Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

# Interpreting estimates of forecast bias

Neil R. Ericsson*

*Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551, USA*
*Department of Economics, The George Washington University, Washington, DC 20052, USA*

ARTICLE INFO

*Keywords:*
Bias
Debt
Federal government
Forecasts
Impulse indicator saturation
Heteroscedasticity
United States

ABSTRACT

This paper 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.

Published by Elsevier B.V. on behalf of International Institute of Forecasters.

## 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. The current paper 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 2 and 3 below show, time-varying bias is more consistent with the combined evidence in Ericsson (2017) and Gamber and Liebner (2017). Section 4 comments further on modeling with IIS.

## 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 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

* Correspondence to: Division of International Finance, Board of Governors of the Federal Reserve System, Washington, DC 20551, USA.
*E-mail addresses:* ericsson@frb.gov, ericsson@gwu.edu.

http://dx.doi.org/10.1016/j.ijforecast.2017.01.001
0169-2070/Published by Elsevier B.V. on behalf of International Institute of Forecasters.

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 Bontemps and Mizon (2008), Davidson, Hendry, Srba, and Yeo (1978), and Mizon and Richard (1986) 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.

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

That is, $w_t$ is normally and independently distributed with a constant mean $\delta_0$ and constant variance $\sigma^2$ 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 $\delta_1$ and $\delta_2$ 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 (1) itself.

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

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

$$\text{Model M1}: \; w_t = \delta_0 + \delta_1 I_{1t} + v_{1t}. \tag{3}$$

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

$$\text{Model M2}: \; w_t = \delta_0 + \delta_2 I_{2t} + v_{2t}. \tag{4}$$

For Model M1, the error $v_{1t}$ is $(\delta_2 I_{2t} + \varepsilon_t)$, so Model M1's mean squared error $\sigma_1^2$ is:

$$\sigma_1^2 = (\sigma^2 + \delta_2^2/T), \tag{5}$$

which is larger than $\sigma^2$, 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 $\sigma_2^2$ is:

$$\sigma_2^2 = (\sigma^2 + \delta_1^2/T), \tag{6}$$

which also is larger than $\sigma^2$.

One possible consequence of model specifications such as M1 and M2 is to shrink $t$-ratios on included variables. As Eqs. (5) and (6) 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 1. 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):

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

[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.

**Table 1**
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)<br>col. #1 | (b)<br>col. #2 | (c)<br>col. #3 | (d)<br>col. #4 | (e)<br>col. #5a | (e)<br>col. #5b |
| $I_{1990}$ | | | | 1.2<br>$\langle 1.5\rangle$ | $4.0^{***}$ | 2.96 |
| $I_{2001}$ | $2.4^{*}$<br>$\langle 2.7^{*}\rangle$ | | $3.5^{**}$<br>$\langle 3.8^{**}\rangle$ | | $4.8^{***}$ | 3.52 |
| $I_{2002}$ | | | $3.1^{**}$<br>$\langle 3.4^{**}\rangle$ | | $4.2^{***}$ | 3.12 |
| $I_{2003}$ | | | $2.6^{*}$<br>$\langle 2.9^{**}\rangle$ | | $3.6^{**}$ | 2.66 |
| $I_{2008}$ | $4.6^{***}$<br>$\langle 4.8^{***}\rangle$ | $3.8^{***}$<br>$\langle 3.9^{***}\rangle$ | $6.4^{***}$<br>$\langle 6.7^{***}\rangle$ | | $8.5^{***}$ | 6.27 |
| $I_{2009}$ | $2.5^{*}$<br>$\langle 2.7^{*}\rangle$ | | $3.6^{**}$<br>$\langle 3.8^{**}\rangle$ | | $4.8^{***}$ | 3.53 |
| $I_{2010}$ | | | $2.6^{*}$<br>$\langle 2.8^{*}\rangle$ | | $3.5^{**}$ | 2.57 |
| $\hat{\sigma}$ | 1.24<br>$\langle 1.28\rangle$ | 1.44<br>$\langle 1.58\rangle$ | 0.94<br>$\langle 0.91\rangle$ | 1.74<br>$\langle 1.88\rangle$ | 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.

(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 1 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 Eqs. (2), (3), and (4) 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 (5) and (6), 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 in Column #5a is rescaled by the ratio of Column #5a'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 1 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 Eq. (6), 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.

Please cite this article in press as: Ericsson, N. R., Interpreting estimates of forecast bias. International Journal of Forecasting (2017), http://dx.doi.org/10.1016/j.ijforecast.2017.01.001

**Table 2**
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 $\sigma_b$ equal to 2.842.

## 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 $w_t$:

$$w_t \sim NID(0, \sigma_a^2), \quad t = 1, \ldots, T_a; \quad \text{and} \tag{7}$$

$$w_t \sim NID(0, \sigma_b^2), \quad t = (T_a + 1), \ldots, T. \tag{8}$$

Based on the empirical setting for debt forecasts as analyzed with bare-bones IIS, Gamber and Liebner choose Eqs. (7)–(8) 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 2'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 $\sigma_a/\sigma_b$ or $\sigma_b/\sigma_a$, 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 2's column "Binomial solution (5%)", which closely matches the previous column, "Monte Carlo (5%)".

