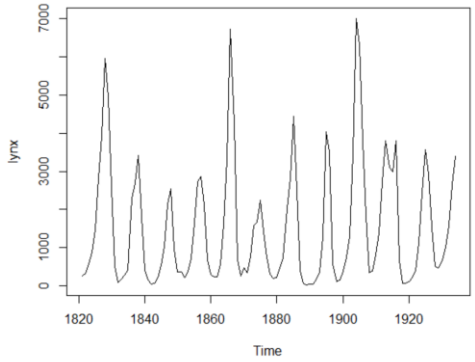
Lecture 17

Go over Exam #2

**Time series** modeling is a statistical method used to analyze data that varies over time. The goal of time series modeling is to identify patterns, trends, and other features in the data that can be used to make predictions about future values.




The basic steps in time series modeling are:

*Data exploration*: The first step in time series modeling is to explore the data and identify any patterns or trends that may be present. This can be done using techniques such as time plots, autocorrelation plots, and spectral analysis.

*Data transformation*: If the data exhibits non-stationarity or other anomalies, it may be necessary to transform the data using techniques such as differencing or logarithmic transformations to make it more amenable to modeling.

*Model selection*: Once the data has been explored and transformed, the next step is to select an appropriate time series model. The most commonly used time series models are the autoregressive integrated moving average (ARIMA) model, the seasonal ARIMA (SARIMA) model, and the exponential smoothing (ETS) model.

*Parameter estimation*: The next step is to estimate the parameters of the selected model using maximum likelihood estimation or other optimization techniques.

*Model validation*: Once the model has been fitted to the data, it is important to validate its performance using techniques such as residual analysis and out-of-sample forecasting.

*Model refinement*: If the model performance is not satisfactory, it may be necessary to refine the model by adjusting the model structure or incorporating additional variables.

Some common applications of time series modeling include forecasting economic indicators, predicting weather patterns, and analyzing stock prices. Time series modeling is a powerful tool for understanding and predicting complex temporal patterns in data, and it has many practical applications in fields such as finance, economics, and engineering.

There are several forecasting models used in time series analysis, each with its own strengths and weaknesses. Here are some common models:

**Autoregressive Integrated Moving Average (ARIMA)**: ARIMA is a popular time series model that can capture both the trend and the seasonality in the data. It is a flexible model that can handle both stationary and non-stationary time series data. ARIMA models are commonly used in economics, finance, and engineering applications.

**Seasonal Autoregressive Integrated Moving Average (SARIMA)**: SARIMA is an extension of the ARIMA model that is designed to handle time series data with seasonal patterns. It is a powerful model for forecasting seasonal time series data, such as quarterly or monthly economic indicators.

**Exponential Smoothing (ETS)**: ETS is a family of models that use weighted averages of past observations to forecast future values. ETS models can handle both trend and seasonality in the data, and are useful for forecasting short-term time series data.

**Vector Autoregression (VAR)**: VAR is a model used for multivariate time series data. It can capture the dependencies between multiple time series variables and can be used for forecasting and causal analysis.

**Neural Networks**: Neural networks are a family of machine learning models that can be used for time series forecasting. They are particularly useful for handling non-linear relationships between variables and can be trained to identify complex patterns in the data.

**Prophet**: Prophet is a time series forecasting model developed by Facebook that uses a combination of seasonal decomposition, trend modeling, and machine learning techniques to forecast time series data. It is designed to be easy to use and can handle multiple seasonality patterns.

These models can be combined and customized to fit specific time series forecasting needs. The selection of a specific model depends on the nature of the data and the forecasting objective.

Extrapolation is a common method used in time series analysis to make predictions about future values based on historical data. Extrapolation involves extending a time series beyond its observed range to make predictions about future values. As with regression, extrapolation far outside the range of the original data is fraught. Some common extrapolation methods used in time series analysis are:

*Trend extrapolation*: Trend extrapolation involves extending the trend of the time series into the future. This method assumes that the underlying trend in the data will continue into the future. Trend extrapolation can be done using linear regression or non-linear regression techniques.

