Time Series Analytics Notes

Introduction

A **time series** is a sequence of observations on a phenomenon taken at successive points in time. Examples include the daily price of a stock like General Motors, or quarterly sales of a company, or the number of clicks on an internet ad each minute. If *Y* denotes the value being observed, like the stock price of GM, then the observations can be listed as a sequence in time: $Y_1, Y_2, ..., Y_t, ...$, going from the first observation (Y_1), to the second (Y_2), …, with a generic observation at time *t* being denoted Y_t.[1]

There are two primary reasons for studying time series:

- To forecast the future
- To explain the present

A **forecast** is simply a numerical guess about a future value, together with a numerical assessment of the uncertainty of the guess.

> Ex. An example of a forecast: I forecast that annual sales of Acme Internet Consultants will be $20 million next year, and I am 90% confident that actual sales will be within the range $19 million - $21 million.
>
> Ex. An example of an explanation of the present: The price of Acme stock depends 75% upon the general economy and 25% upon company-specific developments.

Time series analysis, in both its forecasting and its explanatory aspects, is **supervised**. That is, there is a response (dependent) variable that we want to forecast and/or explain in terms of predictor (independent) variable(s). Both the response and predictor variable(s) are time series. The response variable provides the supervision.

Both forecasting and explanation depend upon having a valid statistical model for the time series. A **statistical model** is a set of **specifications**, or hypotheses, about the way that the time series behaves. The specifications can – and should! – be tested for validity before using the model to make forecasts or to offer explanations.

> Ex. The Random Sample time series model says that the observations are independent and identically distributed (the specifications). You will soon make the acquaintance of the Random Sample model.

To help us develop a time series model, we will have some past data (and also the present) on the supervising response variable *Y* and predicting variable(s) *X* up to a given time *t* to aid in developing the model and making the forecast/explanation. The data may be denoted $y_1, y_2, ..., y_{t-1}, y_t$ and $x_1, x_2, ..., x_{t-1}, x_t$.[2, 3]

---

[1] In time series, *t* is generally used as a subscript letter to stand for a generic unspecified time period. Time series can also be observed so frequently that the observations are essentially continuous. An example is the amplitude of a radar signal.

[2] The distinction between upper case Y_t and lower case y_t is that Y_t denotes a random variable, which has no definite value before it has been observed, and y_t denotes a specific value that has been observed for Y_t. In other words, Y_t is potential data, and y_t is actual data.

[3] Writing the data in this form implicitly assumes that the predictors are observed at the same times as the response variable. Although that is usually the case, the times of observation may not coincide. Non-coincident observations pose a more difficult problem that is usually dealt with by some form of interpolation. We will not explicitly deal with the case of non-coincident observations.

The predicting *X* variables may include the past of *Y* (i.e., $Y_1, Y_2, ..., Y_{t-1}$ if *t* is the present). At time *t* the past data for *Y*, namely, $y_1, y_2, ..., y_{t-1}$, will be known to us and therefore legitimately available to us to help forecast the future of *Y*. More often than not there is correlation between the past of *Y* and its future (or present). This correlation is called **autocorrelation** or **serial correlation**. Exploiting this relationship is a major feature of time series analysis that is not available in statistical analysis of data that are not time series (i.e., cross-sectional data). Time series models that include the past of *Y* are called **dynamic models**.

When we forecast, we use past data to guess the future, based upon a model for the data. When we explain, we again use past data and/or contemporaneous data to build a model to account for a phenomenon. The model that we select for the data is very important. Different models have different ways to forecast and explain.

