0.0	0.3
11	0.5	0.1	0.0	0.1
12	0.1	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0
Probability of retaining 6+ dummies	29.4	35.3	3.5	29.4
Probability of retaining 7+ dummies	17.5	18.2	0.9	14.1
Average number of dummies retained	4.4	4.9	2.6	4.6

Notes. All values for Monte Carlo and binomial calculations are in percent, except for the "average number of dummies retained". Values in the column for "Monte Carlo (5%)" are from Gamber and Liebner (2017, Table 1), rounded to the first decimal in light of the implied uncertainty in their Monte Carlo simulation; see Hendry (1984). Probabilities in the antepenultimate and penultimate rows are calculated from unrounded values. The final column is calculated for the alternative value of σ_b equal to 2.842.

than at least six). The corresponding calculations appear in Table A2's penultimate column, labeled "Binomial solution (1%)". The average number of dummies retained is only 2.6, and the probability of retaining at least seven dummies is under 1%. Pure heteroscedasticity thus appears unlikely to explain the retention of the seven impulse indicators found in practice.

That said, if the difference between the subsample standard deviations σ_a and σ_b were greater, the implied heteroscedasticity could have been a likely explanation for IIS's empirical behavior. Specifically, if σ_b were 2.842 rather than 2.122 (and σ_a unchanged), then the probability of retaining at least six dummies would have been 29.4%, the same value as obtained by Gamber and Liebner. The corresponding calculations appear in Table A2's final column, labeled "Binomial solution (1%) $[\sigma_b = 2.842]$".

A.4 Remarks

Several issues merit additional remarks, including algorithmic implementation, the models considered, power, time-invariant bias, and directions for further research.

First, algorithmic implementation of IIS requires important choices, as Hendry and Doornik (2014) discuss. Choices include the construction of the blocks, model selection criteria, use of diagnostic statistics, path search, block combination and re-selection, iteration, and significance level. These choices may matter under the null hypothesis of correct specification, under the alternative hypothesis, or under both.

For example, under the null hypothesis, too loose a significance level may inadvertently retain many irrelevant dummies, downwardly biasing the estimated residual standard error, and upwardly biasing t-ratios; see Gamber and Liebner (2017). Hendry, Johansen, and Santos (2008) and Johansen and Nielsen (2009, 2013, 2016) consider this issue in detail. Hendry and Doornik (2014, Chapter 15) and Johansen and Nielsen (2016) propose implementable bias corrections. Even simpler, Hendry and Doornik (2014, Chapter 15) recommend a relatively tight significance level of $1/T$ as a rule-of-thumb to help keep such estimation bias minimal. Ericsson (2017) employs an even tighter level of about $0.3/T$ for IIS. So, the seven impulse indicators discussed in Section A.2 above are of substantive interest and do not appear to have been retained spuriously. Relatedly, bare-bones IIS can actually select *more* (and not only fewer) impulse indicators than Autometrics IIS, as Figures 6g and 6h in Ericsson (2017) imply.

Second, the models considered—and those not considered—can affect the model selected. Thus, the results in Section A.2 may depend on differences between bare-bones and Autometrics implementations of IIS, indirectly through which models the two algorithms consider in their selection processes. For instance, if one of the blocks in bare-bones IIS had included 1990 in addition to 1998–2012, bare-bones IIS would have detected the impulse indicator for 1990 at the 1% significance level. When the

null hypothesis is false, the choice of blocks and the implied set of models can strongly influence IIS's ability to detect the alternative. Hence, Autometrics searches over many blocks, including possibly overlapping and unequally sized blocks; see Doornik (2009a).

Third, IIS has power to detect heteroscedasticity—and many other alternatives as well. Applications of IIS reflect that wide-ranging ability: see Hendry (1999) on nonconstancy, Johansen and Nielsen (2009) and Marczak and Proietti (2016) on outliers, Hendry and Doornik (2014, Chapter 15.6) on thick-tailed distributions, Hendry and Santos (2010) on heteroscedasticity and super exogeneity, Ericsson (2011b) on omitted variables and regime changes, Castle, Doornik, and Hendry (2012) on multiple breaks, Pretis, Schneider, Smerdon, and Hendry (2016) on "designer" breaks, and Ericsson (2016) on measurement errors. Gamber and Liebner (2017) underscore the benefits of IIS, stating that "... the IIS technique is useful as an ex-post diagnostic tool for detecting points in time when the model is biased" (Section 4), and that IIS is valuable "... as a general diagnostic tool for detecting model misspecification" (abstract).