*Seasonal extrapolation*: Seasonal extrapolation involves extending the seasonal pattern of the time series into the future. This method assumes that the seasonal pattern in the data will continue into the future. Seasonal extrapolation can be done using seasonal decomposition or seasonal ARIMA models.

*Holt-Winters extrapolation*: Holt-Winters extrapolation is a time series forecasting method that combines trend and seasonality to make predictions about future values. It uses exponential smoothing techniques to capture the trend and seasonal patterns in the data.

*ARIMA extrapolation*: ARIMA models can be used for extrapolation by extending the forecast beyond the end of the data. This involves using the estimated parameters of the model to make predictions about future values.

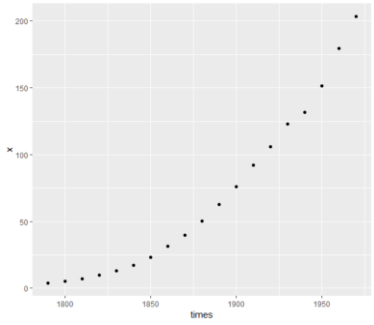
*Neural network extrapolation*: Neural networks can be used for extrapolation by training the model on historical data and using it to make predictions about future values. Neural networks are particularly useful for handling non-linear relationships and complex patterns in the data.

It is important to note that extrapolation carries some risks. Extrapolating too far into the future can result in predictions that are inaccurate or unreliable. It is important to validate the accuracy of extrapolation methods using statistical measures and to consider the limitations and assumptions of the method being used. It is also important to update the model as new information becomes available.

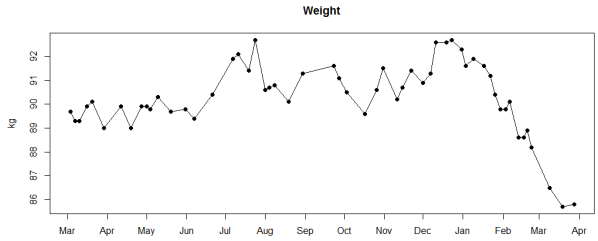
**Regular vs. Irregular Time Series**

The difference between regular and irregular time series is based on the frequency of observations in the time series data.

A regular time series is one in which the observations are equally spaced in time. For example, daily stock prices or monthly sales figures are regular time series data. In a regular time series, the time intervals between consecutive observations are constant and known.




On the other hand, an irregular time series is one in which the observations are not equally spaced in time. For example, a company's revenue might be reported quarterly or yearly, which means that the time intervals between observations are not constant. In an irregular time series, the time intervals between observations can vary and may not be known or consistent.




The distinction between regular and irregular time series is important because it can affect the type of time series analysis that is appropriate for the data. For example, regular time series data can be analyzed using standard techniques such as moving averages or autoregressive models, while irregular

time series data may require specialized techniques such as interpolation or resampling to estimate missing values.

It is also important to note that the frequency of observations can affect the interpretation of the data. For example, if a time series has a strong seasonal pattern, it may be difficult to detect this pattern if the observations are too widely spaced in time. Therefore, it is important to carefully consider the frequency of observations when analyzing time series data.

There are several methods for analyzing and modeling **regular time series** data. Here are some of the most common techniques:

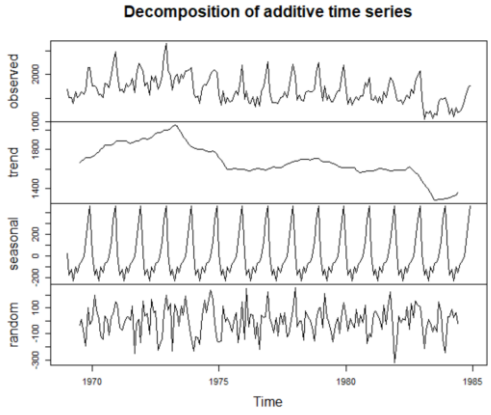
*Moving averages*: Moving averages are used to smooth out fluctuations in the time series data by taking an average of the data over a sliding window. Moving averages are useful for identifying trends and seasonal patterns in the data.

