Notes for Hyndman and Athanasopoulos – Chapter 2

* Frequency of a time-series is the number of observations before the seasonal pattern repeats.
  1. Yearly data: Frequency = 1
  2. Quarterly data: Frequency = 4
  3. Monthly data: Frequency = 12
  4. Weekly data: Frequency = 52
  5. Daily data might have two types of seasonality:
     + Weekly seasonality (frequency = 7)
     + Annual seasonality (frequency = 365)
* Time series patterns
  1. *Trend:* A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear.
  2. *Seasonal:* A seasonal pattern occurs when a time-series is affected by seasonal factors such as the time of the year or the day of the week. Seasonality is always fixed and of known frequency.
  3. *Cyclic:* A cycle occurs when the data exhibit rises and falls that are not of a fixed frequency
* If the fluctuations are not of a FIXED FREQUENCY, then they are cyclic; if the frequency is unchanging and associated with some aspect of the calendar then the pattern is seasonal.
* An accurate forecasting method must take into account the time patterns in the data and be able to capture the patterns properly.
* A seasonal plot is similar to a time plot except that the data are plotted against individual seasons in which the data was observed.
* Scatterplots are useful for visualizing relationship between time-series. It is common in these cases to display the correlation coefficient to measure the strength of the relationship between these two variables.
* Autocorrelation measures the linear relationship between lagged values of a time-series.
* The autocorrelation coefficients are plotted to show the autocorrelation function (AFC). The plot generated is known as the *correlogram*.
* When data has a trend, the autocorrelation for small lags tend to be large and positive because observations nearby in time are also nearby in size (due to changing mean of the series as time goes by).
* When data is seasonal, the autocorrelations will be larger for the seasonal lags (at multiples of the seasonal frequency) than for other lags.
* Time series that show no autocorrelation are called **white noise**.

Notes for Hyndman and Athanasopoulos – Chapter 3

**Simple forecasting methods**

* Average method for forecasting:
  1. The forecasts of all future values are equal to the average of the historical data.
* Naïve method of forecasting:
  1. Naïve forecast consists in setting all forecasts to be the value of the last observation. This is an optimal technique when data follows a random walk.
* Seasonal naïve method:
  1. The forecast is equal to the last observed period from the same season of the past year.
* Drift method:
  1. A variation of the naïve method is to allow for forecasts to increase or decrease over time, where the amount of change over time (**drift**) is set to be the average change seen in historical data. Equivalent to drawing a line between the first and last observations and extrapolating into the future.
* These simple forecasting methods usually serve as benchmark to analyze more complicated models.

**Transformation and adjustment**

* Adjusting the historical data can often lead to a simpler forecasting task. The purpose of these adjustments and transformations is to simply the patterns in historical data by removing known sources of variation.
* Calendar adjustments:
  1. A common source of variation derives from the fact that some months have more days than others.
* Population adjustments:
  1. Any data that is affected by population changes can be adjusted to give per-capita data. This removes the effects of total population changes.
* Inflation adjustments:
  1. Financial time-series are usually adjusted so that all values are stated in dollar values from a particular year.

**Mathematical transformations**

* Logarithmic transformations.
* Power transformations.
* Box-cox transformation: a family of transformations that combines natural logarithms and power transformations. For different values of lambda, the series will change in shape. A good value of lambda makes the seasonal variation about the same across the whole series.
* One issue with mathematical transformations is that the back-transformed forecast will not be the mean of the forecast distribution. Therefore, the result might need to be adjusted to ensure that the mean of the distribution will be recovered. This process is called bias-adjusting.