Fourth, in order to achieve good power against many different alternatives, Hendry and Doornik (2014) intentionally allow Autometrics to beneficially (and temporarily) relax the significance level in "... search[ing] for potentially significant, but as yet omitted, variables" (p. 235). Doing so has little effect under the null hypothesis but may be helpful under alternatives, as Section A.2 highlights.

Fifth, time-*invariant* bias in the government debt forecasts is empirically detectable at the 0.2% significance level when using IIS, even if the retained impulse indicators are thought of as arising purely from "outliers". By contrast, without IIS to robustify estimation and inference, the forecast bias appears insignificant at even the 10% level; cf. the Mincer–Zarnowitz A and A** tests for the CBO in Ericsson (2017, Tables 3 and 7).

Sixth, many directions for further research are highly promising. In particular, generalized saturation offers parsimonious representations of outliers and breaks; see Castle, Doornik, Hendry, and Pretis (2015) on step indicator saturation, and Ericsson (2011b) for a typology of saturation techniques. One saturation technique—multiplicative indicator saturation—embodies a structure similar to that of regime-switching models, while allowing a given regime to differ quantitatively across its multiple occurrences. Highlighting this aspect, test (iii) in Ericsson (2017, Table 7) shows that forecast biases are not equal across different occurrences of the same "event" (or regime), where that event is a peak or a trough. A standard regime-switching model would have difficulty accommodating such heterogeneity, and would have difficulty even detecting turning points as regimes because of their brief nature.

A.5 Conclusions

Gamber and Liebner (2017) raise important issues concerning the interpretation of empirical results, particularly when employing impulse indicator saturation. In the discussion above, the analysis of alternative model specifications and the calculation of empirical power functions highlight consequences for IIS when the null hypothesis is incorrect. Specifically, IIS has power to detect many empirical features, including heteroscedasticity, structural breaks, outliers, and omitted variables. As a practical implication, the evidence in Ericsson (2017) and Gamber and Liebner (2017) supports the interpretation that U.S. government agencies' forecasts of U.S. gross federal debt have time-varying biases.

Appendix B. The Data and the Forecasts

Sections 1–7 above, Gamber and Liebner (2017), and Ericsson (2017) (Appendix A above) analyze data on U.S. government debt (denoted "Debt") and CBO, OMB, and APB forecasts of that debt, as compiled by Martinez (2015). The current appendix lists those data and forecasts in Table B1. See Martinez (2015) and Section 2 above for details, including sources and definitions.

Table B1: U.S. government debt and CBO, OMB, and APB forecasts of that debt.

Year	Debt	CBO	OMB	APB
1983	1381.886	–	–	–
1984	1576.748	1600.	1591.573	1599.
1985	1827.47	1853.	1841.077	1854.
1986	2129.964	2114.	2112.	2110.6
1987	2355.206	2364.	2372.4	2367.2
1988	2600.679	2598.	2581.6	2603.
1989	2865.664	2865.	2868.8	2869.
1990	3206.26	3131.	3113.3	3150.
1991	3598.919	3606.	3617.837	3616.
1992	4002.815	4039.	4080.3	4058.
1993	4351.149	4392.	4396.7	4391.
1994	4643.996	4690.	4676.	4692.
1995	4920.95	4942.	4961.5	4947.
1996	5181.923	5191.	5207.3	5193.
1997	5369.7	5436.	5453.7	5432.
1998	5478.717	5540.	5543.6	5524.
1999	5606.486	5579.	5614.9	5578.
2000	5629.009	5665.	5686.	5674.
2001	5770.249	5603.	5625.	5627.
2002	6198.129	6043.	6137.1	6117.
2003	6758.722	6620.	6752.	6706.
2004	7352.017	7459.	7486.4	7453.
2005	7902.8	7975.	8031.4	7991.
2006	8448.991	8515.	8611.5	8556.
2007	8948.534	8915.	9007.8	8968.
2008	9983.694	9432.	9654.4	9606.
2009	11873.812	11529.	12867.5	12303.
2010	13526.633	13260.	13786.6	13684.
2011	14762.223	15047.	15476.2	15006.
2012	16048.111	16002.	16350.9	16187.
2013	16716.791	17068.	17249.2	16897.

