

MINIMUM SAMPLE SIZE REQUIREMENTS FOR SEASONAL FORECASTING MODELS

Rob J. Hyndman and Andrey V. Kostenko

INTRODUCTION

“How much data do I need?” is probably the most common question our business clients ask us. The answer is rarely simple. Rather it depends on the type of statistical model being used and on the amount of random variation in the data. Our usual answer is “as much as possible” because the more data we have, the better we can identify the structure and patterns that are used for forecasting.

We are sometimes told that “there is no point using data from more than two or three years ago because demand is changing too quickly for older data to be relevant.” This reflects a misunderstanding of what a statistical model does; it is designed to describe the way data changes. A good model will allow for changing trend and changing seasonal patterns. The issue is not whether things have changed, but whether the way things change can be modeled.

Rarely do we have so much data that we can afford to omit any. More often, we are trying to make do with limited historical information. In these days of cheap data storage and high-speed computing, there is no reason not to use all available and relevant historical data when building statistical models. However, we are sometimes asked to forecast very short time series, and it is helpful to understand the minimum sample size requirements when fitting statistical models to such data.

RANDOMNESS AND PARAMETERS

The number of data points required for any statistical model depends on at least two things: the number of model

KEY POINTS

- How much data you need to forecast using a seasonal model depends on the type of model being used and the amount of random variation in the data.
- There are specific *minimum* sample size requirements and we document these for common seasonal forecasting models.
- The minimum requirements apply when the amount of random variation in the data is very small. Real data often contain a lot of random variation, and sample size requirements increase accordingly.
- Some published tables on sample size requirements are overly simplified.

coefficients (parameters) to estimate and the amount of randomness in the data.

Consider the simple scenario of linear regression. Here, $y_t = a + bt + e_t$, where y_t is the observed series at time t , t is a time index (1,2,...) and e_t is a random error term. There are 2 parameters: the intercept a and the slope b .

In Figure 1, next page, we show two data sets, each with only 5 observations. The errors in the right plot have greater variation than the errors in the left plot. In both cases, however, the data were simulated from the same model, using $a = 0$ and $b = 1$. Then the model was estimated from the 5 data points.

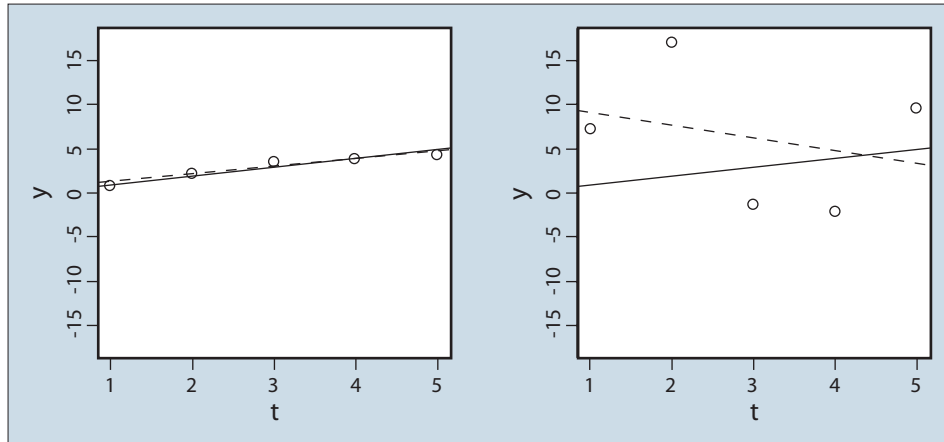


Rob Hyndman is Professor of Statistics at Monash University, Australia, and Editor in Chief of the **International Journal Of Forecasting**. As a consultant, he has worked with over 200 clients during the past 20 years, on projects covering all areas of applied statistics, from forecasting to the ecology of lemmings. He is coauthor of the well-known textbook, **Forecasting: Methods And Applications** (Wiley, 1998), and he has published more than 50 journal articles. Rob is Director of the Business and Economic Forecasting Unit, Monash University, one of the leading forecasting research groups in the world.



Andrey Kostenko is an independent consultant on complex systems. He received his first degree from a state university in Russia and, with the help of a prestigious scholarship from the President of the Russian Federation, an MBA from an institution in the UK. Recently, he was admitted as a PhD student to the Business and Economics Forecasting Unit, Monash University, Australia.

Figure 1. Effect of Small vs. Large Random Variation



The solid line shows the true model and the dashed line shows the fitted model. The left-hand regression is fine because the error variance is small. The right-hand regression is inaccurate because the data contain too much random variation.

To accurately estimate a model where the data contain a lot of random variation, it is necessary to have a lot of data. On the other hand, if the data have little variation, it is possible to estimate a model more or less accurately with only a few observations.

From a purely statistical point of view, **it is always necessary to have more observations than parameters.** So, theoretically, it is feasible to estimate the regression line above with as few as 3 observations. (Any less and the estimated parameters have infinite standard errors, and therefore prediction intervals will be infinitely wide.) For the data in the left-hand plot, a regression through 3 observations still gives reasonable forecasting results. However, for the data in the right-hand plot, one would need thousands of observations to get the same level of accuracy that is possible with 3 observations in the left-hand plot.