*Exponential smoothing*: Exponential smoothing is a technique that uses a weighted average of past observations to make predictions about future values. This method is particularly useful for time series data that has a trend or seasonal pattern.

*Autoregressive models*: Autoregressive (AR) models are used to model the relationship between past observations and future values. In an AR model, the future values of the time series are modeled as a linear combination of past observations.

*Moving average models*: Moving average (MA) models are used to model the relationship between the current value of the time series and past forecast errors. In an MA model, the future values of the time series are modeled as a linear combination of past errors.
*ARIMA models*: Autoregressive Integrated Moving Average (ARIMA) models are a combination of AR and MA models. ARIMA models are used to model time series data that has a trend and/or seasonal pattern.

*Seasonal decomposition*: Seasonal decomposition is a technique that separates the time series data into its seasonal, trend, and residual components. This method is useful for identifying the seasonal pattern in the data.




*Fourier analysis*: Fourier analysis is a mathematical technique used to decompose a time series into its frequency components. Fourier analysis is useful for identifying periodic patterns in the data.

These techniques can be used alone or in combination to analyze and model regular time series data. It is important to choose the appropriate technique based on the characteristics of the data and the research question at hand.

Analyzing and modeling **irregular time series** data can be challenging due to the uneven spacing of observations. Most modeling methods for time series are explicitly for regular time series. We will look at methods for irregular time series in later lectures in more detail, but the field is quite underdeveloped. However, there are several methods that can be used to analyze and model irregular time series data:

*Interpolation*: Interpolation is a method for estimating missing values in a time series by filling in the gaps between observations. There are several interpolation techniques, such as linear interpolation, spline interpolation, and kriging (gaussian process regression).

*Resampling*: Resampling is a technique that involves changing the frequency of the time series data to a regular interval. This can be done by either upsampling (increasing the frequency of the data) or downsampling (decreasing the frequency of the data). Resampling can help to make irregular time series data more amenable to standard time series analysis techniques.

*Kalman filtering*: Kalman filtering is a state-space model that is used to estimate the underlying signal in a time series by combining noisy measurements with a dynamic model of the system. Kalman filtering is particularly useful for analyzing irregular time series data with missing values.

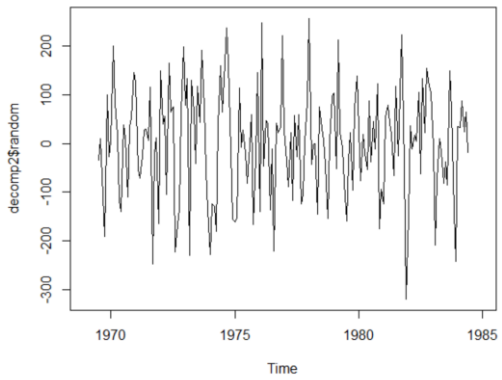
*Local regression methods*: Local regression methods, such as LOESS (locally weighted scatterplot smoothing), are useful for modeling irregular time series data by fitting a smoothed curve to the data. These methods can help to identify trends and seasonal patterns in the data.

*Wavelet analysis*: Wavelet analysis is a mathematical technique used to analyze irregular time series data by decomposing the signal into its frequency components. Wavelet analysis is useful for identifying both short-term and long-term patterns in the data.

These methods can be used alone or in combination to analyze and model irregular time series data. It is important to choose the appropriate method based on the characteristics of the data and the research question at hand.

Stationarity and trends are important concepts in (regular) time series analysis, and they describe different characteristics of the data.

**Stationarity** refers to the property of a time series where the statistical properties of the data, such as the mean, variance, and covariance, are constant over time. In other words, the distribution of the data does not change with time. Stationarity is an important assumption for many time series models, and violating this assumption can lead to unreliable forecasts and model results.




A time series that exhibits a trend is one where the mean value of the data changes over time. This can be a linear trend, where the mean value changes at a constant rate, or a non-linear trend, where the rate of change varies over time. Trends are common in many time series data, and can be caused by a variety of factors, such as economic cycles or technological change.