**Residual diagnostics**

* The residuals in a time series model are what is left over after fitting a model.
* For many models, the residuals are equal to the difference between the observations and the corresponding fitted values.
* Residual analysis is useful for model diagnostics (whether it has adequately captured the information in the data). There are two characteristics of residuals generated by good forecasting methods:
  1. Residuals are uncorrelated. If there is correlation between residuals, it suggests that there is information that could be used in the forecasts.
  2. The residuals have zero mean. Otherwise the forecasts are biased.
* There are two other characteristics that are useful, although not necessary:
  1. Residuals have constant variance.
  2. They are normally distributed.
* Residual autocorrelation can be visually checked using the correlogram. However, it is also advisable to run tests that take into account the possibility that multiple comparisons might give rise to false positives. The box-pierce test is one of such tests (another alternative is the Ljung-Box test). Both tests assume that there is no autocorrelation between values in the series.
* Forecasts should be assessed (in terms of quality) against new information. Therefore, their accuracy can only be truly determined by considering how well a model performs on new data that was not used when fitting the model.
* When choosing models, the usual practice is to separate the available models into two parts:
  1. Training: used to learn model parameters.
  2. Test: used to evaluate the model accuracy.
* The size of the test set is typically about 20% of the total sample. However, it is good to practice for it to be as large as the maximum forecast horizon required.
* A forecast error is the difference between an observed value and its forecast.
* Forecast errors are on the same scale as the data. Therefore, accuracy measures that rely on forecast errors are scale-dependent and cannot be used to make comparison between series that involve different units.
* There are two scale-dependent measures that are frequently used:
  1. Mean absolute error (MAE): equal to the average absolute value of forecast errors.
  2. Root mean squared error (RMSE): equal to the square root of the average squared forecast error.
* A forecast method that minimizes the MAE will lead to forecasts of the median, while forecast methods that minimize the RMSE will lead to forecasts of the mean.
* Percentage error is how much, on a percent scale, the forecast is “wrong”. The mean absolute percentage error is the average of those. This is scale-independent.
* Scaled errors is an alternative to using percentage errors when comparing forecast accuracy across series with different units. They propose scaling the errors based on the *training* MAE from a simple forecast method.
* Cross-validation is another procedure to determine forecasting accuracy: In this procedure, there are a series of test sets, each consisting of a single observation. The corresponding training set consists only of observations that occurred prior to the observation that forms the test set.
* Since we need a minimum number of observations to generate a minimally reliable forecast, the earliest observations are not considered test sets.
* The forecast accuracy is computed by averaging over the test sets.
* A good way to choose the best forecasting model is to find the model with the smallest RMSE computed using time series cross-validation.
* A prediction interval provides an interval within which we expect our forecast to lie with a specified probability.
* The value of prediction intervals is that they express uncertainty in the forecasts.
* One-step prediction intervals:
  1. When forecasting one step ahead, the standard deviation of the forecast distribution is almost the same as the standard deviation of the residuals.
* For multi-step prediction intervals, it is necessary to estimate (the standard deviation of the forecast. For one-step ahead, the standard deviation of the residuals is a relatively good estimate).