This example illustrates that the size of the data set depends on both the model and the level of randomness in the data. To give a reliable answer to questions of sample size, it is essential to take the variability of the data into account.

MINIMUM REQUIREMENTS FOR THREE COMMON METHODS

Regression with Seasonal Dummies

When data are seasonal, a common way of handling the seasonality in a regression framework is to add “dummy

variables.” For example, with quarterly data, we can define three new variables:

- $Q_{1,t} = 1$ when t is in the first quarter and zero otherwise;
- $Q_{2,t} = 1$ when t is in the second quarter and zero otherwise;
- $Q_{3,t} = 1$ when t is in the third quarter and zero otherwise.

Then the regression model is:

$$y_t = a + bt + c_1Q_{1,t} + c_2Q_{2,t} + c_3Q_{3,t} + e_t$$

It is not necessary to include a term for the fourth quarter because it is implicitly defined whenever the first 3 dummies are equal to zero. Note that any one of the quarters could be omitted, not necessarily the fourth quarter. Alternatively, we could omit the intercept and include seasonal dummy variables for all 4 quarters.

Whatever approach is taken, if m is the number of months or quarters in a year, then m parameters are required. A regression usually also has another parameter for the time trend. So the total number of parameters in the model is $m+1$. Consequently, $m+2$ observations is the theoretical minimum number for estimation; that is, 6 observations for quarterly data and 14 observations for monthly data. But this will be **sufficient** only when there is almost no randomness. For most realistic problems, substantially more data are required.

Holt-Winters Methods

Holt-Winters (exponential smoothing) methods require the estimation of up to 3 parameters (smoothing weights) for the Level, Trend, and Seasonal components of the data.

Hence, Holt-Winters forecasting is often thought of as a 3-parameter method. However, the starting values for the Level, Trend, and Seasonal are also parameters that should be estimated.

For monthly data, there are 2 parameters associated with the initial level and initial trend values and 11 extra parameters associated with the initial seasonal components. (We don't need a starting value for the seasonal because it can be calculated from the other 11 by imposing the constraint that the values must sum to zero for the additive model and to 12 for the multiplicative model.)

Similarly, for quarterly data, there are an extra 5 parameters. These are often forgotten because the Holt-Winters equations are usually estimated using heuristic methods (see Makridakis et al., 1998). A better modelling approach (giving more accurate forecasts) is to estimate these starting values along with the smoothing parameters (Hyndman et al., 2002). Just because the starting values are not often estimated along with the other parameters does not mean they can be ignored. They are still parameters which require estimation, even if that is sometimes done using relatively simplistic approaches. In fact, the issue of specifying these parameters becomes particularly important in small sample sizes, especially when low values of the smoothing parameters are used, as they can have a big effect on the forecast values.

In general, for data with m seasons per year, there are $m+1$ initial values and 3 smoothing parameters, giving $m+4$ parameters altogether. Consequently, $m+5$ observations is the theoretical minimum number for estimation: that is, 9 observations for quarterly data and 17 observations for monthly data. Again, these minima are not necessarily adequate to deal with randomness in the data.

ARIMA Models

Seasonal ARIMA models are usually described using the notation $ARIMA(p,d,q)(P,D,Q)_m$ where each of the letters in parentheses indicates some aspect of the model. See Makridakis et al. (1998) for a description of seasonal ARIMA models.

A seasonal ARIMA model has $p+q+P+Q$ parameters. However, if differencing is required, an additional $d+mD$ observations are lost. So a total of $p+q+P+Q+d+mD$ effective parameters are used in the model. For example,

the famous “airline” model of Box et al. (1994) is a monthly $ARIMA(0,1,1)(0,1,1)_{12}$ model and so contains $0+1+0+1+1+12 = 15$ parameters.

Consequently, at least $p+q+P+Q+d+mD+1$ observations are required to estimate a seasonal ARIMA model. For the airline model, at least 16 observations are required. Note that this is actually **less** than for the Holt-Winters method applied to monthly data. There is a widely-held misconception that ARIMA models are complex (and therefore need more data) while exponential smoothing methods such as Holt-Winters are simple (and need less data). The truth is more complicated.

THE CONSEQUENCES OF RANDOMNESS IN THE DATA

Randomness in the data implies that the minimum statistical requirements discussed in the previous section will be insufficient to estimate seasonal models. But how much more data will we need to compensate for randomness? While there unfortunately are no specific rules, the general principle is straightforward.