It is important to note that a model that may be appropriate for explanation may not be appropriate for forecasting. For example, we may develop a model that relates the current price of a share of Dell computer stock Y_t to the level of the Dow Jones Industrial Average (DJIA) X_t to understand (explain) how much of Dell's stock price is determined by general market forces. To do this, we may regress recent prices of Dell's stock $y_1, y_2, ..., y_{t-1}, y_t$ upon corresponding contemporaneous values of the DJIA $x_1, x_2, ..., x_{t-1}, x_t$. **Contemporaneous** data are data that occur at the same time: the predicting *x*'s occur at the same time as the corresponding predicted *y*'s. Such a model may provide an explanation of Dell's stock price. However, it could **not** be used for forecasting because in order to forecast a future price of Dell's stock, say Y_{t+1} (one time period in the future) with this model, we would need to plug in the contemporaneous future value of the DJIA, x_{t+1}, which would not be known when the forecast is made (at time *t*). Models that forecast cannot use predictors that have unknown values for the time period being forecast. However, a model could be developed that would relate recent prices of Dell computer stock $y_2, ..., y_{t-1}, y_t$ to lagged values of the DJIA $x_1, x_2, ..., x_{t-1}$. **Lagged** data are data that have occurred at one or more time periods before the corresponding contemporaneous time series: Each *x* in $x_1, x_2, ..., x_{t-1}$ occurs one time period before the corresponding *x* in $x_2, ..., x_{t-1}, x_t$. So $x_1, x_2, ..., x_{t-1}$ is the first lag of $x_2, ..., x_{t-1}, x_t$. A regression model using lagged DJIA as predictor could forecast Y_{t+1} because the needed plug-in value for lagged DJIA in time period *t*+1 is x_t, which is known at the time *t* when the forecast is made.[4] *The important take-away is that in a forecasting model, all of the values of the predictors must be knowable for the time period being forecast.*[5]

Statistical modeling is a three-step process:

1) Propose the model
2) Validate the model
3) Use the model

---

[4] Note that this model could not forecast two time periods in the future because the plug-in value of DJIA for time *t*+2 would be x_{t+1}, which would not be known at the time (*t*) when the forecast is made. However, a regression model that uses DJIA lagged two time periods as predictor variable would work.

[5] You might reasonably ask, "Could the unknown predictors be estimated?" Yes. That is sometimes done. However, it injects additional uncertainty into the forecast that must be accounted for. And it does not solve the ultimate problem, for the unknown predictor values must *themselves* be estimated with values of *their* predictors that are required to be known for the future time period.

When you propose a model, you nominate one of the many time series models (e.g., random sample, random walk, autoregression, etc. – or one that you have created especially for the situation) as a candidate for the data at hand. Your proposal will likely be the model whose specifications you think will most closely match the data. When you validate your proposal, you test the model's specifications against the data. If the specifications pass, then you go on to step 3. If the specifications fail, then you diagnose why and revise your proposal. In step 3, you make the forecast or provide the explanation *according to the rules of your validated model*. Different models have different rules for forecasting, and your forecasts can go badly astray if you forecast according to the rules of an incorrect model.

From the foregoing introductory remarks, I hope you see the importance that I attach to statistical modeling in the forecasting/explaining of time series. In this course, I will introduce you to a number of different time series models. I will carefully explain the specifications of each, show you how to validate the specifications, and illustrate how to make forecasts and explanations.

Some worldly wisdom: Students sometimes think that validating the model (step 2) means that we must prove that our model is correct. No! It is unlikely that we will ever find a model that fits its specifications perfectly. The famous statistician George Box, who introduced the Box-Jenkins (ARIMA) method of forecasting, said, "All models are wrong, but some are useful." We do not need a *perfect* model. We do need a model that is *good enough*. This generally means that the model is plausible, fits more or less well, with no discernible fatal flaws, and its forecasts seem reasonable, with a tolerable amount of expected error. Do not let perfection become the enemy of the good enough!

It is important to maintain a proper perspective on the forecasting process. A forecast is always based on past data. Forecasting relies on the premise that the past foreshadows the future. In a sense, the future must already be present in some form in the past – but hidden. If the future is completely unrelated to the past (and present), then we have no scientific basis upon which to make a forecast. What time series analysis does is to expose the way the future is hidden in the past. How does it do this? We have no future data. So we cannot regress future data on the past to learn what the relationship is. We have only present and past data. As a proxy for studying how the future relates to the past, we can study how the present relates to the past. Then we can pretend that the relationship between present and the past is the same as the relationship between future and the past. Forecasting assumes that the past relationship will continue into the future, so that the future will be related to the past in the same way that our analysis shows us that the present is related to the past. The future may be more or less related to the past. The more related to the past, the more confident we should be of our forecasts; the less related to the past, the less confident we should be of our forecasts. It is the job of time series analysis to lay bare what the past-present-future relationship is. One final point: A forecast is not complete with the provision of a numerical estimate of the unknown future quantity. The forecast should be accompanied by an assessment of its uncertainty. This can be done most easily by providing a confidence/prediction interval for the forecast. An example: Forecasting that sales next year will total $20 million, with 90% confidence that the actual number will lie between $19 million and $21 million.

After this brief introduction, our next stop will be the Random Sample – our first and most basic time series model, which underlies all other models.