References

Alexander, S. S., and H. O. Stekler (1959) "Forecasting Industrial Production—Leading Series versus Autoregression", *Journal of Political Economy*, 67, 4, 402–409.

Andrews, D. W. K. (1993) "Tests for Parameter Instability and Structural Change with Unknown Change Point", *Econometrica*, 61, 4, 821–856.

Bai, J., and P. Perron (1998) "Estimating and Testing Linear Models with Multiple Structural Changes", *Econometrica*, 66, 1, 47–78.

Bergamelli, M., and G. Urga (2014) "Detecting Multiple Structural Breaks: Dummy Saturation vs Sequential Bootstrapping. With an Application to the Fisher Relationship for US", CEA@Cass Working Paper Series No. WP–CEA–03–2014, Cass Business School, London, April.

Bernanke, B. S. (2011) "Fiscal Sustainability", speech, Annual Conference, Committee for a Responsible Federal Budget, Washington, D.C., June 14.

Bernanke, B. S. (2013) "Chairman Bernanke's Press Conference", transcript, Board of Governors of the Federal Reserve System, Washington, D.C., September 18.

Bontemps, C., and G. E. Mizon (2008) "Encompassing: Concepts and Implementation", *Oxford Bulletin of Economics and Statistics*, 70, supplement, 721–750.

Castle, J. L., M. P. Clements, and D. F. Hendry (2013) "Forecasting by Factors, by Variables, by Both or Neither?", *Journal of Econometrics*, 177, 2, 305–319.

Castle, J. L., M. P. Clements, and D. F. Hendry (2015) "Robust Approaches to Forecasting", *International Journal of Forecasting*, 31, 1, 99–112.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2012) "Model Selection When There Are Multiple Breaks", *Journal of Econometrics*, 169, 2, 239–246.

Castle, J. L., J. A. Doornik, D. F. Hendry, and F. Pretis (2015) "Detecting Location Shifts During Model Selection by Step-indicator Saturation", *Econometrics*, 3, 2, 240–264.

Castle, J. L., N. W. P. Fawcett, and D. F. Hendry (2010) "Forecasting with Equilibrium-correction Models During Structural Breaks", *Journal of Econometrics*, 158, 1, 25–36.

Castle, J. L., D. F. Hendry, and O. I. Kitov (2016) "Forecasting and Nowcasting Macroeconomic Variables: A Methodological Overview", Chapter 3 in EuroStat (ed.) *Handbook on Rapid Estimates*, UN/EuroStat, Brussels, forthcoming.

Choi, H., and H. Varian (2012) "Predicting the Present with Google Trends", *Economic Record*, 88, Special Issue, 2–9.

Chokshi, N. (2013) "Beware Obama's Budget Predictions: Many Forecasts Are Wrong", *National Journal*, April 10 (www.nationaljournal.com).

Chong, Y. Y., and D. F. Hendry (1986) "Econometric Evaluation of Linear Macroeconomic Models", *Review of Economic Studies*, 53, 4, 671–690.

Chow, G. C. (1960) "Tests of Equality Between Sets of Coefficients in Two Linear Regressions", *Econometrica*, 28, 3, 591–605.

Clements, M. P., and D. F. Hendry (1996) "Intercept Corrections and Structural Change", *Journal of Applied Econometrics*, 11, 5, 475–494.

Clements, M. P., and D. F. Hendry (1999) *Forecasting Non-stationary Economic Time Series*, MIT Press, Cambridge.