Notes for Hyndman and Athanasopoulos – Chapter 5

* The basic concept behind time series regression models is that we forecast the time series of interest *y* assuming that it has a linear relationship with other time series x.
* When there are two or more predictor variables, the model is called a **multiple regression model**. The coefficients measure the effect of each predictor after taking into account the effects of all other predictors in the model. Thus, the coefficients measure the *marginal effects* of the predictor variables.
* Assumptions behind the linear regression model:
  1. The model is a reasonable approximation of reality; that is, the relationship between the forecast variable and the predictor variable satisfies the linear equation.
  2. There are three auxiliary assumptions about the errors:
     + They have mean zero; otherwise the forecasts will be systematically biased
     + They are not autocorrelated; otherwise the forecasts will be inefficient, as there is more information in the data that can be exploited.
     + They are unrelated to the predictor variables; otherwise there would be more information that should be included in the systematic part of the model.
     + Fixed predictors.
* We choose the coefficients that minimize the sum of squared residuals.
* The t-statistic and the associated p-values for the coefficients are useful if we are interested in studying the effect of each predictor and making causal claims about it. For forecasting these numbers are not particularly useful.
* A common way to summarize how well a linear regression model fits the data is via the coefficient of determination or R-squared. It reflects the proportion of the variation in the forecast variable that is accounted for by the regression model.
* Another measure of how well the model has fitted the data is the standard deviation of the residuals (the residual standard error). This measure is related to the size of the average error that the model produces.
* The differences between the observed *y* values and the corresponding fitted values *y-hat* are the training-set errors or “residuals”. There are a series of plots that serve to perform diagnostics on the residuals:
  1. ACF plot of residuals to visually detect residual autocorrelation. If autocorrelation is present, then the forecasts are inefficient (there are forecasts that have lower variability).
  2. Histogram of residuals to detect normality.
  3. Residual plot against predictors. We expect the residuals to be randomly scattered without showing any systematic patterns. If the scatterplots show a pattern, then the relationship may be nonlinear, and the model will need to be modified accordingly.
  4. Residual plot against fitted values: a plot of the residuals against the fitted values should also show no pattern. If a pattern is observed, there may be heteroscedasticity (variance not constant).
* Observations that take extreme values compared to the majority of the data are called **outliers.** Observations that have a large influence on the estimated coefficients of a regression model are called **influential observations.**
* More often than not, time series data are non-stationary: that is, the values of the time series do not fluctuate around a constant mean or with a constant variance. Regressing non-stationary time-series can lead to spurious regressions. High R-squared and high residual autocorrelation can be signs of spurious regression.
* Useful predictors for time-series regression models:
  1. Trend: it is common for time-series data to be trending. A linear trend can be modeled by simply setting = t as a predictor.
  2. Dummy variables. Indicator variables that takes the value of 1 (“yes”) or 0 (“no”). The interpretation of the associated coefficient with the dummy variable is that it is a measure of the effect of that category relative to the omitted category.
  3. Intervention variables: it is often necessary to model interventions that may have affected the variable to be forecast.
     + When the effect lasts only for one period, we use a “spike” variable. This is a dummy variable that takes the value of one in the period of intervention and zero elsewhere.
     + When the level shifts, we use a “step variable”. A step variable takes the value of one after the intervention and zero before.
  4. Trading days in sales data. The number of trading days in each month can be included as a predictor.
  5. Distributed lags, such as the ones that measure the effect of advertising.
  6. Fourier series instead of seasonal dummy variables for long seasonal periods.
* Predictive accuracy is a way to determine predictor selection.
* Forecasters should not use R-squared to determine whether a model will give good predictions as it will lead to overfitting. Therefore, it will always choose the model with most variables.
* Adjusted r-squared is a way to select predictors and is equivalent to minimizing the standard error of the regression.
* Another method is performing leave-one-out cross validation and compute the mean squared error.
* Akaike’s information criterion is an estimator of the relative quality of statistical models for a given set of data. The idea is to penalize the fit of the model (SSE) with the number of parameters that need to be estimated. For small series, the AIC tends to select too many predictors, so a bias-corrected version is available.
* Where possible, all potential regression models should be fitted and the best model should be selected based on one of the measures discussed (this is known as “best subsets” regression).
* When using regression models for time series data, we need to distinguish between the different types of forecasts that can be produced, depending on what is assumed to be known when the forecasts are computed.
  1. Ex-ante forecasts: forecasts made using only the information that is available in advance. In order to generate ex-ante forecasts, the model requires forecasts of the predictors.
  2. Ex-post forecasts: forecasts that are made using later information on the predictors. The model from which ex-post forecasts are produced should not be estimated using the data from the forecast period. We assume prior knowledge of the predictor variables (the x variables), but should not assume knowledge of the data that are to be forecast.
* For models that rely on special predictors (seasonal dummies or public holiday indicators, there is no difference between ex-ante and ex-post forecasts, because they rely on predictors known in advance and that are based in calendar variables that repeat themselves.
* Scenario based forecasting: in this setting, the forecaster assumes possible scenarios for the predictor variables that are of interest. Prediction intervals do not include the uncertainty associated with the future distribution of the predictor variables. They assume that the values of the predictors are known in advance.
* The great advantage of regression models is that they can be used to capture important relationships between the forecast variable of interest and the predictor variables. However, ex-ante forecasting requires obtaining forecasts of the predictors and that can be challenging.
  1. An alternative formulation is to use as predictors their lagged values. The predictor set is formed by predictor values that are observed *h* time periods prior to observing *y*.
* The simplest way of modelling a nonlinear relationship is to transform the forecast variable y and/or the predictor variable x before estimating the regression model. While this provides a non-linear functional form, the model is still linear in the parameters.
* There are cases where simply transforming the data will not be adequate and a more general specification is required. (We allow f(x) to be a more flexible nonlinear function of x).
* One of the simplest specifications is to make *f* piecewise linear. We introduce points where the slope of *f*  can change.
* It is important not to confuse correlation with causation, or causation with forecasting. A variable *x* may be useful for forecasting a variable *y*, but that does not mean *x* is causing *y*.
* It is important to understand that correlations are useful for forecasting, even when there is no causal relationship between the two variables, or when the correlation runs in the opposite direction.
* However, often a better model is possible if a causal mechanism can be determined.
* We say that two variables are confounded when their effects on the forecast variable cannot be separated.
* Multicollinearity occurs when similar information is provided by two or more of the predictor variables in a multiple regression. A sign of multicollinearity is an extremely high correlation between a pair of predictors.