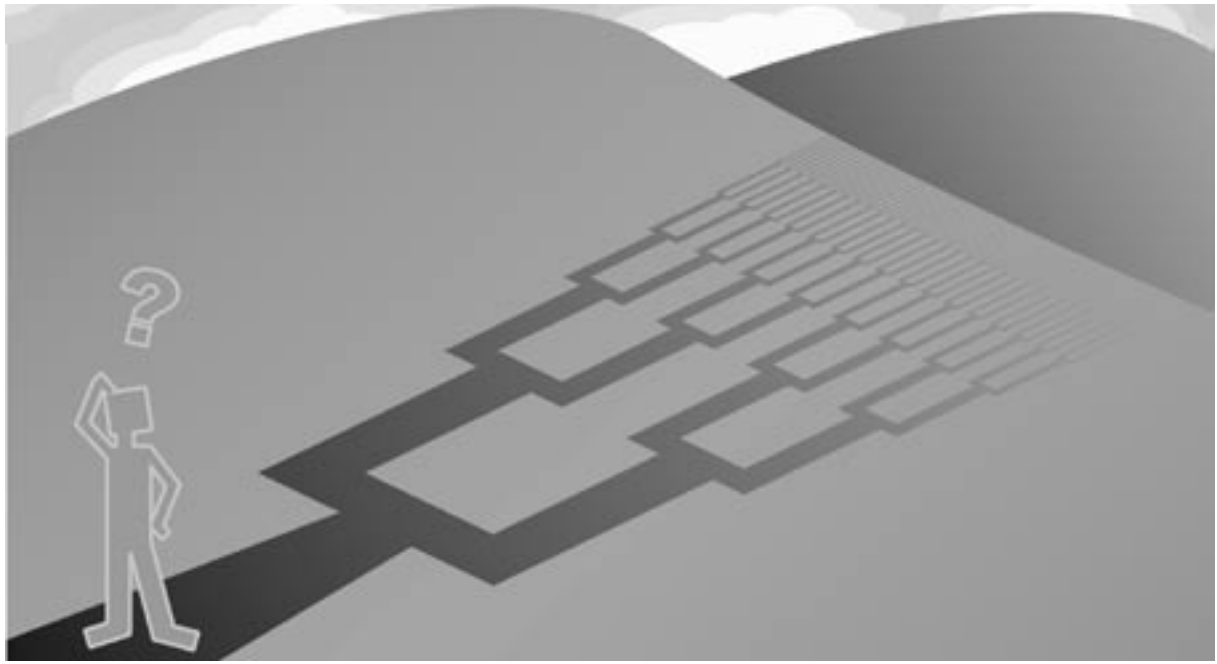
One way to think about the data requirement is in terms of the width of the prediction intervals (margins for error) around the point forecasts. The minimum sample sizes we gave above are the numbers for which the prediction intervals are finite. Smaller sample sizes than this and the PIs are infinite. As sample size (n) increases, the PIs decrease at a rate proportional to the square root of n . If you quadruple the length of the series, you will cut in half the width of a PI.

USING SUPPLEMENTAL INFORMATION

When data are scarce, consider using other information in addition to the available data. As Michael Leonard shows in the next commentary in this Foresight section, it is possible to use analogous time series that exhibit patterns similar to the time series being forecast. Duncan et al. (2001) show how to draw information from analogous time series using Bayesian pooling. Bayesian forecasting is often a productive choice when data are limited because it allows for the inclusion of other information, including expert opinion.

CONCLUSION

Forecasting short time series is fraught with difficulty, and sometimes we are compelled to provide forecasts from fewer data points than we would like to have. We have discussed the



minimum number of observations that are required to estimate some popular forecasting models, but it must be emphasized that for many practical problems substantially more data are required before the resulting forecasts can be trusted.

Some authors report universal minimum data requirements for different forecasting techniques, including those designed for seasonal time series. For example, see Hanke et al. (1998, p.73). Such tables are misleading, because they ignore the underlying variability of the data, and, as we have stressed, data requirements depend critically on the variability of the data.

Having short time series also significantly complicates the development and verification of models intended to produce seasonal forecasts. What may be irrelevant when you have a lengthy time series can be very important for short data samples. For example, when modelling short seasonal time series, the choice of starting values becomes critical.

In short, there is no easy answer. One certainty is that it is always necessary to have more observations than parameters. But since in most practical applications the data exhibit a lot of random variation, it is usually necessary to have many more observations than parameters.

REFERENCES

- Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control* (3rd ed.), Englewood Cliffs, NJ: Prentice-Hall.
- Duncan, G., Gorr, W. & Szczypula, J. (2001). Forecasting analogous time series, in J. S. Armstrong (Ed.), *Principles of Forecasting*, Norwell, MA: Kluwer Academic Publishers, 195-213.
- Hanke, J. E., Reitsch, A. G. & Wichern, D. (1998). *Business Forecasting* (6th ed.), Englewood Cliffs, NJ: Prentice-Hall.
- Hyndman, R. J., Koehler, A. B., Snyder, R. D. & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, 18 (3), 439-454.
- Makridakis, S. G., Wheelwright, S. C. & Hyndman, R. J. (1998). *Forecasting: methods and applications* (3rd ed.), New York : John Wiley & Sons.

Contact Info:

Rob J. Hyndman
Monash University, Australia
Rob.Hyndman@buseco.monash.edu

Andrey V. Kostenko
Independent Complex Systems Consultant, Russia
Andrey.Kostenko@ruenglish.com