In order to model and analyze time series data, it is important to understand whether the data is stationary or exhibits a trend. If the data is stationary, standard time series models can be used to make predictions about future values of the data. If the data exhibits a trend, it may be necessary to remove the trend before modeling the data. This can be done using techniques such as detrending, differencing, or seasonal adjustment.

It is important to note that a time series can exhibit both stationarity and a trend. In this case, the trend component can be removed from the data to make it stationary, allowing standard time series models to be applied to the stationary residuals.

**Autocorrelation** is an important concept in time series analysis because it can help us understand the underlying structure and patterns in the data. Autocorrelation measures the degree of similarity between a time series and a lagged version of itself, and it is a common feature of many time series data.

There are several reasons why autocorrelation is important in time series analysis:

*Identifying trends and seasonality*: Autocorrelation can be used to identify the presence of trends and seasonality in a time series. If a time series exhibits high autocorrelation at specific lags, it suggests that there is some structure or pattern in the data that is repeating over time.

*Model selection*: Autocorrelation can be used to guide model selection in time series analysis. Many time series models, such as ARIMA and SARIMA, rely on the assumption of stationarity and use autocorrelation functions to identify the appropriate order of differencing and lagged terms to include in the model.

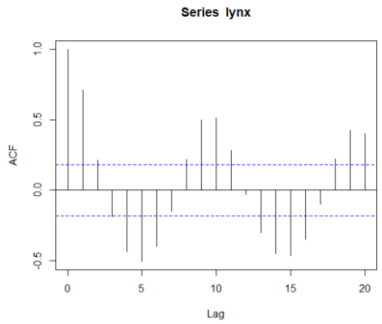
*Forecasting*: Autocorrelation can be used to make forecasts in time series analysis. By modeling the autocorrelation structure of the data, we can make predictions about future values of the time series.

*Checking model assumptions*: Autocorrelation can be used to check the assumptions of time series models. If the residuals of a time series model exhibit significant autocorrelation, it suggests that the model is not adequately capturing the underlying structure of the data.

Testing for randomness in time series data is an important step in time series analysis, as it helps to determine whether the data exhibits any significant patterns or structure that can be modeled or predicted. There are several methods for testing for randomness in time series data, including:

*Visual inspection*: One of the simplest ways to test for randomness in time series data is to visually inspect the data. If the data appears to be random and without any discernible patterns or trends, then it is likely to be random. However, visual inspection alone may not be enough to identify subtle patterns or trends in the data.

*Autocorrelation function (ACF)*: The ACF measures the degree of similarity between a time series and a lagged version of itself. If the ACF shows no significant correlation at any lag, then the data may be considered random. However, it is important to note that some time series may exhibit significant autocorrelation at certain lags even if they are considered random.




*Runs test*: The runs test is a statistical test that checks for randomness in a time series by counting the number of runs, or sequences of consecutive increasing or decreasing values, in the data. If the number of runs is within the expected range for a random process, then the data may be considered random.

*Ljung-Box test*: The Ljung-Box test is a statistical test that checks for the presence of autocorrelation in a time series by comparing the observed autocorrelations to the expected values under the null hypothesis of randomness. If the test statistic is not significant, then the data may be considered random.

*Spectral analysis*: Spectral analysis is a method for analyzing the frequency content of a time series. If the spectral density of the data is flat or white, then it suggests that the data is random.

It is important to note that testing for randomness in time series data is not always straightforward, and different methods may yield different results. In addition, some time series may exhibit complex patterns or structure that cannot be easily identified or modeled using standard tests for randomness.

**First and second differences** are commonly used in time series analysis to transform non-stationary time series data into stationary data, which is a common assumption for many time series models.

Stationary data has constant mean and variance over time, and its statistical properties do not change with time.

**First differences** are obtained by subtracting each observation from its preceding observation. This operation is equivalent to calculating the rate of change or the first derivative of the time series. The first difference can help remove the trend component from the time series data, resulting in a stationary series.

**Second differences** are obtained by taking the first difference of the first difference. This operation is equivalent to calculating the second derivative of the time series. Second differences can help remove not only the trend component but also the seasonality component from the time series data, resulting in a stationary series.

You can continue this process, but it is uncommon.

