Fitting models to short time series

Hyndsight (https://robjhyndman.com/hyndsight)

4 March 2014

forecasting (https://robjhyndman.com/categories/forecasting), R (https://robjhyndman.com/categories/r), statistics (https://robjhyndman.com/categories/statistics), time series (https://robjhyndman.com/categories/time-series)

Following my post on fitting models to long time series (https://robjhyndman.com/hyndsight/long-time-series/), I thought I'd tackle the opposite problem, which is more common in business environments.

I often get asked how *few* data points can be used to fit a time series model. As with almost all sample size questions, there is no easy answer. It depends on the **number of model parameters to be estimated and the amount of randomness in the data**. The sample size required increases with the number of parameters to be estimated, and the amount of noise in the data.

Using least squares estimation, or some other non-regularized estimation method, it is possible to estimate a model only if you have more observations than parameters. (If you use the LASSO, or some other regularization technique, it is possible to estimate a model with fewer observations than parameters.) However, there is no guarantee that a fitted model will be any good for forecasting, especially when the data are noisy.

Some textbooks provide rules-of-thumb giving minimum sample sizes for various time series models. These are misleading and unsubstantiated in theory or practice. Further, they ignore the underlying variability of the data and often overlook the number of parameters to be estimated as well. There is, for example, no justification whatever for the magic number of 30 often given as a minimum for ARIMA modelling.

The only reasonable approach is to first check that there are enough observations to estimate the model, and then to test if the model performs well out-of-sample. With short series, there is not enough data to allow some observations to be witheld for testing purposes. However, the AIC can be used as a proxy for the one-step forecast out-of-sample MSE (see here (https://robjhyndman.com/hyndsight/aic/)). The AIC allows both the number of parameters and the amount of noise to be taken into account.

What tends to happen with short series is that the AIC suggests very simple models because anything with more than one or two parameters will produce poor forecasts due to the estimation error. I applied the `auto.arima()` function from the **forecast** package in R to all the series from the M-competition with fewer than 20 observations. There were a total of 144 series, of which 32 had models with zero parameters (random walks), 95 had models with one parameter, 15 had models with two parameters and 2 series had models with three parameters. For what it's worth, here is the code.

```
library(Mcomp)

n <- unlist(lapply(M1,function(x){length(x$x)}))
n <- n[n<20]
series <- names(n)
nparam <- numeric(length(n))
for(i in 1:length(n))
{
  fit <- auto.arima(M1[[series[i]]]$x)
  nparam[i] <- length(fit$coef)
}
table(nparam)
```

Seasonal models bring their own difficulties because the seasonality usually takes up $m - 1$ degrees of freedom where m is the seasonal period (e.g., $m = 12$ for monthly data). Fourier terms (https://robjhyndman.com/hyndsight/longseasonality/) are one way to reduce the problem — useful whenever the ratio of m to sample size is large. Further comments on seasonality and sample size are in my short *Foresight* paper with Andrey Kostenko: "Minimum sample size requirements for seasonal forecasting models" (https://robjhyndman.com/papers/shortseasonal.pdf), although I wrote that for a statistically unsophisticated audience, so there is no mention of the LASSO or AIC as possible solutions.

« *Previous*: Fitting models to long time series (https://robjhyndman.com/hyndsight/long-time-series/)

Next: Forecasting weekly data » (https://robjhyndman.com/hyndsight/forecasting-weekly-data/)

All Hyndsight posts by date (https://robjhyndman.com/hyndsight/)

4 Comments **Hyndsight** Login

Recommend Tweet Share Sort by Oldest

Join the discussion…

LOG IN WITH

OR SIGN UP WITH DISQUS ?

Name

Giorgio • 5 years ago

I agree that on short time series simple models are usually better.
Thus one could consider candidate models such as random walk or seasonal naive (forecast for July 2014 is the actual value of July 2013).

Such models do not require parameter estimation; thus their out-of-sample error can be estimated on the training data.
A problem is that it is questionable comparing this kind of error with the AIC of e.g. some low-parameterized ARIMA models.

Still I think naive predictor could be considered among the candidates model, when time series are very short.

• Reply • Share ›

saima • 4 years ago

if we have a short multiple time series of weather data for about 20 days, will it be reasonable to assume stationarity and fit VAR model? I did not get very good forecasts using VECM in comparison with VAR in levels.

• Reply • Share ›

saima → saima • 4 years ago

for example, i find one step ahead forecasts for 5 days in a row, and moving the training data of 20 days each time by excluding the first observation and including the latest observation. the time series has different behavior in terms of stationarity. for stationary time series, i use VAR model and for non stationary i use VECM. is this reasonable when the 5 days are in a row means just by the variation of one value time series shifts from stationary to non stationary (checked by ndiffs and other tests). it is quite confusing for me. so, i find forecasts by VAR model for each day assuming stationarity since time series is quite short and does not show any trend and seasonal component. this gives better results than VECM. is this reasonable?

• Reply • Share ›

Mark Adamson • 3 years ago

I'm using a reasonably long time-series for calibration of an ARIMAX model, but then using this model on a short time-series to actually forecast. This is because I have to deal with a changing population.

Presumably I am reasonably safe to use a more complex ARIMA model since the calibration is able to use a long dataset. I make sure that the calibration uses a static population.

I will be doing some back-testing to see for myself, but any thoughts you have would be appreciated

• Reply • Share ›

ALSO ON **HYNDSIGHT**

Forecasting in NYC: 25-27 June 2018

8 comments • a year ago

Avatar Rob J Hyndman — No European plans for 2018. Maybe 2019? As for datacamp, that will have to wait until I finish a new

A brief history of time series forecasting competitions

6 comments • a year ago

Avatar Jeong-Yoon Lee — Thanks for the great review!I'd like to draw your attention to KDD Cup 2018 as well. Since 1997,

 (https://robjhyndman.com/)

Contact

- Department of Econometrics & Business Statistics, Monash University, Clayton VIC 3800, Australia. (https://monash.edu/business/ebs)
- Rob.Hyndman@monash.edu (mailto:Rob.Hyndman@monash.edu)
- @robjhyndman (https://twitter.com/robjhyndman) on Twitter
- robjhyndman (https://github.com/robjhyndman/) on GitHub
- Google Scholar (https://scholar.google.com/citations?user=vamErfkAAAAJ)

Subscribe

To receive updates from this site, you can subscribe to using an RSS feed reader (http://feeds.feedburner.com/ProfessorRobJHyndman) or by email (https://feedburner.google.com/fb/a/mailverify?uri=ProfessorRobJHyndman).

Search

Search the site