Data Models for Forecasting: No Reason to Expect Improved Accuracy

J. Scott Armstrong
jscott@upenn.edu
The Wharton School, University of Pennsylvania
and Ehrenberg-Bass Institute, University of South Australia

Kesten C. Green
kesten.green@unisa.edu.au
University of South Australia Business School
and Ehrenberg-Bass Institute, University of South Australia

January 26, 2019 - Version 22

In the mid-1900s, there were two streams of thought about forecasting methods. One stream—led by econometricians—was concerned with developing causal models by using prior knowledge and evidence from experiments. The other was led by statisticians, who were concerned with identifying idealized "data generating processes" and with developing models from statistical relationships in data, both in the expectation that the resulting models would provide accurate forecasts.

At that time, regression analysis was a costly process. In more recent times, regression analysis and related techniques have become simple and inexpensive to use. That development led to automated procedures such as stepwise regression, which selects "predictor variables" on the basis of statistical significance. An early response to the development was titled, "Alchemy in the behavioral sciences" (Einhorn, 1972). We refer to the product of data-driven approaches to forecasting as "data models."

The M4-Competition (Makridakis, Spiliotis, Assimakopoulos, 2018) has provided extensive tests of whether data models—which they refer to as "ML methods"—can provide accurate extrapolation forecasts of time series. The Competition findings revealed that data models failed to beat naïve models, and established simple methods, with sufficient reliability to be of any practical interest to forecasters. In particular, the authors concluded from their analysis, "The six pure ML methods that were submitted in the M4 all performed poorly, with none of them being more accurate than Comb and only one being more accurate than Naïve2" (p. 803.)

Over the past half-century, much has been learned about how to improve forecasting by conducting experiments to compare the performance of reasonable alternative methods. On the other hand, despite billions of dollars of expenditure, the various data modeling methods have not contributed to improving forecast accuracy. Nor can they do so, as we explain below.

1. Data Models Violate the Golden Rule of Forecasting

The Golden Rule of forecasting is to "Be conservative by using prior knowledge about the situation and about forecasting methods." It was developed and tested by Armstrong, Green and Graefe (2015). The Golden Rule paper provided 28 evidence-based guidelines, which were tested by reviewing papers with relevant evidence. The review identified 105 papers with 150 experimental comparisons. All comparisons supported the guidelines. On average, ignoring a single guideline increased forecast error by more than 40% on average.

One way to incorporate prior knowledge is to use the findings of experiments on which methods work best for the type of situation being forecast. In addition, in practical forecasting situations one can use experts' domain knowledge about the expected directions of trends. In an earlier M-Competition, these two sources of knowledge were implemented by "Rule-based forecasting" (Collopy and Armstrong 1992).

Rule-based forecasting uses domain knowledge to select extrapolation models based on 28 conditions of the data in order to produce combined forecasts. It then uses 99 simple rules to weight each of the forecasting methods. For six-year ahead forecasts, the *ex ante* forecasts provided a 42% reduction compared to those from equal-weights combinations. Other sources of prior knowledge that can be used include decomposition of time-series by level and change (Armstrong and Tessier, 2015), and by causal forces (Armstrong & Collopy, 1993.)

2. Data Models Violate Occam's Razor

Occam's Razor, a principle that was described by Aristotle, states that one should prefer the simplest hypothesis or model that does the job. A review of 32 studies found 97 comparisons between simple and complex methods (Green and Armstrong, 2015). None found that complexity improved forecast accuracy. On the contrary, complexity increased errors by an average of 27% in the 25 papers with quantitative comparisons.

Unsurprisingly, then, all of the validated methods for forecasting are simple (see the Methods checklist at ForecastingPrinciples.com.)

3. Data Models Enable Advocacy, Leading to Unscientific and, Potentially, Unethical Practices

The nature of data modeling procedures is such that researchers can develop data models to provide the forecasts that they know their clients or sponsors would prefer. Doing so helps them to get grants and promotions. They can also use them to support their own preferred hypotheses.

It is not difficult to obtain statistically significant findings. And, of course, all tested relationships become statistically significant if the sample sizes are large enough.

Despite the widespread use of statistical significance testing, the technique has never been validated. On the contrary, decades of research have found that statistical significance testing harms scientific advances. For example, Koning, et al. (2005) concluded that statistical tests showed that combining forecasts did not reduce forecasting errors in the M3-Competition. Those findings were refuted in Armstrong (2007). Research since that time has continued to find gains in accuracy from combining (see Armstrong and Green, 2018, for a recent summary of the research.) As Makridakis et al. (2018, p. 803) concluded from their analysis of the M4-Competition forecasts, "The combination of methods was the king of the M4. Of the 17 most accurate methods, 12 were 'combinations' of mostly statistical approaches." In that case, the combinations were done within methods. Combinations across validated methods have been found to be even more effective at reducing forecast errors due to the different knowledge, information, and biases that different methods bring (Armstrong and Green, 2018.)