Clements, M. P., and D. F. Hendry (2002a) "Explaining Forecast Failure in Macroeconomics", Chapter 23 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 539–571.

Clements, M. P., and D. F. Hendry (2002b) "An Overview of Economic Forecasting", Chapter 1 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 1–18.

Corder, J. K. (2005) "Managing Uncertainty: The Bias and Efficiency of Federal Macroeconomic Forecasts", *Journal of Public Administration Research and Theory*, 15, 1, 55–70.

Davidson, J. E. H., D. F. Hendry, F. Srba, and S. Yeo (1978) "Econometric Modelling of the Aggregate Time-series Relationship Between Consumers' Expenditure and Income in the United Kingdom", *Economic Journal*, 88, 352, 661–692.

Diebold, F. X., and R. S. Mariano (1995) "Comparing Predictive Accuracy", *Journal of Business and Economic Statistics*, 13, 3, 253–263.

Doornik, J. A. (2009a) "Autometrics", Chapter 4 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 88–121.

Doornik, J. A. (2009b) "Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data", draft, Economics Department, University of Oxford, Oxford, September 8 (www.doornik.com/flu/Doornik%282009%29_Flu.pdf).

Doornik, J. A., and D. F. Hendry (2013) *PcGive 14*, Timberlake Consultants Press, London (3 volumes).

Dyckman, T. R., and H. O. Stekler (1966) "Probabilistic Turning Point Forecasts", *Review of Economics and Statistics*, 48, 3, 288–295.

The Economist (2010) "America's Budget Deficit: Speak Softly and Carry a Big Chainsaw", *The Economist*, November 20, leader article.

Engstrom, E. J., and S. Kernell (1999) "Serving Competing Principals: The Budget Estimates of OMB and CBO in an Era of Divided Government", *Presidential Studies Quarterly*, 29, 4, 820–829.

Ericsson, N. R. (1992) "Parameter Constancy, Mean Square Forecast Errors, and Measuring Forecast Performance: An Exposition, Extensions, and Illustration", *Journal of Policy Modeling*, 14, 4, 465–495.

Ericsson, N. R. (2011a) "Improving Global Vector Autoregressions", draft, Board of Governors of the Federal Reserve System, Washington, D.C., June.

Ericsson, N. R. (2011b) "Justifying Empirical Macro-econometric Evidence in Practice", invited presentation, online conference *Communications with Economists: Current and Future Trends* commemorating the 25th anniversary of the *Journal of Economic Surveys*, November.

Ericsson, N. R. (2012) "Detecting Crises, Jumps, and Changes in Regime", draft, Board of Governors of the Federal Reserve System, Washington, D.C., November.

Ericsson, N. R. (2016) "Eliciting GDP Forecasts from the FOMC's Minutes Around the Financial Crisis", *International Journal of Forecasting*, 32, 2, 571–583.

Ericsson, N. R. (2017) "How Biased Are U.S. Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, this issue.

Ericsson, N. R., D. F. Hendry, and K. M. Prestwich (1998) "The Demand for Broad Money in the United Kingdom, 1878–1993", *Scandinavian Journal of Economics*, 100, 1, 289–324 (with discussion).

Ericsson, N. R., and J. Marquez (1993) "Encompassing the Forecasts of U.S. Trade Balance Models", *Review of Economics and Statistics*, 75, 1, 19–31.

Ericsson, N. R., and E. L. Reisman (2012) "Evaluating a Global Vector Autoregression for Forecasting", *International Advances in Economic Research*, 18, 3, 247–258.

Faust, J., and J. S. Irons (1999) "Money, Politics and the Post-war Business Cycle", *Journal of Monetary Economics*, 43, 1, 61–89.

Fildes, R., and H. O. Stekler (2002) "The State of Macroeconomic Forecasting", *Journal of Macroeconomics*, 24, 4, 435–468.

Frankel, J. (2011) "Over-optimism in Forecasts by Official Budget Agencies and Its Implications", *Oxford Review of Economic Policy*, 27, 4, 536–562.

Gamber, E. N., and J. P. Liebner (2017) "Comment on 'How Biased are US Government Forecasts of the Federal Debt?' ", *International Journal of Forecasting*, this issue.