Once a stationary series is obtained by taking first or second differences, various time series models can be applied to analyze and forecast the data. The most common models used for stationary time series data are the Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA (SARIMA) models. These models use the order of differencing required to make the time series stationary as an input parameter, and the model estimates the parameters for the autoregressive and moving average components of the time series.

Autocorrelation is a measure of the similarity between a time series and a lagged version of itself. Autocorrelation is important in time series analysis because it can indicate whether the time series exhibits a pattern or structure that can be modeled or predicted.

In autocorrelation, a correlation coefficient is calculated between the time series and its lagged values. The correlation coefficient ranges from -1 to 1, with values of 0 indicating no correlation, positive values indicating positive correlation, and negative values indicating negative correlation.

If the autocorrelation coefficient is close to 1 or -1, it suggests a strong correlation between the time series and its lagged values, indicating that the time series may exhibit a predictable pattern or trend. If the autocorrelation coefficient is close to 0, it suggests no correlation between the time series and its lagged values, indicating that the time series may be random or unpredictable.

The **autocorrelation function (ACF)** is a commonly used tool for visualizing and analyzing autocorrelation in time series data. The ACF plots the autocorrelation coefficient as a function of the lag, where the lag represents the number of time periods between the original time series and its lagged version. The ACF is a useful tool for identifying the order of autoregressive (AR) and moving average (MA) terms in the Autoregressive Integrated Moving Average (ARIMA) model, which is a popular model used for time series forecasting. A partial autocorrelation function can also be calculated that takes into account other correlations.

The accuracy of time series models is typically measured by comparing the model's predicted values with the actual observed values of the time series. The most common metrics for measuring the accuracy of time series models are listed below, and some are similar to metrics we use in regression analysis.

**Mean Absolute Error (MAE)**: This is the average of the absolute differences between the predicted and actual values. MAE is a simple and easy-to-understand metric, but it does not give more weight to larger errors.

**Root Mean Squared Error (RMSE)**: This is the square root of the average of the squared differences between the predicted and actual values. RMSE gives more weight to larger errors, which can be useful in situations where larger errors are more important to avoid.

**Mean Absolute Percentage Error (MAPE)**: This is the average of the absolute percentage differences between the predicted and actual values. MAPE is useful when the magnitude of the error is important relative to the size of the actual value.

**Symmetric Mean Absolute Percentage Error (SMAPE)**: This is similar to MAPE, but uses the average of the absolute differences between the predicted and actual values as the denominator instead of the actual values. SMAPE is useful when the time series has a wide range of values.

In addition to these metrics, forecast accuracy can also be evaluated using visual methods, such as plotting the actual values and the predicted values on the same graph to visually compare their performance over time.

It is important to note that no single metric can provide a complete picture of the accuracy of a time series model, and it is recommended to use multiple metrics and visual methods to evaluate the performance of the model.

Resources:


1. https://www.investopedia.com/terms/t/timeseries.asp
2. https://www.tableau.com/learn/articles/time-series-analysis
3. https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm
4. https://www.simplilearn.com/tutorials/data-science-tutorial/time-series-forecasting-in-r
5. https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/src/timeseries.html
6. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1179&context=marketing_papers
7. https://rc2e.com/timeseriesanalysis
8. https://www.ibm.com/docs/en/streams/5.3?topic=series-regular-irregular-time
9. https://www.kaggle.com/general/12788
10. https://arxiv.org/abs/2004.08284
11. https://quantitative.cz/wp-content/uploads/2018/09/methods-for-periodic-and-irregular-tomas-hanzak-2014.pdf
12. https://otexts.com/fpp2/decomposition.html
13. https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/
14. https://towardsdatascience.com/time-series-from-scratch-decomposing-time-series-data-7b7ad0c30fe7
15. https://www.influxdata.com/blog/autocorrelation-in-time-series-data/
16. https://www.statistics.com/glossary/differencing-of-time-series/
17. https://www.oreilly.com/library/view/practical-time-series-analysis/9781788290227/7e3122e0-aa26-4d63-b289-67b7f37b42d2.xhtml
18. https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27