Notes for Hyndman and Athanasopoulos – Chapter 6

* Time series data can exhibit a variety of patterns, and it is often useful to split a time series into several components, each representing an underlying pattern category.
* When we decompose a time series into components, we usually combine the trend and the cycle into a single component (trend-cycle component).
* There are, thus, three components in a time series:
  1. A trend-cycle component
  2. A seasonal component
  3. A remainder component
* A time series can be decomposed into a sum of its components (additive decomposition) or as a product of its components (multiplicative decomposition).
* Additive decomposition is most appropriate if the seasonal fluctuations do not vary with the level of the series, whereas multiplicative decomposition is appropriate when the seasonal pattern is proportional to the level of the series.
* Intuitively, the trend-cycle component shows the overall movement in the series, ignoring the seasonality and any small random fluctuations.
* If the seasonal component is removed from the original data, the resulting values represent a data that has been “seasonally adjusted”.
* For additive decomposition, this is equivalent to , whereas for multiplicative data, this is equivalent to .
* Seasonally-adjusted data still contains the remainder component.
* Classical decomposition:
  + The first step is to use a moving-average to estimate the trend-cycle. It estimates the trend cycle at time *t* by averaging the values of the time series within *k periods of t*.
  + The average eliminates some of the randomness in the data, because nearby points are likely to have similar values. Therefore, the result is a smooth trend-cycle component.
  + Each value in a m-MA column is the average of the observations in a window of size m centered on the corresponding observation.
  + The order of the time-series (the m) determines the smoothness of the trend-cycle estimate.
  + Moving-average estimates are usually odd numbered to ensure that the series is symmetric (the missing values on each end of the series would be of the same magnitude).
  + It is, however, possible to make a moving average series of order m symmetric by centering it. Each value now represents a weighted average of the values.
  + If the seasonal period is even and of order *m*, we use a 2 x m MA to estimate the trend cycle.
    1. If the seasonal period is odd and of order m, we use a m-MA to estimate the trend cycle If the seasonal period is even and of order m, we use a 2 x m MA to estimate the trend cycle (all we are doing here is using moving averages to estimate the trend-cycle component, i.e., extracting the information from the data that is “independent” from seasonal variability or unmeasured variation).
  + A combination of moving averages results in weighted moving averages. A major advantage of weighted moving averages is that they yield a smoother estimate of the trend cycle.
  + In classical decomposition, we assume that the seasonal component is constant from year to year.
  + To perform an additive decomposition using the classical method, the procedures are:
    1. If m (frequency) is an even number, compute the trend-cycle component using a 2 x m-MA. If m is odd, compute the trend-cycle component using an m-MA.
    2. Calculate the detrended series (now your series is just seasonality + residual component).
    3. To estimate the seasonal component, average the detrended values for that season. For monthly data, the component for each month is the average of all values for that month. Seasonal components are adjusted that they add to zero.
    4. The remainder component is calculated by subtracting the estimated seasonal and trend-cycle components.
  + Multiplicative decomposition is similar with the exception that subtractions are replaced by divisions. Also, seasonal components are adjusted so that they add to m.
* Problems with classical decomposition:
  + While widely used, classical decomposition has some problems:

1. The estimate of the trend-cycle is unavailable for the first few and last few observations. As a result, the other components for the same observations cannot be estimated.
2. The trend-cycle estimate tends to over-smooth rapid rises and falls in the data (so the variation in the remainder component is not uniform).
3. The assumption of a regular seasonal component might not be adequate for some data.

* X11 decomposition is a method based on classical decomposition but includes extra steps and features in order to overcome the drawbacks of classical decomposition.
* SEATS decomposition is an alternative method for monthly and quarterly data.
* STL decomposition is a method based on LOESS estimation. It has several advantages over other methods:
  1. Can handle any type of seasonality, not only monthly and quarterly.
  2. The seasonal component is allowed to change over time, and the rate of change can be controlled by the user.
  3. Smoothness of the trend cycle can be controlled by the user.
  4. It is robust to outliers depending on the estimation.
* Though it only handles additive decomposition, it is possible to take logs and then back-transform the components.
* A time series decomposition can be used to measure the strength of trend and seasonality in a time-series. For strongly trended data, the seasonally adjusted data should have much more variation than the remainder component, so the quotient of the variation of the residuals to the variation of the sum of residuals and trend component.
* A similar concept applies to the strength of detrended data, but we replace the variation in the quotient with the variation of the seasonal data.
* Time series decomposition can also be used to forecast as we posit that future values can be decomposed into a seasonally adjusted and a seasonal component.
  1. It is usually assumed that the seasonal component is unchanging, so it is forecast by simply taking the last year of the estimated component.
  2. To forecast the seasonally adjusted component, any non-seasonal forecasting method may be used, such as a random walk with drift model, Holt’s method or a non-seasonal ARIMA.