Goldstein, M., G. L. Kaminsky, and C. M. Reinhart (2000) *Assessing Financial Vulnerability: An Early Warning System for Emerging Markets*, Institute for International Economics, Washington, D.C.

Granger, C. W. J. (1983) "Forecasting White Noise", in A. Zellner (ed.) *Applied Time Series Analysis of Economic Data*, Bureau of the Census, Washington, D.C., 308–314.

Granger, C. W. J. (1989) *Forecasting in Business and Economics*, Academic Press, Boston, Massachusetts, Second Edition.

Hendry, D. F. (1984) "Monte Carlo Experimentation in Econometrics", Chapter 16 in Z. Griliches and M. D. Intriligator (eds.) *Handbook of Econometrics*, Volume 2, North-Holland, Amsterdam, 937–976.

Hendry, D. F. (1999) "An Econometric Analysis of US Food Expenditure, 1931–1989", Chapter 17 in J. R. Magnus and M. S. Morgan (eds.) *Methodology and Tacit Knowledge: Two Experiments in Econometrics*, John Wiley and Sons, Chichester, 341–361.

Hendry, D. F. (2006) "Robustifying Forecasts from Equilibrium-correction Systems", *Journal of Econometrics*, 135, 1–2, 399–426.

Hendry, D. F., and J. A. Doornik (2014) *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*, MIT Press, Cambridge, Massachusetts.

Hendry, D. F., and S. Johansen (2015) "Model Discovery and Trygve Haavelmo's Legacy", *Econometric Theory*, 31, 1, 93–114.

Hendry, D. F., S. Johansen, and C. Santos (2008) "Automatic Selection of Indicators in a Fully Saturated Regression", *Computational Statistics*, 23, 2, 317–335, 337–339.

Hendry, D. F., and G. E. Mizon (2014) "Unpredictability in Economic Analysis, Econometric Modeling and Forecasting", *Journal of Econometrics*, 182, 1, 186–195.

Hendry, D. F., and F. Pretis (2013) "Anthropogenic Influences on Atmospheric CO_2", Chapter 12 in R. Fouquet (ed.) *Handbook on Energy and Climate Change*, Edward Elgar, Cheltenham, 287–326.

Hendry, D. F., and C. Santos (2010) "An Automatic Test of Super Exogeneity", Chapter 12 in T. Bollerslev, J. R. Russell, and M. W. Watson (eds.) *Volatility and Time Series Econometrics: Essays in Honor of Robert F. Engle*, Oxford University Press, Oxford, 164–193.

Holden, K., and D. A. Peel (1990) "On Testing for Unbiasedness and Efficiency of Forecasts", *The Manchester School*, 58, 2, 120–127.

Johansen, S., and B. Nielsen (2009) "An Analysis of the Indicator Saturation Estimator as a Robust Regression Estimator", Chapter 1 in J. L. Castle and N. Shephard (eds.) *The Methodology and Practice of Econometrics: A Festschrift in Honour of David F. Hendry*, Oxford University Press, Oxford, 1–36.

Johansen, S., and B. Nielsen (2013) "Outlier Detection in Regression Using an Iterated One-step Approximation to the Huber-skip Estimator", *Econometrics*, 1, 1, 53–70.

Johansen, S., and B. Nielsen (2016) "Asymptotic Theory of Outlier Detection Algorithms for Linear Time Series Regression Models", *Scandinavian Journal of Statistics*, 43, 2, 321–381 (with discussion and rejoinder).

Joutz, F., and H. O. Stekler (2000) "An Evaluation of the Predictions of the Federal Reserve", *International Journal of Forecasting*, 16, 1, 17–38.

Marczak, M., and T. Proietti (2016) "Outlier Detection in Structural Time Series Models: The Indicator Saturation Approach", *International Journal of Forecasting*, 32, 1, 180–202.

Martinez, A. B. (2011) "Comparing Government Forecasts of the United States' Gross Federal Debt", RPF Working Paper No. 2011–002, Research Program on Forecasting, Center of Economic Research, Department of Economics, The George Washington University, Washington, D.C., February.