As Section 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 than at least six). The corresponding calculations appear in Table 2'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 $\sigma_a$ and $\sigma_b$ were greater, the implied heteroscedasticity could have been a likely explanation for IIS's empirical behavior. Specifically, if $\sigma_b$ were 2.842 rather than 2.122 (and $\sigma_a$ 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 2's final column, labeled "Binomial solution (1%) $[\sigma_b = 2.842]$".

Please cite this article in press as: Ericsson, N. R., Interpreting estimates of forecast bias. International Journal of Forecasting (2017), http://dx.doi.org/10.1016/j.ijforecast.2017.01.001

## 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 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 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 (2009).

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 (2011) 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 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 (2011) 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.

## 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.

## Acknowledgments

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 David Hendry, Rob Hyndman, Andrew Kane, Aaron Markiewitz, Jaime Marquez, Andrew Martinez, Felix Pretis, Tara Sinclair, and two anonymous referees for helpful discussions and comments. 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.

## References

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

Castle, J. L., Doornik, J. A., & Hendry, D. F. (2012). Model selection when there are multiple breaks. *Journal of Econometrics*, *169*(2), 239–246.

Castle, J. L., Doornik, J. A., Hendry, D. F., & Pretis, F. (2015). Detecting location shifts during model selection by step-indicator saturation. *Econometrics*, *3*(2), 240–264.

Davidson, J. E. H., Hendry, D. F., Srba, F., & Yeo, S. (1978). Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal*, *88*(352), 661–692.

Doornik, J. A. (2009). Autometrics. In J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics: a Festschrift in honour of David F. Hendry* (pp. 88–121). Oxford: Oxford University Press, (Chapter 4).

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

Ericsson, N.R. (2011). 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. (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.

Gamber, E. N., & Liebner, J. P. (2017). Comment on 'How biased are US government forecasts of the federal debt?' *International Journal of Forecasting*, this issue.

Hendry, D. F. (1984). Monte Carlo experimentation in econometrics. In Z. Griliches & M. D. Intriligator (Eds.), *Handbook of econometrics, Vol. 2* (pp. 937–976). Amsterdam: North-Holland, (Chapter 16).

Hendry, D. F. (1999). An econometric analysis of US food expenditure, 1931–1989. In J. R. Magnus & M. S. Morgan (Eds.), *Methodology and tacit knowledge: two experiments in econometrics* (pp. 341–361). Chichester: John Wiley and Sons, (Chapter 17).

Hendry, D. F., & Doornik, J. A. (2014). *Empirical model discovery and theory evaluation: automatic selection methods in econometrics*. Cambridge, Massachusetts: MIT Press.

Hendry, D. F., Johansen, S., & Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, *23*(2), 317–335. 337–339.

Hendry, D. F., & Santos, C. (2010). An automatic test of super exogeneity. In T. Bollerslev, J. R. Russell, & M. W. Watson (Eds.), *Volatility and time series econometrics: essays in honor of Robert F. Engle* (pp. 164–193). Oxford: Oxford University Press, (Chapter 12).

Johansen, S., & Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In J. L. Castle & N. Shephard (Eds.), *The methodology and practice of econometrics: a Festschrift in honour of David F. Hendry* (pp. 1–36). Oxford: Oxford University Press, (Chapter 1).

Johansen, S., & Nielsen, B. (2013). Outlier detection in regression using an iterated one-step approximation to the Huber-skip estimator. *Econometrics*, *1*(1), 53–70.

Johansen, S., & Nielsen, B. (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.

Marczak, M., & Proietti, T. (2016). Outlier detection in structural time series models: the indicator saturation approach. *International Journal of Forecasting*, *32*(1), 180–202.

Mizon, G. E., & Richard, J.-F. (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica*, *54*(3), 657–678.

Pretis, F., Schneider, L., Smerdon, J. E., & Hendry, D. F. (2016). Detecting volcanic eruptions in temperature reconstructions by designed break-indicator saturation. *Journal of Economic Surveys*, *30*(3), 403–429.

Sargan, J. D. (1988). *Lectures on advanced econometric theory*. Oxford: Basil Blackwell, edited and with an introduction by Meghnad Desai.

Stuart, A., & Ord, J. K. (1987). *Kendall's advanced theory of statistics: distribution theory*, Vol. 1 (5th ed.). New York: Oxford University Press.

Please cite this article in press as: Ericsson, N. R., Interpreting estimates of forecast bias. International Journal of Forecasting (2017), http://dx.doi.org/10.1016/j.ijforecast.2017.01.001