4. Data Models Violate the Guidelines for Regression analysis

We rated data modeling procedures against the *Checklist for forecasting using regression analysis* (at ForecastingPrinciples.com.) By our ratings, typical procedures for estimating data models violate at least 15 of the 18 guidelines in the checklist. In particular, they violate the guidance to limit models to three or fewer causal variables when using non-experimental data. A summary of evidence on the limitations of regression analysis is provided in Armstrong (2012).

5. Data Models Failed when Previously Tested

Before the M4 Competition, an extensive evaluation of published time-series data mining methods using diverse data sets and performance measures found insufficient evidence to conclude that the methods could be useful in practice (Keogh and Kasetty 2003). That paper has been cited over 1,700

times to date. Those who contemplate using data models might want to familiarize themselves with these findings before proceeding.

6. "Knowledge Models" for Forecasting When Data are Plentiful

Benjamin Franklin proposed a simple method that we now refer to as "knowledge models" and, previously, as "index models." The procedure is as follows:

a. Use domain knowledge to specify:
   - all important causal variables,
   - directions of their effects, and
   - if possible, magnitudes of their relationships.

   Variables should be defined to as to be positively related to the thing that is being forecast.
   There is no limit on the number of causal variables that can be used.
   Variables may be as simple as binary, or "dummy" variables to econometricians.

b. Use equal weights and standardized variables in the model unless there is strong evidence of differences in relative effect sizes. Equal weights are often more accurate than regression weights, especially when there are many variables and where prior knowledge about relative effect sizes is poor.

c. Forecast values of the causal variables in the model.

d. Apply the model weights to the forecast causal variable values and sum to calculate a score.

e. The score is a forecast.
   A higher score means that the thing being forecast is likely to be better, greater, or more likely than would be the case for a lower score.
   If there are sufficient data to do so, estimate a single regression model that relates scores from the model with the actual values of the thing being forecast.

Evidence to date suggests that knowledge models are likely to produce forecasts that are more accurate than those from data models in situations where many causal variables are important. One study found error reductions of 10% to 43% compared to established regression models for forecasting elections in the US and Australia (Graefe, Green, and Armstrong, 2019).

Conclusion: Data Models Should Never be Used for Forecasting, and a Suggestion

Forecasters have used data modeling methods of various kinds since the 1960s. Despite the resources that have been devoted to finding support for statistical methods that use big data, we have been unable to find scientific evidence to support their use under any conditions.

Science advances not by looking for evidence to support a favorite hypothesis, but by using prior knowledge and experiments to test alternative hypotheses in order to discover useful principles and methods, as the M4-Competition has done.

Our suggestion for future competitions is that when competitors submit their models and forecasts, they should use prior research to explain the principles and methods that they used, their prior hypotheses on relative accuracy under different conditions, and which category of method their model belongs to. That would allow the competition organizers, commentators, and other researcher to compare the results by category of method taking into account prior evidence, rather than resorting to *ex post* speculation on what can be learned from the results.

References

Armstrong, J.S. (2007). Significance Tests Harm Progress in forecasting. *International Journal of Forecasting,* 23, 321-336

Armstrong, J.S. (2012). **Illusions in regression analysis**, *International Journal of Forecasting*, 28, 689-694.

Armstrong, J. S. & Collopy, F. (1993). Causal Forces: Structuring Knowledge for Time-series Extrapolation, *Journal of Forecasting, 12*, 103-115.

Armstrong, J.S., Collopy, F. & Yokum, J.T. (2005). Decomposition by Causal Forces: A Procedure for Forecasting Complex Time Series, *International Journal of Forecasting*, *21* (2005), 25-36.

Armstrong, J.S. & Green, K. C. (2018). Forecasting methods and principles: Evidence-based checklists, *Journal of Global Scholars in Marketing Science*, 28, 103-159.

Armstrong, J. S, Green, K.C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative, *Journal of Business Research*, 68, 1717-1731.

Armstrong, J.S. & Tessier, T. (2015). Decomposition of time-series by level and change, *Journal of Business Research*, *68* (2015), 1755-1758.

Collopy, F. & Armstrong, J. S. (1992). Rule-Based Forecasting: Development and Validation of an Expert Systems Approach to Combining Time Series Extrapolations, *Management Science*, 38,1394-1414.

Einhorn, H. (1972). Alchemy in the behavioral sciences. *Public Opinion Quarterly, 36,* 367-378.

Graefe, A., Green, K. C. & Armstrong, J. S. (2019). Accuracy gains from conservative forecasting: Tests using variations of 19 econometric models to predict 154 elections in 10 countries. *PLoS ONE*, 14(1), e0209850.

Green, K. C. & Armstrong, J.S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research,* 68, 1678-1685.

Keogh, E. & Kasetty, S. (2003). On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, 7, 349–371.

Koning, A.J., Franses, P. H., Hibon, M., & Stekler, H. O..(2005). The M3-Competition: Statistical tests of the results. *International Journal of Forecasting*, 21, 397-409.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34, 802-808.

Acknowledgements: We thank Robert Fildes, Andreas Graefe, and Eamon Keogh for reviewing this paper.