Martinez, A. B. (2015) "How Good Are US Government Forecasts of the Federal Debt?", *International Journal of Forecasting*, 31, 2, 312–324.

Mincer, J., and V. Zarnowitz (1969) "The Evaluation of Economic Forecasts", Chapter 1 in J. Mincer (ed.) *Economic Forecasts and Expectations: Analyses of Forecasting Behavior and Performance*, National Bureau of Economic Research, New York, 3–46.

Mizon, G. E., and J.-F. Richard (1986) "The Encompassing Principle and its Application to Testing Non-nested Hypotheses", *Econometrica*, 54, 3, 657–678.

National Bureau of Economic Research (2012) "US Business Cycle Expansions and Contractions", webpage, National Bureau of Economic Research, Cambridge, MA, April (www.nber.org/cycles.html).

Nunes, R. (2013) "Do Central Banks' Forecasts Take Into Account Public Opinion and Views?", International Finance Discussion Paper No. 1080, Board of Governors of the Federal Reserve System, Washington, D.C., May.

Podkul, C. (2011) "Bernanke Rejects Alternatives to Raising the U.S. Debt Ceiling", *Washington Post*, July 15, p. A.13.

Pretis, F., L. Schneider, J. E. Smerdon, and D. F. Hendry (2016) "Detecting Volcanic Eruptions in Temperature Reconstructions by Designed Break-indicator Saturation", *Journal of Economic Surveys*, 30, 3, 403–429.

Ramsey, J. B. (1969) "Tests for Specification Errors in Classical Linear Least-squares Regression Analysis", *Journal of the Royal Statistical Society*, *Series B*, 31, 2, 350–371.

Romer, C. D., and D. H. Romer (2008) "The FOMC versus the Staff: Where Can Monetary Policymakers Add Value?", *American Economic Review*, 98, 2, 230–235.

Sargan, J. D. (1988) *Lectures on Advanced Econometric Theory*, Basil Blackwell, Oxford (edited and with an introduction by Meghnad Desai).

Sinclair, T. M., F. Joutz, and H. O. Stekler (2010) "Can the Fed Predict the State of the Economy?", *Economics Letters*, 108, 1, 28–32.

Sinclair, T. M., H. O. Stekler, and W. Carnow (2012) "A New Approach for Evaluating Economic Forecasts", *Economics Bulletin*, 32, 3, 2332–2342.

Stekler, H. O. (1967) "The Federal Budget as a Short-Term Forecasting Tool", *Journal of Business*, 40, 3, 280–285.

Stekler, H. O. (1972) "An Analysis of Turning Point Forecasts", *American Economic Review*, 62, 4, 724–729.

Stekler, H. O. (2002) "The Rationality and Efficiency of Individuals' Forecasts", Chapter 10 in M. P. Clements and D. F. Hendry (eds.) *A Companion to Economic Forecasting*, Blackwell Publishers, Oxford, 222–240.

Stekler, H. O. (2003) "Improving our Ability to Predict the Unusual Event", *International Journal of Forecasting*, 19, 2, 161–163.

Stuart, A., and J. K. Ord (1987) *Kendall's Advanced Theory of Statistics: Distribution Theory*, Volume 1, Oxford University Press, New York, Fifth Edition.

Tsuchiya, Y. (2013) "Are Government and IMF Forecasts Useful? An Application of a New Market-timing Test", *Economics Letters*, 118, 1, 118–120.

Vere-Jones, D. (1995) "Forecasting Earthquakes and Earthquake Risk", *International Journal of Forecasting*, 11, 4, 503–538.

White, H. (1990) "A Consistent Model Selection Procedure Based on m-testing", Chapter 16 in C. W. J. Granger (ed.) *Modelling Economic Series: Readings in Econometric Methodology*, Oxford University Press, Oxford, 369–383.

Yellen, J. L. (2014) "Testimony on 'The Economic Outlook' ", in *Hearing Before the Joint Economic Committee, Congress of the United States, 113th Congress, Second Session*, U.S. Government Printing Office, Washington, D.C., May 7.