Notes for Hyndman and Athanasopoulos – Chapter 7

* Forecasts produced using exponential smoothing are weighted averages of past observations, with the weights decaying exponentially as previous observations get older.
* Choosing between exponential smoothing depend on characteristics of the series being modeled and the way in which these enter the smoothing method.
* Simple exponential smoothing is a method that is suitable for data with no clear trend or seasonal pattern.
* Simple exponential smoothing can be thought as a compromise between naïve forecasting and average forecasting.
* Simple exponential smoothing is a technique that relies on one parameter, alpha. It can be modeled in two ways:
  1. As a weighted average between the most recent observation y\_t and the previous forecast.
  2. In a component form, comprised of a forecast equation and a smoothing equation.
     + The smoothing equation for the level gives estimated level of the series at each period t
* For every application of exponential smoothing method requires the smoothing parameters and the initial values to be chosen.
* The optimal value of alpha and the initial forecast can be determined through the minimization of the sum of squared errors, i.e., the forecast for *y* at time t, given data until *t-1*
* Simple exponential smoothing can be extended to incorporate data with a trend. It involves a forecast equation and two smoothing equations (one for the level, and one for the trend). The forecast function is now trending, depending on the estimate of the trend.
* The forecasts are a linear function of the last estimated value plus h times the last estimated trend value.
* Forecasts generated by exponential smoothing with a trend tend to over-forecast, especially for longer forecast horizons. Therefore, analysts have introduced damped methods, which “dampens” the trend to a flat line sometime in the future. In conjunction with smoothing parameters, the method also includes a damping parameter phi. Forecasts converge to a finite value; therefore, short-run estimations are trended while long-run forecasts are constant.
* A large value of alpha suggests that the level reacts strongly to each new observation, because the forecast is vastly dependent on the immediate past value. A value of phi close to one suggests that damping is small. A small value of the smoothing parameter for the slope indicates that the trend is not changing over time, because past estimates are receiving large coefficients in our estimates.
* Different accuracy measures will hardly indicate that one method dominates all the others.
* The Holt-Winters seasonal method comprises the forecast equation and three smoothing equations - one for level, one for the trend, and one for the seasonal component.
* There are two variations of the method that differ in the nature of the seasonal component:
  1. The additive method is preferred when the seasonal variations are roughly constant through the series. The seasonal component is expressed in absolute terms in the scale of the observed series.
  2. The multiplicative method is preferred when the seasonal variations are changing proportional to the level of the series. The seasonal component is expressed in relative terms, and the series is seasonally adjusted by dividing through each component.
* For the additive method, the component form has:
  1. A level equation that is a weighted average equal to the seasonally-adjusted observation and the non-seasonal forecast for time *t*.
  2. The trend equation is a weighted average of the difference between current and past observation and the estimate of the previous trend level.
  3. The seasonal equation is a weighted average between the current seasonal index, and the seasonal index m periods ago.
* Damping is possible with both additive and multiplicative Holt-Winter’s method. Damping is a way to smooth the trending parameter, so it does not overforecast observations in the far future.
* By considering variations in the combinations of the trend and seasonal components, nine exponential smoothing methods are possible. For instance:
  1. A method with no trend and no seasonality is the simple exponential smoothing.
  2. A method with an additive trend and, but no seasonal component is the holt’s linear method.
* The exponential smoothing methods are algorithms which generate point forecasts.
* A statistical model is a stochastic (or random) data generating process that can produce an entire forecast distribution.
* Each model consists of a measurement equation that describes the observed data, and some state equations that describe how the unobserved components or states (level, trend, seasonal) change over time.
* For each method there exist two models: one with additive errors and one with multiplicative errors. They will generate different prediction intervals, though.
* For a model with additive errors, we assume that the residuals are normally distributed white noise with mean 0 and variance sigma-squared.
* The measurement equation shows the relationship between the observations and the unobserved states.
* The state equation shows the evolution of the state through time. Alpha is a smoothing parameter that determines how much we should weight new information.
* An alternative to estimating the parameters by minimizing the sum of squared errors is to maximize the “likelihood”.
  + - The likelihood is the probability of the data arising from a specified model. The likelihood interprets the data as a function of the parameters.
    - For an additive error model, maximizing the likelihood gives the same results as minimizing the sum of squared errors.
* A great advantage of the ETS statistical framework is that information criterion measures can be used for model selection.
* Models with multiplicative errors are useful when the data are strictly positive but are not numerically stable when the data contain zeros or negative values.
* ETS will, most likely, return point estimates that are different than traditional Holt-Winter’s method, because parameters are estimated using ML and not minimum sum of squares.
* ETS models can also generative prediction intervals.